NPI000047 (PFDA)

by Shivam Ranabhat

Submission date: 06-Mar-2022 02:45PM (UTC+0800)
Submission ID: 1775549089
File name: NPI000047_Shivam_Ranabhat.docx (1.53M)
Word count: 4267
Character count: 20000

1. Introduction and Assumptions

In this report, I had used different data analyzing and manipulation techniques to explore the given weather data set to extract the information which needs to conclude. Different techniques like data pre-processing, data visualization, data exploration, data manipulation were involved during this project which was already discussed in our course. Using all these techniques helps me to improve my visualization, exploration, manipulation skills in R programming. Using which at first, I import the given hourly data set in R Studio, then necessary pre-processing was done to apply appropriate commands so that the data set can be converted into the desired format. In this project, relevant graphics (i.e., bar graphs, scatterplots, histograms, and so on) were used to support the findings through R Studio.

The provided data set for this project is about hourly weather data set which contains two origins i.e., John F Kennedy International Airport (JFK) and LaGuardia Airport (LGA) of US. In the airport, every weather factor is seriously considered when taking the flight. It includes temperature, visibility, wind direction, pressure, and many more. Similarly, all those factors are kept inside a single hourly weather data set which consists of 15 columns and 17, 412 rows. All the descriptions of columns inside the dataset are listed below:

Column(s)	Description
origin	Weather station.
year, month, day, hour	Time of recording.
temp, dewp	Temperature and dewpoint in F.
humid	Relative humidity.
wind_dir, wind_speed, wind_gust	Wind direction (in degrees), speed and gust speed (in mph).
precip	Precipitation, in inches.
pressure	Sea level pressure in millibars.
visib	Visibility in miles.
time_hour	Date and hour of the recording as a POSIXct date.

Table: Column description of Hourly weather data set

2. Aims and Objectives

The major aim of this assignment is to use the R programming language to study the different factors of weather and extract necessary information from hourly weather data set for the year 2013 which cause climate change in two airports.

Objective

i. The primary objective is to import the given data set in R studio.

ii. The secondary objective is to explore, analyze, manipulate and visualize the data set.

3. Background

R programming language is used for different statistical computing and graphics. According to Johnson (2019), R is one of the most used programming languages in data mining. It was developed by Ross lhaka and Robert Gentleman in 1993. In the R programming package, ggplot2 has become the most popular package due to its aesthetic and interactivity. R programming is used in different fields and its purposes vary according to the sectors like finance, banking, healthcare, social media, e-commerce, and different manufacturing companies.

R programming language is used in most large companies like Uber, Facebook, Google, and so on. Similarly, Facebook uses R to update its status, social network graph and used to forecast how their colleagues would interact with R. Ford motors rely on R for statistical analysis and for assisting data-driven decisions. In google, R is used to calculate ROI on advertising campaigns, forecast economic activities, and increase the effectiveness of internet advertising. Microsoft uses R programming for their Xbox matchmaking services as well as a statistical engine with Azure ML architecture. Similarly, In Mozilla Firefox R programming is served as the backbone as it is used to display web activities (Burns, 2019). If such big companies are using the R programming language, we can understand the different applications of the R programming language

4. Data Import and Pre-Processing

4.1.Install Packages

```
install.packages("readxl") #package which include functions to ready excel files
install.packages("colorspace") # package used for manipulating and assessing colors and palettes
install.packages("ggplot2") #package which includes ggplot functions
install.packages("dplyr") #package which includes filter functions)
install.packages("VIM") #package which include functions to read null value
```

Fig 1: Installing packages

4.2. Load Packages

```
library(VIM)
library(readxl)
library(colorspace)
library(ggplot2)
library(dplyr)
```

Fig 2: Load packages

4.3. Read CSV File

```
Weather = read.csv("C:\\Users\\Info-chip\\Documents\\Hourlyweatherdata.csv" )
print(Weather) #To print the excel file
```

Fig: Read Csv file

At first, all the necessary packages should be installed using the install.packages() function, then after installing using library() function it is loaded in the R environment. Then the given file is imported into the Rafter viewing the extension of file format. The given file is in Excel CSV format, after installing the readxl package read.csv() function is used.

Before performing other techniques, necessary packages must be loaded to the R environment. Similarly, “ggplot2”, “VIM”, “colorspace”, “dplyr” packages are first installed using install.packages() function and then loaded using library() function.

4.4. Pre-processing (Filtering origin)

```
JFK = filter(Weather, origin  =="JFK")
LGA = filter(Weather, origin  =="LGA")
```

Fig 3: Data Filtering

As there are two origins in the data set i.e., JKF and LGA airport so, filter all the data based on origin and store it into two variables (JKF for JKF airport and LGA for LGA airport). Using this we can easily extract the necessary information of a particular airport.

5. Analysis

Overview Topics of the analysis

Analysis 1: Comparing overall temperature of JKF and LGA

Analysis 2: Comparing humidity of overall year of JKF and LGA

Analysis 3: Comparing wind direction of the whole year 2013 of both origins

Analysis 4: Comparing wind speed of JKF and LGA for the Year 2013

Analysis 5: Comparing maximum precipitation of both origins for the whole year 2013

Analysis 6: Comparing the dew point of both origins for the whole year 2013

Analysis 7: Comparing wind gust of JFK and LGA for the Year 2013

Analysis 8: Finding the null value of January 2013

Analysis 9: Comparing January temperature below the ice point of both origins

Analysis 10: Comparing wind speed and pressure of spring season

Analysis 11: Comparing the temperature of Dec 22 of JFK and LGA

Analysis 12: Comparing the hourly dew point of 1st January of JFK and LGA

Analysis 13: Comparing the visibility of December of JFK and LGA

Analysis 14: Comparing the maximum pressure of 2013 in both origins

Analysis 1:

```
#---------------------------Analysis 1-----------------------------#

#Study Temperature of JFK and LGA for whole year of 2013 in form of geom smooth
ggplot(weather, aes(x=month, y=temp, na.rm=TRUE, suppressWarnings(expr)))+geom_smooth(size= 1, se=FALSE, aes(color=origin))+
  labs(x="Month", y="Temperature of Both JFK and LGA Airport")+ scale_x_discrete (limits=c(1:12))
summary(JFK$temp)
summary(LGA$temp)
#------------------------------------------------------------#
```

Fig 4: Temperature of JKF and LGA for the overall year 2013

- In this analysis the overall temperature of both origins is compared throughout the year 2013. The main purpose of this analysis is to study the monthly rise and fall in temperature in the year 2013 for both airports.
- In the above code, I have plotted two continuous position variables in a graph using the geom_smooth() function and along the x-axis, I have filled month scaling it from 1 to 12 using scale_x_discrete() function as the data consist from January to December. The warning function is removed using the suppresswarning () function.
- Using summary function(), the minimum temperature, 1st quartile, median, mean, 3rd quartile, and maximum temperature are found as follows:

```
summary(JFK$temp)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
12.02   39.92   53.96   54.47   69.08   98.06
summary(LGA$temp)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
12.02   39.92   55.94   55.76   71.06   98.96
```

Fig 5: summary of overall temperature

Result:

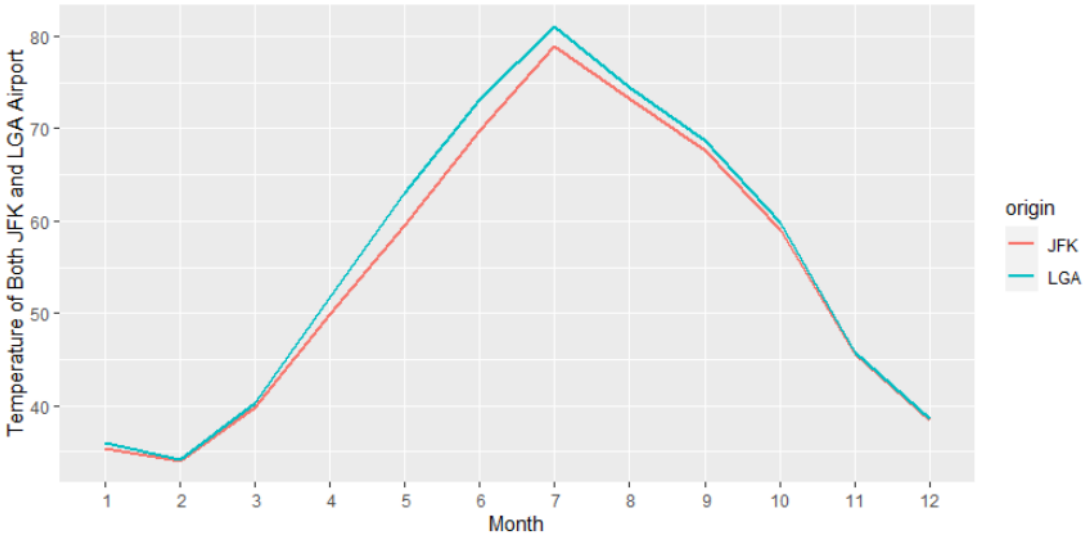

Fig 6: Temperature of JKF and LGA

In the above diagram, the red line represents the temperature of JKF and blue represents the temperature of LGA airport, and 1 to 12 represents the number of months. The temperature of both airports is found to be a minimum in February. The maximum temperature is found in July in both airports. It starts increasing from February and to July and then slowly decreasing from July to December. While comparing both the airports the maximum temperature is found in LGA airport and the minimum temperature is almost the same in both airports.

