HOMEOWNERS INSURANCE

Final Project

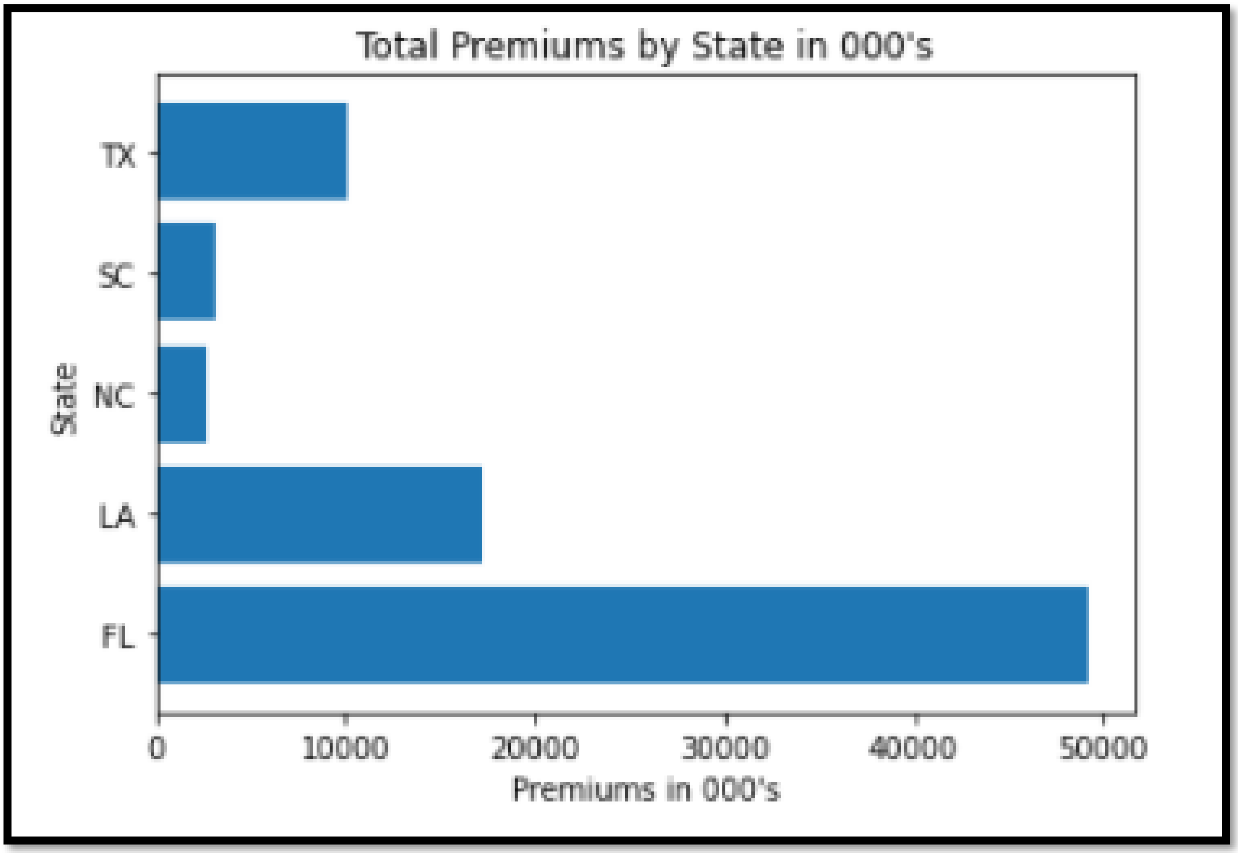
Abed Tabbalat

DSC550

I have been working in the homeowner's insurance industry for 7 years and it has been my career path for a good amount of time. I am sure everyone is aware of how the industry has changed when it comes to renewing a policy and how the prices have been going up. The past 3 years, since the pandemic started, a huge impact on the insurance market took place. In addition, with global warming and the increased frequency of weather events, it hasn't been easy on the insured nor the insurer. I am choosing this topic because it will assist me in my career path and attempt doing premium price predictions based on how variables change.

