 **PORTER DELIVERY TIME ESTIMATION & REGRESSION**

PROBLEM STATEMENT:

Porter is India's Largest Marketplace for Intra-City Logistics. Leader in the country's $40 billion intra-city logistics market, Porter strives to improve the lives of 1,50,000 + driver-partners by providing them with consistent earning & independence. Currently, the company has serviced 5+ million customers.

Porter works with a wide range of restaurants for delivering their items directly to the people. Porter has a number of delivery partners available for delivering the food, from various restaurants and wants to get an estimated delivery time that it can provide the customers on the basis of what they are ordering, from where and also the delivery partners.

EXPECTATION STATEMENT:

This dataset has the required data to train a regression model that will do the delivery time estimation, based on all those feature.

DATA EXPLANATION:

PORTER DATASET consisting of below features :

Market-id : integer id for the market where the restaurant lies

Created-at : the timestamp at which the order was placed

Actual-delivery time : the timestamp when the order was delivered

Store-id : encoded id for different stores

Order protocol : integer code value for order protocol(how the order was placed ie: through porter, call to restaurant, pre-booked, third part etc)

Total-items subtotal : final price of the order

Num-distinct items : the number of distinct items in the order

Min-item price : price of the cheapest item in the order

Max-item price : price of the costliest item in order

Total-onshift partners : number of delivery partners on duty at the time order was placed

Total-busy partners : number of delivery partners attending to other tasks

total-outstanding orders : total number of orders to be fulfilled at the moment.

SUMMARY:

* load the data and understand the features
* feature engineering creating target variable(time taken for each order)
* cleaning the data and visualization
* preparing the data for training
* Linear Regression.
* XGBOOST Regression.
* Random Forest Regression.
* Comparison between all the three models and explaining which one is better.

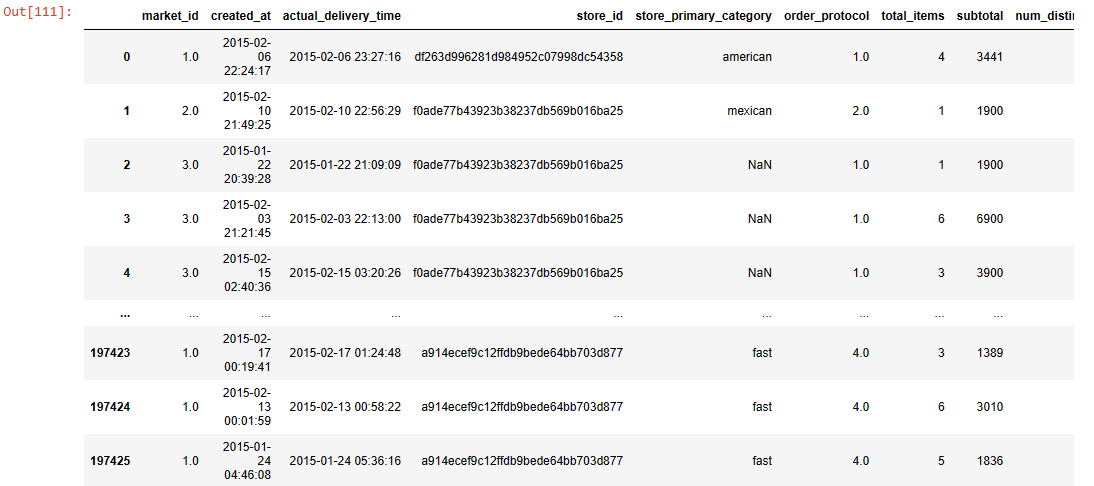
Importing libraries:

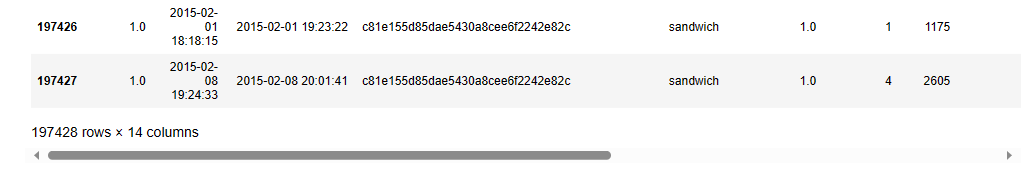


Loading Data & Exploring:

