U.S. Medical Insurance Costs

A Codecademy Portfolio Project

Michelle A. Caler

November 23, 2020
(Updated May 12, 2021)

Outline:

Introduction

Health insurance costs are of concern to policyholders and insurance companies alike; while individual policyholders may be interested in reducing their out-of-pocket costs, health insurance companies may be interested in predicting policyholder costs to make better business decisions about the policies they offer. Additionally, health insurance companies may be interested in seeing if there are patterns in their data as regards policyholders and the charges they incur, to determine for example if certain regions of the United States tend to have higher costs than other regions, or if certain lifestyle choices tend to result in higher charges. These considerations and others will no doubt become amplified in the aftermath of the COVID-19 pandemic, but looking at historical healthcare data could provide insight into cost patterns.

To that end, this Codecademy Portfolio project focuses on finding patterns in the health care costs recorded in one particular toy dataset. The toy dataset used was provided by Codecademy and is an open dataset in the public domain[1]. It was left to the Codecademy Pro user what questions to address and how to go about addressing them.

Two main questions, and one secondary question, were investigated. The first main question asked, "How do health insurance costs, body mass index (BMI), and smoking rates vary by region in the United States?" It is found that the median cost is highest in the Northeastern U.S., while the average cost is highest in the Southeastern U.S. The cost distributions do not seem to be significantly different from each other. BMI is, on average, highest in the Southeastern U.S., which also has the highest smoking rate. The distribution of BMIs for the Southeastern U.S. is significantly different from the distributions of the other regions. However, the smoking rates between regions were not found to be significantly different.

The secondary question asked is related to the first main question; that secondary question asked, "Do policy holders with children tend to have higher BMIs than those without children?" It was found that there is no statistical difference between the BMIs of policy holders without children and those with children.

The second main question asked, "Is it possible to predict individual policyholder costs using a supervised machine learning model?" Not only is it possible to build such a model, but it is possible to construct one with a high goodness of fit ($R^2 = 0.91$) which provides a fairly accurate description of the cost data in the dataset. A k-Nearest

[1] Source: https://www.kaggle.com/mirichoi0218/insurance by way of Codecademy.

Neighbors Regression analysis resulted in the highest goodness of fit. However, due to the limited information in the provided dataset, the model constructed herein should not be applied to larger, more robust datasets.

The remained of this paper is structured as follows. In the Data section, a description of the data, as well as several visualizations of it, is provided. In the Analysis and Results section, the methods used to address the questions posed above are presented, as well as the results of that analysis. The Conclusions section summarizes the project's main findings. Additional details of the statistical and analytical methods utilized in this work can be found in the Appendices, as can data visualizations not presented in the main text.

In what follows, the primary beneficiary shall be referred to as "customer," and the medical costs billed by the beneficiary's health insurance shall be referred to as "charges."

Data

A dataset of health insurance charges, as well as anonymized data of customers, was provided by Codecademy. Included in the 1,338 entries was information about customer age, sex, body mass index (BMI), smoking status, region (in the U.S.) of residence, and charges (in U.S. dollars). No information about year of collection and/or any other timeframe for the data was provided.

To aid in the building of a regression model of the data, sex and smoking status were converted to either a 0 or a 1; details of the conversion process can be found in Appendix A4. An initial predictive equation for the data, provided by Codecademy[2], was used to calculate predicted costs.

To better visualize the data, several figures are included. The first is a histogram of charges; it can be found in Figure 1. This figure was constructed to see what the overall distribution of charges was like.

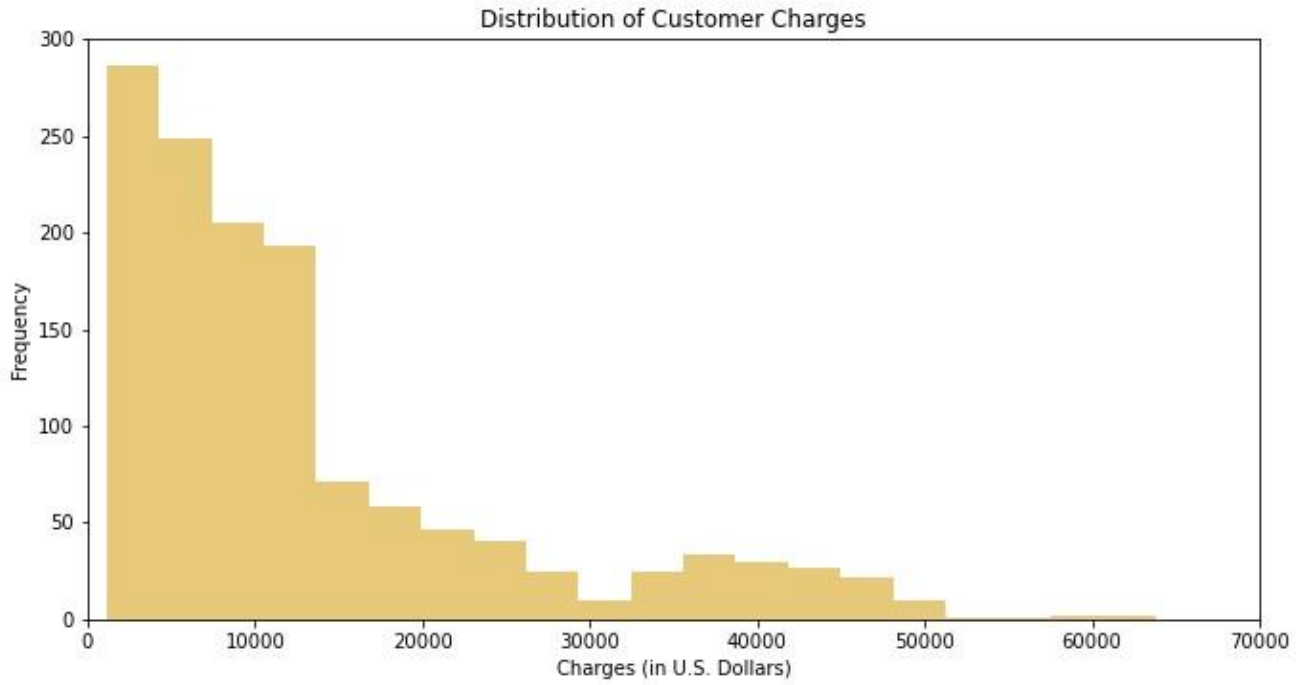

Figure 1: A histogram of customer charges. Two distinct peaks can be seen in the distribution, as well as a steep drop-off after roughly $12,000.

[2] https://www.codecademy.com/paths/data-science/tracks/dscp-python-fundamentals/modules/dscp-python-functions/projects/ds-python-functions-project

It can be readily seen from Figure 1 that most customers incur less than about $12,000 in charges, and many less than $10,000. While there are fewer customers who incur higher charges, there is a significant spread in those charges; in fact, a secondary peak in the distribution of charges occurs at roughly $38,000. So while customers who incur charges in excess of $12,000 are the minority, the charges they do incur can be substantial.

The next visualization to be investigated is a histogram of customer body mass indexes (BMIs), which is found in Figure 2. This figure was constructed to see what the distribution of customer BMIs was like.

