Uber Data Analysis Project

Juan Moctezuma

Abstract

This paper is based entirely on the results of an 8-month period project in which a simple survey along with the recorded data from an independent contractor’s mobile application (Uber partner/driver app) brought together my study’s main dataset. The rest of the tables in my workbook (Microsoft Excel file) contain information extracted from the following websites: Weather Underground and Climate-Data.org. I directly performed a two-question survey to passengers and wrote down the results after completing trips requested by an Uber user. The Uber platform allows the user (passenger) to request a ride to the closest Uber partner, but the driver cannot select the user or their respective destination. This paper examines and discusses people’s opinion about a specific location, and points out the external factors that could highly influence their view. However, other observations were included in this analysis as well. This paper includes a reflection at the end on this discussion file that reviews the project’s outcomes.

*Keywords*: Mobile app, Survey, Uber, Weather

Uber Data Analysis Project

Did you know that San Diego is popularly known as America’s Finest City (Offices.net)? The intention of this study is to compile, visualize and interpret data in order to learn what aspect of the City of San Diego, California individuals considered to be the most appealing. I personally collected this information by asking the following questions: “*Where are you from originally?*” and “*What do you like the most about San Diego?*”. My hypothesis was that most individuals, despite their background or place of origin, like this city because of its ideal climate conditions. This writing sample summarizes not only some of the technical aspects that describe how the project was done, but a brief explanation of non-technical/soft skills reflected during this process as well.

Prior to commencing the data exploratory phase, I had to carefully plan a simple questionnaire that is both short and practical, which would allow me to gain valuable information. Asking the wrong questions would only cause me to redefine the goals of this project, or much worse, waste valuable time or resources. The next step after establishing the appropriate questions was to increase the size of the sample or dataset, but the only way to do so was by interacting with more customers or riders. As time passed by, I gained awareness about other observations such as the fact that most riders are from San Francisco, CA and the US east coast. Other patterns were reflected by the recorded data on the Uber Partner’s mobile app, for instance, the most common drop-off address zip-codes or types of service requested by customers. A few months later, I noticed that a great variety of passengers did, in fact, love the city’s weather, however I decided not to reach a conclusion after 2017 was over.

To increase the randomness and size of the dataset that contains the number of rides (which includes the passengers’ information), I decided to gather data from May to December 2017 and store it in a Microsoft Excel file. I started this project in May because I became an Uber partner/driver in that month. Not all content from the passengers’ information table comes directly from the trip history section in the Uber Partner’s mobile application, part of it was collected from what I compiled while interacting with the riders.

When I began picking up passengers I realized a few situations a couple of days afterwards, and started questioning myself about what to do with the information from passengers that had been aboard my vehicle more than once. This implied that my dataset now contained duplicated information. For the sake of simplicity, I never asked those passengers my survey questions after the first ride. The way I handled it was by writing their name (in the Excel column titled Account Holder Name) in the passengers’ information section, followed by “//Repeated” with the purpose of avoiding to over count the number of votes. For dealing with duplicated data, I found it inconvenient to write a script/code that avoids counting repeated passengers’ data due to the probability, despite being considerably low, of having different individuals with the same name, place of origin and same survey response! This project does not intend to study any specific population, so no demographic questions related to gender, race, age, etc.; were asked and my table doesn’t contain detailed information about my customers. Therefore, no extra data could be used by the script so that it could perform an automated process that could easily detect duplicated passengers. Some rides contain more than one individual, this would imply the necessity of asking every single rider for more information which would be considerably unpractical for the purposes of this project.

Another situation that more commonly occurred was having passengers that were not the account holders, which means that the Uber user ordered a service or trip for someone else. To deal with this, I simply created a format X / Y where X represents the account holder and Y represents the passenger present during the trip. For example, James orders an Uber for Ofelia, then the name would appear as James / Ofelia, where the passenger’s information will belong to Ofelia. But if Y were several passengers then the format would be James / (N/A). If James orders an Uber for Ofelia more than once, then the format would be James / Ofelia //Repeated! The trickiest situation to deal with was having a repeated account holder that was accompanied by other individuals, surprisingly this happened to me only once. If this happened then the format would be “//Repeated\*”, where the number of asterisks represent the number of riders excluding the repeated account holder.

Although coming up with creative ways to handle the riders’ information was quite challenging, improving the quality of my data was probably the most time-consuming process. Organizing the customer’s places of origin required me to make sure each city, state/province or country was in alphabetical order, had no typos, and that each entity was consistent with one another. By entity consistency I refer to labeling each location accurately, for instance entering San Diego, CA into the corresponding cell instead of SD; San Diego, California; or San Diego only). Entering the drop-off zone zip code, was laborious and occasionally difficult since there were unusual circumstances which caused the Uber Partner App to display United States or “Unknown Address” instead of the actual location. Classifying each drop-off address as an area or neighborhood was also challenging since some locations are between the boundaries that separate diverse neighborhoods along the San Diego County. The Uber Partner App provides exact addresses, but my passengers’ information table only includes drop-off zones. Overall, I had to review each column, make sure there were no typographical errors and that each cell contained both the correct information and the format.

