A

PROJECT REPORT ENTITLED

ON

“A Statistical Analysis of Carbon Footprint and Carbon Sequestration in Kolhapur City.”

A PROJECT REPORT SUBMITTED TO

DEPARTMENT OF STATISTICS

SHIVAJI UNIVERSITY

KOLHAPUR.

FOR THE PARTIAL FULFILLMENT OF THE DEGREE

M.Sc. Statistics (Part II)

PRESENTED BY,

Miss. Raut Aishwarya Vitthal.

Miss. Pandhare Rutuja Sudhakar.

Miss. Kurade Jyoti Suresh.

Under the guidance of

Prof. Mr. S. V. Rajguru

CERTIFICATE

This is to certify that the project entitled **"A Statistical Analysis of Carbon Footprint and Carbon Sequestration in Kolhapur City."** as partial fulfilment for the award of the degree of M.Sc. in statistics of Shivaji University, Kolhapur, is a record of bonafide work carried out by them under my supervision and guidance. To the best of my knowledge, the matter presented in the project has not been submitted earlier.

This project is submitted by:

Miss. Raut Aishwarya Vitthal.

Miss. Pandhare Rutuja Sudhakar.

Miss. Kurade Jyoti Suresh.

Place: Kolhapur	Prof. Mr. S.V.Rajguru	Prof. Dr.S.B. Mahadik
Date:	Project Guide	Head of Department,
		Department of Statistics,
		Shivaji University, Kolhapur

ACKNOWLEDGEMENT

We wish to thank the Department of Statistics (Shivaji University, Kolhapur) for giving us an opportunity to do a project.

This report has been prepared under the guidance of Prof. Mr. S.V.Rajguru Sir. We would like to express our profound gratitude towards him for his guidance we constructive throughout this project.

Also, we would like to thank Dr. S. B. Mahadik, Dr. H. V. Kulkarni, Dr. S. D. Pawar and Prof. Mr. S.K. Ganjave, Prof. S.M. Patil, Prof. R.S. Salamwade for their support, suggestions and guidance for this project.

Finally we would like to thank HOD, teachers, non-teaching staff and students of departments for their valuable co-operation in this project.

CONTENTS

Sr. No.	Contents	Pages
1	Introduction	5
2	Objectives	10
3	Data Collection Methodology	11
4	Statistical tools and software's	14
5	Graphical Representation of Carbon Footprint	15
6	Model building	19
7	Graphical Representation of Carbon Sequestration	25
8	Statistical Analysis	29
9	Scope and Limitations	32
10	Major Findings	33
11	References	34
12	Appendix	35

INTRODUCTION

CARBON FOOTPRINT

In last few decades the increase in greenhouse gases (GHG) in the atmosphere has led to climate change and its serious consequences. A **Carbon Footprint** is the total GHG emission causes by an individual, event, organization, services, place or product expressed as carbon dioxide equivalent (CO_2e). Greenhouse gases including the carbon containing gases carbon dioxide and methane can be emitted through the burning of fossil fuels, land clearance and the production and consumption of food, materials, wood, roads, buildings, transportation and other services.

Human activities are one of the main causes of greenhouse gas emission. GHG are gasses that increase the temperature of the earth due to their absorption of infrared radiation. These gases are emitted from fossil fuel usage in electricity, in heat and transportation, as well as being emitted as by product of manufacturing. The most common GHGs are carbon dioxide (CO_2), methane (N_2O) and many fluorinated gases. A GHG footprint is the numerical quantity of these gases that a single entity emits. The Calculation can be ranging from a single person to the entire world.

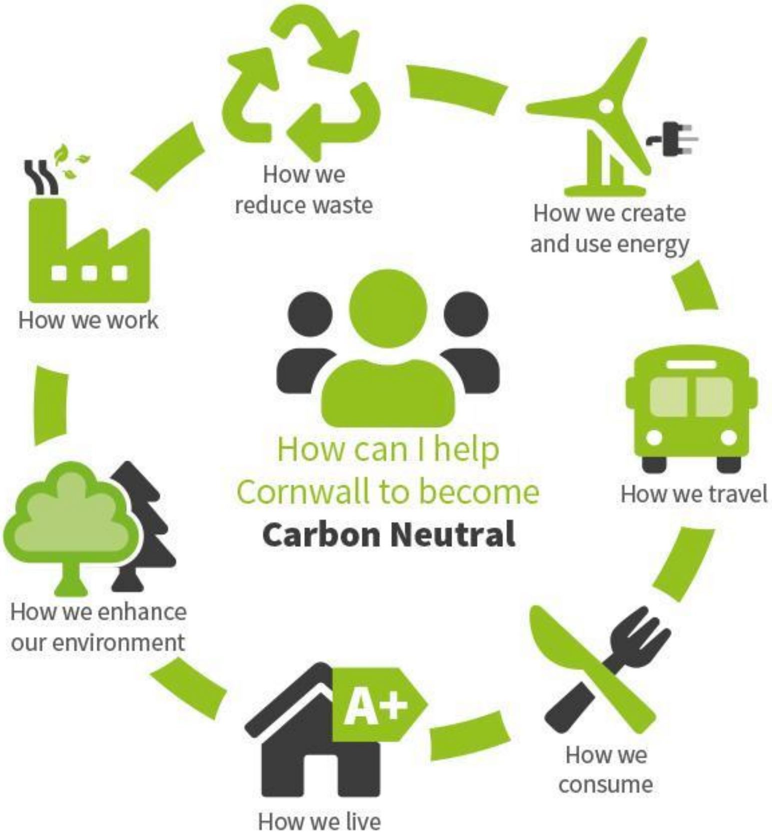
India is the world’s largest emitter of GHG, but not everyone in the country emits equally. The high expenditure households have nearly seven times the carbon footprint of the low expenditure. Household carbon footprint has shown tremendous potential for identification of unforeseen priorities for climate action in individual, countries and globally. Some studies revealed the importance of income equality, social attitudes and religion. The higher figures are found in the urban area than the rural area. Every household have their own consumption patterns. A household carbon footprint consists of not just its own carbon emissions but also the emission created to produce the goods and services it consumes. The carbon footprint is driven mainly by electricity food, firewood, kerosene, cooking gas, public and private transport, etc.

Calculating carbon footprint of industry, product, good and services is a complex task. The “international organization for standardization” has a standard called ISO that has the framework for conducting a study. ISO family of standard provides further sophisticated tools for quantifying, monitoring, reporting and validating of the GHG emission and reduction another method is through the GHG protocol, a set of standards for tracking GHG emission across scope 1, 2&3 emissions within the value chain.

Direct carbon emission or ‘scope 1’ carbon emission comes from sources that are directly from the site that is producing a product or delivering a service

Indirect carbon emission are emission from the sources upstream or downstream from the process being studied also known as scope 2 or scope 3 emissions. Example for upstream emission may include transportation of materials, energy used for the production, waste production from the production unit, etc. example for downstream emission may include product and waste transportation, emission associated with the selling the product, etc. scope 2 emission are the other indirect emission related to purchased electricity, heat and steam used on site. Scope 3 emissions are all other indirect emissions derived from the activities of an organization but from sources which they do not own or control.

