INDIVIDUAL ASSIGNMENT

LORD BUDDHA EDUCATION FOUNDATION

CT127-3-2-PFDA

PROGRAMMNG FOR DATA ANALYSIS

NP2F1909IT

HAND OUT DATE: 04-11-2020

HAND IN DATE: 11-02-2021

WEIGHTAGE: 50%

SUBMITTED BY:

Aaditya Jha

NP000290

SUBMITTED TO:

Mr. R.N. Thakur

Lecturer

INSTRUCTIONS TO CANDIDATES

1. **Submit your assignment to the administration counter.**
2. **Students are advised to underpin their answers with the use of references (sites using the Harvard Name System of Referencing)**
3. **Late submission will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld**
4. **Cases of plagiarism will be penalized**
5. **The assignment should be bound in an appropriate style (Comb Bound or Stapled)**

6. **Where the assignment should be submitted in both hardcopy and softcopy, the softcopy of the written assignment and source code (where appropriate) should be on a CD in an envelope/ CD cover and attached to the hardcopy.**

7. **You must obtain 50% overall to pass this module.**

Acknowledgement

I take this opportunity to thank Asia Pasic University for giving me chance to do this project. I would like to express my profound gratitude and deep regards to my teachers Mr. R.N. Thakur and for the exemplary guidance, monitoring and constant encouragement throughout the course of this project. The project helped me learn how to do proper Data Analysis and I learned about many new things while doing the project.

A special acknowledgement goes to my colleagues who helped me in completing the project by exchanging interesting ideas and sharing their experience.

Table of Contents

List of Figures

1. Introduction

As a part of our in course assignment of Programming for Data Analysis we are assigned to perform data analysis operations i.e. Data Manipulation, Data Exploration and Data Visualization in a Hourly weather dataset. The analysis is to be done using R programming language. R programming was developed in 1993 by Ross Ihaka and Robert Gentlemen. R programming is widely used for statistical analysis. It includes intensive amount of library and packages that helps in performing data analysis, visualization and exploration. It comes with different machine learning algorithms, linear regression, time series and many more tools. R is mainly used for statistical interference, data analysis and machine learning algorithms.

For this assignment we have to use an hourly weather dataset. This data set contains data of weather in hourly basis. In this assignment we have to apply different data visualization, exploration and manipulation techniques. The data set consists meteorological data from two different airport stations they are LaGuardia Airport (LGA) and John F. Kennedy International Airport.

2. Assumptions

For this assignment we are provided with an hourly dataset which includes meteorological data from two different airport stations. The data is recorded for LaGuardia Airport (LGA) and John F. Kennedy International Airport. The dataset consists 17,412 rows and 15 columns. The dataset includes columns like origin of weather station, year, month, day and hour for time of recording the data, Temperature and dewpoint in Fahrenheit, relative humidity, wind direction, wind speed and wind gust speed in miles per hour, precipitation, pressure, visibility and time_hour with full date and time of recording the data. There are several rows with Not Available (NA) values which requires proper data cleaning for precise analysis. For this analysis proper data manipulation, exploration and visualization techniques is to be followed. The result of every analysis is used to take decisions for flights in airports.

3. Analysis 1

```
#Analysis 1
ggplot(data = data, mapping = aes(x = temp, fill = origin)) +
  geom_histogram() +
  labs(title = "Histogram of temperature (JFK)", x="Temperature (F)") +
  theme(panel.grid = element_blank(),
        plot.title = element_text(hjust = 0.5, size = 16)) +
  facet_wrap(~month)
```

The above code is of analysis 1. In this analysis I have used temperature column from the dataset and drawn a histogram. Here I have used ggplot from ggplot2 package for histogram. Here, I have passed 'data' data frame which consists the data imported from csv file. And as an x-axis parameter the temperature column is passed. The fill property is used for coloring histogram bins. The geom_histogram( ) function is used to draw histogram from the data passed from gglpot function. The labs function is used form defining the title, x-axis and y-axis of the plot. The theme function is used to customize the plot. I have removed the panel grid with the property element_black(). And the title of the plot is placed at center with the hjust property in 'plot.title'. The facet wrap function is used to divide plot into different facets or groups. Here we have divided plot into subplots on monthly basis.

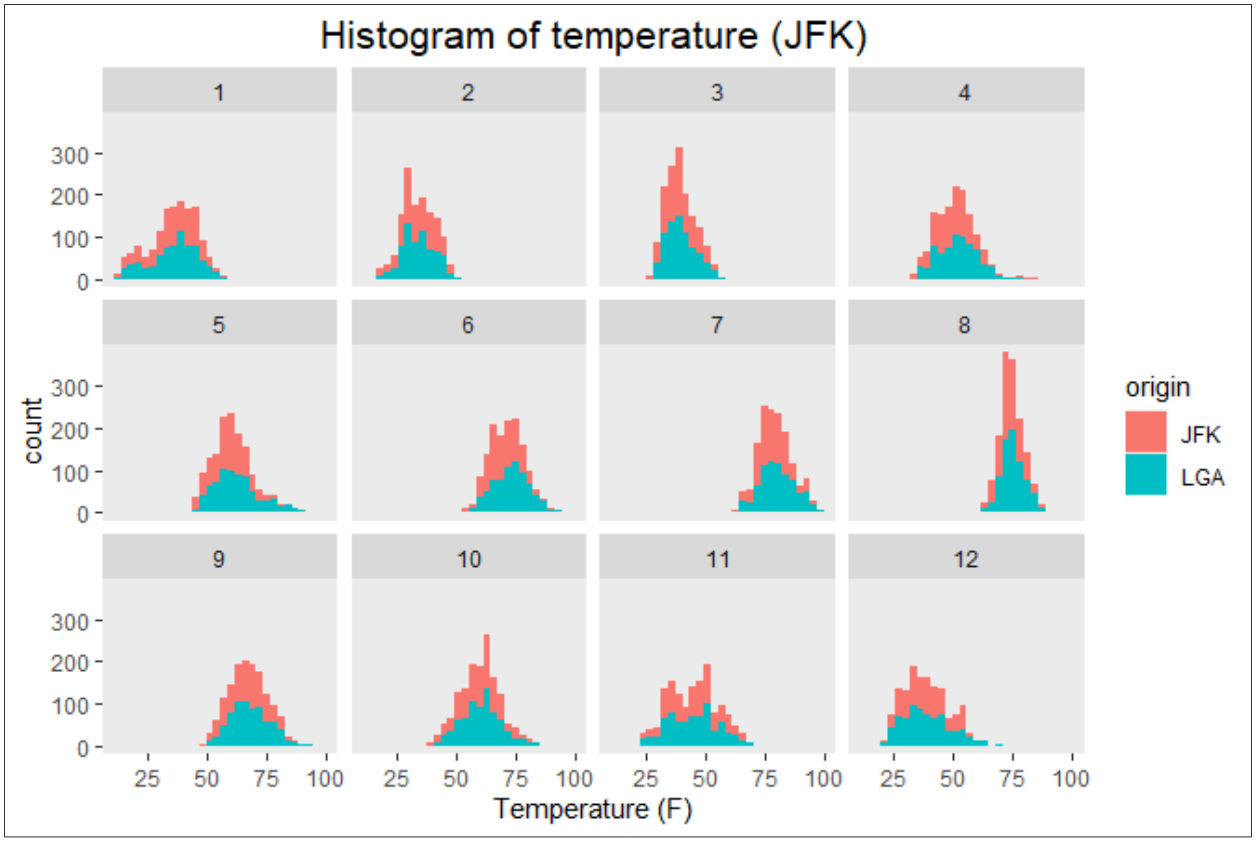

Figure 1 Histogram of Temperature

Output

The above analysis concludes by giving the information about temperature variations in different months. From the above plot we can conclude that the JFK airport is hotter than the LGA airport throughout the year. So we can assume that LGA may be any hill station or cooler place than the JFK. The above histogram shows coolest months are first three months, November and December and rest of all are hotter.

4. Analysis 2

```
#Analysis 2
plot(jfkData$humid, jfkData$temp, main="Scatterplot against Temp & Humid. (JFK)",
     xlab = "Humidity",
     ylab="Temperature", pch = 20, col = "turquoise3")
abline(lm(jfkData$temp ~ jfkData$humid), col = "blue")
legend("topleft", legend = paste("cor = ", round(cor(jfkData$humid, jfkData$temp), 2),
                                 sep = ""),lty = 1, col = "blue")

plot(lgaData$humid, lgaData$temp, main="Scatterplot against Temp & Humid. (LGA)",
     xlab = "Humidity",
     ylab="Temperature", pch = 20, col = "turquoise3")
abline(lm(lgaData$temp ~ lgaData$humid), col = "blue")
legend("topleft", legend = paste("cor = ", round(cor(lgaData$humid, lgaData$temp),
                                                    2), sep = ""),lty = 1, col = "blue")
```

This analysis is for checking the correlations between Humidity and temperature. Here I have drawn two different scatter plots for JFK and LGA. The plot function is used to draw scatter plot. Here two data frames are taken one for JFK and other for LGA. I have also calculated the correlation between humidity and temperature. The abline function is used to draw the correlation equation in the plot. The legend function is used to add legend to the plot. Here I have included the correlation value in the legend of this plot. I have used cor function to calculate correlation between humidity and temperature.

