Predicting Health Insurance Costs in the United States Using Patient Medical History

Zara Nip and Nehal Linganur

05/25/2023

1. Project Overview.

Which factors impact an individual's health insurance premium the most? Which factors are statistically significant? Given a sample person's data, x, can we determine how much they will be charged? We hope to help people understand if they are overpaying or underpaying for health insurance compared to their peers. Government officials may also find this data useful for making conclusions about the overall affordability of health insurance. Please note this report only covers the mainland U.S.

We additionally wanted to make a more comprehensive report of additional Connecticut data, given the first data set covers very broad regions and has already been cleaned up. Because the Connecticut data is publicly available, most information about the patients themselves are omitted. Instead of factors going into a patient's insurance cost, we are interested in the correlation between "reason" and "sub-reason" for insurance companies covering certain costs. Additional material includes how long the ticket was opened and when it was closed.

Our domain spans finance, because of the insurance component; healthcare; and demographics, including sex, age, and number of children.

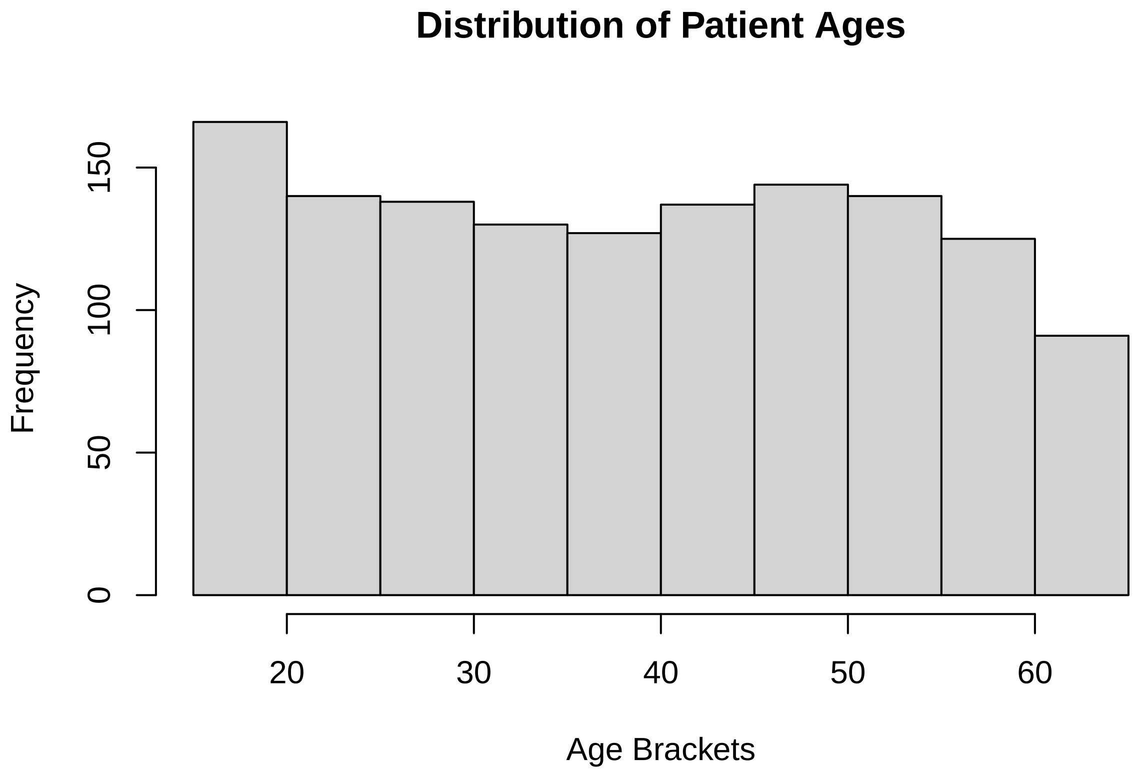
2. Data and Resources Used.

Our main data set is from Kaggle (https://www.kaggle.com/datasets/mirichoi0218/insurance). We initially were unsure of which variables were significant or not, hence our approach to take all factors into account (age, sex, bmi, children, smoker status, region) and finding which ones were statistically significant. The variables are described below.

All empty rows were preeminently ommitted. There was no other data cleaning involved.