Analysis 2:

```
#--------------------------Analysis 2-----------------------------#

#Study Humidity of JFK and LGA for whole year of 2013
#Humidity comparison for JFK and LGA for for whole year of 2013

Comparehumidity=ggplot(weather,aes(x=origin,y=humid))+geom_boxplot()+
  labs(title = "Humidity recorded from JFK and LGA weather Stations",x="weather station",
       y="Humidity");
print(Comparehumidity)
summary(JFK$humid)
summary(LGA$humid)
#------------------------------------------------------------#
```

Fig 7: Comparing humidity of the overall year

- This analysis is about comparing the humidity of both JKF and LGA airports throughout the year 2013. Using this analysis, we can find the difference in humidity found in both airports along with the airport which has maximum and minimum humidity in the whole year 2013.
- In the code of the above figure, I have stored the annual humidity in the "compare humidity" variable. The data is further displayed in the form of the boxplot using the geom_boxplot() function.
- Using the summary() function, the minimum temperature, 1st quartile, median, mean, 3rd quartile, and maximum humidity are found as follows:

```
summary(JFK$humid)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
15.21   49.19   65.80   65.21   82.73  100.00
summary(LGA$humid)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
12.74   45.57   57.30   59.32   73.30  100.00
```

Fig 8: summary of humidity of JKF and LGA

Result

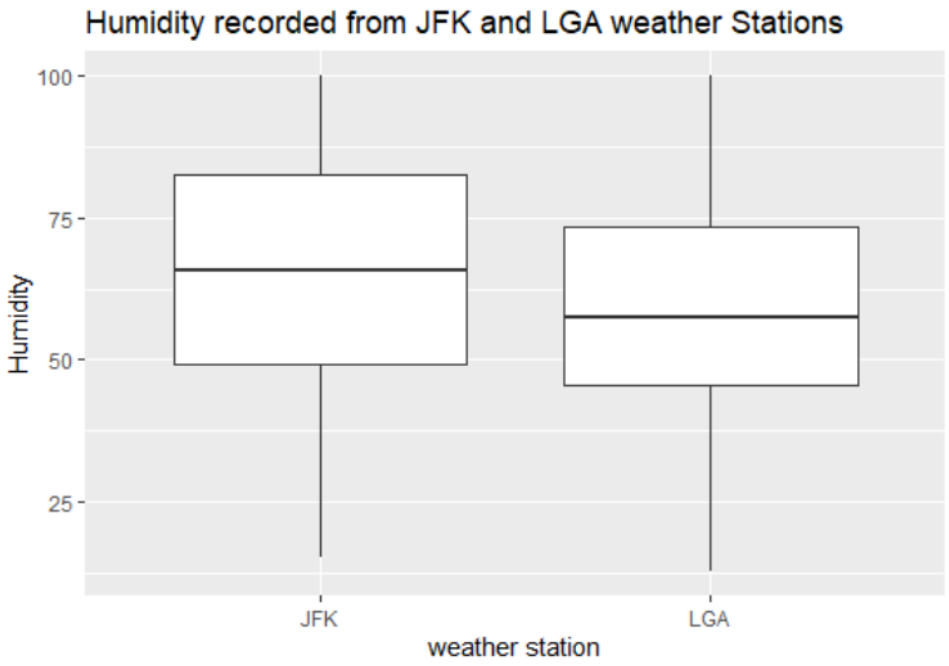

Fig 9: Boxplot of JFK and LGA for the humidity of the 2013 year

From the above figure, the maximum humidity is the same throughout the year 2013, the humidity is minimum in LGA than JFK, the median of humidity is higher in JFK than LGA. Similarly, the 1st and 3rd quartile for humidity in JFK is found to be greater than the LGA.

Analysis 3:

```
#--------------------------Analysis 3-----------------------------#

#Study Wind Direction of JFK and LGA for whole year of 2013
#Wind direction comparison for JFK  for for whole year of 2013
options(warn = -1)

JFK %>% ggplot(aes(wind_dir))+ geom_bar(aes(color=wind_dir,na.rm=TRUE ))+ facet_wrap(~month)+
  labs(title="Wind direction for each day of Year 2013 of JFK Airport", x= "Wind direction of JFK Airport")
summary(JFK$wind_dir)
#Wind direction comparison for LGA  for for whole year of 2013
options(warn = -1)
LGA %>% ggplot(aes(wind_dir))+ geom_bar(aes(color=wind_dir,na.rm=TRUE ))+ facet_wrap(~month)+
  labs(title="Wind direction for each day of Year 2013 of LGA Airport", x= "Wind direction of LGA Airport")
summary(LGA$wind_dir)
#------------------------------------------------------------#
```

Fig 10: Wind Direction of the whole year 2013 of both origins

- In this analysis, the wind direction of the overall year 2013 is compared for both JFK and LGA airports. From this analysis, we can compare the wind direction in the degree of both airports.
- In the code shown above, I have used the option(warn = -1) is used to avoid the warning. Similarly, the data is displayed in the form of a bar graph using the geom_bar() function, and passing the parameter na.rm = TRUE the null value present in the wind_dir column will be removed.
- The overall summary is displayed using the summary() function which contains minimum temperature, 1st quartile, median, mean, 3rd quartile, and maximum temperature as follows:

```
summary(JFK$wind_dir)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.0   140.0   220.0   204.2   290.0   360.0      51
summary(LGA$wind_dir)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.0   100.0   220.0   199.5   300.0   360.0     153
```

Fig 11: Summary of humidity for JFK and LGA

Result:

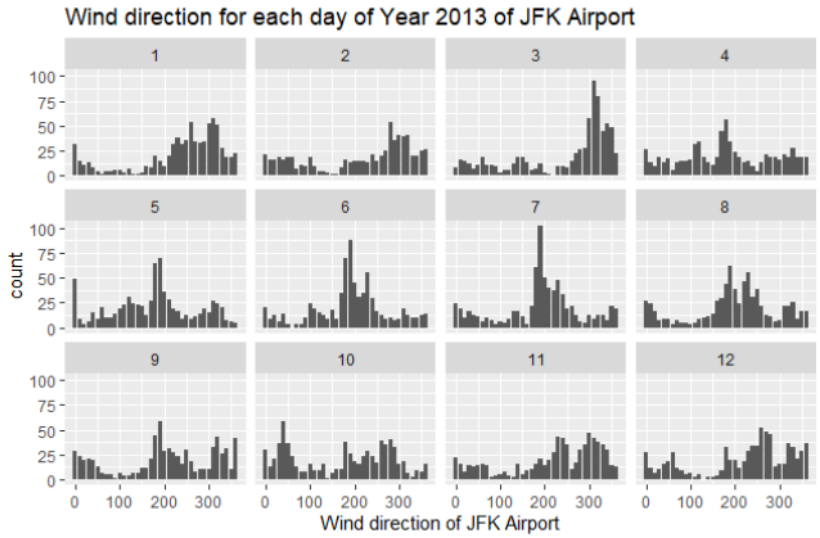

Fig 12: Wind direction of JFK for the year 2013

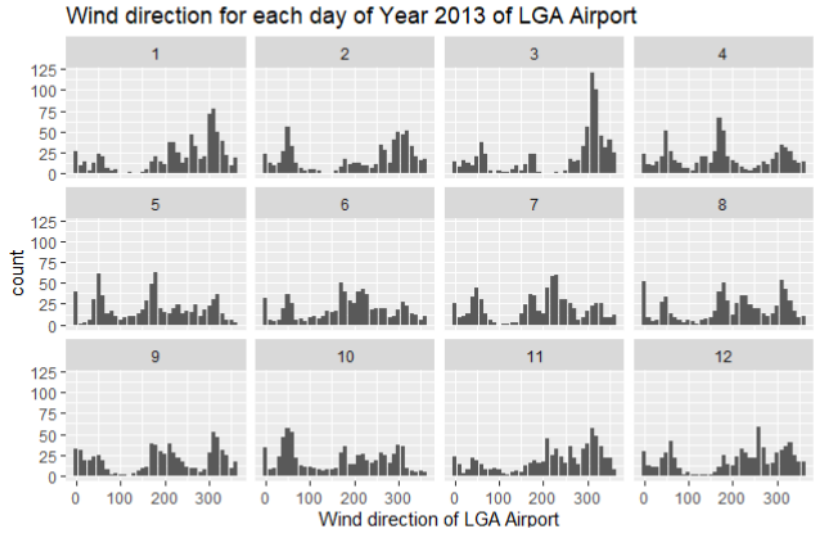

Fig 13: Wind direction of LGA for the year 2013

From the above two figures, the JFK airport has a higher wind direction in comparison to LGA airport. In March, June, and July the wind direction is found to be higher than other months in JFK airport. Similarly, In LGA airport, the wind direction value is found higher only in March. It concluded that the wind direction is found higher in spring seasons rather than others.

Analysis 4

```
#--------------------------Analysis 4-----------------------------#
#Study Wind Speed of JFK and LGA for whole year of 2013
#Wind speed comparison for JFK  for for whole year of 2013
ggplot(JFK,aes(x=day, y=wind_speed, na.rm=TRUE))+ geom_point()+ facet_wrap(~month)+
  labs(title="Wind speed for each day of Year 2013 of JFK Airport",x="Days" ,y= "Wind speed of JFK Airport")
summary(JFK$wind_speed)
#Wind speed comparison for LGA  for for whole year of 2013
ggplot(LGA,aes(x=day, y=wind_speed, na.rm=TRUE))+ geom_point()+ facet_wrap(~month)+
  labs(title="Wind speed for each day of Year 2013 of LGA Airport",x="Days" , y= "Wind speed of LGA Airport")
summary(LGA$wind_speed)
#------------------------------------------------------------------#
```

Fig 14: wind speed of JFK and LGA of 2013

- In this analysis, the overall wind speed of the year 2013 is compared between JFK and LGA. A scatterplot is filled individually for both JFK and LGA so that it would be easy for comparing the wind speed.
- In the above code, along the y-axis wind_speed, is filled in relation with days along the x-axis, and the geom_point() function is used to plot the scatterplot.
- Similarly, a summary() function is used to find out min, max, mean, number of null values, and so on.

```
> summary(JFK$wind_speed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   6.905  10.357  11.468  14.960  42.579       3
> summary(LGA$wind_speed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   6.905  10.357  10.623  13.809  40.277
```

