Awarding Body:

Arden

Programme Name:

MSc Data Analytics and Information Systems Management

Module Name (and Part if applicable):

Data handling and decision making

Assessment Title: TASK 2

Student Number:

STU105262

Tutor Name:

Barry Smith

Word Count:

3886

Please refer to the Word Count Policy on your Module Page for guidance

List of contents: Page No

2.1.1 Explanation of data collection, filtering, and integration procedures.

Objective:

The main aim of analysis is to understand the root cause of the problem related to cancellation and non-availability of cars from airport to city and vice versa.

Data collection: Uber collects the required data from two sources, the application and the backend services used by the application. Passenger and driver data is collected from the application which will have billions of events daily (user activities with the application) which are stored in the uber database. Uber uses this data to handle major problem like increasing funnel conversion, user engagement, etc. to handle, analyze and extract information from these complex data sets we couldn't rely on traditional data-processing systems. So, Uber uses bigdata analytics along with machine learning to make most accurate predictions to meet the customer expectations. (Willis and Trancos, 2021)

Data filtering and Integration procedures: As we are dealing with bigdata it is always difficult to analyze the required data and extract useful information from it. Business decisions cannot be made with the raw data, Uber uses aggregated data collected i.e., data collected from different sources which includes data from application and the backend services which are used by the application. This data collected consists of rider information, Credit card details, driver information, and location information (GPS). As most of the data consists of user's personal information it is anonymized. As the business is growing exponentially in terms of cities uber operates and in number of drivers and people using the service the amount of incoming data is increased and the need to access and analyze all the data in once place to help data managers analyze it and make smarter decisions. Uber uses **Common storage integration** procedures (Data warehouse) to combine this data from these two sources called as **Hadoop data** lake (Shiftehfar, 2021) improving efficiency and stability for its big data platform. By using this integration procedures burden on the host is reduced as it is not handling data queries constantly with clear data appearance and enhanced data analytics.
To assist with decision-making, from financial planning to informing drivers of the optimal location for trip requests at any time. To gain useful insights from the bigdata it is necessary to use filters that will help us to select the small part of data from the data set and perform analysis using machine learning algorithms to make business decisions on surge pricing and supply demand gaps. In the dataset to gain more meaningful insights we can use the **filters** for **Request Timestamp** (time and date when the ride was requested) and pickup location (Airport or city) to understand in which location, date, and time most users have requested for the rides to analyze the supply demand issue. (Eupen, 2021)

2.1.2 Analysis of data representativeness

Data is said to be representative if the small sample data set chosen from a larger group that sufficiently replicates the larger group(bigdata) based on whatever attribute or quality is being investigated. With the help of representative data, we can make decisions on the actual data set. **(A. Ramse and Hewitt, 2021)** Uber uses machine learning algorithms to make predictions on surge pricing and supply demand in context to how the data is generated to make these predictions more accurate using ML algorithms uber is using representative learning. In which the machine learns the representation itself from the provided dataset **(Barla, 2021)**. It's a strategy for determining how the predictive model will perform based on how the data is represented - the features, the distance function, and the similarity function. Thereby representation learning makes it easy to find patterns, anomalies with better understanding of data behavior. By using this method noise is reduced which can be very useful for supervised learning algorithms. The use of representation learning allows for a vast number of input combinations to be achieved. UBER's computational and statistical efficiency will be improved by these methods. For UBER to control surge pricing, demand supply gaps the data representative model is necessary which can analyze the information and forecast which customers would need a ride right away. (Barovick, 2021)

2.1.3 Statement on generalisability and limitations of the integrated dataset.

The degree to which a study's findings can be applied to different situations is referred to as generalisability **(Shantikumar, 2021**). By data generalization uber can identify the data which is trustworthy to make accurate predictions. By using data representative learning complexity of the model is reduced thereby helping to reduce the noise and anomalies leading to overfitting issues **(IBM, 2021).** In general, we'll use a data sample to determine generalization. UBER would benefit from the data representative learning model. This machine learning model is well-suited to the data generalization process.

An integrated data set is a group of data sets that will be combined to generate a single, consistent data set. Data redundancy can result in numerous repeating data in the integrated data, which is a constraint. You must also deal with data that is incomplete. Manual integration, Middleware integration, application-based integration, Uniform Access integration, and data warehousing are all options for data integration. Data integration tools are used to carry out these integrations. On-premises data integration tool, opensource data integration tool, and cloud-based data integration tool are the three types of solutions available. The constraints of integrated data must be investigated. If it is not correctly engaged, it might create an issue by data duplication. **(Talen, 2021).** Is it necessary to resolve data duplication and data redundancy conflicts, which can only be rectified throughout the data integration procedure, to overcome the constraints of data integration, a homogeneous collection of data must be merged to resolve these issues.

2.2.1 Selection and justification of the inferential and/or machine learning models, most relevant to the objectives of your case study.

The main objective is to understand the root cause of the problem related to cancellation and non-availability of cars from airport to city and vice versa. To analyze this problem, we have proposed the machine learning model for uber to analyze the data for more accurate predictions to fulfil the demand supply gaps and surge pricing issue leading to loss of revenue. By analyzing the data visualizations can be created for the business processes which can help uber in cost-cutting and further improvements in making better business decisions. We have proposed the use of machine learning algorithms along with vehicle telematics data and hyperscale real-time matching to predict high demand areas by using heatmaps which can help the drivers to wait in location in high demand.

To grow in the marketplace Uber must leverage machine learning. In here we have proposed a supervised model which uses historical data and make predictions of time and areas and alerts drivers to the corresponding regions with future demand. In representative data model we proposed first we need to select the data set for our supervised learning model which will gain the knowledge from the training dataset (selected training dataset) then we can use the supervised learning model on different datasets from the uber database and analyze the performance of the model predictions. We can define the representative Set as**: (Borovicka, Kordik and Jirina, 2021)**

R* >/ R such that R* is the smallest instance. And ACC(R*), is not equal to ACC®, Where ACC(X) denotes the accuracy obtained by the training set.