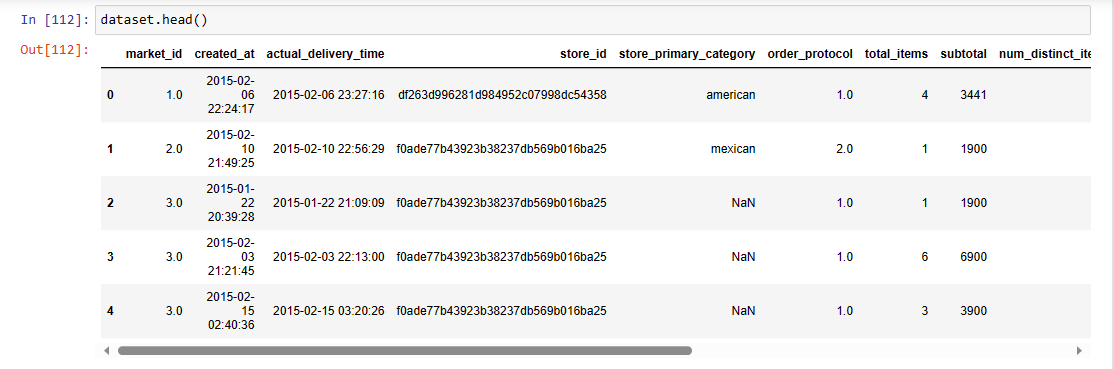


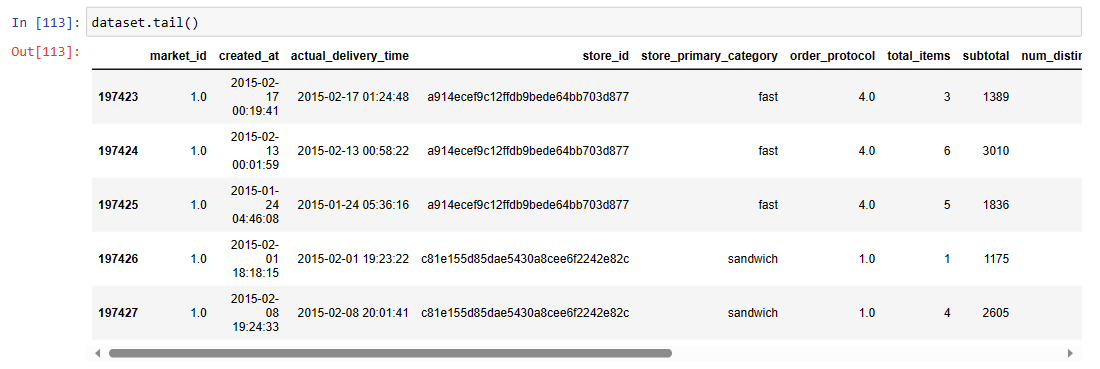


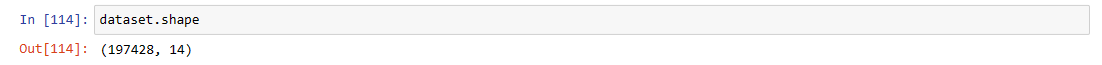




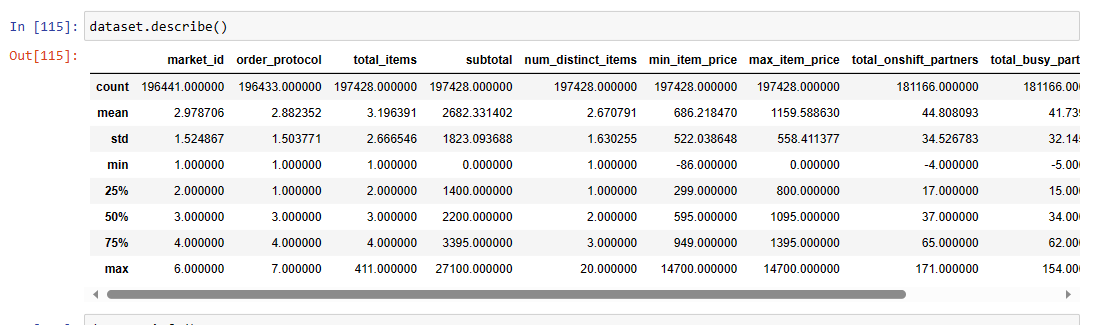
Checking Head ,tail ,null ,shape ,description and NA values from the dataset:



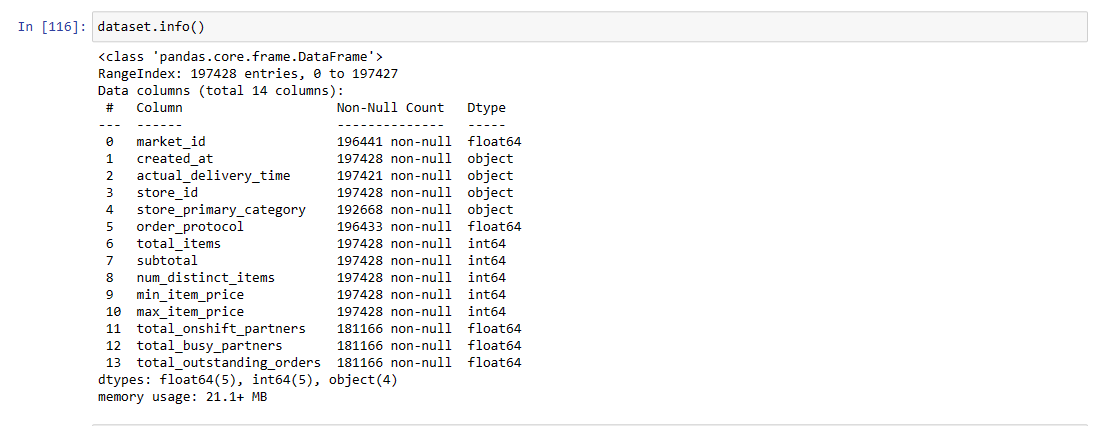




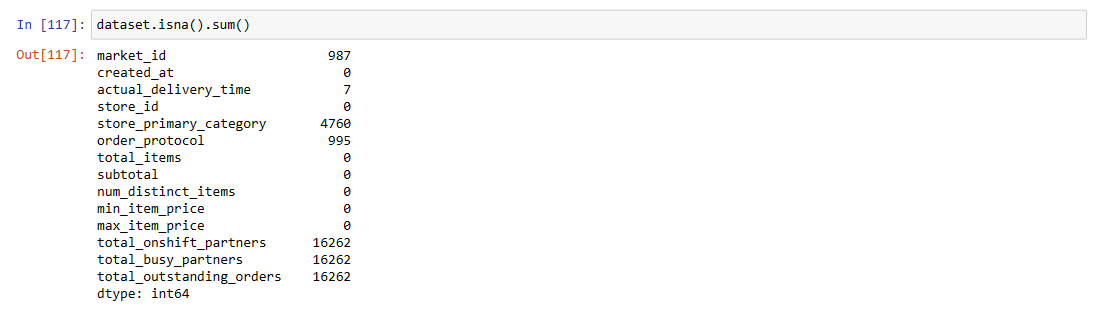
As per shape function there are total 197428 rows & 14 columns in the data.



In data description we can see min values less than 0 that is in negative .I have not operated these values as after removing this negative factors all the used regressor model were giving negative result.

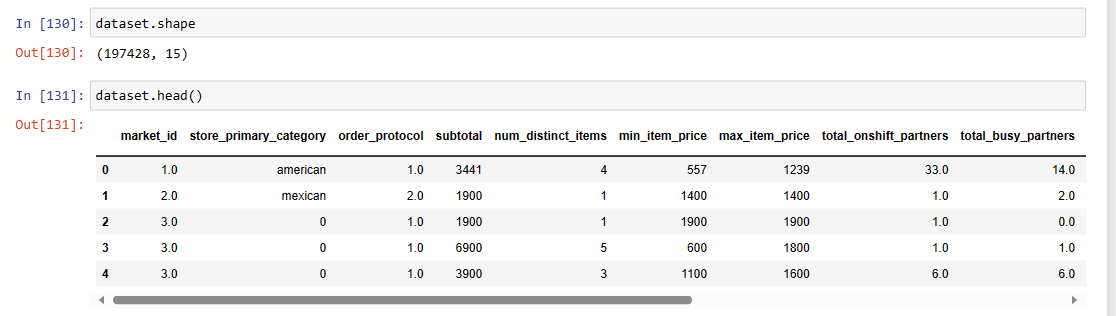


According to above fig ,no null values found in the dataset.



