EDA and Regression Model for US Medical Charges

Jovita Amelinda Kurniawan

Pacmann Project - Statistics for Business

2023

Table of contents

01

Introduction

02

Dataset

03

Analysis

04

Conclusion

01

Introduction

Background

- With increasing age and healthcare cost, **health insurance is an important consideration** when it comes to financial planning.
- **Insured individuals pay premium** to insurance providers in exchange of medical cost coverage.
- Premium sum is determined by insurance providers by assessing risk of insured individuals using various factors.
- However**, risk assessment** and **premium pricing remains a challenge** to insurance providers as there are **a lot of factors that may affect individual risk profiles.**

Objective

- In this project, we are given a dataset of US population medical costs to analyze relationships and trends between variables (like age, dependents, smoking habits, etc) and medical costs.

- By **understanding the influence of these factors on medical costs**, insurance providers can make **better assessment on risk and premium price for their future clients.**

02

Dataset

Dataset

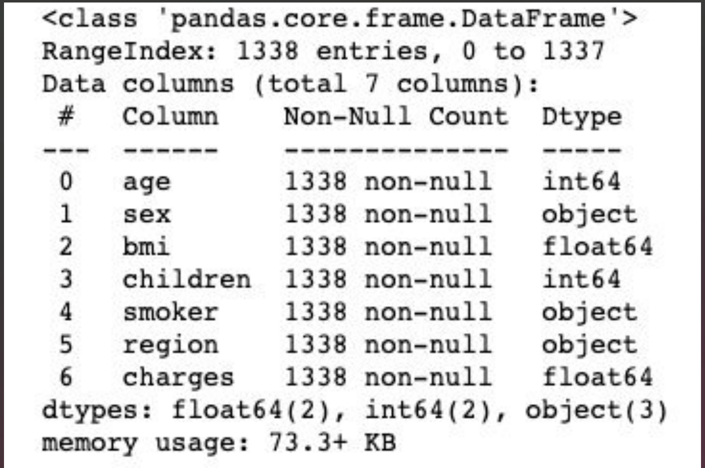
Dataset: Medical Charges of Sample US Population

Total data points: 1338

Null values: 0

Total columns: 7

- Age
- Sex
- BMI
- Children
- Smoker
- Region
- Charges



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Column Description

Column	Definition	Range / Values
Age	Age of primary beneficiary	18 - 64 years old
Sex	Insurance contractor gender	Female, male
BMI	Body mass index, providing an understanding of body weights that are relatively high or low relative to height, objective index of body weight (kg/m2) using the ratio of height to weight. Ideally, a normal BMI is in the range of 18.5 - 24.9	15.96 - 53.13

Column Description

Column	Definition	Range / Values
Children	Number of children covered by health insurance/number of dependents	0 - 5
Smoker	Indicates whether primary beneficiary is a smoker or not	Yes, no
Region	Beneficiary residential area in the US, northeast, southeast, southwest, northwest	Southwest, Southeast, Northwest, Northeast
Charges	Individual medical costs billed by health insurance	USD 1121.87 - 63770.43

