Capstone Technical Report

Team 2: Christian Rodriguez, Philip Waymeyer, Savion Ponce, Victor Pham

Section 1 – Introduction

Team 2 analyzed data related to U.S. healthcare, including the cost of U.S. healthcare premium and the coverage of US demographics. Datasets were grabbed from public online sources, documented in a DataSource file, and described below.

- **US Health Insurance Dataset**: An annual culmination of individual medical costs incurred by people sorted by demographics of age, sex, region, smoker status, and number of children.
- **Indicators of Health Insurance Coverage at the Time of Interview**: Provides data for coverage by age, sex, race, level of education, and state of residence. In addition, it documents changes in insurance coverage between weeks from 23 Apr 2020 to 16 Jun 2020 (a period of 7 weeks). Data was gathered by the U.S. Census Bureau through the Household Pulse Survey conducted online.

  The methodology is documented here by the U.S. Center for Disease Control and Prevention: https://www.cdc.gov/nchs/covid19/pulse/health-insurance-coverage.htm
- **Small Area Health Insurance Estimates:** Provides data for percentage of individuals insured by age, race, sex, state, county, and income level. Income level is reported as a ratio between an individual's income, and the national poverty line. In addition, this income level aggregates smaller ratios depending on the income level label. Data was gathered by the U.S. Census Bureau in 2019.

The full documentation on the methodology can be found here: https://www.census.gov/programs-surveys/sahie/technical-documentation/methodology/demographic-income-model.html

The project had the following initial inquiries into the three datasets, also documented in the ExploratoryQuestions document:

1. What factors affect health costs incurred by health insurance customers? (age, BMI, number of children, region, sex, smoker status)
2. How do factors of education level, U.S. state location, and sex affect insurance status?
3. How does income affect the percentage of individuals insured?
4. What specific demographics have the highest percentage uninsured?
5. How can the project predict health insurance cost based on age, sex, BMI, region, number of children, and smoker status, using a machine learning algorithm?

Section 2.1 – Research Presentation: US Health Insurance dataset

While researching the US Health Insurance dataset, we analyzed certain factors that may affect the individual health costs incurred by healthcare customers. These factors include age, body mass index (BMI), amount of children of the insured, region in the United States where the individual resides, sex, and smoker status. Figure 1 below displays the relationship between age and health costs. When first looking at this scatter plot, it is evident that there are three different groups of healthcare costs with an upward trend. From our exploratory data analysis, we found that healthcare costs have the greatest correlation with smoker status. Therefore, we decided to

plot smoker status along with this scatter plot, as well as the following scatter plot. By plotting this factor, it emphatically helps in depicting not only the relationship between age and healthcare costs, but also, the effect smoker status has on the cost of insurance for individuals, which we will discuss further in depth later on.

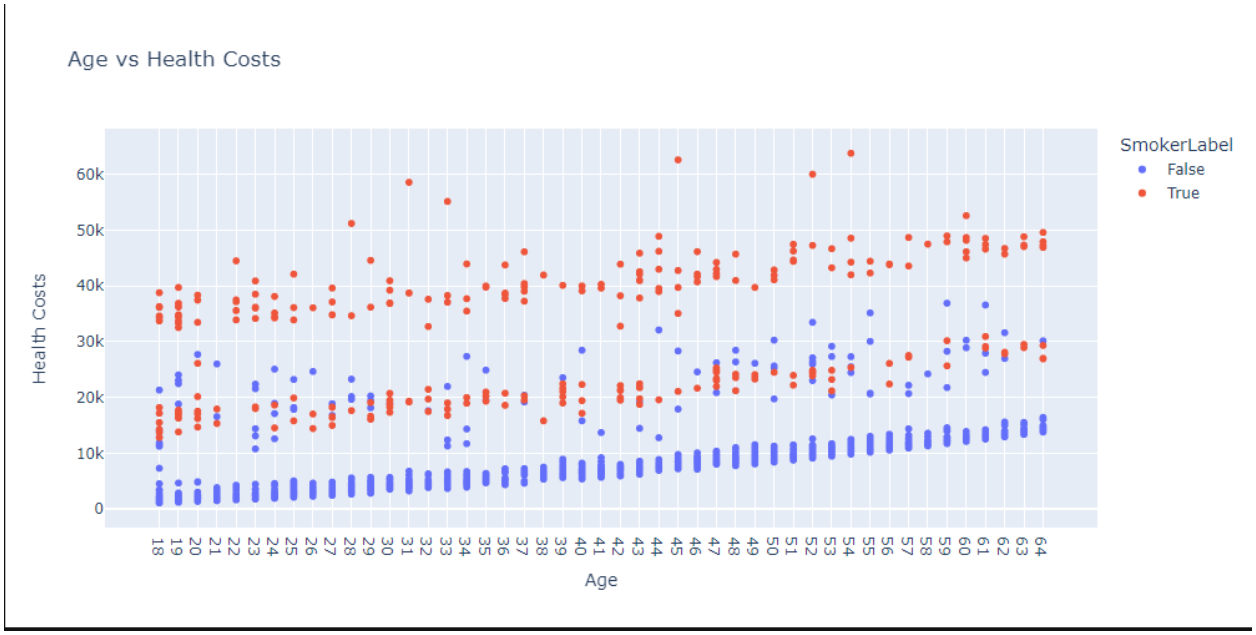

Figure 1: Age vs Health Costs

Figure 1 shows Healthcare Costs vs Age, with Health Costs listed in thousands of USD. Each data point on this plot represents a different combination of factors and its corresponding medical costs. As you can see, as the age of the individual increases, so does the cost. Furthermore, since smoker status has the greatest effect on the cost of health insurance, we wanted to see the results when it was combined with age. The scatter plot depicts that if you are older and a smoker, then you will incur more medical costs than if you are younger in age and a non-smoker.

Similar to the previous graph, we were looking to see if there was any relationship or trends between the body mass index (BMI) of individuals and health insurance costs. Since we are using the same “InsuranceCharges” dataframe as we did for the prior plot, we do not need to filter anything out; we only needed to switch the column that was being graphed from age to BMI to display our findings. The graph is shown below in figure 02.

