**Project Three**

The Group

Our group is made up of five (5) members: Bianca Sosnovski, Elina Azrilyan, Robert Mercier, Asher Dvir-Djerassi and Charls Joseph. The two original members are Bianca and Elina. Robert joined next followed by Asher and Charls.

Kickoff meeting

After some back and forth, the original two members of the group and the third member, decided to meet over Skype. On Monday October 8th, the members of our group met over the online platform for the first meeting in regard to the project. The face-to-face interaction of Skype, fostered meaningful interactions and helpful exchanges. While Skype, Zoom, Facetime and our class platform Gotomeeting may be a common occurrence nowadays, I bring up the online meeting to briefly note that these platforms are efficient ways to gather separate people in different locations into a group setting. As a soft skill, the virtual nature of Skype, initially helped us connect and work through some of our problems in real-time.

The **purpose** of the meeting was to meet and go over the details of the group project - the third of the class.

The **scope** of the project was to come up with a dataset, come up with some analysis to do on the data and answer the question: “Which are the most valued data science skills?” While this sufficed for the time, more meetings with new group members sharpened and enhanced the scope of the project.

It was at this opening meeting the current group members picked the dataset on which we choose to do the analysis. It was at this time, we came up with the answer to the Project Three question: We believe that data wrangling is the most valuable data science skill.

At the time, we were all in agreement that data wrangling is the most important and valuable skill to have. The reason was that we believed it formed the foundation of all the work to be done afterwards. The believed, the term GIGO applied pertaining to if you have Garbage In you will certainly have Garbage Out. There are two wrangling issues that came to the forefront. The first was the organization of the data. The second was the data in each cell. As for the organization, making sure the fields as columns and records as rows are clearly labeled is paramount. Then, with each column potentially being a variable or descriptor and the rows containing data for each, the next part is making sure all the data fits exactly one datum for each cell. This creates a workable foundation from which the rest of the analysis can flow from or as Hadley Wickham defines it: In tidy data 1) each variable forms a column 2) each observation forms a row and 3) each type of observational unit forms a table. Manipulating, modeling and visualizing the data afterward all comes from the critical data wrangling that comes at the beginning. Which led us to believe having the ability to wrangle the data is the most valued data science skill.

Our Dataset

We obtained a dataset that dealt with politics. Before going onto the structure of our dataset, the discussions about the circumstances surrounding of the dataset should be addressed. By its very nature politics is a very divisive topic. Everyone wanted to make sure the topic would not be off-putting or an issue for our members. However, after discussing the dataset, we all came up with the mature point-of-view that the data was agnostic, and the interesting background of the subject matter outweighed any societal flashpoints. In addition, it would provide a good lesson on how to be honest brokers and be neutral about a potentially controversial subject matter. The dataset was also top of mind when our two newer members were introduced to the team. They also agreed the data were very interesting and there was a way forward analyzing the data objectively and without bias.

As for the data itself, it deals with primary results of the 2016 Presidential Election. The original dataset came in three (3) csv files: 1) is a file named “Primary\_Results2016” which contains the results of the Democratic and Republican primaries, by US County and for any presidential candidate 2) is a file named “County\_Facts” which contains the demographic breakdown of the US Counties that voted in the primaries and 3) is a file named “Headers” which labels the columns of the county\_facts spreadsheet. We obtained the datasets from the data science community site Kaggle. The URL for the data is: <https://kaggle.com/benhamner/2016-us-election>. It should be mentioned, around this time we started to question the importance of data wrangling vs some other data science skills. While, still of the utmost importance, the sheer volume of the data and the disjointed files made us have to make some decisions about the direction of the project during this meeting. The idea we came up with was to see if there was a difference between the counties that voted for Hilary Clinton and Bernie Sanders and compare them to the counties that voted for Donald Trump. However, we felt at this time we should just gather all the data and see where it led us. This is a very important takeaway we came up with. The reason is that as a wide-ranging project, we wanted to make sure we were not pigeon-holed into specific topic and we are able to do what we set out to do.

However, the data itself posed an additional issue. The three separate files needed to be combined and although we didn’t want to limit ourselves, we also didn’t want to be overrun with *too* much data. So, while there are other candidates included in the primary\_results spreadsheet, we wanted the data for only the three prime candidates.

That said, the plan was to combine the datasets into a new data frame in R. First, we were going to use the counties\_facts spreadsheet as the base, then pull the headers data into this one to name each column. Next, we found that there is a key field in the remaining sets that matches each county with a unique identifier. This FIPS (Federal Information Processing Standard) identifier is in both the County\_Facts and Primary\_Results2016 spreadsheets. So, to import the extra columns needed, we had to attach the data for the fields: 1) the total number of votes and 2) vote percentages in the primaries. This was done for the Democratic candidates Hillary Clinton and Bernie Sanders and the Republican candidate Donald Trump. These six (6) new columns were to be added to the end of the county\_facts spreadsheet and would have to have headers created. This would create the data frame and complete this part of the project.

The uploading and transformation tasks was assigned to Bianca and Elina. This was the first major undertaking of the project. The two main datasets we worked with were downloaded as CSV files. From here the duo read the individual files as data frames in R using the function "read.csv." In addition, the header file needed to be read into R, as it would be needed as explanation headers for the data frame. The County\_Facts file had about fifty fields that had coded headers and needed to be changed to a readable title, this is what the header file did. This was our working data frame we had from the start.