Since I am part of the Finance department, cutting costs is always the main target for us. What a better way to stop relying on external parties and pay fees to give us premium rate indications. Instead, having an inhouse model that could predict premium prices and what they should be based on historical data.

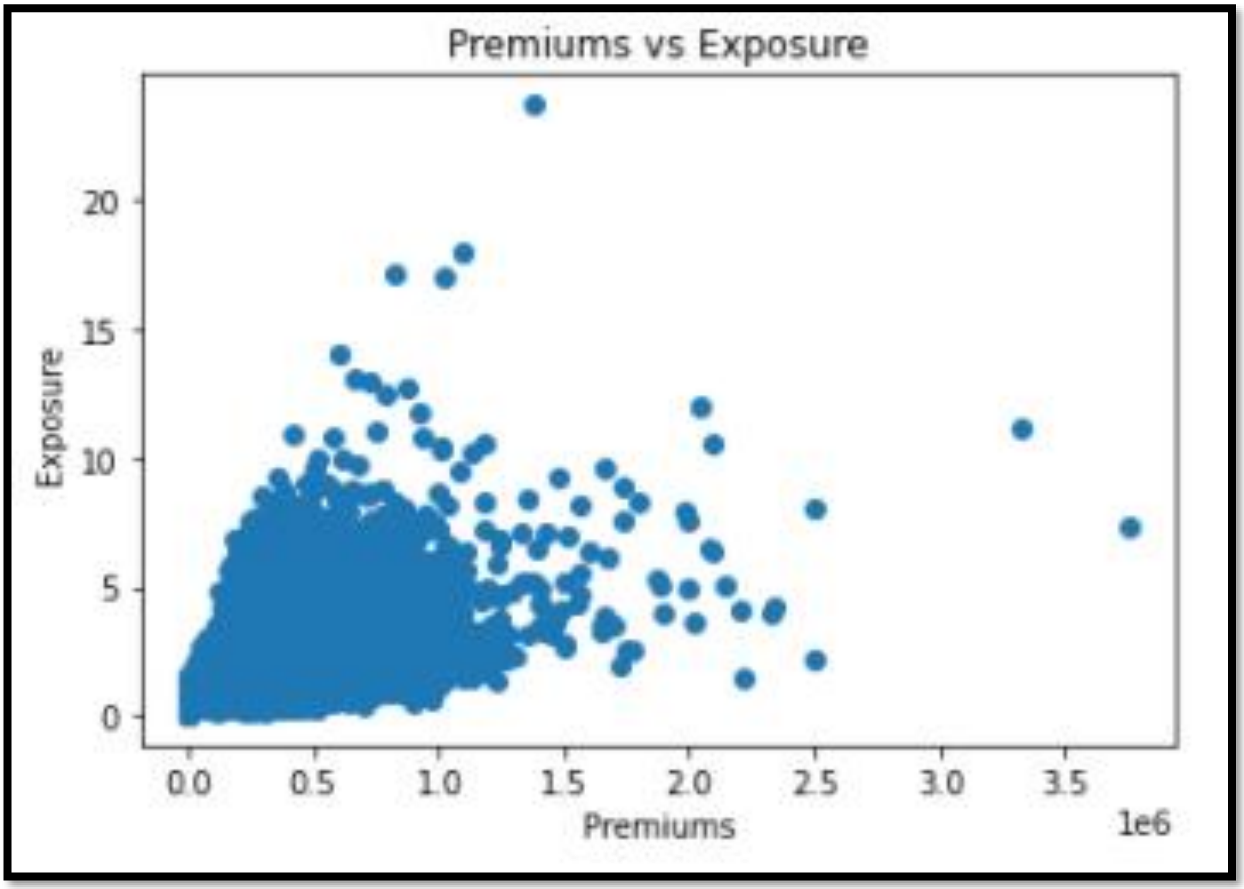
The dataset that I have received consists of a bunch of variables that describe each policy. The company writes in multiple states, but I have chosen to focus on the state of FL since it is the most concentrated for the company. Figure below shows the total premiums by each state:



As the figure shows, the company focuses on FL which has the highest concentration of premiums.

