**Introduction**

The hourly weather dataset is the weather data of two airports in the USA. The two airports in the USA are LGA (LaGuardia Airport) and JFK (John F. Kennedy International Airport). The hourly weather dataset has 15 columns and 17,412 rows in total. The hourly weather data contains data, time, temperature, relative humidity, dew point, wind direction, wind gust speed, speed, precipitation, visibility and sea level pressure of 2013 year.

This project explores the hourly weather dataset and analyzes it, for important information which can help to make decisions or future strategy. The analysis is going to be done in R studio and by using the R programming language. The hourly weather dataset of two airports is going to be analyzed by applying R programming concepts such as data exploration, data manipulation and data visualization. The hourly weather dataset is going to be manipulated and visualized in different graphs such as scatter plot, bar chart, jitter plot, histogram, polar coordinates and line plot to understand the analyzed data. To analyze, manipulate and visualize data the packages and libraries are going to use such as ggplot2, dplyr, tidyr and magrittr.

**Assumption**

In hourly weather dataset there are many missing or NA values in many columns. The missing values in the columns are pressure, wind gust speed, wind direction and pressure. The missing values will affect it while analyzing data, it does not give accurate results or output. I assume that the mean and median imputation method is the best method to replace or fill NA or missing values (ScienceDirect, 2009). I assume that to fill or replace NA or missing values of wind direction, wind speed and wind gust speed column, the mean imputation method is the best method because the values of those columns do not contain extreme values such as 500, 1000. And I assume that to fill or replace NA values of the pressure column. Median imputation method is the best method because this column contains extreme values. I assume that the analyzed hourly meteorological data of two airports will help the airport to make future strategies.

**Aim and Objectives**

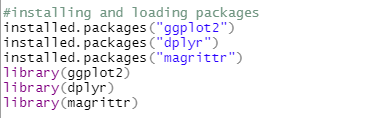
The main aim of this project is to analyze hourly weather dataset for useful output data which helps LGA and JFK airport.

The main objectives of this project are given below:

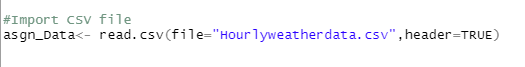
* To analyze data of every column.
* To provide useful data to both airports.
* To find out the relation between two columns of hourly weather dataset.
* To provide clear 2d dimensional visualization of analyzed data.

**Data Pre-Processing**

The packages and libraries used to analyze, manipulate and visualize are installed and loaded in RStudio, which is shown in the figure below.

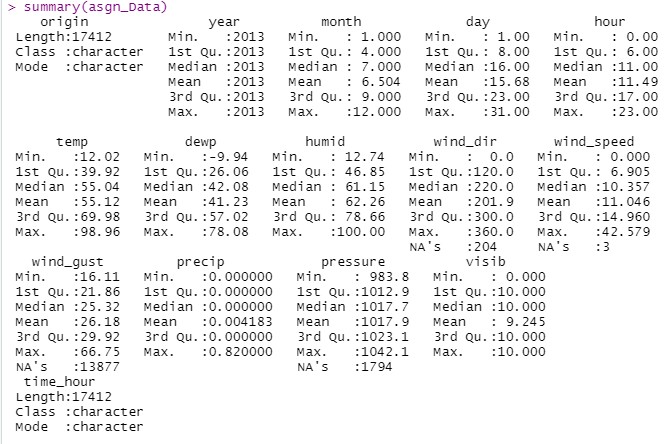


The csv file is imported into RStudio by using the read.csv function, which is shown in the figure below.



Finding the mean, min and max of data by using summary function, which is shown in the figure below.

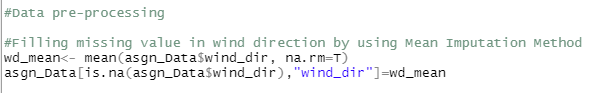




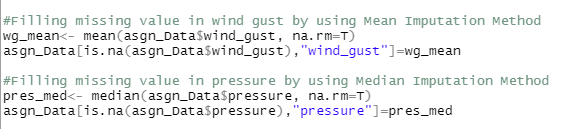
The summary function is used to find mean, median, max, min and NA of the all columns of asgn\_Data. After using the summary function, the four columns have NA values. The four columns with NA values are wind\_dir, wind\_gusr, wind\_speed and speed.

**Replacing NA values**

The hourly weather dataset contains many NA values in many columns, which affect the result of analyzed data. To replace NA values of hourly weather dataset, mean and median imputation method. The code of applying mean and median imputation method in columns which has NA values is shown in the figures below.

