TSA Data Science and Analytics

Washington DC, MD
June, 2026
Team Member IDs:
1098102 (Captain), 1233511

Table of Contents

Introduction

The COVID-19 pandemic produced one of the most significant global economic disruptions in modern history, fundamentally altering consumer behavior, inflation dynamics, and international mobility. While the immediate impacts on global tourism were clear – border closures, travel restrictions, and sharp declines in visitor numbers – the underlying drivers of tourism recovery in the post-pandemic period remain less understood.

As countries reopened and international travel resumed, tourists encountered a global landscape in which relative prices, currencies, and cost-of-living conditions had shifted unevenly across destinations. These changes raised two important questions: which economic factors most strongly influenced tourism recovery following the pandemic, and how did travelers respond to a global environment increasingly shaped by affordability considerations?

Tourism is particularly sensitive to economic conditions. Exchange rates, inflation, and purchasing power influence both destination choice and on-site spending behavior. Traditionally, it is assumed that tourists favor destinations that are cheaper relative to alternatives. However, the pandemic introduced an additional layer of complexity by generating abrupt and uneven economic shocks across countries rather than gradual price changes. As a result, travelers were responding not only to how expensive a destination was, but also to how much its affordability had changed relative to pre-pandemic expectations.

This study examines the relationship between affordability and tourism recovery using international tourism data from 2019 to 2022. Tourism arrivals and tourism receipts are used as measures of tourism activity, while Purchasing Power Parity (PPP) is used as a standardized indicator of relative cost-of-living. The analysis evaluates two related dimensions of affordability: whether absolute affordability levels in 2022 influenced post-pandemic tourism recovery, and whether changes in affordability between 2019 and 2022 played a more significant role in shaping tourism outcomes. By separating affordability into level and change, this study assesses whether tourists responded more strongly to static price differences or to economic disruption and shifting cost-of-living conditions.

Through data preparation, visualization, and predictive modeling, this analysis provides insight into how large-scale economic shocks influence tourism behavior. Understanding these relationships is essential not only for explaining post-pandemic travel patterns, but also for anticipating how future global disruptions may affect international tourism demand. This study addresses three core research questions:

1. Does absolute destination affordability in 2022 explain differences in post-pandemic tourism recovery?
2. Do changes in affordability between 2019 and 2022 better explain tourism recovery outcomes than static price levels?
3. When controlling for tourism volume, does affordability change independently influence tourism spending recovery?

Data Dictionary

Field Name	Data Type	Description	Field Example
Country_std	String	Standardized country name used consistently across all datasets after text normalization and ISO-based matching. This variable serves as the primary key for merging tourism and affordability data at the country level.	United States
Region	String	Geographic region associated with each country, used for organizational grouping and regional comparison in exploratory analysis and visualization.	Americas
Arrivals_2019	Integer	Number of international tourist arrivals in 2019, measured in thousands. Serves as the pre-pandemic baseline level of tourism volume against which post-pandemic recovery is evaluated.	79442
Arrivals_2022	Integer	Number of international tourist arrivals in 2022, measured in thousands. Represents the level of international tourism volume observed during the post-pandemic period and reflects destination demand at a fixed point in time.	50870
Receipts_2019	Integer	International tourism receipts in 2019, measured in millions of U.S. dollars. Represents baseline pre-pandemic tourism spending and provides a reference point for evaluating changes in tourism revenue.	198980
Receipts_2022	Integer	International tourism receipts recorded in 2022, measured in millions of U.S. dollars. Represents total tourism-related spending in the first full year of widespread post-pandemic international travel.	135215
PPP_2019	Float	Purchasing Power Parity (PPP) value for 2019, representing relative price levels compared to the United States prior to the COVID-19 pandemic. Lower values indicate lower relative prices and higher affordability.	.72

PPP_2022	Float	Purchasing Power Parity (PPP) value for 2022, representing relative price levels compared to the United States during the post-pandemic recovery period. Used as a measure of absolute destination cost levels following pandemic-related economic shifts.	.84
Arrivals_Change_Pct	Float	Percentage change in international tourist arrivals between 2019 and 2022. Used as a standardized measure of tourism recovery that normalizes differences in country size and pre-pandemic tourism scale.	-0.36
Receipts_Change_Pct	Float	Percentage change in international tourism receipts between 2019 and 2022. Used to quantify changes in tourism spending recovery and capture shifts in aggregate revenue independent of baseline spending levels.	-0.32
PPP_Change_Pct	Float	Percentage change in Purchasing Power Parity between 2019 and 2022, used as a proxy for relative inflationary pressure and price-level shifts compared to the United States. In this study, this variable captures the magnitude of pandemic-era economic disruption rather than absolute destination affordability.	0.17

Purpose

The primary objective of this analysis is to investigate how affordability influenced international tourism recovery following the COVID-19 pandemic. Specifically, this study examines the relationship between purchasing power parity (PPP) and tourism outcomes, focusing on whether tourism recovery was more strongly driven by a country's absolute affordability level in the post-pandemic period or by changes in affordability that occurred due to pandemic-related economic disruption. By comparing international tourism arrivals and tourism receipts between 2019 and 2022, this project seeks to determine how global travelers responded to shifting economic conditions during the recovery phase of international travel.

The COVID-19 pandemic introduced unprecedented inflationary pressures, currency fluctuations, and cost-of-living changes across countries, fundamentally altering the relative affordability of travel destinations. As international borders reopened, travelers were faced with a global tourism market in which prices and economic conditions differed significantly from pre-pandemic norms. This study explores whether tourists were primarily drawn to destinations that were objectively cheaper after the pandemic, or whether they were more responsive to relative economic changes, such as sudden increases or decreases in affordability compared to pre-pandemic expectations. Distinguishing between these two mechanisms allows for a deeper understanding of how economic shocks influence consumer decision-making in the tourism sector.

The findings of this research are relevant to multiple stakeholders within the global tourism ecosystem. Policymakers and tourism boards may benefit from insights into how inflation and economic volatility affect tourism demand, allowing for more informed recovery and pricing strategies. Tourism-dependent economies can use these findings to better anticipate how future economic disruptions may influence visitor behavior and spending patterns. Additionally, travel industry analysts may gain a clearer understanding of the factors that drive tourism resilience and recovery following large-scale global shocks.

Furthermore, this study contributes to a broader understanding of how sudden economic changes, rather than static economic conditions, shape international consumer behavior. By demonstrating the differing impacts of absolute affordability and affordability change, this research highlights the importance of economic context and expectations in post-pandemic tourism decision-making. These findings are particularly relevant as global economies continue to face inflationary pressures and uncertainty, emphasizing the need to understand how travelers adapt their behavior in response to evolving economic environments.

Methods

Data Sources and Collection

This study utilized country-level international tourism and economic data from two primary open-source sources. International tourism arrivals and tourism receipts data for 2019 and 2022 were obtained from the United Nations World Tourism Organization (UNWTO). Tourism arrivals were reported in thousands of visitors, while tourism receipts were reported in millions of U.S. dollars, allowing analysis of both tourism volume and spending behavior. Affordability data was sourced from the World Bank's Purchasing Power Parity (PPP) dataset, which provides a standardized measure of relative cost-of-living across countries. The years 2019 and 2022 were selected to represent pre-pandemic conditions and the first full year of widespread post-pandemic tourism recovery, respectively.

Data Cleansing and Standardization

Data processing was conducted in Python using the Pandas library. Because tourism and PPP datasets used inconsistent country naming conventions, a multi-stage standardization process was implemented. Text normalization was performed using unicodedata and re to remove accents, punctuation, and formatting inconsistencies, followed by ISO-based country alignment using the pycountry library. Remaining edge cases were resolved using a manually defined override dictionary, and a mismatch log was generated to verify alignment accuracy. After standardization, datasets were merged at the country level, rows containing missing values in key variables were removed, and non-essential columns were excluded.

Data Transformation and Feature Engineering

To enable meaningful cross-country comparison, percentage change variables were calculated for tourism arrivals, tourism receipts, and purchasing power parity between 2019 and 2022. Percentage-based metrics were used to normalize differences in country size and baseline tourism levels. Data were reshaped into a wide-format structure so that each country was represented by a single observation containing both time-period values and derived change metrics. This structure supported direct comparison and analysis.

Exploratory Analysis and Visualization

Exploratory analysis was conducted using Matplotlib to visualize relationships between affordability and tourism recovery. Scatter plots were generated to examine associations between purchasing power parity and percentage changes in tourism arrivals and tourism receipts. Two affordability frameworks were evaluated independently: absolute affordability levels in 2022 and relative affordability change between 2019 and 2022. Linear regression trendlines were added to each plot, and coefficients of determination (R^2 values) were calculated to assess explanatory strength. Distributional characteristics of key variables were examined using box plots to assess symmetry and variability.

Predictive Modeling

A multivariable linear regression model was constructed using Scikit-learn to quantify the combined influence of affordability change and tourism volume on tourism spending recovery. Percentage change in tourism receipts served as the dependent variable, while percentage change in PPP and tourism arrivals was used as independent variables. Model performance was evaluated using R^2 values and residual analysis. Variables that demonstrated no meaningful explanatory power during exploratory analysis, such as absolute affordability levels, were excluded to maintain model interpretability and avoid overfitting.

