Climate Prediction Challenges: Project 1

Hurricane Economic Loss Prediction

Team 3

Arnav Saxena
Jianing Fang
Nan Zhang
Emily Glazer

Overview

- Associations between different hurricane characteristics and the economic loss they cause
- Engineering & experimenting with new hurricane features
  - Power Dissipation Index
  - Sea Surface Temperature
  - Nakamura Clusters (clustering using hurricane track moments)
- Building a linear regression model for predicting economic losses caused by hurricanes
  - Involved handling missing values
  - Encoding categorical data
  - Interpreting feature importance
- Future work: ideas we thought but couldn't implement in the stipulated time

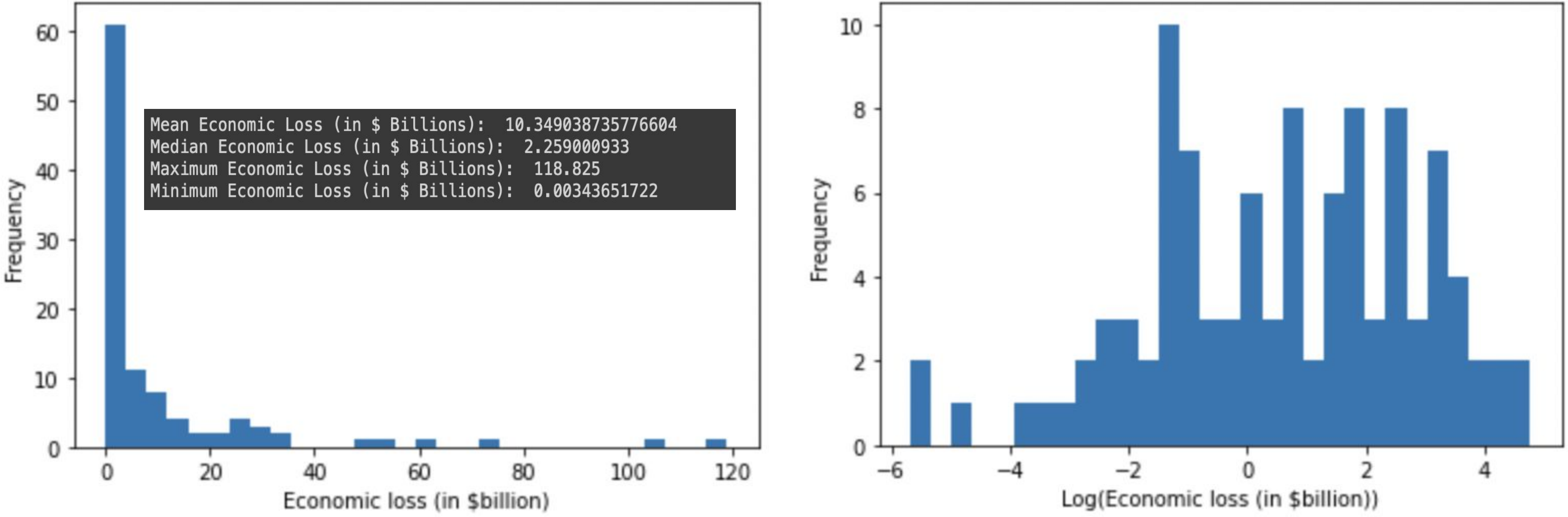
Datasets Used

- **Hurricane Track Data -** IBTrACs best track data. We selected 103 storm records from 1950 - 2017 to match with the hurricane damage dataset. The variables used are longitude, latitude, time, landfall flag, minimum pressure, and maximum windspeed. All variables are from the USA agencies.

- **Normalized Hurricane Damage** - contained damage records of 197 hurricanes causing 206 landfalls in continental US from 1900-2017, costs normalized to 2018 US price level (Weinkle et al. 2018 *Nature Sustainability*). We matched 103 hurricanes with the track dataset to reduce uncertainties in earlier estimates.