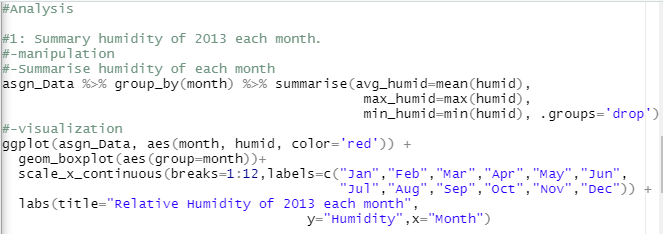


E:\old\Screenshots\Screenshot (676).png

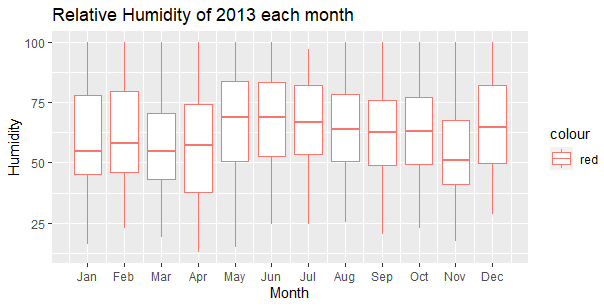


**Analysis Example**

Analysis 1: Summary humidity of 2013 each month

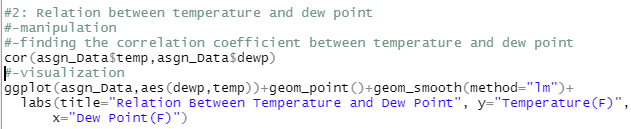


The above code is used to plot a box plot of relative humidity of each month. Box plots demonstrate the distance between the extreme values and the majority of the data (Al., 2019). Box plot helps to show mean, median, min, max, first quartile and third quartile. By using pipe operator data is read from asgn\_Data, humidity data is grouped by month from asgn\_Data by using group\_by function and the data frame is created that contains month, avg\_humid, max\_humid and min\_humid columns by using summarise function. The ggplot and geom\_boxplot function is used to plot a box plot of relative humidity. The x-axis represents month and y-axis represents humidity values in boxplot. scale\_x\_continuous function is used to label x-axis with a 12-month name. The labs function is used to add the title name as “Relative Humidity of 2013 each month” , label x-axis as”Month'' and y-axis as “Humidity”. The box plot is plotted in red color.

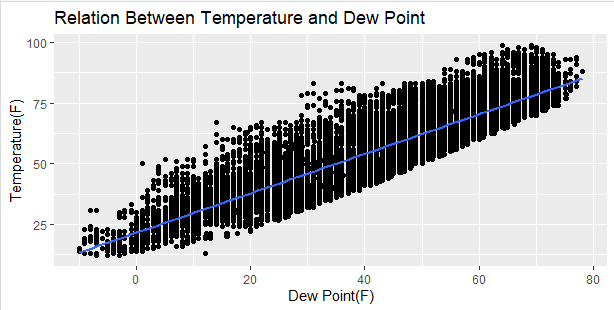


The above figure or graph shows the min, max and average humidity. According to the above figure output, July month has high average humidity and November month has low average humidity at both airports in 2013. April has min humidity value at both airports in 2013. The minimum humidity value of 2013 is 12.74.

Analysis 2: Relation between temperature and dew point

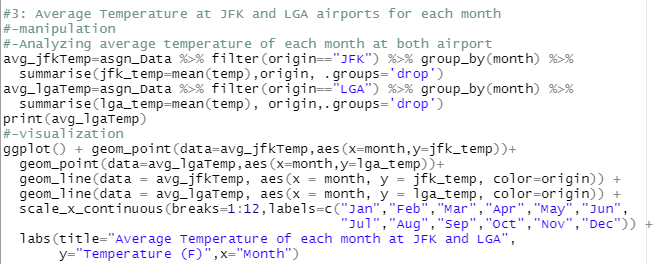


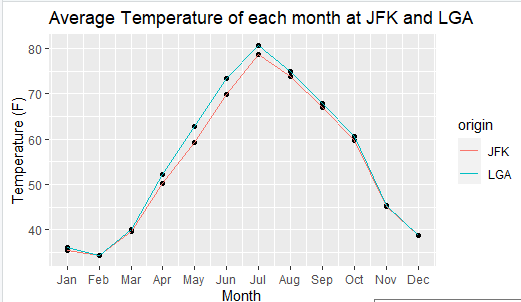
The code shown in above figure, is used to plot a scatter plot between temperature and dew point. Scatter plots help to show the relationship between two column values (Statisticshowto, 2021). The cor function is used to calculate correlation coefficient between temperature and dew point. The $ sign is used to access the values of dew point and temperature. By using ggplot and geom\_point function a scatter plot between temperature and dew point is plotted. In the x-axis dew point and in the y-axis temperature are placed in a scatter plot. The geom\_smooth is used to find the relationship between temperature and dew point. And method=”ln” is used to plot a linear. The labs function is used to add the title name as “Relation Between Temperature and Dew Point”, label x-axis as “Dew Point(F)” and y-axis as “Temperature(F)”.



The above figure or graph shows the relation between temperature and dew point. According to the above figure output, there is a positive linear relationship between temperature and dew point as the regression line is moving upward. The temperature of airport will increase while dew points of airport increase and vice versa. The correlation between temperature and dew point is 0.896, it shows that the relationship between temperature and dew point is very strong.

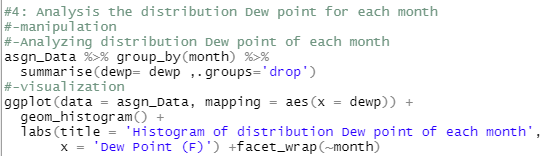
Analysis 3: Average Temperature at JFK and LGA airport for each month.



