FINAL REPORT

1.PROJECT DESCRIPTION

This project is about the data visualization and data analysis methods which were learned throughout the semester in the class which are used on a given dataset. The project contains many visualizations of the data which is generated through Tableau and R and different data analysis methods like linear regression, clustering and random forests using R. This is the final report of the project. The earlier phases of this project are Dataset selection and oral presentation. This is the final phase.

2.INTRODUCTION AND DATASET DESCRIPTION

The data is collected from Kaggle open datasets. This public dataset is part of Airbnb. Airbnb was founded in August 2008 in San Francisco, California. Airbnb, Inc. is an online marketplace for arranging or offering to lodge, especially for homestays and tourism experiences. The company does not own any of the real estate listings. It just acts like a broker and receives a commission from each booking. The company is based in San Francisco, California, United States. Hosts are required to provide government-issued identification before accepting a reservation. The guests have an option to chat with the hosts through a secure messaging system. The Hosts provide rates and other details like the total number of guests allowed, home types, rules, and amenities provided. The host determines to price along with the recommendations from Airbnb.

The dataset is stored in the form of a CSV file. The dataset gives information about the listing activity of lodges for the guest to lodge in Singapore city. The dimensions of the dataset are 7675 Rows and 16 columns.

The different attributes or columns of the datasets are

- id: It is the listing ID given by the company Airbnb.
  Data type: Interval

- name: It is the name of the listing, i.e., Name of the hotel or lodge.
  Data type: Nominal

- host_ID: It is the ID of the host. Hosts are the ones who give lodging facilities to the guests.
  Data type: Interval

- host_name: It is the name of the Host. The person who owns the rooms is the host.
  Data type: Nominal

- neighbourhood_group: It takes about the location in which the listing is located. Example: North Region, West Region, etc.
  Data type: Nominal

- neighborhood: It is the name of the area where the listing is located.
  Data type: Nominal
- latitude: It tells the exact latitude at which the listing is located.
  Data type: Interval
- longitude: Similarly, like latitude, it tells the exact longitude point at which the listing is located.
  Data type: Interval
- room_type: It tells what type of room it is like a Private room or Entire home or Apartment.
  Data type: Nominal
- price: It tells about the price of the room or apartment. The price is mentioned in USD.
  Data type: Ratio
- minimum_nights: It tells about the minimum of nights that the guests need to spend at the host's place.
  Data type: Ratio
- number_of_reviews: It is the count of the total number of reviews given by the guests who had stayed there previously.
  Data type: Ratio
- last_review: It is the date at which the guest gave the last review to the host.
  Data type: Interval
- reviews_per_month: It tells about the average number of reviews per month the hosts get from the guests.
  Data type: Ratio
- calculated_host_listings_count: It is the amount of listing per host.
  Data type: Ratio
- availability_365: It tells the number of days when the listing is available for booking.
  Data type: Ratio

The size of the CSV file is approximately 1MB.

3.RESOURCES REQUIRED TO STUDY THE DATASET:

SOFTWARE RESOURCES

- R STUDIO: For Data exploration, Creating Visualizations, find the summary, and other statistics, Research Questions
- Tableau: For Creating Visualizations.
- MS EXCEL: For storing the data in the form of a table.

METHODS AND TECHNIQUES USED:

1. Linear Regression Statistics give the best predictor to use for predicting the target variable which is the price.
2. Clustering: The data is clustered with the cluster value of 10.
3. Random Forest: The data is classified with the help of Random Forest model

4.RESEARCH QUESTIONS

1. Can a model be built that predicts the price using a best predictor available?
2. Can a model be built that predicts the minimum nights required in a particular neighborhood group /neighborhood?
3. Can a model be built that divides the data into different groups and classify them such that a particular price comes under a particular group?

5.EXPLORATORY ANALYSIS

1.Linear Regression Model:

In the linear regression, we have divided the data frame into 80% of the training data and 20% of the test data with the help of row indexing then created a linear regression model with the help of the training data by using the variables room_type, neighbourhood_group, latitude, longitude, number_of_reviews, availability_365, calculated_host_listings_count, minimum nights as the predictor variables and price as the variable to be predicted.