As we have proposed use of machine learning models to Uber for business development and to answer the supply demand gaps below are the ways these analytics could benefit uber.

Let's have a clear understanding of how statistics will help data analytics. Statistical tools also involve probability. Probability has various kinds of distributions such as uniform distribution, normal distribution, and Poisson distribution. We will look at each of the distribution in detail. A uniform distribution is known to be one of the simplest kinds of distribution. It has a fixed value that occurs only once in a certain range. Anything outside the range would be considered as zero. It can be identified as a piecewise function **(Ugoni, 2021).** The next distribution we will consider is normal. It is also known as Gaussian distribution. In this, kind of distribution we must consider mean and standard deviation. The average value of the data set is spread throughout the chart. It can also be clustered if the results are closer to each other. A Poisson distribution is like a uniform distribution. Here, Skewness is the main factor to determine the Poisson distribution. If the value of skewness is low this means that the graph would be widespread and if the value of skewness is high, it means that the graph would be clustered **(Kissell and Poserin, 2021).** There are many kinds of distributions, but these are enough to get the values covered. The algorithm chosen for Poisson distribution should be robust to analyse all the spatial values. To anticipate the outcomes of UBER, we can use a statistical model. Because UBER creates a large amount of data, we must analyse it to gain insights.

These will aid UBER's success, as well as that of other businesses. To promote UBER, marketing might be quite useful. This should be done on a computer. Excel is used to create the graphs and statistics. These figures depict several characteristics of UBER. Below are the observations from the dataset.

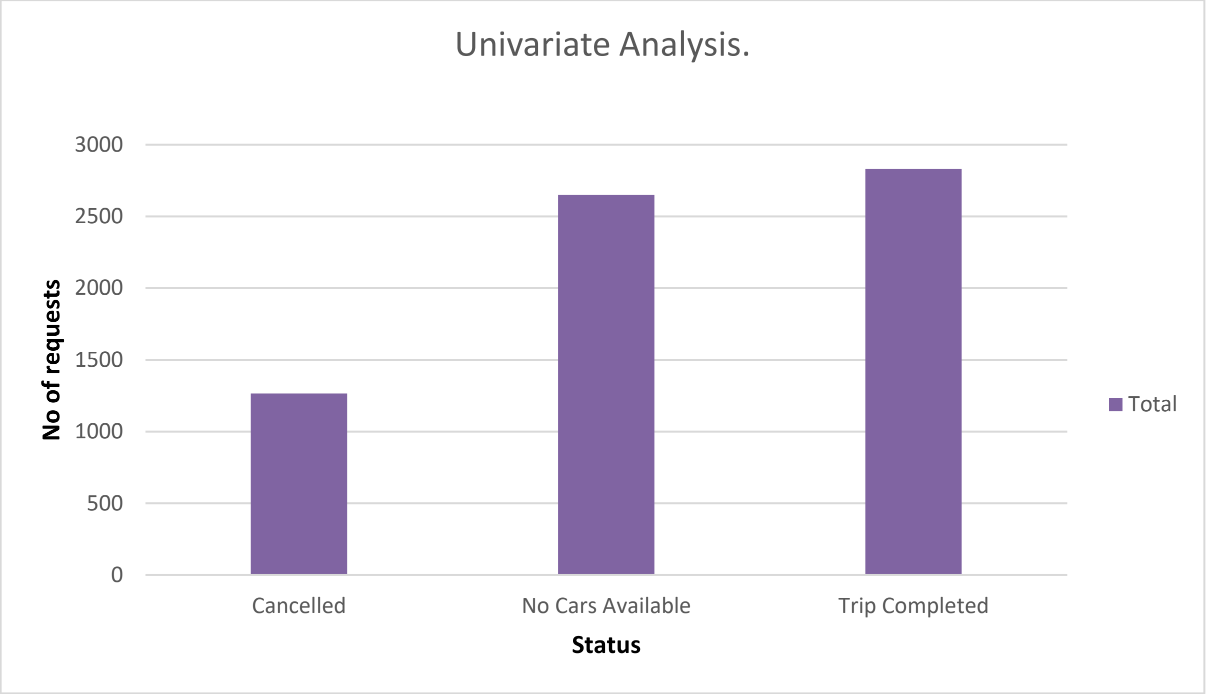


*Figure 1: Bar graph showing distributions for three categories of status (cancelled, no cars available, Trip completed) to count (**Source author self**)*

Univariate Analysis has been performed on the status (non-numerical) columns which has 3 categories or states and below are the findings from it. To perform a univariate analysis for a single categorial variable. A frequency table (pivot table) is constructed between status column and no of requests and a bar graph has been plotted with the observations. Above is the bar graph obtained from the chosen data. The initial outcomes by analyzing the data on excel is that no cars available are way more than no of Cancellations. Uber should think of implementing more cars to fill the gap.

