International tourist's traveling spending in Canada*

Analysis on Canada 2017 International Travel Survey (ITS)

Xuetong Tang

04/27/2022

Abstract

Tourism is a significant component of Canadian economy. This paper generally investigates the relationship between traveler's total spending in Canada and the characteristics of their trip. By using dataset created by the Canadian 2017 International Travel Survey (ITS), we construct some main variables and create histograms analyzing the distribution of traveler's total spending and trip characteristics such as trip reason, total days, carrier class and entry place. Also, multiple linear regression model has been built to show the relationship between total spending and trip characteristics. We find that variable trip reason, total days and entry place all are significant predictors to total spending.

Contents

*Code and data are available at: https://github.com/BellaTang12/Canada_tourism.git

Keywords— Accommodation, Canada Border Services Agency (CBSA), International travel, Transportation, Travel and tourism, Trip, Visitors

1 Introduction

Nowadays, with the development of globalization, international travel has been a common component of people's life. Tourism contributes to countries' GDP and cultural development a lot. In 2019, contribution of travel and tourism to GDP for Canada was 6.5 % ("Canada - Contribution of Travel and Tourism to GDP as a Share of GDP" 2019). Over the last few decades, a rising number of travelers from a growing number of nations have chosen Canada as their business or leisure destination. Almost all visitors visiting Canada in the 1950s came from the United States. The shared border with the United States is still crucial to the Canadian tourism economy; However, in 2015, three out of ten tourists came from outside of Canada (G. of Canada 2018). Considering the significance of tourism in Canada's economy, this paper aims to analyze the relationship between traveler's total spending in Canada and their trip characteristics. In this paper, we use dataset provided by International Travel Survey (ITS) conducted by Statistics Canada.

Statistics Canada has been conducting the International Travel Survey (ITS) since 1920 to satisfy the requirements of the Balance of Payments of the Canadian System of National Accounts (BOP). Over time, questions aimed at gathering extensive information on travellers for market research and industry planning were gradually included to the survey. Today, the ITS provides a comprehensive set of statistics on the number of foreign travelers as well as specific details on their visits, such as expenditure, activities, places visited, and length of stay. In addition to fulfilling BOP requirements, the Tourist Satellite Account (TSA), Canada Border Services Agency (CBSA), Destination Canada, provincial tourism agencies, the United States Department of Commerce, and a variety of private sector sectors are all using the ITS. The ITS is also used to submit reports to international organisations including the World Tourism Organization (WTO), the Organization for Economic Cooperation and Development (OECD), and the Pacific-Asian Tourism Association (PATA). The ITS is based on data gathered from only a part (sample) of the population and thus the results is only estimates of the true values for the travelling population, values which could only be obtained through a census (S. Canada 2018).

The paper would be consisted of mainly four parts, which are data part, model part, result part and the discussion part. In data section, brief analysis based on traveler's trip characteristics would be conducted to have an overview of the distribution of different types of travelers and their trips. In model part, we would build linear regression model based on our research focus which is to analyze the relationship between traveler's total spending and their trip characteristics in Canada and the result would be put in the model result part. Ultimately, the discussion about the phenomenon we analyzed before will revealed in the discussion part at the end of the paper.

By analyzing the dataset, this paper finds that the trip spending by travelers from countries other than United states is obviously higher than travelers from the Unted States. Also, it is witnessed that travelers who travel for holiday or personal pleasure reason have higher total spending on trips than those travelers who travel to Canada for business and other reasons.

All the data analysis in this paper uses R studio (R Core Team 2021) with tidyverse (Wickham et al. 2019), ggplot2 (Wickham 2016), kableExtra (Zhu 2021) and lmtest (Zeileis and Hothorn 2002) packages.

2 Data

2.1 Methodology and Data collection

The dataset used in this paper is provided by International Travel Survey (ITS) conducted by Statistics Canada. The survey was conducted from January 1st to December 31st 2017. The target population of the dataset includes all travelers on entry or re-entry into Canada geographically covering all provinces and territories, census metropolitan area (CMA). The dataset is released on July 21st 2020. There's 12866 observations in the dataset after data cleaning (S. Canada 2018).

The frontier counts and questionnaires are the two main components of the ITS. Both of these approaches rely heavily on the CBSA's cooperation in gathering data on border crossings and distribution of questionnaires. The survey is used to collect data on the characteristics of foreign travellers and journeys on a quarterly basis. The purpose of the trip, the size of the travelling party, the places visited, the activities participated in during the trip, the length of the trip, and the spending on the trip are all included in these information provided by the survey. The Canadian Balance of International Payments is updated using some sections of the survey. In addition, estimates of foreign trip and traveler characteristics on a quarterly and annual basis are used by the federal and provincial governments, the the tourism industry, businesses and the general public. The surveys are mailed back to you. The questionnaires are contained by a mail-back and internet questionnaire survey, as well as an air-exit survey of international and US travelers described as follows (S. Canada 2018).

