**Scope**: Predict the loss ratio (Claims/ Total Insured Value) for flood events induced by Hurricane

**Summary**

The dataset is derived from the system of national insurance flood act. The data sets consist historical data of 42 attributes and 2.5 millions of records. The attributes include information of Policy, Coverage values, Claim Values, location info, building features, elevation, Flood Vulnerability and miscellaneous details.

To identify claim records for flood event induced by Hurricane, a historical data set of hurricane events for 50 years was used. This allowed the data to segregate into two categories

1. Hurricane induced flood

2. Actual flood

**Data Processing**

**Data Cleaning**

Duplicate Records

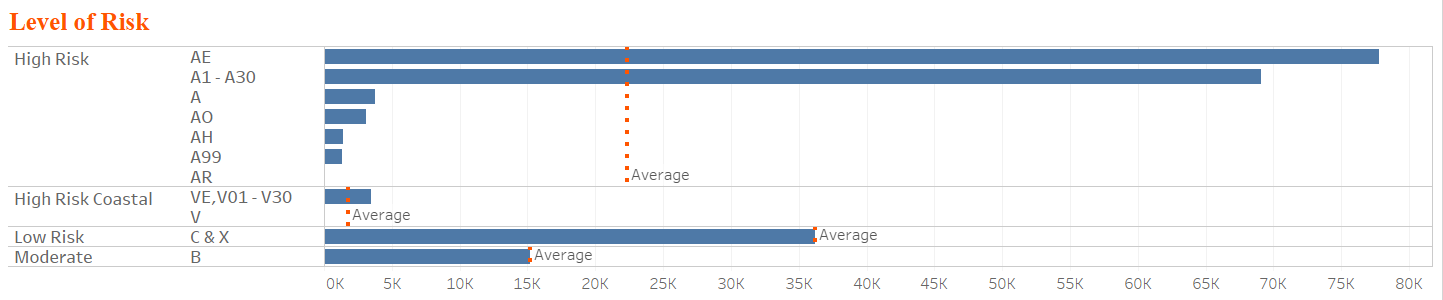
No Claims or negative reported Claims

Limited address information

Converting NA values to 0

**Data Transformation/ Feature Engineering**

Flood zones were provided but were not defined. They were categorized by varying levels of flood risk.



*Figure 1: Average Loss Ratio by level of Risk*

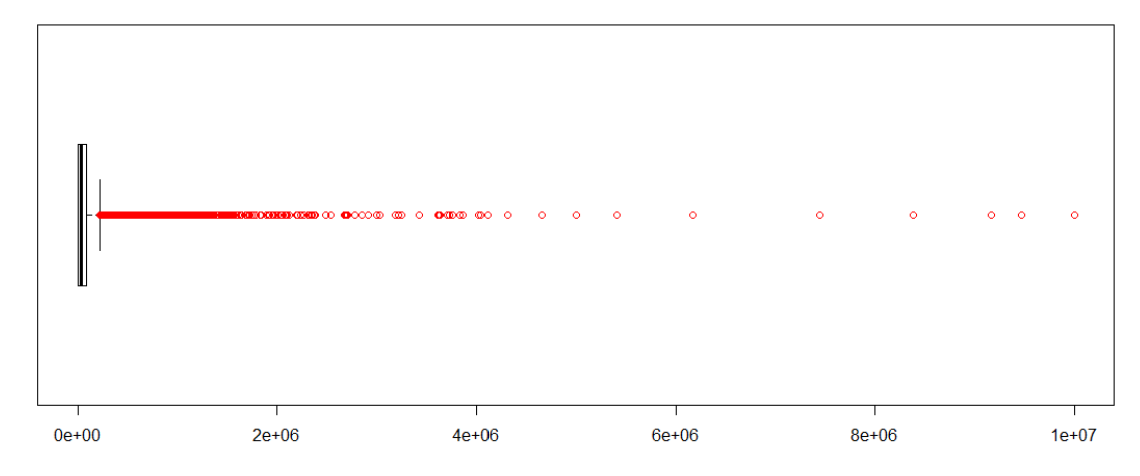
Total Insured Value

Total Claim

Loss Ratio

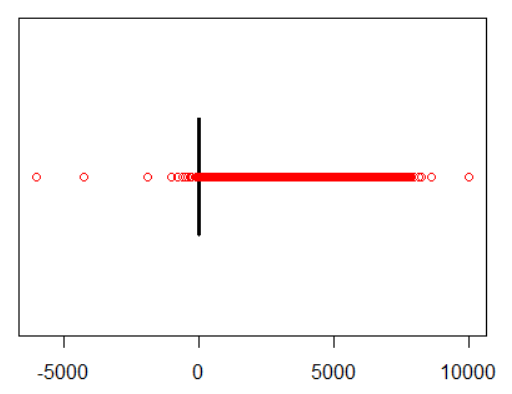
**Outlier Treatment**

As seen below the data had significant number of outliers. We couldn’t treat outliers as they represent high claim amount



*Figure 2: Distribution of Claim Amount*

Elevation Information was provided with 5 variables that had 80% of missing values and had outliers. After reading some document, I observed that the elevation information isn’t important for all the flood Zones.