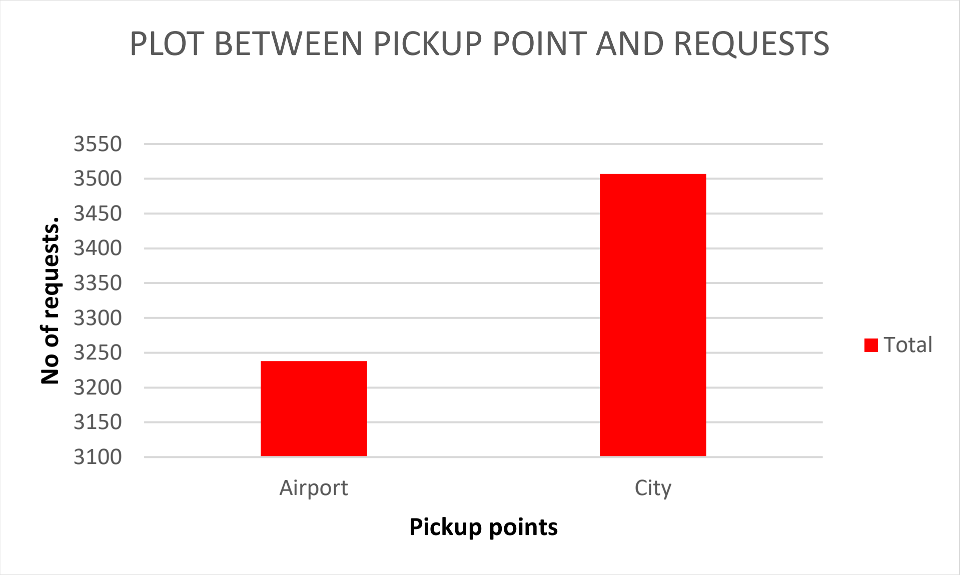


Figure 2 Bar graph showing no. of requests related to pick up points (Source: Author self)

Conclusion from the dataset: The pickup points Airport and City are almost equal times present in the dataset. We can see that the no of requests near the airport are 3240 and the city are 3500. We can analyse that there are more number of requests from airport than city. However, the observed difference is not much.

A bivariate analysis has been drawn for status and pickup point columns below are the findings from the graph.

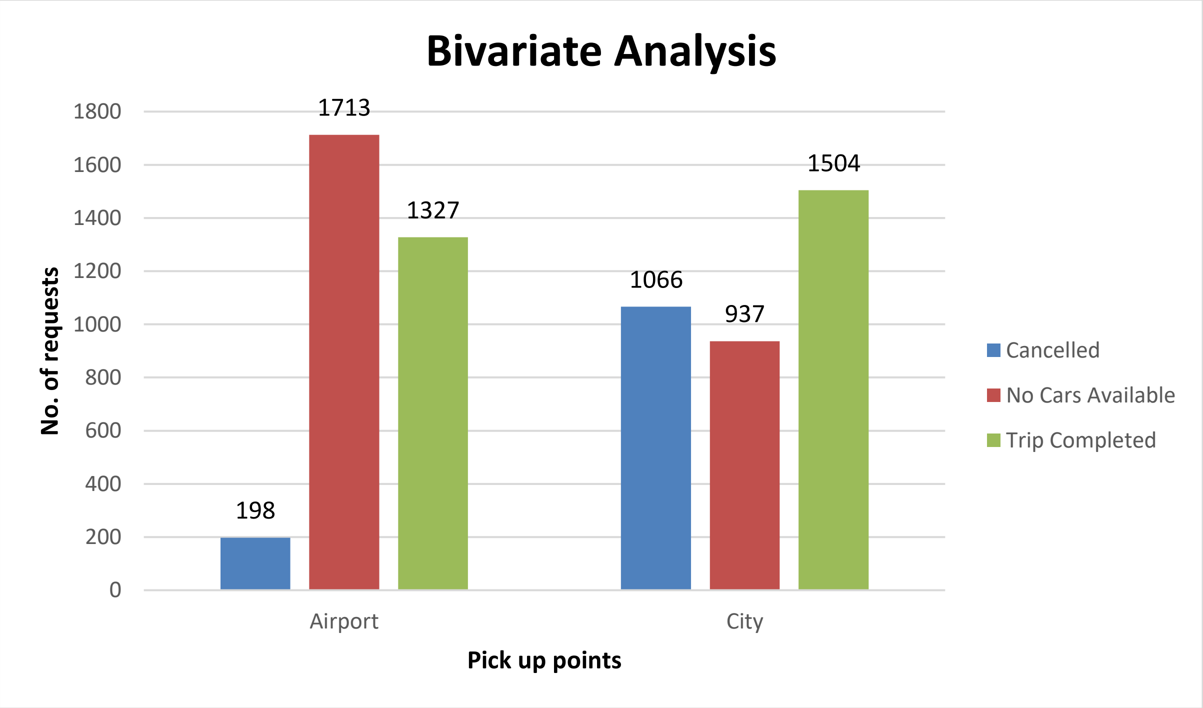


Figure 3: Bar graph showing status count for Airport and city (source: Author self)

Bivariate Analysis conclusion of Status and Pickup point columns: We can see that there are High number of **No cars Available** from airport to city and many cars are **cancelled** from city to Airport. From the above out comes we can conclude that drivers are not interested to go to airport from the city, so we see that there are no cars available at the airport causing an issue for the passengers who are travelling from airport to other parts of the city. The reason might be drivers don't want to travel to city outskirts or they expect the trips which are more feasible to them which are in the city limits. Uber need to educate drivers about the gap and should offer some incentives or extra payments for drivers who have done more trips from airport to city.

