**Data 670 Data Analytics**

***Ameha Eshete***

**Professor Dr. Steve Knode**

**Federal Emergency Management Agency (FEMA)**

**National Flood Insurance Program (NFIP)**

**Premium and Claims Information Analysis**

**October 10th, 2017**

**Executive Summary**

This project is prepared by using the National Flood Insurance Program (NFIP) public dataset obtained from BureauNet website reports for fiscal year 1978 through fiscal year 2017; organized as “claim information” and “Policy information” by state and county. This project is aimed at investigating the correlation and patterns of different attributes that may contribute to claims of total losses submitted based on geographical location and total amount paid comparison on losses for states and counties. Additionally, Policy comparison will be conducted based on the geographical locations and other external factors. The result of analytical model of this project will help to recommend premium rate adjustment and policy classification based on high and low claim losses, as well as helping management in staff and resource allocation and distributions. Fundamentally, the task of this research project requires converting semi-structured datasets to two structure datasets by state and county using R script and then merge the two datasets to final dataset using the common attribute variables. Once the dataset merging and combination in R is completed and become functional the research will explore on discovering and identifying trends and insights thorough various statistical methods. And then, the predictive model analysis will be conducted using R studio along with Tableau for graphical interpretation.

**Table of Contents**

[Project Scope (What problem are you trying to solve?) 4](#_Toc451983127)

[Data Set Description 4](#_Toc451983128)

[Project Insights of Your Data Analysis (Why are you doing this?)](#_Toc451983129) 5

[Project Milestones](#_Toc451983130) 6

[Restate Problem in Data Analytics Terms 7](#_Toc451983131)

[Forecast Outcomes 8](#_Toc451983132)

[Data Compilation Plan 8](#_Toc451983133)

[Data Analysis Methods. 9](#_Toc451983134)

[Data Analysis Strategy. 10](#_Toc451983135)

[High-Level Data Description. 11](#_Toc451983136)

[Data Definition. 11](#_Toc451983137)

[Identify Relationships between Variables. 12](#_Toc451983138)

[Data Profile.](#_Toc451983139) 13

[Data Preparation Requirements.](#_Toc451983140) 14

[Data Execution Plan](#_Toc451983141) 16

[Describe Tools & Applications.](#_Toc451983142) 16

[Describe Descriptive Statistics Used.](#_Toc451983143) 17

[Descriptive Statistics.](#_Toc451983144) 17

[Describe Additional Statistics Used.](#_Toc451983145) 18

[Additional Statistics.](#_Toc451983146) 20

[Data Analytics Plan](#_Toc451983147) 21

[Describe Analytics Used.](#_Toc451983148) 22

[Analytics Results and Findings.](#_Toc451983149) 26

[Recommendations.](#_Toc451983152) 26

**Project Scope**

Ever since “Billion Dollar Betsy” a major category three (3) hurricane disaster hit the Louisiana coast on September 9th, 1965 congress created the National Flood Insurance (NFIP) program. In 1968 this program started to allow homeowners, renters and business owners to purchase insurance against flood losses. In fact, the NFIP program has been here for 30 years now, but technologically it has not been modernized and the program costs too much to operate. As such, this project using the mainframe focus transit dataset available for public review on Bureaunet website will illustrate, find new patterns and give insight on the FEMS/NFIP program team through:

* *Discovering new patterns by combining the Policy and Claims dataset. (BureauNet website).*
* *Determining which variables and attributes have greater influence on the target outcome (Total loss and Policy).*
* *Plotting and illustrating Low and High policy county and total payment based on geographical locations since 1978.*
* *Classifying and give analyzing certain states and counties based on the Total loss and Policy count data.*
* *Analyzing current premium based on actual risk and give suggestion on alternative strategies and methods suing illustration and predictive modeling.*

In turn, providing this analysis enables and gives the NFIP team the tool to provide guidance and strategic planning to set realistic insurance rating, premium enforcement and policy growth targets. Furthermore, data exploration, preparation, and analysis may be performed using one or more of the following analysis tools (Strivastava, 2013):

* *Data manipulation using MS Excel, MS word and Text file*
* *Cleansing, graphical illustration and Predictive modeling development - R studio*
* *Additional Graphic illustration - Tableau software.*

**Data Set Description**

The data used in this project comes with each state and county Claims and Policy information files dating back from 1978 through 2017. For the purpose of this project, the first step requires obtaining each of this analyze and structure them for further data preparation. The second step is to dump all the states Policy information into one text file and bind them using R code. The same applies to the Claims information. Finally, holding the state and county as a common variable will combine Policy and Claims information into one dataset. In general, data analysis will be conducted on each dataset individually as well as the combined dataset.

The first dataset will compromise of seven (7) variables of claims information of each state and county. That is - two (2) character and five (5) integer variables:

1. **State** – The name of the state.
2. **County** – The name of the county.
3. **Total losses** – All losses submitted regardless of the status.
4. **Closed losses** –Losses that have been paid.
5. **Open losses** – Losses that have not been paid in full.
6. **CWOP losses** – Losses that have been closed without payment.
7. **Total Payments** – Total amount paid on losses.

The second dataset compromises of five (5) variables of each states and counties of the Policy information and similarly it has two (2) characters and three (3) integer variables.

1. **State** – The name of the state.
2. **County** – The name of the county.
3. **Policies In Force** – Policies in force on the "as of" date of the report.
4. **Insurance In Force** – The coverage amount for policies in force.
5. **Written Premium In Force** – The premium paid for policies in force.

The last dataset will be a combination of both Policy information and Claims information. Merging these two datasets will result in many missing, zero values rows and changes on the row count; which will be explored, prepared, and analyzed through MS Excel, Tableau Software and R packages (Miner, Nesbet, & Elder, 2009). The predictive model will be built using the combined dataset using Total loss, Policy in Force and Total pay variables. These variables will be utilized through classification and clustering as well as an output for prediction modeling.

