WEATHER DATA ANALYSIS ACROSS US

Avengers Assemble:

Members

- Pramathesh Shukla
- Kashyap Dobariya
- Shibani Maran
- Madhusudhan Gowda

Instructor: Dr. Eli Brown

Table of Contents:

Introduction:

Weather is the day-to-day state of atmosphere and pertains to short term changes in conditions of heat, moisture, and air movement. Weather results from processes that attempt to equalize the differences in the distribution of net radiant energy from sun. In other words, the instantaneous state of atmosphere can be called as weather. It is usually expressed as fine, fair, foggy, cloudy, rainy, sunny, or windy weather.

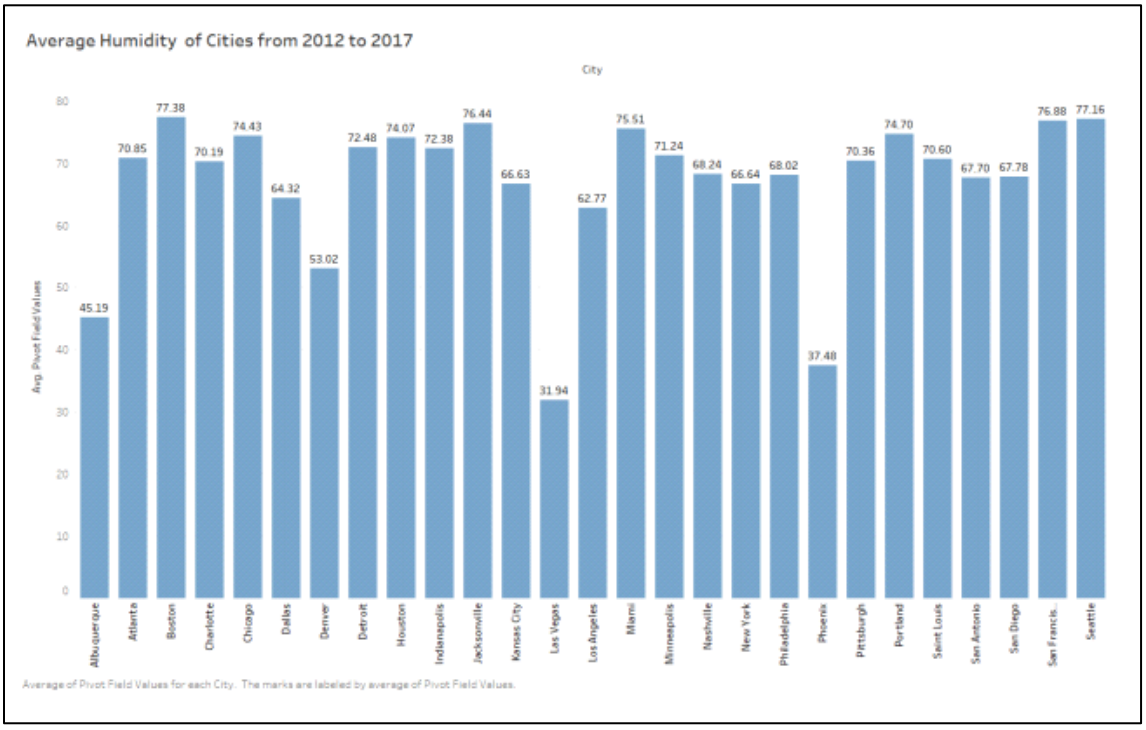
For this project, we selected the historical hourly weather data from 2012-2017. This dataset was derived from Kaggle.com. Each file contains of separate aspect of weather. For an example, humidity, pressure, temperature, wind direction, wind speed, and weather description. We conducted those data to do the visualization with maps, bar chart, line plot, scatterplot, circular, violin plot, and mosaic plot. All the 6 files consist of 45,254 rows and 37 variables in which 36 are cities and 1 daytime. Overall, this is a time series data with geospatial variables. After cleaning and analyzing the data we will use different US cities with humidity, temperature, wind direction, wind speed and weather description.

Exploratory Analysis:

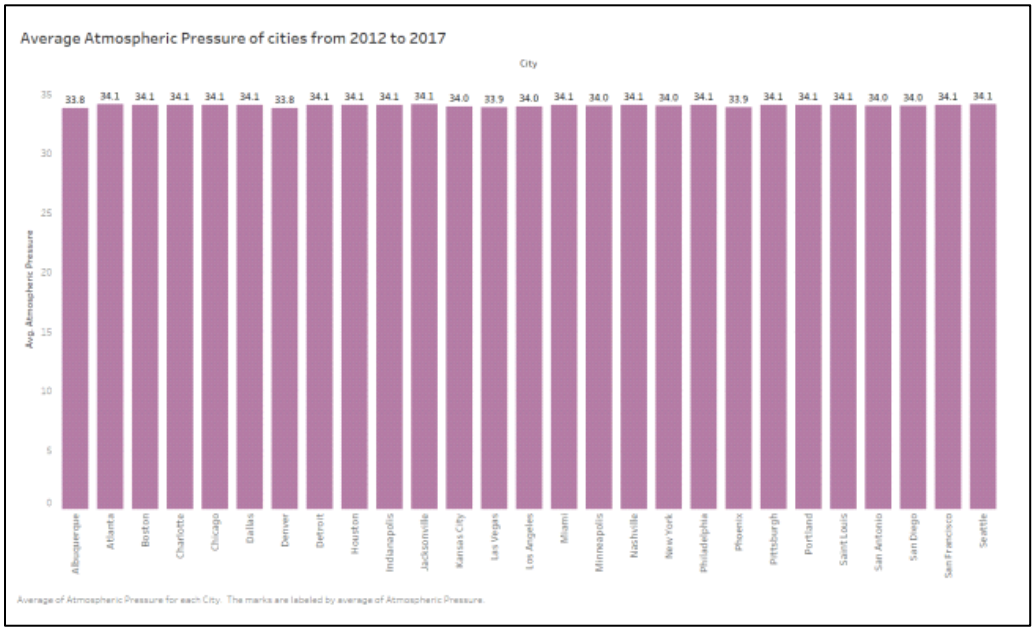
Our group decided to work with time-series data of the weather. As weather is day to day state of the atmosphere and concerns to short term changes in conditions of pressure, temperature, moisture, and air movement at a particular place at any given time and it is highly variable.

As we chose the time series data of 6 files which had different aspects of weather and we took pivot of all the columns to the aggregate city into one. In the next step, we also transformed the data into a scale which are used universally to make sure it was not hard for the audience to understand.