Mobility, shelter and food are the most important consumption factors determining the carbon footprint of a person.


How we reduce waste
How we create and use energy
How we travel
How we consume
How we live
A+
How we enhance our environment
How we work
How can I help Cornwall to become Carbon Neutral

CARBON SEQUESTRATION

Carbon sequestration is the process of capturing and storing atmospheric carbon dioxide. It is one method of reducing the amount of carbon dioxide in the atmosphere with the goal of reducing global climate change. There are different types of carbon sequestration mainly classified as abiotic and biotic sequestration. Abiotic sequestration includes Oceanic Injection, Geological Injection, Scrubbing and Mineral Carbonation. Biotic sequestration include oceanic sequestration and terrestrial sequestration.

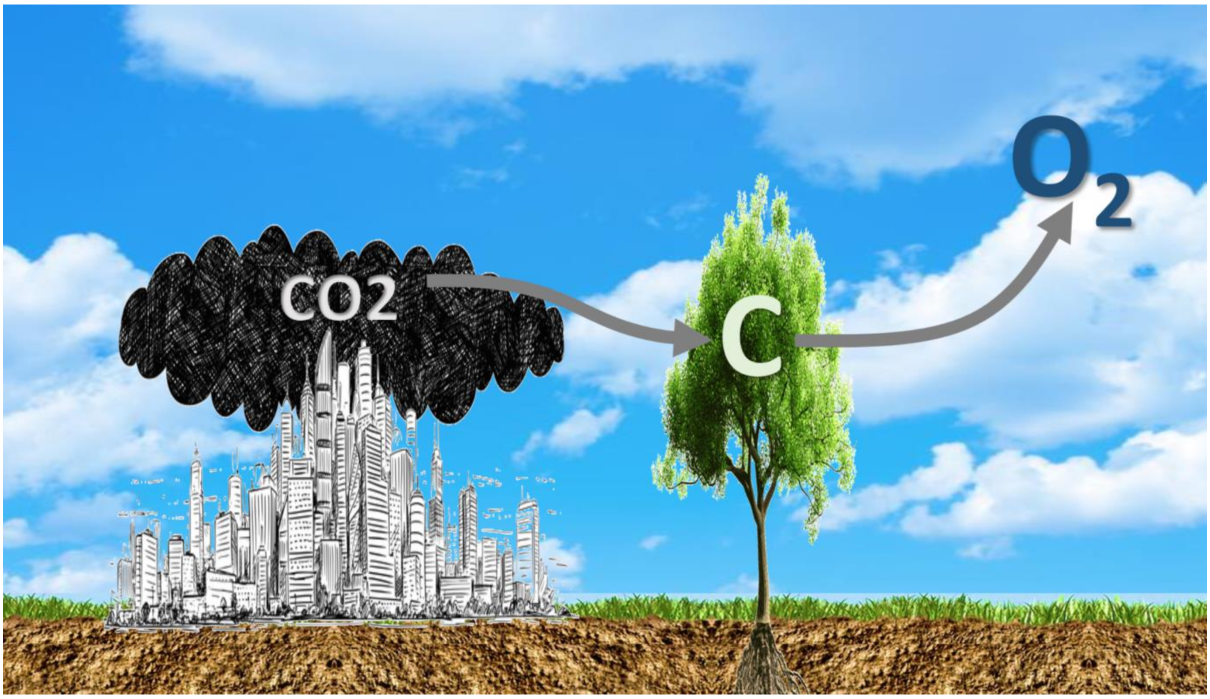
Carbon sequestration is the process by which atmospheric carbon dioxide is taken up by trees, graces and other plants through photosynthesis and stored as carbon in biomass (trunk, branches, foliage, roots) and soil or carbon sequestration is long term storage of carbon in plants, soils, geologic formations, and the oceans. The sink of carbon sequestration forests and wood products helps to offsets sources of carbon dioxide to the atmosphere, such as deforestation, forest fires, and foil fuel emissions. Gardens can act as a good carbon sink.

Trees sequester or store carbon mainly in trees and soil. During the process of photosynthesis trees pull carbon out of the atmosphere to make sugar, but they also release carbon dioxide back into the atmosphere through decomposition. Carbon and other gases within forests are captured and released on a cycle.

CO2 is naturally captured from atmosphere through biological, chemical and physical processes. There are also some artificial processes by which carbon can be captured but still the amount of carbon emission is always greater than the amount of carbon sequestration.

The idea is to stabilize carbon in solid and dissolved forms so that it doesn't cause the atmosphere to warm. The process shows tremendous promise for reducing the human "carbon footprint".

Carbon sink: Reservoirs that return carbon and keep it from entering from Earth's atmosphere are known as carbon sinks.



OBJECTIVE

- To study the relationship between carbon footprint and different lifestyle factors.
- To study the carbon sequestration potential of the gardens.
- To model the relationship between carbon footprint and different lifestyle.
- To study the carbon sequestration and impact of tree species.
- To study various indices and analyse the dominant tree species in selected gardens in Kolhapur city.

DATA COLLECTION METHODOLOGY

- ## Data source:

The data was collected from (1) Professor Priya Vasagadekar of Department of Environment Science (2) Ms. Chaitrali Chavan and Ms. Siddique Sheikh of Department of Environment Science who are currently working on the project related carbon footprint and carbon sequestration under the guidance of Prof. Priya Vasagadekar.

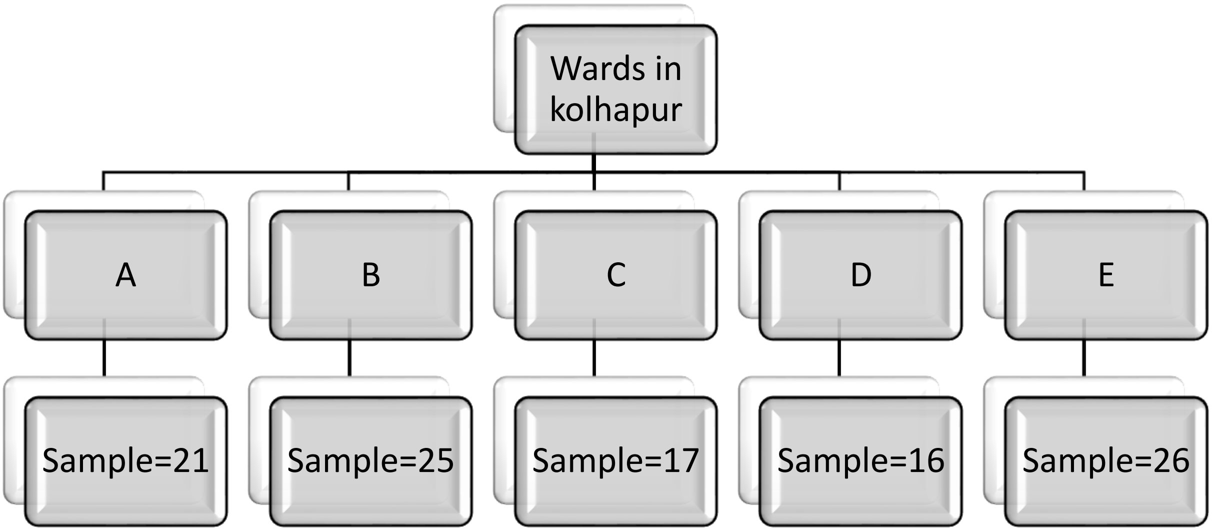
They collect two different data set for this project. First data set was related to carbon footprint and second was related to carbon sequestration. First dataset contain various parameters such as carbon footprint, family member, electricity consumed, two wheelers, four wheelers, mileage of two wheelers, mileage of four wheelers, vehicle maintained, expenditure on grocery, clothes, newspaper, books, etc. Second dataset contain various parameters such as tress species, biomass, carbon stock, carbon sequestration, etc.

