[1]Investigating the Impact of Political Rhetoric on U.S.-Canada Trip Travel Patterns

Sehar Bajwa *Francesca Ye* *Catherine Punnoose*

Abstract

This report aims to investigate the impact of political rhetoric on the U.S.-Canada trip travel patterns and volume. With the recent statements made by President Donald Trump about Canada and the Canadian government, the report will examine travel patterns between the two countries to determine if they are affected by political statements through analysis of travel data and trends across citizens of both countries between the period of January 2000 to January 2025. By understanding current trends and developing an interrupted time-series model for analysis, the report can identify changes in travel counts between Canada and the United States. Broadly, the report asks: "**Can a few words from a U.S. president really change cross-border travel patterns?**" However, the report specifically aims to answer, "**Does hostile international policy have an effect on the travel plans of countries being targeted**?"

I. Introduction

Since the start of Donald J. Trump's second term, a majority of Canadian media outlets find themselves covering American-Canadian relations due to Trump's statements about the potential annexation of Canada and looming tariffs on Canadian goods. A particular industry of concern as highlighted by mainstream media is the tourism sector for both countries as outlets have picked up stories of "snowbird" travellers and other types of Canadian tourists entering the United States as hesitant to visit due to the breakdown of economic relations. (Delouya, 2025) Despite Trump only taking office in January 2025, keen interest and concern have been echoed in discussions of the United States' tourism economy seeing a drop in Canadian visitors due to the tariffs. (Buckley, 2025) Industry organizations and monitors, such as OAG Aviation Worldwide, have noted that passenger flight bookings between Canada and the United States are currently down by 70% in comparison to the same period in 2024. (Grant, 2025) Furthermore, stricter enforcement of border security and threats of annexation towards all non-Americans has also been identified as major motivating factors as to why prospective tourists from across the world are cancelling their travel plans to the United States for alternative locations. (Galloway, 2025) Considering the magnitude and rapidness of the United States' breakdown of international relations with various countries, and in particular, Canada, leaders in the tourism industry have spotlighted these actions as motivating factors for an impending drop through these media outlets.

Throughout the development of these tariffs, the trade relationship between the United States and Canada, Trump has contended that the former subsidizes the latter at a value of US$200 billion annually. However, from an American perspective, the trade deficit with Canada is around US$45 billion (or C$65 billion at the exchange rate), a much smaller figure relative to their Gross Domestic Product at -0.2%. (Ercolao and Foran, 2025) By examining the tourism industry, the monthly number of Canadians returning to Canada from the United States for tourism-related purposes and Americans entering Canada for tourism-related purposes can reveal key industries where the United States have an advantage, Historical data

[1] All data files available at: https://github.com/SEHB2012/TrumpTravel

can also be used to predict future tourism numbers in both directions with models. By developing a model based on historical travel data, these predicted values can be later held against observed values in the future to assess whether claims Trump's policies have had an effect on the travel habits of Canadians.

II. Review of similar research

The article "*Tourist Arrival Forecasting in Sri Lanka: A Bayesian Spline and Interrupted Regression Approach*" written by A.W.L.P. Thilan focuses on using statistical methods to model the complex time series specifically in tourism data and the impact of the COVID-19 Pandemic (Thilan, 2025). The article focuses on the tourism industry in Sri-Lanka. It primarily used the Interrupted regression and Bayesian spline regression to find patterns in tourist arrivals both during and after the pandemic. The study used Interrupted Time Series (ITS) analysis to determine whether the pandemic caused significant changes in tourism while accounting for pre-existing patterns. Results showed that there is a general recovery in tourist arrivals but uncertainty whether the tourism industry will return to pre-pandemic levels or if new patterns will emerge (Thilan, 2025). The interrupted regression approach captured the pandemic's immediate impact on tourism, while Bayesian spline regression identified non-linear effects and long-term recovery trends (Thilan, 2025). This type of analysis relates to the U.S.-Canada relations under President Trump's second term, where the tourism industry has been directly impacted by concerns over tariffs, economic tensions, and border policies. Hence, ITS analysis can be an effective tool for evaluating if policy changes are directly responsible for policy changes. The analysis provides an insight to understand if these effects will persist or stabilize over time. This approach could be valuable in forecasting future trends within the tourism sector.

The "Impact of the COVID-19 pandemic on inbound air travel to Canada" is a study that was published in the *Canada Communicable Disease Report* in April 2024. (Gabriele-Rivet et al., 2024) This study aimed to examine how commercial air travel to Canada was affected by COVID-19-related travel restrictions. Using data from the International Air Transport Association (IATA) spanning March 2017 to February 2023, researchers analyzed travel trends before and during the pandemic. They employed seasonal autoregressive integrated moving average (SARIMA) models and interrupted time series (ITS) analysis to assess the impact of travel restrictions on passenger volumes. Their findings indicated that pre-pandemic inbound travel increased steadily, with seasonal peaks. However, at the onset of COVID-19, travel volumes dropped by 90% and a slow recovery followed. Most notably, their ITS analysis confirmed that major travel restrictions in March 2020 led to an immediate decline, while easing restrictions, starting in August 2021, coincided with a rise in travel volume. (Gabriele-Rivet et al., 2024) This study justifies the use of an interrupted time series (ITS) analysis for assessing the impact of policy by demonstrating its effectiveness in detecting and quantifying shifts in travel patterns due to COVID-19 restrictions. By outlining how ITS effectively captures shifts in travel behavior due to external shocks, such as public health interventions, the study supports the application of ITS in other policy contexts, such as assessing the impact of new visa policies, border security measures, or international relations on travel.

"Testing the efficacy of the economic policy uncertainty index on tourism demand in USMCA: Theory and evidence" is a 2020 study that examined how economic policy uncertainty (EPU) affects the number of tourists traveling from Canada and Mexico to the United States. (Işık et al., 2020) The researchers sourced data from the Federal Reserve Bank of St. Louis and the National Travel and Tourism Office to cover a period spanning January 1996 to September 2017 to determine whether uncertainty in economic policies influences tourism demand beyond traditional factors like income levels and exchange rates. A consumer demand model based on Marshallian theory, which generally considers income and price as key determinants of demand, was developed. An EPU index as a factor and regression and error correction models (ECM) were added to this model in order to study short and long-term effects. Their results showed that higher economic policy uncertainty leads to fewer

tourist arrivals in the United States, with Canadian tourists being more affected than Mexican tourists. (Işık et al., 2020) Although this article does not utilize ITS analysis, it does spotlight how economic policy uncertainty disrupts tourism demand. Therefore, it does support this report's notion that economic and policy factors in addition to their relationship can have an effect on tourism demand from Canadians for the United States, making it a strong methodological precedent for studying the impact of government decisions on tourism.

The article, "*The relationship between joining a US free trade agreement and processed food sales, 2002–2016: a comparative interrupted time-series analysis*" is a study that explores whether joining a U.S. Federal Trade Agreement (FTA) impacts the sales of processed foods in associated countries (Cowling et al., 2020). The study highlights how joining an FTA has historically led to increased sales of ultra-processed foods, processed culinary ingredients, and baby food, demonstrating the significant economic influence of trade policies. The findings of the study show that joining the US FTA can cause negative changes to national dietary consumption as well as increase the population risk of non-communicable diseases (Cowling et al., 2020). This study relates to the topic of Trump's second-term tariffs and strained U.S.-Canada relations as it highlights how government trade policies can influence economic activity. Both cases show how changes in trade agreements can have significant economic consequences, which can affect both the food industry and tourism industry.

III. Data

III.1. Data Profile

The report uses data collected by the Canada Border Services Agency (CBSA). The data is collected through the Frontier Counts program, which counts international travel entries into Canada by air, land, and water (Government of Canada, 2025). The dataset, "International travellers entering or returning to Canada, by type of transportation and traveller type", has collected data since 1972 and is updated quarterly to an online dashboard. The dataset is available on Statistics Canada and this report uses a version released on March 21, 2025.

The data is collected monthly. The CBSA collects data during the month, the data is then processed by Statistics Canada in the following month (Government of Canada, 2025). The target population of the dataset are all international travellers who enter Canada by air, land, or water (Government of Canada, 2025). Due to the large volume, the target population is distributed into the following five categories (2025):

- Canadian residents returning to Canada from the United States of America only
- Canadian residents returning to Canada from countries other than the United States of America,
- United States of America residents entering Canada
- Residents of countries other than the United States of America entering Canada
- "Other" travellers entering Canada, (foreign and resident crew members, diplomats, military personnel, immigrants and former residents) (Government of Canada, 2025)

Data sources for the dataset include Primary Inspection Kiosk (PIK), E311 Declaration Card, NEXUS, Telephone Reporting Centre (TRC)-CANPASS, Integrated Primary Inspection Line (IPIL), E-62 Entry Tally, E63 Commercial and Private Craft/Passenger and Crew Arrivals, and the E63-1 Passenger and Crew Arrivals, Cruise Vessel, Overseas Summary Report. The data undergo monthly quality checks, the dataset is compared with the previous month's set to ensure any errors, inconsistencies, or mistakes are addressed (Government of Canada, 2025). International entries by water are determined by a

combination of data from the water component of TRC-CANPASS, and marine data from E63 or E63-1 forms (Government of Canada, 2025).

III.1.1. Target Population: Air Entry Data

Data related to air entries include international commercial air visitors or returning visitors with a declaration by a Primary Inspection Kiosk (PIK) (Government of Canada, 2025). The data does not include the following as stated by the Government of Canada (2025):

- Visitors who arrived at an airport where the PIK system is not installed
- Those who did not make their declaration at a PIK (e.g., NEXUS travellers)
- Commercial crew members
- Individuals who declared the purpose of their trip as immigration to Canada
- Individuals who declared the purpose of their trip as work (requiring a permit) (Government of Canada, 2025)

International entries by air are determined by a combination of PIK, E311 declaration cards for flights, the air component of NEXUS and TRC-CANPASS, and E63 forms received (Government of Canada, 2025).

III.1.2. Target Population: Land Entry Data

Data related to land entries include international or returning visitors who enter Canada through land ports that are equipped with an automated Integrated Primary Inspection Line (IPIL) system in automobiles, motorcycles or other land vehicles and is licensed in Canada or the United States (Government of Canada, 2025). The data is determined by a combination of the IPIL, land component of NEXUS and TRC-CANPASS, E311 declaration cards for bus and train, E26 tallies, and overseas summary reports (Government of Canada, 2025). The data does not include visitors who used a system different from the IPIL such as NEXUS and E-62. It also excludes different modes of transportation such as truck, bus, or on foot as well as excludes any license plates which are not from the United States or Canada (Government of Canada, 2025).

III.1.3. Target Population: Water Entry Data

International entries by water are determined by a combination of data from the water component of TRC-CANPASS, and marine data from E63 or E63-1 forms (Government of Canada, 2025).

III.2. Variables

The following is a table of all variables referenced throughout this report as provided by the Statistics Canada dataset.

Variable	Raw Type	Description
Traveller type	String • Traveller • Excursionist • Tourist	Type of tourist entrance recorded
Reference period	String Dates formatted MM-YY or YY-MM	Month and year of the reporting period
United States of America residents entering Canada	String • Numerical count	Number of all American residents entering Canada in a given month
United States of America residents, air	String • Numerical count	Number of Americans entering Canada by air in a given month
United States of America residents, land	String • Numerical count	Number of Americans entering Canada by land in a given month
United States of America residents, water	String • Numerical count	Number of Americans entering Canada by water in a given month
Canadian residents returning from the United States of America	String • Numerical count	Number of all Canadians returning to Canada in a given month
Canadian residents returning from the United States of America, air	String • Numerical count	Number of Canadians returning to Canada by air in a given month
Canadian residents returning from the United States of America, land	String • Numerical count	Number of Canadians returning to Canada by land in a given month
Canadian residents returning from the United States of America, water	String • Numerical count	Number of Canadians returning to Canada by water in a given month

III.3. Data Cleaning

Prior to downloading the dataset, Statistics Canada's website allowed for the table in which they present the data to be customized for viewing and downloading purposes. (2025) Using this function, traveller counts for all crew and persons of non-tourism-related purposes entering or returning to Canada were removed as these groups were outside of the report's scope. The reference period was also set to January 2000 as the start and January 2025 as the end of the table's displayed data. *Traveller type* and *Reference period* were set as rows in the table whilst *Geography* and *Traveller characteristics* were set as columns. (Statistics Canada, 2025) This data table was downloaded as a CSV and further cleaning was conducted using Python.