03

Analysis

3.1 Descriptive Statistical Analysis

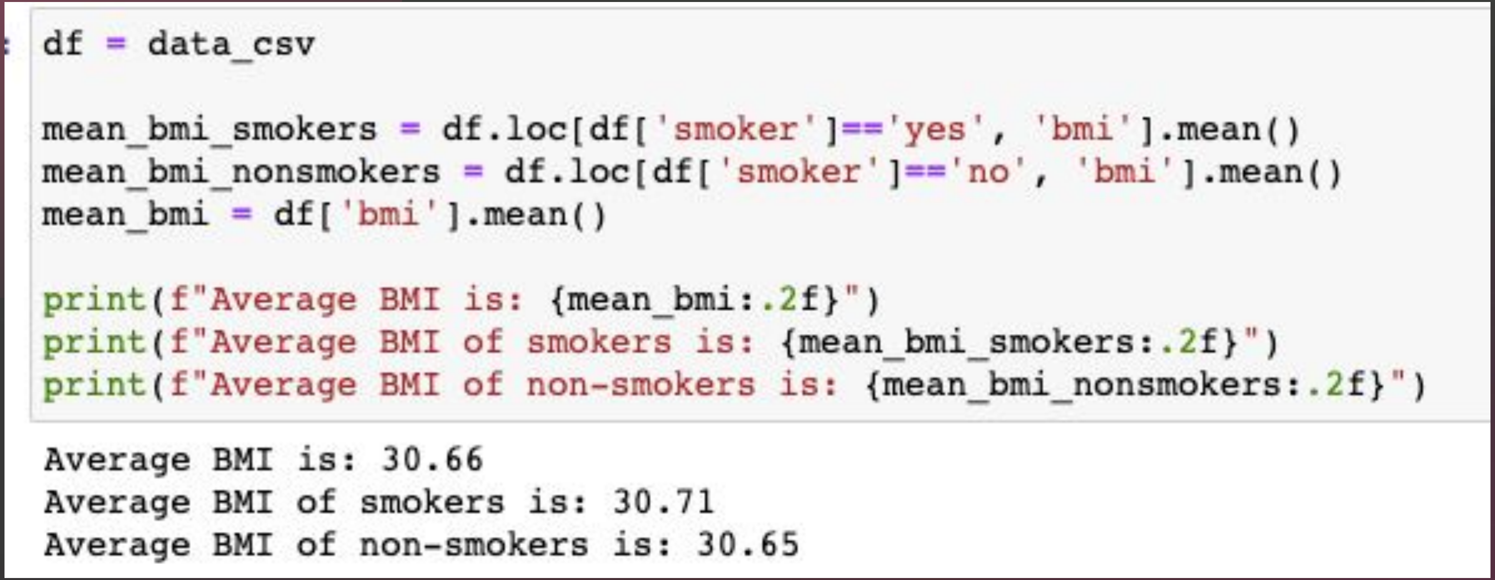
Question selected:

1. What is the average BMI of US population overall and by smoking status
2. Is the variance of charges for smokers equal to non-smokers?
3. Average charges for smokers and non-smokers
4. Is the average charges for smokers higher than non-smokers?
5. Is the average charges of smokers with BMI >25 greater than non-smokers with BMI > 25?

3.1 Descriptive Statistical Analysis

Question 1	**Find the average BMI of smokers**
Method	1. Filter data by smoking status: smoker == ‘yes’ , ‘no’ 2. Use .mean() function to calculate average BMI of smokers, non-smokers and overall
Results	● Average BMI is: 30.66 ● Average BMI of smokers is: 30.71 ● Average BMI of non-smokers is: 30.65
Findings	● According to CDC, those with BMI > 30.0 are considered obese. ● Average BMI of US individuals is 30.66. With those who smoke having slightly higher average of 30.71 and those who don’t slightly lower average of 30.65. ● Hence, US individuals in this sample are likely to be obese even more so for those who smokes.

3.1 Descriptive Statistical Analysis - Method Screenshot



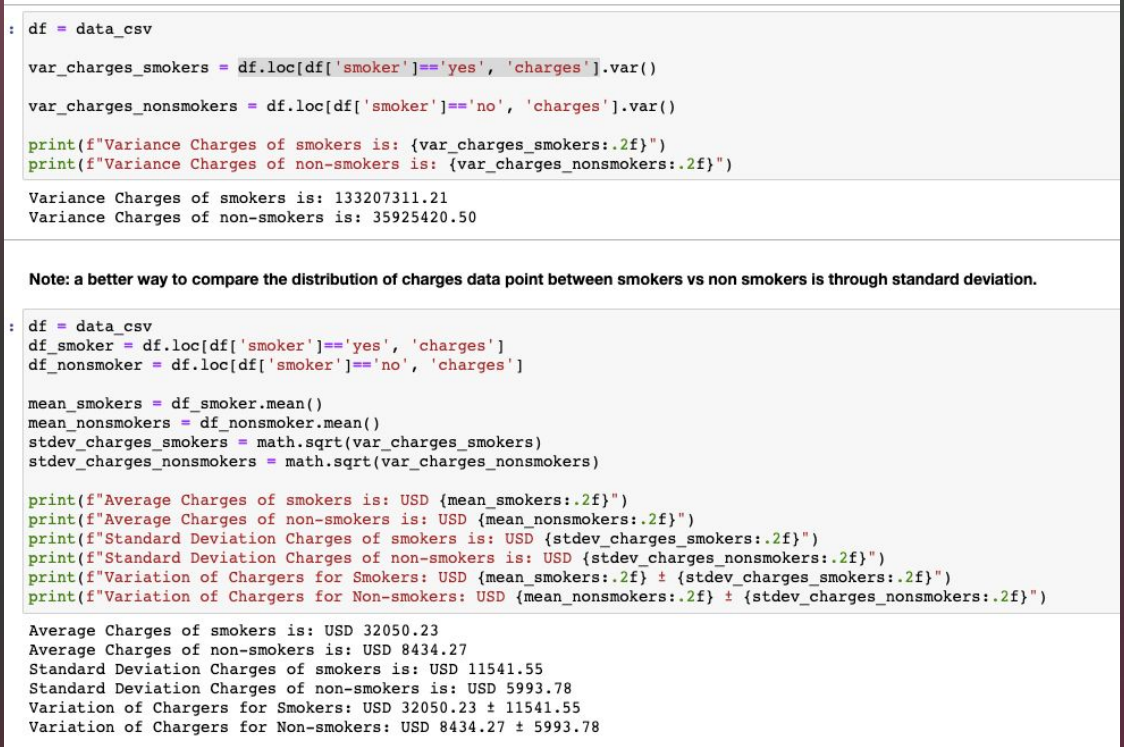
```
df = data_csv

mean_bmi_smokers = df.loc[df['smoker']=='yes', 'bmi'].mean()
mean_bmi_nonsmokers = df.loc[df['smoker']=='no', 'bmi'].mean()
mean_bmi = df['bmi'].mean()

print(f"Average BMI is: {mean_bmi:.2f}")
print(f"Average BMI of smokers is: {mean_bmi_smokers:.2f}")
print(f"Average BMI of non-smokers is: {mean_bmi_nonsmokers:.2f}")
```

```
Average BMI is: 30.66
Average BMI of smokers is: 30.71
Average BMI of non-smokers is: 30.65
```

