Introduction

Given our group was finance and insurance and found out that we were all placed at Northwestern Mutual, one of the largest life insurance companies in the US, we decided to look at data relating to life insurance and try to see if we could find out how many people were covered by life insurance in the US, as well as a few other things. We looked at how many people wanted life insurance but couldn't, what the reasons are for buying or not buying life insurance are, where life insurance companies are paying out, and how much it can cost people to get life insurance. Along the way, we incorporated health insurance data from the US Census Bureau to compare how often it's bought relative to life insurance as well as to help create a machine learning model to predict life expectancy, and see what factors affect it.

Methodology

The data that we used for this project falls into two categories: web-scraped and US census-based. Our scraped data came from a variety of sources, which can all be found in the data sources section of this report, but many of the tables that we used were from https://www.bestliferates.org/statistics, which helped answer several of the questions that we originally had, including how many people in the US were covered by life insurance, why people do or do not buy life insurance, and how many people think that they should have life insurance but do not. We got some sample rates of the monthly cost of a 20-year, $250,000 plan for males and females of varying ages from https://www.investopedia.com/articles/personal-finance/022615/how-age-affects-life-insurance-rates.asp, and the annual payouts of life insurance plans to various categories from the insurance information institute at https://www.iii.org/table-archive/22403. All this data was directly scraped in Azure Databricks

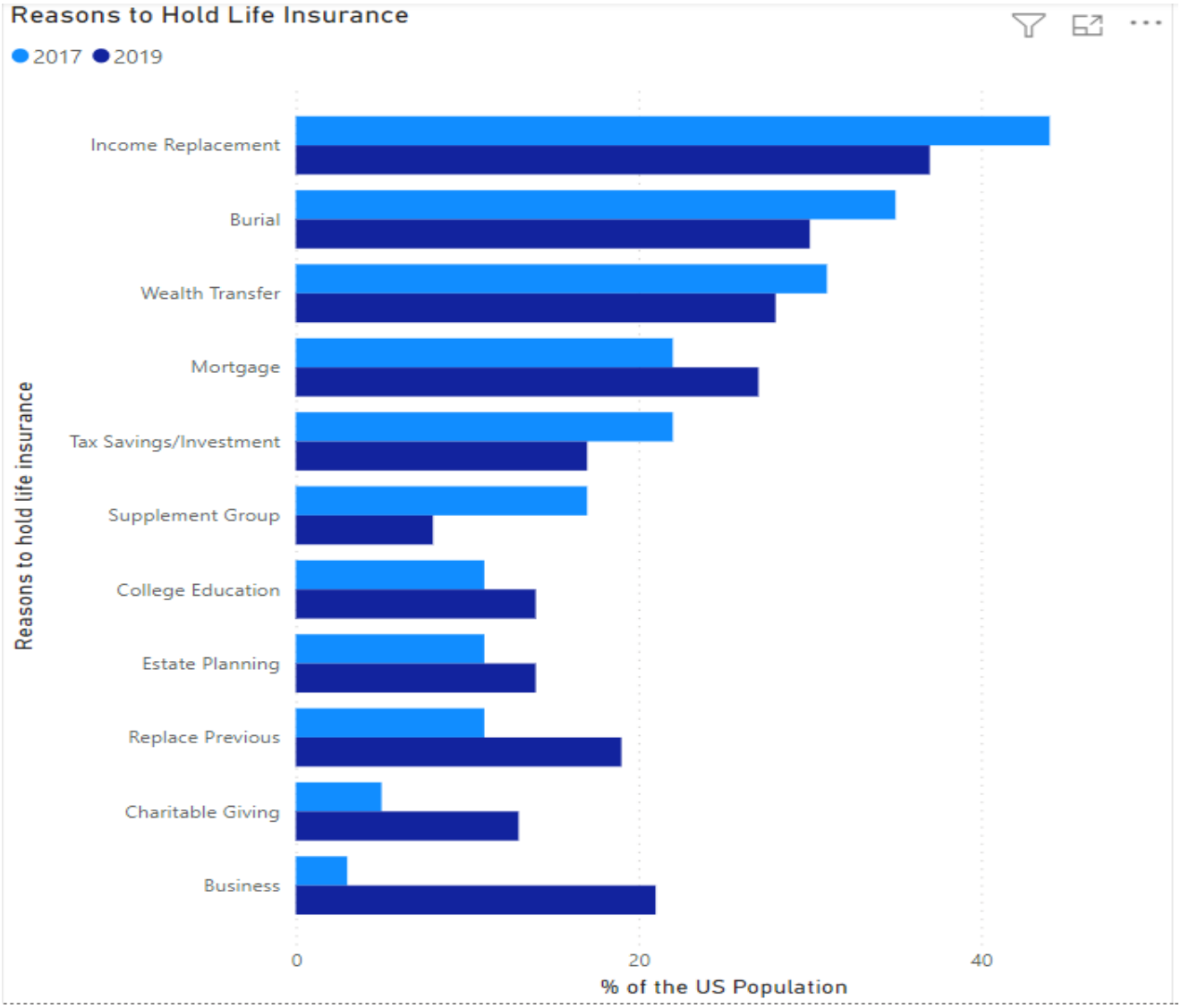
From the US Census, we found percentages of people covered by health insurance at https://www.census.gov/data/tables/time-series/demo/health-insurance/acs-hi.html to compare against life insurance rates, as well as to use along with state median income levels found at https://data.census.gov/cedsci and some helper abbreviation data from https://worldpopulationreview.com/states/state-abbreviations to create a machine learning model to predict life expectancy and compare its accuracy to the life expectancies found at https://www.cdc.gov/nchs/pressroom/sosmap/life_expectancy/life_expectancy.htm.