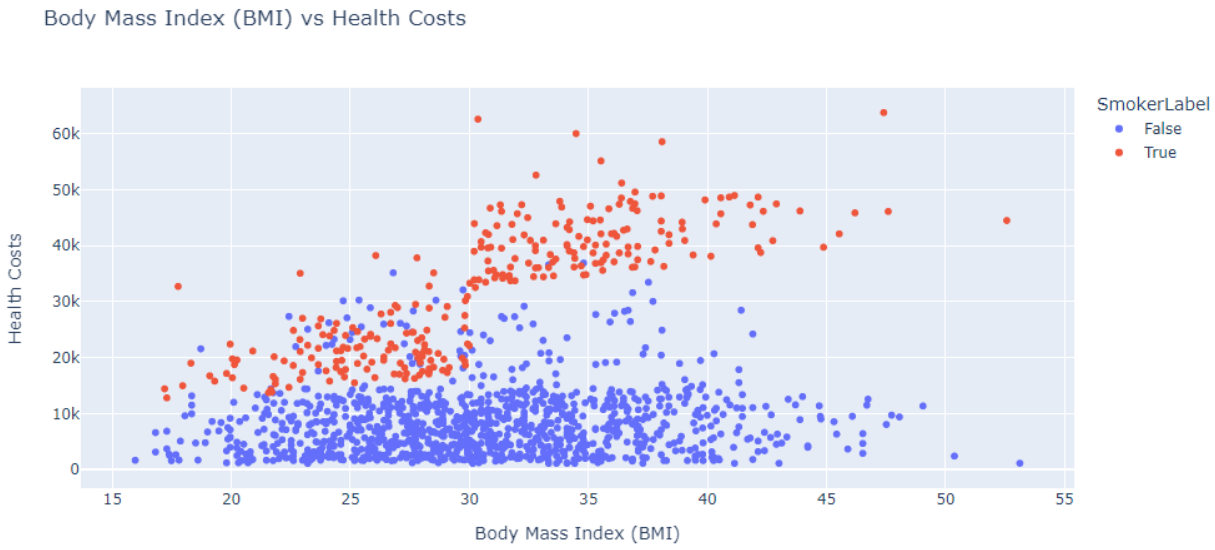

Figure 2: BMI vs Health Costs

Figure 02 displays Health costs versus Body Mass Index. It shows a discontinuity when the individual's BMI hits about 30, there seems to be a break in the amount of cost into two groups. As previously stated and in the previous plot, we saw the smoker status could provide some insight on the difference between the groups. We decided to plot the smoker status along with the BMI, and as you can see, the more expensive of the two groups is almost entirely filled by smokers. Once the BMI hits 30 and the smoker status is true, there is a steep increase in the

healthcare costs incurred by an individual. Furthermore, if the smoker status is true and the BMI is larger, the relationship with healthcare costs is positive. On the other hand, if the smoker status is false and no matter the BMI, there seems to be little to no correlation or relationship with healthcare costs.

In addition to considering healthcare costs, another factor we looked at was the number of children an individual had and its effect on the cost of healthcare. In our dataset there were only six values present in the children column; the values were no children, one child, two children, three children, four children, or five children. There was an overwhelming amount of records for individuals with no children so instead of getting the totals for each group, we decided to compare the average costs for each group. We created a bar graph to display our findings below:

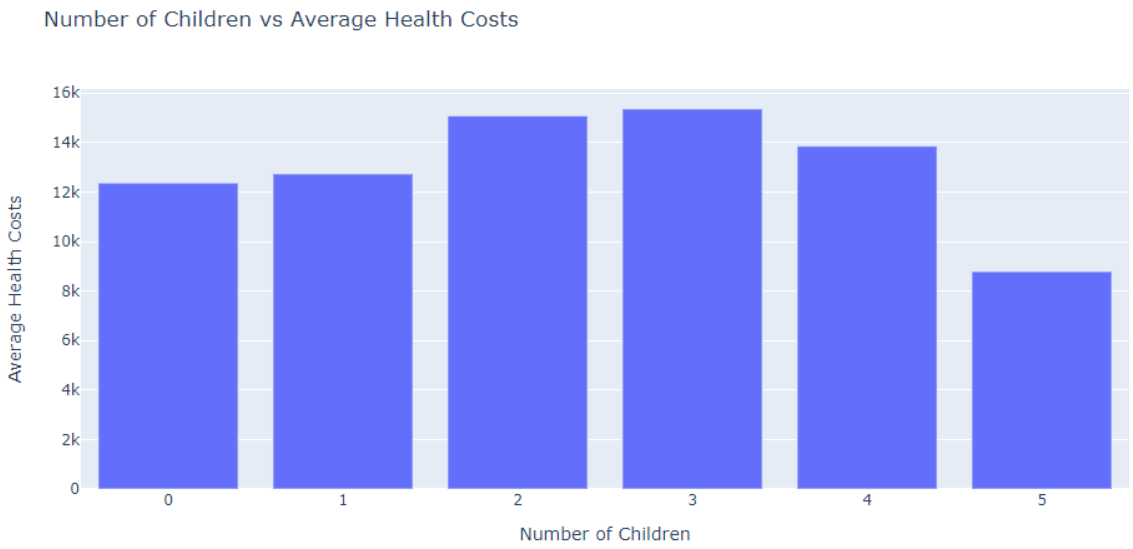

Figure 3: Number of Children vs Average Health Costs

Figure 3 lists the number of children on the x-axis and average healthcare costs, in thousands of USD, on the y-axis, ranging from 0 to $16,000. The averages from groups 0-4 are all similar in amount, with people with three children averaging the highest healthcare cost. People with no children and one child have an average healthcare cost of just over $12,000. Individuals with two or three children have an average of about $15,000. People with four children have an average of about $14,000. Individuals with five children seem to be the least average of healthcare costs, coming in at about $9,000.

