Price Prediction using Regression on UBER Dataset for Boston City

Project for Business Analytics with R

Group Members:

- Abhirami Pillai
- Vaibhav Kumar
- Vinita Shinkar
- Sara Bali
- Kamal Preetham

Professor:
Subhra Rani Patra

Uber Rideshare Dataset Details

Objective

The objective of the study is to analyse the parameters affecting Uber prices in Boston city and predicting the price.

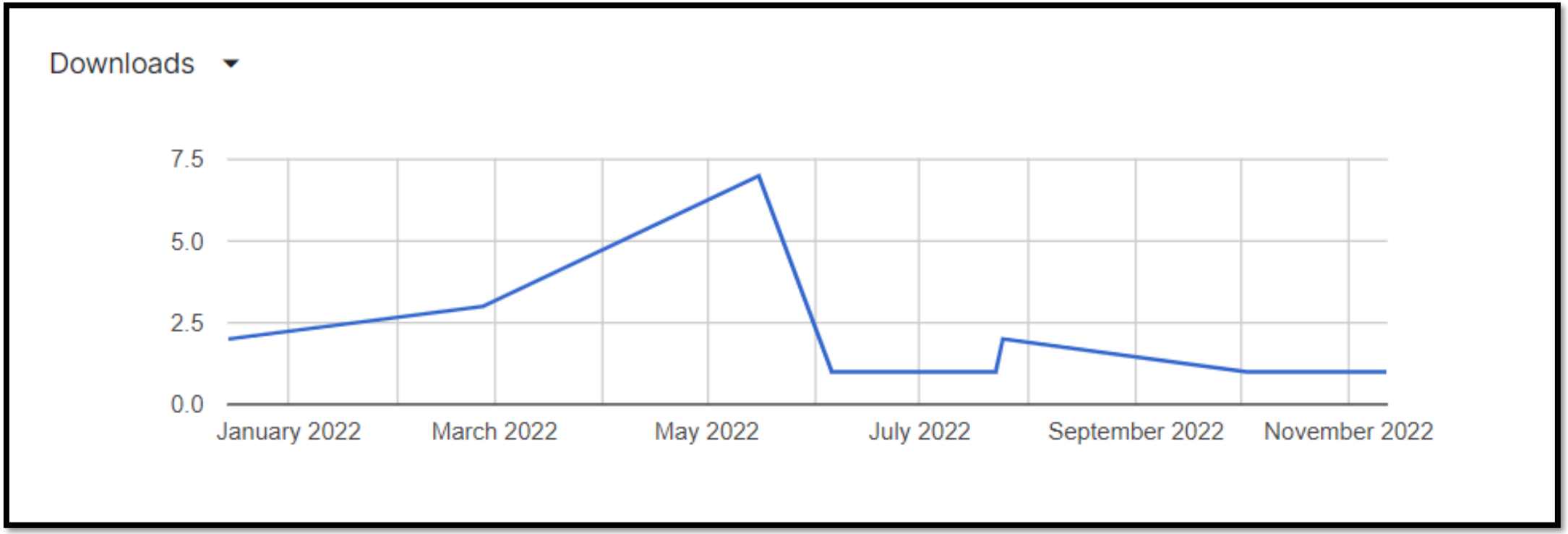
Activity Overview

DATASET STATS

VIEWS	DOWNLOADS
987	68

Details

Data Source: Kaggle
Data Format: CSV
Number of Data Entries: 693071
Number of Parameters: 57



Features in Dataset - 57

ID	Timestamp	Hour	day	month	datetime	timezone	source	destination	cab_type
product_id	price	distance	surge_multiplier	name	latitude	longitude	temperature	apparentTemperature	short_summary
long_summary	precipIntensity	precipProbability	humidity	windSpeed High temperature Low	windGust	windGustTime	visibility	temperature	temperature HighTime
temperatureLow	temperatureLowTime	apparentTemperature High	apparentTemperatureHighTime	apparentTemperatureLow	apparentTemperatureLowTime	icon	dewPoint	pressure	windBearing
cloudCover	uvIndex	visibility.1	ozone	sunriseTime	sunsetTime	moonPhase	precipIntensityMax	uvIndexTime	temperature Min
temperatureMinTime	temperatureMax	temperatureMaxTime	apparentTemperatureMin	apparentTemperatureMin Time	apparentTemperatureMax	apparentTemperatureMax Time			

Libraries and Packages Used

ML Bench

A collection of artificial and real-world machine learning benchmark problems, including, e.g., several data sets from the UCI repository.

