**Project Report**

**Missing Value Analysis**

fare\_amount 24

pickup\_datetime 0

pickup\_longitude 0

pickup\_latitude 0

dropoff\_longitude 0

dropoff\_latitude 0

passenger\_count 55

dtype: int64

Total- 79 missing values. As it had a small percentage omitted the missing values.

**Feature Engineering**

Variables Month,Year,Time(Hrs), Day, Day/Night was derived from pickup\_datetime variable.

**Month-** Month of the cab rental.

**Year-** Year of the cab rental.

**Time-** Time of the cab rental.

**Day-** To classify weekday or weekend cab rental.

**Day/Night-** To classify whether the cab rental was carried on Sunlight/Moonlight.

Using pickup\_longitude, pickup\_latitude, dropoff\_longitude, dropoff\_latitude varibales, Distance\_Km was derived.

**Distance\_Km-** Distance travelled in KM.

**Outlier Analysis**

passenger\_count was kept under 6 and greater than 0.

fare\_amount greater than 0 was taken.

Distance\_Km greater than 0 was taken.

Boxplot outliers of the following variables 'pickup\_longitude','dropoff\_longitude','dropoff\_latitude','fare\_amount' was carried on.

**Converted the Variables to appropriate datatypes**

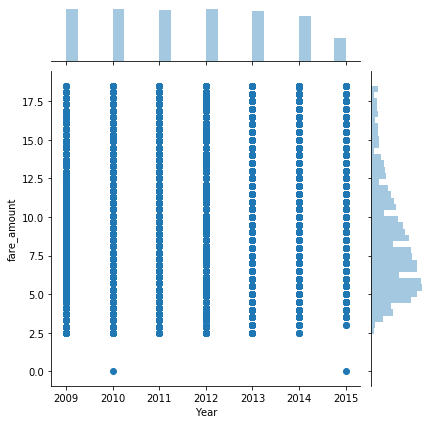
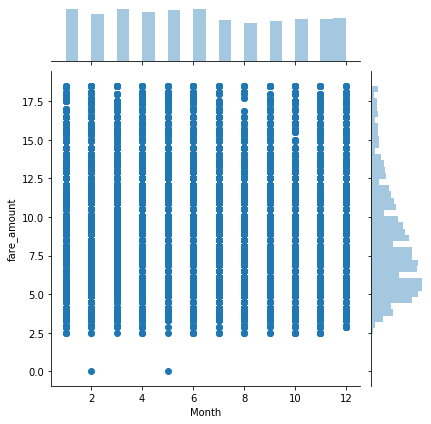
passenger\_count, Month, Year, Day, Day/Night variables are converted into factor variables and other variables were kept in numeric.

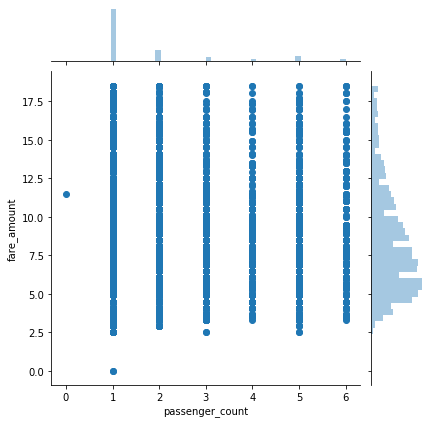
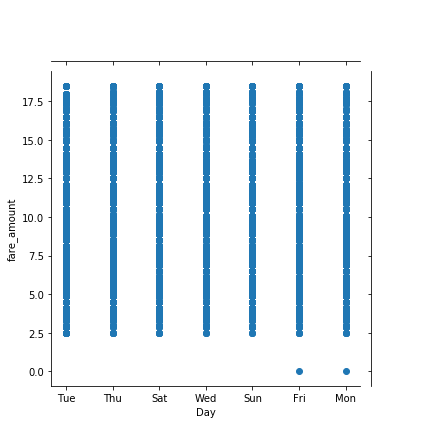
**Exploratory Data Analysis**

Scatter of Distance\_Km and fare\_amount provides us the idea that Distance and fare\_amount are directly propotional.