Age: This variable measures the patient's age at the time of the hospital bill's creation.\

Data Type: numeric
Range/Levels: 24-80

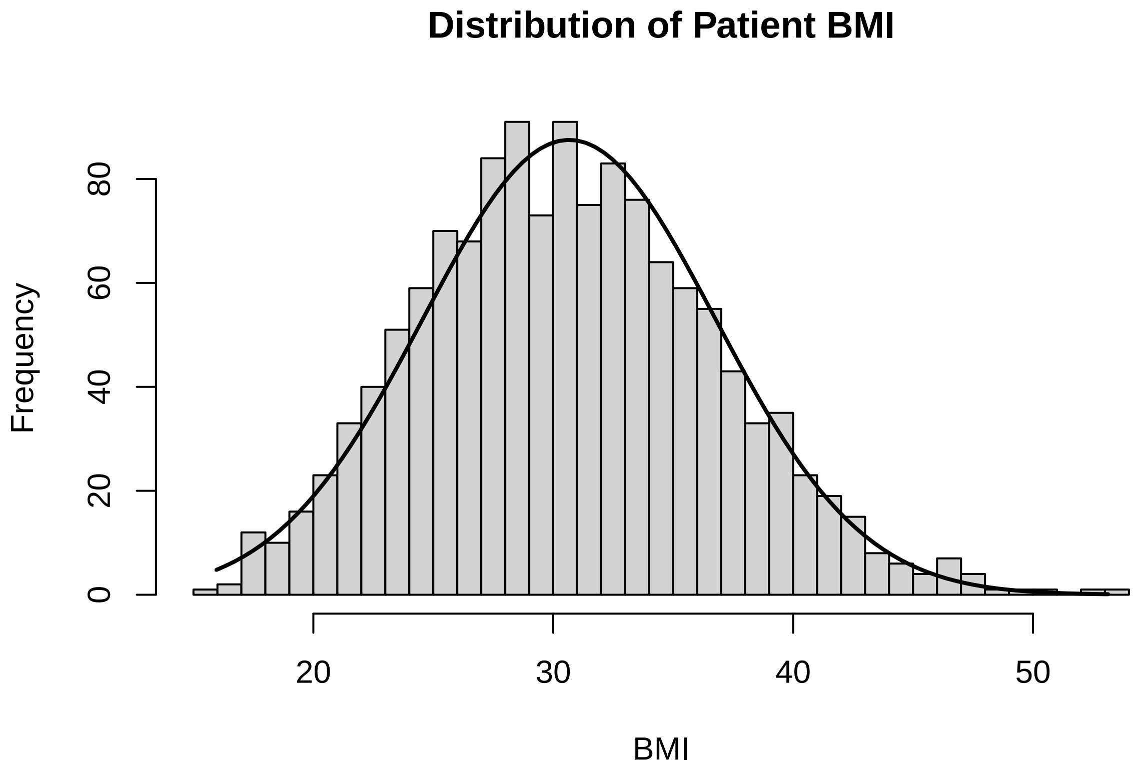


As we see from the graph above, each age group of intervals of 5 has relatively similar frequencies in our data set, and there is no outlying skew to the age.

Sex: This variable measures the patient's sex assigned at birth. Data Type: categorical
Range/Levels: "male", "female"

The table shows the number of females and males in our data set, along with their proportions.

BMI: This variable measures the patient's BMI. Data Type: numeric
Range/Levels: 15.96-53.13



Patient BMI levels follow a normal distribution around BMI = 30.

Children: This variable measures how many children the patient has. Data Type: numeric
Range/Levels: 0-5

```
table(insurance$children)
```

```
##
##   0   1   2   3   4   5
## 574 324 240 157  25  18
```

```
prop.table(table(insurance$children))
```

```
##
##          0          1          2          3          4          5
## 0.42899851 0.24215247 0.17937220 0.11733931 0.01868460 0.01345291
```

Smoker: This variable measures if the patient smokes. Data Type: categorical
Range/Levels: "yes", "no"

About 20% of patients smoke. The CDC (link in "References" section) estimated in 2021 that 11.5% of Americans smoke. We can see a higher proportion of smokers are admitted as patients in hospitals.

Region: This variable measures which region of continental USA the patient is located in.
Data Type: categorical
Range/Levels: "northeast", "northwest", "southeast", "southwest"

The frequency of patients from all four regions is relatively the same, with the southeast region having slightly more patients than the other three regions.