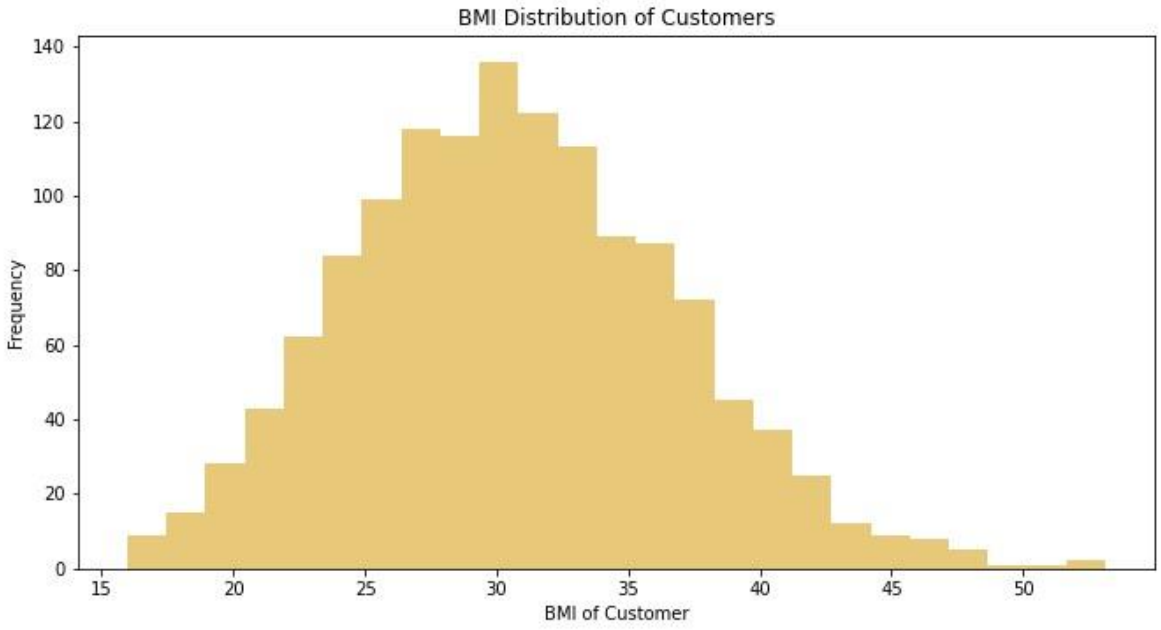

Figure 2: A histogram of customer BMIs. The distribution roughly follows a "bell curve," or normal distribution.

The distribution of customer BMIs appears to approximately follow a bell curve. The peak of that bell curve occurs at a BMI of roughly 30, which falls into the "obese" category as defined by the Centers for Disease Control and Prevention (CDC)[3]. The BMI distribution is slightly asymmetric in that the slope of the distribution seems to be steeper for BMIs lower than 30 than it is for BMIs higher than 30. However, since the CDC defines "overweight" as having a BMI of 25 to 29.9, and "obese" as having a BMI of 30 or higher, it can be readily seen from Figure 2 that a substantial number of customers are at least overweight and a large number of those customers qualify as obese.

[3] https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

The next visualization is a bar graph of customer smoking status; it can be found in Figure 3.

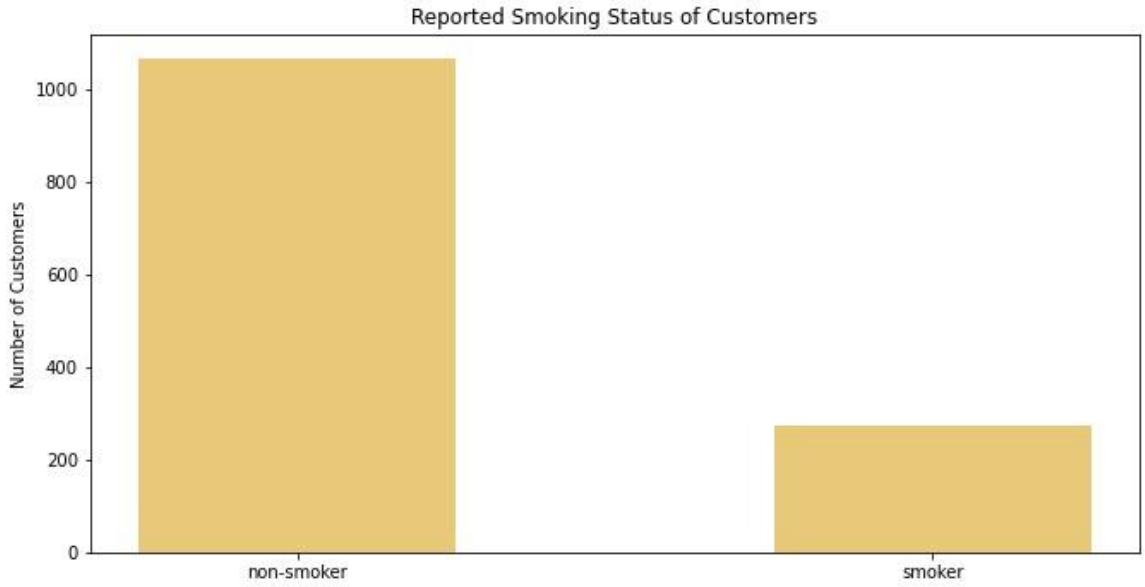

Figure 3: A bar graph of customer smoking status. There are roughly 4 times as many non-smokers as there are smokers.

From Figure 3, it can be seen that there are approximately 4 times as many non-smoking customers as there are customers who smoke.

Lastly, a bar graph of the number of children that customers have is found in Figure 4.

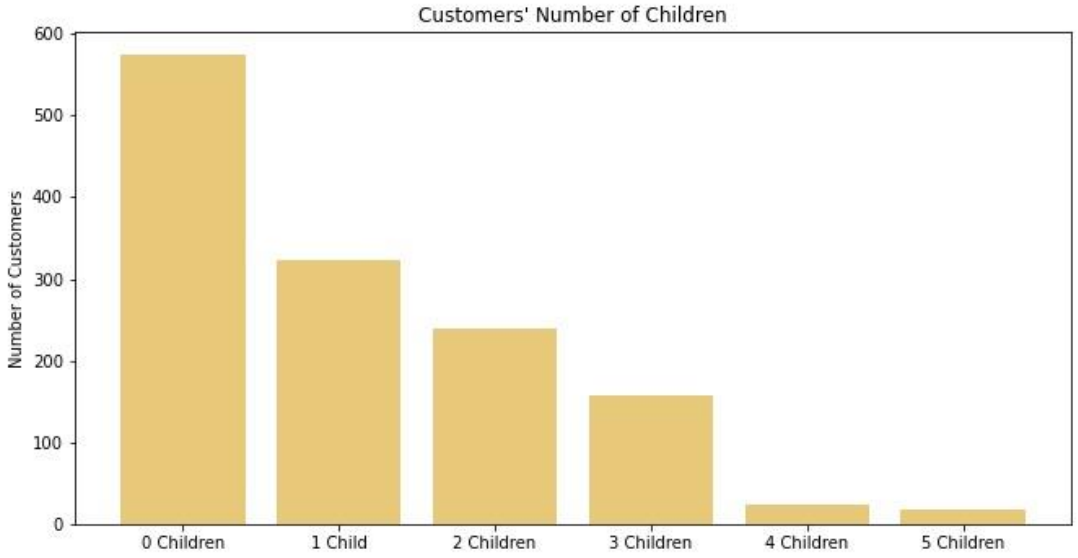

Figure 4: A bar graph of the number of children customers have. Roughly 40% of customers do not have children.

From Figure 4, it can be seen that most customers have at least 1 child, but it is rare for a customer to have more than 3 children. Roughly 40% of customers are childless.

Additional visualizations of the full dataset can be found in Appendix A1.

In the next section, the data visualized above will be used to determine answers to the following questions:

1. How do health insurance charges, body mass index (BMI), and smoking rates vary by region in the United States?

2. Do customers with children tend to have higher BMIs than those without children?

3. Is it possible to predict individual customer charges using a supervised machine learning model?

Analysis and Results

This section provides a description of the analysis done on the data as well as the results of this analysis. It is divided into 3 parts based on the major or minor question being addressed. Details of the data and visualizations of the data can be found in the Data section.

Regional Variation in Charges, Body Mass Index (BMI), and Smoking Rates

This section will be divided into three sub-sections. One will address the analysis of regional variation in charges; one will address the analysis of regional variation in BMI; and one will address the analysis of regional variation in smoking rates.

Regional Variation in Charges

The data was divided into four subsets, one each for the Northeast region, the Southeast region, the Northwest region, and the Southwest region. (Data for the Midwest region was either included in one of the four previously listed regions or was not included in the dataset.) To better visualize the spread in costs, a box plot was made; this plot is included as Figure 5.