**Project Insights**

As discussed above, this project will analyze and find new ways to better define the NFIP process on Policy and Claims dataset obtained from the public FEMA BureauNet website. By using industries current best statistical and analytical tools; new patterns and insights will be presented using the semi-unstructured NFIP Policy and Claims datasets. Furthermore, predictive model analysis will be developed based on a few of the attributes in the combined dataset and it will be utilized using each states/counties historical data dating back fiscal year 1978. Many methods under classification techniques and clustering methods will be used to produce result by categorizing policy count and total payment into Low and High classes. And then, based on the model result of the prediction an insight on future premium rates and resource allocations will be suggested.

Besides utilizing the National Flood Insurance (NFIP) dataset itself external factors will be looked at to justify patterns and geographical locations (zonings) through graphic illustrations of concentrated policy counts and total payouts based on premium and total loss variables. Moreover, such pattern/discoveries help to describe statically difficult viewpoints, increases the NFIP overall data resource management’s efficiency, operation and certainly gives new understanding on a reliable vision plan.

**Project Milestones**

In the state of properly arranging the desire goal in short period of time; a project millstone is laid out. The list below describes the expected completion dates and description of project milestones.

1. **Project Selection.** (October 01-03, 2017). The problem was framed and the executive summary and project scope was prepared.
2. **Analytic Approach.** (October 04-08, 2017). Following obtaining of the dataset from BureauNet website; description and analysis as well as steps to pick and choose different datasets as well as consolidation using R studio to create a single dataset and predictive model will be outlined.
3. **Project Status Presentation**. (October 09, 2017). A presentation of the project scope and the analytical approach techniques to be used will be prepared for peer review and suggestions. The presentation will be conducted via video recording and it lasts between 15-20 minutes.
4. **Data compilation.** (October 23, 2017). A detailed chosen analytical approach and the description of the data content, merging and combination will be conducted. And then, formatting requirement, structuring, and necessary to prepare the data for analysis with MS Excel and R studio.
5. **Data understanding.**  (October 30, 2017). An initial result of discovered patterns and insights through statistical description and data visualization techniques will be provided. A pre-developed predictive analytical approach applications will be detailed.
6. **Analysis evaluation**. (November 14, 2017). The R analysis of the data and steps in developing the predictive analytical model which includes the test and training set results will be described.
7. **Final presentation.** (November 29, 2017). A second presentation of the project which takes about approximately 30 minute will be developed and conducted. In the presentation, describing the application of R techniques and the development of a predictive model based on the NFIP Premiums and Loss data as well as all the selection, cleansing, and analysis steps taken to get to the final point will be presented. Tableau base illustration and future recommendations will be in the final presentation.
8. **Final Report.**  (November 29, 2017). To justify the final analysis report of the NFIP flood Premium and Claim data from year 1978 – 2017 and the predictive model technique used, the final report will describe the output of the class will to be completed and the findings and recommendations will be laid out for future analysis.

# Re-state Problem in Data Analytics Terms

The federal government flood insurance program has helped homeowners, renters and many business owners. However, this program has been debt-ridden, have technological short coming and too costly for tax payers (Katz, D. June 22, 2017). Following the brief explanation of the project insight the dataset analysis; here are shown how the stated problem will be tackled and described using data analytics terms. The project explores two classification model techniques and one clustering method draw statistically processed conclusion of the raw data information. Through this predictive analysis using the initial sourced and the combine dataset; meaningful results and transformation will be developed. The exploring and displaying terms of the dataset are generated by studying recent and historical dataset information.

# Forecast Outcomes

Prediction on flash floods and other natural disaster saves lives. It also reduces the damages on economics loses and protect people from injuries and deaths. The recent scientific advancement has improved our capabilities of predicting hurricane and flash flood warning systems. However, it is continuing to be a challenge to mitigate the after math effect in a timely and less costly manner. These outcomes are driven based on the severity of the disaster and surely analysis is needed to identify patterns in the insurance claims and policy information’s. As a result, this project strives to highlight points that are significant factors in managing policy and premium rate in the operation of the National Flood Insurance Program (NFIP).

The stated above, this project plans to use the NFIP current and historic datasets to be able to show correlations between different variables and help to developed less risky policies and premium rates based on the analysis outcomes.

**Data Compilation Plan**

**Data Analysis Methods**

The data analysis methods used for this project are classification techniques and K-means clustering method. There are many classifier techniques that could be utilized. However, binary and binomial logistic regression will be extensively covered and other classifiers such as Tree methods, Random forest and the SVM (Support Vector machines) will be used to compare the model accuracy results.

Under the classification techniques binary logistic regression method is used to predict the “Low” and “High” policy frequency (concentration) per county. If the policy counts per county falls under ten (10) thousand; it’s categorized as “Low” and if it is greater than or equal to ten (10) thousand it’s categorized as “High”. In the same token for claims total payment variable; if the total paid out is less than ten (10) thousand it’s categorized as “Low” and if it is more than (10) thousand it’s categorized as “High” expenditure. In building the predictive model premium and total loss variables are used as an input variables, and both the primary variables are also utilized as an input and output to each other’s predictive model.

K-Means clustering method will be used to group the policy and total paid data instances based on common characteristics. A pre-determined or unknown cluster analysis will be used to identify the “Low” and “High” claims of total paid and number of policies categorized in to a few number of clusters based on premiums and geographical location. Through placement of this variables based on the analysis of similarities and differences the clustering method shows each instance belonging of one and only one determined sub grouped cluster.

### 

### Data Analysis Strategy

While the main goal of this project is to be able to predict policy frequency and claims payments concentration based on premium and total pay outs, it is also aimed at identifying how the total paid and the number of policies are clustered based on different categories under certain geographical locations. To meet these objective, first the semi-structured data sets named “Policy” and “Claims” were obtained from BureauNet website and using R Studio sorted by state and county. Then by combining the policy and claims dataset a third file named combined policy and claim dataset were created. The data description was provided for both files as well as for the combined dataset in order to give readers better understanding of the datasets information. All the data transformation, analysis and model development put into places are on the combined dataset and will be described and discussed in detail in the remaining sections.

