JEDI team #2 project report - Data Analytic Project Lab

Xing Cheng, Parinintha Kompala, Shrey Khetrapal

1. Background

Life insurance is a contract between you and an insurance company. Essentially, in exchange for your premium payments, the insurance company will pay a lump sum known as a death benefit to your beneficiaries after your death. Financial advisors (known as a distribution) have prioritized selling life insurance to high-net-worth clients for decades since commissions are more significant. During the period of coverage, their policies will help their loved ones financially. But there are times when a company has no choice but to decline to pay a death benefit.

The average monthly cost of life insurance is $26. This is based on Quotacy statistics for a 40-year-old buying a $500,000 term life policy with a 20-year term, which is the most frequent term length and amount sold. Meanwhile, 48% of Americans have no life insurance at all, which means that if the significant breadwinner died, the family would certainly experience substantial financial difficulties.

As part of its ESG strategy, Swiss Re started a Life & Health Sustainability Initiative in 2021 to work with insurance companies (called carriers) to make life insurance available, accessible, and affordable to underprivileged areas. Immigrants, ethnic and racial minorities, women, gig workers, LGBTQ+, rural, and low-income households are among the communities targeted.

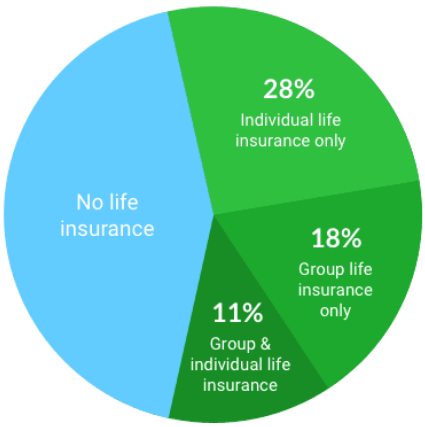


fig1-Piechart of insurance holders in the USA

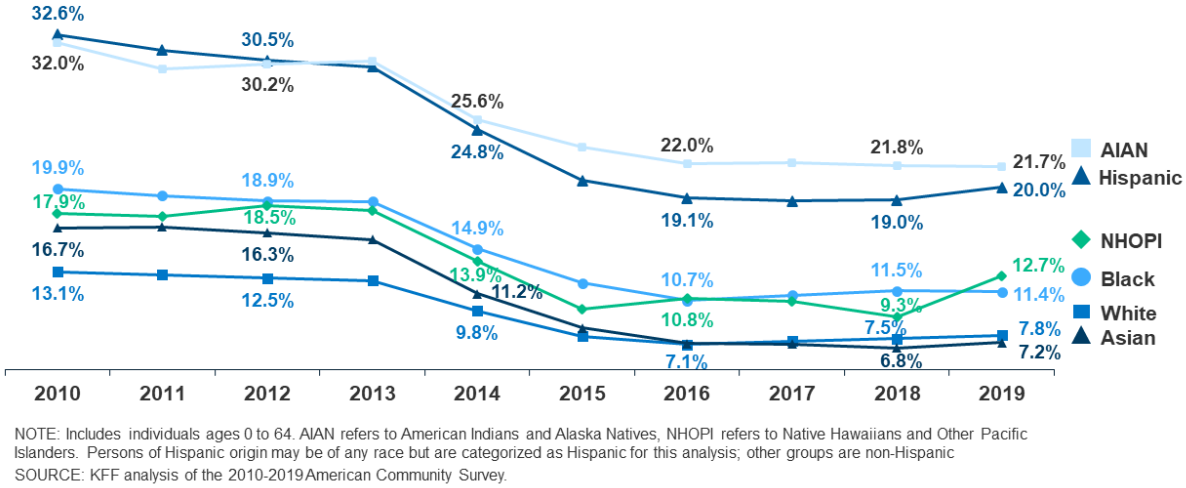
The goal of this project is to provide carriers with a data-driven social inclusion model that will allow them to 1) create profiles of underserved life insurance customers, with a special focus on our seven target communities, and 2) compare those profiles to

carriers' existing customer base to see what groups are missing and where more inclusive sales growth opportunities may exist.