Fig 15: Summary of wind speed of JFK and LGA

Result

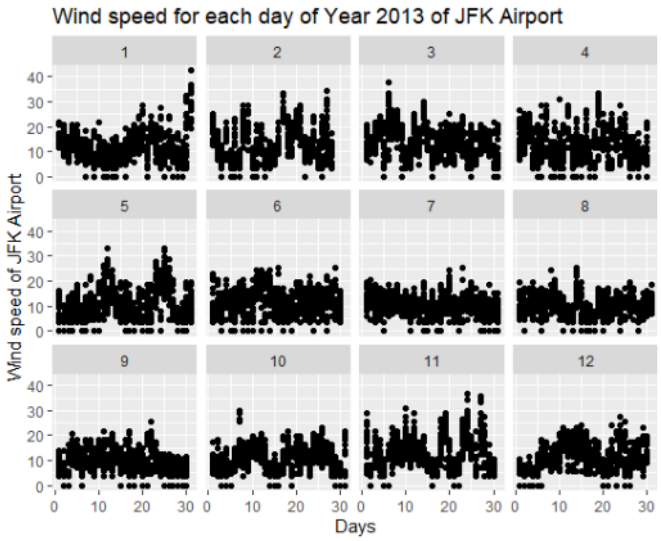

Fig 16: Scatterplot for the wind speed of JFK Airport

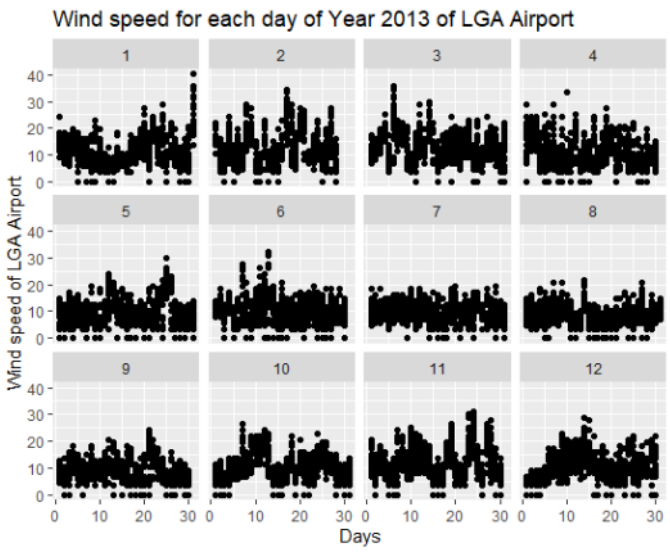

Fig 17: Scatterplot for the wind speed of LGA Airport

By comparing the above plots of two airports it is found that the wind speed of both JFK and LGA is maximum of January but the wind speed is maximum in JFK than LGA airport which is found to be 42.579 mph. But the minimum value is the same for both JFK and LGA which is 0 mph. After January, the wind speed is higher in November month of both airports.

Analysis 5

```
#-------------------------Analysis 5-----------------------------#
#Study Monthly maximum precipitation of JFK and LGA for whole year of 2013
#Monthly maximum precipitation comparison for JFK  for for whole year of 2013
ggplot(JFK %>% group_by(month) %>%summarise(precip = max(precip)),aes(x = month, y = precip)) +
  geom_point()+geom_line() + scale_x_discrete(limits = c(1:12)) +
  labs(title="Maximum Precipitation  of JFK in 2013",x="Month", y="Average")
summary(JFK$precip)
#Monthly maximum precipitation comparison for LGA  for for whole year of 2013
ggplot(LGA %>% group_by(month) %>%summarise(precip = max(precip)),aes(x = month, y = precip)) +
  geom_point()+geom_line() + scale_x_discrete(limits = c(1:12)) +
  labs(title="Maximum precipitation  of LGA in 2013",x="Month", y="Maximum")
summary(LGA$precip)
#-----------------------------------------------------------------#
```

Fig 18: Maximum precipitation of both origins for the whole year 2013

- In this analysis, the monthly maximum precipitation of both JFK and LGA is compared for the whole year of 2013. For this, a line is plotted using the geom_line() function where precipitation along the y-axis and month along the x-axis is plotted. And using summarise () function the precipitation column is selected with maximum values. Scale_x_discrete() function is used to set the number of months that are placed in the vector.
- summary() function is used to retrieve min, max, median, mean, and so on.

```
summary(JFK$precip)
  Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
000000 0.000000 0.000000 0.003985 0.000000 0.660000
summary(LGA$precip)
  Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
000000 0.000000 0.000000 0.004381 0.000000 0.820000
```

Fig 19: Summary of maximum precipitation for the year 2013

Result

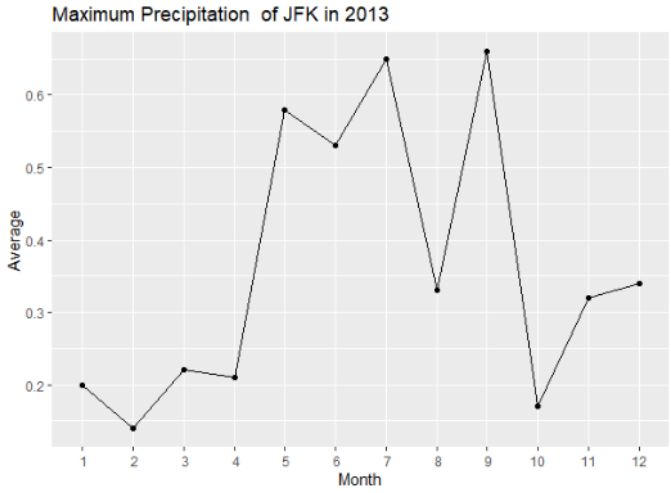

Fig 20: Line plot for maximum precipitation of JFK for 2013

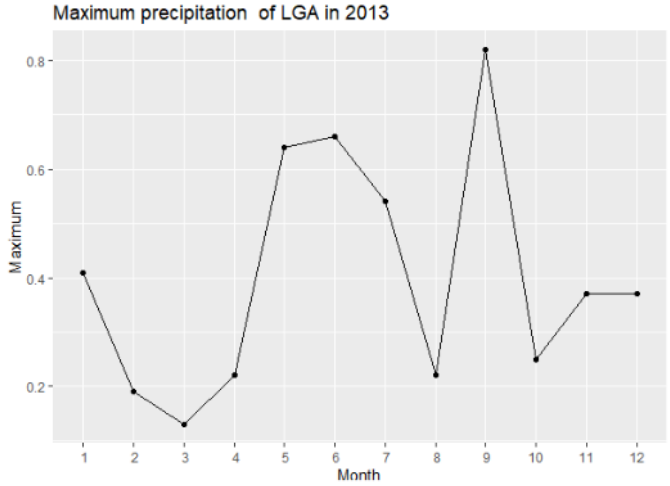

Fig 21: Line plot for maximum precipitation of LGA for 2013

From the above figure, maximum precipitation is found in LGA airport. In LGA airport the precipitation is found maximum in July and September. Similarly, in LGA the maximum precipitation is found in September month which is 0.82 inches. In JFK the maximum precipitation is found to be 0.66 inches. And both airports have the same minimum values. In the rest of the months, the maximum precipitation is normal in both JFK and LGA airports.

Analysis 6:

```
#-------------------------Analysis 6-----------------------------#
#Study  Dew point of JFK and LGA for whole year of 2013
#Dew point of JFK airport of whole year 2013
JFK %>% ggplot(aes(x=month, y=dewp,na.rm=TRUE))+ geom_bar(stat="identity", color="blue", fill="white")+
  scale_x_continuous(breaks = seq(0,12, by=1))+
  labs(title="Dew point for  Year 2013 of JFK Airport", x= "Months")
summary(JFK$dewp)
#Dew point of LGA airport of whole year 2013
LGA %>% ggplot(aes(x=month, y=dewp,na.rm=TRUE))+ geom_bar(stat="identity", color="blue",fill="white")+
  scale_x_discrete(limit=c(1:12))+ labs(title="Dew point for  Year 2013 of LGA Airport", x= "Months")
summary(LGA$dewp)
#-------------------------------------------------------------------#
```

Fig 22: Dew point of both origins for the year 2013

- In this analysis, the dew point of both JFK and LGA is compared for the whole year of 2013. For this, a bar is plotted using the geom_bar() function where the dew point along the y-axis is plotted. The background color is provided white using fill and blue to the bars. Scale_x_discrete() function is used to set the number of months that are placed in the vector.
- Similarly, a summary() function is used to find the min, max, mean, median, and so on as follows:

```
summary(JFK$dewp)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-9.94   26.96   42.98   41.86   57.92   78.08
summary(LGA$dewp)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-7.06   26.06   41.00   40.61   55.94   73.94
```