### High-Level Data Description

The data consists of two data sets merged together into one. According to the BureauNet website, the first file contains about Claims information by each state and county. It states the amount of money paid in total based on total and partial losses in each state and county. The second file contains about policy information by each state and county. It states the count of policies in force, the coverage amount and the premium paid in force for each state and county. This dataset contains information from 1978 to present related to policies and claims that FEMA currently services and the reports are run periodically.

### Data Definition

Initially there are 2,785 instances for the Claims dataset, 2,990 instances for Polices dataset and 3,009 instances for the combined dataset. Once the subgrouping is completed and data exploration and cleansing is done, the final combined dataset total row count equals to 1,534. The combined dataset has four (4) new variables added to it. The first two variables are assigned based on the High and Low sub groups of polices and the other two are based on total payment. As explained above in the data analysis method sections each new variable will hold the Low and High categories of policies and total payment under the set categorical criteria’s. The table below describes each variables definition, datatypes, and the variables dataset source.

*Figure 1: Data Definitions Table – Combined dataset*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Data Set Source** | **Description** | **Variable Type** | **Variable Format** | **# of Variables** |
| **State** | Claim | The name of the state | I | Char | 1,534 |
| **County** | Claim | The name of the county | I | Char | 1,534 |
| **Closed losses** | Claim | All losses submitted regardless of the status | I | Num | 1,534 |
| **Open losses** | Claim | Losses that have not been paid in full | I | Num | 1,534 |
| **CWOP losses** | Claim | Losses that have been closed without payment | I | Num | 1,534 |
| **Total Payments** | Claim | Total amount paid on losses | D | Num | 1,534 |
| **Policies** | Policy | Policies in force on the "as of" date of the report | I | Num | 1,534 |
| **Insurance** | Policy | The coverage amount for policies in force | I | Num | 1,534 |
| **Premium** | Policy | The premium paid for policies in force | I | Num | 1,534 |
| **Payment\_gfx** | Combined | Labeled as between 0 - 10k or 10k -7.3B - based on payment | D | Char | 1,534 |
| **Policies\_gfx** | Combined | Labels as between 1 - 10k or 10k -400k - based on policy count | D | Char | 1,534 |
| **Payment\_gfx\_Class** | Combined | Low and High based payment fgx category | D | Factor | 1,534 |
| **Policies\_gfx\_Class** | Combined | Low and High based on Policy fgx category | D | Factor | 1,534 |

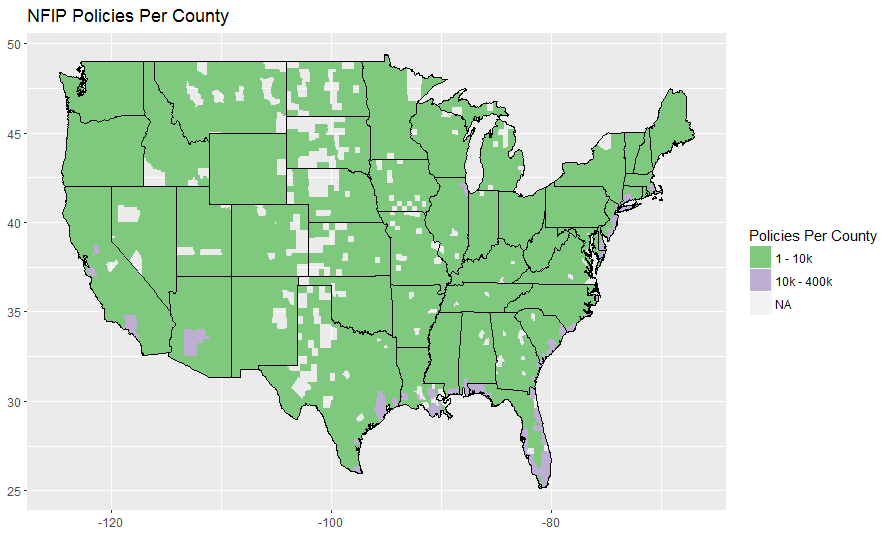
### 

### Identify Relationships between Variables

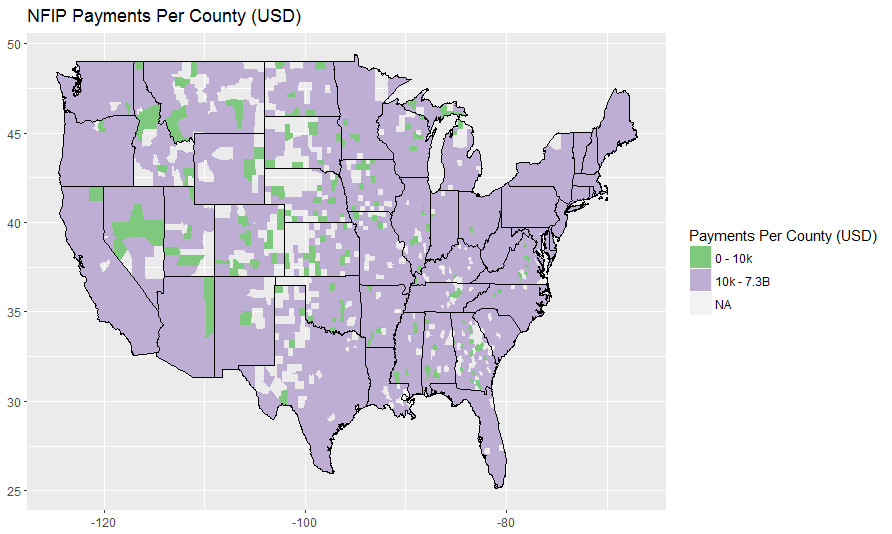
Several relationships between variables has been identified. A few graphical representations of the output variable are shown below. Remember, the aim here is to predict as “Low” when policy concentration is <10k and “High” when >=10k. Also total payment as 0–10k as “Low” and >= 10k as “High” based on other variables and geographical locations

The first bottom two maps show policy and total payment frequency per county using build in longitude and latitude map coordination using ggplot2 package in R studio. This map helps to analyze the correlations between policy concentrations and total payment distribution based on geographical area.