- **collection for carbon footprint:**

In order to collect data for carbon footprint the Kolhapur city area is divided into five wards, they consider each ward as strata and from each stratum sample is selected by convenient sampling method.

They have collected a data from five different wards namely A, B, C, D, E. A random sample of size 105. We collected data of consumption pattern of fuel, LPG, kerosene, electricity. Also consumption of wood and monthly or yearly expenditure on food, electronic devices, newspapers, insurance, education etc.

Several extra details were included in the questionnaires to cover all possible topics.



- ## Data collection for carbon sequestration:

For carbon sequestration, we collect data from eight different gardens of Kolhapur city.

Garden Name	Total trees	Total Species
Maharashtra Garden	67	17
Daphale Garden	15	12
Hutatma Garden	120	21
Kokani Math	25	5
Mahavir Garden	210	23
Rankala Garden	21	6
Shahu Garden	89	16
Hutatma Smarak	64	10

Statistical tools and Software's

- **Statistical tools:**

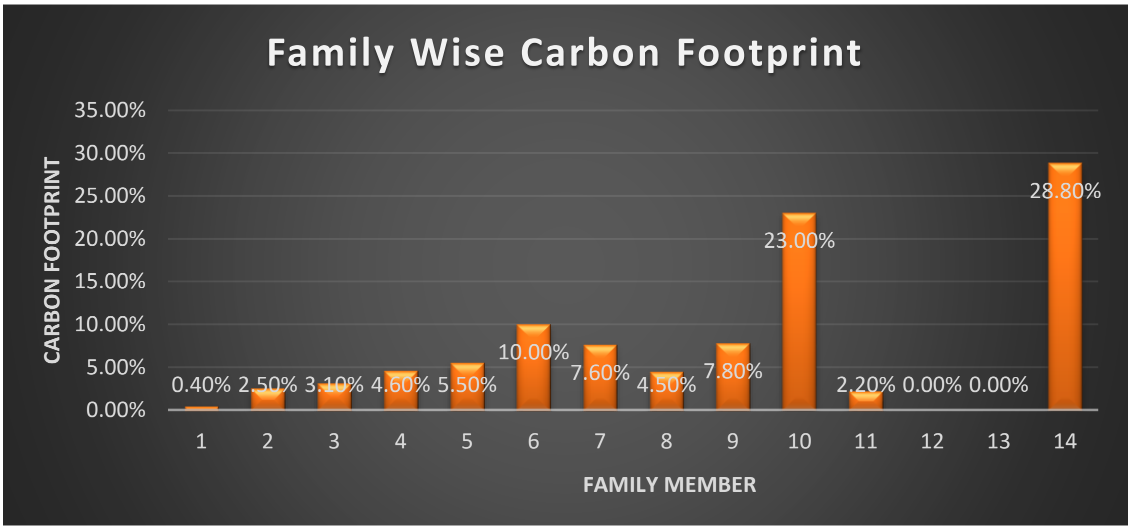
1) Graphical Representation.
2) Data Mining Classifiers: Decision Tree, Random Forest, K-Nearest Neighbours (KNN).
3) Kruskal -Wallis Test.

- **Statistical Software's:**

1) Microsoft Excel
2) R-software
3) Python
4) Minitab

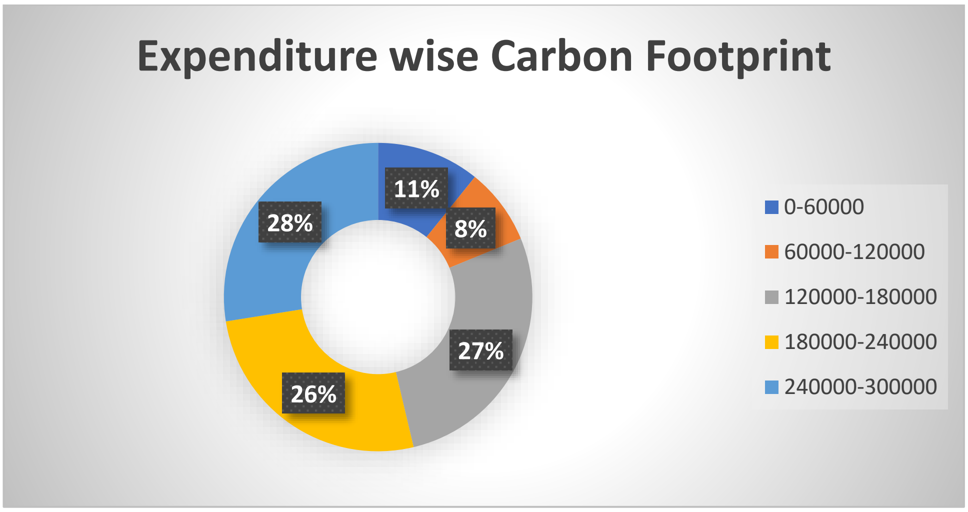
Graphical Representation of Carbon Footprint

1) Carbon footprint with size of family:



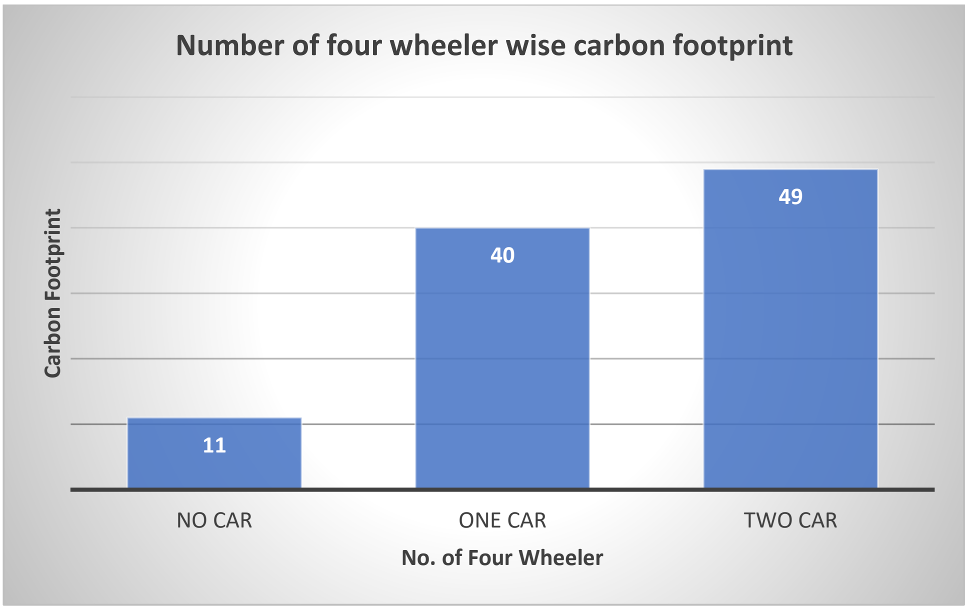
Interpretation: From above chart it is seen that the highest carbon footprint was corresponding to 14 family members.