Results

Overview

The results of this study are presented in a sequential manner designed to build from foundational data validation toward increasingly complex analytical insights. The analysis begins by examining the distributional characteristics of key variables to ensure statistical stability prior to inference. It then explores the relationship between absolute affordability and tourism recovery to assess whether post-pandemic travel behavior reflected price sensitivity. Next, the focus shifts to relative changes in affordability between 2019 and 2022, revealing how inflationary shifts interacted with tourism recovery dynamics. Finally, a multivariable regression model integrates these findings to quantify the combined influence of tourism volume and affordability change on tourism spending recovery. Together, these results support a cohesive narrative suggesting that post-pandemic tourism recovery was more strongly associated with demand-side forces than with affordability constraints.

Distribution of Key Variables

Before examining the relationship between affordability and tourism recovery, the distributions of key variables were evaluated to assess stability, symmetry, and the presence of outliers. Understanding the underlying distribution of the data is essential prior to regression-based analysis, as heavily skewed or unstable variables can distort correlation results and lead to misleading conclusions. The distributions of percentage change in purchasing power parity, tourism arrivals, and tourism receipts were examined.

Figure 1: Distribution of Percentage Change Variables (2019–2022)

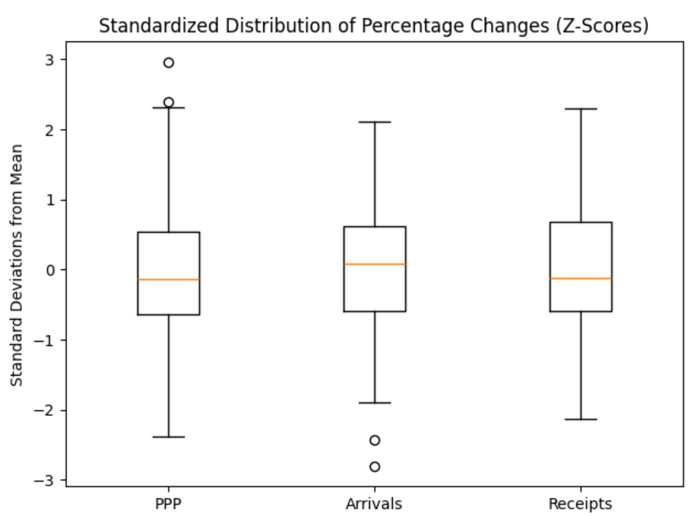


Figure 1: This stacked box plot displays the distributions of percentage change in Purchasing Power Parity (PPP), international tourist arrivals, and international tourism receipts between 2019 and 2022. The visualization is used to assess symmetry, spread, and the presence of extreme outliers prior to regression-based analysis.

Observed Relationship: The distributions of all three percentage change variables are approximately symmetric, with moderate spread and no evidence of extreme outliers dominating the data. While some countries experienced substantial positive or negative changes, the central tendencies of each variable remain well-defined, and no variable exhibits excessive skewness.

Interpretation: The relatively symmetric distributions suggest that the variables are well-suited for regression-based analysis and cross-country comparison. The absence of severe skewness reduces the likelihood that observed relationships are driven by a small number of extreme cases. Establishing this baseline distributional behavior is important for validating subsequent analyses that examine relationships between affordability and tourism recovery outcomes.

Absolute Affordability and Post-Pandemic Tourism Recovery

With the stability and distributional characteristics of the data established, the analysis next evaluates whether absolute affordability levels influenced tourism recovery following the COVID-19 pandemic. Absolute affordability, measured by using Purchasing Power Parity (PPP), represents how expensive or inexpensive a country is relative to others at a fixed point in time. Given the economic disruption caused by COVID-19, including rising inflation and increased cost sensitivity among consumers, it is reasonable to hypothesize that travelers may have favored destinations with lower baseline costs of living. This subsection tests that hypothesis by examining whether countries that were more affordable in 2022 experienced stronger recoveries in international tourist arrivals and tourism receipts.

Figure 2: Purchasing Power Parity Level (2022) vs.
Percentage Change in International Tourist Arrivals (2019–2022)

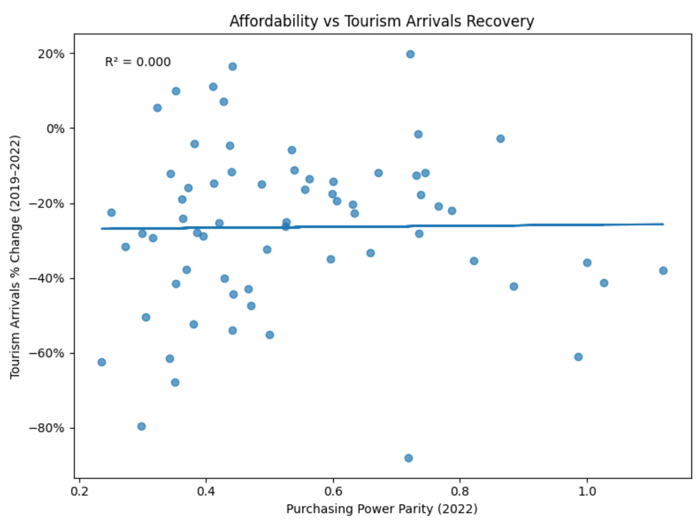


Figure 2 illustrates the relationship between absolute affordability levels in 2022, measured using Purchasing Power Parity (PPP), and the percentage change in international tourist arrivals between 2019 and 2022. A linear regression trendline and coefficient of determination (R^2) are included to evaluate the strength of this relationship.

Observed Relationship: The scatter plot comparing Purchasing Power Parity (PPP) levels in 2022 with the percentage change in international tourist arrivals between 2019 and 2022 reveals no meaningful linear relationship. Data points are widely dispersed across the plot, with no clear upward or downward trend. The fitted regression line is nearly flat, and the resulting coefficient of determination ($R^2 \approx 0.00$) indicates that differences in PPP levels explain virtually none of the variation in tourism arrival recovery across countries. This pattern persists across regions, suggesting that the lack of association is not driven by a small subset of outliers or region-specific effects.

Interpretation: This result suggests that differences in absolute affordability levels in 2022 did not significantly influence tourists' destination choices during the recovery period. While it may be intuitive to assume that travelers would favor countries that were relatively cheaper in absolute terms, the data indicates that affordability shifts alone were not a dominant factor in determining where tourists chose to travel post-pandemic. Instead, non-price considerations – such as travel restrictions, public health policies, perceived safety, flight availability, and pent-up demand for specific destinations – likely played a much larger role in shaping travel behavior. This finding challenges the assumption that inflation-induced affordability changes would meaningfully redirect global tourism flows in the short term.

While international tourist arrivals capture how destination choices vary with affordability, they do not fully reflect the economic impact of tourism. To assess whether affordability influences not only where tourists travel but also how much they spend once they arrive, the next analysis examines the relationship between absolute Purchasing Power Parity levels and changes in tourism receipts. This distinction allows for a more complete evaluation of whether baseline cost-of-living differences affect tourism behavior through spending intensity rather than travel volume alone.

Figure 3: Purchasing Power Parity Level (2022) vs.
Percentage Change in International Tourism Receipts (2019–2022)

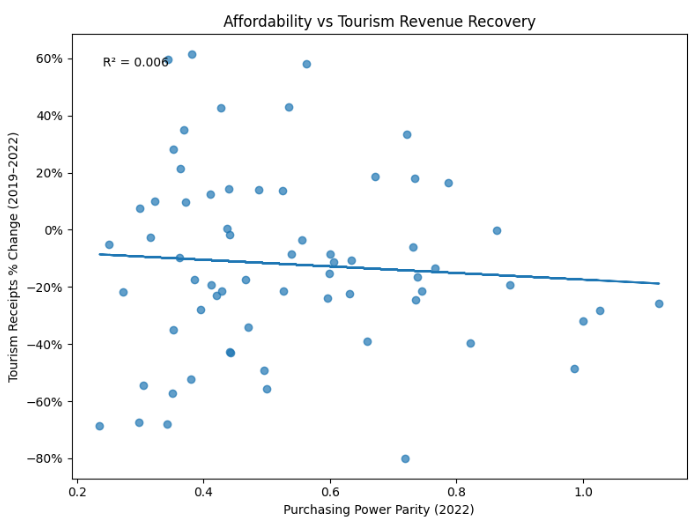


Figure 3 presents the relationship between absolute affordability levels in 2022 and percentage change in international tourism receipts between 2019 and 2022. A linear regression trendline and R^2 value are included to assess explanatory power.

Observed Relationship: The relationship between Purchasing Power Parity levels in 2022 and percentage change in tourism receipts similarly shows no substantive correlation. The scatter plot displays a broad, diffuse distribution of points with no consistent directional pattern, and the regression line again exhibits an almost negligible slope. The R² value (≈ 0.006) confirms that relative affordability changes explain less than one percent of the variation in tourism revenue recovery across countries. Even countries with relatively high or low PPP levels in 2022 do not exhibit higher or lower growth in tourism receipts.

Interpretation: This finding indicates that differences in a country's relative affordability during the pandemic period did not materially affect its post-pandemic tourism recovery rate. While affordability may influence individual purchasing decisions at the micro level, it does not appear to translate into measurable differences in aggregate tourism spending across countries. This reinforces the conclusion that post-pandemic tourism recovery was shaped more by structural and behavioral factors – such as travel demand rebound, length of stay, and tourism capacity constraints – than by relative price competitiveness alone. Together with the arrivals analysis, these results suggest that pandemic-era inflation altered domestic economic conditions without fundamentally reshaping global tourism behavior.