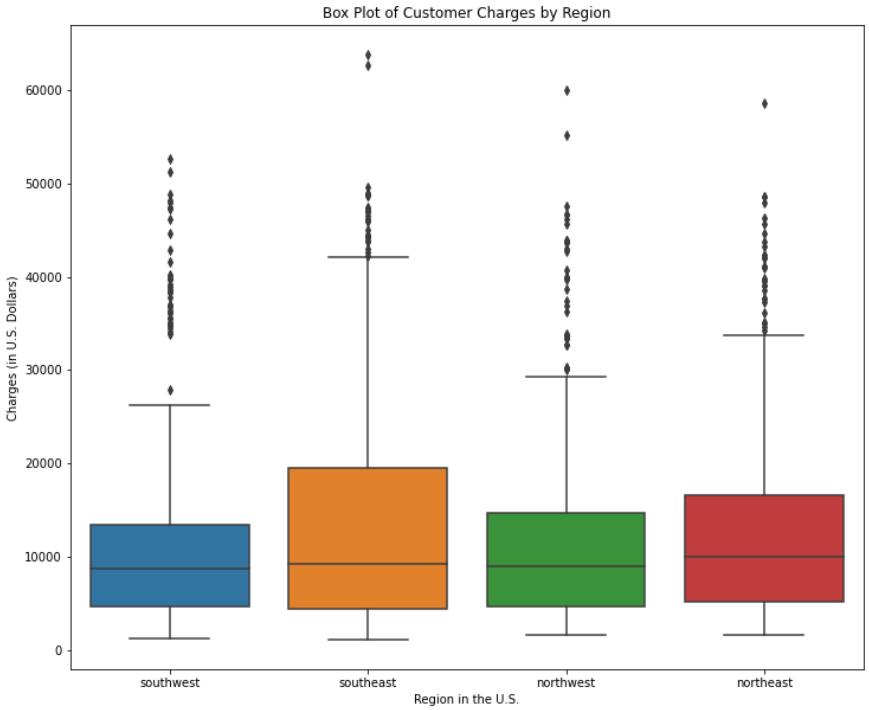

Figure 5: A box plot of customer charges by region in the U.S. The Southeast has the largest spread in the middle of the data, whereas the Northeast has the largest median charge value. In all four regions, there is a significant tail of outlying charges at high charge values.

In Figure 5, each box represents the interquartile range of the associated data subset. 50% of the data lie within this range. It can be readily seen that the Southeast data subset has the largest interquartile range and hence the largest spread in the middle. The solid black line within each box represents the median of the associated data subset; it is the value in the middle of the data subset. The Northeast data subset clearly has the highest median charge value; the other three median values are all similar. The black bars extending from each box represent the first quartile (lower bars) and third quartile (upper bars). 75% of the data falls within that range. Diamonds represent data outliers; it is worth noting that all outliers for all four data subsets lie at the top of the plot, where the largest charges can be found. In conclusion, the spread in the middle of the data is highest for the Southeast, while the median charge is highest for the Northeast; all four data distributions are skewed toward high charges.

To further explore the differences between the charges in each data subset, average values were calculated in addition to precise median values. These values are included in Table 1 below.

Region	Average Charge	Median Charge
Southeast	$14,735.41	$9,294.13
Northeast	$13,406.38	$10,057.65
Northwest	$12,417.58	$8,965.80
Southwest	$12,346.94	$8,798.59

Table 1: Average and Median charges for each region in the U.S.

While the Northeast region has the highest median charge, the Southeast region has the highest average charge. The higher average charge for the Southeast region is consistent with it having the largest interquartile range and is also consistent with a dataset which has a long tail extending to higher charges. Since all four data subsets have outliers only at large charge values, it is unsurprising that average charges are higher than median charges. Visualizations of the distributions of charges themselves can be found in Appendix A2.

To investigate whether or not the distributions of charges are significantly different from each other, a Kolmogorov-Smirnov (K-S) statistical test was run on each pair of distributions. An *a priori* chosen p-value of 0.003 was used as the criterion for rejecting the null hypothesis (that the two distributions being compared were drawn from the same underlying distribution) in favor of the alternative hypothesis (that the two distributions

being compared are not consistent with having been drawn from the same underlying distribution). For the *a priori* chosen p-value, it was found that none of the distributions was significantly different from any of the other distributions. Details of the K-S tests and their outcomes can be found in Appendix A2.

Thus, in conclusion, average charges were seen to be highest in the Southeast, while median charges were seen to be highest in the Northeast. However, there does not seem to be a significant statistical difference between the distributions of charges of the four data subsets.

Regional Variation in BMI

The same data subsets used to investigate regional variations in charges were used to investigate the regional variation in BMI. A visualization of the BMI distributions of the four data subsets was constructed; it is included as Figure 6.

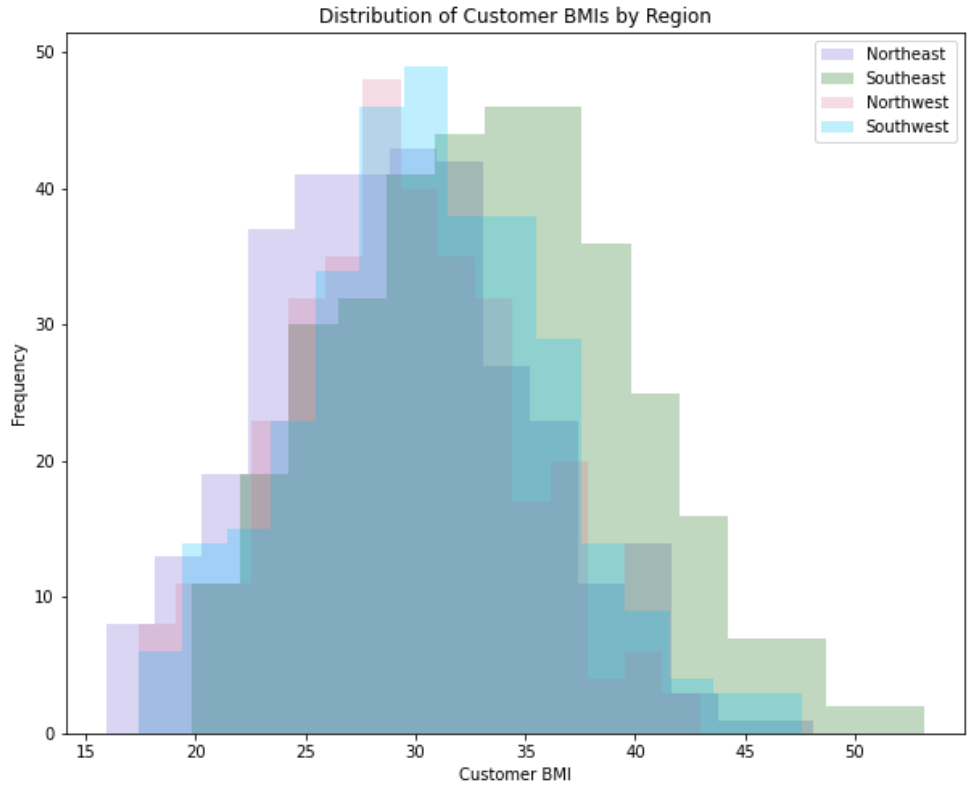

Figure 6: BMI distributions of each region in the U.S. Each BMI data subset is approximately normally distributed. The mean value of the Southeast BMI distribution lies at higher BMI values than that of the other data subsets.

All four BMI distributions are approximately consistent with being normally distributed. The peak of the Southeast's BMI distribution—i.e., the distribution's average value, assuming that it is normally distributed—lies at higher BMI values than the peaks of the other distributions; visually, the Southeast BMI distribution is "shifted to the right" compared to the other distributions. To better illustrate this, the average and standard deviation of each BMI distribution was computed; the resulting values are included in Table 2.