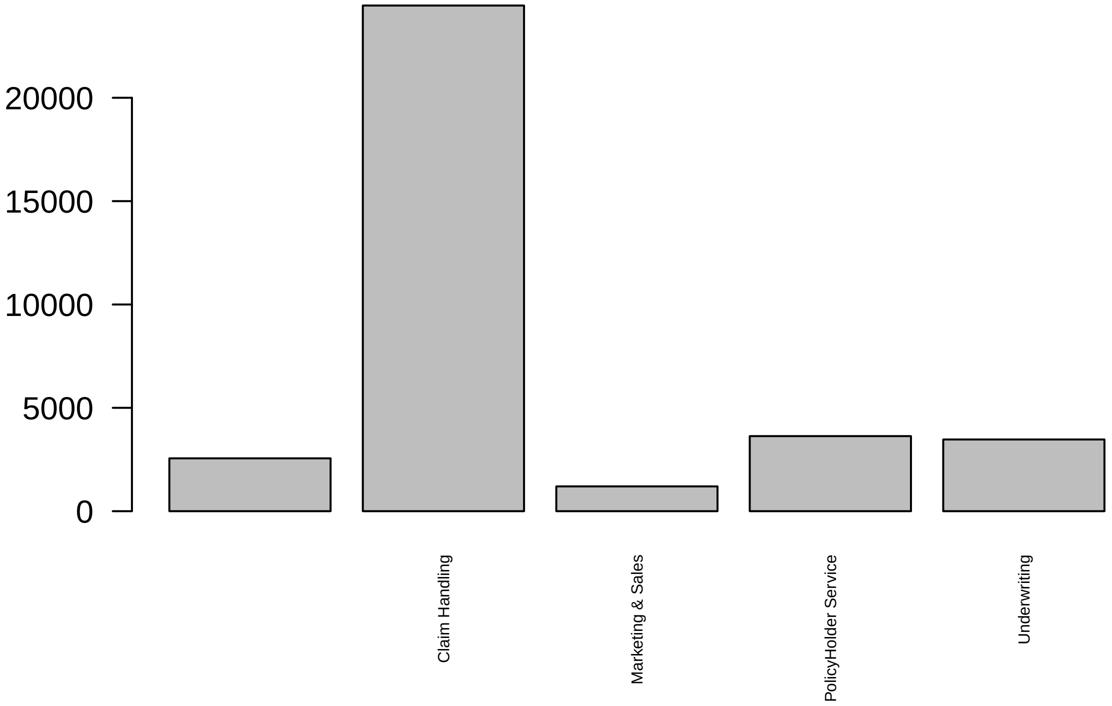
Charges: This is our predictor variable, which we will be using the other factors to estimate
Data Type: numerical
Range/Levels: 1121.874-63770.428

Our Connecticut data is from publicly available government data found at https://catalog.data.gov/dataset/insurance-company-complaints-resolutions-status-and-recoveries. The variables include insurance company, file opening date, file closing date, coverage, reason for filing claim (and a more detailed sub-reason column), recovery amount to patient, and status of insurance covered.

Company: This variable measures the patient's insurance companies. Data Type: categorical
Range/Levels: 753 companies/levels

Opened and Closed: These two variables measure when the patient file was opened and closed.
Data Type: numerical
Range/Levels: Date Opened (01/01/2022 - 12/31/2022) and Date Closed (“ ” - 12/30/2022). Some cases were not closed by the end of the year hence why the value for Closed is empty.

Reason: This variable measures the reason for the insurance claim. Data Type: categorical
Range/Levels: 5 levels: ““,”Claim Handling”, “Marketing & Sales”, “PolicyHolder Service”, and “Underwriting”



SubReason: This variable is a more in depth reason of a blanket reason. : Data Type: categorical : Range/Levels: 176 levels. Examples include “Provider Availability” and “Service Fees”.

These subreasons are matched to a larger more broad “Reason” tag.

Conclusion: We are using this variable to see if the patient's claims were accepted by insurance companies.

Data Type: categorical

Range/Levels: 52 levels from "Accident in Another State" to "Voluntary Reconsideration"

