**Delhivery Business Case Study**

**Problem Statement:**

Delhivery is the largest and fastest-growing fully integrated player in India by revenue in Fiscal 2021. They aim to build the operating system for commerce, through a combination of world-class infrastructure, logistics operations of the highest quality, and cutting-edge engineering and technology capabilities. The company wants to understand and process the data coming out of data engineering pipelines:

• Clean, sanitize and manipulate data to get useful features out of raw fields

• Make sense out of the raw data and help the data science team to build forecasting models on it

**Basic data cleaning and exploration:**

Q) Analyse the data types of the various features available in the data set ?

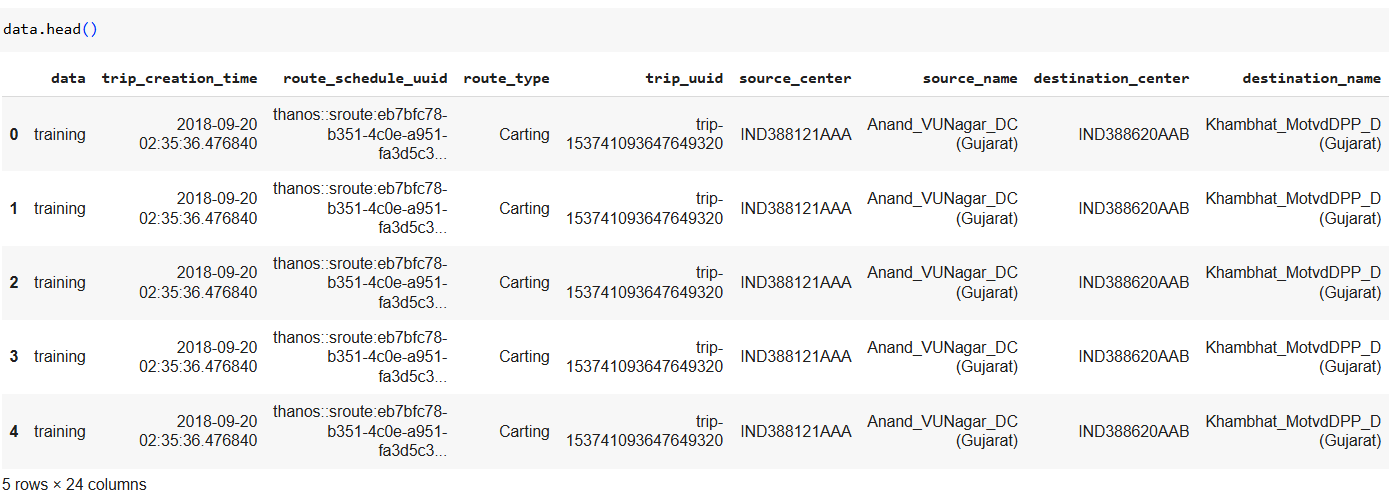
Ans)

A screenshot of a computer program

Description automatically generated

Insights : There are 24 different columns or features available in the data set.Among these, 11 are of Float data type and rest 13 are object data type. Float data type mostly points to the numeric data set and the rest points to categorical types.

Sample Data set:



Q) Check for any missing values in various columns o f the given data set.

Ans)

A screenshot of a computer program

Description automatically generated

Insights : There are a total 20 features that have missing values. Among them, there are two values which has maximum number of missing values. Rest of the values are having one record missing which is negligible when compared to the size of the data set. The two values with maximum values missing are source\_name (213) and destination\_name (167)

Q) What percentage of the total data set are the missing values ?

Ans )

A screenshot of a computer

Description automatically generated

Insights : The percentage of the missing values range from the max of 0.185 to the lowest of 0.000872. Percentage wise though they are very much less than 5% (below which we generally drop the values ) here we can impute for the two most calculated values

Q) Perform the imputation function over the missing values for the variables.

Ans) Since the two variables with most count of missing values – source\_name and destination\_name – are categorical in nature. Since they are categorical in nature, the best method to impute is to replace with the most frequent values. (had they been numerical variables then we could find the presence of outliers and then decide on whether to use median or mean )

A screenshot of a computer code

Description automatically generated

Q) Find the statisticals analysis for the numerical variables present.