Further in NA values test I found there were total16% of data with NA. Hence filled NA with 0 .As first tried to fill data with the mean but it was affecting the final results.



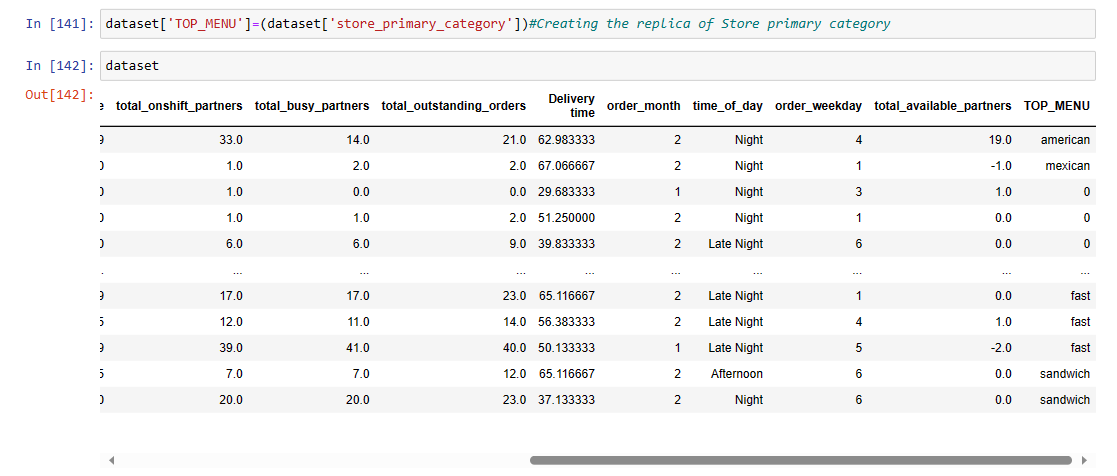


Data Cleaning:-

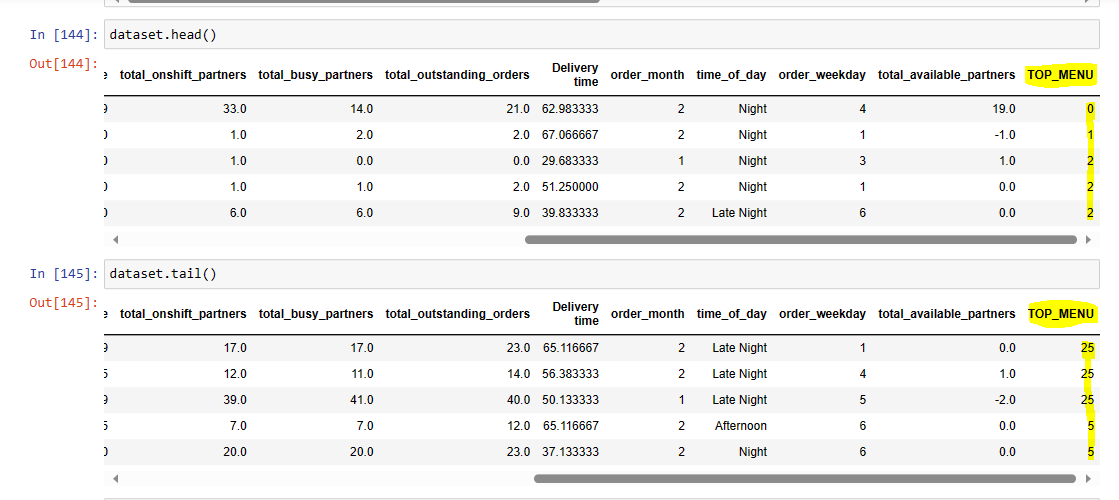


Making a duplicate column of store primary category as Top menu and replacing Top menu with numbers.

