MI 562 Final Project – Weather Prediction

Brian Mallari, Mihir Shah, Jeffrey Wong

As a person who prepares for Rutgers Commencement, you want to know the weather on that day (May 13th) – Rainy or Sunny?

Data collection

a. US National Oceanic and Atmospheric Administration (NOAA)

The US National Oceanic and Atmospheric Administration (NOAA) is an immense and helpful resource for our project. NOAA hosts global historical climate data and provides a user-friendly GUI interface as well as ftp access to access this data. You can access the dataset through here (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/).

b. Climate.gov also provides weather data and relevant information.

c. OpenWeatherMap (http://openweathermap.org) provides historic and current weather data through APIs (http://openweathermap.org/api)

Procedure

- Baseline model (e.g. weighted average climate of past years)
- Extended models with more variables
- Model comparison with other previous days such as 03/14/2016 and 03/14/2017.

Reference

http://www3.cs.stonybrook.edu/~skiena/591/final_projects/white_christmas/

Grading: The project will be worth 100 points and will be graded according to the following rubric.

- **Comprehensiveness and correctness of data processing and computation (e.g., did you try all reasonable correlations? did you look explore the data enough to make informed decisions for processing?): 25**
- **Proper use of tools (e.g., Python, R, databases) and techniques (e.g., clustering, regression): 25**
- **Proper error-checking in the code: 10**
- **Internal documentation (comments in your R and Python code): 10**
- **External documentation (project report): 30**

Phase 1: Pre-planning Phase and gathering of data

We are looking to predict whether the weather will be rainy or sunny on May 13th, 2018 or better known as, commencement day for Rutgers University.

First, our team has to extract weather data in which we will collect from the following :

- US National Oceanic and Atmospheric Administration (NOAA)
- Climate.gov
- Open Weather Map

The next phase will involve creating scripts to wrangle data and showcase only the data we need and can utilize for our evidence and prediction. Analysis will follow up and include that of correlation and regression. For our data, we originally chose to use data from New Brunswick, NJ and Piscataway, NJ. However, the data files consisted of missing values and measurements. Therefore, we found another, more complete dataset with the area of Somerset Airport. As this is in close vicinity of New Brunswick and Piscataway NJ, we were able to utilize this dataset in our weather prediction. These three selected areas are in close vicinity of each other so tests such as correlation, regression, and clustering will be performed.

Phase 2: Scripts for data wrangling + screenshots of data being utilized for New Brunswick, Piscataway, and Somerset Airport , New Jersey.

Script 1 **(562 - final project - weather prediction (Piscataway, New Brunswick, and Somerset Airport).R)** :
Get views of two tables, each one showcasing precipitation values for "May 13" across different years for each location.

- Note: PRCP = Precipitation ; Values = (tenths of mm)

	ID	YEAR	MONTH	ELEMENT	VALUE13
16	US1NJMD0028	2011	5	PRCP	0
40	US1NJMD0028	2012	5	PRCP	0
64	US1NJMD0028	2013	5	PRCP	0
89	US1NJMD0028	2014	5	PRCP	3
113	US1NJMD0028	2015	5	PRCP	0
138	US1NJMD0028	2016	5	PRCP	0
162	US1NJMD0028	2017	5	PRCP	84

New Brunswick, NJ

- As shown in the table above for New Brunswick, NJ, the precipitation value on May 13 in 2011, 2012, 2013, 2015, and 2016 are 0. Meanwhile, the precipitation value for the same date in 2014 is 3 and the precipitation value for the same date in 2017 is 84.

	ID	YEAR	MONTH	ELEMENT	VALUE13
18	US1NJMD0024	2010	5	PRCP	5
43	US1NJMD0024	2011	5	PRCP	0
67	US1NJMD0024	2012	5	PRCP	0
89	US1NJMD0024	2013	5	PRCP	0
114	US1NJMD0024	2014	5	PRCP	5

Piscataway, NJ

- As shown in the table above for Piscataway, NJ, the precipitation value on May 13 in 2011, 2012, and 2013 are 0. Meanwhile, the precipitation value for the same date in 2010 and 2014 is 5.