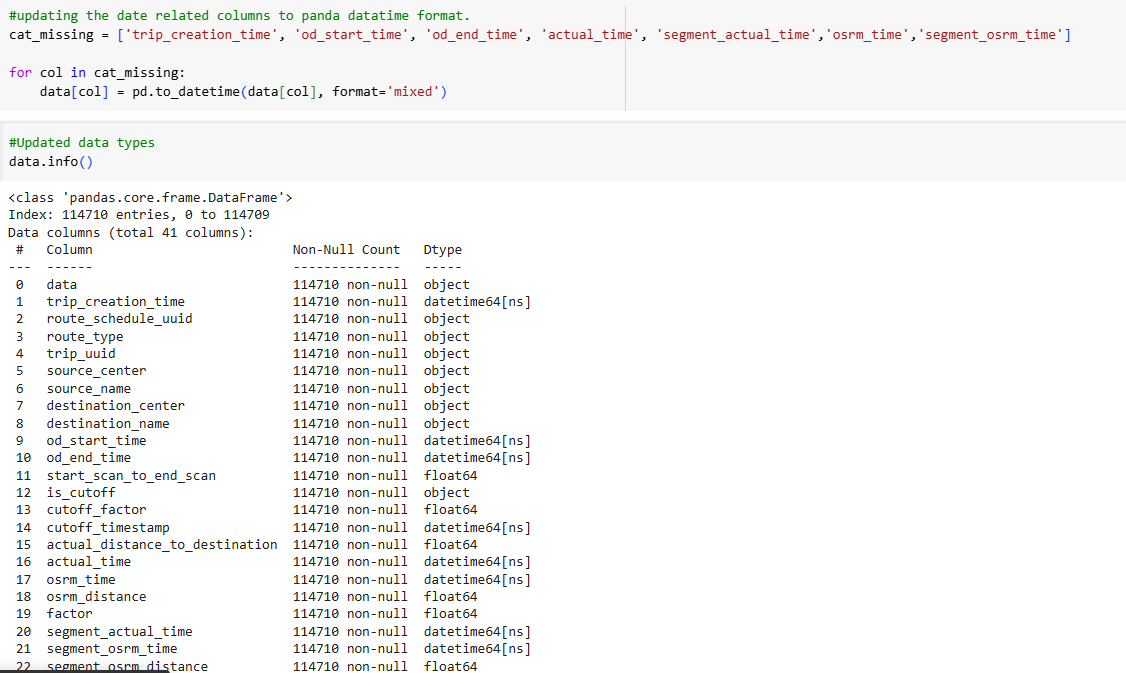
Ans )

A screenshot of a computer

Description automatically generated

Q) Change the data type of the time / date related columns to date time instead of generic object class.

Ans )



Q) Create a unique identifier for different segments of a trip based on the combination of the trip\_uuid, source\_center, and destination\_center and name it as segment\_key. b. You can use inbuilt functions like groupby and aggregations like cumsum() to merge the rows in columns segment\_actual\_time, segment\_osrm\_distance, segment\_osrm\_time based on the segment\_key. c. This way you’ll get new columns named segment\_actual\_time\_sum, segment\_osrm\_distance\_sum, segment\_osrm\_time\_sum.

Ans)

A screenshot of a computer program

Description automatically generated

Insights : segment key is created with the combination of trip id , source and destination. Also an aggregation dictionary is created which will help in mapping the aggregation functions to the columns associated with them.At the end, the data is sorted based upon the last value of the endtime for various trips in ascending order.

Q) For the column Trip\_creation\_time: Extract features like month, year, day, etc. Also Calculate time taken between od\_start\_time and od\_end\_time and keep it as a feature named od\_time\_diff\_hour.

Insights : From trip\_creation\_time we have extracted features like month, year, day etc so that we can have better analysis with these time scales in action. Also the difference in the start and end time that gives the total amount of time taken would be helpful in comparing against the osrm time or the actual time.

A screenshot of a computer

Description automatically generated

Q) For Destination\_name and Source\_name, split the values into various features which is initially in the format of City-place-code (State).

Ans) A screenshot of a computer

Description automatically generated

Insights : Since in the question, area code and places are mentioned, we are extracting the data to that level. In general we would need only the City and state in most of the cases. If data is to be checked at the very granular level, then the data of area code and place will also be helpful.