Relative Affordability Change and Tourism Recovery

With the baseline distribution of the data established and absolute affordability shown to have limited explanatory power, the analysis next shifts to examining relative affordability change between 2019 and 2022. Rather than asking whether tourists favored destinations that were objectively cheaper, this subsection investigates whether countries that experienced larger economic price shifts during the pandemic recovered tourism activity more rapidly. By focusing on percentage changes in Purchasing Power Parity, this analysis captures inflationary disruption and economic volatility during COVID-19, allowing for evaluation of how tourism recovery responded to relative economic change rather than static cost levels.

Figure 4: Percentage Change in Purchasing Power Parity vs. Percentage Change in International Tourism Receipts (2019–2022)

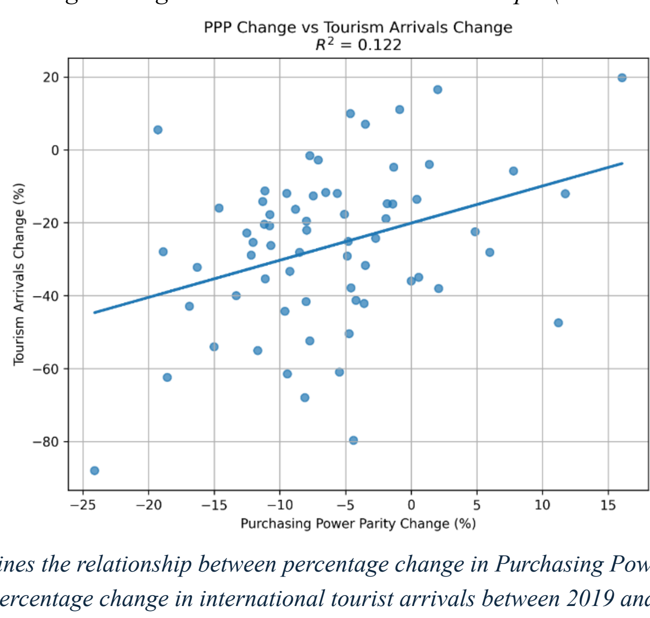


Figure 4 examines the relationship between percentage change in Purchasing Power Parity (PPP) and percentage change in international tourist arrivals between 2019 and 2022. A linear regression trendline and coefficient of determination (R²) are included to assess the strength of association.

Observation: The scatter plot compares the percentage change in Purchasing Power Parity between 2019 and 2022 with the percentage change in international tourist arrivals. A positive percentage change in PPP indicates that a country became relatively more expensive over this period, reflecting higher inflation or faster increases in domestic prices relative to the global baseline. The plot reveals a modest upward-sloping relationship, meaning that countries experiencing larger increases in PPP – those where costs rose more sharply – also tended to see higher percentage growth in tourist arrivals during the post-pandemic recovery, even as prices increased. Although the data points are dispersed, the fitted regression line consistently trends upward, and the R^2 value of approximately 0.12 indicates that relative changes in affordability account for a non-trivial share of the variation in tourism arrival recovery.

Interpretation: This result appears counterintuitive, as higher PPP growth reflects greater inflation rather than improved affordability. However, this relationship suggests that inflation-related price changes were not the dominant factor associated with tourism recovery patterns during this period. Instead, the findings indicate that accumulated travel demand outweighed price sensitivity, with consumers prioritizing travel experiences after prolonged restrictions. This pattern is consistent with the concept of 'revenge travel', in which travelers demonstrate a willingness to absorb higher costs in exchange for long-delayed travel experiences. As a result, destinations that became more expensive still experienced strong tourism rebounds, demonstrating that demand dynamics dominated affordability considerations during recovery.

Figure 5. Percentage Change in Purchasing Power Parity vs.
Percentage Change in International Tourism Receipts (2019–2022)

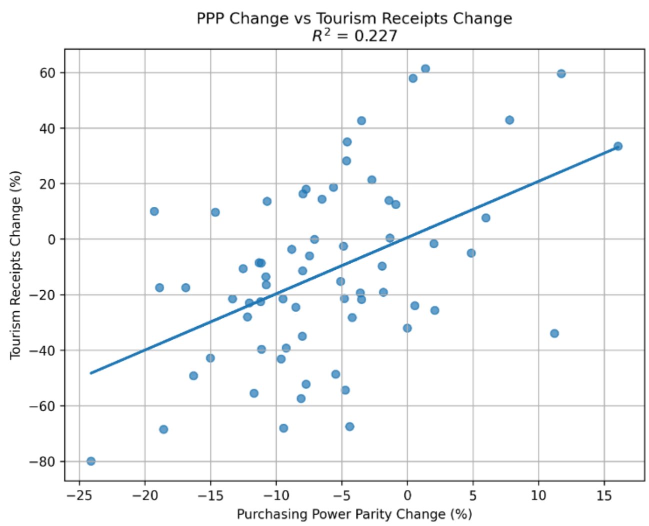


Figure 5 examines the relationship between percentage change in Purchasing Power Parity (PPP) and percentage change in international tourism receipts between 2019 and 2022. A linear regression trendline and coefficient of determination (R^2) are included to evaluate the explanatory strength of affordability change on tourism spending recovery.

Observation: The scatter plot comparing percentage change in Purchasing Power Parity between 2019 and 2022 with percentage change in international tourism receipts reveals a clear and more pronounced positive relationship than observed in the arrivals analysis. A positive change in PPP indicates that a country experienced relatively higher inflation and rising domestic prices over the pandemic and recovery period. The plot shows that countries with larger increases in PPP also tended to experience substantially higher percentage growth in tourism receipts. The regression line exhibits a steeper upward slope than in the arrival's comparison, and the coefficient of determination ($R^2 \approx 0.23$) indicates that nearly one-quarter of the variation in tourism revenue recovery across countries can be explained by changes in PPP. This suggests a stronger and more consistent association between relative price changes and tourism spending recovery outcomes than between price changes and visitor volumes.

Interpretation: This stronger relationship indicates that post-pandemic inflationary dynamics were reflected more directly in tourism spending behavior than in destination choice. While higher prices did not prevent tourists from traveling, they were associated with significantly increased per-visitor expenditure. This implies that travelers returning after the pandemic were less sensitive to price increases and more willing to absorb higher costs in exchange for travel experiences. Such behavior is consistent with demand-driven recovery patterns characterized by pent-up consumption, longer stays, and increased spending on premium services and experiences. Rather than signaling reduced affordability alone, rising PPP values during this period appear to reflect heightened demand pressure within tourism markets, where constrained supply and elevated willingness to spend combined to drive revenue growth. Consequently, the relationship between PPP change and tourism receipts reinforces the conclusion that post-pandemic tourism recovery was shaped primarily by demand intensity rather than affordability considerations.

Multivariable Predictive Modeling of Tourism Spending Recovery

The preceding analyses examined affordability and tourism recovery through isolated, bivariate relationships. While these results provided important insight into how affordability dynamics interacted with tourism outcomes, they did not account for the combined influence of multiple factors operating simultaneously. To address this limitation, a multivariable predictive model was constructed to evaluate how changes in affordability and tourism volume jointly influenced tourism spending recovery following the COVID-19 pandemic. By modeling tourism receipts as a function of both purchasing power parity change and tourist arrivals change, this approach enables a more comprehensive assessment of the mechanisms driving post-pandemic tourism recovery.

Multivariable Regression Framework: A multivariable linear regression model was constructed to disentangle the overlapping effects of tourism volume and economic conditions on tourism revenue recovery. Earlier bivariate analyses demonstrated that changes in Purchasing Power Parity were more strongly associated with tourism receipts than with arrivals, suggesting that spending behavior rather than destination choice was the primary channel through which affordability-related economic changes manifested. This model was therefore designed to evaluate whether affordability change retained explanatory power once differences in tourist arrival growth were explicitly controlled for, allowing for a clearer assessment of how these factors jointly influenced tourism receipts recovery.

Figure 6: Multivariable regression predicting percentage change in tourism receipts.

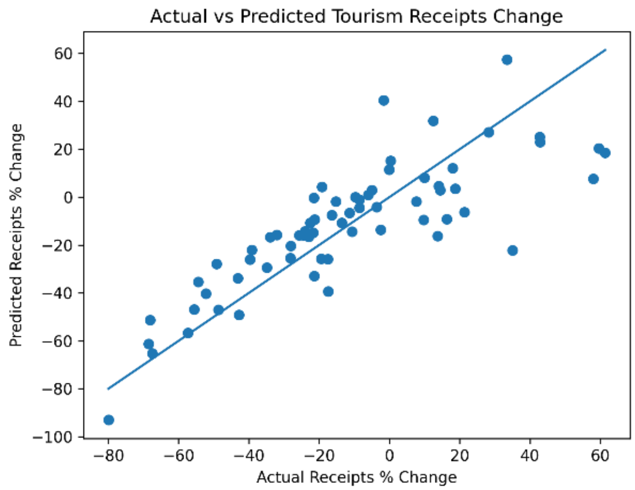


Figure 6: Multivariable linear regression modeling the percentage change in international tourism receipts as a function of changes in Purchasing Power Parity and international tourist arrivals

Model Output and Performance:

- Dependent Variable: Predicted Receipts Percentage Change
- Intercept: 21.2980
- Purchasing Power Parity Coefficient: 0.9751
- Arrivals Change Coefficient: 1.0318
- R^2: 0.667

Coefficient Interpretations:

R^2: The model produced an R^2 value of 0.667, indicating that approximately 66.7% of the variation in percentage change in tourism receipts between 2019 and 2022 is explained by the combined effects of changes in Purchasing Power Parity and changes in tourism arrivals. This relatively high explanatory power suggests that the model captures a substantial portion of the variation in post-pandemic tourism revenue recovery, while the remaining unexplained variation likely reflects country-specific factors such as policy differences, reopening timelines, or destination-specific demand shocks.