Also, get views of three tables, one each for Piscataway, New Brunswick, and Somerset Airport. The years for these three tables will be based off the years the two above tables have in common.

May 13 values

	ID	YEAR	MONTH	ELEMENT	VALUE13
43	US1NJMD0024	2011	5	PRCP	0
67	US1NJMD0024	2012	5	PRCP	0
89	US1NJMD0024	2013	5	PRCP	0
114	US1NJMD0024	2014	5	PRCP	5

Piscataway, NJ

- As shown in the table above for Piscataway, NJ, the precipitation value on May 13 in 2011, 2012, and 2013 are 0. The precipitation value on May 13, 2014 is 5.

	ID	YEAR	MONTH	ELEMENT	VALUE13
16	US1NJMD0028	2011	5	PRCP	0
40	US1NJMD0028	2012	5	PRCP	0
64	US1NJMD0028	2013	5	PRCP	0
89	US1NJMD0028	2014	5	PRCP	3

New Brunswick, NJ

- As shown in the table above for Piscataway, NJ, the precipitation value on May 13 in 2011, 2012, and 2013 are 0. The precipitation value on May 13 in 2014 is 3.

	ID	YEAR	MONTH	ELEMENT	VALUE13
1658	USW00054785	2011	5	PRCP	0
1807	USW00054785	2012	5	PRCP	0
1950	USW00054785	2013	5	PRCP	0
2077	USW00054785	2014	5	PRCP	0

Somerset Airport, NJ

- As shown in the table above for Somerset Airport, NJ, the precipitation value on May 13 in 2011, 2012, 2013, and 2014 are 0.

Script 2 **(562 - final project - weather prediction (Piscataway, New Brunswick, and Somerset Airport - randomly selected date).R)** : Random selected day in year and table pops up for each town/city with precipitation data for that day across “common years” .

Script is to show that even on a randomly selected day the precipitation values among those three locations are similar, (somerset, piscataway, new brunswick) The R script is similar to the previous one created as Script 1. However, this one uses a randomly selected day in the year and tables pop up for each town/city with precipitation data for that day across "common years". Each one showcases precipitation values for "May 13" days in coinciding years for which a measurement was taken.

Excluding 28th, 29th, 30th, 31st days :
The code involves analysis between day 1-28 for each month. This is to avoid any potential invalid entries in the data set (e.g such as trying to pull feb31st which is an invalid date.)

Random day - April 23 values

	ID	YEAR	MONTH	ELEMENT	VALUE23
87	US1NJMD0024	2013	4	PRCP	0
112	US1NJMD0024	2014	4	PRCP	0
182	US1NJMD0024	2017	4	PRCP	18

Piscataway, NJ

- As shown in the table above for Piscataway, NJ, the precipitation value on April 23, 2013 and April 23, 2014 are 0. The precipitation value for the same date in 2017 is 18.

	ID	YEAR	MONTH	ELEMENT	VALUE23
62	US1NJMD0028	2013	4	PRCP	0
87	US1NJMD0028	2014	4	PRCP	0
160	US1NJMD0028	2017	4	PRCP	20

New Brunswick, NJ

- As shown in the table above for New Brunswick, NJ, the precipitation value on April 23 in 2013 and 2014 are 0. The precipitation value for the same date in 2017 is 20.