The model thus generated produced a min_max accuracy of 62%, mean absolute error of 8.543229e+01,mean square error of 3.476426e+04. By running the summary statistics for the model, the adjusted R square found was 0.05443. Here we considered the adjusted R square because there are multiple predictor variables. Here we found that the residuals are not so linear and did not generate imperative results. Hence, we have changed the model by applying log function to the price variable and generated the linear model then by running the summary statistics we have got the adjusted R square value as 0.54 which is a much better value than the previous model. From this model we can observe that the significant predictor variables (with the p-values less than 0.0001) as latitude, number_of_reviews, availability_365, calculated_host_listings_count, minimum_nights. Here, as the neighbhour_group and room_type are categorical variables a linear model could not be perfectly generated, and these variables cannot act as the predictor variables. Here we have even plotted the residuals versus fitted plot which

describes the linearity of the generated model, Normal Q-Q plot which represents the quantiles graph of residuals and the theoretical quantiles, Scale-Location plot which represents the spread of the fitted values , Cook's distance plot which represents the outliers in the data frame. By considering all these plots and the Adjusted R square value the model cannot be said to be perfectly linear, but it has some linearity. By looking at these plots it can also be said as the generated model is homoscedastic.

```
#linear regression
trainingRowIndex <- sample(1:nrow(airbnb), 0.8*nrow(airbnb))  # row indices
for the training data
trainingData <- airbnb[trainingRowIndex, ]  # model training data
testData  <- airbnb[-trainingRowIndex, ]  #creating test data

airbnb.lm <- lm(price ~ room_type + neighbourhood_group
                + latitude + longitude + number_of_reviews +
availability_365
                + calculated_host_listings_count + minimum_nights, data =
trainingData) #generating the linear regression model
pricePred <- predict(airbnb.lm, testData) #Prediction of price
AIC(airbnb.lm) #Akaike Information Criterion

## [1] 92298.89

actuals_preds <- data.frame(cbind(actuals=testData$price,
predicteds=pricePred)) #actual values
correlation_accuracy <- cor(actuals_preds) #calculating correlation accuracy
with the help of actual and predicted
head(actuals_preds) #prints the head of the actuals_preds

##    actuals predicteds
## 9       54  -4.564871
## 12      40 139.687336
## 14      65  64.381007
## 30      26  95.417556
```

Fig 5.1.1

```
#linear model 2
airbnb1 <- airbnb %>% filter(price > 0) #filtering the data for applying log
function
airbnb1.lm <- lm(log(price) ~ room_type + neighbourhood_group
                 + latitude + longitude + number_of_reviews + availability_365
                 + reviews_per_month + calculated_host_listings_count +
minimum_nights, data = airbnb1)
plot(airbnb1.lm)
```

Fig 5.1.2

2. Clustering:

Clustering is used to classify the data into different clusters based on the location and they are differentiated with the help of different colours. Initially a cluster column is created with the help of k means clustering, by using select command which is in the "dplyr" package and formed a cluster of size 10 as there are many locations and the price needs to be classified based on that cluster. The plot generated with these latitudes and longitudes describes a map of singapore. Then we have plotted the clusters based on the room type which classified the locations available for different room types with respect to the median price. Then we have plotted the clusters based on the neighbourhood_group which classified the neighbourhood_groups in different locations with respect to the median price. Here the accuracy is not calculated because the data is not divided into train and test for the clustering, but the summary statistics are generated.

```
# Plotting
hcluster <- left_join(airbnb, airbnb %>%
                    group_by(cluster, room_type) %>%
                    summarise(median_price = median(price))) #using group by
to form clusters with respect to room_type

## Joining, by = c("room_type", "cluster")

hcluster1 <- left_join(airbnb, airbnb %>%
                         group_by(cluster, neighbourhood_group) %>%
                         summarise(median_price = median(price))) #using group
by to form clusters with respect to room_type

## Joining, by = c("neighbourhood_group", "cluster")

hcluster %>%
  ggplot(aes(longitude, latitude, colour = median_price)) +
  geom_point(size = 0.75) +
  scale_colour_gradient(low = 'yellow', high = 'red') +
  labs(x = 'Longitude', y = 'Latitude', colour = 'Median_Price') +
  facet_wrap(~ room_type) +
  theme(legend.position = 'bottom') #plotting the cluster by room_type
```

