Auto Insurance Analysis

Matthew Eng
Department of Computer Science
Rutgers University

OVERVIEW

For Assignment 2: Exploratory Data Analysis, we were instructed to perform a hypothesis test on a dataset of our choosing from Kaggle's extensive selection of datasets. After doing some research and scrolling through their collection of datasets, I ended up choosing one that I felt was relevant, recently published, and can be applied to real-world scenarios: "Insurance Dataset Based on Real-World Statistics" by Sami Alyasin. This dataset regarding auto insurance (https://www.kaggle.com/datasets/samialyasin/insurance-data-personal-auto-line-of-business) contains robust rows of age, marital status, claims frequency, credit score, region, and much more. With multiple upvotes and a usability score of 10, I was compelled to use and analyze this dataset.

NULL / ALTERNATIVE HYPOTHESIS

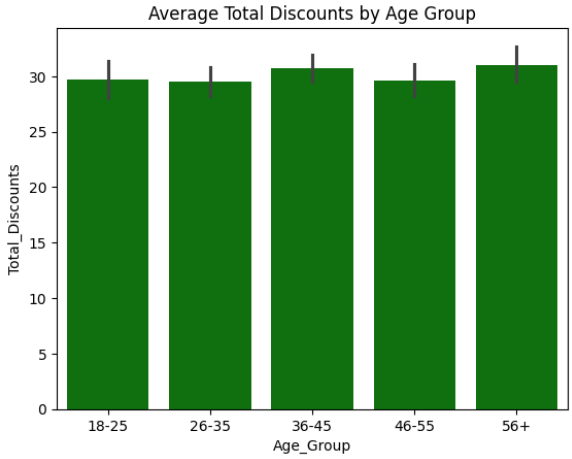
After spending some time doing some basic analysis of the data, I ended up establishing a null and alternative hypothesis that I felt wasn't too simple/obvious and could result in unique findings. For my *null hypothesis*, I claimed that the average frequency of insurance claims is the same for urban policyholders with a credit score >= 600 and rural policyholders with a credit score < 600. My *alternative hypothesis* then claimed that the average frequency of insurance claims is different for urban policyholders with a credit score >= 600 and rural policyholders with a credit score < 600. Some assumptions and notes that I made were that credit score >= 600 is considered fair/good while < 600 is considered poor, urban areas are considered dangerous and more prone to accidents compared to rural areas. I also chose to ignore suburban policyholders

and set the significance level to 0.05 (commonly used). The goal with this hypothesis was to see if a mix of good credit in unsafe areas and bad credit in safe areas led to significant differences in insurance claims. Basically, I was trying to figure out if credit played a more important role in the number of insurance claims that a policyholder makes over region and vise versa. This could then be expanded to see if urban areas are often dangerous due to policyholders with low credit scores or just due to their being lots of vehicles in small areas of space.

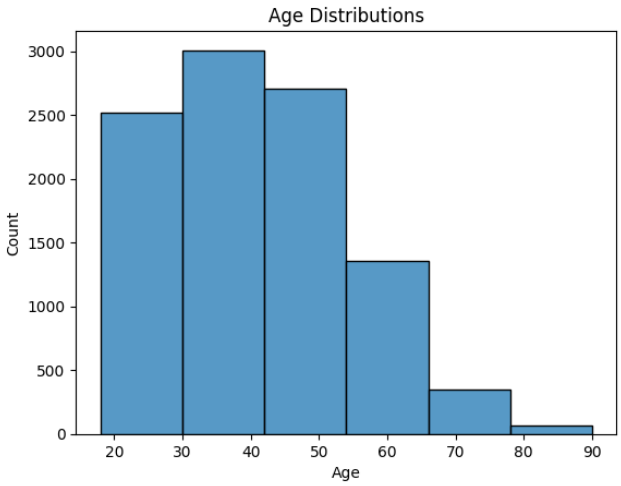
PERMUTATION TEST

For this assignment, I decided to use a permutation test to decide whether we could reject the null hypothesis as permutation tests are generally safer and don't make assumptions regarding mean and variance. I began by filtering the dataset to focus only on the frequency of claims, credit scores, and regions of the policyholders. I then divided the dataset into two groups that were previously described in the hypothesis. After isolating the frequency of claims and generating a difference in means function, I utilized the permutation test function from the scipy.stats Python library. What resulted was an observed difference/statistic of 0.1480 and a p-value of 0.45. Due to the p-value being less than the significance level (0.05), there was no evidence that allowed us to reject the null hypothesis. This meant that there was no difference in the average frequency of claims between urban policyholders with high credit and rural policyholders with low credit.