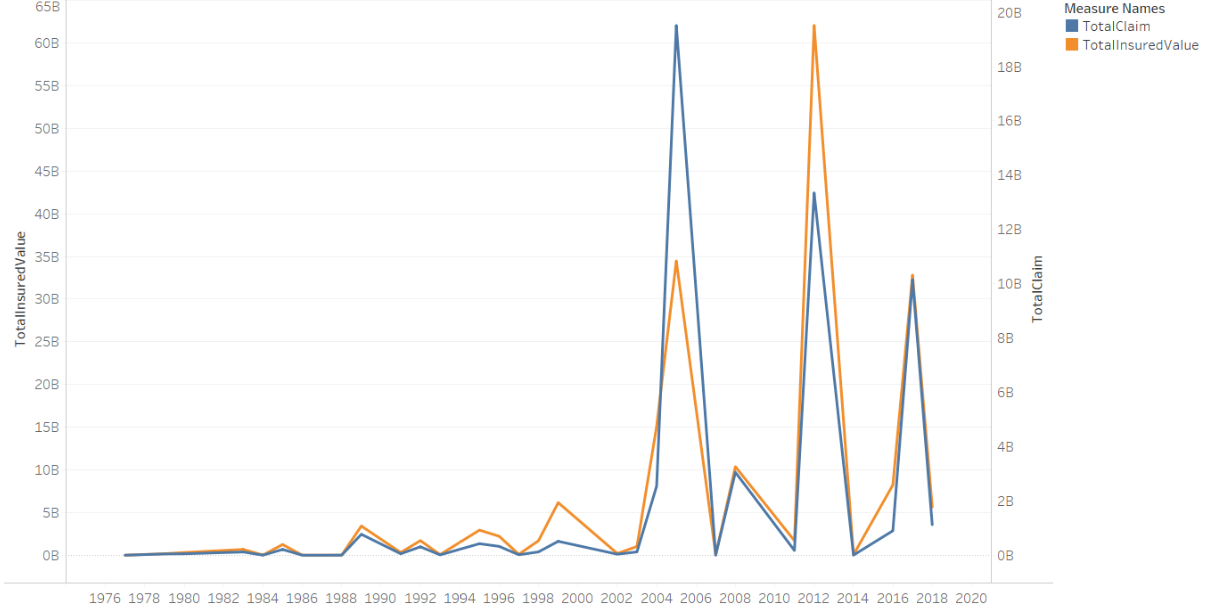
*Figure 3: Distribution of Elevation attributes*

**Missing Value Imputation**

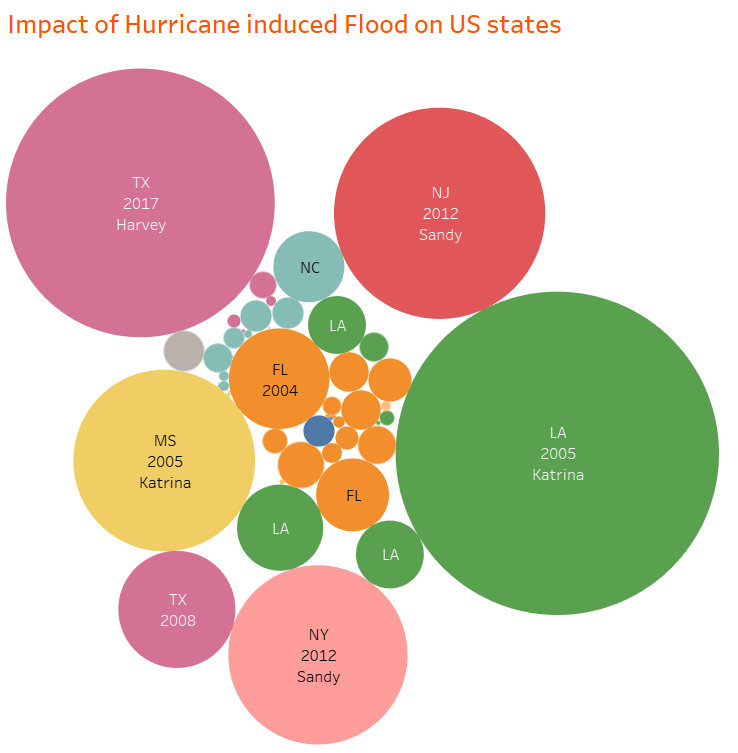
* Single Imputation
* Multiple Imputation

Therefore, the missing values of the elevation attributes are updated with the Interquartile range for Zone based on the level of risk. While the minimum and maximum outliers were scaled with 25th and 75th percentile.

**Exploratory Data Analysis**

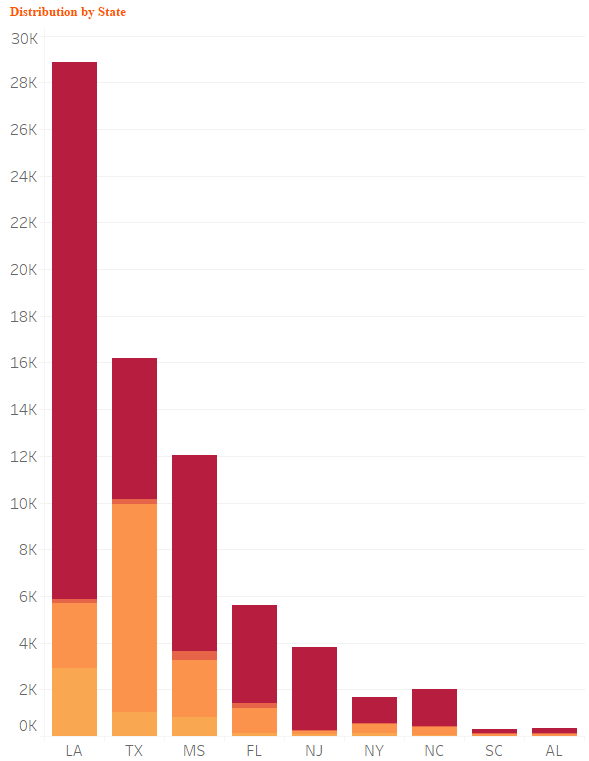
 *Figure 4: Behaviour of Loss Ratio from historical Flood events*

*The highest claims from flood event were reported for 2005 followed by 2012 and 2017*



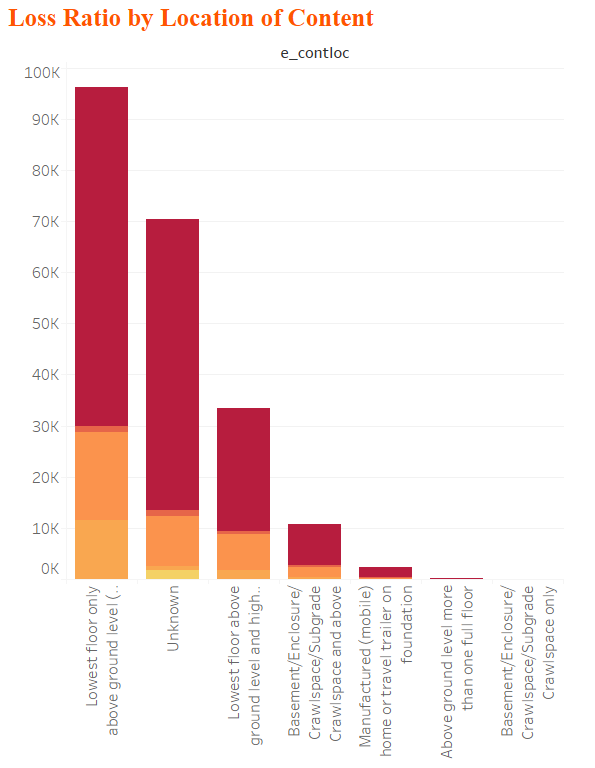
*Figure 5: Impact of Hurricane induced Flood on States*

