Analysis of Insurance Cost on Patient Demographics

Daksh Verma, Evan Patel, Maxine Attobrah, Saurabh Rane, Sujay Shah, Tejas Dighe

May 8, 2022

Abstract

Health insurance costs are a cause of significant concern in the United States, across all age groups. Our aim is to understand the factors affecting the cost of health insurance and whether they can be used to predict the future costs. The primary motivation is to help people plan better for their projected future health insurance costs. We predicted the cost of health insurance based on selected attributes of individuals across age groups of ~18-90 years and found correlations to the cost of insurance. After building several visualizations and models to understand our data, we determined that smoking status, age, and BMI were the biggest influential factor in determining the cost, irrespective of the individual's geographical location in the United States. While we initially believed gender would be a strong predictor of costs due to the differences in health needs, age and BMI were far more influential in helping predict the cost of insurance. Therefore, we believe our analysis will help understand the factors driving costs for insurance. While we can help shape public policy from this data-driven approach, it is important to understand biases and limitations of this methodology. Especially in fields like healthcare where policy decisions can affect millions of people, it is imperative for policy makers to be cognizant of the checks and balances required while using data science. Lastly, the findings can influence behavior in the population to reduce healthcare costs for the future.

Introduction

Use of data science and machine learning in healthcare and public policy

Data Science and Machine Learning are emerging technologies that have been widely used and applied across various products and industries, especially in healthcare. It can play a huge role in industries like finance and life sciences owing to its capability of discovering patterns across large datasets. [1] Data science and machine learning have been used in drug discovery, to predict diseases, decision support, and predicting readmission rates in hospitals, amongst other applications in healthcare[2]. Applied data science has also been used in diverse policy perspectives. They have been used to predicting crime risks, healthcare violations, and effective interventions.[3,4]

Massive amounts of clinical data is available due to the digitalization of medical records. Therefore, we have seen an influx of data science interventions in medicine, particularly to gain insights and knowledge for decision making. However, there are multiple challenges with using data science in healthcare. In an example of using ICU data, where there is a crucial need of evidence-based case management in complex illnesses, data science can help make decisions for better patient outcomes. Despite an increase in the number of publications and studies, very few projects have actually been implemented in the ICU ecosystem[5]. In addition, when data science and machine learning are used to bring policy changes, it can present a whole new set of challenges and benefits. For example, it could induce a strategic shift in behavior. If our models can identify which factors most influence a certain decision, the user or individual could modify their decisions or actions to attain maximum benefit. This could be beneficial in sectors like healthcare, where patients can use this information to change their lifestyle with healthcare costs as a significant motivator. However, in sectors like crime, it would be detrimental as people could change and manipulate signals[6] on factors that classify them as potential criminals or high-risk individuals. Another big challenge using data science for predictions is the inherent inclusion of human biases which manifest as systemic gaps. The datasets can induce a bias within the algorithm depending on how the data was collected and for what purpose it was utilized. Finally, it's not always possible or easy to implement the results that a model provides the user. It could be difficult or impossible to implement certain interventions due to political circumstances, financial constraints, or personal perspectives of key decision makers[7]. This makes the application of data science, especially in the field of healthcare insurance, important but difficult to implement.

Health insurance in the United states: Coverage and factors affecting costs

Health insurance is the mechanism for an individual to pay for their healthcare expenses. Most of the population have private insurance through their workplace or employer, but others could also obtain it through the state or the federal government[9]. Every year, the coverage and rates of health insurance are subject to change based on the economy, population demography change, and policies around healthcare[9]. COVID-19 put forward an unprecedented governance challenge, especially for healthcare and insurance. In the United states where healthcare insurance plays a big role in access to healthcare, the federal government and state governments had fragmented decisions[8]. Pre COVID-19, in 2019, roughly ~27.5 million people in the US did not have health insurance- that's nearly 8.5% of the US population. There was obviously a heavy skew towards private healthcare insurance users with over ~67% people falling in that bucket, since employer based health insurance is the most popular form of insurance in the US. Government programs like 'Medicaid' or 'Medicare' covered ~33% of the population[10].

There are multiple factors that can determine the cost[11]. State and federal laws can dictate what the insurance can and cannot cover or charge. The state of residence can also cause a variance in the costing of insurance. In addition, premiums tend to be lower in urban areas and also it depends from county to county. The type of insurance that an individual buys also determines how much they pay or if the employer's group plan covers a section of the healthcare insurance cost. Moreover, the insurance comes large at scale and therefore is cheaper for large companies. The income level has a big role in the insurance cost, for example the low wage workers tend to pay less through federal or state as they receive subsidies. There are also multiple plans which come across different price ranges. Finally, age and smoking plays an extensive role determining the price of health insurance. For example, the rates hike up with age, and the premium for tobacco users can cost upwards of 50% than non-smokers[12]. There are certain factors that cannot affect the healthcare insurance costs like gender, as insurance companies aren't allowed to charge men and women differently for the same plans[12].

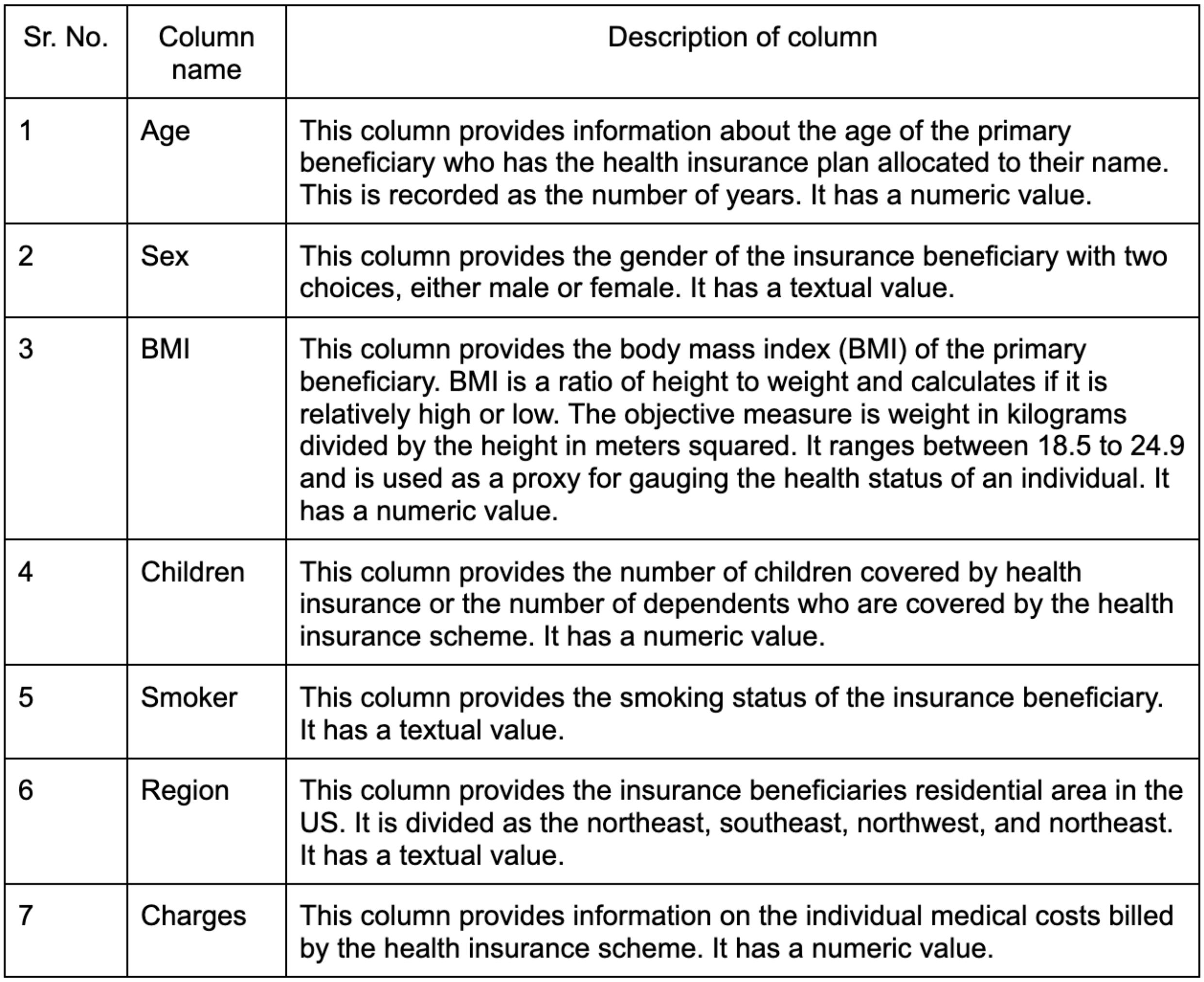
Motivation

The ACA (Affordable Care Act) was created in 2010 and is known as 'Obamacare' which extends to millions of people. However, the ACA also made healthcare insurance mandatory[13].
As of 2020, the average cost of health insurance in the US was $456 per month for a single person and $1152 per month for a family[14]. While there are multiple factors affecting the costs of insurance plans, these prices are high for a big section of the population. Moreover, the most vulnerable populations are most affected by the prices of the insurance premiums. In the US, ~37.2 million people were in poverty in 2020, which was ~3.3 million more than 2019[15]. Similar patterns were observed for age groups under 18, where poverty rates jumped from 14.4% to 16.1% from 2019 to 2020. For the age ranges of 18-64 years, the poverty rates went from 9.4% to 10.4% in 2020, and for people above 65 years of age, it was stagnant at 9.0%. There was a significant change in the poverty levels for Hispanics and Blacks (19.5%)[15]. Therefore, in this paper we wish to explore the various factors that have the highest influence on the cost of healthcare insurance premiums through various modeling efforts and also predict healthcare costs. Through this paper we hope to uncover new correlations or confirm existing beliefs regarding factors that might be in an individual's control to minimize healthcare insurance costs.