Caret

The **caret** package (short for Classification And Regression Training) contains functions to streamline the model training process for complex regression and classification problems.

e1071

e1071 is a package for R programming that provides functions for statistic and probabilistic algorithms like a fuzzy classifier, naive Bayes classifier, bagged clustering, short-time Fourier transform, support vector machine, etc.

Dplyr

The dplyr package in R Programming Language is a structure of data manipulation that provides a uniform set of verbs, helping to resolve the most frequent data manipulation hurdles.

ggplot2

ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

tidyr

'tidyr' contains tools for changing the shape and hierarchy of a dataset, turning deeply nested lists into rectangular data frames, and extracting values out of string columns. It also includes tools for working with missing values.

Methodology of Code

Reading the original data and assigning it to dataframe

Taking the subset of the Cab type and excluding Lyft and considering Uber data

Taking the subset of Original Data and changing required variables for prediction to vectors.

Cleaning the data – Removing unnecessary columns.

Omitting the Null values.

Plotting the relationship between Price and Distance.

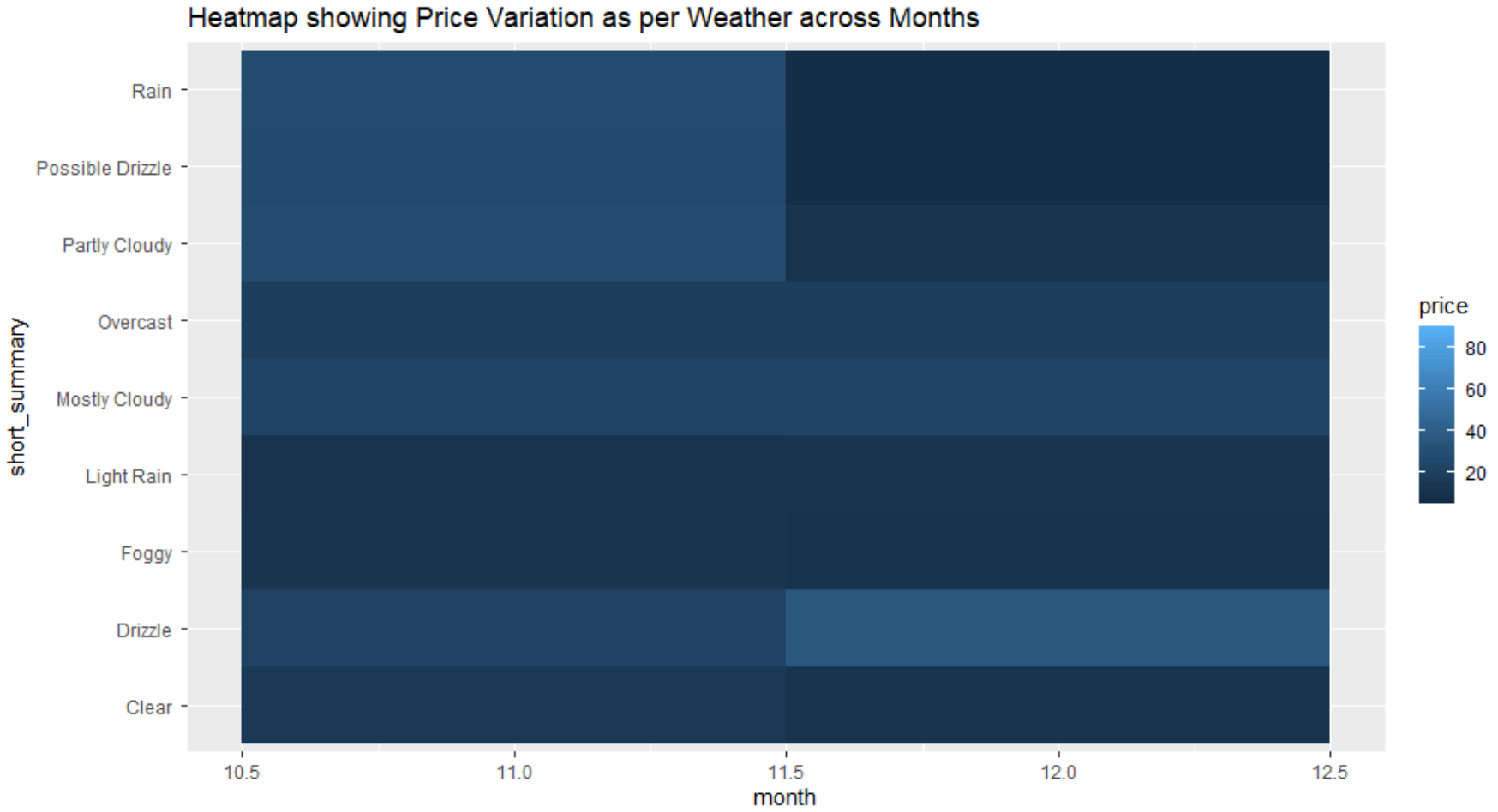
Converting the categorical variables to factors for Regression.