3.1 Descriptive Statistical Analysis

Question 2-4	**Find the variance and standard deviation of charges between smokers & non smokers**
Method	1. Filter data by smoke: smoker == ‘yes’ , ‘no’ 2. Use .var(), .sqrt() of variance to find the variance and standard deviation between smokers and non-smokers charges
Results	● Variance Charges of smokers is: 133207311.21 ● Variance Charges of non-smokers is: 35925420.50 ● Standard Deviation Charges of smokers is: USD 11541.55 ● Standard Deviation Charges of non-smokers is: USD 5993.78 ● Average Charges of smokers is: USD 32050.23 ● Average Charges of non-smokers is: USD 8434.27
Findings	● Smokers has higher mean, variance and stdev for charges than non-smokers. ● Variation of charges for smokers (USD 32050.23 ± 11541.55) is greater than that of non smokers (USD 8434.27 ± 5993.78). ● This indicates the spread of charges for smokers is wider than non-smokers with smokers having higher highs and lower lows when it comes to charges.

3.1 Descriptive Statistical Analysis - Method Screenshot



```
df = data_csv

var_charges_smokers = df.loc[df['smoker']=='yes', 'charges'].var()

var_charges_nonsmokers = df.loc[df['smoker']=='no', 'charges'].var()

print(f"Variance Charges of smokers is: {var_charges_smokers:.2f}")
print(f"Variance Charges of non-smokers is: {var_charges_nonsmokers:.2f}")
```

```
Variance Charges of smokers is: 133207311.21
Variance Charges of non-smokers is: 35925420.50
```

Note: a better way to compare the distribution of charges data point between smokers vs non smokers is through standard deviation.