Most of our data ETL and processing was done in Databricks using various python packages, such as pandas and pyspark, more of which can be found in our ETL report. For some particularly problematic census data, we used power query in Azure Data Factory before sending it back to an Azure Data Lake to be read into Databricks. When all the data has been read into Databricks, the data is cleaned to remove errors and null values, change data types, etc. Afterwards, most of the data is directly imported into an Azure SQL Database, however there was one table which we put through the Kafka Producer Consumer process. For the table of reasons Americans gave for not having life insurance, we used a Kafka producer to send rows of the table as raw text messages to a Kafka topic, and then used a Kafka consumer to get the messages from the topic and convert them back to a spark data frame. We

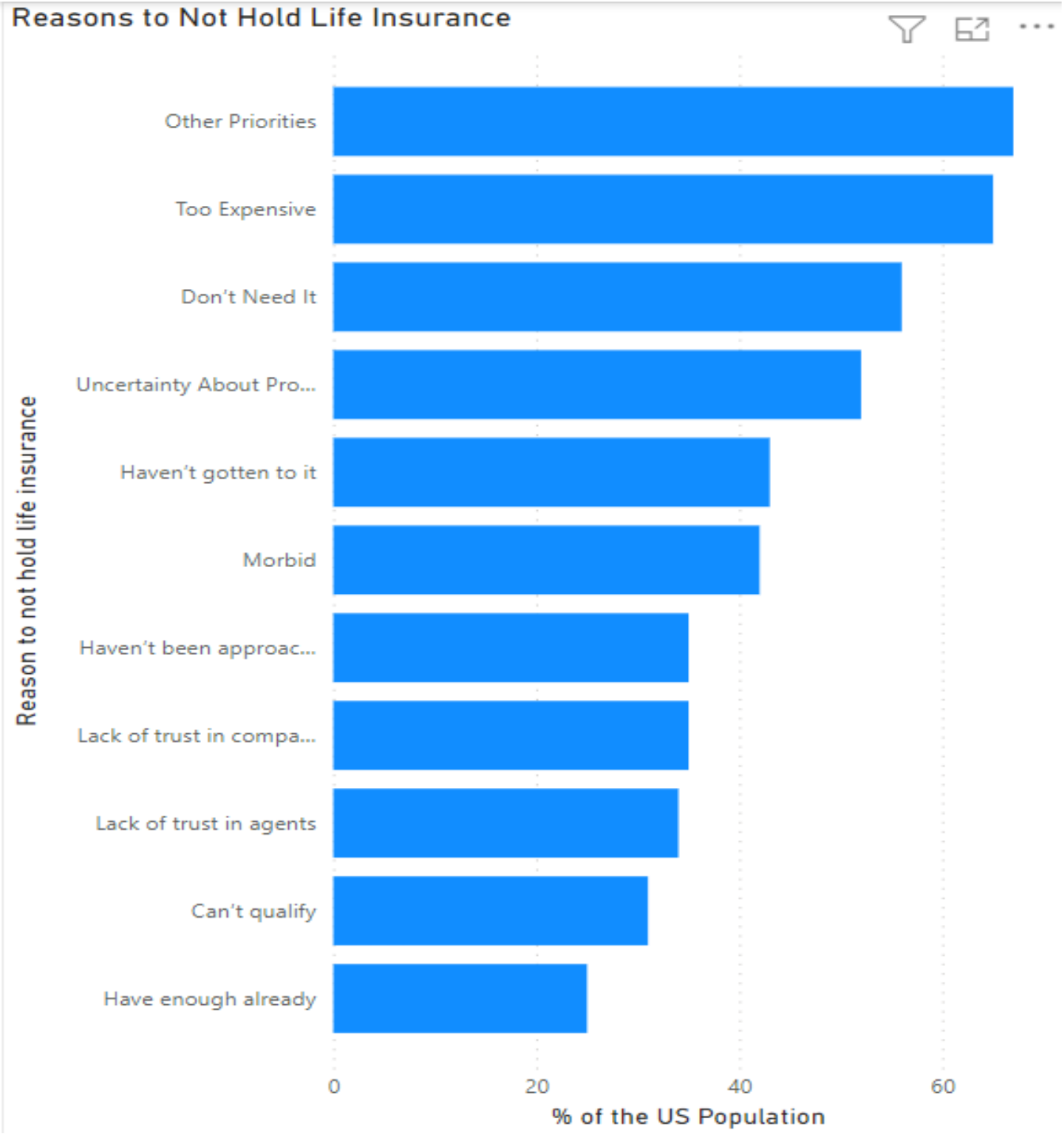
then wrote this data frame table to a data lake as a csv and automated a pipeline in the data factory to run the data brick and copy this newly written data from the data lake into an SQL database.