2.2.2

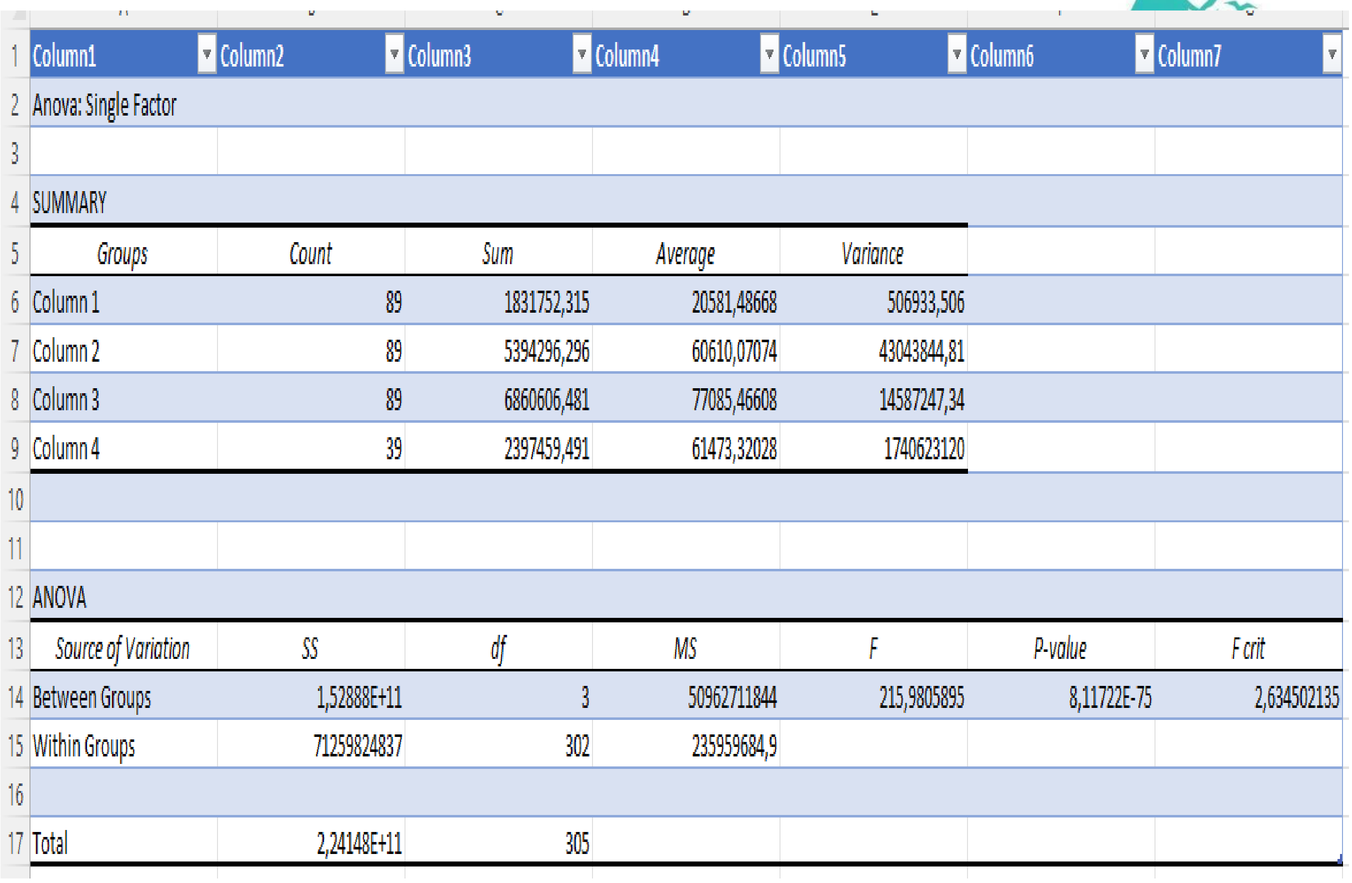
A hypothesis can be formulated to track down the relationship between variables **status column** which is a categorical variable(consisting of three different groups cancelled, no cars available and trip completed)but we are more interested to know if there is really a significant relationship between the status reason **cancellation** and time stamp of the cancellation. In here we can identify cancellation as the dependent on variable time slot.

Null Hypothesis statement H_0 : Frequency of trip cancelled does not depend on the time slot.

Alternative Hypothesis H_a : Frequency of trip cancelled depends on the time slot.

To check the relationship significance between time slot and cancellation data analysis tool Anova: Single factor has been chosen as we have data about one categorical independent variable and one dependent variable and results are formulated.

Anova: Single Factor



	Column1	Column2	Column3	Column4	Column5	Column6	Column7
2	Anova: Single Factor						
3							
4	SUMMARY						
5	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
6	Column 1	89	1831752,315	20581,48668	506933,506		
7	Column 2	89	5394296,296	60610,07074	43043844,81		
8	Column 3	89	6860606,481	77085,46608	14587247,34		
9	Column 4	39	2397459,491	61473,32028	1740623120		
10							
11							
12	ANOVA						
13	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
14	Between Groups	1,52888E+11	3	50962711844	215,9805895	8,11722E-75	2,634502135
15	Within Groups	71259824837	302	235959684,9			
16							
17	Total	2,24148E+11	305				

Table 1: Representing single factor anova analysis (Source: Author own)

From the analysis we notice that p <0,05 and F > F critical we can reject the null at the 0.05 level of significance. $H_{0 \text{ Is}}$ failed. So, there is a relationship between frequency of trips cancelled and time slot.

From the above analysis we have noticed that there are lot of cancellation from airport to city which are time dependent.

Now we will try to understand the time slots (hourly) when there is more cancellation of rides.

let's divide the different time slots to different time periods in a day for a better understanding in which period of the day we are seeing more cancellations. For this we have considered the time slots as below.

Morning-- 6:00 AM to 12:00 AM
Afternoon—12:00 TO 16:00
Evening—16:00 to 22:00
Night—22:00 to 6:00

