University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING
Electrical & Computer Engineering and Computer Science

Electrical & Computer Engineering & Computer Science (ECECS)

TECHNICAL REPORT

SPRING 22

TECHNICAL REPORT

Risk Analysis for Health Insurance Companies
An Exploratory Data Analysis with Visualization & Hypothesis Testing

Group 1
Abiral Shrestha (ashre10@unh.newhaven.edu)
Jyoti Bhandari (jbhan1@unh.newhaven.edu)
Sudip Adhikari (sadhi9@unh.newhaven.edu)

University of New Haven
Electrical & Computer Engineering & Computer Science (ECECS)
West Haven, CT

March 09, 2022

Table of Contents

Executive Summary/Abstract

In this project, a Risk Analysis is conducted to answer questions that health insurance companies might have and would be pleased to have answers to. The health data that our analysis includes features of patients such as age, gender, Body Mass Index(bmi), blood pressure, diabetic condition, number of children, smoking habits, region, and claims made by an insuree. We start with Exploratory Data Analysis where we carry out initial investigations on data to find patterns, detect anomalies, and check assumptions accompanied by visualization. Then we move on to establishing some hypotheses and use statistical methods to test them to reach conclusions eventually. We also examine the correlation between all the fields and extract meaningful insights from the data.

Our analysis shows that there are more outliers in claims data in comparison to the other fields. This aligns with the understanding that most people require basic medication and treatments while only some require expensive medication and treatments. Thus, the various different claim amounts give rise to a high number of outliers. We were also able to find out that gender has no impact on BMI, the number of children has no effect on BMI, and that increase in the number of children doesn't mean an increase in claims made by parents. Also, our findings suggest that smoking habit is not related to age and that males smoke more than females among smokers and so confirm the hypothesis that gender has an effect on smoking habits. More claims were found to be made by people who smoke and we could thus accept the hypothesis that smoking habits have an impact on the increase of claims.

Looking into the data on blood pressure and diabetic patients, we were able to find out that most of the diabetic patients have low blood pressure and that being diabetic and having high blood pressure is not related. Also, higher claims were made by people with higher blood pressure and so, we could confirm that high blood pressure is related to higher claims. We could not precisely conclude the relationship between diabetic patients and an increase of claims, so we accept the hypothesis that being diabetic has no impact on the increase of claims.

Our main objective was to analyze data to see how patients with certain habits, features, and lifestyles can be of risk to the health insurance companies and then accordingly help the companies to better understand their customers and make better decisions.

Highlight of the Project

Eligibilities, Claims (medical data), RXclaims (pharmacy data), lab data, HRA data, and other data sets make up the US healthcare domain so big. In this project, to extract information and determine the link between a large amount of patients' data and their ten characteristics: patientId, age, gender, BMI, blood pressure, diabetic condition, number of children, smoking habits, region, and insuree claims, we performed exploratory data analysis, formulate hypotheses, and test with the statistical methods such as the t-test, chi-square test, ANOVA test, and Pearson correlation test.

According to our research, there are a lot of outliers due to the diverse claim amounts which provide information that based on illness intensity, the majority of people require medication and therapies whereas fewer people require expensive treatments. It showed that gender and rise in the number of children have no bearing on BMI and it does not imply an increase in parental claims. Similarly, the findings indicate that smoking behaviors are unrelated to age and that males smoke more than females among smokers, confirming the theory that gender influences smoking habits. However, individuals who smoke are more likely to file insurance claims. It is also confirmed that people with high blood pressure claim more in the insurance and that the majority of diabetic people have low blood pressure, and that diabetes and high blood pressure are unrelated but definite conclusions about the relationship between diabetic patients and an increase in claims were not completely covered.

With Health Insurance firms spending a lot of money on research to figure out how to understand their customers' claims patterns, this research is confident that it will increase the number of customers for the insurance firm while also reducing loss and boosting profit.

Introduction

Facing risks is an unavoidable part of all our lives and most of us would definitely try to eliminate risks or at least try to minimize them. Understanding the importance of risk management, a ‘Risk Analysis’ to answer questions that can help health insurance companies caught our interest. Having worked on US healthcare data for about 2.5 years, we have some knowledge and experience in this domain that we could utilize for the project.

US healthcare is a huge industry and there are many data sets such as Eligibilities, Claims (medical data), RXclaims (pharmacy data), lab data, HRA data, etc., that make up the US healthcare data. HIPAA Privacy Rule provides federal safeguards for personal health information (PHI) stored by covered companies, as well as a variety of rights for patients. Thus, all those data are not open source and so, we will be working with a small portion of the US healthcare claims data.

Review of Available Research

The leading health insurance companies utilize data and advanced analytics to review risk assessment, enrich the customer experience, and enhance efficiency and decision making throughout the underwriting process. With these insights, actions can accordingly be taken for loss prevention as well. These companies frequently utilize the plethora of third-party data from a wide range of domains, including government data, environmental data, industry-specific data, location data, and more. With all these resources, they have developed agile capabilities to acquire, test, maintain, use, and reuse the data in their models.

The insurance companies still have some way to go before they can become fully data-driven. Some of the challenges are already clear to the industry. Such challenges include:

- Unstructured form of data
- Difficulty in the synthesis and integration of data due to the the siloed nature of data collected
- Fraud cases due to the outdated fraud detection technology that cannot keep pace with modern fraud types and level
- Availability of trained and skilled professionals

Moreover, the most challenging issue focuses on a customer's right to privacy. A multitude of federal and state regulations are enacted to protect consumer privacy and avoid discriminatory practices in the Finance Industry. Successive stringent rules on data collection have been added to regulations, regarding which the insurance legal department should also be up to date on.

Nevertheless, much research has been done on the advantages of claims data in healthcare research. Hundreds of medical research projects have relied on claims data in the last 25 years. Claims data, also known as billing or administrative data, has been used to research a variety of topics, including the use of comorbidity indices to predict risk of death, antiretroviral therapy, psychotropic drug usage, children's mental health services, substance use disorders, the cost effectiveness of lung-volume–reduction surgery, diabetes preventive services, and a variety of others.

