**Analysis of XRides Dataset**

**Introduction**

I have been given a dataset of ~40,000 trips for the XRides company and after carefully analysing the dataset, I have deceived some trends in the booking rates based upon different factors, relationships between time of month and no of bookings and so on.

**The dataset**

id - booking ID

user\_id - the ID of the customer (based on mobile number)

vehicle\_model\_id - vehicle model type.

package\_id - type of package (1=4hrs & 40kms, 2=8hrs & 80kms, 3=6hrs & 60kms, 4= 10hrs

& 100kms, 5=5hrs & 50kms, 6=3hrs & 30kms, 7=12hrs & 120kms)

travel\_type\_id - type of travel (1=long distance, 2= point to point, 3= hourly rental).

from\_area\_id - unique identifier of area. Applicable only for point-to-point travel and

packages

to\_area\_id - unique identifier of area. Applicable only for point-to-point travel

from\_city\_id - unique identifier of city

to\_city\_id - unique identifier of city (only for intercity)

from\_date - time stamp of requested trip start

to\_date - time stamp of trip end

online\_booking - if booking was done on desktop website

mobile\_site\_booking - if booking was done on mobile website

booking\_created - time stamp of booking

from\_lat - latitude of from area

from\_long - longitude of from area

to\_lat - latitude of to area

to\_long - longitude of to area

Car\_Cancellation - whether the booking was cancelled (1) or not (0) due to unavailability of a

car.

**Hypothesis Generation**

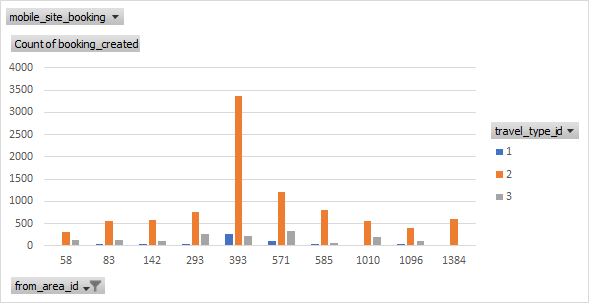
The next step to solve any analytics problems is to list down a set of hypothesis, which in our case are factors that will affect the cost of a taxi trip.

1. Time of day - During peak traffic hours, the taxi fare may be higher.
2. Day of travel - Fare amount may differ on weekday and weekends
3. Trip distance - If the distance to be traveled is within a predefined range fare should ne fixed, otherwise change according to distance.(Eg. If distance travelled is eithin 10kms Rs.40 fixed after that according to distance)
4. Vehicle mode used - Fare would be different for different types of vehicle used.
5. Travel type - Fare may be different based on the locality.(Crowded or Less Crowded or good for two wheelers or not)
6. No of cancellations - Different charges will be applied based upon the no of cancellations by the user.

**Data Cleaning**

As many null values persists in the data, it is essential to clean the data for better outcomes.

1. After looking at the distributions, I found out there were some null values for from\_area\_id. As it is an ambiguity, all rows having such anomaly is deleted from the training data.
2. The range of latitudes and longitudes are between -90 to 90 and -180 to 180 respectively. But in the training data set we observed latitudes and longitudes in range of (22.70,NULL) which is not possible. Thus we remove these rows from our analysis.
3. On analysing the top ten places where the bookings occur maximum, we found out that 89% of all the bookings in a particular area, if of travel type 2. Thus we considered the travel type as point to point only for further analysis and eliminated the attribute. This also helped to handle all the null values corresponding to to\_area\_id.



4. As we are dealing with only point to point travel that is within a city only, we eliminated the from\_city\_id and to\_city\_id from the dataset containing 70% of null values.

5. We further eliminated online\_site\_booking and mobile\_site\_booking attribute from our dataset as it was not of any use to determine how we can increase revenue.

6. Booking\_created attribute was further analysed and splitted into two columns providing the date and time of bookings separately for further analysis.

7. From from\_lat, from\_long, to\_lat, to\_long, we calculated the distance of the trip and recorded the values in a different columns thereby eliminating the longitudes and latitudes attribute.