Q) Do a Grouping and Aggregating at Trip-level

A screenshot of a computer

Description automatically generated

Insights : a trip related dictionary is created with details of distance and time present in it along with the aggregation functions that will be helpful for us. The data is group with respect to the trip id – trip\_uuid.

**Visual Analysis :**

Q) Find any existing outliers in numerical features ? Use box plot for the same.

Ans )

A screenshot of a graph

Description automatically generated

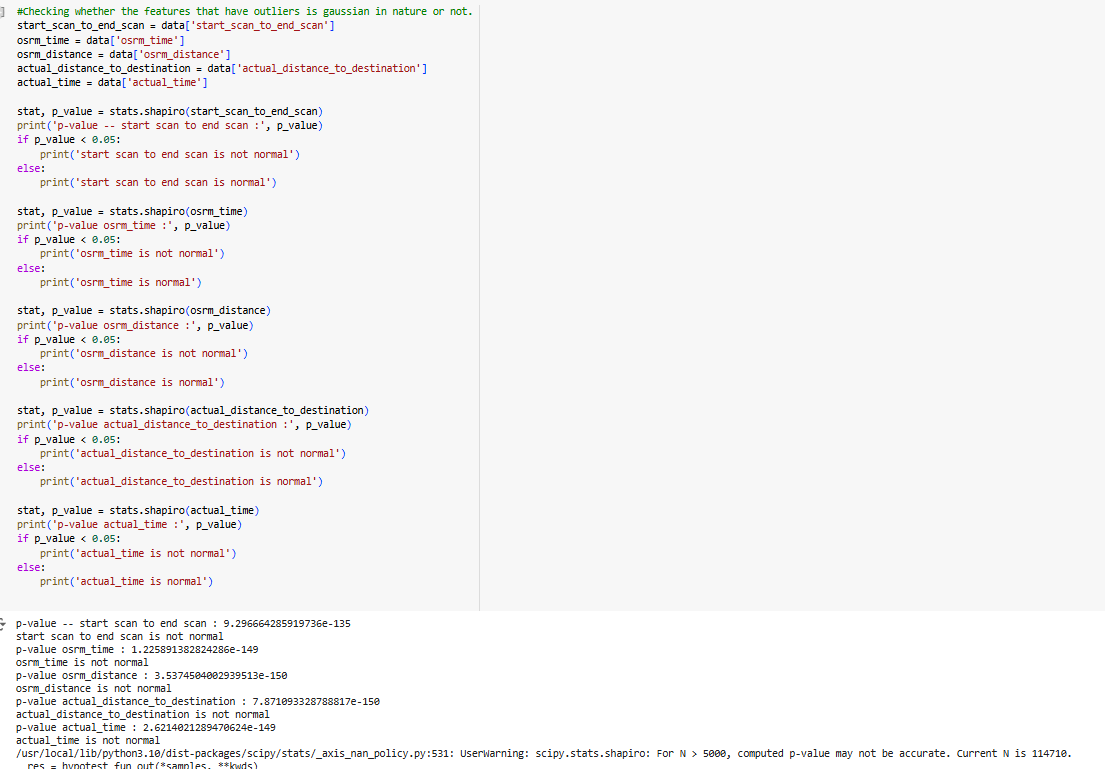
Insights : of all the numeric features, the columns – cut off factor, factor and segment factor are removed from the data since there are already mentioned as unknown in the column profiling section. The Rest of numeric features are plotted for their box plot visualization. It can be easily inferred that all the numeric features are having outliers present in their data. Especially for the column – Actual time , segment\_actual\_time. The rest of the columns too have their own fair share of high outlier composition.

This implies that the values for these columns are most extreme values which would disturb the mean value of the feature. This also says that the features are having skewed distribution.

Recommendations : Since the distance and time components are having so much outliers, efforts are to be taken as to see that the deviation of the delivery time and the routes decided as not as high always in the initial stages. As the distance and time components gets increased the efficiency of delivery will also gets affected. Also as customers want fast delivery across distances and time, we should encourage to route and time in the least possible manner as to impress the customer.

Q) Check whether the distributions are normal or not ?