Splitting the Data into training and testing Running the Regression Models.

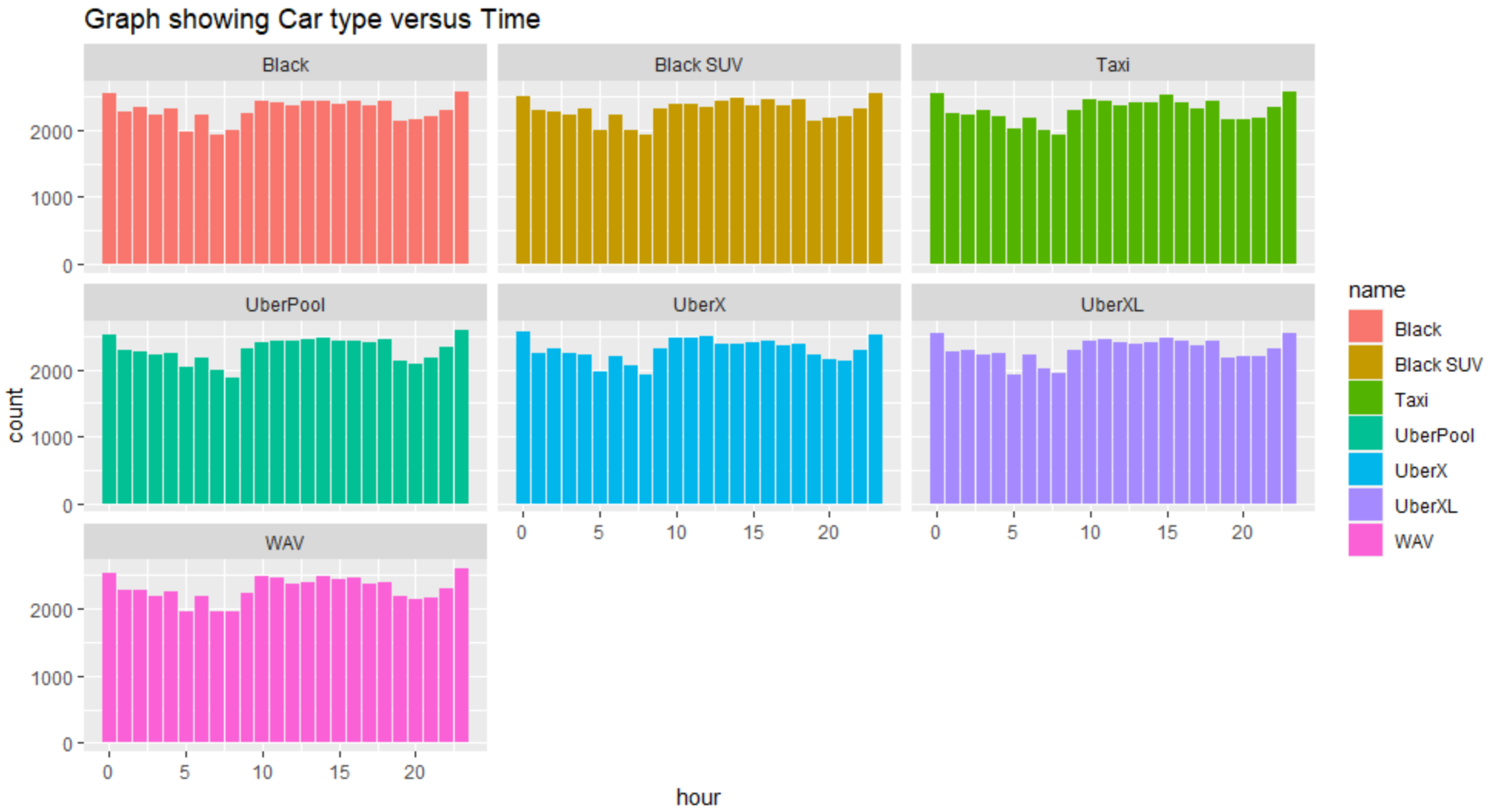
Running the Regression Models.