Arrivals Coefficient: The arrivals coefficient of 1.03 indicates that, holding affordability constant, a one percentage point increase in international tourist arrivals is associated with an approximately 1.03 percentage point increase in tourism receipts. This near proportional relationship confirms that visitor volume is a dominant and predictable determinant of tourism spending, validating the inclusion of arrivals as a control variable in the model rather than a competing explanatory factor.

Purchasing Power Parity Coefficient: The Purchasing Power Parity coefficient of 0.98 indicates that, controlling changes in tourist arrivals, a one percentage point increase in PPP is associated with an approximately 0.98 percentage point increase in tourism receipts. Because higher PPP values during this period reflect rising relative prices rather than improved affordability, this result supports the conclusion that post-pandemic tourism recovery was demand-driven. Tourists were willing to spend more in higher-cost destinations, suggesting that pent-up demand and destination desirability outweighed price sensitivity during the recovery period.

Intercept: The intercept value of 21.30 reflects baseline growth in tourism receipts occurring independently of changes in affordability and tourist arrivals. This component captures structural post-pandemic recovery forces such as the lifting of travel restrictions, reopening of borders, and pent-up consumer demand for travel. Rather than representing a realistic scenario, the intercept reflects baseline tourism spending recovery not directly explained by affordability or tourist arrivals, reinforcing that post-COVID tourism recovery was driven by broader systemic factors in addition to price and volume effects.

Figure 7: Residual plot for multivariable regression predicting percentage change in tourism receipts.

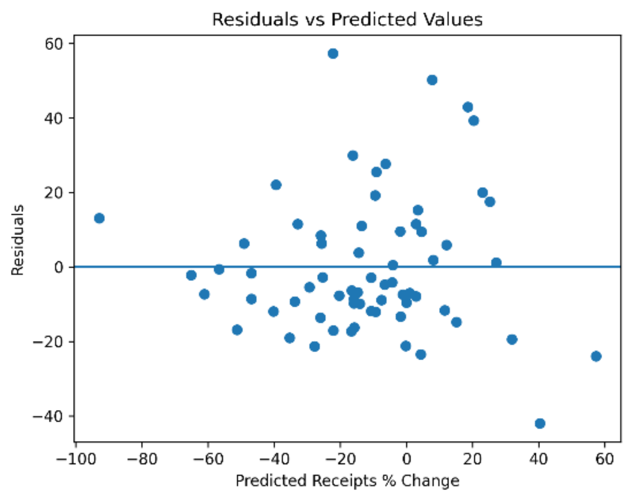


Figure 7: Residual plot for the multivariable regression model predicting percentage change in tourism receipts. Residuals represent the difference between observed and predicted values based on changes in tourism arrivals and Purchasing Power Parity.

Observation: The residual plot displays the differences between the observed and predicted percentage changes in tourism receipts across all countries included in the regression model. The residuals are evenly scattered around zero with no discernible trend, curvature, or clustering as fitted values increase. There is no visible funneling or systematic pattern in the spread of residuals, indicating that prediction errors remain relatively consistent across the range of modeled values.

Interpretation: The absence of a clear pattern in the residual plot suggests that the linear regression model is appropriately specified and does not suffer from major violations of key assumptions such as nonlinearity or heteroskedasticity. This indicates that the relationships between tourism receipts recovery, changes in tourist arrivals, and changes in Purchasing Power Parity are reasonably linear and well captured by the model. As a result, the estimated coefficients and overall explanatory power of the model can be interpreted with confidence, reinforcing the reliability of the predictive findings.

Conclusion

The objective of this study was to evaluate how affordability influenced international tourism recovery following the COVID-19 pandemic and to determine whether tourism outcomes were more strongly associated with absolute cost levels or with relative economic change over time. By integrating international tourism arrivals, tourism receipts, and Purchasing Power Parity (PPP) data across pre-pandemic (2019) and post-pandemic (2022) periods, this project sought to identify the economic mechanisms underlying post-pandemic tourism behavior. Through a structured analytical progression – beginning with data validation, advancing through exploratory analysis, and culminating in predictive modeling – the results collectively provide a cohesive explanation of tourism recovery dynamics in a period of unprecedented disruption.

Distributional analysis established a critical foundation for the study. The roughly symmetric distributions observed across percentage changes in tourism arrivals, tourism receipts, and PPP indicated that extreme skewness or outliers did not dominate the data. This stability supported the use of linear statistical techniques and reinforced confidence that subsequent correlations and regression results reflected genuine relationships rather than artifacts of distorted data. Establishing this baseline was essential before drawing inferences from more advanced modeling stages.

Exploratory analysis revealed a clear distinction between absolute and relative affordability. Absolute affordability, measured by PPP levels in 2022, showed no relationship with tourism recovery outcomes. Countries that were more or less expensive in absolute terms did not systematically experience stronger or weaker growth in either tourism arrivals or tourism receipts. This finding suggests that, in the immediate post-pandemic environment, travelers were not primarily selecting destinations based on static cost-of-living differences. Traditional affordability considerations appeared secondary during the initial recovery phase.

In contrast, relative changes in affordability revealed a meaningful and counterintuitive relationship. Percentage increases in PPP between 2019 and 2022 were positively associated with growth in both tourism arrivals and tourism receipts. Although these relationships were moderate, the consistently upward-sloping trends and non-trivial coefficients of determination demonstrated that countries experiencing greater increases in PPP also tended to experience stronger tourism recovery. At face value, this result appears contradictory, as rising PPP reflects inflation rather than improved affordability. However, this pattern indicates that inflation itself was not the driving force behind recovery outcomes.

Instead, the results suggest that demand-driven forces dominated post-pandemic tourism behavior. Countries experiencing larger PPP increases were often those with stronger baseline tourism demand or more severe pandemic-era disruptions. As travel restrictions eased, pent-up demand and “revenge travel” led consumers to prioritize travel experiences despite rising prices. In this context, relative PPP growth functioned less as a deterrent and more as a proxy for economic disruption intensity and rebound magnitude. Tourists demonstrated a willingness to absorb higher costs in highly desirable destinations, indicating that emotional, social, and experiential motivations outweighed short-term price sensitivity.

The multivariable regression model further clarified these dynamics by jointly evaluating affordability change and tourism volume. The model confirmed that percentage change in tourism arrivals was the dominant driver of tourism receipts growth, reflecting the direct relationship between visitor volume and aggregate spending. Importantly, affordability dynamics remained indirectly embedded within the model through their association with arrivals and broader recovery patterns. The model's R^2 value of approximately 0.67 indicated strong explanatory power, and residual analysis showed no systematic structure, supporting the model's reliability and internal consistency.

While the observed relationships are statistically meaningful, it is important to distinguish correlation from causation. The findings do not imply that inflation caused tourism growth, nor that higher prices attract travelers. Rather, they demonstrate that during an extraordinary recovery period, demand-side forces temporarily overwhelmed traditional affordability constraints. These results highlight how large-scale disruptions can reshape consumer behavior and weaken established economic relationships, at least in the short term.

Overall, this study demonstrates how careful data engineering, rigorous standardization, and layered analytical techniques can uncover nuanced insights into post-crisis economic behavior. By separating absolute affordability from relative economic change and integrating exploratory and predictive analysis, the project provides a clear and cohesive explanation of how global tourism recovered in the wake of COVID-19.

Next Steps

While this study provides a structured and data-driven examination of post-pandemic tourism recovery, several avenues exist to refine and extend the analytical framework in future work. One important next step would be expanding the temporal scope beyond the 2019–2022 window. Including additional post-recovery years would help determine whether the demand-driven patterns observed in this analysis represent a short-term response to pandemic conditions or whether they persist as travel behavior normalizes. A longer time horizon would also allow for the identification of potential inflection points where traditional price sensitivity begins to reassert itself.

Future analysis could also benefit from the inclusion of additional explanatory variables that capture structural and policy-driven influences on tourism recovery. Factors such as exchange rate fluctuations, government travel restrictions, airline capacity constraints, tourism infrastructure investment, or public health policies could improve explanatory power and help isolate the specific mechanisms through which recovery occurred. Incorporating such variables would allow for a more comprehensive assessment of how economic, logistical, and regulatory conditions interacted with consumer demand during the recovery period.

In addition, alternative measures of affordability could be explored to complement or refine the use of Purchasing Power Parity. While PPP provides a broad measure of relative cost-of-living, metrics such as tourism-specific price indices, exchange-rate-adjusted travel costs, or average visitor expenditure benchmarks may better capture the prices tourists face when making travel decisions. Comparing results across multiple affordability measures could help distinguish general inflationary effects from tourism-specific pricing dynamics.

Another meaningful extension would involve stratifying the analysis by region or income level. A global, country-level approach may mask important heterogeneity in recovery dynamics across different economic or geographic contexts. Segmenting countries by income classification, tourism dependency, or regional characteristics could reveal whether the relationships identified in this study vary systematically across different subsets of the global tourism market.

Finally, more advanced modeling techniques could be employed to better capture the complexity of post-crisis recovery while maintaining interpretability. Panel regression models could leverage repeated observations across time, while nonlinear models or interaction terms could account for threshold effects or conditional relationships between affordability, demand, and spending. These approaches would build directly on the current framework, enhancing analytical depth without sacrificing transparency.
Together, these methodological refinements would strengthen the robustness of future analyses while preserving the core structure established in this project, enabling a deeper understanding of how global tourism responds to large-scale economic disruptions.