*In the year 2005, Loss angles and Mississippi reported maximum claims from floods induced by Katrina. Texas reported the second highest claims in 2017 due to Harvey*



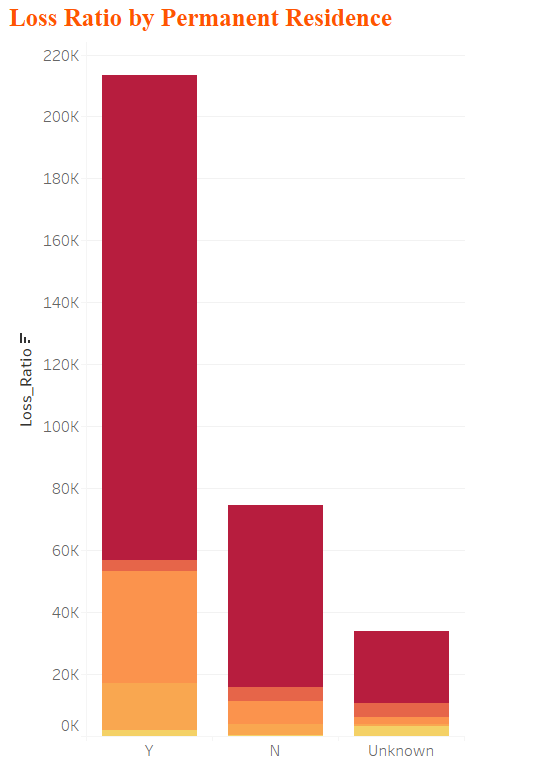
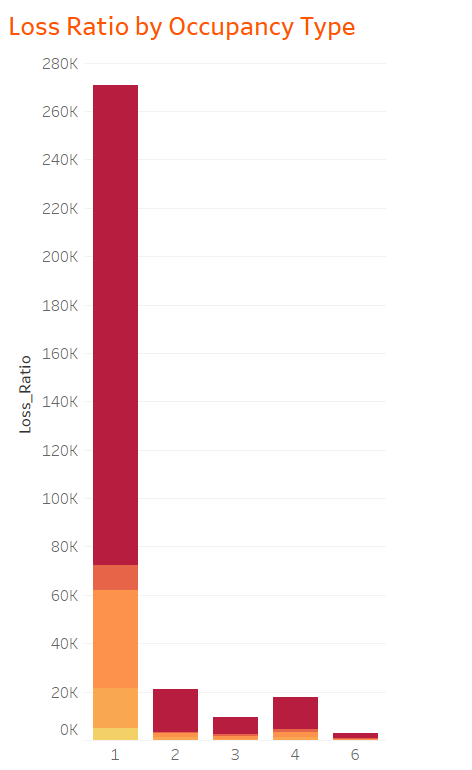
*Figure 6: Distribution of Loss Ratio across States by varying Level of Risk*

*Loss Angeles was severely impacted due to hurricane Katrina*



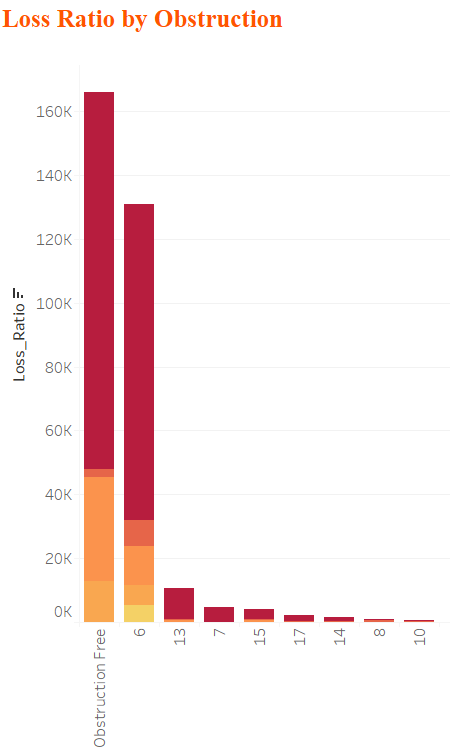
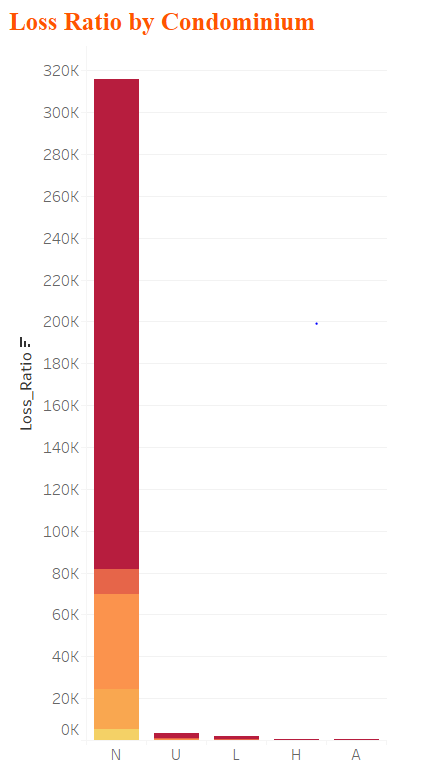
*Figure 7 : Loss Ratio by Location of Content*