Ans)



Insights : This tells us that none of the variables mentioned above are having the gaussian distribution. So the selection of method for outlier treatment should be IQR method since there is skewness in the distribution.

Q) Handle the outliers using the IQR method

Ans) A screenshot of a computer

Description automatically generated

Insights : As the distribution has skewness we are using IQR method for the outlier treatment. What happens here is that we find the IQR values and then remove all the values lying outside the particular range of Q3- Q1. This will help us eliminate the extreme values and we can focus upon the values that are concentrated in the general scenarios.

Q) Perform one-hot encoding on categorical features.

Ans)

A screenshot of a computer

Description automatically generated

Insights : One hot encoding is one of the three methods used for converting categorical variables to numerical variables. In this method, the individual values of the variables are taken as separate features / columns and then we will be assigning Binary values of true or false for then across the other features. A Limitation for this method is that if the number of unique values within the feature is more, then one hot encoding will only increase the complexity by increasing the number of features and space allocation. Hence we are using one hot encoding only for the variables – data ( values – testing and training ) and Route\_type ( FTL, carting).

Ultimately even though it is categorical in nature, computer will understand only in binary values or numbers. Hence one hot encoding is a good method of converting the same

Q) Normalize/ Standardize the numerical features using MinMaxScaler or StandardScaler.

Ans)

Insights : The numerical features since they don’t follow a gaussian behaviour, we will be doing normalization to the data set. Standardisation is done when the behaviour is normal in nature or approx. normal in nature.

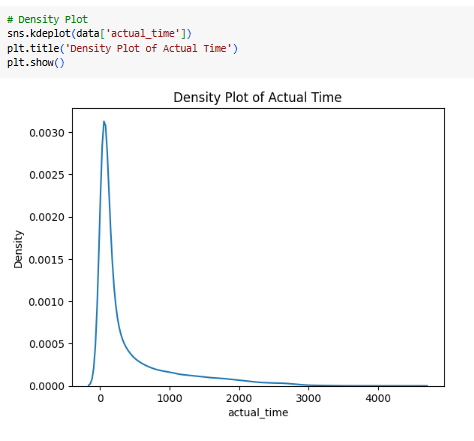
When normalisation is done, we ensure that the whole dataset of the features will lie in the range of 0 and 1. Other wise, multiple features would be spread across different base values which would difficult for analysis. Normalisation brings all the features under the same umbrella so that they can easily cross reference or analysed.

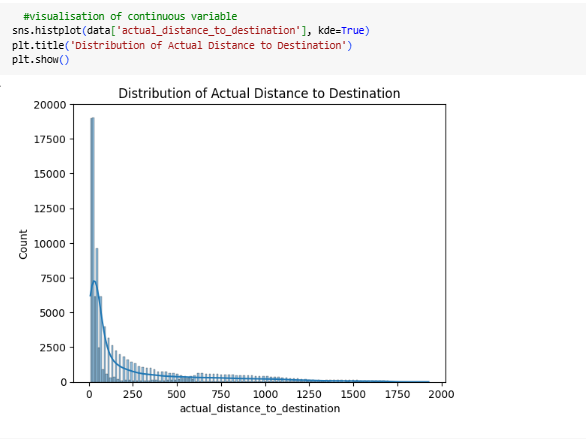
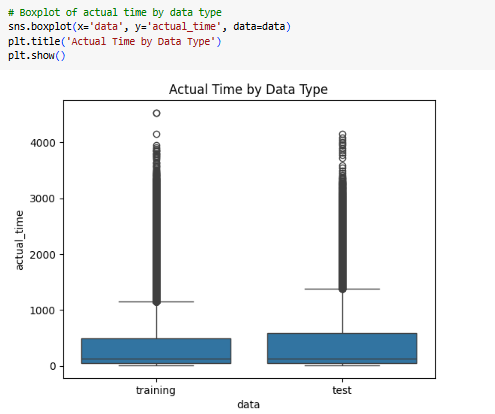
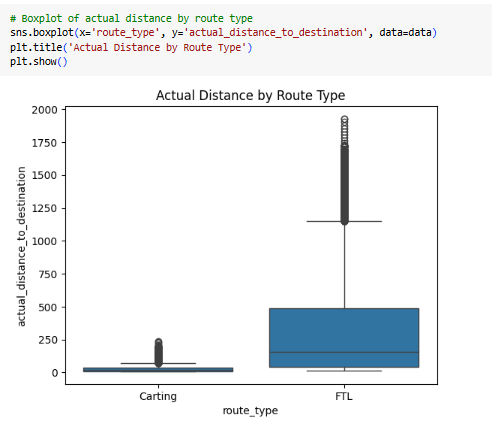
A screenshot of a computer

Description automatically generated

Q) do visual analysis for the continuous variables ?