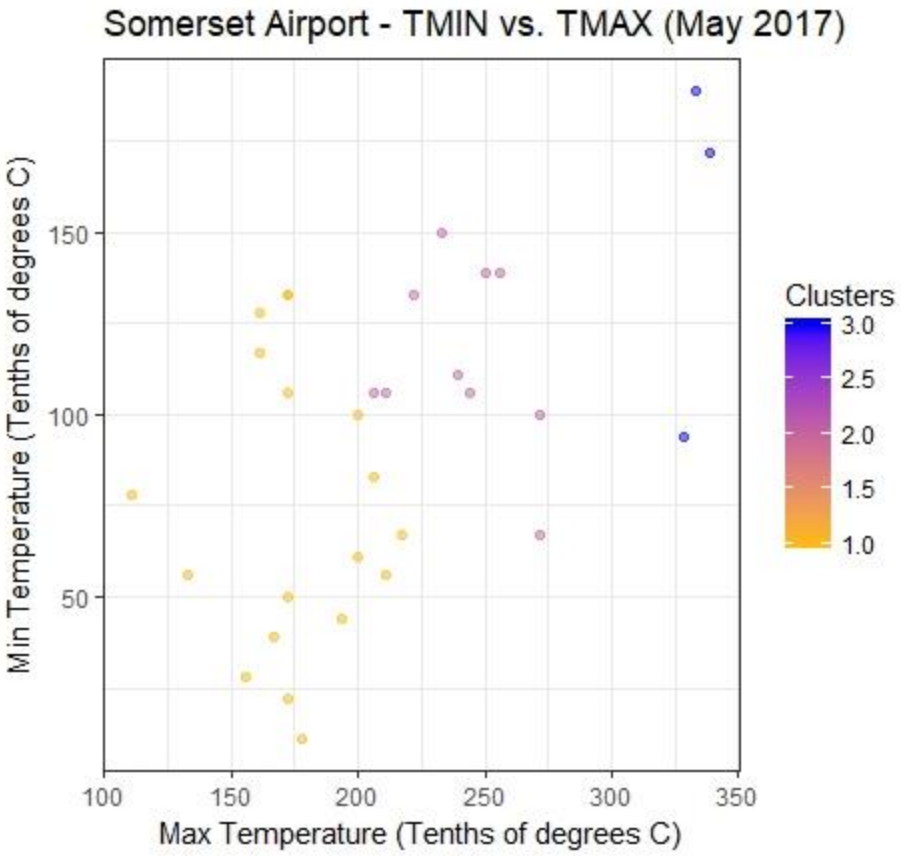
	ID	YEAR	MONTH	ELEMENT	VALUE23
1938	USW00054785	2013	4	PRCP	0
2068	USW00054785	2014	4	PRCP	0
2391	USW00054785	2017	4	PRCP	0

Somerset Airport, NJ

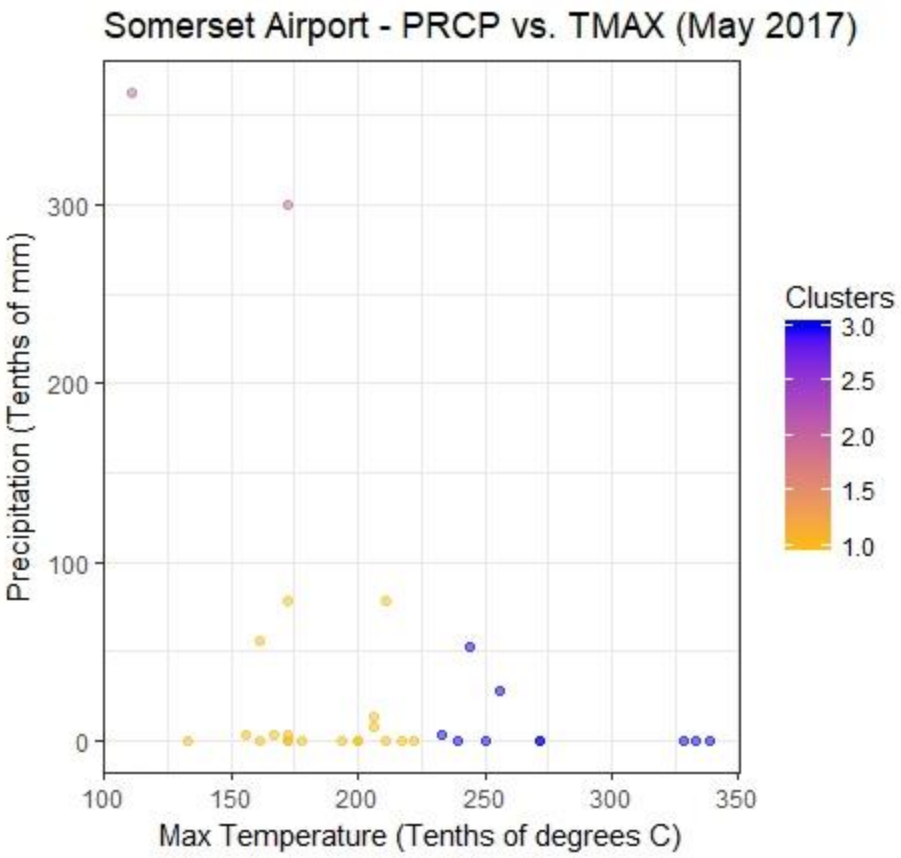
- As shown in the table above for Somerset Airport, NJ, the precipitation value on April 23 in 2013, 2014, and 2017 are 0.

