Project 1

BY: EDGAR ALVARADO, BRAD ANDERSON,
RYAN NUTTALL, JAZMIN HERNANDEZ

Project Proposal

Project Description/Outline:

We are exploring healthcare trends using multiple datasets to view pricing trends and how they vary with different demographics such as age, gender, and insurance provider. We are going to create various charts and graphs to display the different trends we are analyzing. Datasets are being pulled from Kaggle to help us understand these trends

Research Questions to Answer:

1: What trends can we analyze based on Gender?

2: How does insurance provider and geography affect pricing rates?

3: What trends can we analyze based on age?

4: How does smoking affect overall billing?

Datasets to be Analyzed

Healthcare Dataset

- Consists of 10,000 rows, each representing a patient healthcare record
- Contains patient data, admission information, and healthcare services provided
- Information included: Age, Gender, Medical Condition, Admission Dates, Medication, Insurance Provider, and Billing Amount
- **Synthetic Dataset**
- **https://www.kaggle.com/datasets/prasad22/healthcare-dataset**

Insurance Dataset

- Consists of over 1300 rows, each representing an insurance claim
- Contains information on the relationship between personal attributes and medical insurance charges
- Information included: Age, Sex, BMI, Children, Smoker, Region, and Charges
- **Synthetic Dataset**
- **https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance**

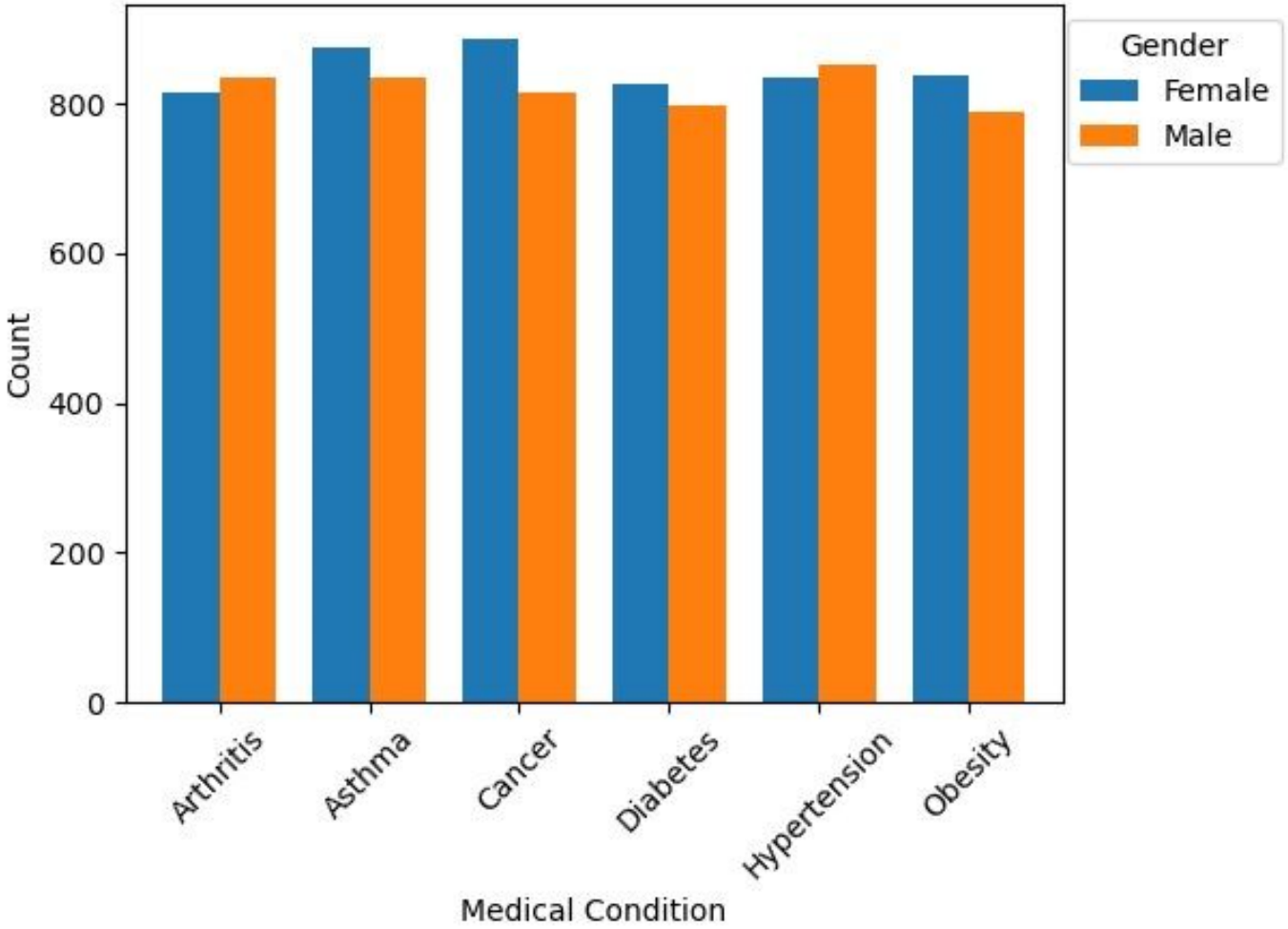
Question 1

What trends can we analyze based on gender?

Background: Some studies have suggested that there might be gender-based differences in healthcare utilization, with women often using more healthcare services than men. This could potentially lead to differences in billing amounts.

Factors contributing to such differences might include variations in health conditions, types of healthcare services utilized, and reproductive health needs. Our healthcare dataset listed only chronic conditions for both genders (Figure 8). Additionally, several studies have also explored the impact of gender on the quality of healthcare services and associated costs. With this in mind we wanted to see if we could find relationships between gender and billing amounts from our datasets.

Figure 8: Medical Condition per Gender



Analysis Question 1.A

How does gender affect pricing (healthcare dataset)?

To get an overview of higher billing amounts across genders, our analysis started with an overall population comparison of mean billing amounts across genders from our healthcare dataset. From this we found the following information on mean billing amounts:

female $25,484

male $25,550

Analysis Question 1.A cont.

How does gender affect pricing (insurance dataset)?

Since the healthcare dataset does not show a significant difference in billing amount (synthetic dataset) by gender, we looked to our insurance charges by geography dataset to see if there was significant variance in mean billing amounts by gender and found the following information and (Figure 13) graph:

female 12569.578844

male 13956.751178

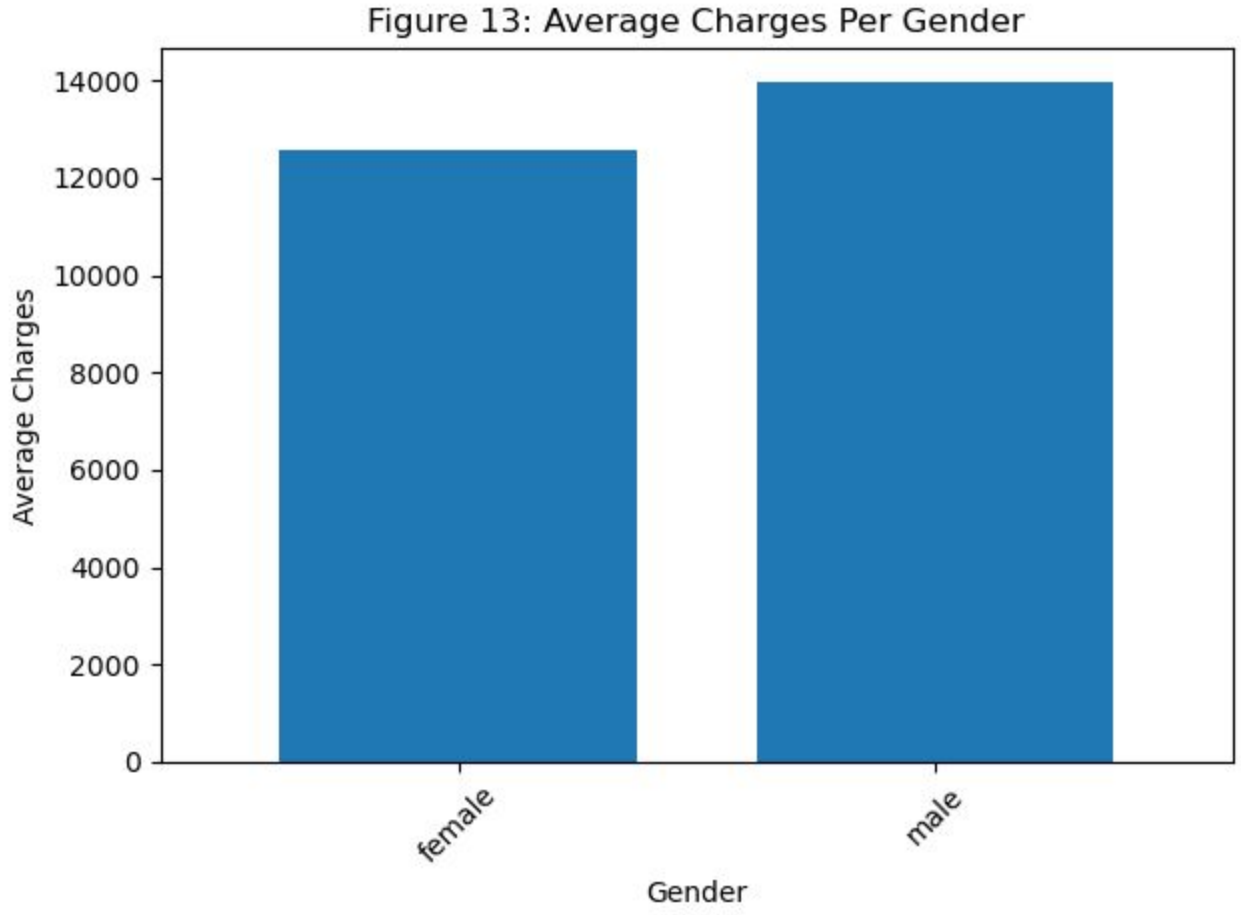


Figure 13: Average Charges Per Gender
14000
12000
10000
8000
6000
4000
2000
0
Average Charges
female
male
Gender

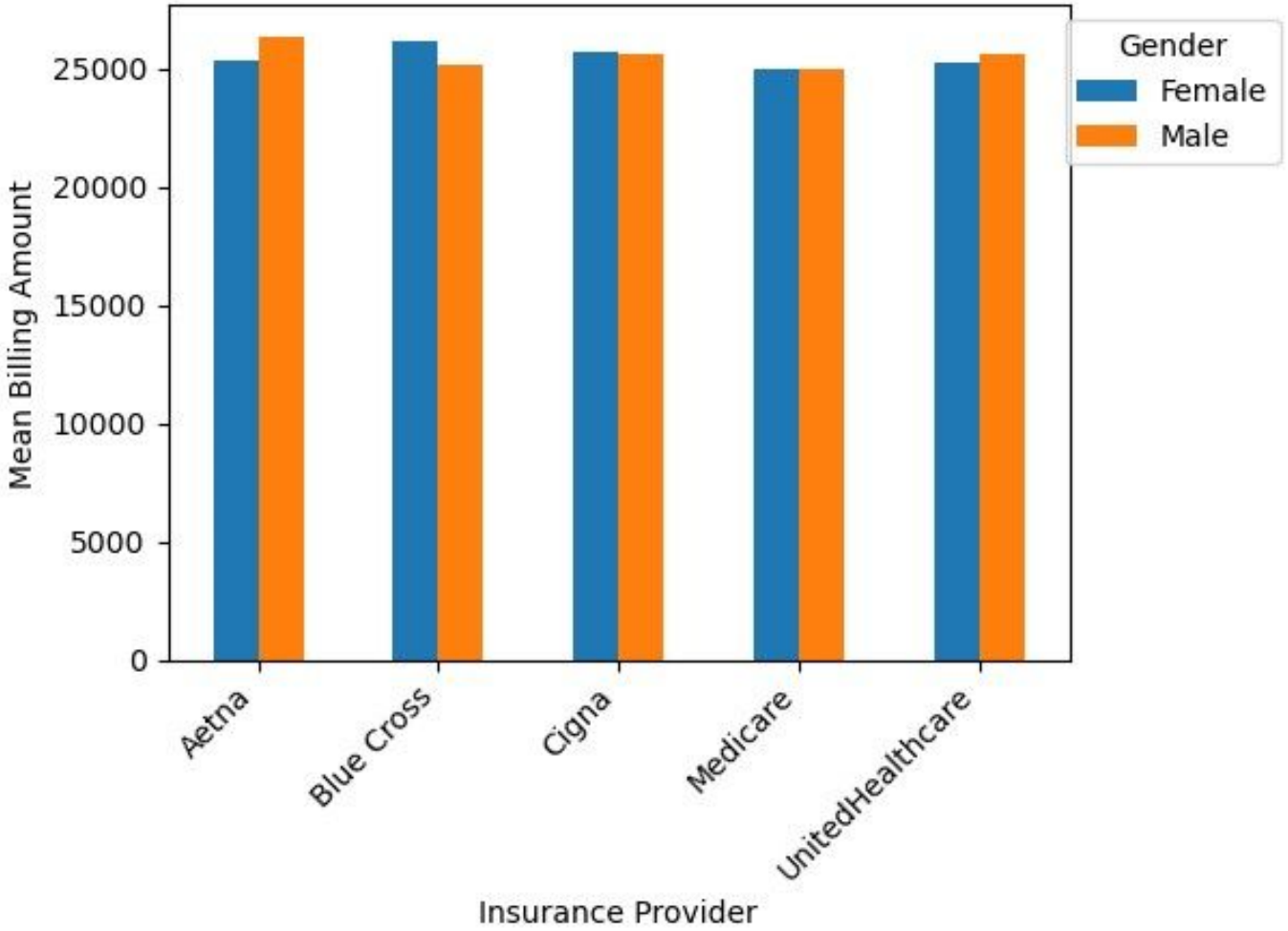
Analysis Question 1.B

How does insurance provider affect pricing by gender?