In Python, blank rows and additional dataset context exported by Statistics Canada were discarded from the file. These *Traveller type* columns were deleted from the file: *International*

travellers entering or returning to Canada, Residents of countries other than the United States of America, air, Residents of countries other than the United States of America, land, Residents of countries other than the United States of America, water, Canadian-resident visitors returning to Canada. The dates given in the *Reference period* column were reformatted into a readable type. All columns containing traveller counts also had the data converted from a *string* to *float* type. All cells in *Traveller type* were then filled with the appropriate value for efficient data processing. These changes were saved to a new CSV file that was used for the analysis of this report.

III.4. Data Insights

To have a general understanding of the dataset multiple visualizations were used. A line chart was used for Figures 2, 3, and 4 as it visually displays the total values of each year/month. This is the most appropriate method to use to view trends over a period of time. *Figure 1* displays trends of both U.S. and Canadian residents entering Canada and their method of entry. To understand the most dominant form of entry for both U.S. residents entering Canada and Canadian residents returning to Canada, a pie chart was used. A pie chart is the most optimal method of visualization as it visually shows the proportions of travellers by air, land, and water. *Figure 2* visually displays the total number of both U.S. and Canadian travellers who entered Canada by either air, land, or water. To view seasonal trends by each entry method, a line chart is used. *Figure 3* shows the trends of air, land, and water entries by both U.S. residents and Canadian residents. To determine the average monthly trends of U.S. residents entering Canada and Canadian residents returning to Canada, a line chart is used. *Figure 4* shows the average number of U.S. residents entering Canada and Canadian residents returning to Canada each month within the past 20 years. A line graph was used to view the annual trends of the U.S. Residents Entering Canada & Canadian Residents Returning from 2000-2025. *Figure 5* visually displays the trends, highlighting significant dips and peaks within the past 20 years.

III.4.1. Comparison of overall travel volumes between US and Canada

Figure 1 displays the Distribution of Entry Methods for U.S. residents entering Canada and Canadian Residents returning from the U.S. from 2000-2025. The graph shows that land travel is the most dominant form of entry between the two countries. The graph also shows that the U.S. residents enter Canada by water more than Canadian residents.

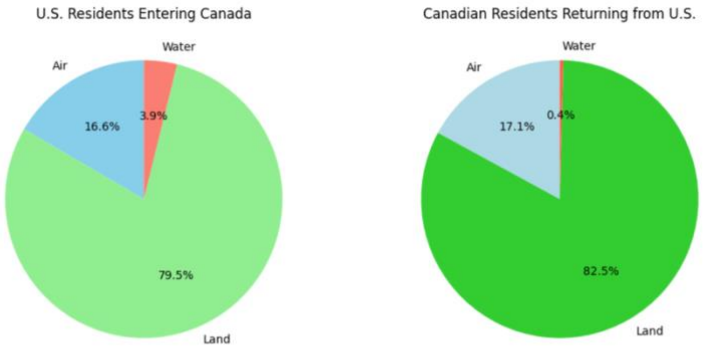


Figure 1: Distribution of Entry Methods for U.S Residents Entering Canada and Canadian Residents Returning from the U.S. from 2000-2025

III.4.2. Seasonal Trends

Figure 2 displays the entry methods by air, land, and water from 2000-2025. One key observation from the graph is how dominant land entry is for both the U.S. and Canada. This indicates that most cross-border travel happens through land. The graph also has multiple peaks. This corresponds to seasonal travel spikes from the summer and holiday season. Another observation is the decline in land travel over time. Although both the U.S. and Canada have a decline in land travel, the U.S. has had a steeper decline following the 2008 recession.

The graph shows that there was a significant decline in all forms of travel in 2020 due to the COVID-19 pandemic. This is likely due to the travel restrictions placed during the pandemic. The graph shows a slow recovery post-pandemic, from 2021 onwards. However, the travel count has not reached pre-pandemic levels. One interesting observation is the gradual increase of air travel. From 2000-2020, air travel for both the U.S. and Canada gradually increased until 2020. However, post-pandemic, air travel rates have almost reached pre-pandemic rates. This shows that air travel is becoming more important post-pandemic. Water travel rates are very low and are insignificant in total cross-border movement. Overall, the graph shows that there are no major dips between the two countries related to major political events.

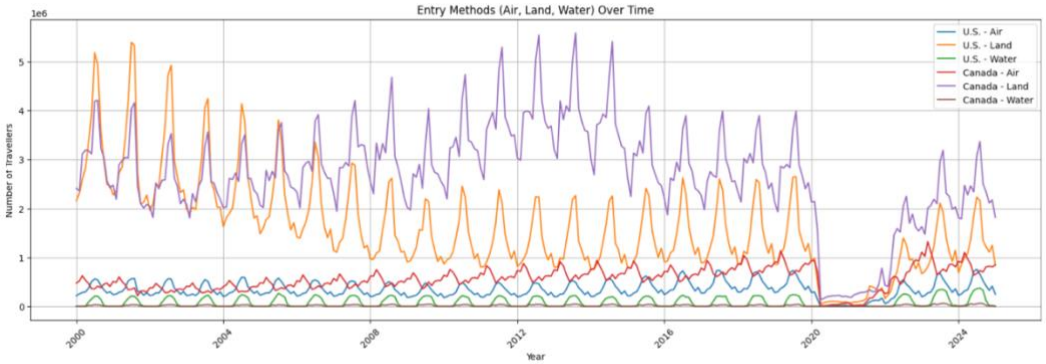


Figure 2: Entry Methods from 2000-2025

III.4.3. Monthly Trends

Figure 3 shows the average monthly trends between United States of America residents entering Canada and Canadian residents returning from the United States of America from 2000-2025. A key observation in the graph is the spike between June to August for both U.S. and Canadian residents. This indicates that most residents (U.S. and Canada) enter Canada during the summer months. This is likely due to the summer weather and summer break period for students. The sharp decline from August to September for both countries indicates that many travellers return to their home countries during the school season.

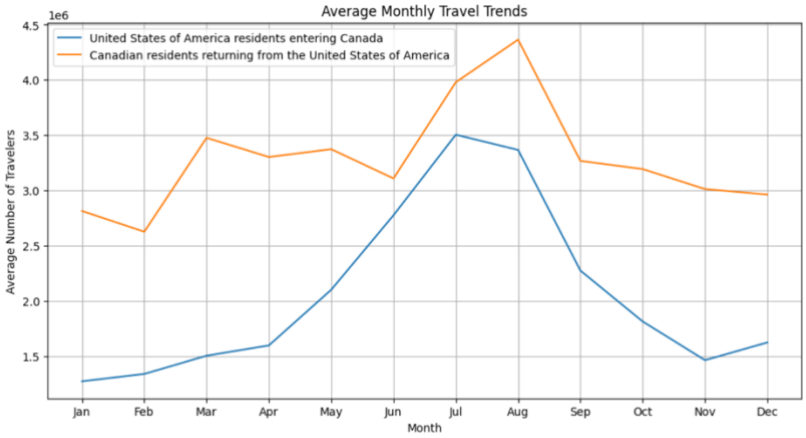


Figure 3: Average Monthly Trends from 2000-2025

III.4.4. Annual Trends

Figure 4 shows the annual trends of the U.S. Residents Entering Canada & Canadian Residents Returning from 2000-2025. The graph visually displays the trends, highlighting significant dips and peaks within the past 20 years. Similar to *Figure 1*, the graph shows significant peaks during the summer months. The graph also shows a sharp drop in 2020, due to the COVID-19 pandemic.

Between 2008-2016, there is a significant gap between the number of Canadian Residents Returning to Canada and the the U.S. Residents Entering Canada. This is likely due to the 2008 recession, where travel declined. This shows that less U.S. residents entered Canada at this time, while many Canadians returned. Looking at 2021-2025, there is a slow increase of travel, however it has not reached pre-pandemic levels.

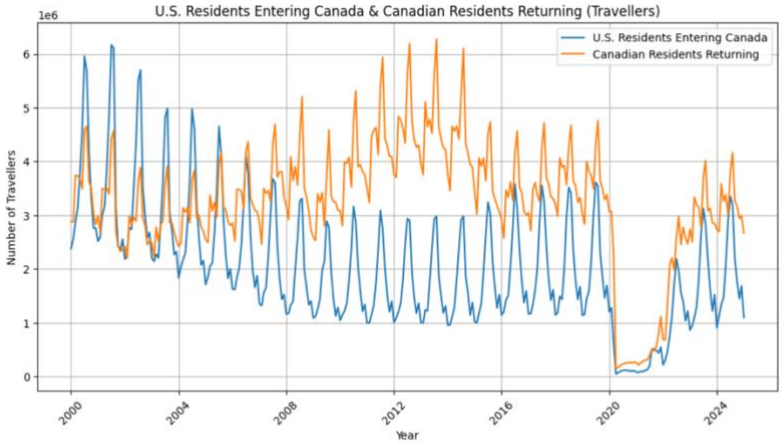


Figure 4: Trend of U.S. Residents Entering Canada & Canadian Residents Returning (Travellers) from 2000-2025

IV. Statistical Testing – Methodology and results

Prior to developing a model for tourism rates across Canadians and Americans as influenced by policy changes, other dimensions of the dataset were explored using a variety of statistical tests to uncover potential connections to the research question outside of modelling. To gain a greater understanding of the relationship between the variables present in the dataset, a variety of t-tests, ANOVA, and two-way ANOVA analysis was performed to answer the following questions:

IV.1. Is there a significant difference between Canadians returning from the United States and Americans entering Canada?

To understand if there is a statistically significant difference between Canadians returning from the United States and Americans entering Canada, a two-tail t-test was selected as the optimal statistical test to provide an analytical answer to the question posed. This is because a two-tailed t-test would allow the ability to discern a significant difference in either direction between two independent groups if it exists. It also checks for deviations in either direction, ensuring that any observed difference is not due to random chance (De Veaux et al., 2021). To conduct this test, the dataset was filtered to look at the number of travellers for both groups without accounting for tourist or excursionist status. After, this data was aggregated by year for both groups and the two-tail t-test was conducted. For the purposes of this test, the hypothesis structure was:

> **Null Hypothesis (H_o):** There is no significant difference between Canadians returning from the United States and Americans entering Canada
>
> **Alternative Hypothesis (H_1):** There is a significant difference between Canadians returning from the United States and Americans entering Canada.

Mean (U.S. residents entering Canada)	23733806.576923076
Mean (Canadian residents returning from the U.S.)	38069855.84615385
T-statistic	-4.199588338025029
P-value	0.00010982799469476195
Significance	There is a significant difference between Americans entering Canada and Canadian residents returning to the United States.

These two-tailed t-test results reveal a significant difference between the number of American residents entering Canada and the number of Canadian residents returning from the United States. On average, 23,733,807 U.S. residents travel to Canada, whereas 38,069,856 Canadian residents return from the United States. The t-test results indicate a t-statistic of -4.20 and a p-value of 0.0001, which is well below the conventional significance threshold of 0.05. This suggests that the observed difference is statistically significant and unlikely to have occurred by chance. The negative t-statistic reflects that the mean number of Canadians returning home is substantially higher than the number of Americans entering Canada. This asymmetry may be influenced by factors such as travel patterns, border policies, or economic and cultural ties between the two countries.

IV.2. Do more Canadian excursionists (short-term) return from the United States than tourists (long-term)? Do more American excursionists (short-term) return to Canada than tourists (long-term)?

To examine whether there is a significant difference between the number of excursionists and tourists for both Canadians and Americans, two-tail t-tests were also conducted on both groups. This test was selected since the travel data presents this as a binary choice which fits into the structure of the t-test and helps discern if a significant difference between these independent variables exists in either direction. (De Veaux et al., 2021) The P-value of the t-test also helps inform interpretation as it provides an idea of the possibility of observed difference occurring due to random chance.

To conduct these tests, the cleaned dataset was filtered for only rows in *Traveller type* that represented excursionists or tourists. Variables were then created by separating the differing rows and filtering for data from Canadian residents returning from the United States. This data was then aggregated by year and the t-test was run. For Americans entering Canada, the same filtering process was conducted but only looking at data for their group.

For the purposes of these tests, the hypothesis structure was:

Canadians returning from the United States

Null Hypothesis (H_0): There is no significant difference between the number of Canadian excursionists and tourists returning from the United States

Alternative Hypothesis (H_1): There is a significant difference between the number of Canadian excursionists and tourists returning from the United States

Americans entering Canada

Null Hypothesis (H_0): There is no significant difference between the number of American excursionists and tourists entering Canada

Alternative Hypothesis (H_1): There is a significant difference between the number of American excursionists and tourists entering Canada

Mean (Canadian Excursionists Returning from the U.S)	21514877.384615384
Mean (Canadian Tourists Returning from the U.S.)	16554978.461538462
T-statistic	2.4735695188291027
P-value	0.016817766903569054
Significance	There is a significant difference between Excursionists and Tourists for Canadian residents returning from the United States

The two-tailed t-test results show a significant difference between Canadian excursionists and Canadian tourists returning from the United States. The t-test results yield a t-statistic of 2.47 and a p-value of 0.0168, which is below the significance threshold of 0.05, indicating that the difference between

these two groups is statistically significant. This suggests that Canadian excursionists are returning from the United States in greater numbers than Canadian tourists. The distinction between excursionists and tourists likely reflects different travel behaviors or motivations, such as the prevalence of cross-border shopping and day trips amongst Canadians.