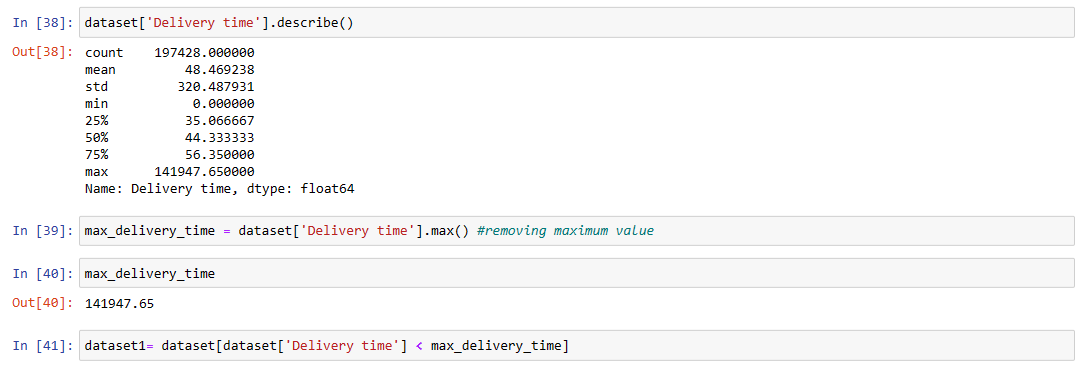




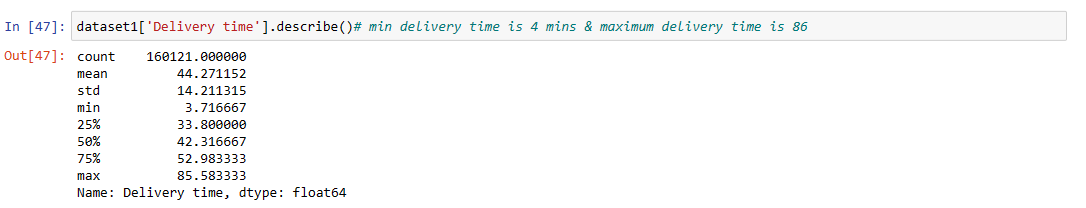




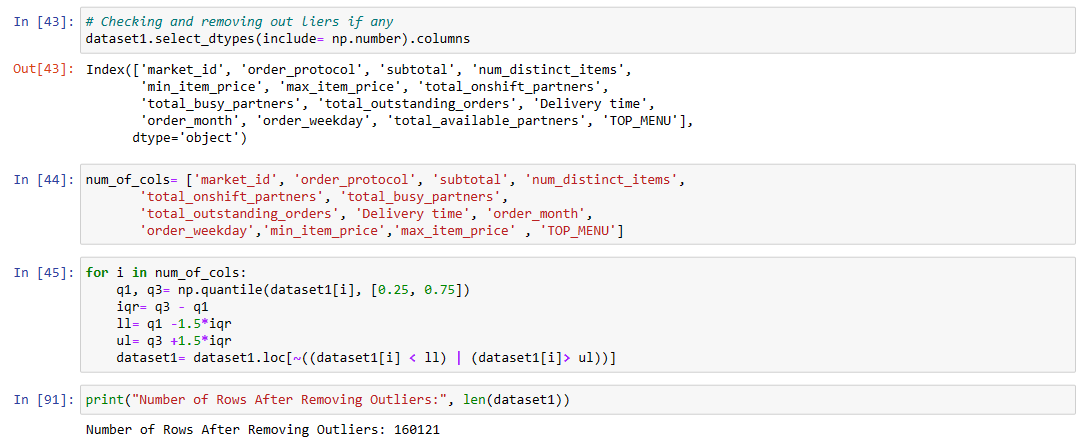
Removing maximum value which is affecting the final results:



After removing maximum value we can see maximum delivery time as 85min and minimum delivery time is 4 min.

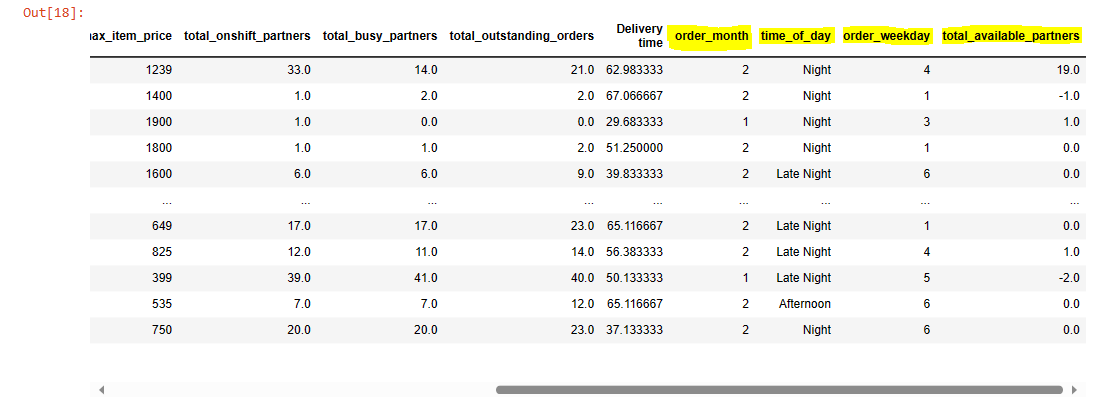


Removing outliers:



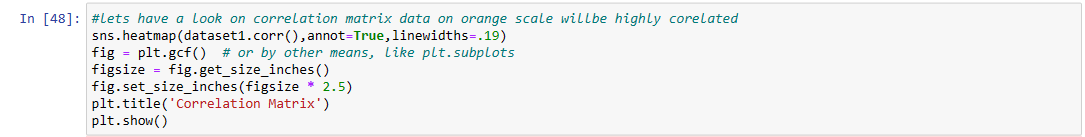
Feature engineering:-

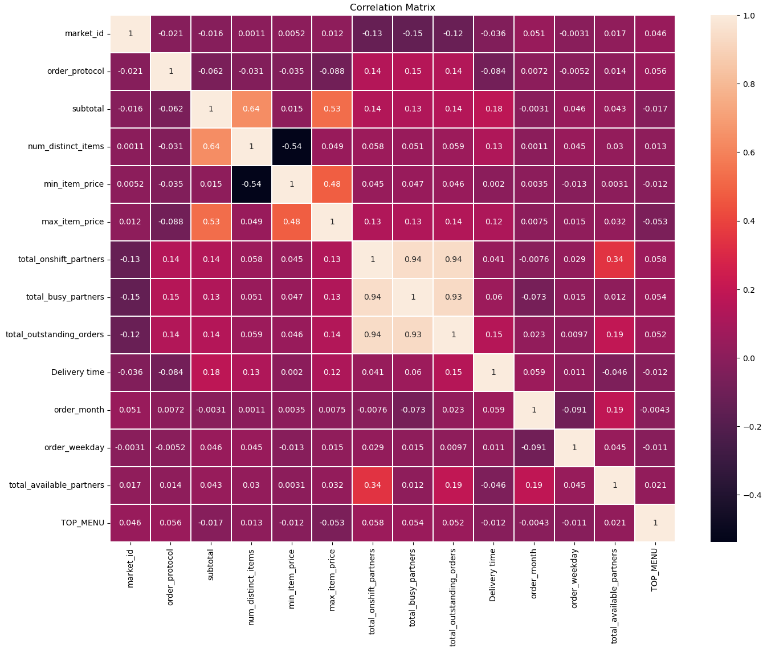


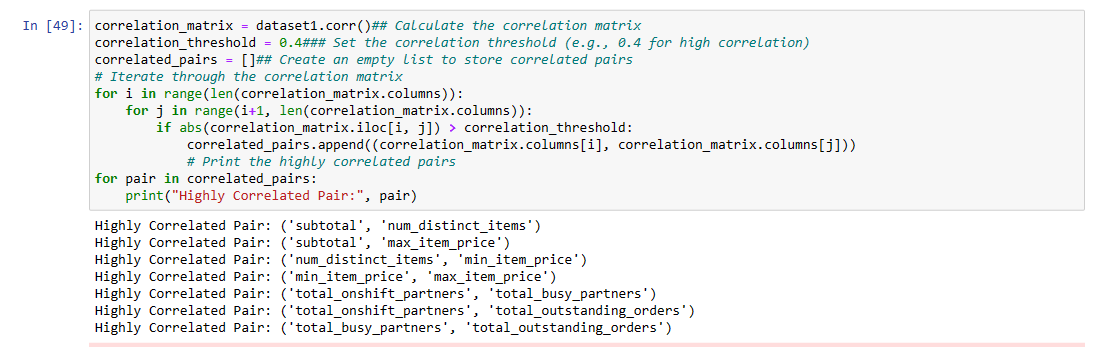


**Visualization:**

Correlation matrix and analysis:



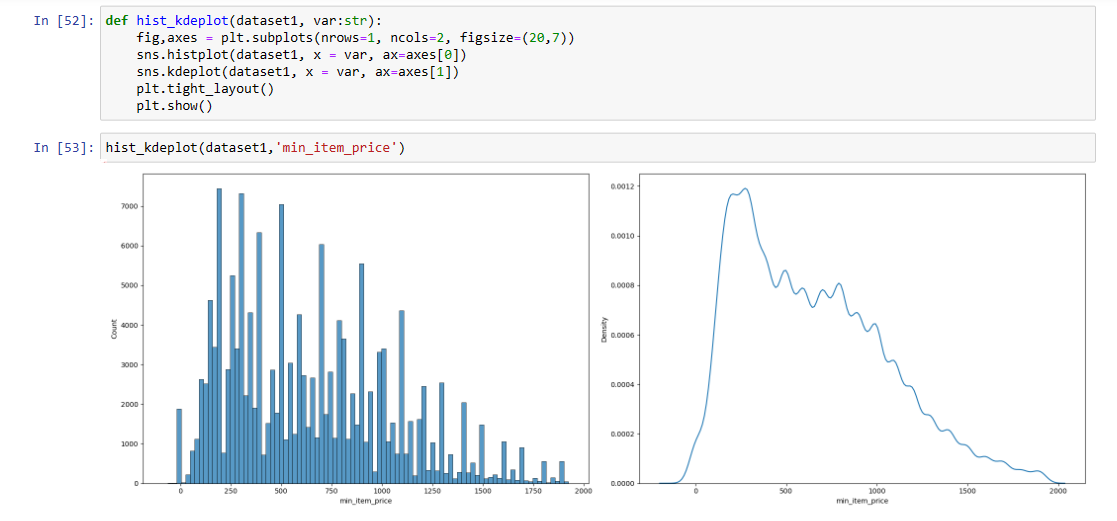


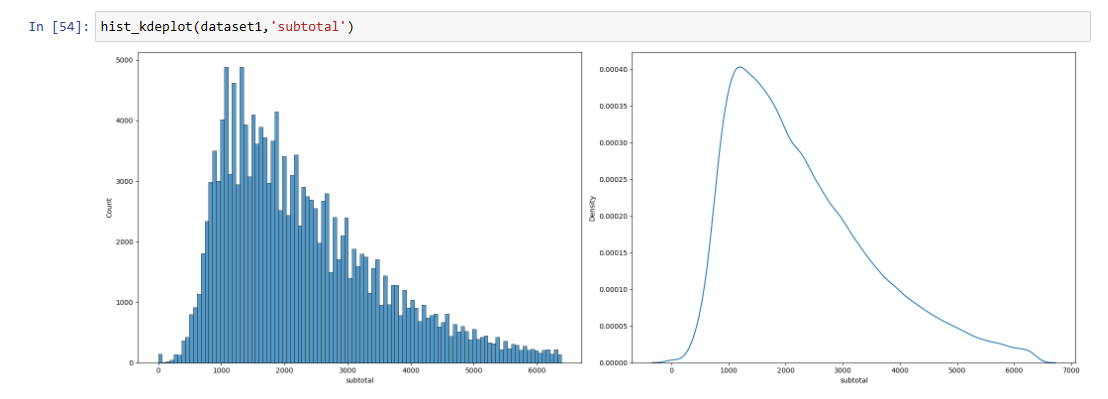


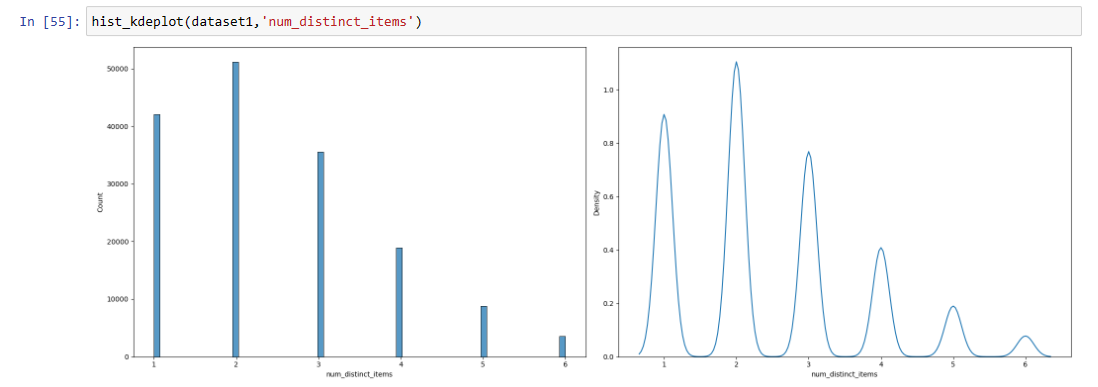
As highly correlated values are mentioned in orange range scale hence took threshold value as 0.4