	TMAX	TMIN	PRCP
TMAX	1.0000000	0.5248476	-0.3704665
TMIN	0.5248476	1.0000000	0.0110118
PRCP	-0.3704665	0.0110118	1.0000000

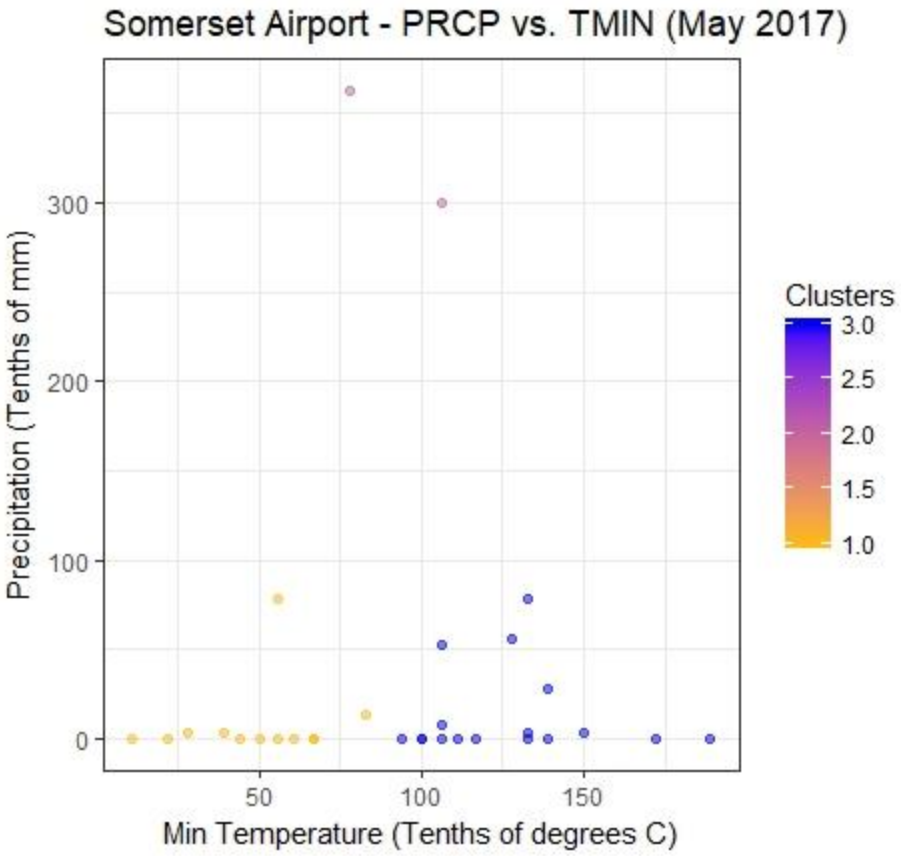
Somerset Airport, NJ - May 2017 - Correlation Matrix