Mean (American Excursionists Entering Canada)	11763477.692307692
Mean (American Tourists Entering Canada)	11970328.884615384
T-statistic	-0.12462204121107444
P-value	0.9013224673219169
Significance	There is no significant difference between Excursionists and Tourists for Americans entering Canada.

These two-tailed t-test results indicate no significant difference between American excursionists and American tourists entering Canada. The t-test results show a t-statistic of -0.12 and a p-value of 0.9013, which is higher than the conventional significance threshold of 0.05. This suggests that the difference between the two groups is not statistically significant, meaning the number of excursionists and tourists entering Canada is essentially the same. This lack of significant difference could imply that both groups share similar travel patterns or that the distinction between excursionists and tourists does not strongly affect the frequency of their visits to Canada.

IV.3. Do travel patterns differ between excursionists and tourists for Canadians returning from the United States and Americans entering Canada?

To understand if there were differences between these groups, a two-way ANOVA was conducted as it allowed for analysis of the effects of two independent categorical variables on a dependent variable. two-way ANOVA not only tests the individual effects of traveler type and direction of travel but also determines if there is an interaction effect. (De Veaux et al., 2021) Specifically, it would help determine whether the difference in travel patterns between excursionists and tourists depends on whether they are Canadians returning or Americans entering. A two-way ANOVA was also selected because it reduces the risk of Type I errors, which is more common in running multiple t-tests, by examining differences between all possible group combinations in one test.

The data was cleaned by filtering for 'Excursionists' and 'Tourists', converting the 'Reference period' to datetime, extracting the year, and aggregating traveler counts for Canadians returning from the U.S. and Americans entering Canada. The dataset was then reshaped for a two-way ANOVA, which tested whether travel patterns significantly differed by traveler type and travel direction, with results interpreted based on the P-value.

For the purpose of this test, the hypothesis structure was:

<u>Main Effects:</u>

Does traveler type (excursionists vs. tourists) affect travel patterns?
Does direction of travel (Canadians returning vs. Americans entering) affect travel patterns?

Interaction Effect:

Does the effect of traveler type depend on the direction of travel?

F-statistic	2.5328236401540507
P-value	0.11459510959507835
Significance	There is no significant difference between Excursionists and Tourists for Canadian residents returning from the U.S. and U.S. residents entering Canada.

The two-way ANOVA results indicate that there is no significant difference between excursionists and tourists for both Canadian residents returning from the U.S. and U.S. residents entering Canada. The F-statistic is 2.53, and the p-value is 0.1146, which exceeds the conventional significance threshold of 0.05. This means that the variation observed between the groups of excursionists and tourists is not statistically significant. Therefore, the type of traveler, excursionist or tourist, does not appear to influence the number of Canadians returning from the United States or the number of Americans entering Canada, suggesting that other factors might be at play in determining travel patterns.

IV.4. Is there a significant difference in the number of Americans entering Canada by air, land, or water? Is there a significant difference in the number of Canadians returning from the United States by air, land, or water*?*

To examine differences in entry type for each of these groups, an ANOVA was selected because it allowed for the comparison of the means for more than two independent groups. ANOVA is appropriate for this context as analysis of three or more categories is required since performing multiple t-tests would increase the risk of Type I error. (De Veaux et al., 2021) The travel modes (air, land, and water) are also distinct independent groups which is a condition of ANOVA. This test also helps determine whether the mode of travel has a significant impact on the number of travelers, without assuming in advance which mode has the highest or lowest values.

The cleaned dataset was filtered to include only Americans entering Canada or Canadians returning from the United States depending. Travel counts for air, land, and water entries were aggregated annually. A one-way ANOVA was then conducted to test for significant differences in entry type, with results interpreted based on the P-value.

The hypothesis structure for each question was:

Americans entering Canada

Null Hypothesis (H_0): There is no significant difference in the number of Americans entering Canada by air, land, or water.

Alternative Hypothesis (H_1): There is a significant difference in the number of Americans entering Canada by air, land, or water.

Canadians returning from the United States

Null Hypothesis (H_0): There is no significant difference between the number of Canadian returning from the United States by air, land, or water

Alternative Hypothesis (H_1): There is a significant difference between the number of Canadian returning from the United States by air, land, or water

F-statistic	73.72523832725544
P-value	1.965835177129121e-18
Significance	There is a significant difference in annual visits by entry type (Air, Land, Water) for Americans entering Canada.

The ANOVA results indicate a significant difference in the number of annual visits by entry type for Americans entering Canada. The F-statistic is 73.73, and the p-value is 1.97×10^{-18}, which is well below the conventional significance threshold of 0.05. This suggests that the type of entry, whether by air, land, or water, has a statistically significant impact on the number of visits. The large F-statistic indicates that the variation between the groups is much greater than the variation within each group. This highlights that the mode of entry is an important factor influencing the frequency of visits by Americans to Canada.

F-statistic	142.88920455189194
P-value	2.619854417565931e-26
Significance	There is a significant difference in annual visits by entry type (Air, Land, Water) for Canadians returning from the United States.

The ANOVA results reveal a significant difference in the number of annual visits by entry type (Air, Land, Water) for Canadians returning from the United States. The F-statistic is 142.89, and the p-value is 2.62×10^{-26}, which is below the standard significance threshold of 0.05. These results indicate that the mode of entry, whether by air, land, or water, has a significant effect on the frequency of visits. The large F-statistic suggests that the differences between the entry types are much greater than the differences within each type, underscoring the importance of entry mode in shaping the patterns of Canadians returning from the United States.

V. Interrupted Time Series Analysis (ITS) Modelling

Following a thorough exploratory data analysis that examined correlations and trends amongst key variables, we chose to employ an Interrupted Time Series (ITS) analysis to investigate whether President Trump's comments and the surrounding political rhetoric have influenced cross-border travel volume between the United States and Canada in a significant manner. The ITS methodology is particularly suited for this type of analysis as it allows for examining data before and after a defined intervention, which in this case is the rise of politically charged rhetoric about Canadian annexation ("the 51st state") and coverage in news media regarding the statistics. This model will thus prove if the coverage numbers are inflated or part of an explainable shift and other factors . The ITS model incorporates both pre- and post-intervention periods to assess any significant shifts in the underlying trend of cross-border travel. The model is specified as:

$$Y_t = \beta_0 + \beta_1 \cdot Time_t + \beta_2 \cdot Intervention_t + \beta_3 \cdot (Time_t \times Intervention_t) + \epsilon_t$$

Where:

- Yt = Outcome at time t
- Time_t = Time trend before the intervention
- tIntervention$t = 0 before, 1 after the event (January 2025)
- Time$t×InterventiontTime_t = Change in trend after intervention
- ϵt = Error term

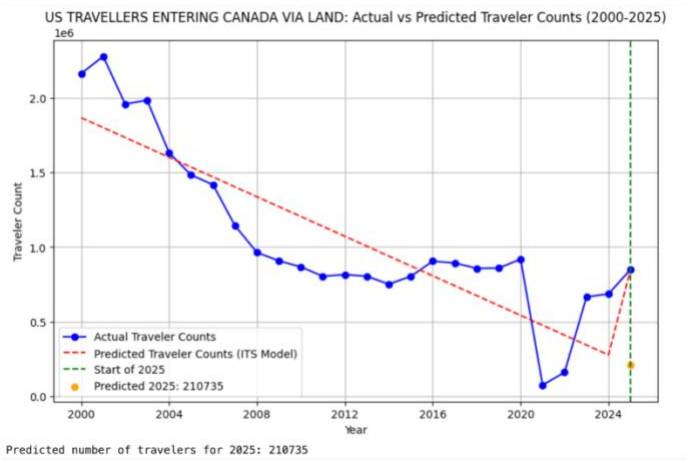
Rather than using data from the entire calendar year, we focused on January data for two primary reasons: first, to account for the cyclical nature of travel, which exhibits significant fluctuations throughout the year, and second, because January 2025 represents the most recent data available from Statistics Canada.

The model was then used to predict the value for 2025 based on pre-existing data, and examine how accurate it is, in addition to referring to relevant parameters. The standardized indicators to analyse the fit of the model are mentioned below:

- **Goodness of Fit**
  - R-squared: A value close to 1 indicates a better model fit.
- **Model Significance**
  - F-statistic: A value greater than 10 is generally considered good, but higher values indicate stronger model fit.
  - Prob (F-statistic): A p-value less than 0.05 suggests that at least one of the predictors in the model is statistically significant.
- **Predictor Significance**
  - Time: A p-value below 0.05 indicates that time (as a continuous variable) is statistically significant in explaining the dependent variable.
  - Intervention: A p-value below 0.05 suggests that the intervention had a statistically significant effect on the dependent variable at the point of intervention.
  - Post-Intervention Time: A p-value below 0.05 indicates that the trend after the intervention is significantly different from the pre-intervention trend.
- **Autocorrelation**
  - Durbin-Watson Statistic: A value close to 2 (typically between 1.5 and 2.5) indicates no significant autocorrelation in the residuals.
- **Normality of Residuals**

- Jarque-Bera Test: A p-value greater than 0.05 indicates that the residuals are likely normally distributed, ensuring that the assumptions of normality are met for regression analysis.

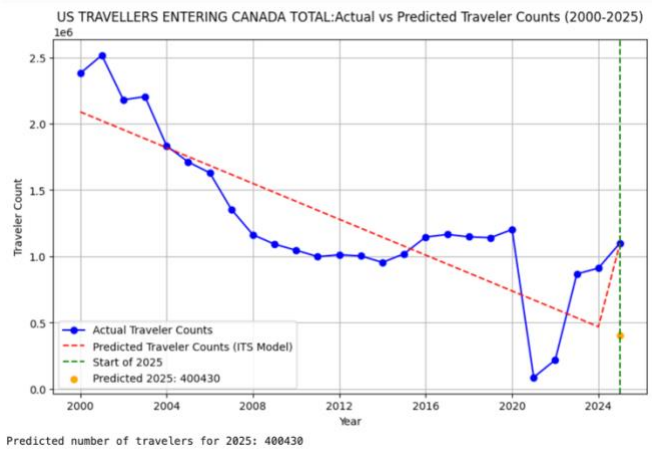
V.1. US travellers entering Canada – via land



R-squared	0.747
F-statistic	33.95
P value (F-statistic)	1.37e-07 = 1.37×10^{-7}
Coefficient for time	-6.62e+04 = -6.62×10^{4}
P value (**Coefficient for time)**	0.000
Coefficient for intervention	0.1599
P value (**Coefficient for intervention)**	0.054
Coefficient for post time	157.482
P value (**Coefficient for intervention)**	0.054
Durbin-Watson Statistic	0.747
Jarque Bera Coefficient	2.024

The model appears to be a good fit, as indicated by a high R-squared value of 0.747 and a highly significant F-statistic ($p < 0.001$). The time variable is a significant predictor ($p = 0.000$), and while the intervention and post-time coefficients are only marginally significant ($p = 0.054$), they still suggest some effect; however, the low Durbin-Watson statistic (0.747) indicates potential autocorrelation.

V.2. US travellers entering Canada – overall



R-squared	0.702
F-statistic	27.07
P value (F-statistic)	9.05e-07 = 9.05×10^{-7}
Coefficient for time	-6.754e+04 = -6.754×10^{4}
P value (**Coefficient for time)**	0.000
Coefficient for intervention	0.1752
P value (**Coefficient for intervention)**	0.064
Coefficient for post time	-350.0886
P value (**Coefficient for intervention)**	0.064
Durbin-Watson Statistic	0.881
Jarque Bera Coefficient	1.749

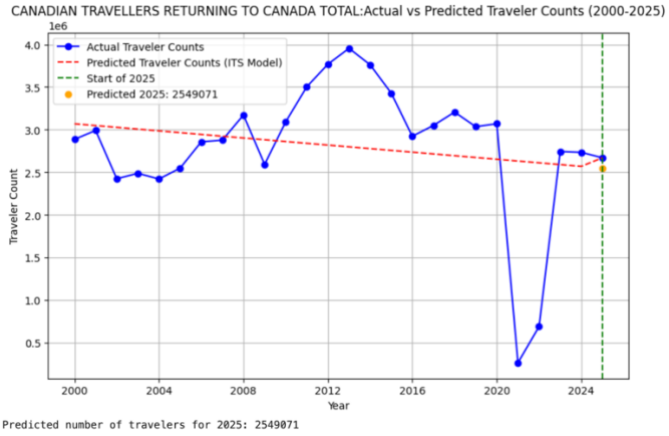
The model demonstrates a good fit with an R-squared value of 0.702 and a highly significant F-statistic ($p < 0.001$), indicating that the predictors jointly explain a substantial portion of the variation in travel. The time variable is a strong and significant predictor ($p = 0.000$), while the intervention and post-time effects are marginally significant ($p = 0.064$), and the Durbin-Watson statistic (0.881) suggests some positive autocorrelation.

