**UNIVERSITY OF MUMBAI**

**DEPARTMENT OF STATISTICS**

**VIDYANAGARI MUMBAI – 400098**

****

**CERTIFICATE**

**This is to certify that the following team members as given below, of PGDASS Statistics has successfully completed the project entitled “TRANSPORT-TENSION IN MUMBAI” during the academic year 2018-2019.**

**The Team Comprises of: -**

* **Rakesh Pawar**
* **Aarti Yadav**
* **Dattu Gaidhankar**
* **Shyam Sunder Chaurasiya**
* **Pratibha Kokare**
* **Biraj Paul**
* **Rohan Gaikwad**

**This work is to the best of our knowledge and belief is original.**

**Dr. Santosh Gite**

**(Head of Department & Project Guide)**

**ACKNOWLEDGEMENT**

From the conception of topic to the final presentation, we are extremely thankful to our guide **Dr. Santosh Gite** (*Head of Department of Statistics at University of Mumbai*) for his immense support and guidance throughout the process of makingthis entire project. It was only with the help of his suggestions & corrections, at every step that we were able to complete the project on time & achieve our objectives to our satisfaction.

In addition, we would like to express our gratitude to all the professors and non-teaching staff of the Department of Statistics for their co-operation as well as providing us with requisite amenities over the duration of the project work.

Last but not the least; we would like to give our special thanks to our respondents who diligently and honestly filled our survey purely for academic purposes and taking the time out of their schedule to answer our questions genuinely.

**INTRODUCTION**

**Public transport in Mumbai** involves the transport of millions of its citizens by train, road, and water. Over 88% of the commuters in [Mumbai](https://en.wikipedia.org/wiki/Mumbai) use public transport. Mumbai has the largest organized bus transport network among major Indian cities.

Mumbai's public transport consists primarily of [rapid transit](https://en.wikipedia.org/wiki/Rapid_transit) on exclusive suburban railway lines augmented by [commuter rail](https://en.wikipedia.org/wiki/Commuter_rail) on main lines serving outlying suburbs, the bus services of the three municipalities making up the metropolitan area, public [taxis](https://en.wikipedia.org/wiki/Taxi) and [auto rickshaws](https://en.wikipedia.org/wiki/Auto_rickshaw), as well as ferry services. A [metro](https://en.wikipedia.org/wiki/Rapid_transit) and a monorail system have recently been inaugurated.  A commercial [seaplane](https://en.wikipedia.org/wiki/Seaplane) service was introduced in 2014.[[2]](https://en.wikipedia.org/wiki/Public_transport_in_Mumbai#cite_note-2)

Mumbai is undoubtedly crowded and stuffed with traffic, but even then, a civilized city. Getting in and around is simple if few rules are religiously followed. The excellent public transport network is easily accessible and fairly cheap. Those big red double-decker buses, the black-and-yellow taxis are efficient and convenient to get you to any place; also, the rates are perfectly reasonable. The suburban train system of Mumbai works flawlessly the entire day except 1 am to 4 am. This capital city of Maharashtra has the largest organized bus transport system amongst other Indian states. This disciplined transportation arrangement makes travelling in Mumbai through public transport pretty comfortable. The people of Mumbai even prefer travelling by public transport, because the traffic caps their idea of churning the wheels of their own vehicle.

**The Common modes of transport in Mumbai are as under:**

[**Local Trains**](https://www.mumbai.org.uk/travel-tips/local-trains.html)

Mumbai has two lines that provide services to the suburban traffic from eastern and western suburbs. Local trains are said to be the lifelines of Mumbai, as over half a million make their way to their destinations. These locals run round the clock (except 1 am-4 am) every few minutes and facilitate you with both first- and second-class travel. The two types of trains that run on these lines are fast locals and slow locals. Fast locals stop at selective stations, whereas the slow ones take a halt at every station. The first-class tickets cost almost four times the fare of second-class coaches. The travelling part in second class is a tough cookie during the peak hours; these trains also have separate compartments for ladies. But then locals happen to be the fastest way to travel.

**Buses**

The bus service of Mumbai is supposed to be the best in India. It is run by Bombay Electric Supply and Transport Company (BEST), whose network is so vast that it links literally every nook and corner of the city. The red double-decker buses are comfortable, cheap and safe. The top of these double-decker buses is the best to explore the city in the rush hour. It provides you with a fine view of the actual Mumbai.

But after such best service by the Best Buses, in about a decade, the number of Mumbaiites using the BEST bus system has declined from 4.2 million to under 3 million, and almost all of the transport undertaking’s 500-plus routes have run into losses. Is there a co-relation? Perhaps, yes. Should the shift towards two-wheelers, four wheelers, ola-uber, Auto and Taxi have been much more covenant for traveling.

**Taxis**  
The most relaxing way to travel is the black and yellow taxis, which can be spotted everywhere, down the lane. The cabbies are friendly and helpful. The fares are to be paid by the meter only. A tip for the driver is optional. And if you are out of those suited booted ones, for the business meets and do not want the Mumbai heat to spoil your cool; hire a Cool Cab. These are air conditioned cabs, undoubtedly comfortable and expensive.  
  
**Auto Rickshaws**

If you are in a hurry, you can even have an auto rickshaw ride to your destination, which is more relaxing and covenant. These auto rickshaws can be seen in suburbs of Mumbai. The skilled drivers will zip their way out through the narrow lanes. Comparatively cheaper than taxis but higher then BEST buses, these can also be considered as an option to commute in Mumbai.

The other modes of transportation include Personal Vehicles & Ola-Uber for better convenience.

# CHANGING THE FACE OF THE TRANSPORT SYSTEM IN MUMBAI!!

Who amongst us has not wondered at the sudden exponential rise of two-wheelers on Mumbai’s streets in recent years? Well, official data tells us that there are now nearly two million two-wheelers on the city’s roads, a staggering rise of more than 58% in the last five to six years, which makes it an average of 975 two-wheelers for every kilometre.

This rise has brought a host of issues such as rash and reckless riding, driving in the wrong direction, jumping signals, bike races and other violations. Road accidents involving two-wheelers have increased too.

In about a decade, the number of Mumbaiites using the BEST Bus system has declined from 4.2 million to under 3 million, and almost all of the transport undertaking’s 500-plus routes have run into losses. Is there a co-relation? Perhaps, yes. Should the shift towards other modes of transport have been anticipated and proved to be more convenient

Similarly, the pattern of commuter movement on Mumbai’s famed suburban rail network has been changing over the last decade. Once a strictly north-south linear movement, commuters have been thronging railway stations not known to carry the city’s professional workforce. The older business districts in south Mumbai – Fort, Nariman Point, Ballard Estate – are now less crowded than they used to be at the start of this century.

The crux is that Mumbai’s transport needs and demands have kept changing; its transport systems and options have not. The number of people using BEST Bus has reduced, being the cheapest mode of transport for public, right after Train which also being the best mode of transport and most frequently used, Thereby other modes of transport are chosen over BEST buses, being more cheaper, for the common public who are using the transport services more frequently as their daily mode of transport.

**OBJECTIVES**

Mumbai’s public road transport needs and demands have kept changing; but the improvement has not, which in turn increased the number of other modes of transport which are more preferred over public transport, thereby increasing the level of stress and traffic with higher rates.

The different modes of transport used in Mumbai are Train, Bus, Auto, Taxi, Ola-Uber and Personal Vehicle etc.

Based on the primary research[[1]](#footnote-1), as mentioned in page\_\_\_\_ *“Introduction”*the Prime reason to take up this agenda was to identify the usability of mainly three modes of transports which are highly used by common people i.e Auto, Bus and Train. Identify the reason behind the reduction in usage of BEST Buses**.** BEST Buses being the cheapest mode of transport for public (10-50 Rs on average) after Train.Trains are used along with other modes of road transport as one of the options, being one of the cheaper and frequent mode of transport (5-20Rs on average)usedby the people for long distance travel.

The Statistical analysis of most frequently used public mode of transport Train, bus and Autois identify relationship between the socio-economic factors of population and choosing specific mode of transport. It also aims to understand the association between Train, Bus and Auto, problems faced by common people while travelling by public transport and Identify means to improve it. This survey helps us to understand ground level realities faced everyday by population.

Therefore, our objectives are as follows: -

* Identify and analyze the socio-demographic and geographical factors that affect the population’s choice for a particular mode of transport.
* Statistical analysis of time & Money spent and choice of mode of transport.
* Identify association among Train, Bus, Auto and reason for specific choice of a particular mode of transport.
* Statistical analysis of factors increasing the irritation level& causing inconvenience to the population.
* Identify ways to improve the public transport.

**RESEARCH DESIGN**

To meet aim and objectives of the study, it is important that the researcher select the most appropriate research design. The research design identifies the procedures by which the study population will be selected. The research design refers to the overall strategy that you choose to integrate the different components of the study in a coherent and logical way, thereby, ensuring you will effectively address the research problem; it constitutes the blueprint for the collection, measurement, and analysis of data.

1. **Research Methodology:**

* **Defining the objectives of the project**
* Defining the problem clearly. (Ex. – the Prime reason to take up this agenda was to identify the usability of mainly three modes of transports which are highly used by common people i.e Auto, Bus and Train).
* The scope of the project was chalked out to determine the project plan.
* The plan of action was developed which included the start and the end dates of various steps of the project.
* **Questionnaire Design**
* As per our objectives, a questionnaire was prepared. It was not the final questionnaire as changes would be made (if necessary) after the pilot survey.
* **Pilot Survey**
* We collected primary data by means of an offline survey.
* Pilot survey was done on the questionnaire with a sample size of 50 on the basis of which changes were made in the questionnaire. The questionnaire contains 30 questions which gives factors affecting final conclusion/decision.
* The pilot survey was conducted for the purposes of framing our objectives better and to get an idea about the data we can expect from the full-fledged survey.
* **Preparing final questionnaire**
* A final questionnaire was prepared using all the input from various sources.
* **Collection of data**
* The data was collected using Google Forms.
* The response of people was recorded by open-ended questions; multiple choice questions, grade scale, likert scale and rank scale.
* The filled up information was later obtained to give the required interpretation and findings of most preferred transport modes and factors behind its popularity.
* **Performing data analysis**
* Different statistical tests (Ex. – Multiple Linear Regression, Binary logistic regression, Decision Tree, etc.) were conducted to test the different hypothesis using various software.
* **Arriving at the conclusions**
* From the results obtained after performing the tests, conclusions were given accordingly.

1. **Research approach:**

There are two main approaches to research:

Qualitative approach is concerned with assessment of attitudes, opinions and behavior. We employed qualitative research to fulfill our objectives\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Quantitative approach emphasizes objective measurements and the statistical, mathematical or numerical analysis of data. We employed quantitative research to fulfill our objectives \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Types of research design:**

* **Exploratory research:**
* Exploratory research is the type of research conducted for a problem that has not been clearly defined. Exploratory research helps to determine the best research design, data collection method and selection of subjects. It also helps in framing of our objectives.
* The results of exploratory research are not usually useful for decision making by them, but they can provide significant insight into a given situation.
* We went forward with the exploratory research by conducting a pilot survey to check whether there were any issues in the questionnaire, to reframe our objectives better and to get an idea of the kind of responses we would receive especially, in the open-ended questions.
* **Descriptive research:**
* Descriptive research refers to research that provides an accurate description of characteristics of a particular individual, situation, or group. Descriptive research is also known as statistical research.
* In short, descriptive research deals with something that can be counted and studied, which has an impact in the lives of the people it deals with. After our pilot survey our actual main research was descriptive research based on the objectives we framed after the exploratory research.

1. **Target population:**

The respondents will be the users who are using Transport mediums in their day-to-day life for modes of Road Transport. Our respondents were primarily belonging to an urban setting.

* **Gender:** Male and Female
* **Age:** All age groups

1. **Sample size:**

A large sample size helps to negate the biasness if any that arise due to sampling. Our aim was to collect as many responses as we could within a month’s time from as many sources as we could, the selection was based on convenience sampling due to time and cost limitation/constrain.

1. **Statistical software used:**

* Excel
* R
* SPSS
* Minitab

1. **Visualization software used:**

* Tableau

**QUESTIONNAIRE DESIGN**

1. Where do you live? \*

* Mumbai
* Mumbai Suburban
* Navi Mumbai
* Thane

2. Gender \*

* MALE
* FEMALE
* OTHERS

3. Marital Status \*

* MARRIED
* UNMARRIED

4. Qualification \*

* SECONDARY SCHOOL (SSC)
* HIGHER SECONDARY (HSC)
* GRADUATE
* POST GRADUATE
* Other

5. Employment Status \*

* Student
* Employed
* Unemployed
* Self-Employed

6. Employment sector \*

* Government
* Private
* Unemployed
* Business

7. Annual income (In Lakh) \*

* 0
* 1 to 2.5
* 2.5 to 5
* 5 to 7.5
* Above 7.5

8. Age \*

* < 18
* 18-23
* 24-28
* 29-34
* 35-38
* 38-42
* Above 42

9. What mode of transportation do you use most often? (Multiple choice) \*

* Bike
* Bus
* Train
* Auto
* Taxi
* Walking
* Ola-Uber
* Personal Vehicle

10. How often do you use public transport? \*

* Daily
* Weekly
* Several times a month
* Rarely
* Never

11. What is the main reason why you choose your particular mode of transport? \*

* Time Saving
* Reliability And safety
* Cost
* Comfort

12. Number of vehicles owned in the house? \*

* 0
* 1
* 2
* 3
* 4
* More than 4

13. On an average how much time do you spend while travelling every day ? (In Hrs) \* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

14.On an average how much do you spend while travelling every day ? (In Rs.) \*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

15. How much money do you spend if public transport is not working/delayed that

Day?(In Rs.) \*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

16. How much time do you spend travelling if public transport is not Working/delayed that day? (If you not travel then write zero) \*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

17. Daily Trip length considering both sides ( In km) \*

* <5
* 5-10
* 10-20
* 20-30
* 30-40
* 40-60
* Above 60

18. What is the distance from home to office or college if you are going by walking? (In km) \*

.

* 0-1
* 1-2
* Above 2

19. At what time do you generally travel? \*

* 5-7 AM
* 7-9 AM
* 9-11 AM
* 12 PM and Above

20. What time do travel for coming? \*

* 12-3 PM
* 3-6 PM
* 6-7 PM
* 7-8 PM
* 8-9 PM
* 9-11 PM
* 11 & Above

21. Which factors affect or irritate you most while travelling ?(Please use arrow on right to rank the option 1 to 10 ) \*

1. Traffic
2. Rush / crowd
3. Delay in running
4. Noise Level
5. Anxiety
6. Limited Space
7. Travel Expense
8. Delay In time
9. Long distance
10. People Behavior

22. Which factors are responsible for increasing the road traffic?(Please use the

Arrows on right side to rank the option 1 to 10) \*

1. Hawkers
2. Width of road
3. Illegal parking
4. Current work of metro or mono construction
5. People not following traffic rules or lane discipline
6. Heavy vehicles
7. Rash driving
8. Increasing number of private vehicles
9. Automated traffic signal
10. Road accidents

23. How much would you rate the train service? \*

24. How much would you rate the public bus transport service? \*

25. Do you go for your respective job/college during strike days? \*

* Yes
* No

26. Does travelling hours increases in case of rainy season or any natural calamity? \*

* Yes
* No

27. If yes then how much is the delay in time (in Hr) from the normal traveling Time? \*

* 0-1
* 1-2
* Above 2 Hr.

28.Do you think, due to the upcoming metro and mono project will this effect

Positively on transportation in Mumbai? \*

* Yes
* No

29. Would implementation of GPS system in buses help improve the public Transport? \*

* Yes
* No

**DATA COLLECTION**

Our objectives demanded primary data and hence a questionnaire was prepared with the help of Google Forms. The preliminary step of data collection was to carry out a pilot survey of suitable size. This helped us to formulate our objectives and better understand the structure of data.

The next step was to send the questionnaire to as many people as possible. Becausewhen it comes to data, the higher the number of responses the better. The data was collected through various mediums like e-mails, offline surveys, and lastly, with the help of social networking sites like WhatsApp and Facebook.

We decided to close the survey a month after it went live. We finally closed it once we received 255 responses approximately a month after we started it. We then entered the entire data in Microsoft Excel. Among the 255 responses about 9 responses were discarded due to false and incomplete information, fake/ spam responses. The final data we were left with (246 responses) was then coded for further analysis.

**DATA CLEANING**

Preparing the data for analysis is very essential process as data needs to be cleaned before bringing it into an actionable form where all the final analysis can be conducted on the sample. Typically, in a survey data there are ample of issues which can occur if the data is not cleaned properly as responses can be duplicate, can contain some bias or the respondent might not have taken the survey seriously. The aim of data cleaning is to remove these overt biases and invalid responses to get the good quality of data for performing accurate analysis.

**Steps followed during data cleaning:**

1. **Importing the data into Excel**

Excel is a very efficient tool for data cleaning as it has many functions in it, which makes the process less taxing. Our survey was designed and conducted on Google forms. After reaching the deadline of the survey, we closed the survey and downloaded the data in an Excel file and went forward with data cleaning. We had collected 255 responses by the deadline.

1. **Removal of unnecessary responses**

This was the major issue faced by us during cleaning as our survey had many open-ended questions. So, many non-serious respondents took unnecessary advantage of that and gave erroneous responses. We had to go through the open-ended questions very thoroughly and we came across very absurd responses which were then eliminated. The responses which are not related to the survey needs to be removed for improving the quality of data on which analysis will be done since better the quality of data, the more accurate will be the analysis.

1. **Investigating the outlier cases**

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

Following statistical& graphical methods are used to look at the data patterns and identify outliers

* 1. Leverage plot
  2. Cook’s distance
  3. Box whisker plot

1. **Code the responses**

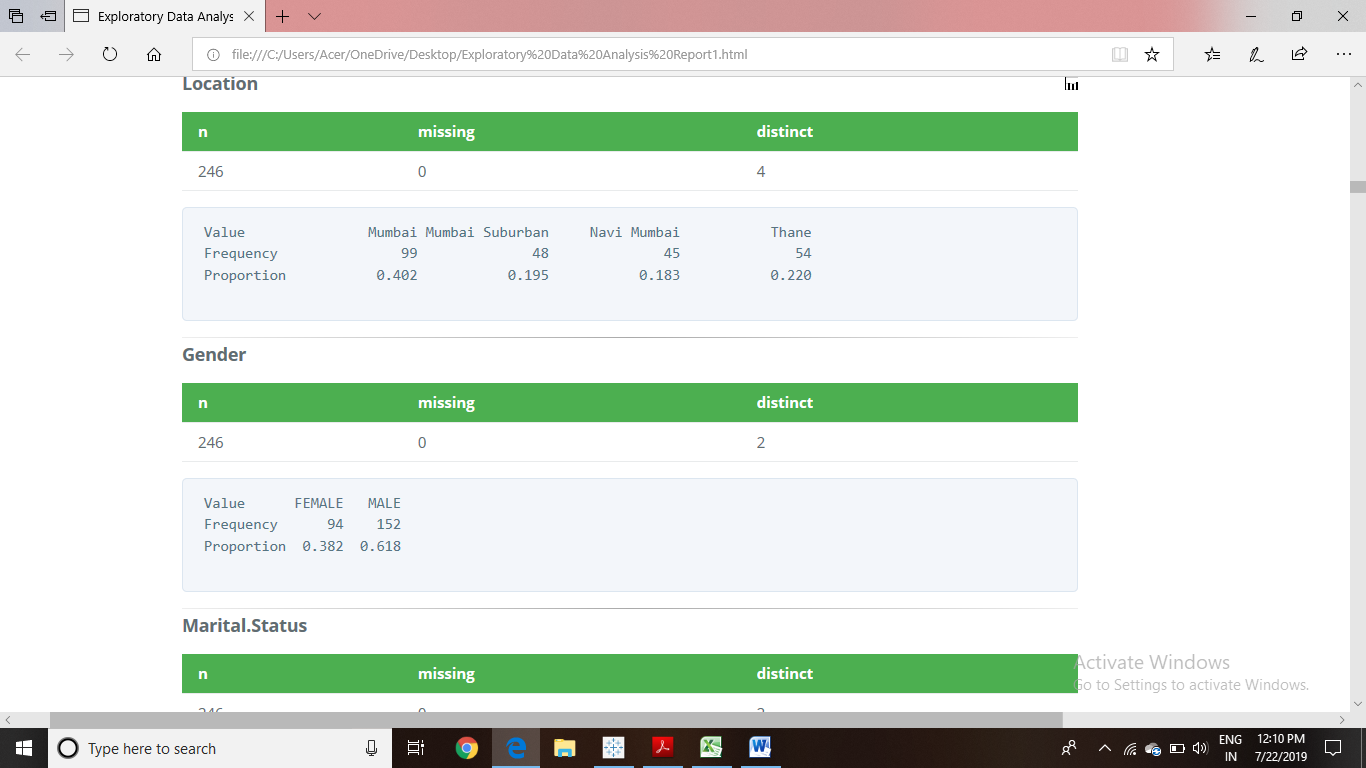
After all the above process was done, we came to the last step, which was making the data compatible with the tools on which we would run our analysis. The first step was to label a variable name into the required format. Now, the second step was to code the open-ended data into categories. Then, thirdly, we coded the data using whole numbers both in open and close ended questions. For example, in question where the respondent had to choose between yes and no we coded “yes” as 1 and “No” as 2. This made our data compatible with tools which we had used for further analysis.

**DESCRIPTIVE STATISTICS**

**DESCRIPTIVE ANALYSIS / GRAPHICAL REPRESENTATION:**

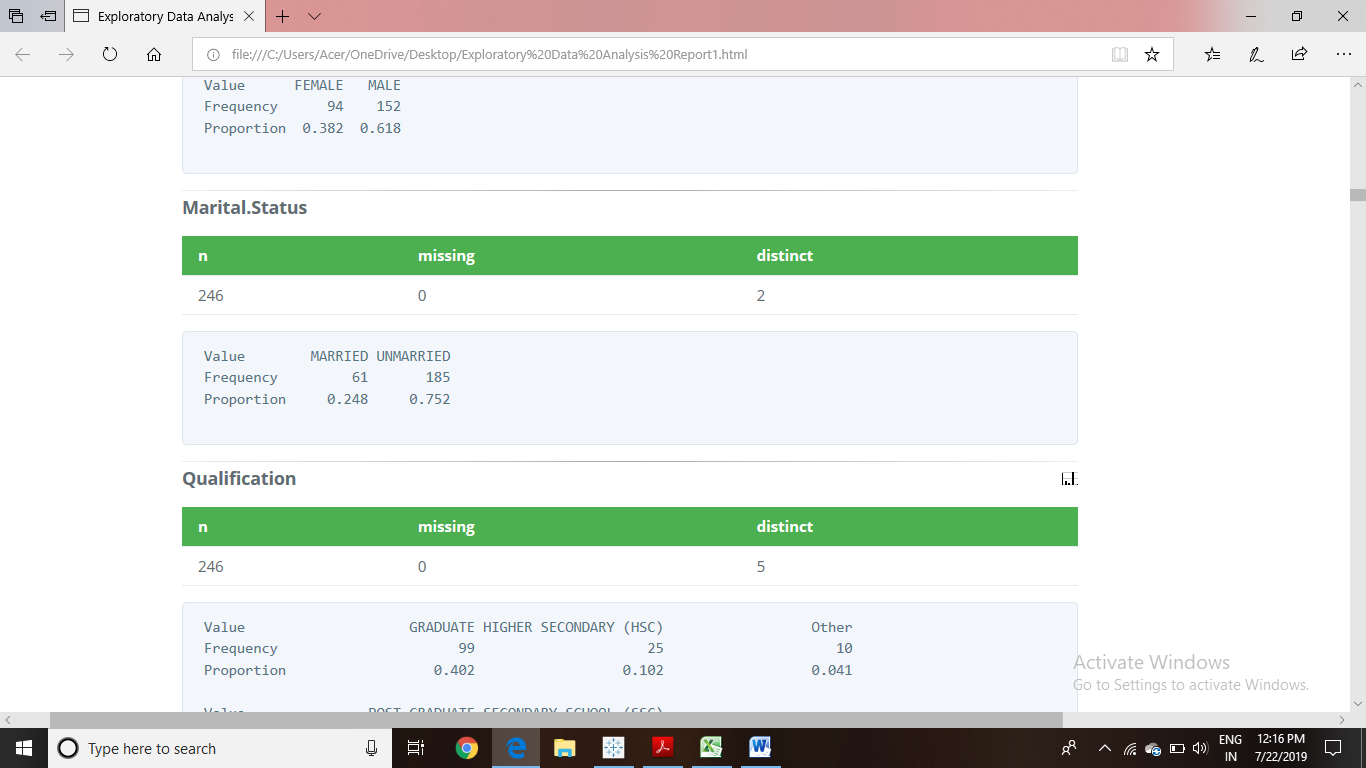
In our dataset contain 246 observation and 68 variables.

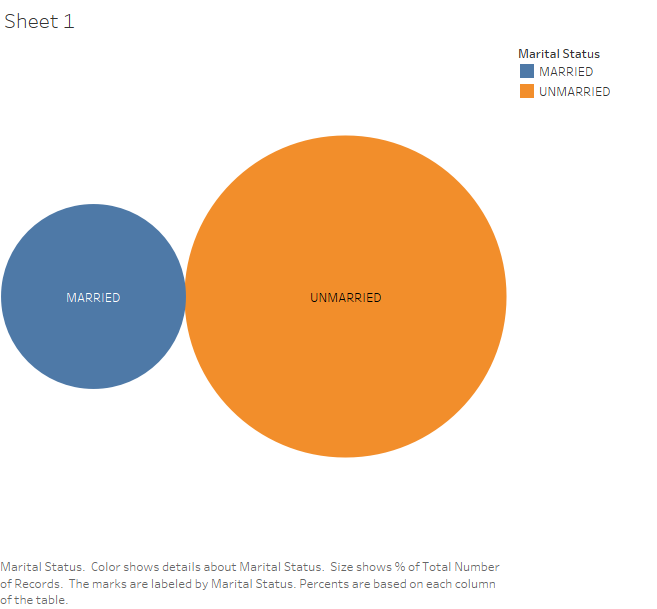
1. **GENDER:**



From the above pie chart there are total 62% males and 38% females, Are daily use the transport system.