International Tourism Recovery and Affordability After COVID-19

Member IDs: 1098102 (Captain), 1233511

National TSA Conference | June 2026 | Washington DC, MD

Introduction

The COVID-19 pandemic produced one of the most significant global economic disruptions in modern history, sharply reducing international tourism and altering inflation dynamics, cost-of-living conditions, and global mobility. While the immediate collapse of tourism during the pandemic was well documented, the economic factors that shaped tourism recovery after international travel resumed remain less well understood.

As countries reopened, travelers encountered a global tourism market in which relative prices and affordability conditions had changed unevenly compared to pre-pandemic expectations. This raised an important question: did affordability influence where tourism recovered, or were recovery patterns driven by broader demand-side forces following prolonged travel restrictions?

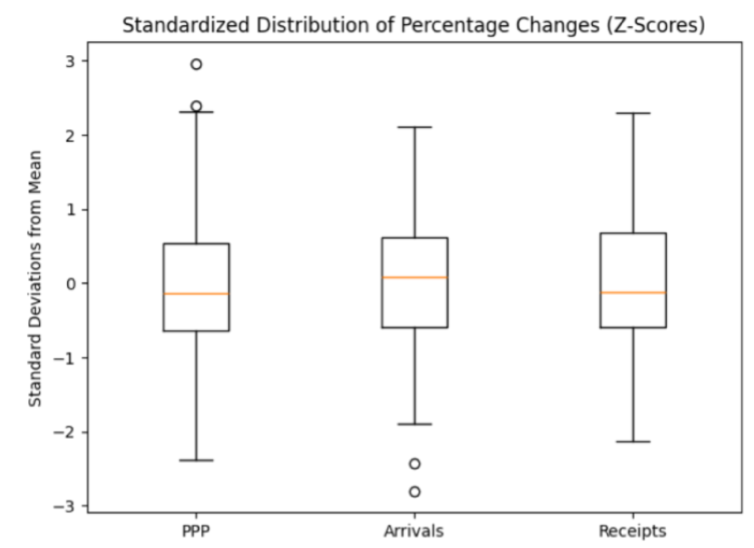
Purpose

The purpose of this study is to evaluate how affordability influenced international tourism recovery following the COVID-19 pandemic. Specifically, the analysis examines whether tourism recovery outcomes were more strongly associated with absolute destination affordability levels in 2022 or with changes in affordability between 2019 and 2022 caused by pandemic-era economic disruption.

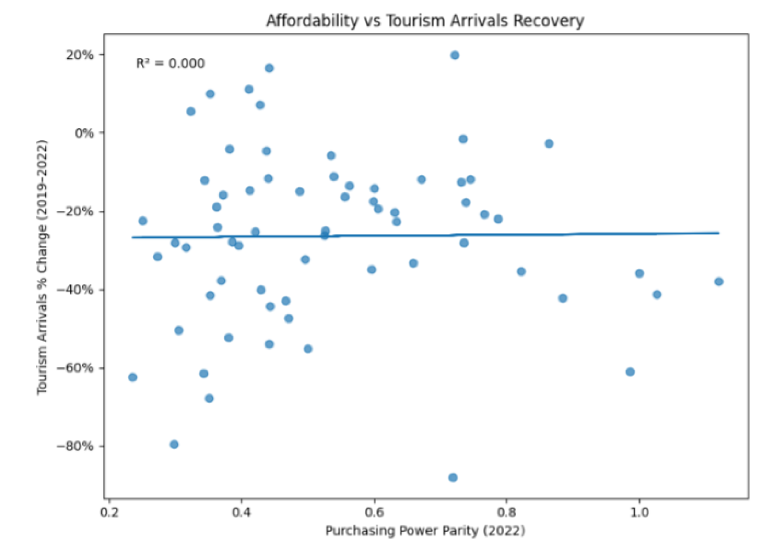
Methods

- Collected international tourism arrivals and tourism receipts data for 2019 and 2022 from the United Nations World Tourism Organization
- Collected Purchasing Power Parity data from the World Bank as a standardized measure of relative cost of living
- Standardized country names using ISO-based matching and merged datasets at the country level
- Calculated percentage change variables to normalize differences in country size and baseline tourism levels
- Used scatter plots and linear regression to examine relationships between affordability and tourism recovery
- Constructed a multivariable regression model to evaluate the combined influence of tourism volume and affordability change on tourism spending recovery

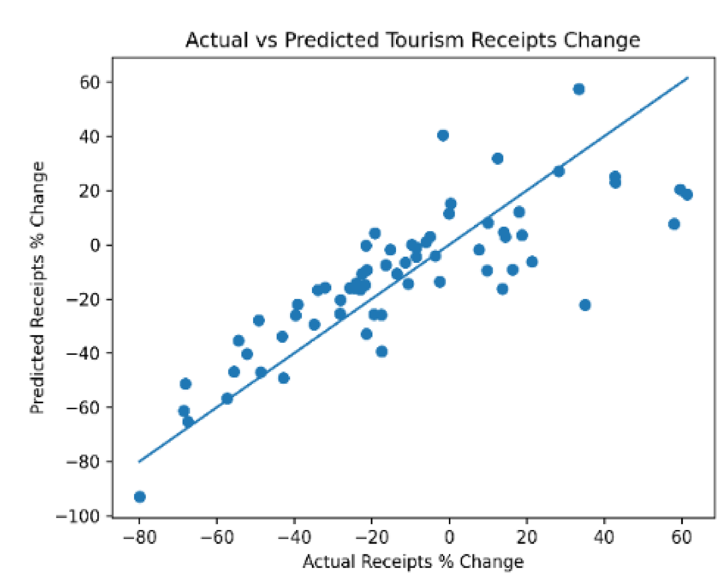
Results



Distribution of percentage change in Purchasing Power Parity, international tourist arrivals, and international tourism receipts between 2019 and 2022. Key tourism and affordability variables are approximately symmetric and not dominated by extreme outliers, supporting reliable regression-based analysis.

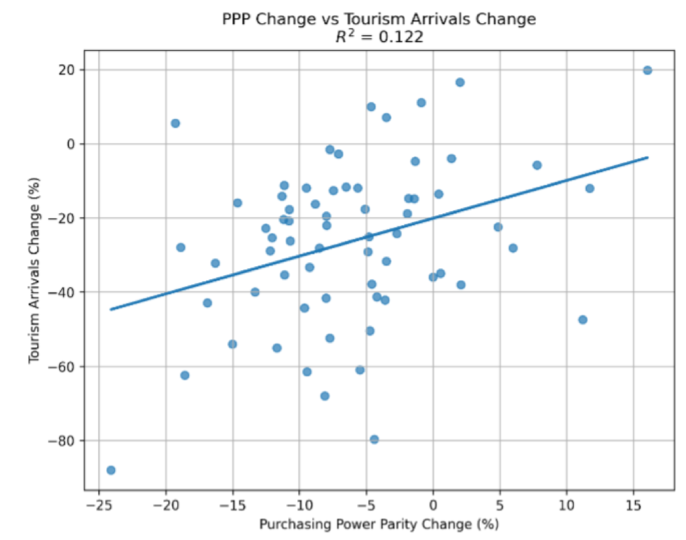


Purchasing Power Parity level in 2022 versus percentage change in international tourist arrivals between 2019 and 2022. Absolute destination affordability in the post-pandemic period shows no meaningful relationship with tourism arrival recovery.

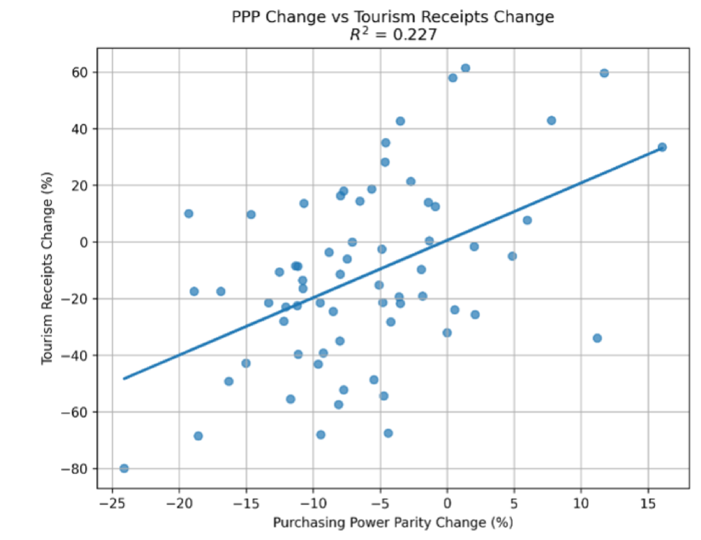


Multivariable linear regression model predicting percentage change in tourism receipts using changes in Purchasing Power Parity and tourist arrivals between 2019 and 2022. Both variables produced positive coefficients, indicating tourism revenue recovery was positively associated with visitor growth and relative price increases in the post-pandemic period.

Percentage change in Purchasing Power Parity versus percentage change in international tourist arrivals between 2019 and 2022. Countries experiencing larger affordability shifts often saw stronger recovery in tourist arrivals despite rising prices.



Percentage change in Purchasing Power Parity versus percentage change in international tourism receipts between 2019 and 2022. Changes in affordability are positively associated with tourism spending recovery, indicating demand-driven behavior rather than price sensitivity.