Similar to the previous graph, we decided to observe which region contained the most capital in healthcare within the dataset. A difference between the two graphs is that instead of finding the average of each region like before for the number of children, we thought it would be interesting to find the total amount for each region. We have constructed the graph below to display this:

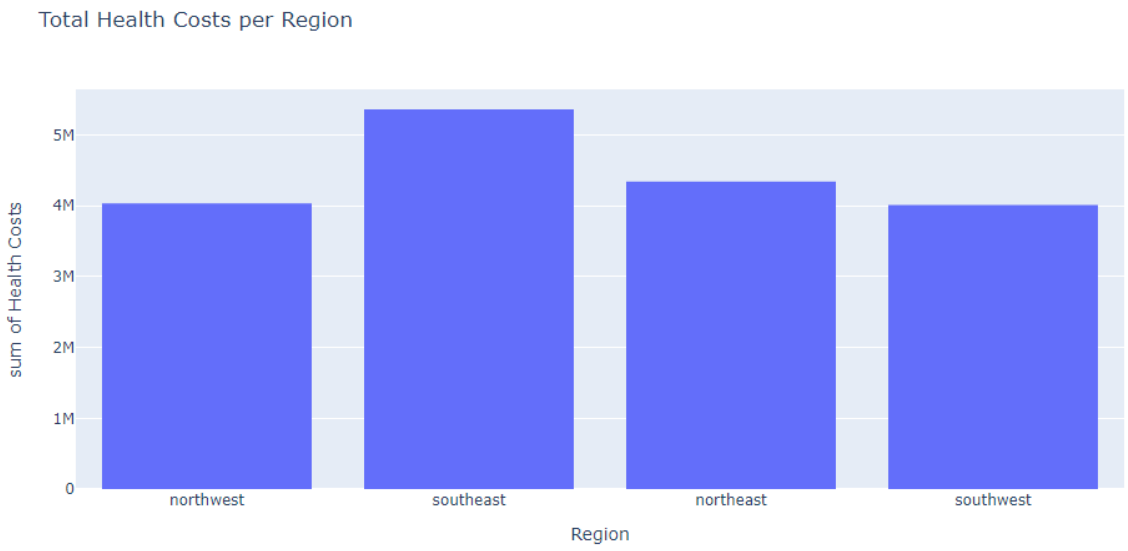

Figure 4: Total Health Costs per Region

Figure 4 depicts the sum of healthcare costs per region, in millions of USD, on the y-axis, and the regions northwest, southeast, northeast, and southwest on the x-axis. The northwest, northwest, and southwest regions are all similar in the total amount of healthcare costs amongst individuals, coming in at around four million dollars. The southeast region contains more than a million dollar gap from the next highest when it comes in at about $5.3 million. After looking at this graph it seems as if the region of where an individual lives does not have much effect on the total cost of healthcare expenses. The only observation that could be taken from this is that if you live in or near the Bible Belt, then you might incur more healthcare costs.

The final two factors that we did observations to see if they had any effect on healthcare cost, were sex and smoker status. We constructed box plots for each category to get the five-number summary of each column; the five-number summary is the minimum, first quartile, median, third quartile, and maximum. After glancing at the sex box plot, what seemed interesting was the large amount of outliers for both female and male healthcare costs. The majority of individual costs for females range from about $1,600 to $28,000, and the costs for males range from about $1,100 to $40,000. For both plots, there are many outliers that are above the 75th percentile for the column. Since these outliers, fall outside of the normal distribution, we believe that other factors have other than sex, have cause the healthcare cost to increase.

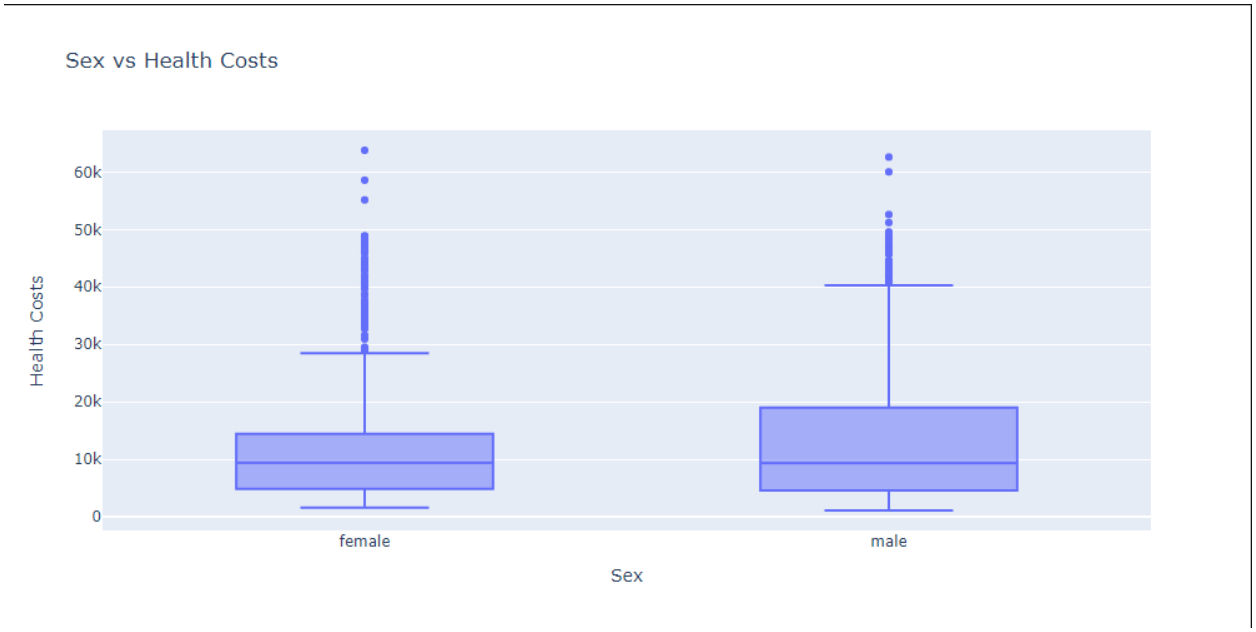

Figure 5: Sex vs Health Costs

Figure 6: Smoking Status vs Health Costs

Like the “Sex vs. Healthcare Costs” box plot, figure 6 compares smoking status with healthcare costs. This gives us the five-number summary of the smoker column in the

“InsuranceCharges” dataframe. A contrast with this box plot and the prior, is that the difference between smoker and non smoker is dramatically larger than the difference of the prior box plot that described sex. We ran a t-test to see if there was a statistical significance between the health costs of smokers and non-smokers. The null hypothesis was that charges of smoker and non-smoker are the same and we set our significance level to five percent. Using scipy.stats ttest_ind, we calculated a p-value of 8.2^{-283} which suggests that charges of smokers and non-smokers are not the same since the p-value was below our significance level of 0.05. The majority of individual’s cost for non-smokers range from about $1,100 to $22,000, and the majority of smokers' costs range from about $13,000 to about $63,000. Also, there seems to be many outliers for non-smokers, but for smokers there are none. Since there are many outliers for non-smokers, this shows that the other factors are causing the price for certain individuals to increase above the 75% quartile of the distribution of non-smokers. On the other hand, due to the fact that there are no outliers being represented on the smoker plot, this shows that yes, other factors are influencing the price of healthcare, but the status of being a smoker has the greatest influence on that individual’s cost for healthcare.