- Mail-back and electronic questionnaires: This is a sample poll in which only a portion of the foreign traveller population receives invitation cards. On entry or re-entry to Canada, authorities from Canada Border Services issue out an invitation card to travelers. The following people will receive cards: 1. Visitors from the United States or other nations; 2. Canadian residents returning from visits outside the country; After the trip, the recipients of the invitation card are asked to complete the electronic questionnaire online using the web link supplied on the card. Based on the previous year's traffic, a stint distribution system has been developed to survey international travelers. A stint is a period of time, usually several days, during which invitation cards are delivered to all qualified travelers as defined above. Each port of entry participating in this scheme is given a particular number of cards and a start date for each of its stints. Officers give the cards to the appropriate travelling population on a continual basis on the start date until all cards have been distributed. In 2017, around 400,000 invitation cards were given to all Canadian and international travelers (S. Canada 2018).

- Air Exit Survey of Visitors to Canada: The Air Exit Survey (AES) collects additional questionnaires from US and international travellers returning straight to their country of origin by commercial air. The AES focuses on the countries from where we receive the most visits such as the United Kingdom, France, Germany, and Australia as well as a number of growing markets such as China, Japan, India, and South Korea. While they wait for their return flights to these target overseas countries, Statistics Canada interviewers conduct personal interviews with overseas travellers. Every month, during a collecting period of 5 to 7 days, these interviews are done at international airports in five locations (Halifax, Montreal, Toronto, Calgary, and Vancouver). Around 7,500 interviews were completed in 2017. The interviewing crew is made up of interviewers with various language abilities, allowing interviews to be performed in the travelers' native language where possible. The questionnaire is accessible in ten different languages. Commercial flight travelers from the United States fill out the AES at international airports in Toronto, Vancouver, Montreal, Ottawa, Calgary, Edmonton, and Halifax. Approximately 2,500 interviews were done in 2017 (S. Canada 2018).

The replies collected from questionnaire surveys are treated as a simple random sample from the total traffic in each stratum for estimating purposes (port or group of ports, by type of traffic, by quarter). The

statistics may be skewed by "distribution bias," which occurs when not all types of travelers are represented in the distribution, or "non-response bias," which occurs when the people who respond are not necessarily representative of the travelling population (S. Canada 2018).

2.2 Data cleaning

In the process of data cleaning, all "valid skip," "Don't know," "Refusal" and "Not stated" value in the questionnaire answer are removed. Also, all missing values are removed. Then, we create a new dummy variable *carrier_class* which indicates the type of fare used in transportation when entering Canada.

We select the following variables from the dataset for our investigation:
- 1. *total_spending* (response variable): Sum of global spending in all places visited reported by respondents. We want to investigate the relationship between total spending of travelers during one visit and the characteristics of travelers in Canada.
- 2. *total_days* (numerical explanatory variables): The total number of days spent in Canada in one visit reported by respondents.

We also choose 3 dummy explanatory variables to discuss different conditions in our research problem:
- 3. *trip_reason*: The purpose of trip to Canada reported by respondents which includes 6 levels: 1. Holidays, leisure or recreation; 2. To visit friends or relatives; 3. Other personal - pleasure; 4. Personal reason - other; 5. Attend a conference, convention or trade show; 6. Other business reason.
- 4. *entry_place*: The route of entry into Canada reported by respondents which includes 3 levels: 1. From the United States only; 2. Directly from another country other than the United States; 3. From another country via the United States.
- 5. *carrier_class*: The type of fare used in transportation when entering Canada which includes First class, Business class, Economy class, Charter and Travel reward program.

There is no similar dataset because this survey was conducted from January 1st to December 31st 2017. It takes a long time to collect the data and it include a large number of populations. The survey is not conducted every year so that there is almost no similar dataset.