*Lowest floor above the ground were highly impacted*



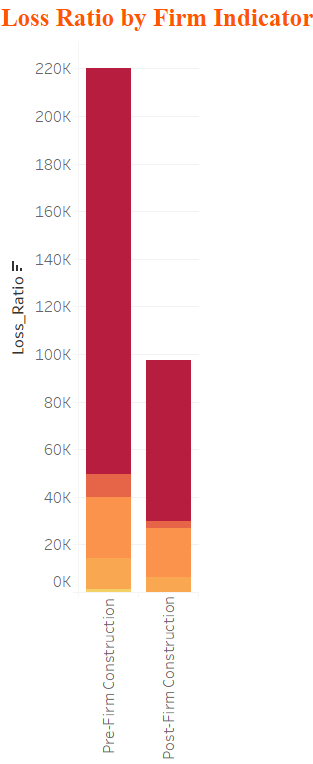
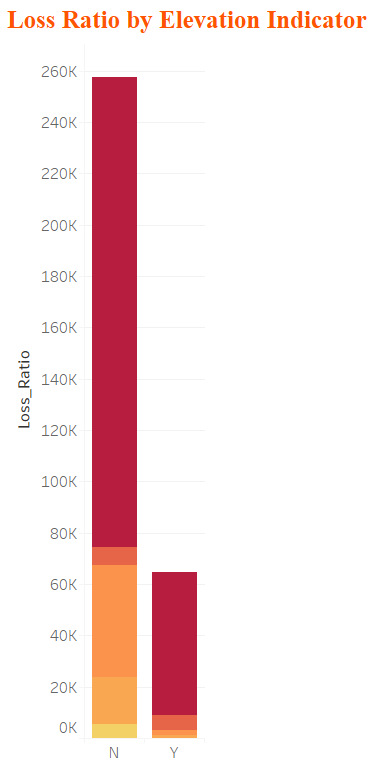
*Figure 8: Permanent Residents were reported with high claims*

*Figure 7: Single Family Occupancy was the most affected*

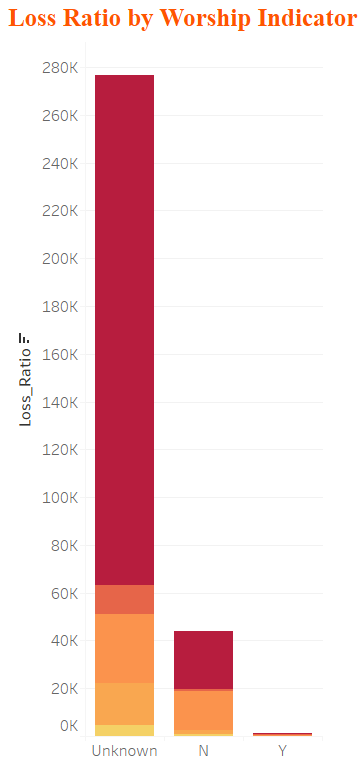
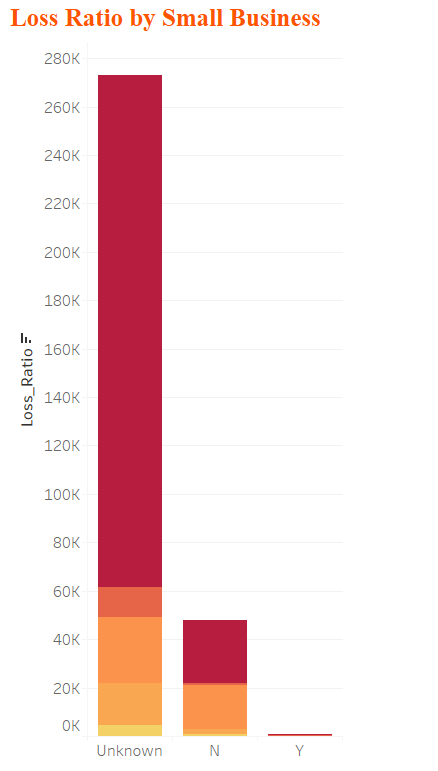
*Figure 9: Properties with no obstruction were the most damaged*

*Figure 10: Condominium observed less impact*



*Figure 12: Construction before 1975 had higher claims*

*Figure 11: Locations with no Elevation suffered more*

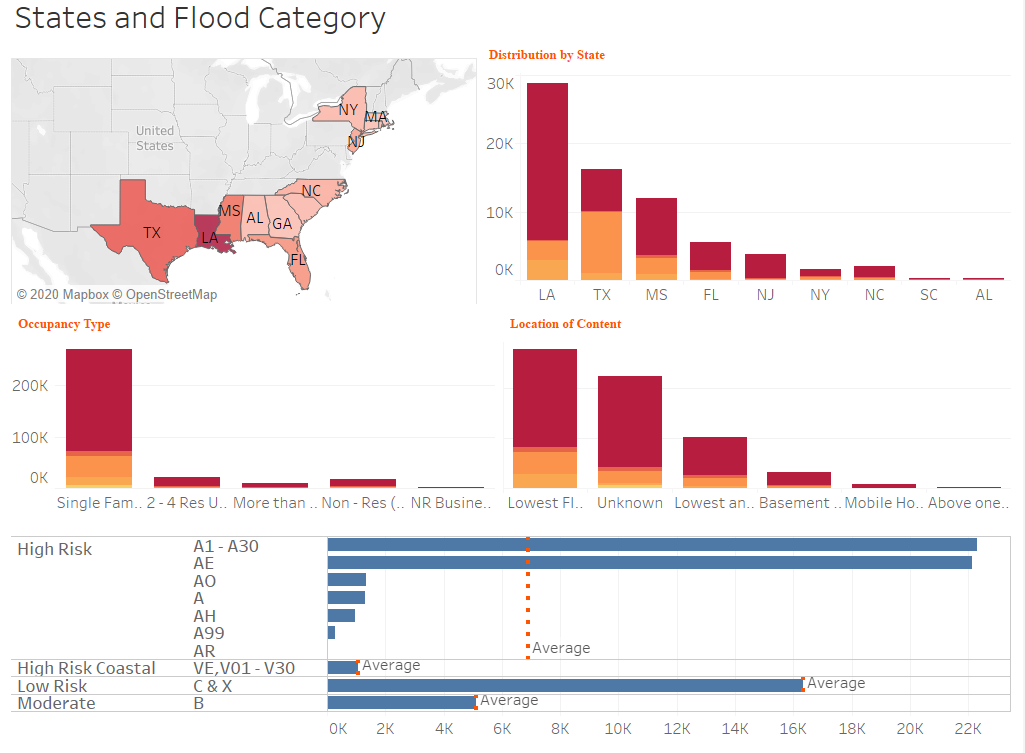


*Figure 14*

*Figure 13*

Since the above parameters are driven by unknown, their contribution towards the Loss ratio is undefined

After extracting records reported for flood claims induced by Hurricane, the No of records remaining for further analysis was 844k. A Hurricane can event may cause flood after few days from the beginning of the event. We assumed that the duration could range from 5 to 7 days. Based on the no of days and the affected states, records with flood claims were extracted.