Region	Average BMI	Standard Deviation
Southeast	33	7
Southwest	31	6
Northeast	29	6
Northwest	29	5

Table 2: Average BMI and standard deviation for each region in the U.S.

Consistent with Figure 6, the Southeast was found to have the biggest BMI, followed by the Southwest. Interestingly, the Northeast and Northwest had identical average BMIs. An explanation for why standard deviation is reported as a whole number can be found in Appendix A3.

To determine if the BMI distributions of the data subsets are significantly different from each other, an ANOVA test was conducted, the details of which can be found in Appendix A3. It was found that one distribution was indeed different from at least one other distribution. To better ascertain the details of which distribution(s) differ from each other, a subsequent Tukey Range test was performed, with p-value of 0.003 chosen beforehand as the threshold for rejecting the null hypothesis (which in this case is that each pair of regional BMI distributions have the same mean). The results of this statistical test can be found in Table 3.

Regions Being Compared	p-value	Reject Null?
Northeast – Northwest	0.9	No
Northeast – Southwest	0.01	No
Northwest – Southwest	0.01	No
Southeast – Northeast	0.001	Yes
Southeast – Northwest	0.001	Yes
Southeast – Southwest	0.001	Yes

Table 3: Outcome of the Tukey Range Test performed on pairs of regional BMI distributions.

From this, it can be seen that the Southeast's BMI distribution is significantly different from the other three BMI distributions. The other three distributions are not statistically significant from each other at the significance level chosen ahead of time.

In conclusion, it is found that the distribution of BMIs in the Southeastern U.S. is significantly different from the BMI distributions of the other U.S. regions. The BMI distributions of the remaining three regions are not significantly different from each other. The average BMI is highest in the Southeast, and lowest in the North.

Regional Variation in Smoking Rate

The same data subsets used in the previous two sections were used to investigate the regional variation in smoking rates. A bar graph of the fraction of customers who smoke and the fraction of customers who do not smoke in each region is included as Figure 7.

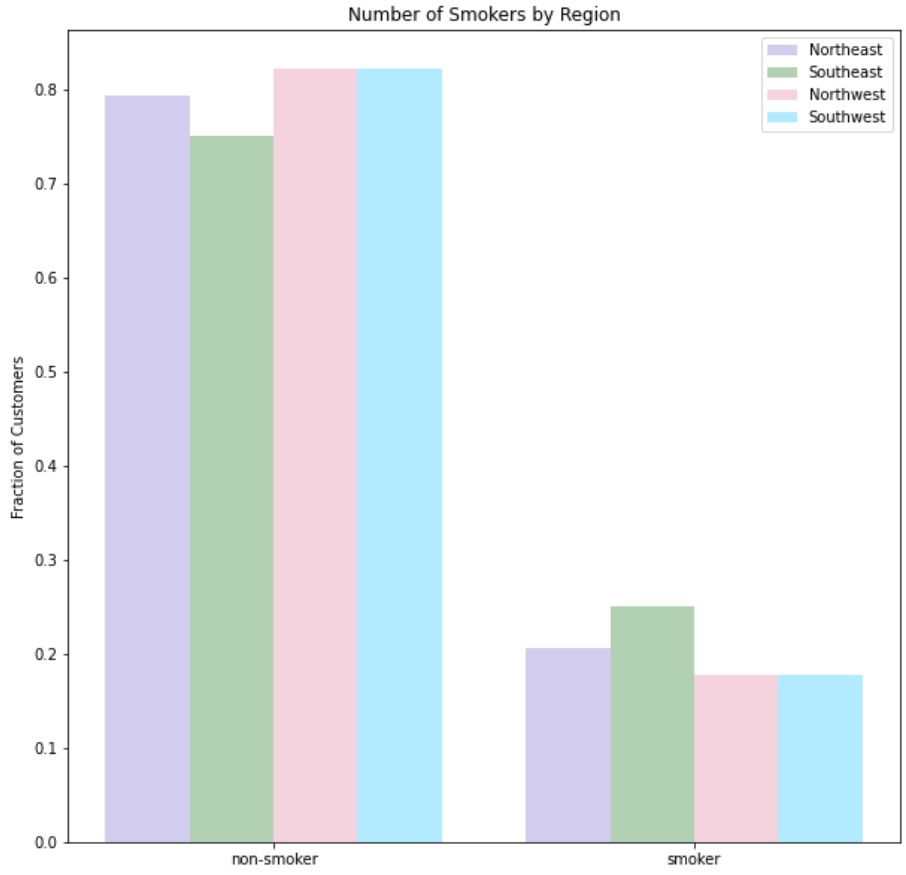

Figure 7: Bar graph of the fraction of non-smoking customers and customers who smoke in each region of the U.S.

Interestingly, the fractions of non-smoking customers are identical for the Northwest and Southwest. The Southwest has the largest fraction of customers who smoke, whereas the Northeast has the second highest fraction.

To determine if the fraction of customers in the Southeast who smoke is significantly different from the fractions of customers who smoke in each of the other regions, and if the fraction of customers in the Northeast who smoke is significantly different from the fractions of customers who smoke in the Northwest and Southwest, a Chi-Squared statistical test was performed on each pair of regions (save the Northwest and Southwest, as their numbers of non-smoking customers and customers who smoke were identical). A p-value of 0.003 was chosen ahead of time to decide whether or not to reject the null hypothesis, which in this case is that there is no significant difference between the smoking rates of the pair of regions being considered. The results of the Chi-Squared test are reported below in Table 4.

Regions Being Compared	p-value	Reject Null?
Northeast – Southeast	0.2	No
Northeast – Northwest	0.4	No
Northeast – Southwest	0.4	No
Southeast – Northwest	0.03	No
Southeast – Southwest	0.03	No

Table 4: Outcome of the pairwise Chi-Squared test performed on reported smoking rates.

From this, it can be seen that the smoking rates of the four different regions are not significantly different from each other.

In conclusion, while the fraction of customers who smoke is highest in the Southeast, there does not appear to be a significant statistical difference in the smoking rates among the four U.S. regions.

Variation in BMI for Parents v. Non-Parents

To investigate whether or not customers with children have bigger BMIs than childless customers, the data was split into two subsets. Customers with 0 children were compiled into one subset, whereas customers with 1 or more children were compiled into another. The distributions of BMIs for these two data subsets are included as Figure 8 below.

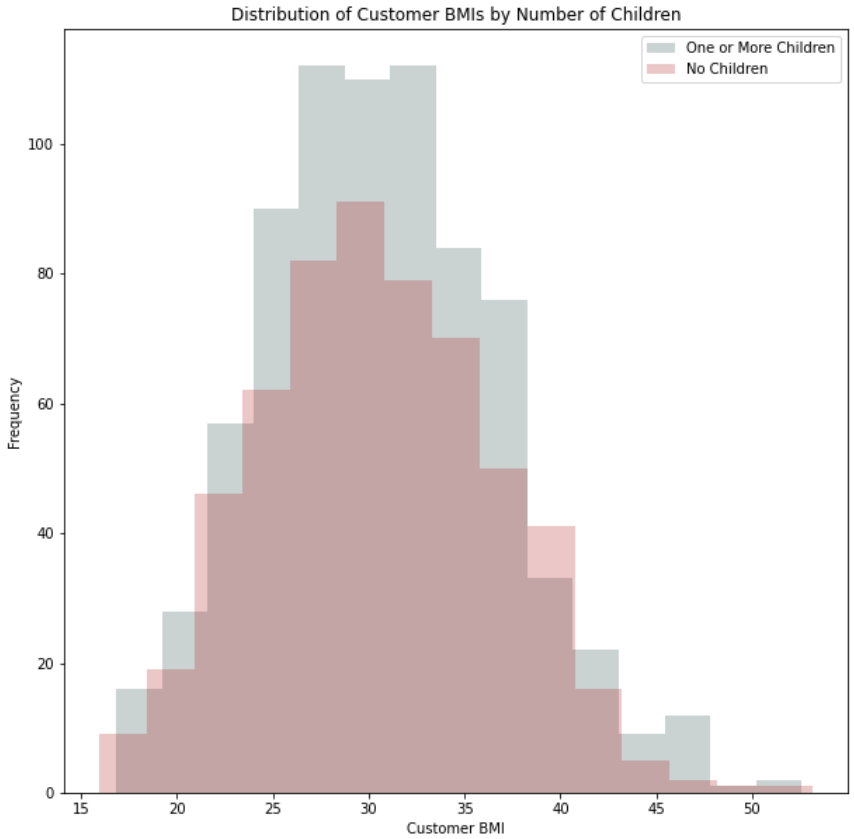

Figure 8: BMI distributions of parents and non-parents. Each BMI data subset is approximately normally distributed.