Data

Information on the dataset

This data is called the ‘Medical Cost Personal Datasets’ and the paper uses this publicly available dataset to build models. The data for this paper was obtained from Kaggle, which primarily hosts multiple datasets across all domains including healthcare, finance, manufacturing, and agriculture, amongst others. The dataset is also associated with a book called ‘Machine Learning by R’ written by Brett Lantz. It seems that Packt Publishing used this dataset or similar dataset for teaching or conducting experiments. The dataset is small in size with only ~1339 rows, seems fairly straightforward, and needs no cleaning or pre-processing. Figure 1 is a table that contains the columns along with their brief descriptions.



Sr. No.	Column name	Description of column
1	Age	This column provides information about the age of the primary beneficiary who has the health insurance plan allocated to their name. This is recorded as the number of years. It has a numeric value.
2	Sex	This column provides the gender of the insurance beneficiary with two choices, either male or female. It has a textual value.
3	BMI	This column provides the body mass index (BMI) of the primary beneficiary. BMI is a ratio of height to weight and calculates if it is relatively high or low. The objective measure is weight in kilograms divided by the height in meters squared. It ranges between 18.5 to 24.9 and is used as a proxy for gauging the health status of an individual. It has a numeric value.
4	Children	This column provides the number of children covered by health insurance or the number of dependents who are covered by the health insurance scheme. It has a numeric value.
5	Smoker	This column provides the smoking status of the insurance beneficiary. It has a textual value.
6	Region	This column provides the insurance beneficiaries residential area in the US. It is divided as the northeast, southeast, northwest, and northeast. It has a textual value.
7	Charges	This column provides information on the individual medical costs billed by the health insurance scheme. It has a numeric value.

Figure 1: Columns from the ‘Medical Cost Personal Datasets’ dataset.

Cleaning and transformation

The dataset we chose was already cleaned and transformed into rows of data with the columns above. There was no missing data, which was useful for utilizing all of the dataset.

Exploratory data analysis

First, we explored the data to see if we can find any biases in the data. We first analyzed the age of people in the dataset to find that the dataset contains a wide spectrum of age groups. However, the data dataset has more younger people than older people. We will comment on the effect of this bias in the discussion section if it is of any significance to our modeling efforts or results. Similarly, we found another bias in the quantum of individuals who smoke, as there are more non-smokers in the dataset. Finally, there are roughly the same number of men and women within this dataset (Refer to Figures 2, 3, and 4). Second, we looked at the relation between the age and payment for health insurance. We drew a line plot and observed that as a person gets older they are more likely to pay more in insurance regardless of their gender (Refer to appendix). We also looked at the age v/s cost of insurance and we constructed a scatter plot for this purpose (Refer to figure 5). There was no clear observation whether the sex of an individual affects the cost of insurance. It seemed that from this graph gender may not play a role in the cost of medical insurance. It's clear that some further analysis is needed (Refer to appendix).

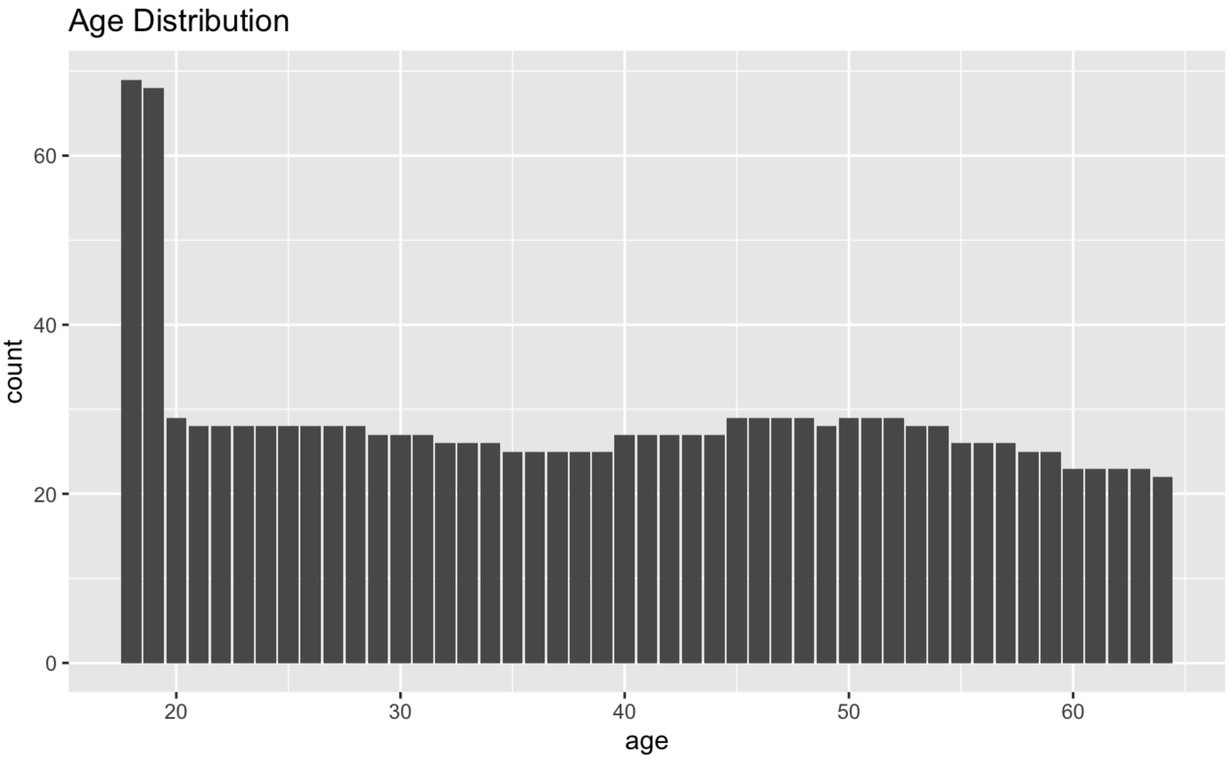


Figure 2: Age Distribution

In addition, we looked at the cost of Insurance (Female vs Male) and the bar plot in the figure 6 in the appendix shows that men might be paying higher in insurance most times than women. We then created histograms based on gender and both histograms are skewed to the right. This means that the majority of the people in the dataset are paying lower insurance costs. After taking the log transformation of the data we can clearly see that after about 9.6 there is a small amount of people paying higher insurance costs which is skewing the original data to the right (refer to appendix). We also looked at the cost of insurance based on age and smoking status. In the previous graph we noticed that there are alot of data points on top of each other. After using geom_jitter() and taking the log transform of the cost we can see very clearly that the people paying the most are smokers. The cost of insurance seems to be more influenced on whether or not a person smokes Majority of the people that do not smoke pay less in insurance. However, there are more people that do not smoke that have high insurance costs. Further investigation is needed to see why (refer to Figure 5).