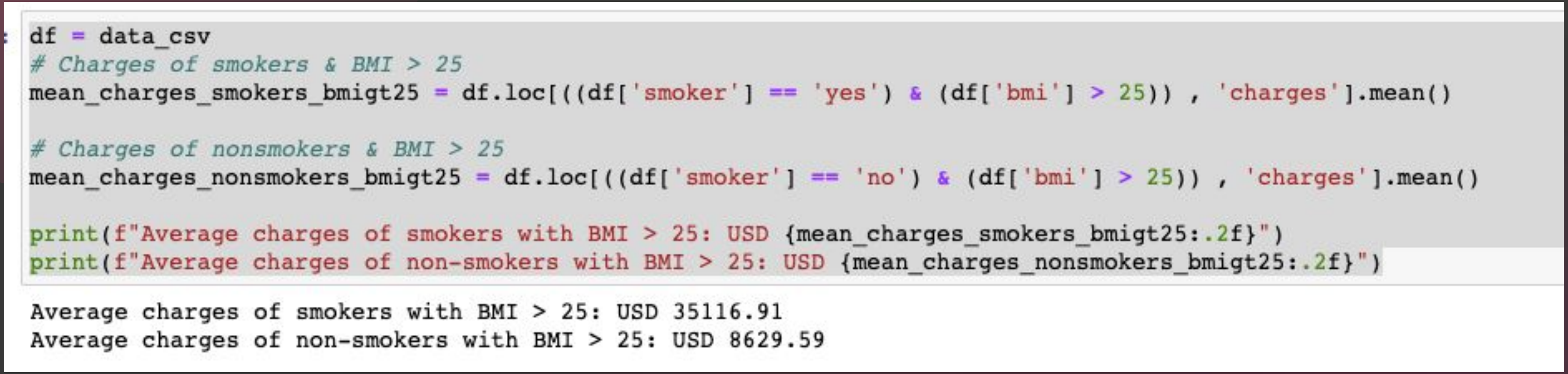
```
df = data_csv
df_smoker = df.loc[df['smoker']=='yes', 'charges']
df_nonsmoker = df.loc[df['smoker']=='no', 'charges']

mean_smokers = df_smoker.mean()
mean_nonsmokers = df_nonsmoker.mean()
stdev_charges_smokers = math.sqrt(var_charges_smokers)
stdev_charges_nonsmokers = math.sqrt(var_charges_nonsmokers)

print(f"Average Charges of smokers is: USD {mean_smokers:.2f}")
print(f"Average Charges of non-smokers is: USD {mean_nonsmokers:.2f}")
print(f"Standard Deviation Charges of smokers is: USD {stdev_charges_smokers:.2f}")
print(f"Standard Deviation Charges of non-smokers is: USD {stdev_charges_nonsmokers:.2f}")
print(f"Variation of Chargers for Smokers: USD {mean_smokers:.2f} ± {stdev_charges_smokers:.2f}")
print(f"Variation of Chargers for Non-smokers: USD {mean_nonsmokers:.2f} ± {stdev_charges_nonsmokers:.2f}")
```

```
Average Charges of smokers is: USD 32050.23
Average Charges of non-smokers is: USD 8434.27
Standard Deviation Charges of smokers is: USD 11541.55
Standard Deviation Charges of non-smokers is: USD 5993.78
Variation of Chargers for Smokers: USD 32050.23 ± 11541.55
Variation of Chargers for Non-smokers: USD 8434.27 ± 5993.78
```

3.1 Descriptive Statistical Analysis

Question 5	**Is the average charges of smokers with BMI >25 greater than non-smokers with BMI > 25?**
Method	1. Filter data by smoke: smoker == ‘yes’ & bmi > 25 and smoker ==’no’ & bmi > 25 2. Use .mean() function to find average of the two groups
Results	● Average charges of smokers with BMI > 25: USD 35116.91 ● Average charges of non-smokers with BMI > 25: USD 8629.59
Findings	● From previous calculation, avg charges for smokers and non-smokers are USD 32050.23 and USD 8434.27 respectively. ● When BMI of sample is more than 25, the average charges for both smokers and nonsmokers becomes higher. With average charges for smokers with BMI > 25: USD 35116.91 greater than non-smokers with BMI > 25: USD 8629.59.

3.1 Descriptive Statistical Analysis - Method Screenshot



```
df = data_csv
# Charges of smokers & BMI > 25
mean_charges_smokers_bmigt25 = df.loc[((df['smoker'] == 'yes') & (df['bmi'] > 25)) , 'charges'].mean()

# Charges of nonsmokers & BMI > 25
mean_charges_nonsmokers_bmigt25 = df.loc[((df['smoker'] == 'no') & (df['bmi'] > 25)) , 'charges'].mean()

print(f"Average charges of smokers with BMI > 25: USD {mean_charges_smokers_bmigt25:.2f}")
print(f"Average charges of non-smokers with BMI > 25: USD {mean_charges_nonsmokers_bmigt25:.2f}")
```

```
Average charges of smokers with BMI > 25: USD 35116.91
Average charges of non-smokers with BMI > 25: USD 8629.59
```