Fig 5.2.1

3. Random Forest:

The Random Forest represents the node purity i.e which variables split the data correctly. For this "randomForest" library is imported and then we have used the variables room_type, neighbourhood_group, latitude, longitude, availability_365, minimum nights to generate the model as the other variables are not suitable for generating the Random Forest model as they contain nominal values and NA values. Then the values are predicted, and a data frame is created with the help of cbind function which consists of the actual values and the predicted values the Root mean square error is calculated which accounts to be 356.75. Then the node purity is calculated with the help of function importance then the node purity plot is generated which depicts that latitude, longitude and neighbourhood shows the highest node purity and neighbourhood_group shows the lowest node purity. The error vs trees plot is also generated which represents the model. Then the summary statistics are generated which did not provide imperative results.

```
#RandomForest
airbnb_rf <- randomForest(price ~ neighbourhood_group + neighbourhood +
latitude + longitude + room_type + minimum_nights+ availability_365 ,data=
trainingData) # creates random forest model
test_rf <- predict(airbnb_rf, testData)  #predicts the airbnb data
airbnb_frame <- data.frame(cbind(actual= airbnb$price, predicted=test_rf))
#creates a data frame using the cbind function

## Warning in cbind(actual = airbnb$price, predicted = test_rf): number of
## rows of result is not a multiple of vector length (arg 2)

airbnb_frame$error <- airbnb_frame$actual - airbnb_frame$predicted
#calculates the error
rmse_rf <- sqrt(mean(airbnb_frame$error^2)) #calculates the rmse value from
the above created column
print(rmse_rf) #prints the rmse value

## [1] 385.9999
```

Fig 5.3.1

6.RESULTS

Some Tableau Visualizations

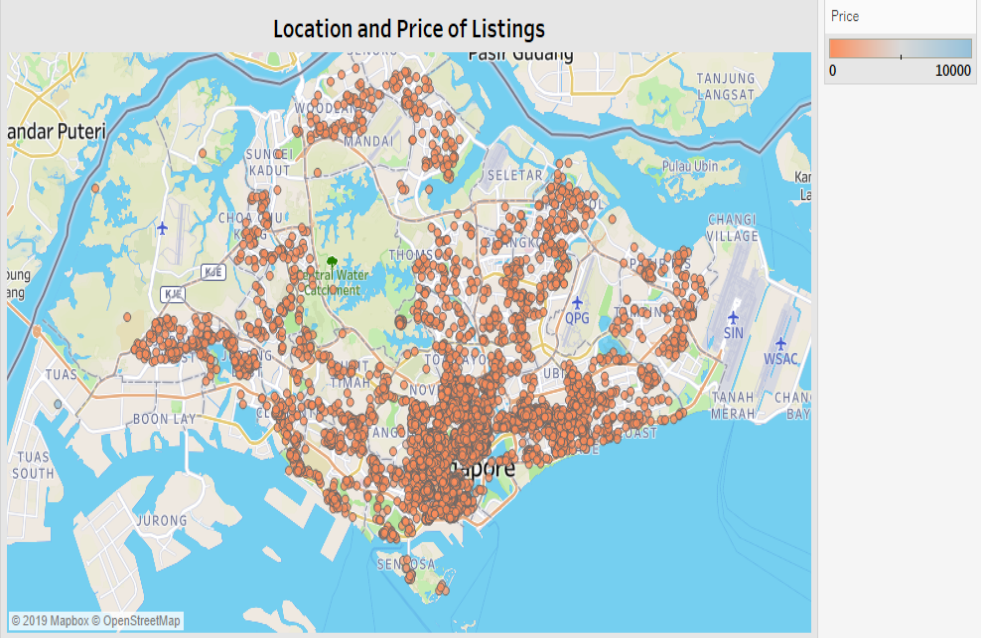

Fig 6.1

The below geographical plot shows the price in different regions of Singapore. It can also be said that most listings are available in the region Kalang