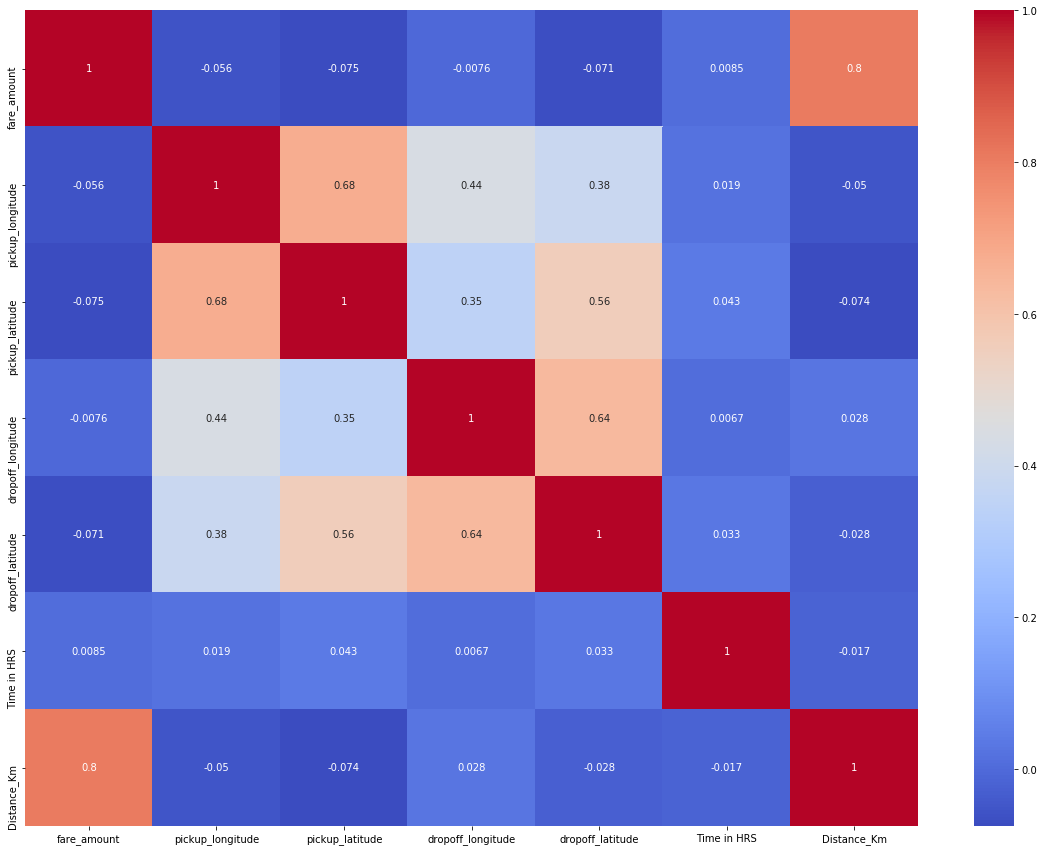
The distribution of fare\_amaount with factor variables was carried out. We can infer that the none of the factor variable has any proportionality with fare\_amount. Fare\_amount was equally distributed among every factors of each factor variables.