Since there was no significant difference in the mean of Billing Amount by Gender from the healthcare dataset, we wanted to delve further to see if any variable cost driver would be present from the Insurance Provider that someone would have. From the healthcare dataset we generated this data and (Figure 1) graph:

Insurance Provider	Gender	Billing Amount
Aetna	Female	$25,335
	Male	$26,341
Blue Cross	Female	$26,178
	Male	$25,158
Cigna	Female	$25,724
	Male	$25,588
Medicare	Female	$25,008
	Male	$24,996
UnitedHealthcare	Female	$25,200
	Male	$25,617

Figure 1: Mean Billing Amount by Insurance Provider and Gender

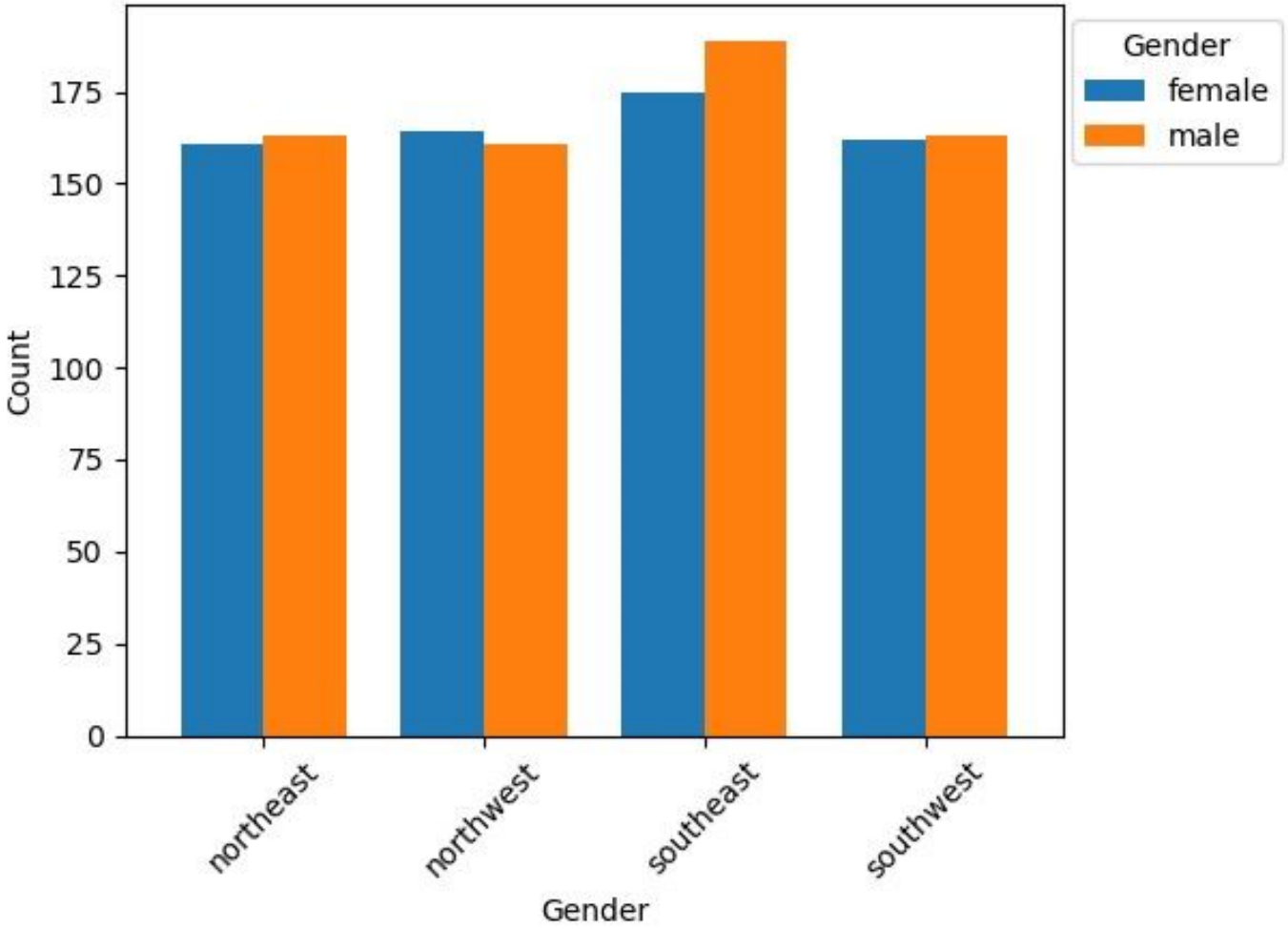


Analysis Question 1.C

How does region and gender count affect pricing?

We plotted the gender count distribution for each region (Figure 14), with the southeast region exhibiting the highest average cost. As depicted in the graph, the number of males in the southeast region surpasses that of any other region, potentially influencing the region's elevated average cost - this could be one of the reasons for an elevated increasing in male billing charges.

Figure 14: Gender Counts Per Region



Question 2

How does insurance provider and geography affect pricing rates ?

Background: Healthcare providers often negotiate contracts with insurance companies to determine the reimbursement rates for specific medical services. These negotiated rates can vary between different insurance providers and healthcare facilities. We also wanted to analyze data to see if we could find relationships between insurance provider and geography since we were working with a couple of distinct datasets. Lastly, we

Analysis Question 2.A

How does insurance provider affect pricing rates?

We first started by grouping Medical Condition and Insurance Provider and calculating the mean, minimum, and maximum of the billing amount. (head(5) info below).

	Medical Condition	Insurance Provider	Mean Billing	Min Billing	Max Billing
0	Arthritis	Aetna	24694.858132	1009.417327	49745.81198
1	Arthritis	Blue Cross	25989.515740	1271.433037	49838.14629
2	Arthritis	Cigna	25189.873894	1256.479072	49936.07375
3	Arthritis	Medicare	24206.381013	1042.981212	49985.97307
4	Arthritis	UnitedHealthcare	25704.428629	1308.848408	49573.39990

Analysis Question 2.A cont.

How does insurance provider affect pricing rates?

Next we wanted to find each maximum Billing Amount by Insurance Provider and Medical Condition. From this we gathered the following data. It would appear that Aetna has the highest billing amounts across various medical conditions but bear in mind this is a synthetic dataset.

	Medical Condition	Insurance Provider	Mean Billing	Min Billing	Max Billing
3	Arthritis	Medicare	24206.381013	1042.981212	49985.97307
5	Asthma	Aetna	24761.515983	1032.263087	49974.29914
14	Cancer	UnitedHealthcare	25222.992960	1020.337790	49994.98474
15	Diabetes	Aetna	26703.077342	1282.493591	49954.96833
20	Hypertension	Aetna	26420.830930	1376.234851	49995.90228
28	Obesity	Medicare	25259.825969	1000.180837	49974.16046

Analysis Question 2.B

How does geography affect pricing?

Earlier in question one we explored how gender affected billing amount (and Figure 13). We could see that there is quite a large variance in the billing amount - and according to our analysis of the insurance dataset, male was higher than female - so this prompted us to analyze regional differences to see if this could explain the variance in billing rates. Graphing our insurance dataset found (Figure 10) that highest billing was found in the Southeast region.