So, what is the problem that we are trying to solve? Predicting premium prices. What impacts this? Multiple variables. The dataset provided contains a series of variables but has been simplified for this exercise. That said, I would expect the model accuracy output to not be where it needs to be.

The way to tackle this exercise is to find how each variable correlates with the premium price. What I predict is the reinsurance cost to be the biggest driver of the premium price. The reason for that is because the reinsurance treaties are usually the highest expense in any insurance companies. Reinsurance is placing an insurance policy on the insurance company policies that gets activated when a certain criterion is met. For example, CAT (catastrophe) treaty, activates when a natural disaster event occurs, for FL, it is usually hurricane and thunderstorms. The figure below shows the premium prices vs the exposure (reinsurance cost) within our dataset:



The graph above clearly shows a positive correlation shaped as a cone between the premium price and the reinsurance cost, but we can confirm it after the test is run.

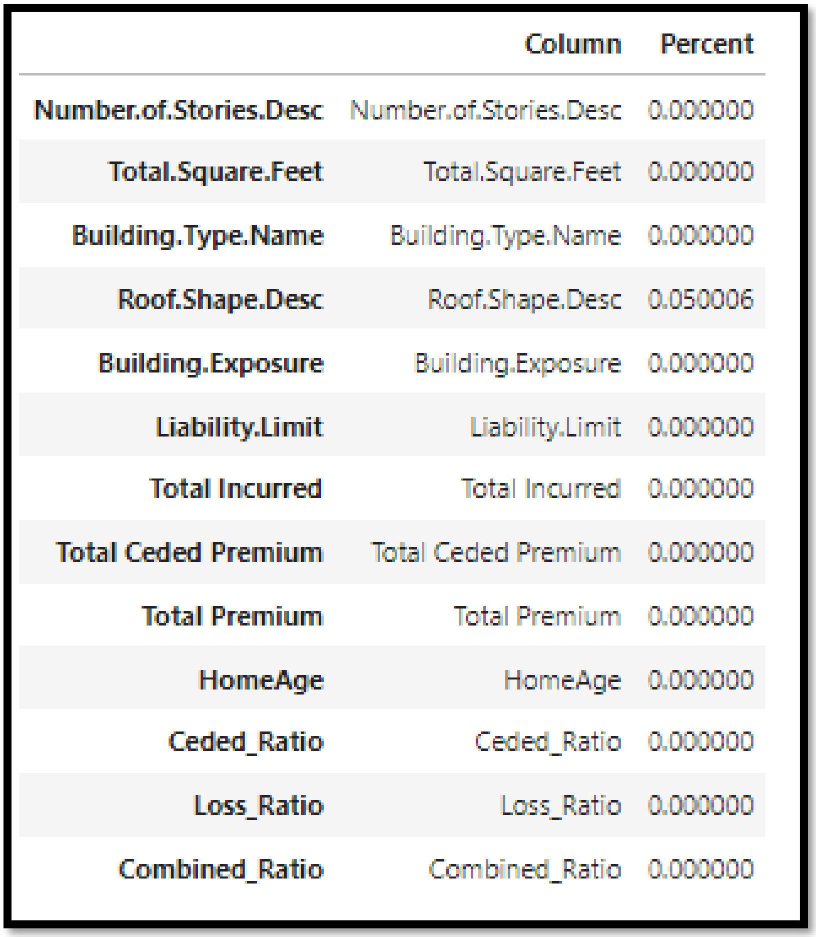
Once the correlation test is done, a linear regression model can be used to predict what premium prices should be based on all the variables. The variables in the dataset are the following:

- **WPolicyNum** → Policy Number
    - Will not be used in the model
- **Company** → There are 2 subsidiaries within the company
    - Will not be used in the model
- **Product** → Identifies the product type
    - Since the dataset is all about homeowners, this field is unnecessary
- **Territory** → territory factor
    - Will not be used in the model
- **Property.State** → Shows what state the policy is written in
    - Will be used to filter for FL only policies
- **Number.of.Stories.Desc** → How many stories does the home contain
    - Will be used as a variable in the regression model
- **Total.Square.Feet** → Home square footage
    - Will be used as a variable in the regression model
- **Year.Build** → Shows the year the home was built in
    - Will be used to create a numeric variable that calculates the age of the home based off the year
- **Building.Type.Name** → Shows the type of the building
    - Will be used as a variable in the regression model
- **Roof.Shape.Desc** → Roof type
    - Will be used as a variable in the regression model
- **Building.Exposure** → Exposure metric to the home that determines the risk on having a claim
    - Will be used as a variable in the regression model
- **Liability.Limit** → Cap dollars on the amount paid on liability
    - Will be used as a variable in the regression model
- **Total Incurred** → Total amount paid on claims filed for a specific policy
    - Will be used as a variable in the regression model
- **Total Ceded Premium** → Total charge to the reinsurer for a specific policy
    - Will be used as a variable in the regression model

The first step taken in data preparation for the model is filtering for the state of FL. Once that has been completed a few additional variables were created, these variables are all calculations based on existing variables. These variables are the following:

- **HomeAge** → Subtracts current year (2022) from the Year.Built variable to obtain a numeric age factor
- **Ceded_Ratio** → Divides Total Ceded Premium over Total Premium to obtain the ratio
- **Loss_Ratio** → Divides Total Incurred over Total Premium to obtain the ratio
- **Combined_Ratio** → Adds Ceded_Ratio and Loss_ratio together to see the overall ratio

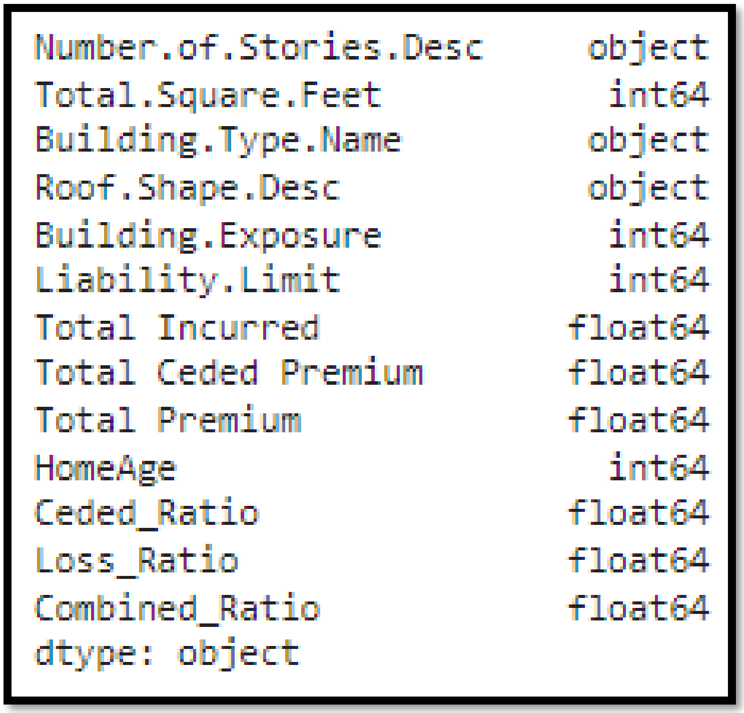
Once all the new variables are added to the dataset, performing an exercise to drop all unwanted variables would come to place. Now the dataset's external shape is complete. From this point we perform a check on the quality of the data. Checking for NaNs is necessary. However, depending on the result of how much missing data we have, different actions would occur. Below shows the results of percentage missing values within the dataset:



	Column	Percent
Number.of.Stories.Desc	Number.of.Stories.Desc	0.000000
Total.Square.Feet	Total.Square.Feet	0.000000
Building.Type.Name	Building.Type.Name	0.000000
Roof.Shape.Desc	Roof.Shape.Desc	0.050006
Building.Exposure	Building.Exposure	0.000000
Liability.Limit	Liability.Limit	0.000000
Total Incurred	Total Incurred	0.000000
Total Ceded Premium	Total Ceded Premium	0.000000
Total Premium	Total Premium	0.000000
HomeAge	HomeAge	0.000000
Ceded_Ratio	Ceded_Ratio	0.000000
Loss_Ratio	Loss_Ratio	0.000000
Combined_Ratio	Combined_Ratio	0.000000