Analysis

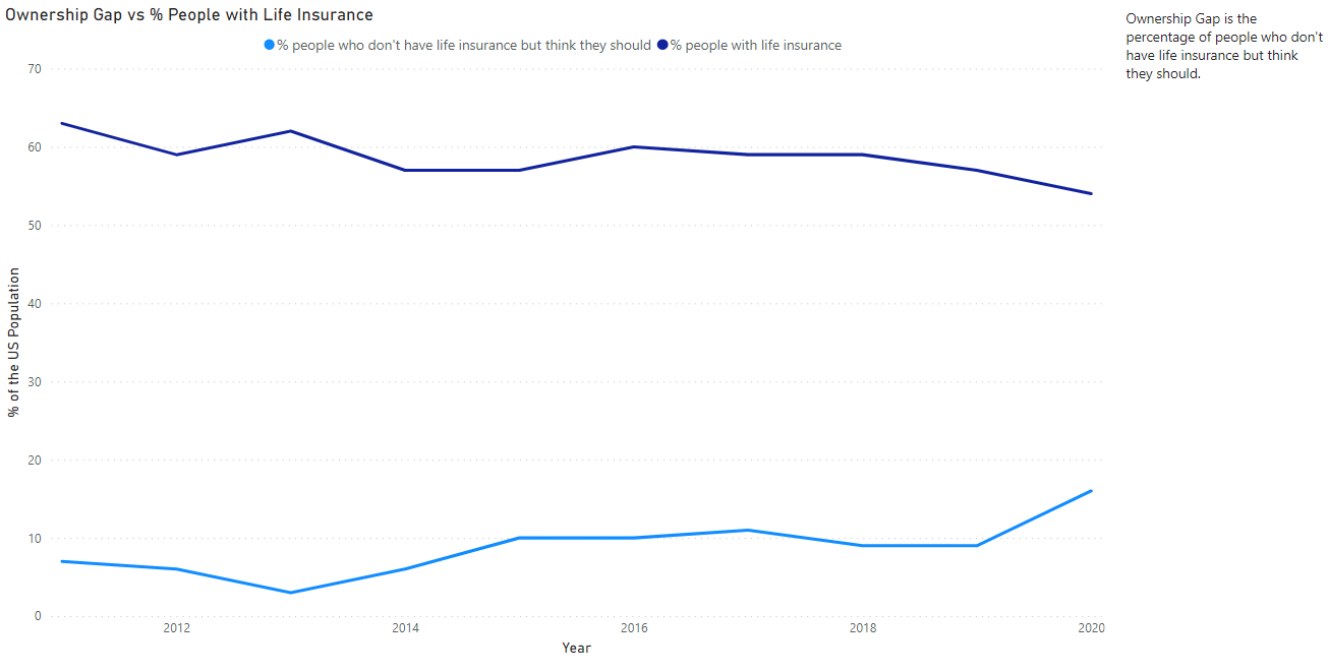
As seen below, across both years surveyed, income replacement, burial costs and transferring wealth stand out as the most cited reasons for one to purchase life insurance, which seems to be in line with the common conception of life insurance as a way to alleviate some of the economic burden on one's loved ones after their death. However, it's also notable that all these reasons showed a decrease in popularity from 2017 to 2019, as well as some previously uncommon reasons increasing in popularity, such as business, which may suggest a change in the way the public perceives life insurance. As such, it may be a good idea to investigate why this is, so that any advertisements can better cater to these new needs accordingly