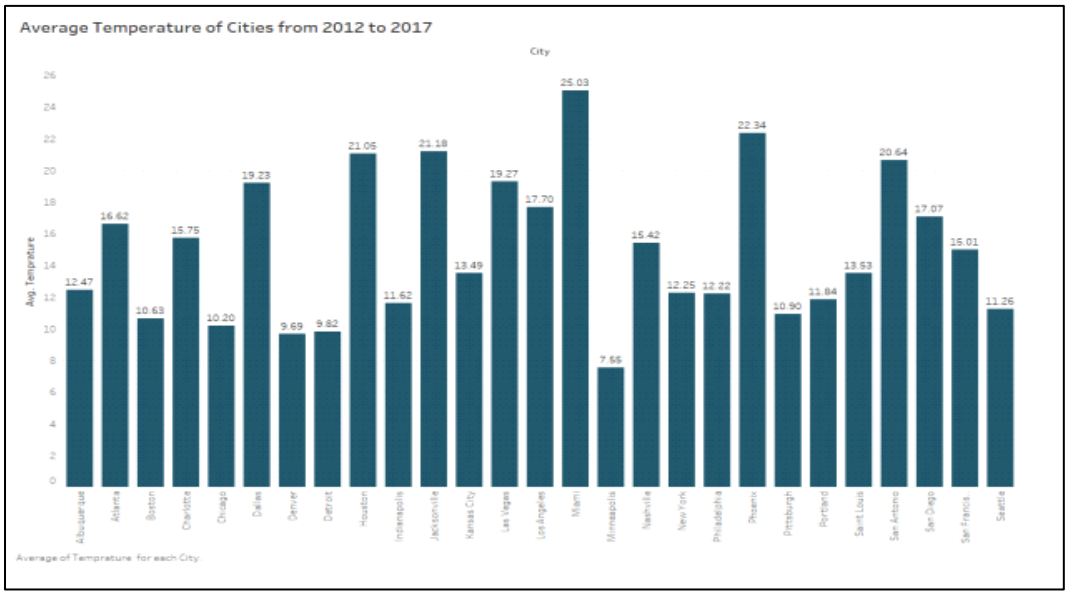
After pre-processing, we started to build some basic plots and gather the information about the dataset. So, our first plot was created by taking the average of all the humidity data collected from 2012 to 2015 city wise. The bar plot of the cities versus the Average humidity is displayed below. From the below graph we can see that Las Vegas has the lowest average humidity, and this is because it is a city in the desert with low rainfall.



The second bar plot is displayed below which consists of the average atmospheric pressure vs the cities. The pressure was converted to be more readable format from Inch of Mercury (inHg) to Atmospheric pressure unit(atm). The bar plot shows different cities of united stated having the average atmospheric pressure captured daily from 2012 to 2017. Which is almost equal in every city in other words in whole United States.



The Third bar plot is displayed below which consists of the average temperature vs the cities. The temperature was converted to be more readable format from Kelvin (K) to Celsius (C) unit. The bar plot shows different cities of united stated having the average temperature captured daily from 2012 to 2017.



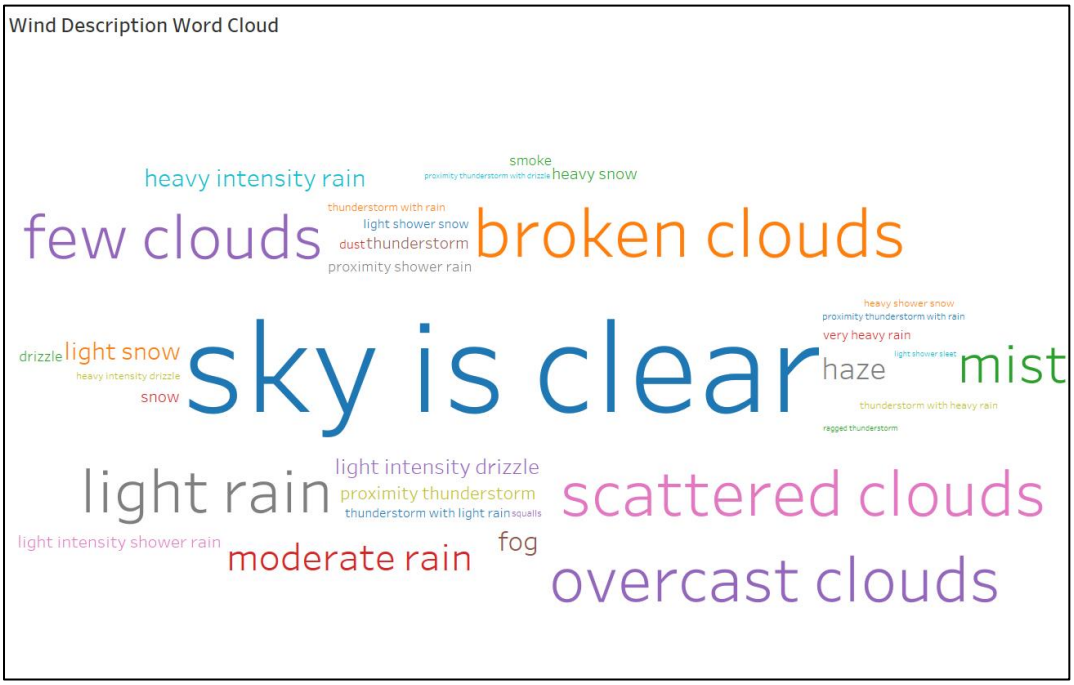
After trying a few plots with the basic information which gave us insights about how the data was recorded and we only used simple combination of attributes such as temperature vs cities, pressure vs cities. We then planned on plotting few visualizations using different graphs to explore combinations of few attributes to compare and get insights about the patterns. For e.g., will the wind speed vary with the temperature in a certain area and finding the relationship between temperature and pressure in the same cities.

Visualization

Visualization – 1: Graphs built on categorical data of Weather Description.

Out of total 7 files of the main dataset one of the files is Weather Description. Which contains the categorical variables only, in other words the file contains the description of the weather at the time noted in each city. So, data consist of only categorical variables and in order to mold that information in easily interpretable format we have generated two graphs which works were well with this kind of data, which are displayed below.

The word cloud possesses all the words used in the dataset and the size of the words in the graph suggest the frequency of each separate word used. In other words, we can say more the usage of a particular word the bigger the font size of that word in the graph. For example, if was look at word cloud graph we can see word **sky is clear** has the largest font size of all, that suggests the maximum usage of that word in the dataset. Similarly, scattered clouds, overcast clouds, broken clouds are also noticed more frequently in the dataset and hence their font size is bigger then mist or haze. The Word cloud graph is displayed below.

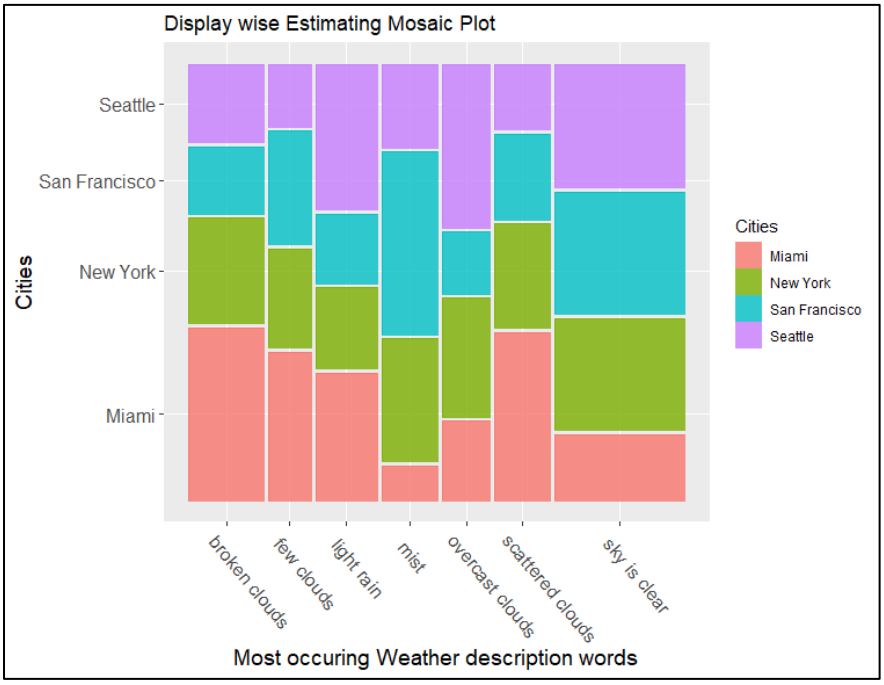


(Word Cloud of Weather Description Data)

The second graph generated is **Mosaic plot** using the same weather description data to get some categorical information out of it. The mosaic plot was trimmed and only few variables were selected in order to remove some clutter and give some clarity. The plot possesses the four cities on y-axis and the type of the description used for that cities weather on x-axis. Note the weather description selected for the cities is also trimmed and the most occurring words out of all the minor description.