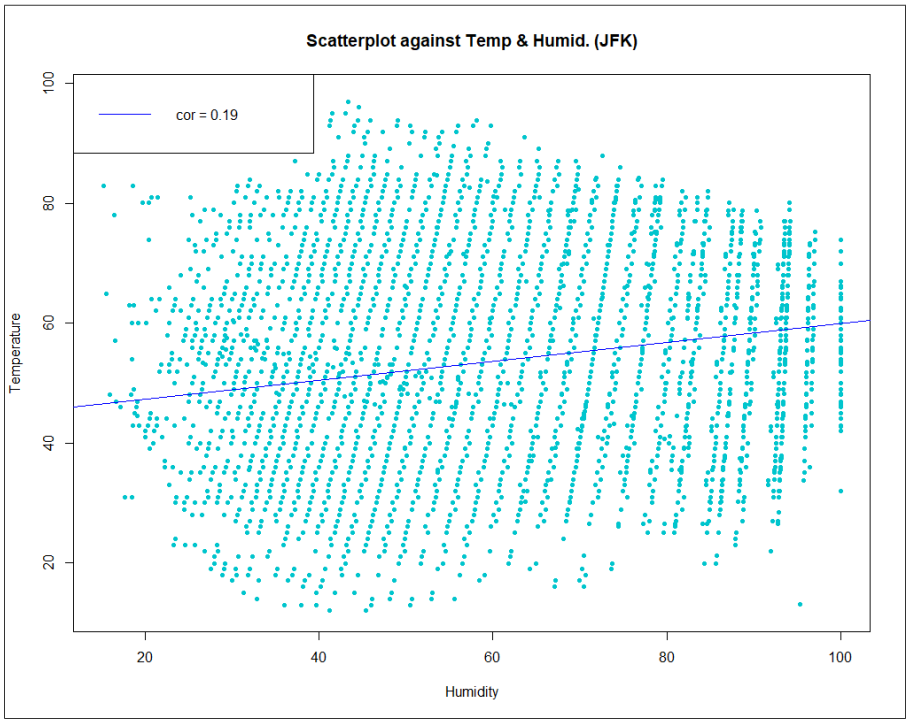

Figure 2 Scatterplot for Humidity and temperature (JFK)

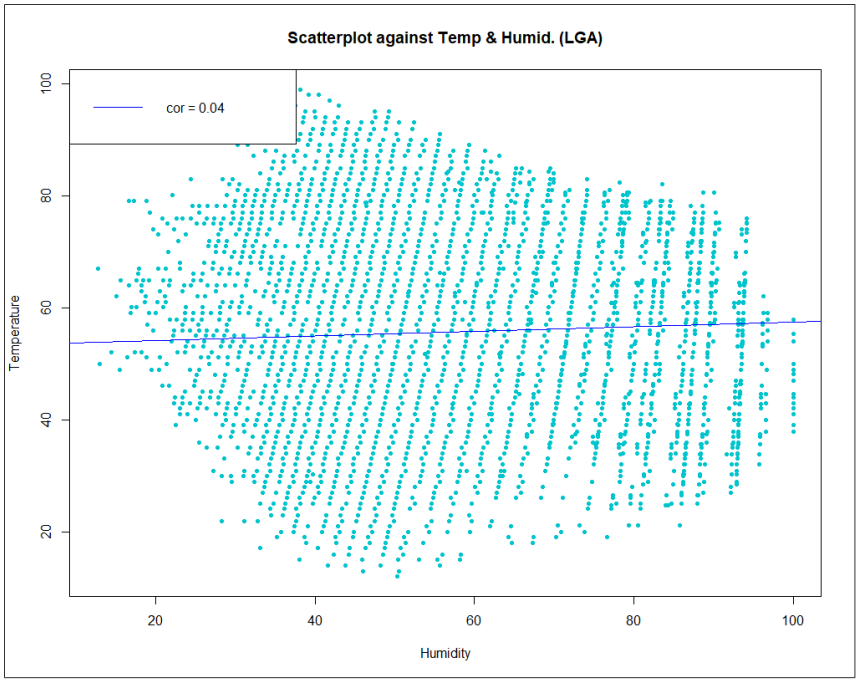

Figure 3 Scatterplot for Temperature and Humidity (LGA)

Output

The above analysis outputs two scatter plots. These plots gives us idea about the relationship betwee Temperature and humidity. Here the regression line clearly states about the relations of the temperature against humidity. The correlation between humidity and temperature of LGA is 0.04 and JFK is 0.19. This shows that temperature and humidity doesn't have relations between each other. From this analysis we can conclude that humidity and temperature are independent of each other.

5. Analysis 3

```
#Analysis 3
plot(jfkData$temp, jfkData$dewp, main="Scatterplot against Temp & Dewp. (JFK)",
     xlab = "Temperature",
     ylab="Dewpoint", pch = 20, col = "turquoise3")
abline(lm(jfkData$dewp ~ jfkData$temp), col = "blue", lwd=2)
legend("topleft", legend = paste("cor = ", round(cor(jfkData$temp,
                          jfkData$dewp), 2), sep = ""),lty = 1, col = "red")

plot(lgaData$temp, lgaData$dewp, main="Scatterplot against Temp & Dewpoint (LGA)",
     xlab = "Temperature",
     ylab="Dewpoint", pch = 20, col = "turquoise3")
abline(lm(lgaData$dewp ~ lgaData$temp), col = "blue", lwd=2)
legend("topleft", legend = paste("Corr. = ",
                                 round(cor(lgaData$temp, lgaData$dewp), 2),
                                 sep = ""),lty = 1, col = "red")
```

The above code is for analysis 3. Here two scatterplots are drawn for both LGA and JFK. The scatterplots are drawn on temperature and dew point columns. The plot function is used to draw scatterplots. Here, abline is used to draw the regression line in the plot. Regression line is calculated between dew point and temperature using the lm method. Then for legend I have included the correlation value of temperature and dew point.

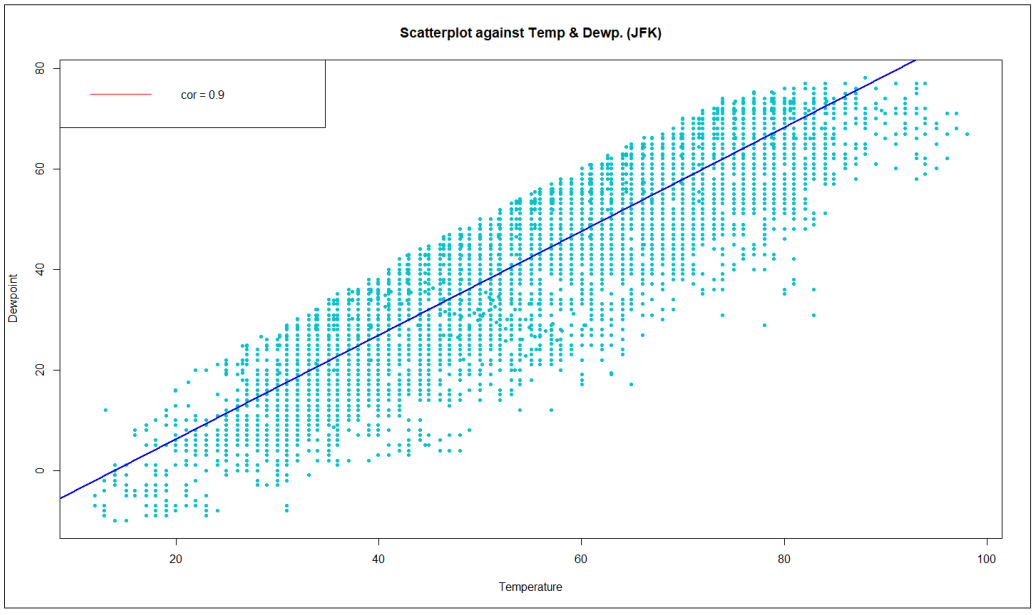

Figure 4 Scatterplot against Temperature and Dew point (JFK)

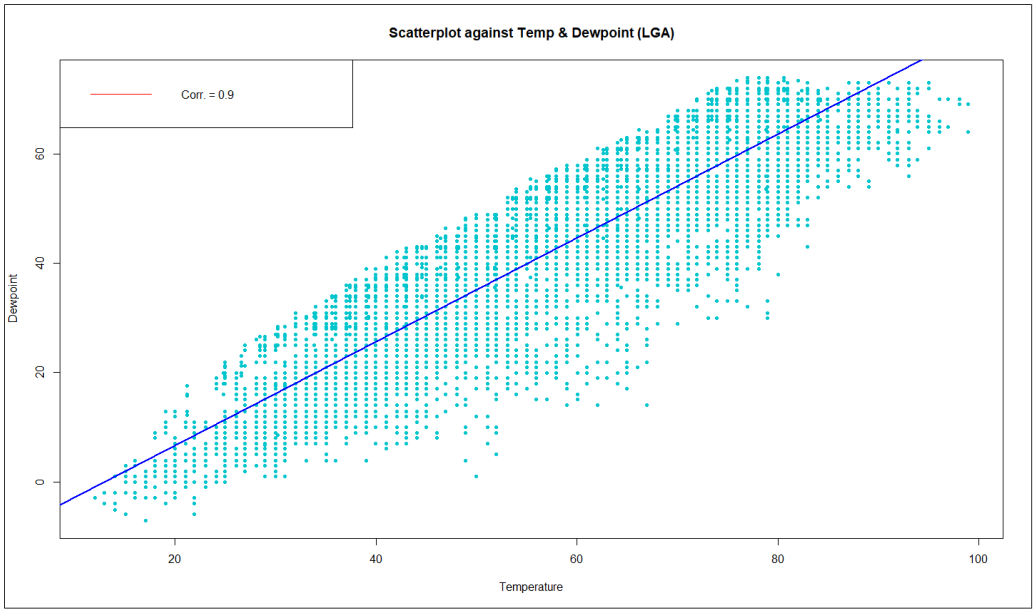

Figure 5 Scatterplot of Temperature and Dew point (LGA)

Output

The above analysis is concluded with the finding of linear relationships between Temperature and dewpoint for both LGA and JFK airports. Here, the value of correlation of JFK and LGA both is 0.9. From this we can conclude that dewpoint and temperature are highly dependent to each other.