3. Analysis.

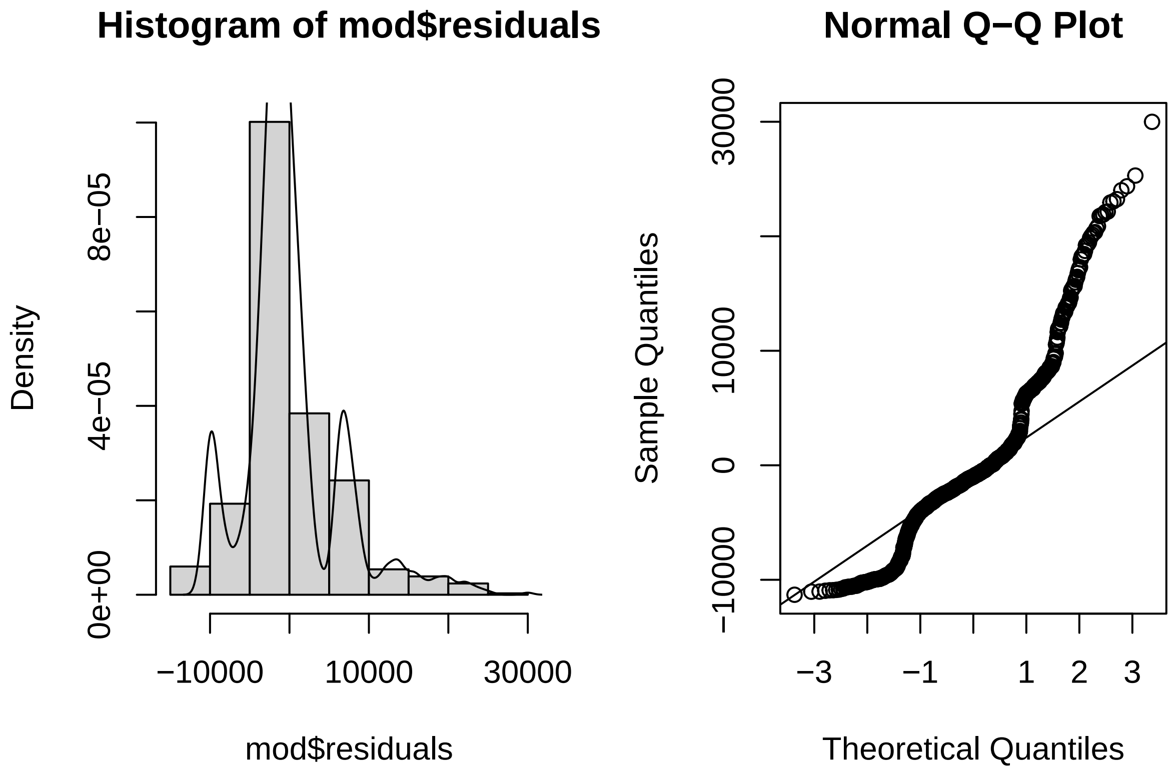
In order to identify which variables are statistically significant and which variables provide the best explanations as to what effects how much a person is charged for insurance, we used a variety of techniques to analyze the variables. Below each analysis technique is a brief explanation of the results and analysis process.

Linear Regression Models

We used a linear regression model to evaluate which variables are statistically significant with respect to the "charged" variable. We tested all variables against the charged variable and the results show that the age, bmi, children, and smokeryes variables are the statistically significant ones. This result allows us to conclude that these variables are the ones that affect the "charged" variable the most. We can also conclude that about 75% of the variance is explained by this regression model.

###Finding residuals and QQ plots for all factors

The histogram of the residuals is used to check whether the variance is normally distributed. The symmetric, bell-shaped curve histogram around zero shows that the normality distribution is likely to be assumed. This also lets us know that there aren't many observations that are skewed to either side and that the data is evenly distributed. The QQ plot gives us a visual representation of the distribution of the error points. It tells us whether or not a data set came from a theoretical distribution. If the fit is accurate, the error points should be evenly distributed around 0, and our QQ plot below shows the error points normally distributed with a mean of 0. Even though there is a slight deviation in the QQ plot, we can ignore this since the data set is very large and may havve some discrepancies.