Health and life insurance companies have recently used them as data to make decisions on whether or not to insure certain individuals. Many big companies such as Deerhold, Cotiviti etc. are working on the US healthcare domain to give risk-assessment services and decision analytics within the area of U.S healthcare that helps customers better understand and manage their risk. Thus, if managed and done properly, the pros of data science and analytics far outweigh the cons.

This leads us to our research question: **How are different fields in data correlated and how can that relationship between data be used to make better decisions for health insurance companies?**

Looking at the claims data we have, we can pose the following questions and perform hypothesis testing to find out the answers:

Questions	**Hypothesis**
1. Is there a relationship between Body mass index and gender?	Ho = "Gender has no impact on bmi." Ha = "Gender has an impact on bmi."
2. Are people with more children claiming more?	Ho = "Increase in number of children doesn't mean increase in claims made by parents." Ha = "Increase in number of children mean increase in claims made by parents."
3. Does the proportion of smokers significantly differ in different genders?	Ho = "Gender has no effect on smoking habits." Ha = "Gender has an effect on smoking habits."
4. Are age and smoking habits correlated?	Ho = "Smoking habits is not related to age." Ha = "Smoking habits is related to age."
5. Does the number of children have any effect on the BMI of females?	Ho = "No. of children has no effect on bmi of females." Ha = "No. of children has an effect on bmi of females."
6. Is the patient having higher blood pressure also diabetic?	Ho = "Having high blood pressure and being diabetic are not related." Ha = "Having high blood pressure and being diabetic are related."
7. If a patient who smokes has made higher claims?	Ho = "Smoking habits has no impact on increase of claims." Ha = "Smoking habits has an impact on increase of claims."
8. If a patient who is diabetic has made higher claims?	Ho = "Being diabetic has no impact on increase of claims." Ha = "Being diabetic has an impact on increase of claims."
9. If a patient who has high blood pressure has made higher claims?	Ho = "High blood pressure is not related to higher claims." Ha = "High blood pressure is related to higher claims"

Methodology

We researched through a lot of data resources and came across https://data.world/ that provides good sources of data. We as a team have some experience working with US healthcare data so we decided to work on the healthcare domain. US healthcare is a huge industry and there are many data sets such as Eligibilities, Claims (medical data), RXclaims (pharmacy data), lab data, HRA data, etc., that make up the US healthcare data. HIPAA Privacy Rule provides federal safeguards for personal health information (PHI) stored by covered companies, as well as a variety of rights for patients meaning all those data are not open source so, we are taking a little sneak peek into a small part of the US healthcare claims data. We came across the insurance data that contains 1340 patients' data and their 10 features: patientId, age, gender, Body Mass Index(bmi), blood pressure, diabetic condition, number of children, smoking habits, region and claims made by insuree. Please refer the following diagram:

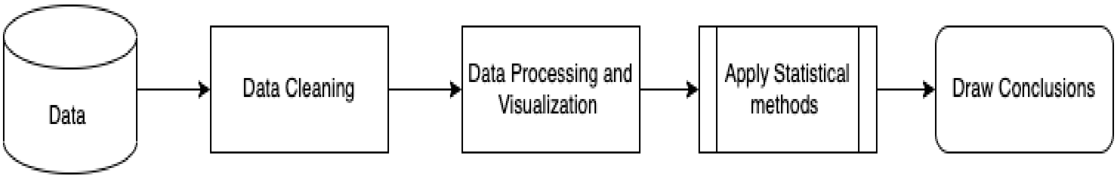


Fig. i: Data Science Process

Data Cleaning:

On the initial analysis, we found 5 null values on the field age and 3 null values on the field region. So, for the first phase of the data analysis i.e., data cleaning, we first identified the missing data. We could identify that the missing data in age can be filled in using the method mean or median as there were no outliers in the age data. We used median of the age data and filled up the 3 missing values. For the region data, there was no way to fill in the missing data, so we replaced the null values with the value ‘Unknown’.

Data Visualization:

For visualizing the data, we have used seaborn library.

Data analysis and Hypothesis Testing:

We have 9 hypotheses to determine relationship between different variables and based on the requirement of the variables included and the hypotheses we have used the following statistical methods to test the hypotheses:

Independent t-test:

This is a parametric test used to test for a statistically significant difference in the means between 2 groups. This is for the comparison of categorical and continuous variables. We can perform Independent t-test using the library scipy.stats.ttest_ind in python.

For the independent t-test, there are null and alternative hypotheses.
The independent t-null test's hypothesis is that the population means of the two unrelated groups are equal (meaning they are independent):

H0: u1 = u2

In most circumstances, we are aiming to see if we can prove that we can reject the null hypothesis and accept the alternative hypothesis, which is that the population means are not equal (meaning they are dependent):

HA: u1 ≠ u2

To do so, we must first determine a significance level (also known as alpha) that allows us to reject or accept the alternative hypothesis. This value is most typically set to 0.05.

Chi-square (χ2) test of Independence:

It is used to decide whether a relationship exists between two variables of a population. It is useful when analyzing survey results of 2 categorical variables.

H_0: The two categorical variables have no relationship.
H_1: There is a relationship between two categorical variables.

Chi-square test of independence assumptions are:

- The two samples are independent
- No expected cell count is = 0
- No more than 20% of the cells have and expected cell count < 5

We used the method scipy.stats.chi2_contingency to implement this test in python. To do so, we must first determine a significance level (also known as alpha) that allows us to reject or accept the alternative hypothesis. This value is most typically set to 0.05.

ANOVA test:

ANOVA, or Analysis of Variance, can be regarded as a generalization of t-tests for more than two groups. The independent t-test is used to compare the means of two groups in a condition. When we wish to compare the means of a condition between more than two groups, we employ ANOVA.

Before we can conduct any tests, we must first specify the null and alternate hypotheses:

Null hypothesis: No significant differences exist between the groups.
Alternate Hypothesis: A significant difference exists between the groups.