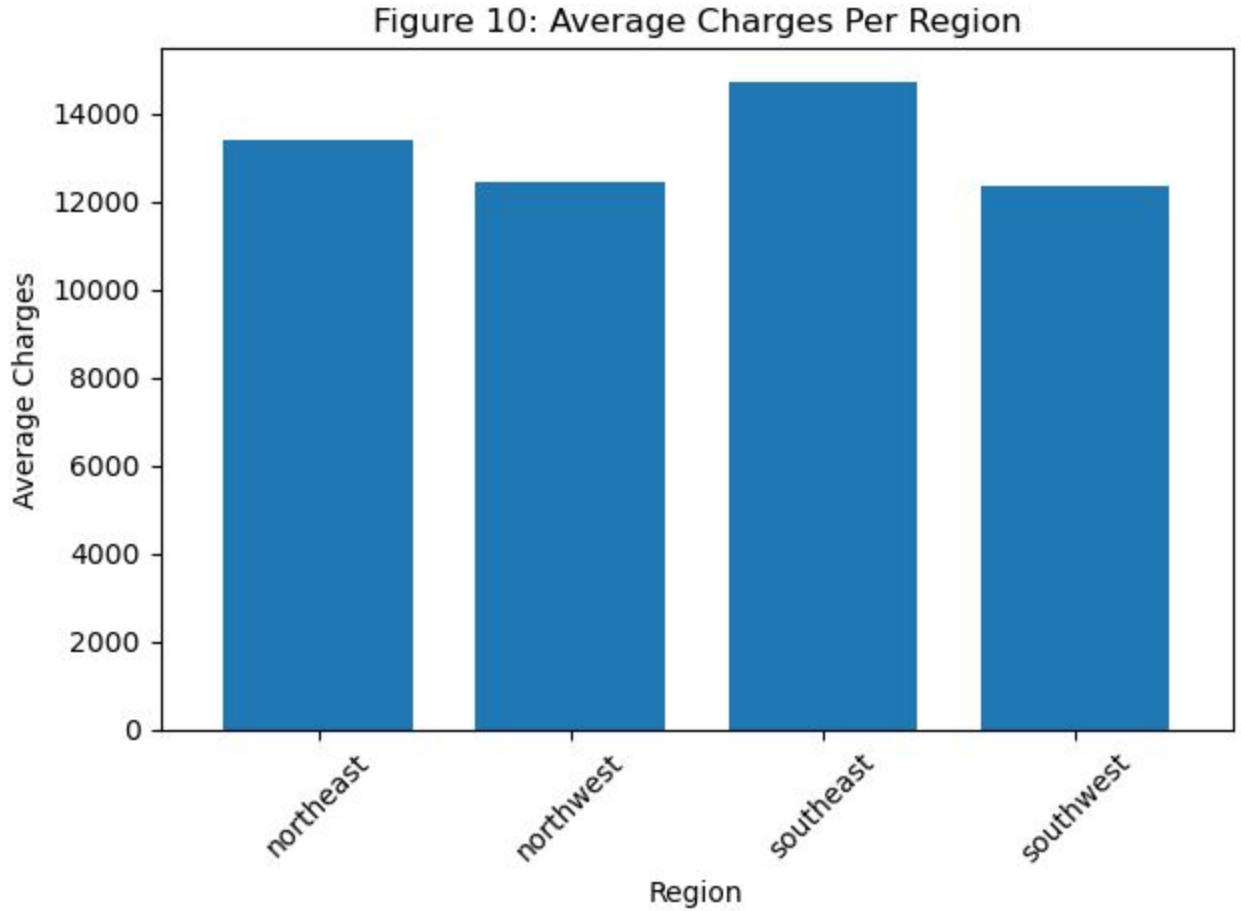


Figure 10: Average Charges Per Region
Average Charges
14000
12000
10000
8000
6000
4000
2000
0
northeast
northwest
southeast
southwest
Region

Analysis Question 2.C

What else could contribute to the southeast region having higher rates?

Looking to our insurance dataset and geography, we sought to check if BMI (body mass index) would be a factor in billing charges by location. BMI is commonly used to assess whether a person is underweight, normal weight, overweight, or obese. However, it is important to note that BMI does not directly measure body fat or health. From this we found again that the southeast region looks to have a higher BMI (Figure 9).

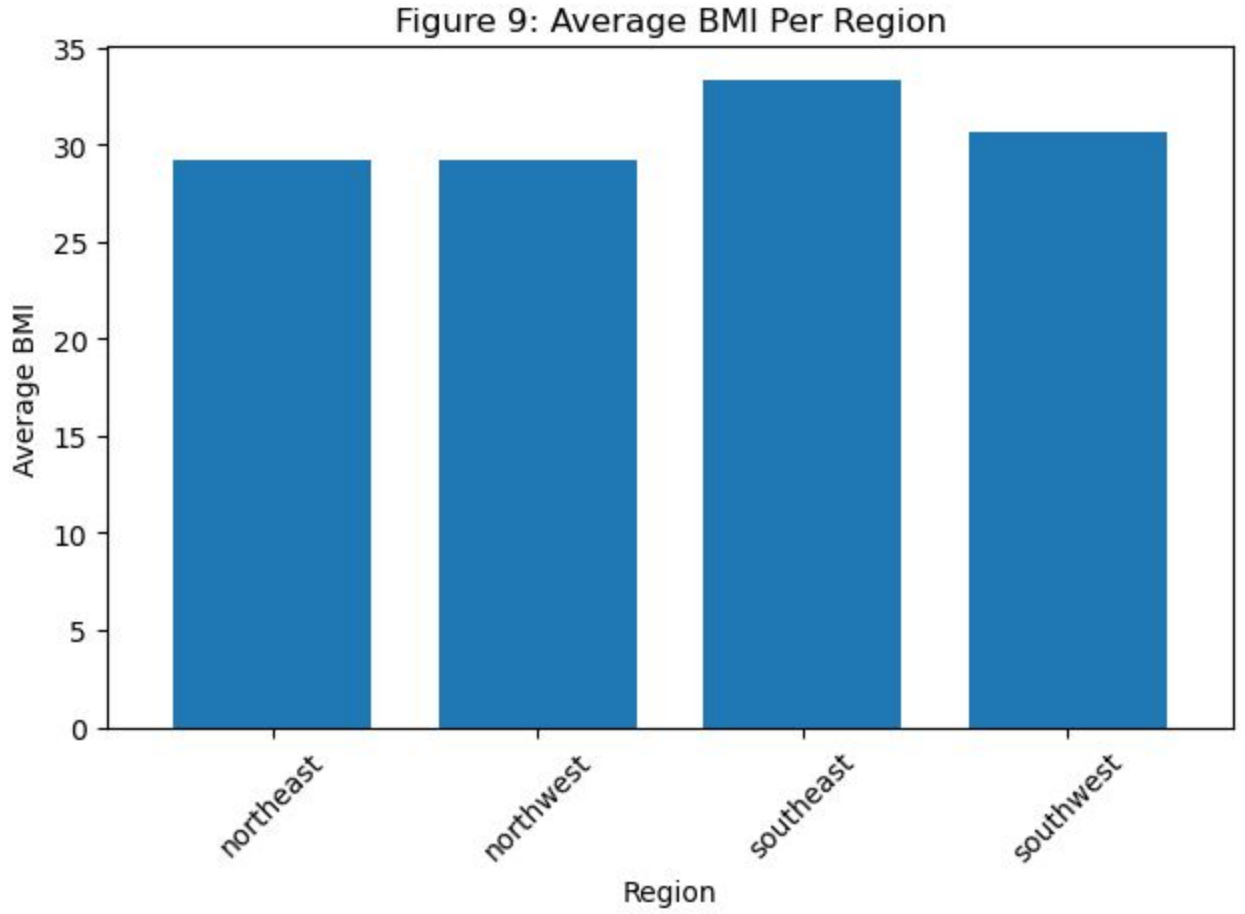


Figure 9: Average BMI Per Region
35
30
25
20
15
10
5
0
Average BMI
northeast
northwest
southeast
southwest
Region

Question 3

What trends can we analyze based on age?