6. Analysis 4

```
#Analysis 4
ggplot(data = data, mapping = aes(x = humid, y = dewp, color = origin)) +
  geom_point(alpha = 0.2)+ stat_smooth(method = "lm") +
  labs(title = 'Scatter Plot of Dew Point against humidity',
       x = 'Humidity ', y = 'Dew Point') +
  theme(legend.position = c(0.9, 0.2))
  legend("topleft", legend = paste("Corr. (JFK) = ",
  round(cor(jfkData$humid, jfkData$dewp),
  2), sep = ""), lty = 1, col = "red")
```

The above analysis is done by making scatterplot using ggplot geom_point function against humidity and dew point. Here data is passed as data frame in the data attribute and for mapping humidity and dewp columns are passed. The color attribute is used to differentiate color between JFK and LGA. The labs function is used to add title to the plot and also x-axis and y-axis label. The theme function is used to customize the plot.

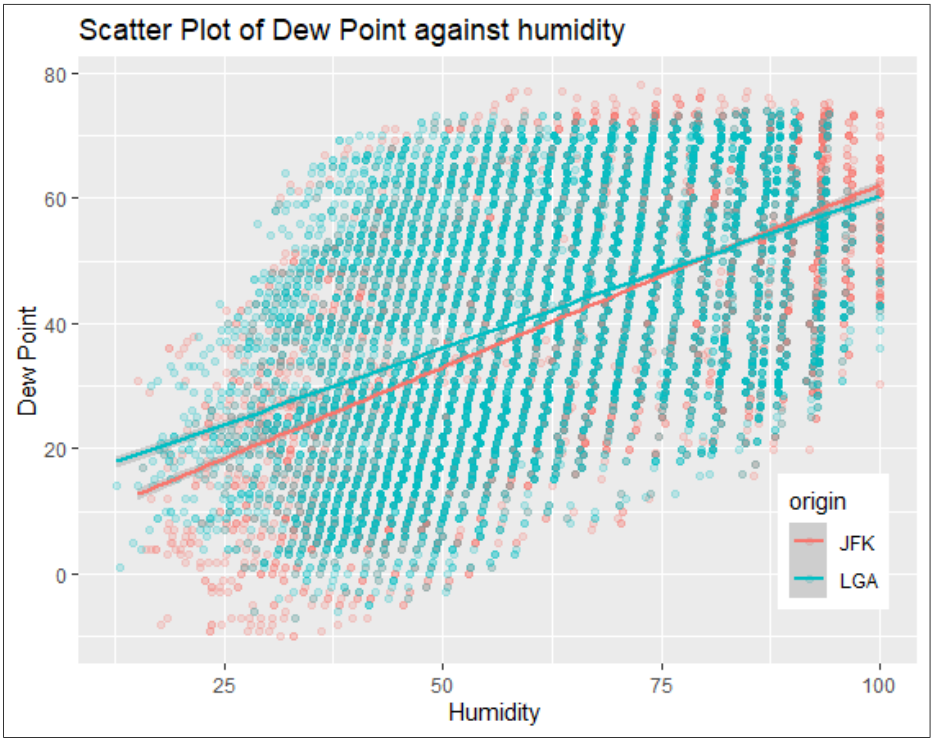

Figure 6 Scatterplot of Dew point against Humidity

Output

The above analysis is shows us that the dew point and humidity are dependent on each other. By looking at the regression line on the graph we can assume that they are both dependent on each other.

7. Analysis 5

```
ggplot(data = data, mapping = aes(x = humid,  fill = origin)) +
  geom_histogram() +
  labs(title = "Histogram of Humidity", x = "Humidity", y = "Count") +
  theme(panel.background = element_rect(fill = "grey94"),
        panel.grid = element_blank()) +
  facet_wrap(~origin)
```

The above code produces histogram for humidity. The histogram is divided into two facets using facet_wrap. One facet is for jfk and other for lga. The geom_histogram is used to draw histogram in the plot. The ggplot function takes data frame and the humid column in that data frame. The labs function defines the title and the axis labels. Theme function is used to customize the grid of the plot. The grid layout is removed from background and filled with grey94 color.

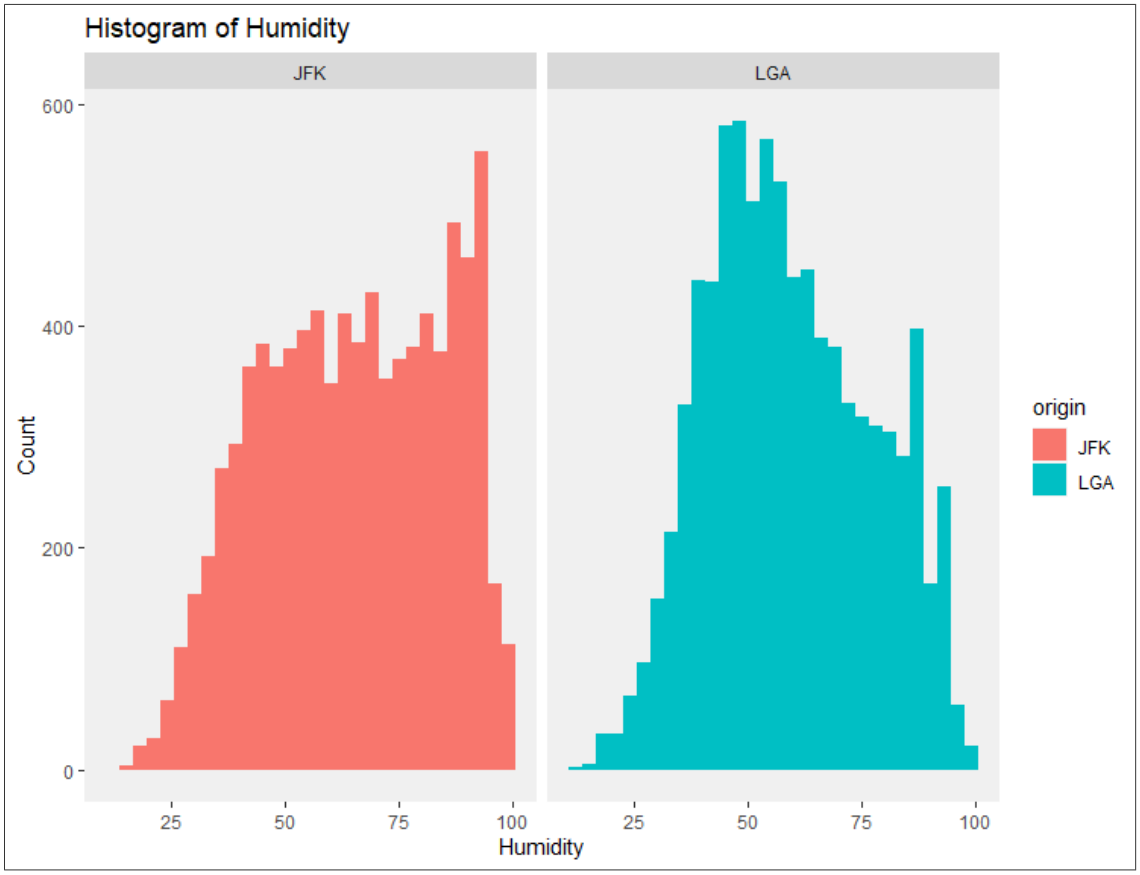

Figure 7 Histogram of Humidity

Output

The above analysis gives information about the humidity variations in both jfk and lga airports. This shows that humidity is very often high in both the airports. The above diagram shows that the average humidity of JFK is around 90-95 and that of LGA is around 40-50.

8. Analysis 6

```
#Analysis 6
prep<-na.omit(data)
prep
prep<-na.omit(prep)
prep

ggplot(data = prep, mapping = aes(x = na.omit(pressure) )) +
  geom_histogram(color = "black", fill = "aquamarine2") +
  labs(title = "Histogram of Pressure", x="Pressure") +
  facet_wrap(~origin)
```

The above code is for displaying histogram of Pressure data from both airports. Firs NA values are omitted in the data frame using na.omit function. This plot is divided into two facets based on the origin i.e. lga and jfk. Labs is used to add title and axis labels in the plot.