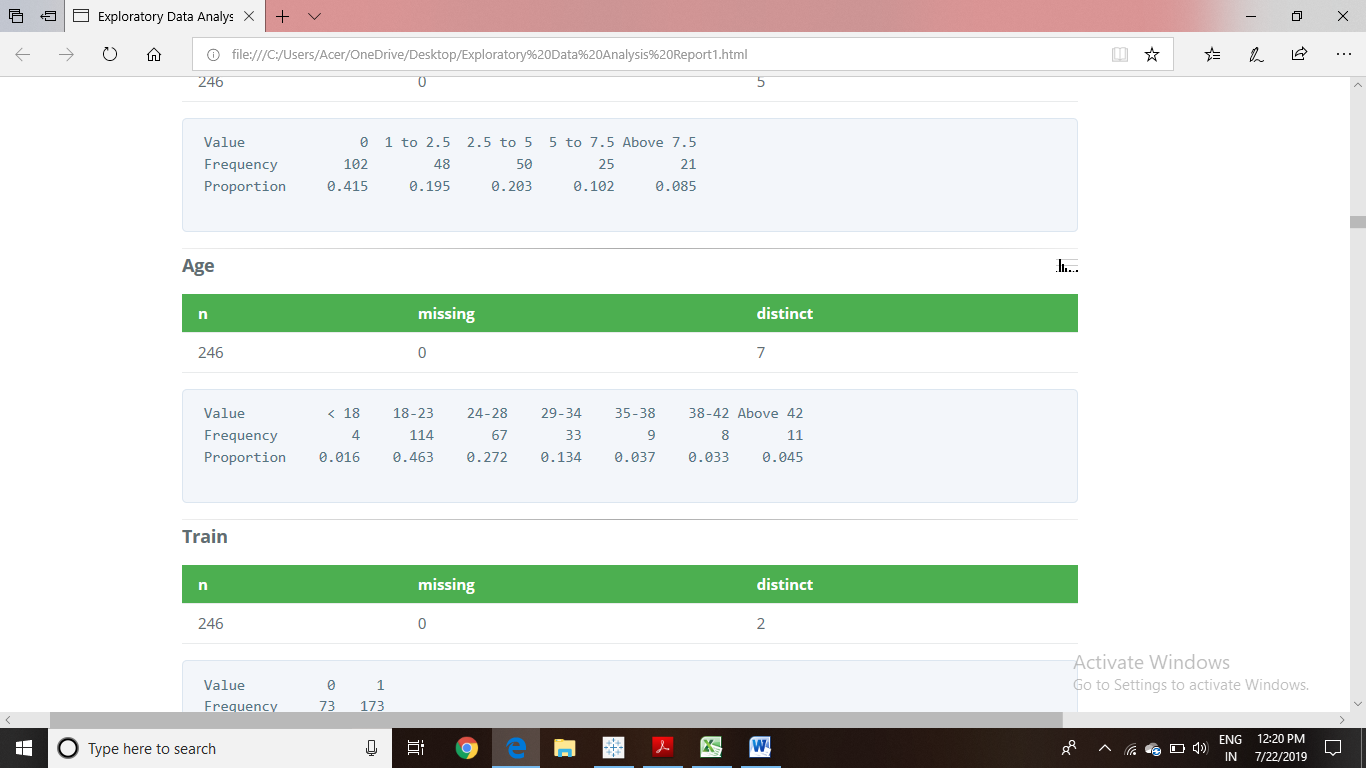
2.**MARITAL STATUS:**





Total 75.2% individual who daily travel are unmarried and only 24.8% are married.

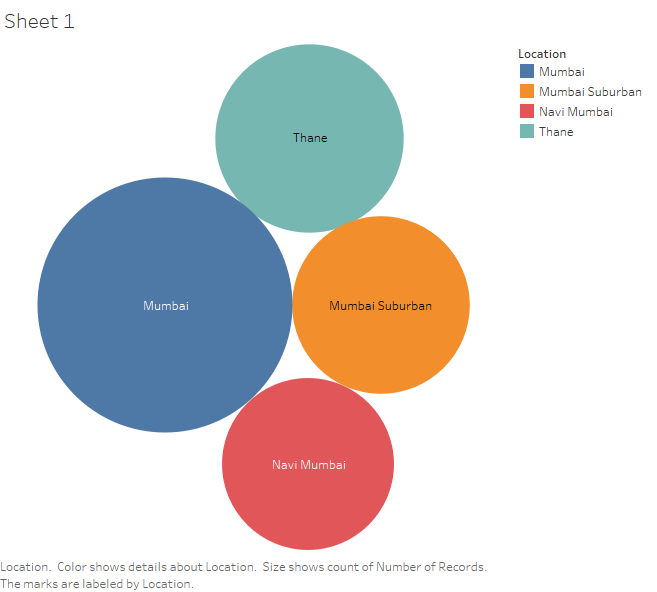
3.**AGE:**



The people who has Age in between 18-23 travel most.

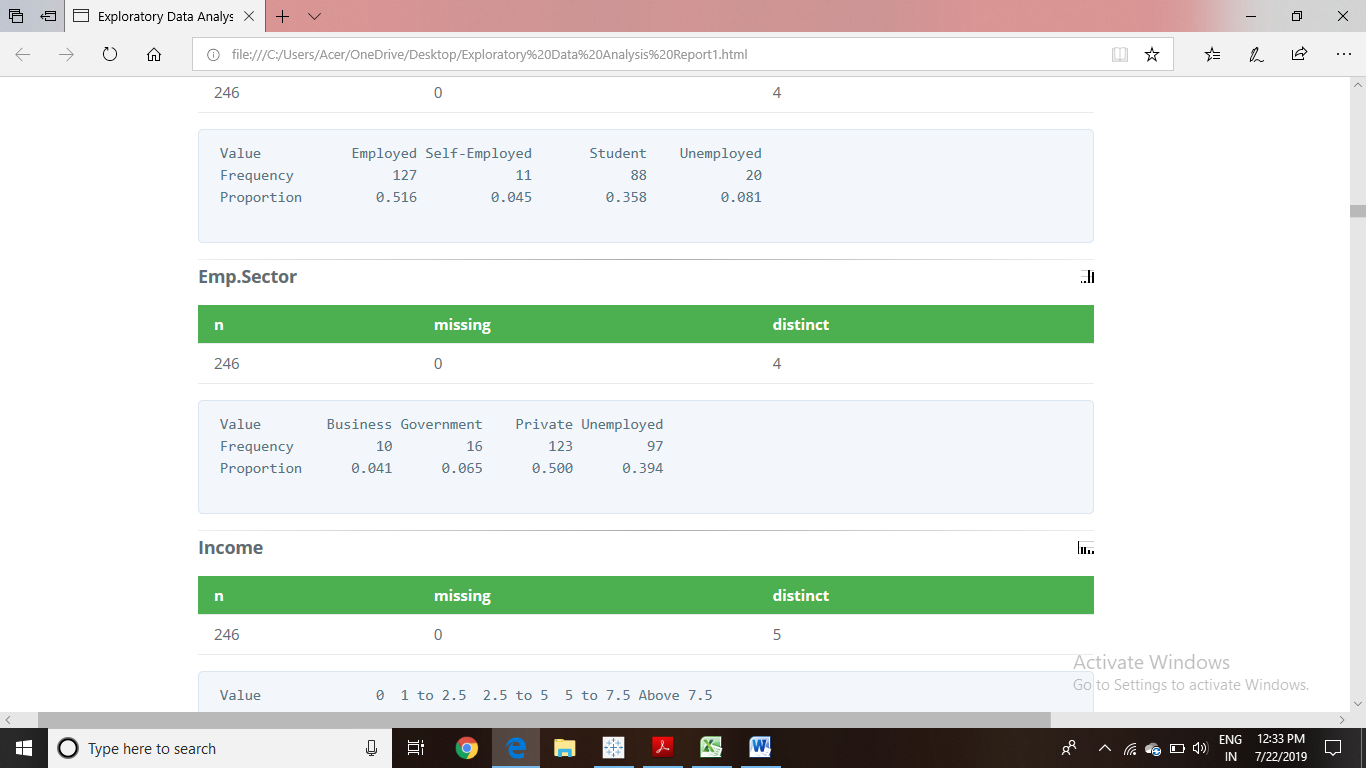
4.**LOCATION :**





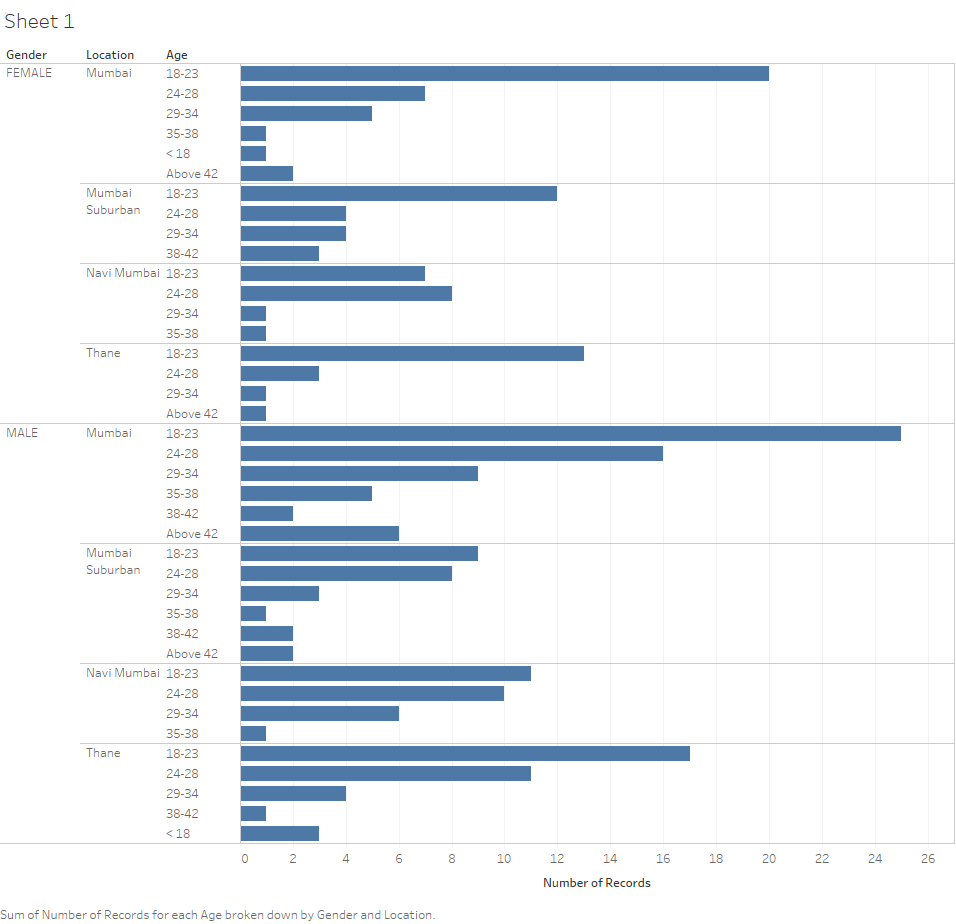
Peoples who are living in mumbai travel more than the people who are living in other region.

5.EMPLOYMENT SECTOR:



Most of the people are doing job in private sector.

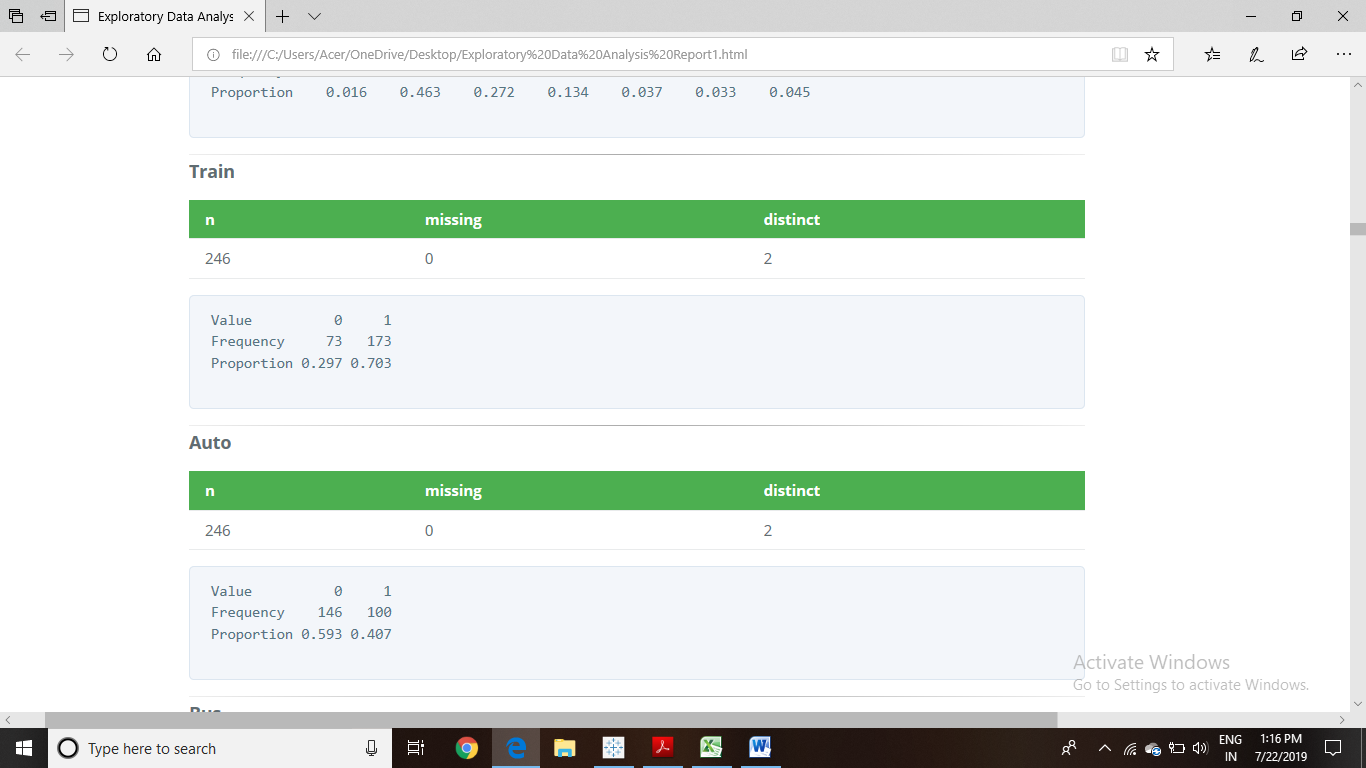
7.Gender with location & age:



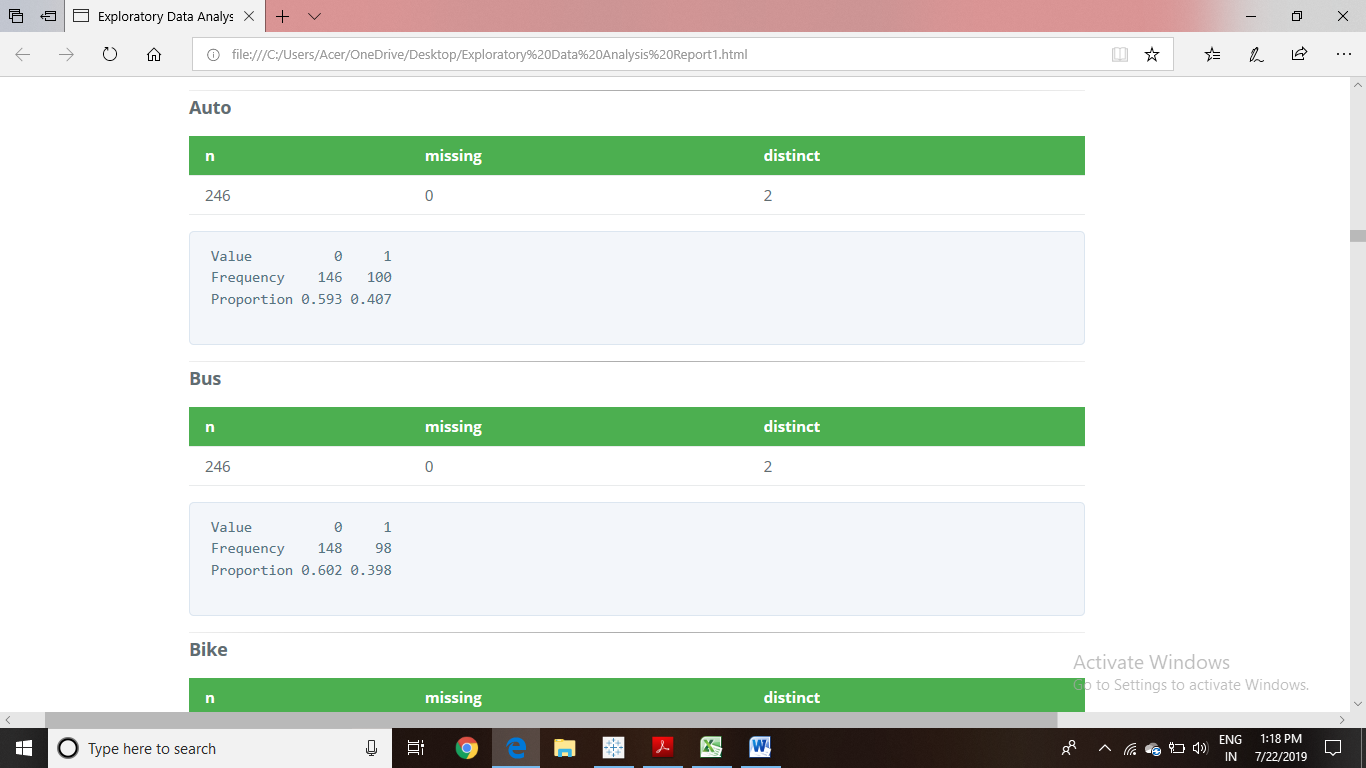
The no of females from age group 18-23 for location mumbai, mumbai-subarban and thane are more and for location navi-mumbai the age group of female from 24-28 are more who travel daily.

The no of males from age group 18-23 for all locations who travels more.

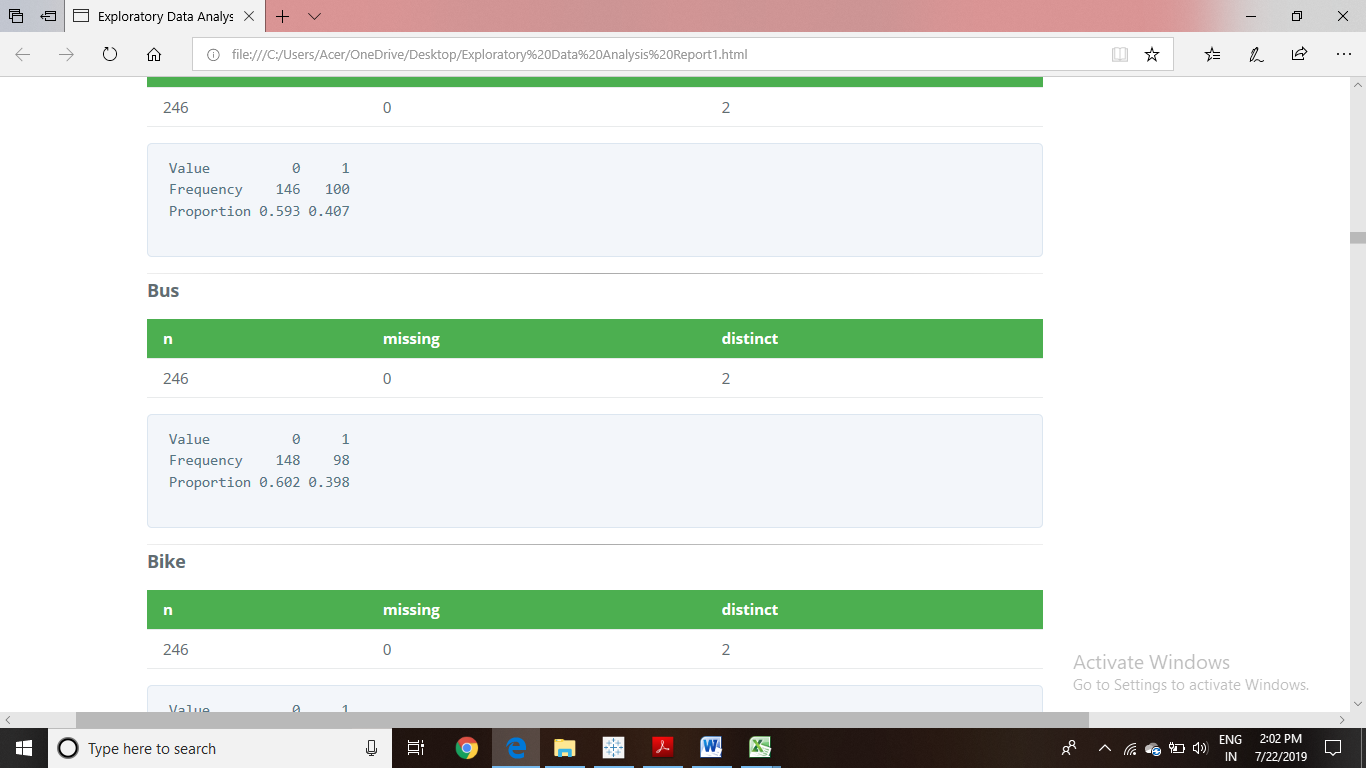
8. Highly used transport system:



From abovechart,There are 70.3% people who prefer train for transportation.

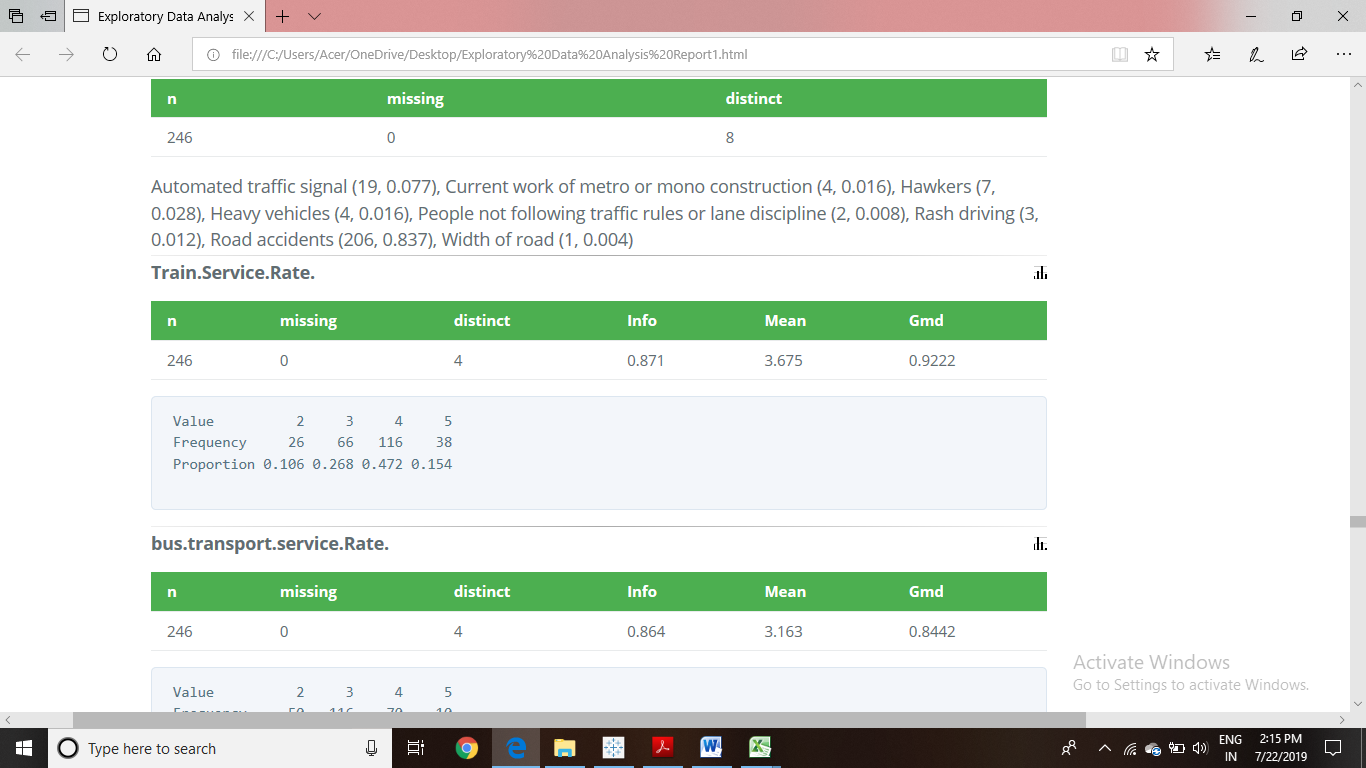


From above diagram 40.7% people uses auto for trnsportation.



From above diagram 39.8% people uses bus for transportation.

9. TRAIN RATING STAR



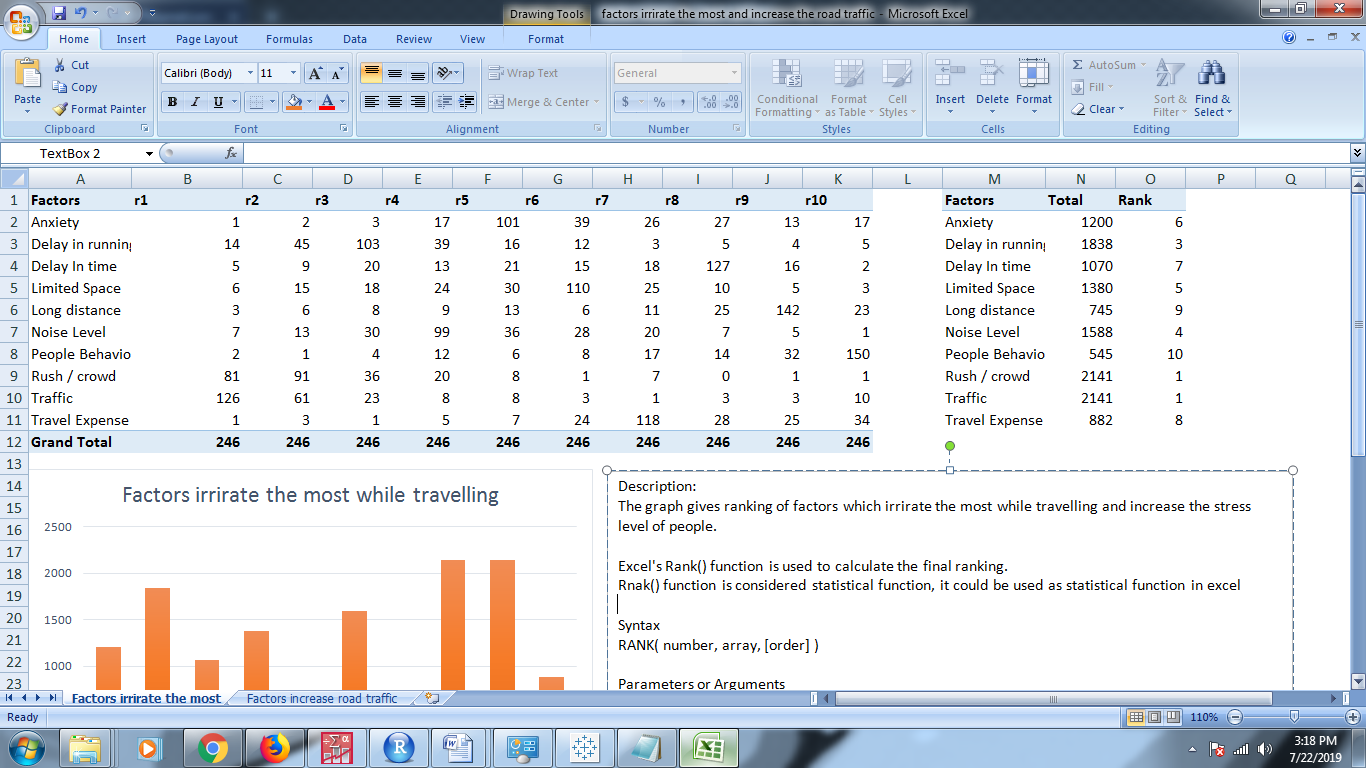
47.2% people agreed with train service on rating scale.