Essentially, ANOVA compares two forms of variation: difference between sample means and variation within each sample. In python we used the method scipy.stats.f_oneway to perform this test.

Pearson correlation coefficient:

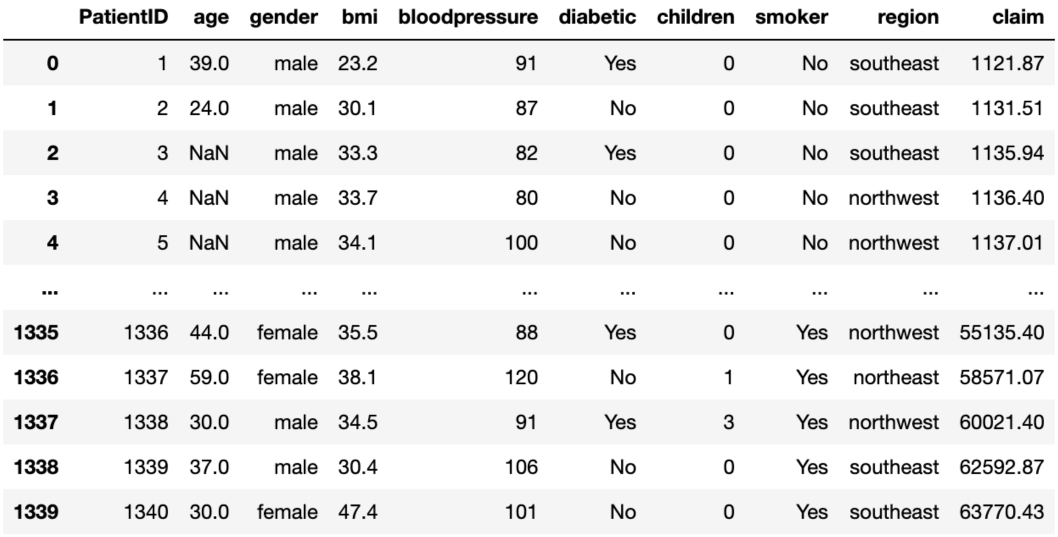
The Pearson correlation coefficient (named after Karl Pearson) is a measure of the strength of a linear relationship between two sets of data. It determines the relationship between two continuous variables.

The Pearson's correlation coefficient is derived by dividing the covariance of two variables by the product of each data sample's standard deviation. It's the process of converting the covariance between two variables into a score that can be understood.

The null hypothesis is that the correlation coefficient does not differ significantly from zero. In the population, there is no substantial linear relationship (correlation) between x and y. Our alternate hypothesis is that the population correlation coefficient differs from 0 in a substantial way. In the population, there is a largely linear relationship (correlation) between x and y.

Result Section

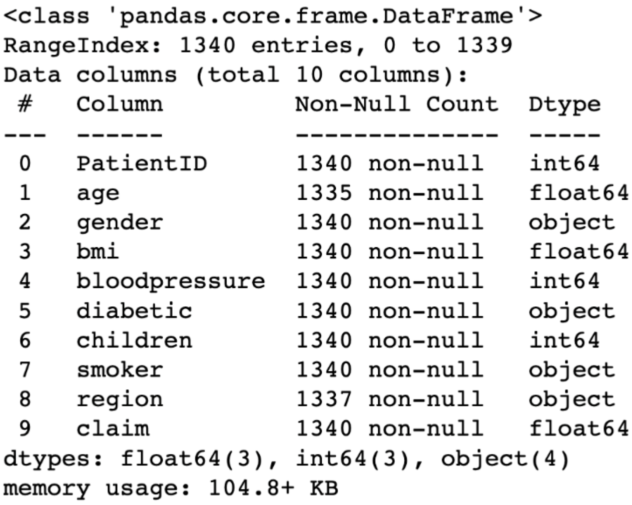
Fig. 1 shows us what our data set looks like along with the total number of samples and its features. With this, we can get an idea of what kind of data we are working with and then move on to data wrangling.



	PatientID	age	gender	bmi	bloodpressure	diabetic	children	smoker	region	claim
0	1	39.0	male	23.2	91	Yes	0	No	southeast	1121.87
1	2	24.0	male	30.1	87	No	0	No	southeast	1131.51
2	3	NaN	male	33.3	82	Yes	0	No	southeast	1135.94
3	4	NaN	male	33.7	80	No	0	No	northwest	1136.40
4	5	NaN	male	34.1	100	No	0	No	northwest	1137.01
...	...	...	...	...	...	...	...	...	...	...
1335	1336	44.0	female	35.5	88	Yes	0	Yes	northwest	55135.40
1336	1337	59.0	female	38.1	120	No	1	Yes	northeast	58571.07
1337	1338	30.0	male	34.5	91	Yes	3	Yes	northwest	60021.40
1338	1339	37.0	male	30.4	106	No	0	Yes	southeast	62592.87
1339	1340	30.0	female	47.4	101	No	0	Yes	southeast	63770.43