And in this project, we will be concentrating on the black community people.

2. Problem statement

- Identify whether race is a factor for people with high insurance gap
- Define the characteristics of policyholders by zip code, discover underserved groups, and locate them by zip code.



NOTE: Includes individuals ages 0 to 64. AIAN refers to American Indians and Alaska Natives, NHOPI refers to Native Hawaiians and Other Pacific Islanders. Persons of Hispanic origin may be of any race but are categorized as Hispanic for this analysis; other groups are non-Hispanic
SOURCE: KFF analysis of the 2010-2019 American Community Survey.

fig2-Uninsured rates for the Nonelderly population by race and ethnicity, 2010-2019

Life insurance is available to anyone, but the cost or premium level can vary greatly based on the risk level an individual presents based on factors like age, health, and lifestyle. Life insurance applications generally require the customer to provide medical records and medical history and submit to a medical exam. Some types of life insurance such as guaranteed life don't require medical exams but generally have much higher premiums and involve an initial waiting period before taking effect and offering a death benefit.

From the graph above (fig 2) we have a notion that few communities are underserved and that we here in this project will try to find exactly where these communities are living, by zip codes.This "protection gap" is especially difficult to close in low-income and marginalized populations, where risk knowledge and insurance price can be significant roadblocks. 4 As a result, numerous government bodies have found themselves in the

position of acting as primary insurers. This is an unsustainable position for budget-strapped government agencies as well as vulnerable communities and residents who face years of waiting for assistance to get back on their feet following a disaster.

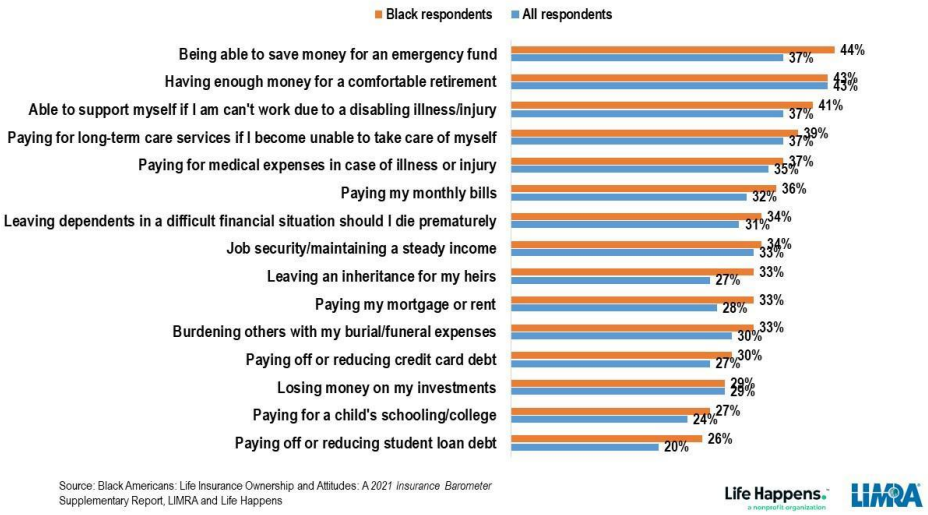


fig3-Black respondents to a financial survey

From this graph, you can clearly say that people from the black community are more active and secure, and organized about their finances. Eighty-one percent of Black respondents indicate they have life insurance, either through work or individually, compared to **70 percent** of white respondents.

Initially, African-Americans could purchase life insurance policies on equal footing with whites. That all changed in **1881**. In March of that year Prudential, one of the country's largest insurers, announced that policies held by black adults would be worth one-third less than the same plans held by whites.

And also its 2021 Barometer Study says that 56 percent of **Black Americans** have now purchased life insurance policies in the last year, which is the highest rate among all racial groups.