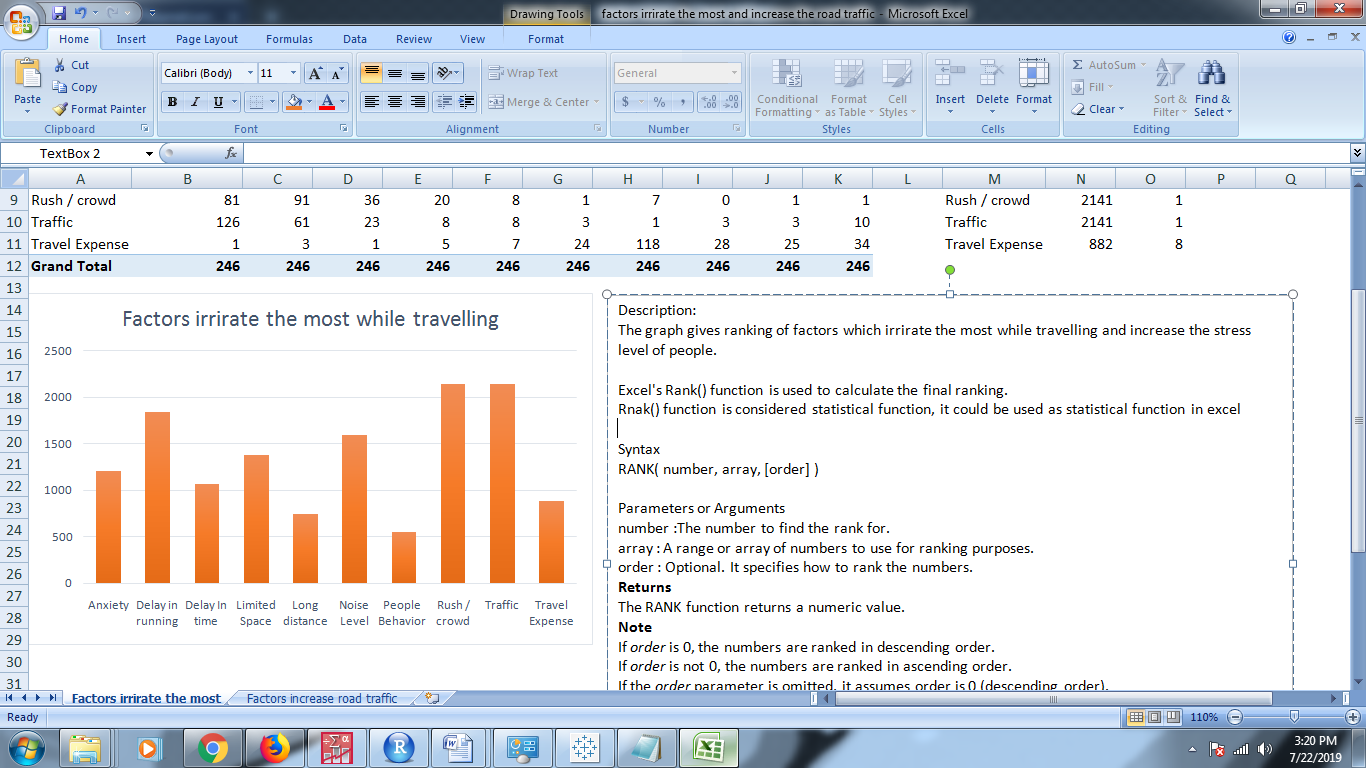
10. BUS RATING STAR



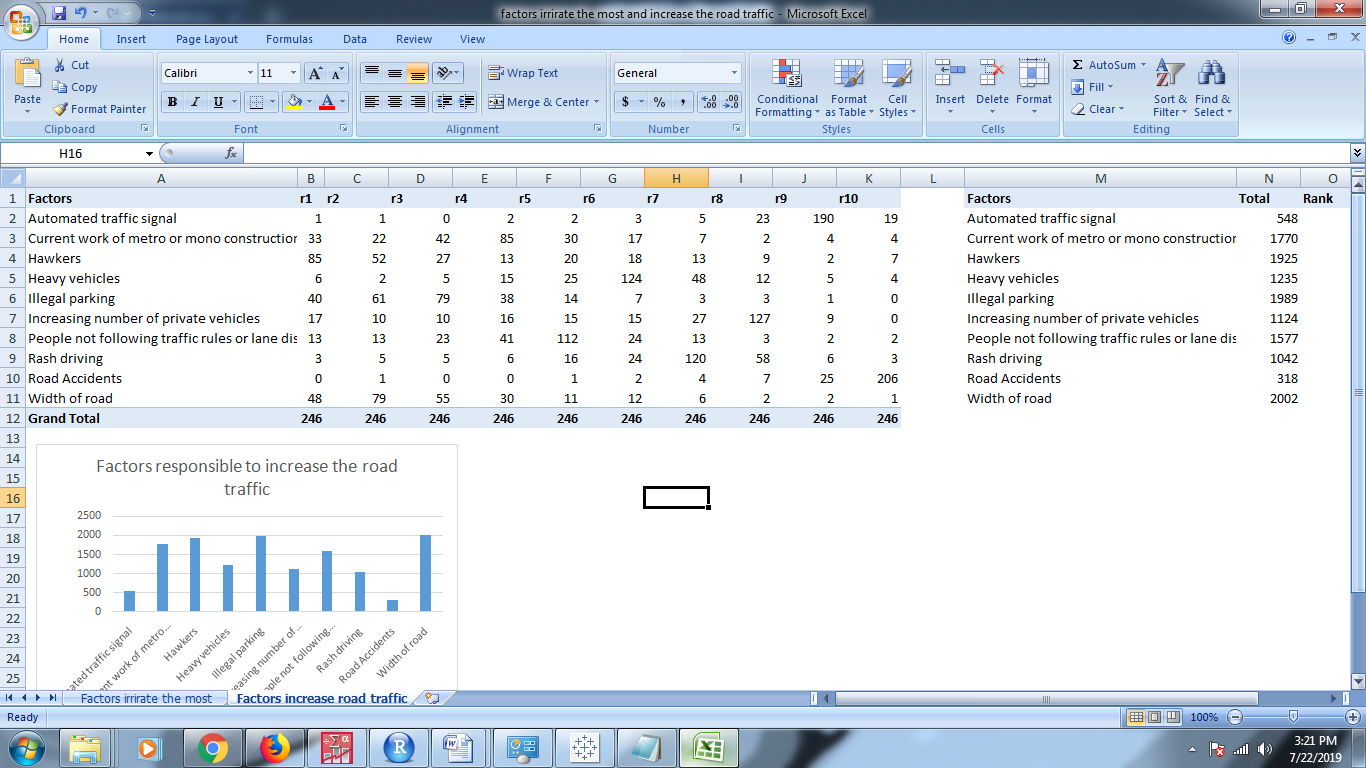
47.2% people feel bus service is a neutral on rating scale.

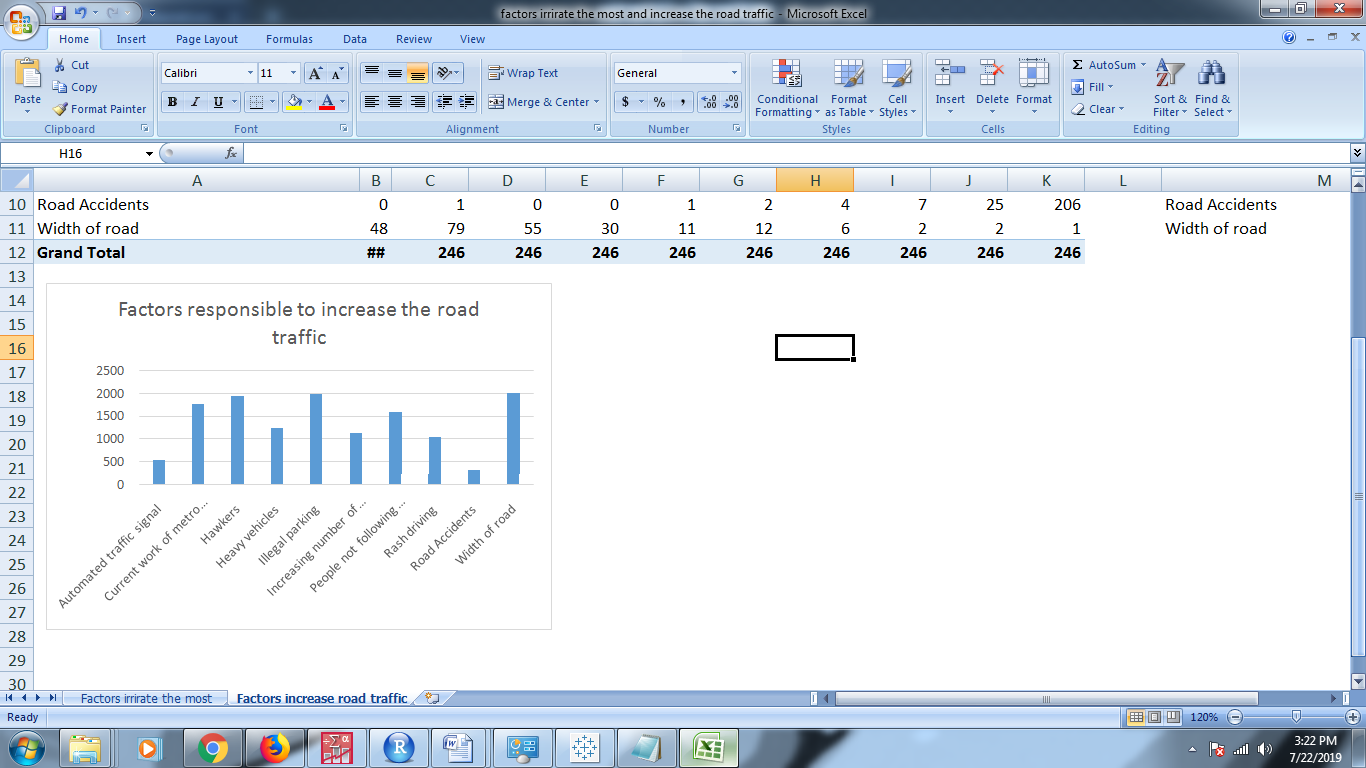
**FACTORS WHICH IRRITATES THE MOST WHILE TRAVELING**

****

****

**FACTORS RESPONSIBLE FOR INCREASE IN ROAD TRAFFIC**

****

****

**OBJECTIVE-1**

* **Identify and analyze the socio-demographic and geographical factors that affect the population’s choice for a particular mode of transport.**

1. To identify and analyze the socio-demographic and geographical factorsthat influences the population for choosing Train as the medium of transport.
2. To identify and analyze the socio-demographic and geographical factors that influences the population for choosing Auto as the medium of transport.
3. To identify and analyze the socio-demographic and geographical factorsthat influences the population for choosing Bus as the medium of transport.

**Note:** we choose to analyze these four mediums, because the frequency/proportion of population who choose the different mode of transport was found to be highest under these four categories.

**OBJECTIVE-2**

* **Statistical analysis on the socio-demographic factors and choice of mode of transport, influencing the Time spend & Money spend by the people for traveling.**

1. Statistical analysis on the socio-demographic factors and choice of mode of transport, influencing the Time spend by the people for traveling.
2. Statistical analysis on the socio-demographic factors and choice of mode of transport, influencing the Money spend by the people for traveling.

**BINARY LOGISTIC REGRESSION**

Logistic regression is part of a category of statistical models called generalized linear models.

Logistic regression is a predictive analysis, like linear regression. Logistic regression allows one to predict a discrete outcome from a set of variables that may be continuous, discrete and dichotomous or a mix of any of these. Here the dependent variable is dichotomous such as presence/ absence.

Binary logistic regression is a form of regression which is used when the dependent is a binary and the independents are of any type. Continuous variables are not used as dependents in logistic regression. Unlike logit regression, there can be only one dependent variable.

The goal of an analysis using logistic regression method is find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables and to determine the percent of variance in the dependent variable explained by the independents; to rank the relative importance of independents; to assess interaction effects; and to understand the impact of covariate control variables.

We used binary logistic regression since our dependent variable is categorical taking 2 values viz. using cashless payment systems and not using cashless payment systems.

The binary Regression model is,

Where,

P = Pr (Y=1 | X) = Conditional probability that the outcome is present

Y: Response Variable

X: Vector of Independent Variables

It models the logit-transformed probability as a linear relationship with the Predictor variables. More formally, let **y** be the binary outcome variable Indicating failure/success with 0/1 and **p** be the probability of **y** to be 1, p=prob(**y**=1). Let x1,….,xk be a set of predictor variables. Then the logistic regression of y on x1,…,xk estimates parameter values for 𝛽1,…, 𝛽k via maximum likelihood method of the following equation.

𝐿𝑜𝑔𝑖t (𝑝) = log = 𝛽o+𝛽1𝑥1+ 𝛽2𝑥2+ . . . +𝛽k𝑥k

P: Probability that Y=1

Y: Dependent Variable

Note that LHS of the model can lie between −∞ 𝑡𝑜 ∞.

We have 28 variables out of which 2 are continuous and 26 are categorical, for model building some are considered as dependent variables and rest being independent variables depending upon the model to be building. The variables are Location, Gender, Marital Status, Qualification, Employment Sector, Income, Age, Daily Usage, Time Saving, Cost, Reliability, Comfort, No. of Vehicles, Daily Trip(KM), Walking Distance(KM), Time-by-go, Time-by-come, Train, Auto, Bus, Bike, Ola-Uber, Personal Vehicle, Taxi, Walking.

**ASSUMPTIONS**

Logistic regression does not make many of the key assumptions of linear regression and general linear models that are based on ordinary least squares algorithms – particularly regarding linearity, normality, homoscedasticity, & Measurement.

Firstly, it does not need a linear relationship between the dependent and independent variables. Logistic regression can handle all sorts of relationships, because it applies a non-linear log transformation to the predicted odds ratio. Also, the independent variables do not need to be multivariate normal –although multivariate normality yields a more stable solution. Also, the error terms (the residuals) do not need to be multivariate normally distributed. In addition to that, homoscedasticity is not needed. Logistic regression does not need variances to be heteroscedastic for each level of the independent variables. Lastly, it can handle ordinal and nominal data as independent variables. The independent variables do not need to be metric (interval or ratio scaled). However, some other assumptions still apply.

Secondly, since logistic regression assumes that P(Y=1) is the probability of the event occurring, it is necessary that the dependent variable is coded accordingly. That is, for a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.

Thirdly, the model should be fitted correctly. Neither over fitting nor under fitting should occur. That is only the meaningful variables should be included. A good approach to ensure this is to use a stepwise method to estimate the logistic regression.

Fourthly, the error terms need to be independent. Logistic regression requires each observation to be independent. That is that the data-points should not be from any dependent samples design, e.g., before-after measurements, or matched pairings.

Also the model should have little or no multicollinearity. That is that the independent variables should be independent from each other. However, there is the option to include interaction effects of categorical variables in the analysis and the model. If multicollinearity is present, centering the variables might resolve the issue, i.e. deducting the mean of each variable. If this does not lower the multicollinearity, a factor analysis with orthogonally rotated factors should be done before the logistic regression is estimated.

Fifthly, logistic regression assumes linearity of independent variables and log odds. Whilst it does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds. Otherwise the test underestimates the strength of the relationship and rejects the relationship too easily, that is being not significant (not rejecting the null hypothesis) where it should be significant. A solution to this problem is the categorization of the independent variables. That is transforming metric variables to ordinal level and then including them in the model. Another approach would be to use discriminant analysis, if the assumptions of homoscedasticity, multivariate normality, and absence of multicollinearity are met.

Lastly, it requires quite large sample sizes. Because maximum likelihood estimates are less powerful than ordinary least squares (e.g., simple linear regression, multiple linear regression); whilst OLS needs 5 cases per independent variable in the analysis, ML needs at least 10 cases per independent variable, some statisticians recommend at least 30 cases for each parameter to be estimated.

**TREE BASED LEARNING**

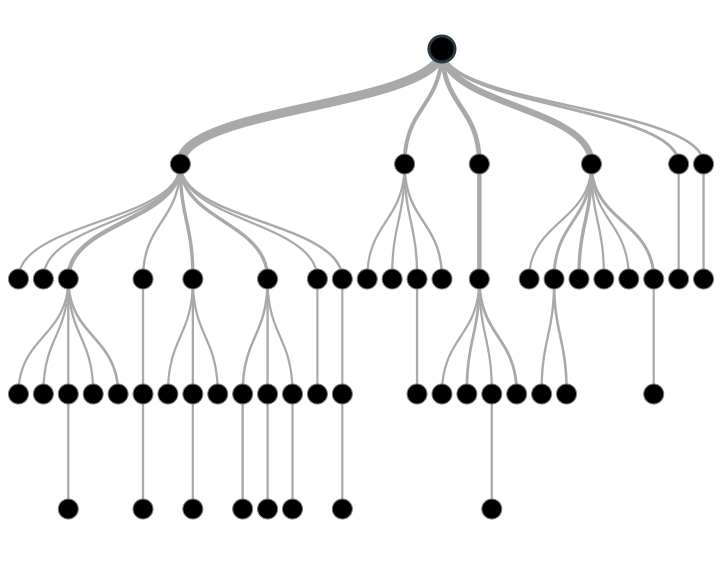
Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression).

Methods like decision trees, random forest, gradient boosting are being popularly used in all kinds of data science problems. Hence, for every analyst (fresher also), it’s important to learn these algorithms and use them for modeling.

This tutorial is meant to help beginners learn tree based modeling from scratch. After the successful completion of this tutorial, one is expected to become proficient at using tree based algorithms and build predictive models.

## What is a Decision Tree ? How does it work ?

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.



Tree based modeling in R and Python

### Types of Decision Trees

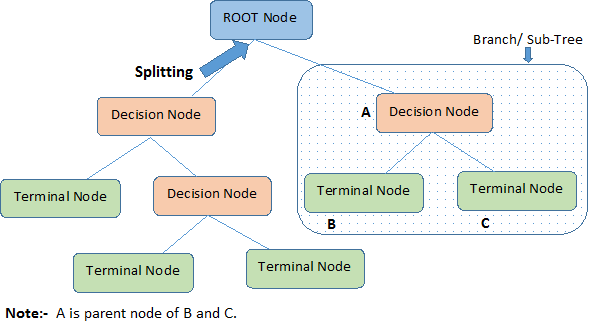
Types of decision tree is based on the type of target variable we have. It can be of two types:

1. **Categorical Variable Decision Tree:**Decision Tree which has categorical target variable then it called as categorical variable decision tree. Example:- In above scenario of student problem, where the target variable was “Student will play cricket or not” i.e. YES or NO.
2. **Continuous Variable Decision Tree:**Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree.

### Important Terminology related to Decision Trees

Let’s look at the basic terminology used with Decision trees:

1. **Root Node:**It represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:**It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:**When a sub-node splits into further sub-nodes, then it is called decision node.
4. **Leaf/ Terminal Node:**Nodes do not split is called Leaf or Terminal node.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/01/Decision_Tree_2.png)

1. **Pruning:**When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
2. **Branch / Sub-Tree:**A sub section of entire tree is called branch or sub-tree.
3. **Parent and Child Node:**A node, which is divided into sub-nodes is called parent node of sub-nodes where as sub-nodes are the child of parent node.

These are the terms commonly used for decision trees. As we know that every algorithm has advantages and disadvantages, below are the important factors which one should know.

### Advantages

1. **Easy to Understand**: Decision tree output is very easy to understand even for people from non-analytical background. It does not require any statistical knowledge to read and interpret them. Its graphical representation is very intuitive and users can easily relate their hypothesis.
2. **Useful in Data exploration:**Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables. With the help of decision trees, we can create new variables / features that has better power to predict target variable. You can refer article ([Trick to enhance power of regression model](https://www.analyticsvidhya.com/blog/2013/10/trick-enhance-power-regression-model-2/)) for one such trick.  It can also be used in data exploration stage. For example, we are working on a problem where we have information available in hundreds of variables, there decision tree will help to identify most significant variable.
3. **Less data cleaning required:**It requires less data cleaning compared to some other modeling techniques. It is not influenced by outliers and missing values to a fair degree.
4. **Data type is not a constraint:**It can handle both numerical and categorical variables.
5. **Non Parametric Method:**Decision tree is considered to be a non-parametric method. This means that decision trees have no assumptions about the space distribution and the classifier structure.

### Disadvantages

1. **Over fitting:** Over fitting is one of the most practical difficulty for decision tree models. This problem gets solved by setting constraints on model parameters and pruning (discussed in detailed below).
2. **Not fit for continuous variables**: While working with continuous numerical variables, decision tree looses information when it categorizes variables in different categories.

## 2. Regression Trees vs Classification Trees

We all know that the terminal nodes (or leaves) lies at the bottom of the decision tree. This means that decision trees are typically drawn upside down such that leaves are the the bottom & roots are the tops (shown below).



Both the trees work almost similar to each other, let’s look at the primary differences & similarity between classification and regression trees:

1. Regression trees are used when dependent variable is continuous. Classification trees are used when dependent variable is categorical.
2. In case of regression tree, the value obtained by terminal nodes in the training data is the mean response of observation falling in that region. Thus, if an unseen data observation falls in that region, we’ll make its prediction with mean value.
3. In case of classification tree, the value (class) obtained by terminal node in the training data is the mode of observations falling in that region. Thus, if an unseen data observation falls in that region, we’ll make its prediction with mode value.
4. Both the trees divide the predictor space (independent variables) into distinct and non-overlapping regions. For the sake of simplicity, you can think of these regions as high dimensional boxes or boxes.
5. Both the trees follow a top-down greedy approach known as recursive binary splitting. We call it as ‘top-down’ because it begins from the top of tree when all the observations are available in a single region and successively splits the predictor space into two new branches down the tree. It is known as ‘greedy’ because, the algorithm cares (looks for best variable available) about only the current split, and not about future splits which will lead to a better tree.
6. This splitting process is continued until a user defined stopping criteria is reached. For example: we can tell the the algorithm to stop once the number of observations per node becomes less than 50.
7. In both the cases, the splitting process results in fully grown trees until the stopping criteria is reached. But, the fully grown tree is likely to overfit data, leading to poor accuracy on unseen data. This bring ‘pruning’. Pruning is one of the technique used tackle overfitting. We’ll learn more about it in following section.

## 3. How does a tree decide where to split?

The decision of making strategic splits heavily affects a tree’s accuracy. The decision criteria is different for classification and regression trees.

Decision trees use multiple algorithms to decide to split a node in two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable. Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

The algorithm selection is also based on type of target variables. Let’s look at the four most commonly used algorithms in decision tree:

### Gini

Gini  says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.

1. It works with categorical target variable “Success” or “Failure”.
2. It performs only Binary splits
3. Higher the value of Gini higher the homogeneity.
4. CART (Classification and Regression Tree) uses Gini method to create binary splits.

**Steps to Calculate Gini for a split**

1. Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
2. Calculate Gini for split using weighted Gini score of each node of that split.

### Chi-Square

It is an algorithm to find out the statistical significance between the differences between sub-nodes and parent node. We measure it by sum of squares of standardized differences between observed and expected frequencies of target variable.

1. It works with categorical target variable “Success” or “Failure”.
2. It can perform two or more splits.
3. Higher the value of Chi-Square higher the statistical significance of differences between sub-node and Parent node.
4. Chi-Square of each node is calculated using formula,
5. Chi-square = ((Actual – Expected)^2 / Expected)^1/2
6. It generates tree called CHAID (Chi-square Automatic Interaction Detector)

**Steps to Calculate Chi-square for a split:**

1. Calculate Chi-square for individual node by calculating the deviation for Success and Failure both
2. Calculated Chi-square of Split using Sum of all Chi-square of success and Failure of each node of the split

### Information Gain:

Look at the image below and think which node can be described easily. I am sure, your answer is C because it requires less information as all values are similar. On the other hand, B requires more information to describe it and A requires the maximum information. In other words, we can say that C is a Pure node, B is less Impure and A is more impure.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/01/Information_Gain_Decision_Tree2.png)

Now, we can build a conclusion that less impure node requires less information to describe it. And, more impure node requires more information. Information theory is a measure to define this degree of disorganization in a system known as Entropy. If the sample is completely homogeneous, then the entropy is zero and if the sample is an equally divided (50% – 50%), it has entropy of one.

Entropy can be calculated using formula:-Entropy, Decision Tree

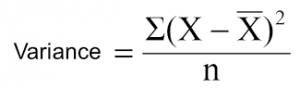
Here p and q is probability of success and failure respectively in that node. Entropy is also used with categorical target variable. It chooses the split which has lowest entropy compared to parent node and other splits. The lesser the entropy, the better it is.

**Steps to calculate entropy for a split:**

1. Calculate entropy of parent node
2. Calculate entropy of each individual node of split and calculate weighted average of all sub-nodes available in split.

### Reduction in Variance

Till now, we have discussed the algorithms for categorical target variable. Reduction in variance is an algorithm used for continuous target variables (regression problems). This algorithm uses the standard formula of variance to choose the best split. The split with lower variance is selected as the criteria to split the population:

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/01/Varince.png)

Above X-bar is mean of the values, X is actual and n is number of values.

**Steps to calculate Variance:**

1. Calculate variance for each node.
2. Calculate variance for each split as weighted average of each node variance.

## What are the key parameters of tree modeling and how can we avoid over-fitting in decision trees?

Overfitting is one of the key challenges faced while modeling decision trees. If there is no limit set of a decision tree, it will give you 100% accuracy on training set because in the worse case it will end up making 1 leaf for each observation. Thus, preventing overfitting is pivotal while modeling a decision tree and it can be done in 2 ways:

1. Setting constraints on tree size
2. Tree pruning

Lets discuss both of these briefly.

### Setting Constraints on Tree Size

This can be done by using various parameters which are used to define a tree. First, lets look at the general structure of a decision tree:

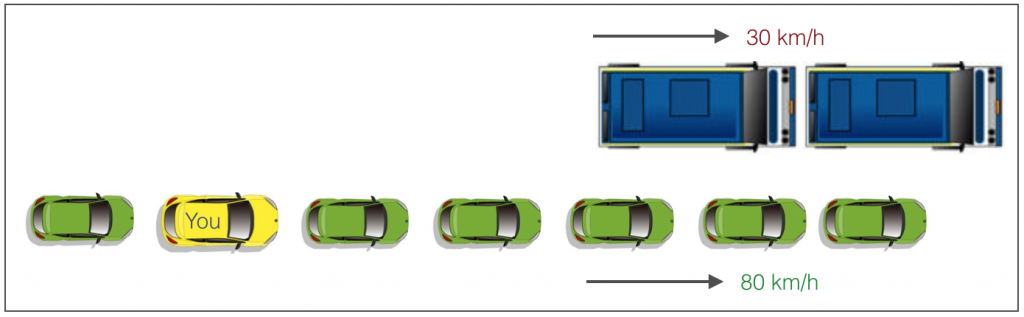
[](https://www.analyticsvidhya.com/wp-content/uploads/2016/02/tree-infographic.png)

The parameters used for defining a tree are further explained below. The parameters described below are irrespective of tool. It is important to understand the role of parameters used in tree modeling. These parameters are available in R & Python.

1. **Minimum samples for a node split**
   * Defines the minimum number of samples (or observations) which are required in a node to be considered for splitting.
   * Used to control over-fitting. Higher values prevent a model from learning relations which might be highly specific to the particular sample selected for a tree.
   * Too high values can lead to under-fitting hence, it should be tuned using CV.
2. **Minimum samples for a terminal node (leaf)**
   * Defines the minimum samples (or observations) required in a terminal node or leaf.
   * Used to control over-fitting similar to min\_samples\_split.
   * Generally lower values should be chosen for imbalanced class problems because the regions in which the minority class will be in majority will be very small.
3. **Maximum depth of tree (vertical depth)**
   * The maximum depth of a tree.
   * Used to control over-fitting as higher depth will allow model to learn relations very specific to a particular sample.
   * Should be tuned using CV.
4. **Maximum number of terminal nodes**
   * The maximum number of terminal nodes or leaves in a tree.
   * Can be defined in place of max\_depth. Since binary trees are created, a depth of ‘n’ would produce a maximum of 2^n leaves.
5. **Maximum features to consider for split**
   * The number of features to consider while searching for a best split. These will be randomly selected.
   * As a thumb-rule, square root of the total number of features works great but we should check upto 30-40% of the total number of features.
   * Higher values can lead to over-fitting but depends on case to case.