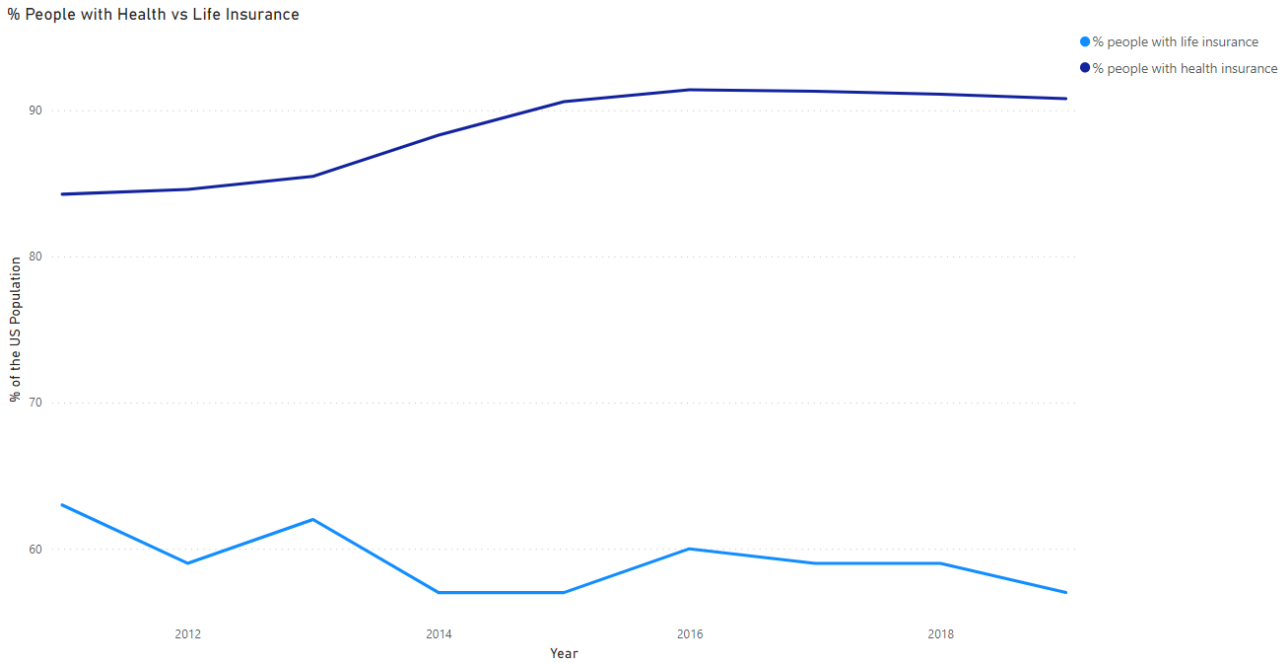


As for why Americans don't have life insurance, it mostly boils down to not being satisfied with the value that they'd get out of it. They don't perceive it as being worth the money that they're putting in, either because the premiums are too much, as seen by the "Too Expensive" category being the second most common, or that they have other things that are more worth their money or time, as seen by entries 1, 3 and 5. Interestingly enough, the idea of life insurance being too morbid took the number 6 slot, but overall this data was a bit limited, as it didn't specify a year or have separate groups for how these responses varied by demographic, especially age, so these may be things to look into.