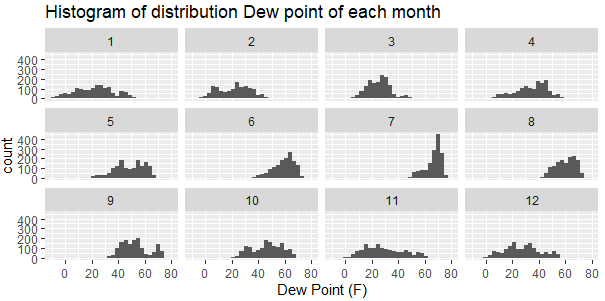
The above code is used to plot a line chart of average temperature at both airports for each month. Line charts help to display data that is updated over time. Line charts display quantitative data over a specific time interval (Peters, 2022). The average temperature data of JFK and LGA airport is stored in separate variables. By using filter function, the temperature of both airports are filtered, by using group function the filter data are grouped by month and summarise average temperature with origin by using summarise function. The ggplot, geom\_point and geom\_line function is used to plot a line chart of average temperature at both airports for each month. The geom\_point function is used to point to the average temperature of each month. A geom\_line function is used to display temperature data over a month interval. In the x-axis month and in the y-axis temperature are placed in a line chart. scale\_x\_continuous function is used to label x-axis with a 12-month name. The labs function is used to add the title name as “Average Temperature of each month at JFK and LGA”, label x-axis as”Month'' and y-axis as “Temperature”.

The above figure or graph shows the average temperature at both airports for each month. According to the above figure output, the July month has high average temperature and February month has low average temperature at both airports. The July average temperature is 81.96 F at JFK and 79.06 F at LGA airport. The February average temperature is 12.08 F at JFK and 12.98 F at LGA airport.

Analysis 4: Analysis of distribution Dew point for each month.

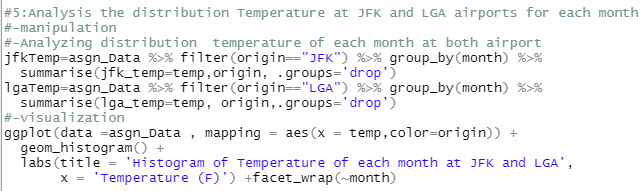


The above code is used to plot a histogram of distribution Dew point for each month. Histogram is the best tool for summarizing continuous data on an interval scale and it is suitable for large datasets (CHEN, 2021).  By using pipe operator dew point values are read from asgn\_Data, dew point values are grouped by month from asgn\_Data by using group\_by function and new data frame is created with month and dew columns by using summarise function. The ggplot and geom\_histogram function is used to plot a histogram of distribution Dew point for each month. In the x-axis of histogram, the dew point is placed. The labs function is used to add the title name as “Histogram of distribution Dew point of each month” , label x-axis as”Dew Point (F)''. The facet\_wrap function is used for visualizing data in 2d according to month.

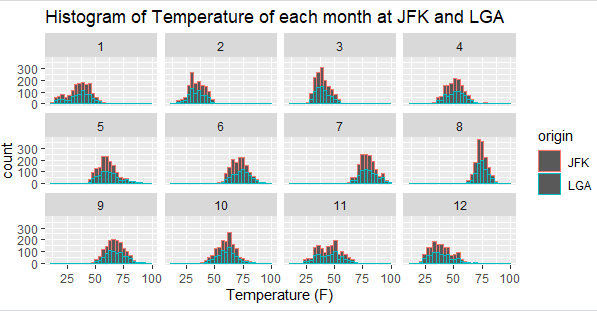


The above figure or graph shows a histogram of distribution Dew point for each month. According to the above figure output, the July month has a high dew point between 40F to 80F and February month has a low dew point between 10F to 65F. From dew point we can analyze the visibility. In February there is a high risk of worse visibility caused by thick fog.

Analysis 5: Analysis of distribution temperature at JFK and LGA airports for each month.

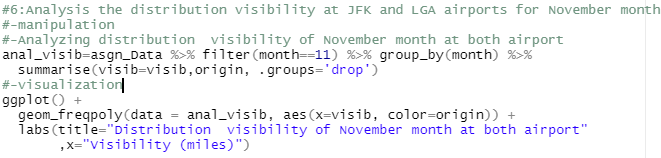


The above code is used to plot a histogram of distribution temperature at JFK and LGA airport for each month. The histogram is selected to plot the histogram of distribution temperature at both airports because it is suitable for a large number of dataset. The histogram is used to summarize distribution temperature on a month interval scale. The temperature values of both airports are stored in separate data frames.  By using pipe operator temperature values are read from asgn\_Data. By using filter function, the temperature of both airports are filtered, by using group function the filter data are grouped by month and a new data frame is created with month, lga\_temp and origin columns by using summarise function. The geom\_histogram and ggplot function is used to plot a histogram of distribution temperature at JFK and LGA airport for each month. In the x-axis of histogram, the temperature is placed and color is used to separate temperature values based on origin. The labs function is used to add the title name of the graph as “Histogram of Temperature of each month at JFK and LGA”, label x-axis as “Temperature (F)”. The facet\_wrap function is used for visualizing data in 2d according to month.