Fig 23: Summary of dew point for the year 2013

Result

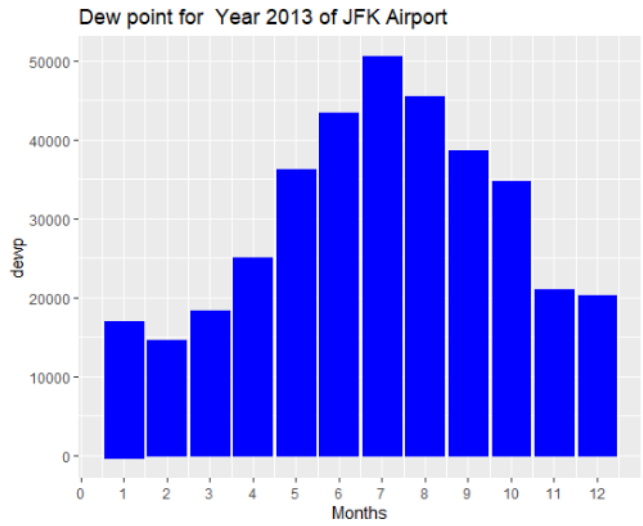

Fig 24: Bar chart for the monthly dew point of JFK

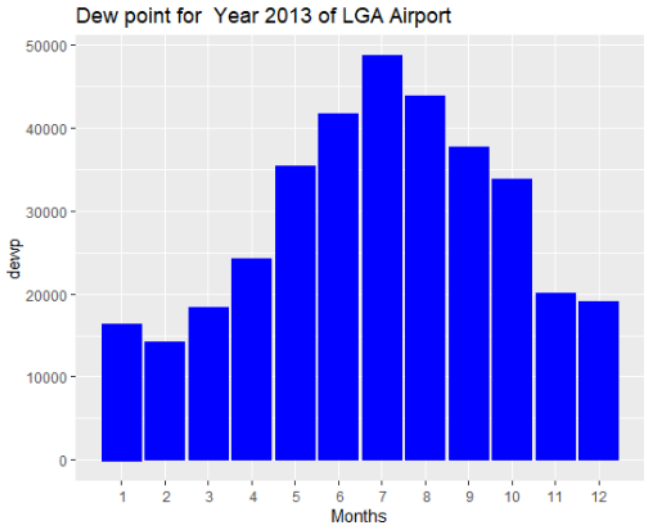

Fig 25: Bar chart for the monthly dew point of LGA

From figures 24 and 25, we can compare the overall monthly dew point of JFK and LGA. The maximum value of dew point is found in July of both airports where JFK has a higher dew point with a value of 78.78 degrees Fahrenheit than LGA whose maximum value of dew point is 73.94 degrees Fahrenheit. Similarly, the minimum dew point is found in February where the value is -9.94 degrees Fahrenheit in JFK is greater than LGA airport having a dew point of -7.06 degrees Fahrenheit.

Analysis 7:

```
#-------------------------Analysis 7-----------------------------#
#Study  wind gust of JFK and LGA for whole year of 2013
#Wind gust of JFK airport of whole year 2013
ggplot(JFK %>% group_by(month) %>%summarise(wind_gust, na.rm=TRUE),aes(x = month, y = wind_gust)) +
  geom_point()+geom_line() + scale_x_discrete(limits = c(1:12)) +
  labs(title="Wind gust  of JFK in 2013",x="Month", y="Overall wind gust")
summary(JFK$wind_gust)
#wind gust of LGA airport of whole year 2013
ggplot(LGA %>% group_by(month) %>%summarise(wind_gust,na.rm=TRUE),aes(x = month, y = wind_gust)) +
  geom_point()+geom_line() + scale_x_discrete(limits = c(1:12)) +
  labs(title="Wind gust  of LGA in 2013",x="Month", y="Overall wind gust")
summary(LGA$wind_gust)
#-----------------------------------------------------------------#
```

Fig 26: Wind gust comparison for JFK and LGA

- In this analysis, the wind gust of the whole year 2013 for both JFK and LGA airports is compared in the form of a line graph. In the above analysis ggplot() function is used to plot the graph which is grouped by month using group () function, summarise() function is used to select the desired column, geom_line() and geom_point() is used to fill the graph in the form of the line with points and scale_x_discrete is used to provide the numbers for the month along the x-axis.
- summary () function is used to find the overall summary i.e., min, max, median, number of null values, and so on.

```
summary(JFK$wind_gust)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
16.11   23.02   26.47   27.56   31.07   66.75    7199
summary(LGA$wind_gust)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
16.11   20.71   24.17   25.14   27.62   62.14    6678
```

Fig 27: Summary of Wind gust for both Airports

Result

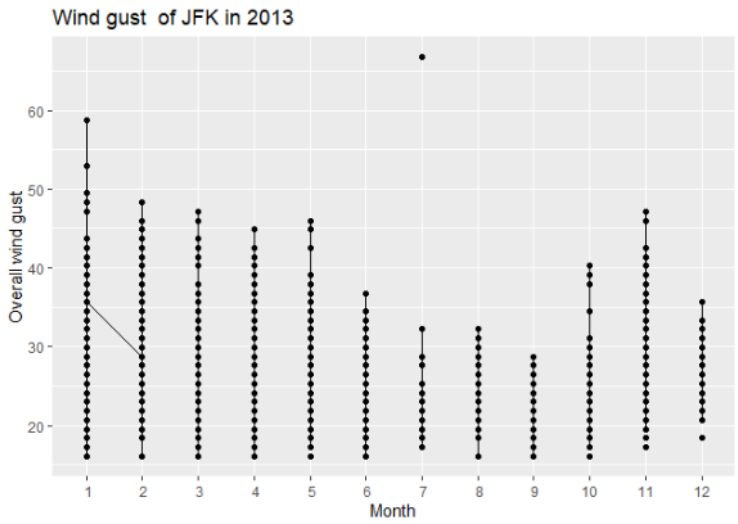

Fig 28: Line plot with points of wind gust in JFK

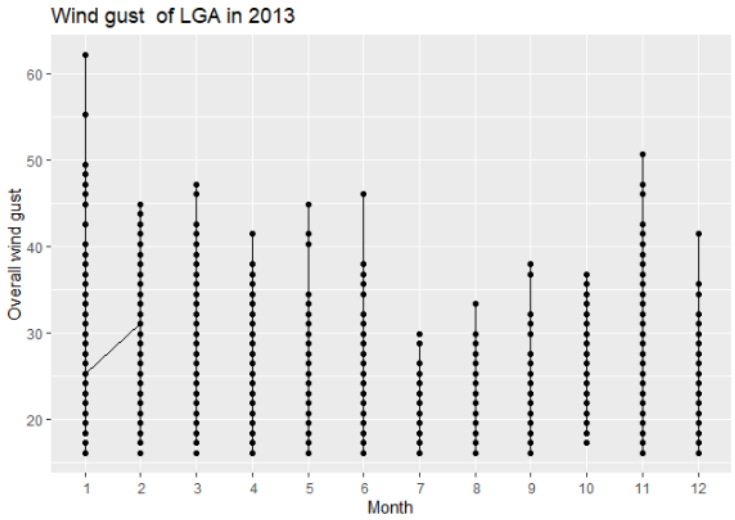

Fig 29: Line plot with points of wind gust in LGA

After comparing both figures 28 and 29 it is found that the wind gust is higher in JFK in July which is 66.75 mph. Similarly, in LGA airport, the maximum value of wind gust is found in the month of is 62. 14 mph. In December both wind gusts are found slightly higher after the overall maximum value. Similarly, the minimum value for wind gust is the same for both JFK and LGA whose value is 16.11 mph.

Analysis 8:

```
#-------------------------Analysis 8-----------------------------#
#Finding null value of JFK for  January of 2013
JFKJan = filter(Weather, origin  =="JFK", month=="1")
JFKNA = JFKJan %>% select(temp:visib) %>% aggr(prop=T, numbers=T)
summary(JFKNA)
#Finding null value of LGA for  January of 2013
LGAJan = filter(Weather, origin  =="LGA", month=="1")
LGANA = LGAJan %>% select(temp:visib) %>% aggr(prop=T, numbers=T)
summary(LGANA)
#-----------------------------------------------------------------#
```

Fig 30: Finding the null value of January 2013 of JFK and LGA

- This analysis is about finding the null values and the columns where the maximum and a minimum number of values are missing. After this, both JFK and LGA are compared about the number of null values.
- In this analysis, pre-processing is performed to filter out the origin and their data using the filter () function from the weather data.
- To plot the graph of proportion aggr () function is used which is enabled due to installing and loading the VIM package. After providing TRUE to both proportion and number the graph is generated which shows the data both in proportion and number of missing values.
- Similarly, the summary () function is used to find the missing numbers along with their columns as follows:

```
Missings per variable:
  Variable Count
      temp     0
      dewp     0
     humid     0
  wind_dir     1
wind_speed     0
 wind_gust   600
    precip     0
  pressure    76
     visib     0
```

```
Missings per variable:
  Variable Count
      temp     0
      dewp     0
     humid     0
  wind_dir     7
wind_speed     0
 wind_gust   508
    precip     0
  pressure    86
     visib     0
```

Fig 31: Summary of null values of JFK and LGA

Result

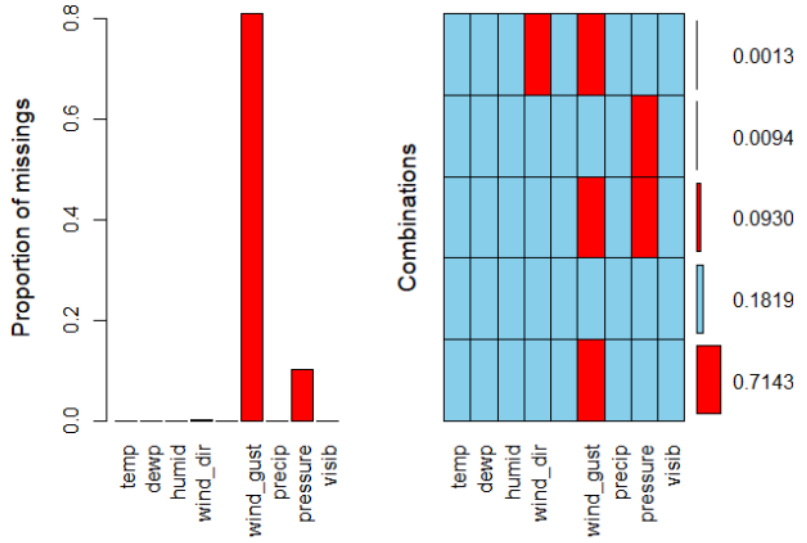

Fig 32: Proportion graph of missing values of JFK

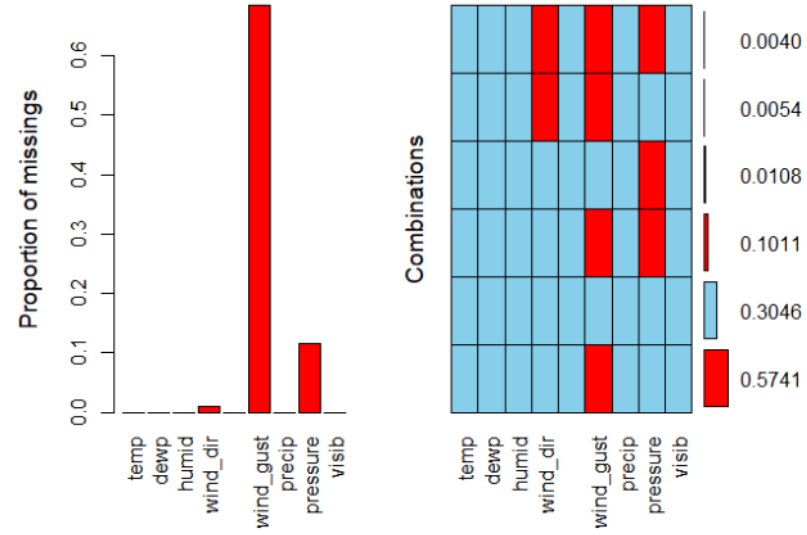

Fig 33: Proportion graph of the missing value of LGA

From figures 32 and 33, it is found that most null values are found in JFK airport. The null values are a little less in LGA compared to JFK. Similarly, the values are mostly missing in wind_gust and then pressure and wind_dir in both airports. The missing value of wind_dir is found more in LGA than in JFK airport.