The mosaic plot gives a lot of good information. When we look at mosaic plot the thickness of the bar suggest the frequency of the described word overall in all the cities. Whereas the length

of each separate color in each bar suggest the frequently usage of that particular word in that city. For example, orange color is of Miami city, and the broken clouds are more frequently used in Miami as compared to the blue color of San Francisco city in the first bar with label broken clouds. Note the color legend displayed below suggest the color dedicated to each city.

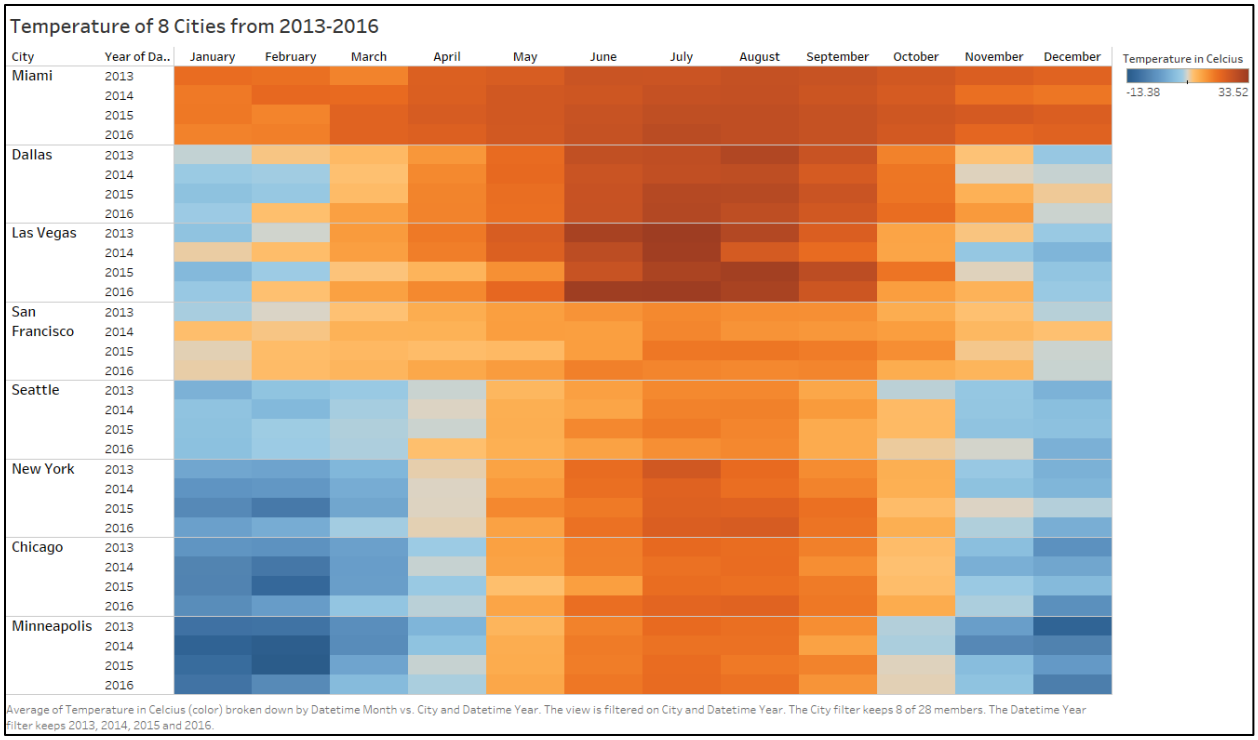


(Mosaic plot of Weather Description)

Visualization – 2: Graph suggesting the temperature patterns in US.

The second type of visualization generated is Heat map suggesting the temperature pattern of 8 major cities of united states from 2013 to 2016, also divided monthly. For this graph the dataset used is temperature one of the 7 files in main dataset. Note the original temperature was present in the Kelvin scale we transformed it to the Celsius format for easy interpretation. The cities were pivoted and then graph is generated. In the graph the x axis is the months from January to December, on other hand on y axis is having two separate axis the first is cities and each city is having years from 2013 to 2016 separately. To add to it the legend suggests the temperature range in Celsius and the center is placed at the comfortable temperature of 15 degrees centigrade. The blue color suggests cold temperature and the red color suggest the hotter days or high temperature.

The clear pattern can be inferred from the below displayed heat map. During the summers the temperature across all the cities is high but a clear pattern is immerged that can be seen in cities Miami, Dallas, Las Vegas it the most reddish color from June to September says highest temperatures during summers in this cities as compared to other cities. To add to it Miami has the very nonvolatile temperature because from January to December it has very stable weather because there is drastic color change yearly when compared to Chicago it has very low temperature during winter but in summer it is red so hot days.



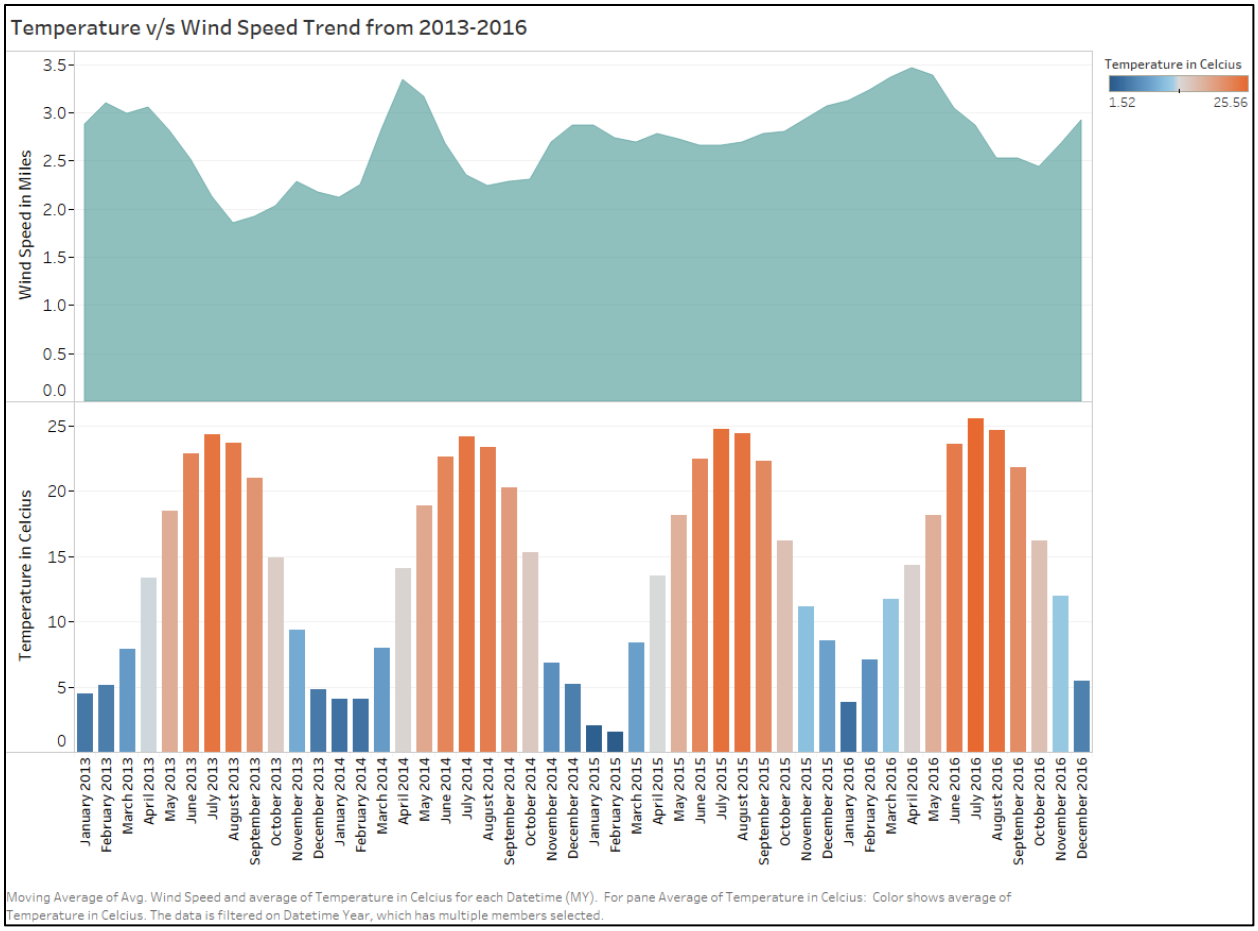
(Heat Map of US cities for temperature pattern)