2) Expenditure wise carbon footprint:



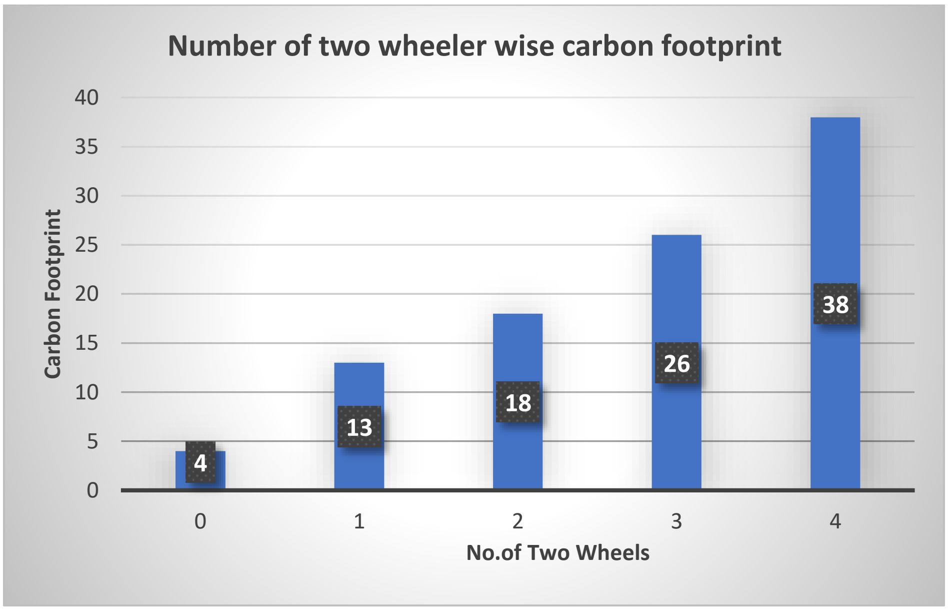
Interpretation: From above chart it is seen that maximum expenditure produce maximum carbon footprint.

3) Four wheeler wise carbon footprint:



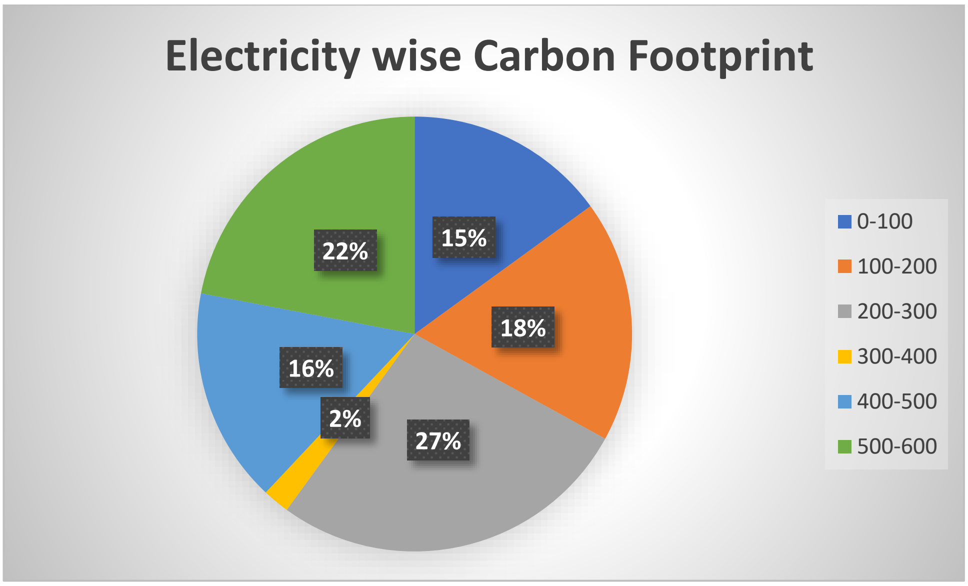
Interpretation: From above column chart it is seen that, as number of four wheeler usage increases then the carbon footprint increases.

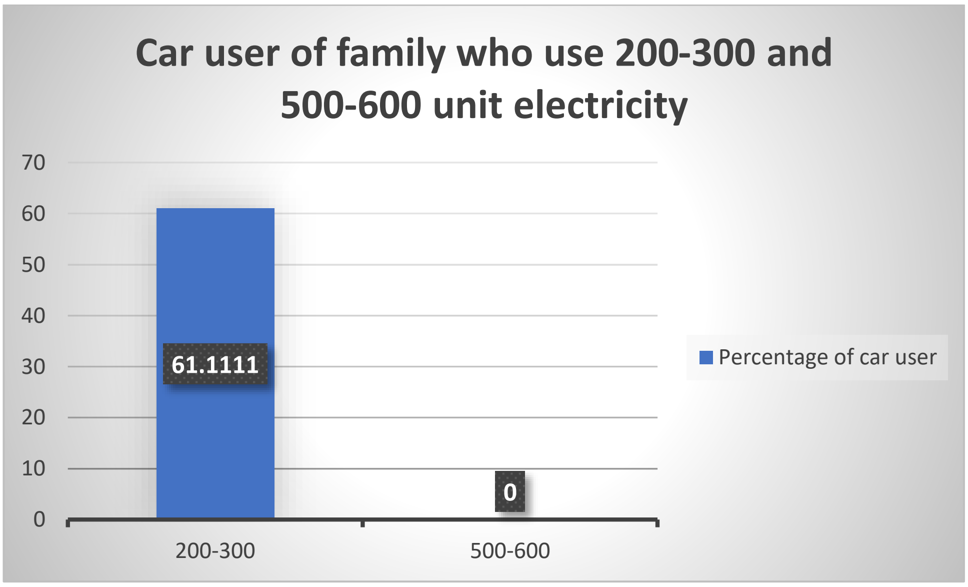
4) Two wheeler wise carbon footprint:



Interpretation: From above column chart it is seen that, as number of two wheelers usage increases then the carbon footprint increases.