Similar to the regional BMI distributions, the BMI distributions of childless customers and customers with children are approximately normally distributed. The peak values of the two distributions do not appear to be different.

To determine if the two BMI distributions are consistent with having been drawn from the same underlying distribution, a 2-Sample T-Test was conducted. A p-value of 0.003, chosen before the test was performed, served as the threshold for rejecting the null hypothesis previously stated. Upon conducting the 2-Sample T-Test, a p-value of 0.6 was obtained; thus, the two BMI distributions are consistent with each other.

In conclusion, there is no statistical difference between the BMI distribution of customers with children and that of customers without children. A customer with children is no more likely to have a large BMI than a customer without children.

Predicting Individual Customer Charges

Predicting the costs a customer will incur is of particular interest to health insurance companies. To that end, two supervised machine learning models were fit to the full data set provided by Codecademy; these models aimed to predict customer charges based on overall trends in the data. One model was a Multi-Linear Regression model; the other was a k-Nearest Neighbors Regression model. Plots exploring the dependence of charges upon age, sex, BMI, number of children, and smoking status were constructed before models were fit; these can be found in Appendix A4.

Before fitting the models, the full data set was split into three parts: 80% of the data served to "train" the model (i.e., fit optimal model parameters); 10% of the data was used to "validate" the model (i.e., choose the number of nearest neighbours which maximized the model's accuracy); and 10% of the data served to "test" the constructed models for goodness of fit. Details of the model training process can be found in Appendix A4. Once the models were trained, they were used to predict the charges corresponding to data in the "test" set. A visualization of model performances is provided in Figure 9.

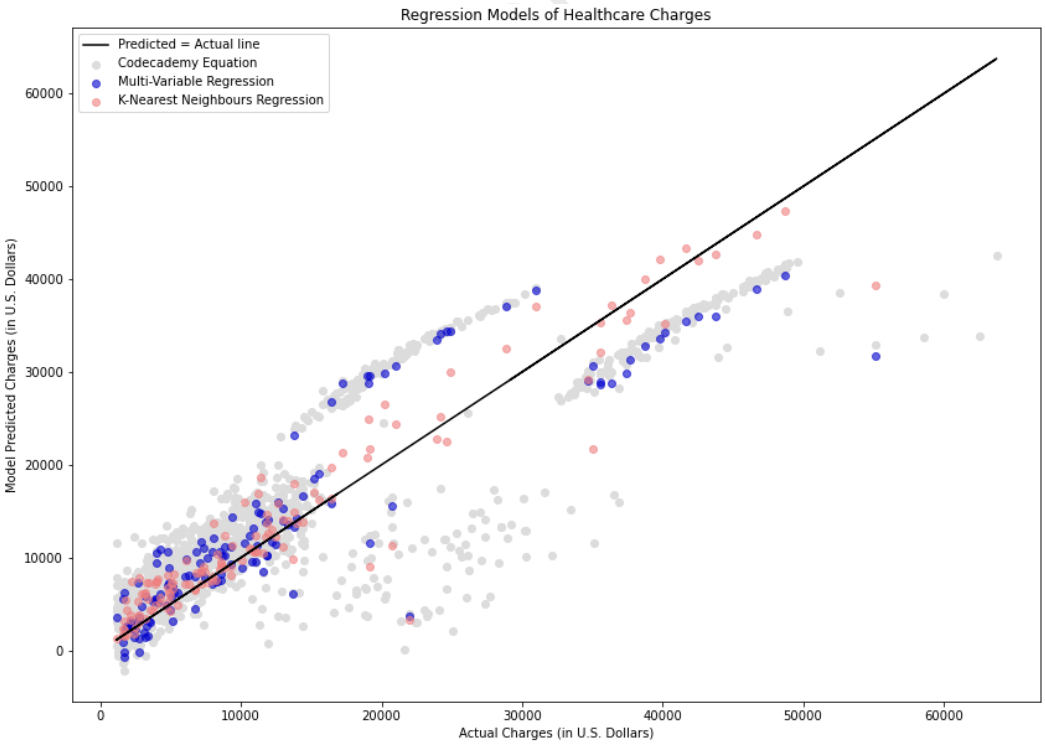

Figure 9: Multi-Linear Regression (blue) and k-Nearest Neighbors Regression (pink) model performance. The solid line marks where predicted charges and actual charges are exactly equal. Grey dots illustrate predictions from the Codecademy predictive equation discussed in the Data section of this paper.

It is clear that the k-Nearest Neighbors Regression model (pink) does a better job of predicting the data than does the Multi-Linear Regression model (blue). There is interesting structure in the predictions made by the Multi-Linear Regression model; a brief discussion of this can be found in Appendix A4.

In Table 5, the coefficient of determination (a.k.a. "goodness of fit", denoted R^2 in the table below) for the two models is provided.

Model	R^2
Multi-Linear Regression	0.81
k-Nearest Neighbors Regression	0.91

Table 4: Coefficient of Determination for each model fit to the full dataset.

The k-Nearest Neighbors Regression model resulted in a better goodness of fit than the multi-linear regression model did.

In conclusion, two supervised machine learning models, designed to predict customer charges were successfully fit to the data. The k-Nearest Neighbors Regression model resulted in a better goodness of fit ($R^2 = 0.91$) than did the Multi-Linear Regression model ($R^2 = 0.81$).

Further comparison between the Multi-Linear Regression model fit here and the Codecademy predictive equation discussed in the Data section can be found in the Jupyter Notebook file included in this repository.

Conclusions

This Codecademy Portfolio Project examined a database of U.S. Healthcare Costs. After exploration of the data, the following questions were raised and successfully answered.

First was asked, “How do health insurance costs, body mass index (BMI), and smoking rates vary by region in the United States?” It was found that, statistically, health care costs do not vary significantly by region. In terms of the reported data, average costs were found to be highest in the Southeastern U.S., whereas median costs were found to be highest in the Northeastern U.S. The Southeastern U.S. was found to have a significantly different distribution of BMIs than the other regions of the U.S., with an average BMI that was higher than any of the others. The BMI distributions of the other regions were found, statistically, to be similar. While the fraction of policyholders who smoke was found to be highest in the Southeastern U.S., the rate of smoking vs. non-smoking was not found to be significantly different between this region and any other region. Thus, it seems as though BMI does significantly vary by region in the U.S., but costs and smoking rates do not (statistically speaking).

Next was asked, “Do policy holders with children tend to have higher BMIs than those without children? They do not. Statistically, there is no difference in the distribution of BMIs between policyholders with children and policyholders without children.

Finally, the question, “Is it possible to predict individual policyholder costs using a supervised machine learning model?” was asked. Two such supervised machine learning models were successfully fit to the full dataset. The k-Nearest Neighbors Regression model fit to the data had a better goodness of fit than did the Multi-Linear Regression Model fit to the data; it also resulted in more accurate predictions.

As regards this last point, given the small size (1,338 policyholders) and limited policyholder data available with this dataset, it should be cautioned that the predictive models constructed herein should not be applied to any other dataset. There are many factors which this data does not account for which could affect healthcare costs—pre-existing conditions, treatment of accidents, etc.—and a more robust model would take these factors into account. Also, since the year the data were compiled was not provided, it is not known how representative these reported costs are of healthcare costs in 2020. The data used herein should be seen as a toy dataset used to investigate several interesting questions, but not applicable to broader healthcare data as a whole.