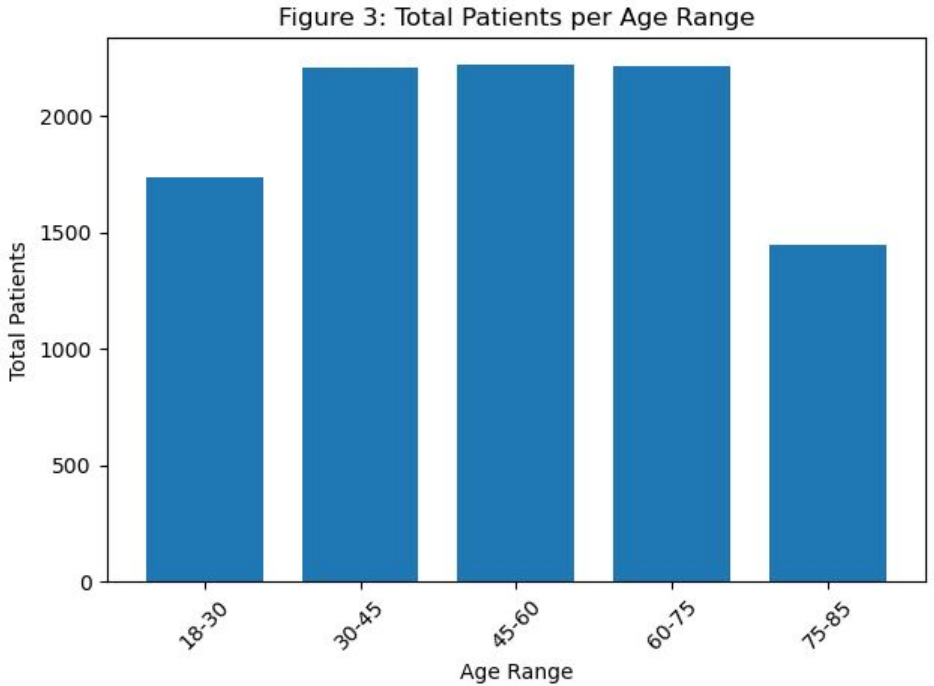
Background: In the healthcare world a patient's age can be used to make a large number of predictions. Knowing a patient's age can assist healthcare professionals in a number of ways including medical diagnosis, provided treatment, and estimated recovery time. We wanted to analyze the data based on a patient's age to see if we could uncover trends to help inform decisions.

Analysis Question 3 Summary Stats

- Created bins of age groups so we could compare the results for each age range
- Converted the Admission Date and Discharge Date columns to datetime so that we could calculate the duration of a patient's stay in the hospital
- Started the analysis by calculating some summary statistics on the age column to get an idea of what we were working with
- Calculated the 68-95-99.7 rule using standard deviation
    - Roughly 68% of the data is between 32.0 and 71.0
    - Roughly 95% of the data is between 12.0 and 91.0
    - Roughly 99.7% of the data is between -7.0 and 110.0

Summary Statistics

	Age
count	10000.000000
mean	51.452200
std	19.588974
min	18.000000
25%	35.000000
50%	52.000000
75%	68.000000
max	85.000000



Analysis Question 3.A

How does age affect a patient's length of stay in the hospital?

As a person gets older the time it takes for patients to recover usually increases. We were curious what this might look like within our dataset. For this analysis we calculated the average length of stay at the hospital across each of the age ranges and plotted the results on a bar graph.

We found that the average length of stay was relatively the same per age range. This could indicate that age does not play a large factor in length of stay at the hospital. Again this dataset is synthetic. We were expecting the average length of stay would increase as the age increased.

Age Range	Average Stay (days)
18-30	15.503163
30-45	15.410222
45-60	15.621573
60-75	15.563883
75-85	15.744813

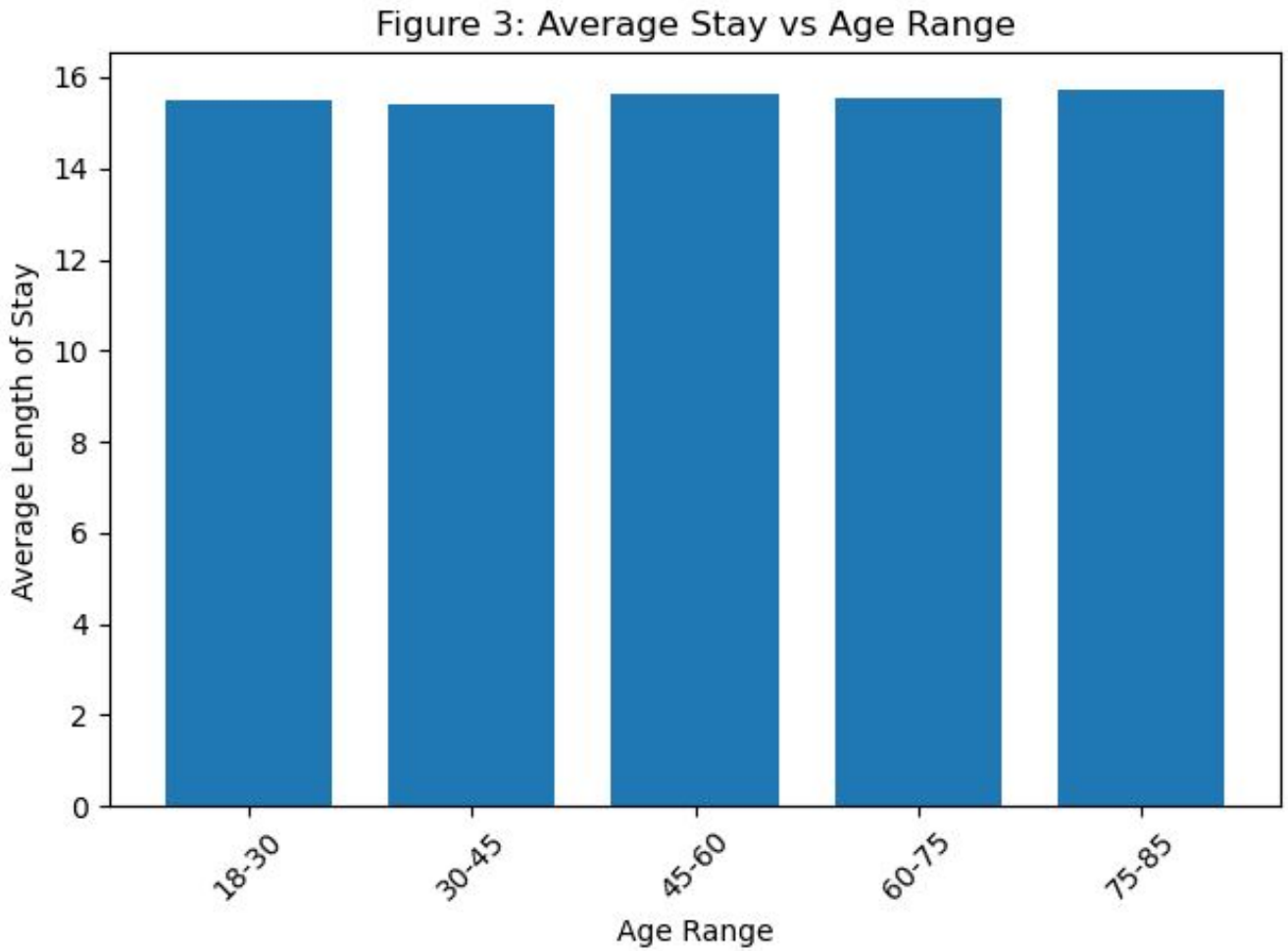


Figure 3: Average Stay vs Age Range
Average Length of Stay
0
2
4
6
8
10
12
14
16
18-30
30-45
45-60
60-75
75-85
Age Range

Analysis Question 3.B

How does age impact medical condition?

We were interested in discovering if patients were more or less susceptible to any medical conditions as they got older. For this analysis we calculated the total count of each medical condition recorded across the difference age ranges and plotted the results in a bar graph.