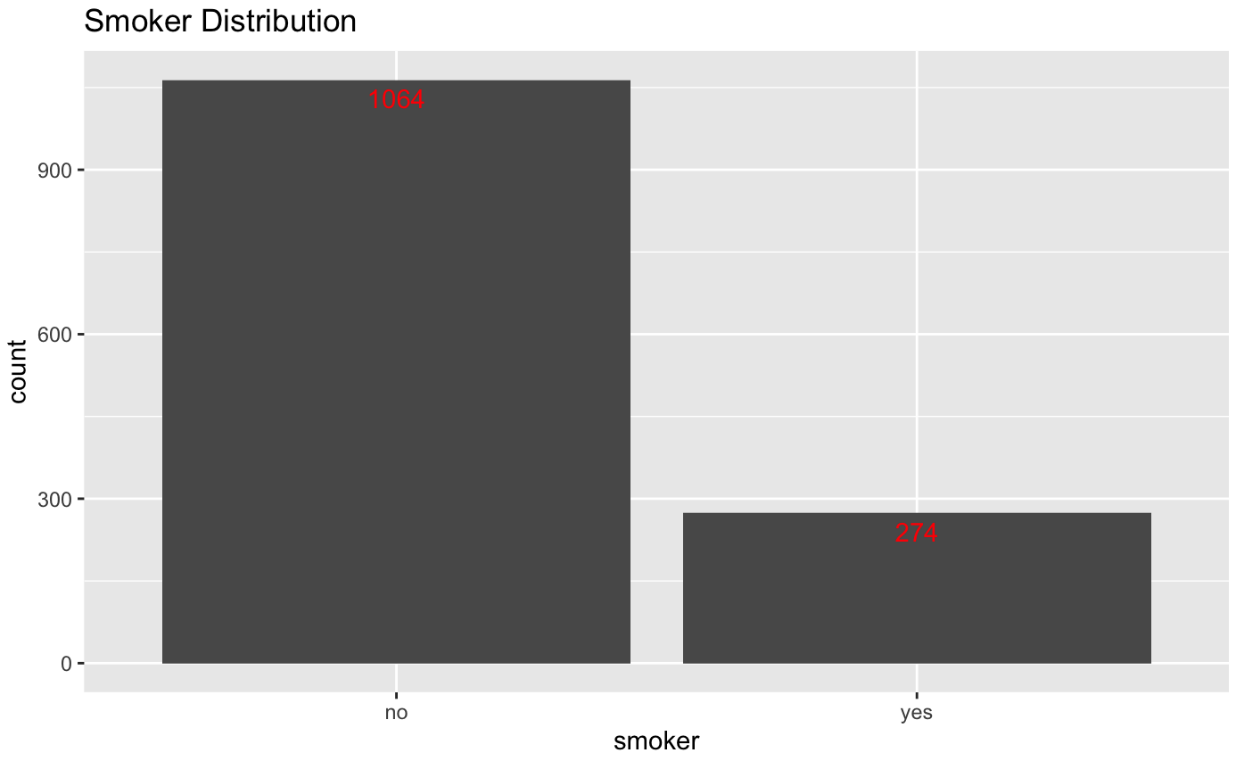


Figure 3: Smoker Distribution

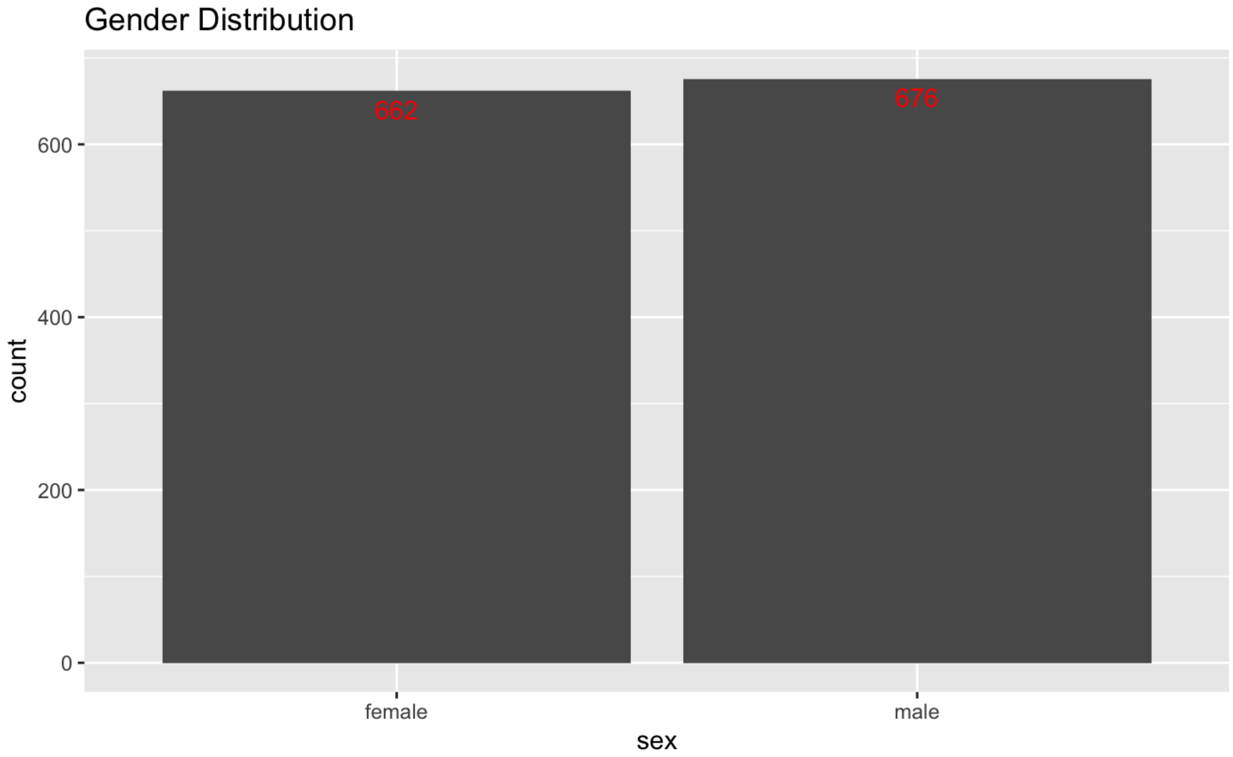


Figure 4: Gender Distribution

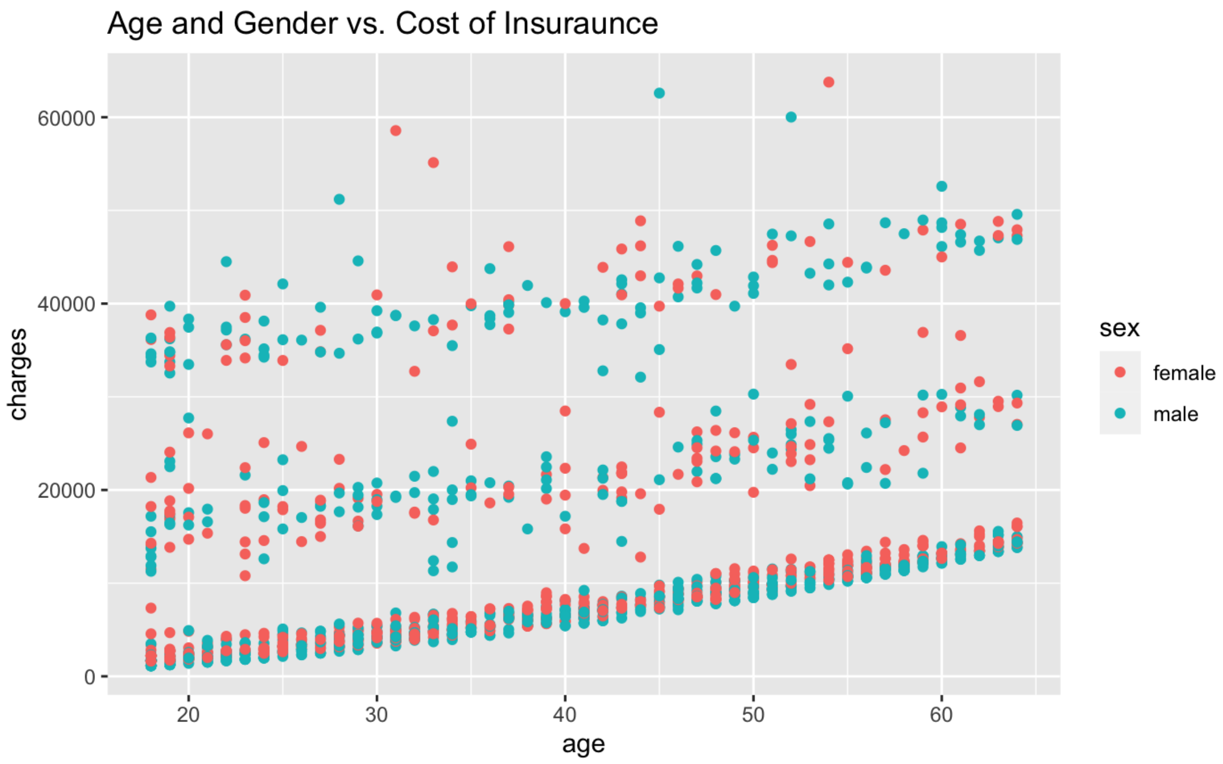
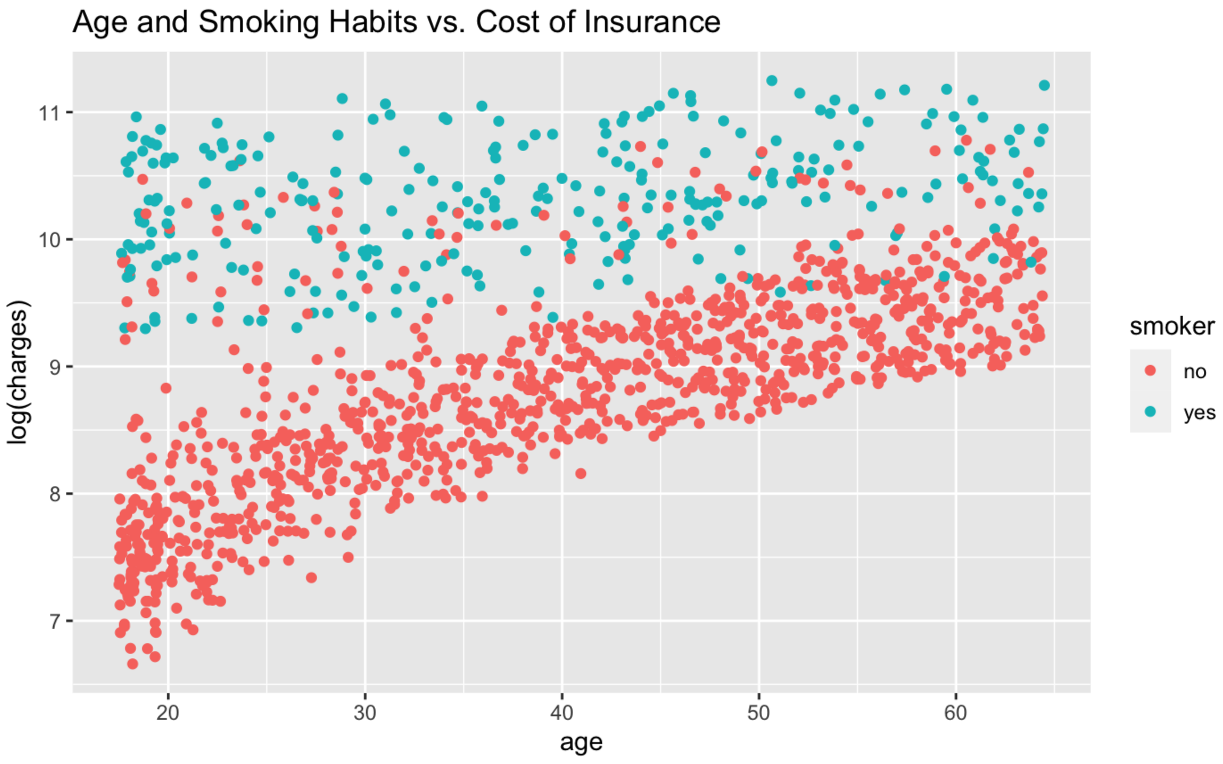