Section 2.2 – Research Presentation: Indicators of Health Insurance Coverage

The Indicators of Health Insurance Coverage (Indicators) dataset includes data on the percentage of selected demographics which are privately insured, publicly insured, or uninsured. This data was obtained by the CDC in a 7 week interview process during the spring of 2020. While researching this dataset, there was a clear trend seen with education level and percentage

of people uninsured. Further researching also brought up questions about sex and location, so visualizations were made to answer these questions as well.

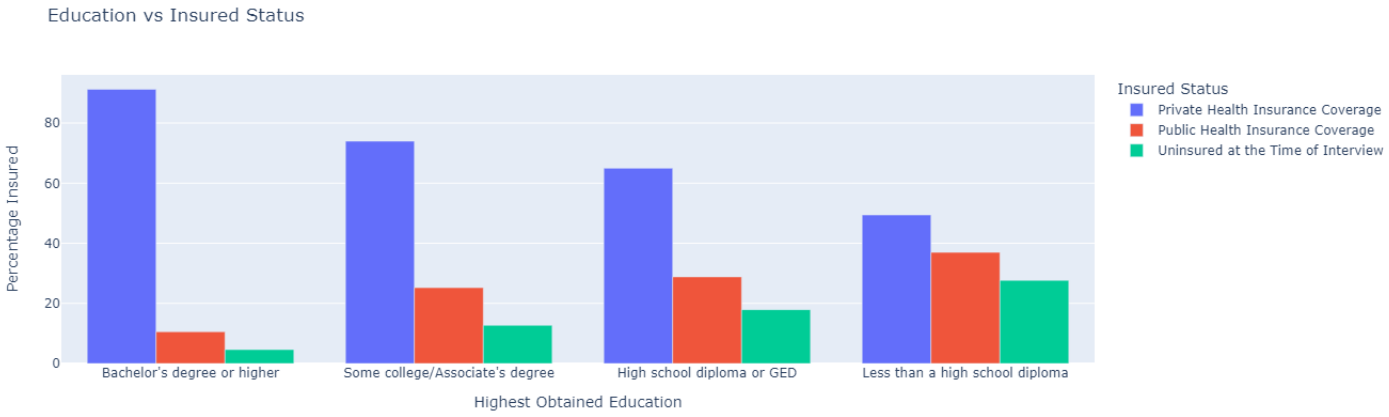

Figure 7: Education vs Insured Status

Figure 7 was the first visualization made with the Indicators dataset, and it shows the percentage for each level of education broken down by insured status. This bar chart shows that with a higher education level, the percentage of people with private insurance increases, while the percentage that are publicly insured or uninsured decreases. This observation is important because it suggests to insurance companies that there may be more success in marketing towards people with less education.

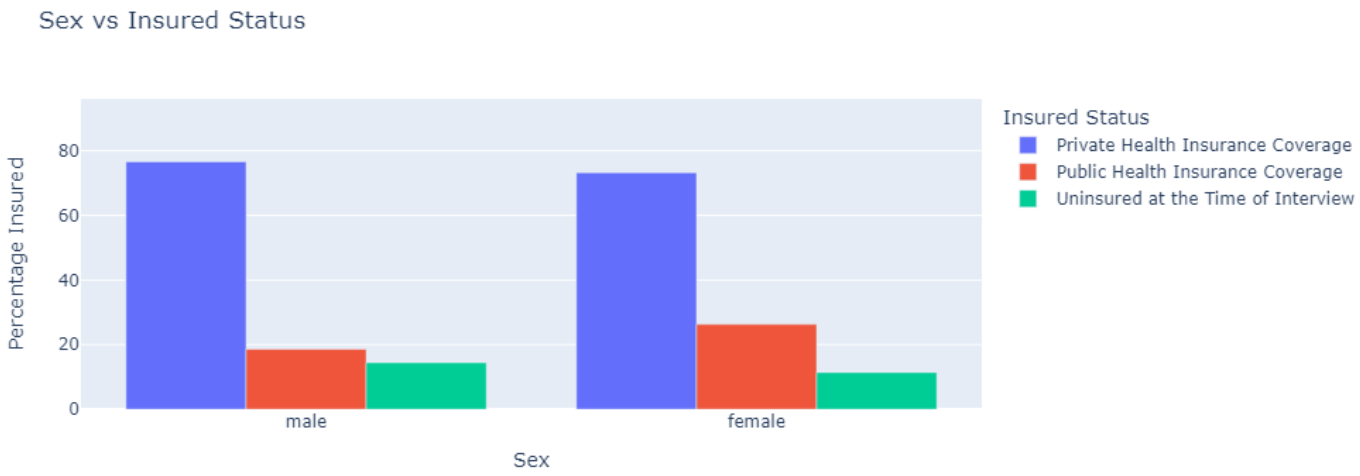

Figure 8: Sex vs Insured Status

The next visualization made was figure 8, which shows the percentage of male vs female broken down by insured status. This bar chart shows that males had a slightly higher percentage of privately insured people and uninsured people, while females had a slightly higher percentage of publicly insured people. This graph may suggest that marketing towards males over females could be more successful, though it is a minor difference so it's not a strong suggestion**.**