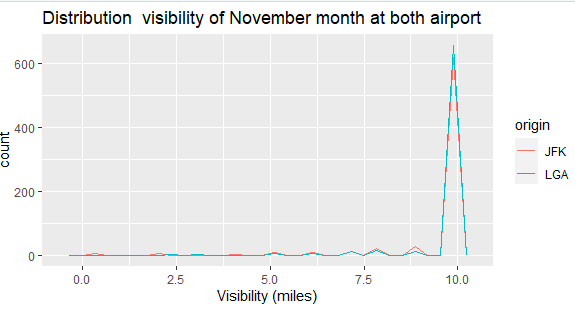


The above figure or graph shows a histogram of distribution Dew point for each month. According to the above figure output, both airports have high temperature in June month and low temperature in February month. Based on the above figure output, the JFK airport has higher temperature than LGA airport in every month. The JFK airport airbus needs a long runway for take-off because in hot weather airbus need high speed to take off.

Analysis 6: Analysis of distribution visibility at JFK and LGA airports for November month.

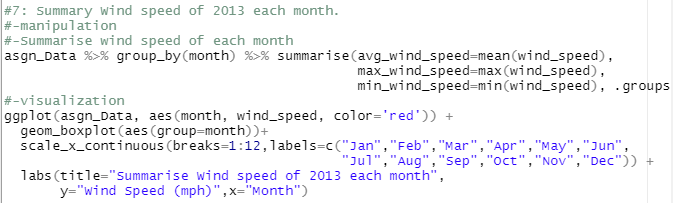


The above code is used to plot a frequency polygon chart of distribution visibility at both airports for November month. A frequency polygon is the best graph for figuring out how distributions are shaped (Libretexts, 2022). The anal\_visib variable is created to store visibility values at both airports for November month. The pipe operator is used to read data from asgn\_Data. A filter function is used to filter month which values are equal to 11 and group\_by function is used to group filter values. The summarise function is used to create a new data frame that contains month, visib and origin columns. In frequency polygon graph x-axis represents the values of visibility of November month and color is used to separate visibility values based on origin. By using labs function the title name of the graph as “Distribution visibility of November month at both airport”, label x-axis as”Visibility (miles)'' are added.

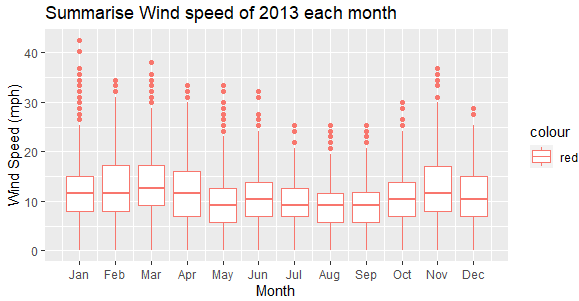


The above figure or graph shows a frequency polygon graph of distribution visibility of November month at both airports. According to the above figure output, both airports have a high number of 10 miles’ visibility values. JFK airport is more visible than LGA airport because the LGA contains a smaller number of visibility values such as between 0 to 5 miles. Both airports are visible to land and take off in November month because both contain less numbers between 0 to 1 miles.

Analysis 7: Summary of wind speed of 2013 each month.

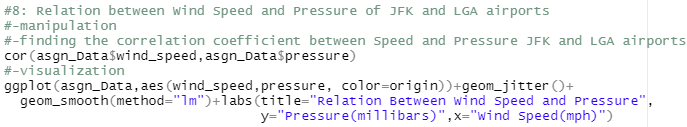


The above code is used to plot a box plot of wind speed of 2013 each month. The box plot is used to summarize wind speed of 2013 each month because the box plot is famous for summary mean, median, min, max, Q1 and Q3 (Miller, 2017). By using pipe operator wind speed values are read from asgn\_Data. By using filter function, the wind speed values are filtered from asgn\_Data, by using group function the filter data are grouped by month and a new data frame is created with month, avg\_wind\_speed, max\_wind\_speed and min\_wind\_speed columns by using summarise function. The ggplot and geom\_boxplot function is used to plot a box plot of wind speed of 2013 each month. The x-axis represents month and y-axis represents wind speed values of asgn\_Data in the box plot. The  scale\_x\_continuous function with breaks =1:12 is used to label x-axis with a 12-month name. The labs function is used to add the title name as “Summarise Wind speed of 2013 each month”, label x-axis as”Month'' and y-axis as “Wind Speed (mph)”. The box plot is plotted in red color by using color = ‘red’.

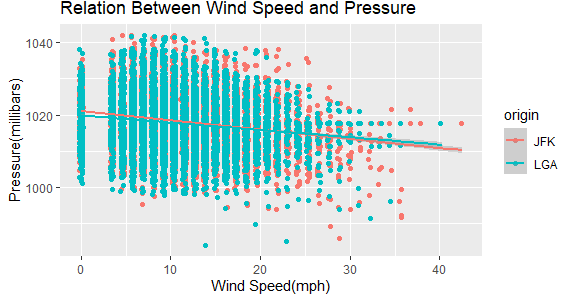


The above figure or graph shows a box plot of wind speed of 2013 each month. According to the above figure output, March month has a higher average value and May month has low average value of wind speed in 2013. The average value of March month in 2013 is 13.587mph and average value of May month in 2013 is 9.107mph. The January month has a max speed value of 42mph in 2013. Above 40mph wind speed will affect landing and take-off, so in January there is a high chance of canceling flight.

Analysis 8: Relation between speed and pressure of JFK and LGA airports.