**Modified attributes**

id - booking ID

user\_id - the ID of the customer (based on mobile number)

vehicle\_model\_id - vehicle model type.

package\_id - type of package (1=4hrs & 40kms, 2=8hrs & 80kms, 3=6hrs & 60kms, 4= 10hrs

& 100kms, 5=5hrs & 50kms, 6=3hrs & 30kms, 7=12hrs & 120kms)

from\_area\_id - unique identifier of area. Applicable only for point-to-point travel and

packages

to\_area\_id - unique identifier of area. Applicable only for point-to-point travel

from\_date - time stamp of requested trip start

to\_date - time stamp of trip end

booking\_date - date of booking

Booking\_time - time of booking

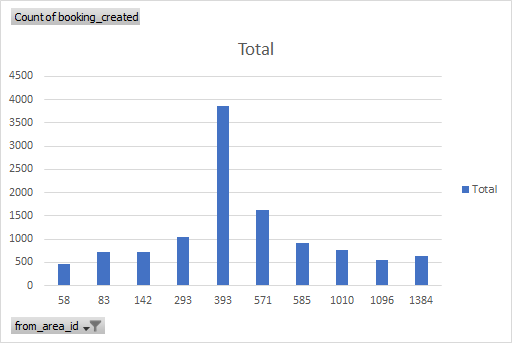
Distance - distance of trip

Car\_Cancellation - whether the booking was cancelled (1) or not (0) due to unavailability of a

car.

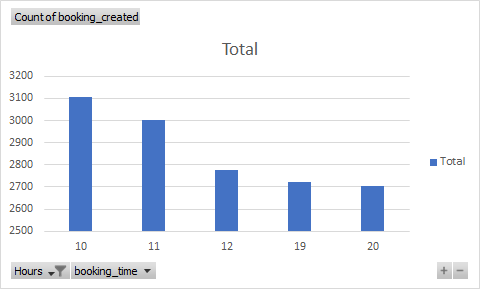
**Exploratory Analysis**

1. First we anaLysed the pattern by counting the no of bookings against each unique from\_area\_ids and recorded the observations in the form of a graph. Below is the graph

For the top ten areas with maximum no of bookings starting from 3858 bookings in area id 393, followed by 1631 bookings in the area\_id 571 and so on.

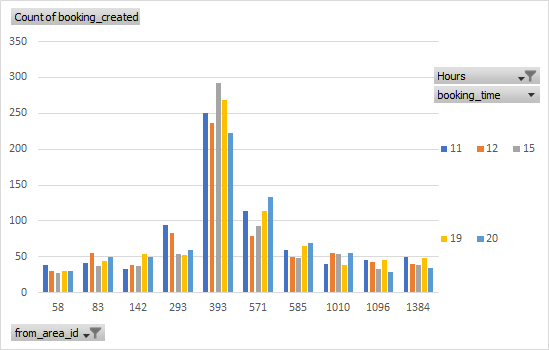
|  |  |
| --- | --- |
| Row Labels | Count of booking\_created |
| 58 | 466 |
| 83 | 719 |
| 142 | 727 |
| 293 | 1052 |
| 393 | 3858 |
| 571 | 1631 |
| 585 | 911 |
| 1010 | 768 |
| 1096 | 542 |
| 1384 | 628 |
| Grand Total | 11302 |

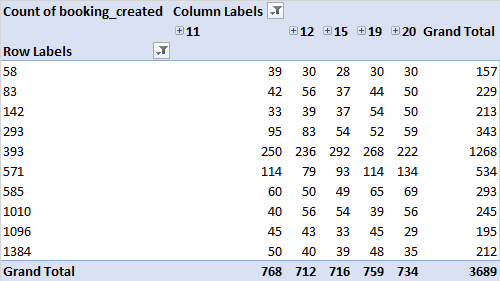
2. Next the no of bookings are analysed on the basis of hours. We see that between 10-12am, most of the bookings occur followed by 7-8 pm in the evening. The graph below depicts the top five hours having maximum no of bookings from every area.