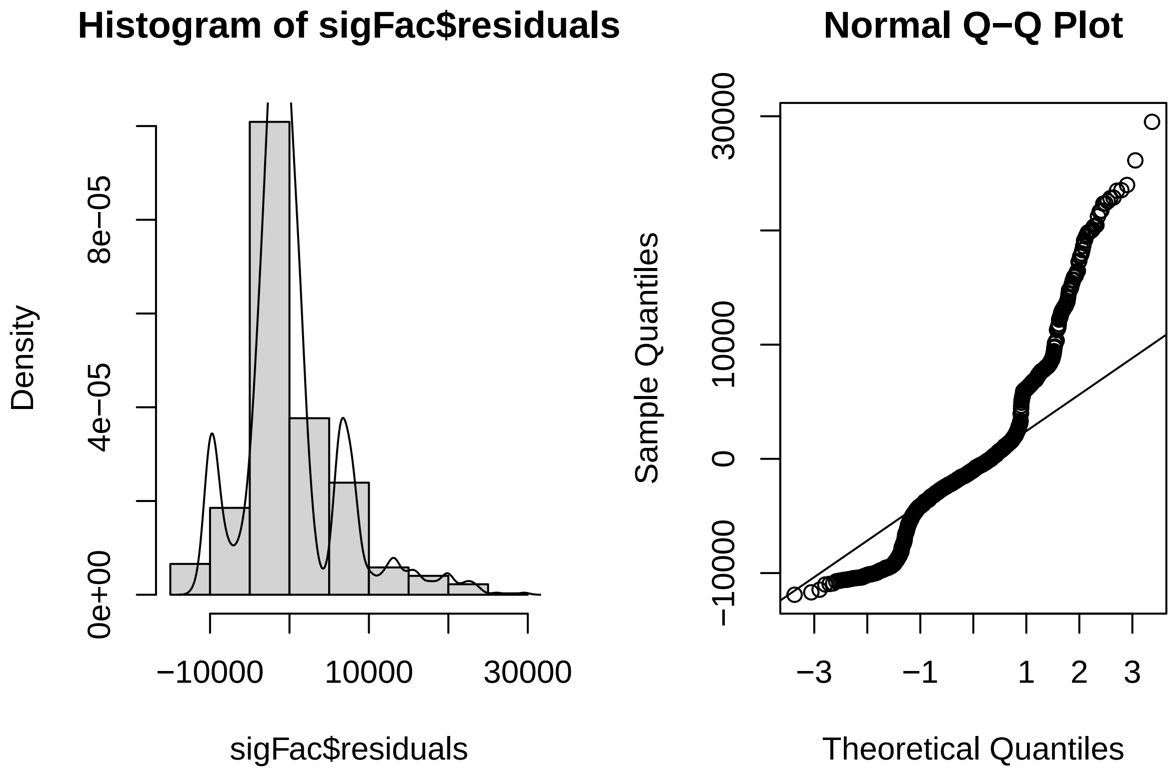


###Using only statistically significant factors: age, bmi, children, smoker (yes).

Using only the statistically significant variables allows us to further interpret which variables provide the biggest effect to the "charged" variable. Using the linear regression model again, we can check the statistic significance for these variables and see how much of the variance is explained by the model. The results obtained from this linear regression model show that all variables are extremely significant with a p-value < 0.001, and about 75% of the variance is explained by the model.

Residual Plots

The following residual plots provide us with information about the error terms for the significant variables determined from the previous model. We can see from both plots that the error terms are normally distributed with a mean of 0, which indicates that we can assume normal distribution for the data set.