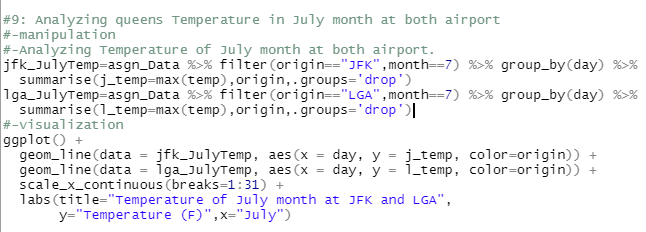


The above code is used to plot a jitter plot between wind speed and pressure of JFK and LGA airport. Jitter plot is same as scatter plot but jitter plot visualizes the data or relation between two columns of dataset detaily (Loop, 2011). The cor function is used to calculate correlation coefficient between wind speed and pressure. The $ sign is used to access the values of pressure and wind speed. By using ggplot and geom\_jitter function a jitter plot between wind speed and pressure is plotted. The x-axis of the jitter plot represents the values of wind speed and the y-axis of the jitter plot represents the values of pressure of hourly weather dataset and color is used to separate wind speed values and pressure values based on origin. The geom\_smooth is used to find the relationship between wind speed and pressure. And method= “ln” is used to plot a linear. The labs function is used to add the title name as “Relation Between Wind Speed and Pressure”, label x-axis as “Wind Speed(mph)” and y-axis as “Pressure(millibars)”.

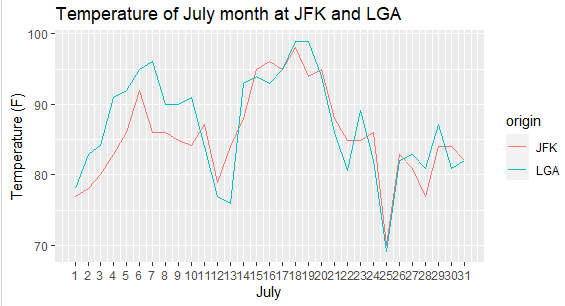


The above figure or graph shows a jitter plot of relation between speed and pressure of JFK and LGA airport. According to the above figure output, there is a negative linear relationship between pressure and wind speed as the regression line is moving downward. The pressure of the both airports will decrease while wind speed of the both airports increases and vice versa. The correlation between pressure and wind speed is -0.183, it shows that the relationship between pressure and wind speed is very weak.

Analysis 9: Analyzing queen temperature in July month at both airports.

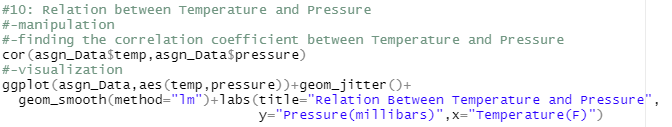


The above code is used to plot a line chart of queen temperature in July month at both airports. The temperature values of July month at both airports are stored in separate data frames.  By using pipe operator temperature values are read from asgn\_Data. By using filter function, the temperature of both airports are filtered and month that is equal to 7 is also filtered, by using group function the filter data are grouped by day and a new data frame is created with day, j\_temp and origin columns for jfk\_JulyTemp and day, l\_temp and origin columns for lga\_JulyTemp by using summarise function. The geom\_line and ggplot function is used to plot a line chart of queen temperature in July month. In the x-axis of the line chart the day is placed, in the y-axis j\_temp and y\_temp are placed and color is used to separate temperature values based on origin. scale\_x\_continuous function is used to label x-axis from 1 to 31 by using breaks=1:31. The labs function is used to add the title name of the graph as “Temperature of July month at JFK and LGA”, label y-axis as “Temperature (F)” and x\_axis as “July”.

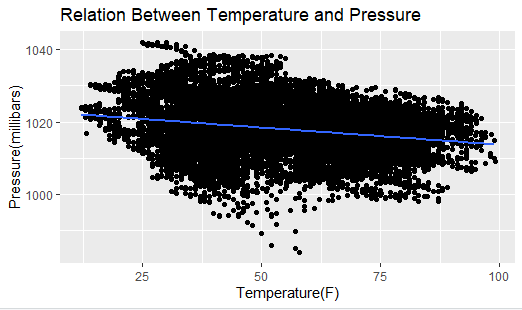


The above figure or graph shows a line chart of queen temperature in July month at both airports. According to the above figure output, the LGA has a higher temperature than JFK. Based on the above figure output, on the 18th day of July month both airports had maximum temperature which is more than 95F.  And on the 22th day of July month both airports have a minimum temperature which is less than 70F.

Analysis 10: Relation between temperature and pressure.

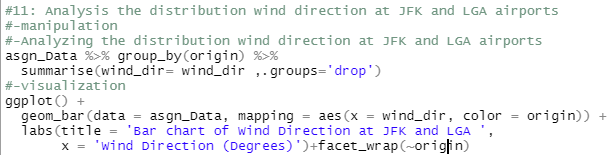
****

The above code is used to plot a Scatter plot between temperature and pressure. A scatter plot is suitable to compare or show the relationship between two columns of a dataset. The cor function is used to calculate correlation coefficient between temperature and pressure. And the $ sign is used to access the values of pressure and temperature. By using ggplot and geom\_point function a scatter plot between temperature and pressure is plotted. The x-axis of the scatter plot shows the values of temperature and the y-axis of the scatter plot shows the values of pressure of hourly weather dataset. The geom\_smooth is used to find the relationship between temperature and pressure. And method= “ln” is used to plot a linear. The labs function is used to add the title name as “Relation Between Temperature and Pressure”, label x-axis as “Temperaure(F)” and y-axis as “Pressure(millibars)”.