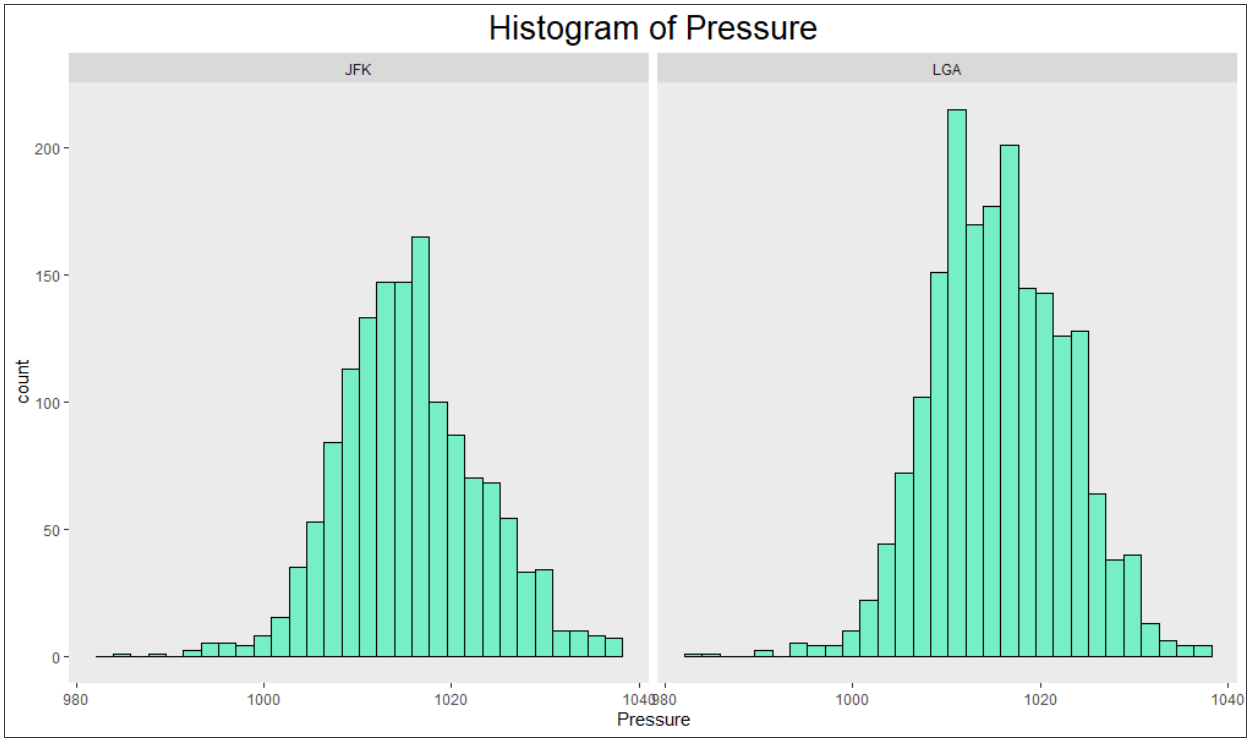

Figure 8 Histogram of Pressure

Output

The above analysis gives us information about the precipitation variations in both the airports. The above graph shows that the average pressure of JFK is around 1010-1015 and that of LGA is around 1008-1010.

9. Analysis 7

```
#Analysis 7
prep
ggplot(data = prep, aes(x = wind_speed, y = wind_gust, color = origin)) +
  geom_point(alpha = 0.2) + stat_smooth(method = "lm") +
  labs(title = "Scatter Plot of Wind Speed vs Wind gust Speed",
       x = "Wind Speed (mph)", y= 'Wind Gust speed (mph)')
```

The above code is for seventh analysis where scatterplot is drawn between wind speed and wind gust. The gglplot function takes x and y axis data as well as color parameter to differentiate both the airports data with each other. Here labs is used to add title to the plot. The alpha attribute of geom_point functions defines opacity level of the scattered points. I have used method "lm" for regression line in the plot.

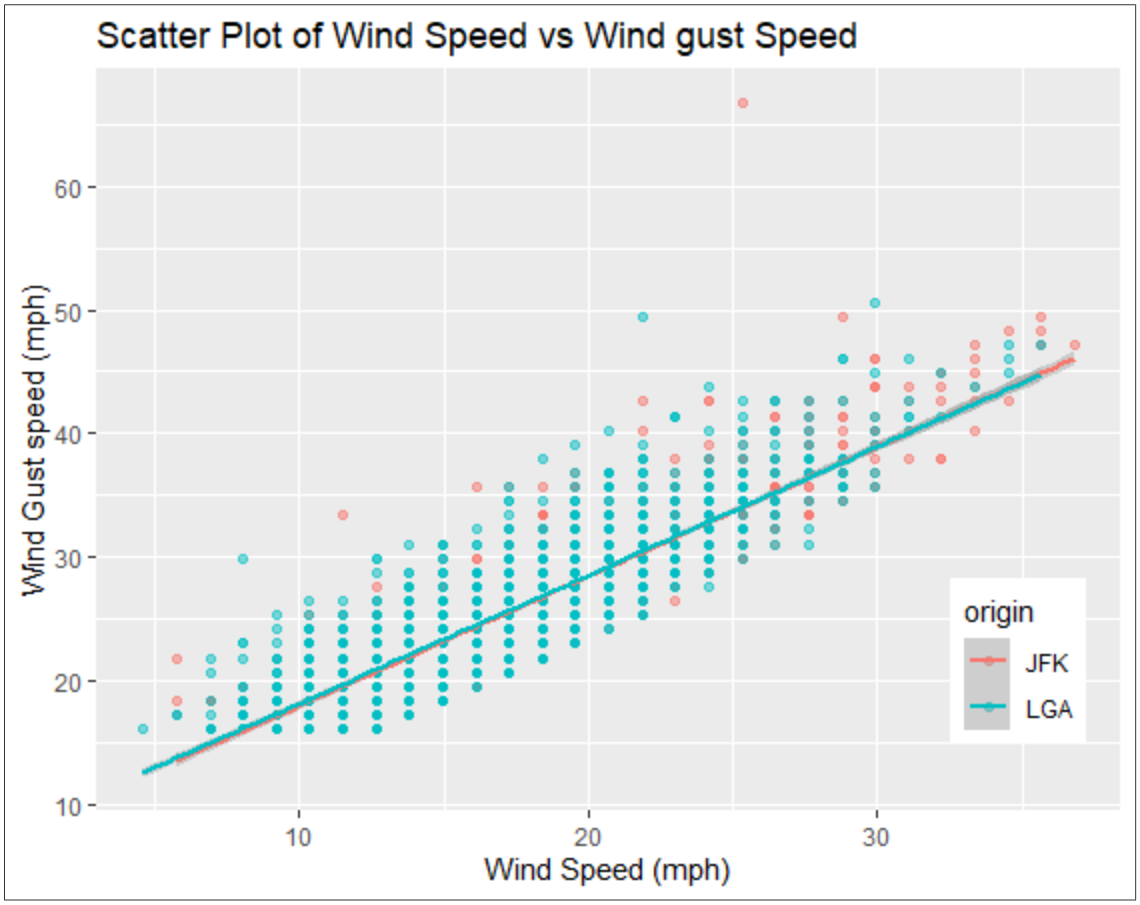

Figure 9 Scatterplot of Wind Speed vs Wind gust speed

Output

From above analysis it is clear that wind speed and wind gust have relations with each other. With the increase in wind speed the wind gust also increases. There may be several flight delays if wind gust is higher. Here we can compare wind speed and wind gust between both the airports.

10. Analysis 8

```
# Analysis 8
ggplot(data = prep, aes(x = wind_speed, y = wind_gust, color = origin, shape = origin)) +
  geom_point(alpha = 0.5) + stat_smooth(method = "lm", se=FALSE) +
  labs(title = "Scatter Plot of Wind Speed vs Wind gust Speed",
       x = "Wind Speed (mph)", y= 'Wind Gust speed (mph)') +
  facet_wrap(~month)
```

The above code produces scatterplot against wind speed and wind gust. The whole plot is divided into 12 facets or groups on the basis of months. The color and origin parameters in the ggpplot defines color and shapes of the scatter points in the graph. The alpha property in geom_plot method is used to define the opacity of the dots in graph.