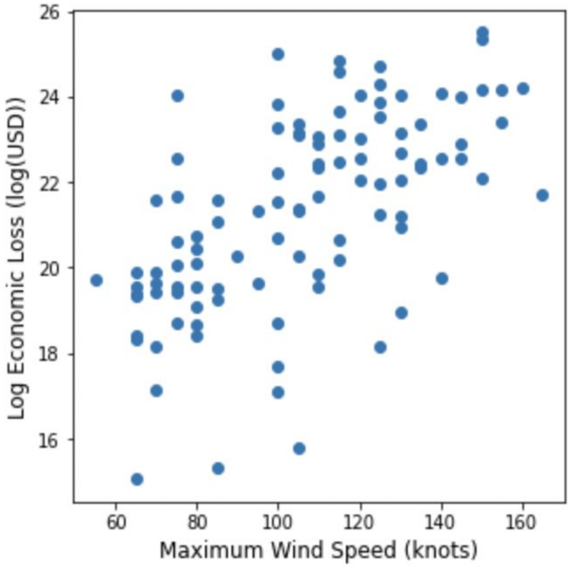
- **NOAA Optimum Interpolation Sea Surface Temperature v2** - global gridded dataset at 1 degree resolution, available since 1981. Only long term monthly means was available from 1961-1981. We used the best available data.

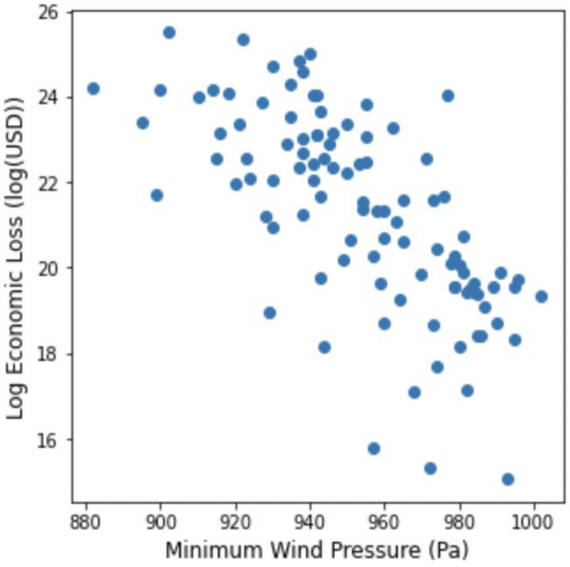
Target variable - Economic Loss

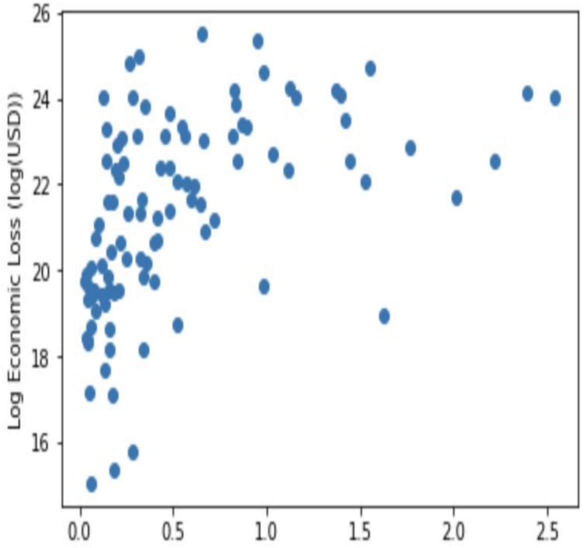


- ➢ The histogram on the left shows that **the economic loss is highly positively skewed.** The maximum loss goes up to $120 B where as the mean and median hovers around ~$10 and ~$2 respectively
- ➢ Hence we decided to **apply log transformation to the economic loss** as otherwise the tail might have acted as outlier thereby impacting our regression model