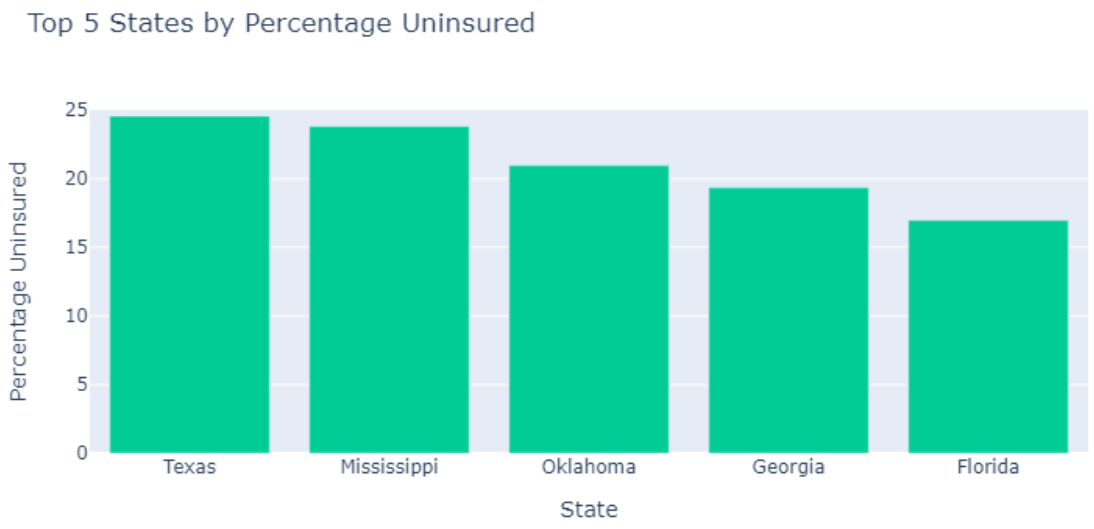

Figure 9.1: Top 5 States by Percentage Uninsured

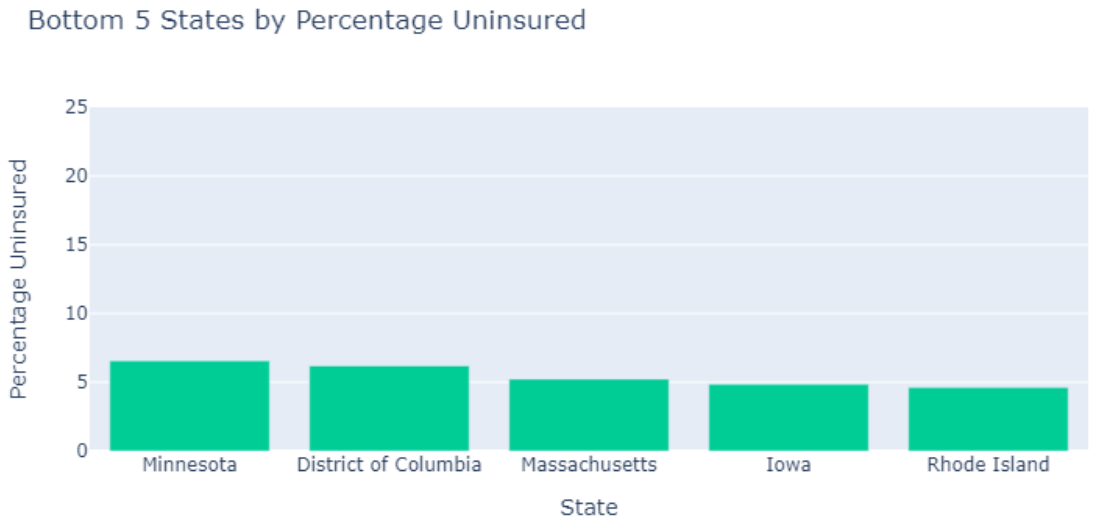

Figure 9.2: Bottom 5 States by Percentage Uninsured

Figures 9.1 and 9.2 show the top end and bottom end of the percentage of people uninsured by state. These graphs show that states like Texas and Mississippi in figure 9.1 have

nearly 5 times the percentage of uninsured people as compared to states like Rhode Island and Iowa in figure 9.2. This pair of graphs could prove to be useful to nationwide insurance companies who wish to market towards states with a smaller number of insured residents.

Section 2.3 – Research Presentation: Small Area Health Insurance Estimates (SAHIE) dataset

The SAHIE 2019 dataset presents data on percentage of selected demographics insured as estimated and gathered by the U.S. Census Bureau. This particular dataset was chosen to appropriately reflect the amount of certain populations that are insured, allowing insurance companies to focus marketing and advertising to uninsured groups that are more likely to sign onto a plan as new customers.

After cleaning, documented within the accompanying ETL file, the SAHIE file contains the following breakout groups, with respective available group labels:

- Age
    - Under 65 years
    - 18 to 64 years
    - 21 to 64 years
    - 40 to 64 years
    - 50 to 64 years
    - Under 19 years
- Sex
    - Male

    - Female
    - Both sexes
- Race
    - Hispanic or Latino, any race
    - Non-Hispanic black (single race)
    - Non-hispanic white (single race)
    - All races
- U.S. State
- Geographic breakdown
    - State geographic identifier
    - County geographic identifier
- County (left N/A if the breakdown is by state)
- Income
    - All income levels
    - At or below 138% of poverty
    - At or below 200% of poverty
    - At or below 250% of poverty
    - At or below 400% of poverty
    - Between 138 - 400% of poverty

Note that an individual can qualify for multiple categories, i.e. an individual with an income of 150% of the national poverty line will be included in calculating the "At or below 200%", "250%", and "400% of poverty" statewide percentages and consequent means.

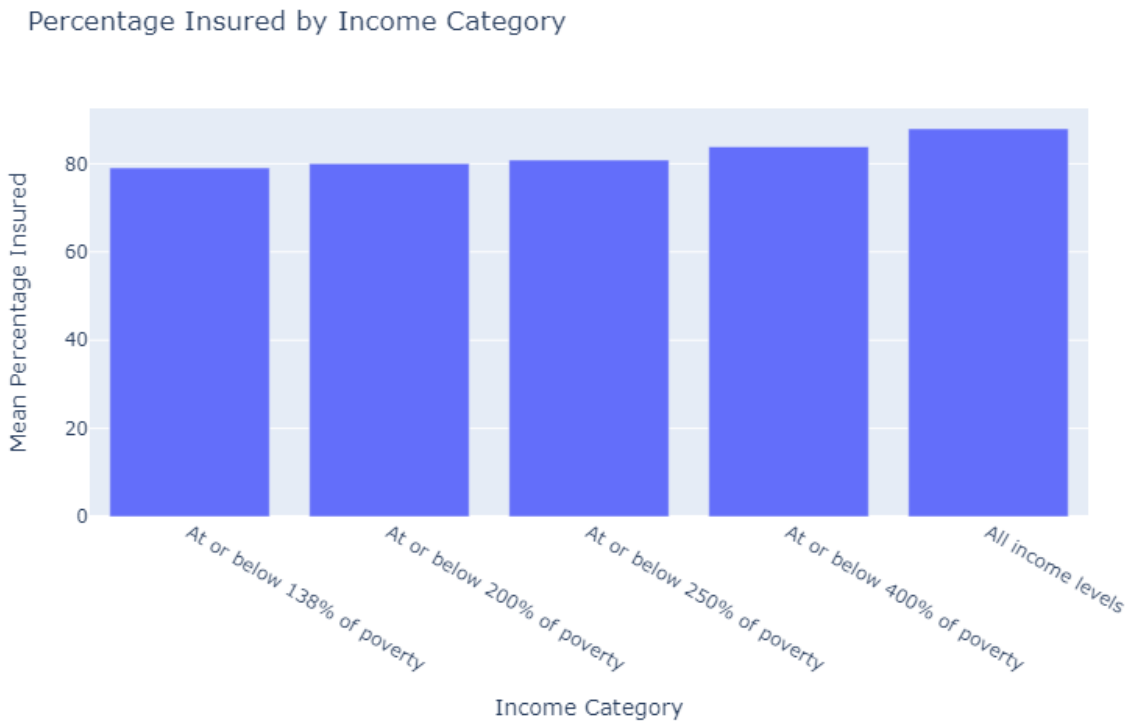

Figure 10: Percentage Insured by Income Category

Figure 10 shows the percentage of the population insured averaged over the four income groups, as well as for all income levels. The label "All income levels" includes all explicit income levels and individuals making more than 400% of poverty.