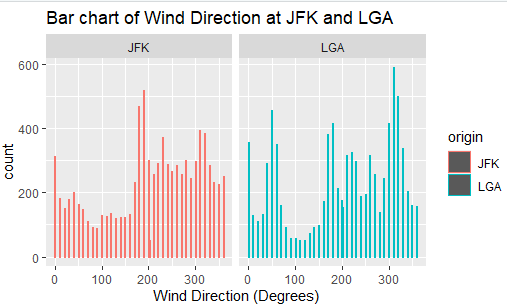
****

The above figure or graph shows a scatter plot of relation between temperature and pressure. According to the above figure output, there is a negative linear relationship between pressure and Temperature as the regression line is moving downward. The pressure of the both airports will decrease while temperature of the both airports increases and vice versa. The correlation between pressure and temperature is -0.236, it shows that the relationship between pressure and wind speed is weak.

Analysis 11: Analysis of distribution wind direction at JFK and LGA airports.

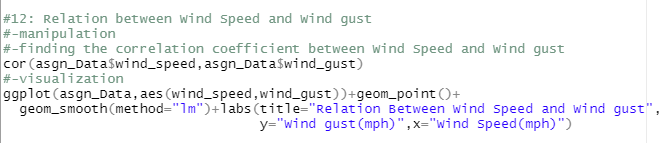
****

The above code is used to plot a bar graph of distribution wind direction at JFK and LGA airports. Bar graph is the graph tool that shows the data of a variable in a rectangle bar (CUEMATH, 2015). By using pipe operator data is read from asgn\_Data, wind direction data is grouped by origin from asgn\_Data by using group\_by function and the data frame is created that contains wind direction and origin columns by using summarise function. The geom\_bar and ggplot function is used to plot a bar chart of distribution wind direction at JFK and LGA airport. In the x-axis of the bar chart the data of wind direction is shown and color is used to separate wind direction values based on origin. The labs function is used to add the title name of the graph as “Bar chart of Wind Direction at JFK and LGA”,  x\_axis as “Wind Direction (Degrees)”. The facet\_wrap function is used for visualizing data in 2d according to origin.

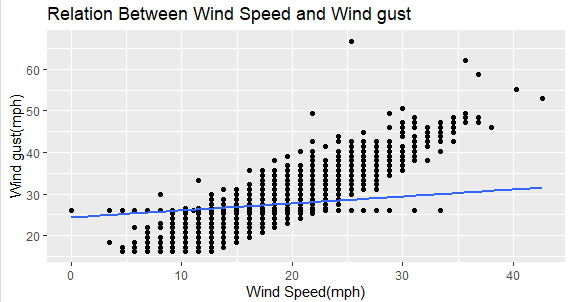
****

The above figure or graph shows a bar chart of wind direction at JFK and LGA. According to the above figure output, the maximum wind blows from 320 degrees and 310 degree and minimum wind blows from 100 degrees and 110 degrees in LGA airports. The maximum wind blows from 180 degrees and 190 degree and minimum wind blows from 90 degrees and 80 degrees in JFK airports.  According to the figure output, in LGA airport the wind blows more than JFK airport.

Analysis 12: Relation between wind speed and wind gust speed.

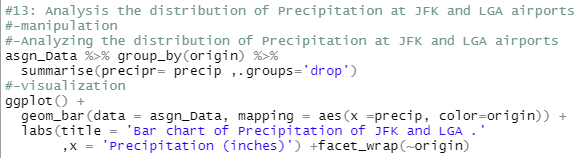
****

The above code is used to plot a Scatter plot between wind speed and wind gust speed.  The cor function is used to calculate correlation coefficient between wind speed and wind gust speed. And the $ sign is used to access the values of wind speed and wind gust speed from the hourly weather dataset. By using ggplot and geom\_point function a scatter plot between wind speed and wind gust speed is plotted. The x-axis of the scatter plot shows the values of wind speed and the y-axis of the scatter plot shows the values of wind gust speed of hourly weather dataset. The geom\_smooth is used to find the relationship between wind speed and wind gust speed. And method= “ln” is used to plot a linear. The labs function is used to add the title name as “Relation Between Wind Speed and Wind gust”, label x-axis as “Wind Speed(mph)” and y-axis as “Wind gust(mph)”.

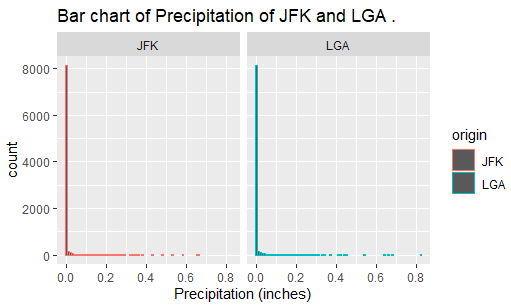
****

The above figure or graph shows a scatter plot of relation between wind speed and wind gust speed. According to the above figure output, there is a positive linear relationship between wind speed and wind gust speed as the regression line is moving upward. The wind speed of the airport will increase while wind gust speed of the airports increases and vice versa. The correlation between wind speed and wind gust speed is 0.3523, it shows that the relationship between pressure and wind speed is weak.