### Tree Pruning

As discussed earlier, the technique of setting constraint is a greedy-approach. In other words, it will check for the best split instantaneously and move forward until one of the specified stopping condition is reached. Let’s consider the following case when you’re driving:

[](https://www.analyticsvidhya.com/wp-content/uploads/2016/04/graphic.png)

There are 2 lanes:

1. A lane with cars moving at 80km/h
2. A lane with trucks moving at 30km/h

At this instant, you are the yellow car and you have 2 choices:

1. Take a left and overtake the other 2 cars quickly
2. Keep moving in the present lane

Lets analyze these choice. In the former choice, you’ll immediately overtake the car ahead and reach behind the truck and start moving at 30 km/h, looking for an opportunity to move back right. All cars originally behind you move ahead in the meanwhile. This would be the optimum choice if your objective is to maximize the distance covered in next say 10 seconds. In the later choice, you sale through at same speed, cross trucks and then overtake maybe depending on situation ahead. Greedy you!

This is exactly the difference between normal decision tree & pruning. A decision tree with constraints won’t see the truck ahead and adopt a greedy approach by taking a left. On the other hand if we use pruning, we in effect look at a few steps ahead and make a choice.

So we know pruning is better. But how to implement it in decision tree? The idea is simple.

1. We first make the decision tree to a large depth.
2. Then we start at the bottom and start removing leaves which are giving us negative returns when compared from the top.
3. Suppose a split is giving us a gain of say -10 (loss of 10) and then the next split on that gives us a gain of 20. A simple decision tree will stop at step 1 but in pruning, we will see that the overall gain is +10 and keep both leaves.

Note that sklearn’s decision tree classifier does not currently support pruning. Advanced packages like xgboost have adopted tree pruning in their implementation. But the library rpart in R, provides a function to prune. Good for R users!

## 

## Are tree based models better than linear models?

“If I can use logistic regression for classification problems and linear regression for regression problems, why is there a need to use trees”? Many of us have this question. And, this is a valid one too.

Actually, you can use any algorithm. It is dependent on the type of problem you are solving. Let’s look at some key factors which will help you to decide which algorithm to use:

1. If the relationship between dependent & independent variable is well approximated by a linear model, linear regression will outperform tree based model.
2. If there is a high non-linearity & complex relationship between dependent & independent variables, a tree model will outperform a classical regression method.
3. If you need to build a model which is easy to explain to people, a decision tree model will always do better than a linear model. Decision tree models are even simpler to interpret than linear regression.

## What are ensemble methods in tree based modeling ?

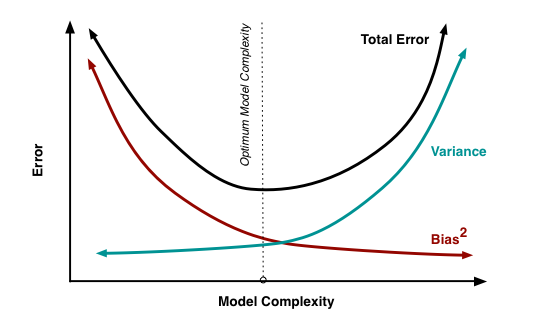
The literary meaning of word ‘ensemble’ is group. Ensemble methods involve group of predictive models to achieve a better accuracy and model stability. Ensemble methods are known to impart supreme boost to tree based models.

Like every other model, a tree based model also suffers from the plague of bias and variance. Bias means, ‘how much on an average are the predicted values different from the actual value.’ Variance means, ‘how different will the predictions of the model be at the same point if different samples are taken from the same population’.

You build a small tree and you will get a model with low variance and high bias. How do you manage to balance the trade off between bias and variance?

Normally, as you increase the complexity of your model, you will see a reduction in prediction error due to lower bias in the model. As you continue to make your model more complex, you end up over-fitting your model and your model will start suffering from high variance.

A champion model should maintain a balance between these two types of errors. This is known as the **trade-off management** of bias-variance errors. Ensemble learning is one way to execute this trade off analysis.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/07/model_complexity.png)

Some of the commonly used ensemble methods include: Bagging, Boosting and Random Forest. In this tutorial, we’ll focus on Bagging, Random Forest and Boosting in detail.

## What is Bagging? How does it work?

Bagging is a technique used to reduce the variance of our predictions by combining the result of multiple classifiers modeled on different sub-samples of the same data set. The following figure will make it clearer:



The steps followed in bagging are:

1. **Create Multiple DataSets**:
   * Sampling is done with replacement on the original data and new datasets are formed.
   * The new data sets can have a fraction of the columns as well as rows, which are generally hyper-parameters in a bagging model
   * Taking row and column fractions less than 1 helps in making robust models, less prone to overfitting
2. **Build Multiple Classifiers:**
   * Classifiers are built on each data set.
   * Generally the same classifier is modeled on each data set and predictions are made.
3. **Combine Classifiers:**
   * The predictions of all the classifiers are combined using a mean, median or mode value depending on the problem at hand.
   * The combined values are generally more robust than a single model.

Note that, here the number of models built is not a hyper-parameters. Higher number of models are always better or may give similar performance than lower numbers. It can be theoretically shown that the variance of the combined predictions are reduced to 1/n (n: number of classifiers) of the original variance, under some assumptions.

There are various implementations of bagging models. Random forest is one of them and we’ll discuss it next.

## What is Random Forest ? How does it work?

Random Forest is considered to be a panacea of all data science problems. On a funny note, when you can’t think of any algorithm (irrespective of situation), use random forest!

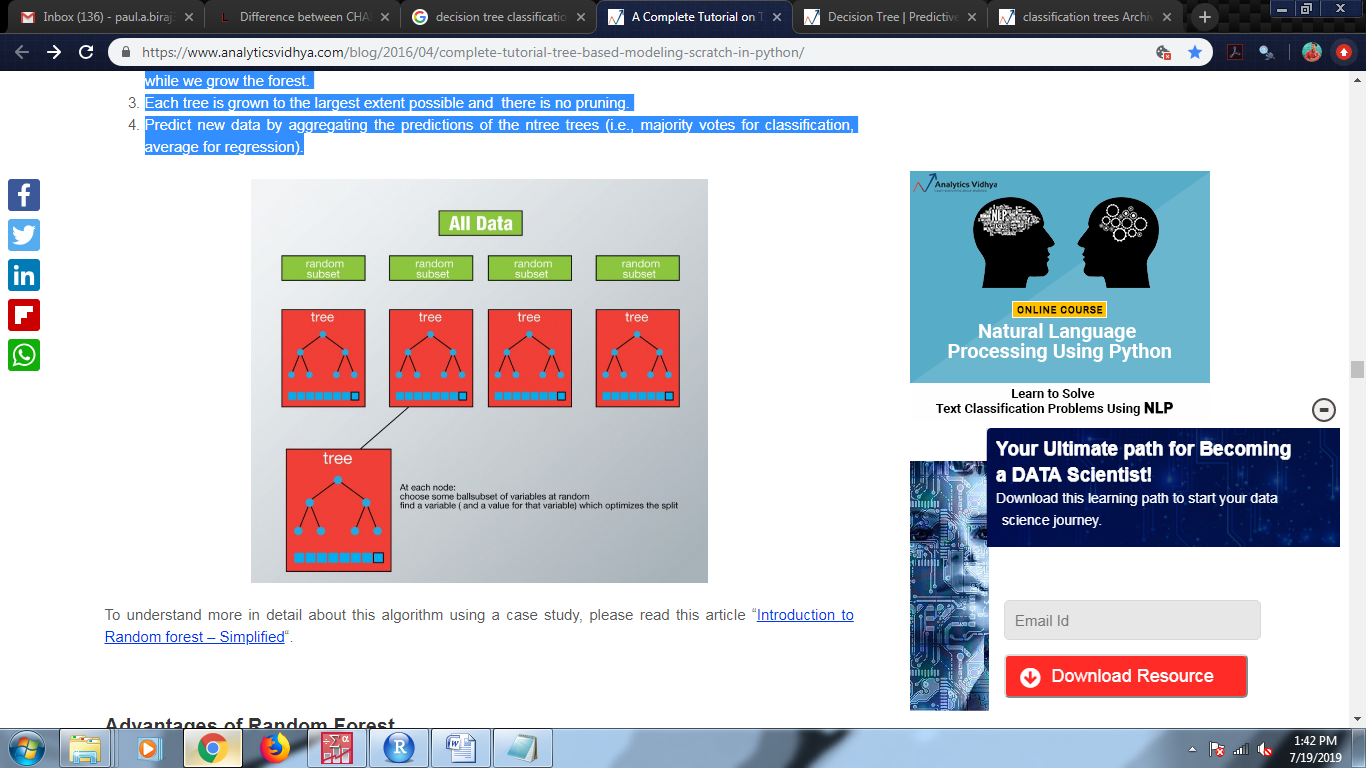
Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential [steps of data exploration](https://www.analyticsvidhya.com/blog/2015/02/data-exploration-preparation-model/), and does a fairly good job. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.

### How does it work?

In Random Forest, we grow multiple trees as opposed to a single tree in CART model (see comparison between CART and Random Forest here, [part1](https://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-forest-1/) and [part2](https://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model/)). To classify a new object based on attributes, each tree gives a classification and we say the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest) and in case of regression, it takes the average of outputs by different trees.

It works in the following manner. Each tree is planted & grown as follows:

1. Assume number of cases in the training set is N. Then, sample of these N cases is taken at random but *with replacement*. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest.
3. Each tree is grown to the largest extent possible and  there is no pruning.
4. Predict new data by aggregating the predictions of the ntree trees (i.e., majority votes for classification, average for regression).



### Advantages of Random Forest

* This algorithm can solve both type of problems i.e. classification and regression and does a decent estimation at both fronts.
* One of benefits of Random forest which excites me most is, the power of handle large data set with higher dimensionality. It can handle thousands of input variables and identify most significant variables so it is considered as one of the dimensionality reduction methods. Further, the model outputs **Importance of variable,**which can be a very handy feature (on some random data set).
* It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
* It has methods for balancing errors in data sets where classes are imbalanced.
* The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.
* Random Forest involves sampling of the input data with replacement called as bootstrap sampling. Here one third of the data is not used for training and can be used to testing. These are called the **out of bag** samples. Error estimated on these out of bag samples is known as out of bag error. Study of error estimates by Out of bag, gives evidence to show that the out-of-bag estimate is as accurate as using a test set of the same size as the training set. Therefore, using the out-of-bag error estimate removes the need for a set aside test set.

### Disadvantages of Random Forest

* It surely does a good job at classification but not as good as for regression problem as it does not give precise continuous nature predictions. In case of regression, it doesn’t predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy.
* Random Forest can feel like a black box approach for statistical modelers – you have very little control on what the model does. You can at best – try different parameters and random seeds!

## What is Boosting ? How does it work?

Definition: The term ‘Boosting’ refers to a family of algorithms which converts weak learner to strong learners.

Let’s understand this definition in detail by solving a problem of spam email identification:

How would you classify an email as SPAM or not? Like everyone else, our initial approach would be to identify ‘spam’ and ‘not spam’ emails using following criteria. If:

1. Email has only one image file (promotional image), It’s a SPAM
2. Email has only link(s), It’s a SPAM
3. Email body consist of sentence like “You won a prize money of $ xxxxxx”, It’s a SPAM
4. Email from our official domain “[Analyticsvidhya.com](http://analyticsvidhya.com/)” , Not a SPAM
5. Email from known source, Not a SPAM

Above, we’ve defined multiple rules to classify an email into ‘spam’ or ‘not spam’. But, do you think these rules individually are strong enough to successfully classify an email? No.

Individually, these rules are not powerful enough to classify an email into ‘spam’ or ‘not spam’. Therefore, these rules are called as **weak learner**.

To convert weak learner to strong learner, we’ll combine the prediction of each weak learner using methods like:

* Using average/ weighted average
* Considering prediction has higher vote

### How does it work?

Now we know that, boosting combines weak learner a.k.a. base learner to form a strong rule. An immediate question which should pop in your mind is, ‘How boosting identify weak rules?‘

To find weak rule, we apply base learning (ML) algorithms with a different distribution. Each time base learning algorithm is applied, it generates a new weak prediction rule. This is an iterative process. After many iterations, the boosting algorithm combines these weak rules into a single strong prediction rule.

Here’s another question which might haunt you, ‘How do we choose different distribution for each round?’

For choosing the right distribution, here are the following steps:

Step 1:  The base learner takes all the distributions and assign equal weight or attention to each observation.

Step 2: If there is any prediction error caused by first base learning algorithm, then we pay higher attention to observations having prediction error. Then, we apply the next base learning algorithm.

Step 3: Iterate Step 2 till the limit of base learning algorithm is reached or higher accuracy is achieved.

Finally, it combines the outputs from weak learner and creates a strong learner which eventually improves the prediction power of the model. Boosting pays higher focus on examples which are mis-classiﬁed or have higher errors by preceding weak rules.

**DATA SET INFORMATION**

|  |  |  |
| --- | --- | --- |
| **DATA SET** | | |
| **Data set** | **246** Samples | **28** Variables |
| **Train Set** | **164** Samples | **28** Variables |
| **Test Set** | **82** Samples | **28** Variables |

**Dataset splitting:** It is standard in statistical modeling/Machine learning analysis to split data into training and test sets. The reason forth it is very straightforward: if you try and evaluate your system on data you have trained it on, you are doing something unrealistic. The whole point of a Statistical modeling/machine learning system is to be able to work with unseen data, if you know you are going to see all possible values in your training data, you might as well just use some form of lookup. However, the two-way split is over-simplistic. Real Statistical modeling/ML typically involves three phases:1. Training2. Development (also known as Validation or Tuning)3. Testing (aka Evaluation). However for our convenience and dataset being not large enough to partition into three parts, we have split into two parts i.e **splitting data set randomly into Train Set and Test Set, with ratio 2/3** on the basis of the dependant variable being continuous or categorical.

* **training set**—a subset to train a model.
* **test set**—a subset to test the trained model.

You could imagine slicing the single data set as follows:

**Slicing a single data set into a training set and test set.**

Make sure that your test set meets the following two conditions:

* Is large enough to yield statistically meaningful results.
* Is representative of the data set as a whole. In other words, don't pick a test set with different characteristics than the training set.

Assuming that your test set meets the preceding two conditions, your goal is to create a model that generalizes well to new data. Our test set serves as a proxy for new data. For example, consider the following figure. Notice that the model learned for the training data is very simple. This model doesn't do a perfect job—a few predictions are wrong. However, this model does about as well on the test data as it does on the training data. In other words, this simple model does not overfit the training data.

**Never train on test data.** If you are seeing surprisingly good results on your evaluation metrics, it might be a sign that you are accidentally training on the test set. For example, high accuracy might indicate that test data has leaked into the training set. **For example:** consider a model that predicts whether an email is spam, using the subject line, email body, and sender's email address as features. We apportion the data into training and test sets, with an 80-20 split. After training, the model achieves 99% precision on both the training set and the test set. We'd expect a lower precision on the test set, so we take another look at the data and discover that many of the examples in the test set are duplicates of examples in the training set (we neglected to scrub duplicate entries for the same spam email from our input database before splitting the data). We've inadvertently trained on some of our test data, and as a result, we're no longer accurately measuring how well our model generalizes to new data.

**VARIABLE INFORMATION**

The variables having categories are converted to factors with levels and the variables which are continuous are kept as it is. The variables design are defined corresponding to the variable as under.

|  |  |  |
| --- | --- | --- |
| **CLASS** | **STRING VALUE** | **LEVELS** |
| **Location** | Mumbai | 1 |
| Mumbai Suburban | 2 |
| Navi Mumbai | 3 |
| Thane | 4 |
| **Gender** | Male | 1 |
| Female | 2 |
| **Marital Status** | Married | 1 |
| Unmarried | 2 |
| **Qualification** | Secondary School(SSC) | 1 |
| Higher Secondary(HSC) | 2 |
| Graduate | 3 |
| Post Graduate | 4 |
| Other | 5 |
| **Employment Status** | Employed | 1 |
| Self- Employed | 2 |
| Student | 3 |
| Unemployed | 4 |
| **Employment Sector** | Business | 1 |
| Government | 2 |
| Private | 3 |
| Unemployed | 4 |
| **Income(Rs.)** | 0 | 1 |
| 1 to 2.5(in Lakh) | 2 |
| 2.5 to 5(in Lakh) | 3 |
| 5 to 7.5(in Lakh) | 4 |
| Above 7.5(in Lakh) | 5 |
| **Age** | <18 | 1 |
| 18-23 | 2 |
| 24-28 | 3 |
| 29-34 | 4 |
| 35-38 | 5 |
| 38-42 | 6 |
| Above 42 | 7 |
| **Daily Usage** | Daily | 1 |
| Weekly | 2 |
| Rarely | 3 |
| Several Time a month | 4 |
| Never | 5 |
| **Time Saving** | Yes | 1 |
| No | 2 |
| **Cost** | Yes | 1 |
| No | 2 |
| **Reliability and Safety** | Yes | 1 |
| No | 2 |
| **Comfort** | Yes | 1 |
| No | 2 |
| **No. of Vehicles** | 0 | 1 |
| 1 | 2 |
| 2 | 3 |
| 3 | 4 |
| 4 | 5 |
| More than 4 | 6 |
| **Daily Trip(Km)** | <5 Km | 1 |
| 5-10 Km | 2 |
| 10-20Km | 3 |
| 20-30Km | 4 |
| 30-40 Km | 5 |
| 40-60Km | 6 |
| Above 60Km | 7 |
| **Walking Dist(Km)** | 0-1Km | 1 |
| 1-2Km | 2 |
| Above 2Km | 3 |
| **Time-by-go** | 5-7 AM | 4 |
| 7-9 AM | 5 |
| 9-11 AM | 6 |
| 12PM & Above | 7 |
| **Time-by-come** | 12-3 PM | 1 |
| 3-6 PM | 2 |
| 6-7 PM | 3 |
| 7-8 PM | 4 |
| 8-9 PM | 5 |
| 9-11 PM | 6 |
| 11 & Above | 7 |
| **Train** | Yes | 1 |
| No | 2 |
| **Auto** | Yes | 1 |
| No | 2 |
| **Bus** | Yes | 1 |
| No | 2 |
| **Bike** | Yes | 1 |
| No | 2 |
| **Ola-Uber** | Yes | 1 |
| No | 2 |
| **Personal Vehicle** | Yes | 1 |
| No | 2 |
| **Taxi** | Yes | 1 |
| No | 2 |
| **Walking** | Yes | 1 |
| No | 2 |
| **Time** | Numeric/Integer | Continuous |
| **Money** | Numeric/Integer | Continuous |
| **Delay Time Spend** | Numeric/Integer | Continuous |
| **Delay Money Spend** | Numeric/Integer | Continuous |

**Note:**

* Time-by-go – What time people usually travel while going for their respect work/activity.
* Time-by-come- time people usually travel while coming from their respect work/activity.
* Daily Trip (Km)- The distance travel on average while traveling.
* Walking distance (Km)- The distance walked on average while traveling.
* Delay Time Spend – The amount of Time Spend(Hrs) if delayed due any reason while traveling.
* Delay Money Spend – The amount of Money Spend(Rs) if delayed due any reason while traveling.
* Time – The amount of Time Spend(Hrs) for traveling.
* Money – The amount of Money Spend(Rs) for traveling.

**OBJECTIVE-1**

1. **To identify and analyze the socio-demographic and geographical factors that influences the population for choosing Train as the medium of transport.**

**Variables:** The variables used for the above objective are all the socio-demographic and geographical variables as mentioned above in “*Variable information*”, excluding the continuous variable (Time, Money, Time Spend and Money Spend), taking Train as dependant variable and rest as independent variables.

**MODEL BUILDING**

**LOGISTIC REGRESSION METHOD:**

The model is build using all the socio- demographic and geographical variables as independent variable, dependant variable being train.

The logit function is :

𝐿𝑜𝑔𝑖t (𝑝) = log = 𝛽o+∑m,ni≠j=1𝛽ij𝑥ij

Where,

𝛽o = Intercept

𝛽ij = Co-efficient of the jth level of the ith variable.

i = is the variable; i=1, 2, 3,…,m

j = is the levels of categorical variable, j=1, 2, 3,…,n

Each variable has total number of 256 observations in total.

**Output :**

| **Classification Tablea,b** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Train | | Percentage Correct |
|  | 0 | 1 |
| Step 0 | Train | 0 | 0 | 73 | .0 |
| 1 | 0 | 173 | 100.0 |
| Overall Percentage | |  |  | 70.3 |
| a. Constant is included in the model. | | | | | |
| b. The cut value is .500 | | | | | |

**Overall Percentage** – This gives the percent of cases for which the dependent variables was correctly predicted given the model.  In this part of the output, this is the null model.  70.3%.

| **Case Processing Summary** | | | |
| --- | --- | --- | --- |
| UnweightedCasesa | | N | Percent |
| Selected Cases | Included in Analysis | 246 | 100.0 |
| Missing Cases | 0 | .0 |
| Total | 246 | 100.0 |
| Unselected Cases | | 0 | .0 |
| Total | | 246 | 100.0 |
| a. If weight is in effect, see classification table for the total number of cases. | | | |

| **Variables in the Equation** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 0 | Constant | .863 | .140 | 38.220 | 1 | .000 | 2.370 |

**Wald and Sig**. – This is the Wald chi-square test that tests the null hypothesis that the constant equals 0.    
This hypothesis is rejected because the p-value (listed in the column called “Sig.”) is smaller than the critical p-value of .05 (or .01).  Hence, we conclude that the constant is not 0.

**Omnibus Tests of Model Coefficients**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Chi-square | df | Sig. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Step 6 | Step | 6.583 | 1 | .010 |
| Block | 141.086 | 14 | .000 |
| Model | 141.086 | 14 | .000 |

**Conclusion :** The model is statistically significant because the p-value is less than 0.000.

| **Model Summary** | | | |
| --- | --- | --- | --- |
| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
| 1 | 220.583a | .273 | .389 |
| 2 | 198.496a | .336 | .477 |
| 3 | 190.085a | .358 | .509 |
| 4 | 172.225a | .403 | .573 |
| 5 | 164.675a | .421 | .599 |
| 6 | 158.091a | .436 | .620 |

**R square is 62%**

**Hosmer and Lemeshow Test:**

| **Hosmer and Lemeshow Test** | | | |
| --- | --- | --- | --- |
| Step | Chi-square | Df | Sig. |
| 1 | .000 | 2 | 1.000 |
| 2 | 2.541 | 3 | .468 |
| 3 | 1.674 | 5 | .892 |
| 4 | 6.125 | 8 | .633 |
| 5 | 4.779 | 8 | .781 |
| 6 | 5.679 | 8 | .683 |

H0:Model is good fit

H1:Model is not good fit

Since,p-value > 0.05, we fail to reject H0 and conclude that our model is a good fit.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | |
|  | | B | S.E. | Wald | Df | Sig. | Exp(B) |
| Step 6f | Dailyusage |  |  | 36.867 | 4 | .000 |  |
| Dailyusage(1) | 1.725 | .684 | 6.368 | 1 | .012 | 5.615 |
| Dailyusage(2) | -20.170 | 40192.970 | .000 | 1 | 1.000 | .000 |
| Dailyusage(3) | -3.161 | .943 | 11.233 | 1 | .001 | .042 |
| Dailyusage(4) | .751 | .782 | .922 | 1 | .337 | 2.119 |
| TimeSaving | 2.494 | .510 | 23.919 | 1 | .000 | 12.112 |
| Cost | 1.115 | .446 | 6.242 | 1 | .012 | 3.050 |
| Timebycome |  |  | 18.150 | 6 | .006 |  |
| Timebycome(1) | -1.308 | .986 | 1.758 | 1 | .185 | .270 |
| Timebycome(2) | -1.535 | .869 | 3.120 | 1 | .077 | .216 |
| Timebycome(3) | 1.331 | .752 | 3.135 | 1 | .077 | 3.785 |
| Timebycome(4) | .463 | .668 | .481 | 1 | .488 | 1.589 |
| Timebycome(5) | 2.114 | .823 | 6.603 | 1 | .010 | 8.283 |
| Timebycome(6) | .928 | .742 | 1.562 | 1 | .211 | 2.529 |
| Bike | -1.414 | .502 | 7.936 | 1 | .005 | .243 |
| Walking | 2.056 | .608 | 11.420 | 1 | .001 | 7.815 |
| Constant | -2.577 | .948 | 7.395 | 1 | .007 | .076 |

**Interpretation**

Odds ratio is a measure of association. It approximates how much more likely it is for outcome to be present among the different levels of independent variables.

* People who travels daily and several time in month are more likely to use train than people who never use train.
* People select train 3 times more likely for cost than comfort
* People who travel by train is 8.2 times more likely in time 8-9 than 11 and above
* People who use bike are 0.243 time more likely to use train than those people who do not use bike
* People who walk are7.815 time more likely use train than those people who do not walk.

**Classification:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| observed | | |  | | Predicted | | | | |
|  | | | Train | | | Percentage Correct | |
|  | | | 0 | 1 | |
| Step 6 | Train | 0 | | 48 | | | 25 | | 65.8 |
| 1 | | 11 | | | 162 | | 93.6 |
| Overall Percentage | | |  | | |  | | 85.4 |

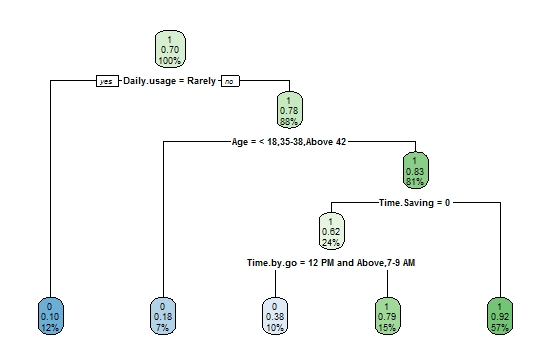
Accuracy of this is 85%

**ROC Curve:**

****

| **Area Under the Curve** |
| --- |
| Test Result Variable(s):Predicted probability |
| Area |
| .887 |

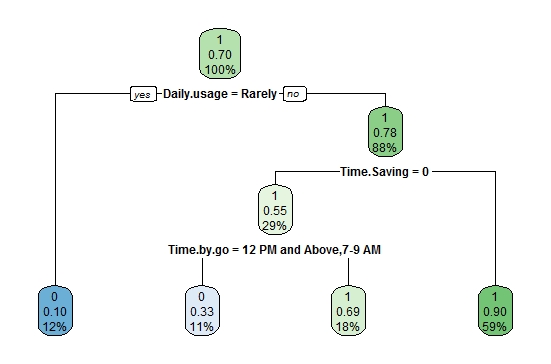
**CLASSFICATION METHOD:**



**Decision Tree**

As we can see from the above tree, there are 4 variables retained from the Full-Model tree for the classification of ‘Train’(dependant variable/binary variable),on the basis of independent variables which are significant for the model, where ‘1’ being; using Train and ‘0’ being not using Train. Accuracy gained, 74.39% for test set; where the tree takes into account all variables and observations with no conditions applied to its splitting and branches.

**Model Tree:** Daily usage, Age, Time saving (1-yes, 0-No), Time-by-go is retained.



**Decision Tree(With Min-Split=40)**

In order to prevent over fitting and prune the tree, we tried a minimum split of 30, 40, 50 as the criteria for splitting, where the best tree with better accuracy and efficiency was found for min-split of 40 with accuracy 79.26% for test set, thereby making the tree more simpler and efficient to understand.

