Interactive Data Science HW2

Assignment Goals

1. Inspecting trends of travellers from different countries during different seasons, years and times of the year.
2. Comparing and analyzing the length of stay at any hotel based on whether the travellers have children or not.
3. Finding trends of cancellations with respect to number of days before the booking was made, and also what type of customers are more prone to cancelling in what type of stays.

Dataset Description

Our dataset is the **Hotel booking demand** dataset [Dataset Link]. This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

The data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.

Questions and Explanations

What is the trend for travellers from different countries through the different years or the various months of the year?

Hypothesis

We aim to analyze if there is a seasonal trend to travellers from different countries. For example, say a country like Canada, we have a hypothesis that there would be more travellers in the summer months of November, December, January and February when it gets really cold in Russia, since people are looking for respite from cold weather. Also, we are looking to find if there are particular years when there were more travellers from a country. We see that the hypothesis is not quite true.

Explanation

There are other more powerful factors governing the most popular months for travel. We find that travellers from most countries like the United States, Russia, Brazil, France etc. prefer to travel in the summer months of April, May, June, and July. A possible explanation for this is that children have school vacations during that time and so families are able to travel together. Another noticeable trend is that contrary to our hypothesis, December and January, which are generally the coldest months have least travellers. A possible explanation for this is that December and January are christmas and new year months. Hence, there are very few business travellers and people prefer to spend time at home with their families. Also, we find that there is much less data for the year 2015 compared to the years 2016 and 2017. For the years 2016 and 2017, the number of travellers are more or less similar across countries.

How does the length of stay vary depending on the number of adults and children who are travelling together?

Hypothesis

Our aim is to analyze if there is a correlation between length of stay and number of adults and children travelling together. Typically we expect that the length of stay will be longer when a person is travelling with family ie. with more adults and children compared to when he/she is travelling alone.

Explanation

From the data we find that when only 1 adult is traveling, the stay nights are less than 3 for over 60% of the observations. As the number increases to 3 adults, less than 30% of the observations correspond to less than 3 night stays. These results are for when there are no children. It indicates that groups tend to travel for longer length compared to single adults. With 1 child and 1 adult, more than 70% of the stays are 3 days or longer. With 3 children and 1 adult, more than 75% of the stays are 5 days or longer. Similar trend is seen with 2 adults and 1 child, and 2 adults and 2 children. With 1 child and 2 adults, around 67.5% of stays are 3 nights or more, while with 2 adults and 2 children, around 67.5% of stays are 3 nights or more.

What is the trend of cancellations with respect to the number of days before which a reservation is made?

Hypothesis

Here, our aim is to analyze if there is a correlation between the booking gap (ie. the number of days between reservation and arrival) and the cancellations. Typically we expect that the earlier someone has booked a stay, the more likely they are to cancel it due to the emergence of unknown circumstances. In this graph, we also stack the data based on various other variables - customer type, hotel. These variables give us more insight into the cancellations - for example, customer types that are 'groups' seem to be less likely to cancel a reservation if it is nearer to their arrival date than if it is farther.

Explanation

From this dataset, we find out that maximum cancellations are done if the time frame between reservation and arrival date is shorter. \ Typically, we would expect people to cancel if their arrival dates are farther away. But that is not what we observe from this data, thus proving our hypothesis wrong. From the graph, we can also observe that transient customers are more prone to cancel their reservations. We expect this, because transient customers are not a part of any group or contract and so they are more likely to change their mind and cancel their reservations. Next, we observe that city hotels had more cancellations compared to resorts. This is again typical because city hotels have more business travellers, whose plans may change. Whereas, resort bookings are usually planned vacations, hence less likely to be cancelled.

Interactions and design decisions

1. The map contains the number of travellers from each country. We can get the country name and the number of travellers by hovering over the country location. Also the density of the color indicates the percentage of travellers from the country compared to the total number of travellers overall. This design gives a visual representation of the relative number of travellers from the different countries without knowing the exact numerical values. Also, simply by choosing different years, and months, we can form a visual image about the relative number of travellers from various countries. Choosing a world map helps in visualizing the geographical data better compared to other forms of visualization.
2. Length of stay is an important data analysis feature for hotel stays. We plot the top-10 values of the stay only since the data is very sparse for larger values of stay. Also, we can change the number of adults and children travelling with the help of the slider. We selected the range for the number of adults and children keeping in mind the values for which sufficient data was available. Also, choosing the barplot helps in relative comparison between the different night stay durations
3. We have used a stacked histogram to find the trend between number of cancellations and type of customer/hotel

Why? This design decision was made because a stacked histogram is the best way to represent the number of cancellations made in particular range of days, and we can use stacks to show the influence of another variable over it - that is, how much part of the bar belongs to a particular hotel or a particular customer type. We have also added a zoom in-out feature to change the granularity or the range of the number of days.

Development Process

We worked in a team of 2 for this project. The first step of the assignment was to determine the appropriate dataset which could provide us with interesting and innovative questions to answer. We looked up the various Kaggle datasets and came up with the present data since it provided geographical data, along with plenty of other useful columns regarding hotel bookings. Several other hotel booking datasets were available as well but this was the most comprehensive in terms of the features and observations, hence we selected it. We began with first exploring the dataset, examining the various features available in the data, and performing exploratory data analysis. Then, we began figuring out the pertinent questions that could be answered from the data.

Once the questions were determined, we tried to see what features and feature correlations could be useful in answering those questions. The next step was to figure out the best visualizations that could answer the proposed questions. In total, the assignment took us a total of 25 hours. Much of the time was spent on finding the appropriate dataset, preprocessing the data to get it in the desired format for plotting the geographical map.

Components of the Assignment

1. hotel_bookings.csv : Original Dataset from Kaggle
2. hotel_bookings_mod1.csv : Preprocessed data
3. country_codes.csv : Used for obtaining the country IDs for plotting
4. streamlit_app.py : Streamlit visualization code
5. Report.pdf : Analysis of the various components of the assignment