Visualization – 3:Graps built for relation between Temperature and Wind speed.

The third visualization is generated to capture the relationship of the two aspects of the weather temperature and the wind speed. So, to get this type of relationship first the temperature dataset the pivoting is applied to it to merge all the city columns to one single column just of cities and other is of the temperatures of that city. Same with the wind speed pivoting has been applied to this dataset too. After both the dataset has been joined in the tableau via the Date column because all the dataset consists of hourly data from 2012 to 2017.

To generate the visualization and to reduce the clutter some filters were applied to the both the datasets. The first was the to generate the just specific city visualization all the remaining cities were excluded and other filer was universal to this to any kind of this plot was the Year 2012 and 2017 were removed because the two years just has few months data was only captured. So, in order to get whole yearly pattern this filtration was applied.

The first graph is the aggregation(average) of wind speed and the temperature of all the cities. The wind speed is in Miles and the temperature is in Celsius. The top area chart is of wind speed and the bottom bar chart is of temperature. The x-axis is month/year type pf axis of every month from 2013 to 2016. The divergent color legend is displayed in the right side of the graph which shows the average temperature range of all the cities and the center is taken as the center value o 10. The clear relationship is inferred from the graph which is negative correlation between these two aspects which means, when temperature increase the wind speed decreases and the vice versa. Only the year 2015 is kind of not following this pattern but other then that all other years follow this pattern.



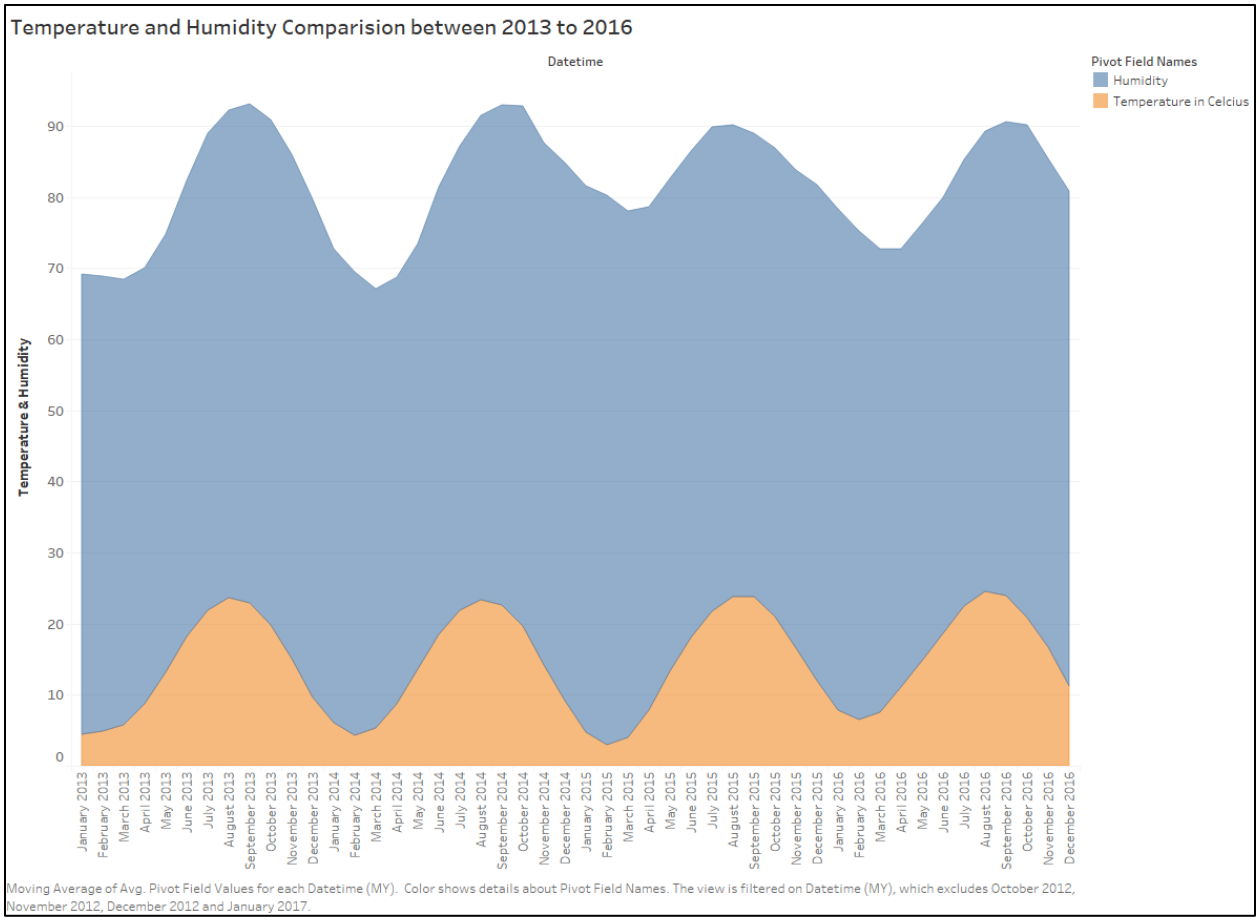
(Wind Speed v/s Temperature Comparison of All cities aggregated)

The other graphs are also generated with the separate city data only to compare the relationship between two and get how this aspect varies with time yearly. The separate graph generated with the similar pattern as above of temperature and wind speed are of cities Chicago, New York, and Los Angeles. Note this graph also inferred the same information that this two has negative correlation between them. These graphs were displayed in the other graphs section of the report below.

Visualization – 4: Graphs built for relation between Humidity and Temperature.

The fourth visualization shows the relation between Humidity and weather temperature. We posses the date and time on x axis and temperature and humidity on y axis. Blue color in graph shows the Humidity, where orange color shows the Temperature in Celsius. We conducted date and time from January 2013 to December 2016.

In graph, there is lot of fluctuations of humidity. Because when the temperature rises, humidity also rises and when temperature decreases, humidity also decreases. We can see from the visualization that in January 2013 temperature to December 2013, when temperature rises along with that humidity rises and when temperature goes down, humidity also goes down.