Given the limited nature of the information provided in this dataset, there are many open questions regarding it and the broader predictive analysis performed in this work. One key question regards the BMI distribution of customers; in all regions of the U.S., the average BMI reported falls at least into the "overweight" category, and half the time into the "obese" category as defined by the Centers of Disease Control and Prevention[4]. What factors contribute to the high average BMIs seen in the U.S.? It is impossible to know from the data. Further, as noted in Appendix A4, there is a complicated dependence of charges incurred upon age; efforts to investigate an underlying cause for this dependence were unsuccessful, but better characterizing that dependence could lead to better predictive models. As regards better predictive models, it is possible that factors unaccounted for by the data in this dataset, such as having a pre-existing condition, stress levels, and education levels (among others), could impact customer charges incurred; it would be interesting to repeat this analysis on a more detailed dataset to see if a more accurate predictive model could be constructed.

Acknowledgements

This project was a Codecademy Portfolio Project, so the author would like to thank the curriculum developers at Codecademy for gathering up the resource needed to complete this project, as well as the structure they created to help track progress through it. Codecademy references the following Kaggle user as being the compiler of the data set:

https://www.kaggle.com/mirichoi0218/insurance

According to the Kaggle page, this is an open database released in the public domain. The author would like to thank the Kaggle user for the work they did in cleaning up the data base and posting it to their Kaggle page.

[4] https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

Appendices

Included in these appendices are figures which are supplemental to the main text as well as more in-depth descriptions of statistical tests and other analytical methods used.

Appendix A1: Additional Visualizations of the Full Dataset

Histograms of customer charges (Figure 1) and customer BMI (Figure 2) can be found in the Data section of this paper; bar graphs of customer smoking status (Figure 3) and customers' number of children (Figure 4) can be found in the same section. Included here are two additional plots of customer ages (Figure A1.1) and customer sex (Figure A1.2).

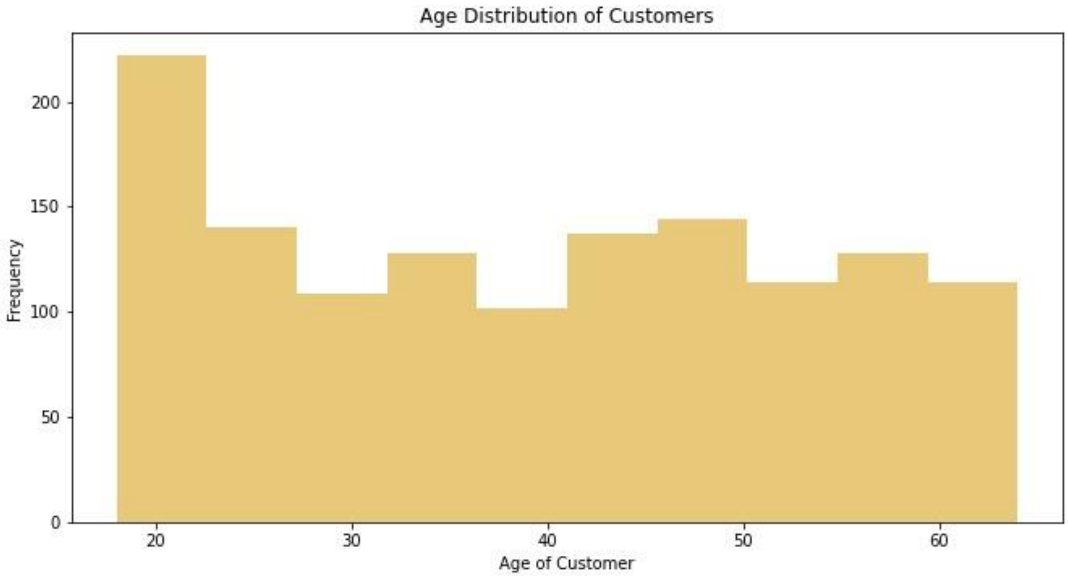

Figure A1.: A histogram of customer ages. The distribution of ages is approximately flat, although there is a small excess of customers with ages between 18 and roughly 22.

The distribution of ages is roughly flat, although there is a small excess of customers with ages between 18 and roughly 22. Thus, it is not expected that the results of this work will be affected by one particular age group being overrepresented or underrepresented in the data.

Figure A1.2 can be found on the next page.

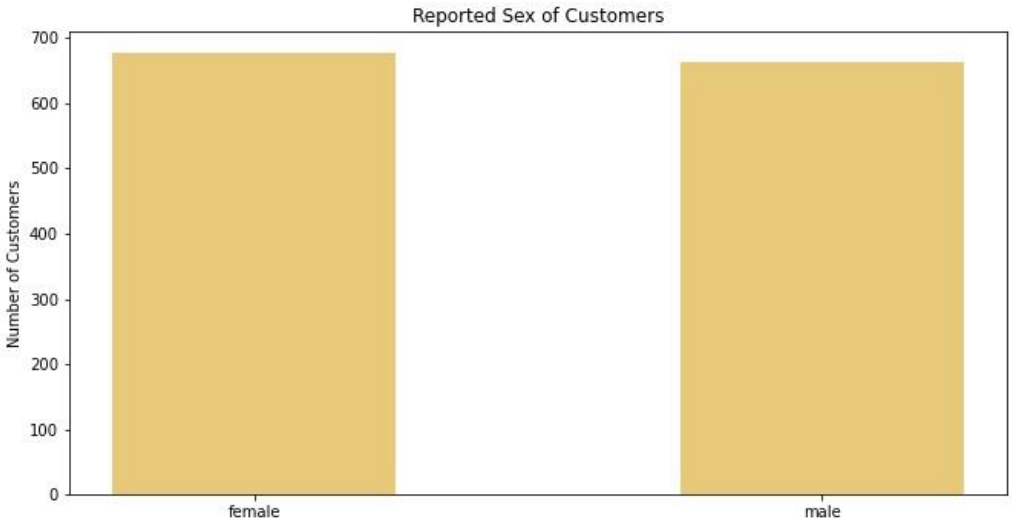

Figure A1.2: A bar chart of the number of female and male customers in the dataset. There are slightly more females than males.

There are approximately equal numbers of females and males in the dataset, with slightly more females than males. Thus, it is not expected that this work will be affected by the under- or overrepresentation of one gender compared to the other.

Appendix A2: Regional Variation in Charges Supplemental Material

A box plot of the spread in customer charges broken down by region was provided in Figure 5. While that box plot makes the spread in data easy to see, some detail is lost in its making; thus, a histogram of charges for each region was constructed. This histogram is included below in Figure A2.1.