Analysis 13: Analysis of the distribution of precipitation at JFK and LGA airports.

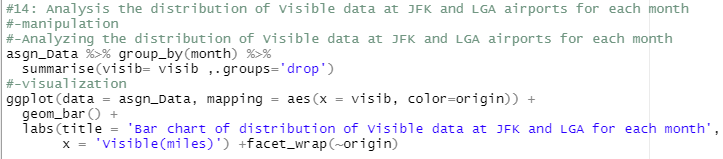
****

The above code is used to plot a bar chart of distribution of precipitation at JFK and LGA airport. By using pipe operator data is read from asgn\_Data, precipitation data is grouped by origin from asgn\_Data by using group\_by function and the data frame is created that contains prescripr and origin columns by using summarise function. The geom\_bar and ggplot function is used to plot a bar chart of distribution precipitation at JFK and LGA airport. In the x-axis of the bar chart the data of precipitation is shown and color is used to separate wind precipitation values based on origin. The labs function is used to add the title name of the graph as “Bar chart of Precipitation of JFK and LGA”,  x-axis as “Precipitation (inches)”. The facet\_wrap function is used for visualizing data in 2d according to origin.

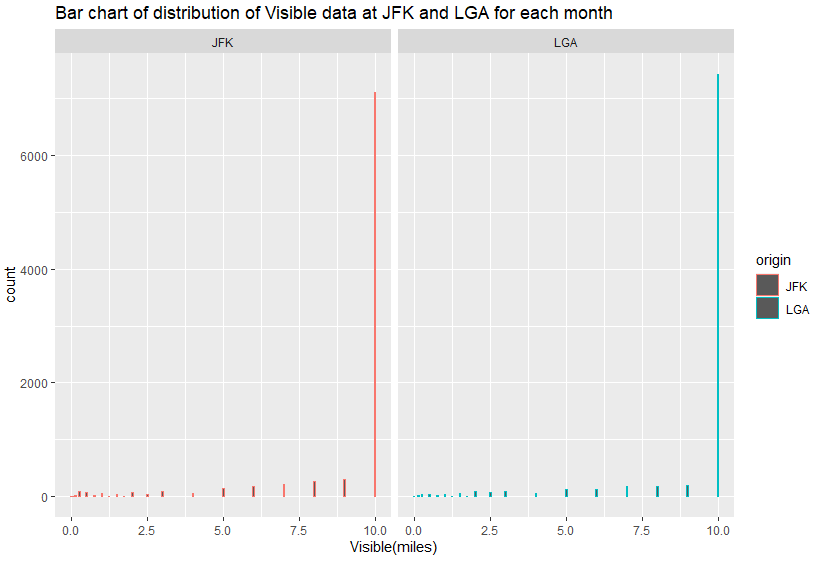
****

The above figure or graph shows a bar chart of Precipitation at JFK and LGA. According to the above figure output, both airports had contained a maximum number of 0.0 inches’ precipitation in hourly weather dataset. This means, there is a very little chance of rain at both airports.

Analysis 14: Analysis of the distribution of visible at JFK and LGA airports for each month.



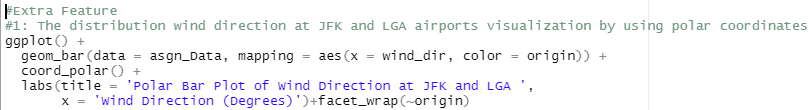
The above code is used to plot a bar chart of distribution of visible at JFK and LGA airport for each month. By using pipe operator data is read from asgn\_Data, visible values are grouped by origin from asgn\_Data by using group\_by function and the data frame is created that contains visib and origin columns by using summarise function. The geom\_bar and ggplot function is used to plot a bar chart of distribution visible at JFK and LGA airport. In the x-axis of the bar chart the data of visible is shown and color is used to separate wind precipitation values based on origin. The labs function is used to add the title name of the graph as “Bar chart of Precipitation of JFK and LGA”,  x-axis as “Visible (miles)”. The facet\_wrap function is used for visualizing data in 2d according to origin.



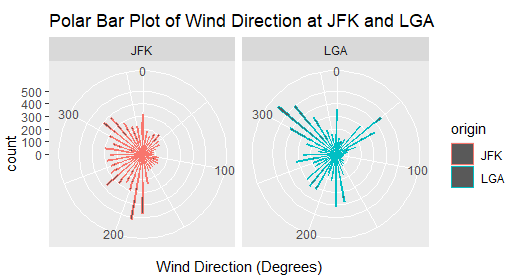
The above figure or graph shows a bar chart of visible data at JFK and LGA for each month. According to the above figure output, both airports contained a maximum number of 10.0 miles visible in hourly weather dataset. Both airports are properly visible. This means, there is a very less chance of flights being canceled at both airports.