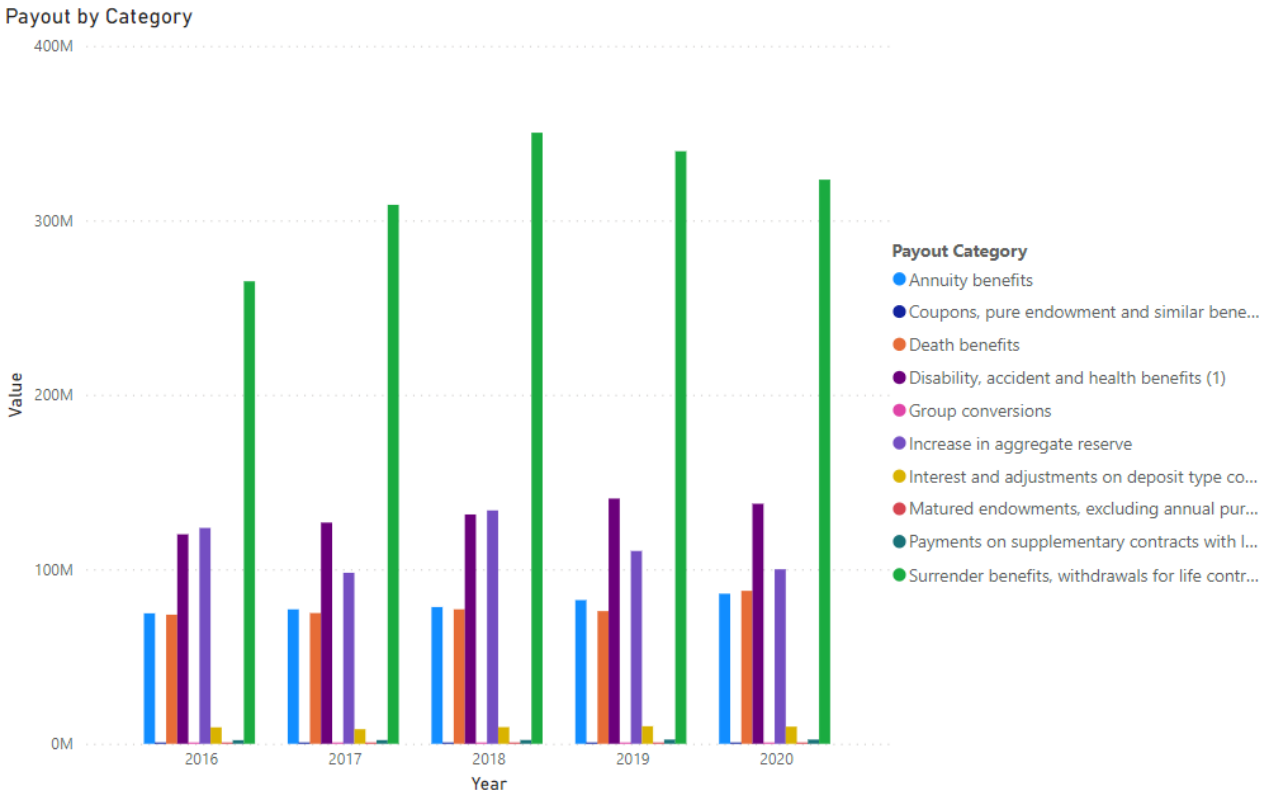


The 2 graphs shown below exhibit the general trends that from 2011 to 2020, the percentage of people with life insurance coverage goes down, while health insurance coverage goes up along with the ownership gap, or number of people who think they should have life insurance but do not. As such, the number of Americans overall who believe they should have life insurance, found by adding life insurance coverage and ownership gap, roughly stays the same. It may be that due to more common health insurance coverage, fewer Americans think that they need life insurance as well, and various other similar speculations could be made, but this information isn't really suited towards making any conclusive deductions. It is worth noting, however, that there is a sizable jump in ownership gap between 2019 and 2020 where the Covid-19 pandemic really hit the world in full force.