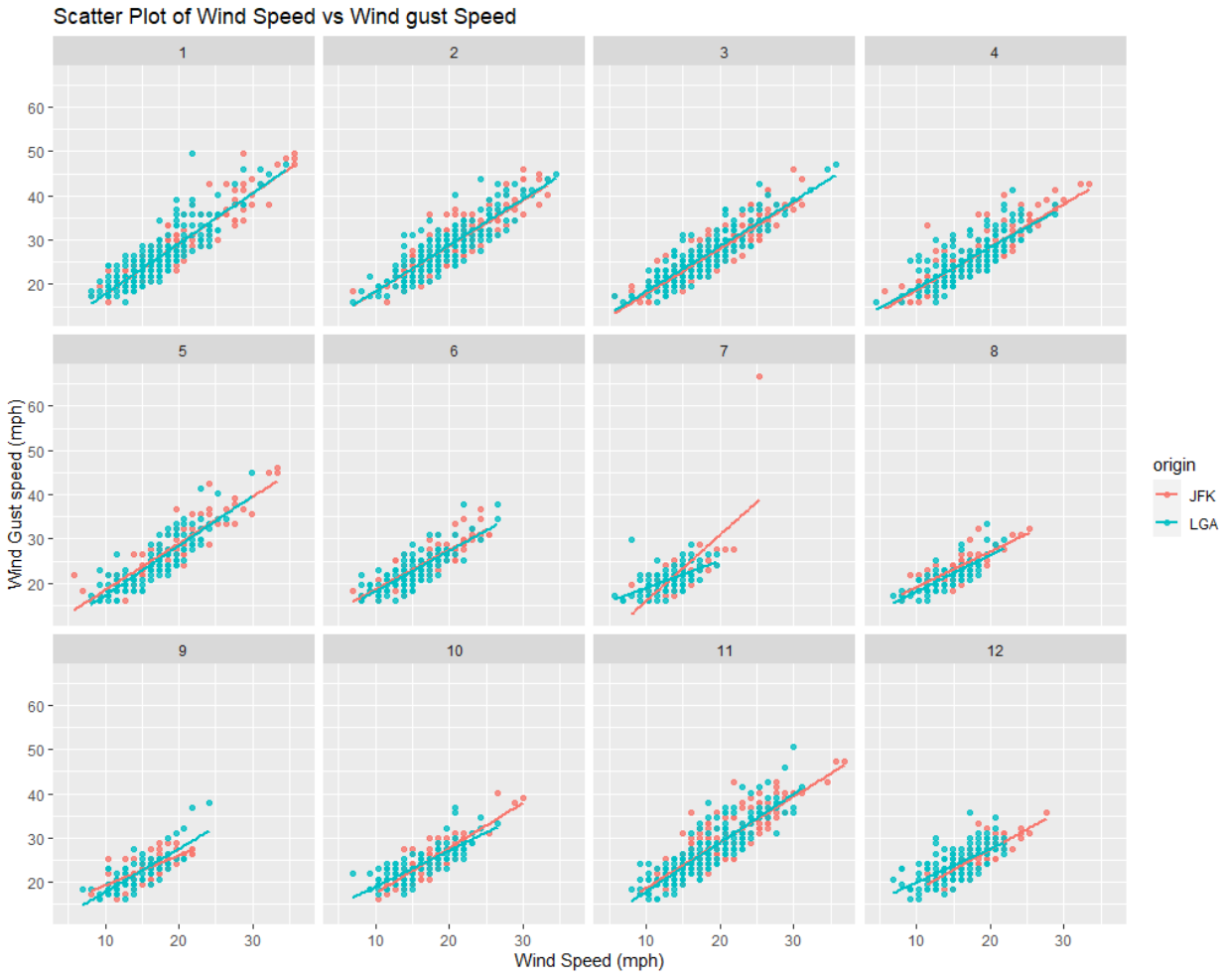

Figure 10 Scatterplot of Wind Speed and Wind Gust speed (Monthly basis)

Output

The above analysis provides visual information about the wind speed and wind gust relations on monthly basis. We can differentiate the data between jfk and lga with the colors. The above graph shows relationship between wind gust and wind speed. From the graph we can analyze that wind gust and wind speed have some relations between them. Wind gust speed and wind speed is higher in the month of January to March.

11. Analysis 9

```
#Analysis 9
ggplot(data = prep, mapping = aes(x = visib, color = origin, fill = origin)) +
  geom_histogram() +
  labs(title = "Histogram of Visibility in Miles", x = "Visibility (miles)") +
  facet_wrap(~month)
```

The above code gives us histogram for visibility on both airports. The complete plot is divided into 12 different facets for 12 months. Data for lga and jfk are differentiated using colors and fill properties. Facet wrap is used for dividing plot intro sub groups.

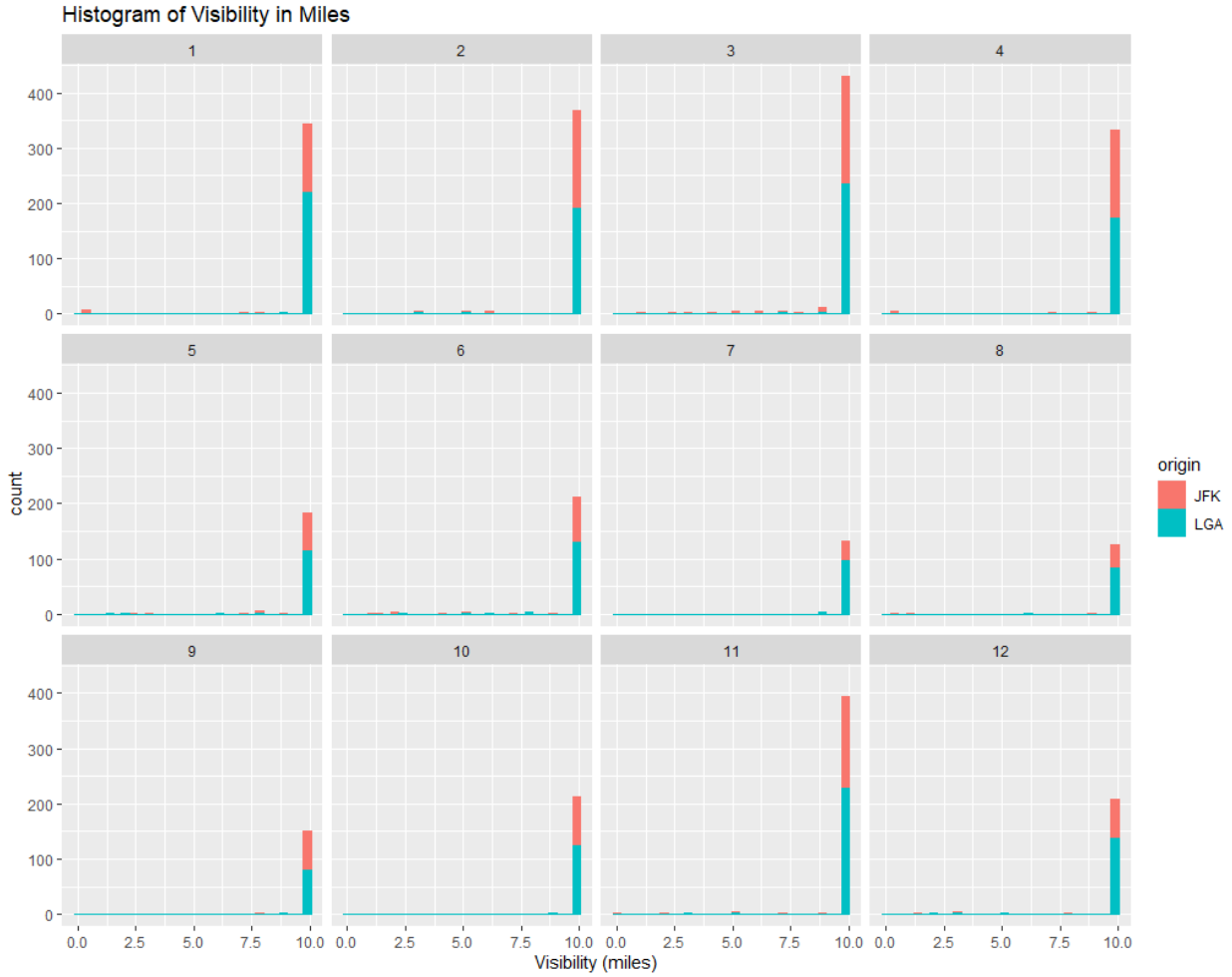

Figure 11 Histogram of Visibility in Miles

Output

The above result shows the visibility of both the airports. The visibility is less in mid quarter year than other months. It tells us that there might occur issues to flights when the visibility is low. The highest visibility falls in the month of March and November. And the lowest of visibility is in the month of July to September.

12.Analysis 10

```
#Analysis 10
ggplot(data = prep, mapping = aes(x=factor(month), y = humid, na.rm = TRUE,
  color = month, fill = month)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(alpha = 0.4,color = "tomato") +
  labs(title = "Box plot Humidity against Month", x = "Month", y = "Humidity") +
  facet_wrap(~origin)
```

The above code is used to build a boxplot for Humidity against Month. The code divides graph into two facets, one for JFK and the other for LGA. Data is passed in the ggplot method as data property and then mapped using mapping method. Here month is used as factor that affects the boxplot. The color palette is defined based on month.

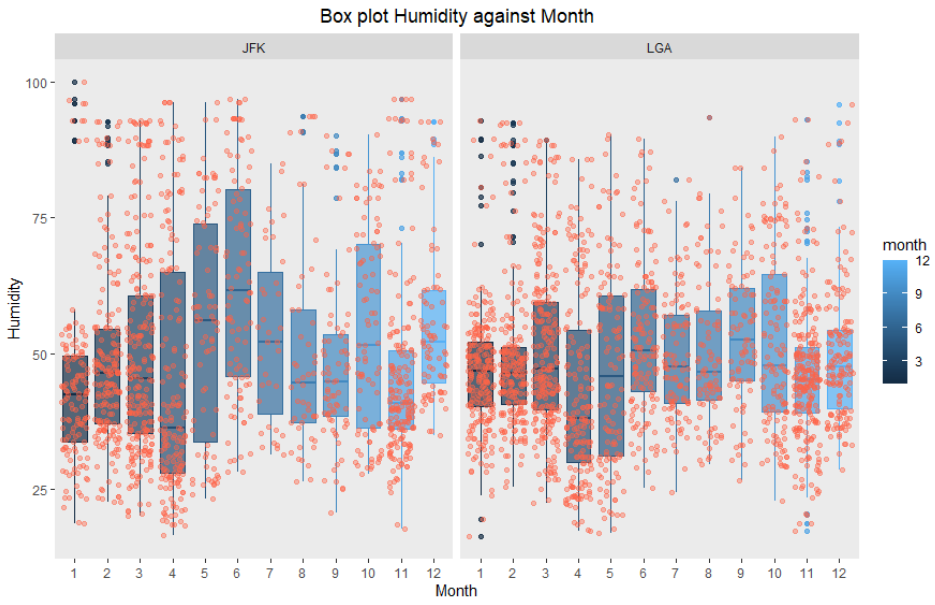

Figure 12 Box plot of Humidity against Month

Output

The above graph tells us that humidity is high during the first three months. So we can assume that humidity is higher in winter season than in the summer. Also, on comparison with jfk, lga has higher humidity. Box plot is used to show skewness of the data.