2.3 Data visualization

Table 1: Summary table of traveler's total spending in Canada

median	mean	sd	Q1	Q3	IQR	max	min
965	1892.158	2710.761	325	2340	2015	52000	0

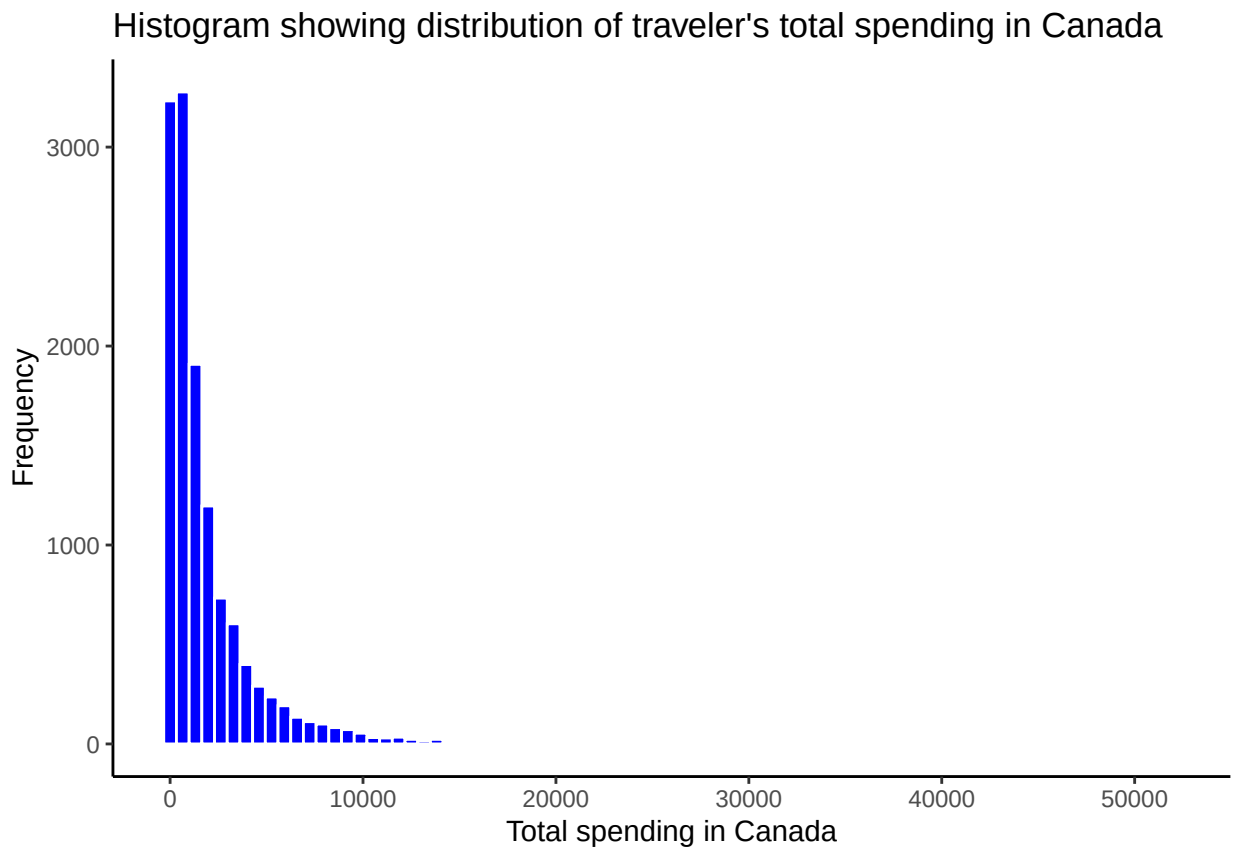

Figure 1: A histogram

To give a brief glance of our sample, a histogram 1 of *total_spending* could be created to reveal the general distribution of tourists' total spending in Canada. In this histogram, we could see that the distribution is right-skewed, which illustrates that the majority of total expenditure are approximately between 0 to 10000, and the highest total spending of the chosen respondents is 52000. Also the table 1 shows that the average total spending is 1892.158.

Figure 2: A barplot

To analyze traveler's entry route into Canada, the barplot 2 above illustrates the distribution of traveler's entry place into Canada with each color representing different trip purpose. From the barplot, we find that the majority of travelers come directly from countries other than the United States. Around 5000 travelers come from the United States and only less than 2000 travelers come from another country via the United States. What's more, the colors representing different trip reasons show that a large portion of travelers who come directly from countries other than the United State come to Canada to visit friends or relatives while the majority of travelers from the United States come to Canada for business.