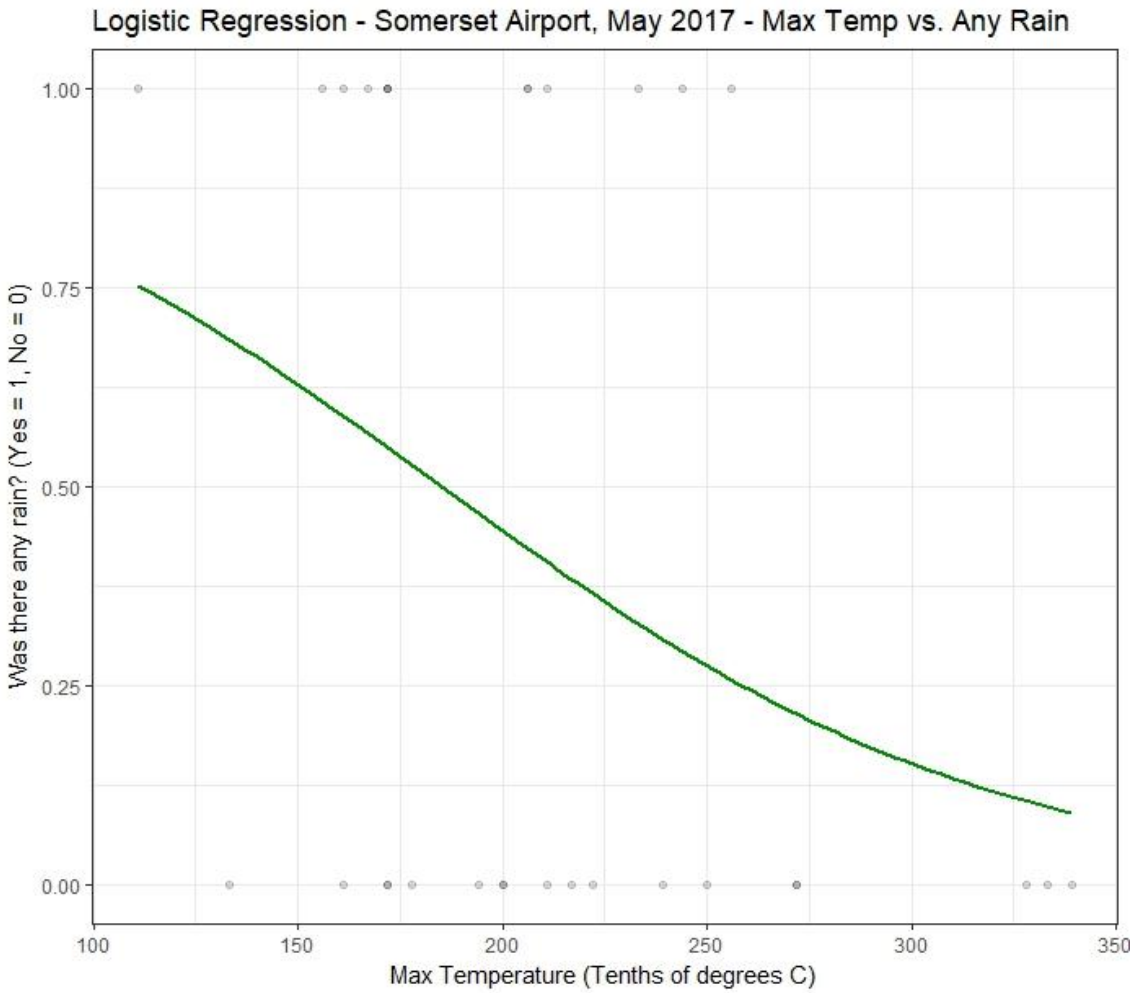
- Reference script (**562 - final project - weather prediction (Somerville Somerset AP, NJ only, May 2017, Correlation and Scatterplots with kmeans).R**)
- This plot includes 3 clusters. When n=1.0, the minimum and maximum temperatures are lower. When n=2.0, the minimum temperature is roughly between 100-150 tenths of degrees celsius and the maximum temperature is roughly between 200-300 tenths of degrees celsius. When n=3.0, the minimum temperature is mostly above 150 tenths of degrees celsius and the maximum temperature is over 300 tenths of degrees celsius.