Ans ) The visualization says that the distributions are not normal. They are having a skewness in the data.

****

****

The below graph says that the delivery time taken for the FTL routes are more when compared to the carting scenario. This is true since the carting would be for a segment of the total distance covered for the FTL.

**A screenshot of a computer screen

Description automatically generated**

**Hypothesis Testing:**

Q) Perform hypothesis testing / visual analysis between :

The given features since they are not normally distributed, we would be using KS test for the hypothesis analysis since it a non parametric test and it compares overall values.

* actual\_time aggregated value and OSRM time aggregated value.

Ans)

A screen shot of a computer

Description automatically generated

* Insights : In this question – the Null Hypothesis is that as per hypothesis testing, actual\_time aggregated value and OSRM time aggregated value are not significantly related to each other. The alternate hypothesis would be these two values are significantly related to each other.

After testing, it is found that the p-value is less than the alpha value, hence we reject the null hypothesis. This means that there is significant relation between the actual time agg. And OSRM time agg.

Recommendations : While OSRM is the system devised time and actual time being the value is from the real world scenario, we can understand that the OSRM is working as expected since any change or deviation shown in the OSRM value also shows that the actual value has changed. Hence more routing and planning can be done on the basis on the OSRM time which would be approximated to the real life values. By providing proper inputs about the time, customers will be able to plan their pickup and drop of the consignments and will also have more faith in the organization since the estimated time is more or less similar to the actual time taken.

* actual\_time aggregated value and segment actual time aggregated value.

Ans) A computer screen shot of a computer code

Description automatically generated

Insights : In this question – the Null Hypothesis is that as per hypothesis testing, actual\_time aggregated value and segment actual time aggregated value are not significantly related to each other. The alternate hypothesis would be these two values are significantly related to each other.

After testing, it is found that the p-value is less than the alpha value, hence we reject the null hypothesis. This means that there is significant relation between the actual time agg. And segment actual time agg.

Recommendations : The segment osrm time is the calculated time for the segment. The Actual time taken is the aggregation for the segment times. This means that for all non FTL time calculations, the valuation are almost similar to that of the actual time taken to travel. Hence we can focus more on the segment travel which gives the opportunity to serve more cities at the same time while the transit for the major destination is also covered in the time calculated.

* OSRM distance aggregated value and segment OSRM distance aggregated value.

Ans)

A screenshot of a computer

Description automatically generated

Insights : In this question – the Null Hypothesis is that as per hypothesis testing, OSRM distance aggregated value and segment OSRM distance aggregated value are not significantly related to each other. The alternate hypothesis would be these two values are significantly related to each other.

After testing, it is found that the p-value is less than the alpha value, hence we reject the null hypothesis. This means that there is significant relation between the OSRM distance agg. And segment OSRM distance agg.

Recommendations : The segment OSRM distance would be segment level distance calculated by the system. When more destinations are covered, the spread of the company will increase there by increasing the number of orders created and thus overall revenue generation

* OSRM time aggregated value and segment OSRM time aggregated value

Ans)

**A screen shot of a computer

Description automatically generated**

Insights : In this question – the Null Hypothesis is that as per hypothesis testing, OSRM time aggregated value and segment OSRM time aggregated value are not significantly related to each other. The alternate hypothesis would be these two values are significantly related to each other.

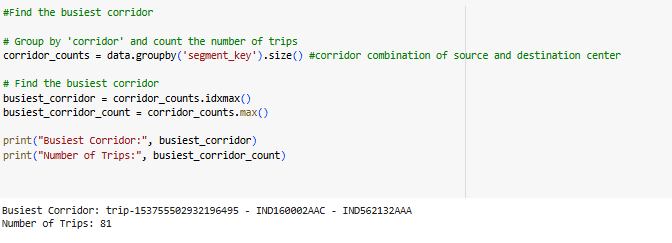
After testing, it is found that the p-value is less than the alpha value, hence we reject the null hypothesis. This means that there is significant relation between the OSRM time agg. And segment OSRM time agg.