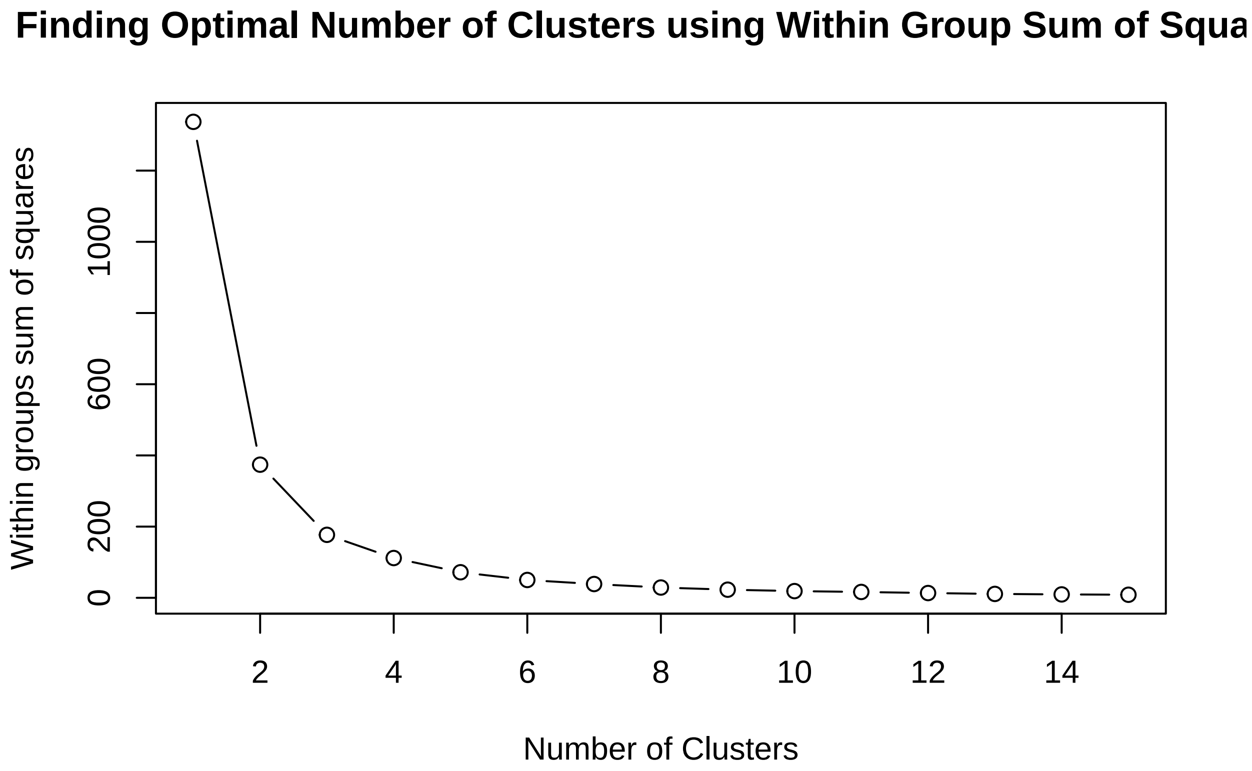


Testing Correlation With Clustering Our previous model stated that age, bmi, children, smoker (yes) were all significant factors. We are now going to perform a clustering model to see if this still holds true.

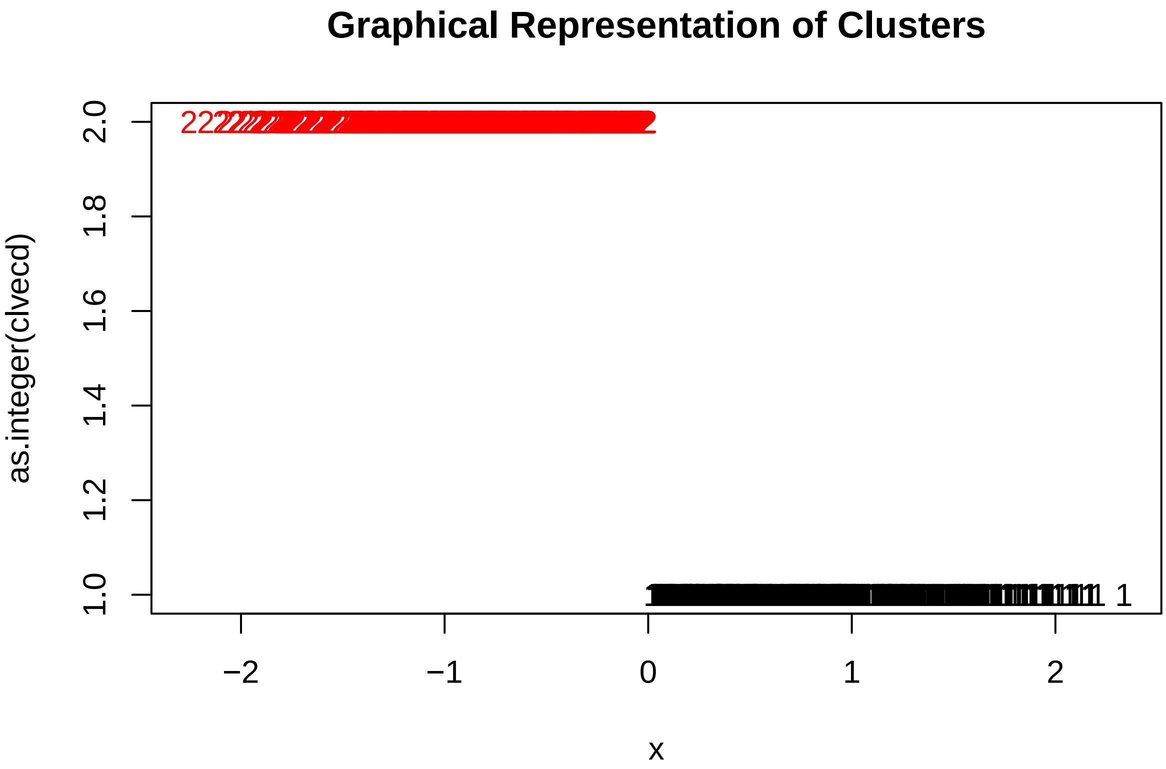
The following code chunk gives us the max scaled age value.

From the output obtained below, we can see that two clusters were developed. The size of each cluster falls in the 600 range and we also computed the within clusters sum of squared distances and between clusters sum of squared distances.

The within clusters distances is smaller than the between clusters distances, which is the desired output. Since the sum os squared distances within clusters is smaller than between clusters, this shows that the data within clusters is tightly grouped and very similar to each other. A larger between clusters value shows that data from different clusters differ a lot from other clusters and shows high variability in the observations.



The plot above shows the optimal number of clusters formed using the within group sum of squared differences. The steep drop at 2 clusters indicates that 2 clusters is the optimal number of clusters.