13.Analysis 11

```
#Analysis 11
ggplot(data = data, mapping = aes(x = factor(month), y = wind_speed, na.rm = TRUE,
                                  color = origin)) +
  geom_boxplot(alpha = 0.8) +
  labs(title = "Box plot Wind Speed against Month", x = "Month",
       y = "Wind Speed (Mph)") +
  facet_wrap(~origin) +
  # guides(fill = guide_legend(reverse = TRUE)) +
  scale_fill_brewer(palette = "Greens", name = "Origin") +
  theme(plot.title = element_text(hjust = 0.5, size = 20),
        axis.ticks = element_blank(),
        panel.border = element_blank(),
        panel.grid = element_blank(),
        panel.background = element_rect(fill = "bisque"),
        axis.text = element_blank(),
        axis.title = element_text(size = 14),
        legend.title = element_text(size = 16))
```

The above code is for generating boxplot of windspeed against month. geom_boxplot function of ggplot2 package is used to make the boxplot. I have passed month and wind_speed as the x and y parameter in the ggplto function. The na.rm is used to remove any row with NA values. The theme function is used to customize the graph. plot.title is used to adjust size and position of the title. The graph is divided into two facets one for jfk and the other for the lga.

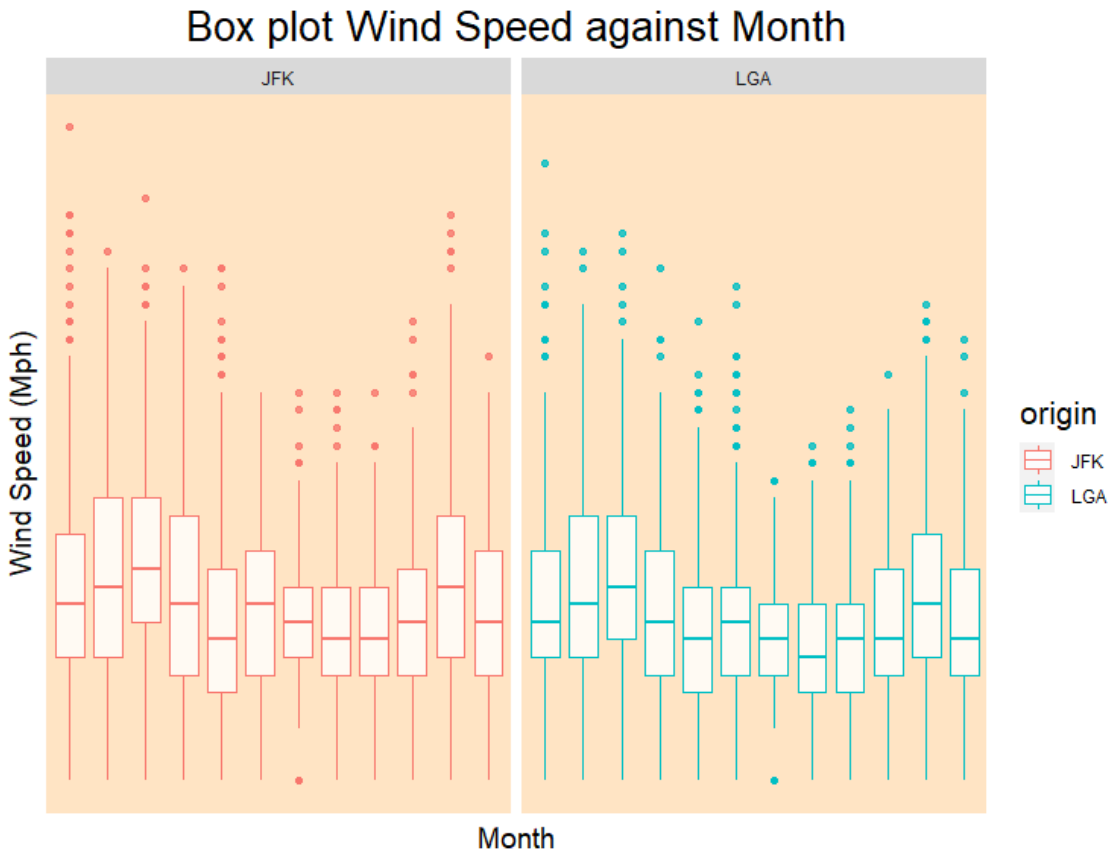

Figure 13 Box plot Wind speed against Month

Output

The above graph gives result on Wind speed against month. Boxplot helps us analyze the skewness of the data distribution. We can analyze the Q1 (First Quartile), Q3 (Third Quartiles), median, minimum and maximum values of each distribution.

14.Analysis 12

```
#Analysis 12
ggplot(data = data, mapping = aes(x = factor(month) , y = wind_gust, na.rm = TRUE,
  color = origin)) +
  geom_boxplot(alpha = 0.8) +
  labs(title = "Box plot Wind Gust against Month", x = "Month",
       y = "Wind Gust Speed (Mph)") +
  facet_wrap(~origin) +
  # guides(fill = guide_legend(reverse = TRUE)) +
  scale_fill_brewer(palette = "Greens", name = "Origin") +
  theme(plot.title = element_text(hjust = 0.5, size = 20),
        axis.ticks = element_blank(),
        panel.border = element_blank(),
        panel.grid = element_blank(),
        panel.background = element_rect(fill = "lightblue1"),
        axis.text = element_blank(),
        axis.title = element_text(size = 14),
        legend.title = element_text(size = 16))
```

This analysis is done by producing boxplot for wind gust speed against month. The plot is divided into two facets for both the airports. geom_boxplot function is used to make a box plot chart. Here month is taken as a factor for the boxplot. The NA values are removed using 'na.rm = TRUE'.

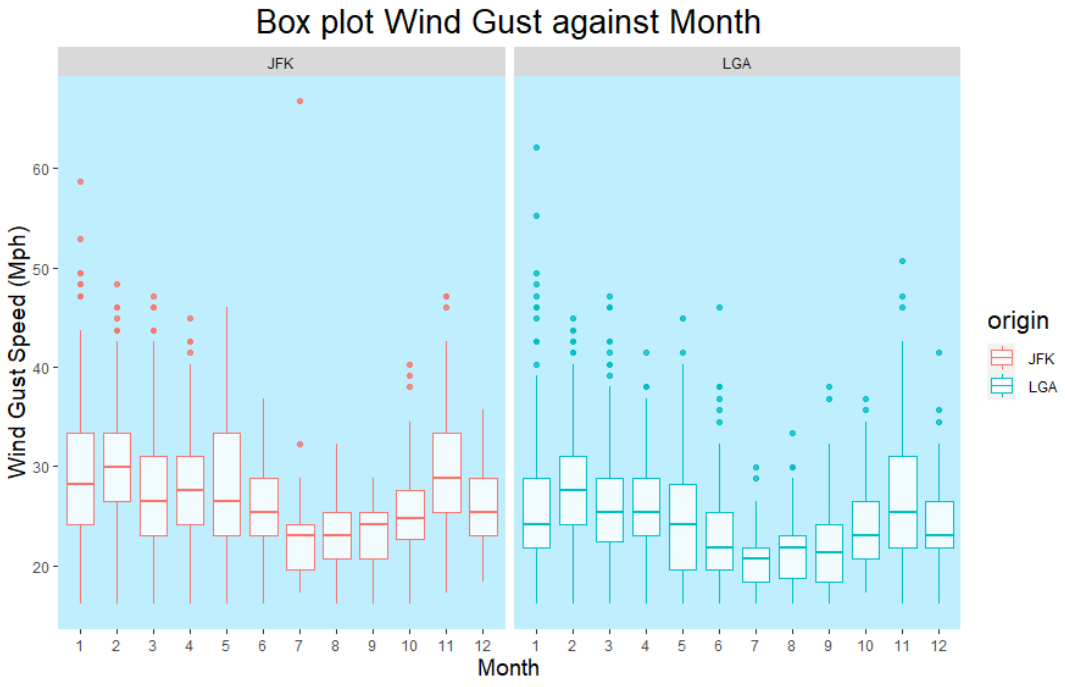

Figure 14 box plot Wind Gust speed against Month

Output

The above graph shows information of Wind Gust speed with respect to Month. The graph helps in analyzing the skewness of the data. Here the graph shows that the average wind gust speed is about 24-26 in JFK airport throughout the year and that of LGA is around 21-23.