Age Range	Max Count	Min Count
18-30	Hypertension (311)	Arthritis (272)
30-45	Hypertension (389)	Diabetes (346)
45-60	**Cancer (392)**	Hypertension (340)
60-75	Asthma (384)	**Cancer (353)**
75-85	**Cancer (261)**	Arthritis (229)

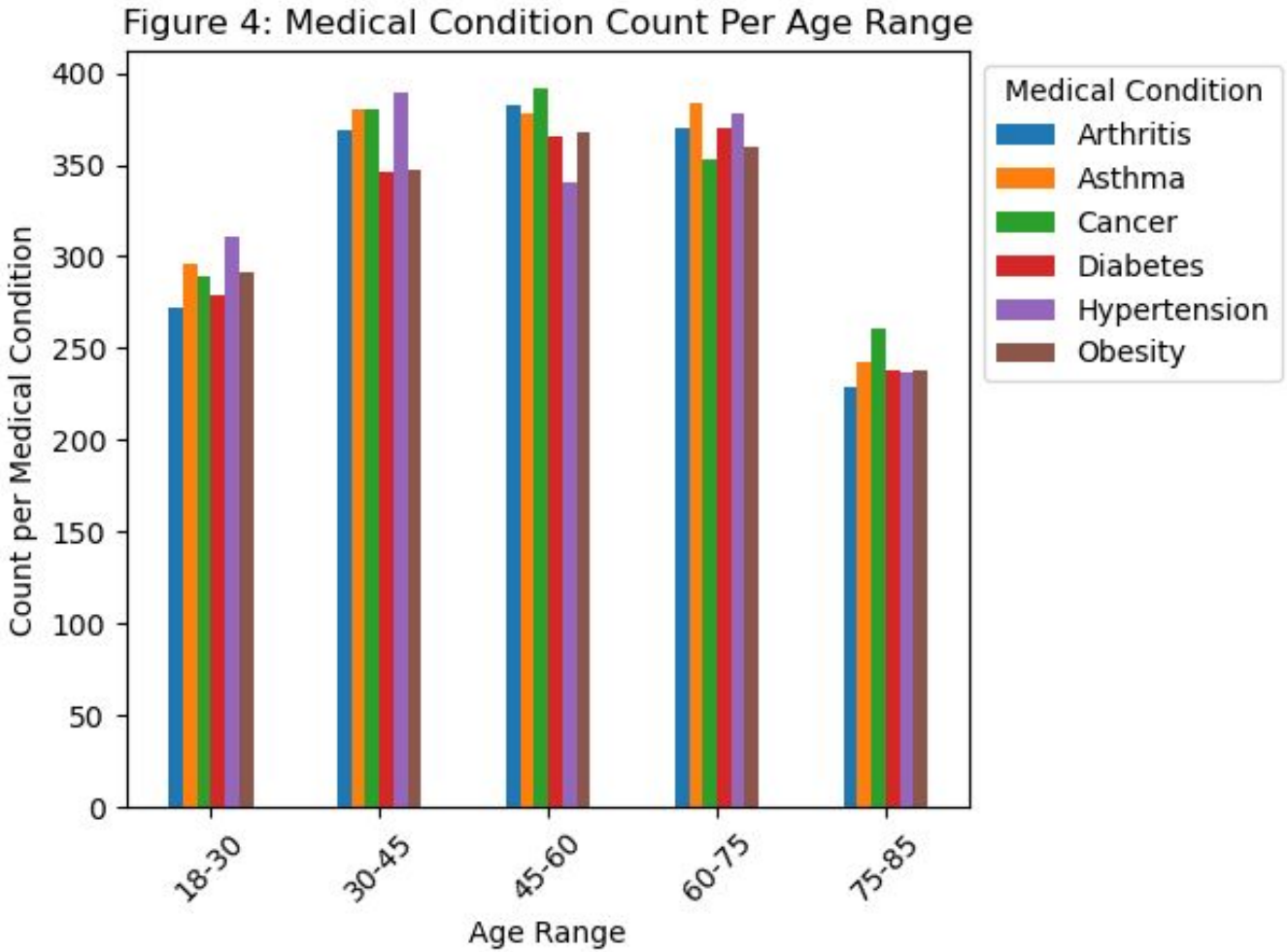


Figure 4: Medical Condition Count Per Age Range
Count per Medical Condition
0
50
100
150
200
250
300
350
400
18-30
30-45
45-60
60-75
75-85
Age Range
Medical Condition
Arthritis
Asthma
Cancer
Diabetes
Hypertension
Obesity

Analysis Question 3.C

How does age impact prescribed medication?

We thought it might be interesting to see what medications were most and least commonly used across the age groups. Age can play a factor in patients' response to medication. It's important to know what medication is dangerous to patients at certain ages. For this analysis we calculated the total count of each medication used across the difference age ranges and plotted the results in a bar graph

Age Range	Max Count	Min Count
18-30	Ibuprofein (368)	Paracetamol (329)
30-45	**Pencillin (485)**	Ibuprofein (420)
45-60	**Penicillin (467)**	Aspirin (417)
60-75	Lipitor (470)	Paracetamol (408)
75-85	**Penicilin (307)**	Lipitor (276)

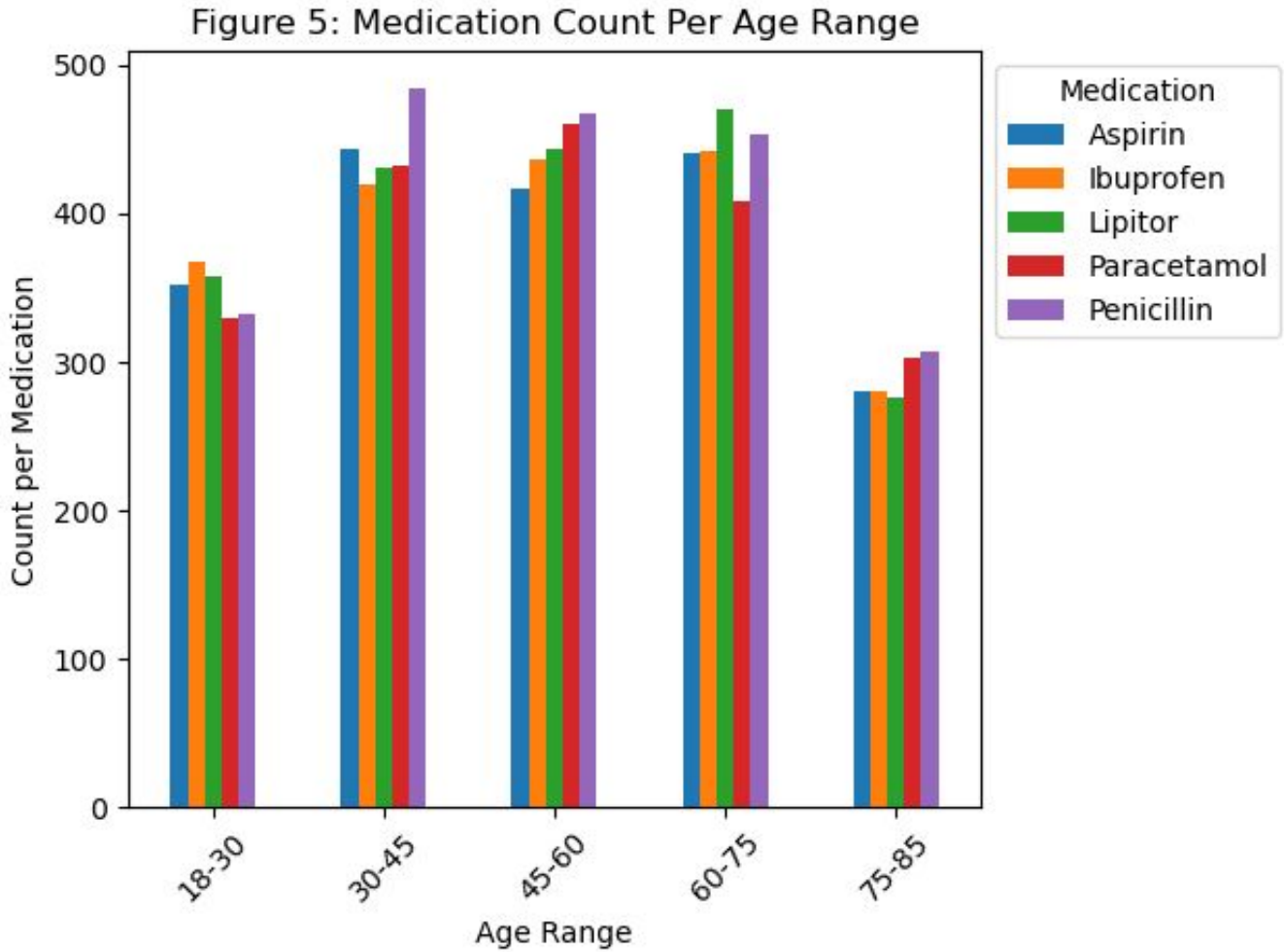


Figure 5: Medication Count Per Age Range
Count per Medication
500
400
300
200
100
0
18-30
30-45
45-60
60-75
75-85
Age Range
Medication
Aspirin
Ibuprofen
Lipitor
Paracetamol
Penicillin

Analysis Question 3.D

How does age affect pricing?