Analysis 9:

```
#--------------------------Analysis 9-----------------------------#
#Study  temperature below ice point of JFK for January of 2013
#Pre-processing
icepointJFK= JFKJan%>% group_by(day)%>% filter(origin=="JFK")%>%filter(temp <= 32)%>%
  select(origin,month,day,temp)
ggplot(icepointJFK, aes(x=day, y=temp,)) + geom_point() + facet_wrap(~month)+
  labs(title="Temperature below ice point for January of Year 2013 of JFK",
       x= "Days of January")
summary(icepointJFK$temp)

#Study  temperature below ice point of LGA for January of 2013
#Pre-processing
icepointLGA= LGAJan%>% group_by(day)%>% filter(origin=="LGA")%>%filter(temp <= 32)%>%
  select(month,day,temp)
ggplot(icepointLGA, aes(x=day, y=temp,)) + geom_point() + facet_wrap(~month)+
  labs(title="Temperature below ice point for January of Year 2013 of LGA ",
       x= "Days of January")
summary(icepointLGA$temp)

#-----------------------------------------------------------------#
```

Fig 34: Comparing January temperature below ice point of both origins

- In this analysis, the temperature of January below ice point i.e., less than or equals 32 degrees Fahrenheit. At first, I filter the origin including the temperature which is less than or equal to 32, and using the select () function I select origin, month, day, and temp and stored it in a variable i.e., “icepointJFK” for JFK airport and “icepointLGA” for LGA airport.
- Similarly, after completing the pre-processing I use the ggplot () function where the day is kept along the x-axis and filtered temp along the y-axis, and using geom_point () the graph is plotted in the form of point.
- summary () function is used to find the overall summary i.e., min, max, median, number of null values, and so on.

```
summary(icepointJFK$temp)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
12.02   19.04   24.98   24.10   30.02   32.00
summary(icepointLGA$temp)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
12.02   19.04   24.08   24.25   30.02   32.00
```

Fig 35: Summary of temperature below ice point

	month	day	temp
1	1	1	32.00
2	1	1	30.02
3	1	1	28.94
4	1	1	26.96
5	1	2	26.06
6	1	2	26.06
7	1	2	24.98
8	1	2	24.98
9	1	2	24.08
10	1	2	23.00
11	1	2	23.00
12	1	2	23.00
13	1	2	24.98
14	1	2	26.96
15	1	2	28.94
16	1	2	30.02
17	1	2	30.92
18	1	2	32.00
19	1	2	32.00
20	1	2	30.92
21	1	2	30.92
22	1	2	30.92
23	1	2	30.92
24	1	2	30.02
25	1	2	30.02
26	1	3	28.94
27	1	3	28.94

Showing 1 to 27 of 255 entries, 3 total columns

	month	day	temp
1	1	1	32.00
2	1	1	30.02
3	1	1	28.94
4	1	1	28.04
5	1	2	26.96
6	1	2	26.96
7	1	2	26.06
8	1	2	24.98
9	1	2	24.08
10	1	2	24.08
11	1	2	24.08
12	1	2	24.08
13	1	2	24.98
14	1	2	26.96
15	1	2	28.04
16	1	2	28.94
17	1	2	30.92
18	1	2	32.00
19	1	2	32.00
20	1	2	32.00
21	1	2	30.92
22	1	2	30.92
23	1	2	30.92
24	1	2	30.92
25	1	2	30.02
26	1	3	30.02
27	1	3	30.02

Showing 1 to 27 of 240 entries, 3 total columns

Table 1: Table generated from pre-processing

Result

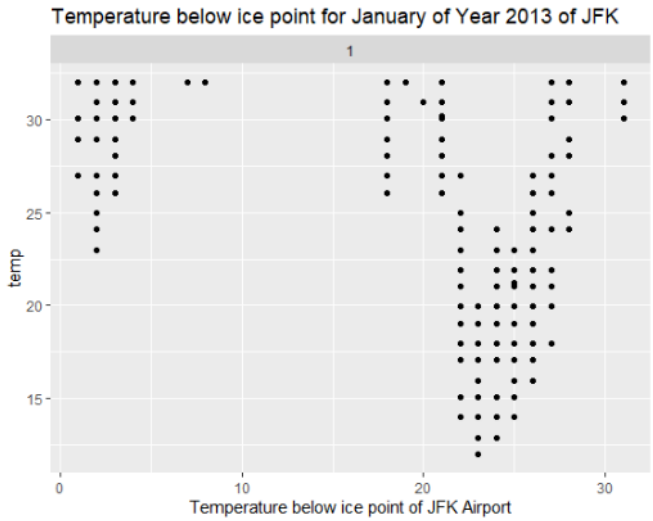

Fig 36: Point plot for temperature below ice point of JFK

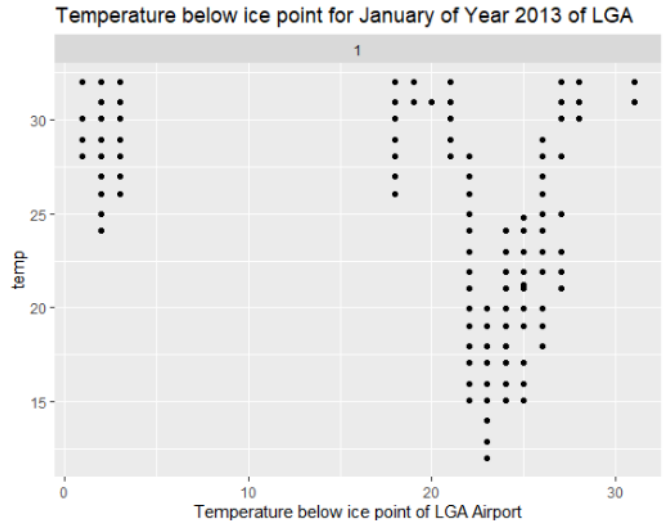

Fig 37: Point plot for temperature below ice point of LGA

From above figures 36 and 37, the temperature ice point of both airports is found almost similar. The minimum value of temperature below ice point is also the same for both airports. The temperature below the ice point does not found from day 5 to day 15 in LGA airport from day 5 to 15 in JFK airport. In both airports, the temperature below ice points is mostly found on day 25 to day 27. So, we can say that the temperature is cold at the end of January.

Analysis 10:

```
#--------------------------Analysis 10-----------------------------#
#Study  the relation of temperature and visibility  of spring season  2013
# Wind speed and pressure In JFK Airport
springJFK= JFK%>% group_by(month)%>% filter(origin=="JFK")%>%filter(month =="3"|month =="4"|month =="5" )%>%
  select(origin,month,day,wind_speed, pressure)

ggplot(springJFK,aes(wind_speed,pressure,color=origin))+geom_point()+geom_smooth(colour = "blue", size=3)+
  facet_wrap(~month)+
  labs(title="Wind Speed VS Pressure comparison for JFK",
       x="Wind Speed", y="Pressure ")
summary(springJFK[,4:5])
# Wind speed and pressure In LGA Airport
springLGA= LGA%>% group_by(month)%>% filter(origin=="LGA")%>%filter(month =="3"|month =="4"|month =="5" )%>%
  select(origin,month,day,wind_speed, pressure)

ggplot(springLGA,aes(wind_speed,pressure,color=origin))+geom_point()+geom_smooth(colour = "blue", size=3)+
  facet_wrap(~month)+
  labs(title="Wind Speed VS Pressure comparison for LGA",
       x="Wind Speed", y="Pressure ")
summary(springLGA[,4:5])
```

Fig 38: Comparing wind speed and pressure of spring season in JFK and LGA

- In this analysis the relation of wind speed and pressure is compared for the spring season i.e., March, April, and May. As the spring season is known as the windy season so, the effect of wind speed and pressure generated is shown in this analysis.
- For pre-processing, the filter () function is used to filter the origin and month then using the select () function, origin, month, day, wind speed, and pressure is selected and stored in a variable (springJFK for JFK and springLGA for LGA). Then, using the ggplot() function it is filled in a graph using geom_smooth() in the form of the regression line.
- Similarly, the summary () function is used to find the overall summary i.e., min, max, median, number of null values, and so on.

```
> summary(springJFK[,4:5])
   wind_speed        pressure
 Min.   : 0.000   Min.   : 998
 1st Qu.: 6.905   1st Qu.:1012
 Median :11.508   Median :1017
 Mean   :12.340   Mean   :1018
 3rd Qu.:17.262   3rd Qu.:1024
 Max.   :37.976   Max.   :1038
 NA's   :1        NA's   :212
> summary(springLGA[,4:5])
   wind_speed        pressure
 Min.   : 0.000   Min.   : 997.9
 1st Qu.: 6.905   1st Qu.:1011.5
 Median :10.357   Median :1016.5
 Mean   :11.297   Mean   :1017.3
 3rd Qu.:14.960   3rd Qu.:1023.4
 Max.   :35.674   Max.   :1038.0
                  NA's   :250
```