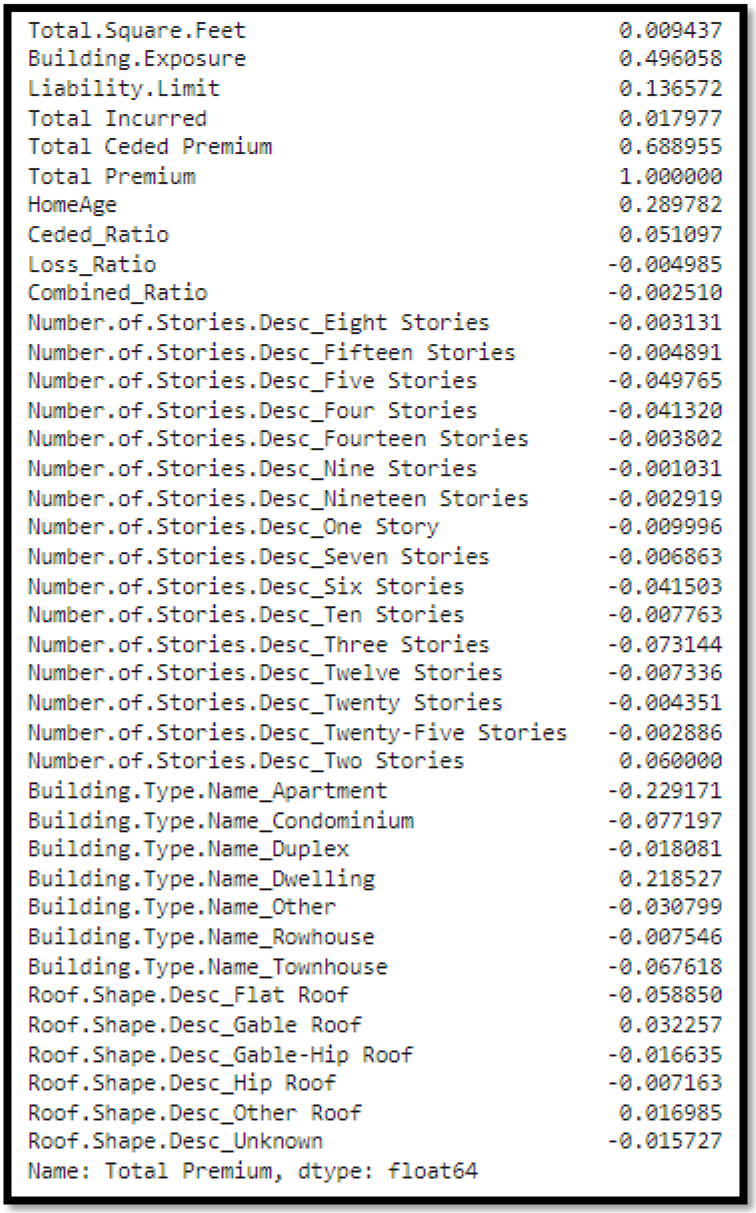
Only 5% of the data is missing from the Roof.Shape.Desc column and the rest of the variables have no missing data. Since our total rows are over 40K, it would be a good call just to drop all the rows containing missing data from the dataset.

The final data wrangling exercise would be applying the dummy function to all object variables to convert the metrics to numeric. Below shows the list of all the variables that will be used in the model and their data types, any column with 'object' will be converted to numeric using the dummy function:



```
Number.of.Stories.Desc     object
Total.Square.Feet           int64
Building.Type.Name         object
Roof.Shape.Desc            object
Building.Exposure           int64
Liability.Limit             int64
Total Incurred            float64
Total Ceded Premium       float64
Total Premium             float64
HomeAge                     int64
Ceded_Ratio               float64
Loss_Ratio                float64
Combined_Ratio            float64
dtype: object
```