**Model Tree:** Daily usage, Time saving (1-yes,0-No), Time-by-go is retained.

**ENSEMBLE MODELS AND THERE ACCURACY**

In order to reduce Bias-Variance error for the Tree Models, we have used ensemble methods involving group of predictive models to achieve a better accuracy and model stability for Bias-Variance error **trade-off management. The models and there outputs are as given in the table below.**

|  |  |  |
| --- | --- | --- |
| **Model** | **Test Set –Accuracy (%)** | **Confusion Matrix** |
| **Decision Tree** | 74.39 | |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 15 | 12 | | 1 | 9 | 46 |   1 – Train  0 – No-Train  A/P-Actual vs Predicted |
| **Decision Tree**  **(With MinSplit=40)** | 79.26 | |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 13 | 9 | | 1 | 11 | 49 |   1 – Train  0 – No-Train  A/P-Actual vs Predicted   |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 14 | 7 | | 1 | 10 | 51 |   1 – Train  0 – No-Train  A/P-Actual vs Predicted |
| **Bagging** | 75.60 |  |
| **Random Forest** | 80.48 | |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 15 | 5 | | 1 | 9 | 53 |   1 – Train  0 – No-Train  A/P-Actual vs Predicted   |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 13 | 5 | | 1 | 11 | 53 |   1 – Train  0 – No-Train  A/P-Actual vs Predicted |
| **Boosting**  **(ntree = 4500**  **shrink=0.001)** | 82.92 |  |

**INTERPRETATION:**

For the above objective for analyzing the socio-demographic and geographical factors that influences the population for choosing Train as the medium of transport, we have used Logistic regression model firstly, where the R square value was found to be 62% with few important variables retained such as time saving , cost, daily usage and several time in a month, with model accuracy 85% and area under the ROC curve was found to be 0.887.

For the same objective we have tried other models such as decision tree models along with ensembling models for comparing the results of each model where the data set was divided into train set and test set and the model was run on train set and accuracy was computed from the test set, thereby finding out which model is working much better in terms of giving accuracy for classification of the dependant variable i.e train in our case, in satisfying the objective.

The models such as Decision Tree with min-split 40 was giving 79.26% having the variables retained such as daily usage, time saving, time by go and age.

Boosting technique showed highest accuracy with 82.92% among the ensembling models.

**OBJECTIVE-1**

1. **To identify and analyze the socio-demographic and geographical factors that influences the population for choosing Auto as the medium of transport.**

**MODEL BUILDING**

**Variables:** The variables used for the above objective are all the socio-demographic and geographical variables as mentioned above in “*Variable information*”, excluding the continuous variable (Time, Money, Time Spend and Money Spend), taking Auto as dependant variable and rest as independent variables.

**LOGISTIC REGRESSION METHOD:**

The model is build using all the socio- demographic and geographical variables as independent variable, dependant variable being Auto.

The logit function is :

𝐿𝑜𝑔𝑖t (𝑝) = log = 𝛽o+∑m,ni≠j=1𝛽ij𝑥ij

Where,

𝛽o = Intercept

𝛽ij = Co-efficient of the jth level of the ith variable.

i = is the variable; i=1, 2, 3,…,m

j = is the levels of categorical variable, j=1, 2, 3,…,n

Each variable has total number of 256 observations in total.

**OUTPUT:**

| **Case Processing Summary** | | | |
| --- | --- | --- | --- |
| UnweightedCasesa | | N | Percent |
| Selected Cases | Included in Analysis | 246 | 100.0 |
| Missing Cases | 0 | .0 |
| Total | 246 | 100.0 |
| Unselected Cases | | 0 | .0 |
| Total | | 246 | 100.0 |
| a. If weight is in effect, see classification table for the total number of cases. | | | |

| **Classification Tablea,b** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Auto | | Percentage Correct |
|  | 0 | 1 |
| Step 0 | Auto | 0 | 146 | 0 | 100.0 |
| 1 | 100 | 0 | .0 |
| Overall Percentage | |  |  | 59.3 |

**Overall Percentage** – This gives the percent of cases for which the dependent variables was correctly predicted given the model.  In this part of the output, this is the null model.  59.3%

| **Variables in the Equation** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 0 | Constant | -.378 | .130 | 8.500 | 1 | .004 | .685 |

**Wald and Sig**. – This is the Wald chi-square test that tests the null hypothesis that the constant equals 0.    
This hypothesis is rejected because the p-value (listed in the column called “Sig.”) is smaller than the critical p-value of .05 (or .01).  Hence, we conclude that the constant is not 0.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Omnibus Tests of Model Coefficients** | | | | |
|  | | Chi-square | Df | Sig. |
| Step 4 | Step | 6.627 | 1 | .010 |
| Block | 235.811 | 26 | .000 |
| Model | 235.811 | 26 | .000 |

**Conclusion :** The model is statistically significant because the p-value is less than 0.000.

| **Model Summary** | | | |
| --- | --- | --- | --- |
| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
| 1 | 313.375a | .074 | .100 |
| 2 | 299.855b | .124 | .167 |
| 3 | 288.727b | .163 | .219 |
| 4 | 277.619b | .200 | .269 |
| 5 | 259.654c | .256 | .345 |
| 6 | 255.370c | .269 | .363 |

**R square is 36.3%**

| **Hosmer and Lemeshow Test** | | | |
| --- | --- | --- | --- |
| Step | Chi-square | df | Sig. |
| 1 | .000 | 0 | . |
| 2 | 4.711 | 2 | .095 |
| 3 | 11.338 | 5 | .045 |
| 4 | 7.295 | 6 | .294 |
| 5 | 10.017 | 8 | .264 |
| 6 | 9.577 | 7 | .214 |

H0:Model is good fit

H1:Model is not good fit

Since,p-value > 0.05, we fail to reject H0 and conclude that our model is a good fit.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 6 | Gender(1) | -1.059 | .317 | 11.185 | 1 | .001 | .347 |
| Qualification |  |  | .691 | 4 | .952 |  |
| Qualification(1) | -.846 | 20039.904 | .000 | 1 | 1.000 | .429 |
| Qualification(2) | 21.966 | 11450.624 | .000 | 1 | .998 | 34.660 |
| Qualification(3) | 21.536 | 11450.624 | .000 | 1 | .998 | 22.642 |
| Qualification(4) | 21.502 | 11450.624 | .000 | 1 | .999 | 21.825 |
| Train(1) | -1.231 | .378 | 10.597 | 1 | .001 | .292 |
| OlaUber(1) | -1.546 | .564 | 7.515 | 1 | .006 | .213 |
| Taxi(1) | -1.289 | .663 | 3.773 | 1 | .052 | .276 |
| Walking(1) | -1.122 | .336 | 11.130 | 1 | .001 | .325 |
| Constant | -17.569 | 11450.624 | .000 | 1 | .999 | .000 |

**Interpretation:**

Odds ratio is a measure of association. It approximates how much more likely it is for outcome to be present among the different levels of independent variables.

* Male uses Auto more likely than female
* People who use train are 0.292 time more likely use auto than those people who not use train
* People who use ola uber are0.213 time more likely to use auto than those people who do not use ola uber
* People who use taxi are0.276 time more likely to use auto than those people who do not use taxi
* People who walk are0.325timemorelikelyuseautothanthosepeoplewhonotwa

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
| **Classification Table** | | | | | | | | | |
|  | Observed | | | Predicted | | | | | |
|  | Auto | | | | Percentage Correct | |
|  | 0 | | 1 | |
| Step 4 | | Auto | 0 | | 132 | | 14 | | 90.4 |
| 1 | | 11 | | 89 | | 89.0 |
| Overall Percentage | | |  | |  | | 89.8 |

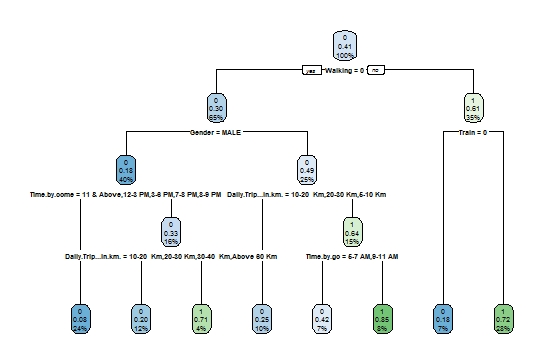
**Accuracy of this is 89.8%**

**ROC curve:**

****

| **Area Under the Curve** |
| --- |
| Test Result Variable(s):Predicted probability |
| Area |
| .798 |

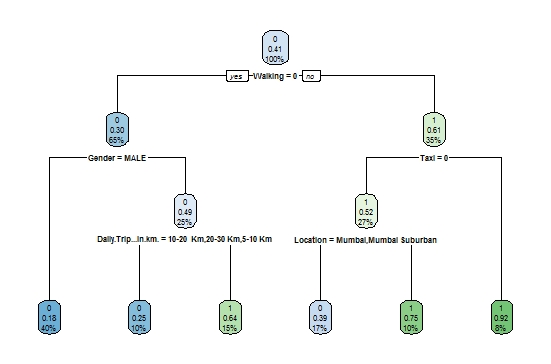
**CLASSFICATION METHOD :**



**Decision Tree**

As we can see from the above tree, there are 5 variables retained from the Full-Model tree for the classification of ‘Auto’ (dependant variable/binary variable), on the basis of independent variables, which are significant for the model; where ‘1’ being, using Auto and ‘0’ being not using Auto. Accuracy gained, 65.85% for test set; where the tree takes into account all variables and observations with no conditions applied to its splitting and branches.

**Model Tree:** Walking(1-yes,0-No),Gender, Time-by-go, Daily Trip(Km),Train(1-yes,0-No).



**Decision Tree (With MinSplit=40)**

In order to prevent over fitting and prune the tree, we tried a minimum split of 30, 40, 50 as the criteria for splitting, where the best tree with better accuracy and efficiency was found for min-split of 40 with accuracy 69.51% for test set, thereby making the tree more simpler and efficient to understand. **Model Tree:** Walking (1-yes,0-No), Gender, Taxi(1-yes,0-No), Daily Trip(Km), Location is retained.

**ENSEMBLE MODELS AND THERE ACCURACY**

In order to reduce Bias-Variance error for the Tree Models, we have used ensemble methods involving group of predictive models to achieve a better accuracy and model stability for Bias-Variance error **trade-off management. The models and there outputs are as given in the table below.**

|  |  |  |
| --- | --- | --- |
| **Model** | **Test Set –Accuracy (%)** | **Confusion Matrix** |
| **Decision Tree** | 65.85 | |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 40 | 19 | | 1 | 9 | 14 |   1 – Auto  0 – No-Auto  A/P-Actual vs Predicted |
| **Decision Tree**  **(With MinSplit=40)** | 69.51 | |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 41 | 20 | | 1 | 8 | 13 |   1 – Auto  0 – No-Auto  A/P-Actual vs Predicted   |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 41 | 17 | | 1 | 8 | 16 |   1 – Auto  0 – No-Auto  A/P-Actual vs Predicted |
| **Bagging** | 65.85 |  |
| **Random Forest** | 64.63 | |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 42 | 22 | | 1 | 7 | 11 |   1 – Auto  0 – No-Auto  A/P-Actual vs Predicted |
| **Boosting**  **(ntree = 4500**  **shrink=0.001)** | 63.41 | |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 39 | 20 | | 1 | 10 | 13 |   1 – Auto  0 – No-Auto  A/P-Actual vs Predicted |

**INTERPRETATION:**

For the above objective for analyzing the socio-demographic and geographical factors that influences the population for choosing Auto as the medium of transport, we have used Logistic regression model firstly, where the R square value was found to be 36.3% with few important variables retained such as walking, Gender, train, ola-uber, with model accuracy 89.8% and area under the ROC curve was found to be 0.798.

For the same objective we have tried other models such as decision tree models along with ensembling models for comparing the results of each model where the data set was divided into train set and test set and the model was run on train set and accuracy was computed from the test set, thereby finding out which model is working much better in terms of giving accuracy for classification of the dependant variable i.e Auto in our case, in satisfying the objective.

The models such as Decision Tree with min-split 40 was giving 69.51% being highest among the decision tree models having the variables retained such as walking, gender, time-by-go, daily trip(Km),train.

**OBJECTIVE-1**

1. **To identify and analyze the socio-demographic and geographical factors that influences the population for choosing Bus as the medium of transport.**

**MODEL BUILDING**

**Variables:** The variables used for the above objective are all the socio-demographic and geographical variables as mentioned above in “*Variable information*”, excluding the continuous variable (Time, Money, Time Spend and Money Spend), taking Bus as dependant variable and rest as independent variables.

**LOGISTIC REGRESSION METHOD:**

The model is build using all the socio- demographic and geographical variables as independent variable, dependant variable being Bus.

The logit function is :

𝐿𝑜𝑔𝑖t (𝑝) = log = 𝛽o+∑m,ni≠j=1𝛽ijk𝑥ijk

Where,

𝛽o = Intercept

𝛽ij = Co-efficient of the jth level of the ith variable.

i = is the variable; i=1, 2, 3,…,m

j = is the levels of categorical variable, j=1, 2, 3,…,n

Each variable has total number of 256 observations in total.

**OUTPUT:**

| **Case Processing Summary** | | | |
| --- | --- | --- | --- |
| UnweightedCasesa | | N | Percent |
| Selected Cases | Included in Analysis | 246 | 100.0 |
| Missing Cases | 0 | .0 |
| Total | 246 | 100.0 |
| Unselected Cases | | 0 | .0 |
| Total | | 246 | 100.0 |
| a. If weight is in effect, see classification table for the total number of cases. | | | |

| **Classification Tablea,b** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Bus | | Percentage Correct |
|  | 0 | 1 |
| Step 0 | Bus | 0 | 148 | 0 | 100.0 |
| 1 | 98 | 0 | .0 |
| Overall Percentage | |  |  | 60.2 |
| a. Constant is included in the model. | | | | | |
| b. The cut value is .500 | | | | | |

**Overall Percentage** – This gives the percent of cases for which the dependent variables was correctly predicted given the model.  In this part of the output, this is the null model.  60.2%

| **Variables in the Equation** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 0 | Constant | -.412 | .130 | 10.020 | 1 | .002 | .662 |

**Wald and Sig**. – This is the Wald chi-square test that tests the null hypothesis that the constant equals 0.    
This hypothesis is rejected because the p-value (listed in the column called “Sig.”) is smaller than the critical p-value of .05 (or .01).  Hence, we conclude that the constant is not 0.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Omnibus Tests of Model Coefficients** | | | | |
|  | | Chi-square | df | Sig. |
| Step 4 | Step | 13.414 | 6 | .037 |
| Block | 44.612 | 11 | .000 |
| Model | 44.612 | 11 | .000 |

**Conclusion :** The model is statistically significant because the p-value is less than 0.000.

| **Model Summary** | | | |
| --- | --- | --- | --- |
| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
| 1 | 317.723a | .052 | .070 |
| 2 | 310.220a | .080 | .109 |
| 3 | 299.596a | .119 | .161 |
| 4 | 286.182a | .166 | .224 |
| **R square is 22.4%** | | | |

**Hosmer and Lemeshow Test:**

| **Hosmer and Lemeshow Test** | | | |
| --- | --- | --- | --- |
| Step | Chi-square | df | Sig. |
| 1 | .000 | 0 | . |
| 2 | .018 | 1 | .895 |
| 3 | 7.437 | 7 | .385 |
| 4 | 7.494 | 8 | .484 |

H0:Model is good fit

H1:Model is not good fit

Since,p-value > 0.05, we fail to reject H0 and conclude that our model is a good fit.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Classification Tablea** | | | | | | | | | |
|  | Observed | | | Predicted | | | | | |
|  | Bus | | | | Percentage Correct | |
|  | 0 | | 1 | |
| Step 4 | | Bus | 0 | | 117 | | 31 | | 79.1 |
| 1 | | 47 | | 51 | | 52.0 |
| Overall Percentage | | |  | |  | | 68.3 |

**Accuracy of this is 68.3%**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | |
|  | | B | S.E. | Wald | Df | Sig. | Exp(B) |
| Step 4d | Location |  |  | 11.677 | 3 | .009 |  |
| Location(1) | -.712 | .396 | 3.229 | 1 | .052 | .491 |
| Location(2) | -1.548 | .472 | 10.783 | 1 | .001 | .213 |
| Location(3) | -1.057 | .464 | 5.185 | 1 | .023 | .348 |
| Cost | 1.207 | .306 | 15.523 | 1 | .000 | 3.343 |
| DailyTripInkm |  |  | 12.342 | 6 | .055 |  |
| DailyTripInkm(1) | .486 | .804 | .365 | 1 | .545 | 1.626 |
| DailyTripInkm(2) | 1.528 | .552 | 7.649 | 1 | .006 | 4.607 |
| DailyTripInkm(3) | 1.123 | .569 | 3.895 | 1 | .048 | 3.074 |
| DailyTripInkm(4) | 1.598 | .525 | 9.265 | 1 | .002 | 4.943 |
| DailyTripInkm(5) | .745 | .552 | 1.823 | 1 | .177 | 2.107 |
| DailyTripInkm(6) | 1.006 | .571 | 3.106 | 1 | .078 | 2.736 |
| Taxi | 1.226 | .474 | 6.685 | 1 | .010 | 3.409 |
| Constant | -1.539 | .506 | 9.243 | 1 | .002 | .215 |

**Interpretation:**

Odds ratio is a measure of association. It approximates how much more likely it is for outcome to be present among the different levels of independent variables.

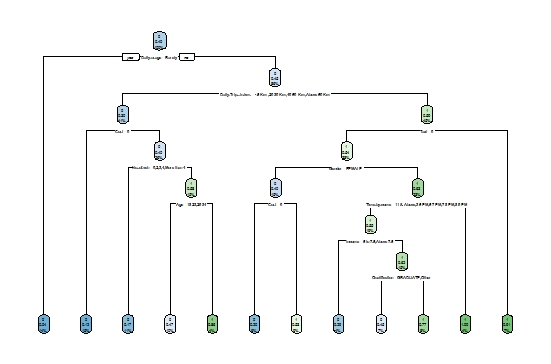
* People who travels from Mumbai, Mumbai Suburban, Navi Mumbai are more likely to use bus than people from thane .
* People select bus 3.3 times more likely for cost than comfort
* People who travels by bus on weekly,several times in months and rarely are more likely use bus than who never travels by bus.
* People who use taxi are3.409 time more likely to use bus than those people who do not use taxi

**ROC curve:**

****

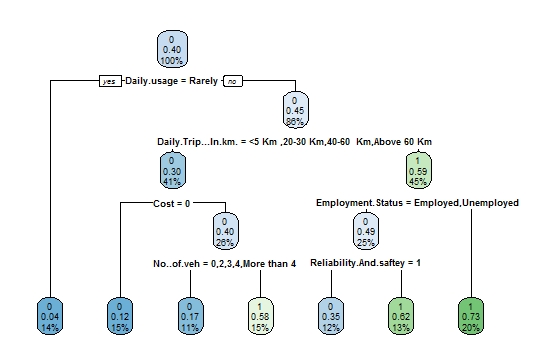
| **Area Under the Curve** |
| --- |
| Test Result Variable(s):Predicted probability |
| Area |
| .690 |

**CLASSFICATION METHOD :**



**Decision Tree**

As we can see from the above tree, there are many variables retained from the Full-Model tree for the classification of ‘Bus’(dependant variable/binary variable),on the basis of independent variables, tree being not understandable due to presence of two many splits, where ‘1’ being; using Bus and ‘0’ being not using Bus. Accuracy gained, 58.53% for test set; but inefficient, where the tree takes into account all variables and observations with no conditions applied to its splitting and branches.

****

**Decision Tree (With MinSplit=40)**

In order to prevent over fitting and prune the tree, we tried a minimum split of 30, 40, 50 as the criteria for splitting, where the best tree with better accuracy and efficiency was found for min-split of 40 with accuracy 52.43% for test set, thereby making the tree more simpler and efficient to understand; where ‘1’ being; using Bus and ‘0’ being not using Bus.

**Model Tree:** Daily usage, Daily Trip (Km), cost(1-yes,0-No),Employment status, No. of vehicle, Reliability and safety(1-yes,0-No) is retained

**ENSEMBLE MODELS AND THERE ACCURACY**

In order to reduce Bias-Variance error for the Tree Models, we have used ensemble methods involving group of predictive models to achieve a better accuracy and model stability for Bias-Variance error **trade-off management. The models and there outputs are as given in the table below.**

|  |  |  |
| --- | --- | --- |
| **Model** | **Test Set –Accuracy (%)** | **Confusion Matrix** |
| **Decision Tree** | 58.53 | |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 38 | 23 | | 1 | 11 | 10 |   1 – Bus  0 – No-Bus  A/P-Actual vs Predicted |
| **Decision Tree**  **(With MinSplit=40)** | 52.43 | |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 36 | 23 | | 1 | 13 | 10 |   1 – Bus  0 – No-Bus  A/P-Actual vs Predicted   |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 28 | 18 | | 1 | 21 | 15 |   1 – Bus  0 – No-Bus  A/P-Actual vs Predicted |
| **Bagging** | 56.09 |  |
| **Random Forest** | 59.75 | |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 42 | 26 | | 1 | 7 | 7 |   1 – Bus  0 – No-Bus  A/P-Actual vs Predicted   |  |  |  | | --- | --- | --- | | A/P | 0 | 1 | | 0 | 42 | 26 | | 1 | 7 | 7 |   1 – Bus  0 – No-Bus  A/P-Actual vs Predicted |
| **Boosting**  **(ntree = 4500**  **shrink=0.001)** | 59.75 |  |

**INTERPRETATION:**

For the above objective for analyzing the socio-demographic and geographical factors that influences the population for choosing Bus as the medium of transport, we have used Logistic regression model firstly, where the R square value was found to be 22.4% with few important variables retained such as cost, taxi, Mumbai-suburban, daily trip(Km)10-20 & daily trip 30-40, with model accuracy 68.3% and area under the ROC curve was found to be 0.690.

For the same objective we have tried other models such as decision tree models along with ensembling models for comparing the results of each model where the data set was divided into train set and test set and the model was run on train set and accuracy was computed from the test set, thereby finding out which model is working much better in terms of giving accuracy for classification of the dependant variable i.e Bus in our case, in satisfying the objective.

The models such as Decision Tree with min-split 40 was giving 58.53% having the variables retained such as cost, Employment status, No. of vehicle, Reliability and safety.

Boosting and random forest technique showed highest accuracy with 59.75% among the ensembling models.

**OBJECTIVE-2**

1. **To identify and analyze the socio-demographic factors and choice of mode of transport, influencing the Time spend by the people for traveling.**

**ONE SAMPLE T TEST:**

The one sample *t*-test is a statistical procedure used to determine whether a sample of observations could have been generated by a process with a specific mean. Suppose you are interested in determining whether an assembly line produces laptop computers that weigh five pounds. To test this hypothesis, you could collect a sample of laptop computers from the assembly line, measure their weights, and compare the sample with a value of five using a one-sample *t*-test.

**The assumptions of the two-sample t-test are:**

* • The dependent variable must be continuous (interval/ratio).
* • The observations are independent of one another.
* • The dependent variable should be approximately normally distributed.
* • The dependent variable should not contain any outliers

* **DELAY TIME SPEND**

Notation:

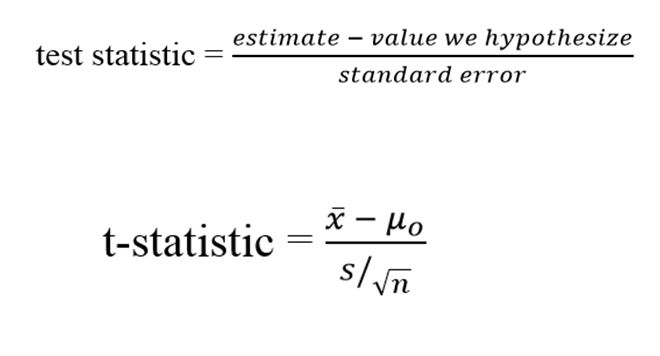
µ0: mean of delay Time spend(In Hrs)

*Equal variances are assumed for this analysis.*

Hypothesis:

*Ho: delay time spend = 0(µ0=3.227) vs H1: delay time spend≠0 (µ0>3.227).*

*Test statistic:*



X bar= sample mean of delay time spend

S=sample variance

n=no. of observations.

**Result:**

**One-Sample T: DelayTime :**

Test of mu = 3.227 vs > 3.227

95% Lower

Variable N Mean StDev SE Mean Bound T P

DelayTime 246 3.2277 0.7537 0.0481 3.1484 0.01 0.494

**Interpretation:**

For above one sample t test we are getting p value greater than 0.05. Hence we accept H0 i.e., the average delay time for transportation is 3.2277 Hrs.

***ANALYSIS OF VARIANCE (TIME):***

ANOVA is a statistical technique that assesses potential differences in a scale-level dependent variable by a nominal-level variable having 2 or more categories. An Analysis of Variance (one way ANOVA) willbe conducted to determine if there a significant difference on the dependent variable by independent variable. One way ANOVA is an appropriate statistical analysis when the purpose of research is to assess if mean differences exist on one continuous dependent variable by an independent variable with two or more discrete groups. In ANOVA dependent variable must be continuous (interval or ratio) level of measurement. Then independent variable In ANOVA must be categorical (nominal or ordinal) variable.