**Additional Feature**

**Polar Bar plot of wind direction at both airports by using polar coordinates.**

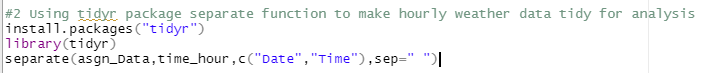


The above code is used to plot a polar bar plot of wind direction at JFK and LGA airport. By using polar coordinates, the polar car of wind direction is plotted. Polar coordinates are used to connect bar chart coxcomb chart. The geom\_bar and coord\_polar function is used to plot a polar bar plot of wind direction at JFK and LGA airport. In the x-axis of the polar bar plot the data of wind direction is shown and color is used to separate wind direction values based on origin. The labs function is used to add the title name of the graph as “Polar Bar Plot of Wind Direction at JFK and LGA '‘, x-axis as “Wind Direction (Degrees)”. The facet\_wrap function is used for visualizing data in 2d according to origin.



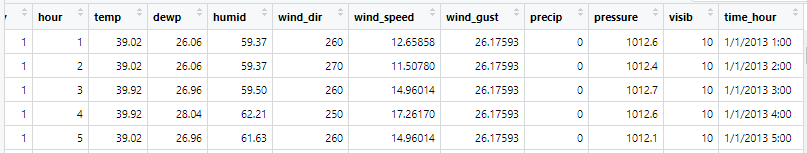
The above figure or graph shows a polar bar plot of wind direction at JFK and LGA. The above figures show the clear or detailed visualization of wind direction. By using polar bar plot, we can see from what degree the wind is blowing maximum and minimum. From that analysis the airport can make a future strategy related to the wind direction case.

**Using tidyr package to make data tidy.**

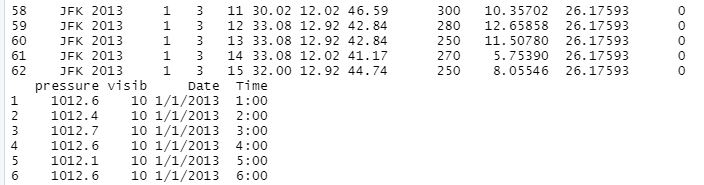
****

The tidyr package is used in this project to make hourly weather data tidy. In the above figure, the tidyr package is installed and used to make data tidy. The separate function of the tidyr package is used to separate time\_hour columns into Date and Time columns from asgn\_Data. The Date and Time columns are useful for analysis related to time and date.

Before using separate function.



After using separate function.



**Future enhancement**

* Analyze data for weather forecasting.
* Provide 3d dimension visualization of analyzed data.
* Analyze visual and pressure data detail.
* Provide more useful analyzed data by using other packages.

**Conclusion**

The 'hourly weather' project is done for weather analysis of two different airports (i.e, JFK and LGA). The project is done in R programming using a free and open source IDE called RStudio. Throughout the analysis, I explored, manipulated and analysed the provided CSV dataset, to retrieve meaningful information out of it. Distinctive R packages (built-in and installed) were used in manipulating and analysing the dataset. A total of 14 variables were provided with different weather information. Unavailable observations were replaced or eliminated with data pre-processing so that results precision can be maintained. The analysis came out helpful as it can assist airport management to make sensible decisions according to the weather of JFK and LGA airports.

**References**

* ScienceDirect. (2009, September 21). *Imputation Method - an overview | ScienceDirect Topics*. Missing Data. Retrieved February 25, 2022, from https://www.sciencedirect.com/topics/mathematics/imputation-method#:%7E:text=Mean%20imputation%20(MI)%20is%20one,Jamshidian%20and%20Bentler%2C%201999).
* Al., E. O. I. B. (2019, March 30). *Box Plots | Introduction to Statistics*. Introduction to Statistics. Retrieved March 1, 2022, from https://courses.lumenlearning.com/introstats1/chapter/box-plots/
* Statisticshowto. (2021, June 8). *Scatter Plot / Scatter Chart: Definition, Examples, Excel/TI-83/TI-89/SPSS*. Statistics How To. Retrieved March 2, 2022, from https://www.statisticshowto.com/probability-and-statistics/regression-analysis/scatter-plot-chart/
* Peters, K. (2022, February 18). *Understanding a Line Graph*. Investopedia. Retrieved March 3, 2022, from https://www.investopedia.com/terms/l/line-graph.asp
* CHEN, J. A. M. E. S. (2021, August 18). *Histogram Definition*. Investopedia. Retrieved March 5, 2022, from https://www.investopedia.com/terms/h/histogram.asp#:%7E:text=A%20histogram%20is%20a%20graphical,into%20logical%20ranges%20or%20bins.
* Libretexts. (2022, January 6). *2.5: Frequency Polygons*. Statistics LibreTexts. Retrieved March 6, 2022, from https://stats.libretexts.org/Bookshelves/Introductory\_Statistics/Book%3A\_Introductory\_Statistics\_(Lane)/02%3A\_Graphing\_Distributions/2.05%3A\_Frequency\_Polygons
* Miller, T. (2017, July 7). *What Is a Box Plot and When to Use It*. Chartio. Retrieved March 6, 2022, from https://chartio.com/resources/tutorials/what-is-a-box-plot/
* Loop, T. O. W. R. D. O. (2011, July 6). *To jitter or not to jitter: That is the question*. The DO Loop. Retrieved March 6, 2022, from https://blogs.sas.com/content/iml/2011/07/06/to-jitter-or-not-to-jitter-that-is-the-question.html
* CUEMATH. (2015, June 21). *Bar Graph - Properties, Uses, Types | How to Draw Bar Graph?* Retrieved March 2, 2022, from https://www.cuemath.com/data/bar-graphs/