The data is ready for modeling, and the first step towards our regression is applying a correlation test of all the variables to the total premiums, below shows the results:



```
Total.Square.Feet                              0.009437
Building.Exposure                              0.496058
Liability.Limit                                0.136572
Total Incurred                                 0.017977
Total Ceded Premium                            0.688955
Total Premium                                  1.000000
HomeAge                                        0.289782
Ceded_Ratio                                    0.051097
Loss_Ratio                                    -0.004985
Combined_Ratio                                -0.002510
Number.of.Stories.Desc_Eight Stories          -0.003131
Number.of.Stories.Desc_Fifteen Stories        -0.004891
Number.of.Stories.Desc_Five Stories           -0.049765
Number.of.Stories.Desc_Four Stories           -0.041320
Number.of.Stories.Desc_Fourteen Stories       -0.003802
Number.of.Stories.Desc_Nine Stories           -0.001031
Number.of.Stories.Desc_Nineteen Stories       -0.002919
Number.of.Stories.Desc_One Story              -0.009996
Number.of.Stories.Desc_Seven Stories          -0.006863
Number.of.Stories.Desc_Six Stories            -0.041503
Number.of.Stories.Desc_Ten Stories            -0.007763
Number.of.Stories.Desc_Three Stories          -0.073144
Number.of.Stories.Desc_Twelve Stories         -0.007336
Number.of.Stories.Desc_Twenty Stories         -0.004351
Number.of.Stories.Desc_Twenty-Five Stories    -0.002886
Number.of.Stories.Desc_Two Stories             0.060000
Building.Type.Name_Apartment                  -0.229171
Building.Type.Name_Condominium                -0.077197
Building.Type.Name_Duplex                     -0.018081
Building.Type.Name_Dwelling                    0.218527
Building.Type.Name_Other                      -0.030799
Building.Type.Name_Rowhouse                   -0.007546
Building.Type.Name_Townhouse                  -0.067618
Roof.Shape.Desc_Flat Roof                     -0.058850
Roof.Shape.Desc_Gable Roof                     0.032257
Roof.Shape.Desc_Gable-Hip Roof                -0.016635
Roof.Shape.Desc_Hip Roof                      -0.007163
Roof.Shape.Desc_Other Roof                     0.016985
Roof.Shape.Desc_Unknown                       -0.015727
Name: Total Premium, dtype: float64
```

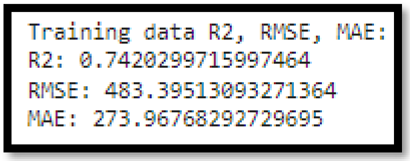
The top three highest correlation results are the following:

- Ceded Premiums (Positive correlation of 0.689)
- Building Exposure (Positive correlation of 0.496)
- Home Age (Positive correlation of 0.289)

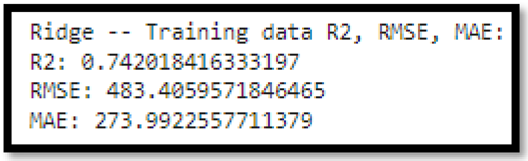
The results of the correlation test confirm the predictions that were mentioned earlier. The reinsurance cost (ceded premiums) is the biggest driver towards premiums.

The data will be split in training and test set, since the pool of rows are large, the split will be 50% between the training and test dataset. Once that is complete, the regression model will be applied to the training dataset for the model to learn, and then will be tested through the dataset and compare the results with the actual premium amounts. The model used will be Linear Regression and applied to Ridge Regression model. If the results are close, then we can know the factors affecting the results are the variables.

Below shows the results of the linear regression model and the ridge regression model:



```
Training data R2, RMSE, MAE:
R2: 0.7420299715997464
RMSE: 483.39513093271364
MAE: 273.96768292729695
```



```
Ridge -- Training data R2, RMSE, MAE:
R2: 0.742018416333197
RMSE: 483.4059571846465
MAE: 273.9922557711379
```

The model came out to be 74% accurate and the RMSE is high in both scenarios. This means that to price premiums, more variables are needed to lower the RMSE down. The result makes sense because there are very important variables that are highly correlated with the premium price such as distance to coast. Therefore, the model cannot be deployed with the current data provided, as more fields are needed to be able to deploy the model.

To summarize, the insurance world is highly complex, the model chosen would be a great fit if all the necessary variables are factored in. A simple correlation test can identify the impact of these variables to the premium price and once that is determined, a regression model can be highly significant and predict accurate prices to policies.