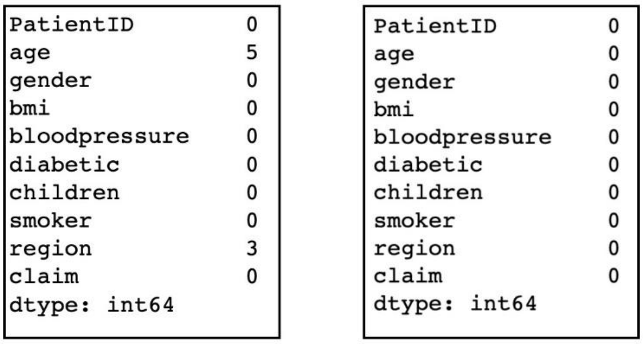
Fig. 1: Data Set Representation

We can observe the data information from Fig. 2 and can see that there are some null values in the data. We can observe the null value information before and after data wrangling from Fig. 3.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1340 entries, 0 to 1339
Data columns (total 10 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   PatientID      1340 non-null   int64
 1   age            1335 non-null   float64
 2   gender         1340 non-null   object
 3   bmi            1340 non-null   float64
 4   bloodpressure  1340 non-null   int64
 5   diabetic       1340 non-null   object
 6   children       1340 non-null   int64
 7   smoker         1340 non-null   object
 8   region         1337 non-null   object
 9   claim          1340 non-null   float64
dtypes: float64(3), int64(3), object(4)
memory usage: 104.8+ KB
```

Fig. 2: Data Information

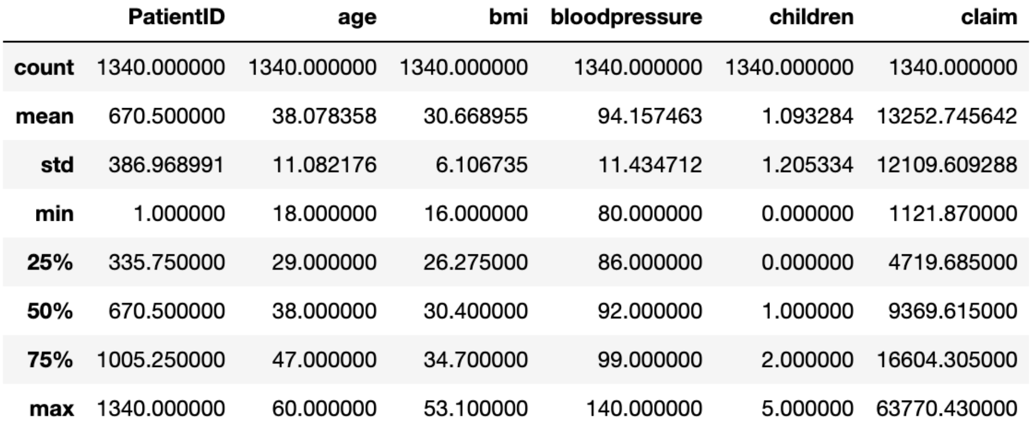


```
PatientID        0
age              5
gender           0
bmi              0
bloodpressure    0
diabetic         0
children         0
smoker           0
region           3
claim            0
dtype: int64
```