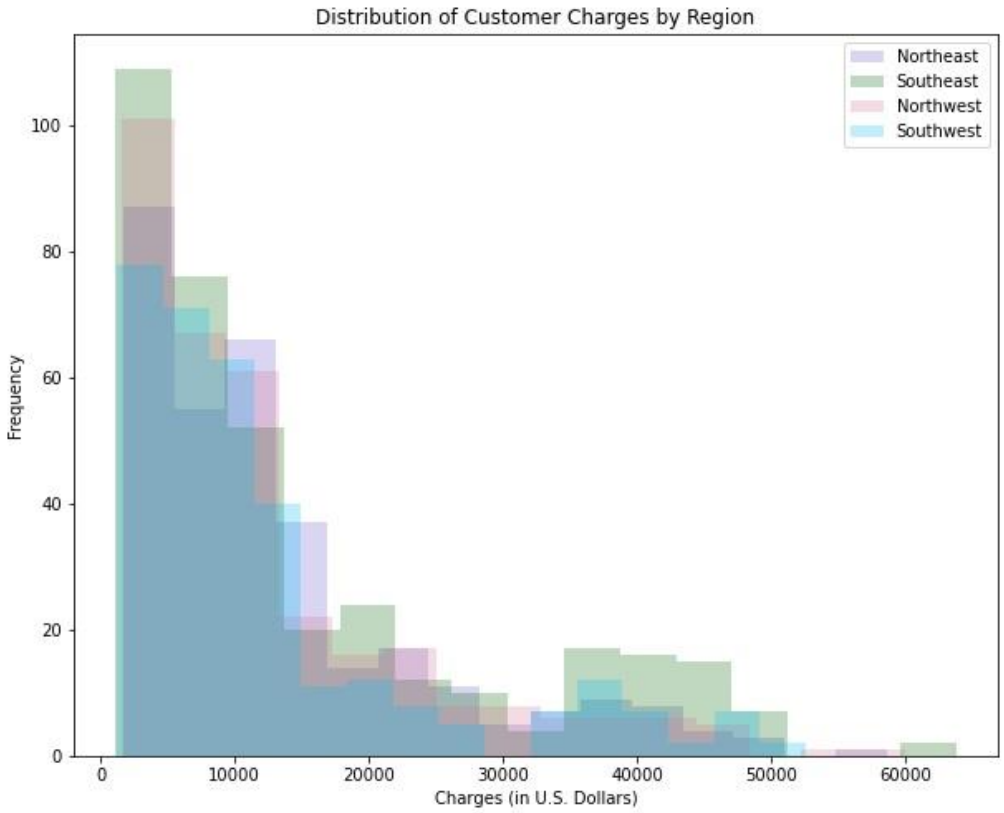

Figure A2.1: A histogram of customer charges broken up by region. The secondary peak of the distribution of charges for the Southeast region is noticeably higher than the secondary peaks in the distributions of the other regions. In addition, the drop-off in charges between ~$12,000 and ~$20,000 is not as steep for the distribution of charges for the Northeast as it is for the other regions.

To determine if these distributions differed significantly from each other, a series of pairwise Kolmogorov-Smirnov (K-S) tests was performed. A K-S test was chosen because the distributions depicted in Figure A2.1 clearly do not follow a normal distribution; since the K-S test is non-parametric, it does not assume a particular shape for the input distributions, making it ideal to compare the distributions in Figure A2.1. In each case, the null hypothesis can be stated as, “the two distributions being tested are in fact drawn from the same continuous distribution.” Before the pairwise K-S tests were

run, a p-value of 0.003 was chosen as the criterion for rejecting that null hypothesis. The results of these tests are included in Table A2.1.

Regions Being Compared	p-value	Reject Null?
Northeast – Southeast	0.2	No
Northeast – Northwest	0.2	No
Northeast – Southwest	0.05	No
Southeast – Northwest	0.06	No
Southeast – Southwest	0.03	No
Northwest – Southwest	0.7	No

Table A2.1: Outcome of the K-S tests performed on pairs of regional distributions of charges.

None of the p-values obtained from the pairwise K-S test performed falls below the *a priori* chosen value of 0.003; thus, in none of the cases should the null hypothesis be rejected. It should be noted that a p-value of 0.003 is a rather conservative value; had a larger threshold p-value been chosen, the interpretation of the results may well have been different.

Appendix A3: Regional Variation in BMI Supplemental Material

In Table 2, the standard deviation in BMI was reported to 1 significant figure. This may strike the reader as unusual, so a justification is warranted. First, it should be noted that the numbers of customers per region were roughly equal at about 325 per region, although the Southeast did have 40 more customers than did the other regions. This is a relatively small sample size, and so the uncertainty in any calculated figure (here, the average BMI) cannot be known with great accuracy. The uncertainty in each BMI measurement is unknown; however, the standard deviation of the BMI measurements can be interpreted as the uncertainty in a single BMI measurement[5]. (Note that the previous statement does assume that sources of uncertainty are small and random, as well as that the same quantity is being measured many times using the same method each time.) Given the relatively small number of measurements, it is customary to report their uncertainty to only 1 significant figure; further, the calculated average is then rounded to the decimal place of the uncertainty. This explains why the standard deviation in BMI was reported to 1 significant figure, as well as why average BMIs were rounded to the same decimal place as the standard deviation (i.e., the ones place).

Meanwhile, in the discussion of regional variation in BMI, it was reported that an ANOVA test was conducted to determine if the BMI distributions of the regional data were significantly different from each other. An ANOVA test was chosen in this case because the input data for an ANOVA test are assumed to be normally distributed, and the distributions illustrated in Figure 6 are approximately normal. The ANOVA test also assumes that the ratios of the standard deviations of the distributions are approximately 1; in all cases but 1, this was found to be approximately so. This can be seen in Table A3.1 below.

Regions Being Compared	Ratio of σs
Northeast – Southeast	0.92
Northeast – Northwest	1.1
Northeast – Southwest	1.0
Southeast – Northwest	1.3
Southeast – Southwest	1.1
Northwest – Southwest	0.90

Table A3.1: Ratios of standard deviations (denoted in the table as σ) of pairs of regional BMI distributions.

[5] Taylor, John R. *An Introduction to Error Analysis, Second Edition.* Sausalito, CA, University Science Books, 1997. See Section 4.3 and more specifically page 101.

In Table A3.1, standard deviation is abbreviated with the symbol σ. Two digits are reported for the purposes of illustration, not because they are both significant. The ratio of the standard deviation of the Southeast BMI distribution to the Northwest BMI distribution is found to be slightly higher than 1 at 1.3 rather than 0.9 or 1.1; however, the data were still included in the ANOVA test. When interpreting the results of the ANOVA test, as well as the resultant Tukey Range test, this caveat should be borne in mind.

When performing the ANOVA test, a p-value of 0.003 was chosen ahead of time as the threshold for indicating a significant difference between at least one distribution and at least one other distribution. Upon performing the test, a p-value of 1.9×10^{-24} was obtained. This is clearly smaller than the *a priori* chosen p-value of 0.003; thus, a significant difference between at least one distribution and one other distribution was detected. Since a significant difference was detected, a follow-up Tukey Range test was performed as in the main text. The outcome of this statistical test can be found in Table 3 and accompanying text.

Appendix A4: Predicting Individual Customer Charges with Regression Models

As noted in the Data section, in order to build a regression model of the charges data customer sex and smoking status needed to be converted to numerical values. Since the data for each category was binary—"female" and "male" for sex and "yes" or "no" for "smoker"—it was natural to convert this data to "0" or "1". Following the convention set forth by Codecademy[6], the sex "female" was converted to 0 and the sex "male" was converted to 1. Further, "yes" for "smoker" was converted to 1, while "no" for "smoker" was converted to 0. The Pandas method *map* was used to carry out the conversion. The specific code used to accomplish this can be found in the Jupyter Notebook file uploaded to this repository.

Prior to the construction of any regression models, plots exploring the dependence of charges upon age, sex, BMI, number of children, and reported smoking status were made. The resulting visualizations are included on the next page as Figure A4.1. A description of the trends seen in this figure is as follows. Given how sex and smoking status were mapped to values of "0" or "1" depending on the categorical data reported, the appearance of two vertical lines in the plots of charges vs. age and charges vs. smoking status is expected. There is no discernable structure to the plot of charges vs. BMI. Three distinct, positively-trending lines can be seen in the plot of charges vs. age. As detailed in the Jupyter Notebook accompanying this report and found in this repository, preliminary efforts to understand the structure of this plot were unsuccessful. Nonetheless, it can be inferred from Figure A4.1 that age and charges are positively correlated; no other clear trends can be found in the remaining plots.

Once all available numerical data were gathered, they were split using the Scikit-Learn pre-processing module *train_test_split* into testing, training, and validation sets. 10% of the data were set aside as the "test" set, 10% were set aside as the "validation" set, and the remaining 80% formed the training set. Once this splitting was accomplished, a Multi-Linear Regression model was fit to the data. This model resulted in a coefficient of determination of 0.813, as reported in Table 4. Values for the model coefficients and their associated features can be found in Table A4.1, which follows on the next page.