*In our model we have taken dependent variable* ***TIME*** *and independent variables are* **LOCATION, GENDER, MARITAL STATUS, QUALIFICATION, EMPLOYMENT STATUS, EMPLOYMENT SECTOR, INCOME, AGE, NO.OF VEHICLES AND DAILY TRIP-LENGTH.**

**Since the model is given as,**

**Y (ijkmnpqsr) = µ+Ai+Bj+Ck+Dm+En+Fp+Gq+Hs+Ir+Jl+ϵ(ijkmnpqsr).**

**Y= Money spend**

**µ=**General mean effect

**Ai** = Effect due to ith level of location i=1,2,3,4.

**Bj** = Effect due to jth level of gender j=1,2,3.

**Ck**= Effect due to kth level of marital status k=1,2.

**Dm**= Effect due to mth level of qualification m=1,2,3,4,5.

**En**= Effect due to nth level of employment status n=1,2,3,4.

**Fp**= Effect due to pth level of employment sector p=1,2,3,4.

**Gq**= Effect due to qth level of income q=1,2,3,4,5.

**Hs**= Effect due to sth level of age s=1,2,3,4,5,6,7.

**Ir**= Effect due to rth level of no.of vehicles r=1,2,3,4,5,6.

**Jl=** Effect due to lth level of Daily trip length l=1,2,3,4,5,6,7.

***(Part-1 ) Assumptions:***

***1****. The population from which samples are drawn should be normally distributed.*

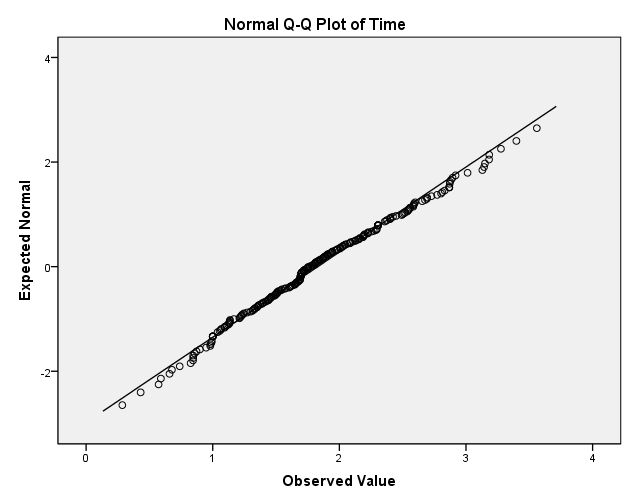
*Checking normality assumption is necessary, since the conclusion and interpretation is based on normality assumption. Here by using Kolmogorov-smirnov test and Q-Q plot we are going to check normality assumption.*

***Hypothesis for Kolmogorov-smirnov test***

*Ho : Data is normally distributed.*

*H1: Data is not normally distributed.*

| **Tests of Normality** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
| Statistic | df | Sig. | Statistic | df | Sig. |
| Time | .049 | 246 | .200\* | .992 | 246 | .241 |
| a. Lilliefors Significance Correction | | | | | | |
| \*. This is a lower bound of the true significance. | | | | | | |

**

***Conclusion****: From above kolmogorov-smirnov test and Q-Q plot we can say that the data are normally distributed.*

***2****.* ***Homogeneity of variance:***

*Homogeneity means that the variance among the groups should be approximately equal.*

*For Testing Equality of Variances, Levene’s test can be used.*

Ho: *The error variance of the dependent variable is equal across groups.*

H1: *The error variance of the dependent variable is* not *equal for at least one across groups.*

| **Test of Homogeneity of Variances** | | | |
| --- | --- | --- | --- |
| Time | | | |
| Levene Statistic | df1 | df2 | Sig. |
| 1.633 | 3 | 242 | .182 |

**Conclusion:**

*Since p-value is 0.182 i.e. greater than 0.05, the assumption of Homoscedasticity accepted.*

***3*** *.****Independence of cases:***

*The sample cases should be independent of each other. For checking independency we are using Durbin-Watson test. The test says that if the value lies between 1.5 to 2.5 then there is* ***NO-AUTOCORRELATION*** *in the observations.*

|  |  |
| --- | --- |
| *Model* | *Durbin-Watson* |
| *1* | *1.747* |

***Conclusion:*** *Since the Durbin-Watson test statistic value is 1.747 which lies between 1.5 to 2.5. Hence our model is free from auto-correlation.*

***4. Checking multicollinearity***

Model with collinear predictors can indicate how well the entire bundle of predictors predicts the [outcome variable](https://en.wikipedia.org/wiki/Dependent_variable#Use_in_statistics), but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others. Hence we go for this assumption,

|  |  |  |
| --- | --- | --- |
| ***Model*** | ***VIF*** | ***Condition Index*** |
| ***Constant*** |  |  |
| ***Location*** | 1.105 | 3.916 |
| ***Gender*** | 1.031 | 4.557 |
| ***Marital status*** | 1.983 | 6.303 |
| ***Qualification*** | 1.186 | 7.972 |
| ***Emp status*** | 1.340 | 8.664 |
| ***Emp sector*** | 1.381 | 9.445 |
| ***Income*** | 1.923 | 10.757 |

|  |  |  |
| --- | --- | --- |
| ***Age*** | 2.544 | 14.654 |
| ***No.of vehicles*** | 1.063 | 16.271 |

|  |  |  |
| --- | --- | --- |
| **Daily trip-length** | 1.117 | 37.979 |

**Conclusion: VIF values for all independent variables are less than 5 therefore our model is free from *multicollinearity.***

***Part (2)***

***Ho:*** *Average effect due to the different levels of independent variables on spending*

*Time is not significant.*

***H1:*** *Average effect due to the different levels of independent variables on spending time is significant.*

| **Tests of Between-Subjects Effects** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Dependent Variable:Time | | | | | |
| Source | Type III Sum of Squares | Df | Mean Square | F | Sig. |
| Corrected Model | 48.106a | 36 | 1.336 | 6.301 | .000 |
| Intercept | 44.221 | 1 | 44.221 | 208.510 | .000 |
| Location | .370 | 3 | .123 | .582 | .627 |
| Gender | .186 | 1 | .186 | .879 | .350 |
| MaritalStatus | .000 | 1 | .000 | .002 | .961 |
| Qualification | 1.577 | 4 | .394 | 1.859 | .119 |
| EmploymentStatus | .511 | 3 | .170 | .803 | .494 |
| EmpSector | .705 | 3 | .235 | 1.107 | .347 |
| Income | 2.273 | 4 | .568 | 2.680 | .033 |
| Age | 1.917 | 6 | .320 | 1.507 | .177 |
| No.ofveh | 1.871 | 5 | .374 | 1.764 | .122 |
| DailyTriplengthkm | 30.330 | 6 | 5.055 | 23.835 | .000 |
| Error | 44.325 | 209 | .212 |  |  |
| Total | 916.000 | 246 |  |  |  |
| Corrected Total | 92.431 | 245 |  |  |  |
| a. R Squared = .520 (Adjusted R Squared = .438) | | | | | |

**Conclusion: From the above table we observed that P-value for Income and Daily trip-length are significant (<0.05) and for remaining all the variables it’s insignificant (>0.05). Hence we conclude that levels of income and daily trip-length for individual who spending time on transport is not equal. And for all remaining variables it is insignificant for at least one level.**

**MULTIPLE REGRESSION:**

Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable. Regression analysis provides detailed insight that can be applied to further improve products and services.

In our model we have consider, dependent variable as time spend for transportation. And independent variable location,gender,marital status,qualification,emp status,emp sector,income,age,daily usage,time saving,cost,reliability and safety,comfort,no.of veh,daily trip length,walking dist,time by go,time by come,money spend ,train,auto,bus,bike,ola-uber,personal vehicle,taxi,walking.

Yij= β0+ βi+ęij

Where,

β0-Intercept of dependent variable when all independant variables values consider zero.

And,

βi-coefficient of independent(predictor) variables.

ęij -Random error.

i=1,2,3,……………..26.(variables)

j=1,2,…………..,246.(Observations)

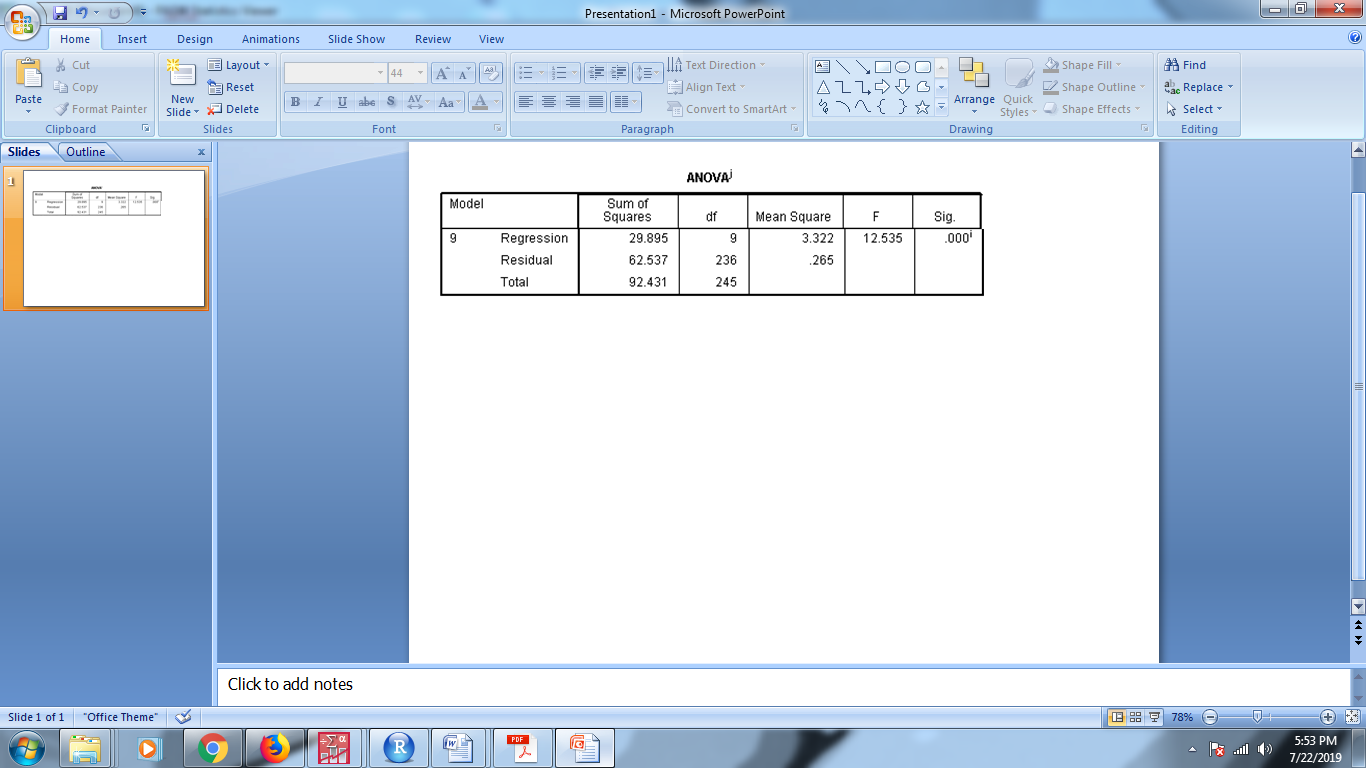
**Assumption:**

1. The population from which samples are drawn should be normally distributed Checking normality assumption is necessary, since the conclusion and interpretation is based on normality assumption. Here by using Kolmogorov-smirnov test and Q-Q plot we are going to check normality assumption.

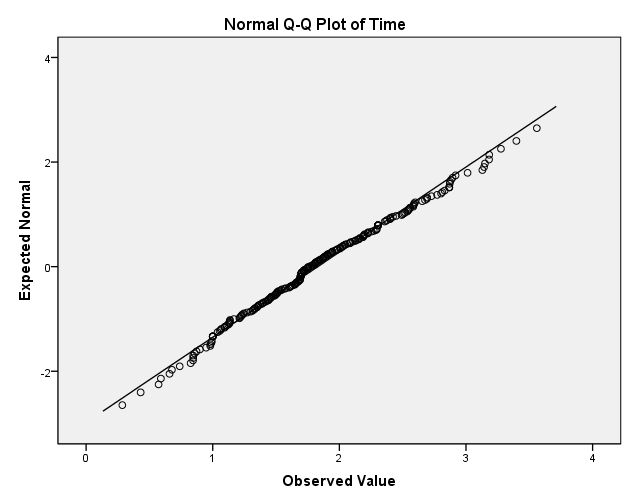
**Hypothesis for Kolmogorov-smirnov test**

Ho : Data is normally distributed.

H1: Data is not normally distributed.



| **Tests of Normality** |
| --- |



**Conclusion**: From above kolmogorov-smirnov test and Q-Q plot we can say that the data are normally distributed.

**2**. **Homogeneity of variance:**

Homogeneity means that the variance among the groups should be approximately equal.

For Testing Equality of Variances, Levene’s test can be used.

Ho: The error variance of the dependent variable is equal across groups.

H1: The error variance of the dependent variable is not equal for at least one across groups.

| **Test of Homogeneity of Variances** | | | |
| --- | --- | --- | --- |
| Time | | | |
| Levene Statistic | df1 | df2 | Sig. |
| 1.633 | 3 | 242 | .182 |

**Conclusion:** Since p-value is 0.182 i.e. greater than 0.05, the assumption of Homoscedasticity accepted.

**3** .**Independence of cases:**

The sample cases should be independent of each other. For checking independency we are using Durbin-Watson test. The test says that if the value lies between 1.5 to 2.5 then there is **NO-AUTOCORRELATION** in the observations.

|  |  |
| --- | --- |
| *Model* | *Durbin-Watson* |
| *1* | *1.770* |

**Conclusion:** Since the Durbin-Watson test statistic value is 1.770 which lies between 1.5 to 2.5. Hence our model is free from auto-correlation.

**4. Checking multicollinearity**

Model with collinear predictors can indicate how well the entire bundle of predictors predicts the [outcome variable](https://en.wikipedia.org/wiki/Dependent_variable#Use_in_statistics), but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others. Hence we go for this assumption.

|  |  |  |  |
| --- | --- | --- | --- |
| **1.** | **1.15** | **14.** | **1.496** |
| **2.** | **1.273** | **15.** | **1.273** |
| **3.** | **1.215** | **16.** | **1.536** |
| **4.** | **2.142** | **17.** | **1.312** |
| **5.** | **1.336** | **18.** | **1.767** |
| **6.** | **1.460** | **19.** | **1.310** |
| **7.** | **1.520** | **20.** | **1.334** |
| **8.** | **2.445** | **21.** | **1.281** |
| **9.** | **2.775** | **22.** | **1.241** |
| **10.** | **1.920** | **23.** | **1.181** |
| **11.** | **1.346** | **24.** | **1.197** |
| **12.** | **1.184** | **25.** | **1.321** |
| **13.** | **1.392** | **26.** | **1.432** |

VIF values for each variable is as follow,

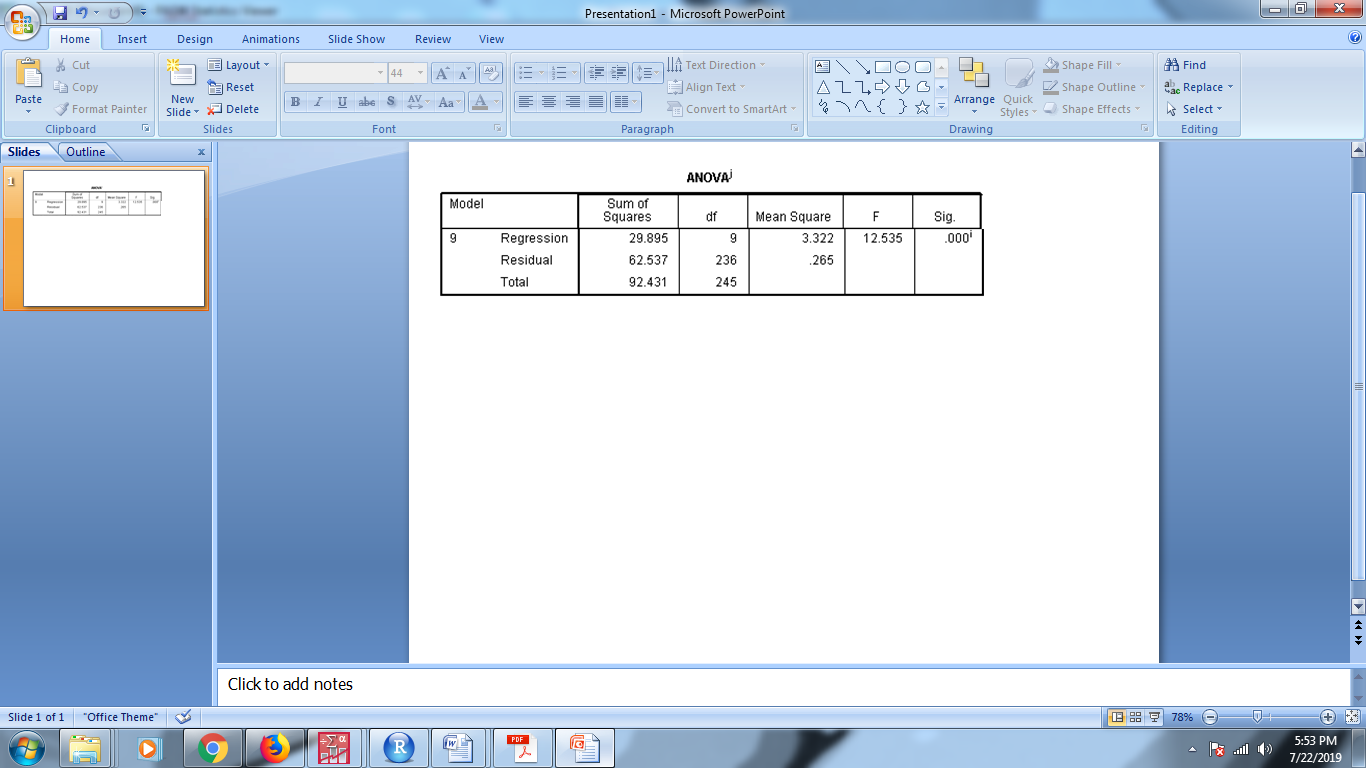
**Conclusion: VIF values for all independent variables are less than 5 therefore our model is free from *multicollinearity.***

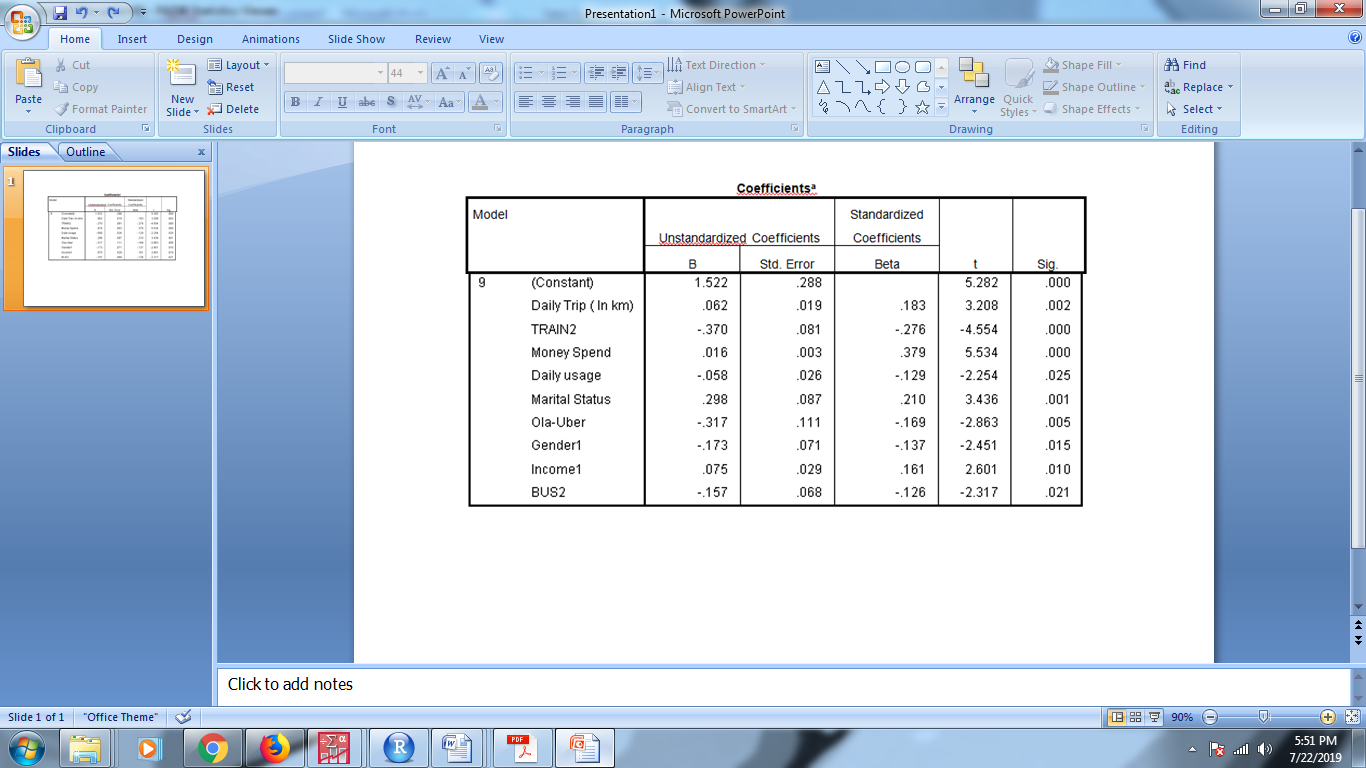
**Time:**

**Hypothesis:**

**Ho: Average means of independent variables are equal. (βi = 0)**

**H1: Average means of independent variables are not equal. (βi ≠ 0)**





| **Model Summary** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Model | | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| dimension0 | 1 | .321 | .103 | .099 | .5828736 |
| 2 | .397 | .158 | .151 | .5660379 |
| 3 | .474 | .225 | .215 | .5442057 |
| 4 | .496 | .246 | .234 | .5377481 |
| 5 | .512 | .262 | .247 | .5331482 |
| 6 | .526 | .276 | .258 | .5289784 |
| 7 | .540 | .292 | .271 | .5245129 |
| 8 | .555 | .308 | .285 | .5194930 |
| 9 | .569 | .323 | .298 | .5147681 |

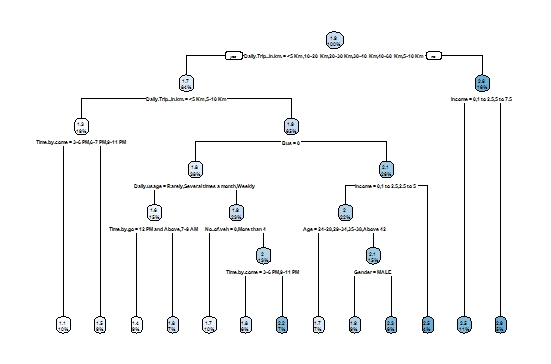
**Interpretation**: From above anova table we are getting above variables significant.

Hence the model we getting as,

**Spending Time**= 1.522+0.62\*Daily trip length-0.370\*Train + 0.16\* Money spend -0.58\*Dailyusage+0.298\*Maritalstatus-0.317\*ola-uber-0.173\*Gender1+0.075\*Income1-0.157\*Bus2.

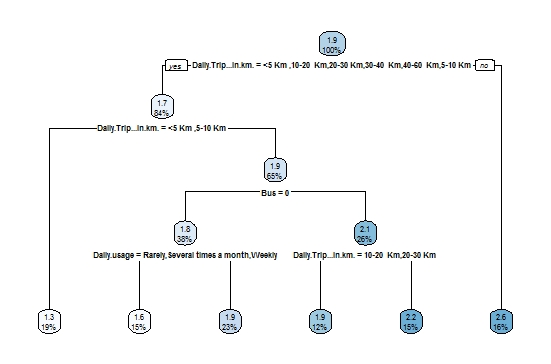
**Goodness of fit**:For the above model we are getting R2-square is 0.298 and it indicates 29.8 percent of variation of dependent variable is explain by independent variables presents in the model.

**REGRESSION TREE METHOD:**



**Decision Tree**

As we can see from the above tree, there are too many variables retained from the Full-Model tree for the classification of ‘Time’(dependant variable/continuous variable),on the basis of independent variables which are significant for the model, thereby tree being inefficient and difficult to understand, R square value being 0.047 for test set; where the tree takes into account all variables and observations with no conditions applied to its splitting and branches.

****

**Decision Tree (With Min-Split=40)**

In order to prevent over fitting and prune the tree, we tried a minimum split of 30, 40, 50 as the criteria for splitting, where the best tree with better accuracy and efficiency was found for min-split of 40 with R square value being 0.183 for test set, thereby making the tree more simpler and efficient to understand.

**Model Tree:** Daily Trip (Km) being at top, Bus (1-yes, 0-No), Daily usage is retained.

**ENSEMBLE MODELS AND THERE ACCURACY**

In order to reduce Bias-Variance error for the Tree Models, we have used ensemble methods involving group of predictive models to achieve a better accuracy and model stability for Bias-Variance error **trade-off management. The models and there outputs are as given in the table below.**

|  |  |
| --- | --- |
| **Model** | **R square(Test set)** |
| **Decision Tree** | 0.047 |
| **Decision Tree**  **(With MinSplit=40)** | 0.183 |
| **Bagging** | 0.257 |
| **Random Forest** | 0.228 |
| **Boosting**  **(ntree = 4500**  **shrink=0.001)** | 0.246 |

**INTERPRETATION:**

For the above objective we have conducted statistical analysis on time spent and choice of mode of transport with respect to socio demographic and geographical variables.

One-sample T-test for time spend during delay, conclude that the average delay time spent for transportation is 3.2277 Hrs, being H0 accepted.

From Analysis of Variance table we could conclude that income and daily trip(Km) found to be significant with respect to Time.

From Multiple regression method taking time as dependant variable we could conclude that Daily Trip(Km), Train, Money Spend, Daily usage, Marital Status, Ola-Uber, Gender, Income, Bus are found to be significantly related to time, R square value is 0.29.

For the same objective we have tried other models such as decision tree models along with ensembling models for comparing the results of each model where the data set was divided into train set and test set and the model was run on train set and accuracy was computed from the test set, thereby finding out which model is working much better in terms of giving accuracy with respect to time , being dependant in our case, in satisfying the objective.

The models such as Decision Tree with min-split 40 was giving R square value as 0.183 having the variables retained such as daily trip(Km), Bus, Daily usage.

Bagging technique showed highest R square value i.e. 0.257 among the ensembling models.

**OBJECTIVE-2**

1. **To identify and analyze the socio-demographic factors and choice of mode of transport, influencing the Money spend by the people for traveling.**

**ONE SAMPLE T-TEST:**

* **DELAY money spend**

Notation:

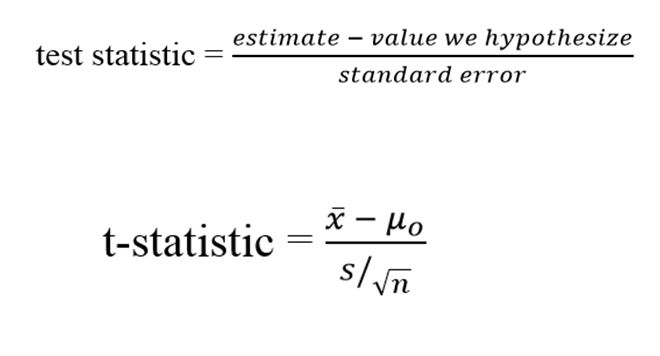
µ0: mean of delay money spend(In Hrs)

*Equal variances are assumed for this analysis.*

Hypothesis:

*Ho: delay money spend = 0(µ0=120.177) vs H1: delay money spend≠0 (µ0>120.177).*

*Test statistic:*



X bar= sample mean of delay money spend

S=sample variance

n=no. of observations.

Result:

**One-Sample T: Delay Money Spend**

Test of mu = 120.177 vs > 120.177

95% Lower

Variable N Mean StDev SE Mean Bound T P

Delay Money Spend 246 120.18 29.25 1.87 117.10 0.00 0.500

**Interpretation:**

For above one sample t test we are getting p value is greater than 0.05. Hence we accept H0 i.e., the average delay money for transportation is 120.177 Rs.

***ANALYSIS OF VARIANCE(MONEY):***

ANOVA is a statistical technique that assesses potential differences in a scale-level dependent variable by a nominal-level variable having 2 or more categories. An Analysis of Variance (one way ANOVA) willbe conducted to determine if there a significant difference on the dependent variable by independent variable. One way ANOVA is an appropriate statistical analysis when the purpose of research is to assess if mean differences exist on one continuous dependent variable by an independent variable with two or more discrete groups. In ANOVA dependent variable must be continuous (interval or ratio) level of measurement. Then independent variable In ANOVA must be categorical (nominal or ordinal) variable.