V.3. Canadian travellers returning to Canada - Land

R-squared	0.160
F-statistic	2.196
P value (F-statistic)	0.134
Coefficient for time	-3.58e+04 = -3.58×10^4
P value (**Coefficient for time)**	0.055
Coefficient for intervention	0.0261
P value (**Coefficient for intervention)**	0.881
Coefficient for post time	-52.1371
P value (**Coefficient for post time)**	0.881
Durbin-Watson Statistic	0.715
Jarque Bera Coefficient	4.777

The model does not appear to be a strong fit, with a low R-squared value of 0.160 and a non-significant F-statistic (p = 0.134), suggesting the predictors do not explain much of the variation in travel. Additionally, all coefficients are statistically insignificant (p > 0.05), and the low Durbin-Watson value (0.715) indicates positive autocorrelation in the residuals.

V.4. Canadian travellers retuning to Canada – Overall



R-squared	0.036
F-statistic	0.4323
P value (F-statistic)	0.654
Coefficient for time	-2.08e+04 = -2.08×10^4
P value (**Coefficient for time)**	0.371
Coefficient for intervention	0.0304
P value (**Coefficient for intervention)**	0.892
Coefficient for post time	444.947
P value (**Coefficient for post time)**	0.892
Durbin-Watson Statistic	0.912
Jarque Bera Coefficient	14.262

The model does not provide a good fit, as indicated by the very low R-squared value (0.036) and a high p-value for the F-statistic (0.654), suggesting that the predictors do not significantly explain the variance in cross-border travel. All coefficients have high p-values (above 0.35), indicating that time, intervention, and post-time have no statistically significant effect in this model.

Overall, the US travelling volumes better fit the model and are more statistically significant, however the Canadian Models do a very good job at predicting the 2025 number of travellers, though one must not ignore that this could be due to overfitting.

VI. Discussion

VI.1. Limitations

Although this report aims to make a robust assessment on the effect of Donald Trump's international and economic policy on cross-border tourism between the United States and Canada, there are several limitations that must be acknowledged when considering the results and following interpretations produced. Broadly, it is important to highlight that there are a lot of confounding factors that can not be fully isolated from the travel dataset used in this report as currency rates, inflation, and preference towards alternate destinations outside of these two countries could also have an effect on individual decisions regarding travel. While mainstream news media have emphasized the magnitude of Trump's comments and actions towards Canada, it is possible that there are other factors that play a role in individual choices that are reflected in the overall group trends that were analyzed in this report.

In appraising the ITS model that was developed specifically, there is incomplete 2025 data to compare the model's predictions with. Since the travel counts for the full year are not out yet, it is not possible to make strong long-term conclusions about Trump's effect on cross-border tourism. This report contends that the model developed could be used in the future to compare observed monthly counts to predicted counts; however, that avenue is not possible at the current time. Additionally, COVID-19 and related travel restrictions have an effect on this data set which is reflected in the predictions made by the model due to its values being extreme outliers in monthly travel counts. Thus, it is difficult to isolate Trump's influence on travel as aforementioned, travel also still appears to be recovering from the effect of COVID-19. Focusing on the ITS model itself, there are also limitations in its accuracy of application to this scope and data set. The ITS model assumes that trends in the data are consistent over time, which may not hold true when seasonal fluctuations are present. Thereby potentially distorting the analysis. Additionally, the model assumes normally distributed errors, which is often an unrealistic assumption for real-world data. In application, errors may not follow a normal distribution, and it could lead to biased estimates. These limitations should be carefully considered when interpreting the results derived from the ITS model.

VI.2. Next Steps

There are also several next steps that could be taken to expand on the analysis conducted in this report to further assess the effect of Trump on cross-border travel. This includes expanding on the ITS model that was developed for this report by expanding the number of covariates. Specifically, this means to integrate macroeconomic indicators (e.g., exchange rates, inflation, GDP) and sentiment analysis from consumer travel surveys to control for confounding influences on travel plans. By integrating these covariates such as the aforementioned, the model can more accurately isolate the effect of the primary variables being studied. Thus, reducing the risk of omitted variable bias and improving the model's ability to capture the true relationships within the data. The ITS model could also benefit from robustness checks to validate the model and assess the stability of intervention effects. This can be done by conducting placebo tests using pre-2025 data to ensure that any observed effects are not simply due to underlying trends or random fluctuations. Placebo tests also test the validity of the intervention's impact to ensure that the changes observed post-intervention are attributable to the intervention itself, and not to other factors or modeling errors. Similarly, it is crucial to validate the data with Q3/Q4 2025 travel data when it becomes available to confirm trend persistence and refine post-intervention estimates. Moreover, it provides an opportunity to refine the model's post-intervention estimates based on more comprehensive data to improve the precision and accuracy of the conclusions. By taking these steps, the ITS model developed in this report could improve its validity and precision.

Outside of improving the ITS model, other models that better fit the data set could also be explored. As previously noted, the Seasonal Autoregressive Integrated Moving Average (SARIMA) model could be used for this data since it is specifically designed to handle seasonality in time-series data. Modeling the time series data using SARIMA enables the ability to capture both trend and seasonal patterns with adjustments for seasonal lags and periods. (Dagum, 1986) A Negative Binomial regression model would also be worth applying this particular dataset to since it is able to account for count data that exhibits overdispersion. (Hilbe, 2011) Based on preliminary work for this report's ITS model, the data set exhibited high overdispersion where the variance exceeded the mean and thus ruled out the possibility of a Poisson distribution. By fitting a Negative Binomial regression model to this data set, it would allow flexible variance structure and better predictions for overdispersed data. These models, or a hybrid model utilizing approaches from ITS, SARIMA, and Negative Binomial could be useful in accurately capturing the underlying dynamics of travel between Canadians and Americans. Thus, exploring different modeling approaches could also be suitable next steps to examining the research question outlined.

VII. References

Buckley, C. (2025, March 30). *'Catastrophic': Canadian bookings for U.S. travel drop sharply*. CTVNews. https://www.ctvnews.ca/canada/article/where-have-all-the-snowbirds-gone-canadians-cold-on-us-travel/

Cowling, K., Stuart, E. A., Neff, R. A., Vernick, J., Magraw, D., & Pollack Porter, K. (2020). The relationship between joining a US free trade agreement and processed food sales, 2002-2016: a comparative interrupted time-series analysis. *Public Health Nutrition*, *23*(9), 1609–1617. https://doi.org/10.1017/S1368980019003999

Dagum, E. B., & Canada. Statistics Canada. Methodology Branch. (1986). *Seasonal adjustment for forecasting*. Statistics Canada.

Delouya, S. (2025, March 30). *Canada's snowbirds reconsider calling the US their second home | CNN business*. CNN. https://www.cnn.com/2025/03/30/business/canada-snowbirds-trump-trade-war-tariffs/index.html

De Veaux, R. D. et al. (2021). *Stats: Data and Models 4th Canadian edition* (4th ed.). Pearson.

Ercolao, M., & Foran, A. (2025, January 21). *Setting the record straight on Canada-U.S. trade*. TD Canada Trust. https://economics.td.com/ca-canada-us-trade-balance

Gabriele-Rivet, V., Rees, E., Rahman, A., & Milwid, R. M. (2024). Impact of the COVID-19 pandemic on inbound air travel to Canada. *Canada Communicable Disease Report*, 50(3–4), 106–113. https://doi.org/10.14745/ccdr.v50i34a04

Galloway, L. (2025, April 1). *"a hostile state": Why some travellers are avoiding the US*. BBC News. https://www.bbc.com/travel/article/20250328-the-people-boycotting-travel-to-the-us

Grant, J. (2025, March 26). *Canada - US Aviation: Airlines Respond to Weakening Demand*. OAG. https://www.oag.com/blog/canada-us-airline-capacity-aviation-market

Government of Canada. (2025, March 7). *Frontier Counts (FC)*. Statistics Canada. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=1565966#a1

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge University Press.

Statistics Canada. International travellers entering or returning to Canada, by type of transportation and traveller type. (2025, March 21).

Işık, C., Sirakaya-Turk, E., & Ongan, S. (2020). Testing the efficacy of the economic policy uncertainty index on tourism demand in USMCA: Theory and evidence. *Tourism Economics : The Business and Finance of Tourism and Recreation*, 26(8), 1344–1357. https://doi.org/10.1177/1354816619888346

McDowall, D., McCleary, R., & Bartos, B. J. (2021). *Interrupted time series analysis*. Oxford University Press.

Thilan, A. W. L. P. (2025). Tourist Arrival Forecasting in Sri Lanka: A Bayesian Spline and Interrupted Regression Approach. *Sri Lankan Journal of Applied Statistics*, 26 (1), 46-71. https://doi.org/10.4038/sljas.v26i1.8157