Conclusion

Post-pandemic tourism recovery was driven primarily by demand rather than affordability constraints. Travelers returned to international travel despite rising costs, prioritizing experiences over price sensitivity. As a result, tourism spending rebounded even in destinations that became more expensive during the recovery period. These findings highlight how large-scale economic disruptions can temporarily weaken traditional price-based consumer behavior, reshaping tourism recovery dynamics in the short term.

Next Steps (TPACK)

- Extend the analysis beyond 2022 by including additional post-recovery years to evaluate whether the demand-driven recovery patterns identified in this study persist as international travel stabilizes and the effects of pent-up demand diminish over time
- Incorporate additional economic, logistical, and policy-related variables such as exchange rate fluctuations, border closure and reopening timelines, and government travel restrictions to better isolate the factors that contributed to differences in tourism recovery across countries
- Compare recovery dynamics across regions, income classifications, and levels of tourism dependence to assess whether the observed demand-driven effects were consistent globally or varied based on economic structure and reliance on international tourism

References

- United Nations World Tourism Organization. (2023). International tourism highlights: 2023 edition – The impact of COVID-19 on tourism (2020–2022). UN Tourism. https://www.eunwto.org/doi/pdf/10.18111/9789284424986
- World Bank. (2024). Purchasing power parity (PPP). World Bank Open Data. https://data.worldbank.org/indicator/PA.NUS.PPP
- McKinney, W. Camp; pandas development team. (2024). pandas: Powerful data structures for data analysis in Python. https://pandas.pydata.org
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science Camp; Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55
- Pedregosa, F., Viroqua, G., Gram-fort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://scikit-learn.org
- Py-country development team. (2024). Py-country: ISO country, subdivision, language, currency and script definitions. https://pypi.org/project/pycountry/

Bibliography

United Nations World Tourism Organization. (2023). *International tourism highlights: 2023 edition – The impact of COVID-19 on tourism (2020–2022)*. UN Tourism. https://www.e-unwto.org/doi/pdf/10.18111/9789284424986

World Bank. (2024). *Purchasing power parity (PPP)*. World Bank Open Data. https://data.worldbank.org/indicator/PA.NUS.PPP

McKinney, W., & pandas development team. (2024). *pandas: Powerful data structures for data analysis in Python*. https://pandas.pydata.org

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://scikit-learn.org

pycountry development team. (2024). *pycountry: ISO country, subdivision, language, currency and script definitions*. https://pypi.org/project/pycountry/

Appendix

A.1 Regional Tourism Data Consolidation

Purpose:

The purpose of this project was to consolidate international tourism arrivals and tourism receipts data from multiple geographic regions into a single, standardized dataset suitable for cross-country and longitudinal analysis.

Code Description: This script programmatically loads all regional tourism CSV files from a single directory, identifies each file's region based on its filename, and separates arrivals and receipts data. The data are reshaped into a consistent long format by year, merged across regions, and combined into a unified country-level dataset containing tourism arrivals and tourism receipts for 2019 and 2022. The final consolidated dataset is exported for use in subsequent affordability and predictive analyses

```
            df = df.melt(
                id_vars=["Country code", "Country/Group name"],
                value_vars=["2019", "2022"],
                var_name="Year",
                value_name="Arrivals"
            )

            df["Region"] = region
            arrivals_list.append(df)

        elif "Receipts" in filename:
            df = df[[
                "Country code",
                "Country/Group name",
                "2019",
                "2022"
            ]]

            df = df.melt(
                id_vars=["Country code", "Country/Group name"],
                value_vars=["2019", "2022"],
                var_name="Year",
                value_name="Receipts"
            )

            df["Region"] = region
            receipts_list.append(df)

arrivals_df = pd.concat(arrivals_list, ignore_index=True)
receipts_df = pd.concat(receipts_list, ignore_index=True)

final_df = pd.merge(
    arrivals_df,
    receipts_df,
    on=["Country code", "Country/Group name", "Year", "Region"],
```

```
import pandas as pd
import glob
import os

DATA_FOLDER = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\Tourism Data"

arrivals_list = []
receipts_list = []

for file in glob.glob(os.path.join(DATA_FOLDER, "*.csv")):
    filename = os.path.basename(file)

    if "Europe" in filename:
        region = "Europe"
    elif "Americas" in filename:
        region = "Americas"
    elif "Africa" in filename:
        region = "Africa"
    elif "Asia" in filename:
        region = "Asia"
    elif "Middle East" in filename:
        region = "Middle East"
    else:
        continue

    df = pd.read_csv(file, encoding="latin1")

    if "Arrivals" in filename:
        df = df[[
            "Country code",
            "Country/Group name",
            "2019",
            "2022"
        ]]

```

```
    how="outer"
)

final_df = final_df.sort_values(["Region", "Country/Group name", "Year"])

OUTPUT_FILE = r"C:\Users\lucas\OneDrive\Desktop\25-26 TSA Project\merged_tourism_data.csv"
final_df.to_csv(OUTPUT_FILE, index=False)

print("Final dataset created successfully.")
```

A.2 Country Name Standardization and Validation

Purpose:

The purpose of this project was to standardize country names across the tourism dataset and the Purchasing Power Parity (PPP) dataset to enable accurate country-level merging in subsequent analysis.

Code Description:

This script applies a multi-step country name standardization pipeline that cleans formatting inconsistencies, removes accents and punctuation, and converts names to a uniform title-case structure. Country names are then aligned to ISO-recognized standards using the pycountry library, with a small set of documented manual overrides applied to resolve remaining edge cases. The script generates a diagnostic file listing unmatched country or regional names, allowing verification of standardization accuracy before merging datasets.

```python
import pandas as pd
import unicodedata
import re
import pycountry

TOURISM_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\merged_tourism_data.csv"
PPP_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\Purchasing_Price_Parity_Data.csv"

tourism_df = pd.read_csv(TOURISM_FILE, encoding="latin1")
ppp_df = pd.read_csv(PPP_FILE, encoding="latin1")

def basic_normalize(name):
    """Basic text cleaning and title-casing"""
    if pd.isna(name):
        return None

    name = str(name)
    name = unicodedata.normalize("NFKD", name)
    name = "".join(c for c in name if not unicodedata.combining(c))
    name = re.sub(r"[^\w\s]", "", name)
    name = re.sub(r"\s+", " ", name).strip()
    return name.title()

def iso_normalize(name):
    """
    Convert country name to ISO-recognized name using pycountry.
    Falls back to cleaned name if no ISO match is found.
    """
    if name is None:
        return None

    try:
        country = pycountry.countries.lookup(name)
        return country.name
    except LookupError:
        return name

MANUAL_OVERRIDES = {
    "Brunei Darussalam": "Brunei",
    "Congo Rep": "Congo",
    "Egypt Arab Rep": "Egypt",
    "Somalia Fed Rep": "Somalia",
    "Venezuela, Bolivarian Republic of": "Venezuela",
    "Venezuela Rb": "Venezuela",
    "Utd Arab Emirates": "United Arab Emirates",
    "Yemen Rep": "Yemen"
}

def apply_manual_override(name):
    if name in MANUAL_OVERRIDES:
        return MANUAL_OVERRIDES[name]
    return name

tourism_df["Country_clean"] = tourism_df["Country/Group name"].apply(basic_normalize)
ppp_df["Country_clean"] = ppp_df["Country Name"].apply(basic_normalize)

tourism_df["Country_std"] = tourism_df["Country_clean"].apply(iso_normalize)
ppp_df["Country_std"] = ppp_df["Country_clean"].apply(iso_normalize)

tourism_df["Country_std"] = tourism_df["Country_std"].apply(apply_manual_override)
ppp_df["Country_std"] = ppp_df["Country_std"].apply(apply_manual_override)

tourism_countries = set(tourism_df["Country_std"].dropna().unique())
ppp_countries = set(ppp_df["Country_std"].dropna().unique())

missing_in_ppp = sorted(tourism_countries - ppp_countries)
missing_in_tourism = sorted(ppp_countries - tourism_countries)

mismatch_df = pd.DataFrame({
    "Tourism_only": pd.Series(missing_in_ppp),
    "PPP_only": pd.Series(missing_in_tourism)
})

OUTPUT_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\country_name_mismatches_new3.csv"
mismatch_df.to_csv(OUTPUT_FILE, index=False)

print("☑ Country name standardization complete.")
print("☑ Manual overrides applied successfully.")
print("☑ country_name_mismatches_new3.csv created.")
```

A.3 Tourism and Affordability Dataset Construction

Purpose:

The purpose of this project was to merge standardized international tourism data with Purchasing Power Parity (PPP) data to create a single, country-level dataset containing both tourism activity and affordability measures for pre-pandemic and post-pandemic years.