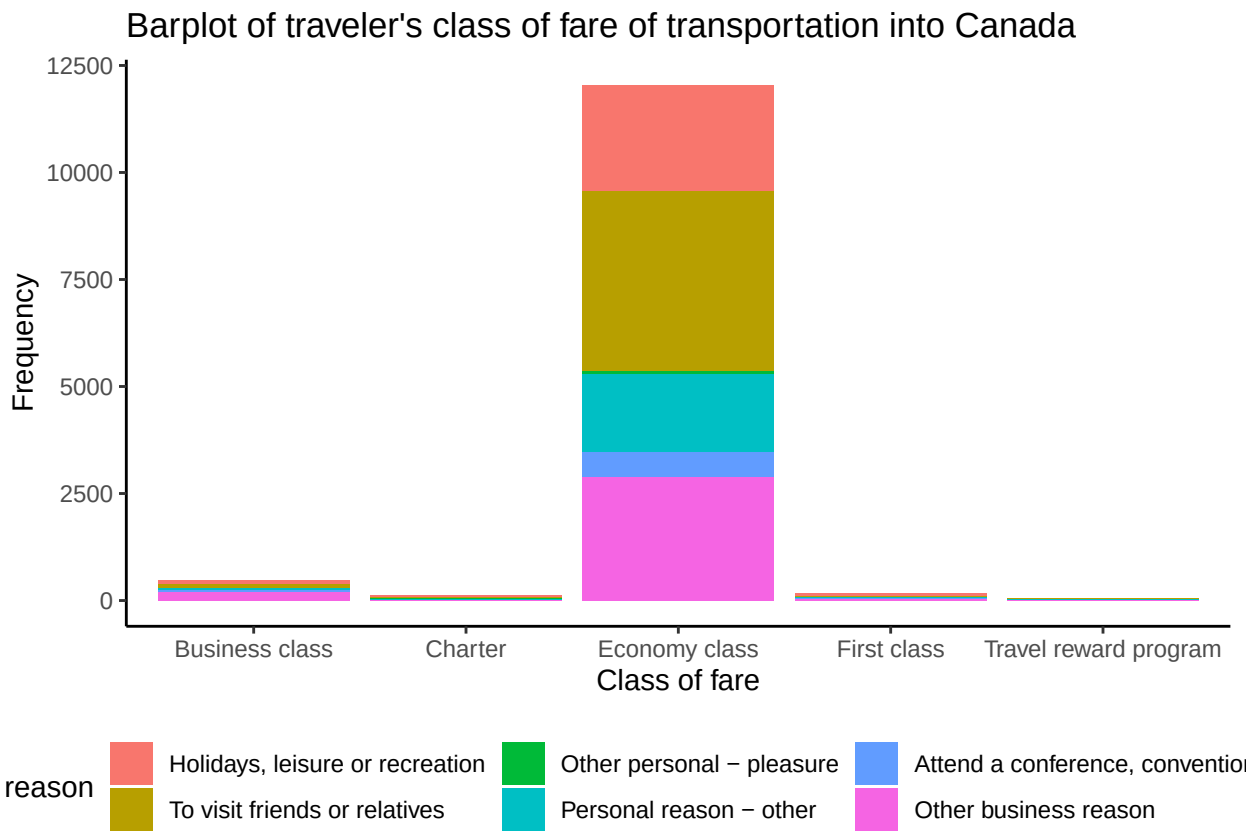

Figure 3: A barplot

To analyze traveler's class of fare of transportation into Canada, the barplot 3 above illustrates the distribution of traveler's class of fare of transportation into Canada with each color representing different trip reason. From the plot, we could see that the majority of travelers choose economy class when entering Canada, especially those travelers come for business and visiting friends or relatives.

