Term Project - Step 2

Tabbalat, Abed

2021-05-21

How to import and clean my data

As mentioned, I will be working with insurance data. There are 3 datasets obtained that contain significant amount of columns and rows. The datasets are a combination of CSV and .XLSX (Excel). Importing CSV files can be done with base R and I will use the readxl library to import the excel files into R. Cleaning up the data will be tricky as the dataset is large. The data contains many variables and I have identified my unique identifier that is common between the 3 data sets. The common identifier will be the policy number.

The datasets will be imported with a "raw" tag in order to keep the original data the same so that if I need to circle back it. The datasets are called:

- `df_claim_raw`
- `df_score_raw`
- `df_exposure_raw`

`df_claim`

The policy ID is set up to add a zero after the last dash "-" in order to keep the string length consistent, however the other datasets do not have this so I will reformat the Policy number column by splitting them with the dash "-" as the delimiter and then removing the additional character that isn't needed and then concatenating them back into one column that will match the policy number characteristics in the other datasets.

In addition, I have noticed duplicate values that exists in the claims dataset that need to be summarized and this is where `library(dplyr)` is helpful by using the `group_by()` function and the `summarise()` function to condense them into unique values.

Finally, I have identified in the claims dataset, the only column required for the model will be the Incurred Loss column, all other components will not be needed and hence, I have reshaped the dataset to only show policy number and total incurred losses.

`df_score`

This dataset includes details on the reinsurance CAT premium regarding the policy holders. I will only need `Total Ceded Premium` column as it is the total value of all the components that are in the table. Therefore, I have eliminated the other variables and the dataset will only contain `Policy number` and `Total Ceded Premium`.

`df_exposure`

This dataset has been determined to be the master. The table contains the written premium detail by policy, and all the other variables that can be chosen to be used within the model. Therefore, using the `left_join()` function I will merge the columns that are needed from the other 2 sets into this one so that the analysis can happen all in one table.

What does the final data set look like?

`df_claim`

Table 1: Reformatted Claims Table

WPolicyNum	Total Incurred
09-0010101453-1-20	10688.11
09-0010101453-1-24	18942.59
09-0010103170-5-21	1641.66
09-0010103502-2-22	6675.00
09-0010103645-4-22	2165.18
09-0010103884-3-23	24931.25

`df_score`

Table 2: Reformatted CAT Score Table

WPolicyNum	Total Ceded Premium
09-0010140659-8-24	262.60400
09-0010142560-1-24	524.91360
09-0010143579-0-24	283.38689
09-0010144369-4-24	789.62456
09-0010144442-4-24	75.43641
09-0010144446-8-24	947.80824

`df_exposure`

This table is too large to be printed in a document.

Questions for future steps

- Did I miss any components from that I eliminated from the dataset cleanup?
- Will my choices be a good fit to the model?
- Would there be a different way to write my code where it looks cleaner?

What information is not self-evident?

I believe the insurance world is tricky as there are plenty of metrics involved and it can get very complicated and comprehensive. The main exercise that needs to be computed is calculating a total combined ratio for the products to determine what areas are causing the product to not be profitable. Of course, when it comes to claims, there are plenty of factors that happens such as liability, hurricane damages, and regular damages (like pipe bursts). Another non-self-evident area is how many variables are accounted for to determine what the premium for a specific policy is, the dataset shows over 50 variables, but do we really know if those were factored in the premium? More digging and asking question need to happen to determine this.

What are different ways you could look at this data?

The data is comprehensive enough that there are many angles to look at it. I am focusing on the product/state combination mostly to see if the general geographical area can be a factor in how the products are performing. The data can be looked at and base plenty of variables that are available such as county or year built as an example.

How do you plan to slice and dice the data?

I sliced the data into 3 components:

- Product/State combination
- Home construction type
- Product segmentation

How could you summarize your data to answer key questions?

After slicing and dicing the data, I will perform the combined ratio calculation that requires a fixed expense ratio of 30%, calculating the Incurred Loss ratio and the CAT ratio, the sum of those 3 components will result to give out the combined ratio. Below shows an example of what the dataset looks like by Product/State:

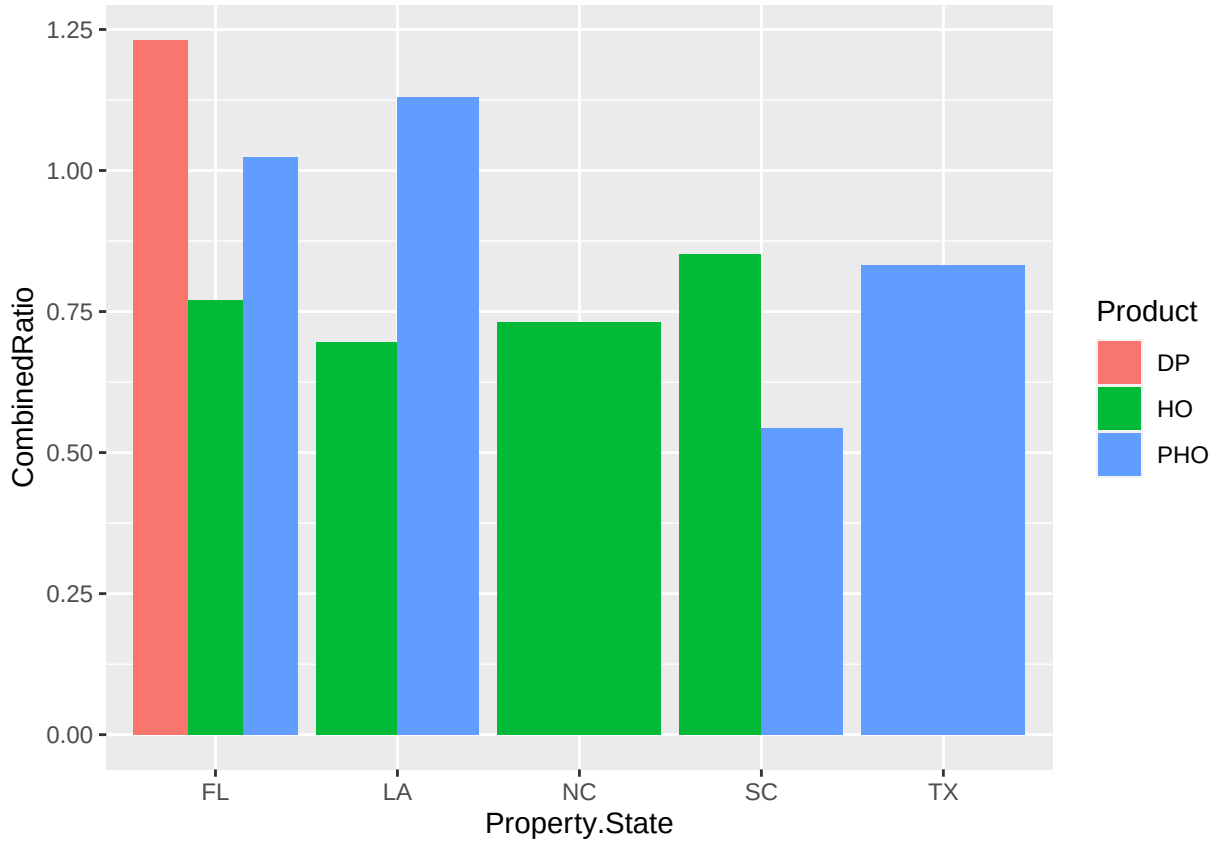
```
## `summarise()` has grouped output by 'Product'. You can override using the `.groups` argument.
```

Table 3: Summary Calculations By Product & State

Product	Property.State	Total Premium	Total Incurred	Total Ceded Premium	LossRatio	CATRatio	CombinedRatio
DP	FL	662476	204892.1	412264.01	0.3092823	0.6223078	1.2315901
HO	FL	18085866	3498277.0	5008000.92	0.1934260	0.2769014	0.7703274
HO	LA	7869043	865112.8	2254177.56	0.1099388	0.2864615	0.6964002
HO	NC	2618547	443266.0	685011.66	0.1692794	0.2615999	0.7308793
HO	SC	3116780	1291671.2	428250.32	0.4144249	0.1374015	0.8518264
PHO	FL	30507967	8722226.3	13344938.61	0.2859000	0.4374247	1.0233247
PHO	LA	9329294	2087277.9	5653167.72	0.2237337	0.6059588	1.1296925
PHO	SC	67492	0.0	16359.81	0.0000000	0.2423963	0.5423963
PHO	TX	10034275	2145257.1	3197426.04	0.2137929	0.3186504	0.8324434

What types of plots and tables will help you to illustrate the findings to your questions?

After trial and error, since we are comparing qualitative data to quantitative data, the best plots to present for this research would be bar charts. I will perform 3 different bar charts for each component, below shows the combined ratio to Product/State bar chart that is taken from the sample table above:



Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I will be performing a multiple linear regression model to determine if there is any relation between Loss Incurred and the written premium with the other variables in the exposure dataset using the R2 output. Then I will perform a multiple linear regression model to the CAT premium and determine the relationship between CAT premium and written premium with other variables in the exposure dataset. This will give me a good idea on how those 2 attributes impact the written premium.

Questions for future steps.

- Will this analysis give a final answer?
- Is the analysis clear enough for the reader?
- Where there any variables missed in the model?