Fig 39: Summary of wind speed and pressure of spring season

Result

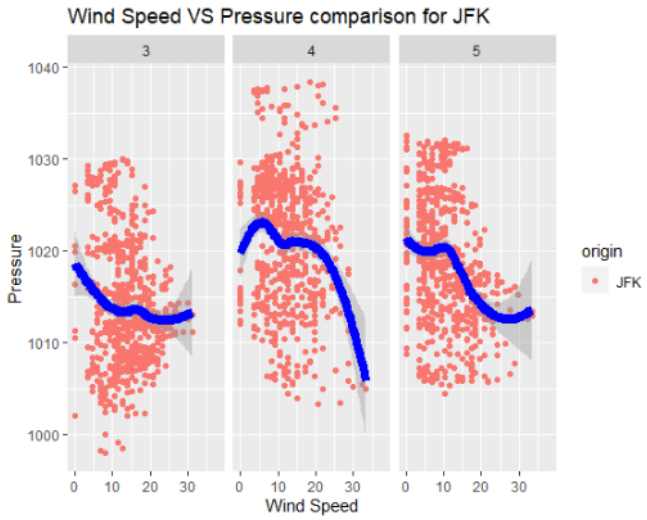

Fig 40: Wind speed vs pressure of spring season of JFK

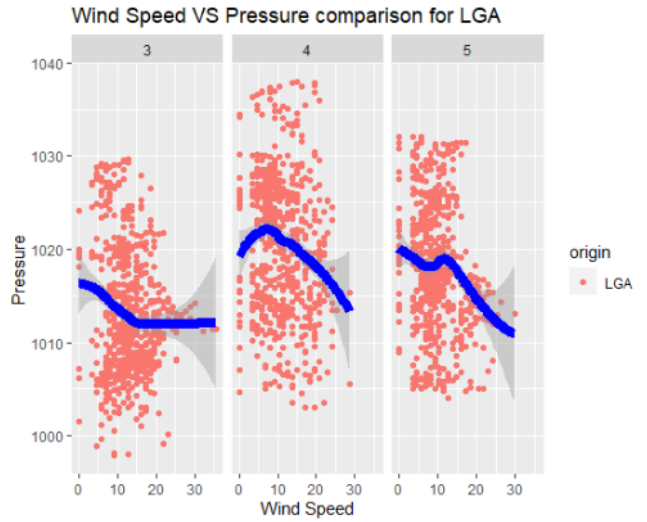

Fig 41: Wind speed vs pressure of spring season of LGA

From above figures 40 and 41, when there is minimum wind speed the pressure increases gradually. In JFK the wind speed and pressure are found maximum than LGA. In April of JFK, the lowest pressure is found where the wind speed the pressure gradually decreases in the comparison of March and May. In May of LGA, the pressure is found to be minimum than others. We can conclude that there is an inversely proportional relationship between the wind speed increases pressure.

Analysis 11:

```
#--------------------------Analysis 11-----------------------------#
#Study  the  temperature of December 22 of 2013 JFK airport
#Pre-processing
JFKDec= JFK%>% group_by(day)%>% filter(origin=="JFK")%>% filter(month=="12")%>% filter(day=="22")%>%
  select(origin, month, day,hour, temp)
ggplot(JFKDec, aes(x=hour, y=temp))+ geom_line()+ geom_point()+ scale_x_discrete(limits=c(1:23))+
 labs(title="Temperature of 22 dec of JFK",
       x="Hour", y="Temperature ")
summary(JFKDec$temp)
#Study  the  temperature of December 22 of 2013 LGA airport
LGADec= LGA%>% group_by(day)%>% filter(origin=="LGA")%>% filter(month=="12")%>% filter(day=="22")%>%
  select(origin, month, day,hour, temp)
ggplot(LGADec, aes(x=hour, y=temp))+ geom_line()+ geom_point()+ scale_x_discrete(limits=c(1:23))+
  labs(title="Temperature of 22 dec of LGA",
         x="Hour", y="Temperature ")
summary(LGADec$temp)
#-------------------------------------------------------------------#
```

Fig 42: Comparing temperature of Dec 22 of JFK and LGA

- In this analysis the hourly rise and fall of temperature of the coldest day of year i.e., December 22 are compared between JFK and LGA.
- For pre-processing, the filter () function is used to filter the origin, month, and day then using the select() function, origin, month, day, hour, and temperature are selected and stored in a variable (JFKDec for JFK and LGADec for LGA). Then, using the ggplot() function it is filled in a graph in the form of a line using geom_line() and the scale_x_discrete () function is used to provide the number for an hour from 1 to 23 as shown in figure 42.
- Similarly, the summary () function is used to find the overall summary i.e., min, max, median, number of null values, and so on.

```
summary(JFKDec$temp)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
51.98   53.47   55.22   55.29   57.02   60.80
summary(LGADec$temp)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
53.06   55.31   60.98   60.49   64.04   69.08
```

Fig 43: Summary of the temperature of Dec 22

	origin	month	day	hour	temp
1	JFK	12	22	0	51.98
2	JFK	12	22	1	51.98
3	JFK	12	22	2	51.98
4	JFK	12	22	3	53.06
5	JFK	12	22	4	51.98
6	JFK	12	22	5	53.06
7	JFK	12	22	6	53.96
8	JFK	12	22	7	53.60
9	JFK	12	22	8	53.96
10	JFK	12	22	9	55.40
11	JFK	12	22	10	55.40
12	JFK	12	22	11	59.00
13	JFK	12	22	12	59.00
14	JFK	12	22	13	57.92
15	JFK	12	22	14	57.02
16	JFK	12	22	15	55.94
17	JFK	12	22	16	55.04
18	JFK	12	22	17	55.94
19	JFK	12	22	18	57.92
20	JFK	12	22	19	60.80
21	JFK	12	22	20	55.94
22	JFK	12	22	21	57.02
23	JFK	12	22	22	55.04
24	JFK	12	22	23	53.96

	origin	month	day	hour	temp
1	LGA	12	22	0	55.04
2	LGA	12	22	1	53.06
3	LGA	12	22	2	55.40
4	LGA	12	22	3	55.04
5	LGA	12	22	4	55.04
6	LGA	12	22	5	53.60
7	LGA	12	22	6	55.40
8	LGA	12	22	7	55.04
9	LGA	12	22	8	55.94
10	LGA	12	22	9	59.00
11	LGA	12	22	10	68.00
12	LGA	12	22	11	68.00
13	LGA	12	22	12	68.00
14	LGA	12	22	13	69.08
15	LGA	12	22	14	60.08
16	LGA	12	22	15	60.98
17	LGA	12	22	16	62.06
18	LGA	12	22	17	60.98
19	LGA	12	22	18	64.04
20	LGA	12	22	19	64.04
21	LGA	12	22	20	64.04
22	LGA	12	22	21	64.04
23	LGA	12	22	22	62.96
24	LGA	12	22	23	62.96

Table 3: Table generated after pre-processing

Result

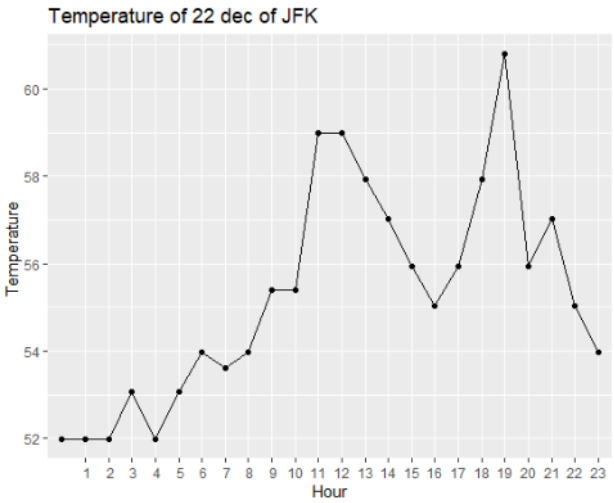

Fig 44: Line plot for an hourly temperature of Dec 22 in JFK

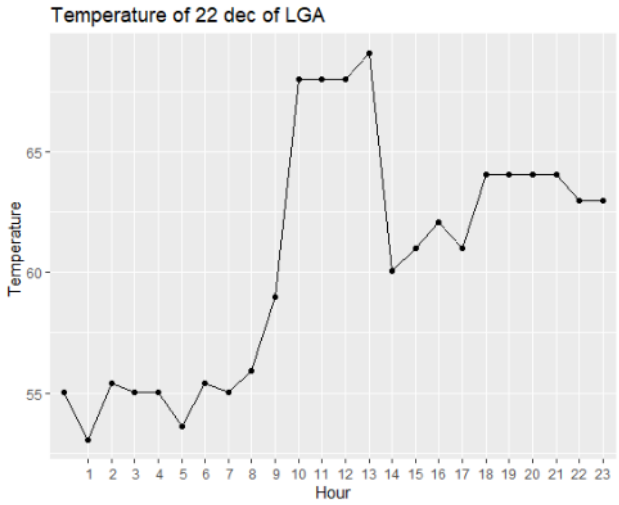

Fig 45: Line plot for an hourly temperature of Dec 22 in LGA

From the above figures 44 and 45, it is found that on the coldest day throughout the year i.e., December 22 the temperature in LGA airport is higher than JFK. At 13 o'clock, the temperature is found higher i.e., 69.08 degrees Fahrenheit in LGA. Similarly, in JFK at 19 o'clock, the temperature is found higher with the value of 60.80 degrees Fahrenheit. From the above diagram, we can also find the rise and fall of temperature on December 22.