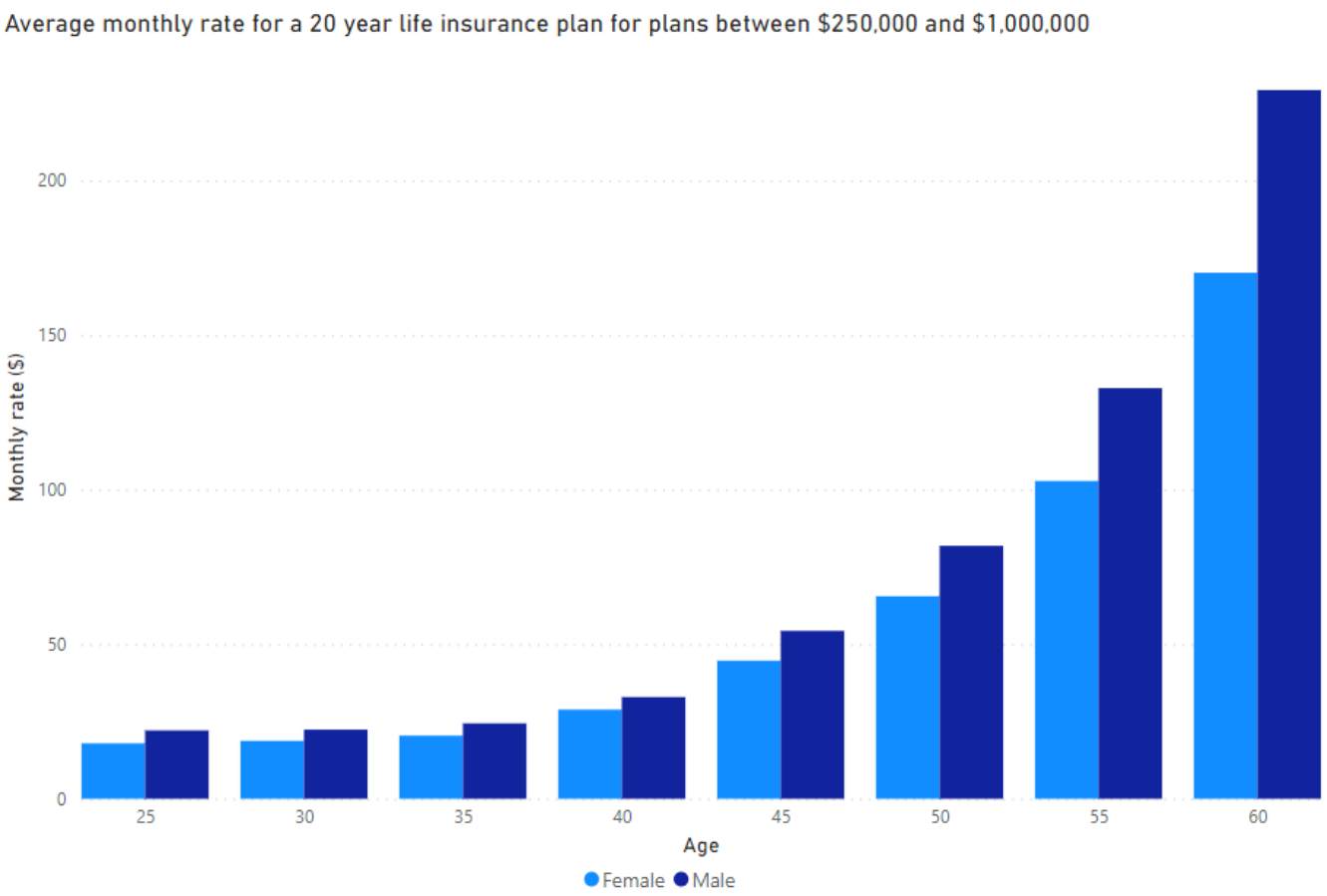




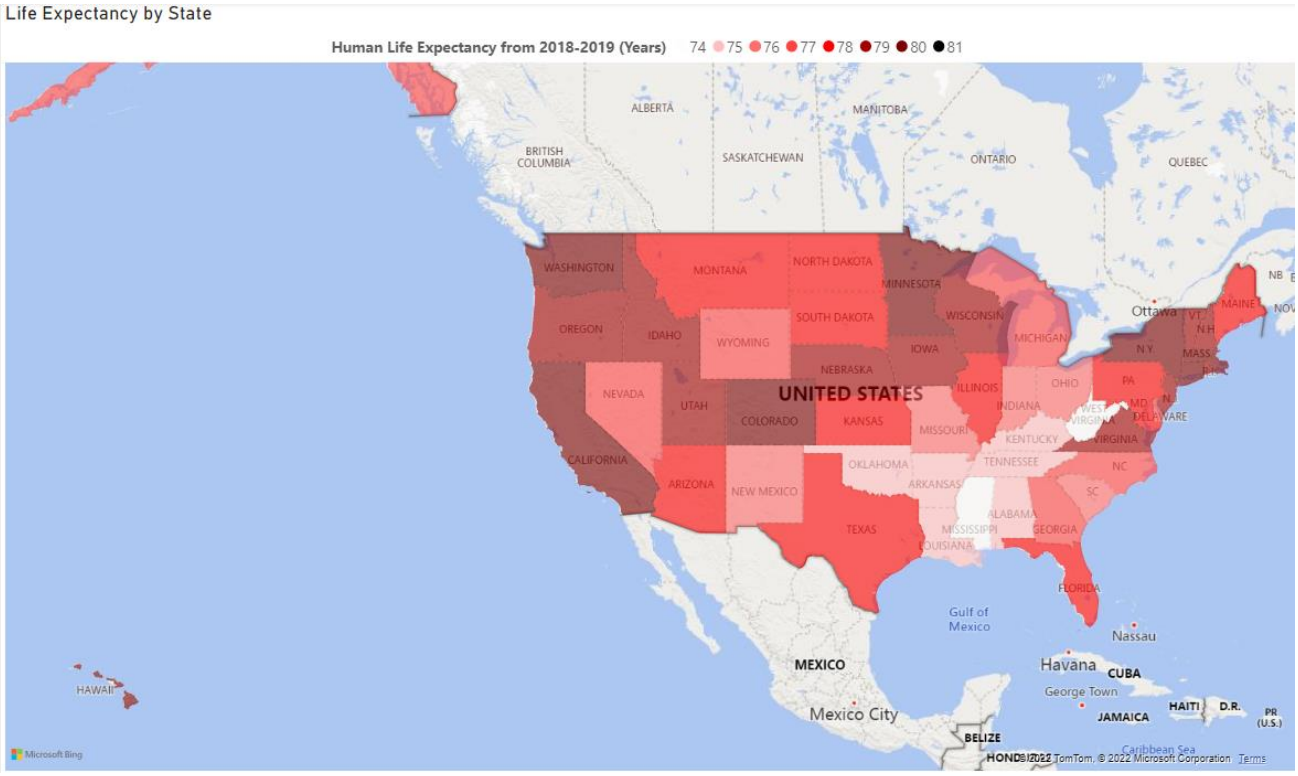
This graph showing the classifications of payouts by life insurance companies demonstrates that far and away the most payout money goes to people who surrender the benefits of the life insurance policy and withdraw earlier than their plan is active, as shown in green. Additionally, death benefits, shown in orange and what is hypothetically the main point of buying life insurance plans, doesn't even come close. This could mean one of two things: either that the reasons given for not holding life insurance are completely justified, as most people don't ever see the full benefit of their contracts, or that there's no reason to not buy life insurance, as it gives the customer peace of mind while in addition to the option to pull out and get some of the value of the contract back even if they don't die. Ultimately, this depends on what the customer's personal preferences and needs are.



As one would probably expect, it can be seen below that life insurance costs are generally higher for men than women and increase as one ages. As these are the costs of a 20-year plan averaged by various plan payout values, it's unsurprising that there are particularly sharp increases between 50-55 and 55-60, as these would last until one is approximately hitting the average US life expectancy by age 80. Still, it is a bit interesting that the male rates increase so much faster than the female ones, but this again may be due to differences in life expectancy, which will be covered later



The main things that stick out when seeing the life expectancies of various states shown below, are that the rural south is a bit low, while the western coast is markedly higher. While there are a multitude of possible explanations for this, such as regional values, state laws and regulations etc., all of these would require further study, as this graph doesn't contain any reasons or causes for the data shown.