Code Description:

This script applies the validated country name standardization logic from Project 2 to align tourism and PPP datasets using consistent country identifiers. PPP data were filtered to include only the years 2019 and 2022 and reshaped to match the year-based structure of the tourism dataset. The datasets were then merged at the country-year level, excluding countries or regions without complete data coverage. The resulting consolidated dataset was sorted and exported as the foundational input for all subsequent data restructuring, cleaning, and analytical modeling.

```python
import pandas as pd
import unicodedata
import re
import pycountry

TOURISM_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\merged_tourism_data.csv"
PPP_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\Purchasing_Price_Parity_Data.csv"

tourism_df = pd.read_csv(TOURISM_FILE, encoding="latin1")
ppp_df = pd.read_csv(PPP_FILE, encoding="latin1")

def basic_normalize(name):
    if pd.isna(name):
        return None

    name = str(name)
    name = unicodedata.normalize("NFKD", name)
    name = "".join(c for c in name if not unicodedata.combining(c))
    name = re.sub(r"[^\w\s]", "", name)
    name = re.sub(r"\s+", " ", name).strip()

    return name.title()


def iso_normalize(name):
    if name is None:
        return None

    try:
        return pycountry.countries.lookup(name).name
    except LookupError:
        return name

MANUAL_OVERRIDES = {
    "Brunei Darussalam": "Brunei",
    "Congo Rep": "Congo",
    "Egypt Arab Rep": "Egypt",
    "Somalia Fed Rep": "Somalia",
    "Venezuela Bolivarian Republic Of": "Venezuela",
    "Venezuela Rb": "Venezuela",
    "Utd Arab Emirates": "United Arab Emirates",
    "Yemen Rep": "Yemen"
}

def apply_manual_override(name):
    return MANUAL_OVERRIDES.get(name, name)

tourism_df["Country_std"] = (
    tourism_df["Country/Group name"]
    .apply(basic_normalize)
    .apply(iso_normalize)
    .apply(apply_manual_override)
)

ppp_df["Country_std"] = (
    ppp_df["Country Name"]
    .apply(basic_normalize)
    .apply(iso_normalize)
    .apply(apply_manual_override)
)

ppp_df = ppp_df[["Country_std", "2019", "2022"]]

ppp_long = ppp_df.melt(
    id_vars="Country_std",
    value_vars=["2019", "2022"],
    var_name="Year",
    value_name="PPP"
)

ppp_long["Year"] = ppp_long["Year"].astype(int)

final_df = pd.merge(
    tourism_df,
    ppp_long,
    on=["Country_std", "Year"],
    how="inner"
)

final_df = final_df.sort_values(["Region", "Country_std", "Year"])

OUTPUT_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\final_tourism_affordability_dataset.csv"
final_df.to_csv(OUTPUT_FILE, index=False)

print("☑ Final tourism + PPP dataset created successfully.")
```

A.4 Data Reshaping and Structural Preparation

Purpose:

The purpose of this project was to restructure the merged tourism and affordability dataset into a wide format that enables direct year-to-year comparison across countries.

Code Description:

This script filters the merged dataset to include only the years 2019 and 2022 and pivots the data so that each country appears in a single row. Tourism arrivals, tourism receipts, and purchasing power parity values for each year are separated into distinct columns. This restructuring aligns all relevant variables within the same observation, establishing a consistent data structure required for data cleaning, change calculations, and statistical analysis in subsequent projects.

```python
import pandas as pd

DATA_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\final_tourism_affordability_dataset.csv"
df = pd.read_csv(DATA_FILE, encoding="latin1")

df = df[df["Year"].isin([2019, 2022])]

df_wide = df.pivot_table(
    index=["Country_std", "Region"],
    columns="Year",
    values=["Arrivals", "Receipts", "PPP"],
    aggfunc="first"
)

df_wide.columns = [
    f"{metric}_{year}" for metric, year in df_wide.columns
]

df_wide.reset_index(inplace=True)

OUTPUT_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\wide_tourism_ppp_data.csv"
df_wide.to_csv(OUTPUT_FILE, index=False)

print("Data reshaped successfully.")
print("One row per country.")
print("wide_tourism_ppp_data.csv created.")
```

A.5 Data Cleaning and Change Metric Construction

Purpose:

The purpose of this project was to clean the reshaped dataset and construct standardized change metrics that quantify tourism and affordability changes between 2019 and 2022.

Code Description:

This script converts all relevant tourism and purchasing power parity variables to numeric format and removes countries with missing values in any required field to ensure valid year-to-year comparisons. Absolute and percentage change variables are then calculated for tourism arrivals, tourism receipts, and purchasing power parity. Percentage change metrics were prioritized to normalize differences in scale and units across countries. Raw year-specific columns are removed, producing a streamlined, analysis-ready dataset used for visualization and predictive modeling.

```python
INPUT_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\wide_tourism_ppp_data.csv"
df = pd.read_csv(INPUT_FILE, encoding="latin1")

NUMERIC_COLS = [
    "Arrivals_2019", "Arrivals_2022",
    "Receipts_2019", "Receipts_2022",
    "PPP_2019", "PPP_2022"
]

for col in NUMERIC_COLS:
    df[col] = pd.to_numeric(df[col], errors="coerce")

df_clean = df.dropna(subset=NUMERIC_COLS).copy()

df_clean["Arrivals_Change_Abs"] = df_clean["Arrivals_2022"] - df_clean["Arrivals_2019"]
df_clean["Receipts_Change_Abs"] = df_clean["Receipts_2022"] - df_clean["Receipts_2019"]
df_clean["PPP_Change_Abs"] = df_clean["PPP_2022"] - df_clean["PPP_2019"]

df_clean["Arrivals_Change_Pct"] = (
    (df_clean["Arrivals_2022"] - df_clean["Arrivals_2019"]) / df_clean["Arrivals_2019"]
) * 100

df_clean["Receipts_Change_Pct"] = (
    (df_clean["Receipts_2022"] - df_clean["Receipts_2019"]) / df_clean["Receipts_2019"]
) * 100

df_clean["PPP_Change_Pct"] = (
    (df_clean["PPP_2022"] - df_clean["PPP_2019"]) / df_clean["PPP_2019"]
) * 100

df_clean = df_clean.drop(columns=NUMERIC_COLS)

df_clean = df_clean.sort_values("Country_std")

OUTPUT_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\final_analysis_ready_dataset.csv"
df_clean.to_csv(OUTPUT_FILE, index=False)
```

A.6 Bivariate Analysis and Relationship Visualization

Purpose:

The purpose of this project was to visually examine bivariate relationships between changes in affordability and changes in tourism performance following the COVID-19 pandemic.

Code Description:

This script generates scatter plots comparing percentage changes in purchasing power parity with percentage changes in tourism arrivals and tourism receipts across countries. Linear regression trendlines and corresponding coefficients of determination (R^2 values) are added to each plot to quantify the strength and direction of observed relationships. All visualizations are programmatically generated using Python and saved as image files to ensure consistent formatting, reproducibility, and direct integration into the research paper.

```python
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import os

DATA_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\final_analysis_ready_dataset.csv"
OUTPUT_DIR = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\Project_6_Graphs"

os.makedirs(OUTPUT_DIR, exist_ok=True)

df = pd.read_csv(DATA_FILE)

def regression_plot(x, y, x_label, y_label, title, filename):
    plot_df = df[[x, y]].dropna()

    X = plot_df[x].values.reshape(-1, 1)
    Y = plot_df[y].values

    model = LinearRegression()
    model.fit(X, Y)
    Y_pred = model.predict(X)
    r2 = r2_score(Y, Y_pred)

    plt.figure(figsize=(8, 6))
    plt.scatter(X, Y, alpha=0.7)
    plt.plot(X, Y_pred, linewidth=2)
    plt.xlabel(x_label)
    plt.ylabel(y_label)
    plt.title(f"{title}\n$R^2$ = {r2:.3f}")
    plt.grid(True)

    plt.savefig(os.path.join(OUTPUT_DIR, filename), dpi=300, bbox_inches="tight")
    plt.close()
```

```python

regression_plot(
    x="PPP_Change_Pct",
    y="Arrivals_Change_Pct",
    x_label="Purchasing Power Parity Change (%)",
    y_label="Tourism Arrivals Change (%)",
    title="PPP Change vs Tourism Arrivals Change",
    filename="PPP_vs_Arrivals.png"
)

regression_plot(
    x="PPP_Change_Pct",
    y="Receipts_Change_Pct",
    x_label="Purchasing Power Parity Change (%)",
    y_label="Tourism Receipts Change (%)",
    title="PPP Change vs Tourism Receipts Change",
    filename="PPP_vs_Receipts.png"
)

regression_plot(
    x="Arrivals_Change_Pct",
    y="Receipts_Change_Pct",
    x_label="Tourism Arrivals Change (%)",
    y_label="Tourism Receipts Change (%)",
    title="Tourism Arrivals Change vs Receipts Change",
    filename="Arrivals_vs_Receipts.png"
)

regression_plot(
    x="PPP_Change_Pct",
    y="Receipts_Change_Pct",
    x_label="Purchasing Power Parity Change (%)",
    y_label="Tourism Receipts Change (%)",
    title="Affordability and Tourism Revenue Relationship",
    filename="PPP_vs_Receipts_Strength.png"
)
```

A.7 Distribution Analysis of Key Variables

Purpose:

The purpose of this project was to examine the distributional properties of key variables to validate the statistical quality of the dataset prior to predictive modeling.