```
PatientID        0
age              0
gender           0
bmi              0
bloodpressure    0
diabetic         0
children         0
smoker           0
region           0
claim            0
dtype: int64
```

Fig. 3: Null Values Before and After Data Wrangling

After data wrangling, we can then observe some statistical information regarding the data set as shown in Fig. 4.



	PatientID	age	bmi	bloodpressure	children	claim
count	1340.000000	1340.000000	1340.000000	1340.000000	1340.000000	1340.000000
mean	670.500000	38.078358	30.668955	94.157463	1.093284	13252.745642
std	386.968991	11.082176	6.106735	11.434712	1.205334	12109.609288
min	1.000000	18.000000	16.000000	80.000000	0.000000	1121.870000
25%	335.750000	29.000000	26.275000	86.000000	0.000000	4719.685000
50%	670.500000	38.000000	30.400000	92.000000	1.000000	9369.615000
75%	1005.250000	47.000000	34.700000	99.000000	2.000000	16604.305000
max	1340.000000	60.000000	53.100000	140.000000	5.000000	63770.430000

Fig. 4: Statistical Information on the Data Set

Fig. 5 shows us the normal distribution of some of the features among which we can see that bmi is more normally distributed in comparison to others.

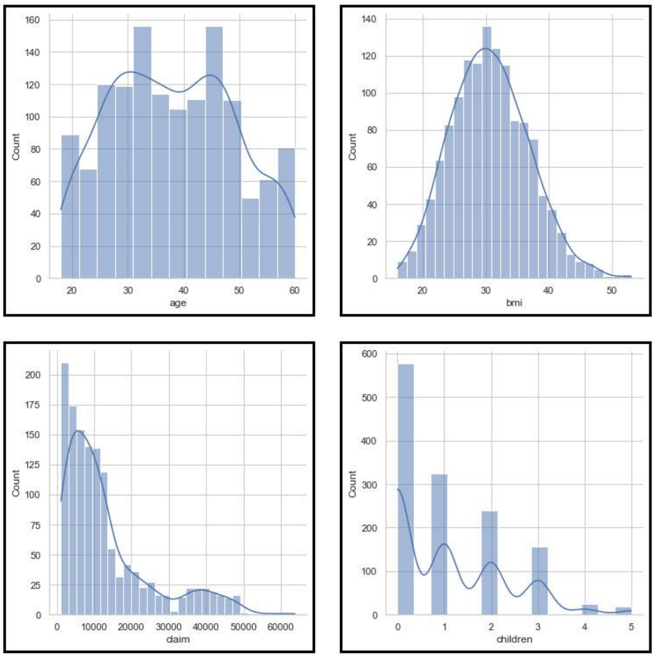


Fig. 5: Normal Distributions

We can observe the outliers in the data set from Fig. 6. We can see that *bmi* and *blood pressure* have some outliers while *claim* has a lot more outliers in comparison to the others.

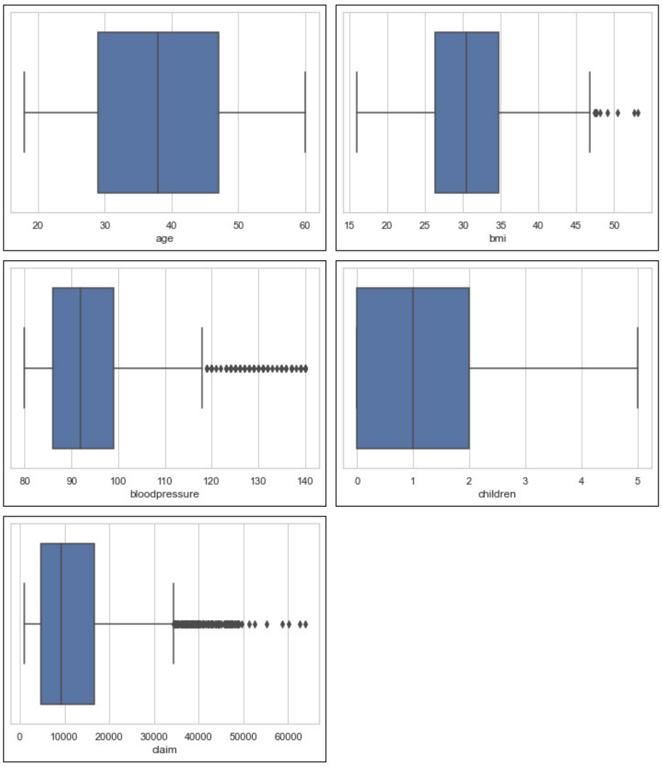


Fig. 6: Outliers

Fig. 7 provides the visualization for the first question: Is there a relationship between Body mass index and gender?

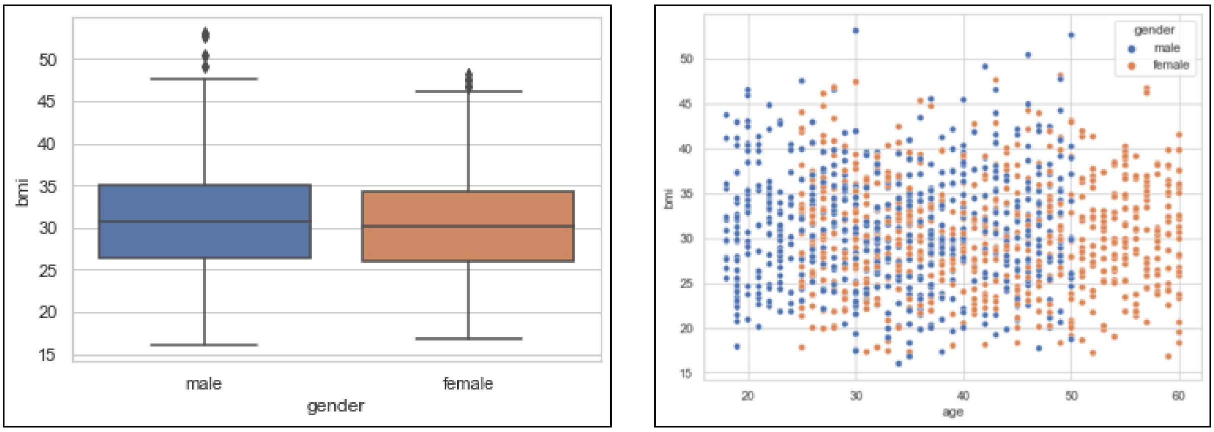


Fig. 7: Box Plot and Scatter Plot of Relationship between BMI and Gender

Fig. 8 provides the visualization for the second question: Are people with more children claiming more?

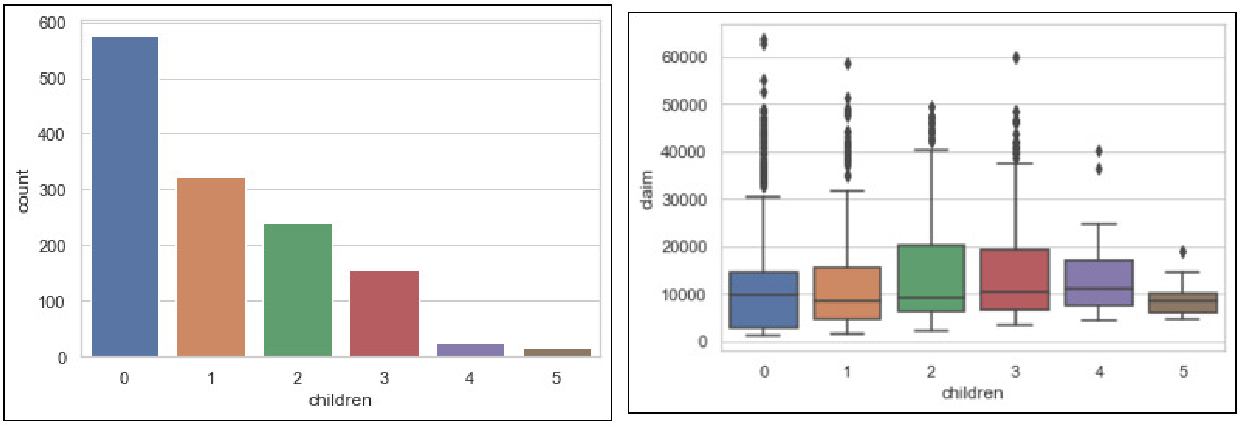


Fig. 8: Bar Graph of the Patient's Number of Children and Box Plot of Relationship between Claim and Patient's Number of Children

Fig. 9 provides the visualization for the third question: Are gender and smoking habits correlated?

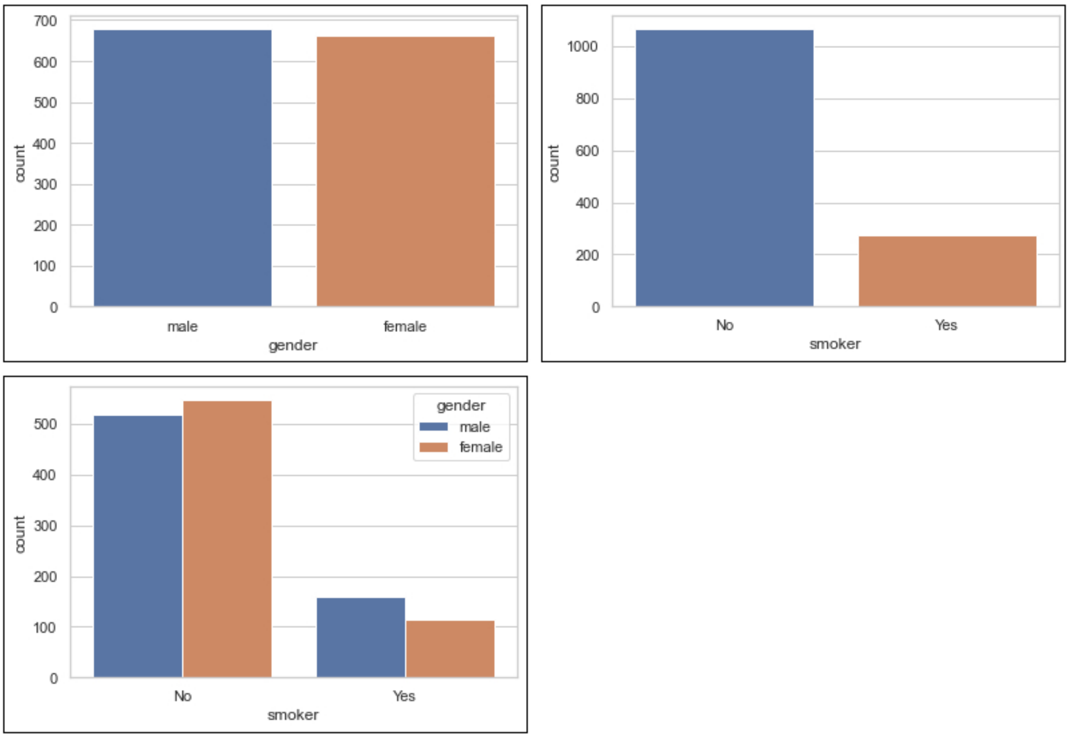


Fig. 9: Bar Graph of Patient's Gender, Smoking Preference, and Smoking Preference according to Gender

Fig. 10 provides the visualization for the fourth question: Are age and smoking habits correlated?

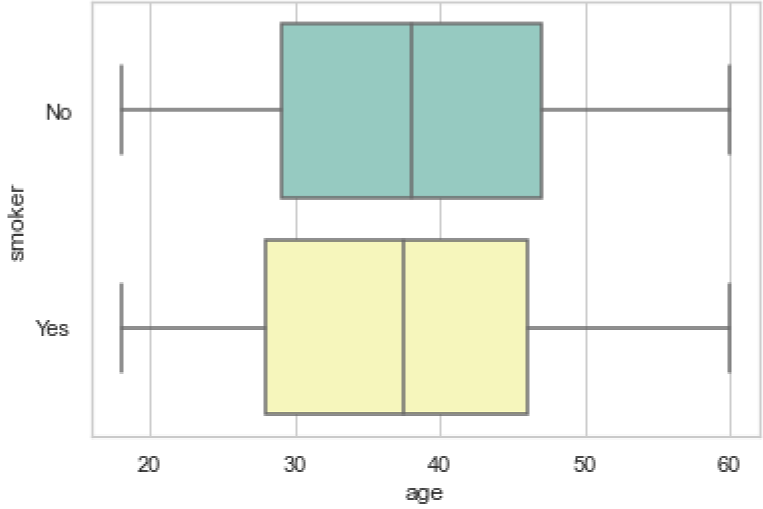


Fig. 10: Box Plot of Relationship between Smoking Preference and Age

Fig. 11 provides the visualization for the sixth question: Is the patient having higher blood pressure also diabetic?

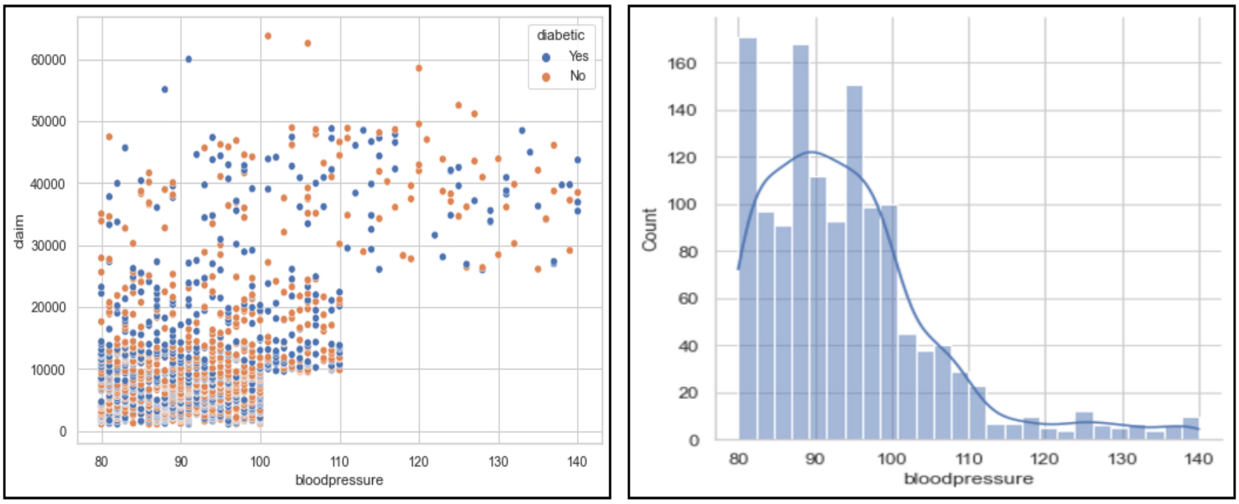


Fig. 11: Scatter Plot of Relationship between Claim and Blood Pressure of Diabetic and Non-Diabetic Patients, and Graph for Patient's Blood Pressure

Fig. 12 provides the visualization for the seventh question: If a patient who smokes has made higher claims?