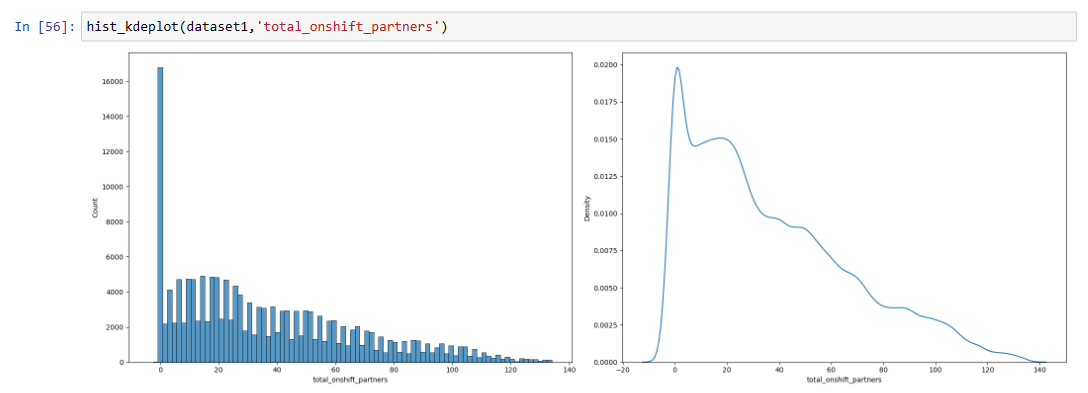
Above is the list of highly correlated features in the given snapshot.

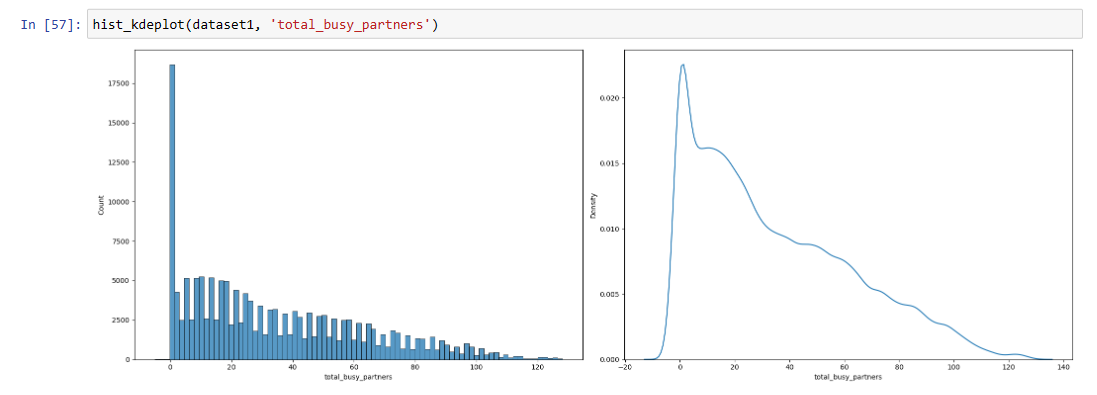
Histogram plot and density plot.

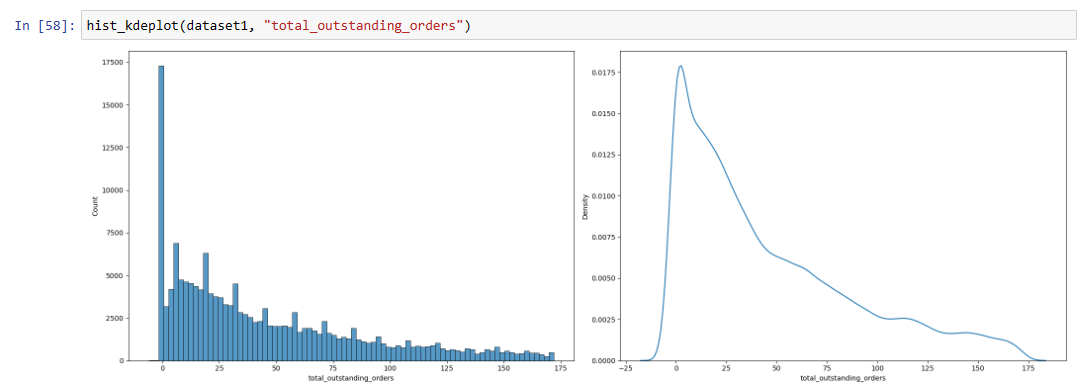






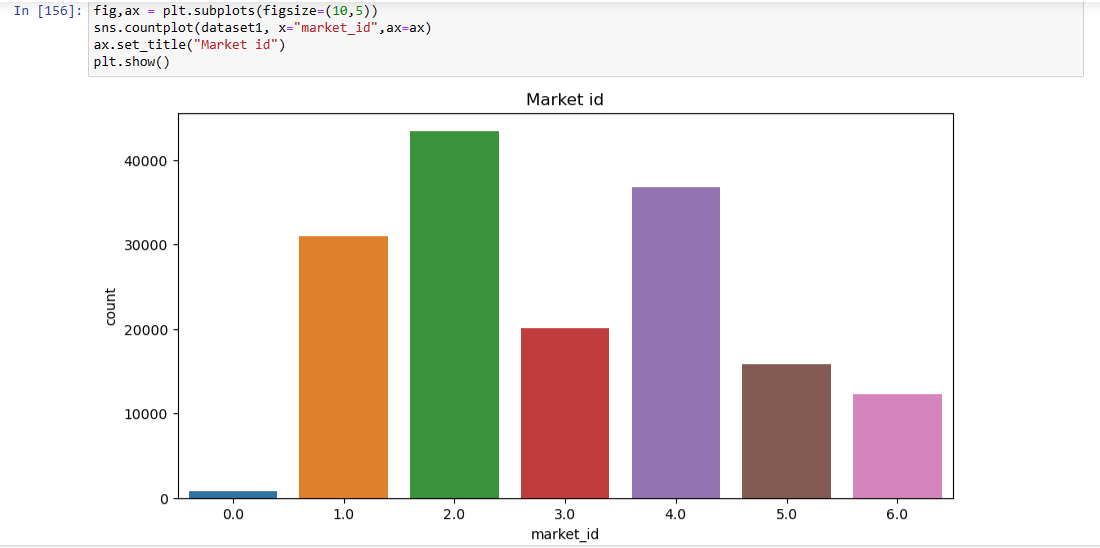






Num\_of\_distinct\_items is normally distributed and rest all are skewed to right.

MARKET ANALYSIS.