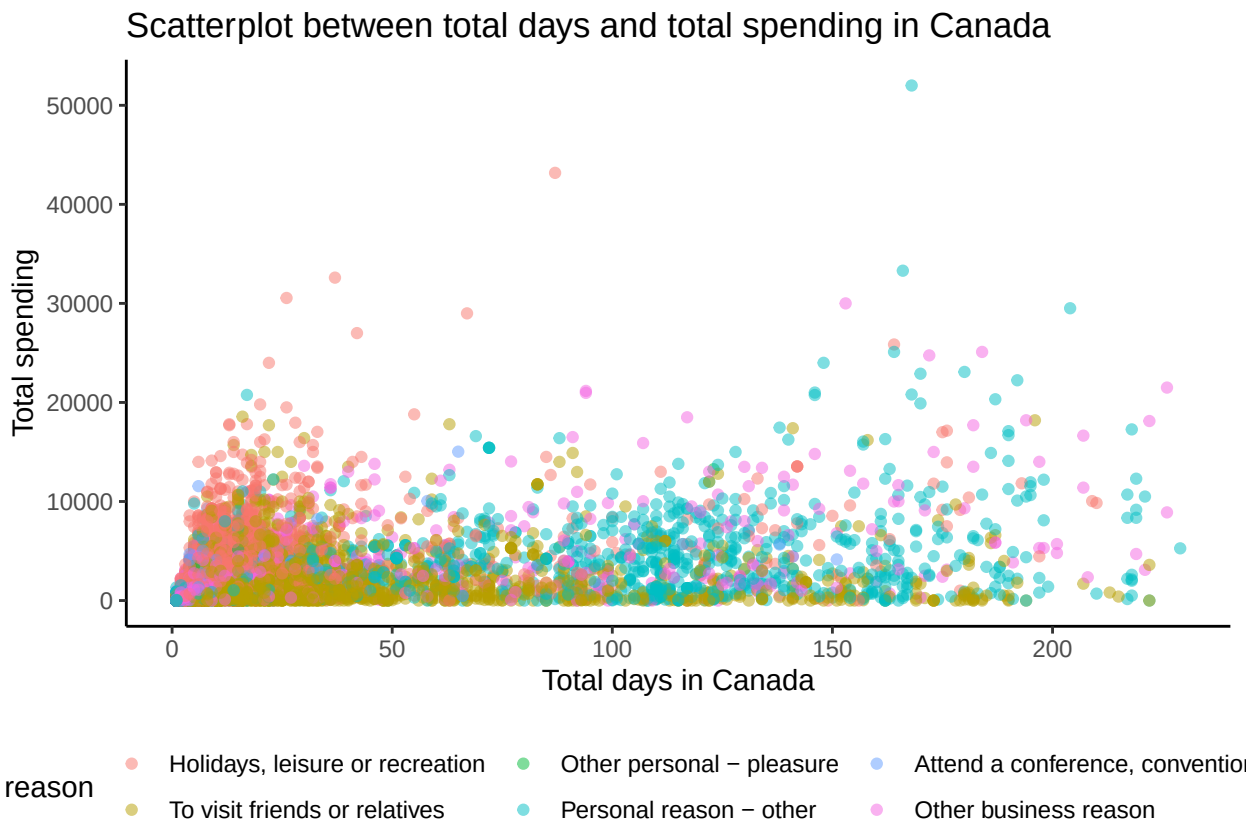

Figure 4: A scatterplot

To give a brief view of the relationship between traveler's total days and total spending in Canada, we construct the scatterplot 4 above. In the scatterplot, different trip reasons are shown by different colors. From the plot, we find that travelers who visits for holidays, leisure or recreation usually stay less than 50 days in Canada while travelers who visits to see friends or relatives may stay longer. On the other hand, the total staying time in Canada for travelers who visits for business reasons or other personal reasons is more evenly distributed. What's more, from the plot, we cannot see a explicit relationship between total days in Canada and total spending in Canada.