One of our major questions now is " WHY ARE THE BLACK COMMUNITY STILL CONSIDERED AS UNDERSERVED COMMUNITY?"

Black Americans are more likely to carry life insurance than the average American, but their coverage does little more than provide for funeral costs.

So in this project, the factor that we will be considering along with the ethic ratio will be the income.

3. Data Scraping and data preprocessing

1. The first dataset consists of the race data of the state of CA, which include the total income and income by race and total population and population by race.

2. The second dataset consists of the policyholder data of the state CA, which includes the face amount, zip code, age, and term.

3.1 Census data

Total counts, sample surveys, and administrative records are all used to obtain census data at regular periods. Census data is summarized after it is gathered or generated to show counts or estimates of groups of people for various geographic areas.

As we needed the race data and as one of the drawbacks is that the insurance companies do not collect race data, we scraped the census data that had race details for the state of California from the internet.

Our Census data contained the population count of the people from the black community and the white community zip-code vise and the income data of the people from the black community and the white community zip-code vise.

3.1.1 Missing Data

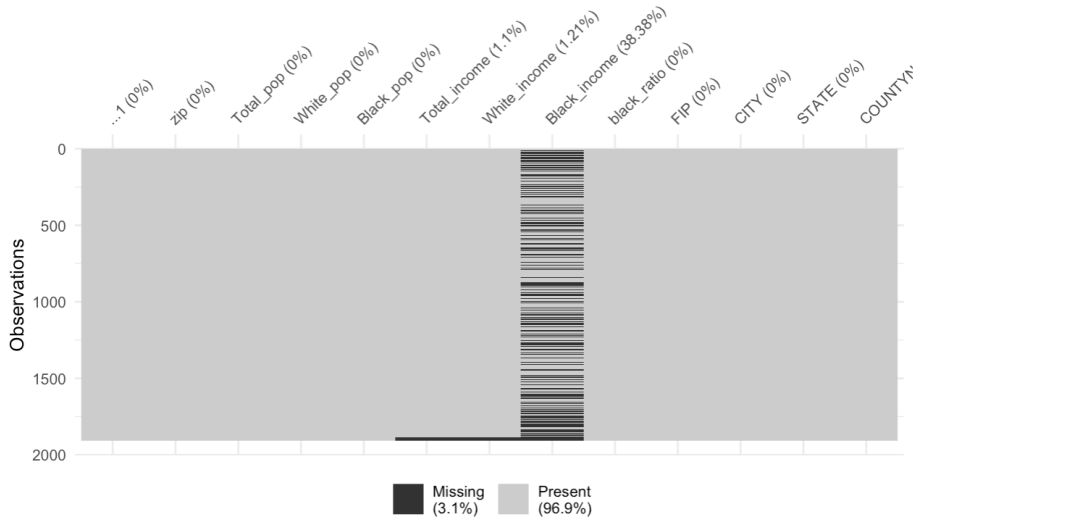


fig4.Missingness chart of the census dataset

The missing data from this particular dataset was manageable. Even after removing the NA values, we had sufficient data to perform the analysis.

3.1.2. Data Preprocessing

Removed the top1%, Bottom 1%, that had outliers from the census dataset and sorted by income, and filtered out the Top 10% and Bottom 10%. At the end of all the filtering and processing, we obtained the high-income and low-income datasets.