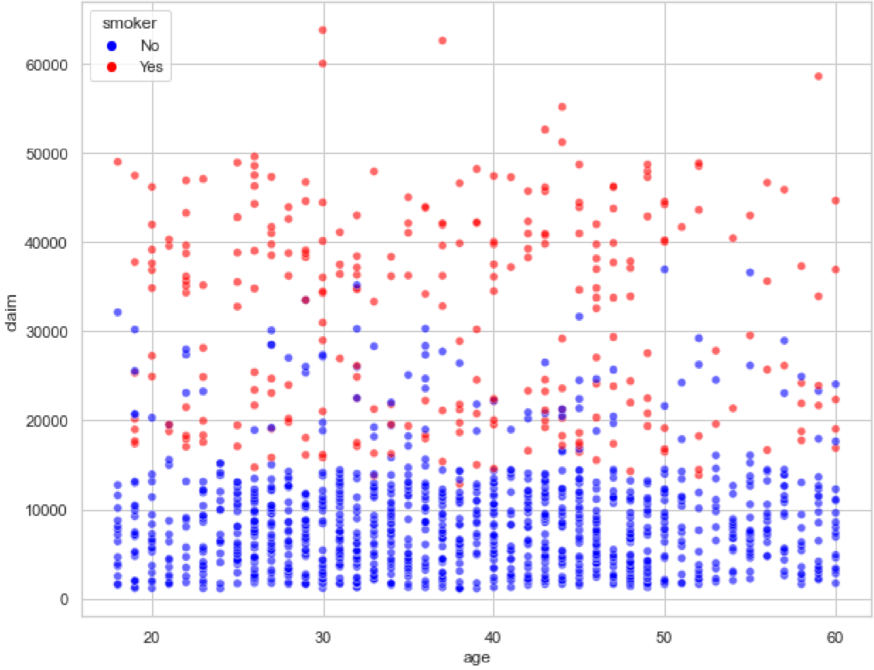


Fig. 12: Scatter Plot of Relationship between Claim and Age of Smokers and Non-Smokers

Fig. 13 provides the visualization for the eighth question: If a patient who is diabetic has made higher claims?

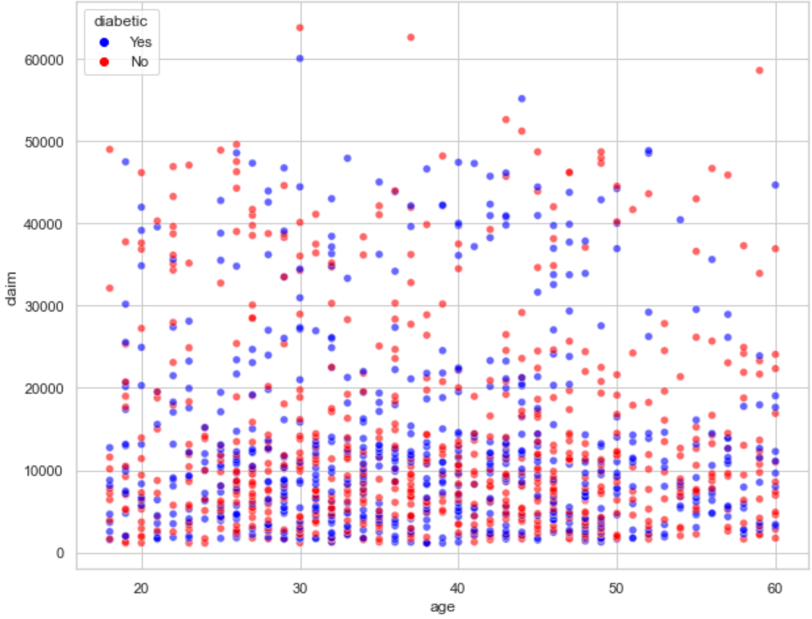


Fig. 13: Scatter Plot of Relationship between Claim and Age of Diabetic and Non-Diabetic Patients

Fig. 14 provides the visualization for the eighth question: If a patient who has high blood pressure has made higher claims?

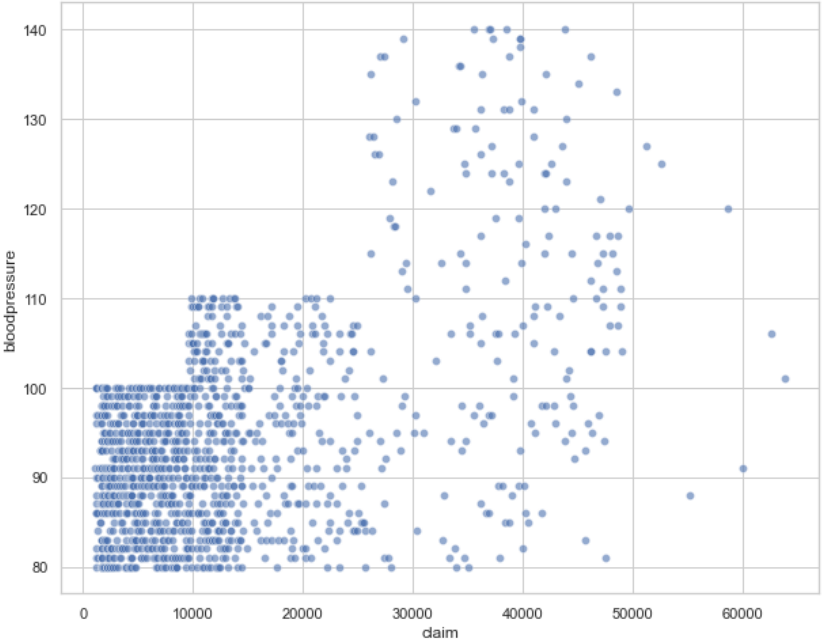


Fig. 14: Scatter Plot of Relationship between Blood Pressure and Claim

Discussion

From our findings and results, we can observe various things and then analyze them accordingly. From Fig. 6, we can see that there are more outliers in claims data in comparison to the other fields. This aligns with the understanding that most people require basic medication and treatments while only some require expensive medication and treatments in accordance with the severity of their illness. Thus, the various different claim amounts relative to each patient's personal needs give rise to a high number of outliers.

To answer our first question: "**Is there a relationship between Body mass index and gender?"**, we look at Fig. 7. Looking at the box plot only, it is not certain if gender has any impact on bmi, so we also take a look at the scatter plot which seems to show that gender has no impact on bmi, but it is still not very clear. So, we do a t-test to determine this and get the result: **0.08672083693208645 > 0.05**, from which we can accept the null hypothesis: **Gender has no impact on bmi**.

To answer our second question: **"Are people with more children claiming more?"**, we look at Fig. 8. Looking at the bar graph of the patient's number of children, we can see that many of them have no children and only a few of them have 4 or 5 children. From the box plot, we can also see that people with no children are claiming more than the people with more children. For instance, people with 5 children have made less claims than people with 0 children. This could be due to the fact that the data set contains more data on people with no or less children in comparison to people with 4 or 5 children, thus making it seem like people with no children claim more than the people with more children. Performing a Chi-square test to test the hypothesis gives the result: **P-value = 0.4291156531605151 and 0.4291156531605151 > 0.05**. Hence, we accept the null hypothesis: **Increase in the number of children doesn't mean increase in claims made by parents.**