VIII. Appendix

ITS Code

```
# -*- coding: utf-8 -*-
"""ITSanalysis.ipynb

Automatically generated by Colab.

Original file is located at
    https://colab.research.google.com/drive/1VHkRO0ilt2QRfES-ISkxPmGp7FPF-zNJ
"""

# US TRAVELLERS ENTERING CANADA VIA LAND: Actual vs Predicted Traveler Counts (2000-2025)
import pandas as pd
import statsmodels.api as sm
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
import seaborn as sns

 #Data for travelers from 2000-2025
data = {
   'Date': [
        '2000-01-01', '2001-01-01', '2002-01-01', '2003-01-01', '2004-01-01', '2005-01-01', '2006-01-01', '2007-01-01',
        '2008-01-01', '2009-01-01', '2010-01-01', '2011-01-01', '2012-01-01', '2013-01-01', '2014-01-01', '2015-01-01',
        '2016-01-01', '2017-01-01', '2018-01-01', '2019-01-01', '2020-01-01', '2021-01-01', '2022-01-01', '2023-01-01',
        '2024-01-01', '2025-01-01'
    ],
    'Travelers': [
        2161951, 2277514, 1958254, 1983565, 1632375, 1482636, 1417424, 1143393, 963554, 907468, 866690, 803987, 815191,
        804093, 751461, 803338, 904807, 893109, 857092, 858662, 918575, 74962, 160707, 666019, 685639, 849495
    ]
}

# Create the DataFrame
df = pd.DataFrame(data)

# Convert the Date column to datetime
df['Date'] = pd.to_datetime(df['Date'])

# Add time variables
df['time'] = np.arange(1, len(df) + 1)  # Sequential time (1, 2, 3, ..., 26)
df['intervention'] = np.where(df['Date'].dt.year >= 2025, 1, 0)  # Binary intervention variable for 2025
```

```
df['post_time'] = np.where(df['Date'].dt.year >= 2025, df['time'] - 2024, 0)  # Time since 2025

# OLS Regression: Y = β0 + β1 * time + β2 * intervention + β3 * post_time + ε
X = df[['time', 'intervention', 'post_time']]
X = sm.add_constant(X)  # Add intercept
y = df['Travelers']

# Fit the model
model = sm.OLS(y, X)
results = model.fit()

# Predict the values for the full dataset (including 2025)
df['predicted'] = results.predict(X)

# Plot the actual vs predicted data for 2025
plt.figure(figsize=(10, 6))
plt.plot(df['Date'], df['Travelers'], label='Actual Traveler Counts', color='b', marker='o')
plt.plot(df['Date'], df['predicted'], label='Predicted Traveler Counts (ITS Model)', color='r',
linestyle='--')

# Highlight the start of 2025
plt.axvline(pd.to_datetime('2025-01-01'), color='g', linestyle='--', label='Start of 2025')

plt.title('US TRAVELLERS ENTERING CANADA VIA LAND: Actual vs Predicted
Traveler Counts (2000-2025)')
plt.xlabel('Year')
plt.ylabel('Traveler Count')
plt.legend()
plt.grid(True)
plt.show()

# Create a new DataFrame to predict the value for 2025
new_data = pd.DataFrame({
    'const': [1],
    'time': [26],  # This is the time point for 2025
    'intervention': [1],  # For post-2025 period
    'post_time': [1]  # Time since 2025 (relative to the intervention)
})

# Predict for 2025 using the model
predicted_2025 = results.predict(new_data)[0]

# Add 2025 prediction to the plot
plt.figure(figsize=(10, 6))
plt.plot(df['Date'], df['Travelers'], label='Actual Traveler Counts', color='b', marker='o')
plt.plot(df['Date'], df['predicted'], label='Predicted Traveler Counts (ITS Model)', color='r',
linestyle='--')

# Highlight the start of 2025 and the predicted value for 2025
plt.axvline(pd.to_datetime('2025-01-01'), color='g', linestyle='--', label='Start of 2025')
```

```
plt.scatter(pd.to_datetime('2025-01-01'), predicted_2025, color='orange', label=f'Predicted 2025: {predicted_2025:.0f}')

plt.title('US TRAVELLERS ENTERING CANADA VIA LAND: Actual vs Predicted Traveler Counts (2000-2025)')
plt.xlabel('Year')
plt.ylabel('Traveler Count')
plt.legend()
plt.grid(True)
plt.show()

# Print the predicted value for 2025
print(f"Predicted number of travelers for 2025: {predicted_2025:.0f}")

# Print the regression results summary
print(results.summary())

# Diagnostic Checks
# 1. Residuals vs Fitted Plot
plt.scatter(df['predicted'], results.resid)
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Residuals vs Fitted Values')
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.grid(True)
plt.show()

# 2. Q-Q Plot for Normality of Residuals
sm.qqplot(results.resid, line ='45')
plt.title('Q-Q Plot of Residuals')
plt.show()

# 3. Levene's Test for Homogeneity of Variance
stat, p_value = stats.levene(df['Travelers'], df['predicted'])
print(f"Levene’s test statistic: {stat}, p-value: {p_value}")

# 4. Durbin-Watson Test for Autocorrelation
from statsmodels.stats.stattools import durbin_watson
dw_stat = durbin_watson(results.resid)
print(f'Durbin-Watson Statistic: {dw_stat}')

# 5. Influence Plot
sm.graphics.influence_plot(results)
plt.title('Influence Plot')
plt.show()

# 6. Histogram of Residuals
plt.hist(results.resid, bins=10, edgecolor='k')
plt.title('Histogram of Residuals')
plt.xlabel('Residuals')
```

```
plt.ylabel('Frequency')
plt.show()

# 7. Scale-Location Plot
standardized_residuals = results.get_influence().resid_studentized_internal
plt.scatter(df['predicted'], np.sqrt(np.abs(standardized_residuals)))
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Scale-Location Plot')
plt.xlabel('Fitted Values')
plt.ylabel('Sqrt(Standardized Residuals)')
plt.grid(True)
plt.show()

"""IGNORE"""

#US TRAVELLERS ENTERING CANADA VIA AIR: Actual vs Predicted Traveler Counts (2000-2025)
import pandas as pd
import statsmodels.api as sm
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
import seaborn as sns

#Data for travelers from 2000-2025
data = {
  'Date': [
      "2000-01-01", "2001-01-01", "2002-01-01", "2003-01-01", "2004-01-01", "2005-01-01", "2006-01-01", "2007-01-01",
      "2008-01-01", "2009-01-01", "2010-01-01", "2011-01-01", "2012-01-01", "2013-01-01", "2014-01-01", "2015-01-01",
      "2016-01-01", "2017-01-01", "2018-01-01", "2019-01-01", "2020-01-01", "2021-01-01", "2022-01-01", "2023-01-01",
      "2024-01-01", "2025-01-01"
    ],
    'Travelers': [
      216987, 238824, 220961, 220739, 200893, 228299, 211623, 210382, 198782, 183404,
      178505, 193304, 195813, 198202, 201965, 215393, 240179, 271319, 289486, 281378,
      283023, 11487, 57701, 198179, 223574, 248628
    ]
}

# Create the DataFrame
df = pd.DataFrame(data)

# Convert the Date column to datetime
df['Date'] = pd.to_datetime(df['Date'])

# Add time variables
df['time'] = np.arange(1, len(df) + 1)  # Sequential time (1, 2, 3, ..., 26)
```

```
df['intervention'] = np.where(df['Date'].dt.year >= 2025, 1, 0)  # Binary intervention variable for 2025
df['post_time'] = np.where(df['Date'].dt.year >= 2025, df['time'] - 2024, 0)  # Time since 2025

# OLS Regression: Y = β0 + β1 * time + β2 * intervention + β3 * post_time + ε
X = df[['time', 'intervention', 'post_time']]
X = sm.add_constant(X)  # Add intercept
y = df['Travelers']

# Fit the model
model = sm.OLS(y, X)
results = model.fit()

# Predict the values for the full dataset (including 2025)
df['predicted'] = results.predict(X)

# Plot the actual vs predicted data for 2025
plt.figure(figsize=(10, 6))
plt.plot(df['Date'], df['Travelers'], label='Actual Traveler Counts', color='b', marker='o')
plt.plot(df['Date'], df['predicted'], label='Predicted Traveler Counts (ITS Model)', color='r', linestyle='--')

# Highlight the start of 2025
plt.axvline(pd.to_datetime('2025-01-01'), color='g', linestyle='--', label='Start of 2025')

plt.title('Actual vs Predicted Traveler Counts (2000-2025)')
plt.xlabel('Year')
plt.ylabel('Traveler Count')
plt.legend()
plt.grid(True)
plt.show()

# Create a new DataFrame to predict the value for 2025
new_data = pd.DataFrame({
    'const': [1],
    'time': [26],  # This is the time point for 2025
    'intervention': [1],  # For post-2025 period
    'post_time': [1]  # Time since 2025 (relative to the intervention)
})

# Predict for 2025 using the model
predicted_2025 = results.predict(new_data)[0]

# Add 2025 prediction to the plot
plt.figure(figsize=(10, 6))
plt.plot(df['Date'], df['Travelers'], label='Actual Traveler Counts', color='b', marker='o')
plt.plot(df['Date'], df['predicted'], label='Predicted Traveler Counts (ITS Model)', color='r', linestyle='--')

# Highlight the start of 2025 and the predicted value for 2025
```

```
plt.axvline(pd.to_datetime('2025-01-01'), color='g', linestyle='--', label='Start of 2025')
plt.scatter(pd.to_datetime('2025-01-01'), predicted_2025, color='orange', label=f'Predicted
2025: {predicted_2025:.0f}')

plt.title('Actual vs Predicted Traveler Counts (2000-2025)')
plt.xlabel('Year')
plt.ylabel('Traveler Count')
plt.legend()
plt.grid(True)
plt.show()

# Print the predicted value for 2025
print(f"Predicted number of travelers for 2025: {predicted_2025:.0f}")

# Print the regression results summary
print(results.summary())

# Diagnostic Checks
# 1. Residuals vs Fitted Plot
plt.scatter(df['predicted'], results.resid)
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Residuals vs Fitted Values')
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.grid(True)
plt.show()

# 2. Q-Q Plot for Normality of Residuals
sm.qqplot(results.resid, line ='45')
plt.title('Q-Q Plot of Residuals')
plt.show()

# 3. Levene's Test for Homogeneity of Variance
stat, p_value = stats.levene(df['Travelers'], df['predicted'])
print(f"Levene’s test statistic: {stat}, p-value: {p_value}")

# 4. Durbin-Watson Test for Autocorrelation
from statsmodels.stats.stattools import durbin_watson
dw_stat = durbin_watson(results.resid)
print(f'Durbin-Watson Statistic: {dw_stat}')

# 5. Influence Plot
sm.graphics.influence_plot(results)
plt.title('Influence Plot')
plt.show()

# 6. Histogram of Residuals
plt.hist(results.resid, bins=10, edgecolor='k')
plt.title('Histogram of Residuals')
plt.xlabel('Residuals')
```

```
plt.ylabel('Frequency')
plt.show()

# 7. Scale-Location Plot
standardized_residuals = results.get_influence().resid_studentized_internal
plt.scatter(df['predicted'], np.sqrt(np.abs(standardized_residuals)))
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Scale-Location Plot')