Temperature and Humidity Comparision between 2013 to 2016
Datetime
Pivot Field Names
Humidity
Temperature in Celcius
Temperature & Humidity
0
10
20
30
40
50
60
70
80
90
January 2013
February 2013
March 2013
April 2013
May 2013
June 2013
July 2013
August 2013
September 2013
October 2013
November 2013
December 2013
January 2014
February 2014
March 2014
April 2014
May 2014
June 2014
July 2014
August 2014
September 2014
October 2014
November 2014
December 2014
January 2015
February 2015
March 2015
April 2015
May 2015
June 2015
July 2015
August 2015
September 2015
October 2015
November 2015
December 2015
January 2016
February 2016
March 2016
April 2016
May 2016
June 2016
July 2016
August 2016
September 2016
October 2016
November 2016
December 2016
Moving Average of Avg. Pivot Field Values for each Datetime (MY). Color shows details about Pivot Field Names. The view is filtered on Datetime (MY), which excludes October 2012, November 2012, December 2012 and January 2017.

ANALYSIS AND DISCUSSION:

We have a dataset showing temperature, humidity, and wind of different cities of United States. The dataset we used is from 2013 to 2016. We found many things that could be done with visuals for the dataset. For an example, we created a cloud plot which shows the weather information. We also build a heatmap, mosaic plot with the data. From the cloud plot we got to know that most of the time sky is clear as "sky is clear" is the biggest word in that plot. Bigger word shows the frequency of weather. From the mosaic plot we got to know about the categorical information about the data.

Heatmap gives the proper information of different cities temperature over the year between 2013 to 2016. We compared each cities with their temperature showing the colder (blue color) and hottest (red color) in the heatmap. We also compared wind speed and the temperature using our dataset. From that visualization we got to know that when temperature rises, wind speed decreases. And when temperature decreases, wind speed increases. We got to know the relation between temperature and weather. We found that when temperature increases humidity also increases. When temperature decreases, humidity also decreases. Temperature and humidity has a relation with each other.

APPENDIX: INDIVIDUAL REPORTS:

KASHYAP DOBARIYA:

For our project we took the dataset from the Kaggle platform which consists of various files consisting of different aspects of weather stored. Our main focus was to explore the weather patterns of across the Whole united states. Also this data was time series and also possesses the location coordinates so we thought of building some Geographical graphs but due to the limitation of the cities in the united stated we explored different aspects ad built some visualization which are displayed below.

Talking about myself in terms of contribution. First of all, I cleaned the data the data was not that bad just a few missing values were to be handled. And being the time series data to fill the missing values I have used the forward fill to do so. After, at the time of the exploratory data analysis phase of the project I helped group in terms of building few useful visualizations to get some insights of the data.
In the main visualization building phase of the project I have built the visualization -3 the comparison of the temperature versus the temperature to dig and get some interesting patterns out of these two aspects of the data. And to do so I have gone through few iterations to refine the graphs in terms of the message. Also, after the presentations and getting the feedback from the professor and peer members of class, taking that feedback into considerations I have changed accordingly to improve readability of the visualization.

Other then that I helped other group members in generating the visualization when they are having the trouble. Like I have group member to remove the clutter from the mosaic plot.

In terms of Learning for the course, I have learnt a lot specifically in generating powerful visualizations. The most important part in learning process was the building the visualizations in R using the library GGPLOT2. It has helped a lot in other subjects too. To add to it, in terms of tools I have learn tableau as well. So that is also a tool that I can add to my resume. Other then that the important gain was to get familiar with the visualization design paradigms. Data, audience, message, what type of visualizations are effective to the human perception what color should be more effective etc. Overall taking this course will helm me a lot in future in my job as well.

PRAMATHESH SHUKLA:

In the group, my role was to help the group in any way I could. If the group needed visuals for the exploratory or the final explanatory part, I would create one and see if it were helpful to the project.

I was responsible for creating a Heatmap, which shows the US cities temperature between 2013 and 2016. It shows the cities like San Francisco, Dallas, Seattle, Las Vegas, New York, Chicago, Miami, Minneapolis, Los Angeles. I divided it into the month for better understanding. For easy interpretation I divided Kelvin into the Celsius.

In the graph the x axis is the months from January to December, on other hand on y axis is having two separate axis the first is cities and each city is having years from 2013 to 2016 separately. In the graph red color suggest the hotter days and blue color suggest the colder days. During the summers the temperature across all the cities is high but a clear pattern is immerged that can be seen in cities Miami, Dallas, Las Vegas it the most reddish color from June to September says highest temperatures during summers in this cities as compared to other cities.

I learned a lot about data visualization in the final project. I learned how to use Tableau and R from watching and reading the tutorials made by professor. From Tableau, I learned how to manipulate data, how to join two or more datasets and use them to create visuals. I also learned how to do the table calculations. In R, I was able to use ggplot2 extensively to create many different visuals covered in the course.

SHIBANI MARAN:

With respect to our historical hourly weather dataset over the years 2012-2017, my contribution towards the group work is to create some plots which helps to analyze the relationship between temperature and humidity. After making few combinations with different graphs, I decided to clean the dataset by neglecting the years which has null values. Consequently, I used the area plots to represent the comparisons between temperature and humidity over the years 2013-2016.

In the area chart I created, to understand the relationship between temperature and humidity, I used the different cities of USA over the years 2013-2016. I calculated the temperature field from kelvin to Celsius scale. The relationship between them is that temperature therefore directly relates to the amount of moisture the atmosphere can hold. The color scale we used here is to differentiate the temperature and humidity. The key takeaway is that when temperature increases, the humidity also increases, when temperature decreases the humidity also decreases. In addition to those visualization, I chose a couple of US cities namely

Los Angeles and Minneapolis to verify the positive correlation between them. I have plotted area graph likely to previous graph. I used the variables like temperature and humidity of these two cities. These two charts show the dependency of both variables. We can observe that this graph clearly proves the correlation.

By the end of this course, I was able to create different plots and make some interpretation based on the outcome. Also, provided some insightful information on interpreting data, message and understanding who our target audience are. Having had the opportunity to work on both Tableau and R helped me to visualize the data more efficiently. Working as a team on the group project, had a great hands-on experience in collaborating and understanding our dataset.

MADHUSUDAN GOWDA:

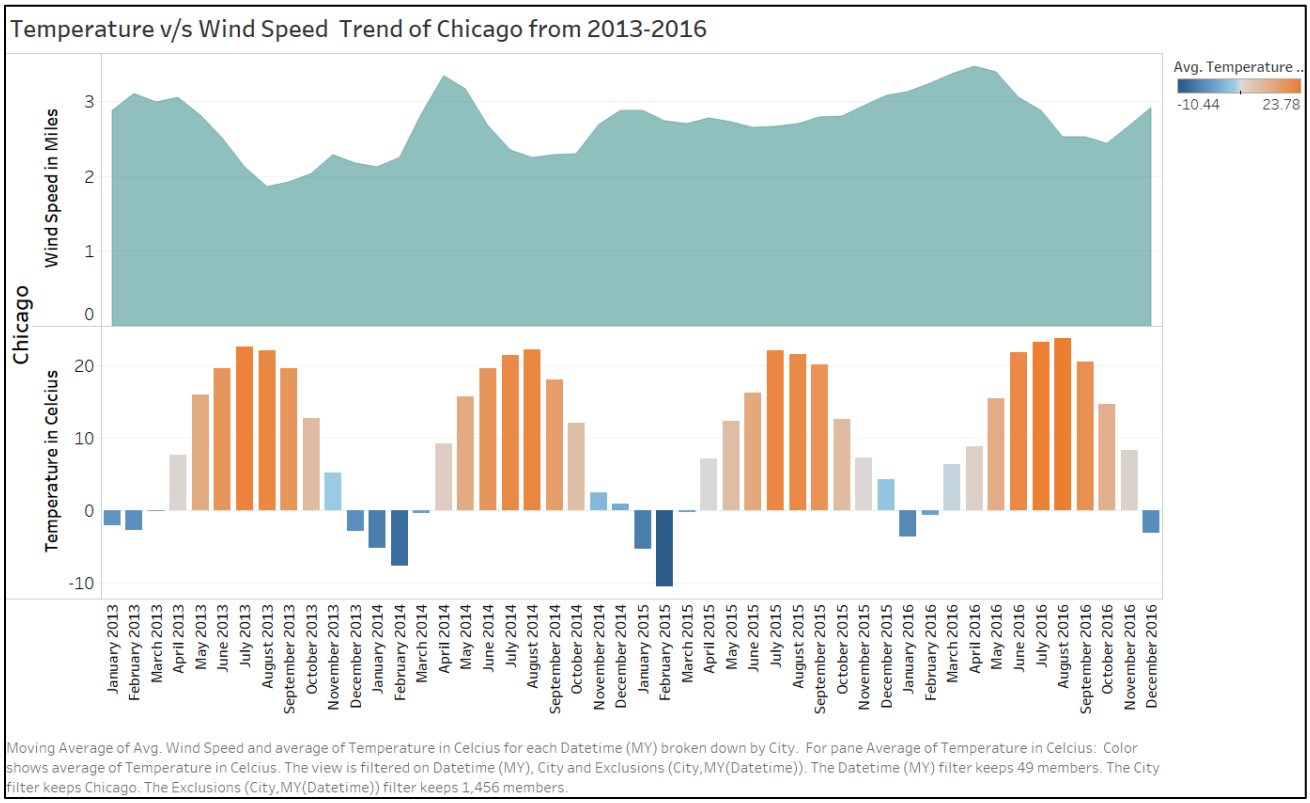
In the group, I made some plots based on the team's requirement such as the world cloud map and mosaic plot to display the highest frequency of the word which indicates the most common type of weather in the entire data.