Comparing the efficiency of the Models.

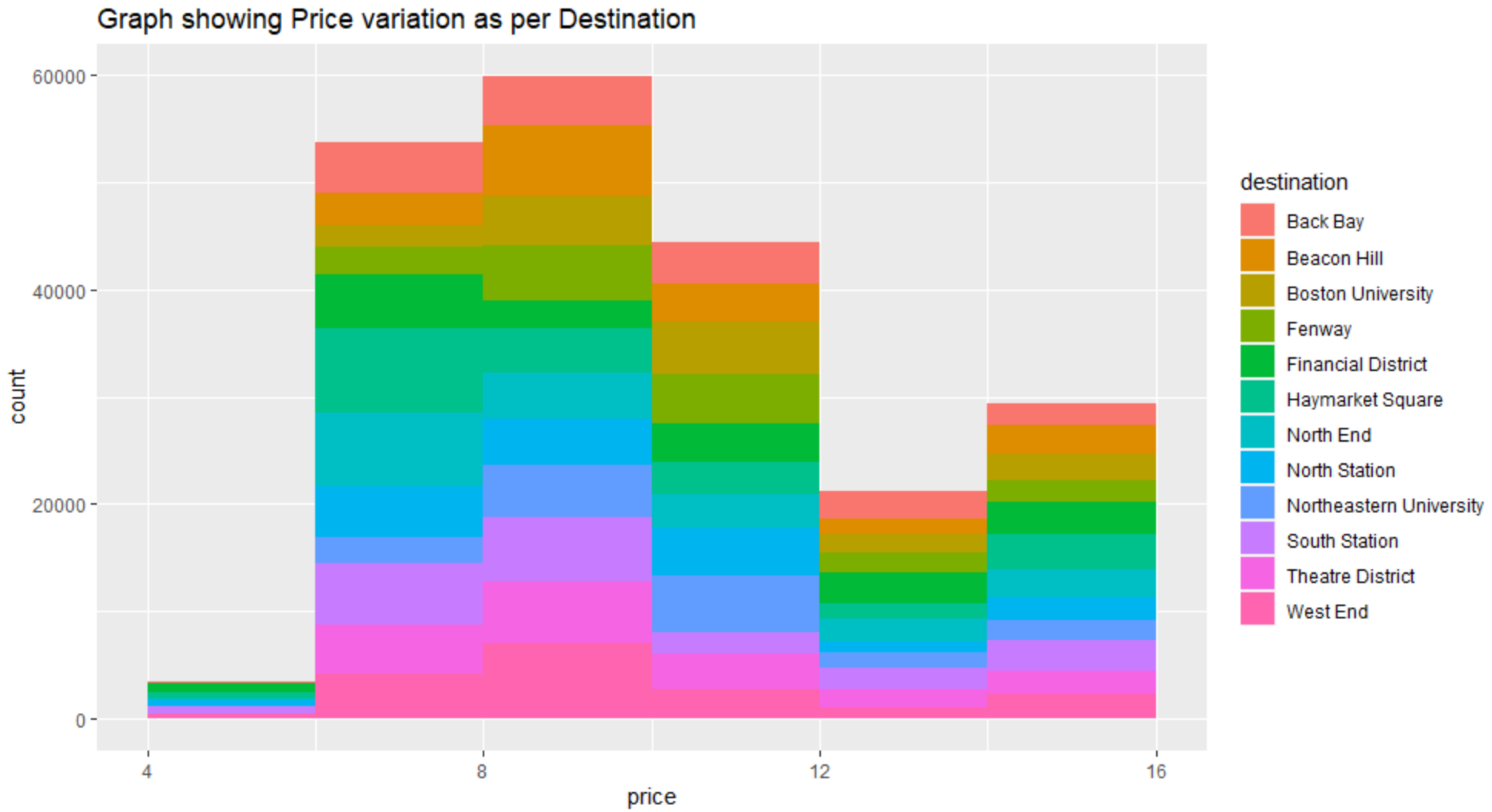
Visualisations



Highest prices are observed during the times when weather consists of Drizzle and Rain and Partly Cloudy. Least prices are observed when the weather is clear.