3 Result

3.1 Relationship between total days in Canada and total spending

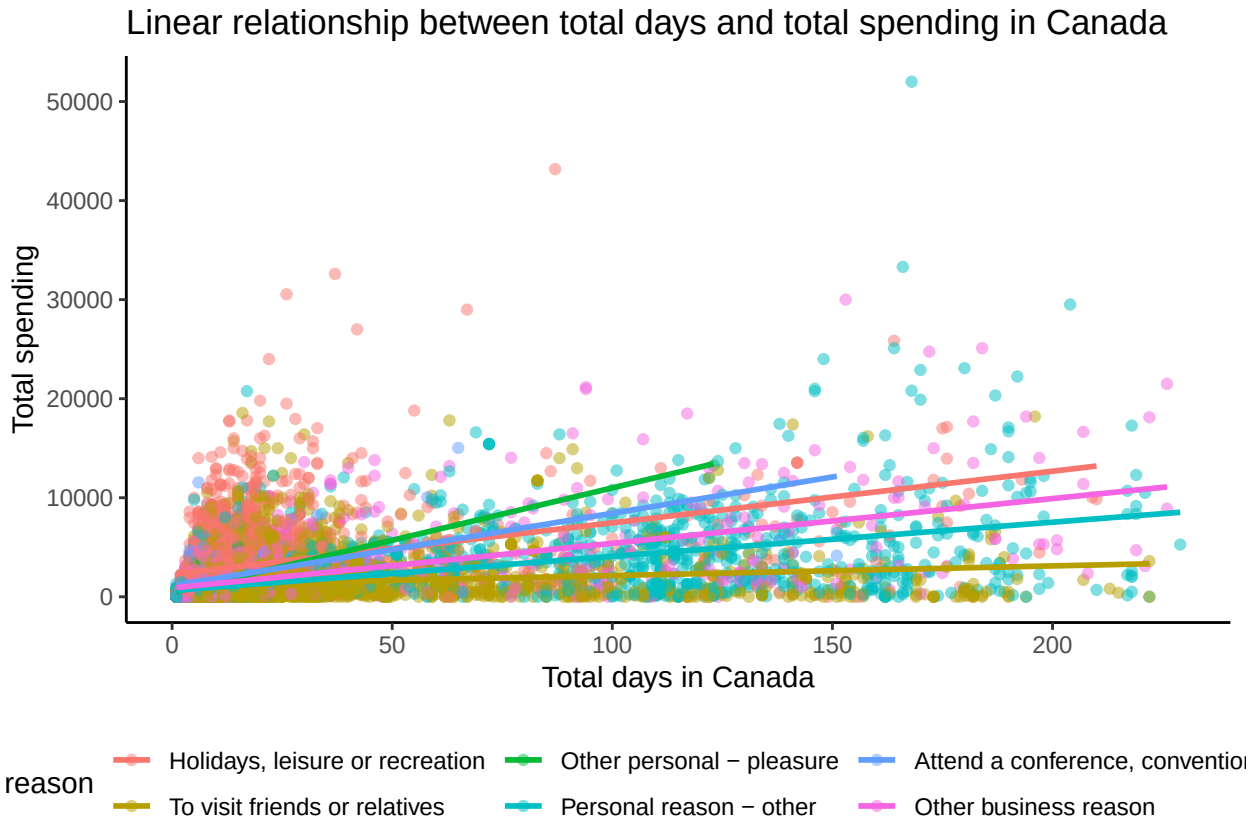

Figure 5: A scatterplot

To analyze the relationship between traveler's total days and total spending in Canada under different trip reason, we construct the linear regression model 5 above. The regression lines shown with different color represents different trip reasons. From the graph, we could clear see that, in general, as traveler's total days in Canada increases, their total spending increases. More specifically, When total days in Canada remain constant, travelers who travel for personal pleasure spend the most; Travelers who travel to attend a conference, convention ot trad spend the second highest. Travelers who travel for holidays spend the third highest. Travelers who travel for business spend the forth highest. Travelers who travel for other personal reason spend the second least. Travelers who travel to visit friends or relatives spend the least.

3.2 Model

Multiple linear regression is a statistical approach that predicts the outcome of a dependent variable using two or more independent variables. Here, to have a better view on the whole picture of our research problem to discuss the relationship between traveler's total spending and the characteristics of their trip, we build a multiple linear regression model which conveys the relationship between *total_spending* and all explanatory variables we considered in our research including *total_days*, *carrier_class*, *entry_place*, *trip_reason*. We also generate one reduced model which only include *total_days*, *entry_place* and *trip_reason* as explanatory variables. By comparing the full model and reduced model, we will know which model predict traveler's total spending during their trip in Canada better.

3.2.1 Model selection

In this paper, we use the likelihood ratio test for model selection. The likelihood-ratio test assesses the goodness of fit of two competing statistical models based on the ratio of their likelihoods. Likelihood describes the probability of the observed data as a function of the parameters of our chosen statistical model. In likelihood ratio test, our null hypothesis (H_0) is: the simpler model (reduced model) explains the data just as well as the more complicated model (full model). Our alternative hypothesis (H_a) is: the simpler (reduced model) model does not explain the data as well as the complicated model (full model).

We choose a model based on p-values generated in the test. If p-value is less than 0.05, it means that there's strong evidence against the null hypothesis which indicates that the full model is better. In contrast, if p-value is larger than 0.05, it means that there's no evidence against the null hypothesis which indicates that we should use the reduced model.

Table 2: Summary of likelihood ratio test

#Df	LogLik	Df	Chisq	Pr(>Chisq)
14	-118196.4	NA	NA	NA
10	-118213.2	-4	33.6112	9e-07

Above table 2 is the likelihood ratio test outcome on these two models. Comparing model_full and model_reduced, model_full is better since the p-value we get from the likelihood ratio test is 9e-07, it is a value much smaller than 0.05, which means we have strong evidence against the null hypothesis which indicates that the full model is better. Thus, We choose the reduced model as our final model for further analysis.

Thus, our final model is a multiple linear regression model which conveys the relationship between *total_spending* and explanatory variables including *total_days*, *entry_place*, *trip_reason*. The equation of the multipe linear regression model is shown as below:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \epsilon_i$$

where:
- β_0 is the term of intercept
- $\beta_1 ... \beta_4$ are the coefficients for each explanatory variable
- $i = 1, \ldots, n$ where n is the number of observations in the sample data
- ϵ_i is the i^{th} error term for this regression model

3.2.2 Result for the model

Table 3: Summary of simple regression model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3041.54	53.45	56.90	0
total_days	26.81	0.64	42.03	0
entry_placeOther country via United States	-357.46	71.38	-5.01	0
entry_placeUnited States	-985.97	50.27	-19.61	0
trip_reason2	-2050.28	58.75	-34.90	0
trip_reason3	-1550.86	219.32	-7.07	0
trip_reason4	-1660.46	73.52	-22.59	0
trip_reason5	-1110.26	103.90	-10.69	0
trip_reason6	-1268.91	63.39	-20.02	0

Above is the summary table 3 of the model. We could interpret the estimated value of coefficients in the summary table as follows:

- Interpretation for numerical explanatory variable *total_days*:
  The estimated coefficient β_1 of variable *total_days* is 26.81. This means that when all else remain constant, as *total_days* increases by one unit, the expected value of *total_spending* changes by 26.81 unit.

- Interpretation for dummy variable *entry_place*:
  The estimated coefficient β_2 is -357.46. This means that when all else remain constant, as traveler's entry place is other country via United States, the expected value of traveler's *total_spending* is 357.46 units lower than those travelers who enter directly from countries other than United States. The estimated coefficient β_3 is -985.97. This means that when all else remain constant, as traveler's entry place is directly United States, the expected value of traveler's *total_spending* is 985.97 units lower than those travelers who enter directly from countries other than United States.