The data suggests that individuals with higher income are more likely to be insured. Individuals at or below 138% **IPR (income-poverty ratio)** are 79% likely to be insured, while individuals at or below 400% IPR are 84% likely to be insured. Over all income categories, any given individual is 88% likely to be insured.

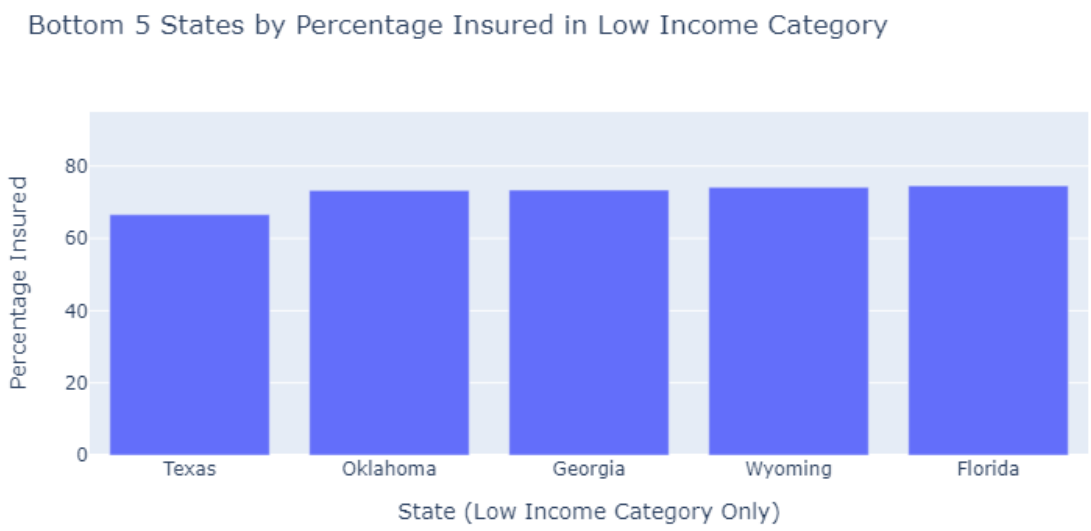

Figure 11.1: Bottom 5 States by Percentage Insured in Low Income Category

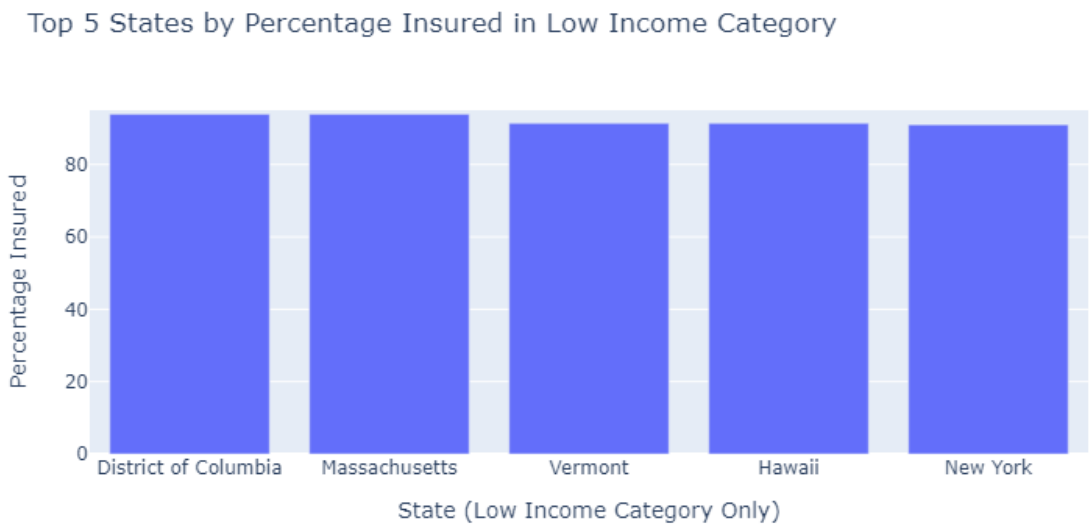

Figure 11.2: Top 5 States by Percentage Insured in Low Income Category

Within the lowest recorded income category (individuals at or below 138% IPR), the top and bottom 5 states for mean insurance were selected, depicted in figures 11.1 and 11.2. No

calculations were necessary; constraints were placed to include all ages under 65, both sexes, and all available races, while framing the data on the state level and at the aforementioned income level.

The lowest insured states range from having 66.6% to 78.1% of their citizens insured, while the highest insured states range 88.8% to 93.9%.