*Figure 15: EDA for Hurricane induced Flood*

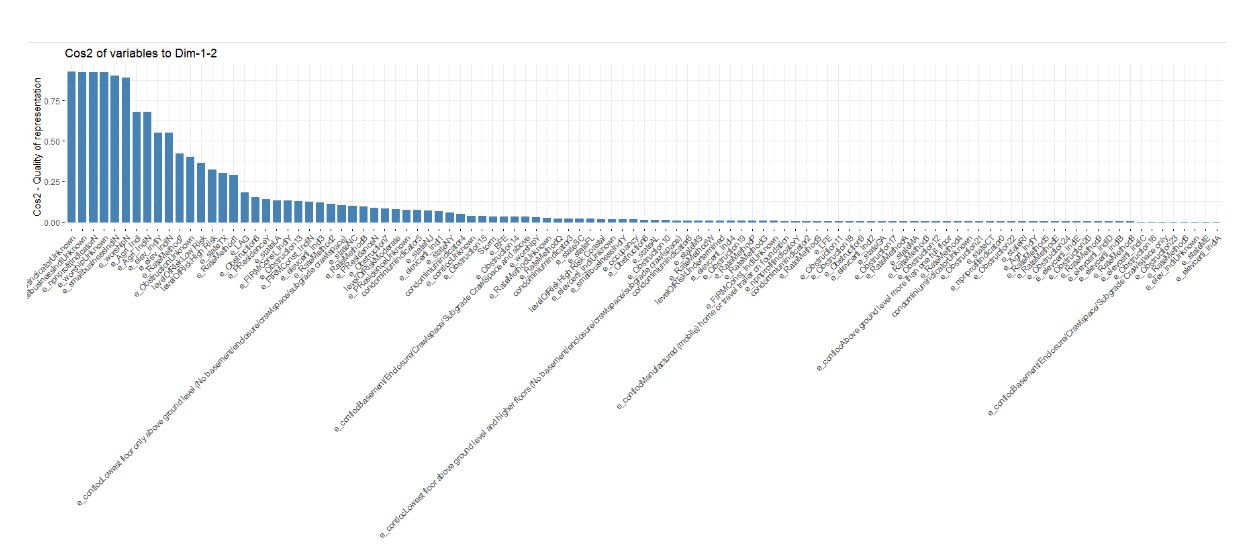
One Hot encoding

There were 113 attributes that were to be employed for building the model

Feature Selection

To identify significant variables, step-AIC was used and we were able to reduce the independent variables to 80

Unsupervised technique called Principle Component Analysis was used identify the key metrics



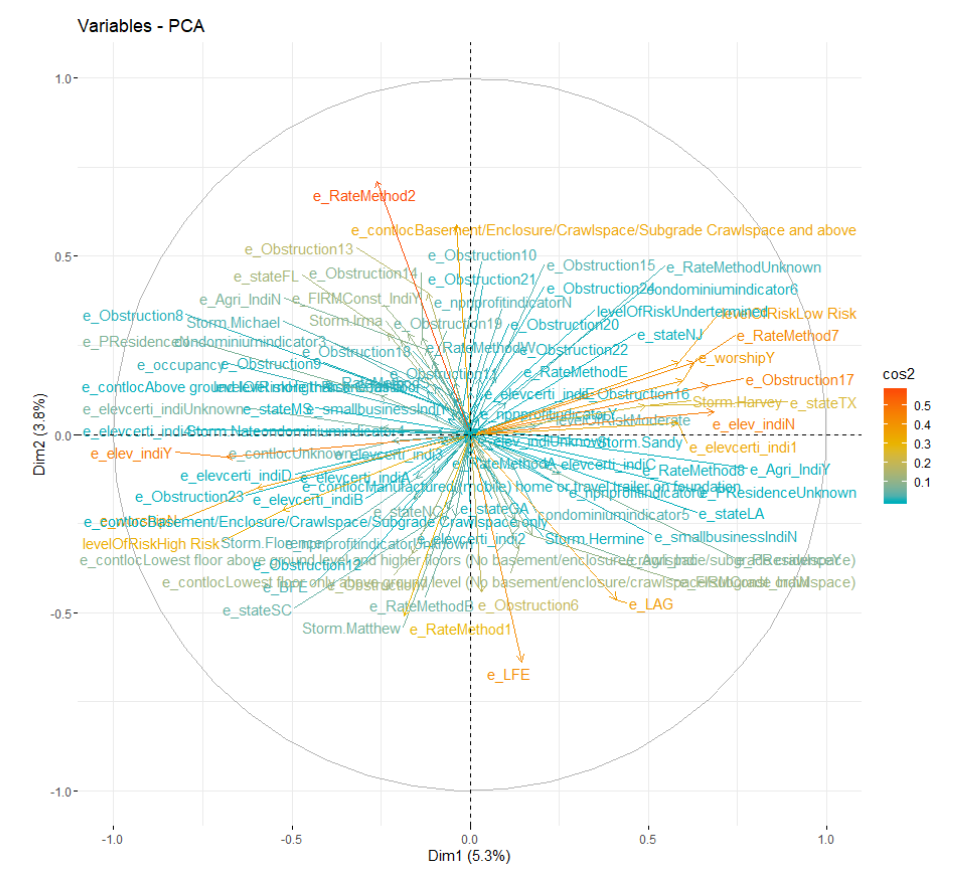
*Figure 16: PCA with 80 variables*

*Contribution of attributes to dimension 1 and 2*

*A high cos2 indicates a good representation of the variable on the principal component. In this case the variable are positioned close to the circumference of the correlation circle.*

*A low cos2 indicates that the variable is not perfectly represented by the PCs and are close to the center of the circle.*

*Variables that are closed to the center of the plot are less important for the first components.*