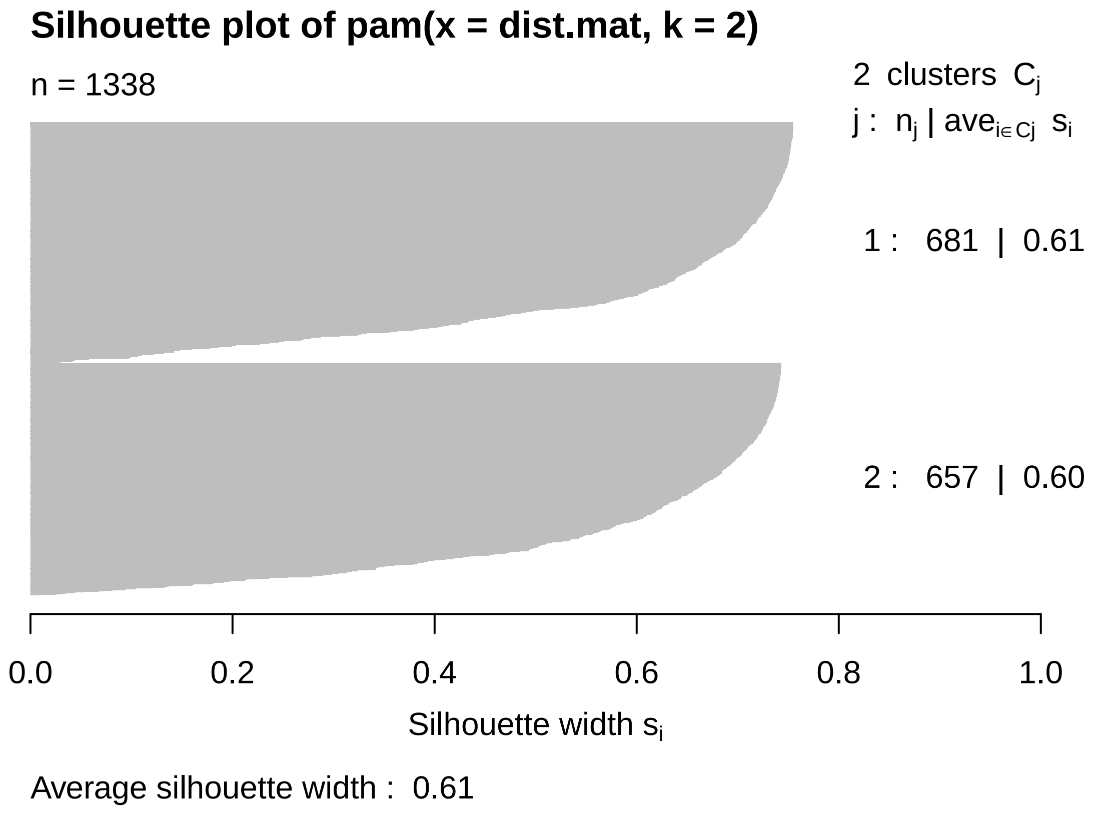
The above plot shows the distribution of the 2 clusters.

The NBClust function was not included due to knitting issues. Returned k = 2 as optimal amount of clusters.

###Optimal number of clusters using PAM

We identified the optimal number of clusters using k-means and concluded that 2 clusters was the optimal number. However using just one clustering method is not always reliable, so we used PAM to corroborate the optimal number of clusters. The result indicates that the optimal number of clusters is 2.

The plots generated below show a silhouette plot and the optimal number of clusters with the number of data points in each plot.