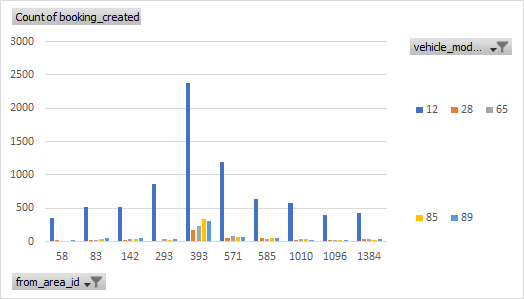
|  |  |
| --- | --- |
| Row Labels | Count of booking\_created |
| 10 | 3106 |
| 11 | 3001 |
| 12 | 2775 |
| 19 | 2724 |
| 20 | 2705 |
| Grand Total | 14311 |

3. Next, I analysed the data and figured out each area at each hour of day gets how many bookings. As we earlier found out that data with from\_area\_id 393 produces maximum bookings, looking closely at the bookings produced at each hour, the results show that the maximum no of bookings are done between 3-4 pm, followed by 7 till late night. Next area having the maximum no of bookings is the area with from\_area\_id 571 having the total of 1631 bookings. The maximum no of bookings for this area occur at night starting from 8pm. Below is the graph depicting top 10 areas with maximum no of bookings and each area having top 5 hours of bookings.



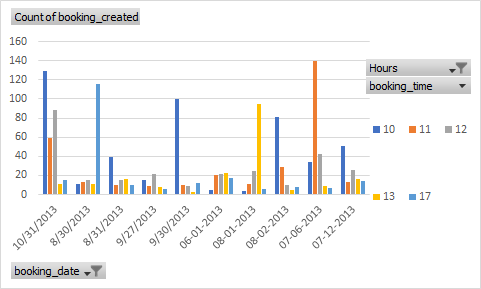


4. Next analysing the mode of vehicle preferred by people in top 10 bookings producing, area, we found at that vehicle\_mode\_type 12 is highly preferred by all the users while all the other modes are negligible preferred.



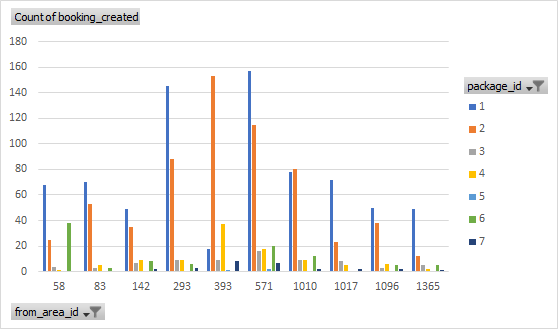
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Count of booking\_created | Column Labels |  |  |  |  |  |
| Row Labels | 12 | 28 | 65 | 85 | 89 | Grand Total |
| 58 | 360 | 19 | 16 | 17 | 22 | 434 |
| 83 | 516 | 21 | 31 | 34 | 54 | 656 |
| 142 | 513 | 21 | 35 | 44 | 49 | 662 |
| 293 | 858 | 16 | 42 | 29 | 47 | 992 |
| 393 | 2381 | 179 | 234 | 338 | 315 | 3447 |
| 571 | 1196 | 59 | 85 | 74 | 76 | 1490 |
| 585 | 640 | 51 | 46 | 53 | 51 | 841 |
| 1010 | 583 | 32 | 37 | 38 | 26 | 716 |
| 1096 | 393 | 21 | 30 | 19 | 32 | 495 |
| 1384 | 426 | 35 | 45 | 32 | 43 | 581 |
| Grand Total | 7866 | 454 | 601 | 678 | 715 | 10314 |

5. Next analysis was on the basis of day, date of the year. Here we found out that the no of bookings were highest in the months of July, August, September, October and majority of them were at late nights starting from 8-9 p.m.