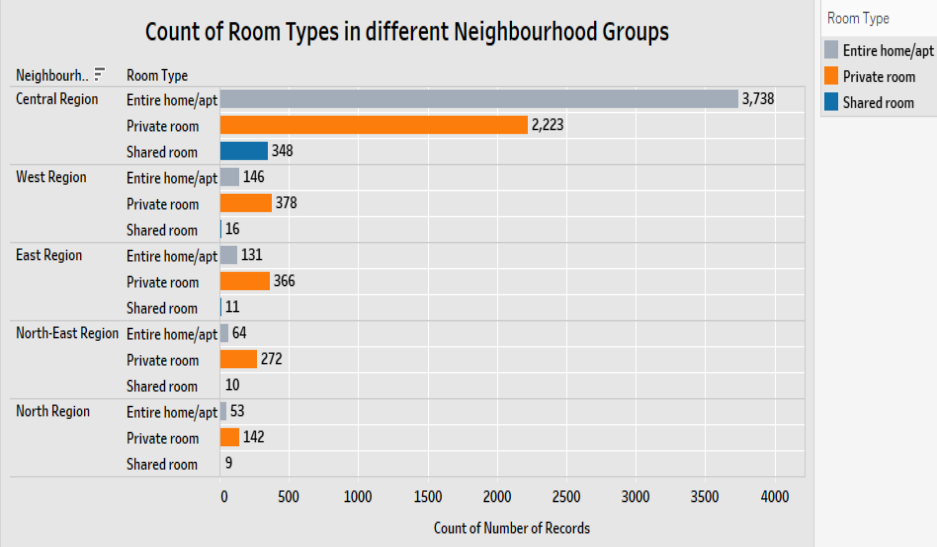

Fig 6.2

The above graph represents count of room types in different Neighbourhood groups

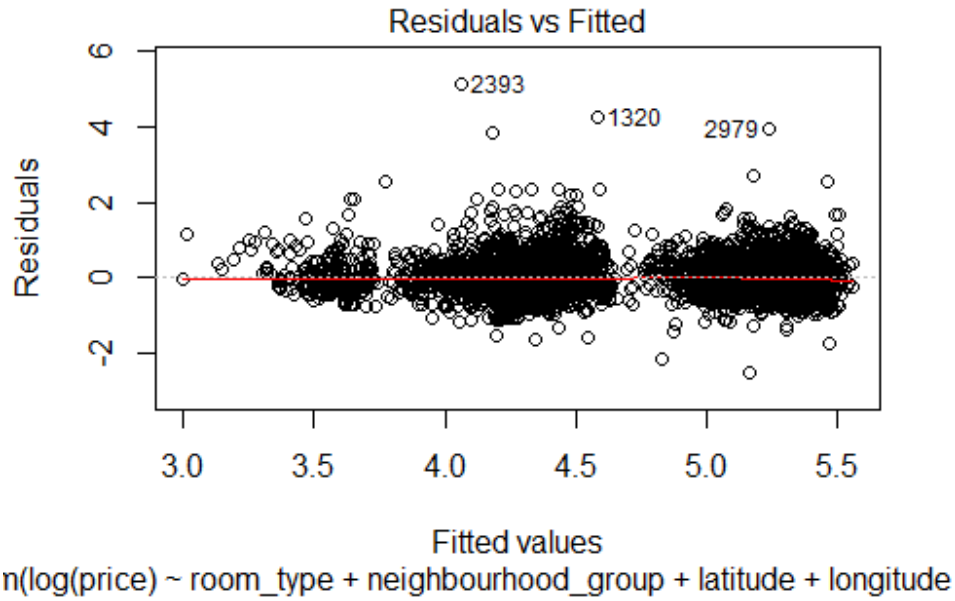

Fig 6.3

The above residual plot represents that there is no great linearity in the model.

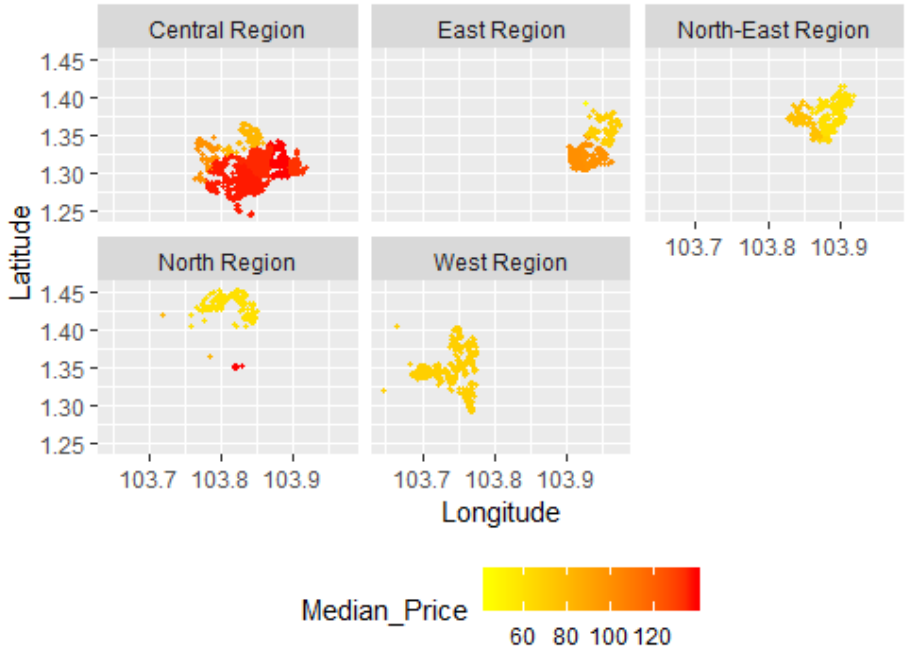

Fig 6.4

The above clustering plot represents the prices in different neighbourhood groups