***Image 1:***

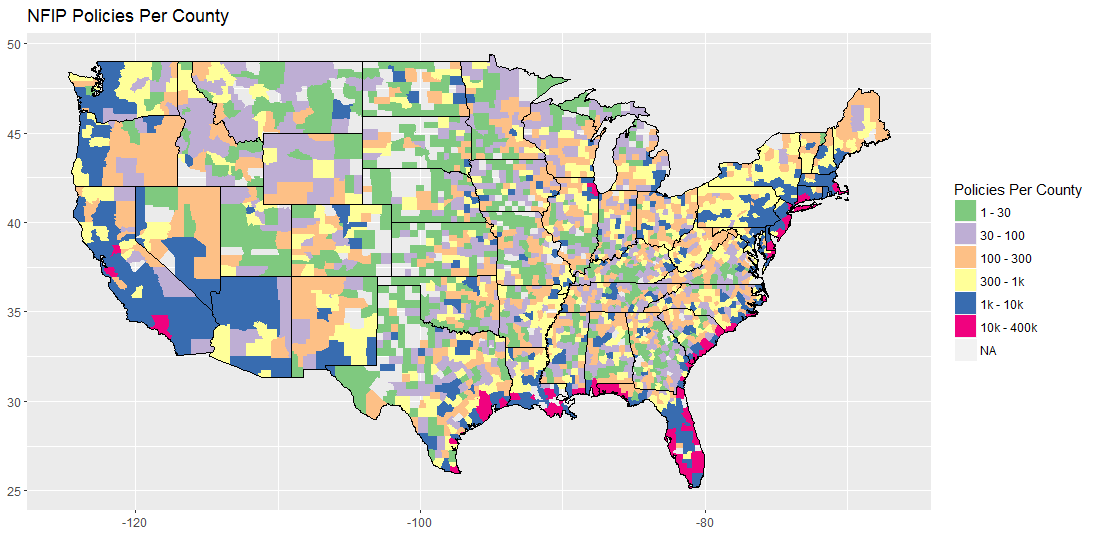


***Image 2***:

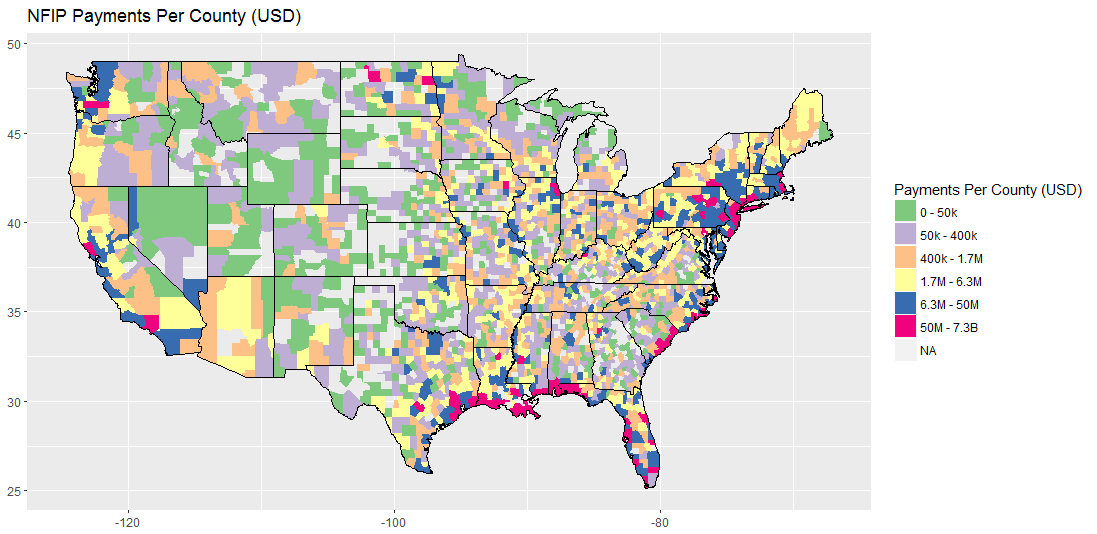


The second bottom two maps shown below classifies the policy per county and total payment per county in comprehensive way verses the above two maps. The above two maps emphasize the importance of identifying the Low and High within the 10k thresholds based on geographical locations.

***Image 3***:



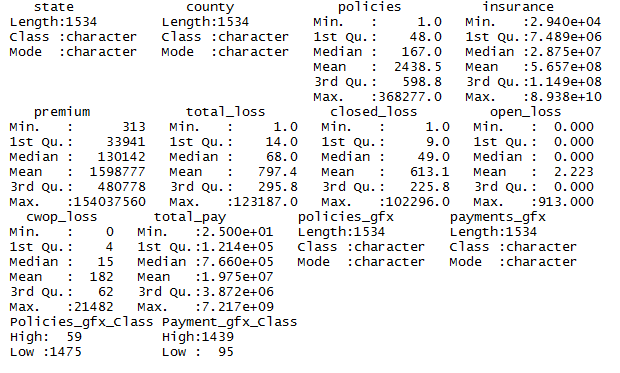
***Image 4***:



### Data Profile

The first figure below shows all the variables of the final dataset in one go, where all the basic statistical results produced. The output shows the number instances, the minimum, maximum values, mean, median and missing values which in this case already eliminated using the case function on R studio.

*Figure 2: Combined dataset basic statistics*

******

A few things to make some of the above statistics results clear: the value of the State variables count shows 1,534 instead of 50 because the rows are repeated so they could be aligned with the county. “Policies\_gfx” and “Payment\_gfx” variables are only used to create the High and Low subgrouping of policies and payment variables. Therefore, they will be eliminated before descriptive analysis and predictive modeling are conducted.