Code Description:

This script generates distribution visualizations for percentage changes in purchasing power parity, tourism arrivals, and tourism receipts across countries. Histograms are used to assess the spread and symmetry of each variable, and a standardized stacked box plot is created to compare distributions on a common scale. These visual checks confirm that the dataset is not dominated by extreme skew or outliers, supporting the validity of subsequent regression-based analysis.

```python
import pandas as pd
import matplotlib.pyplot as plt
import os

INPUT_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\final_analysis_ready_dataset.csv"
df = pd.read_csv(INPUT_FILE, encoding="latin1")

OUTPUT_FOLDER = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\Project_7_Distributions"
os.makedirs(OUTPUT_FOLDER, exist_ok=True)

variables = {
    "PPP_Change_Pct": "Purchasing Power Parity % Change",
    "Arrivals_Change_Pct": "Tourism Arrivals % Change",
    "Receipts_Change_Pct": "Tourism Receipts % Change"
}

for col, label in variables.items():
    plt.figure()
    plt.hist(df[col].dropna(), bins=30)
    plt.title(f"Distribution of {label} (2019-2022)")
    plt.xlabel(label)
    plt.ylabel("Number of Countries")
    plt.tight_layout()

    filename = f"{col}_histogram.png"
    plt.savefig(os.path.join(OUTPUT_FOLDER, filename))
    plt.close()

standardized_df = pd.DataFrame()
for col in variables.keys():
    standardized_df[col] = (df[col] - df[col].mean()) / df[col].std()

plt.figure()
plt.boxplot(
    [standardized_df[col].dropna() for col in variables.keys()],
    labels=["PPP", "Arrivals", "Receipts"],
    showfliers=True
)

plt.title("Standardized Distribution of Percentage Changes (Z-Scores)")
plt.ylabel("Standard Deviations from Mean")
plt.tight_layout()

plt.savefig(os.path.join(OUTPUT_FOLDER, "stacked_standardized_boxplot.png"))
plt.close()
```

A.8 Multivariable Predictive Modeling of Tourism Revenue Recovery

Purpose:

The purpose of this project was to construct a predictive model that quantifies how changes in affordability and tourism volume jointly influenced tourism revenue recovery across countries following the COVID-19 pandemic.

Code Description:

This script builds a multiple linear regression model using percentage change in tourism receipts as the dependent variable and percentage changes in purchasing power parity and tourism arrivals as explanatory variables. The model is trained using country-level data from 2019 to 2022, and model performance is evaluated using the coefficient of determination (R^2) and residual diagnostics. Predicted values, residual plots, and the final regression equation are programmatically generated and exported to ensure transparency, reproducibility, and clear documentation of model results.

```python
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import os

INPUT_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\final_analysis_ready_dataset.csv"
OUTPUT_FOLDER = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\Project 8 - Predictive Model"

os.makedirs(OUTPUT_FOLDER, exist_ok=True)

df = pd.read_csv(INPUT_FILE, encoding="latin1")

X = df[["PPP_Change_Pct", "Arrivals_Change_Pct"]]
y = df["Receipts_Change_Pct"]

model = LinearRegression()
model.fit(X, y)

df["Receipts_Predicted"] = model.predict(X)

r2 = r2_score(y, df["Receipts_Predicted"])

coef_ppp = model.coef_[0]
coef_arrivals = model.coef_[1]
intercept = model.intercept_

print("----- MODEL RESULTS -----")
print(f"Intercept: {intercept:.4f}")
print(f"PPP Coefficient: {coef_ppp:.4f}")
print(f"Arrivals Coefficient: {coef_arrivals:.4f}")
print(f"R-squared: {r2:.3f}")

results_text = (
    "Multiple Linear Regression Results\n"
```

```python
        "---------------------------------\n"
        f"Intercept: {intercept:.4f}\n"
        f"PPP Coefficient: {coef_ppp:.4f}\n"
        f"Arrivals Coefficient: {coef_arrivals:.4f}\n"
        f"R-squared: {r2:.3f}\n\n"
        "Model Equation:\n"
        "----------------\n"
        "Receipts % Change =\n"
        f"{intercept:.4f}\n"
        f"+ ({coef_ppp:.4f} × PPP % Change)\n"
        f"+ ({coef_arrivals:.4f} × Arrivals % Change)\n"
    )

    RESULTS_FILE = os.path.join(OUTPUT_FOLDER, "model_results.txt")

    with open(RESULTS_FILE, "w") as file:
        file.write(results_text)

    print("☑ Model results and equation saved to model_results.txt")

    plt.figure()
    plt.scatter(y, df["Receipts_Predicted"])
    plt.plot(
        [y.min(), y.max()],
        [y.min(), y.max()]
    )
    plt.xlabel("Actual Receipts % Change")
    plt.ylabel("Predicted Receipts % Change")
    plt.title("Actual vs Predicted Tourism Receipts Change")

    plt.savefig(os.path.join(OUTPUT_FOLDER, "actual_vs_predicted.png"), dpi=300)
    plt.close()

    df["Residuals"] = y - df["Receipts_Predicted"]

    plt.figure()
    plt.scatter(df["Receipts_Predicted"], df["Residuals"])
    plt.axhline(0)
    plt.xlabel("Predicted Receipts % Change")
    plt.ylabel("Residuals")
    plt.title("Residuals vs Predicted Values")

    plt.savefig(os.path.join(OUTPUT_FOLDER, "residuals_plot.png"), dpi=300)
    plt.close()
```

A.9 Data Preparation for Absolute Affordfability Analysis

Purpose:

The purpose of this project was to restructure an earlier-stage dataset to support a new analytical direction examining absolute affordability as a factor in post-pandemic tourism recovery.

Code Description:

Using the wide-format dataset produced in Project 4, this script isolates only the variables required for absolute affordability analysis, including tourism arrivals and receipts for 2019 and 2022 and purchasing power parity values for 2022. This step intentionally reverses portions of the transformation performed in Project 5, where raw year-specific values were removed in favor of change metrics, to preserve PPP as a level-based variable. All selected fields are converted to numeric data types, countries with incomplete data are excluded, and percentage change variables for tourism arrivals and tourism receipts are recalculated. The resulting dataset is exported as a dedicated file used as the input for the absolute affordability analysis conducted in Project 10.

```python
import pandas as pd

INPUT_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\wide_tourism_ppp_data.csv"

df = pd.read_csv(INPUT_FILE, encoding="latin1")

required_columns = [
    "Country_std",
    "Region",
    "Arrivals_2019",
    "Arrivals_2022",
    "Receipts_2019",
    "Receipts_2022",
    "PPP_2022"
]

df = df[required_columns]

numeric_cols = [
    "Arrivals_2019", "Arrivals_2022",
    "Receipts_2019", "Receipts_2022",
    "PPP_2022"
]

for col in numeric_cols:
    df[col] = pd.to_numeric(df[col], errors="coerce")


df_clean = df.dropna().copy()


df_clean["Arrivals_Change_Pct"] = (
    (df_clean["Arrivals_2022"] - df_clean["Arrivals_2019"])
    / df_clean["Arrivals_2019"]
)

df_clean["Receipts_Change_Pct"] = (
    (df_clean["Receipts_2022"] - df_clean["Receipts_2019"])
    / df_clean["Receipts_2019"]
)


df_clean = df_clean.sort_values("Country_std")


OUTPUT_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\project_9_affordability_clean_data.csv"
df_clean.to_csv(OUTPUT_FILE, index=False)
```

A.10 Absolute Affordability Dataset Preparation

Purpose:

The purpose of this project was to evaluate whether absolute destination affordability levels in 2022 influenced post-pandemic tourism recovery, independent of affordability changes over time.

Code Description:

Using the analysis-ready dataset produced in Project 9, this script performs bivariate analysis to examine the relationship between Purchasing Power Parity levels in 2022 and post-pandemic tourism recovery outcomes. Scatter plots are generated comparing PPP (2022) against percentage changes in international tourism arrivals and tourism receipts between 2019 and 2022. For each visualization, a linear regression line is fitted using Python's Scikit-learn library, and the coefficient of determination (R^2) is calculated and displayed directly on the plot to quantify explanatory strength. All graphs are saved as image files to ensure reproducibility and consistent presentation. This project formally tests whether destination cost levels alone influenced tourism recovery patterns following COVID-19 and provides a direct empirical contrast to the relative affordability analysis conducted earlier.

```python
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn.linear_model import LinearRegression
from matplotlib.ticker import PercentFormatter
import os

INPUT_FILE = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\project_9_affordability_clean_data.csv"
df = pd.read_csv(INPUT_FILE, encoding="latin1")

OUTPUT_FOLDER = "C:\\Users\\lucas\\OneDrive\\Desktop\\25-26 TSA Project\\Project 10 – Affordability Analysis"
os.makedirs(OUTPUT_FOLDER, exist_ok=True)

def scatter_with_regression(x, y, x_label, y_label, title, filename):
    X = x.values.reshape(-1, 1)
    Y = y.values

    model = LinearRegression()
    model.fit(X, Y)
    y_pred = model.predict(X)

    r2 = model.score(X, Y)

    plt.figure(figsize=(8, 6))
    plt.scatter(X, Y, alpha=0.7)
    plt.plot(X, y_pred)

    plt.xlabel(x_label)
    plt.ylabel(y_label)
    plt.title(title)
    plt.gca().yaxis.set_major_formatter(PercentFormatter(1.0))

    plt.text(
        0.05, 0.95,
        f"R² = {r2:.3f}",
        transform=plt.gca().transAxes,
        verticalalignment="top"
    )

    plt.tight_layout()
    plt.savefig(os.path.join(OUTPUT_FOLDER, filename))
    plt.close()

scatter_with_regression(
    x=df["PPP_2022"],
    y=df["Arrivals_Change_Pct"],
    x_label="Purchasing Power Parity (2022)",
    y_label="Tourism Arrivals % Change (2019–2022)",
    title="Affordability vs Tourism Arrivals Recovery",
    filename="ppp_vs_arrivals_change.png"
)

scatter_with_regression(
    x=df["PPP_2022"],
    y=df["Receipts_Change_Pct"],
    x_label="Purchasing Power Parity (2022)",
    y_label="Tourism Receipts % Change (2019–2022)",
    title="Affordability vs Tourism Revenue Recovery",
    filename="ppp_vs_receipts_change.png"
)
```