3.2 Categorical Variable Analysis

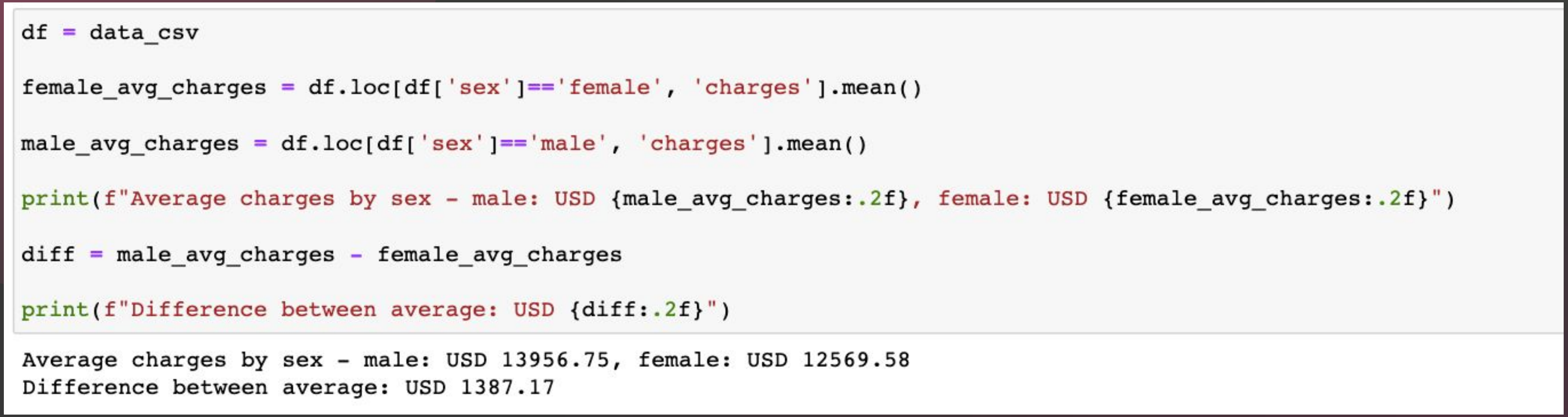
Question selected:

1. Which gender has higher average charges?
2. Is proportion of individuals who are smokers greater than non-smokers?
3. What is the probability of an individual is a female given that she is a smoker?
4. What is the probability of an individual is a male given that he is a smoker?
5. Distribution of Charges by Region

3.2 Categorical Variable Analysis

Question 1	**Which gender has higher average charges?**
Method	1. Filter data by sex: male, female 2. Use .mean() function to find average charges for the two groups
Results	● Average charges by sex - male: USD 13956.75 ● Average charges by sex - female: USD 12569.58
Findings	● Average charges for male is higher than female by USD 1387.17. ● Hence, male in US sample population tend to pay higher medical charges as compared to female.

3.2 Categorical Variable Analysis - Method Screenshot



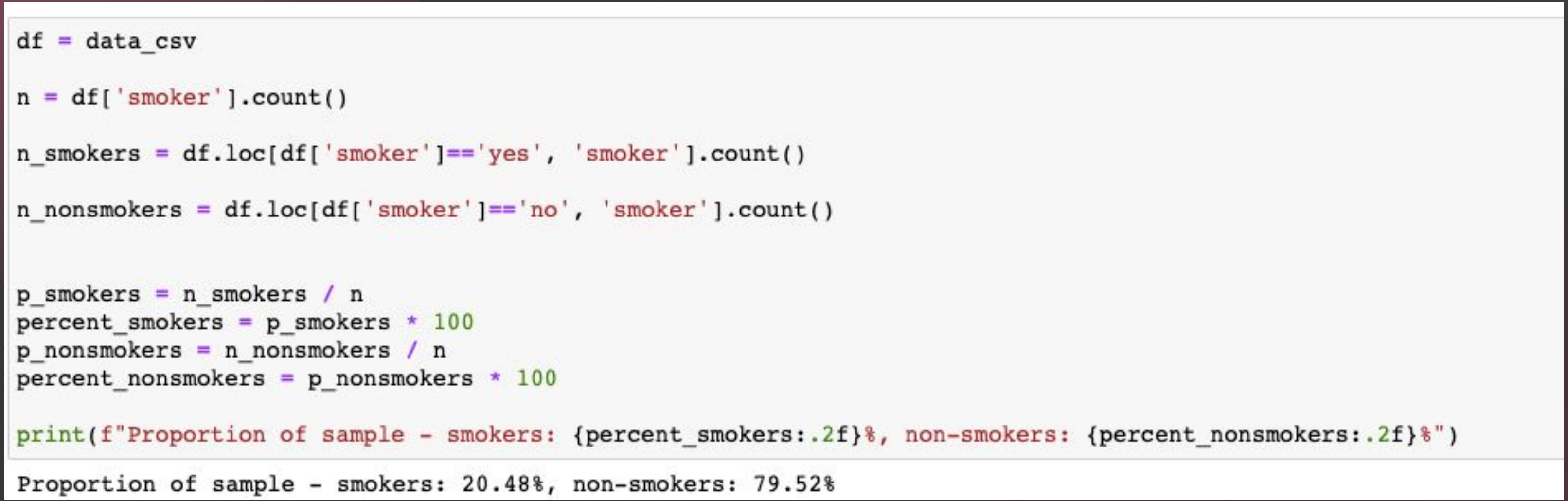
```
df = data_csv

female_avg_charges = df.loc[df['sex']=='female', 'charges'].mean()

male_avg_charges = df.loc[df['sex']=='male', 'charges'].mean()

print(f"Average charges by sex - male: USD {male_avg_charges:.2f}, female: USD {female_avg_charges:.2f}")

diff = male_avg_charges - female_avg_charges

print(f"Difference between average: USD {diff:.2f}")
```

```
Average charges by sex - male: USD 13956.75, female: USD 12569.58
Difference between average: USD 1387.17
```

