RS21 Data Science Interview Responses

**How did you go about getting to know the data?**

I began reviewing the data the old-fashioned way, with pen & paper. I wrote down all of the folder names and the subfolders and files within them. I used this as a map/cheat sheet of the data throughout my analyses. This also helped me get an organized overview of the data sets and their overarching themes. One of the themes I immediately noticed was that all of the data was collected for Albuquerque and surrounding areas. Then, I began digging through the Excel sheets and collecting the dimensions of each data frame. After converting the public transportation file to an Excel document, I was able to quickly notice that this data set, along with the Twitter, Facebook and health indicator datasets all had geolocation as a variable. My initial thoughts were to combine these data sets among matching geolocations to get a larger data set. However, I did not move forward with this initiative as I found an easier way to visualize the data using ArcGIS. With the hdma\_ABQ dataset, I ran exploratory data analysis and data munging as I decided to use this data set for my classification task. I looked through the variables to see which ones were sparse and/or repetitive at first glance and deleted those. Afterwards, I pulled the data frame into python and looked at the dimensions, five summary statistics, first five columns and other characteristics through the exploratory data analysis & data munging phases of my code and project. I was able to take a closer look at variable relationships & distributions via correlation plots, frequency plots and box & whisker plots.

**Why did you choose your particular hypothesis?**

My hypothesis was that the hdma\_ABQ data set could be used by a real estate company to classify their customers into counties that they would want to live in based on information provided by the client via an initial survey. The rest of the data could be visualized on a map to help the customer get a quick and comprehensive idea of the health, transportation & entertainment circumstances within the county that they have been matched with. I chose this hypothesis because I noticed that the data sets were varied but all had one major commonality, location variables. The hdma\_ABQ dataset could have also been used to classify whether or not a customer would qualify for a loan, but I realized that this task did not mesh well with the rest of the data sets.

**How did you approach testing that hypothesis?**

I approached testing the hypothesis via running 3 different algorithms for classification tasks on the data set and measuring their outcomes with accuracy and precision scores. I mainly used the sklearn, pandas and numpy packages in python to run these algorithms.

**What were the results and how could those results be useful to a client?**

The results were that the classification task succeeded with an 81% accuracy from the logistic regression model. This means that the model allows the real estate agency to use its client information to classify the county that they will most likely want to live in correctly, 81% of the time. I was able to gain this score even after removing variables that a customer who walks into the real estate agency is unable to provide regarding a certain county such as minority population and population. The logistic regression model and the random forest were close in accuracy. The neural net had the lowest accuracy.

**What were some of the biggest challenges you faced during this exam and how did you tackle them?**

The biggest challenges I faced during the exam was my initial encounter with the data. I was unfamiliar with .GEOJSON files, NDVI and shapefiles. I researched these files and realized that they are compatible with ArcGIS. I ended up downloading a free trial of the software and realizing that this software was also very compatible with the 4 data sets that had geolocation data as well. I did not end up using the NDVI rasters as they were not valuable to my hypothesis and I was unable to download the .GEOJSON files and was therefore unable to use them as well. However, the hospitals shapefiles meshed very nicely with the health indicator data set and I was able to use these in conjunction to create a visualization of the ABQ area. However due to my unfamiliarity with the initial file types listed, I focused on working with the data sets I was more comfortable with throughout the beginnings of the process until I gained a good understanding of what I wanted from the data and how my hypothesis would pan out.

**If you had five months to develop meaningful insights using this data rather than five days, what other pieces of information and/or analytical techniques would you include? What deeper insights or hypotheses would you explore and how?**

If I had 5 months, I would like to build a classification survey which would ask the client for pertinent data, run the linear regression & match them with a county within 15 minutes of arriving in the real estate agent's office. Afterwards, they could browse this county's entertainment, health and transportation features on a map that is instantaneously created & personalized for them after they input their selected interests (health, transportation, entertainment). I would also like run the same classification task on a more granular output such as neighborhood instead of county after collecting the pertinent dataset. Additionally, I would like to learn how to utilize ArcPy and the map functionalities in matplotlib. These skills would be useful in automating the data mappings. Lastly, I would have liked to run a sentiment analysis on the content of the tweets. I could use this data to then classify whether or not one region possibly has cyberbullying. I began this analysis but was unable to finish. Code is provided.