- Interpretation for dummy variable *trip_reason*:
  The estimated coefficient β_4 is -2050.28. This means that when all else remain constant, as traveler's trip purpose is to visit friends or relatives, the expected value of traveler's *total_spending* is 2050.28 units lower than those travelers whose trip reason is holidays, leisure or recreation. The estimated coefficient β_5 is -1550.86. This means that when all else remain constant, as traveler's trip purpose is other personal pleasure, the expected value of traveler's *total_spending* is 1550.86 units lower than those travelers whose trip reason is holidays, leisure or recreation. The estimated coefficient β_6 is -1660.46. This means that when all else remain constant, as traveler's trip purpose is other personal reason, the expected value of traveler's *total_spending* is 1660.46 units lower than those travelers whose trip reason is holidays, leisure or recreation. The estimated coefficient β_7 is -1110.26. This means that when all else remain constant, as traveler's trip purpose is to attend a conference, convention or trade show, the expected value of traveler's *total_spending* is 1110.26 units lower than those travelers whose trip reason is holidays, leisure or recreation. The estimated coefficient β_8 is -1268.91. This means that when all else remain constant, as traveler's trip purpose is other business reason, the expected value of traveler's *total_spending* is 1268.91 units lower than those travelers whose trip reason is holidays, leisure or recreation.

From above estimation of linear model, we find that there's a positive relationship between total days travelers spent in Canada and their total spending. Moreover, different trip reasons influence traveler's total

spending a lot. Travelers who travel for holiday and leisure reason on average spend the most during their visit to Canada. In contrast, travelers who visits to see friends or relatives spend the least. Also, we can see that travelers who travel for personal pleasure on average spend more than travelers who travel for other personal reason. On the other hand, the entry place of travelers also matters. Traveler who enter directly from countries other than United States usually spend the most while travelers who enter directly from United States tend to spend the least.

4 Discussion

4.1 What this paper done

This paper aims to investigate the relationship between traveler's total spending during their trip in Canada and their trip characteristics. To analyze this research question, we set *total_spending* as our response variable and choose variables *total_days* which indicates how many days traveler has spend in Canada, *carrier_class* which indicates traveler's class of fare of transportation into Canada, *trip_reason* which indicates traveler's trip purpose, *entry_place* which indicate from where travelers enter into Canada.

By using these variables in our dataset, we first construct histogram and barplots which provides us with brief information of distribution on variables *total_spending*, *entry_place*, *trip_reason* and *carrier_class*. We also build scatter plot on *total_spending* and *total_days* to have an overview of the relationship between them. To discuss the factors which influence tourists' spending, we do simple linear regression on variable *total_spending* and *total_days* separated by different *trip_reason*.

Then, to focus on our research question more specifically, we build multiple linear regression models which uses variable *total_spending* as response and include all variables *total_days*, *entry_place*, *trip_reason* and *carrier_class* as predictors. By model selection, we find that model without *carrier_class* can give use more precise estimation. Thus, we choose the reduced model which uses *total_spending* as response and only variables *total_days*, *entry_place* and *trip_reason* as predictors. We find that in general there's positive relationship between tourists' total days and their total spendings in Canada. Furthermore, various trip motives have a significant impact on a traveler's total spending. Travelers who come to Canada for vacation and pleasure spend the most money. Travelers who visit friends or family, on the other hand, spend the least. We can also see that travellers who travel for personal enjoyment spend more on average than tourists who go for other reasons. Travelers' entrance point is also very important. Travelers entering directly from countries other than the United States spend the most, while those entering directly from the United States spend the least.

4.2 Findings and discussion

According to the linear regression model, the scatterplots we made previously, our research question could be discussed. From above analysis on our research question we find that as tourists spend more days in Canada, their total spending increase. This make sense since as travelers stay longer, they need to spend more on food and hotel. We also find that travelers come from countries other than United States spend the most. In recent years, more and more Asian people travel to Canada for holidays. It is shown that China has accounted for the majority of the rise of Asian travellers since 1990. The number of Chinese tourists increased by 12.3% each year on average between 1990 and 2015. This compares to a total of 2.6 percent for Asia and 2.1 percent for all international travelers (G. of Canada 2018). We also see that travelers from United States spend the least, which may related to decline in the share of U.S. tourists. In 1947, Americans made up 98 percent of all visitors to Canada. By 1990, that percentage had dropped to 80%, and by 2015, Americans made up only 70% of overseas visitors (G. of Canada 2018).

What's more, we find that tourists who come for holidays and personal pleasure on average spend more than those who come to visit friends or relatives or for other business reasons. This is to some extent matches our common sense. On holidays, traveler travel more to enjoy the holiday and have incentives to spend more on shopping and other expansive services. When travelling for one's own pleasure, people always have more willingness to pay. In contrast, when trip purpose is to visit friends or relatives, people won't pay much attention on consumption which leads their total spending during the trip to be low.