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Count of booking\_created | Column Labels |  |  |  |  |  |
|  | 10 | 11 | 12 | 13 | 17 | Grand Total |
| Row Labels |  |  |  |  |  |  |
| 10/31/2013 | 129 | 59 | 88 | 11 | 15 | 302 |
| 8/30/2013 | 11 | 13 | 15 | 11 | 116 | 166 |
| 8/31/2013 | 39 | 10 | 15 | 16 | 10 | 90 |
| 9/27/2013 | 15 | 9 | 21 | 8 | 6 | 59 |
| 9/30/2013 | 100 | 10 | 9 | 3 | 12 | 134 |
| 06-01-2013 | 5 | 20 | 21 | 23 | 17 | 86 |
| 08-01-2013 | 4 | 11 | 25 | 95 | 6 | 141 |
| 08-02-2013 | 81 | 29 | 10 | 5 | 8 | 133 |
| 07-06-2013 | 34 | 140 | 42 | 9 | 7 | 232 |
| 07-12-2013 | 51 | 13 | 26 | 16 | 14 | 120 |
| Grand Total | 469 | 314 | 272 | 197 | 211 | 1463 |

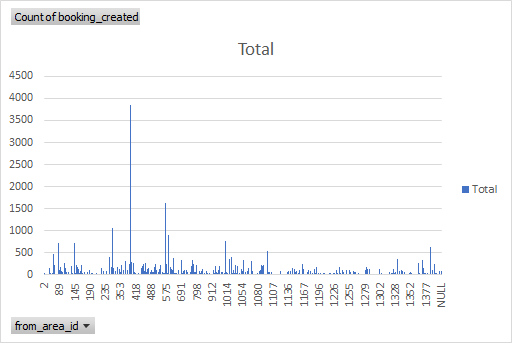
6. Analyzing data on the basis of package\_id, it is seen that the most frequent package used is package 1 that is of 4 hrs and 40kms. This is also evident from the fact that mostly point to point travelling is seen to be practiced by users often.



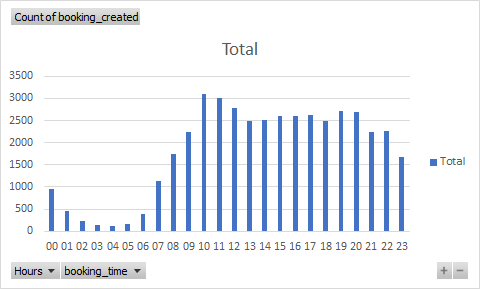
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Count of booking\_created | Column Labels |  |  |  |  |  |  |  |
| Row Labels | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Grand Total |
| 58 | 68 | 25 | 4 | 1 |  | 38 |  | 136 |
| 83 | 70 | 53 | 3 | 5 |  | 3 |  | 134 |
| 142 | 49 | 35 | 7 | 9 |  | 8 | 2 | 110 |
| 293 | 145 | 88 | 9 | 9 |  | 6 | 3 | 260 |
| 393 | 18 | 153 | 9 | 37 | 1 |  | 8 | 226 |
| 571 | 157 | 115 | 16 | 18 | 2 | 20 | 7 | 335 |
| 1010 | 78 | 80 | 9 | 9 |  | 12 | 2 | 190 |
| 1017 | 72 | 23 | 8 | 5 |  |  | 2 | 110 |
| 1096 | 50 | 38 | 3 | 6 |  | 5 | 2 | 104 |
| 1365 | 49 | 12 | 5 | 2 |  | 5 | 1 | 74 |
| Grand Total | 756 | 622 | 73 | 101 | 3 | 97 | 27 | 1679 |

Generalized graphs based on whole training data.

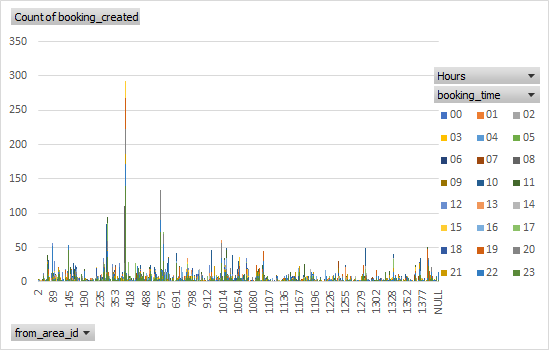
1. From\_area\_id v/s No of Bookings



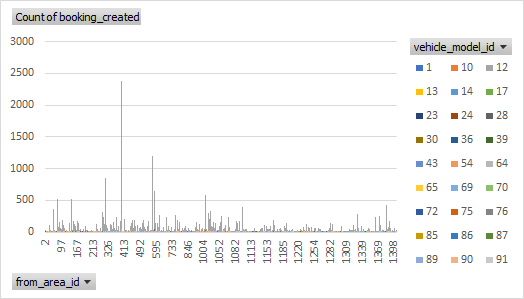
2. No of Bookings in each hour of a day.