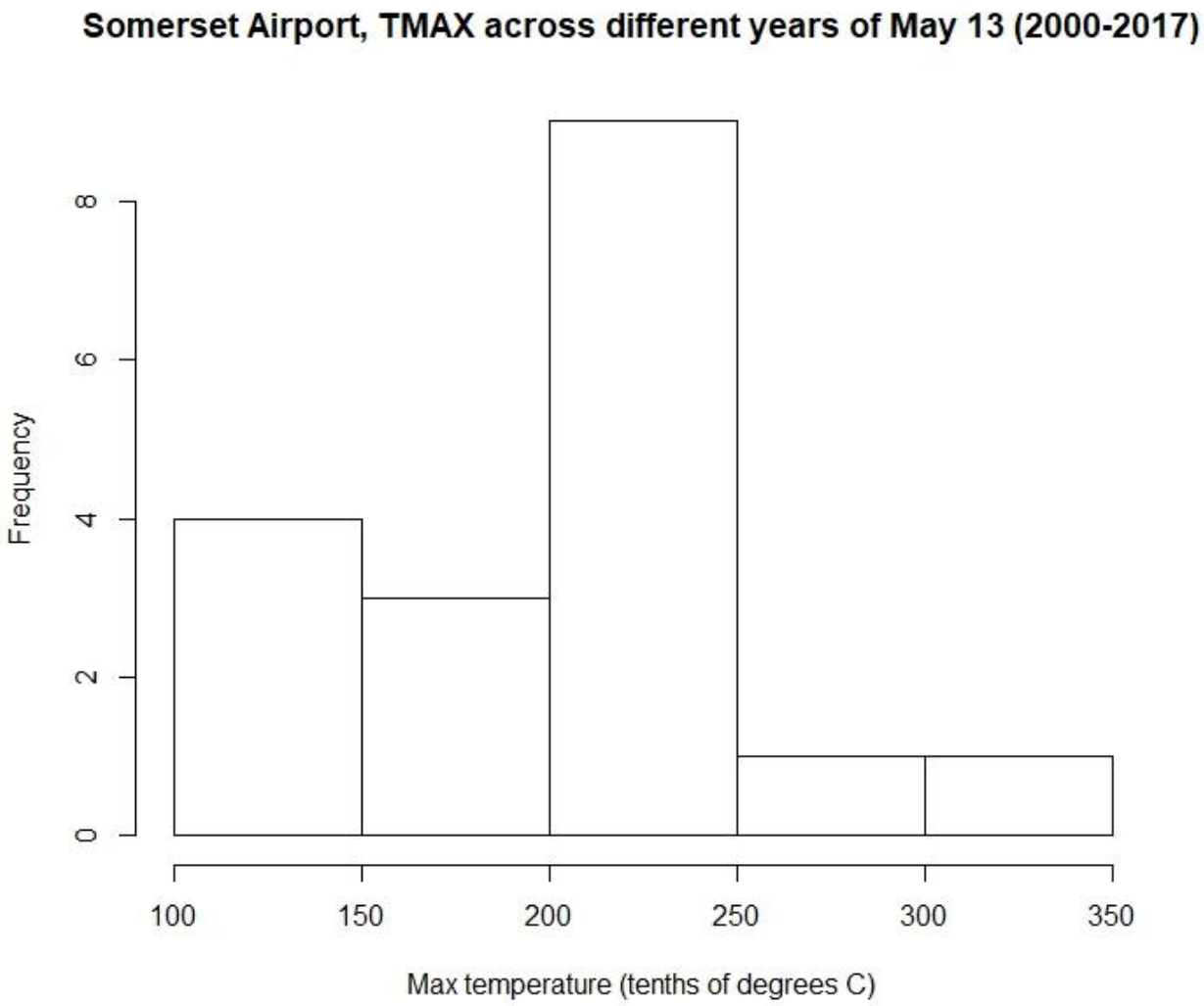
- Reference script (**562 - final project - weather prediction (Somerville Somerset AP, NJ only, May 2017, Correlation and Scatterplots with kmeans).R**)
- This plot includes 3 clusters. For the precipitation of all three clusters, majority of the data points are at 0 tenths of mm. When n=1.0, the maximum temperature is between 125-225 tenths of degrees celsius. When n=2.0, the maximum temperature is between 225-275 tenths of degrees celsius. When n=3.0, the maximum temperature is over 325 tenths of degrees celsius.