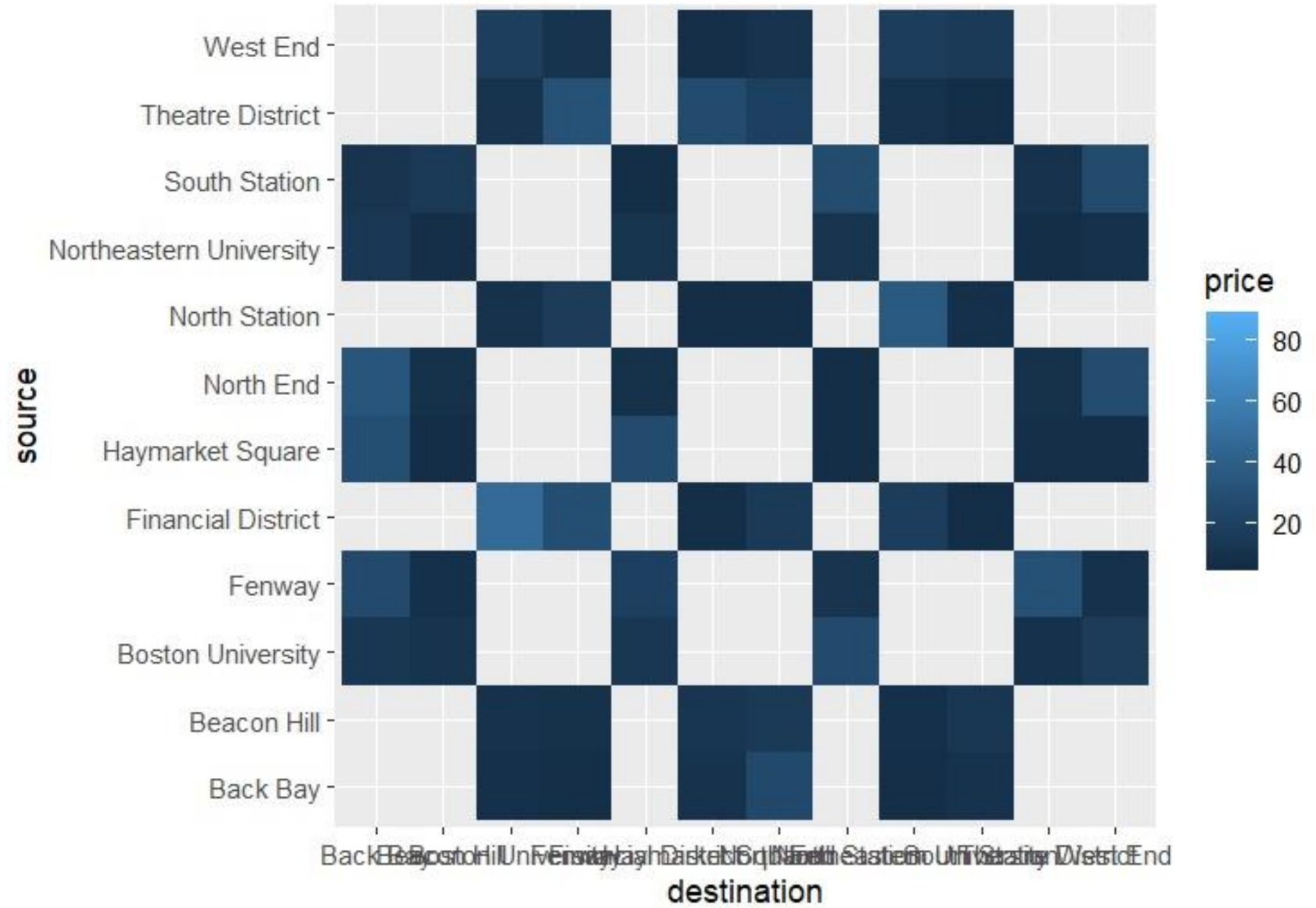


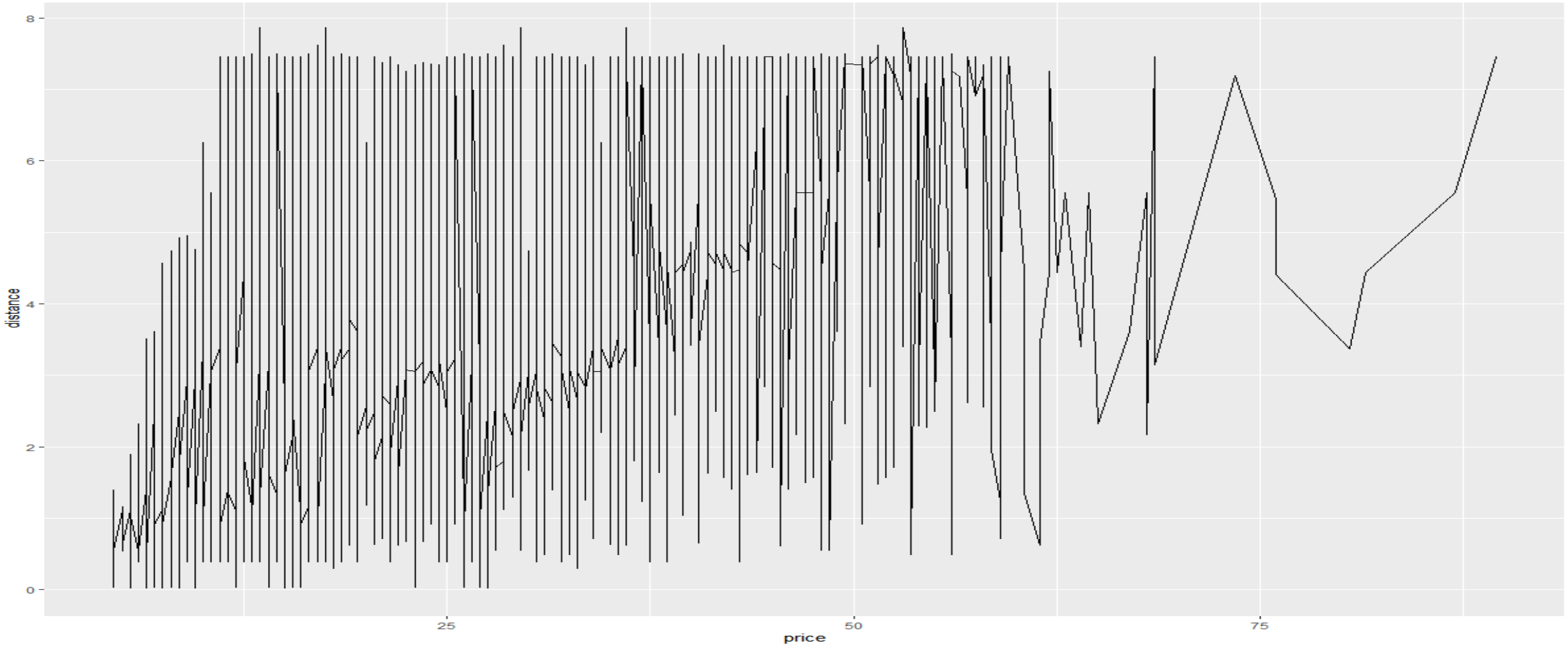
The above graphs show the variation of Car types over the duration of the day. It can be interpreted that highest vehicle usage is during mid hours of the day.