3.1.3 Data Visualization

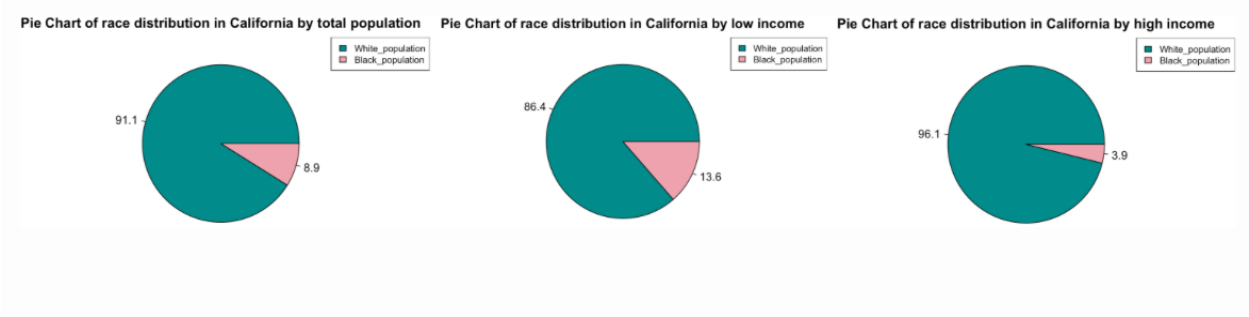


fig 5.Pie Chart of race distributions

Since the Great Recession of 2007, economic criticism has alternated between praising the recovery and raising concerns about rising inequality. Inequality in income is frequently cited as a reflection of the country's fundamental economic problems as well as government policies that disproportionately benefit the wealthy. Wealth, specifically the expanding wealth difference between people of different races and ethnicities, is an alternative indication of the nation's social and economic health.
No race or ethnic group constitutes a majority of California's population: 39% of state residents are Latino, 35% are white, 15% are Asian American or Pacific Islander, 5% are Black, 4% are multiracial, and fewer than 1% are Native American or Alaska Natives, according to the 2020 Census.
Also, the black ratio of the US is 13% and from that point of view, the above graphs are quite acceptable regarding the contribution of black people’s income towards the entire state’s income. The pie charts above show the race distribution in CA by total income, low income, and high income.

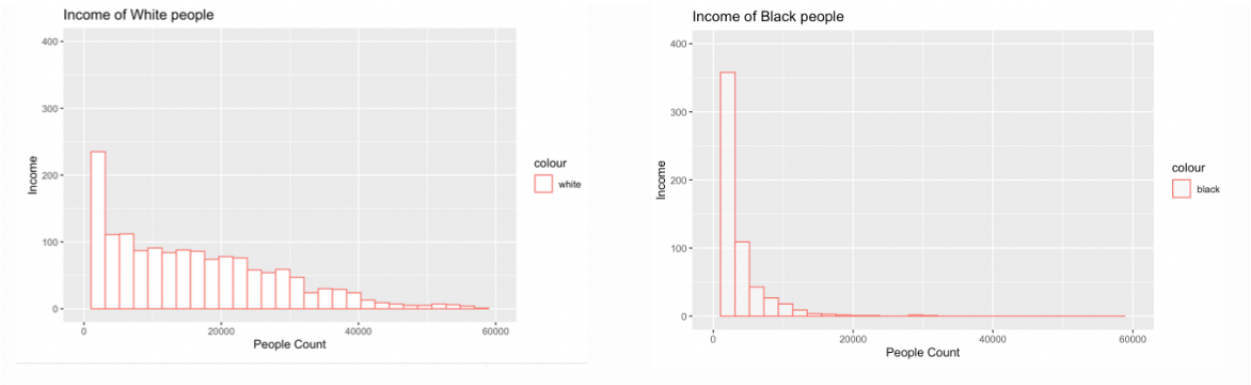


fig6.Distribution graph of income by race of California

From the histogram above we can talk about the distribution of income of both the races and in both the graph, they are skewed rightwards. On the right side of the graph, the frequencies of observations are lower than the frequencies of observations on the left side.
The black people's income is more skewed than that of the white people, this clearly says that black people's median income is way lower than the white people's median income.

3.2 Policyholder data

A policyholder is a person who owns an insurance policy in the insurance sector. You are the one who purchased the policy and can make changes to it as a policyholder. Policyholders are also in charge of ensuring that their premiums are paid.
Policyholder Data refers to the Policyholders' Personal Information, information, records, and data, whether financial or otherwise, held by, under the direct or indirect control of, or in possession of the Intermediary for the purpose of rendering the Intermediary Services under the terms of this Agreement or processed by the Intermediary in performing its obligations under this Agreement, including.