[6] https://www.codecademy.com/paths/data-science/tracks/dscp-python-fundamentals/modules/dscp-python-functions/projects/ds-python-functions-project

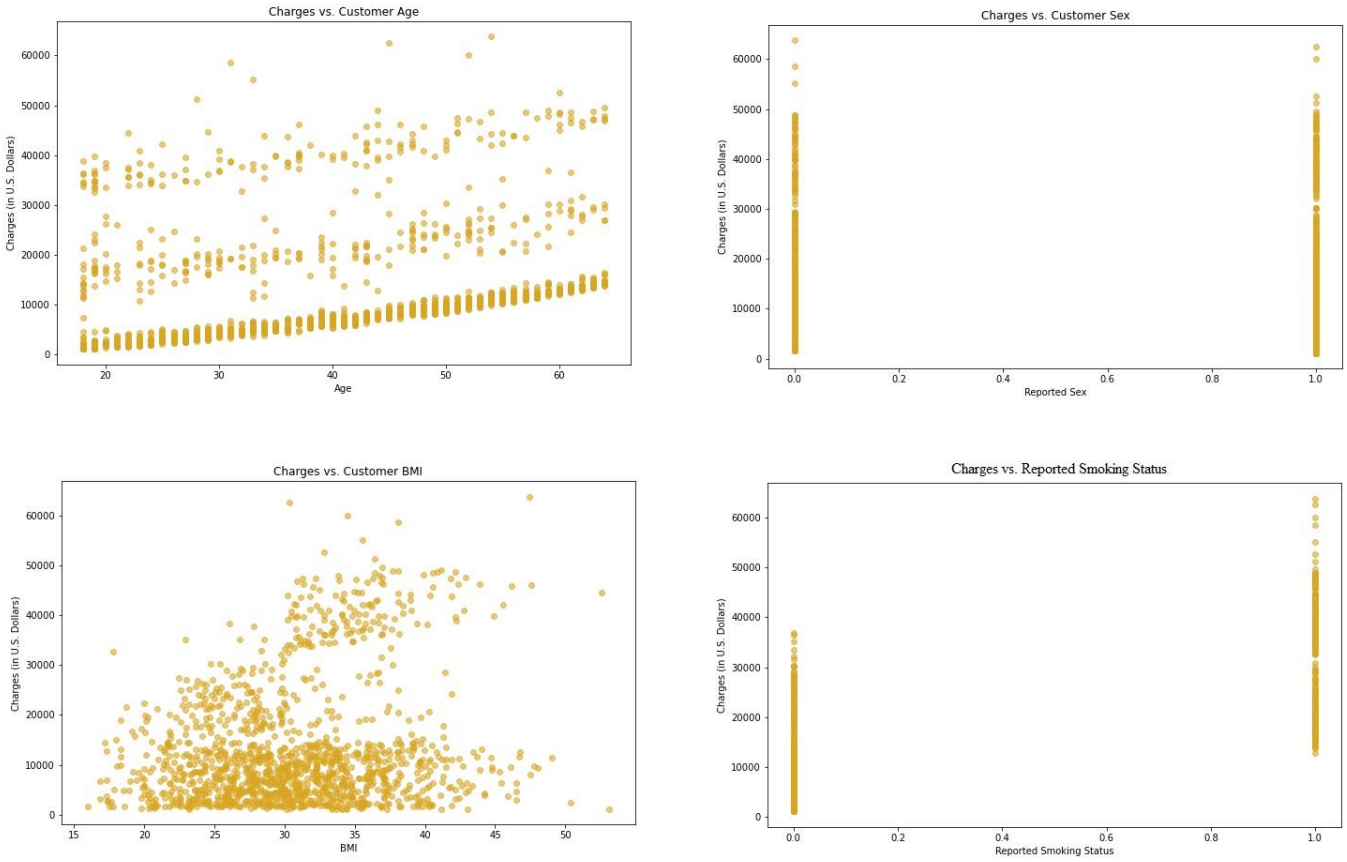

Figure A4.1: X-Y scatter plots of charges vs. age (top left), sex (top right), BMI (bottom left), and smoking status (bottom right). The plots of charges vs. sex and charges vs. smoking status are vertical lines, as expected; while there is no clear structure to the charges vs. BMI plot, three distinct curved lines can be seen in the charges vs. age plot. It is not known why the plot of charges vs. age has the structure that it does.

Feature	Coefficient
Age	260
Sex	–136
BMI	322
Number of Children	535
Smoking Status	23,800
y-Intercept:	–12,100

Table A4.1: Coefficients corresponding to each feature in the Multi-Linear Regression model fit to the full dataset.

Prior to constructing the k-Nearest Neighbors Regression model, the numerical features in the training, validation, and test sets were scaled to unit variance, after removal of the mean, using the Scikit-Learn pre-processing module *StandardScaler*. From there, a series of regression models was fit treating *k*, the number of neighbors used to predict charges, as a free parameter. Values of *k* between 1 and 100 were probed. The optimum *k* was chosen to be the one which resulted in the highest coefficient of determination for the validation set. For this particular k-Nearest Neighbors Regression model, *k* was determined to be 12, and the resulting coefficient of determination for the test set was 0.908. A plot of coefficient of determination vs. *k* for the k-Nearest Neighbors Regression model-building process is included as Figure A4.2.

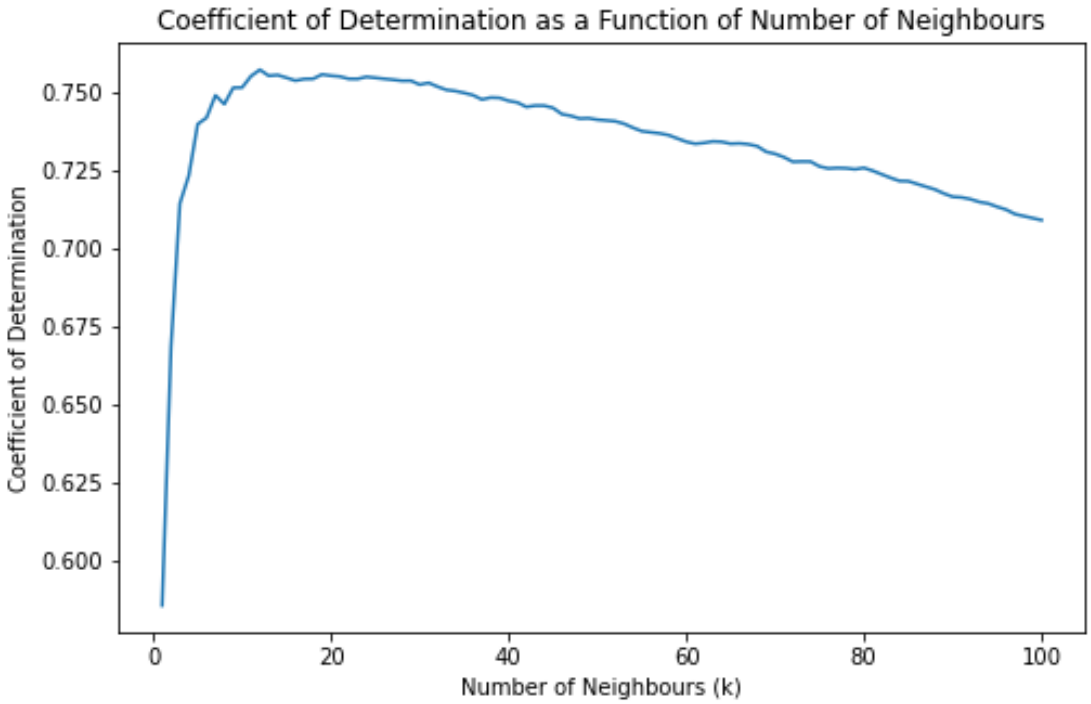

Figure A4.2: Coefficient of determination vs *k* for k-Nearest Neighbors Regression models fit to the training set as described above. As expected, coefficient of determination rises, reaches a peak, and then falls with increasing *k*.

A plot of predicted charges vs. actual charges can be found in the main text as Figure 9.