Analysis 12:

```
#--------------------------Analysis 12-----------------------------#
#Study  the   dew point of 1st January of 2013 JFK And LGA airport
#Pre-processing for JFK
option(warn=-1)
Jan_dewpointJFK = JFK%>% group_by(day)%>% filter(origin=="JFK")%>% filter(month=="1")%>%filter(day=="1")%>%
  select(origin,month, day, hour,dewp)
ggplot(Jan_dewpointJFK, aes(x=hour, y=dewp))+ geom_line()+geom_point()+ scale_x_discrete(limits=c(1:23))+
  labs(title="Dew point of 1st dec of JFK",
         x="Hour", y="Dew point ")
summary(Jan_dewpointJFK$dewp)

#Pre-processing for LGA
option(warn=-1)
Jan_dewpointLGA = LGA%>% group_by(day)%>% filter(origin=="LGA")%>% filter(month=="1")%>%filter(day=="1")%>%
  select(origin,month, day, hour,dewp)
ggplot(Jan_dewpointLGA, aes(x=hour, y=dewp))+ geom_line()+geom_point()+ scale_x_discrete(limits=c(1:23))+
  labs(title="Dew point of 1st dec of LGA",
       x="Hour", y="Dew point ")
summary(Jan_dewpointLGA$dewp)
#-------------------------------------------------------------#
```

Fig 46: Comparing the dew point of 1st January of JFK and LGA

- In this analysis the hourly rise and fall of the dew point of the 1st January are compared between JFK and LGA.
- For pre-processing, the filter () function is used to filter the origin, month, day, hour, and dewp, and are selected using the select () function are selected and stored in a variable (Jan_dewpointJFK for JFK and Jan_dewpointLGA for LGA). Then, using the ggplot() function it is filled in a graph in the form of a line with point using geom_line() and geom_plot(), and the scale_x_discrete () function is used to provide the number for an hour from 1 to 23 as shown in figure 46. Option() function is used to remove the warning message by passing warn = -1 parameter inside it.
- Similarly, the summary () function is used to find the overall summary i.e., min, max, median, number of null values, and so on.

```
summary(Jan_dewpointJFK$dewp)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 8.06   14.77   26.06   21.76   26.96   28.04
summary(Jan_dewpointLGA$dewp)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
12.02   16.52   24.98   21.96   26.06   28.40
```

Fig 47: Summary of the dew point of 1st January

	origin	month	day	hour	dewp
1	JFK	1	1	1	26.06
2	JFK	1	1	2	26.06
3	JFK	1	1	3	26.96
4	JFK	1	1	4	28.04
5	JFK	1	1	5	26.96
6	JFK	1	1	6	26.96
7	JFK	1	1	7	28.04
8	JFK	1	1	8	26.96
9	JFK	1	1	9	26.96
10	JFK	1	1	10	28.04
11	JFK	1	1	11	26.96
12	JFK	1	1	13	26.60
13	JFK	1	1	14	24.08
14	JFK	1	1	15	23.00
15	JFK	1	1	16	17.96
16	JFK	1	1	17	17.06
17	JFK	1	1	18	14.00
18	JFK	1	1	19	14.00
19	JFK	1	1	20	14.00
20	JFK	1	1	21	10.94
21	JFK	1	1	22	10.94
22	JFK	1	1	23	8.06

	origin	month	day	hour	dewp
1	LGA	1	1	1	26.06
2	LGA	1	1	2	26.06
3	LGA	1	1	3	26.06
4	LGA	1	1	4	26.06
5	LGA	1	1	5	24.98
6	LGA	1	1	6	24.98
7	LGA	1	1	7	26.06
8	LGA	1	1	8	26.06
9	LGA	1	1	9	26.06
10	LGA	1	1	10	26.06
11	LGA	1	1	11	26.06
12	LGA	1	1	12	28.40
13	LGA	1	1	13	24.98
14	LGA	1	1	14	24.08
15	LGA	1	1	15	21.92
16	LGA	1	1	16	21.02
17	LGA	1	1	17	17.06
18	LGA	1	1	18	15.98
19	LGA	1	1	19	14.00
20	LGA	1	1	20	15.08
21	LGA	1	1	21	14.00
22	LGA	1	1	22	12.02
23	LGA	1	1	23	12.02

Table 4: Table generated after pre-processing

Result

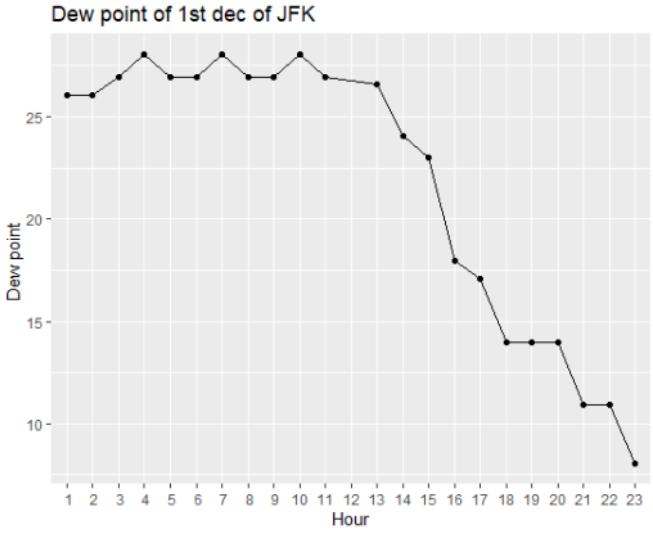

Fig 48: Line plot for an hourly dew point of JFK

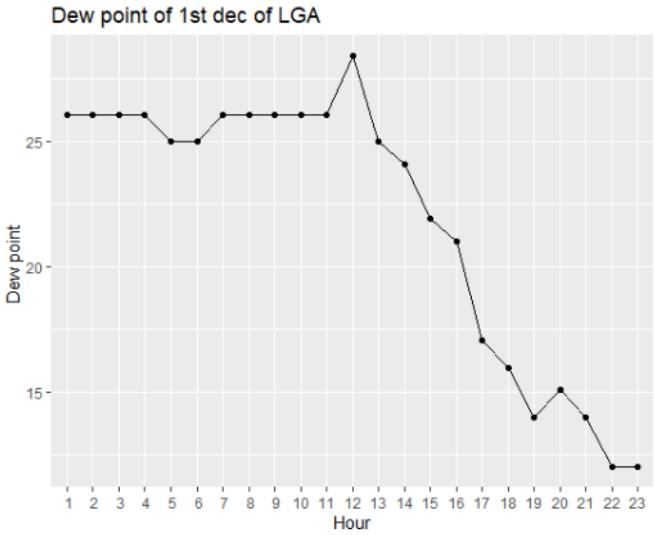

Fig 49: Line plot for an hourly dew point of LGA

From the above figures 48 and 49, it is found that the dew point on 1st January in LGA airport is higher than JFK. At 12 o'clock, the dew point is found higher i.e., 28.40 degrees Fahrenheit in LGA. Similarly, in JFK at 4, 7, 11 o'clock, the dew point is found constant and higher with the value of 20.04 degrees Fahrenheit. From the above diagram, we can also find the rise and fall of dew point on 1st January.

Analysis 13

```
#--------------------------Analysis 13-----------------------------#
#Study  the  visibility of December of 2013 JFK And LGA airport
#Pre-processing for JFK airport
JFKDecm= JFK%>% group_by(day)%>% filter(origin=="JFK")%>% filter(month=="12")%>%
  select(origin, month, day,hour, visib)
ggplot(JFKDecm, aes(x=day, y=visib)) + geom_point()+scale_x_discrete(limits=c(1:30))+
  labs(title="Visibility dec of JFK",
       x="Hour", y="Visibility ")
summary(JFKDecm$visib)
#Pre-processing for  LGA airport
LGADecm= LGA%>% group_by(day)%>% filter(origin=="LGA")%>% filter(month=="12")%>%
  select(origin, month, day,hour, visib)
ggplot(LGADecm, aes(x=day, y=visib)) + geom_point()+scale_x_discrete(limits=c(1:30))+
  labs(title="Visibility dec of JFK",
       x="Hour", y="Visibility ")
summary(LGADecm$visib)
#-------------------------------------------------------------------#
```

Fig 50: Comparing the visibility of December of both origins

- In this analysis the monthly rise and fall of the visibility of December are compared between JFK and LGA.
- For pre-processing, the filter () is used to filter the origin, and month then using the select function, origin, month, day, hour, and visibility are selected and stored in a variable (JFKDecm for JFK and LGADecm for LGA). Then, using the ggplot() function it is filled in a graph in the form of a scatterplot using geom_point() and the scale_x_discrete () function is used to provide the number for an hour from 1 to 30 as shown in figure 50.
- Similarly, the summary () function is used to find the overall summary i.e., min, max, median, number of null values, and so on.

```
summary(JFKDecm$visib)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.060   9.000  10.000   8.601  10.000  10.000
summary(LGADecm$visib)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   8.000  10.000   8.565  10.000  10.000
```