Figure 5: Healthcare Costs vs. Age and Gender



We plotted the cost of insurance based on age and smoking conditions. From the box plots we can see that there is a big difference between what smokers and non-smokers have to pay regardless of age (refer to appendix). The minimum insurance cost for smokers is higher than what 75% of non-smokers pay for insurance (refer to Figure 6).

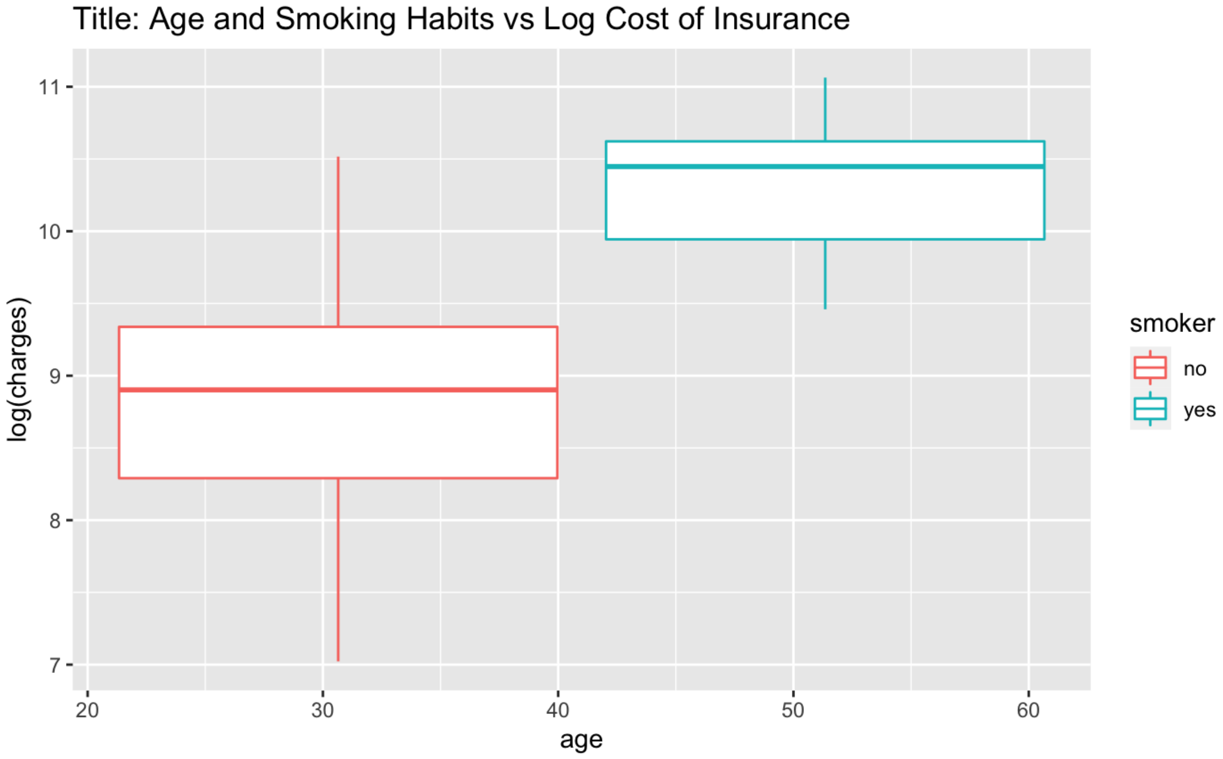


Figure 6: Health Insurance vs. Age and Smoking Habits Boxplot

Moving ahead, we wanted to check if the number of children has an affect on health insurance. We created a face wrap graph for this purpose. In this graph we break down by the number of children a person has. Each subplot represents how many children a group of people have. Here we can see that for those that do not smoke the insurance price is influenced by the number of children that they have. The less children a person has the more likely they are to have a higher insurance cost (refer to appendix).

We finally conclude our EDA by showing heatmaps of the correlation matrix (refer to Figure 7). The heatmap shows which variables have the strongest correlation with a person's insurance cost. Irrespective of location, the strongest influencer is the smoker variable. The next strongest influencer is BMI. Initially, it was observed that gender had a stronger influence on charges than the number of children but as we took a deeper dive into each region that was not the case. The influence of the number of children and gender on charges depends on which region (refer to appendix) a person is in.

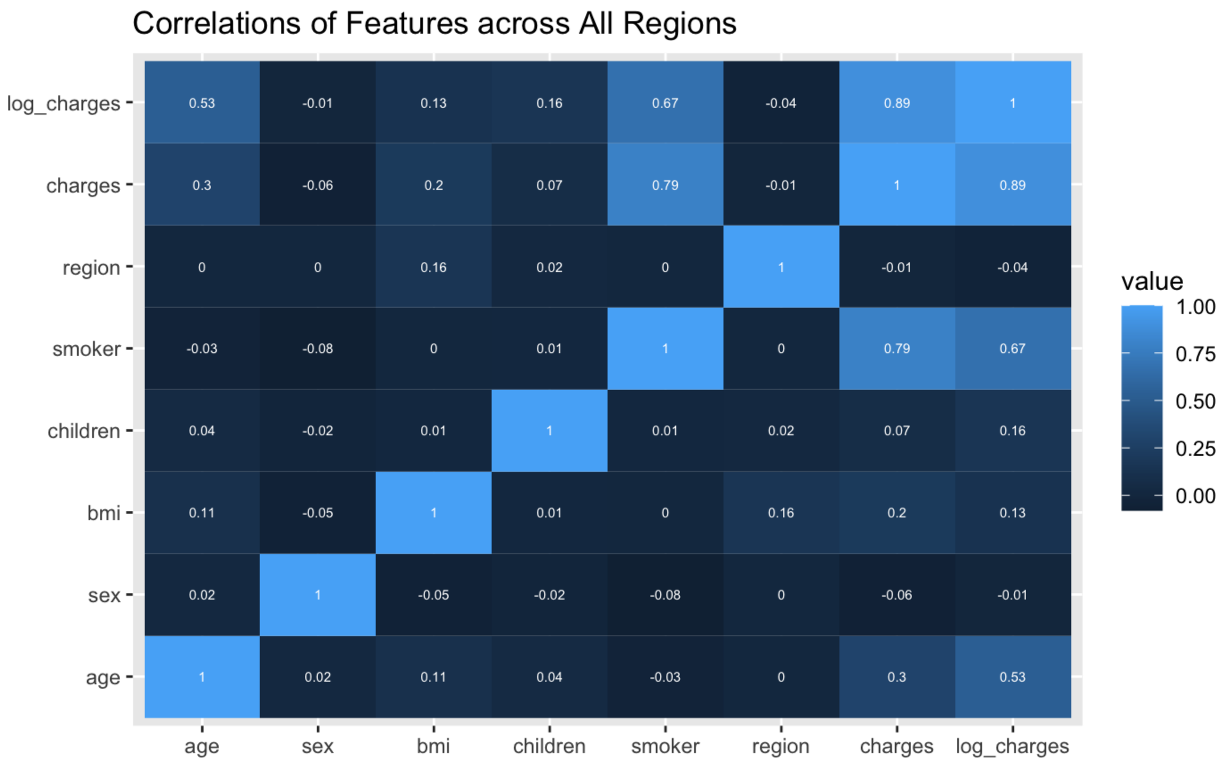


Figure 7: Features Correlation

Methods

Research Questions and Preliminary Hypotheses

Based on the research data and the data set we obtained, we analyzed the factors that may affect the insurance charges. We hypothesized that factors within a degree of patient's control like BMI and smoking habits would correlate with insurance costs. This hypothesis is supported by the correlation chart from the EDA section where smoking and BMI have the strongest correlations to costs. In addition, we would like to expose another factor with a wide range that can be a potentially overwhelming factor in determining costs. From the EDA, the best candidate appears to be age as the feature with the second strongest correlation. We also hypothesized that the number of children and demographics do not impact the medical insurance charges. The data does not show much variation with respect to changing demographics. And the data is insufficient to analyze the variation in insurance prices based on the number of children. Therefore, based on all these factors, we came up with the following questions to conduct the research:

1. Can average insurance prices be predicted based on gender, age, BMI, and smoking habits?
2. Which model is the best predictor of insurance prices given these strongly correlated features?