*In our model we have taken dependent variable* **MONEY** *and independent variables are* **LOCATION, GENDER, MARITAL STATUS, QUALIFICATION, EMPLOYMENT STATUS, EMPLOYMENT SECTOR, INCOME, AGE ,NO.OF VEHICLES and DAILY TRIP LENGTH.**

**Since the model is given as,**

**Y (ijkmnpqsr) = µ+Ai+Bj+Ck+Dm+En+Fp+Gq+Hs+Ir+Jl+ϵ(ijkmnpqsr).**

**Y= Money spend**

**µ=**General mean effect

**Ai** = Effect due to ith level of location i=1,2,3,4.

**Bj** = Effect due to jth level of gender j=1,2,3.

**Ck**= Effect due to kth level of marital status k=1,2.

**Dm**= Effect due to mth level of qualification m=1,2,3,4,5.

**En**= Effect due to nth level of employment status n=1,2,3,4.

**Fp**= Effect due to pth level of employment sector p=1,2,3,4.

**Gq**= Effect due to qth level of income q=1,2,3,4,5.

**Hs**= Effect due to sth level of age s=1,2,3,4,5,6,7.

**Ir**= Effect due to rth level of no.of vehicles r=1,2,3,4,5,6.

**Jl=** Effect due to lth level of daily trip-length l=1,2,3,4,5,6,7.

**(Part-1 ) Assumptions:**

**1**. The population from which samples are drawn should be normally distributed.

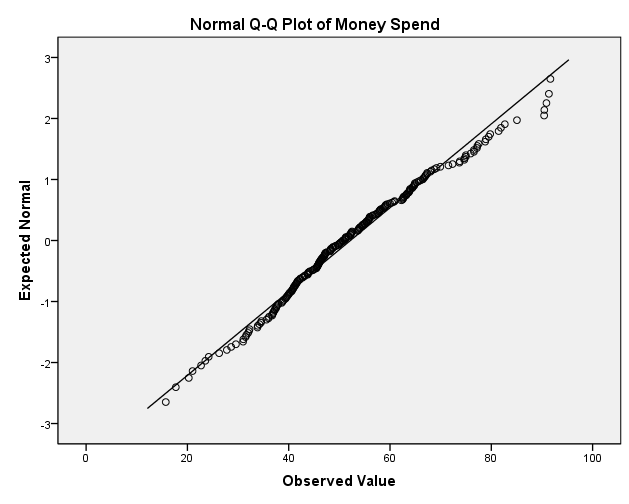
Checking normality assumption is necessary, since the conclusion and interpretation is based on normality assumption. Here by using Kolmogorov-smirnov test and Q-Q plot we are going to check normality assumption.

**Hypothesis for Kolmogorov-smirnov test**

Ho : Data is normally distributed.

H1: Data is not normally distributed.

| **Tests of Normality** | | | |
| --- | --- | --- | --- |
|  | Kolmogorov-Smirnova | | |
| Statistic | df | Sig. |
| Money Spend | .054 | 246 | .080 |
| a. Lilliefors Significance Correction | | | |

**

**Conclusion**: From above kolmogorov-smirnov test and Q-Q plot we can say that the data are normally distributed.

**2**. **Homogeneity of variance:**

Homogeneity means that the variance among the groups should be approximately equal.

For Testing Equality of Variances, Levene’s test can be used.

Ho: The error variance of the dependent variable is equal across groups.

H1: The error variance of the dependent variable is not equal for at least one across groups.

| **Test of Homogeneity of Variances** | | | |
| --- | --- | --- | --- |
| Money Spend | | | |
| Levene Statistic | df1 | df2 | Sig. |
| .410 | 3 | 242 | .746 |

**Conclusion:**

Since p-value is 0.746 i.e. greater than 0.05, the assumption of Homoscedasticity accepted.

**3** .**Independence of cases:**

The sample cases should be independent of each other. For checking independency we are using Durbin-Watson test. The test says that if the value lies between 1.5 to 2.5 then there is **NO-AUTOCORRELATION** in the observations.

|  |  |
| --- | --- |
| Model | Durbin-Watson |
| 1 | 2.056 |

**Conclusion:** Since the Durbin-Watson test statistic value is 2.056 which lies between 1.5 to 2.5. Hence our model is free from auto-correlation.

**4. Checking multicollinearity**

Model with collinear predictors can indicate how well the entire bundle of predictors predicts the [outcome variable](https://en.wikipedia.org/wiki/Dependent_variable#Use_in_statistics), but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others. Hence we go for this assumption,

|  |  |  |
| --- | --- | --- |
| ***Model*** | ***VIF*** | ***Condition Index*** |
| ***Constant*** |  |  |
| ***Location*** | 1.068 | 3.727 |
| ***Gender*** | 1.028 | 4.343 |
| ***Marital status*** | 1.966 | 6.070 |
| ***Qualification*** | 1.181 | 7.648 |
| ***Emp status*** | 1.339 | 8.946 |
| ***Emp sector*** | 1.369 | 10.162 |
| ***Income*** | 1.865 | 13.904 |

|  |  |  |
| --- | --- | --- |
| ***Age*** | 2.537 | 15.450 |
| ***No.of vehicles*** | 1.062 | 35.749 |

**Conclusion: VIF values for all independent variables are less than 5 therefore our model is free from *multicollinearity.***

***Part (2)***

***Ho:*** *Average effect due to the different levels of independent variables on spending money is not significant.*

***H1:*** *Average effect due to the different levels of independent variables on spending money is significant.*

| **Tests of Between-Subjects Effects** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Dependent Variable:Money Spend | | | | | |
| Source | Type III Sum of Squares | Df | Mean Square | F | Sig. |
| Corrected Model | 33402.911a | 36 | 927.859 | 10.466 | .000 |
| Intercept | 49732.925 | 1 | 49732.925 | 561.001 | .000 |
| Location | 472.987 | 3 | 157.662 | 1.778 | .152 |
| Gender | 2373.146 | 1 | 2373.146 | 26.770 | .000 |
| MaritalStatus | 1669.058 | 1 | 1669.058 | 18.827 | .000 |
| Qualification | 2296.080 | 4 | 574.020 | 6.475 | .000 |
| EmploymentStatus | 1637.313 | 3 | 545.771 | 6.156 | .000 |
| EmpSector | 1539.549 | 3 | 513.183 | 5.789 | .001 |
| Income | 2764.096 | 4 | 691.024 | 7.795 | .000 |
| Age | 4949.026 | 6 | 824.838 | 9.304 | .000 |
| No.ofveh | 2655.260 | 5 | 531.052 | 5.990 | .000 |
| DailyTriplengthkm | 6547.366 | 6 | 1091.228 | 12.309 | .000 |
| Error | 18527.927 | 209 | 88.650 |  |  |
| Total | 721615.752 | 246 |  |  |  |
| Corrected Total | 51930.838 | 245 |  |  |  |
| a. R Squared = .643 (Adjusted R Squared = .582) | | | | | |

**MULTIPLE REGRESSION METHOD:**

In our model we have consider, dependent variable as time spend for transportation. And independent variable location,gender,marital status,qualification,emp status,emp sector,income,age,daily usage,time saving,cost,reliability and safety,comfort,no.of veh,daily trip length,walking dist,time by go,time by come,money spend ,train,auto,bus,bike,ola-uber,personal vehicle,taxi,walking.

Yij=β0+βi+ęij

Where, β0-Intercept of dependent variable when all independant variables values consider as zero.

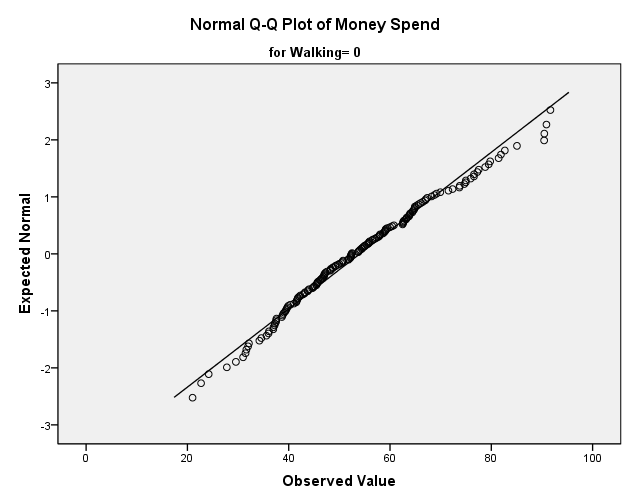
And, βi-coefficient of independent(predictor) variables.

ęij-Random error.

i=1,2,3,……………..26.(variables) j=1,2,…………..,246.(Observations)

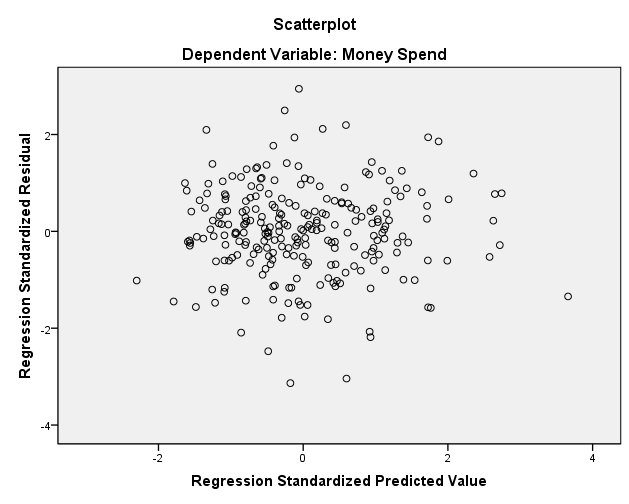
**Assumption**:

1. Normality:



**Conclusion:**From above Q-Q plot we can say that the data is normal.

1. homoschedasticity:



Conclusion from above scatter diagram,we conclude that data follows homoschedasticity.

1. Autocorrelation:

**MODEL SUMMARY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | R | R-Square | Adj R square | Std. Error of the Estimate | Durbin-Watson |
| 1 | .806a | .650 | .594 | 9.278510584 | 1.991 |

All the assumption are valid with our data hence now we can go for further analysis.

1. Multicollinearity:

VIF VALUES

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| variable | VIF | Variable | VIF | variable | VIF | variable | VIF | variable | VIF |
| 1 | 1.058 | 6 | 1.071 | 11 | 1.009 | 16 | 1.043 | 21 | 1.197 |
| 2 | 1.013 | 7 | 1.018 | 12 | 1.011 | 17 | 1.067 | 22 | 1.251 |
| 3 | 1.032 | 8 | 1.129 | 13 | 1.059 | 18 | 1.015 | 23 | 1.121 |
| 4 | 1.043 | 9 | 1.037 | 14 | 1.050 | 19 | 1.006 | 24 | 1.078 |
| 5 | 1.019 | 10 | 1.016 | 15 | 1.098 | 20 | 1.046 | 25 | 1.034 |

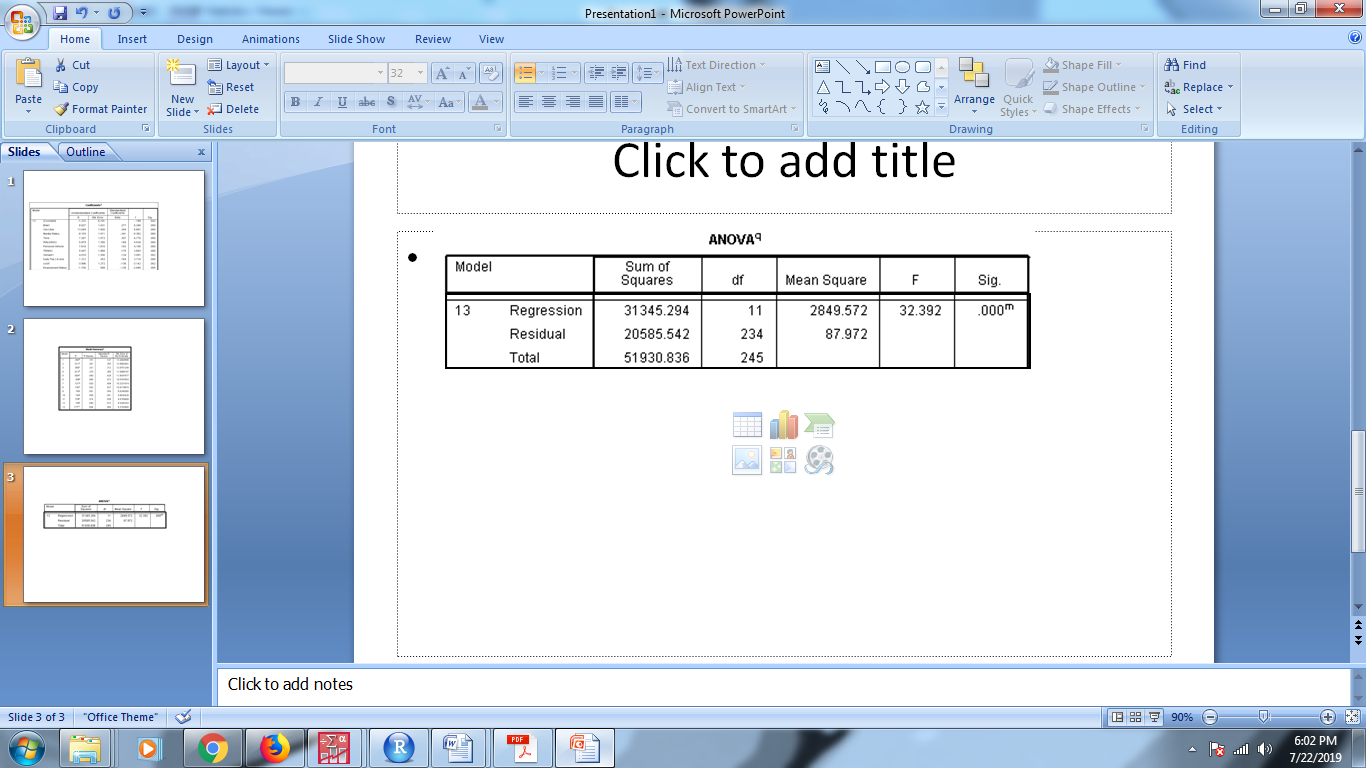
All the VIF values are less than 5 hence, the model is free from multicollinearity.

**we have checked all the assumption and all the assumtions are valid hence we are going for testing the hypothesis.**

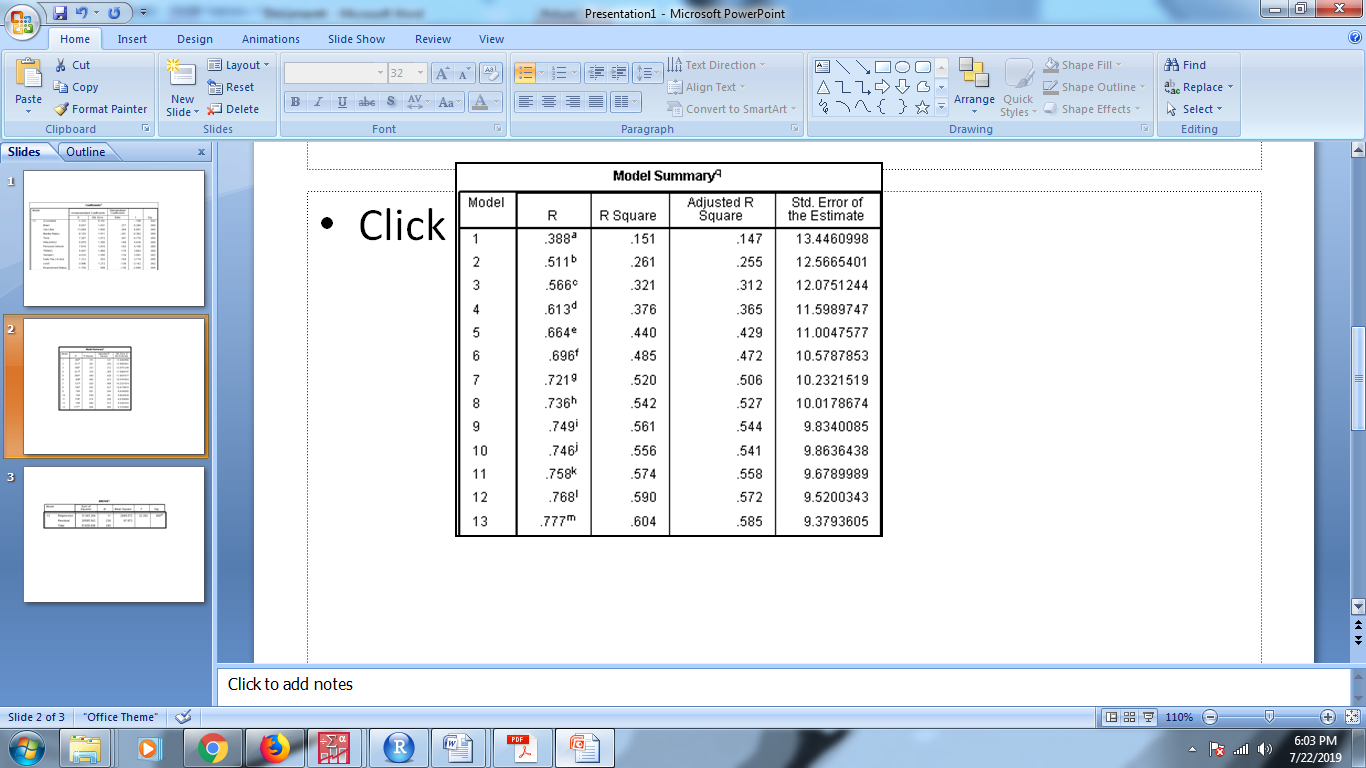
**Hypothesis:**

**Ho: Average means of independent variables are equal. (βi = 0)**

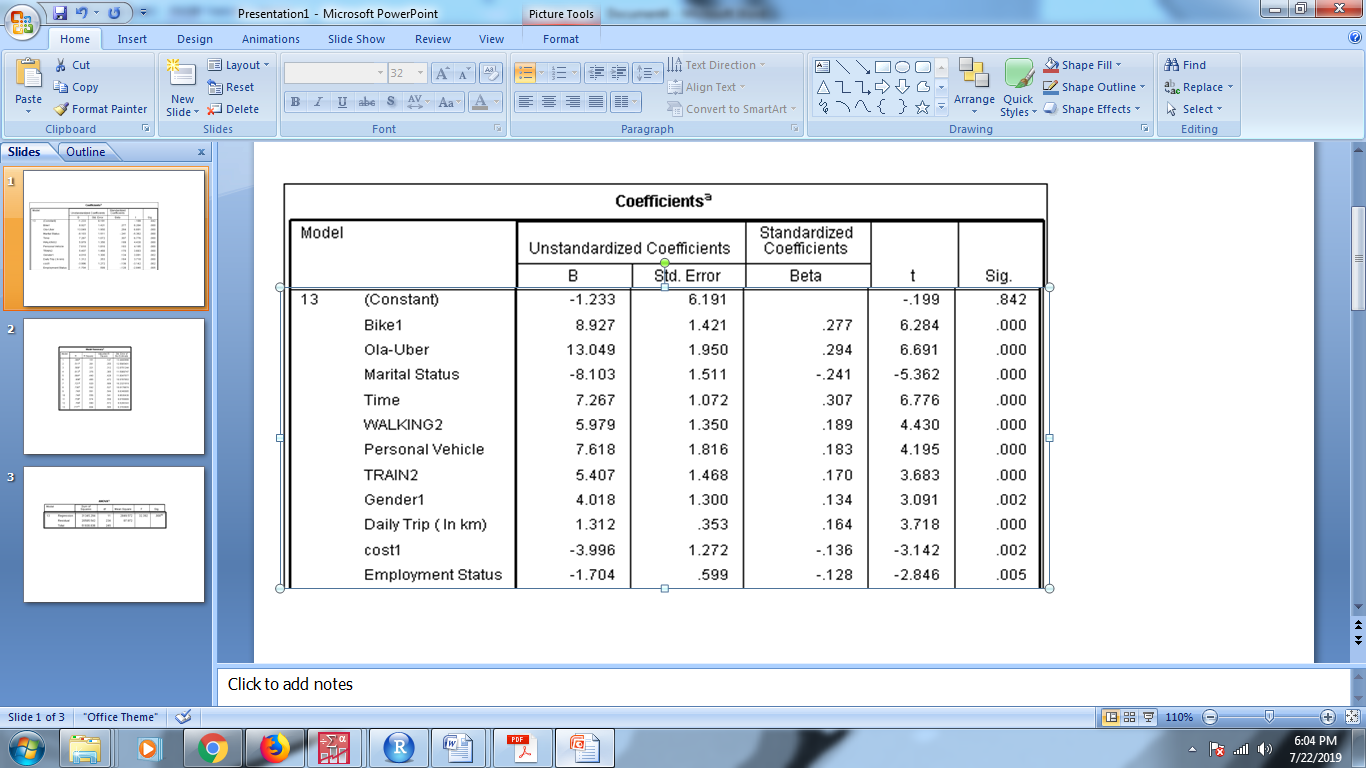
**H1: Average means of independent variables are not equal. (βi ≠ 0)**



Since p value is less than 0.05 model is significant.



R Square is 58.5%



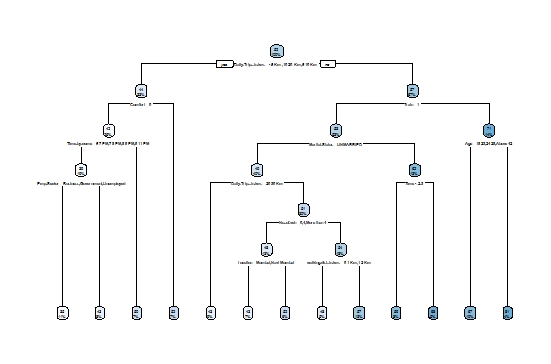
All variables are significant.

Hence the model is given as,

**Money**= -1.233+9.927\*bike(1) + 13.049\*ola-uber – 8.103\*marital\_status +7.267\*time + 5.979\*walking2 + 7.619\* personal\_vehicle + 5.407\*train2 + 4.018\*gender1 + 1.312\* dailytrip -3.996\*cost1 -1.702\*emp\_status

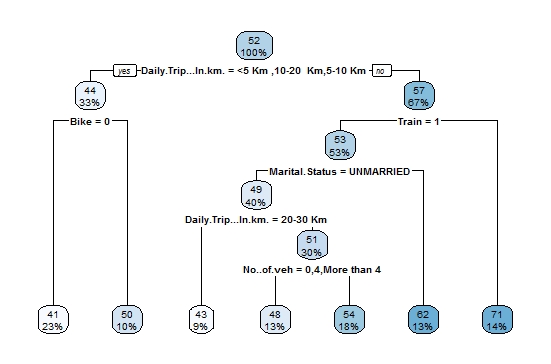
**Goodness of fit** – from the above model we are getting R-square is 0.585 which is good. hence in our model total 58.5% of variation is explained by the independent variable which are present in the model.

**REGRESSION TREE METHOD :**

****

**Decision Tree**

As we can see from the above tree, there are too many variables retained from the Full-Model tree for the classification of ‘Money’ (dependant variable/continuous variable),on the basis of independent variables which are significant for the model, thereby tree being inefficient and difficult to understand, R square value being 0.070 for test set; where the tree takes into account all variables and observations with no conditions applied to its splitting and branches.

****

**Decision Tree(With MinSplit=40)**

In order to prevent over fitting and prune the tree, we tried a minimum split of 30, 40, 50 as the criteria for splitting, where the best tree with better accuracy and efficiency was found for min-split of 40 with R square value being 0.142 for test set, thereby making the tree more simpler and efficient to understand.

**Model Tree:** Daily Trip (Km), Bike (1-Yes,0-No),Train(1-Yes,0-No),Marital Status, No. of Vehicles.

**ENSEMBLE MODELS AND THERE ACCURACY**

In order to reduce Bias-Variance error for the Tree Models, we have used ensemble methods involving group of predictive models to achieve a better accuracy and model stability for Bias-Variance error **trade-off management. The models and there outputs are as given in the table below.**

|  |  |
| --- | --- |
| **Model** | **R square(Test set)** |
| **Decision Tree** | 0.070 |
| **Decision Tree**  **(With MinSplit=40)** | 0.142 |
| **Bagging** | 0.469 |
| **Random Forest** | 0.464 |
| **Boosting**  **(ntree = 4500**  **shrink=0.001)** | 0.55 |

**INTERPRETATION:**

For the above objective we have conducted statistical analysis on Money spent and choice of mode of transport with respect to socio demographic and geographical variables.

One-sample T-test for Money spend during delay, conclude that the average delay time spent for transportation is found to be insignificant, being H0 accepted.

From Analysis of Variance table we could conclude that gender, marital status, qualification, employment status, employment sector, age, no. of vehicle, daily trip length(km) found to be significant with respect to Money.

From Multiple regressions method taking Money as dependant variable, the where variable retained are bike, Ola-Uber, Marital Status, Time, Walking, Personal Vehicle, Train, Gender, Daily Trip(Km), cost and employment status thereby R square value found to be 0.585.

For the same objective we have tried other models such as decision tree models along with ensembling models for comparing the results of each model where the data set was divided into train set and test set and the model was run on train set and accuracy was computed from the test set, thereby finding out which model is working much better in terms of giving accuracy with respect to Money , being dependant in our case, in satisfying the objective.

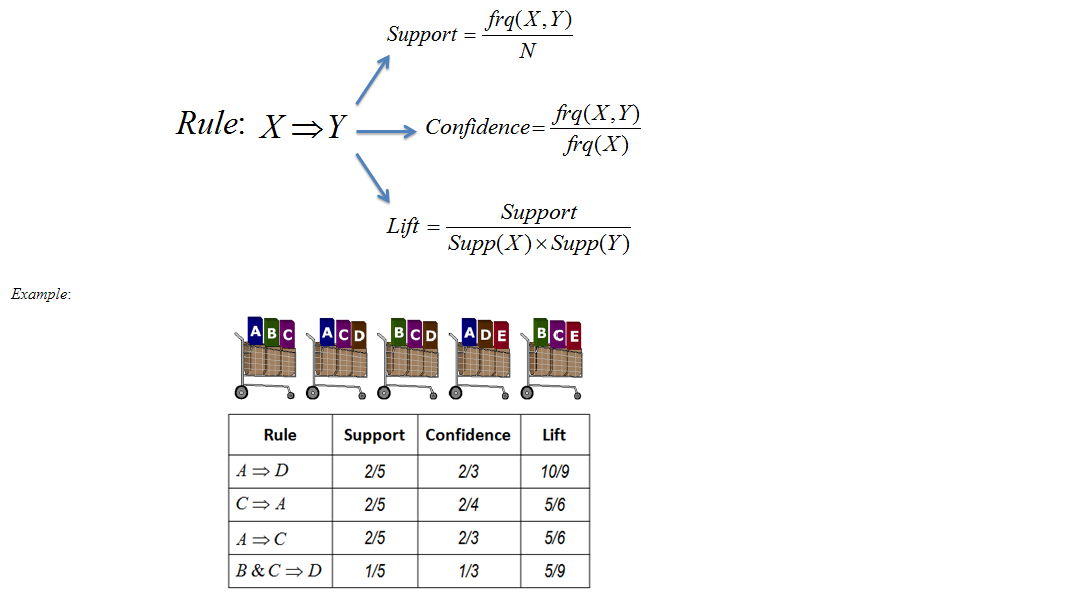
The models such as Decision Tree with min-split 40 was giving R square value as 0.142 having the variables retained such as daily trip(Km), Bus, Daily usage.

Boosting technique showed highest R square value i.e. 0.55 among the ensembling models.

**OBJECTIVE-3**

* **Identify association among Train, Bus, Auto and reason for specific choice of a particular mode of transport.**

**Market Basket Analysis — Association Rules**



Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy.

**Association Rules** are widely used to analyze retail basket or transaction data, and are intended to identify strong rules discovered in transaction data using measures of interestingness, based on the concept of strong rules.

**An example of Association Rules**

* Assume there are 100 customers
* 10 of them bought milk, 8 bought butter and 6 bought both of them.
* bought milk => bought butter
* support = P(Milk & Butter) = 6/100 = 0.06
* confidence = support/P(Butter) = 0.06/0.08 = 0.75
* lift = confidence/P(Milk) = 0.75/0.10 = 7.5

Note: this example is extremely small. In practice, a rule needs the support of several hundred transactions, before it can be considered statistically significant, and datasets often contain thousands or millions of transactions.

**Goal of Association Rule Mining:**

When you apply Association Rule Mining on a given set of transactions T your goal will be to find all rules with:

1. Support greater than or equal to min\_support
2. Confidence greater than or equal to min\_confidence

**APRIORI Algorithm**

In this part of the tutorial, you will learn about the algorithm that will be running behind R libraries for Market Basket Analysis. This will help you understand your clients more and perform analysis with more attention. If you already know about the APRIORI algorithm and how it works, you can get to the [coding part](https://www.datacamp.com/community/tutorials/market-basket-analysis-r#code).