Recommendations : The system calculated time value for the segment and for the total trip are related i.e. the calculations for the individual segment and the cumulative ones are per the design. Since we can predict the segment time we can get an idea of the total time for the trip. As this value is in tandem with the actual time, we should be able to predict a more or less accurate value for the time taken for the consignment to reach the destination. This would improve the faith among the customers and will also help in more engagement as the real world timings are more or less similar to that of the time predicted by the system

Q) Discuss the patterns observed in the data along with what you can infer from them?

* + Find the busiest corridor within the data set?

Ans)



Insights: as per analysis, the busiest corridor is found to be the Chandigarh\_Mehmdpur\_H (Punjab) [ IND160002AAC]-- Bangalore\_Nelmngla\_H (Karnataka) [IND562132AAA] which conducts a trip of 81 in the time period of the data set shared.

Recommendations : Since this is the busiest corridor, more number of logistics could be arranged in this section. More number of transport vehicles could be allocated and also we could identify the time period ( whether festive related or not ) and then price accordingly. We could also check for the presence of rest areas for the drivers and accompany person so that they get ample rest during the transit enabling safer commute and on time delivery. We can also introduce the slot booking scenarios for major clients who have done atleast sizeable delivery during all these time, so that they don’t miss out the opportunity to ship the consignment through us.

We can also try increasing the segment destinations in this route as this would be most used route for the time period. So this inturn will increase the spread of the company as well as increase the revenue too.

Q) What is the Average distance covered and time taken ?

Ans)

A screen shot of a computer code

Description automatically generated

Q) Check from where most orders are coming from?

Ans)

A screenshot of a computer

Description automatically generatedA screenshot of a computer

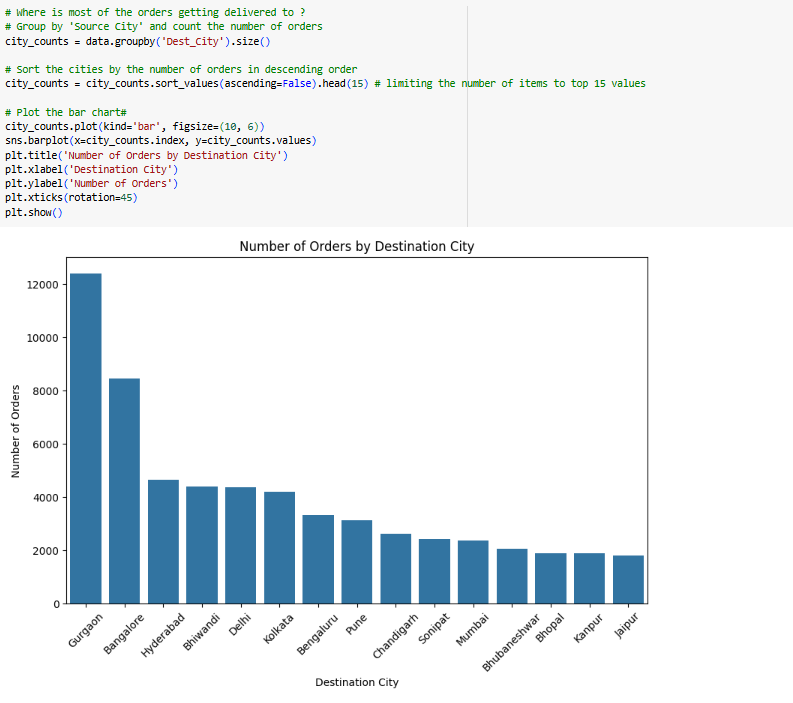
Description automatically generated

Insights : As per the data given above Gurgaon has topped the list of creating maximum number of delivery orders. It is followed by Bangalore and then by Bhiwandi.The number of orders from Gurgaon are exceptionally high as 18933 while the second most orders accounts for only 8403.