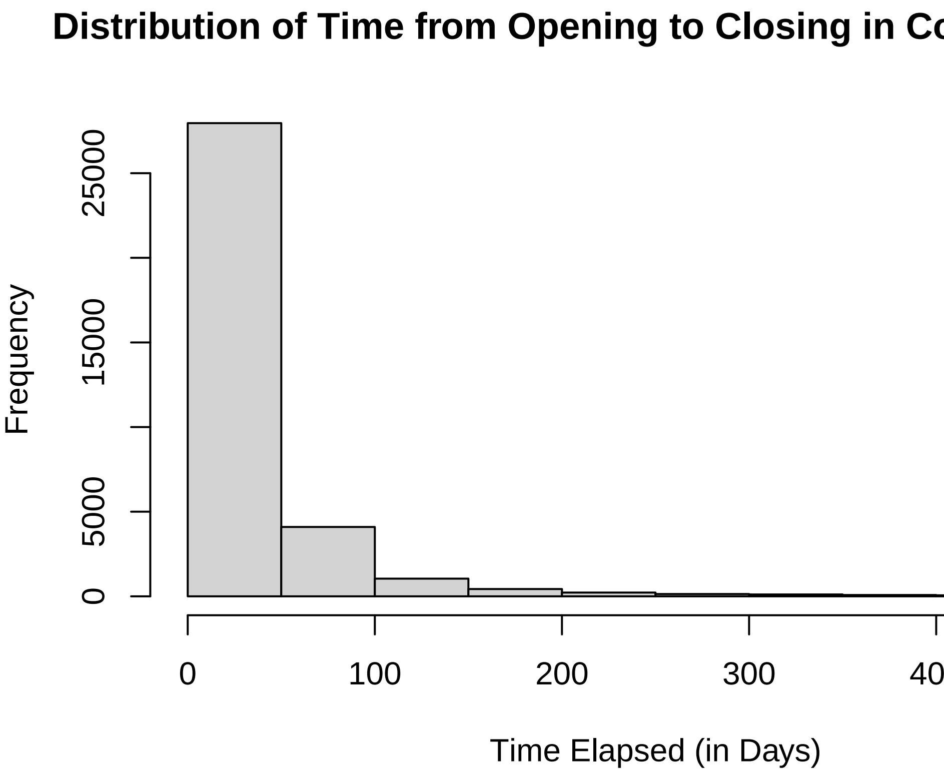
###Chi-Squared Test

The chi-squared test performed below tells us the relation between the "charges" variable in the insurance.df and the 2 clusters created. The p-value tells us the statistical significance between these two conditions. The results obtained below are a perfect example of data that isn't exactly desired. We can see that the p-value is greater than 0.05 so it may not be statistically significant. However our models use real data and this may hint at a correlational relationship between the two variables. A p-value less than 0.7 for real data can still be considered to hint at some sort of relationship. A reason for this extremely weak correlation can be due to the fact that there are many variables that are statistically significant. Since there isn't a single strong correlation between "charged" and another variable, it might have been difficult to obtain clusters that gave a strong statistical significance in the chi-square test.

How different are the clusters?

As we saw from the chi-square test results above, there isn't a very strong statistical significance between the "charged" variable and the clusters. This shows that the clusters may be very different to each other. Another way we can visualize this assumption is by calculating the randIndex. A value close to 1 shows that the sets are the same and a value closer to -1 means they are different. The value obtained from the code below shows a value that is extremely far from 1. This shows that the sets may be very different from each other.

##Connecticut Data Exploration As described in the first section, we want to use the open and closing time and reason (with its respective subcategories), to draw a conclusion with the "Conclusion" variable.



###Open and Closing Time Distribution
As we see from the histogram, most cases are closed in 50 days or less. The great majority of cases are closed before 100 days, with values tapering off after 100 days.

Which "Reasons" had the shortest time elapsed?

Claim Handling was responsible for the largest proportion (56.41%) of all cases that were closed in less than 50 days. This makes sense as Claim Handling has the largest amount of cases in general. On a percentage basis, this is larger than the other categories. On the other hand, uncategorized reasons and "Marketing & Sales" had the highest proportion of cases closed in greater than 50 days.

What about exploring by subreason?

Due to the large amount of data, we found that sorting by subreason was not useful. In insurance cases, there are many other factors besides "reasons" which may impact the results.

Summary and Conclusions.

Our project focuses on identifying certain variables that affect the rate at which people will get charged for their insurance plan. We obtained a data set which contained the demographics for individuals and we created a variety of models to identify which variables are significant in determining the rate charged for specific insurance plans. Certain modelling techniques such as linear regression models and residual plots allowed us to clearly identify which variables are statistically significant in finding out what affects the "charged" variable. However the results we obtained from clustering may not be as desirable since our data set is extremely large and there may be many other external factors not included in the data set that affect the rate at which people get charged for their insurance plan. Our chi-square test outputted a result that is difficult to interpret, but overall we can say that certain factors such as age, bmi, children, and smoking status may affect the rate charged for insurance plans for individuals.

Through our Connecticut data, we were able to compare TimeElapsed to reasons for filing claims. We found that while most cases were solved in less than 50 days, a large proportion of these cases fell under the “Claim Handling” category. The other categories had a much smaller proportion of cases solved within 50 days. We also wanted to see if certain subreasons would have longer wait times. However, our results were messy and we were unable to draw anything conclusive.

Further research should focus on more specific areas - such as company. Because of company policies, we believe there may be inconsistencies in time taken to resolve a claim conflict.

References

- https://www.statology.org/overlay-normal-curve-histogram-in-r/
- https://www.datacamp.com/tutorial/make-histogram-basic-r

Academic References

- https://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm#:~:text=This%20mea