3.2 Categorical Variable Analysis

Question 2	**Is proportion of individuals who are smokers greater than non-smokers?**
Method	1. Filter data by ‘smoker’: yes for smoker, no for non-smoker 2. Use probability to find proportion by dividing count of smoker / count of sample and similarly for non-smoker
Results	● Proportion of sample - smokers: 20.48%, non-smokers: 79.52%
Findings	● In this sample data of US population, there are much more non-smokers than smokers. We can say there are almost 8 non-smokers and only 2 smokers for every 10 individuals in US.

3.2 Categorical Variable Analysis - Method Screenshot



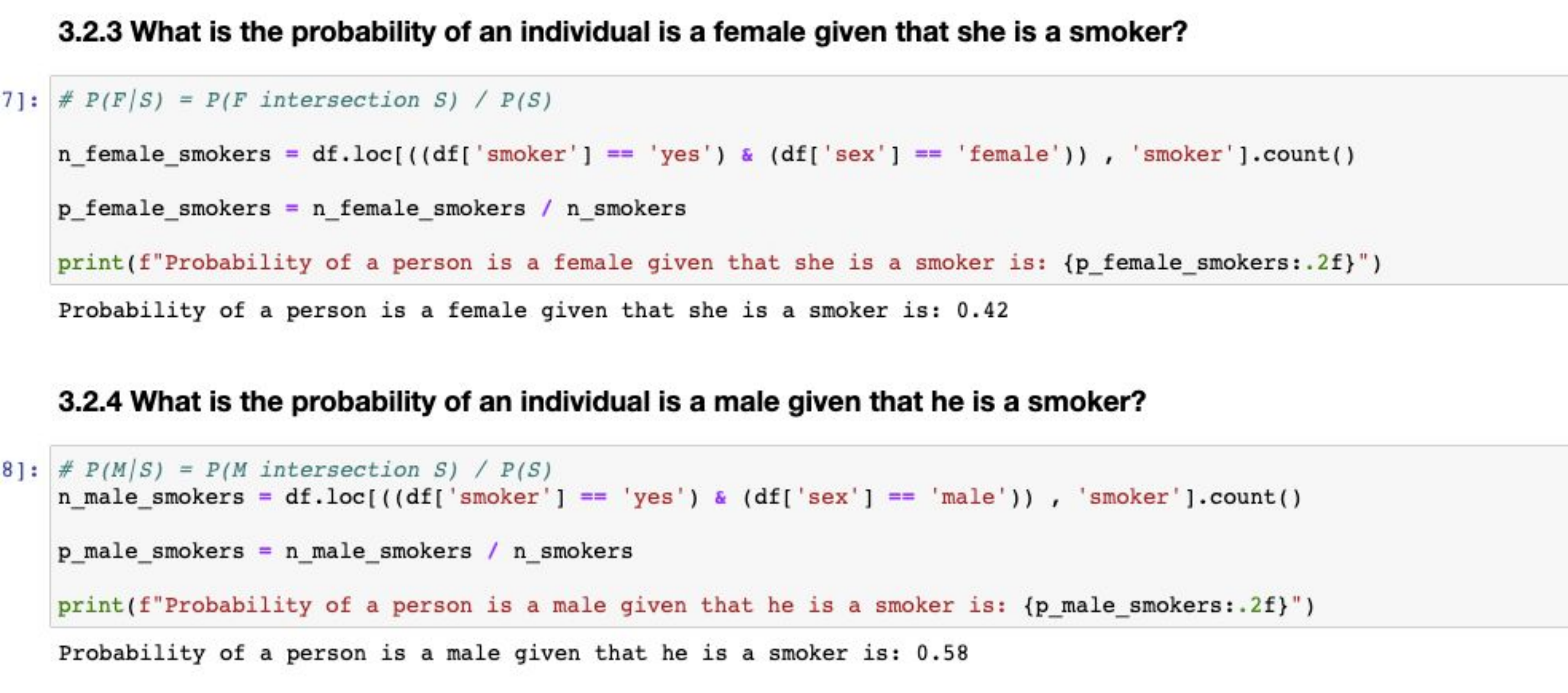
```python
df = data_csv

n = df['smoker'].count()

n_smokers = df.loc[df['smoker']=='yes', 'smoker'].count()

n_nonsmokers = df.loc[df['smoker']=='no', 'smoker'].count()


p_smokers = n_smokers / n
percent_smokers = p_smokers * 100
p_nonsmokers = n_nonsmokers / n
percent_nonsmokers = p_nonsmokers * 100

print(f"Proportion of sample - smokers: {percent_smokers:.2f}%, non-smokers: {percent_nonsmokers:.2f}%")
```

```
Proportion of sample - smokers: 20.48%, non-smokers: 79.52%
```