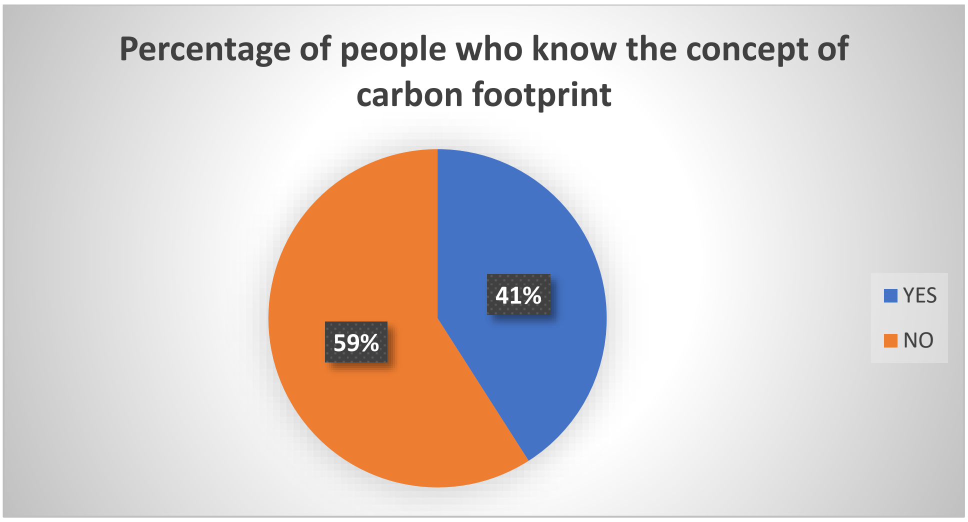
5) Usage of electricity with carbon footprint:





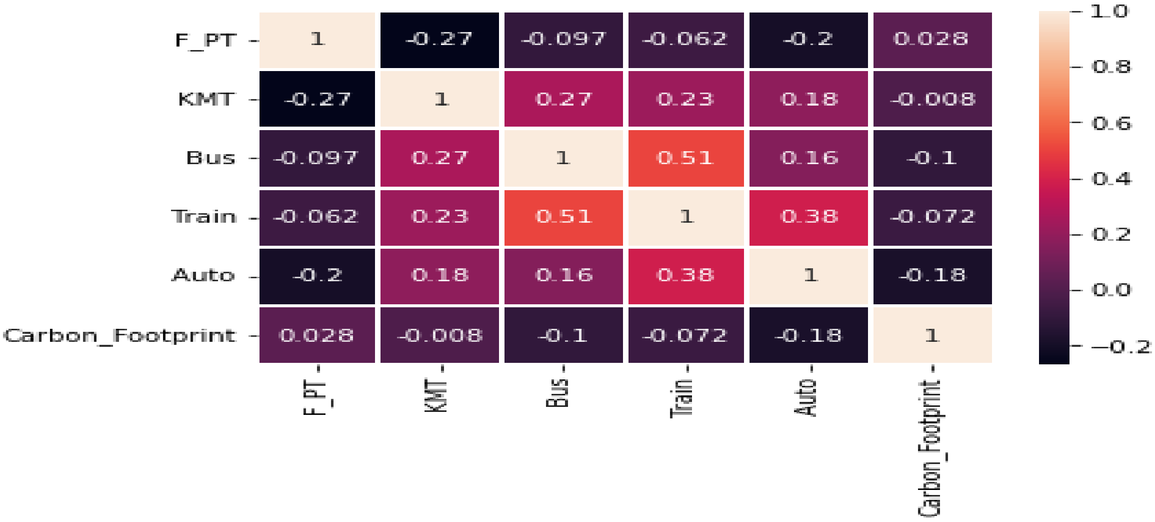
Interpretation: From above bar chart it is seen that, family who use electricity between 200 to 300 units has produce more carbon footprint. This is due to the family who use 200-300 unit electricity uses more cars as compare to other.

6) Carbon footprint concept:



Interpretation: From above pie chart it is seen that, 59% of people who don't know the concept of carbon footprint, it shows that less number of people knows the concept of carbon footprint.

7) Heat map of Public Transport:



From above heat map we say that, there is negative correlation between the carbon footprint and bus, KMT, train, auto. It states that, use of public transport can decrease the carbon footprint.

Model Building

First I delete those observations that contain outlier observation. After deleting the outlier observation our data contains 103 observations and 36 variables.

After changes in data we use k-nearest neighbours, decision trees and random forest. This model has been selected for this study because of their popularity in the recent literature. I first give a short description of these classification models.

1) K-Nearest Neighbour:

K-Nearest Neighbour is one of the simplest machine learning algorithms based on supervised learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into category that is most similarity to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. K-NN algorithm at the training phase just stores the dataset and when it gets new data, and then it classifies that data into a category that is much similar to the new data.

2) Decision Tree:

Decision tree are powerful classification algorithm that are becoming increasing more popular with the growth of data mining in the field of information systems. In the literature there are so many popular decision tree algorithms like ID3, C4.5 and C5 etc. As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, they use mathematical algorithms like information gain, Gini index to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups. This step is repeated at each leaf node until the complete tree is conducted. The objective of the splitting pair that maximizes the homogeneity of the resulting two or more subgroup samples. The most commonly used mathematical algorithm for splitting includes Entropy based information gain used in ID3, C4.5 and C5. Gini index used in CART. In this study I choose to use J48 algorithm as our decision tree method which is a simple C4.5 decision tree for classification.

3) Random Forest:

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithm. It utilizes ensemble learning which is a technique that combines many classifiers to provide solution to complex problem.

Features of random forest algorithm:

- It is more accurate than the decision tree algorithm.
- It provides an effective way of handling missing data.
- It solves the issue of over fitting in decision trees.

In every random forest tree a subset of features is selected randomly at the node's splitting point.

The main difference between the decision tree algorithm and the random forest algorithm is that establishing root nodes and segregating nodes is done in the latter.

The random forest employs the bagging method to generate the required prediction. Bagging involves using different samples of data (training data) rather than just one sample. A training dataset comprises observations and features that are used for making predictions. The decision trees produce different outputs, depending on the training data fed to the random forest algorithm. These outputs will be ranked and the highest will be selected as the final output.

4) Kruskal-Wallis Test:

Kruskal-Wallis test, proposed by Kruskal and Wallis in 1952, is one of the non-parametric tests that are used as a generalised form of the Mann-Whitney U test. It is used to test the null hypothesis which states that 'k' number of sample has been drawn from the same population or the identical population with the same or identical median. If S_j is the population median for the j^{th} group or sample in Kruskal-Wallis test, then the null hypothesis in mathematical form can be written as $S_1=S_2=...=S_k$. Obviously, the alternative hypothesis would be that S_i is not equal to S_j. This means that at least one pair of groups or samples has different pairs.

The test statistic of Kruskal-Wallis test is,

$$H = \left(\frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(n+1)$$

Where,

k= number of comparison groups,

n= total sample size,

n_j=sample size in the j^{th} group,

R_j=sum of the ranks in the j^{th} group

A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference. If the p-value is less than or equal to significance level, we reject the null hypothesis and conclude that not all the group medians are equal.

5) Root Mean Squared Error(RMSE):

RMSE is standard deviation of the errors which occur when a prediction is made on a dataset. This is a same as MSE (Mean Square Error) but the root of the value is considered while determing the accuracy of the mode. It has the useful property of being the same unit as response variable. Lower value of RMSE indicates better fit. RMSE is good major of how accurately the models predict the response.

a) K-Nearest Neighbour:

K-Nearest Neighbour is one of the simplest machine learning algorithms based on supervised learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into category that is most similarity to the available categories.

