Modeling the Characteristics of Airport Travelers: Milestone Report

Melanie Malinas
Springboard Data Science Career Track
Capstone Project 1

Introduction and Problem Statement:

Travel and tourism is a huge industry. According to the World Travel and Tourism Council, the world travel and tourism sector grew 3.9 percent in 2018, outpacing the global GDP growth of 3.2 percent, contributing $8.8 trillion and 319 million jobs to the global economy. In 2018, it generated 10.4 percent of all global economic activity. There is therefore a huge opportunity to use data science to advance the travel and tourism industry.

According to the same report, 21.5 percent of travel and tourism spending in 2018 was on business. If you look at just airlines, however, business travelers represent a huge portion of an airline's profitability. Business travelers account for 12% of an airline's passengers, but are usually twice as profitable - in fact, on some flights, business travelers make up 75% of an airline's profits. This is because businesses these days are willing to pay more to book last-minute flights for their employees, or book non-stop flights or first-class/business class seats. It would therefore be useful to use data science be able to distinguish between business and non-business travelers for the purpose of better targeting these different types of travelers.

For this project, I will be using the San Francisco International Airport (SFO) Customer Survey from 2017 to build a model of the profiles of business vs. non-business travelers. This survey asked over 2,500 SFO travelers, mostly in person at their gates, about their experience with SFO as well as information about their trip and demographic information. Using this information, I will statistically analyze the characteristics of SFO travelers as well as use machine learning to predict who is a business traveler and who is a non-business traveler.

Data Cleaning:

I acquired this survey data freely from DataSF (San Francisco Open Data), a project of the city of San Francisco. DataSF has many different datasets from different sectors of San Francisco. The survey data was downloadable in CSV format, which made it simple to open and work with. Each row in the survey was a different survey respondent, and each column was a question on the survey. A data dictionary described the different questions and the codes for each answer. Almost all of the data was categorical data, with the answers coded as 1, 2, 3, etc. with 0 for blank/non-response.

The data also had weights for each respondent that were given to try and make the survey more representative of the population. For various different questions, I created dictionaries with weighted counts for each answer. For example, when the data was weighted, the number of pleasure travelers went down and the number of business travelers went up.

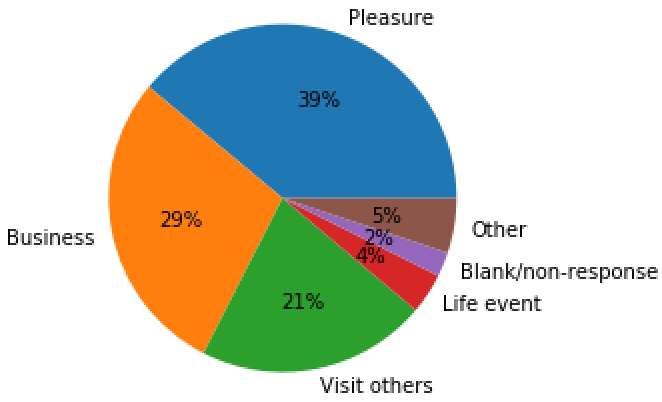
Another step I took to clean the data was to rename several of the columns. I renamed 'Q2PURP1' to 'Purpose', 'Q3GETTO1' to 'Transportation', etc. This made the column names better formatted and easier to read. I also mapped the actual descriptions of the answers from the data dictionary onto their codes to make the data more readable.

Other than that, this data was pretty clean, well-organized, and easy to work with.

Exploratory Analysis:

For the first part of my project, I did exploratory data visualization and statistical analysis. I found several interesting trends in the data that could help distinguish between business and non-business travelers.

Looking at the weighted dictionary of counts, I found that about 29% of travelers were business travelers. I created a new column classifying all travelers as either business or non-business travelers.



I then looked at correlations between variables. Because the data is categorical data, it is not possible to directly compute correlations. I therefore used something called Cramer's V to compute the correlations. Cramer's V is based on the chi-squared statistic:

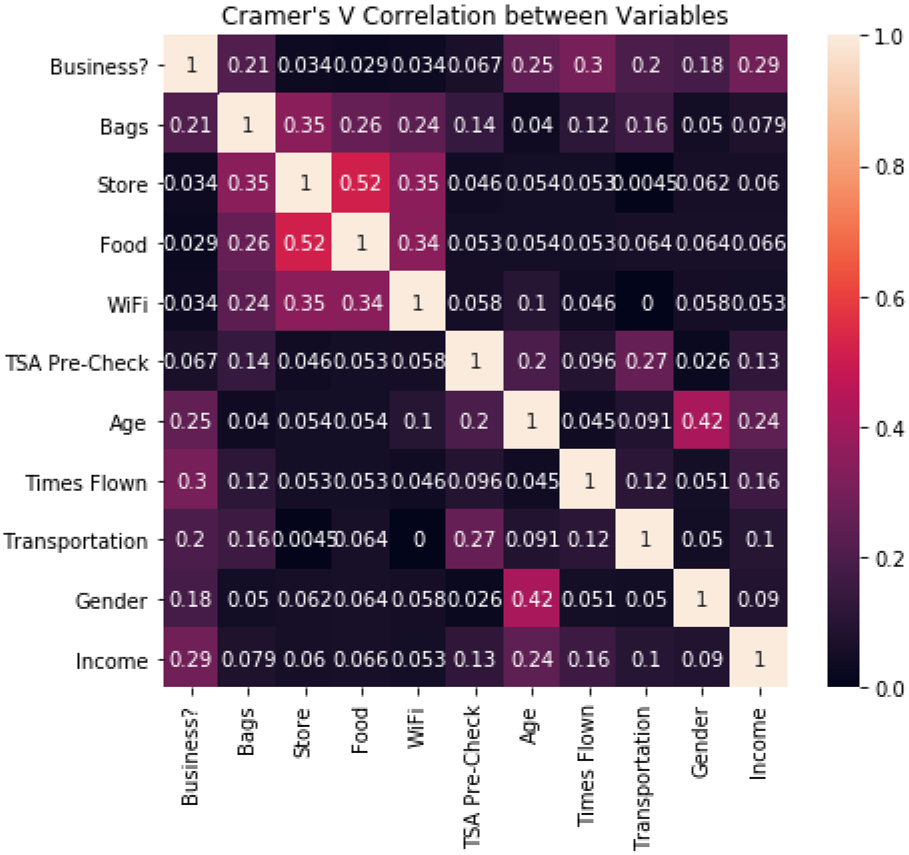
$$V = \sqrt{\frac{\varphi^2}{\min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

Here, φ is the phi coefficient, χ^2 is the chi-squared statistic, n is the number of observations, and k and r are the number of columns and rows, respectively.

The columns of interest that I looked at were business/non-business (Business?), whether bags were checked (Bags), whether someone went to a store (Store), whether someone bought food (Food), whether someone used the free WiFi (WiFi), whether someone went through TSA

Pre-Check (TSA Pre-Check), the traveler's age (Age), the number of times flown out of SFO in the past year (Times Flown), the type of transportation used to get to the airport (Transportation), the gender of the traveler (Gender), and the traveler's household income (Income).

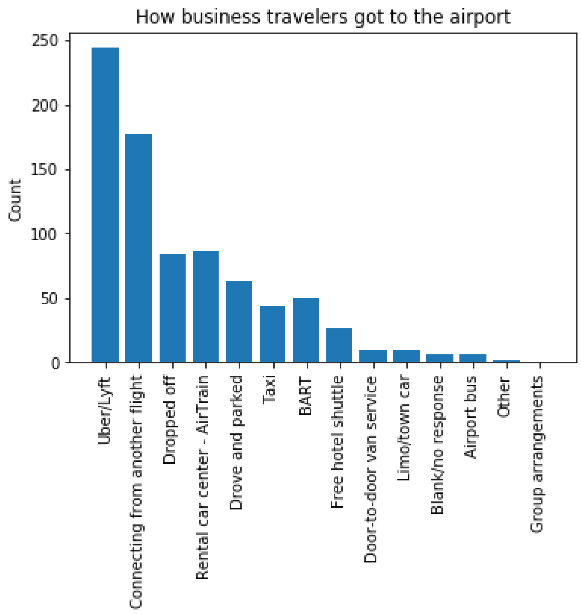
I then created a heatmap of the Cramer's V values:

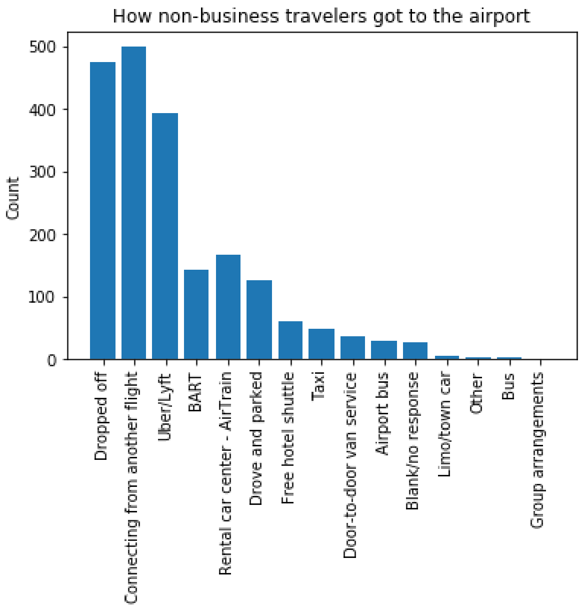


The highest correlation was between whether someone went to a store and whether they bought food, with a Cramer's V value of 0.52. The next highest correlation was between gender and age. Examining the gender and age breakdown more closely, it appears that there are more female travelers between the ages of 18 and 34, whereas there are more male travelers between the ages of 35 and 54. At ages 55 and over it appears that the gender breakdown of travelers is about equal.

In terms of correlations between whether someone was a business traveler and the other categories, I saw small correlations between that and Bags, Age, Times Flown, Transportation, Gender, and Income. These are therefore good categories to explore more in depth.

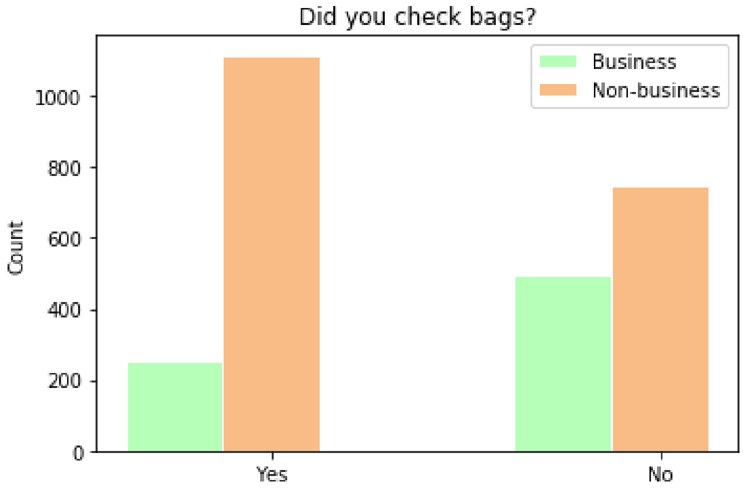
The next thing I did in my exploratory analysis was to compare business travelers and non-business travelers in my columns of interest. I made graphs comparing how business and non-business travelers got to the airport:





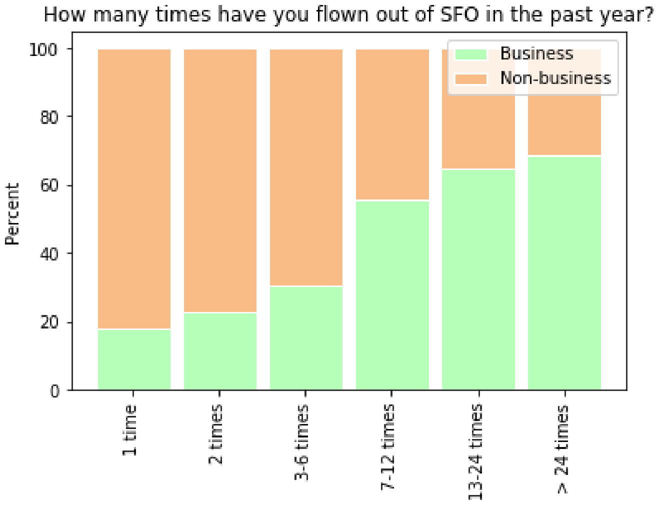
As you can see, the most common way that business travelers got to the airport was using a ridesharing service such as Uber or Lyft, whereas the most common way that a non-business traveler got to the airport was being dropped off. The proportion of business travelers using Uber or Lyft was statistically significantly higher than the proportion of non-business travelers using Uber or Lyft ($p<0.00001$). This therefore is a good category to use in predicting whether or not someone is a business traveler.

I also examined the breakdowns between business and non-business travelers on whether the traveler checked a bag, and found that most business travelers did not check a bag, while most non-business travelers checked a bag. This difference was statistically significant.