plt.xlabel('Fitted Values')
plt.ylabel('Sqrt(Standardized Residuals)')
plt.grid(True)
plt.show()

#US TRAVELLERS ENTERING CANADA OVERALL: Actual vs Predicted Traveler Counts (2000-2025)# JANUARY ONLY 2001 TO JANUARY 2025
import pandas as pd
import statsmodels.api as sm
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
import seaborn as sns

#Data for travelers from 2000-2025
data = {
   'Date': [
        "2000-01-01", "2001-01-01", "2002-01-01", "2003-01-01", "2004-01-01", "2005-01-01", "2006-01-01", "2007-01-01",
        "2008-01-01", "2009-01-01", "2010-01-01", "2011-01-01", "2012-01-01", "2013-01-01", "2014-01-01", "2015-01-01",
        "2016-01-01", "2017-01-01", "2018-01-01", "2019-01-01", "2020-01-01", "2021-01-01", "2022-01-01", "2023-01-01",
        "2024-01-01", "2025-01-01"
    ],
    'Travelers': [
        2379373, 2516617, 2179481, 2204612, 1833581, 1711345, 1629252, 1353963, 1162479, 1090995,
        1045322, 997427, 1011258, 1002423, 953524, 1018877, 1145168, 1164610, 1146690, 1140166,
        1201690, 86456, 218558, 866679, 910621, 1100257
    ]
}

# Create the DataFrame
df = pd.DataFrame(data)

# Convert the Date column to datetime
df['Date'] = pd.to_datetime(df['Date'])

# Add time variables
df['time'] = np.arange(1, len(df) + 1)  # Sequential time (1, 2, 3, ..., 26)
```

```
df['intervention'] = np.where(df['Date'].dt.year >= 2025, 1, 0)  # Binary intervention variable for 2025
df['post_time'] = np.where(df['Date'].dt.year >= 2025, df['time'] - 2024, 0)  # Time since 2025

# OLS Regression: Y = β0 + β1 * time + β2 * intervention + β3 * post_time + ε
X = df[['time', 'intervention', 'post_time']]
X = sm.add_constant(X)  # Add intercept
y = df['Travelers']

# Fit the model
model = sm.OLS(y, X)
results = model.fit()

# Predict the values for the full dataset (including 2025)
df['predicted'] = results.predict(X)

# Plot the actual vs predicted data for 2025
plt.figure(figsize=(10, 6))
plt.plot(df['Date'], df['Travelers'], label='Actual Traveler Counts', color='b', marker='o')
plt.plot(df['Date'], df['predicted'], label='Predicted Traveler Counts (ITS Model)', color='r', linestyle='--')

# Highlight the start of 2025
plt.axvline(pd.to_datetime('2025-01-01'), color='g', linestyle='--', label='Start of 2025')

plt.title('US TRAVELLERS ENTERING CANADA TOTAL: Actual vs Predicted Traveler Counts (2000-2025)')
plt.xlabel('Year')
plt.ylabel('Traveler Count')
plt.legend()
plt.grid(True)
plt.show()

# Create a new DataFrame to predict the value for 2025
new_data = pd.DataFrame({
    'const': [1],
    'time': [26],  # This is the time point for 2025
    'intervention': [1],  # For post-2025 period
    'post_time': [1]  # Time since 2025 (relative to the intervention)
})

# Predict for 2025 using the model
predicted_2025 = results.predict(new_data)[0]

# Add 2025 prediction to the plot
plt.figure(figsize=(10, 6))
plt.plot(df['Date'], df['Travelers'], label='Actual Traveler Counts', color='b', marker='o')
plt.plot(df['Date'], df['predicted'], label='Predicted Traveler Counts (ITS Model)', color='r', linestyle='--')
```

```
# Highlight the start of 2025 and the predicted value for 2025
plt.axvline(pd.to_datetime('2025-01-01'), color='g', linestyle='--', label='Start of 2025')
plt.scatter(pd.to_datetime('2025-01-01'), predicted_2025, color='orange', label=f'Predicted
2025: {predicted_2025:.0f}')

plt.title('US TRAVELLERS ENTERING CANADA TOTAL:Actual vs Predicted Traveler
Counts (2000-2025)')
plt.xlabel('Year')
plt.ylabel('Traveler Count')
plt.legend()
plt.grid(True)
plt.show()

# Print the predicted value for 2025
print(f"Predicted number of travelers for 2025: {predicted_2025:.0f}")

# Print the regression results summary
print(results.summary())

# Diagnostic Checks
# 1. Residuals vs Fitted Plot
plt.scatter(df['predicted'], results.resid)
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Residuals vs Fitted Values')
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.grid(True)
plt.show()

# 2. Q-Q Plot for Normality of Residuals
sm.qqplot(results.resid, line ='45')
plt.title('Q-Q Plot of Residuals')
plt.show()

# 3. Levene's Test for Homogeneity of Variance
stat, p_value = stats.levene(df['Travelers'], df['predicted'])
print(f"Levene’s test statistic: {stat}, p-value: {p_value}")

# 4. Durbin-Watson Test for Autocorrelation
from statsmodels.stats.stattools import durbin_watson
dw_stat = durbin_watson(results.resid)
print(f'Durbin-Watson Statistic: {dw_stat}')

# 5. Influence Plot
sm.graphics.influence_plot(results)
plt.title('Influence Plot')
plt.show()

# 6. Histogram of Residuals
plt.hist(results.resid, bins=10, edgecolor='k')
```

```
plt.title('Histogram of Residuals')
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.show()

# 7. Scale-Location Plot
standardized_residuals = results.get_influence().resid_studentized_internal
plt.scatter(df['predicted'], np.sqrt(np.abs(standardized_residuals)))
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Scale-Location Plot')
plt.xlabel('Fitted Values')
plt.ylabel('Sqrt(Standardized Residuals)')
plt.grid(True)
plt.show()

#CANADIAN TRAVELLERS RETURNIGN TO CANADA VIA LAND: Actual vs Predicted Traveler Counts (2000-2025)
import pandas as pd
import statsmodels.api as sm
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
import seaborn as sns

#Data for travelers from 2000-2025
data = {
  'Date': [
      "2000-01-01", "2001-01-01", "2002-01-01", "2003-01-01", "2004-01-01", "2005-01-01", "2006-01-01", "2007-01-01",
      "2008-01-01", "2009-01-01", "2010-01-01", "2011-01-01", "2012-01-01", "2013-01-01", "2014-01-01", "2015-01-01",
      "2016-01-01", "2017-01-01", "2018-01-01", "2019-01-01", "2020-01-01", "2021-01-01", "2022-01-01", "2023-01-01",
      "2024-01-01", "2025-01-01"
   ],
   'Travelers': [
      2412575, 2480518, 2042884, 2076722, 2018639, 2066091, 2331740, 2307662, 2551879, 2023341,
      2454010, 2820717, 3033350, 3207616, 3000522, 2630092, 2206077, 2266722, 2333912, 2130547,
      2132891, 192318, 428880, 1633224, 1811362, 1822502
   ]
}

# Create the DataFrame
df = pd.DataFrame(data)

# Convert the Date column to datetime
df['Date'] = pd.to_datetime(df['Date'])
```

```
# Add time variables
df['time'] = np.arange(1, len(df) + 1)  # Sequential time (1, 2, 3, ..., 26)
df['intervention'] = np.where(df['Date'].dt.year >= 2025, 1, 0)  # Binary intervention variable for 2025
df['post_time'] = np.where(df['Date'].dt.year >= 2025, df['time'] - 2024, 0)  # Time since 2025

# OLS Regression: Y = β0 + β1 * time + β2 * intervention + β3 * post_time + ε
X = df[['time', 'intervention', 'post_time']]
X = sm.add_constant(X)  # Add intercept
y = df['Travelers']

# Fit the model
model = sm.OLS(y, X)
results = model.fit()

# Predict the values for the full dataset (including 2025)
df['predicted'] = results.predict(X)

# Plot the actual vs predicted data for 2025
plt.figure(figsize=(10, 6))
plt.plot(df['Date'], df['Travelers'], label='Actual Traveler Counts', color='b', marker='o')
plt.plot(df['Date'], df['predicted'], label='Predicted Traveler Counts (ITS Model)', color='r', linestyle='--')

# Highlight the start of 2025
plt.axvline(pd.to_datetime('2025-01-01'), color='g', linestyle='--', label='Start of 2025')

plt.title('Actual vs Predicted Traveler Counts (2000-2025)')
plt.xlabel('Year')
plt.ylabel('Traveler Count')
plt.legend()
plt.grid(True)
plt.show()

# Create a new DataFrame to predict the value for 2025
new_data = pd.DataFrame({
    'const': [1],
    'time': [26],  # This is the time point for 2025
    'intervention': [1],  # For post-2025 period
    'post_time': [1]  # Time since 2025 (relative to the intervention)
})

# Predict for 2025 using the model
predicted_2025 = results.predict(new_data)[0]

# Add 2025 prediction to the plot
plt.figure(figsize=(10, 6))
plt.plot(df['Date'], df['Travelers'], label='Actual Traveler Counts', color='b', marker='o')
plt.plot(df['Date'], df['predicted'], label='Predicted Traveler Counts (ITS Model)', color='r', linestyle='--')
```

```
# Highlight the start of 2025 and the predicted value for 2025
plt.axvline(pd.to_datetime('2025-01-01'), color='g', linestyle='--', label='Start of 2025')
plt.scatter(pd.to_datetime('2025-01-01'), predicted_2025, color='orange', label=f'Predicted
2025: {predicted_2025:.0f}')

plt.title('CANADIAN TRAVELLERS RETURNING TO CANADA VIA LAND: Actual vs
Predicted Traveler Counts (2000-2025)')
plt.xlabel('Year')
plt.ylabel('Traveler Count')
plt.legend()
plt.grid(True)
plt.show()

# Print the predicted value for 2025
print(f"Predicted number of travelers for 2025: {predicted_2025:.0f}")

# Print the regression results summary
print(results.summary())

# Diagnostic Checks
# 1. Residuals vs Fitted Plot
plt.scatter(df['predicted'], results.resid)
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Residuals vs Fitted Values')
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.grid(True)
plt.show()

# 2. Q-Q Plot for Normality of Residuals
sm.qqplot(results.resid, line ='45')
plt.title('Q-Q Plot of Residuals')
plt.show()

# 3. Levene's Test for Homogeneity of Variance
stat, p_value = stats.levene(df['Travelers'], df['predicted'])
print(f"Levene’s test statistic: {stat}, p-value: {p_value}")

# 4. Durbin-Watson Test for Autocorrelation
from statsmodels.stats.stattools import durbin_watson
dw_stat = durbin_watson(results.resid)
print(f'Durbin-Watson Statistic: {dw_stat}')

# 5. Influence Plot
sm.graphics.influence_plot(results)
plt.title('Influence Plot')
plt.show()

# 6. Histogram of Residuals
```

```
plt.hist(results.resid, bins=10, edgecolor='k')
plt.title('Histogram of Residuals')
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.show()

# 7. Scale-Location Plot
standardized_residuals = results.get_influence().resid_studentized_internal
plt.scatter(df['predicted'], np.sqrt(np.abs(standardized_residuals)))
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Scale-Location Plot')
plt.xlabel('Fitted Values')