VISUALIZATIONS

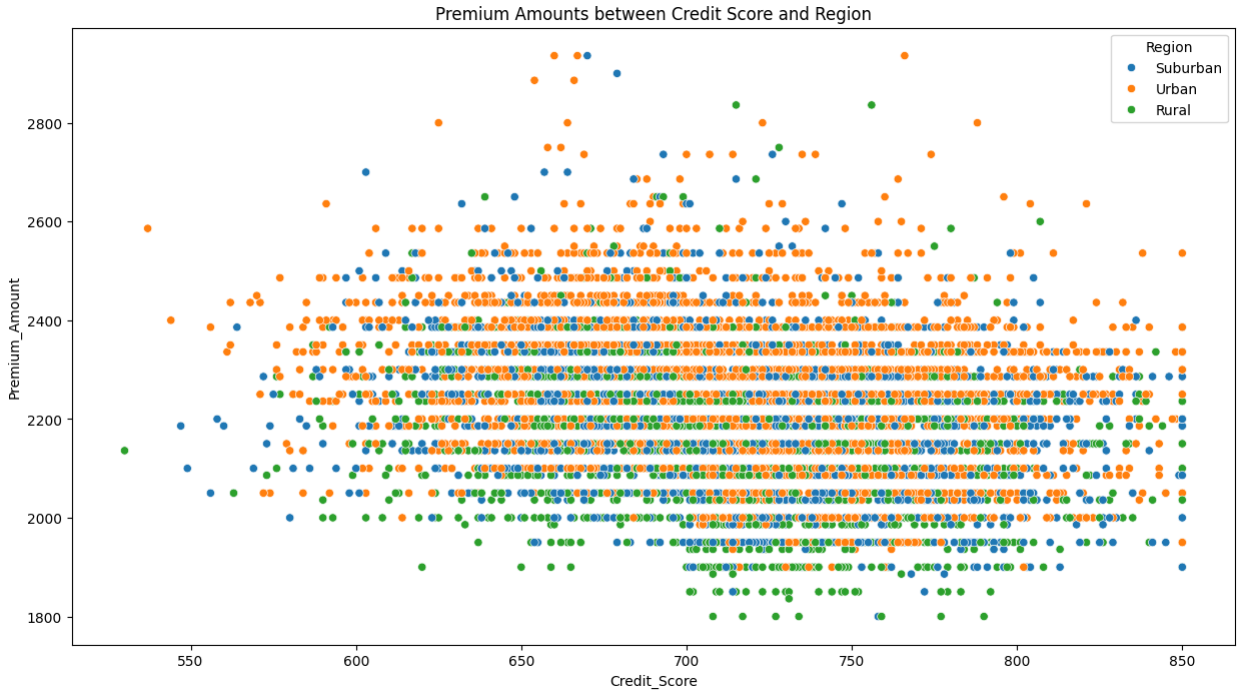


This first visualization is a bar plot that compares the average total discounts between different age groups. I formed groups for the different ages to increase readability. With this visualization, I'm claiming that there is no significant difference in average total discounts between different age groups. Meaning, all ages receive similar average total discounts. As the graph shows, each group is at about 30 in terms of discounts. So, age won't really factor into the number of total discounts that you receive as a policyholder.



This second visualization is a histogram that displays the distribution of the different age groups. My claim for this visualization is that there aren't enough seniors (especially 70+) in this dataset

to make accurate claims that involve age. Now this immediately ties back to my first claim about total discounts. Even though I said all age groups receive similar number of discounts, this could be false if there were more seniors in this dataset. As the visualization shows, there aren't even 500 seniors that fall into the 65-ish to 75-ish age group. This visualization also shows that there were about 3000 that fall into the 30 to 40 age group. While there are statistically less seniors alive, there should still be more data on them to be able to make conclusive arguments that involve age groups.



My final visualization compares premium amounts between different credit scores and regions. My claim with this visualization is that premiums amounts increase for policyholders in heavily populated areas. As can be seen in the scatterplot, most of the rural policyholders occupy the bottom half of the plot in terms of premium amounts. They pay smaller premium amounts compared to those in suburban or urban areas that pay more expensive premium amounts. This can especially be seen in the "outliers" towards the top that mainly consist of non-rural

policyholders and end up paying the most out of everyone. The plot also shows that there is no significant difference of premium amounts between different credit scores. This means that credit score doesn't really factor into the amount of premiums that you'll end up paying. Overall, all of these visualizations offer different insights and support different claims and contribute to the process of data analysis.