After compiling the data, counting my passengers’ votes, and creating a pivot table with the survey’s results, I confirmed that most people love the city’s weather! I reviewed the customers’ opinion, and then decided to classify each response into the following choices: weather (W), people (P), not available (N/A), and other (O). Other encompasses aspects related with local cuisine, entertainment, local beaches, lack of traffic congestion in comparison with any great metropolis, bars and nightlife, etc. A variety of cities, states or countries were selected after a pivot table displayed the number of passengers, native to those regions, that voted for the weather. I had to set a standard because most of my passengers not only come from diverse places, but in several occasions, they did not specify their city of origin.

Consequently, I had to create two sections, one which displayed numerical results based on weather conditions (from May to December 2017) from the top list of selected cities and another one with generic information related with Köppen-Geiger climate classification (Rubel, 2017). The only way I could be able to compare a city’s clime with places that commonly have an enormous landmass is by displaying a table that described the top list for selected states or countries into several climate subclasses. Surprisingly most selected places did not have the same classification as San Diego, CA, which is labeled as cold semi-arid (climate-data.org, 2017). Anyhow, the survey results demonstrated that most passengers voted for weather and considered it San Diego’s best attribute. Climate data collected during this 8-month period indicated that this city’s humidity levels were not excessive, average temperature was approximately 20ºC or 68ºF and precipitation was low.

Several pivot tables were created after extracting data from the passenger’s information table (in tabular format), and then both, pie and bar charts (including one 3D chart) were developed out of these. Some pivot tables are connected to slicers but not all of them depending on how information was organized. Also, two graphs containing temperature and humidity with respect to time (in this case months) were created from the table containing the cities’ weather data (climate-data.org, 2017). The purpose of creating a dashboard was to allow the reader or user to easily see my results. Developing interactive charts not only allowed myself to explore, summarize, analyze and present data arranged by pivot tables, but to practice a common application related to business intelligence as well.

Reflection

This project initially consisted mostly of data entry during several months. Anyhow, it allowed me to improve my organizational skills because I had to constantly re-plan certain steps as unexpected situations occurred, prioritize activities and set myself deadlines to achieve my project’s goals. This initial phase also allowed me to develop efficient interpersonal skills due to the constant social interaction with people from diverse backgrounds.

The data exploration phase allowed me to use many of Microsoft Excel’s tools such as: absolute and relative cell references, pivot tables, filters, functions, formulas, conditional formatting (including the use of Boolean operators or logical operation), pivot tables, slicers, graphs, charts, one Macro and basic VBA programming. During this phase of my project I also had the chance to review and apply concepts from Statistics and Mathematics.

When I first started compiling the survey’s information, I wasn’t immediately aware of all the patterns that could be observed! Therefore, I decided to share these observations by inserting charts in the data visualization section/worksheet of my project’s workbook (Excel file).

Conclusion

Overall, the Uber Data Analysis Project experience constantly tested my critical thinking/problem-solving skills as I encountered challenges. As I mentioned before, there are occasions in which unexpected situations occurred, and finding solutions to these issues was not necessarily an obvious or easy task to perform. My creative thinking skills not only allowed myself to find solutions to those challenges mentioned previously, as well, but to come up with this unique project! This activity also required close attention to detail to maintain both, data and ideas consistent with one another and to preserve the quality of my work. Furthermore, I believe data obtained from these types of surveys could be extremely useful for further studies or research in advertising campaign planning that could ultimately benefit a city’s touristic sector. Thank you for taking the time to read this writing sample.

References

Computer Hope. (2017, June). *Logical Operation*. Retrieved from https://www.computerhope.com/jargon/l/logioper.htm.

Climate-Data.org. (2017). Retrieved from <https://en.climate-data.org>.

Offices.net. (2018). *City Slogans & Nicknames of US Cities*. Retrieved from

https://offices.net/city-slogans.htm

Rubel, F. (2017). *World Maps of Köppen-Geiger climate Classification*. Retrieved from

http://koeppen-geiger.vu-wien.ac.at.

Schuele, C. and Lee, M. *Demographics*. Retrieved from

<https://www.britannica.com/topic/demographics>.

Uber Profile. (2017). *Uber Profile*. Retrieved from <https://www.partners.uber.com>.

Weather Underground. (2017). Retrieved from https://www.wunderground.com.