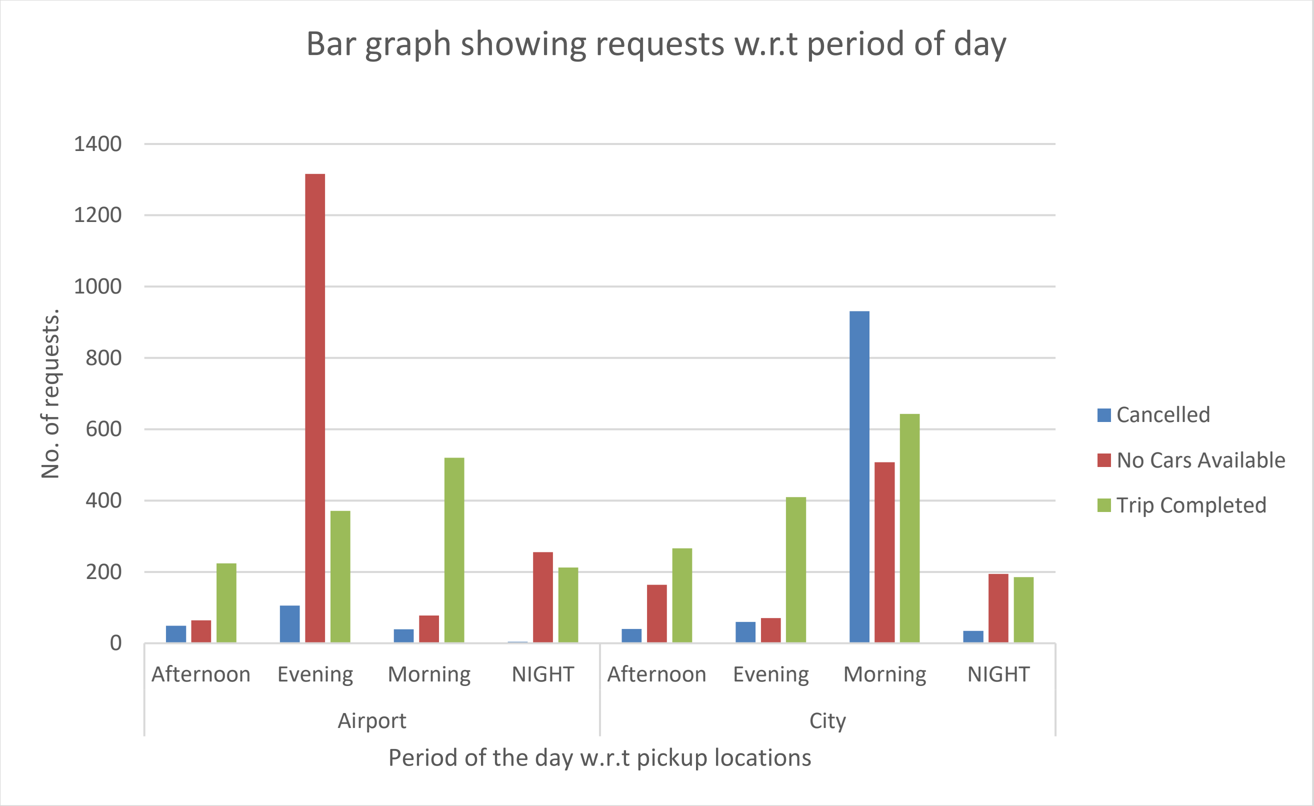


Figure 4 Graphical analysis of status W.R.T time period

Conclusion from the graph: From the above graph we see that there are more number are cancellation in **Morning** in the city and more no cars available in the **evening** at airport. This might happen because drivers might retire early from work causing a supply gap in the evening. Morning cancellations from time slots 5 Am to 9 Am The requests are most likely being cancelled by the drivers owing to the morning rush because it is office hours, and because the destination is an airport, which is too far away, the driver would prefer to earn more for shorter trips within the city. To overcome this uber can offer incentives and special gift coupons for drivers who are willing to complete the trips from city to airport in the morning this can motivate drivers not to cancel the trips to airport in the morning hours.

2.2.3 Recommendations to solve the supply demand gaps from above analysis:

- Uber could offer higher performance bonuses to drivers who travel from the city to the airport during rush hour (Mornings). which might be motivating to drivers and can help in reducing the number of cancellations.
- Setting goals for drivers to take City to Airport or Airport to City trips, and then appropriately compensating them with draws and gift cards.
- Orientation classes can be conducted for all the drivers and insights from the analysis in the form of heat maps and notifications through application can be provided so that, the drivers will be aware of the issues and act accordingly.
- To meet the no cars available gaps in the evening. We can inform the drivers about the shortage in that time and asking them not to sign off early in the busy hours and there is a need of more cars and drivers to full fill the above gap. Uber can attract drivers by offering Sign On bonus for new drivers. (Nicole, 2021)

These were some of the possible outcomes of Big Data analysis. Decisions are made based on this analysis. The technical examination of data is aided by understanding basic statistical principles. This is used to predict the results of the analytics. We can also examine the structure of the data and how to use various data analytics approaches using statistics. To summarize, data analytics must be used throughout a company to compare it to other organizations. The method of data analytics is beneficial for converting large amounts of unprocessed data into refined and valuable information. UBER is one of the most popular ride-hailing services. Machine learning may also be used to save costs by allowing smarter judgments to be made. To obtain more clients, the peak factor might be decreased. Customers will be drawn in by the special deals that UBER will make available through digital marketing. In addition to the searched phrases, machine learning extracts terms that might improve UBER's dynamics. It's vital to remember that human mistakes might occur while changing a large data set, which is why these technologies are used to handle Big Data. To deal with supply demand, surge prices etc. we can conclude that we can full fill these gaps by leveraging **machine learning** in UBER. (Qasim Shabbir and Gardezi, 2021)

2.3 Data visualization.

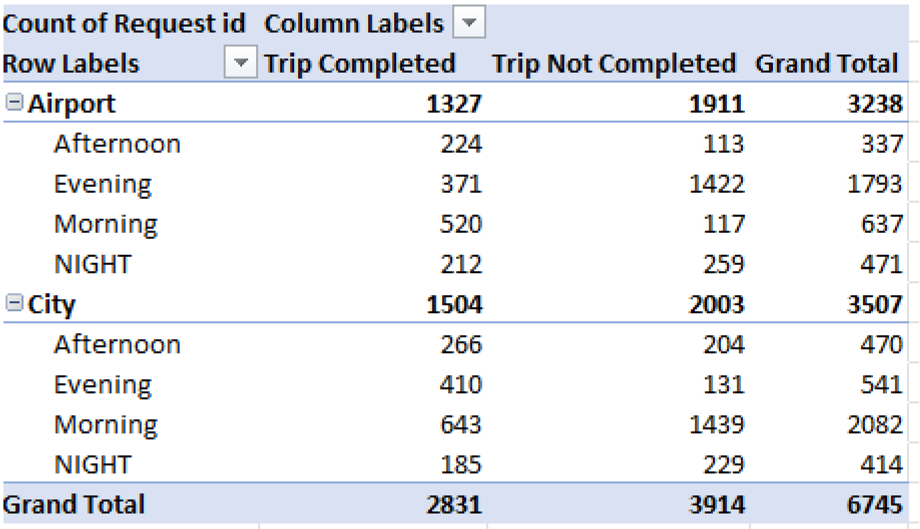
From the question 2.2 we have drawn various conclusions on supply demand gaps between airport to city trips and vice versa. In here we will try to scaleup the analysis and identify the gaps in detail for more detailed understanding of supply demand gaps.