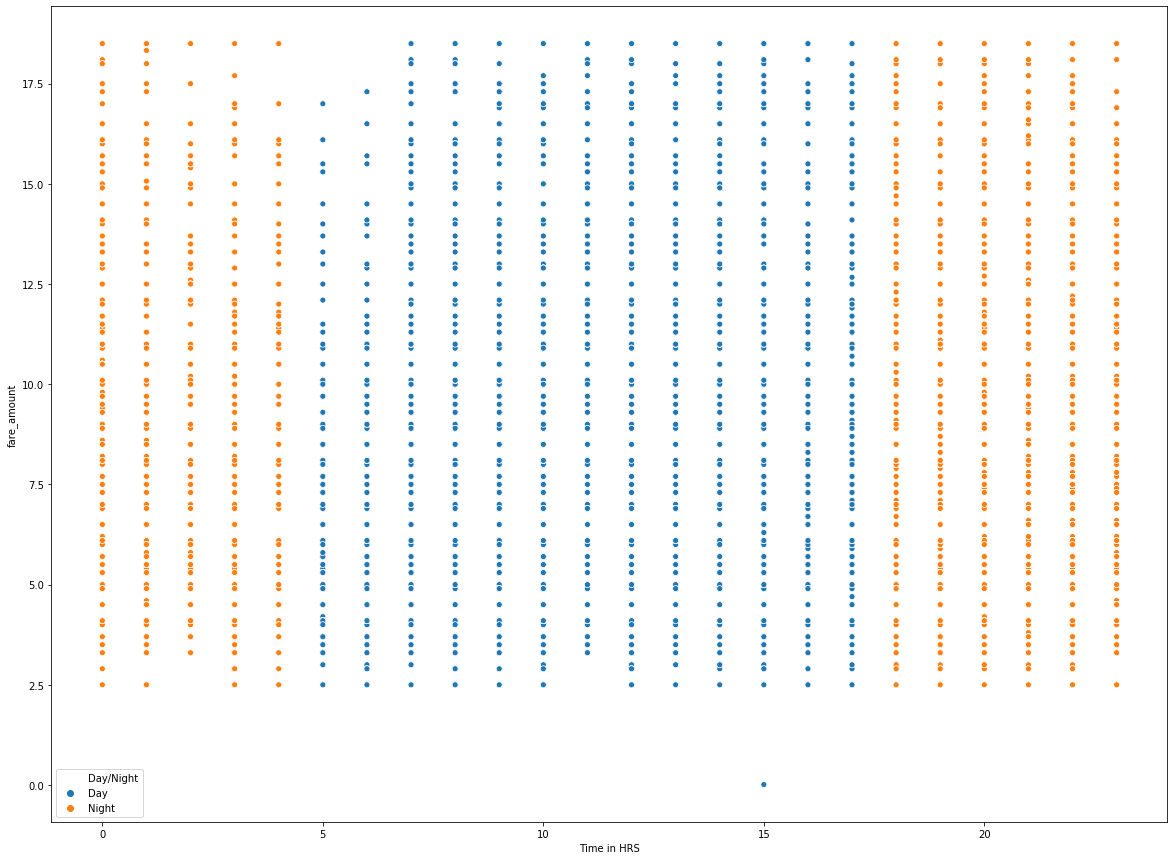
**Correlation Heat Map**



From Heat map we could conclude that there isn’t any multicollinearity.

**Dimensionality Reduction**

From the heat map analysis, the following variables are omitted.  
**'pickup\_datetime','pickup\_longitude','pickup\_latitude','dropoff\_longitude','dropoff\_latitude’** as they didn’t contribute much information and has some collinearity with other variables.



From the above graph the Time in Hrs doesn’t have any peak hours as the the fare amount is equally distributed. Also, we have another factor variable ‘Day/Night’ which gives us the info of sunlight and moonlight travel. Hence omitted the Time in Hrs variable.

**Creating Dummy Variables**

Created dummy variables for the following factor variables for better analysis.

**'Month','Year','Day','Day/Night','passenger\_count'.**

**Creating Models**

train\_cab.csv was split in train and test for evaluation purpose.

Error Metric used: - MAPE

|  |  |
| --- | --- |
| **In Python** | **In R** |
| Knn Regressor: -  Error percent: -34.48% | Knn Regressor: -  Error percent: -30.04 % |
| Decision Tree: -  Error percent: - 24.70% | Decision Tree: -  Error percent: - 20.30% |
| Linear Regression: -  Error percent: - 17.91% | Linear Regression: -  Error percent: - 18.2% |
| Random Forest: -  Error percent: -19.85% | Random Forest: -  Error percent: - 11.17 % |

From the error percentage of above analysis, LR model was applied on test.csv in python and Random Forest was applied on test.csv in R. The following predictions are written into the below csv files.

**Instructions to run and deploy code**

**Deploying the scripts through Widows Task Scheduler**

* Make sure the R and python interpreter paths are available on system environment (CMD).
* Keep the R script and .py file in particular directories.
* Open Task Schedulers -> and create new task
* Provide the heading.
* Under the Action tab provide the following paths
* Path of the Python and R scripts interpreter i.e. path of .exe file under Program/Script tab.
* Path of the program file of python or R under Start in (optional)
* Give the name of the program file under Add arguments.
* After this go to Triggers tab and create a new Trigger.
* Provide the appropriate date and time in which the script should run.
* The program would run on the given time or you could run the program temporarily by just clicking run.