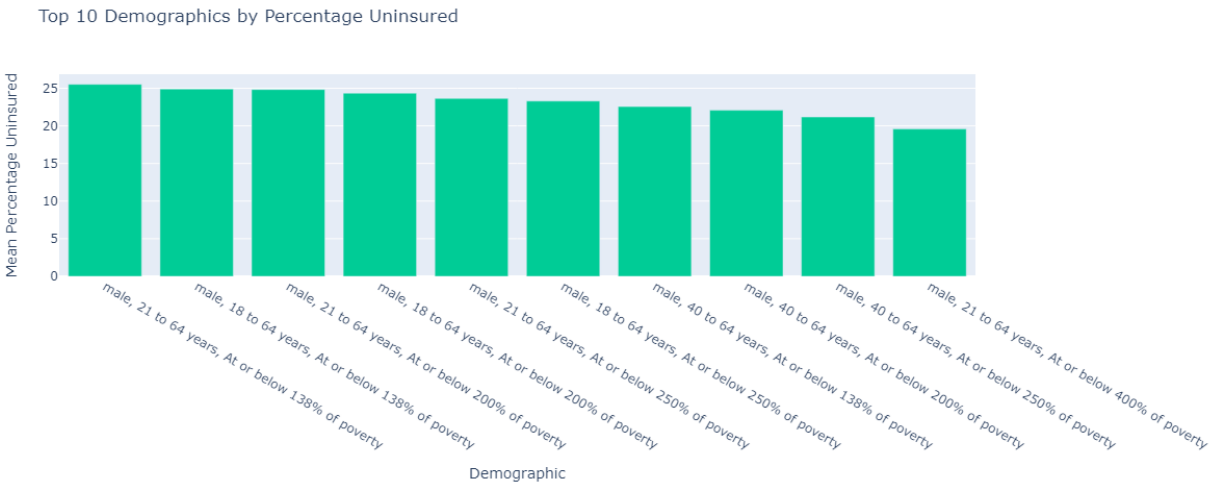

Figure 12: Top 10 Demographics by Percentage Uninsured

The final graph continues to describe the typical uninsured population by looking at the most uninsured demographic by sex, age range, and income level, while averaging over all states. Race as a demographic factor was constrained and omitted due to the small and incomprehensive number of labels, though it was noted during exploratory data analysis that the "Hispanic/Latino, any race" label had the lowest insurance chance of the three options of race; further datasets would be needed to fully explore this possibility. Aggregated labels were excluded to frame the data using the highest resolution demographic; as a result, labels such as "Both sexes", "All income levels", and "Under 65 years" were deselected.

The most uninsured groups tended to be males with less than 200% IPR, with an age range of 18-64 and 21-64. In sum, young adult males with low income are the most likely to be uninsured.

Section 2.4 - Machine Learning Model

Our goal was to generate a machine learning model to predict the health costs that will be incurred by future health insurance customers. The process began examining and cleaning the data as outlined in the ETL report. After cleaning the data, we turned categorical values (Sex, Region) into a series of boolean values using pandas get_dummies. We then split the dataset into a train and test set using sklearn's train_test_split with a test_size of 20%. For our base model, we started with sklearn's LinearRegression algorithm on our training data set. The model scored an R^2 value of 0.74 when scored against the test data and the performance can be seen in figure 13 below. Data points along the diagonal are fitted well with the model. As we can see from figure 13, the model does not perform very well in the range of about $15k to $30k. In this range the model both over-predicts and under-predicts medical costs incurred by individuals.

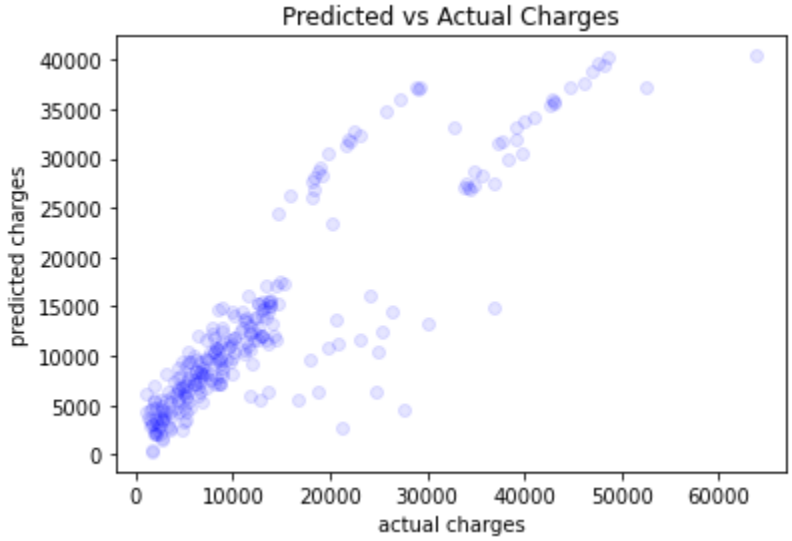

Figure 13: Predicted vs. Actual Charges accrued by individuals for medical costs

Since our model was using sex to predict health costs, we explored how the model would do without sex in order to avoid ethical concerns. We created a base model as the one described above, but without feeding it sex labels. The linear regression model with no sex labels performed the same as the linear regression model with sex labels (with an R^2 score of 0.74). Since the two models performed that same on the the same train and test data, we proceeded to use the model that did not use sex labels to avoid ethical issues.

We then moved on to improve the model by tuning hyperparameters. We tried using a Lasso model and running GridsearchCV with five folds and a logspace range of alphas from 10^{-3} to 10^{3}. When running the GridsearchCV, we noticed that a wide range of alphas below ten had little-to-no difference on the cross-validation score. However, there was a steep drop-off when the alpha crossed over one hundred as seen in figure 14 below.