We used the Insurance Dataset for this analysis as we wanted to see the effect age had on medical insurance cost rather than what the hospital billed the insurance. We plotted the age and amount charged on a scatter plot and calculated the linear regression for the relationship between age and cost. We plotted the linear equation on the scatter plot and calculated the r-value to determine the correlation.

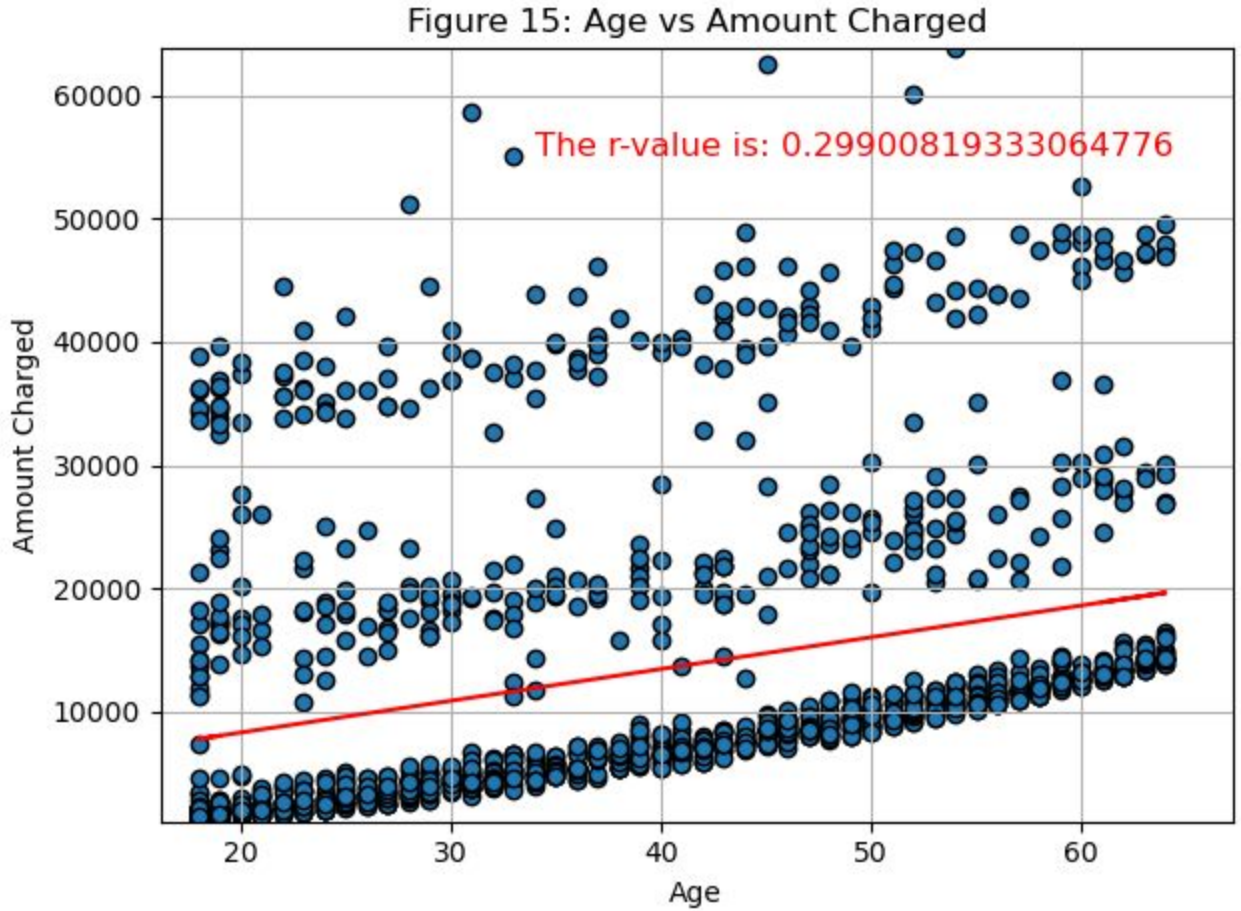


Figure 15: Age vs Amount Charged
The r-value is: 0.29900819333064776
Amount Charged
60000
50000
40000
30000
20000
10000
20
30
40
50
60
Age

Question 4

How does smoking affect billing?

We wanted to analyze if smoking status would affect the billing rates to patients from their insurance provider and then validate whether the difference in billing rates had any statistical significance.

Hypothesis: If there is an affect on billing to the patient from the insurance provider based on smoking status, then we will see an increase in billing rates to smokers vs non-smokers for similar claims by their insurance provider.

Null Hypothesis:

A patient's smoking status does not affect billing rates by insurance provider.

Alternative Hypothesis:

A patient’s smoking status increases billing rates by insurance provider.