A table is created between trips completed and trips not completed with respect to pick up point and request time slot.

From the below table we can notice that there is **demand** (total No. of requests)**6745** out of which we see that only 2831 trips are completed **(Supply**) we can determine the Gap as follows:

GAP = DEMAND-SUPPLY = 3914. (IDENTIFIED GAP)

Clearly, we can see the huge gap of 3914 requests which are not addressed by UBER. Either in the form of **Cancellations** by drivers and **No cars available**.



Count of Request id	Column Labels		
Row Labels	Trip Completed	Trip Not Completed	Grand Total
⊟ **Airport**	**1327**	**1911**	**3238**
Afternoon	224	113	337
Evening	371	1422	1793
Morning	520	117	637
NIGHT	212	259	471
⊟ **City**	**1504**	**2003**	**3507**
Afternoon	266	204	470
Evening	410	131	541
Morning	643	1439	2082
NIGHT	185	229	414
Grand Total	**2831**	**3914**	**6745**

Figure 5 Graphical analysis of status W.R.T to time period.

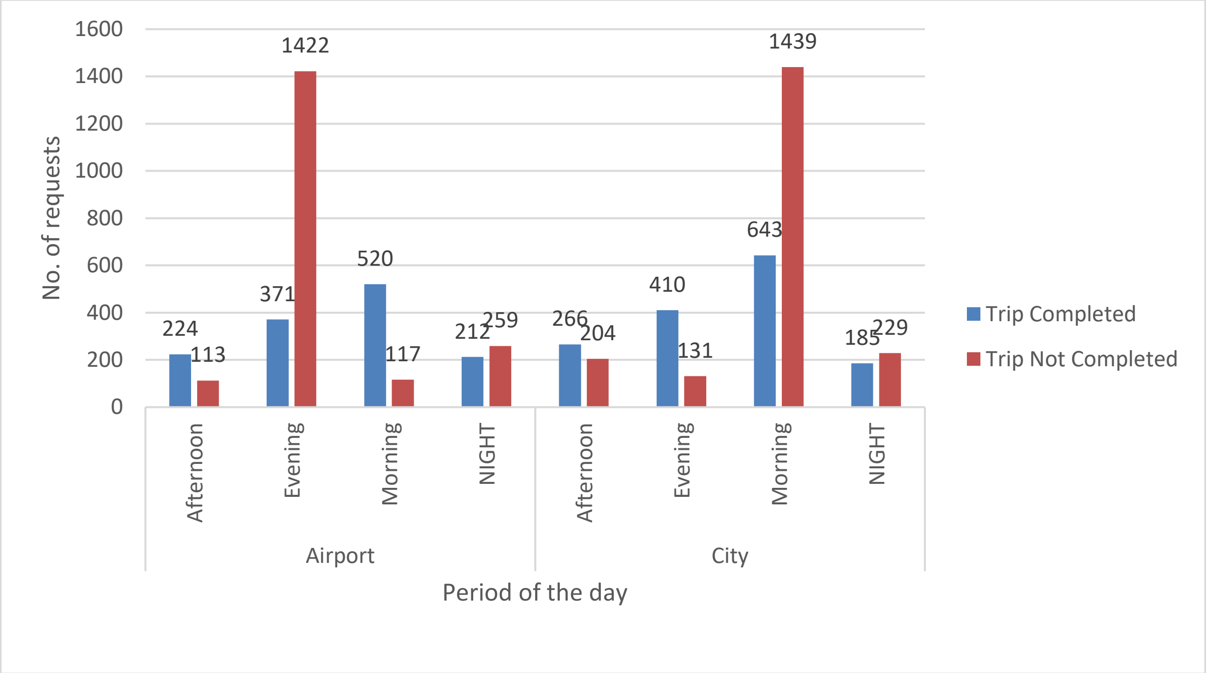


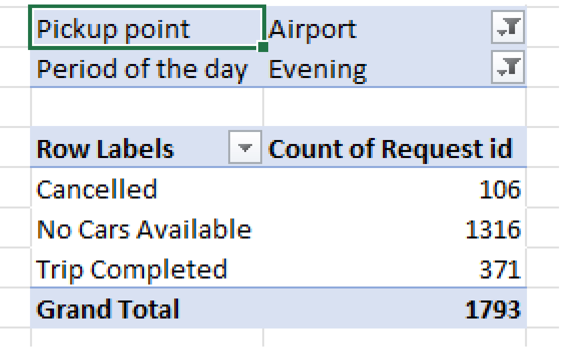
Figure 6: Bar graph showing GAP for trips completed and not completed W.R.T to pick up point and time (Source: Author self)

From the above graph we can notice the huge gap at airports in the evening and mornings in the city. In the morning we see that there is a lot of demand from city to airport with lot of cancellations as the trip from city to airport usually takes long time, depending on the flight patterns, the driver will have a longer idle time after he arrives at the airport. In the mornings, many planes depart the city, while fewer flights arrive. Furthermore, returning empty from the airport to the city is not cost effective. On the other hand, we have noticed a huge demand from Airport to city in the evening due to no cars available out of 1422 requests only 371 are met showing high gap of 1051. This is because during the evening, many planes, particularly international flights, begin to arrive at the airport. The cars are in great demand because of this. As the day progresses, the Uber driver partners begin to retire for the day, resulting in a high level of automobile non-availability.