- Reference script (**562 - final project - weather prediction (Somerville Somerset AP, NJ only, May 2017, Correlation and Scatterplots with kmeans).R**)
- This plot includes 3 clusters. For the precipitation of all three clusters, majority of the data points are at between 0-100 tenths of mm. When n=1.0, the minimum temperature is less than 100 tenths of degrees celsius. When n=2.0, the minimum temperature is between 75-125 tenths of degrees celsius. When n=3.0, the minimum temperature is over 100 tenths of degrees celsius.



- Reference script (**562 - final project - weather prediction (Somerville Somerset AP, NJ only, May 2017, logistic regression).R**)
- Based on this logistic regression for Somerset Airport in May 2017, the regression line is going downwards. As the maximum temperature increases, the chances of rain in May 2017 decreases.



- Reference script (**562 - final project - weather prediction (Somerville Somerset AP, NJ only, May 13 TMAX).R**)
- As presented in the histogram above, the maximum temperature between 200-250 tenths of degrees celsius is the most frequent, which appears at 9. Meanwhile, the frequency for the max temperature between 100-150 is 4 and the frequency for max temperature between 150-200 tenths of degrees celsius is 3. Finally, the frequency is the lowest when the max temperature is much higher at between 250-350 tenths of degrees celsius.

Phase 3 : Analysis Explanation

The weather file for Piscataway only included precipitation and snowfall values. Moreover, there were instances where no measurement was taken (coded as -9999). The weather data for New Brunswick was then evaluated as a substitute since New Brunswick is just across the Raritan River from Piscataway. However, the New Brunswick data had the same problems - only precipitation and snowfall. Moreover, some measurements were not taken as well for some of the days. The missing values required for additional research of data.

We came across weather data for Somerset Airport which was then evaluated due to its proximity to both Piscataway and New Brunswick. This data file file included more data, such maximum temperature, minimum temperature, and more variables for measurement. As seen from the comparison of precipitation values for common years between Piscataway, New Brunswick, and Somerset Airport, the data for Somerset Airport can serve as a viable proxy for the other two locations for the sake of analysis.

Our focus for data was placed on all of May 2017. The assumption was that annual climate patterns would be relatively uniform throughout the month of May, thus providing allowing all 31 days of data to be a useable for analysis. Moreover, 2017 was was most likely to reflect the most recent effects of climate change. With respect to data, focus was placed on maximum temperature, minimum temperature, and precipitation since each of these values was most likely to be recorded across the different years, thus providing a common element between the months and years should analysis need to be extended to include data recorded further in the past. Moreover, snowfall was not expected to occur in the month of May.

From the correlation matrix, it can be seen that there is a moderate, positive linear relationship between the maximum temperatures and the minimum temperatures. There is a weak, negative linear relationship between maximum temperatures and total precipitation. Also, there is a very weak, positive linear relationship between minimum temperatures and precipitation values. The bivariate scatterplots confirm these relationships.