The RMSE of the data is,

Train dataset	Test dataset
0.87	1.04

Accuracy=78%

b) Decision Tree:

Decision tree are powerful classification algorithm that are becoming increasing more popular with the growth of data mining in the field of information systems.

The RMSE of the data is

Train dataset	Test dataset
0.59	0.98

Accuracy=71%

c) Random Forest:

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithm. It utilizes ensemble learning which is a technique that combines many classifiers to provide solution to complex problem.

Train dataset	Test dataset
0.19	0.76

Accuracy=80.6%

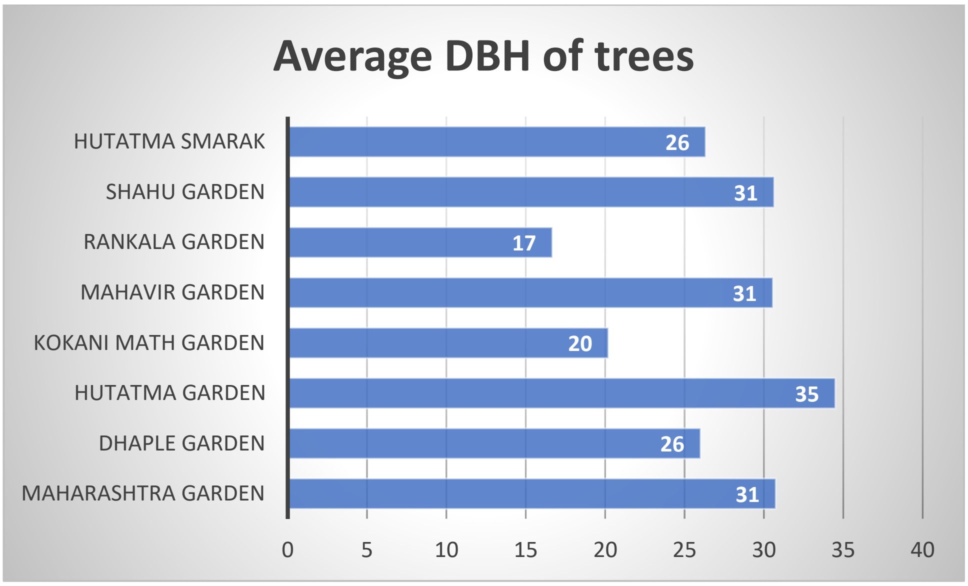
The performance of above all classifiers is compared in following table,

	Performance measure		
Classifiers	Accuracy	train dataset RMSE	test dataset RMSE
K-NN	78	0.87	1.04
Decision Tree	71	0.59	0.98
Random Forest	80.6	0.19	0.76

From above table it is clear that accuracy of random forest larger than all other classifiers (80.6%) as well as the RMSE of train and test dataset is of random forest is smaller than all other classifiers.

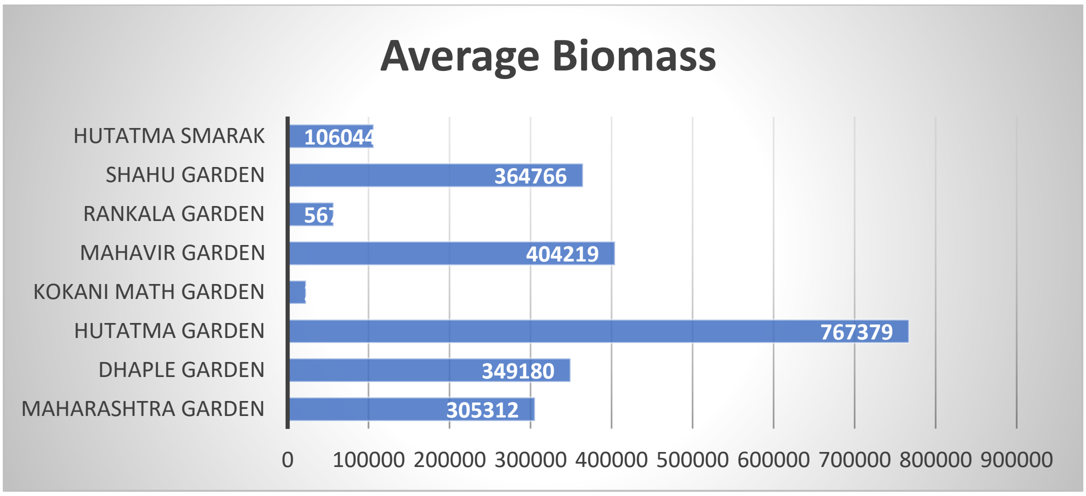
Graphical Representation of Carbon Sequestration

1) Diameter breast height (DBH):



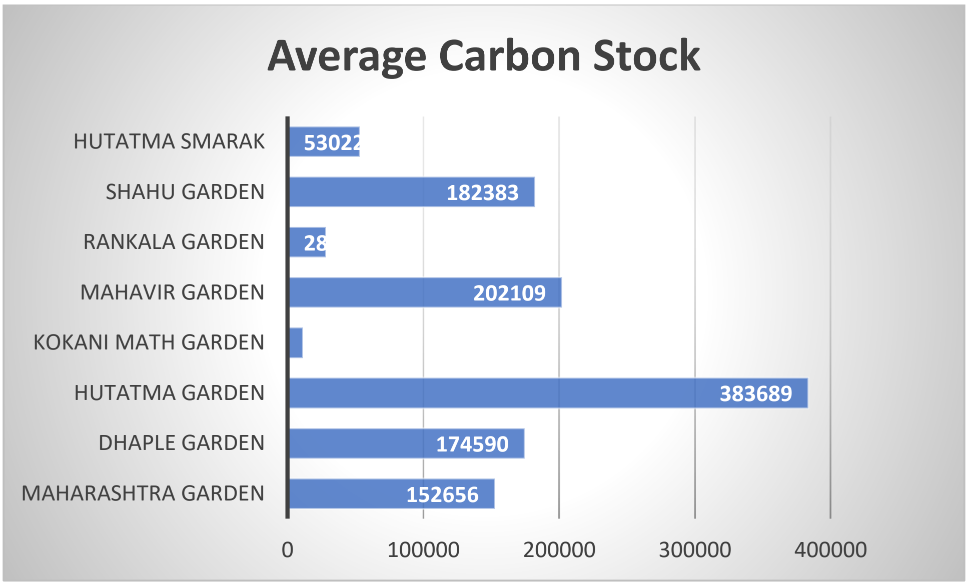
Interpretation: The average DBH (Diameter Breast Height in ft) of tree species in different gardens demonstrate that Hutatma garden has the highest DBH, compared to other gardens. This is due to bigger tree species encountered as compare to other gardens.