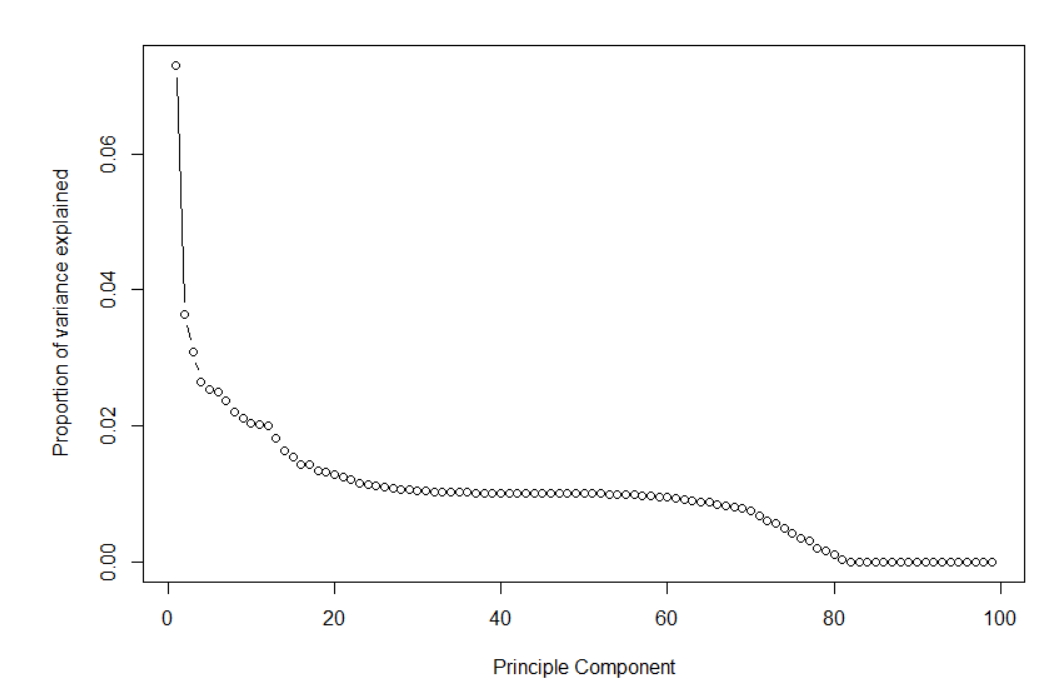


*Figure 17: Correlation Circle*

*variables with low cos2 values will be colored in “white”*

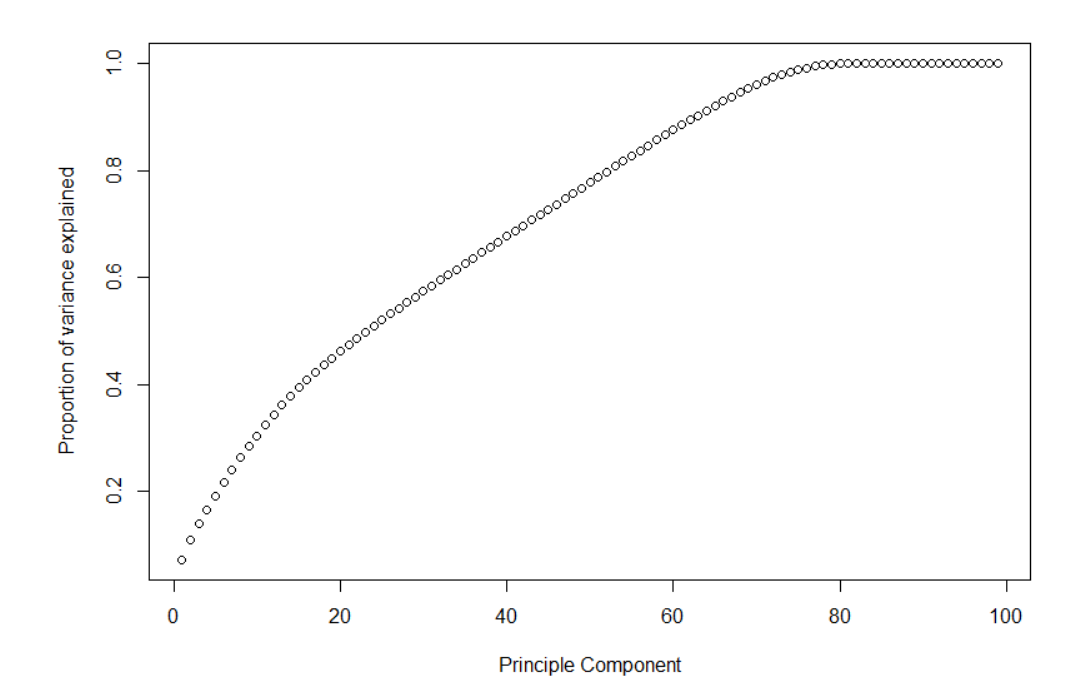
*variables with mid cos2 values will be colored in “blue”*

*variables with high cos2 values will be colored in red*



*Figure18: Elbow plot*

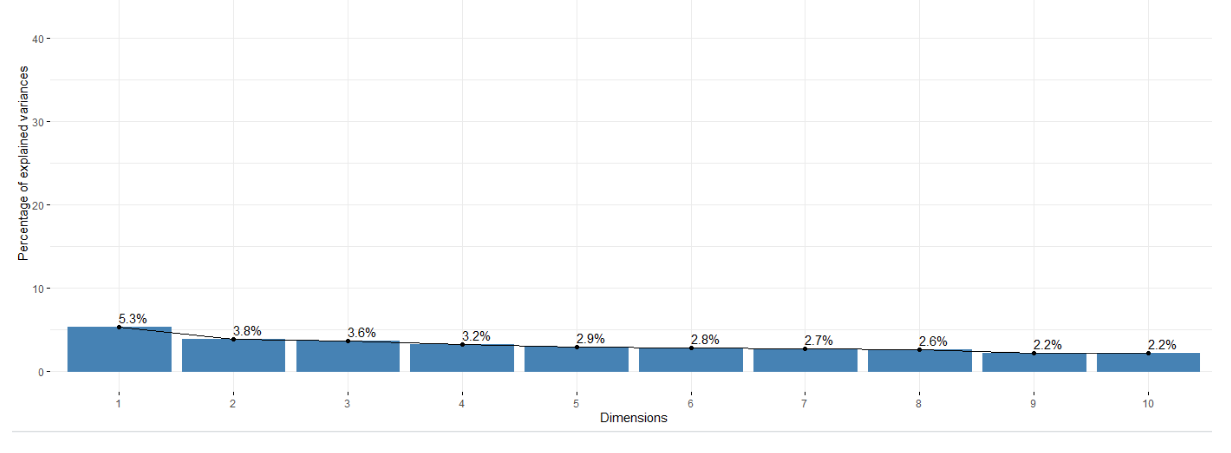
*The plot above shows that ~ 80 components explains around 98.4% variance in the data set. In order words, using PCA we have reduced 98 predictors to 80 without compromising on explained variance.*