Max Wind Speed, Minimum Pressure, Power Dissipation Index





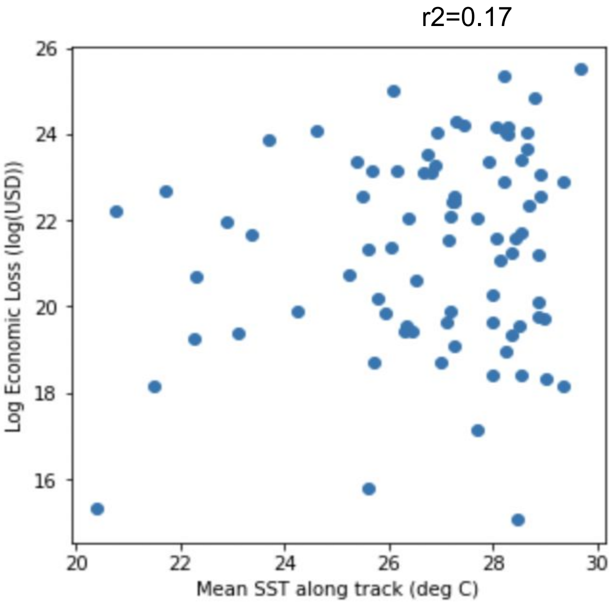


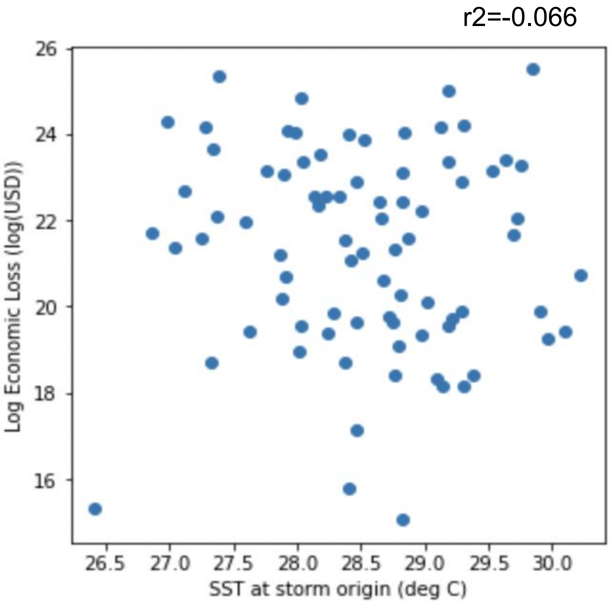
Power Dissipation Index(m^3/s^2)

- ➢ PDI provides an integrated measure of storm intensity by combining wind speed and lifespan
- ➢ Correlate with SST and low-level vorticity

$$\mathrm{PDI} \equiv \int_1^n V^3\, dt,$$

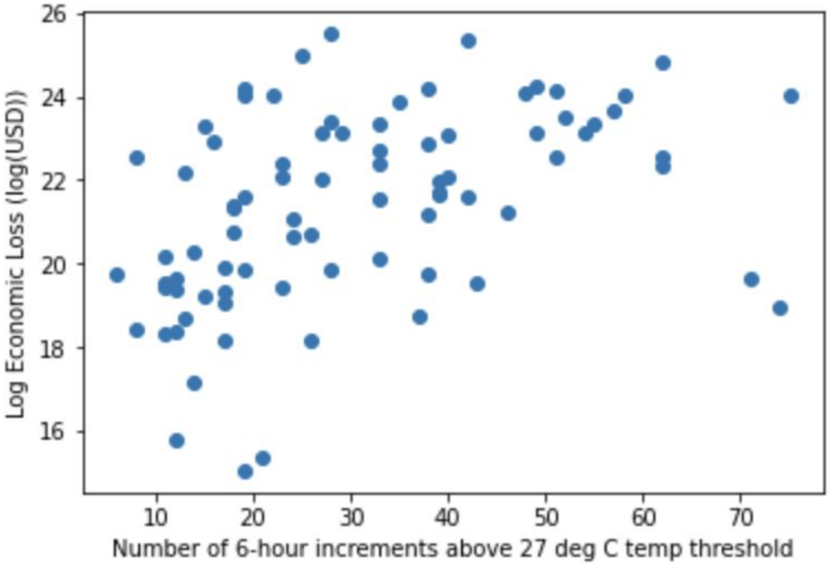
SST: Mean SST, Origin SST, & Threshold





- ➢ Sea surface temperature provides a measure of the potential available energy for storm formation but has an indirect impact on the costs of hurricane damage.
- ➢ Need to differentiate between mean energy state and individual extreme event intensity.
- ➢ Highlights the value in developing satellite observational networks for hurricane early warning and designing robust methods for wind and precipitation nowcasting during an impact.