### Data Preparation Requirements

Data preparation is the core part of preparing a statistical predictive analysis project paper. MS excel and Tableau were greatly used to explore, clean and prepare the final dataset. However, in the beginning and the final stage of preparing the dataset; R studio were greatly utilized besides building the classification and clustering predictive model. Here are the steps taken to prepare the dataset:

* Explore – MS Excel, Tableau and R Studio
* Modify – MS Excel and R studio
* Model – R studio
* Assess – R studio

**Explore:** In the first stage R modification work started using the initial dataset by merging, subtracting and combining the Policy and Claims dataset. Then, the dataset was exported as a text file and migrated to excel for sorting, category validation and spot checking of the newly created subcategorized columns, and geographical locations.

**Modification:** Along with exploring and modifying; transforming the dataset to normalization stage, converting the output variable of Policy and Total Payment sub-group from characters datatype to factor and eliminate missing values as appropriate for the predictive models to take place.

**Model:** Select independent attributes to include in the models; mainly “premium” and “total loss” are used as an input but other attributes are used as needed. Split the data set into test and training. The training data sets has 304 observations and the test has 1,385 observations to build and test the predictive models.

**Assess:** Evaluate the accuracy of the models and revise initial model settings as needed. Consider different set of attributes and/or transformations of values to evaluate the impact on predictive model results and iterate predictive models in terms of independent attributes, their transformed and imputed values, types of models, and model settings

**Data Execution Plan**

This section explains the chosen tools and applications and the analytics and approach taken to solve the problem. Descriptive statistics is used in the context of steps to format, combine, and identify the semi-structured, structured /unstructured relevance of the problem. Explain data transformation, and adding of new variables, and data cleaning techniques, such as provisions for missing data, and other additional statistics and formatting requirements.

### Describe Tools & Applications.

There were three (3) tools and applications utilized for this project. The first is R studio which is used to sort and merge the initial dataset. MS Excel and Tableau were used to explore and help to cleanse and prepare a new sub grouped columns and eliminating outliers. Excel played a common role of moving the data between R and Tableau in preparing the final dataset and creating visualization.

R Studio were not only used to transform the initial dataset but to prepare, explore and also graph the attributes. The primary variables and output attributes scatter plot with geographical location was developed using R. Mainly, R was used to build models to analyze classification techniques and clustering methods. In addition to this model, it was also used to create a few other machine learning models to predict project successfully. Machine learning models included logistics regression, K-mean clustering, Tree methods and Random forest. Various functions and methods were developed to select the independent variables for predictive models.

MS Excel and Tableau was used to extract and transfer as well as visualize and characterize the data attribute values, relationships and data distributions. Excels easy data filter and sorting methods helped to explore the data row-by row and helped to transfer the base work for R modeling and Tableau visualizations to take place. On the other hand, Tableau software has an easy way of filtering and coloring the dataset to help identify patterns and relationships through enhanced visualization’s. Therefore; it was utilized to develop bar charts, and line graphs to compliment initial R findings.

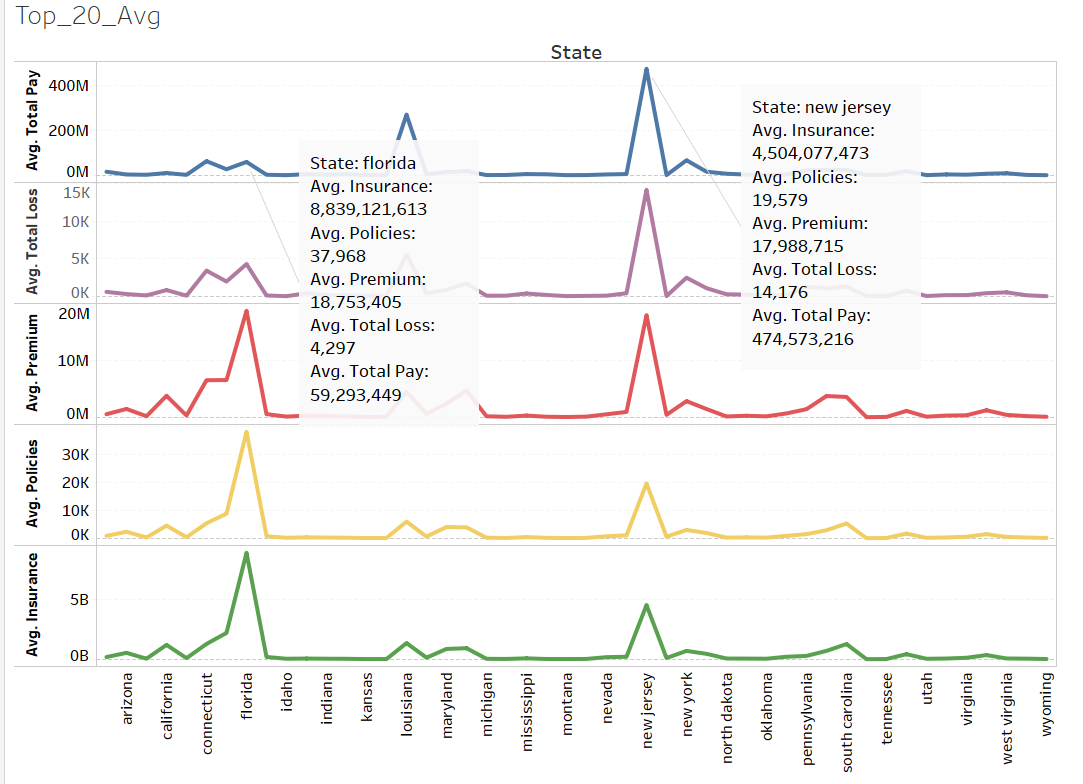
### Describe Descriptive Statistics Used

Given the vast information being generated related to the subject matter of NFIP (National Flood Insurance Program); I have taken two simpler approaches to produce the descriptive statistics. The first approach uses *measuring the central tendency* method to describe the central positon of frequency distribution for a set of data. The Highs and Lows of “policy” frequency and “payment” category and other variables along the median, mode and mean results. The second approach is *Measure of Spread* statistic method that summarizes the results in range, quartiles and deviation result. And the last part focuses on computing the information value (IV) or weight of evidence (WoE) result on categorical and continuous variables to help determine best variable selection for model building.