New Group Members

After coming up with the specs for the project, we were contacted by Asher to join the group. Originally, Asher and another student were to be paired with us, but that student found another group. In the interim time, Asher connected with Charls and the two were set to join our new, expanded, five-person group. Again, with the loaded topic of politics, we wanted to make sure they were okay with the topic which they were. From the beginning, Asher and Charls added value to the team by trying to nail down GitHub repo sessions and commenting on the goals requested for project three. However, with our team working different parts of the assignment, now what was needed was another meeting with the new members to get everyone on the same page.

Meeting #2

As another entry to the soft skills record, Asher created a Doodle poll to find out what time worked best for everyone. It worked by giving the team members 1.5 hour time increments to vote on a meeting time, this way everyone was able to see everyone’s schedules for an acceptable time for the most people. The that worked us was on Friday October 12th, so at 4:30 our team had another meeting. This one consisted of Elina, Robert, Asher and Charls. Bianca was out on business and had completed the task of merging the county demographics and column headers together as stated above. In addition, Bianca is the keeper of the GitHub repository and in charge of its management. During the meeting the team discussed a wide range of topics and organized a set of tasks for the group members.

Firstly, Asher and Charls were to work on merging the datasets. While the data had been read into R, the two main data sets had not been merged. Within the next couple of days the two were to finish this process and give the group access to the final combined set. Furthermore, we changed the structure of the data set to include the Republican candidate Ted Cruz. Since, we had all the candidate data we thought it might be beneficial to have Mr. Cruz in the set to get the data more consistent between the Democrats and the Republicans. Since there were two Democratic candidates, it would give a sense of balance to do the same on the Republican side.

Secondly, each of the members with the exception of Robert was to take a crack at answering the question: **What are the key elements that explain which candidate wins?** This new direction sprang from the idea that it wasn’t enough to just merge the data. There was some form of analysis that had to be done on the new dataset. Additionally, with the expanded group, in order to avoid duplication of work, we decided we could either all answer one individual question regarding the data or everyone could come up with a different question to of ask of the dataset. Because the latter could take us into many different directions that could prove difficult to bring back together, the former idea was selected. We thought that it could be interesting to see the different ways we would view the data in light of the question. The question was broad enough that we could get some interesting answers. Robert’s job was to do the write up for the project.

Thirdly, we were going to revisit the question of: Which are the most valued data science skills. After discussing it, we thought we may have jumped the gun on answering the question. Without going through the entire process, it was a little premature to answer the question. It started to come off as we knew the answer to the question before going through the proper process. As data scientist, it is important to gather all the relevant information before making a conclusion and the group felt like we were starting shape the project to lead to our ready-made conclusions. We simply decided to wait to answer the question.

Fourthly, we decided on some of the details of the presentation. Some of these items included trying to use visuals when possible. Coming up with timelines of when to have the sections of the rmd file complete. Even, the timing of the presentation and some basic outlines of who will talk when and for how long.

Lastly, we decided that we could communicate via email and GitHub and no more meetings were needed for the time being. We could wrap up the data analysis by October 18th and complete the writeup for the due date on October 21st. We were free to schedule a meeting, if needed but the next Skype meeting wouldn’t need to be scheduled until after the date of the 21st.

Key Data Challenge:

For this assignment, we prepared data that could be used downstream to analyze the variety of factors that explain the differential primary election results by state, with counties as the unit of observation. As stated above, we sought to allow for such analysis – analysis which we too conducted in the final section of this report – by incorporating demographic information of all 50 states with the primary election results of those states. While we were successful in doing this for the bulk of the states, for a select few the primary election results and the demographic information of those states could not be appropriately merged. However, this did not hinder our ability to very thorough interrogate the question: “What are the key elements that explain primary election results of a candidate by state?”

Fundamentally, this challenge of merging demographic and primary election results was due to the fact that for 11 states, primary election results were not reported on the county level. However, we chose to use demographic information which was organized on the county level. So as to provide the most consistent analysis across states and analysis able to disentangle factors behind election results at the smallest unit possible, we made the decision to exclude those 11 states. Instead, our focus remained on the 39 states where demographic information and election results could be merged by county.

To merge the two datasets, we used a unique county level identifier called the Federal Information Processing Standard Publication (FIPS) code. Since the FIPS code is classified strictly on an individual county basis and 11 states organized their primary results into cities, towns or congressional districts (results manifested into 8-digit codes that where not associated with FIPS), these 11 states were not merged.

Learnings:

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 Analysis: The analysis can go many ways. However, we decided to even up the parties looking at two candidates for each party. Although, the amount of competition in each party, the timing for when the other various candidates left the primaries and makeup of the states makes the dynamic of getting true apples-to-apples comparison between the parties tenuous. However, our analysis does show differences between the two candidates that received their party’s respective nominations.

The Visuals: this last point is an issue that we discussed in our October 12th meeting. As an aid to presenting, visuals are key to bringing the data analysis alive. In the business world, company stakeholders always look for graphs, visualization in order to showcase trends and patterns. Visualization plays a major role in data science projects and is one of the most valued skills by clients both internal and external. While there are many visualization tools in R, Charls led the adoption of using Tableau as well. While R is known for its robust analytic feature, Tableau is one of the top data visualization packages.