The policyholder data that was used in this project contained the CA state's policyholder from Insured connect.

3.2.1 Missing Data

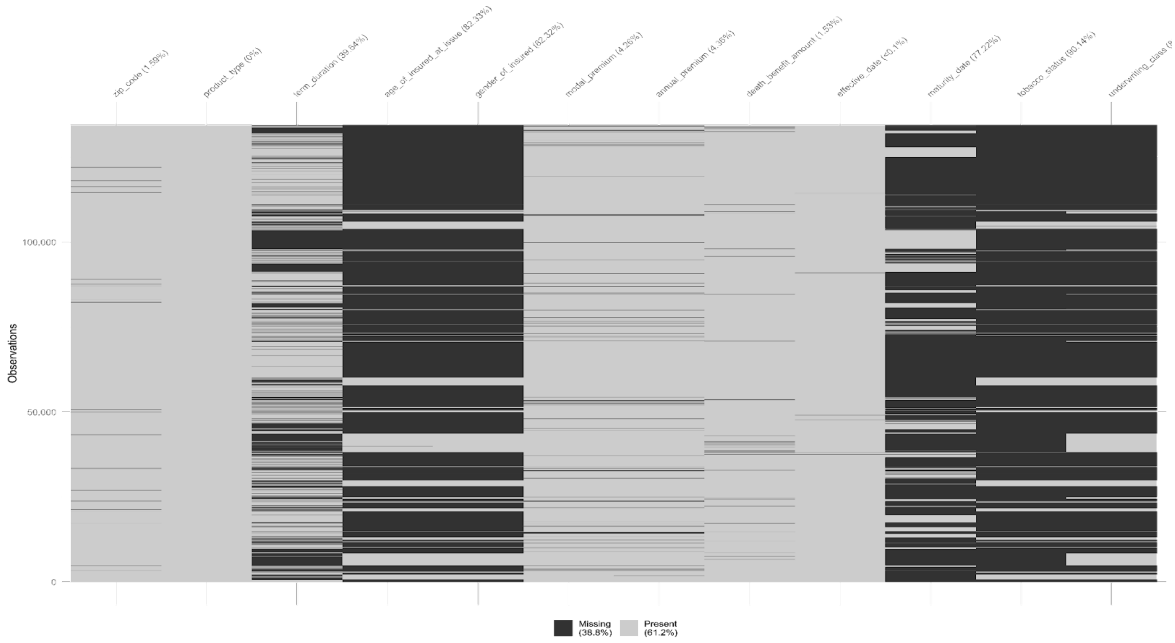


fig 7.Missingness chart of the policyholder dataset

From the image above you can clearly see that there were so much missing data as this is real-life data.
We did a lot of preprocessing to overcome this and those steps will be explained as we move forward in the paper.

3.2.2 Data Preprocessing

Initially, we had 1,34,159 observations. From the missing count graph above we can see that a lot of preprocessing should be done on the data to move ahead with the analysis.

Age is one of the variables that we need to perform our analysis on, so we removed the NA values from the age and also the values which were zeros in that column. After this, we were left with 23,737 observations.

There are many other types of life insurance plans available, including innovative variations on classic coverage, but most consumers choose between standard term and whole life policies, which each have their own set of benefits. Term life insurance is less expensive and easier to understand, but it does not continue forever, whereas whole life insurance does not expire and is more expensive.
Term life insurance is appropriate for most people because they will not want coverage during their retirement years, but it is not appropriate for everyone. People with long-term dependents or more sophisticated financial planning needs might consider whole life insurance. Our dataset had many policies but we need only ’the term policy’, after filtering for that we had 5796.

We then filtered out the face amounts that were 0 and had 5032 observations to do our analysis.