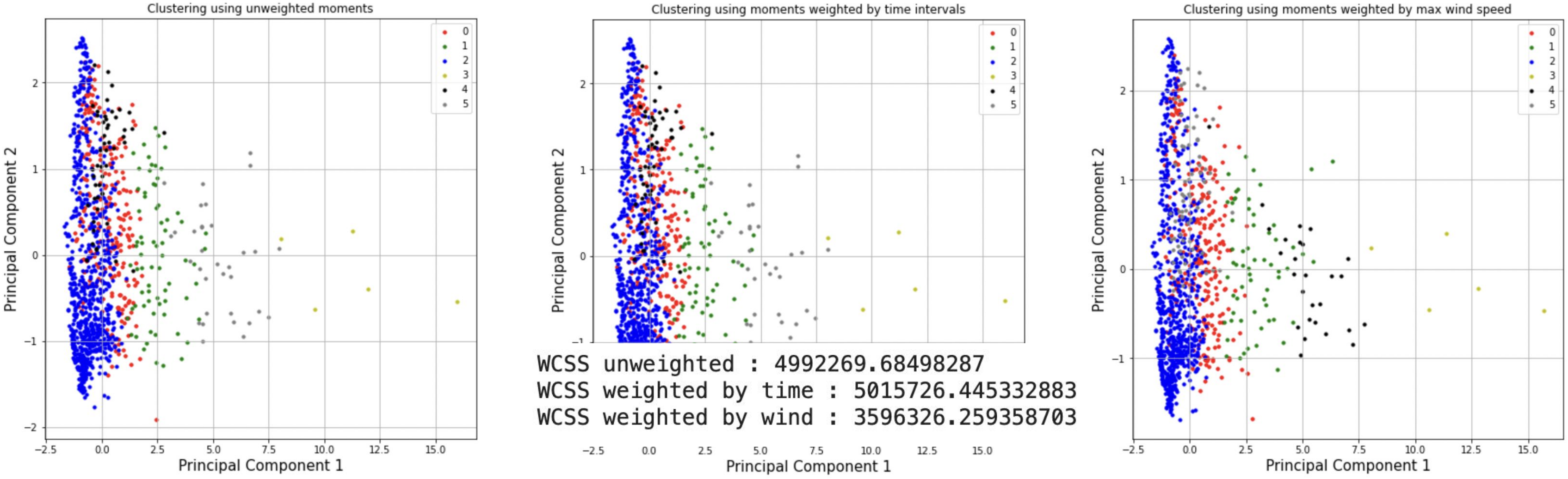
SST: Threshold Crossing Rate



r2=0.44

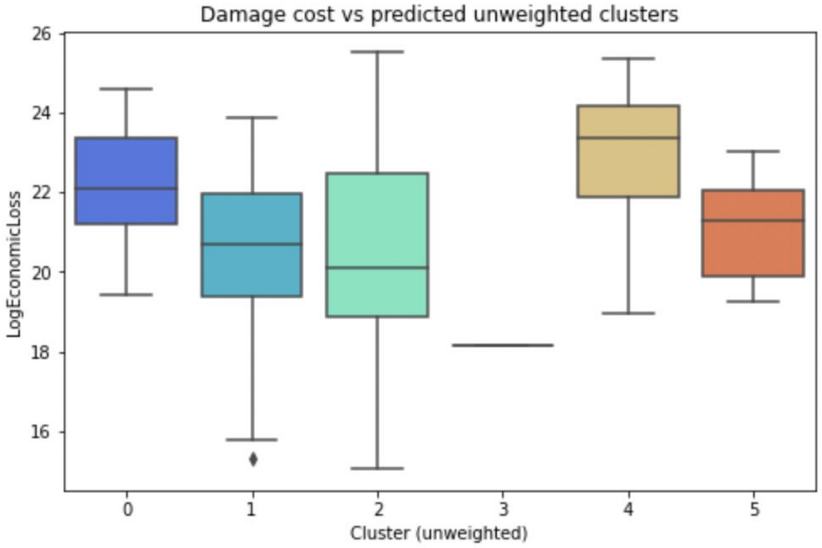
- ➢ As mean SST along track and origin SST did not appear to be significantly correlated with economic loss, we defined a new variable: number of times (in 6-hour increments) the SST crossed a temperature of 27 degrees.
- ➢ The 6-hour interval was chosen because that is the maximum time between measurements in the dataset.
- ➢ The 27 degrees C threshold was chosen because:
  - – From the previous slide, *most* hurricanes seem to originate where SST is at least 27 degrees C.
  - – Through trial-and-error, this threshold temperature had a moderately high correlation with log economic loss (0.44).