### Descriptive Statistics

In *measuring the central tendency* based on the summary of the final dataset statistical results: the median value of a “policy” count across the entire country stands at 136 with a minimum count of 2 policies in six (6) different states and a maximum count of 368,277 which is located in the state of Florida. Total pay has a minimum of zero (0) payout value, 552,712 median value, and a maximum value of 7.5 billion. A zero total pay values represent that there were a claim submitted but no payment has been granted due so many different reasons. Looking at the line graph below, the top twenty (20) states an average value of each variables; gives some insight to how some of the average insurance premiums fluctuations from state to state. Let us look at an example:

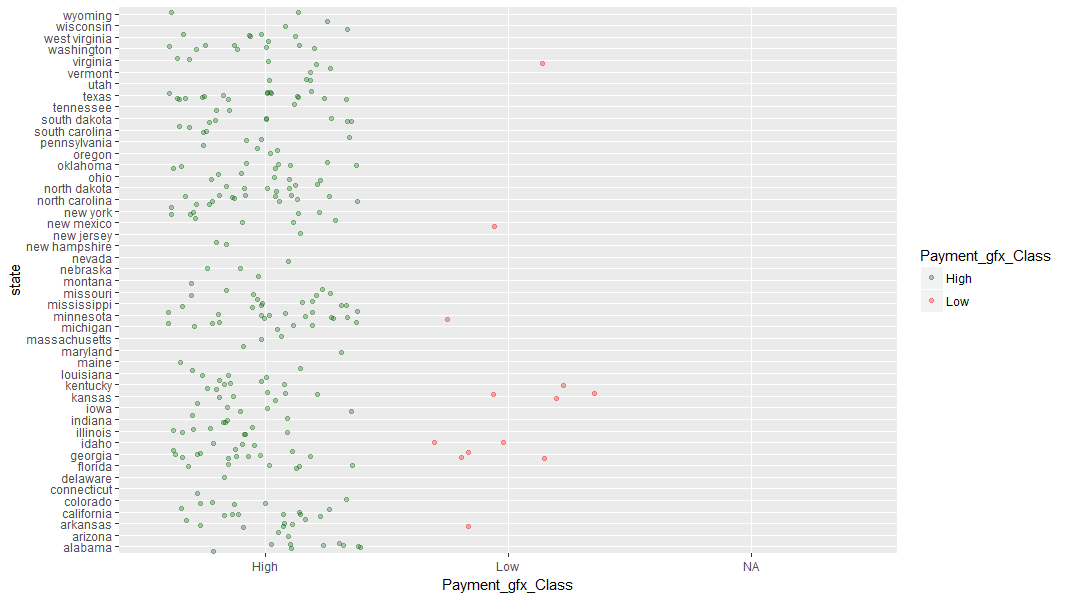
***Image 5: Top 20 average state results each variables***

******

Florida pays a monthly $38 ($462 yearly) flood insurance premium verses New Jersey $70 ($840). One would wonder how New Jersey has an expensive insurance premium rate than Florida. However, New Jersey has an average of 14,176 total loses verses Florida’s 4,297. There could be many variables why the premiums differ greatly from one state to another. Another example shows that Colorado pays 50% more than Florida but looking at the variables it is hard to justify the premium rate. Although, the amount of Polices sold in Florida greatly differs that of Colorado but the Total Loss in Florida tremendously exceeds Colorado. This is one area where NFIP has to do further actuary rate analysis to lower insurance risks and help discover the huge premium rate contracts.

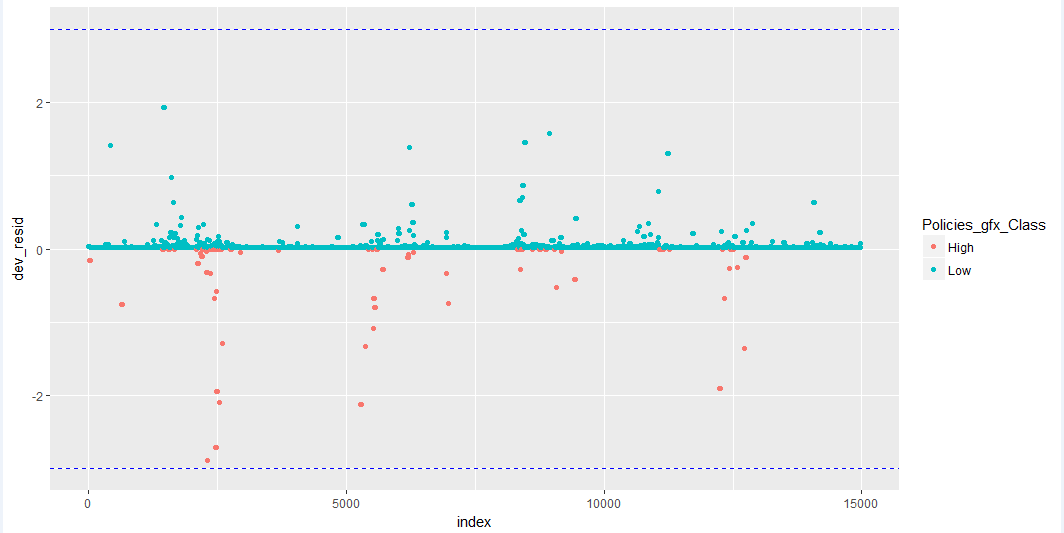
Using the *Measure of Spread* method we could easily find the range and quartiles statistic results by running a summary function in R studio. However, let’s look at outputs that shows a classification models portion of the “Payment” instances that belongs to the High and Low class value as based on the respective States. This result shows the number of instances belonging to each class and the percentage of output showing from the entire dataset. A High frequency group of 1,625 and Low frequency group of 59 which is 96.5% and 3.5% respectively distributed across the states.