15. Analysis 13

```
#Analysis 13
ggplot(data = data, mapping = aes(x = factor(month) , y = dewp, na.rm = TRUE,
                                  color = origin)) +
  geom_boxplot(alpha = 0.8) +
  labs(title = "Box plot Dewpoint against Month (F)", x = "Month", y = "Dewpoint (F) ") +
  facet_wrap(~origin) +
  # guides(fill = guide_legend(reverse = TRUE)) +
  scale_fill_brewer(palette = "Greens", name = "Origin") +
  theme(plot.title = element_text(hjust = 0.5, size = 20),
        axis.ticks = element_blank(),
        panel.border = element_blank(),
        panel.grid = element_blank(),
        panel.background = element_rect(fill = "lightblue1"),
        axis.text = element_blank(),
        axis.title = element_text(size = 14),
        legend.title = element_text(size = 16))
```

The above code produces two boxplot diagrams for JFK and LGA. The boxplot is of dew point against month. Here, NA values are removed using 'na.rm' attribute. The facet wrap function splits the diagram in two both for jfk and lga. The theme function is used for customizing the plot. Here plot title is displayed in the center and size. Axis ticks are removed from the plot. The background color is changed using the element_rect function.

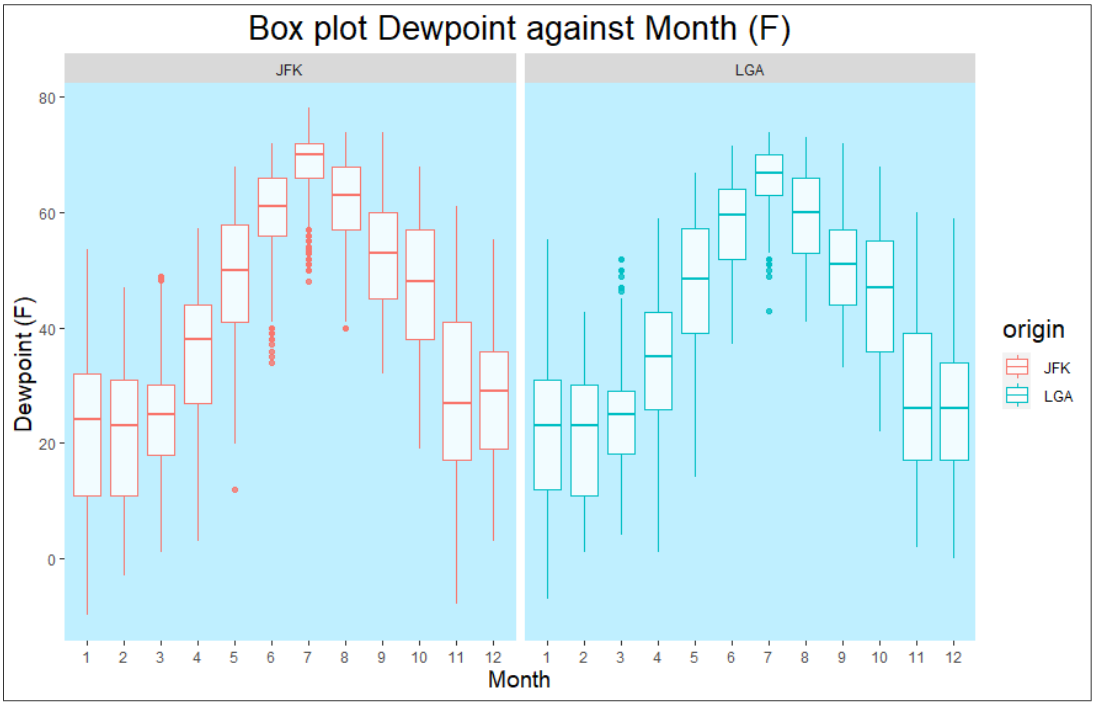

Figure 15 Box plot Dew point against Month (F)

Output

The above graph shows information of Wind Gust speed with respect to Month. The graph helps in analyzing the skewness of the data. The above graph shows that the data is highly negative skewed.

16. Analysis 14

```
#Analysis 14
precipData<- subset(data, precip > 0)
precipData
ggplot(data = precipData, mapping = aes(x = factor(visib) , y = precip, na.rm = TRUE)) +
  geom_boxplot(alpha = 0.8) +
  labs(title = "Box plot Precipitation against Visibility", x = "Visibility",
       y = "Precipitation (F) ") +
  facet_wrap(~origin) +
  # guides(fill = guide_legend(reverse = TRUE)) +
  scale_fill_brewer(palette = "Greens", name = "Origin") +
  theme(plot.title = element_text(hjust = 0.5, size = 20),
        axis.ticks = element_blank(),
        panel.border = element_blank(),
        panel.grid = element_blank(),
        panel.background = element_rect(fill = "lightblue1"),
        axis.text = element_blank(),
        axis.title = element_text(size = 14),
        legend.title = element_text(size = 16))
```

The code is producing boxplot for precipitation against visibility. Data is taken only from those rows whose precipitation value is greater than 0. Then the subset dataframe is passed in to the data property of the ggplot method.

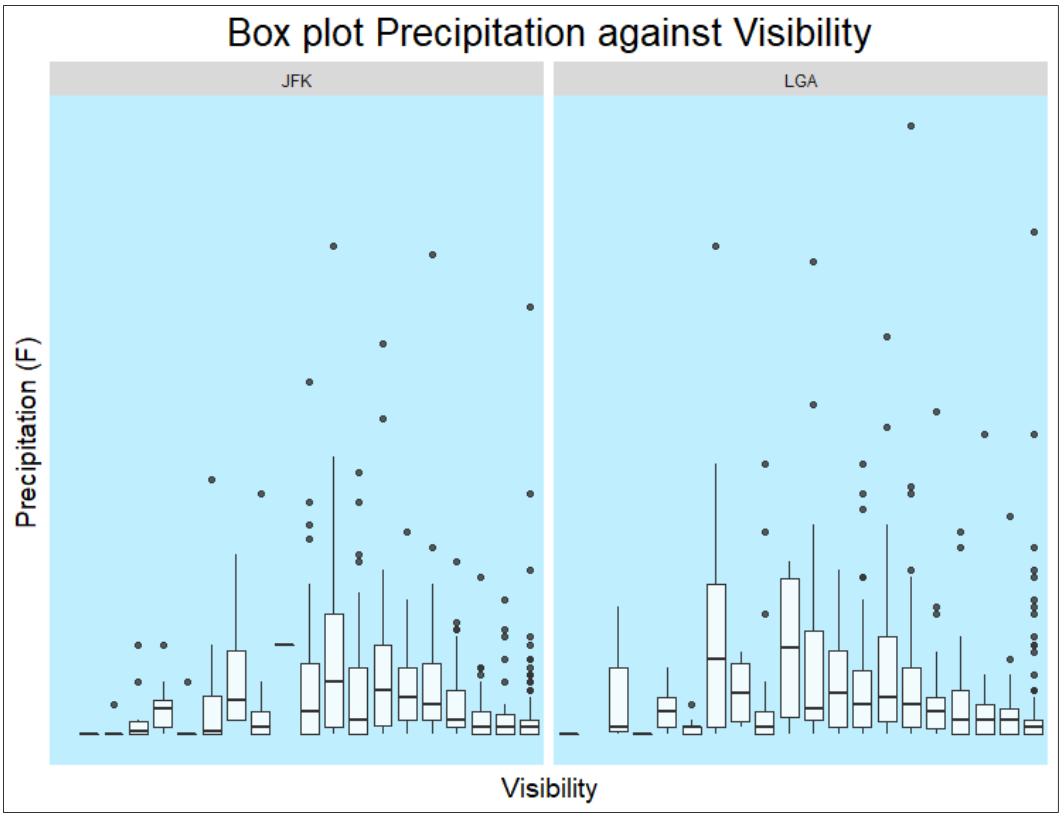

Figure 16 Boxplot for Precipitation against visibility

Output

The above graph shows information of Wind Gust speed with respect to Month. The graph helps in analyzing the skewness of the data. The above graph helps in determining Quartiles, lower and upper quartiles, minimum and maximum values.

17.Extra feature 1

```
# Extra feature 1
ggplot(data = data, aes(x = wind_speed)) +
  geom_density( aes(fill = origin, color = origin, alpha = 0.1) ) +
  labs(title = "Weight density graph of Wind Speed(Mph)",
       x = "Wind Speed (Mph)", y = "Density") +
  theme(plot.title = element_text(hjust = 0.5, size = 16),
        panel.grid = element_blank(),
        # panel.background = element_rect(fill = "lightblue")
        )
```

The above graph produces a density graph. This analysis is made as an extra feature in this assignment. For building this graph I have used geom_density method of ggplot2 package. Here, data from the csv file is passed as the data property and the value of x-axis is given as wind speed. The geom_density then builds plot. Here both airports are given different colors using fill and color property and the alpha is used to define opacity level of the graph.