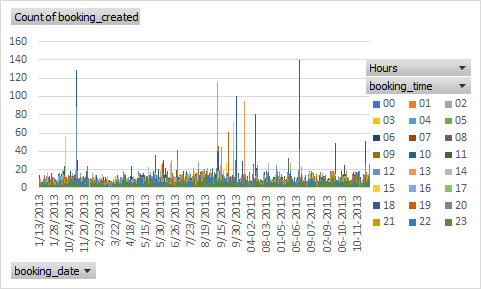
3. No of bookings corresponding to each hour in every unique from\_area\_id in the training data.



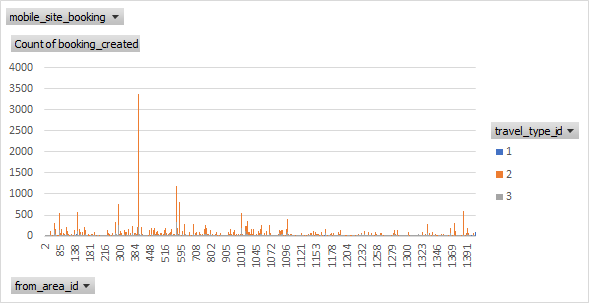
4. No of bookings corresponding to each vehicle\_mode\_id. The maximum bookings are done for vehicle\_mode = 2.

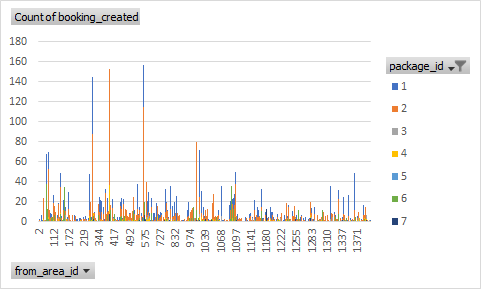


5. No of bookings in each hour at unique dates extracted from booking\_created attribute of the dataset.



6. Bookings corresponding to each travel\_type\_id. Mostly travel type 2 is used as depicted from the graph.





**Conclusions and Solutions**

On the basis of the above analysis, there are several conclusions and strategy that can be opted to increase the utilization rates of the cabs and to increase the revenue overall.

1. Surge Pricing

When demand increases, the algorithm can be adjusted with per mile price increase. This can most effectively be done in areas where there are maximum no of times and at times when bookings rate is higher, i.e in peak hours. Limiting the no of cars in its fleet can help increase the revenue to a higher level.

2. The timefall of July, October witness the highest no of bookings and thus price hikes can

Done at that time of the year.

3. In areas having less bookings at peak hours, promotional strategies can be applied to

attract more people thereby increasing the overall revenue.

4. As seen from the analysis, the most used vehicle mode was 12, so price for that vehicle

Can be increased at peak hours plus when there's less availability of these vehicle

Types.

5. Different conclusions about the lifestyle of people can be drawn by carefully analyzing their

Areas of interest n the type of vehicle mode they choose, package id they are often

Interested in, no of bookings outside and inside cities they book every year and based

Upon the results, different fares can be allocated to them,

6. Keeping a track of how many times each user cancels a trip, each time 2% increase can

Be applied as the multiple cancellation within a short period of time..

7. To increase its customers, XRides can save the frequently visited places by a person and

Can offer exclusive discounts and cashbacks for these locations.

8. Pooling of vehicles can help the company increase its customers and at the same time

Saving on company’s fuel charges for transportation. It also provides a good solution for

Tackling heavy demand and less supply.