2) Biomass Production:



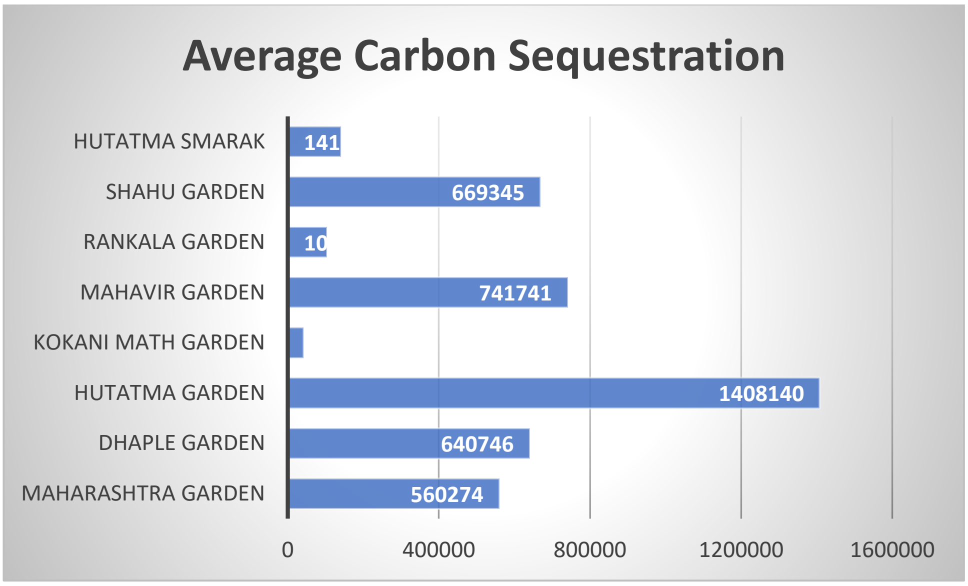
Interpretation: The above bar chart shows that average tree biomass density at different gardens. The average trees biomass density is highest in Hutatma Garden followed by Mahavir Garden although maximum numbers of tree species were present in Mahavir Garden.

3) Carbon Stored at different garden:



Interpretation: The above bar chart shows that average carbon store of tress species at different gardens. Referring to the average carbon stored in tree species, Hutatma Garden has highest average carbon stored.

4) **Carbon sequestration at different gardens**

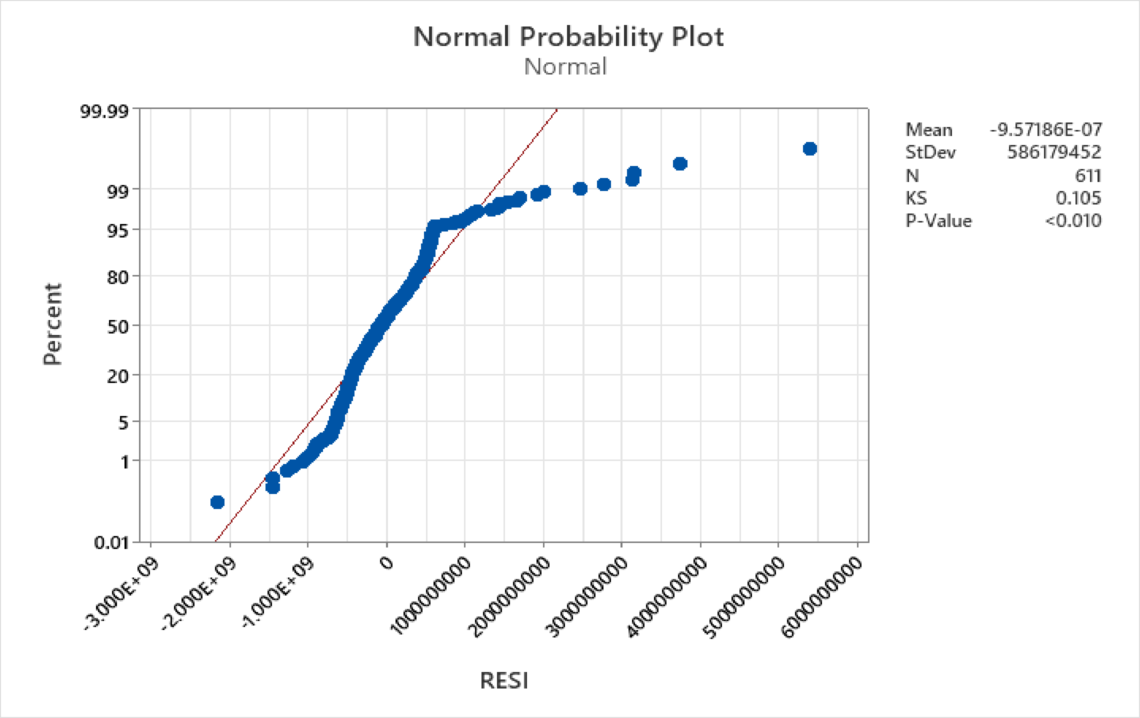


Interpretation: The above bar chart shows that average carbon sequestration of tress species at different gardens. The highest amount of carbon sequestration in Hutatma Garden was due to high biomass density and carbon stored.

Statistical Analysis

1) Kolmogorov-Smirnov Normality test for carbon sequestration (kg) versus height(cm) and width(cm):

The Null Hypothesis of this test is that the carbon sequestration and height, width is normally distributed.



From the above graph, p-value is less than the significance level 0.05. We can conclude that the data is not normally distributed.

2) Non parametric ANOVA (Kruskal-Wallis test):

Kruskal-Wallis test, proposed by Kruskal and Wallis in 1952, is a non-parametric method for testing whether samples are originated from the same distribution. The null hypothesis of the Kruskal-Wallis test is that, the mean ranks of tree species are same. As the non-parametric equivalent one-way

ANOVA, Kruskal-Wallis test is called one-way ANOVA on ranks. Unlike the analogous one-way ANOVA, the non-parametric Kruskal-Wallis test does not assume a normal distribution of the underlying data.

Here, the Null Hypothesis for this test is the mean ranks of the ranks of the tree species are the same.

```
data:  carbon sequestration (kg)  by factor(Tree name)
Kruskal-Wallis chi-squared = 75.542, df = 9, p-value = 1.235e-12
```

From the outputs generated by above test we see that the carbon sequestration (kg) for different Tree species is significantly different from each other.

3) Dominant Species:

- Top 10 tree species that store and sequester high amount of mean carbon for above 30cm DBH.

DBH>30		
Common name	Carbon store	Carbon sequestration
Bhabul	1003559	3683063
Rain tree	641439.5	2354083
Gulmohar	418242.6	1523572
karanj	241290.9	885537.7
kassod	309053.1	809728.5
Putranjiva	164672	604346.3
Royal palm	161809.6	593841.3
Suru	156367.4	573868.5
Ashoka	151490.4	552667.4
kadamb	133100.9	395770.3

Among above tree species in the above table, Bhabul demonstrated the highest amount of carbon stored 1003559 with tonne and carbon sequestration of 3683063 tonne. The tree species that has the lowest carbon stored and sequestration is the Kadamb with 133100.9 tonne and 395770.3 tonne.

- Top 10 tree species that store and sequester high amount of mean carbon for 5 to 30 cm DBH.