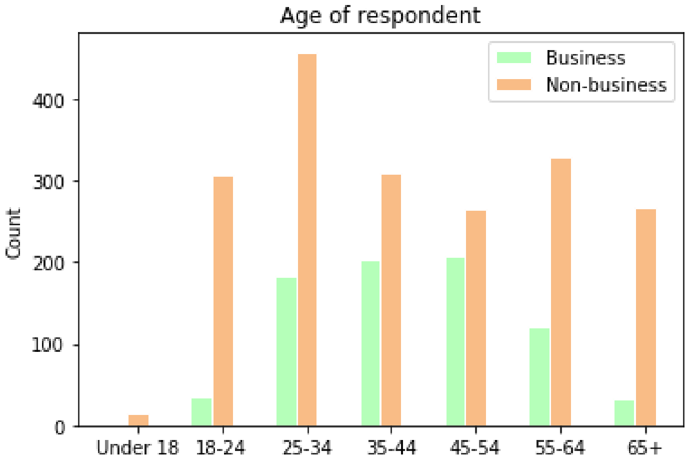
I also examined the variables of whether someone had visited a store at the airport, bought food, or used the free WiFi, and I found that there were no significant differences between business and non-business travelers in these categories.

I then looked at how many times a traveler had flown out of SFO in the past year. For this I created a stacked bar chart, where each bar shows the percentage of each category that are business travelers vs. non-business travelers:



As the number of times flown out of SFO increases, the proportion of travelers who are business travelers increases. The proportion of travelers flying more than 13 times per year is significantly different between business and non-business travelers.

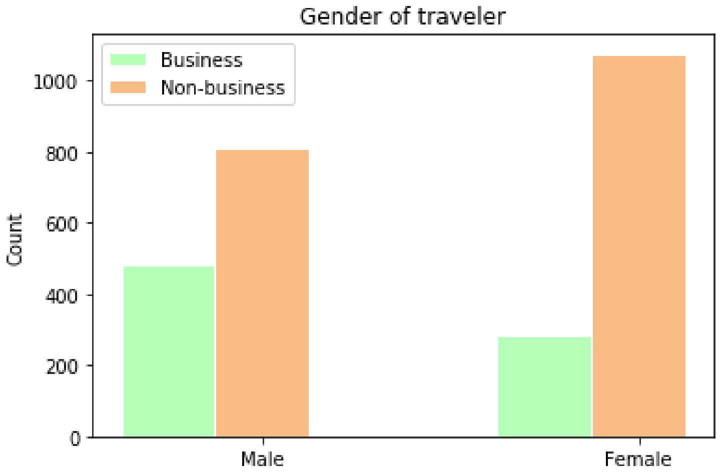
I also looked at the distribution of ages of the travelers:



Business travelers appear to be mostly in the range of 25-64, with very few being age 18-24 or 65 and over. By contrast, the non-business travelers have large cohorts of travelers ages

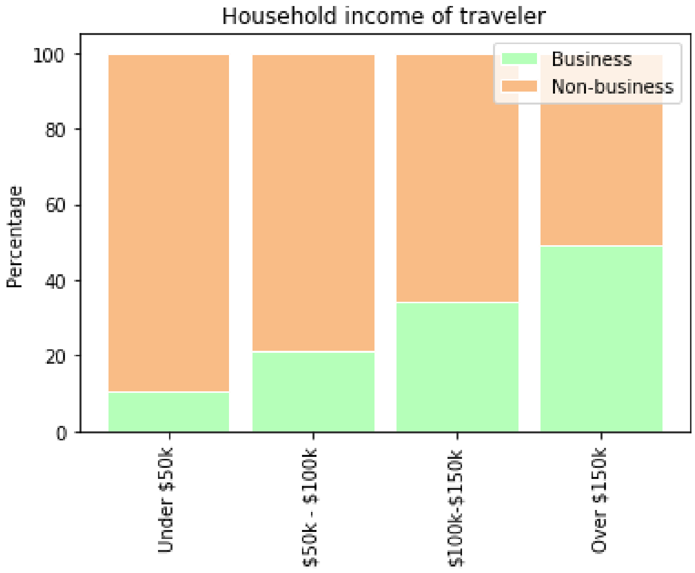
18 to 24 and 65 and over. However, the average age of the travelers, around 43 years, was not significantly different between business and non-business travelers.

I then looked at the gender of the travelers:



Business travelers were majority male, and non-business travelers were majority female, and this difference was statistically significant. It is perhaps not surprising that the business travelers were majority male, as it seems more likely to have male businessmen that travel for work. However, it was surprising to me that the non-business travelers were majority female. I am not sure why this is or if it is actually representative of SFO travelers. I would think that the weighting would have corrected for this, but it did not.

Finally, I analyzed household income levels for business and non-business travelers using a stacked bar plot:



Perhaps unsurprisingly, I found that as the household income level went up, the proportion of business travelers also went up. The average household income for a business traveler was about $134,000, which was statistically significantly different from the average household income for a non-business traveler, $97,000.

The categories of Bags, Age, Times Flown, Transportation, Gender, and Income are therefore all good variables for predicting whether someone is a business traveler, which is what I will do in the next section using machine learning.