In here, we will try to separate the findings in different stages with the pickup points for a superior arrangement.

We have noticed a huge supply gap at the airports from the above graph, let's try to find the reason for those gaps to make appropriate business decisions by using a pivot table.

Table 2 showing Trip status at airport.



Pickup point	Airport
Period of the day	Evening
Row Labels	**Count of Request id**
Cancelled	106
No Cars Available	1316
Trip Completed	371
Grand Total	**1793**

From the above table we can see that the major reason for the meeting the supply is **no cars available**.

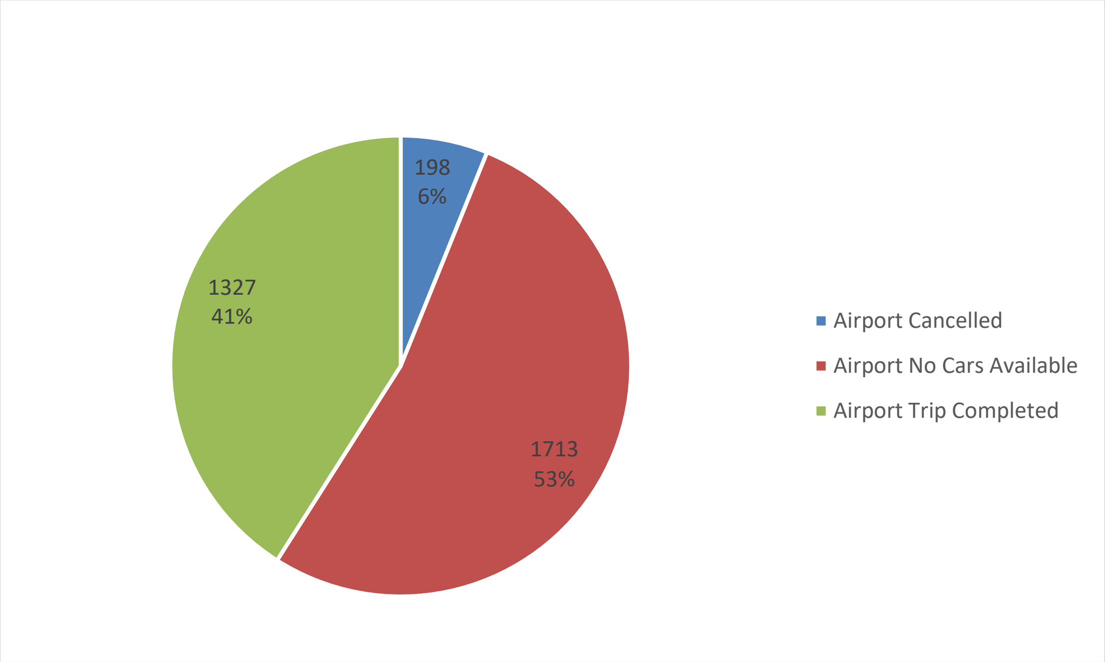
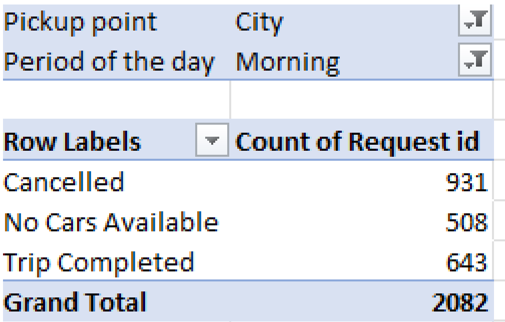


Figure 7 pie chart showing the gaps at airport. (Source: Authors self)

Clearly, we can see that there is gap of 53 % due to no cars available and 6 % cancellations from drivers. Uber needs to create a stand at airport where cabs can be available all the time, this will decrease the **No Cars Available Status.**

Now, Lets analyse the gap in the city. From the Figure:6 we can notice that there are lot of requests in the morning from city to airport. Out of 1439 requests only 643 are accepted and the rest are trip not completed. So, in the city especially in the morning we see a gap of 796 requests. A pivot table has been drawn to understand the reason for Trip not completed and below are the results from it.

Table 3 Showing trip status in city (Morning)(Source: Authors own)



Pickup point	City
Period of the day	Morning
Row Labels	**Count of Request id**
Cancelled	931
No Cars Available	508
Trip Completed	643
Grand Total	**2082**

From the above table we can see that there is huge gap due to cancellations in the morning. And It has been represented in graphical way using pie chart for better understanding.

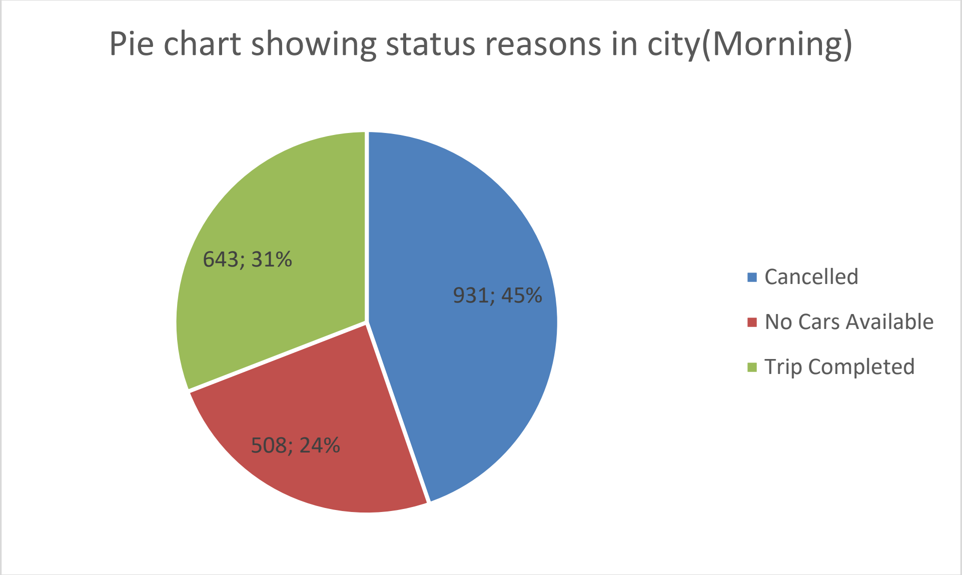


Figure 8 pie chart showing reasons for GAP in city (Morning) (Source: Author Self)