4.3 Weekness of this paper

There's also some limitations of this paper. First, we only choose predictor that can be found in the survey to conduct the model estimating the relationship between total spending and traveler's trip characteristics. In reality, there are many factors other than the variable we chosen could contribute to traveler's spending in Canada. Our model don't include these effects. Second, the data collected is subject to non-response problems. When fulling out the questionnaires, many tourist choose to choose option "Don't know" to save time or for other reasons. This causes the non-response problem and make our dataset biased.

4.4 Next step

In further analysis, more characteristics of travelers will be include in our model in order to analyze the relationship between traveler's trip characteristics and their spending in Canada. Some aspects worth analyzing includes traveler's total income, gender, shopping willingness and so on. By discussing different factors that influence traveler's spending in Canada, we could design some different type of travel programs which attracts different population groups. By combining various services and trip experiences together, those travel plan could effectively provide travelers more incentives to spend more. In such way, Canadian tourism is developed and can contribute to Canada's economy and GDP much more.

5 Appendix

5.1 Data sheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - Statistics Canada has been conducting the International Travel Survey (ITS) since 1920 to satisfy the requirements of the Balance of Payments of the Canadian System of National Accounts (BOP).
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Statistics Canada created this dataset for Canadian government
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - Canadian government
4. *Any other comments?*
    - None

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - Individual. This dataset is provided by International Travel Survey (ITS)
2. *How many instances are there in total (of each type, if appropriate)?*
    - 12866
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
    - It is a sample. The ITS is based on data gathered from only a part (sample) of the population and thus the results is only estimates of the true values for the travelling population, values which could only be obtained through a census. The larger set is all travelers entering Canada.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
    - Each instance is consist of one survey response. It is recoded as numerical variables and dummy variables in the dataset.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - Each individual in the dataset ias an visiter id. For example: 100001
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

- No

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - No

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

   - No

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

   - No

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - It is self-contained.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - No

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No

16. *Any other comments?*

    - No

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was directly observable. The frontier counts and questionnaires are the two main components of the ITS.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The data was collected by questionnaires and frontier counts.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - It is collected by questionnaires and frontier counts. Every month, during a collecting period of 5 to 7 days, these interviews are done at international airports in five locations (Halifax, Montreal, Toronto, Calgary, and Vancouver).

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The interviewing crew is made up of interviewers with various language abilities, allowing interviews to be performed in the travelers' native language where possible. The questionnaire is accessible in ten different languages.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - Every month, during a collecting period of 5 to 7 days. Yes.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - I get the data from third party, Statistics Canada.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - Yes, they complete the quetionaires and interviews.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Yes but details not provided.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - No

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - No
12. *Any other comments?*
    - None

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - No
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - No
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - No
4. *Any other comments?*
    - None

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Statistics Canada has been conducting the International Travel Survey (ITS) since 1920 to satisfy the requirements of the Balance of Payments of the Canadian System of National Accounts (BOP).
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - http://odesi2.scholarsportal.info/webview/
3. *What (other) tasks could the dataset be used for?*
    - None
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - None
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - None
6. *Any other comments?*

- None

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
   - Yes
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - Data Liberation Initiative (DLI) (Statistics Canada) , https://www.statcan.gc.ca/eng/dli/dli
3. *When will the dataset be distributed?*
   - April 21, 2020
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
   - Copyright (c) Statistics Canada DLI License
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
   - No
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
   - No
7. *Any other comments?*
   - None

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - Statistics Canada
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - Data Liberation Initiative (DLI) (Statistics Canada) , https://www.statcan.gc.ca/eng/dli/dli
3. *Is there an erratum? If so, please provide a link or other access point.*
   - None
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
   - No, it won't be updated.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

- There's no limits anounced by Statistics Canada.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

   - Yes, it would be supported by Statistics Canada.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

   - Others can contact Data Liberation Initiative (DLI) (Statistics Canada) , https://www.statcan.gc.ca/eng/dli/dli to extend/augment/build on/contribute to the dataset.

8. *Any other comments?*

   - None

Reference

"Canada - Contribution of Travel and Tourism to GDP as a Share of GDP." 2019. https://knoema.com/atlas/Canada/topics/Tourism/Travel-and-Tourism-Total-Contribution-to-GDP/Contribution-of-travel-and-tourism-to-GDP-percent-of-GDP.

Canada, Government of. 2018. "The Evolution of Canadian Tourism, 1946 to 2015." https://www150.statcan.gc.ca/n1/pub/11-630-x/11-630-x2017001-eng.htm.

Canada, Statistics. 2018. "International Travel Survey, 2017: US and Overseas Visitors to Canada." https://CRAN.R-project.org/doc/Rnews/.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10. https://CRAN.R-project.org/doc/Rnews/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.