ML Model

Life Insurance is very closely associated with death, so much so that in a survey referenced above about 40% of people surveyed saying they didn't have any listed thinking it morbid as a reason for that lack. Because of this, we decided to attempt to replicate a CDC life expectancy model using Machine Learning, as having a good idea how likely an individual is to die in a coming period would greatly impact term life insurance rates.

We had found a breakdown of newborn Life Expectancy by State for 2018 and 2019, and using that as the dependent variable, built a model using Sex, Race, and Ethnicity Percentages for the various states for the two years as well as Median Household Income and % of the State's Population with Health Insurance as the independent variables (all of this data coming from various tables from the US Census Bureau). These variables were chosen from discoveries made in the initial research stage, the male/female ratio because of the listed pricing disparity between the genders as outlined in a graph above, the race and ethnicity data from various other CDC sources which seem to indicate differences exists (https://www.cdc.gov/nchs/data/nvsr/nvsr68/nvsr68_07-508.pdf), and the census data as wildcards that at first blush seem likely to have an impact (having health insurance may help you take better care of yourself, and having money often helps problems). Once all the data was collected and merged, the last step taken was standardizing the data by casting the data in each column to its z-score.

Taking this standardized data we built several ML models, beginning with a basic Linear Regression Model. This first model had an R-squared value of .726, which while not great, does seem to indicate that at least some of the independent variables have some predictive value for such a model. To improve upon this result we decided to see how an SVR model would perform on the same data, and how well such a model could be tuned. The default SVR model yielded an R-squared of .787, already a

minor improvement on the Linear Regression, but still with decided room for improvement. Additionally, the Residual vs Fit Plot(fig1) appears to indicate a slight positive correlation between the residuals and predicted values, which could indicate the existence of some unaccounted-for trend or could be caused by the small sample size upon which our models were built. After some grid search hyperparameter tuning of the model we ended up with a tuned SVR model with an R-squared value of .941. This is much better than the default SVR model already discussed and examining the Residual vs Fit Plot(fig2) seemed to show the previous possible positive correlation to have waned, though not disappeared entirely, and a new issue in an apparent magnitude increase of the residuals as predicted values rose could also indicate some unexamined connection.

fig1

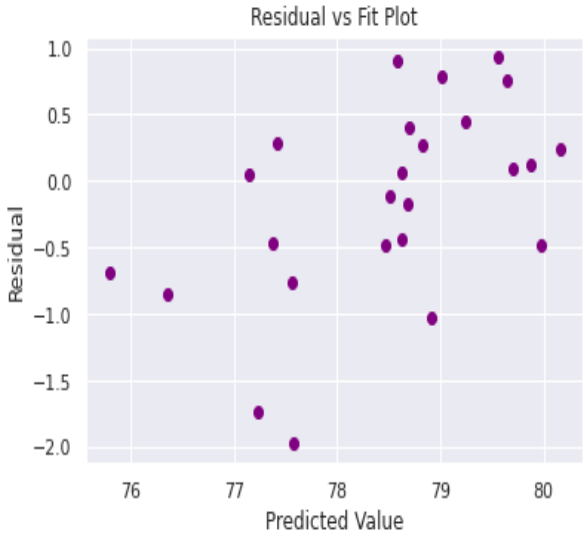
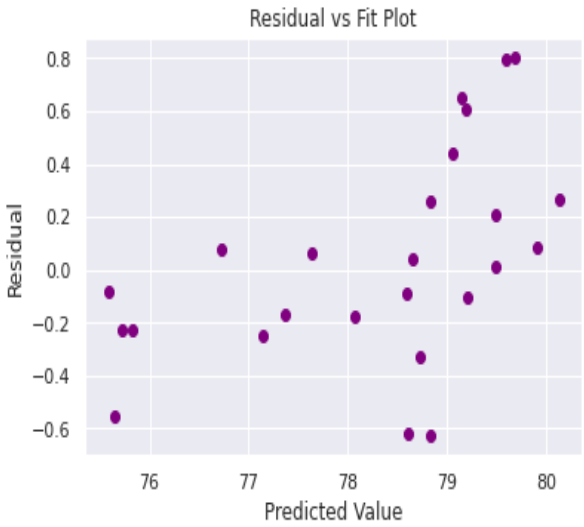