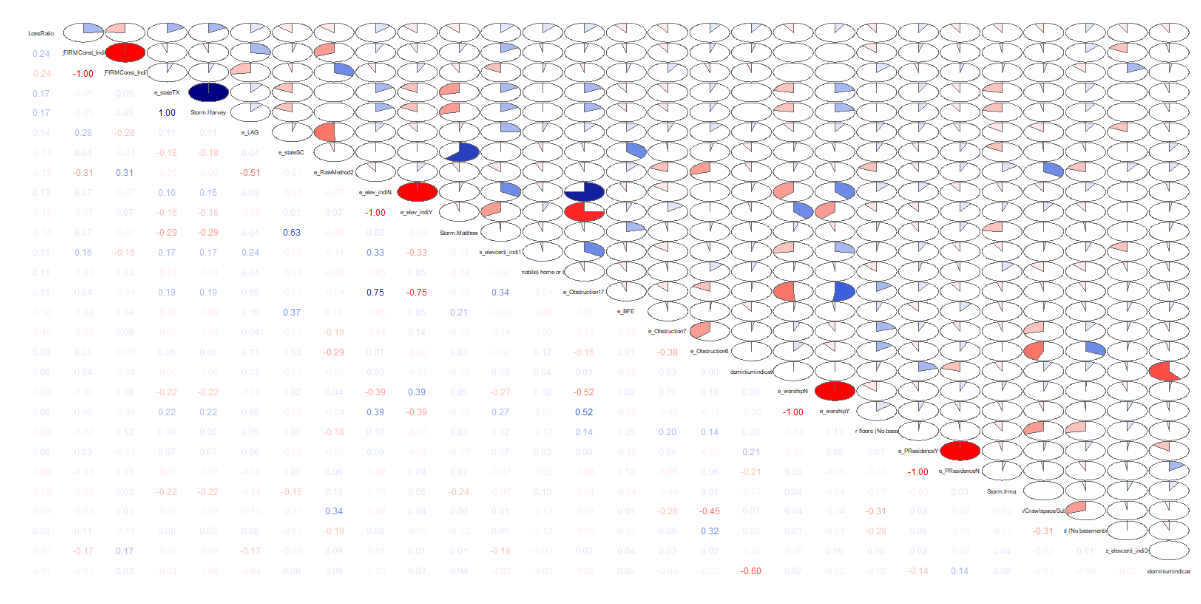
*Figure 19: Screeplot for 80 variables*

*This plot shows that 80 components results in variance close to ~ 98%. Therefore, in this case, we’ll select number of components as 80 [PC1 to PC30].*

The analysis identified “unknown” attributes for small business, Workship, Obtruction type and Firm Indicator. Having Unknown as PC would not add value to the model. Therefore, records with the above mentioned unknown attributes were excluded from modeling. The number of records were reduced to approximately 36000 be significant in idenfying the as STEP AIC

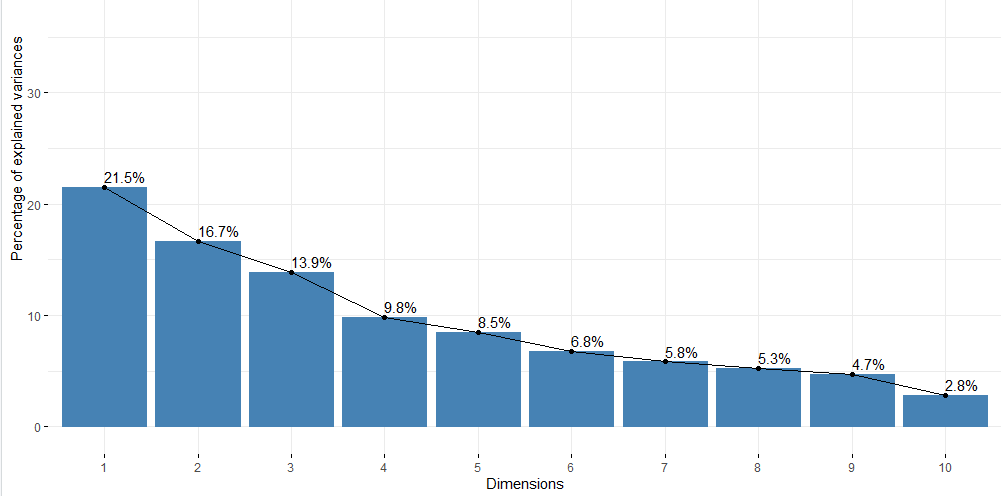


*Figure 20: Proportion of variance explained by each variable*



*Figure 21: Correlation Plot*

Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients.