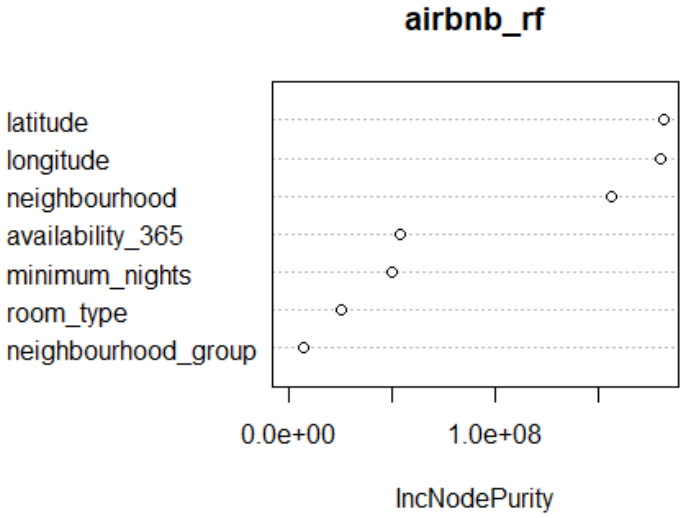

Fig 6.5

The above plot represents the node purity with the help of Random Forests

7.CONCLUSIONS

VISUALIZATION CONCLUSIONS

1. Central Region neighbourhood has most number (6309) of hotel listings which is very much greater than the sum of all the listings in the remaining regions.
2. Central Region is also the most expensive neighbourhood group followed by West Region which is not well behind.
3. Entire Home/Apartment is the most common type of rooms in the city of Singapore.
4. Kallang has the most number of listings in neighbourhood.
5. North Region neighbourhood has the least number of listings.

DATA ANALYSIS CONCLUSIONS

1. Linear Regression model gave only the accuracy of 62% and did not show predominant results but the model generated is somewhat linear.
2. The Clustering model has classified different neighbourhood groups based on the price.
3. The Random Forest model has given the values of Node purity and the summary statistics have not shown good results. It might be because of the vagueness of the dataset chosen.

8.CHALLENGES

1. They were many categorical variables in the data which is hard to analyse.
2. We have encountered Linearity Problem while predicting price variable using Linear Regression Technique.

3. We made several attempts in choosing the predictors for performing linear regression analysis.
4. For generating the linear regression model, we have applied log functions to the price variable and used filter command such that the price value is greater than zero.
5. The dataset is so vague that any analysis method did not show predominant results.

9.FURTHER ANALYSIS

The data set selected has the price value to be predicted from categorical variables, Logistic regression using dichotomous variables, Support vector machine could be applied for getting better accuracy and consistent results.

10.SOLUTIONS OF RESEARCH QUESTIONS

1. The Linear Regression model can be built with predictor variables as latitude, number_of_reviews, availability_365, calculated_host_listings_count, minimum_nights.
2. As the minimum nights is the best predictor variable it can predict the price.
3.Clustering analysis can be used for classifying the neighbourhood groups and room type with respect to price.

11.REFERENCES

1. I Putu Angga K, Singapore Airbnb, 29 August 2019, https://www.kaggle.com/jojoker/singapore-airbnb
2. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL https://www.R-project.org/.
3. Taiyun Wei and Viliam Simko (2017). R package "corrplot":Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot
4. Hadley Wickham (2017). tidyverse: Easily Install and Load the Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse
5. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.
6. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller(2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr

12.TEAMWORK

Report: Nikhith, Sumanth and Kaushil
Tableau Visualizations: Kaushil, Nikhith, Sumanth
R Script:
Linear Regression: Nikhith
Clustering: Sumanth
Random Forest: Kaushil
Dataset Selection: Sumanth and Nikhith