Choice of models based on data

Linear Regression Model using OLS

We wanted to see if there is a linear relationship between our independent variables (age, bmi and smoker) and our dependent variable (charges) so we decided to use a linear regression model.

A linear regression model makes 5 assumptions which are the Gauss Markov assumptions. These assumptions are: Linearity: The relationship between the dependent variable and independent variable is linear

Non-colinearity: The dependent variables are independent of each other Homoscedasticity: There is not a huge change for each error of the independent variables Random sampling: The data is randomly sampled Exogeneity: Regressors are not correlated with the error term

In addition, to test the level of generalization in the model, we experimented with several splits of training to test data ratios to determine the level of generalization to best model to balance bias-variance in the model.

Random Forest

The second model chosen was a random forest regression model to predict the insurance charges based on a select subset of features in the dataset. Random forest is a type of ensemble method that combines multiple decision trees that uses a technique called bagging to use random subsets of features at each decision node on each tree, then combines the results to predict regression values. Random forest models are useful because the volume of uncorrelated models helps reduce variance by avoiding overfitting training data. A random forest model can help better understand which variable(s) are best to split on first and which variables are good predictors only after another variable is determined. The training to test data ratio was fixed to focus on the search for best hyperparameters for the model.

Models

Expected model performance

OLS and the Gauss Markov conditions

Linearity: We can not test that the independent variables and the dependent have a linear relationship. However, we believe that it is reasonable to assume that these parameters will have a linear relationship

Random sampling: There is no information on data that states whether the data was randomly sampled.

Non-colinearity: From our correlation matrix we did not see a huge correlation between any of the dependent variables.There could be some omitted variables that could affect some of the predictor variables but the cost of insurance and vice versa.

Homoscedasticity: We assume that there is no huge change in error for each predictor variable

Exogeneity: We assume that there is not relationship between each predictor variable and the residuals

Random Forest and Hyperparameters

Using an ensemble method like random forest is expected to produce a low MSE due to a concentrated subset of features to predict insurance cost. In addition, the choice to perform k-cross validation on different hyperparameter values is expected to further improve the model by determining the number of decision trees and the number of variables to split on at each level in the model. It's expected that selecting the right hyperparameters will lower the Mean Squared Error while predicting the insurance cost for the test dataset. Finally, we decided to keep the split to 80% train and 20% test based on our judgment to keep it standardized.

Results

Summary of Models

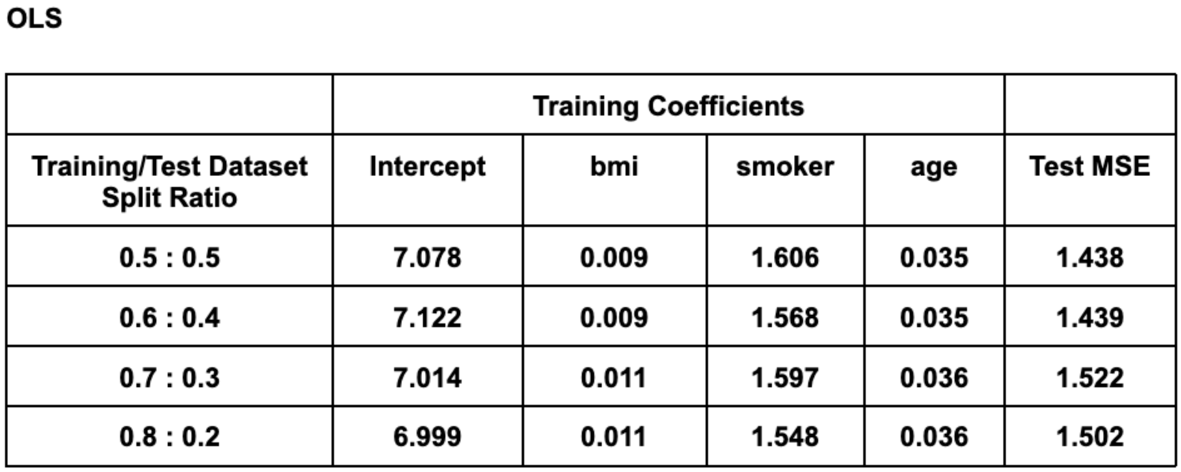
OLS

To prevent the model from having too high variance as a result of overfitting to the training dataset, the OLS was tested with different proportions of training/test to find the right tradeoff between bias and variance. According to our results, the optimal ratio was between 50-60% of the dataset being used for training, resulting in the lowest test MSE. Using the log of the dependent variable helped scale down the features size and made comparison of the coefficients much easier to decipher. The linear regression resulted in coefficients on BMI, Smoker, and Age were approximately 7, 0.01, and 0.03 respectively. The MSE of the optimal OLS for log of the healthcare costs was 1.438. This shows that the linear regression was perhaps not a good model for the data given the higher magnitude of MSE and shows that there might be a different model that will predict better.

Random Forest

The training dataset was sampled pseudorandomly from the dataset and was 80% of the original dataset. To select hyperparameters, we tested the set of (100, 200, 300, 400, 500) for the number of trees in the random forest and (1, 2, 3) features compared at each decision node. For each configuration, we performed a 5-fold cross validation, trained the model on the data minus the fold and then validated on the data in the fold. From this, we calculated the MSE for each configuration and selected the hyperparameters with the lowest MSE. While the number of trees had negligible MSE differences, the best fit for our validation was consistently 2 variables compared at each decision split. For the test dataset, the number of trees selected for the model changed frequently. Included below are sample results of one run of the 5-fold cross validation. Finally, after selecting the hyperparameters the model was tested against the test dataset and our test MSE was consistently around ~0.091.

Tables and Plots of Results



OLS

	Training Coefficients				
Training/Test Dataset Split Ratio	Intercept	bmi	smoker	age	Test MSE
0.5 : 0.5	7.078	0.009	1.606	0.035	1.438
0.6 : 0.4	7.122	0.009	1.568	0.035	1.439
0.7 : 0.3	7.014	0.011	1.597	0.036	1.522
0.8 : 0.2	6.999	0.011	1.548	0.036	1.502

