**Group 6 Project 1 Narrative**

Weather data is prevalent, yet messy. Most likely not the easiest project to kick off a data science career, an analysis of Hurricane Harvey seemed a challenging and rewarding project that quite literally hits close to home. Many lessons were learned in this short journey with regard to data sourcing and acquisition, cleaning and parsing, visualizing and conveying – Data has a story to tell. We greatly enjoyed using some newly acquired skills to piece together our own small bit of this story. The realization grasped fully that there is vast power and much to be learned in this exciting field.

The Gulf of Mexico Coastal Ocean Research “is the heart of data collection for ocean and control waters of the Gulf”. This resource was used to pull rig and weather station web-accessible-folder data in .csv format for cleaning and analysis. The Harvey path was collected using data found via the National Hurricane Center. The actual data pull relied on a mixture of availability of sensors, stations, and rigs in a proximity that was both cohesive to showing results and contained enough completeness of data acquisition in the actual timeframe of the hurricane passage.

All .csv data (air temperatures, water temperatures, air pressures, water salinities, and wind speeds) was obtained and sorted by platform/station. It was then merged via an outer join technique via acquisition date. Redundant columns were removed. Much of the data actually aligned quite nicely due the to the apparent standardization of half-hourly interval. Even though it aligned nicely, the cleaning process was not as straightforward.

Of the potential 27,000+ weather attributes in four months selected over seven sensor points within proximity of the hurricane, our purpose was to determine the most significant attribute contributing to its intensity (in our evaluation = wind speed). We’re considering this to be our dependent variable because the hurricane is classified by meteorologist accordingly between a category 1 to 5 depending predominately on wind speed. Therefore, our analysis consists on making comparisons between the data retrieved by each sensor point with relation to both the actual path and intensity of the hurricane. Our initial null hypothesis (H0) was that sea water temperature was the most significant attribute unless our study proves otherwise.  We compared this among the following results by identifying any anomalies within data retrieved considered as outliers, the correlation coefficient (r) and its p-value, regression over four months interval depicting trends, and dispersion during a closer interval period to the hurricane.

Conclusive evidence presented itself in the correlation calculations within regard to this specific data set: air pressure mathematically related to windspeed inversely with most weight, with regard to the data station that held enough information. The p-values reinforced this data by showing very high significance, with same relation. Based on these calculations, this disproves our null hypothesis. The Air Pressure Firefly diagram is used as a specific tribute to data we would like to highlight: Sonar data that was acquired separately completely overlays expected speeds and their associated lower pressures, as put forth be meteorologists for category 1-5 hurricanes. Unexpectedly, the wind speed data collected by the rigs / weather stations did not exceed these speeds, but extra study could prove that these were sustained winds (as opposed to measured gusts). Even given this fact, the data utilized for the analysis, that was collected at rig level, still greatly correlated inversely to air pressure with the highest significance.

As extra commentary, the 42043 location did not have enough wind speed data (damaged sensor) to come to this specific conclusion; Sea Water Temperature showed a high impact for this locale. Interestingly, for the TABS-X location in which the eye of the hurricane passed directly over, Sea Water Temp again showed strongest correlation. These were interesting findings when considering the dataset as a whole. A Sea Water Temperature Firefly diagram was also produced to prove that the temperatures recorded in the days leading up to Harvey landfall were significantly above the expected values considered as “elevated risk” for hurricane development by meteorologists. It should be noted that we did find several calculations to return 0, meaning no correlation, not significance, and/or no data present.

As stated, there was quite a learning curve for gathering and cleaning the data. Weather data is rarely complete – we received a solid lesson in using what is available in the time allotted for a minimum viable product, then revisiting and adding depiction with any remaining time. It would be very interesting to continue the study and enlarge the dataset with some items that were found slightly too late in the project (as an example, sonar data) to include. Some interesting studies to continue this project could encompass:

* Using regression studies to compare historical data/seasonality to the Harvey- specific data.
* Continue slice and study segments of the data via K-means clustering – distancing of centroids could tell a great story.
* Revisit the incomplete data above, using more sources and make better analysis.
* Add new measurements of intensity such as size/development speed of the hurricane.

There is a rejuvenation of confidence that comes along with successfully utilizing newly acquired skills that, prior to project, could have been in question regarding the solidity of retention. This was a great experiment and we are excited to continue honing these skills and adding the to data analysis toolbox.