***Image 6: Scoter plot of Payment sub category distribution along states***

******

Below, another distributed deviation index scatterplot indicating how much equally “policy” High category and Low category distributed along the deviation line. As we can see the High and Low of policy does not follow a straight line which indicated non-linear relationship but this inspection alone is not a sufficient indication of non-linearity. It shows that the fitted line does not seem to follow pattern observed across all points. While we see a significant p-value (very close to zero), the model generated does not yield a strong coefficient of determination.

***Image 7: Residual distribution of the policy category***



### Describe Additional Statistics Used

### N/A

### Additional Statistics.

In addition to the above descriptive statistics analysis this section shows the result conducted on finding an Information Value (IV) or Weight of Evidence (WoE) Score. The IV or WoE range could be anywhere between from 0% to 100%, where 0% is the worst possible combination of IVs and 100% is the perfect combination of IVs for that species. It helps to explore the data and screen the variables importance. In this case both continuous and categorical variables outputs have resulted in an all 0% scores. Several factors could be explained as a reason. First, the dataset has been subcategorized into High and Low group between <10k and >=10K which reduced the amount of data to work with. Second, when the dataset where split into test and training groups not only reduced the overall dataset count but resulted in less than 5% bin subcategory cases. All in all, the result supports the above deviation conclusion of a less coefficient determination.

***Image 8: Policies and Total Pay IV and WOE result***

|  |  |
| --- | --- |
| policies  VARS IV  1 Policies\_gfx\_Class 0  2 state 0  3 county 0  4 policies 0  5 insurance 0  6 premium 0  7 total\_loss 0  8 closed\_loss 0  9 open\_loss 0  10 cwop\_loss 0  11 total\_pay 0 | total pay  VARS IV  1 Payment\_gfx\_Class 0  2 state 0  3 county 0  4 policies 0  5 insurance 0  6 premium 0  7 total\_loss 0  8 closed\_loss 0  9 open\_loss 0  10 cwop\_loss 0  11 total\_pay 0 |

# Data Analytics Plan

### Describe Analytics Used

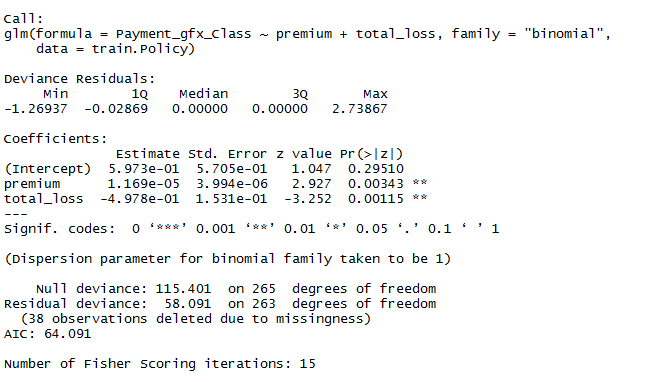
Various methods of predictive models have been tested using the final dataset. Beside the initial stage of major data exploration, combination, cleansing and preparation work the final dataset only requires picking out important variables, and splitting the dataset into training and test to produce meaningful results and help to discover new patterns. As a first choice the classification technique of logistic regression involving assigning categorical variables to a specifics class is the best choice to determine Total payment classes of less than (Low) or greater than (High) 10 thousand dollar. Support Vector Machine (SVM) classification will also be used to reassure and compare the classification model accuracy of the logistic regression model. Additionally, the K-mean clustering method will be utilized to see how “Polices” sub-category of the High and Low are clustered along each rows of all variables.

# Analytics Results and Findings

This section describes the prediction models that created and tested with different techniques and the analytical results. Starting with two classification techniques of logistic regression and random forest followed by K-mean clustering. For classification model the data was split into 70% training and 30% testing sets. Let us explore each model:

*Classification - Logistic Regression*

The first model we are looking at below is a prediction model of Total Payment sub category using Premium and Total Loss as an input variable. Multiple logistic regression result shows the coefficient value of Premium that has a maximum likelihood of probabilities to determine counties to receive a payment that are greater or less than 10k.

*Figure 3: Payment Logistic Regression model* ******

This means that the Premium variable is associated with an increase in probability of High and Low Payment. Statistically, a one-unit increase in Premium increases the log odds of High Total payment by 1.6 units. Total Loss is less significant in classification High and Low on the Payment class. This might be due to so many other factors like Insurance, Claims, Premium itself and geographical location along with Policy distribution. Below is a confusion matrix result of the first model.

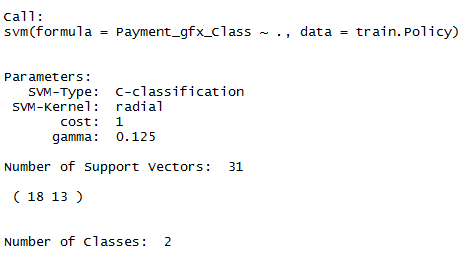
|  |  |  |
| --- | --- | --- |
| **Pred** | **High** | **Low** |
| <10000k | 260 | 79 |
| >=10000K | 854 | 0 |

There is a very interesting result here. Prediction on the test dataset shows a High payment of 854 cases verses 79 Low payment and the overall accuracy (how often the classifier is correct) is 21%; which is very low accuracy. This tells me that 10 thousand Payment variable sub category might not be the best choice of class or the amount of data to do this analysis is not sufficient.

*Classification - SVM*

As the second choice multi-class Support Vector Machine (SVM) model were developed. Given around 3500 distinct class and 305 separate observations it would be expensive iteration to run one-versus-one approach (OVO), I chose to develop one-verses-all (OVA) approach. Having to test the Tree method and Random forest models, SVM comes a little closer in balancing as true class but still challenging given the 1:3500 prevalence rate. Explore the final model selected after tuning:

*Figure 4: Payment SVM model*

**

Out of those 263 cases, the classifier predicted "High" 250 times, and "Low" 13” times. In reality, 250 flood cases will have received above 10 thousand dollars of total payment, and 13 locations do not.