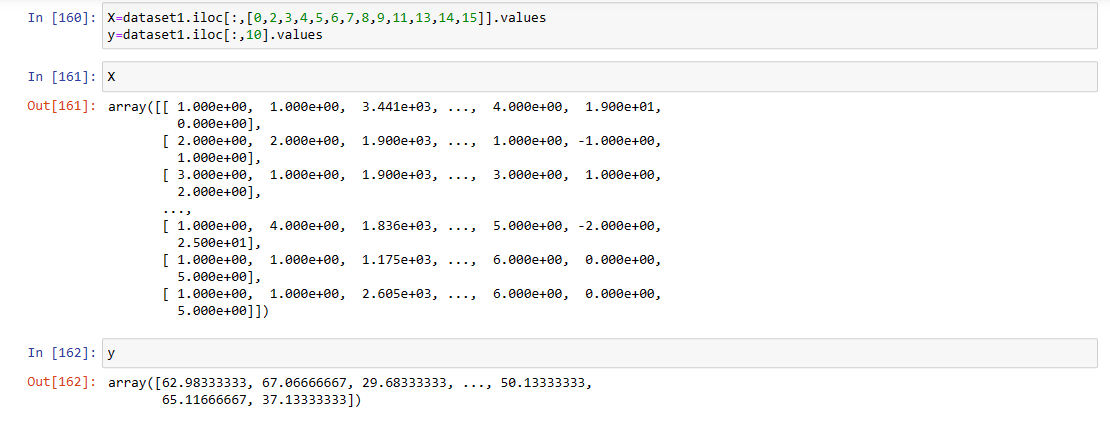


As per above figure market 2.0 has good business empire.

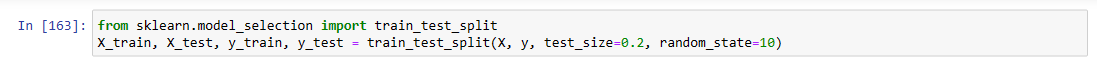
Selecting X & y

y as delivery time.

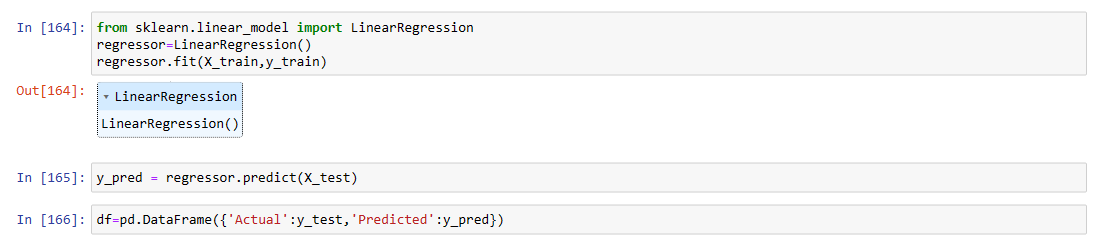
X as rest all features except delivery time.

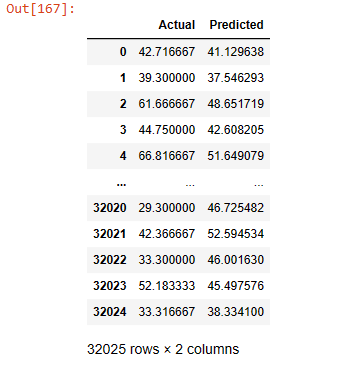


Training the data:

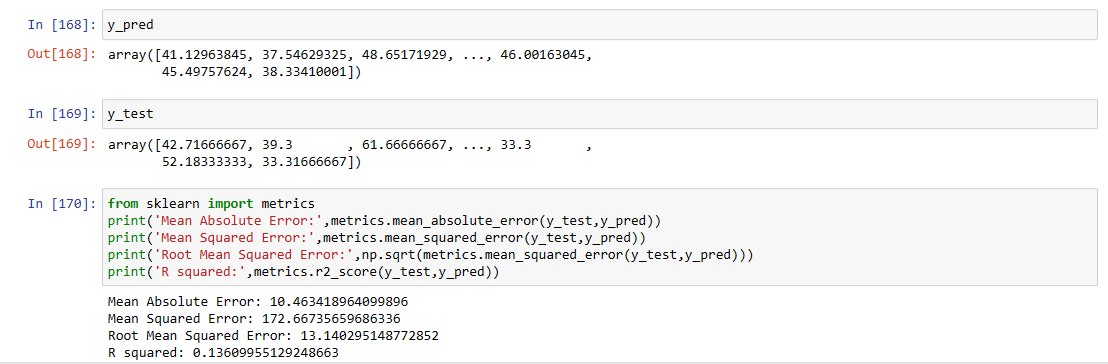


Applying linear regression:

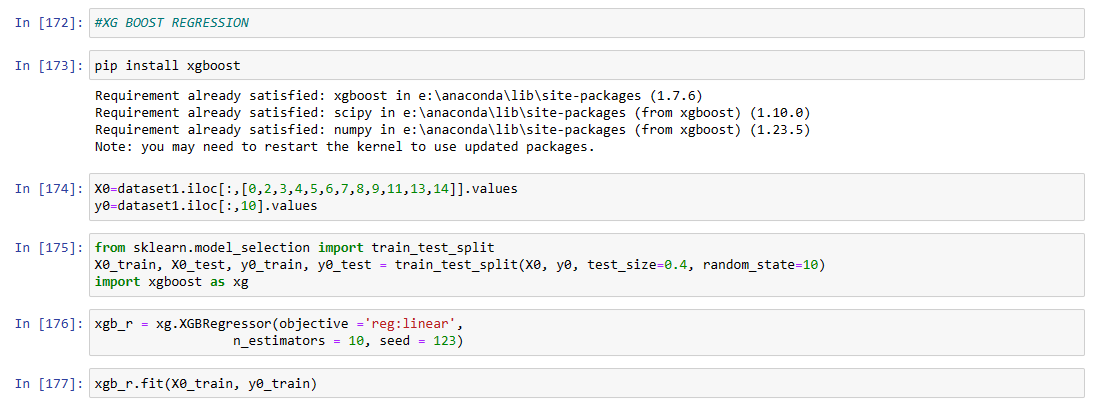


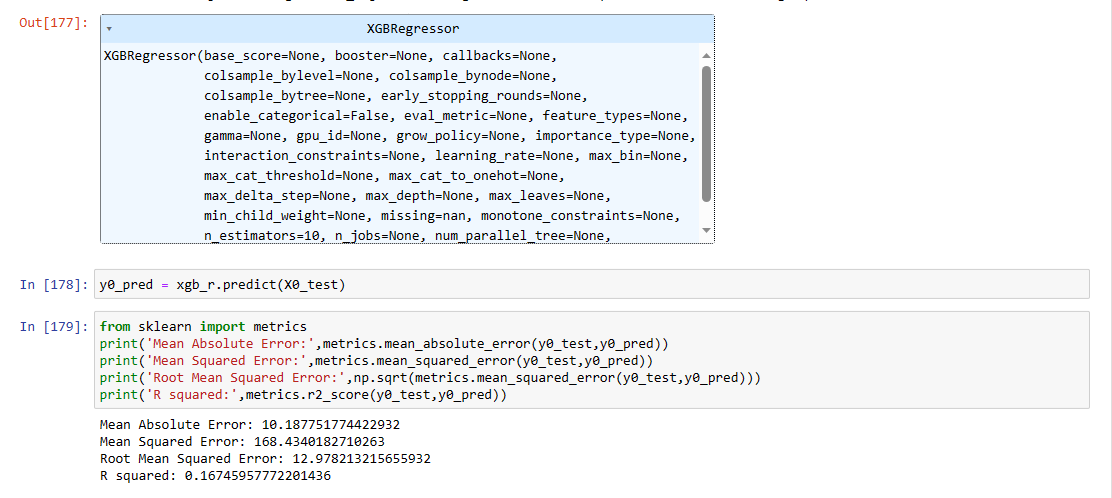


We can see the slight difference between actual and predicted value.



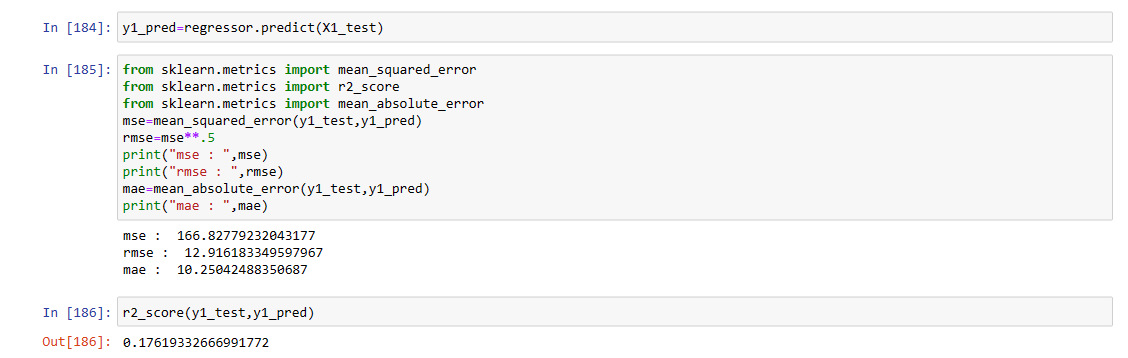
Applying XGBOOST Regression:





Applying random forest Regression:





Why metrics:

**Root Mean Square Error (RMSE), Mean Square Error (MSE), Mean Absolute Error (MAE), and R-squared (R²) are all evaluation metrics commonly used in regression problems to assess the performance of predictive models. Each metric provides different insights into how well a regression model fits the data and makes predictions.**

Metrics information:

(MSE)Mean squared error :( MSE = Σ(yᵢ - ŷᵢ)² / n) MSE is the measure of average square difference between the predicted values and actual values.

(MAE)Mean absolute error :( Σ|yᵢ - ŷᵢ| / n) MAE measures of the average absolute difference between predicted and actual values.

(RMSE)Root mean Squared error :( √(Σ(yᵢ - ŷᵢ)² / n)) RMSE is square root of MSE and provides the measure of the typical magnitude of the error.

(R Square):( R² = 1 - (SSR / SST))(R-squared measures the proportion of the variance in the dependent variable(target) that is explained by the independent variables in the model.

where:

* SSR: Sum of Squared Residuals (errors)
* SST: Total Sum of Squares (total variability in the data)

Why linear ,XGboost and random forest regressor.

For the continuous data we require all above regressions.

Conclusion:

Random forest Regressor:

* + RMSE: 12.91
  + MAE: 10.25
  + R²: 17.7%
  + It achieved lower RMSE and MAE values compared to Linear Regression.
  + Its R² value is higher than that of Linear Regression.

XGBoost regressor:

* + RMSE: 12.98
  + MAE: 10.19
  + R²: 16.7%
  + The XGBoost Regressor still shows relatively lower RMSE and MAE values compared to Linear Regression.
  + Its R² value is slightly lower than that of Random Forest.

Linear Regressor:

* + RMSE: 13.14
  + MAE: 10.46
  + R²: 13.6%
  + Linear Regression appears to have higher RMSE and MAE values compared to both Random Forest and XGBoost Regressors.
  + Its R² value is the lowest among the three.

Summary:

The Random Forest Regressor has achieved the highest R² value of 17.7% among the three models. This indicates that it explains the most variance in the target variable.

Both Random Forest and XGBoost Regressors have shown lower errors (RMSE and MAE) compared to Linear Regression, indicating better predictive accuracy.

The choice between Random Forest and XGBoost depends on factors such as computational efficiency, interpretability, and other considerations relevant to your specific problem.

Ultimately, the Random Forest Regressor appears to have performed slightly better overall based on the combination of R² and lower error values.

Earlier data was capturing RMSE for all the models till 18 but after removing outliers and cleaning the data I have succeeded to reach RMSE as 12 and R2 score for all the model in positive.

RMSE,MAE,MSE should carry lower values which are nearly equal to the actual values. Here in this case due to unstructured data even after making many of the changes in the dataset I landed up on 10 to 12 units higher and positive values.

Features like driving time estimation from restaurant to consumer is not given it may also affect the results .

Real world example for model used.

Linear regression: Can be suitable if use it for house price prediction.

XGBoost(Extreme Gradiant Boosting) regression: Can be used for customer churn(subscription) prediction (most of used in telecom company).

Random forest regression: Can be used for crop yield prediction.