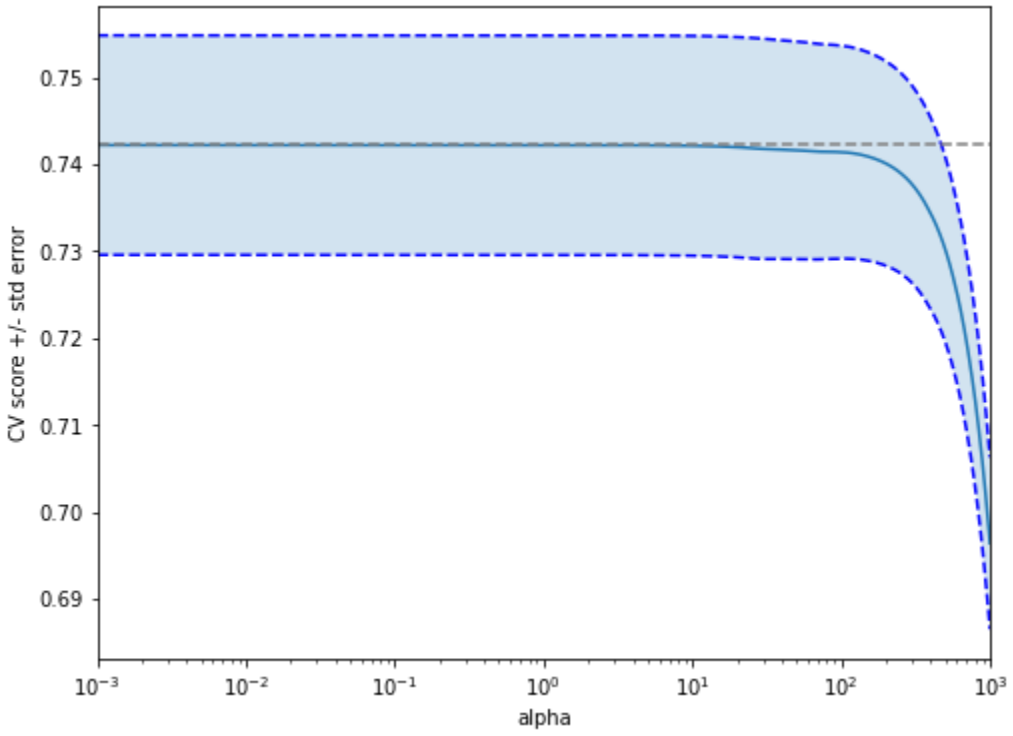

Figure 14: Cross Validation model performance as a function of alpha

The GridsearchCV suggests lower alphas perform better. Having an alpha value of zero is just a typical sklearn LinearRegression() which is what was used in our base model. The next step we took to improve the machine learning model was to apply feature engineering. We applied a pipeline to our test data before feeding it into the LinearRegression algorithm. The pipeline consisted of sklearn's PolynomialFeatures, which takes in the dataset and creates a new data matrix with all polynomial combinations of features with degree less than or equal to the specified degree. For example, if we have our training data, X = [a, b], then PolynomialFeatures(degree=2) would return X' = [1, a, b, ab, a^2, b^2]. Once again, we ran a GridsearchCV with polynomials in a range from zero to eight and seven cross validations. The best model performance was that with PolynomialFeatures(degree=3) which had an R^2 value of 0.87 when scored against the test data.

The test data can be seen in figure 15 and is plotted as predicted costs vs actual costs. Additional features such as smoker status are observed by marker shape; figure 15.1 has a color gradient corresponding to BMI values while figure 15.2 has the same color gradient corresponding to age. It is evident from both these plots that the model performed significantly better where there was more data (as expected). However, it is surprising to note that the model performed well in the higher range (i.e. above $30k) while it did not perform very well in the mid-range (i.e. ~$13k-$23k) even though the mid-range has more data points. This could be because the higher range has more defined grouping (e.g. has more smokers) while the middle range has a higher mix of values from BMI, Age, and smoker status.

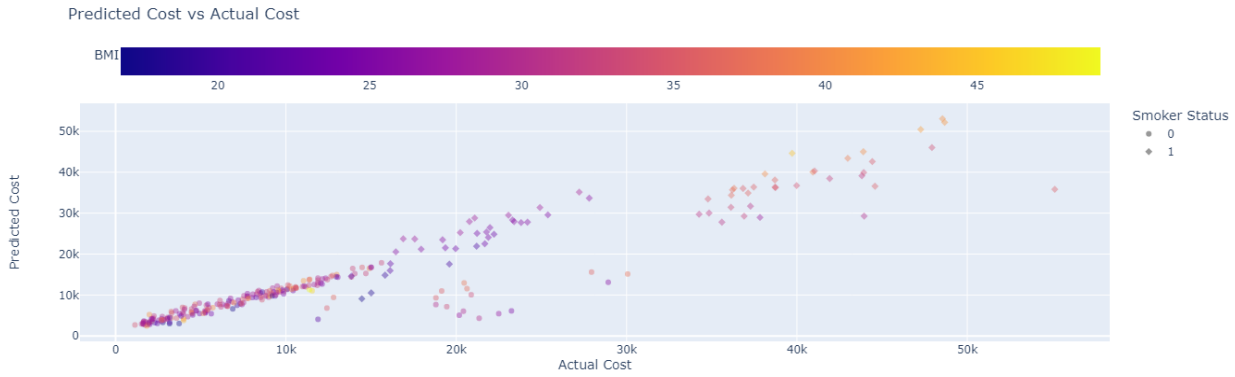

Figure 15.1: Predicted vs Actual Costs accrued by individuals for medical costs. Smokers are designated by diamond markers and non-smokers are circle markers. Marker colors correspond to BMI with purple (dark) being the low end of the BMI spectrum and yellow (light) being the high end of the BMI spectrum.

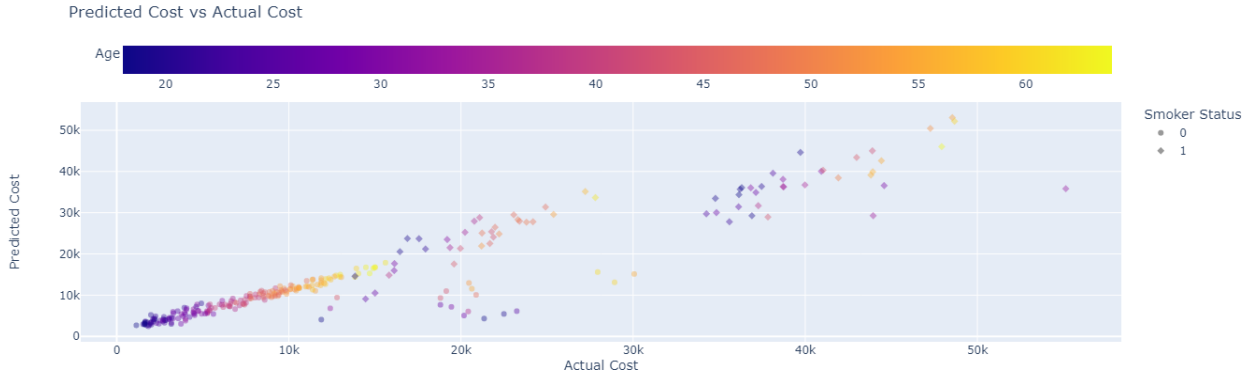

Figure 15.2: Predicted vs Actual Costs accrued by individuals for medical costs. Smokers are designated by diamond markers and non-smokers are circle markers. Marker colors correspond to age with purple (dark) being the low end of the age spectrum and yellow (light) being the high end of the age spectrum.

Section 3 – Conclusion

Through Team 2's research, insurance companies are advised to consider which demographics/locations have both a high rate of uninsured people and low health costs. Through

these strategic decisions, profit margins can be maximized through effective marketing towards these demographics. In particular, marketing towards low-income young men nationally, groups with lower education, and focusing on less insured states such as Texas would prove to be effective. In respect to company costs, young non-smokers have the lowest health cost, which indicates a market for capitalizing profit. For a better grasp on the predicted health costs for a potential customer, our designed model can predict costs by inserting corresponding attributes.