Fig 51: Summary of visibility of December

	origin	month	day	hour	visib
1	JFK	12	1	0	10.00
2	JFK	12	1	1	10.00
3	JFK	12	1	2	10.00
4	JFK	12	1	3	10.00
5	JFK	12	1	4	10.00
6	JFK	12	1	5	10.00
7	JFK	12	1	6	10.00
8	JFK	12	1	7	7.00
9	JFK	12	1	8	10.00
10	JFK	12	1	9	10.00
11	JFK	12	1	10	10.00
12	JFK	12	1	11	10.00
13	JFK	12	1	12	10.00
14	JFK	12	1	13	10.00
15	JFK	12	1	14	10.00
16	JFK	12	1	15	10.00
17	JFK	12	1	16	10.00
18	JFK	12	1	17	10.00
19	JFK	12	1	18	10.00
20	JFK	12	1	19	10.00
21	JFK	12	1	20	10.00
22	JFK	12	1	21	10.00
23	JFK	12	1	22	10.00
24	JFK	12	1	23	10.00
25	JFK	12	2	0	10.00
26	JFK	12	2	1	10.00
27	JFK	12	2	2	10.00

howing 1 to 27 of 715 entries, 5 total columns

	origin	month	day	hour	visib
1	LGA	12	1	0	10.00
2	LGA	12	1	1	10.00
3	LGA	12	1	2	10.00
4	LGA	12	1	3	10.00
5	LGA	12	1	4	10.00
6	LGA	12	1	5	10.00
7	LGA	12	1	6	10.00
8	LGA	12	1	7	10.00
9	LGA	12	1	8	10.00
10	LGA	12	1	9	10.00
11	LGA	12	1	10	10.00
12	LGA	12	1	11	10.00
13	LGA	12	1	12	10.00
14	LGA	12	1	13	10.00
15	LGA	12	1	14	10.00
16	LGA	12	1	15	10.00
17	LGA	12	1	16	10.00
18	LGA	12	1	17	10.00
19	LGA	12	1	18	10.00
20	LGA	12	1	19	10.00
21	LGA	12	1	20	10.00
22	LGA	12	1	21	10.00
23	LGA	12	1	22	10.00
24	LGA	12	1	23	10.00

owing 1 to 25 of 715 entries, 5 total columns

Table 5: Table generated after pre-processing

Result

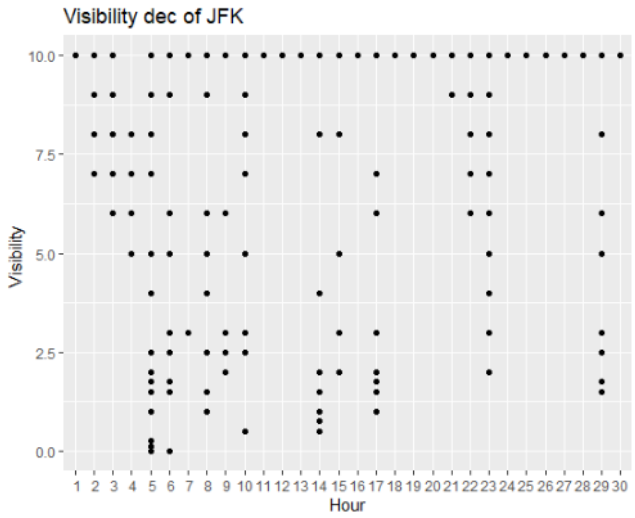

Fig 52: Scatterplot for Visibility of December in JFK

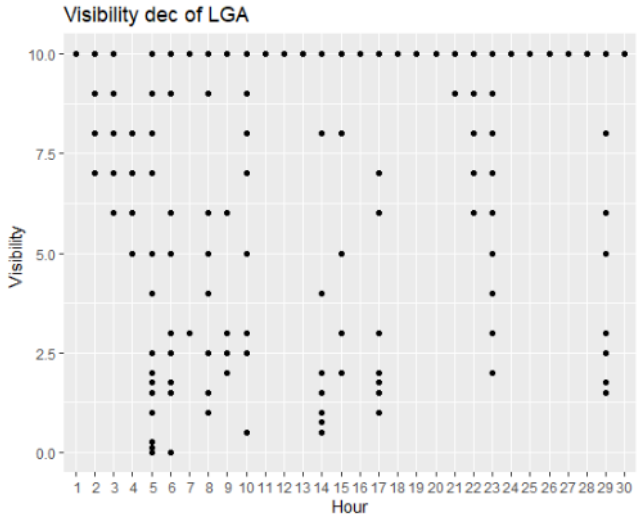

Fig 53: Scatterplot for Visibility of December in LGA

From the above figures 52 and 53, it is found that the maximum visibility in LGA and JFK are the same. The maximum value of visibility in both airports is found to be 10 miles which are found on all the days except the 4th of December. But the minimum visibility is found in LGA which is 0.00 miles and 0.060 miles in JFK on the 5th of December in both LGA and JFK.

Analysis 14:

```
#---------------------------Analysis 14-----------------------------#
#Study  the  maximum pressure of 2013 JFK And LGA airport
# for JFK airport
option(warn=-1)
ggplot(JFK %>% group_by(month) %>%summarise(pressure = max(pressure, na.rm=TRUE)),aes(x = month, y = pressure)) +
  geom_point()+geom_line() + scale_x_discrete(limits = c(1:12)) +
  labs(title="Maximum pressure  of JFK in 2013",x="Month", y="Maximum Pressure")
summary(JFK$pressure)
# for LGA airport
option(warn=-1)
ggplot(LGA %>% group_by(month) %>%summarise(pressure = max(pressure, na.rm=TRUE)),aes(x = month, y = pressure)) +
  geom_point()+geom_line() + scale_x_discrete(limits = c(1:12)) +
  labs(title="Maximum pressure  of LGA in 2013",x="Month", y="Maximum Pressure")
summary(LGA$pressure)
#-------------------------------------------------------------------#
```

Fig 54: Comparing the maximum pressure of 2013 in both origins

- In this analysis the monthly maximum pressure of 2013 is compared between JFK and LGA.
- In the above analysis ggplot() function is used to plot the graph which is grouped by month using group() function, summarise() function is used to select the pressure column with monthly maximum value, geom_point() and geom_line() is used to fill the graph in the form of the line with points and scale_x_discrete is used to provide the numbers for the month along the x-axis. Similarly, the Option ()function is used to remove the warning message by passing warn = -1 parameter inside it.
- Similarly, the summary () function is used to find the overall summary i.e., min, max, median, number of null values, and so on.

```
summary(JFK$pressure)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
985.7  1013.1  1017.9  1018.2  1023.3  1042.1     831
summary(LGA$pressure)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
983.8  1012.6  1017.4  1017.7  1022.8  1041.9     963
```

Fig 55: Summary of the pressure of JFK and LGA

Result

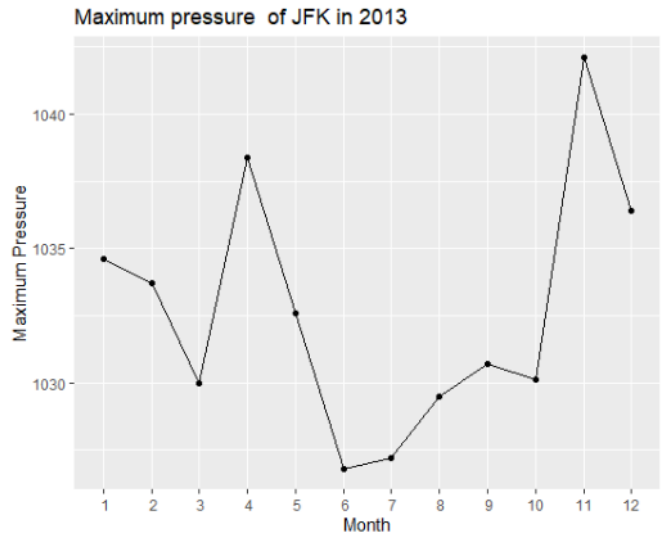

Fig 56: Line plot of monthly maximum pressure of JFK

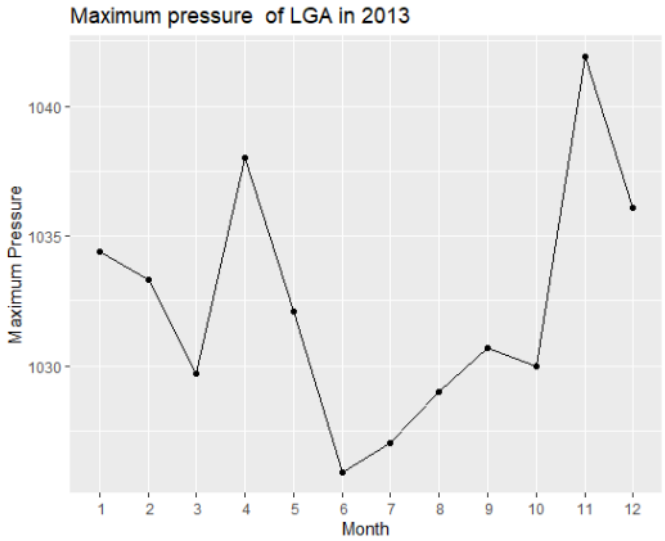

Fig 57: Line plot of monthly maximum pressure of LGA

From the above figures 56 and 57, it is found that the maximum pressure in LGA and JFK is found in November. The maximum value of pressure in JFK is found to be 1042.1milli bars which are higher than 1041.9 mill bars of LGA. But the least value of maximum pressure is found in June month of both airports (985.7 in JFK and 983.8 in LGA). Similarly, we can see the rise and fall of maximum pressure in both airports through the whole year of 2013.

5. Future Recommendation

R is found as an important language for analyzing the data set provided in this assignment and R studio is found as an excellent tool for data analysis which involves visualization and manipulation. R is also found to be a flexible programming language that has no coding rules, it results in some complex pre-processing for extracting a particular data. R studio and the file generated in the R extension are found to be less concerned with memory allocation which occupies less space and helps me to work in R studio without any problem. But the unnecessary errors which don't affect the results in the R command line reduce the user experience. Similarly, during the analysis process, it was found the late response for the visualization. These two problems were faced during the analysis process in my assignment which needs to upgrade. If these problems will be solved in the future, then R studio and R programming will provide a better experience to the users.

6. Conclusion

At the end of this assignment, I was able to better comprehend the basics of R programming. It assisted me to understand the concept of data exploration, pre-processing, modification, and visualization which results in developing my analysis techniques and improved the outcome of the data analysis for decision making. Fourteen analyses in this project also assisted me to know the relationship between different weather factors which helps in the field of weather forecasting. Similarly, the techniques employed in R studio have been learned from this assignment which aids to develop a career in the field of data science.

NPI000047 (PFDA)

ORIGINALITY REPORT

2%	1%	0%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	1%
2	Submitted to Cardinal Stritch University Student Paper	1%
3	Submitted to Arab Open University Student Paper	<1%
4	data-flair.training Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		