From the above pie chart, we see that out of 2082 requests in the morning 45 % of the trips were cancelled by the drivers. And rest are no cars available. This might be due to the fact the drivers are not motivated to go to airport from the city as it is a long and tiring trip and in return, they might not find the passenger to city as most of the international flites leave in early morning. Further, drivers might think that they can complete the short rides in the city which is profitable as it is a peak hour where most people will be rushing to their respective workplaces.

2.4

Decisions taken in 2.2 and 2.3 can be combined to the framework to make uber more powerful. These considerations can help with strategic planning. Machine learning evaluates data using in-depth patterns to manage surge pricing and demand supply gaps. Surge pricing is one of the most difficult problems that the machine learning algorithm can tackle. The constant process of planning and monitoring an organization's needs is known as **strategic management (Ronda-Pupa, 2021).** To speed up organizational efficiency, a firm must use strategic management. A strategic plan provides a company with future projections and forecasts. Things can be planned in a timely manner by the organization. This planning gives the employees a sense of direction. Similarly, UBER is required to make revolutionary adjustments as projected by data analysis.) The faults that UBER is experiencing must be replaced by the conclusions drawn from the analysis. **Strategic Management** can be divided into below stages which are essential in applying strategic management to Uber.

Firstly, we need to understand the present state of Uber, furthermore we need to understand and analyze the internal and external structure. Formulate the action plans as per the decisions taken while analyzing the data and evaluate the action plans as per the performance of the model.

By analyzing the data to know the reason for not meeting supply demand we have came across 2 different reasons.

1. No cars available in evenings at airport to city
2. Huge number of cancellations in the mornings from city to airport.

Action Plan:

We can either increase the number of cars. I.e., increase the supply or Uber has to leverage the existing supply to handle no cars available at airport. One direction could be usage of surge price technique which will increase the number of drivers as they would be attracted towards good margins for the rides. Thereby, increasing the number of cars to meet the gap. Cab pool can be introduced from airport to city at peak hours where one car can handle multiple requests at the same time to meet the supply demand. Incentives, gift cards, lucky draws can be offered to drivers taking airport to city and city to airport routes which also increase the availability of cars at airports. Drivers should be informed not to retire early in the evenings, surcharge can be offered to drivers who work late night which will avoid more drivers not to retire early in the evenings. A permanent stand for Uber can be created at the airport which will in turn increases the availability of cars at the airport.

Regarding the problem on number of cancellations in the mornings from city to airport Uber should offer incentives to drivers who take up early morning rides from city to airport this motivates drivers to take up the ride thus avoiding number of cancellations. Along with the above recommendations knowledge transfer sessions should be conducted for the drivers regarding the insights obtained from over all analysis. By using heatmaps drivers can check the location of high demand and wait in those particular regions ahead to avoid waiting times which can lead to cancellations.

References:

A.Ramse y, C. and Hewitt, A., 2021. *A Methodology for Assessing Sample Representativeness*. [online] ResearchGate. Available at: <https://www.researchgate.net/publication/239820911_A_Methodology_for_Assessing_Sample_Representativeness> [Accessed 29 November 2021].

Borovicka, T., 2021. *Selecting Representative Data Sets*. [online] Cdn.intechopen.com. Available at: <https://cdn.intechopen.com/pdfs/39037/InTech-Selecting_representative_data_sets.pdf> [Accessed 29 November 2021].

Borovicka, t., Kordik, P. and Jirina, M., 2021. *selecting Representative Data Sets*. www.intechopen.com. [Accessed 29 November 2021].

Eupen, C., 2021. *The impact of data quality filtering of opportunistic citizen science data on species distribution model performance*. sciencedirect.com. [Accessed 28 November 2021].

Education, I., 2021. *What is Overfitting?* [online] Ibm.com. Available at: <https://www.ibm.com/cloud/learn/overfitting> [Accessed 29 November 2021].
Talend.com. 2021. [online] Available at: <https://www.talend.com/resources/data-integration-methods/> [Accessed 29 November 2021].

Kissell, R. and Poserin, J., 2021. *Advanced Math and Statistics*. [online] ScienceDirect. Available at: <https://www.sciencedirect.com/topics/mathematics/poisson-distribution> [Accessed 1 December 2021].

Kissell, R. and Poserin, J., 2021. *Advanced Math and Statistics*. [online] ScienceDirect. Available at: <https://www.sciencedirect.com/topics/mathematics/poisson-distribution> [Accessed 1 December 2021].

Qasim Shabbir, M. and Gardezi, S., 2021. *Application of big data analytics and organizational performance: the mediating role of knowledge management practices*. [online] Available at: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00317-6> [Accessed 1 December 2021].

Ronda-Pupo, G., 2021. *Strategic Management Journal*. [online] researchgate. Available at: <https://www.researchgate.net/publication/260478505_Strategic_Management_Journal> [Accessed 1 December 2021].

Willis, G. and Tranos, E., 2021. *Using 'Big Data' to understand the impacts of Uber on taxis in New York City*. [online] researchgate.net. Available at: <https://www.researchgate.net/publication/345126370_Using_'Big_Data'_to_understand_the_impacts_of_Uber_on_taxis_in_New_York_City> [Accessed 28 November 2021].