Weekly Progress Report: Oct 16 - Oct 20

Contributors: Ryan Choe, Rakin Hussain, Sid Taneja

Progress/Accomplishments:

- Sid created a new dataset in which only columns with <20% missing values and locations with <20% missing values are still present.
  - Filtered dataset from 67 columns to 25 columns
  - Filtered dataset from ~166k rows (238 locations) to ~131k rows (168 locations)
- Ryan and Rakin began to explore missing value imputation upon this new dataset - specifically KNN imputation. Ryan pursued two options for imputing (filling with previous values and mean values) just for elementary experimentation. Ryan also used this resource to help look at potential imputing techniques to explore later. Rakin helped Ryan develop the KNN model. N_nearest neighbors was chosen as k=sqrt(n), where n=# observations in the dataset.
- Rakin included a variance inflation factor (VIF) analysis to check the selected predictors for signs of multicollinearity.

Problems/Challenges:

- Trouble deciding a threshold to filter with missing values. Decided on 20% after seeing that any higher values removed important columns, but lower values kept too many columns.
- Trouble with implementing KNN Imputation for missing values (syntax issues). After a lot of troubleshooting/debugging we were able to figure out. May require further guidance to determine a more appropriate n_neighbors value, if applicable.

Plan/Next Steps:

- Finish imputation of values and dealing with missingness. Will need to look for feedback from Professor Shi and TA Lining.
- Create visualizations using previously made functions by the team for the intermediate demo.
- Determine how to deal with multicollinearity in accordance with VIF results.
- Craft Intermediate Demo and Midterm Slides, accordingly

Overall, the team has explored various techniques into handling missing values. By the end of next week, the group will have finished most (if not all) of the EDA and will begin

exploring the development of prediction models. The group is planning to pursue XGBoost and Decision Tree models, but must ensure the proper time series formatting.