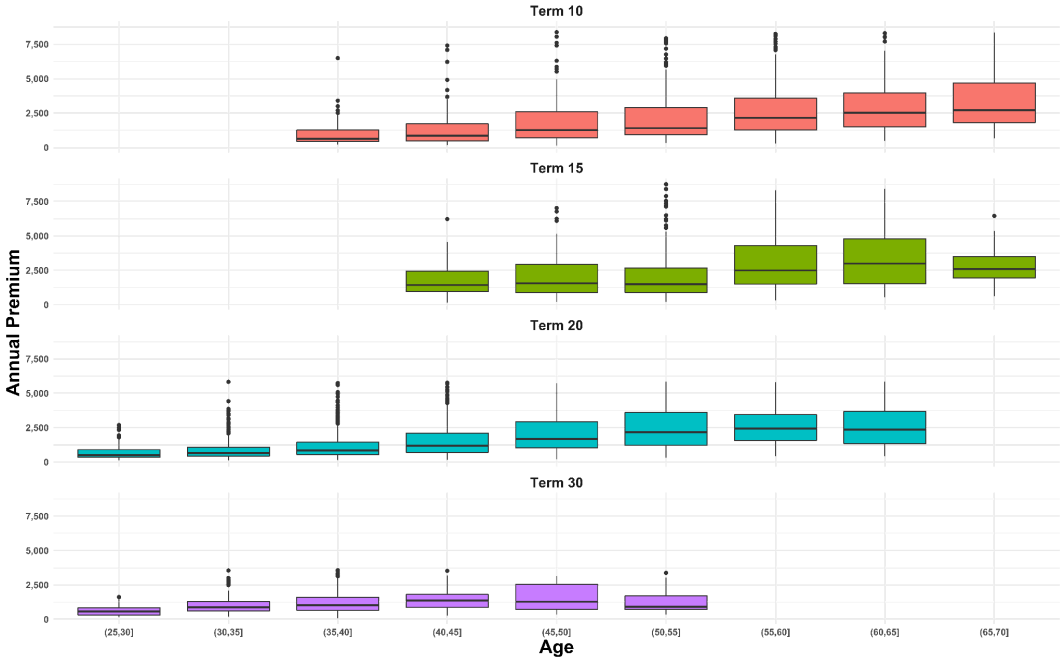
3.2.3 Analyzing trends

We analyzed different trend patterns between the following variables :

1. Annual Premium
2. Death benefit amount
3. Age
4. Gender
5. Policy Term Length

The first analysis is between Annual premium and age stratified by term on the y-axis. We can see that there is an increasing trend between the annual premium and age and if you want to compare the term policy rates for one age in particular we can use the Y axis i.e. one vertical line and compare the annual premiums for a group based on term policies.

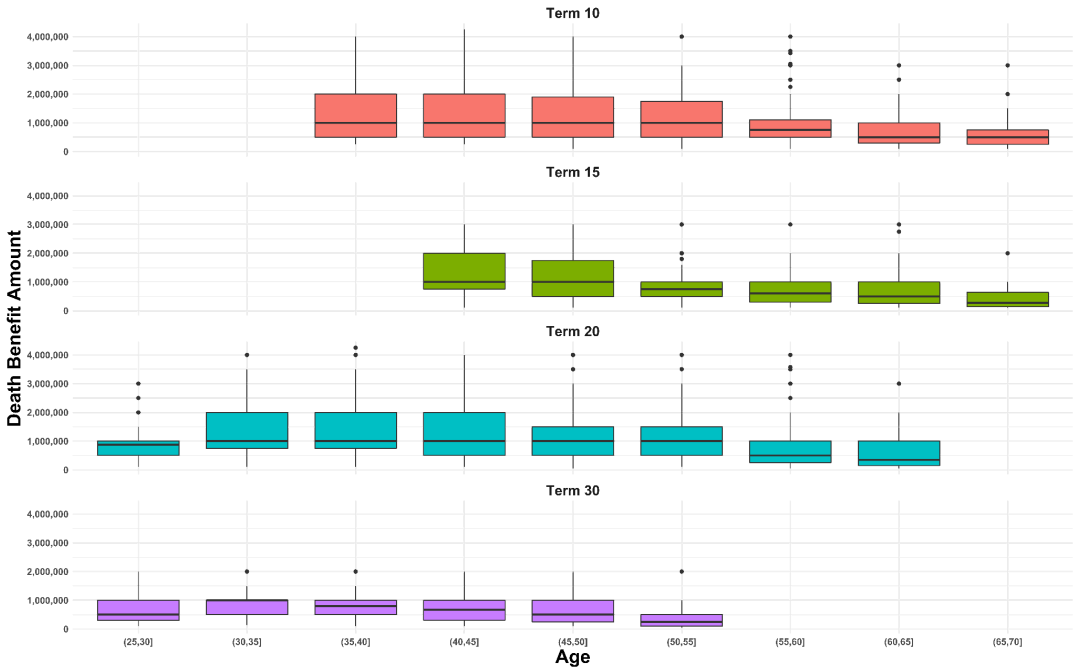
It is a very information graph and we can measure a lot