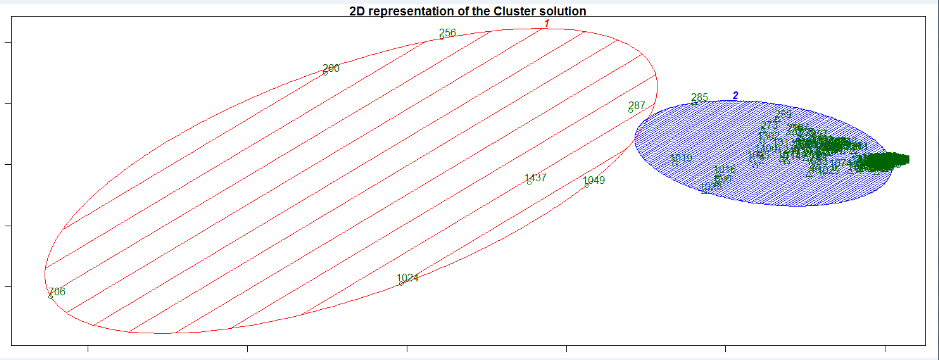
|  |  |  |
| --- | --- | --- |
| **Pred** | **High** | **Low** |
| High | 250 | 13 |
| Low | 0 | 0 |

As observed in the confusion matrix summary above SVM classification prediction over all accuracy stands at 95% and a misclassification rate of 0.08 which is about equivalent to 1 minus. Given the fact the original dataset had a few “Low” misclassified data, this was a direct performance result, relatively better in terms of giving a high predictions accuracy achievement compared with logistic regression model.

*K-mean Cluster Analysis*

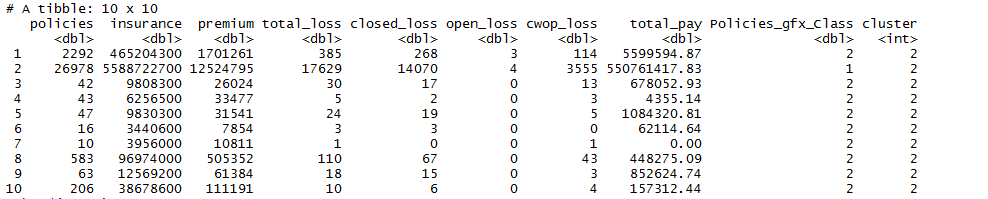
The last model in this project is K-means cluster. Using the Policy sub-category count of High (>=10k) and Low (<10k) as the resulting classifier it groups each variables instance into a placement in a cluster based on similarities and differences. Below shown a 2-part cluster configuration.

*Figure 5: 2D Cluster plot*



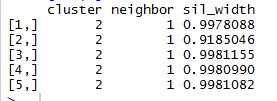
One of the advantage of using K-mean clustering algorithm is very easy to understand. The goal to assign and group each instance close to the center and separate the distance between the clusters itself. On the initial 2 part cluster configuration one of the cluster distribution are very wide hence difficult to show within feasible margin of R illustration. The table below shows the top ten (10) values of Policy clustered according to their developmental patterns.

*Figure 6: Cluster Rows – K = 2*



Let’s test the above separation distance between the resulting clusters. The Silhouette computation (function) shows how close each point in one cluster is to points in the neighboring clusters. This in general helps to assess the parameters of the cluster numbering; to choose optimal value for number of clusters. To find the Silhouette coefficients (near +1 indicator) we feed the cluster of each observation and the distance matrix to silhouette() and then look at the first few lines for result.

*Figure 7: Silhouette results*



The mean value of Silhouette result column equals to **0.995** whichindicates how good the clustering current case is.

### Recommendations

For this study, the relatively large data was not used. Also, considering the very wide spectrum of complicated flood incident variables and government restrictions; I believe the scope of the project could be widen and include many other factors into the analysis. Otherwise, beside the logistic regression analysis other type of SVM, Clustering or Neural Network methods could be tested and see if they are better in classifying policy risk and predicting Total Loss and Claim variables. For this study, the relatively wide verity of the variable clustering in few k values was a bit difficult.

### Conclusions

In conclusion besides geographically locating the sub-categorical Total Payments and Policy count results; cluster and classification models have not sufficiently resulted in identifying primary variables to easily discover policy/premium risks and new patterns. Although the goal could be achieved with additional essential variables added to the dataset and trial of different type of model development. In general this paper tried to show a result by subgrouping the Policy and Payment into two different categories and focused on important attributes like Premium and Total loss.

**References**

Gopi, Goteti. “RationShop/Nfip.” GitHub, © 2017 GitHub, Inc, 21 Oct. 2013,

github.com/RationShop/nfip.

Holladay, S. and Schwartz, J. (2017). Flooding the Market. [Online]

Policyintegrity.org. Available at: http://policyintegrity.org/documents/FloodingtheMarket.pdf [Accessed 4 Oct. 2017].

Katz, D. (June 22, 2017). The National Flood Insurance Program: Drowning in Debt and Due for

Phase-out. Retrieved October 29, 2017, from <http://www.heritage.org/government>

regulation/report/the-national-flood-insurance-program-drowning-debt-and-due-phaseout

Miner, G., Nisbet, R., & Elder, J. (2009). Handbook of Statistical Analysis and Data Mining. Academic Press.

National Flood Insurance Program. (n.d.). Retrieved October 04, 2017, from

<https://bsa.nfipstat.fema.gov/reports/data_definitions.html>

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2004). Selection of K in K-means clustering. DOI:10.1234/09544605X8298. Retrieved from <https://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf>

Sarma, K. S. (2007). Variable Selection and Transformation of Variables in SAS® Enterprise Miner™ 5.2. Retrieved from <http://www.lexjansen.com/nesug/nesug07/sa/sa17.pdf>

Srivastava, T. (2013). Extracting right variables for your Regression model. Retrieved November 29, 2014, from <http://www.analyticsvidhya.com/blog/2013/12/extracting-variables-regressionmodel-2/>