DBH<30		
Common name	Carbon store	Carbon sequestration
Rain tree	79855.495	290355.44
Gulmohar	73313.378	266571.7
Ashoka	72112.575	262309.99
Suru	70933.625	257955.57
Kassod	61726.785	224326.47
Karanj	43655.739	157817.2
Fishtail palm	40827.054	148112.81
Areca palm	35674.597	130540.93
Australian babhul	35426.147	127983.55
Kadamb	35109.007	126225.71

Among above tree species in the above table, Rain Tree demonstrated the highest amount of carbon stored with 79855.495 tonne and carbon sequestration of 290355.44 tonne. The tree species that has the lowest carbon stored and sequestration is the Kadamb with 35109.007 tonne and 126225.71 tonne.

Scope and Limitation:

- **Scope:**

The studied analysis can be used as a base for the Green Audit of the city as well as can be used for the smart city project and development activities.

- **Limitations:**

1. The results of statistical tools comes from this particular analysis are not valid universally.
2. At the time of data collection some information was not available (missing data) also there are some outliers are resent in our data.
3. We can’t able to build the model because our data does not satisfy the assumptions of model building.

Major Findings and Conclusions:

- **Major Finding:**

In accordance with the SEMMA methodology, after obtaining the sample data we can started by understanding variable and relation with carbon footprint and carbon sequestration. Also skewness of the data has significant effect on analysis. Also dealing with missing values and outliers is an important step of data cleaning. For carbon footprint we further modelled our data using various classification techniques K-NN, Decision Tree, Random Forest. Then when assessing the various techniques, we used RMSE to compare the results. For carbon sequestration we used Kruskal-Wallis test for comparing the mean rank of trees species.

- **Conclusion:**

In conclusion, for carbon footprint we can show that it is possible to use a superior analytics algorithm through the use of different sampling method to correctly classify carbon footprint. We believe that Random forest classifier classifies in term of RMSE and accuracy. The accuracy of random forest is larger than all other classifiers (80.6%) and RMSE of random forest (0.76) is smaller than all other classifiers. For carbon sequestration we can show that the impact of tree species on carbon sequestration. Trees species has significant impact on carbon sequestration. We can see that from the top 10 species we say that, Ashoka tree has most carbon sequestration capacity when DBH>30 and Areca Palm tree has most carbon sequestration capacity, when DBH<30.

Reference:

- Richard O. Duda, Peter E. Hart, David G. Stork
  - Pattern Classification

    Second Edition

- Jiawei Han, MichelineKamber, Jain Pei
  - Data Mining: Concepts and Techniques.

    Third Edition

- Gopal K. Kanji
  - 100 Statistical Tests

Appendix:

1) K-NN code in python:

```
import pandas as pd
jm=pd.read_csv("C:/Users/pcc/Desktop/suga.csv")
print(jm)
type(jm)
import matplotlib.pyplot as plt
jm["Carbon_Footprint"].hist(bins=3)
plt.show()
correlation_matrix = jm.corr()
correlation_matrix["Carbon_Footprint"]
X = jm.drop("Carbon_Footprint", axis=1)
X = X.values
y = jm["Carbon_Footprint"]
y = y.values
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=12345)
from sklearn.neighbors import KNeighborsRegressor
knn_model = KNeighborsRegressor(n_neighbors=6)
knn_model.fit(X_train, y_train)
from sklearn.metrics import mean_squared_error
from math import sqrt
train_preds = knn_model.predict(X_train)
mse = mean_squared_error(y_train, train_preds)
rmse = sqrt(mse)
rmse
test_preds = knn_model.predict(X_test)
mse = mean_squared_error(y_test, test_preds)
rmse = sqrt(mse)
```

```
rmse
from sklearn.model_selection import GridSearchCV
parameters = {"n_neighbors": range(1, 50)}
gridsearch = GridSearchCV(KNeighborsRegressor(), parameters)
gridsearch.fit(X_train, y_train)
GridSearchCV(estimator=KNeighborsRegressor(),param_grid={'n_neighbors':
range(1, 50),'weights': ['uniform', 'distance']})
train_preds_grid = gridsearch.predict(X_train)
train_mse = mean_squared_error(y_train, train_preds_grid)
train_rmse = sqrt(train_mse)
test_preds_grid = gridsearch.predict(X_test)
test_mse = mean_squared_error(y_test, test_preds_grid)
test_rmse = sqrt(test_mse)
train_rmse
test_rmse
from sklearn.metrics import accuracy_score
print(accuracy_score(y_test, test_preds))
```

2) **Decision Tree code in python**:

```
pip install sklearn
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
import pandas as pd
jm=pd.read_csv("C:/Users/pcc/Desktop/suga.csv")
jm.head()
jm.shape
jm.columns
inputs=jm.iloc[:,:-1]
inputs.columns
```

```
target = jm['Carbon_Footprint']
feature_col=['Electricity', 'Two_Wheeler', 'Four_Wheeler',
    'Transportation_vehicle', 'Mileage_Two_Wheeler', 'Mileage_Four_Wheeler',
    'Mileage_Transportation_Vehicle', 'Vehicle_Maintenance',
    'EXP_Grocery', 'EXP_Clothes', 'EXP_Newspaper', 'Mob_recharge',
    'Exp_computer', 'Exp_TV', 'Exp_furniture', 'Insurance ',
    ' Education_cost', 'Cost_family_event', 'Exp_hotel']
x=jm[feature_col]
y=target
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=1)
clf=DecisionTreeClassifier()
clf=clf.fit(x_train,y_train)
clf
y_pred=clf.predict(x_test)
y_pred
accuracy=metrics.accuracy_score(y_test,y_pred)
accuracy
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.tree import plot_tree
from sklearn.tree import plot_tree
plt.figure(figsize=(50,50))
plot_tree(clf,feature_names=feature_col,class_names='Carbon_Footprint',filled=True)
plt.show()
from sklearn.metrics import mean_squared_error
from math import sqrt
train_preds = clf.predict(x_train)
mse = mean_squared_error(y_train, train_preds)
rmse = sqrt(mse)
rmse
```

```
test_preds = clf.predict(x_test)
mse = mean_squared_error(y_test, test_preds)
rmse = sqrt(mse)
rmse
```

3) Random Forest code in python:

```
import pandas as pd
from sklearn import metrics
from sklearn.model_selection import train_test_split
jm=pd.read_csv("C:/Users/pcc/Desktop/suga.csv")
print(jm)
jm.describe()
x=jm.drop('Carbon_Footprint',axis='columns')
y=jm['Carbon_Footprint']
feature_names=x.columns
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import RandomForestClassifier
x_train, x_test, y_train, y_test=train_test_split(x,y,test_size=0.3)
my_model=RandomForestClassifier()
my_model.fit(x_train,y_train)
y_test_pred=my_model.predict(x_test)
print(metrics.accuracy_score(y_test,y_test_pred))
from sklearn.metrics import mean_squared_error
from math import sqrt
train_preds = my_model.predict(x_train)
mse = mean_squared_error(y_train, train_preds)
rmse = sqrt(mse)
rmse
test_preds = my_model.predict(x_test)
mse = mean_squared_error(y_test, test_preds)
```

```
rmse = sqrt(mse)
rmse
```

4) Kruskal-Wallis test in R:

```
> Data=read.csv("C://Users//Lenovo//Downloads//bts.csv")
> Data
>fit=kruskal.test(carbon.seq.kg~factor(common.name),data=Data)
> fit
```