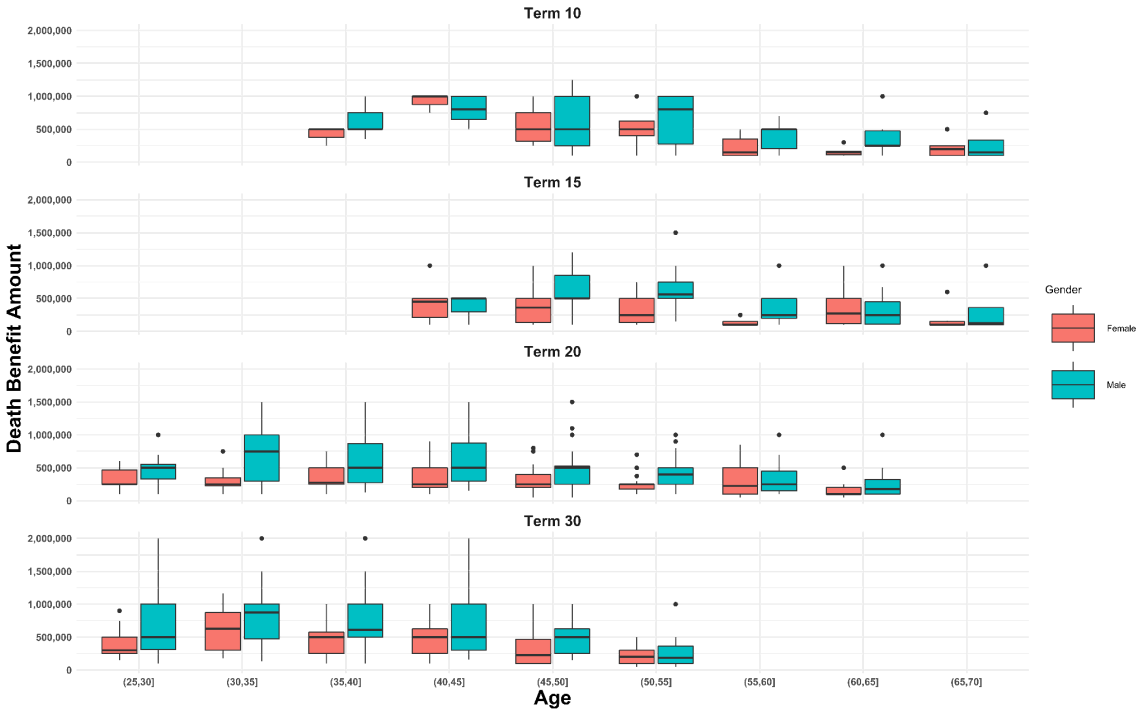
The second analysis is between the death benefit amount and age stratified by the time policies and as you can clearly see that when we had an increasing trend for the annual premium graph here it is a decreasing trend if you look from an overview.