Figure 8: OLS Performance

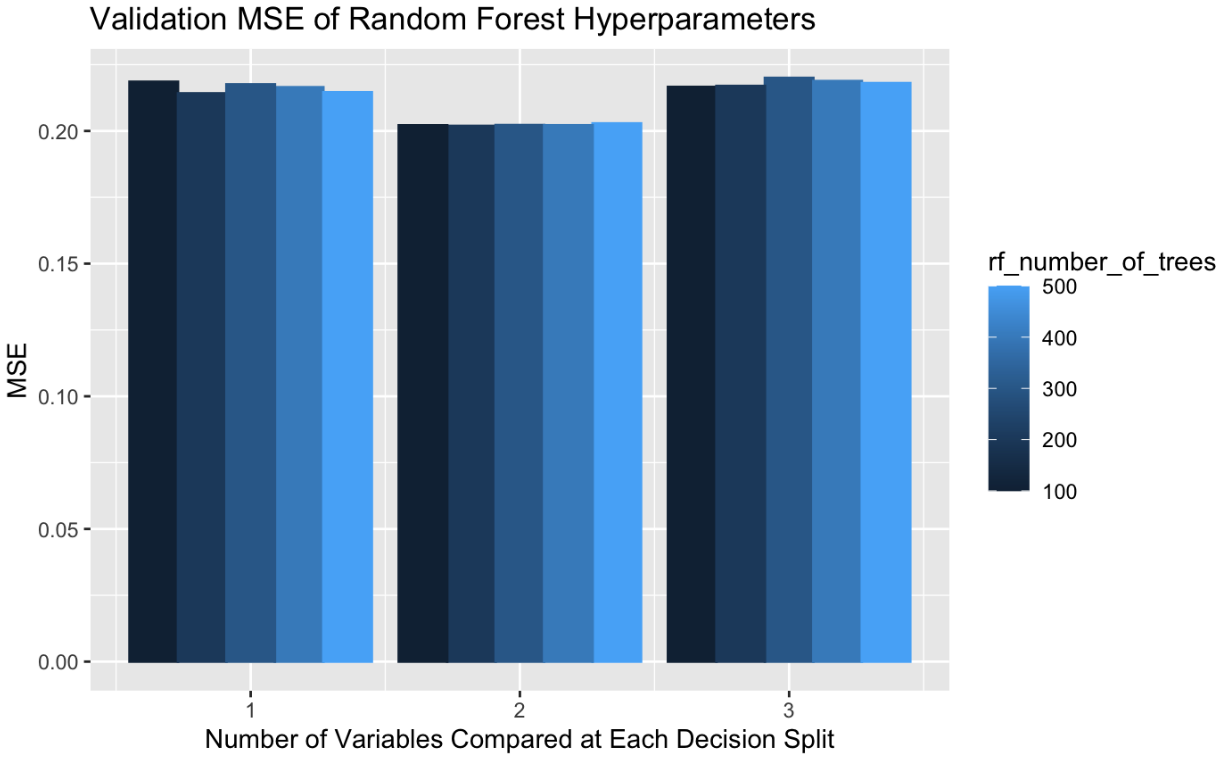


Figure 9: Random Forest Performance

Discussion

Hypothesis Evaluation

Our preliminary hypothesis was that gender would be a strong contributing factor to the cost of healthcare insurance. Literature represents that insurance companies cannot charge separately based on gender and therefore our analysis would determine if men or women are more prone to face medical issues that reflect in the healthcare insurance costing. However, our hypothesis was dismissed as gender did not turn out to be as influential and did not rank in the top three things that affected healthcare costs. Healthcare costs were mostly affected by the smoking status of the individual, followed by age, and then BMI.

It was interesting to find BMI to be one of the influential variables because BMI is not considered to be the best indicator of good health and is only a crude measure of the health status of an individual. BMI does not account for the body fat percentage or distribution. Moreover, there is speculation that BMI ranges were determined from data from white Europeans and therefore is not generalizable to all populations. It is said that blood pressure, waist circumference, and cholesterol levels are better indicators of health status of an individual[17]. But, BMI considers the weight of an individual, which is in control of the individual and therefore is an encouraging finding for our paper. Since there is a clear correlation between obesity and illnesses and it translates to increased healthcare costs, BMI change could contribute in reducing healthcare insurance prices for an individual for the future, irrespective of their age.

Our findings regarding age and smoking being definitive factors that affect the healthcare insurance pricing, aligns with the literature review.

Tradeoffs

There are limitations and drawbacks of the dataset itself. First, the dataset has a low size and therefore any models built on top of it would not be dependable or representative of the population in any manner. There are four sections of the US that are all represented within roughly ~1339 rows. Moreover, the data is also skewed in the representativeness across age groups. Age groups of 18 and 19 years have ~68 and ~69 counts whereas the other age groups are between ~22 to ~30 counts. There is a good balance between the genders as men and women are equally represented in the dataset. However, there could be a difference in the data collection method, where they could have asked people their preferred gender instead of providing only two options of 'men' and 'women' because policies or business decisions have a direct effect on underrepresented communities like the transgender people[16]. Similarly, the four regions have been represented at 324 (northeast), 325 (northwest), 325(southwest), and 364 (southeast) counts. However, these samples are very small to infer anything at a larger scale and it is difficult to generalize any finding with such a small sample of people in the dataset. Finally, the dataset has not provided any information on the methodology of data collection, storage, processing, or purpose of creating this dataset. In the absence of this information it is difficult to understand the context and nature of the data. For example, if this is self reported data, it could be possible that some smokers could have answered the questions as being 'non-smokers' due to the stigma surrounding smoking as a habit. In such a case, the dataset would be non-dependable as the accuracy of the information would be questionable.

Conclusion

Implications in Business and Policy

This paper confirms and highlights a few factors that affect the prices of healthcare insurance. However, there are considerations to be made while using these findings to drive policy or business decisions. The paper does not include, represent, or analyze the factors that could differentially affect transgender individuals. Our current findings suggest that gender does not have any correlation with healthcare costs, but in the absence of data on transgender populations, this finding remains invalidated for direct policy application. Similarly, this paper also does not consider collecting the race or ethnicity of the individual, while there is evidence that healthcare costs are different across races. In the absence of data on race, the findings from this paper would be incomplete for policy application. There is a clear correlation between smoking and the cost of healthcare insurance. Despite this finding being evident in many papers, including ours, there has been minimal effect on the smoking industry. Our analysis clearly signifies that smoking across all ages will give rise to increasing healthcare insurance costs, but it would be useless unless the corporate legal structures surrounding the tobacco and smoking industry make changes to restrict access or availability. Another implication on policy is around the effect of being in a certain geography, and our findings are currently for the overall US population. There would be regional differences and often subjects like healthcare are split between the federal government and the state governments, private and public insurance providers regionally, and also lifestyle of people in varying geographies. Therefore, it would be incorrect to extrapolate findings from this paper and apply them to build policies or make business decisions, as it might result in unanticipated consequences. Finally, if the results from these papers are used to drive awareness around how people might be able to reduce healthcare costs, and it does not yield desired results for the masses, it could dent the policy maker's reputation. Undertaking policy and business decisions should be made on the basis of strong reliable data sources, minimal self reported data, and large representative volumes of diverse covariates. It's equally important to scrutinize the findings with subject matter experts, both in a qualitative and quantitative manner.

Future Work and Limitations

It would be interesting to explore at an older age, if BMI continues to affect the cost of healthcare insurance. Our current data analysis shows that both age and BMI play an important role in healthcare insurance pricing. However, at an early age it is clear that weight can be reduced and therefore avoid illnesses that could later cause a bump in pricing. However, at an older age, it would be interesting to find out how much BMI affects the pricing, and if it doesn't, what factors influence the pricing. This would help understand what targeted interventions might be required for the eldery population. There is always the limitation of the effectiveness of the intervention at a certain age and also the intervention compliance of the individual. It would also be interesting to further understand the changes in factors based on geography, race, ethnicity, and gender. The current limitation of the paper is the volume and variance in the data for the above mentioned factors. Finally, any future work on healthcare and policy should include subject matter experts, stakeholders in the healthcare value chain (doctors, patients, providers, etc.), and data science should be used to support decision making and not replace human expertise in policy designing.

