Project Three – Executive Summary

Group Members: Bianca Sosnovski, Elina Azrilyan, Robert Mercier, Asher Dvir-Djerassi and Charls Joseph.

Project Dataset: Results from the 2016 Presidential Primaries by county.

Files: 1) **"Primary_Results2016"**: Results of the Democratic and Republican primaries, by US County and each presidential candidate; 2) **"County_Facts"**: The demographic breakdown of the US Counties that voted in the primaries; and 3) **"Headers":** Labels of the columns in the county_facts spreadsheet.

Data Source: From the data science community at Kaggle. URL: https://kaggle.com/benhamner/2016-us-election.

Project Description: Check for the differences between the counties that voted for the top presidential candidates

Project Summary:

Primary Question: "Which are the most valued data science skills?"
Secondary Question: "What are the key elements that explain primary election results of a candidate by state?"

Team Tools: R – for Analysis; Tableau – for some visualizations; Skype – for group meetings; Doodle – for synchronizing meetup times; GitHub for coding and making changes; plain old email.

Loading process: The two datasets we worked with were downloaded as CSV files. We read the individual files as data frame in R using the function "read.csv".

Transforming process: Since the original datasets were separated, the files had to be merged. The most efficient way of merging them was using the unique identifier, FIPS code, in each of the two main files.

Key Data Challenges:

1) While we wanted to view the results for all 50 states, one state did not have primary results. Minnesota has counties profile results but not primary results, therefore it was not included.
2) However, the most significant challenge was getting matches for all the counties in the Primary Results and the County facts files, by no fault of our own. The FIPS code was the unique identifier needed. However, since the FIPS code is classified strictly on an individual county basis, some of the polling data did not match. While the county demographics file based on individual county info had the correct 4-digit and 5-digit codes (4-digits codes did not have a 'leading' zero), 11 states broke out the primary results into city, town or district results. These results manifested themselves into 8-digit codes that where not associated with FIPS. This was also the reason for the fourth dataset the ANSI county codes for subdivisions. This said as an example, Illinois has 102 counties, however Cook county is divided into Chicago and the outlying suburbs for a total of 103. Instead of inferring too much and trying to decode the multitude of rules for each state we simply struck out any state without all the matches. There were ten (10) states total and one other state (New Hampshire) which was missing the FIPS codes entirely.

Thoughts on Data Challenges:

1) One thing that became clear is that your final data frame can only be as good as the individual parts. While we obtained a trove of useful data in the two sets and having the FIPS code helped greatly, if say the FIPS codes are missing in an entire state, as in the case in New Hampshire there is nothing we can do. The choice comes down to looking up the information manually which 1) defeats the purpose of using an organized dataset 2) is inefficient 3) could lead to mistakes on our end.
2) Having a standardized way of receiving or reordering the primary data would have been helpful. While most of the results were reported by county, the ten states that reported differently affected the merging of the data and therefore the final data frame. Though there was enough data that these omissions wouldn't systematically affect the results, it is interesting to note of the twelve states: Alaska, Connecticut, Illinois, Kansas, Maine, Massachusetts, North Dakota, New Hampshire, Rhode Island, Vermont and Wyoming that were omitted all six (6) New England states are on the list. The other states are concentrated in the Midwest to West and Alaska.
3) Using the County_Fact dataset as the primary table could have alleviated some of the issues, but also could have caused a lot more headaches. As mentioned, since the FIPS code was used as the primary identifier and the FIPS is based off the individual counties this dataset might have been the better table to base everything else from. However, whatever data set you use to relate to the table having that data being reported by county (and therefore FIPS) would be paramount. Instead of the example of Kansas where the data received was from congressional district rather than county, the data would have to be reported in similar fashion.

Recommendations for Future Analysis:

As a next level analysis from the current data it would be interesting to:

1) Add other candidates to the mix, especially on the Republican side, since there were more.
2) Cluster the counties into similar types and analyze those clusters' propensity to vote on each of the candidates

Or with additional data it would be interesting to:

1) Compare the primary results to the general election results.

Conclusion: The primary question we were asked to answer is "Which are the most valued data science skills?" Although we had our own ideas of the most valued data science skills before the assignment, it was important for us to go through the entire process to make sure we didn't have any pre-conceived notions. While there were other vital skills that were required, the most valuable data skill, not because it was the "hardest" but because it was the most unpredictable, is data wrangling. It is exactly the unpredictability of working with large datasets that makes wrangling so valuable. In huge data sets it is often unpredictable how the data will be organized throughout. Every issue must be dealt with to be able to get the fun part: analysis of the data. The secondary question, "What are the key elements that explain which candidate wins?", is integrated in the code that follows.