Looking at the histogram for maximum temperatures recorded at Somerset Airport in May 2017, it can be seen that the range with the highest frequency of values is the 200-250 range. This range of max temperatures includes both days with precipitation and days without precipitation as seen from the logistic regression plot. However, looking at the scatterplot for precipitation and maximum temperature, the amounts of precipitation within this temperature range all fall below 100 tenths of millimeters, which is also depicted by the k-means cluster analysis with two of the three clusters towards the bottom of the scatterplot accounting for most of the precipitation values.

Phase 4 : Conclusion

Based on the results obtained from the tables, graphs, and plots, we predicted that the weather will be sunny on May 13, the day of the Rutgers University commencement where all graduates will gather with families, friends, and other fellow graduates for the most exciting moment one last time on campus at Piscataway, NJ. Moreover, looking at our analysis, our prediction is that it will not rain on May 13. Majority of our weather data is based off of May 2017, due to the assumption that annual climate change would be relatively uniform throughout this timeframe, thus allowing us to explore the data for all 31 days in May. There will also be no snowfall in May.

Phase 5: References

1. Information on importing a .dly file into R was taken from the following websites:
    a. https://stackoverflow.com/questions/40874719/r-how-to-increase-efficiency-of-my-import-from-100k-flat-files-dly
    b. https://www.researchgate.net/post/How_can_I_read_GHCN_precipitation_daily_files_format_dly_with_MATLAB_or_R
2. Weather data taken from version 3.24-upd-2018042004 (i.e, an update that started at 2018042004 [yyyymmddhh] UTC; yyyy=year; mm=month; dd=day; hh=hour) of GHCN Daily
3. Information on changing the name of one column in a date frame in R was taken from the following webpage: https://stat.ethz.ch/pipermail/r-help//2012-May/312920.html
4. Information on selecting columns in an R data frame using just a string was taken from the following webpage: https://stackoverflow.com/questions/25923392/select-columns-based-on-string-match-dplyrselect
5. Information on adding columns to a pre-existing data frame in R was taken from the following webpage: https://stackoverflow.com/questions/25357194/how-to-add-multiple-columns-to-a-data-frame-in-one-go
6. Information on taking a subset of a data frame in R was taken from the following webpage: https://www.statmethods.net/management/subset.html
7. Information on airports near Piscataway, NJ, was taken from the following webpage: http://www.allplaces.us/afz.cgi?s=08846&rad=30
8. Information on the warning message, “longer object length is not a multiple of shorter object length,” was taken from the following webpage: https://stackoverflow.com/questions/12040197/longer-object-length-is-not-a-multiple-of-shorter-object-length
9. Information on how to select rows from a data frame based on values in a vector was taken from the following webpage:

https://stackoverflow.com/questions/12040197/longer-object-length-is-not-a-multiple-of-shorter-object-length

10. Information on how to find common elements in multiple vectors was taken from the following webpage: https://stackoverflow.com/questions/3695677/how-to-find-common-elements-from-multiple-vectors
11. Information on randomly selecting an element from a vector in R was taken from the following webpages:
    a. https://stackoverflow.com/questions/9390965/select-random-element-in-a-list-of-r
    b. https://stat.ethz.ch/R-manual/R-devel/library/base/html/sample.html
12. Information on using the set.seed() function in R was taken from the following webpage: https://stackoverflow.com/questions/13605271/reasons-for-using-the-set-seed-function
13. Code for clustering was borrowed from Brian Mallari's homework assignment pertaining to Yelp data and RStudio for the Week 13 of 562 - Problem Solving with Data.
14. Information on doing logistic regression was taken from the sample code provided by Dr. Michael Lesk for his class, 561 – Data Analytics.
15. Code for logistic regression was borrowed from Brian Mallari's post for the discussion thread on logistic regression for 561 – Data Analytics.