To answer our third question: **"Are gender and smoking habits correlated?"**, we look at Fig. 9. Looking at the bar graph of the patient's gender, smoking preference, and smoking preference according to gender, we can see that there is no significant difference in the male to female ratio in the data and that there are more non-smokers in the sample data than smokers, with higher number of males being smokers than females. Performing a Chi-square test to check the proportion of smokers as per gender, we get the results: **P-value = 0.007121650243180834 and 0.007121650243180834 < 0.05**. Hence, we reject the null hypothesis and accept the hypothesis: **Gender has an effect on smoking habits**.

To answer our fourth question: **“Are age and smoking habits correlated?”**, we look at Fig. 10. Looking at the box plot of the relationship between smoking preference and age, we can see that smoking habits are not age specific since some of the people within the same age group are smoking whereas others are not. Performing a Chi-square test to check the proportion of smokers as per age, we get the results: **P-value = 0.3211436086661882 and 0.3211436086661882 > 0.05**. Hence, we accept the null hypothesis: **Smoking habits are not related to age**.

To answer our fifth question: **“Does the number of children have any effect on the BMI of females?”**, we perform an ANOVA test to see the relationship between 3 groups, female with zero children, female with 2 children, and female with 4 children. From the test, we get the results: **0.7148335064686346 and 0.7148335064686346 > 0.05**. Hence, we accept the null hypothesis: **Number of children has no effect on bmi of females**.

To answer our sixth question: **“Is the patient having higher blood pressure also diabetic?”**, we look at Fig. 11. Looking at the scatter plot of relationship between claim and blood pressure of diabetic and non-diabetic patients, and graph for patient’s blood pressure, we can see that most of the diabetic patients have blood pressure ranging from 80 to 100. It looks like most of the diabetic patients have low blood pressure and being diabetic and having high blood pressure do not seem to be related. Performing an independent t-test to test the hypothesis, we get the results: **0.6425962639741668 and 0.6425962639741668 > 0.05**. Hence, we accept the null hypothesis: **Having high blood pressure and being diabetic are not related**.

To answer our seventh question: **“If a patient who smokes has made higher claims?”**, we look at Fig. 12. Looking at the scatter plot of the relationship between claim and age of smokers and non-smokers, we can see that more claims have been made by people who smoke. Performing an independent t-test to test the hypothesis, we get the results: **2.914915316662231e-283 and 2.914915316662231e-283 < 0.05**. Hence, we reject the null hypothesis and accept the hypothesis: **Smoking habits have an impact on increase of claims**.