fig2

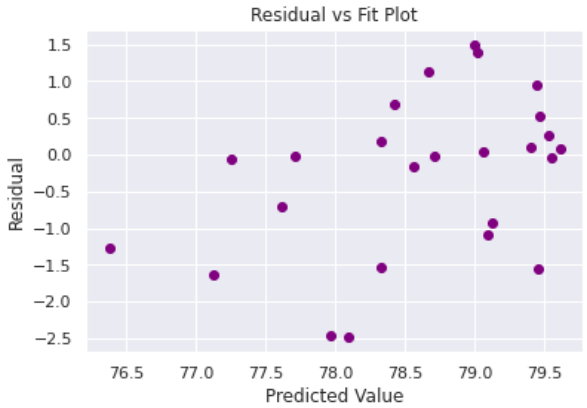


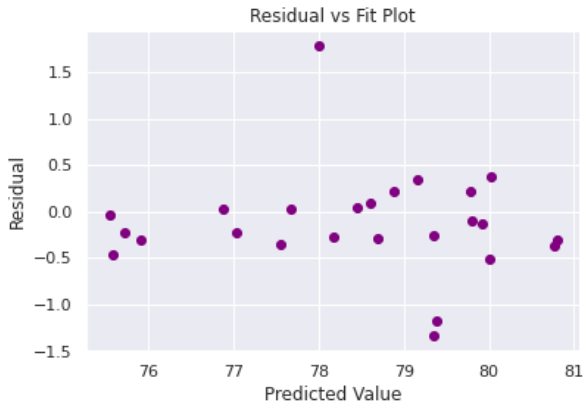
Returning to the Linear Regression Model, and examination of the coefficients would appear to indicate some factors having a greater impact on the end results than others, the greatest of which was one of our wildcards, Median Household Income. Looking at correlations between the independent variables and the dependent variable, the highest correlation was also with Median Household Income, and the other wildcard, % with Health Insurance, as also one of the higher correlations, so we decided to see how good a model could be made without these two variables.

Following the same process as before, a Linear Regression, default SVR, and tuned SVR model were built with R-squared values of .508, .534, and .879 respectively. Once again, the default SVR model's Residual vs Fit Plot(fig3) seemed to have a positive correlation between the residuals and predictions, which was significantly reduced in the tuned model but not altogether eliminated. Additionally, the tuned SVR model's Residual vs Fit Plot(fig4) has a few rather extreme outliers, likely causing the lessened R-squared value. This is quite likely a sampling error and would likely be better mitigated by a larger initial dataset.

fig3

fig4





Our models seem to work better when given information about Median Household Income and % of population with health insurance, and while at its best it does seem to replicate the CDC's projection data with a fair amount of accuracy, there are a number of drawbacks. First and foremost, as we were only able to find two years' worth of projection data, our model was built and tested on only 100 data points, so it is almost certainly the case that it is underfit. Secondly, as seen in the Residual vs Fit Plots, the residuals don't look terribly random, seemingly having a positive correlation with predicted values, and thus implying a possible trend the model fails to account for. As previously mentioned, this could be simply a consequence of the likely underfitting, or could indicate a need for a different model approach. Another couple of drawbacks to our models come from the impetus for their creation: use in a possible industry pricing model. Our models do not project life expectancy at an individual level but rather for a population with given characteristics, and were built to predict newborn life expectancy and not the life expectancy remaining for a group of adults. Additionally, even if taking what our models are built to predict into account, even the best one we made only matches the predictions at about 94% accuracy, which while great for a score on a math test, is not as high as would be desirable when used for making decisions representing hundreds of thousands or more dollars per instance.

To improve upon our models, the most important addition would be more data so the models can be trained better. Adding to that, as there appear to be still unexplained trends, adding more variables would likely improve the accuracy as well, provided more overall data can be obtained so as not to have too complex a model for the available data. If truly trying to build an actuarial death table model we would likely need to blow up the whole process and start fresh with very different data that we were unable to find, in large part because a lot of what we would be looking for is personal health data which is not publicly available, hence why we did what we did instead.

Conclusion

Overall, while we successfully answered the questions we had set out to in the beginning, there are certainly some large limitations to our findings. Our most recent data was from approximately 2020, and given how much the global pandemic of COVID-19 has changed the state of the world, there are likely some very significant differences between the world our data portrays and the world as it exists in 2022. Additionally, not having access to any sort of more detailed, propriety data regarding life insurance companies, policies, etc. severely limited the scope of our findings. Although we were able to demonstrate some various trends regarding life insurance and health insurance, we were unable to

pinpoint any of the reasons behind what we found. As such, this would be a logical next step for any further exploration of the matter.

Data Sources

https://www.bestliferates.org/statistics

https://www.investopedia.com/articles/personal-finance/022615/how-age-affects-life-insurance-rates.asp

https://www.iii.org/table-archive/22403

https://www.cdc.gov/nchs/pressroom/sosmap/life_expectancy/life_expectancy.htm

https://data.census.gov/cedsci

Search for the following tables:

- S1901 (ACS 5-year estimates 2019&2018)
- S2701 (ACS 5-year estimates 2019&2018)

https://www.census.gov/data/tables/time-series/demo/health-insurance/acs-hi.html

Get HI-05_ACS excel file for appropriate years

https://worldpopulationreview.com/states/state-abbreviations