Appendix

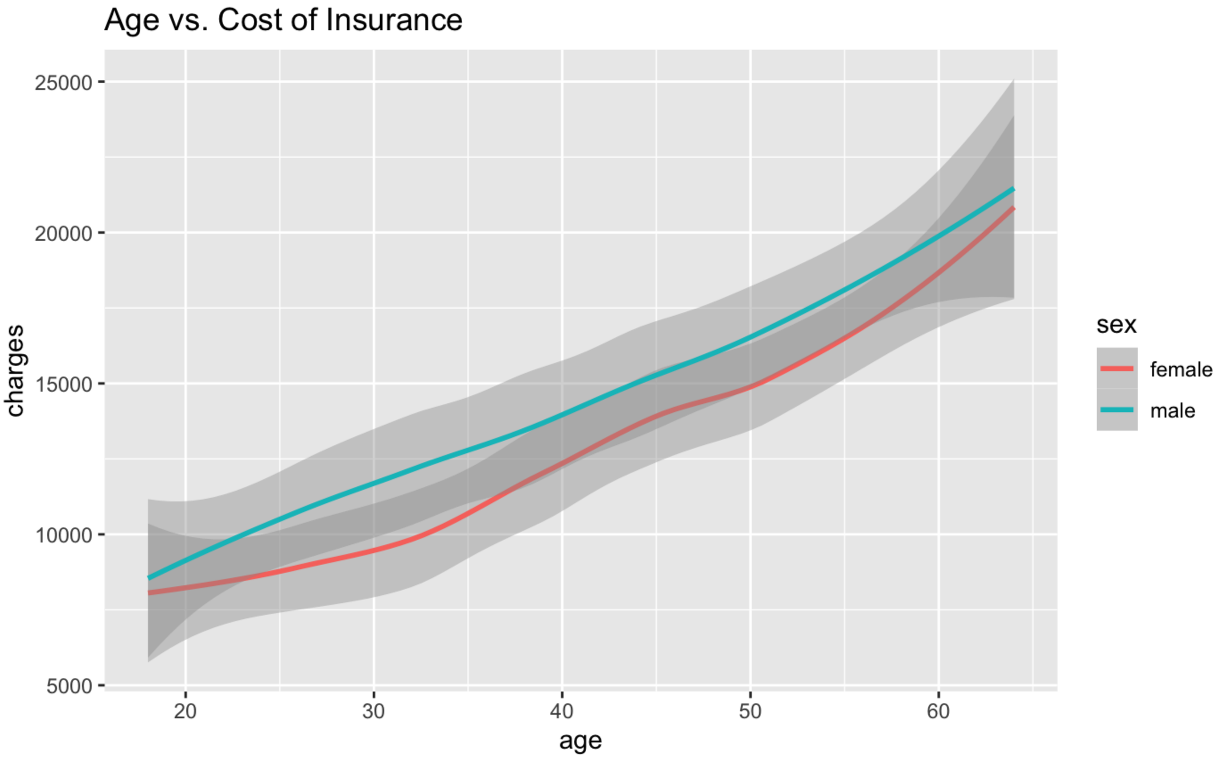


Figure 10: Cost Differences by Age and Gender

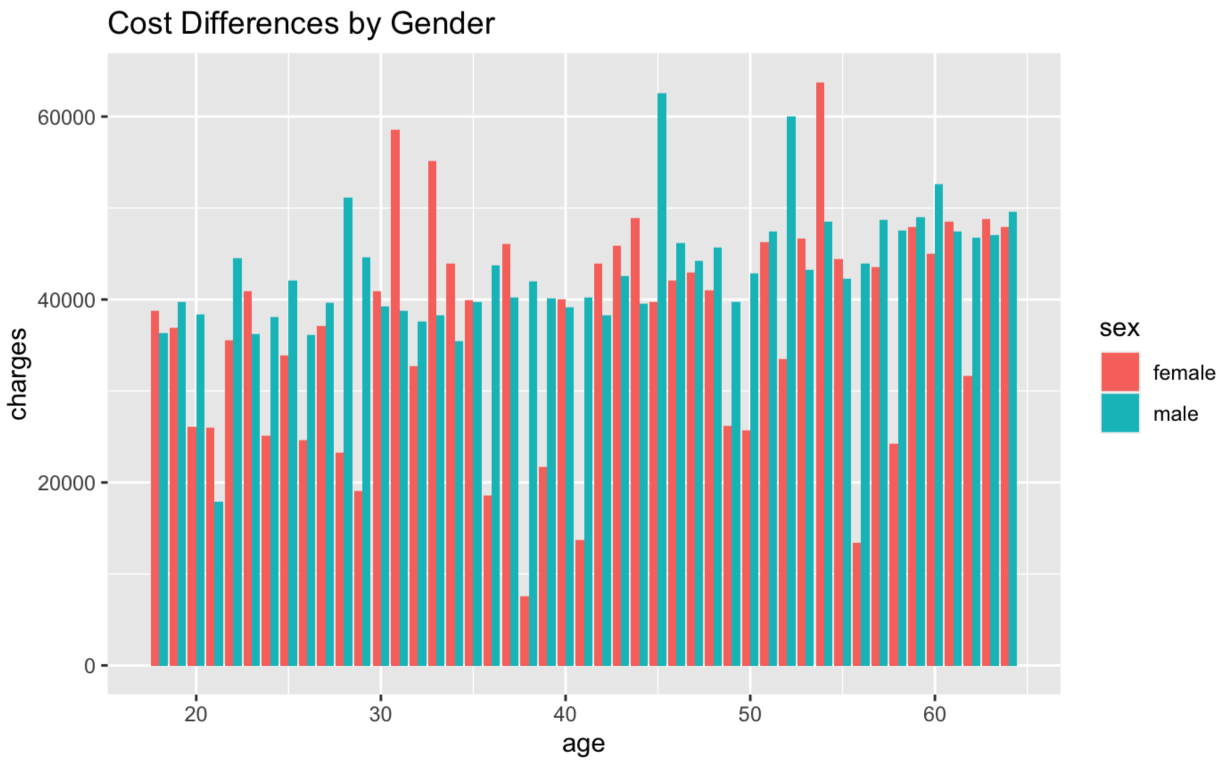


Figure 11: Cost Differences by Age and Gender

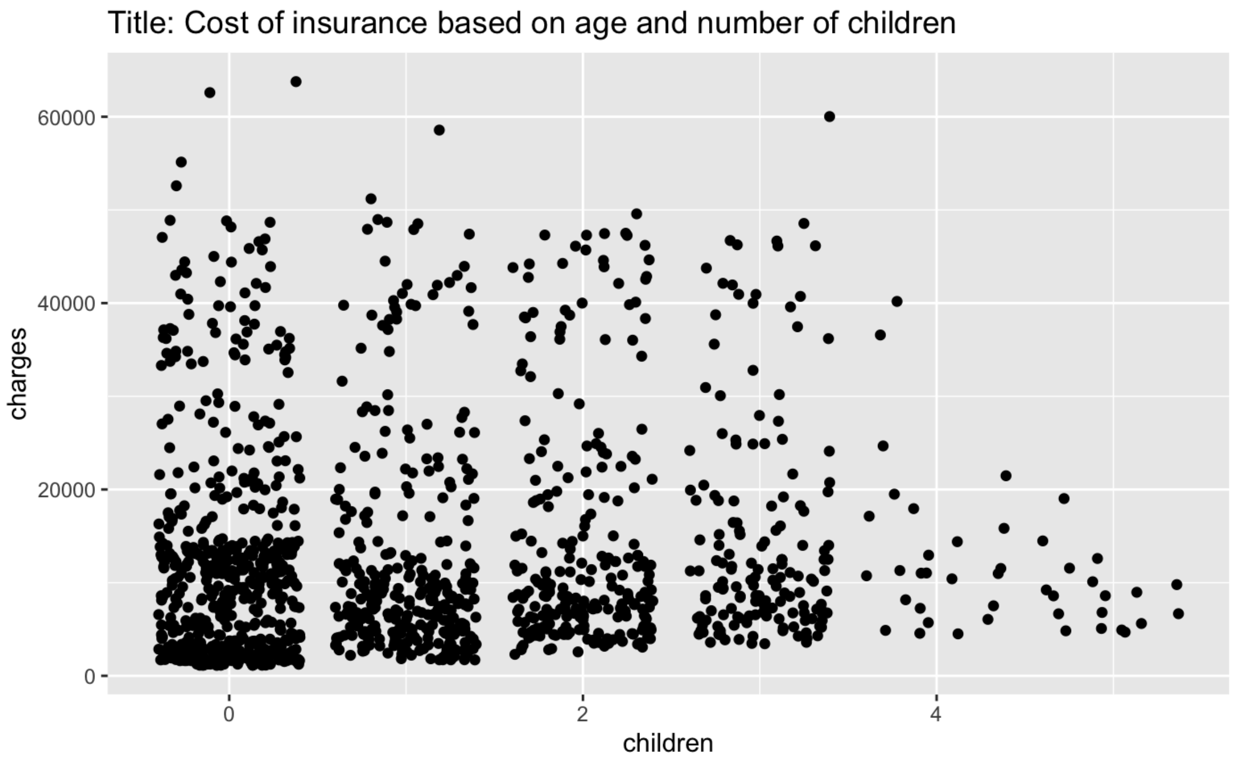
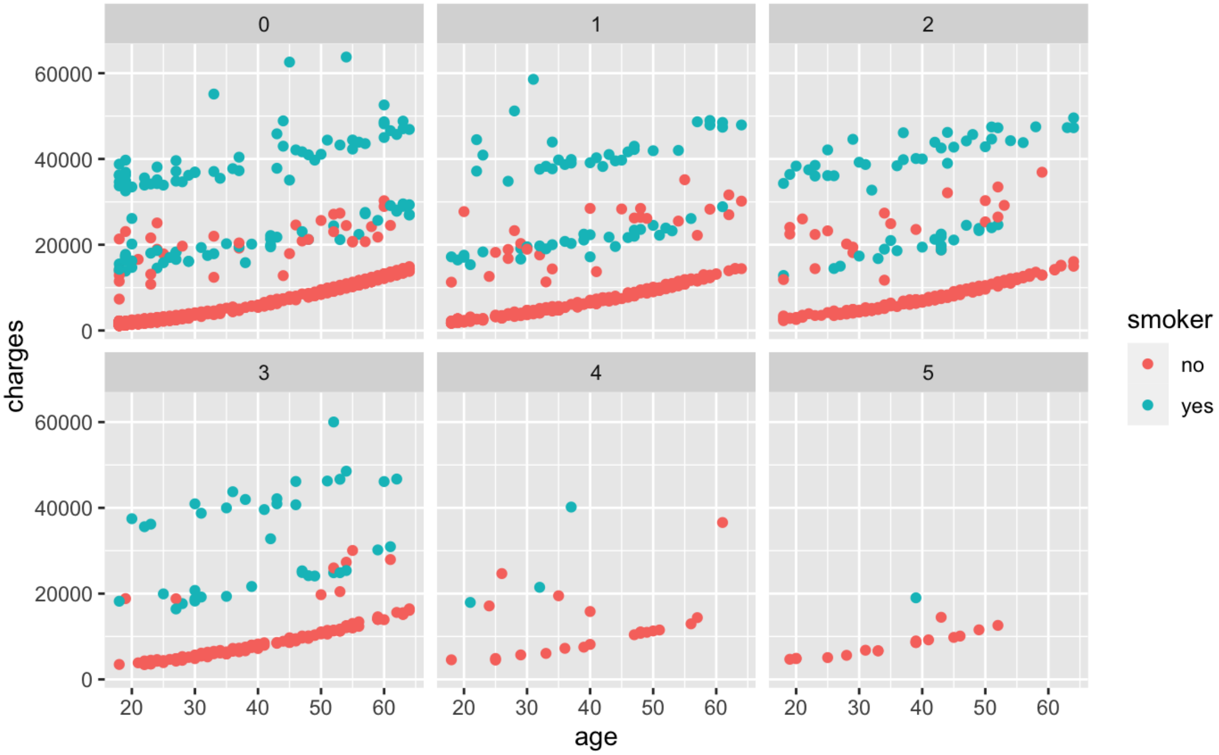
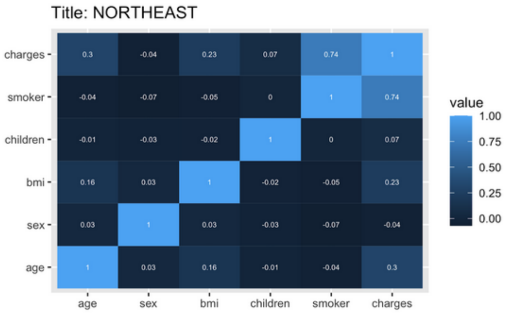
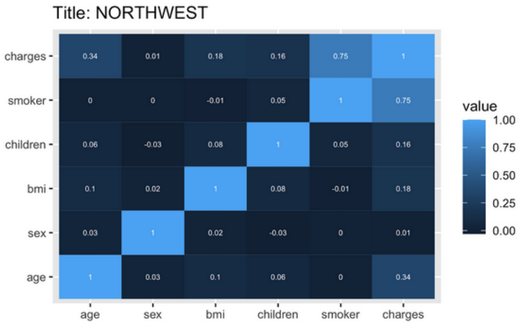