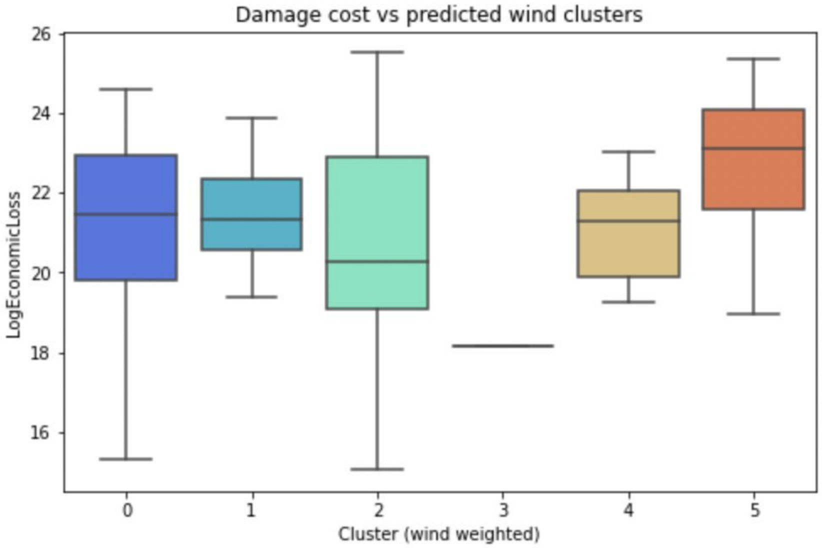
Nakamura Clusters



- Idea was to use the clustering methodology used in Nakamura et al. (2009) to encode the hurricane track information in a single variable and explore if this variable held any predictive power
- We experimented weighing the X & Y coordinates with time interval between adjacent observations and max wind speed along with testing a non-weighted moments calculation
- Unweighted and time weighted clusters were almost identical since almost all coordinates values were available across each time step and hence the weight ended up having negligible effect
- Clusters weighted by max wind speed ended up having the least WCSS