To answer our eighth question: **“If a patient who is diabetic has made higher claims?”**, we look at Fig. 13. Looking at the scatter plot of the relationship between claim and age of diabetic and non-diabetic patients, we cannot precisely conclude the relationship between diabetic patient and increase of claim. Performing an independent t-test to test the hypothesis, we get the results: **0.7496665799159535 and 0.7496665799159535 > 0.05**. Hence, we accept the null hypothesis: **Being diabetic has no impact on increase of claims**.

To answer our ninth question: **“If a patient who has high blood pressure has made higher claims?”**, we look at Fig. 14. Looking at the graph we can see that higher claims have been made by people with higher blood pressure. Performing Pearson's correlation coefficient test to test the hypothesis, we get the results: **stat=0.531, p=0.000, P-value= 1.6938057599059636e-98 and 1.6938057599059636e-98 < 0.05**. Hence, we reject the null hypothesis and accept the hypothesis: **High blood pressure is related to higher claims**.

With the data available to us, we have wrangled and then analyzed the data to come up with suitable hypotheses and visualizations. The results have been communicated along with the visualizations and the analysis.

Conclusion

Study and analysis are imperative to avoid risk and maximize profit. US healthcare being the big domain of research and the increase in the health sector data has provided a number of ways to understand the data behavior and implement the best measure before making any business decision. There is no doubt that the health insurance company will make the most of it with the help of study on patient health data on practice, characteristics and way of life using the right tool. As done on this project, now it is known that smoking habits can lead to increase in claims or charges so insurance company can enroll their insuree in wellness program to control the smoking habits. Similarly, they can segregate the risky patients and enroll them in certain exercise program to make sure they are doing everything to mitigate the risk. This will help the patients in preventing

Health Insurance companies are finding a way to understand the claiming behavior of the customer and are investing a lot of money in research. However, more research and understanding are still needed on big data sets to maintain the proper results.

References

- https://www.mckinsey.com/industries/financial-services/our-insights/how-data-and-analytics-are-redefining-excellence-in-p-and-c-underwriting
- https://www.mastersindatascience.org/industry/insurance
- http://www.mash.dept.shef.ac.uk/Resources/MASH-WhatStatisticalTestHandout.pdf
- https://data.world/sumitrock/insurance/workspace/file?filename=insurance_data.csv
- https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/
- https://online.stat.psu.edu/stat501/lesson/1/1.9
- https://medium.datadriveninvestor.com/p-value-t-test-chi-square-test-anova-when-to-use-which-strategy-32907734aa0e
- https://openpublichealthjournal.com/contents/volumes/V2/TOPHJ-2-11/TOPHJ-2-11.pdf