Analysis Question 4.A

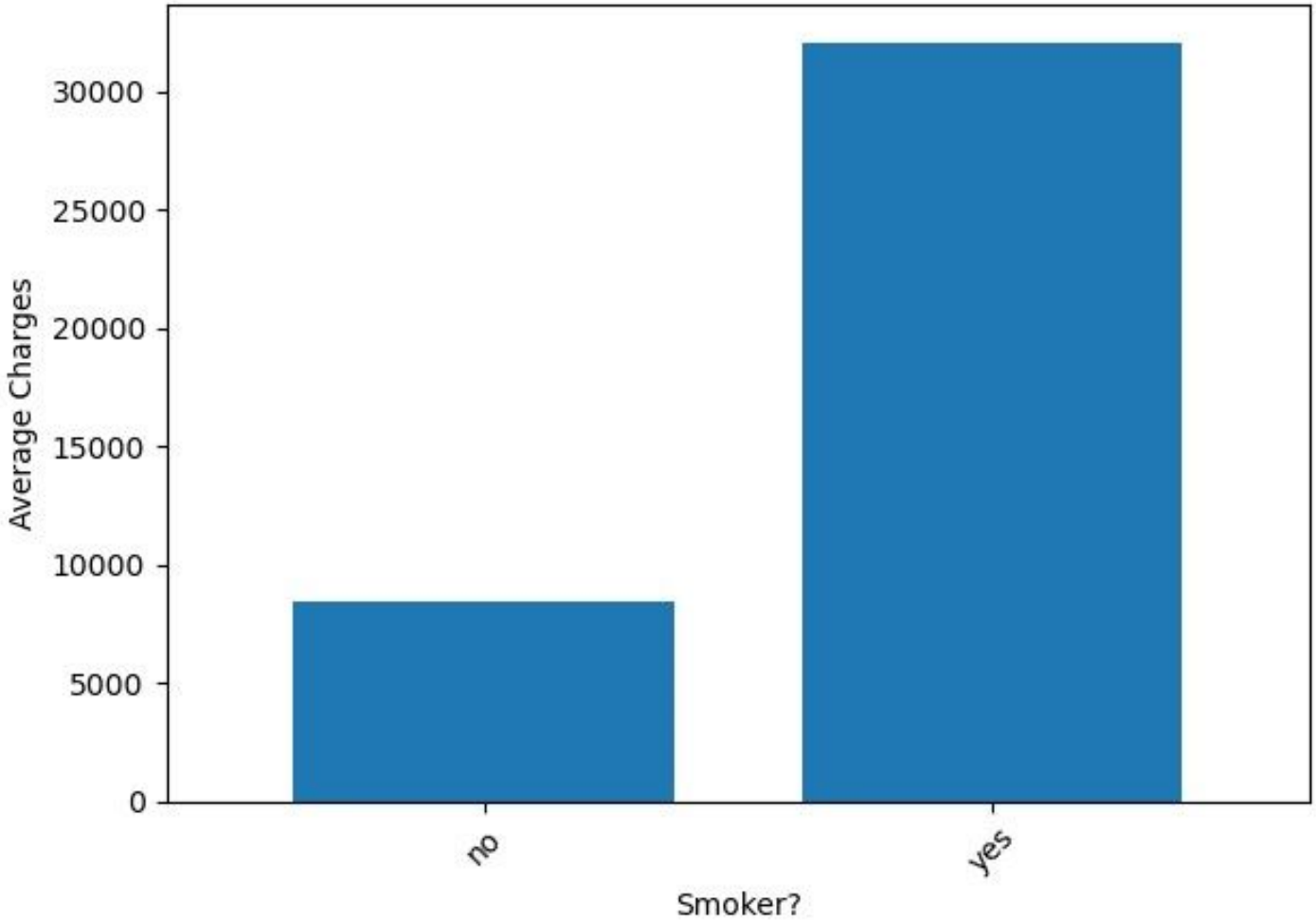
Using a t-test we analyzed the significance in difference in billings rates from insurance provider to patients who smoke and those who are non-smokers.

- We found that non-smokers on average were billed **$8,434 USD** per insurance claim.
- We found that smokers on average were billed **$32,050 USD** per insurance claim.

The T-Test resulted in a **p value=5.88946444671698e-103**

- We concluded that the T-Test results show enough significance to reject the Null Hypothesis and accept the Alternative Hypothesis

Figure 11: Average Charges When Smoking

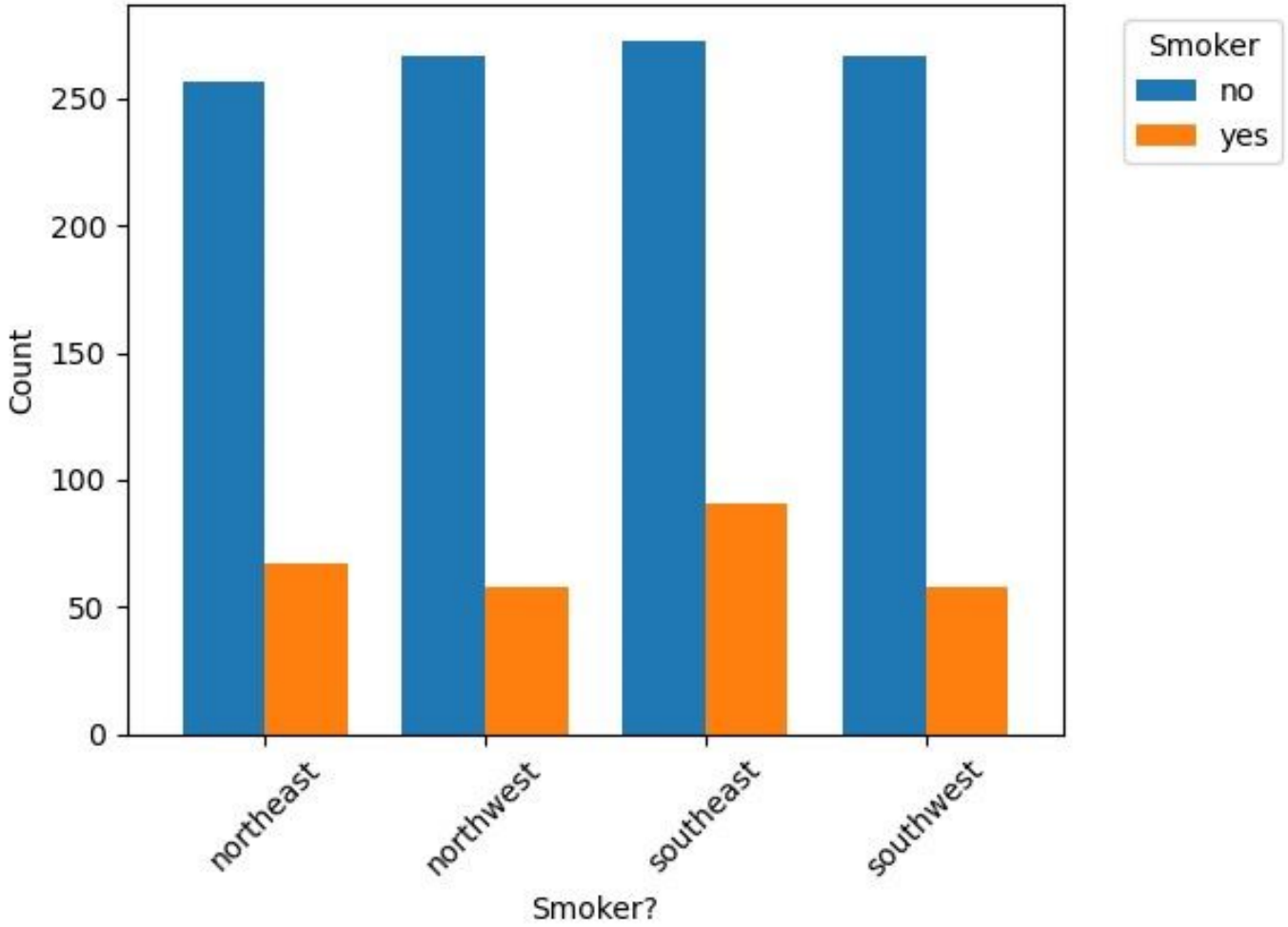


ANALYSIS Question 4.B

We wanted to further analyze billing rate trends for smokers and non-smokers by their insurance providers, and specifically if there is any variance by regions.

In earlier questions, (1.B and 2.C) we see an increase in billing rates for the southeast region of the united states. While in earlier questions we see an increase in the male population (1.B) as well as an increase in average patient BMI (2.C) as possible contributing factors for increased billing rates, in figure 14 we also see an increase in the overall smoking population in the southeast region as well. This increase in smokers is a variable that may contribute to the higher patient billing rates we see in this region compared to the rest of the united states.

Figure 12: Smoking Status Per Region



Conclusions

- Synthetic datasets made it difficult to find any meaningful insights in our analysis.
- Key takeaways:
  - There is a clear trend for increased billing rates among all ages and genders in the southeast region of the United States (insurance dataset).
    - Increased male population in the southeast may be a contributing factor to higher billing rates.
    - Higher average BMI in the population of the southeast region is another possible contributor to the higher billing rate trend seen throughout our analysis.
  - Most of the patients (68%) in our dataset were between the ages of 32-71.
    - This could indicate people younger than 30 visit hospitals less frequently. (good health)
    - There are most likely less people over 75 due to lower population rates. (mortality)
  - Most significant finding was the effect that smoking had on medical insurance charges.
    - Non-Smoker average billing rate: **$8,434.00**
    - Smoker average billing rate: **$32,050.00**