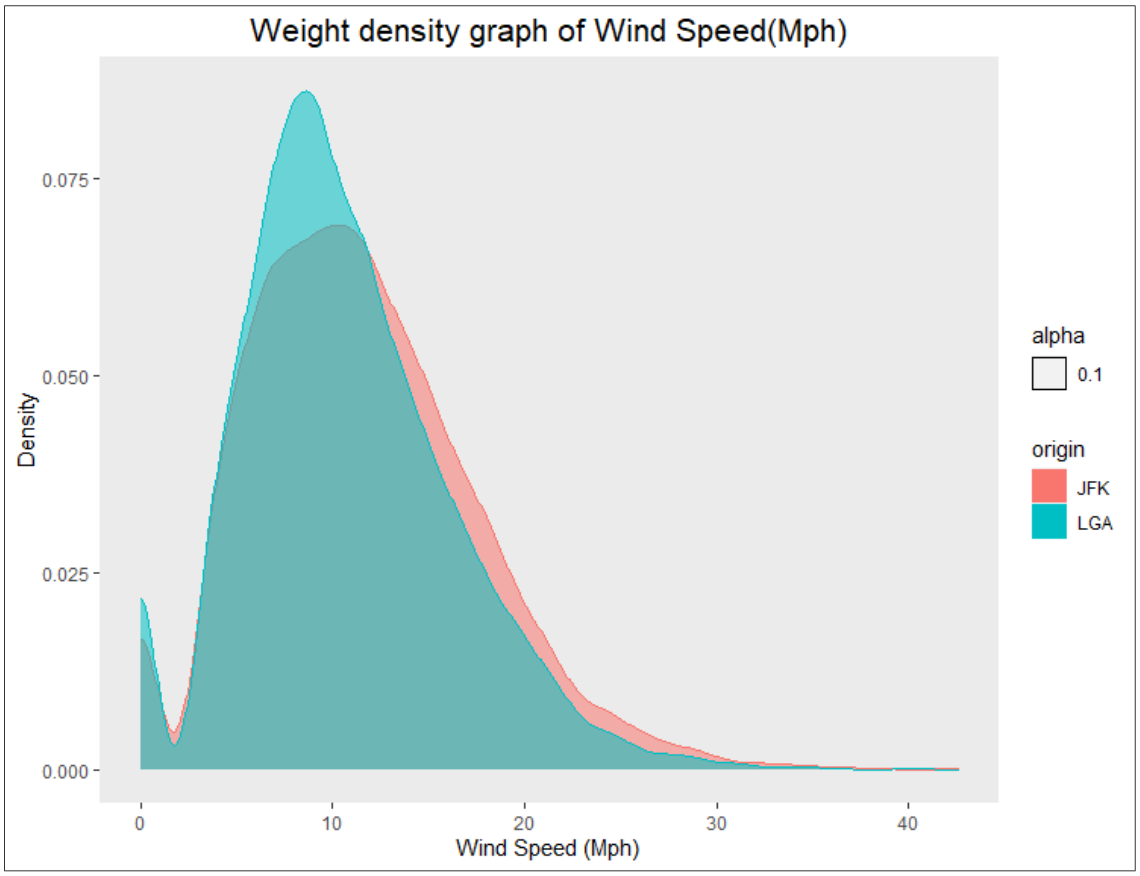

Figure 17 Weight density graph of wind Speed (Mph)

Output

This graph shows the density level of the wind speed. From this graph we can analyze that the average wind speed for JFK airport is around 12-15 and that of LGA is around 8-10. This shows that on average JFK has higher wind gust speed than the LGA airport.

18.Extra Feature 2

```
#Extra feature 2
ggplot(data = data, aes(x = wind_speed, y = temp)) +
  geom_bin2d(bins = 70) +
  scale_fill_continuous(type = "viridis") +
  labs(title = "Heat map of Temperature against Wind Speed") +
  theme_bw()
```

As the second extra feature I have made heat map for temperature against Wind speed. I have passed wind_speed as x-axis value and temp as the y-axis value. Here geom_bind2 is used for making the heatmap. The bins property is used to define the number of bins to be included in the graph. The scale_fill_continuos function is used to define the color and the heat map type. I have used the 'viridis' type for this heat map.

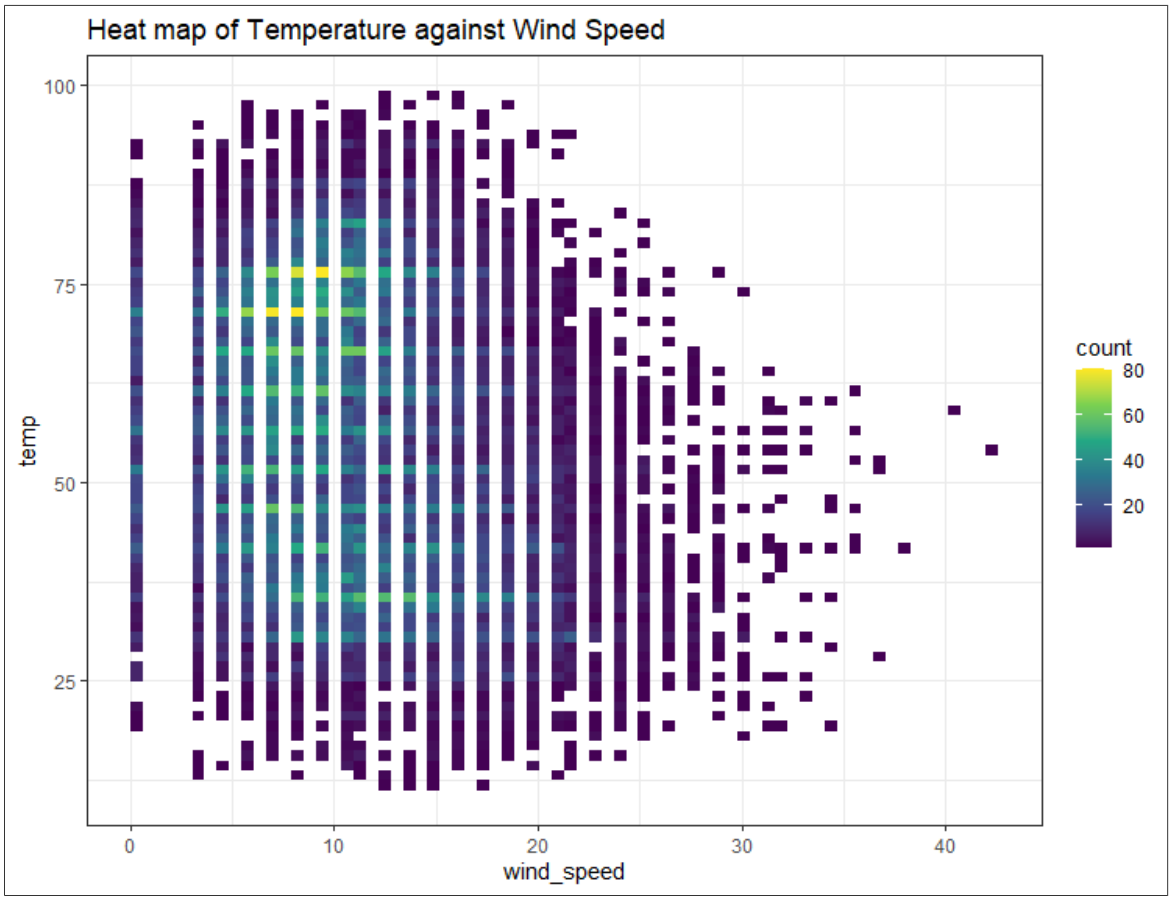

Figure 18 Heatmap of temperature against Wind Speed

Output

The above graph shows us the relation between wind speed and temperature. Here, Temperature is inversely proportional to wind speed. This means when temperature increases the wind speed decreases and when wind speed is higher the temperature decreases.

19. Conclusion

In this assignment I have done 14 analysis and 2 extra features are added on the hourly weather data of two airport stations i.e. LGA and JFK. The analysis included data manipulation, cleaning, data exploration and data visualization. For data visualization I have used many graphs like Scatterplot, Boxplot and histograms. We were assigned with the dataset of hourly weather to perform analysis and make decisions according to the outcomes that comes from the analysis. The decisions will be used for airport to manage air traffic. As an extra feature I have added two graphs which are Density plot and Heat map. Both of them are highly encouraged by professional data analysts to use as they have many advantages and are clear for understanding.

References

Mcleod, S., 2019. Box Plot | Simply Psychology. [online] Simplypsychology.org. Available at: <https://www.simplypsychology.org/boxplots.html#:~:text=In%20descriptive%20statistics%2C%20a%20box,(or%20percentiles)%20and%20averages.> [Accessed 7 January 2021].

Medium. 2017. *Understanding Boxplots*. [online] Available at: <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51> [Accessed 18 December 2020].

Yi, M., 2019. *A Complete Guide to Scatter Plots*. [online] Chartio. Available at: <https://chartio.com/learn/charts/what-is-a-scatter-plot/> [Accessed 20 January 2021].

Holtz, Y., 2019. *Histogram | the R Graph Gallery*. [online] R-graph-gallery.com. Available at: <https://www.r-graph-gallery.com/histogram.html> [Accessed 20 February 2021].

Datavizcatalogue.com. 2018. *Density Plot - Learn about this chart and tools to create it*. [online] Available at: <https://datavizcatalogue.com/methods/density_plot.html> [Accessed 20 December 2021].

Name: Aaditya Jha **TP: NP000290**
Mark:

Individual Assignment - Design, Documentation, and Implementation (80%):

Question No.	Topic	Question Vs Taxonomy						PLO
		Cognitive Level						
		1	2	3	4	5	6	
		SQ	SQ	SQ	SQ	SQ	SQ	
1	Report Format and System Documentation				37%			2
2	Implementation				63%			2
	Total				**100%**			

Individual Assignment - Presentation (20%):

Question No.	Topic	Question Vs Taxonomy					PLO
		Affective Level					
		1	2	3	4	5	
		SQ	SQ	SQ	SQ	SQ	
1	Assignment Presentation			100%			5
	Total			**100%**			