Association Rule Mining is viewed as a two-step approach:

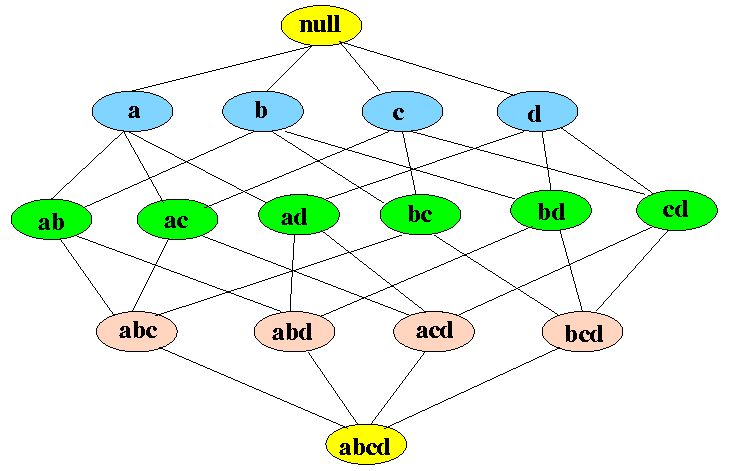
1. **Frequent Itemset Generation:** Find all frequent item-sets with support >= pre-determined min\_support count
2. **Rule Generation:** List all Association Rules from frequent item-sets. Calculate Support and Confidence for all rules. Prune rules that fail min\_support and min\_confidence thresholds.

Frequent Itemset Generation is the most computationally expensive step because it requires a full database scan. Among the above steps, Frequent Item-set generation is the most costly in terms of computation.

Above you have seen the example of only 5 transactions, but in real-world transaction data for retail can exceed up to GB s and TBs of data for which an optimized algorithm is needed to prune out Item-sets that will not help in later steps. For this APRIORI Algorithm is used. It states:

*Any subset of a frequent itemset must also be frequent. In other words, No superset of an infrequent itemset must be generated or tested*

It is represented in Itemset Lattice which is a graphical representation of the APRIORI algorithm principle. It consists of k-item-set node and relation of subsets of that k-item-set.



You can see in above figure that in the bottom is all the items in the transaction data and then you start moving upwards creating subsets till the null set. For *d* number of items size of the lattice will become 2d2d. This shows how difficult it will be to generate Frequent Item-set by finding support for each combination. The following figure shows how much APRIORI helps to reduce the number of sets to be generated:



If item-set *{a,b}* is infrequent then we do not need to take into account all its super-sets. Let's understand this by an example. In the following example, you will see why APRIORI is an effective algorithm and also generate strong association rules step by step. Follow along on with your notebook and pen!

As you can see, you start by creating *Candidate List* for the 1-itemset that will include all the items, which are present in the transaction data, individually. Considering retail transaction data rom real-world, you can see how expensive this candidate generation is. Here APRIORI plays its role and helps reduce the number of the Candidate list, and useful rules are generated at the end. In the following steps, you will see how we reach the end of Frequent Itemset generation, that is the first step of Association rule mining.

Your next step will be to list all frequent itemsets. You will take the last non-empty Frequent Itemset, which in this example is L2={I1, I2},{I2, I3}. Then make all non-empty subsets of the item-sets present in that Frequent Item-set List. Follow along as shown in below illustration:

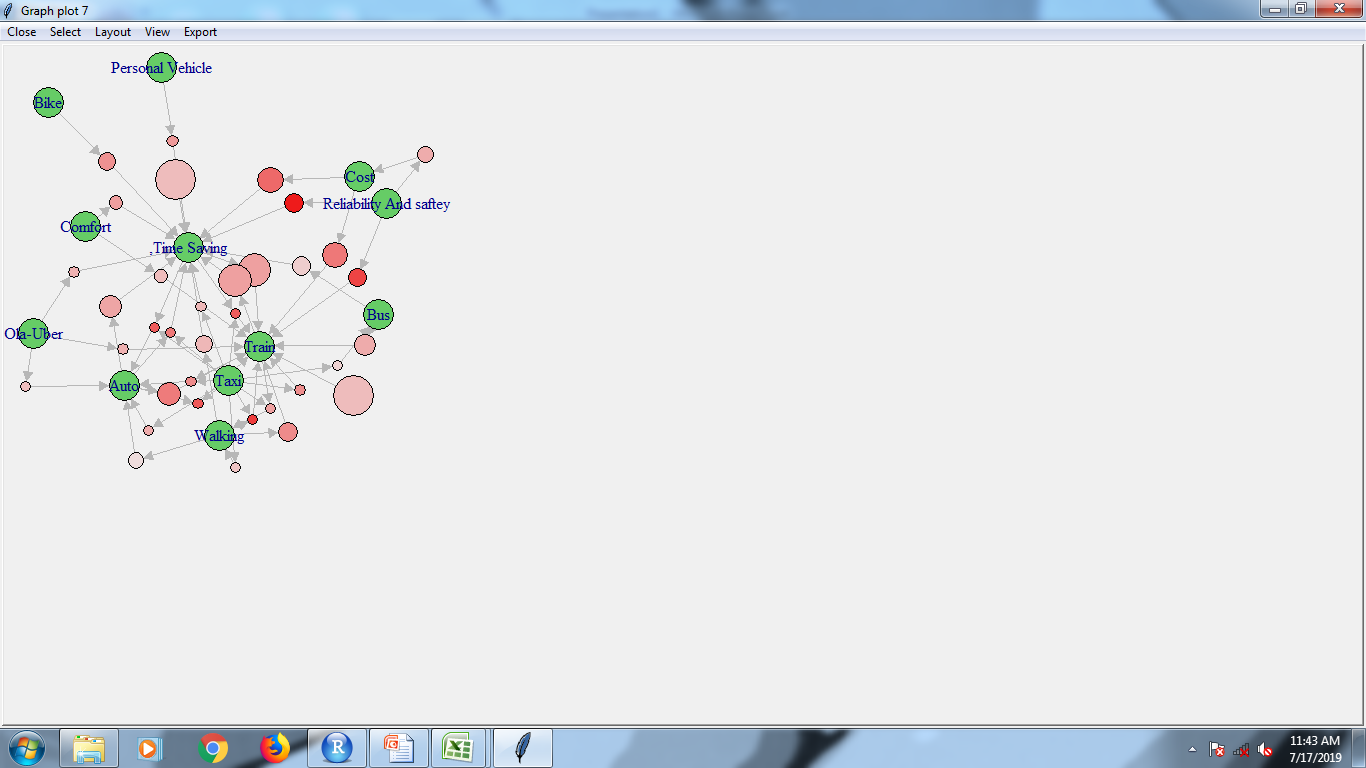
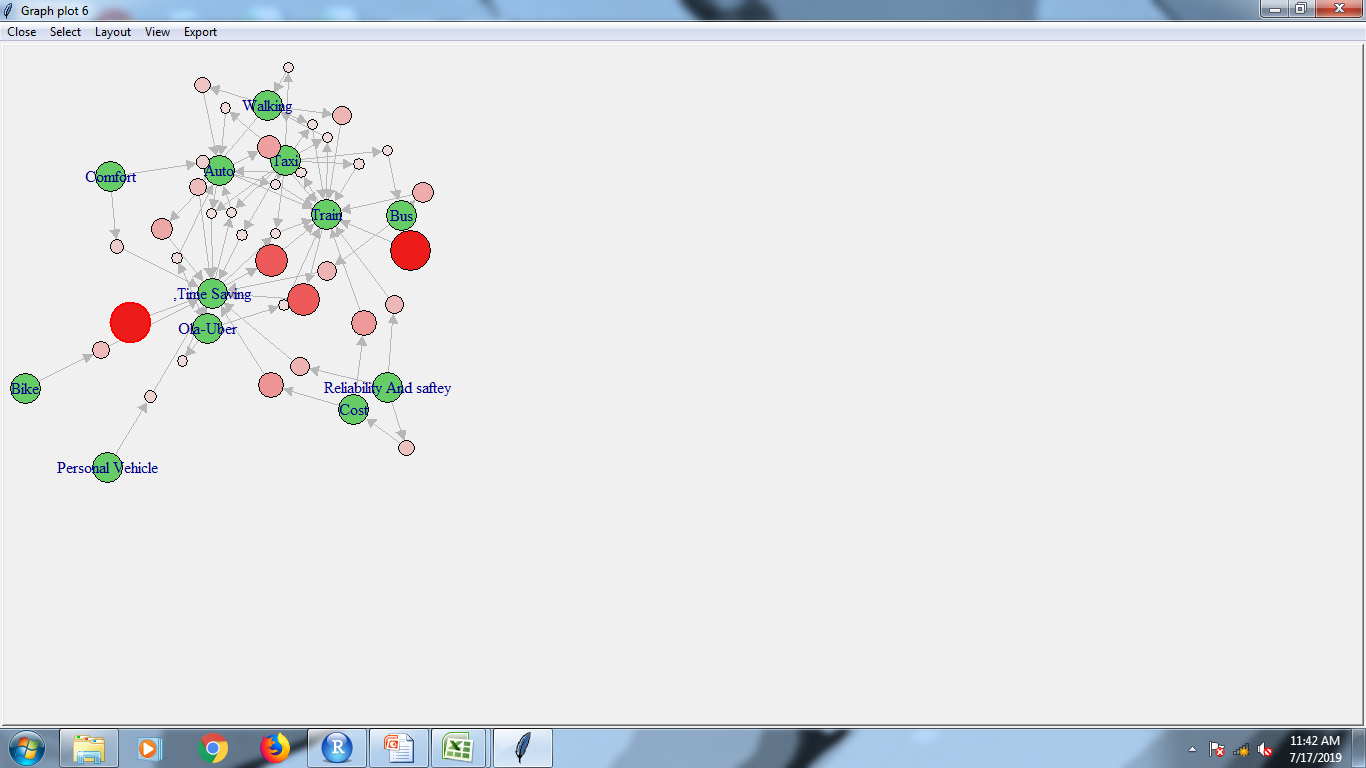
You can see above there are four strong rules. For example, take I2=>I3I2=>I3 having confidence equal to *75%* tells that 75% of people who bought I2 also bought I3.

**OBJECTIVE-3**

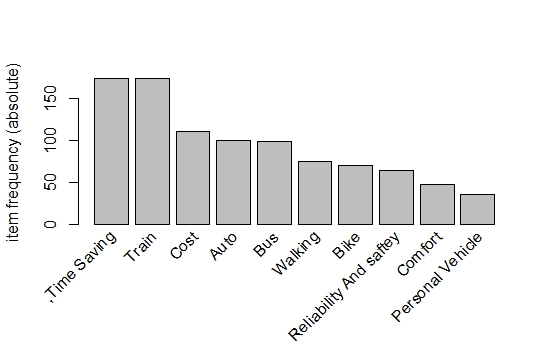
* **Identify association among Train, Bus, Auto and reason for specific choice of a particular mode of transport.**

**Association rule of Market Basket Analysis(APRIORI):**

We have find out the assoiciation between the modes of transport and reason of choice by using APRIORI algorithm for Market Basket Analysis (Association Rule). The results are found as under.

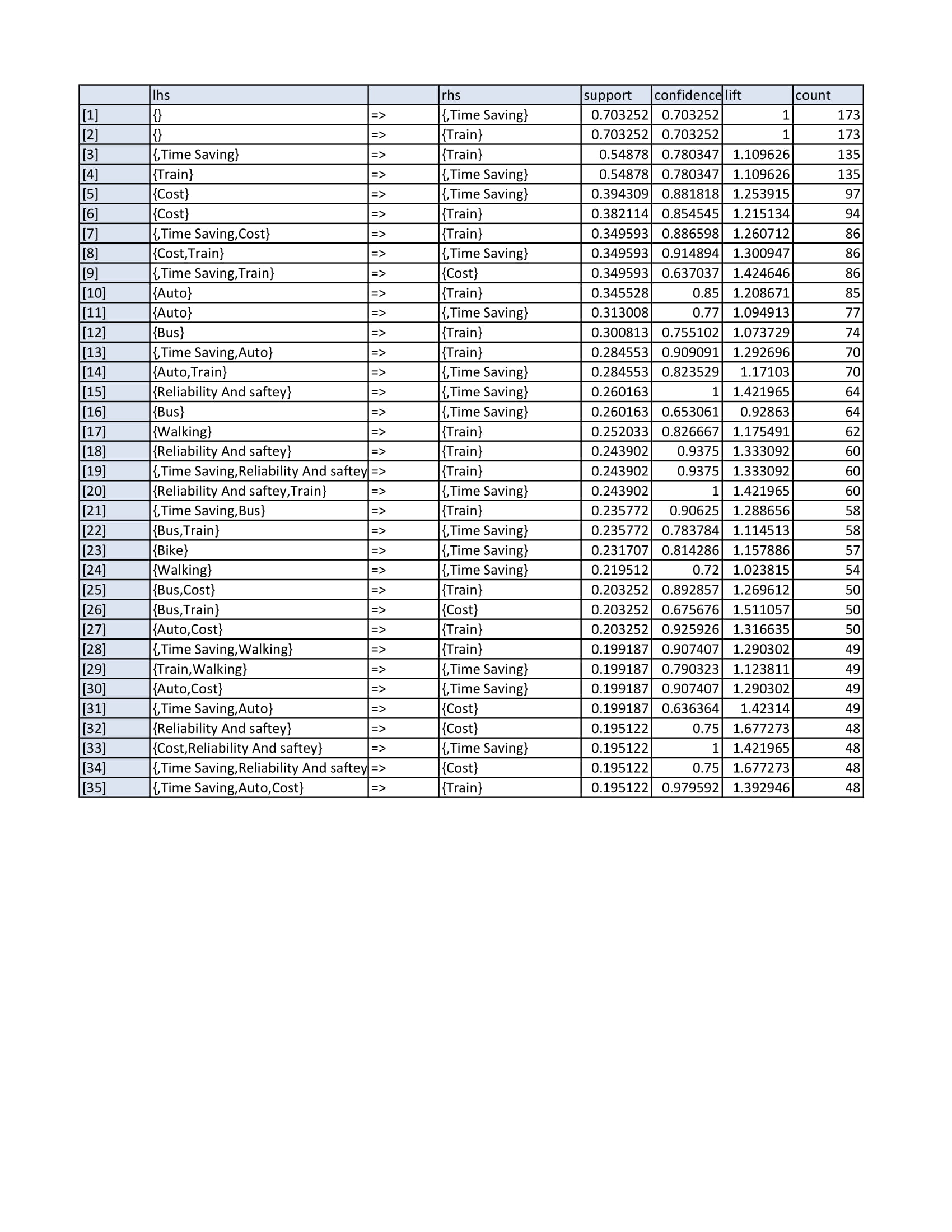
 

**Fig. By Confidence(Reachability Plot)** **Fig. By Support(Reachability Plot)**



**Fig. Item Frequency Plot(Variables)**

**Table : Inspected Rules** -(Sorted by Support)



**INTERPRETATION:**

From the above table we could inspect that:

* Train is used mostly for time saving and cost saving with respective support and confidence being significantly large as compare to other associated variables.
* Auto is used mostly for time saving with respective support and confidence being significantly large as compare to other associated variables.
* Auto is used provided the person uses Train is mostly for time saving with respective support and confidence being significantly large as compare to other associated variables.
* Bus and Auto is used provided the person uses Train is very similar with almost equal support and confidence approximately.

Being the people using **Bus & Auto** is used provided the person uses **Train** is very similar but then **Bus** is used for **Cost Saving** whereas Auto is used for **Time Saving** for traveling. So we could see that people are using **Auto** more than **Bus** for **time saving** being **cost higher** just to save **Time** **but not Cost** whereas for the same time of travel they aren’t using **Bus** for which traveling **cost is cheaper then Auto**.

**Objective 4**

* **Statistical analysis of factors increasing the stress level& causing inconvenience to the population.**

1. PARETO Analysis for the factors increasing the irritation level to the population while traveling.
2. PARETO Analysis for the factors causing inconvenience to the population while traveling.

**PARETO ANALYSIS**

Pareto analysis is a statistical technique in decision making used for the Selection of limited number of tasks that produce significant overall effect. It uses the Pareto principle (also known as the 80/20 rule) the idea that by Doing 20% of work you can generate 80% of benefit of doing the entire job. Take quality improvement, for example, a vast majority of problems (80%) Are produced by a few causes (20%).This technique is also called as the “Vital Few and Trivial Many”.

We can apply 80-20 rule to almost anything.

- 80% of customer’s complaints arise from 20% of your product and

Services.

- 20% of your product and services account for 80% of your profit.

**PROCEDURE:**

* Create a vertical bar chart with causes on X-axis and frequency on Y-axis.
* Arrange the bar chart in descending order of cause importance that is, that cause with the highest frequency comes first.
* Calculate the cumulative frequency for each cause in descending order.
* Create the cumulative frequency percentage for each cause in Descending order.
* Create a second Y-axis with percentages descending in increments Of 10 from 100% - 0%.
* Plot the cumulative frequency percentage for each cause on X-axis.
* Join the points to form a curve.
* Draw a line at 80% on Y-axis running parallel to X-axis. Then drop the line at point of intersection with the curve on X-axis.Thispoint on X-axis separates the important causes on left from less important causes on right.

**OBJECTIVE-4**

1. **PARETO Analysis for the factors increasing the irritation level to the population while traveling.**

**Pareto used for factors affect or irritate most while travelling:**

|  |  |
| --- | --- |
| **FACTORS** | **RESPONSE** |
| Traffic | 2141 |
| Rush/Crowd | 2141 |
| Delay in running | 1838 |
| Noise Level | 1588 |
| Anxiety | 1200 |
| Limited Space | 1380 |
| Travel Expense | 882 |
| Delay in time | 1070 |
| Long Distance | 745 |
| People Behaviour | 545 |



**CONCLUSION:** Our analysis shows that, factors affect or irritate most while travelling are Traffic, Rush/crowd, Delay in running, Noise level,anxiety,limited space and travel expense.

**OBJECTIVE-4**

1. **PARETO Analysis for the factors causing inconvenience to the population while traveling.**

**Pareto used for factors responsible for increasing road traffic:**

|  |  |
| --- | --- |
| **FACTORS** | **RESPONSE** |
| Hawkers | 1925 |
| Widht of road | 2002 |
| Illegal Parking | 1989 |
| Current working of Mono & Metro | 1770 |
| People not follow traffic rules | 1577 |
| Heavy vehicles | 1235 |
| Rash Driving | 1042 |
| Increaseing no.of pvt vehicles | 1124 |
| Automated signals | 548 |
| Road accidents | 318 |
|  |  |



**Conclusion:** Our analysis shows that, factors responsible for increasing road traffic are Hawkers, width of road, illegal parking, current working of mono/metro consrtuction,people not follow traffic rules, heavy vehicles, and rash driving.

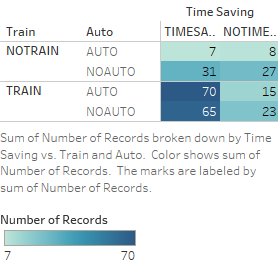
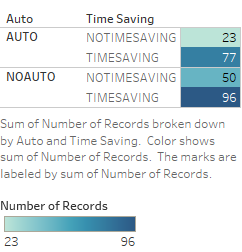
**OBJECTIVE-5**

* **Identify ways to improve the public transport.**

**SOFTWARE:**TABLEAU

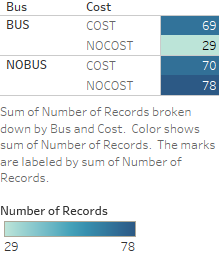
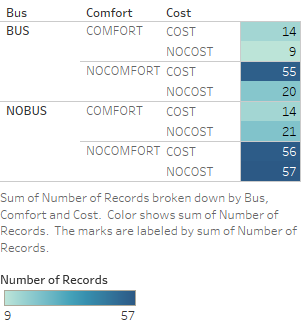
**CORSS TABULATIONS:**

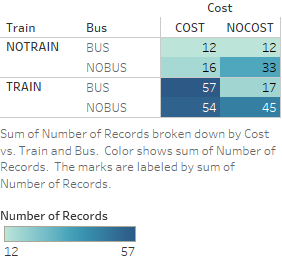
**TRAIN AND AUTO:**

** **

* Above cross tabulation analysis of Train/Auto with Time saving as well as Auto and Time saving suggests that people use Auto in relation to Trains to save Time.

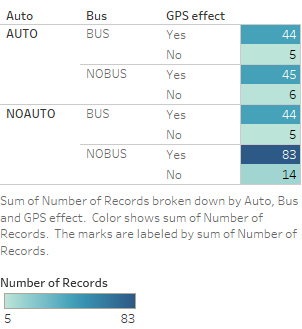
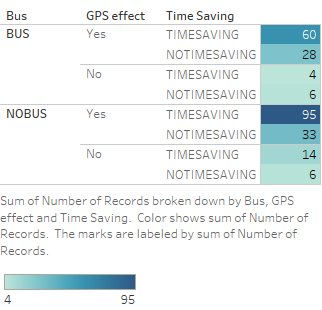
**TRAIN AND BUS:**

** **

****

* Above analysis suggests that Buses seems to be uncomforatble but used for cost saving options as compared to Train and Auto for daily commuters.
* According to the above analysis which relates people travelling by Train and Bus taking cost effect into account, it becomes less costly choice as compared to Train and Auto combination for everyday commuters.

**GPS EFFECT**

** **

* Above analysis suggests that people who do not travel by Bus and Auto seems to be very positive using GPS system in Buses.
* Using GPS system in Buses would help attract daily commuetrs to use Buses instead of their own vehicles which in turn may help to reduce the traffic and stress level of passengers.
* Bus in relation to time savings suggests that GPS will help to save time.

**Interpretation:**

* From the above analysis it is evident that Auto and Train combination works poorly for daily commuters in terms of cost saving but works best for time saving.
* Whereas Bus and Train combination adds to cost saving (decrease in cost) , but troublesome commute of daily passengers in terms of comfort and time saving.
* Having GPS systems in place is certainly going to help in comfortable journey as well as increase in Time saving for BEST Buses, where cost is already seen to be much lesser as compared to Auto for daily commuters.

**FINAL CONCLUSION**

The focus of the project was not only confined to one or two aspects of public transport but also on multiple factors which influence its use and areas of concern. Since our aim was to find out the challenges faced by people travelling by public transport in Mumbai, we ran statistical analysis to find out the major factors and concerns surrounding it. Thus, our project was divided into five objectives assessing the factors influencing the use of different modes of transport and how we can strategize them effectively to help make daily commute easier for everyone.

• Objective one was focused on identifying and analyzing sociodemographic factors influencing the preference of public transport. We used logistic regression, Decision Tree models and other assembling models to interpret the final accuracy of the model as a whole is retained, respectively. For Train we could retain in common few variables which are related are daily usage and time saving and cost. For Auto we could retain in common few variables which are related are walking, gender, time saving and train. For Bus we could retain in common few variables which are related are Cost, Daily Trip(10-40Km), Reliability and safety.

• Objective two was to analyze the time spend, money spend and its relationship with the different socio-demographic and geographical variables. This analysis helped us to understand the way people travel and use the transport system one after the other with respect to time and money. The variables retained as related to time spend are Daily Trip(Km) Bus, Train, money spend, Marital Status and so on whereas the variables retained as related to Money spend are Bike, Ola-Uber, Marital Status, Time, Walking, Personal Vehicle, Train Gender, Daily Trip(Km),Cost and Employment status.

• Objective three was to indentify association among Train, Bus, Auto and Reason for specific choice of a particular mode of transport, and we could conclude that Commuters using Bus and Auto, provided the Commuter uses Train, is very similar but then Bus is used for Cost Saving, whereas Auto is used for Time Saving for traveling. So we could see that commuters who are using public transport are using Auto more than Bus for time saving; being more costly, just to save Time but not cost whereas for the same time of travel they aren’t using Bus for which traveling cost is cheaper than Auto, due to factors responsible for inconvenience such as Increase in traveling hours, Uncomfortable for traveling and so on.

• In Objective four we wanted to analyze the factors which cause inconvenience or increase the stress level of people travelling by public transport. We used simple bar charts to represent the frequency distribution and Pareto analysis to identify top factors which cause inconvenience or increase the stress level of commuters by 80%. We found out that Traffic, Rush and people behavior increase the stress level in commuters whereas factors such as metro construction work causes maximum inconvenience in daily commute.

• Our final objective was to identify major concerns among people travelling by public transport and identify ways/recommendations to improve the efficiency and comfort while keeping the money spent and time taken to lowest. We have found the 80% people agree that a GPS system in BEST buses could improve the efficiency and reduce the waiting time which in turn will reduce the stress level of daily commuters.

So, after reviewing the analysis and objectives of our project, we feel we have found the requisite factors which affect the choice of mode of transport, major concerns and areas of improvement.

**RECOMMENDATIONS**

It is fast becoming necessary to persuade personal motor vehicle users to shift to public transport to mitigate the negative impacts of road congestion, deteriorating air quality and increasing carbon emissions. Such a shift can happen only if public transport offers the conveniences that personal vehicles allow—namely, on-demand availability, door-to-door connectivity, safety and comfort.

* Today’s public transport systems, which follow fixed routes and schedules, can’t offer such conveniences. Poor quality and shabby looking vehicles are not inviting to commuters. Inadequate capacity has led to overcrowding, making them unsafe and uncomfortable for women, senior citizens and children. Under such circumstances, a shift in preference is unlikely.
* Cities need to increase the number of public transport vehicles significantly to ensure safe, comfortable, frequent and crowd-free commutes to all.
* It is time for the government to widen the definition of public transport to include small buses, vans and pooled vehicles that offer on-demand services. Ongoing studies indicate that bus aggregator systems—a model that uses technology (mobile apps) to allow passengers to book seats in buses operating on routes within city limits, pay fares online and track location—have managed to pull people out of their private vehicles and bring about a modal shift.

**SOFTWARE USED:**

* R
* SPSS
* MINITAB
* TABLEAU (for visualization)
* MS OFFICE

**BIBLIOGRAPHY**

**Books:**

1. Fundamental of Mathematical Statistics by Kapoor and Gupta
2. Fundamental of Applied Statistics by Kapoor and Gupta
3. Decision Trees and Random Forest by Chirs Smith
4. Krzanowski, W.J., and F.H.C Marriott Multivariate analysis.London Edward Arnold, 1994. Print.
5. Cody, Ronald P., and Jeffrey K. Smith. Applied statistics.
6. Hosmer, David W., Stanley Lemeshow, and Rodney X. Studivant. Applied logistic regression. Hoboken, NJ: Wiley, 2013. Print.

**Websites:**

1. <https://en.wikipedia.org/wiki/Public_transport_in_Mumbai>
2. <https://www.hindustantimes.com/topic/transport-in-mumbai>
3. <https://www.internations.org/mumbai-expats/guide/living-in-mumbai-15323/transportation-in-mumbai-2>
4. <https://indiasomeday.com/en/your-guide-to-mumbais-public-transport-system/>
5. <https://www.aboutmumbai.com/Transport.asp>
6. <https://www.slideshare.net/aziz_khan/public-transport-problems-in-mumbai>
7. <https://www.internations.org/mumbai-expats/guide/living-in-mumbai-15323/transportation-in-mumbai-2>
8. <https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/>

1. https://www.hindustantimes.com/mumbai-news/the-real-problem-within-best-s-mess/story-VMn9PBaTFrIDmwSWg04vSL.html [↑](#footnote-ref-1)