plt.ylabel('Sqrt(Standardized Residuals)')
plt.grid(True)
plt.show()

"""CANADIAN RESIDENT TESTS"""

#CANADIAN TRAVELLERS RETURNING TO CANADA VIA AIR: Actual vs Predicted Traveler Counts (2000-2025)
import pandas as pd
import statsmodels.api as sm
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
import seaborn as sns

#Data for travelers from 2000-2025
data = {
  'Date': [
      "2000-01-01", "2001-01-01", "2002-01-01", "2003-01-01", "2004-01-01", "2005-01-01", "2006-01-01", "2007-01-01",
      "2008-01-01", "2009-01-01", "2010-01-01", "2011-01-01", "2012-01-01", "2013-01-01", "2014-01-01", "2015-01-01",
      "2016-01-01", "2017-01-01", "2018-01-01", "2019-01-01", "2020-01-01", "2021-01-01", "2022-01-01", "2023-01-01",
      "2024-01-01", "2025-01-01"
    ],
    'Travelers': [
      474372, 510152, 381235, 411363, 402378, 482633, 525694, 569567, 619980, 569630,
      638197, 682657, 734872, 747132, 760867, 799317, 716712, 783361, 872392, 904756,
      936866, 72649, 261143, 1107843, 919454, 845669
    ]
}

# Create the DataFrame
df = pd.DataFrame(data)

# Convert the Date column to datetime
df['Date'] = pd.to_datetime(df['Date'])
```

```
# Add time variables
df['time'] = np.arange(1, len(df) + 1)  # Sequential time (1, 2, 3, ..., 26)
df['intervention'] = np.where(df['Date'].dt.year >= 2025, 1, 0)  # Binary intervention variable for 2025
df['post_time'] = np.where(df['Date'].dt.year >= 2025, df['time'] - 2024, 0)  # Time since 2025

# OLS Regression: Y = β0 + β1 * time + β2 * intervention + β3 * post_time + ε
X = df[['time', 'intervention', 'post_time']]
X = sm.add_constant(X)  # Add intercept
y = df['Travelers']

# Fit the model
model = sm.OLS(y, X)
results = model.fit()

# Predict the values for the full dataset (including 2025)
df['predicted'] = results.predict(X)

# Plot the actual vs predicted data for 2025
plt.figure(figsize=(10, 6))
plt.plot(df['Date'], df['Travelers'], label='Actual Traveler Counts', color='b', marker='o')
plt.plot(df['Date'], df['predicted'], label='Predicted Traveler Counts (ITS Model)', color='r', linestyle='--')

# Highlight the start of 2025
plt.axvline(pd.to_datetime('2025-01-01'), color='g', linestyle='--', label='Start of 2025')

plt.title('Actual vs Predicted Traveler Counts (2000-2025)')
plt.xlabel('Year')
plt.ylabel('Traveler Count')
plt.legend()
plt.grid(True)
plt.show()

# Create a new DataFrame to predict the value for 2025
new_data = pd.DataFrame({
    'const': [1],
    'time': [26],  # This is the time point for 2025
    'intervention': [1],  # For post-2025 period
    'post_time': [1]  # Time since 2025 (relative to the intervention)
})

# Predict for 2025 using the model
predicted_2025 = results.predict(new_data)[0]

# Add 2025 prediction to the plot
plt.figure(figsize=(10, 6))
plt.plot(df['Date'], df['Travelers'], label='Actual Traveler Counts', color='b', marker='o')
```

```
plt.plot(df['Date'], df['predicted'], label='Predicted Traveler Counts (ITS Model)', color='r', linestyle='--')

# Highlight the start of 2025 and the predicted value for 2025
plt.axvline(pd.to_datetime('2025-01-01'), color='g', linestyle='--', label='Start of 2025')
plt.scatter(pd.to_datetime('2025-01-01'), predicted_2025, color='orange', label=f'Predicted 2025: {predicted_2025:.0f}')

plt.title('Actual vs Predicted Traveler Counts (2000-2025)')
plt.xlabel('Year')
plt.ylabel('Traveler Count')
plt.legend()
plt.grid(True)
plt.show()

# Print the predicted value for 2025
print(f"Predicted number of travelers for 2025: {predicted_2025:.0f}")

# Print the regression results summary
print(results.summary())

# Diagnostic Checks
# 1. Residuals vs Fitted Plot
plt.scatter(df['predicted'], results.resid)
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Residuals vs Fitted Values')
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.grid(True)
plt.show()

# 2. Q-Q Plot for Normality of Residuals
sm.qqplot(results.resid, line ='45')
plt.title('Q-Q Plot of Residuals')
plt.show()

# 3. Levene's Test for Homogeneity of Variance
stat, p_value = stats.levene(df['Travelers'], df['predicted'])
print(f"Levene’s test statistic: {stat}, p-value: {p_value}")

# 4. Durbin-Watson Test for Autocorrelation
from statsmodels.stats.stattools import durbin_watson
dw_stat = durbin_watson(results.resid)
print(f'Durbin-Watson Statistic: {dw_stat}')

# 5. Influence Plot
sm.graphics.influence_plot(results)
plt.title('Influence Plot')
plt.show()
```

```
# 6. Histogram of Residuals
plt.hist(results.resid, bins=10, edgecolor='k')
plt.title('Histogram of Residuals')
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.show()

# 7. Scale-Location Plot
standardized_residuals = results.get_influence().resid_studentized_internal
plt.scatter(df['predicted'], np.sqrt(np.abs(standardized_residuals)))
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Scale-Location Plot')
plt.xlabel('Fitted Values')
plt.ylabel('Sqrt(Standardized Residuals)')
plt.grid(True)
plt.show()

#CANADIAN TRAVELLERS RETURNIGN TO CANADA OVERALL: Actual vs
Predicted Traveler Counts (2000-2025)
import pandas as pd
import statsmodels.api as sm
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
import seaborn as sns

#Data for travelers from 2000-2025
data = {
  'Date': [
      "2000-01-01", "2001-01-01", "2002-01-01", "2003-01-01", "2004-01-01", "2005-01-
01", "2006-01-01", "2007-01-01",
      "2008-01-01", "2009-01-01", "2010-01-01", "2011-01-01", "2012-01-01", "2013-01-
01", "2014-01-01", "2015-01-01",
      "2016-01-01", "2017-01-01", "2018-01-01", "2019-01-01", "2020-01-01", "2021-01-
01", "2022-01-01", "2023-01-01",
      "2024-01-01", "2025-01-01"
   ],
   'Travelers': [
      2887227, 2990873, 2424297, 2488296, 2421217, 2548897, 2857631, 2877375,
3172026, 2593090,
      3092336, 3503674, 3768389, 3954853, 3761567, 3429575, 2922914, 3050181,
3206391, 3035434,
      3069835, 264999, 690166, 2744340, 2734155, 2670643
   ]
}

# Create the DataFrame
df = pd.DataFrame(data)

# Convert the Date column to datetime
```

```
df['Date'] = pd.to_datetime(df['Date'])

# Add time variables
df['time'] = np.arange(1, len(df) + 1)  # Sequential time (1, 2, 3, ..., 26)
df['intervention'] = np.where(df['Date'].dt.year >= 2025, 1, 0)  # Binary intervention variable for 2025
df['post_time'] = np.where(df['Date'].dt.year >= 2025, df['time'] - 2024, 0)  # Time since 2025

# OLS Regression: Y = β0 + β1 * time + β2 * intervention + β3 * post_time + ε
X = df[['time', 'intervention', 'post_time']]
X = sm.add_constant(X)  # Add intercept
y = df['Travelers']

# Fit the model
model = sm.OLS(y, X)
results = model.fit()

# Predict the values for the full dataset (including 2025)
df['predicted'] = results.predict(X)

# Plot the actual vs predicted data for 2025
plt.figure(figsize=(10, 6))
plt.plot(df['Date'], df['Travelers'], label='Actual Traveler Counts', color='b', marker='o')
plt.plot(df['Date'], df['predicted'], label='Predicted Traveler Counts (ITS Model)', color='r', linestyle='--')

# Highlight the start of 2025
plt.axvline(pd.to_datetime('2025-01-01'), color='g', linestyle='--', label='Start of 2025')

plt.title('CANADIAN TRAVELLERS RETURNING TO CANADA TOTAL:Actual vs Predicted Traveler Counts (2000-2025)')
plt.xlabel('Year')
plt.ylabel('Traveler Count')
plt.legend()
plt.grid(True)
plt.show()

# Create a new DataFrame to predict the value for 2025
new_data = pd.DataFrame({
    'const': [1],
    'time': [26],  # This is the time point for 2025
    'intervention': [1],  # For post-2025 period
    'post_time': [1]  # Time since 2025 (relative to the intervention)
})

# Predict for 2025 using the model
predicted_2025 = results.predict(new_data)[0]

# Add 2025 prediction to the plot
plt.figure(figsize=(10, 6))
```

```
plt.plot(df['Date'], df['Travelers'], label='Actual Traveler Counts', color='b', marker='o')
plt.plot(df['Date'], df['predicted'], label='Predicted Traveler Counts (ITS Model)', color='r',
linestyle='--')

# Highlight the start of 2025 and the predicted value for 2025
plt.axvline(pd.to_datetime('2025-01-01'), color='g', linestyle='--', label='Start of 2025')
plt.scatter(pd.to_datetime('2025-01-01'), predicted_2025, color='orange', label=f'Predicted
2025: {predicted_2025:.0f}')

plt.title('CANADIAN TRAVELLERS RETURNING TO CANADA TOTAL:Actual vs
Predicted Traveler Counts (2000-2025)')
plt.xlabel('Year')
plt.ylabel('Traveler Count')
plt.legend()
plt.grid(True)
plt.show()

# Print the predicted value for 2025
print(f"Predicted number of travelers for 2025: {predicted_2025:.0f}")

# Print the regression results summary
print(results.summary())

# Diagnostic Checks
# 1. Residuals vs Fitted Plot
plt.scatter(df['predicted'], results.resid)
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Residuals vs Fitted Values')
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.grid(True)
plt.show()

# 2. Q-Q Plot for Normality of Residuals
sm.qqplot(results.resid, line ='45')
plt.title('Q-Q Plot of Residuals')
plt.show()

# 3. Levene's Test for Homogeneity of Variance
stat, p_value = stats.levene(df['Travelers'], df['predicted'])
print(f"Levene’s test statistic: {stat}, p-value: {p_value}")

# 4. Durbin-Watson Test for Autocorrelation
from statsmodels.stats.stattools import durbin_watson
dw_stat = durbin_watson(results.resid)
print(f'Durbin-Watson Statistic: {dw_stat}')

# 5. Influence Plot
sm.graphics.influence_plot(results)
plt.title('Influence Plot')
```

```
plt.show()

# 6. Histogram of Residuals
plt.hist(results.resid, bins=10, edgecolor='k')
plt.title('Histogram of Residuals')
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.show()

# 7. Scale-Location Plot
standardized_residuals = results.get_influence().resid_studentized_internal
plt.scatter(df['predicted'], np.sqrt(np.abs(standardized_residuals)))
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Scale-Location Plot')
plt.xlabel('Fitted Values')
plt.ylabel('Sqrt(Standardized Residuals)')
plt.grid(True)
plt.show()
```