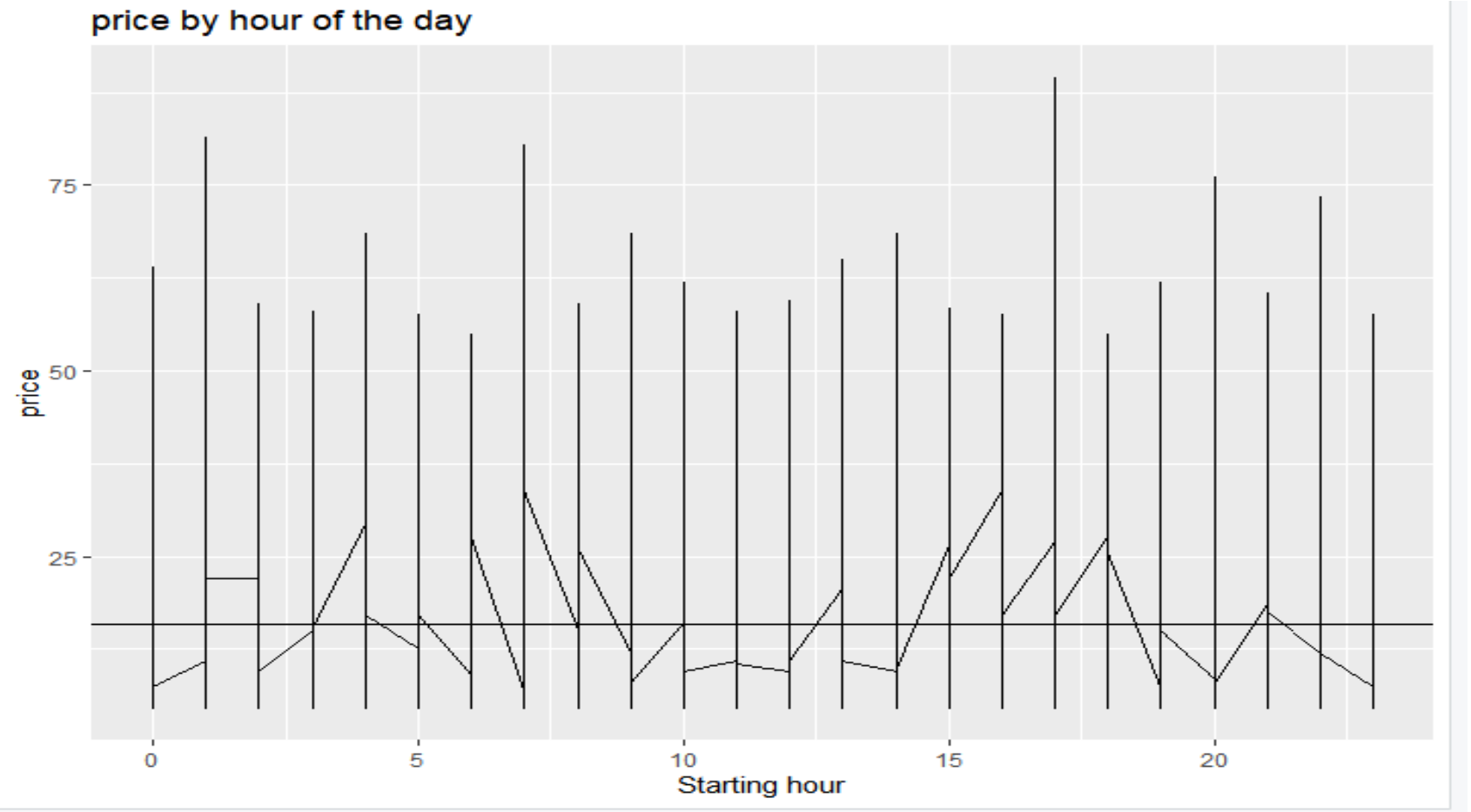
Highest prices are observed in the areas of City centre and the areas that are part of the downtown..


West End
Theatre District
South Station
Northeastern University
North Station
North End
Haymarket Square
Financial District
Fenway
Boston University
Beacon Hill
Back Bay
source
destination
name
Black
Black SUV
UberPool
UberX
UberXL
WAV


West End
Theatre District
South Station
Northeastern University
North Station
North End
Haymarket Square
Financial District
Fenway
Boston University
Beacon Hill
Back Bay
source
destination
price
80
60
40
20



- There is no clear linear relation between price and the distance travelled.
- We can conclude that ride pricing is affected by other variables.
- It is the aim of this project to find those variables.



- We can observe the phenomenon of "rush hour" from this chart.
- And also price fluctuations in every hour of the day
- We can conclude that hour of the day has say in the ride pricing.

Multiple Linear Regression

Why? - Multiple linear regression is used to estimate the relationship between two or more independent variables (hour, distance) and one dependent variable (price).