The highest frequency shows the biggest word and the lower the size suggests less frequency. For e.g. 'Sky is Clear' is the biggest word that is seen in the plot which indicates that it is the common type of weather throughout the year and words like 'tornado' are lower in size which suggests that it only happens a few times in a year.

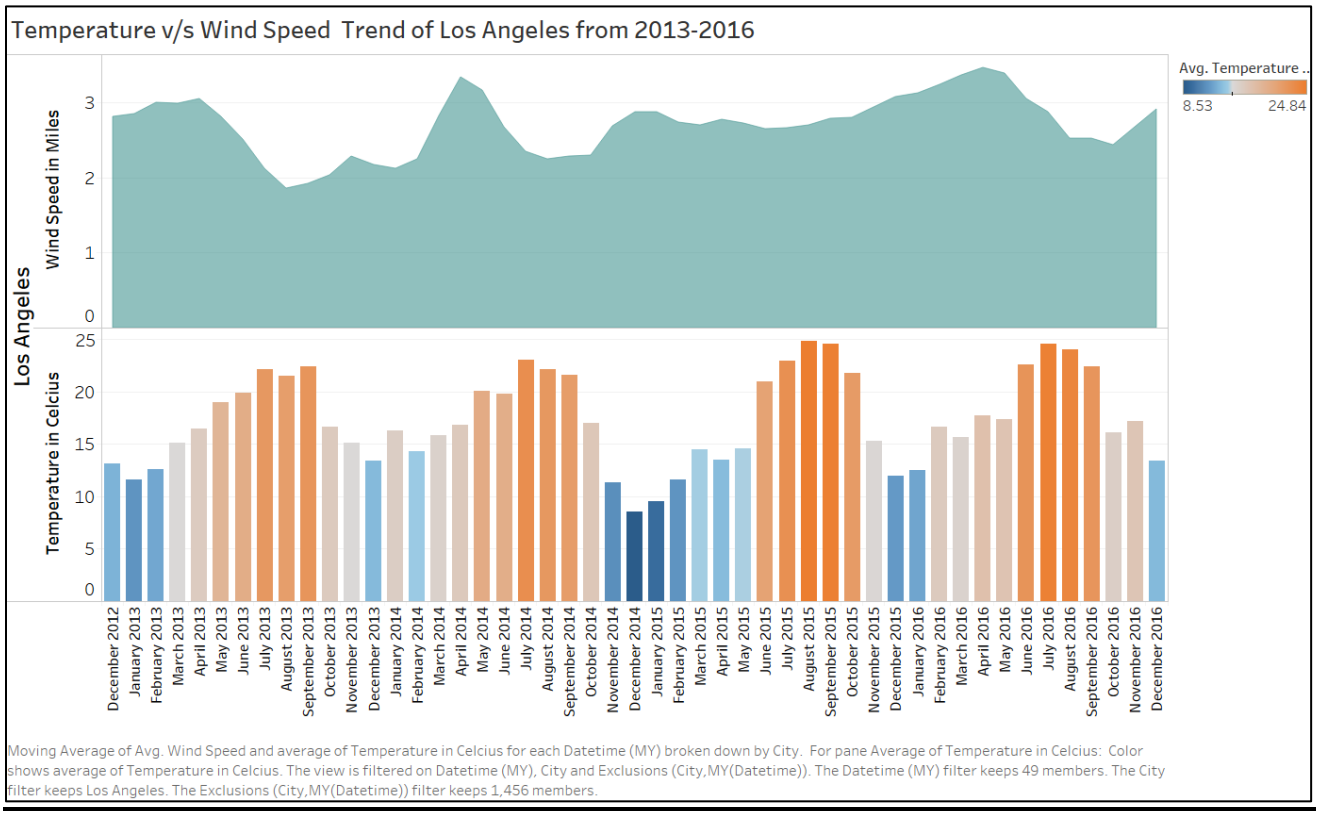
In the Mosaic plot, we used different cities on the Y-axis and weather description on the x-axis. The width of the weather description suggests more frequent use of that word. As we can see from the plot 'Sky is Clear' is the weather description that is seen and is the word used in common for a long time in all the cities. Snow and Drizzle are the weather descriptive words that are used very less and only seen in Minneapolis and New York. I also tried several ways to display to display the data, however few graphs got more complex for this time series of data.

In the overall course and with the final project, I had a great opportunity to work with the best people and to learn to see the data freely, understand it, and to use it in the appropriate chart to make sure that the message reaches the audience without a lot of confusions. The whole perspective of seeing the data has been changed and gained a lot of knowledge in using Tableau and R to build graphs. Even though it was my first time using both the R and Tableau it was a great experience and the tutorial videos helped a lot to explore so many things on tableau. Finally, this class thought me how to use the data more appropriately and create different kinds of plots to analyze the data. It will be a great add on to my skills.