Test code

```
# -*- coding: utf-8 -*-
"""statistical_tests_code.ipynb

Automatically generated by Colab.

Original file is located at
    https://colab.research.google.com/drive/1IDn4f3TS8wH_EA0dukDYIXGbykFCj-wG
"""

# Import Packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from statsmodels.stats.outliers_influence import variance_inflation_factor
from scipy import stats

# Clean Dataset

# Reads CSV file and and delete rows
df = pd.read_csv('raw_data.csv', skiprows = 9)

# Delete the last 22 rows
df = df.iloc[:-22]

# Clean column names
df.columns = df.columns.str.strip().str.replace(r'\s+', ' ', regex=True)
```

```
# Delete columns not relevant to the report
df = df.drop(columns=[
    "International travellers entering or returning to Canada",
    "Residents of countries other than the United States of America, air",
    "Residents of countries other than the United States of America, land",
    "Residents of countries other than the United States of America, water",
    "Canadian-resident visitors returning to Canada"
], errors="ignore")

# Function to parse mixed date formats
def fix_reference_period(date_str):
    try:
        if "-" in date_str:
            parts = date_str.split("-")

            if parts[0].isdigit():
                year = int(parts[0]) + 2000
                month = parts[1]
            else:
                month = parts[0]
                year = int(parts[1])
                if year < 100:
                    year += 2000

            # Convert to datetime and format as YYYY-MM
            return pd.to_datetime(f"{month} {year}", format="%b %Y").strftime("%Y-%m")

    # Debugging and blank value caveat
    except Exception as e:
        print(f"Error converting {date_str}: {e}")
        return None

# Apply function to fix dates
df["Reference period"] = df["Reference period"].apply(fix_reference_period)

# Change data to float values
# Apply relevant traveller type to all rows
df["Traveller type"] = df["Traveller type"].ffill()

# Save cleaned data as a new CSV file
df.to_csv('clean_data.csv', index = False)

# Two-tail t-test for Canadians Returning from the US vs Americans Entering Canada
(Annually)

# Filter rows where the 'Traveller type' is 'Travellers'
df_filtered = df[df['Traveller type'] == 'Travellers'].copy()

# Convert 'Reference period' to datetime format
```

```
df_filtered['Reference period'] = pd.to_datetime(df_filtered['Reference period'],
errors='coerce')

# Check for any none time values after conversion
if df_filtered['Reference period'].isna().sum() > 0:
    print(f"There are {df_filtered['Reference period'].isna().sum()} invalid date entries.")
    print(df_filtered[df_filtered['Reference period'].isna()])

# Extract the year from the 'Reference period' column
df_filtered['Year'] = df_filtered['Reference period'].dt.year

# Aggregate the data by year
annual_data = df_filtered.groupby('Year').agg({
    'United States of America residents entering Canada': 'sum',
    'Canadian residents returning from the United States of America': 'sum'
}).reset_index()

# Create variables for the t-test based on the annual aggregated data
us_residents_entering_canada = annual_data['United States of America residents entering
Canada']
canadian_residents_returning_usa = annual_data['Canadian residents returning from the
United States of America']
# Run the t-test
t_statistic, p_value = stats.ttest_ind(us_residents_entering_canada,
canadian_residents_returning_usa)

# Calculate the means of both groups
mean_us_entering_canada = us_residents_entering_canada.mean()
mean_canadians_returning_usa = canadian_residents_returning_usa.mean()

# Print the means
print(f'Mean of U.S. residents entering Canada: {mean_us_entering_canada}')
print(f'Mean of Canadian residents returning from the U.S.:
{mean_canadians_returning_usa}')

# Print the results of the t-test
print(f'T-statistic: {t_statistic}')
print(f'P-value: {p_value}')

# Interpretation of results
if p_value < 0.05:
    print("There is a significant difference between U.S. residents entering Canada and
Canadian residents returning to the U.S.")
else:
    print("There is no significant difference between U.S. residents entering Canada and
Canadian residents returning to the U.S.")

# Two tail t-test for Excursionist vs Tourist Canadians Returning from the US (Annually)

# Clear the object
df_filtered = df.copy()
```

```
# Filter rows where the 'Traveller type' is 'Excursionists' or 'Tourists'
df_filtered = df[df['Traveller type'].isin(['Excursionists', 'Tourists'])].copy()

# Convert 'Reference period' to datetime format
df_filtered['Reference period'] = pd.to_datetime(df_filtered['Reference period'],
errors='coerce')

# Check for any none time values after conversion
if df_filtered['Reference period'].isna().sum() > 0:
   print(f"There are {df_filtered['Reference period'].isna().sum()} invalid date entries.")
   print(df_filtered[df_filtered['Reference period'].isna()])

# Extract the year from the 'Reference period' column
df_filtered['Year'] = df_filtered['Reference period'].dt.year

# Filter for Canadian residents returning from the U.S. based on Excursionists and Tourists
df_canadian_return_excursionists = df_filtered[(df_filtered['Traveller type'] == 'Excursionists')
&
                                    df_filtered['Canadian residents returning from the United States of
America'].notna()]

df_canadian_return_tourists = df_filtered[(df_filtered['Traveller type'] == 'Tourists') &
                                 df_filtered['Canadian residents returning from the United States of
America'].notna()]

# Aggregate the data by year for both Excursionists and Tourists
annual_excursionists = df_canadian_return_excursionists.groupby('Year').agg({
   'Canadian residents returning from the United States of America': 'sum'
}).reset_index()

annual_tourists = df_canadian_return_tourists.groupby('Year').agg({
   'Canadian residents returning from the United States of America': 'sum'
}).reset_index()

# Create variables for the t-test
canadian_residents_returning_excursionists = annual_excursionists['Canadian residents
returning from the United States of America']
canadian_residents_returning_tourists = annual_tourists['Canadian residents returning from
the United States of America']
# Calculate the means for both groups
mean_canadian_residents_returning_excursionists =
canadian_residents_returning_excursionists.mean()
mean_canadian_residents_returning_tourists = canadian_residents_returning_tourists.mean()

# Print mean results
print(f'Mean of Canadian Excursionists Returning from the U.S.:
{mean_canadian_residents_returning_excursionists}')
print(f'Mean of Canadian Tourists Returning from the U.S.:
{mean_canadian_residents_returning_tourists}')
```

```
# Run the t-test
t_statistic, p_value = stats.ttest_ind(canadian_residents_returning_excursionists,
canadian_residents_returning_tourists)

# Print the results of the t-test
print(f'T-statistic: {t_statistic}')
print(f'P-value: {p_value}')

# Interpretation of results
if p_value < 0.05:
    print("There is a significant difference between Excursionists and Tourists for Canadian
residents returning from the U.S.")
else:
    print("There is no significant difference between Excursionists and Tourists for Canadian
residents returning from the U.S.")

# Two tail t-test for Excursionist vs Tourist Americans Entering Canada (Annually)

# Clear object
df_filtered = df.copy()

# Filter rows where the 'Traveller type' is 'Excursionists' or 'Tourists'
df_filtered = df[df['Traveller type'].isin(['Excursionists', 'Tourists'])].copy()

# Convert 'Reference period' to datetime format
df_filtered['Reference period'] = pd.to_datetime(df_filtered['Reference period'],
errors='coerce')

# Extract the year from the 'Reference period' column
df_filtered['Year'] = df_filtered['Reference period'].dt.year

# Filter for American residents entering Canada based on Excursionists and Tourists
df_us_residents_entering_canada_excursionists = df_filtered[(df_filtered['Traveller type'] ==
'Excursionists') &
                                                            df_filtered['United States of America residents entering
Canada'].notna()]

df_us_residents_entering_canada_tourists = df_filtered[(df_filtered['Traveller type'] ==
'Tourists') &
                                                       df_filtered['United States of America residents entering
Canada'].notna()]

# Aggregate the data by year for both Excursionists and Tourists
annual_excursionists = df_us_residents_entering_canada_excursionists.groupby('Year').agg({
    'United States of America residents entering Canada': 'sum'
}).reset_index()

annual_tourists = df_us_residents_entering_canada_tourists.groupby('Year').agg({
    'United States of America residents entering Canada': 'sum'
}).reset_index()
```

```
# Create variables for t-test
us_residents_entering_canada_excursionists = annual_excursionists['United States of America residents entering Canada']
us_residents_entering_canada_tourists = annual_tourists['United States of America residents entering Canada']

# Calculate the means for both groups
mean_us_residents_entering_canada_excursionists = us_residents_entering_canada_excursionists.mean()

mean_us_residents_entering_canada_tourists = us_residents_entering_canada_tourists.mean()

# Print mean results
print(f'Mean of American Excursionists Entering Canada: {mean_us_residents_entering_canada_excursionists}')
print(f'Mean of American Tourists Entering Canada: {mean_us_residents_entering_canada_tourists}')

# Run t-test
t_statistic, p_value = stats.ttest_ind(us_residents_entering_canada_excursionists, us_residents_entering_canada_tourists)

# Print the results of the t-test
print(f'T-statistic: {t_statistic}')
print(f'P-value: {p_value}')

# Interpretation of results
if p_value < 0.05:
    print("There is a significant difference between Excursionists and Tourists for U.S. residents entering Canada.")
else:
    print("There is no significant difference between Excursionists and Tourists for U.S. residents entering Canada.")

# Two-way ANOVA for Excursionists and Tourists for Canadian residents returning from the U.S. and U.S. residents entering Canada (Annually)

# Clear object
df_filtered = df.copy()

# Filter rows where the 'Traveller type' is 'Excursionists' or 'Tourists'
df_filtered = df[df['Traveller type'].isin(['Excursionists', 'Tourists'])].copy()

# Convert 'Reference period' to datetime format
df_filtered['Reference period'] = pd.to_datetime(df_filtered['Reference period'], errors='coerce')

# Extract the year from the 'Reference period' column
df_filtered['Year'] = df_filtered['Reference period'].dt.year
```

```
# Aggregate the data by year for both groups (Canadians and Americans)
annual_data = df_filtered.groupby(['Year', 'Traveller type']).agg({
    'United States of America residents entering Canada': 'sum',
    'Canadian residents returning from the United States of America': 'sum'
}).reset_index()

# Combine the two columns (Canadian and U.S. residents) into one column for the ANOVA
# Create a new column indicating 'Travel Direction' (either returning Canadians or entering
Americans)
annual_data_melted = annual_data.melt(id_vars=['Year', 'Traveller type'], value_vars=[
    'United States of America residents entering Canada',
    'Canadian residents returning from the United States of America'
], var_name='Travel Direction', value_name='Travelers')
# Run a two-way ANOVA
f_statistic, p_value = stats.f_oneway(
    annual_data_melted[annual_data_melted['Traveller type'] == 'Excursionists']['Travelers'],
    annual_data_melted[annual_data_melted['Traveller type'] == 'Tourists']['Travelers']
)

# Print the results of the ANOVA
print(f'F-statistic: {f_statistic}')
print(f'P-value: {p_value}')

# Interpretation of results
if p_value < 0.05:
    print("There is a significant difference between Excursionists and Tourists for Canadian
residents returning from the U.S. and U.S. residents entering Canada.")
else:
    print("There is no significant difference between Excursionists and Tourists for Canadian
residents returning from the U.S. and U.S. residents entering Canada.")

# ANOVA of Entry Type (air, land, water) of Americans Entering Canada (Annually)

# Filter rows where the 'Traveller type' is 'Travellers' and the 'Reference period' is for U.S.
residents entering Canada
df_filtered = df.copy()
df_filtered['Reference period'] = pd.to_datetime(df_filtered['Reference period'],
errors='coerce')
df_us_residents = df_filtered[(df_filtered['Traveller type'] == 'Travellers') &
                              (df_filtered['United States of America residents, air'].notna())].copy()

# Create a new column for the year by extracting from 'Reference period'
df_us_residents['Year'] = df_us_residents['Reference period'].dt.year

# Aggregate the data by year and sum the travel counts for each entry type (Air, Land, Water)
annual_entry_data = df_us_residents.groupby('Year')[[
    'United States of America residents, air',
    'United States of America residents, land',
    'United States of America residents, water']].sum().reset_index()

# Perform one-way ANOVA for entry types (Air, Land, Water)
```

```
f_statistic, p_value = stats.f_oneway(
   annual_entry_data['United States of America residents, air'],
annual_entry_data['United States of America residents, land'],
   annual_entry_data['United States of America residents, water']
)

# Print the results
print(f'F-statistic: {f_statistic}')
print(f'P-value: {p_value}')

# Interpretation of results
if p_value < 0.05:
   print("There is a significant difference in annual visits by entry type (Air, Land, Water).")
else:
   print("There is no significant difference in annual visits by entry type (Air, Land, Water).")

# ANOVA of Entry Type (air, land, water) of Canadians Returning from the US (Annually)

# Filter rows where the 'Traveller type' is 'Travellers' and the 'Reference period' is for
Canadian residents returning from the U.S.
df_filtered = df.copy()
df['Reference period'] = pd.to_datetime(df['Reference period'], errors='coerce')
df_filtered['Reference period'] = pd.to_datetime(df_filtered['Reference period'],
errors='coerce')
df_canadian_residents = df_filtered[(df_filtered['Traveller type'] == 'Travellers') &
                           (df_filtered['Canadian residents returning from the United States of
America, air'].notna())].copy()

# Create a new column for the year by extracting from 'Reference period'
df_canadian_residents['Year'] = df_canadian_residents['Reference period'].dt.year

# Aggregate the data by year and sum the travel counts for each entry type (Air, Land, Water)
annual_entry_data_canadians = df_canadian_residents.groupby('Year')[[
   'Canadian residents returning from the United States of America, air',
   'Canadian residents returning from the United States of America, land',
   'Canadian residents returning from the United States of America,
water']].sum().reset_index()

# Perform one-way ANOVA for entry types (Air, Land, Water) for Canadian residents
f_statistic, p_value = stats.f_oneway(
   annual_entry_data_canadians['Canadian residents returning from the United States of
America, air'],
   annual_entry_data_canadians['Canadian residents returning from the United States of
America, land'],
   annual_entry_data_canadians['Canadian residents returning from the United States of
America, water']
)

# Print the results
print(f'F-statistic: {f_statistic}')
print(f'P-value: {p_value}')
```

```
# Interpretation of results
if p_value < 0.05:
    print("There is a significant difference in annual visits by entry type (Air, Land, Water) for Canadian residents.")
else:

    print("There is a significant difference in annual visits by entry type (Air, Land, Water) for Canadian residents.")
else:
    print("There is no significant difference in annual visits by entry type (Air, Land, Water) for Canadian residents.")

# In[ ]:


# Get Data
df = pd.read_csv('C:/Users/cathe/Desktop/INF412_Final_Project_Data/clean_data.csv')


# In[11]:


# Line graph showing United States of America residents entering Canada and Canadian residents returning from the United States of America

# Filter for 'Travellers' only
df_filtered = df[df["Traveller type"] == "Travellers"].copy()

# Collect needed columns
cols_to_convert = [
    "United States of America residents entering Canada",
    "Canadian residents returning from the United States of America"
]

# Removes any commas from string values and convert each column into an integers
for col in cols_to_convert:
    df_filtered[col] = df_filtered[col].str.replace(",", "").astype(int)

# Ensure the Reference period is in datetime format
df_filtered["Reference period"] = pd.to_datetime(df_filtered["Reference period"])

# Plotting the data
# Size of Graph
plt.figure(figsize=(12, 6))
plt.plot(df_filtered["Reference period"], df_filtered["United States of America residents entering Canada"],
         label="U.S. Residents Entering Canada")
plt.plot(df_filtered["Reference period"], df_filtered["Canadian residents returning from the United States of America"],
         label="Canadian Residents Returning") # Format Graph
```

```
plt.xlabel("Year")
plt.ylabel("Number of Travellers")
plt.title("U.S. Residents Entering Canada & Canadian Residents Returning (Travellers)")
plt.legend()
plt.xticks(rotation=45)
plt.grid(True)
plt.show()


# In[ ]:


# Entry Methods Over Time

# Filter for 'Travellers' only
df_filtered = df[df["Traveller type"] == "Travellers"].copy()

# Collect needed columns
travel_modes = [
    "United States of America residents, air",
    "United States of America residents, land",
    "United States of America residents, water",
    "Canadian residents returning from the United States of America, air",
    "Canadian residents returning from the United States of America, land",
    "Canadian residents returning from the United States of America, water",
]

# Removes any commas from string values and convert each column into an integers
for col in travel_modes:
    df_filtered[col] = df_filtered[col].str.replace(",", "").astype(int)

# Ensure the Reference period is in datetime format
df_filtered["Reference period"] = pd.to_datetime(df_filtered["Reference period"])
# U.S. residents entering Canada
plt.plot(df_filtered["Reference period"], df_filtered["United States of America residents, air"],
label="U.S. - Air", linestyle="solid")
plt.plot(df_filtered["Reference period"], df_filtered["United States of America residents,
land"], label="U.S. - Land", linestyle="solid")
plt.plot(df_filtered["Reference period"], df_filtered["United States of America residents,
water"], label="U.S. - Water", linestyle="solid")

# Canadian residents returning from the U.S.
plt.plot(df_filtered["Reference period"], df_filtered["Canadian residents returning from the
United States of America, air"], label="Canada - Air", linestyle="solid")
plt.plot(df_filtered["Reference period"], df_filtered["Canadian residents returning from the
United States of America, land"], label="Canada - Land", linestyle="solid")
plt.plot(df_filtered["Reference period"], df_filtered["Canadian residents returning from the
United States of America, water"], label="Canada - Water", linestyle="solid")

# Format Line Graph
plt.xlabel("Year")
```

```
plt.ylabel("Number of Travellers")
plt.title("Entry Methods (Air, Land, Water) Over Time")
plt.legend()
plt.xticks(rotation=45)
plt.grid(True)
plt.show()


# In[ ]:


# Average Monthly Travel Trends

# Ensure the Reference period is in datetime format
df_filtered["Reference period"] = pd.to_datetime(df_filtered["Reference period"])

# Convert needed columns
travel_modes = [
    "United States of America residents entering Canada",
    "Canadian residents returning from the United States of America"
]

# Removes any commas from string values and convert each column into an integers
for col in travel_modes:
    df_filtered[col] = df_filtered[col].str.replace(",", "").astype(int)

# Extract month and group by month to calculate the average
df_filtered["Month"] = df_filtered["Reference period"].dt.month
monthly_avg = df_filtered.groupby("Month")[travel_modes].mean()

# Plot the line chart
plt.figure(figsize=(12, 6))
# plotting monthly average
for col in travel_modes:
    plt.plot(monthly_avg.index, monthly_avg[col], label=col)

# Structure Graph
plt.xlabel("Month")
plt.ylabel("Average Number of Travelers")
plt.title("Average Monthly Travel Trends")
plt.xticks(range(1, 13), ["Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct",
"Nov", "Dec"]) # renaming by month
plt.legend()
plt.grid(True)
plt.show()


# In[ ]:


# PIE CHARTS
```

```
# Filter for 'Travellers' only
df_filtered = df[df["Traveller type"] == "Travellers"].copy()

# Collect all travel modes
travel_modes = [
    "United States of America residents, air",
    "United States of America residents, land",
    "United States of America residents, water",
    "Canadian residents returning from the United States of America, air",
    "Canadian residents returning from the United States of America, land",
    "Canadian residents returning from the United States of America, water",
]

# Removes any commas from string values and convert each column into an integers
for col in travel_modes:
    df_filtered[col] = df_filtered[col].str.replace(",", "").astype(int)

# Preparing data for plotting the sum across all years (2000-2025)
labels = ["Air", "Land", "Water"]

# United States of America residents
us_data = [
    df_filtered[travel_modes[0]].sum(),
    df_filtered[travel_modes[1]].sum(),
    df_filtered[travel_modes[2]].sum(),
]

# Canada residents
canada_data = [
    df_filtered[travel_modes[3]].sum(),
    df_filtered[travel_modes[4]].sum(),
    df_filtered[travel_modes[5]].sum(),
]

# Create Pie Charts
fig, axes = plt.subplots(1, 2, figsize=(12, 6))

# Structure output so that pie charts are beside each other
# Use autopct="%1.1f%%" to display the percentages with one decimal point
# U.S. Residents Pie Chart (left)
axes[0].pie(us_data, labels=labels, autopct="%1.1f%%", startangle=90, colors=["skyblue",
"lightgreen", "salmon"])
axes[0].set_title("U.S. Residents Entering Canada")

# Canadian Residents Pie Chart (right)
axes[1].pie(canada_data, labels=labels, autopct="%1.1f%%", startangle=90, colors=["lightblue",
"limegreen", "tomato"])
axes[1].set_title("Canadian Residents Returning from U.S.")

# Plot graphs
```

```
plt.show()
```