Nakamura Clusters Deep Dive: Clusters vs Target





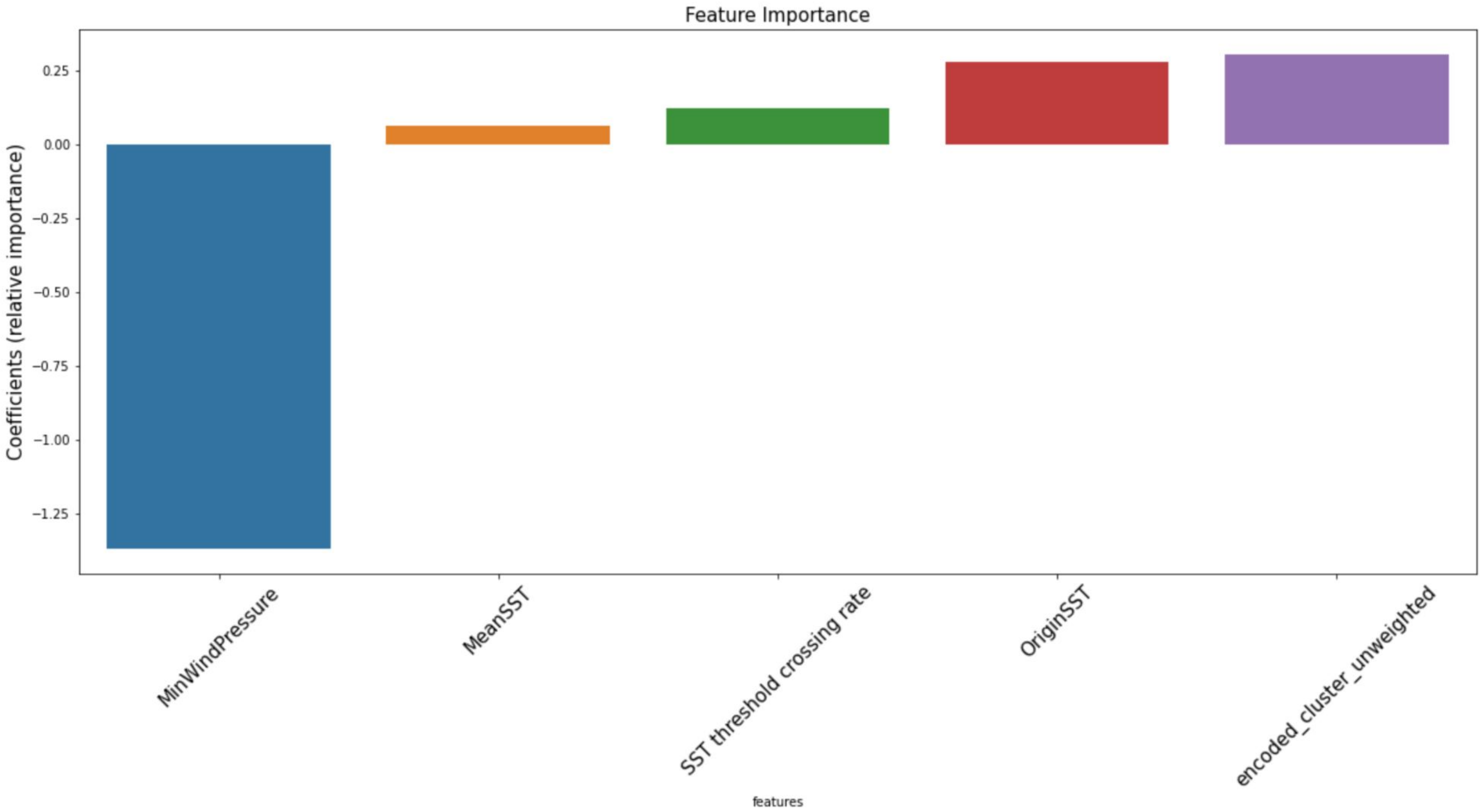
Linear Regression: Multicollinearity

Linear Regression: Results

```
Coefficients: [-1.36856594  0.06117912  0.27770411  0.12184377  0.30561396]
Intercept:  21.31688310267105

Evaluation:
r2 socre:  0.7064054439049541
mean_sqrd_error:  1.339101444148785
```

Linear Regression: Feature importance



Columbia University
in the City of New York

Summary

- Started with 9 features including **max wind speed**, **min wind pressure**, **PDI**, **different representations of SST**, our own implementation of **Nakamura clusters**
- Showcased almost all these features were **moderate to highly correlated with the economic loss**
- Dealt with multicollinearity by getting rid of certain features
- Found that **min wind pressure was the strongest predictor** of economic log loss with the nakamura clusters and SST both being a distant second

Future Work

Future Work - What if…

- **…we had incorporated other features in the model that are also relevant for predicting economic loss?**
  - features such as
    - *duration of landfall*
    - *lifespan* of *a hurricane*
    - *site of landfall and the relevant social/economic indices*
    - *time of impact* (as a proxy for sea level during different times of the day)

- **…we were able to find additional earth system datasets that we couldn't in this universe?**
  - Wanted to collect more data on seemingly important metrics **such as *wind shear*, *aerosol concentrations***
  - With a **larger dataset** we might have wanted to **explore complex models** such as random forest, XGboost out for figuring out the importance of different features we have used

- **…we had more time at hands?**
  - Analyzed the clusters and compared their properties with that of Nakamura's
  - Applied PCA to correlated columns them and ensured that we didn't lose information which we definitely did while deleting some correlated columns

Appendix

Our linear regression model in a snapshot

- **Handling missing data:**
  - ST data (mean, origin, and threshold) was missing ~20% of the observations
  - Hence we imputed the values using KNN imputer