This makes sense because the death benefit amount decreases as a person grows older while the annual premium of the person increases as the person gets older.

And again if you have to compare the debt benefit amounts for one age for different terms we can look at one particular vertical line and compare the median and distributions of the different box plots for different age groups.



The third analysis is between the death benefit amount and the age group stratified by the term policies on the y-axis but now the graph says even more because we have stratified by gender as well. Without looking at the immediate trends of the box plots if you look from an overview we can see that the red boxes are lower than the blue boxes, the red boxes signify the death benefit amount for the females while the blue boxes signify the death benefit amount for males. The question here arises that why do females have lower death benefit amounts compared to males when we are looking at a similar distribution of samples and data. So based on this data we can clearly say that females are also an underserved group when it comes to insurance.



4. K-Nearest Neighbor similarity model

The basic logic of the methodology is inspired from A/B testing(randomized controlled experiment). To discover the driver factor of life insurance purchasing power, which is black community ratio in zip code level in this project, we need to control other confounders which have an effect on life insurance purchasing power, like income, population, etc. When two zip codes are very similar on all other factors, we can conclude that any purchasing power difference is caused by the black community ratio, which is the only difference between two zip codes.

Since income is the most important driver factor, we separate the zip codes to decile by zip code per capita income. Then we selected the bottom 10% zip codes as low-income areas. In another dimension, all low income zip codes are separated into two groups, high black community ratio and low black community ratio, using median black community ratio as threshold.

K-Nearest Neighbor(KNN) algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. If we use multiple census factors to represent each zip code on multi-dimensional space as a data point, the similarity between zip codes can be measured by the distance between data points.
KNN was used to match low black community ratio zip codes with the most similar high black ratio zip codes. In this project, we used Euclidean distance as the distance

measurement. And because we only want to find the most similar zip code to form pairs, the parameter k of KNN is equal to 1.

After the KNN step, we got 23 pairs of zip codes. Within each pair, the two zip codes have similar demographic factors (population, population per capita income, black community income, white community income), however, the black community ratio has major differences. Then we estimated purchasing power using the average premium amount from the policyholder dataset and computed the purchasing power gap between two zip codes. If the purchasing power gap is positive, we can conclude that the low black community ratio zip code has higher purchasing power than the high black community ratio zip code. And a bigger purchasing power gap stands for the higher potential of the high black community ratio zip code.

5. Challenges

So we faced many challenges but the three of them that was the major concern for us were the

The first ones that raise data are not collected by the insurance companies and our target market for this project was different races, so, we had to rope in census data and make an estimate regarding the ethnic background of a location.

Secondly, there was a lot of missing data in both the census data set and the policyholder data set. So while filtering out data we only took the ZIP Codes which had at least 50 data points in them so they were many ZIP Codes that had only 1 datapoint and the annual premium amount for that person was very high which acted like an outlier for us so to get rid of that we summarize the data set in a way that only the ZIP Codes that have at least 50 data points are available for us and we only use those for further analysis.

The problem with census data is that it's collected on a county level and the dataset that we had from the policyholders contains data at a ZIP Code level. We found ZIP Code-level data from a third-party census website but insurance companies would also have resources to deploy third-party companies to get the actual ZIP Code-level data for better analysis in comparing the policyholder data set and their own census data

6.Future scope

1. The future score of this project is that we can add more factors such as sex, age and education level to the analysis along with income and race for the better performance and estimation.
2. The beauty of the method that we applied is that we can use it for any state that we want to so for example if you want to target Virginia and apply the same analysis all we want is the policyholder data from Virginia and we already have the census data so just follow the same step 1, 2 and 3 and we'll have pairs of similar ZIP Codes ready for further analysis.
3. Also, when aiming at hispanic communities, we would only need the hispanic data from the census and the same clustering method can be applied and fine tuned according to preference.
4. Use of other clustering methods such as random forest could enhance the model's performance