Figure 12: Cost Differences by Age and Children

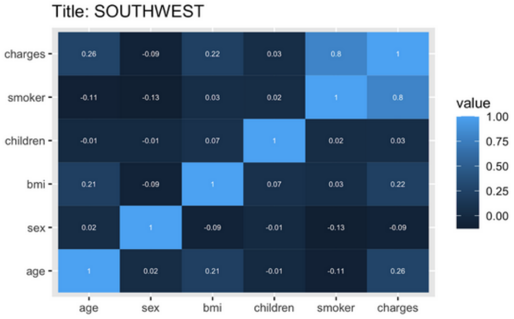


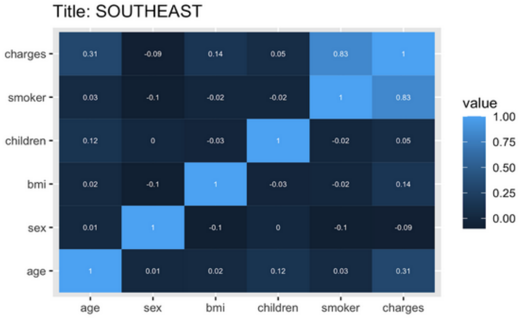
In the northwest and northeast children have a stronger influence on charges than gender.

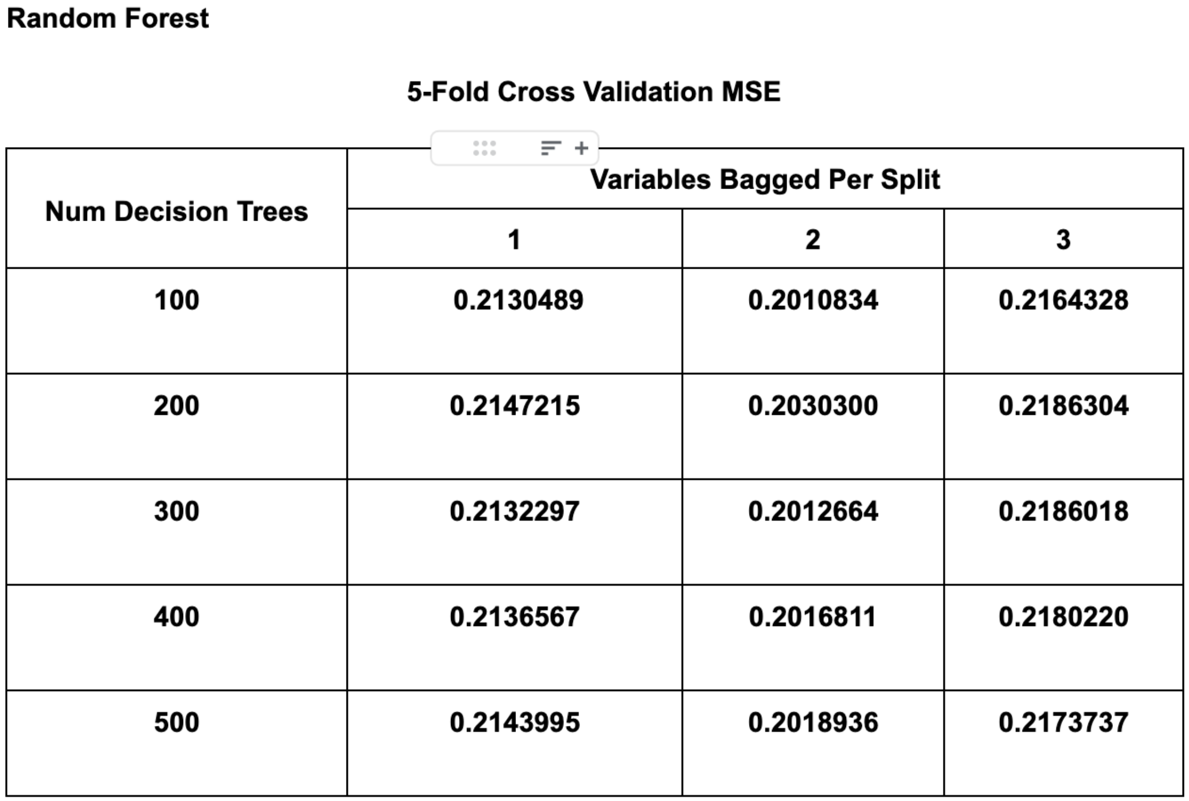




In the southwest and southeast gender has a stronger influence on charges than the number of children.







Random Forest

5-Fold Cross Validation MSE

Num Decision Trees	Variables Bagged Per Split		
	1	2	3
100	0.2130489	0.2010834	0.2164328
200	0.2147215	0.2030300	0.2186304
300	0.2132297	0.2012664	0.2186018
400	0.2136567	0.2016811	0.2180220
500	0.2143995	0.2018936	0.2173737

References

1. Machine Learning in Healthcare: A Review
   https://ieeexplore.ieee.org/abstract/document/8474918?casa_token=vs5rB2oeF7sAAAAA:d47DY_CRrzveZzUNJSDaYnlNqZLN054lFu8Pz1LMc-OpzV-IZBKteFOzGKMOTSpUZyjoU7Yi
2. Key Advances in Clinical Informatics; Chapter 19 - Machine Learning in Healthcare
   https://www.sciencedirect.com/science/article/pii/B9780128095232000194
3. Kang, J. S., Kuznetsova, P., Luca, M. & Choi, Y. Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. In Proc. 2013 Conference on Empirical Methods in Natural Language Processing 1443–1448 (Association for Computational Linguistics, 2013).
4. Chandler, D., Levitt, S. D. & List, J. A. Predicting and preventing shootings among at-risk youth. Am. Econ. Rev. 101, 288–292 (2011).
5. Big Data and Data Science in Critical Care
   https://www.sciencedirect.com/science/article/abs/pii/S0012369218307256
6. Athey, S. Beyond prediction: using big data for policy problems. Science 355, 483–485 (2017).
7. O'Neil, C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (Broadway Books, New York, USA, 2016).
8. States Divided: The Implications of American Federalism for COVID-19
   https://onlinelibrary.wiley.com/doi/full/10.1111/puar.13243?casa_token=DJZythyJAP0AAAAA%3AmVCbaBPWY0ENhQIxCpUccxkpHpnXGu2F5QTt41hl16B0YtkuR8_RTu6oINxvFO7YsCeZ3fEqfrsEy9U
9. Health Insurance Coverage in the United States: 2020
   https://www.census.gov/content/dam/Census/library/publications/2021/demo/p60-274.pdf
10. Should All Americans Have the Right (Be Entitled) to Health Care?
    https://healthcare.procon.org/

11. Factors affecting healthcare costs
https://www.investopedia.com/how-much-does-health-insurance-cost-4774184
12. U.S. Centers for Medicare & Medicaid Services. “How insurance companies set health premiums
https://www.healthcare.gov/how-plans-set-your-premiums/
13. Affordable Care Act
https://www.investopedia.com/terms/a/affordable-care-act.asp#citation-11
14. Cost of health insurance in the US
https://www.ehealthinsurance.com/resources/individual-and-family/how-much-does-individual-health-insurance-cost
15. Census report on income and poverty in the US
https://www.census.gov/library/publications/2021/demo/p60-273.html
16. Laws and transgender population
https://www.lgbtmap.org/equality-maps/healthcare_laws_and_policies
17. BMI is not a good indicator of health
https://www.medicalnewstoday.com/articles/265215#Waist-size-linked-to-diabetes-risk,-regardless-of-BMI