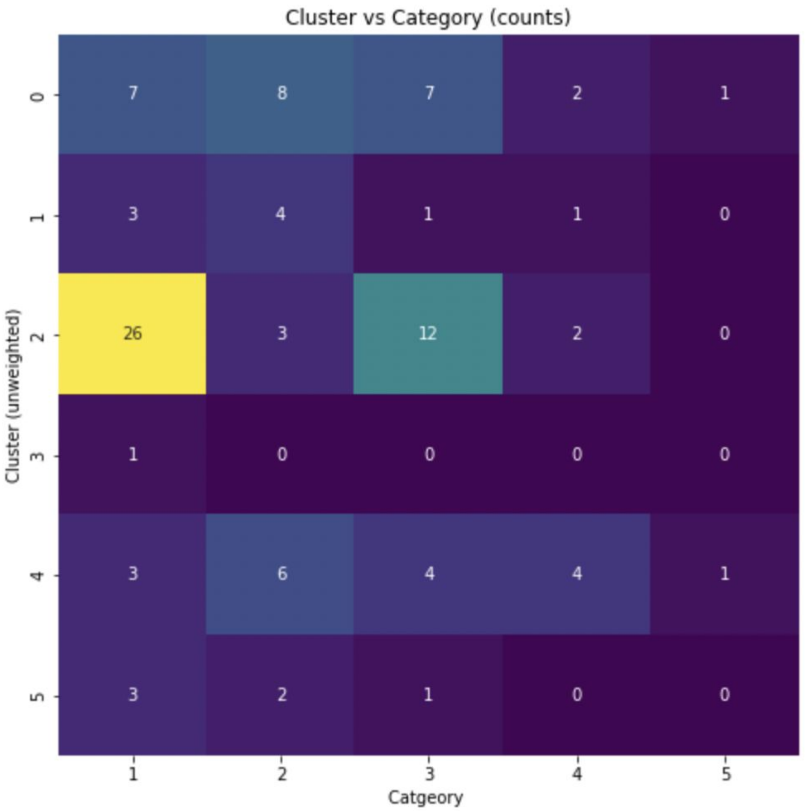
- **Feature Scaling:**
  - Applied StandardScaler() on continuous variables

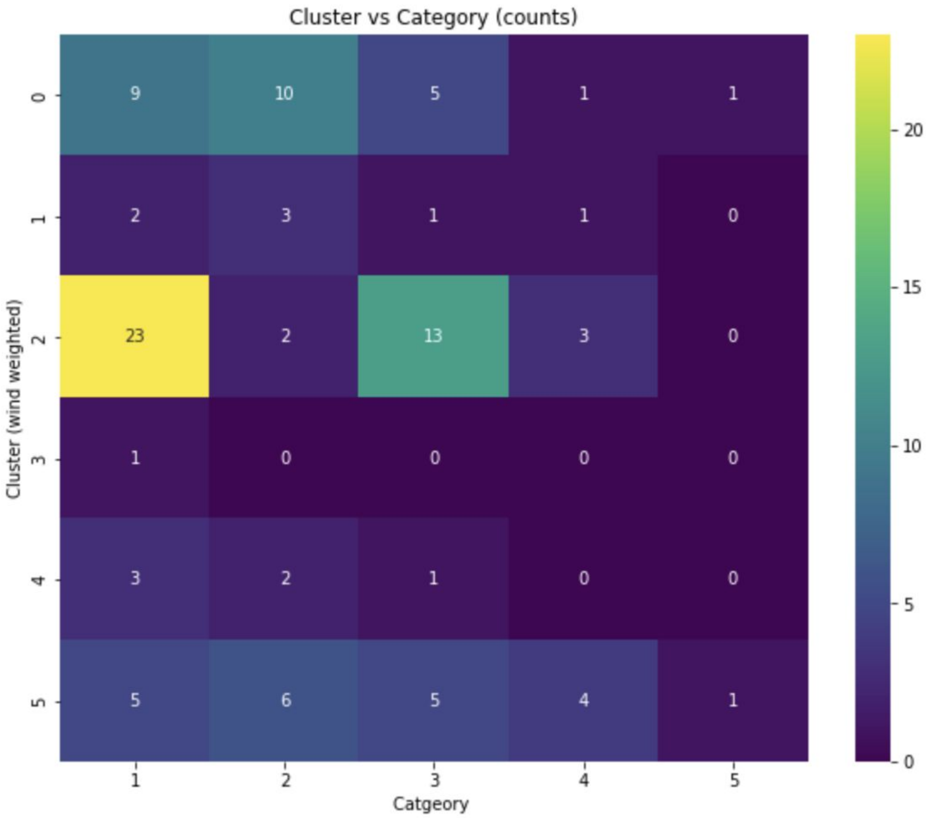
- **Encoding categorical variable (Clusters):**
  - Used target encoding to encode clusters and mapped the clusters to median log(economic losses) for each cluster

- **Multicollinearity:**
  - Observed multicollinearity among obviously correlated variables such as PDI-Max Wind Speed, and unweighted clusters and time weighted clusters. Have ignored one variable each in these cases

- **Training:**
  - Used a dev-test split of 80:20
  - Applied K-fold cross validation with K=4 for training

Appendix: Nakamura Clusters Deep Dive: Clusters vs Intensity Category





Nakamura Clusters Deep Dive: Clusters vs Target