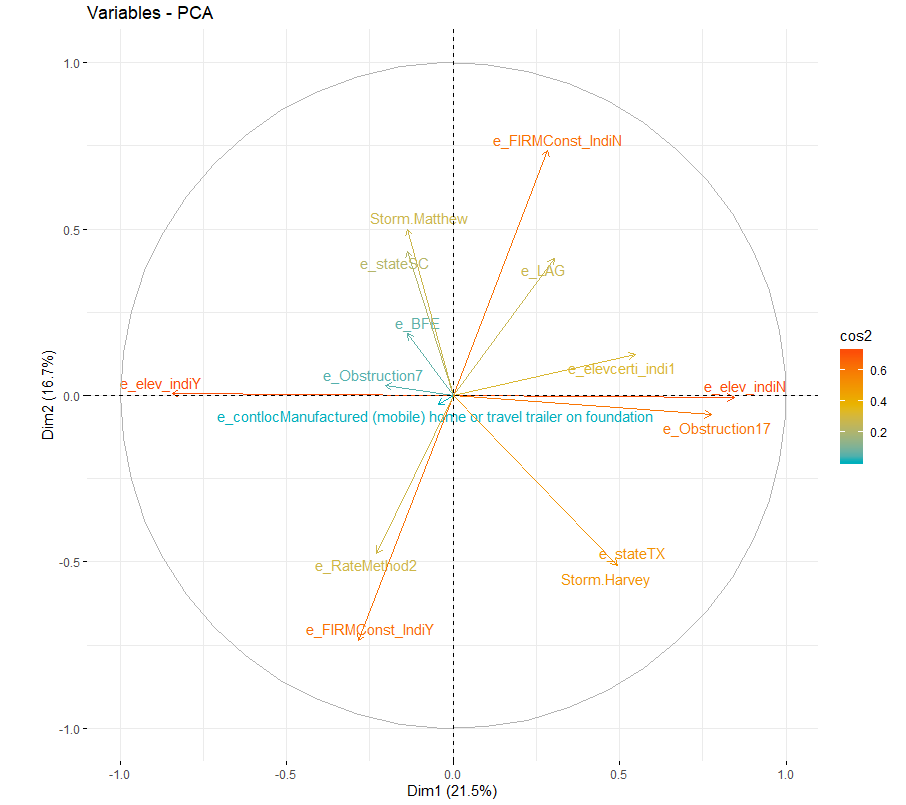


*Figure 22: Scree pLot*

*88% of the information(variance) is explained by first 8 principle components*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Dim.1 | Dim.2 | Dim.3 | Dim.4 | Dim.5 |
| e\_elev\_indiY | 0.71 | 0.00 | 0.19 | 0.03 | 0.00 |
| e\_elev\_indiN | 0.71 | 0.00 | 0.19 | 0.03 | 0.00 |
| e\_stateTX | 0.24 | 0.26 | 0.19 | 0.26 | 0.01 |
| Storm.Harvey | 0.24 | 0.26 | 0.19 | 0.26 | 0.01 |
| e\_LAG | 0.09 | 0.17 | 0.10 | 0.08 | 0.19 |
| e\_FIRMConst\_IndiN | 0.08 | 0.54 | 0.19 | 0.01 | 0.12 |
| e\_FIRMConst\_IndiY | 0.08 | 0.54 | 0.19 | 0.01 | 0.12 |
| e\_RateMethod2 | 0.05 | 0.23 | 0.12 | 0.00 | 0.32 |
| e\_stateSC | 0.02 | 0.19 | 0.23 | 0.31 | 0.00 |
| Storm.Matthew | 0.02 | 0.25 | 0.25 | 0.13 | 0.00 |

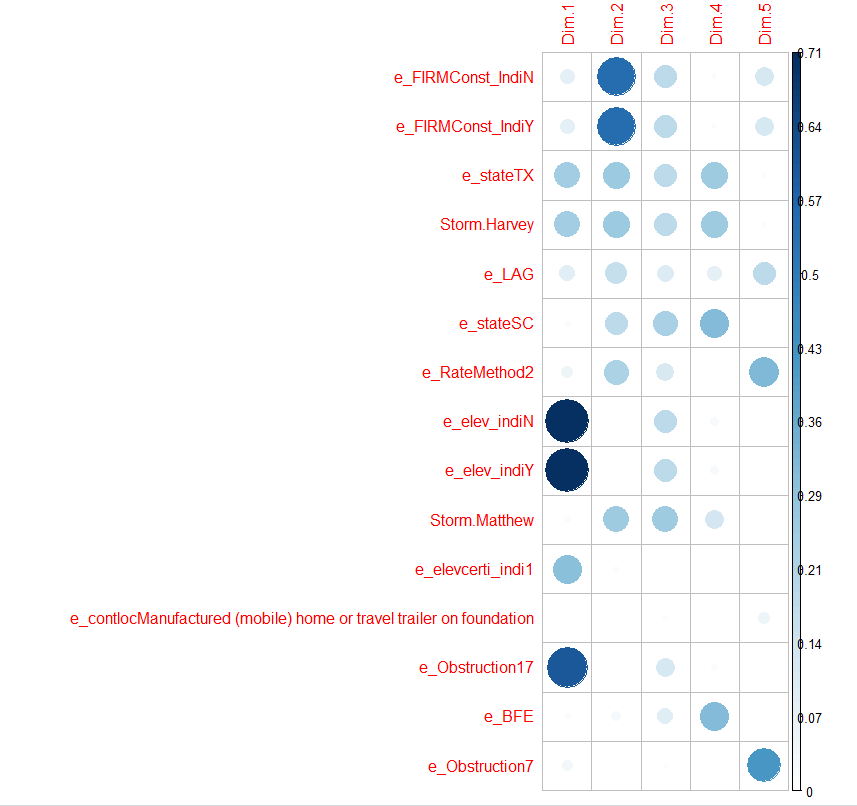
Table 1: Factor Loading table shows that PC1(dim 1) load heavily of elevation indicator



*Figure 23: Correlation Circle*

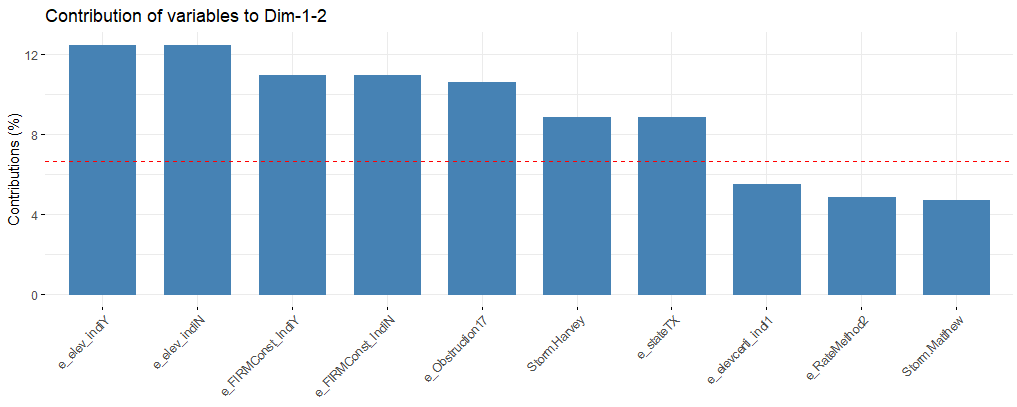
Shows the relationship between the variables

* Positively correlated variables are grouped together.
* Negatively correlated variables are positioned on opposite sides of the plot origin (opposed quadrants).
* The distance between variables and the origin measures the quality of the variables on the factor map. Variables that are away from the origin are well represented on the factor map



*Figure 24: Bar graph of correlation*

Contribution of variable to individual Dimension



*Figure 25: Combined contribution of variable to dimension 1 and 2*