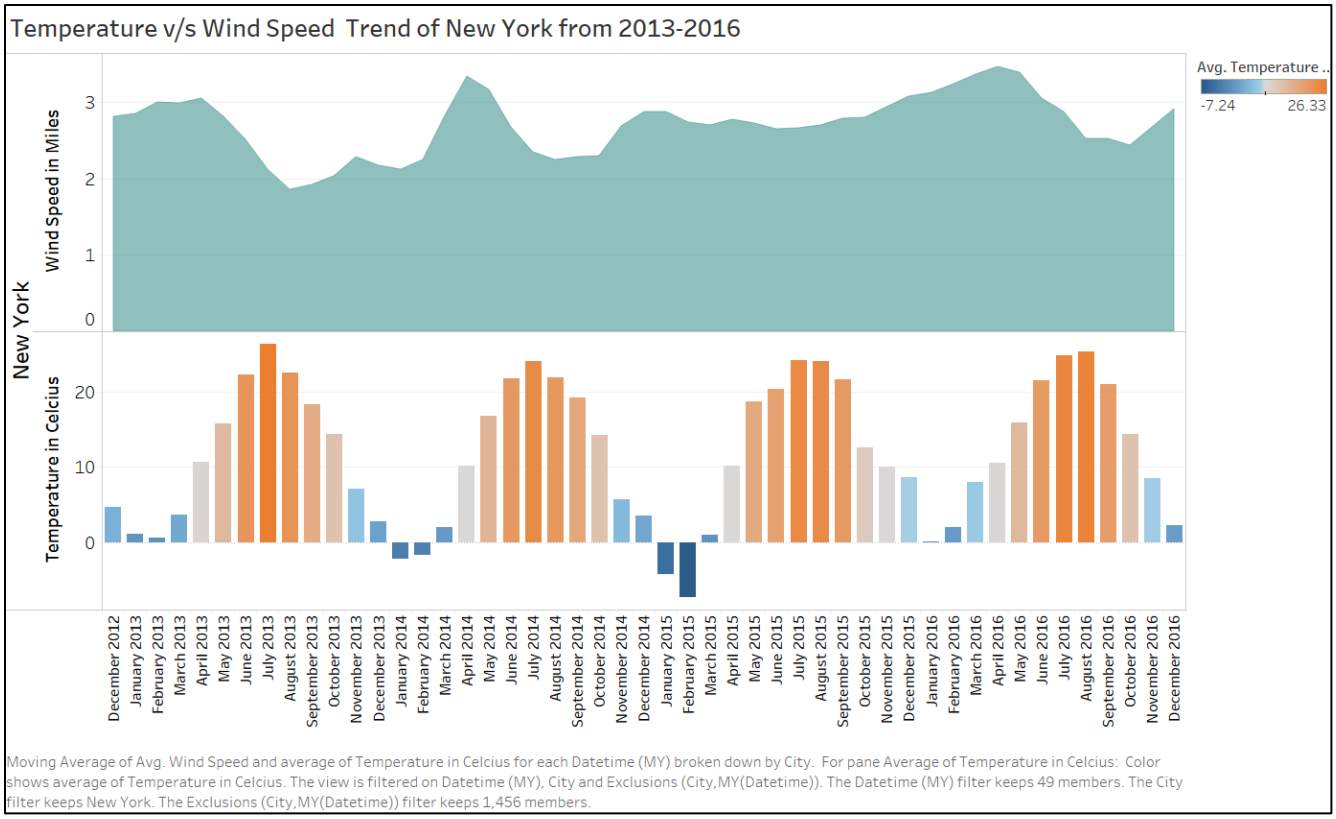
PLOTS AND GRAPHS:



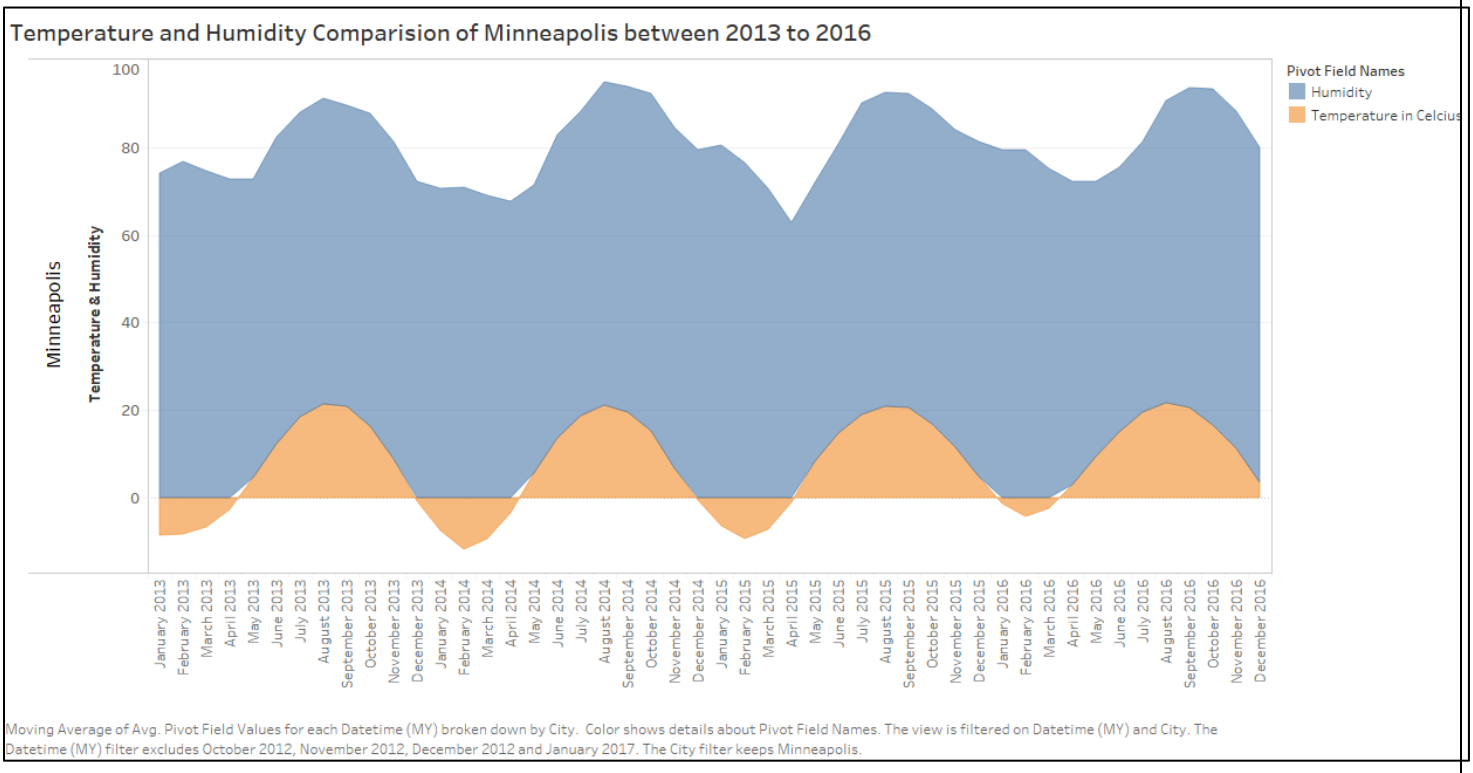
(Wind Speed v/s Temperature Comparison Chicago)