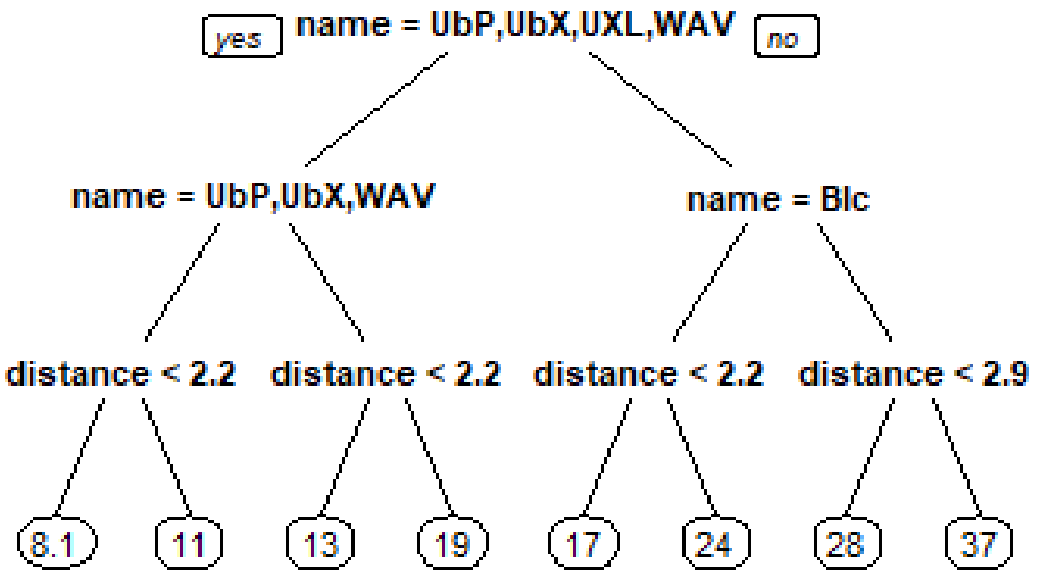
The Estimate column is the estimated effect, also called the regression coefficient or r^2 value. The estimates in the table tell us that for every one percent increase in hour (time) there is an associated 4 percent increase in the price

```
> model <- lm(price ~ hour+name+distance+surge_multiplier+temperature+short_summary+source+destination, data =
 train)
> summary(model)

Call:
lm(formula = price ~ hour + name + distance + surge_multiplier +
    temperature + short_summary + source + destination, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-11.434  -1.388  -0.276   1.025  54.648
```

```
Coefficients: (2 not defined because of singularities)
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                    1.464e+01  4.139e-02  353.733  < 2e-16 ***
hour                           4.791e-04  7.882e-04    0.608   0.5433
nameBlack SUV                  9.761e+00  1.730e-02  564.331  < 2e-16 ***
nameUberPool                  -1.177e+01  1.731e-02 -679.881  < 2e-16 ***
nameUberX                     -1.076e+01  1.733e-02 -620.860  < 2e-16 ***
nameUberXL                    -4.832e+00  1.732e-02 -278.941  < 2e-16 ***
nameWAV                       -1.076e+01  1.732e-02 -620.869  < 2e-16 ***
distance                       2.534e+00  5.387e-03  470.475  < 2e-16 ***
```

Decision Tree Regression

- Decision tree is a type of algorithm in machine learning that uses decisions as the features to represent the result in the form of a tree-like structure.
- Regression trees are used when the dependent variable is continuous.




```
> print(head(results))
         pred real
14 18.090664 16.0
17 27.853622 26.0
26  9.980967  8.5
29  9.947185  8.5
40 17.324210 15.0
63 18.222327 16.5
```

- From the adjoining table, we can observe the difference between predicted price and real prices.

Random Forest Regression

Random forest is a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of overcoming over-fitting problem of individual decision tree.

In other words, random forests are an ensemble learning method for classification and regression that operate by constructing a lot of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.

```
#random forest
require(randomForest)
rf.fit <- randomForest(price ~ hour+name+distance+surge_multiplier+temperature+short_summary+source+dest
rf.pred <- predict(rf.fit,test)

head(rf.pred)
mse <- sum((rf.pred - test$price)^2)/7000

var.y <- sum((test$price - mean(test$price))^2)/6999

rsq <- 1 - mse/var.y

rsq
```

Model Creation

- Various Machine learning algorithms are developed, and accuracy is compared.
- Linear regression, Decision Tree and Random forest algorithms are used for creating model.
- R SQUARE,MSE and RMSE are used as metrics.

The model is trained on 70% data and R square, MSE and RMSE are compared

Algorithms	R SQUARE	MSE	RMSE
Random Forest Regressor	0.94	1.86	1.36
Decision Tree Regressor	0.89	21.7	4.65
Linear Regression	0.92	5.77	2.403

Future Scope

- Analysis for the data is performed and 92% accuracy is achieved.
- The accuracy of the algorithm can be increased by increasing data and by hyper parameter tuning
- Deep learning algorithms can be developed, and the accuracy can be compared.
- Categorical features can be handled in other ways to check the accuracy of the model.

Thank You