Recommendations : This gives us the insight that we should be able to increase our reach and spread in the area of Gurgaon. We can further check the values for order count at area code level which will help us understand more about the distribution of order creation and create offices so that the spread is increased as per the need in the city.

Q) Check where does most of the orders are delivered to ?

Ans)



A screenshot of a phone

Description automatically generated

Insights : The destination city that gets more number of delivery is Gurgaon followed by Bangalore and then Hyderabad. When we compare with the values of Source City we can understand that both in Gurgaon and Bangalore Intra city orders are created more. When compared to the total order count, the inter city count is minimal to these two cities.

Recommendations : This information tells us that the spread within the city of Gurgaon and Bangalore has to be increased. Since most of the orders are intra city orders. As discussed earlier, area code level analysis can be done which will help us understand the demand areas for both source and destinations and thus we can strategically place the offices.

Say if opening an office in a vicinity is not profitable enough then we could arrange for pickup at door step which can come with a cost for the same. If the customers are given a pleasant experience on the consignment creation to delivery, they would be ready to pay for the extra amount for the door step pickup facility. Alternatively, we can also provide them the pointers to the near by office from where they can book the consignment if they are ready to go in person.

Q) Are there any differences in performance in weekdays and weekends ?

Ans)

A screenshot of a computer screen

Description automatically generated

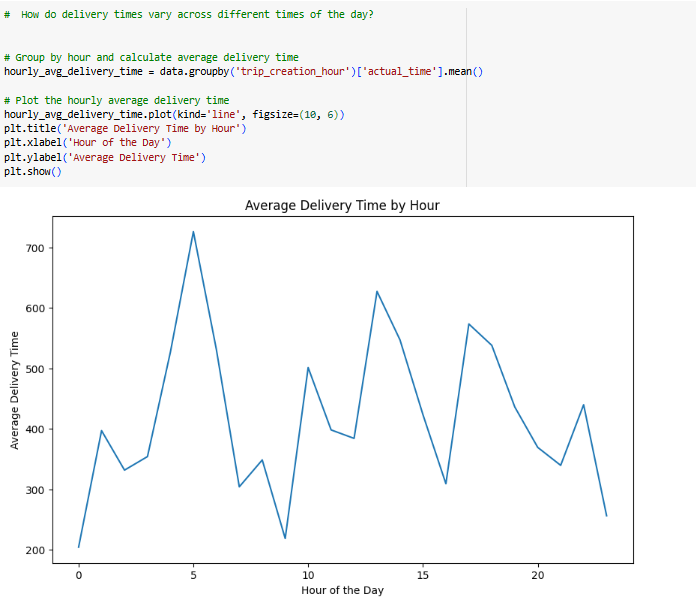
Insights : The above visualization confirms that there is no change in performance for the delivery of the consignment whether it is a weekday or a weekend. It can be noted that the highest numbers are shown for Wednesday and weekends of Saturday and Sunday as as good as any other working day.

Recommendations : Since performance across the weekends are similar to that of weekdays, we can ensure multiple shifts for the employees in office, especially during the weekends so that they also get their fair share of time with family at the same time, company is also generating revenue. Special incentives could be introduced for those working on weekends

Q)

Q) How does the delivery time vary across different times of the day ?\

Ans )



Insights: This visualization lets us know that most of the consignments start to the destination office in the early morning time and next would be afternoon time. The Early morning peak should be due to the long distance travels that would be starting to the destination in the early morning time so that the traffic is low. The Afternoon peak should be from bundling the consignments from the morning section.

Recommendations : Drivers and accompanying person to be ensured that they are not sleepy and they have got enough sleep previous night as they are starting very early to avoid traffic congestions.

Q) what is the variation of deliveries across the months?

A screenshot of a graph

Description automatically generated

Insights : this suggests that we have the data only for two months and among these two months the 9th month – September has more number of deliveries. This could be due to festive seasons which would be coming up in October month.

Recommendations : To tackle the increased volume of transactions, we can equip our software to bear more load as well as the transport vehicles should be finished with the maintenance work so that we avoid the delays due to mechanical failures of the vehicles. We should also equip with more buffer areas for the warehouse so that suppose there is an overflow for the existing resource we should be able to manage the same

Q)

A screenshot of a computer screen

Description automatically generated