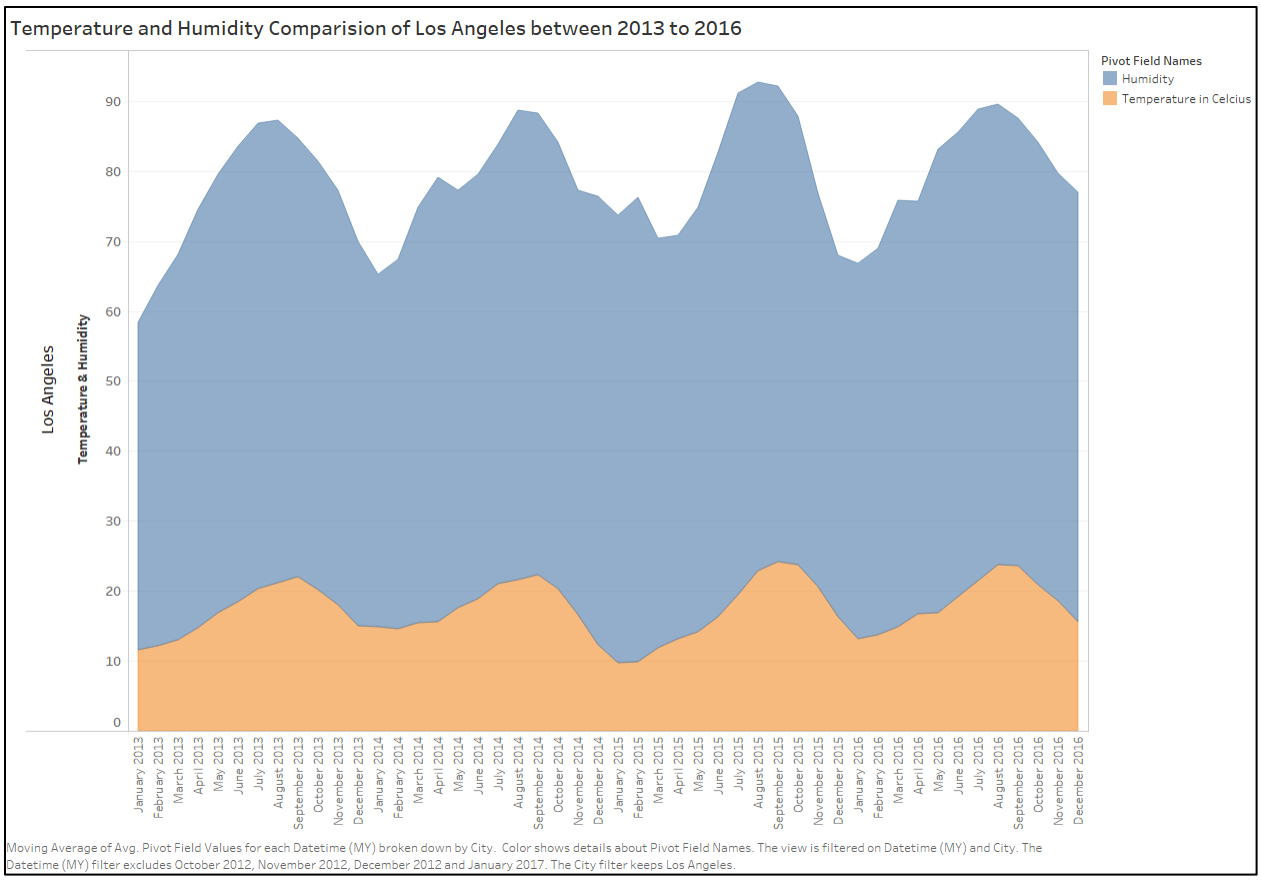
(Wind Speed v/s Temperature Comparison Los Angeles)



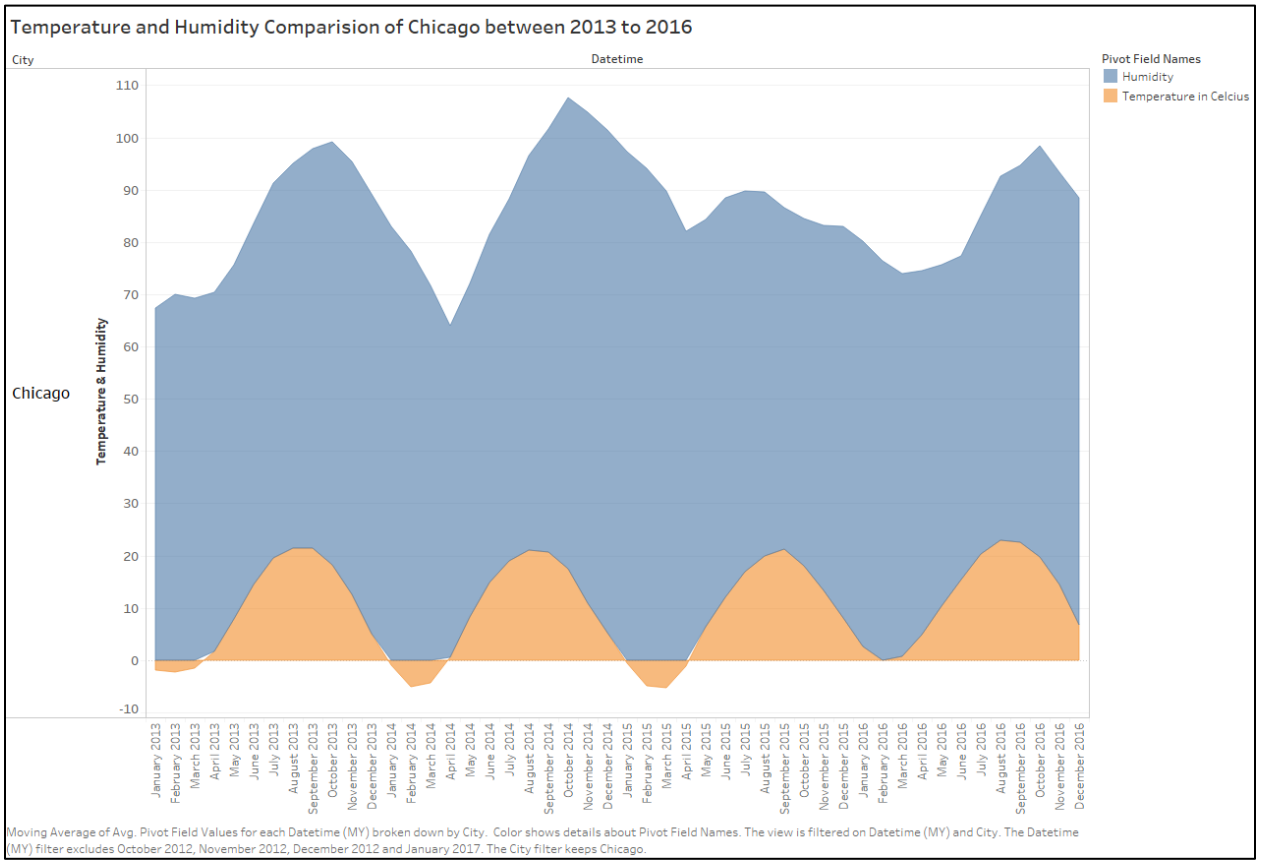
(Wind Speed v/s Temperature Comparison New York)



(Humidity v/s Temperature Comparison Minneapolis)



(Humidity v/s Temperature Comparison Los Angeles)



(Humidity v/s Temperature Comparison Chicago)

R Code:

```
#################################################################
wd <- read.csv("wind_direction.csv")
head(wd)

which(is.na(wd))
lapply(wd,function(x) { length(which(is.na(x)))})

wd2<- pivot_longer(wd, cols = 3:4,names_to = "Type" )
head(wd2)

write.csv(wd2,"F:/DSC-465 Data Visualization/pROJECT/Updated Data/wd2.csv")

###################
wdes <- read.csv("weather_description.csv")
write.csv(wed2,"F:/DSC-465 Data Visualization/pROJECT/Updated Data/wed2.csv")

wed2<-wed2[!(wed2$Pivot.Field.Values=="heavy snow"),]
wed2<-wed2[!(wed2$Pivot.Field.Values=="light snow"),]
wed2<-wed2[!(wed2$Pivot.Field.Values=="drizzle"),]
wed2<-wed2[!(wed2$Pivot.Field.Names=="Minneapoli"),]
wed2<-wed2[!(wed2$Pivot.Field.Values=="proximity thunderstorm"),]
wed2<-wed2[!(wed2$Pivot.Field.Values=="moderate rain"),]


p <- ggplot(data=wed2) +
  geom_mosaic(aes(x=product(Pivot.Field.Values),fill=Pivot.Field.Names,na.rm=TRUE))+
  theme(axis.text.x=element_text(angle=-50, hjust= .1)) +
  labs(x = "Most occuring Weather description words", y= "Cities", title = "Display wise Estimating Mosaic Plot") +
  scale_fill_discrete(name="Cities")

p +theme(axis.text=element_text(size=12),
         axis.title=element_text(size=14))
```

(R code for mosaic plot)