3.2 Categorical Variable Analysis

Question 3-4	**What is the probability of an individual is a female given that she is a smoker?** **What is the probability of an individual is a male given that he is a smoker?**
Method	To find probability of female given that she is a smoker: 1. Find count of female and smoker = ‘yes’ using .count() function 2. Find count of smoker = ‘yes’ using .count() function 3. Use conditional probability by dividing 1 / 2 To find probability of male given that he is a smoker: 1. Find count of male and smoker = ‘yes’ using .count() function 2. Find count of smoker = ‘yes’ using .count() function 3. Use conditional probability by dividing 1 / 2
Results	● Probability of a person is a female given that she is a smoker is: 0.42 ● Probability of a person is a male given that he is a smoker is: 0.58
Findings	● As probability of smoker being a male is higher than that of female, it is likely that there are more male who is a smoker than female in the US sample population.

3.2 Categorical Variable Analysis - Method Screenshot



3.2.3 What is the probability of an individual is a female given that she is a smoker?

```
7]: # P(F|S) = P(F intersection S) / P(S)

n_female_smokers = df.loc[((df['smoker'] == 'yes') & (df['sex'] == 'female')) , 'smoker'].count()

p_female_smokers = n_female_smokers / n_smokers

print(f"Probability of a person is a female given that she is a smoker is: {p_female_smokers:.2f}")
```

```
Probability of a person is a female given that she is a smoker is: 0.42
```

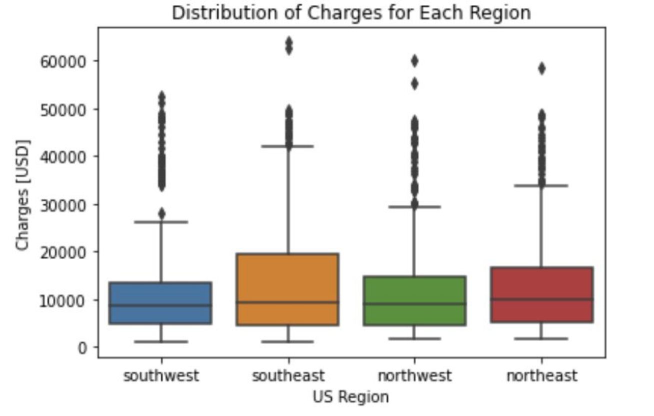
3.2.4 What is the probability of an individual is a male given that he is a smoker?

```
8]: # P(M|S) = P(M intersection S) / P(S)
n_male_smokers = df.loc[((df['smoker'] == 'yes') & (df['sex'] == 'male')) , 'smoker'].count()

p_male_smokers = n_male_smokers / n_smokers

print(f"Probability of a person is a male given that he is a smoker is: {p_male_smokers:.2f}")
```

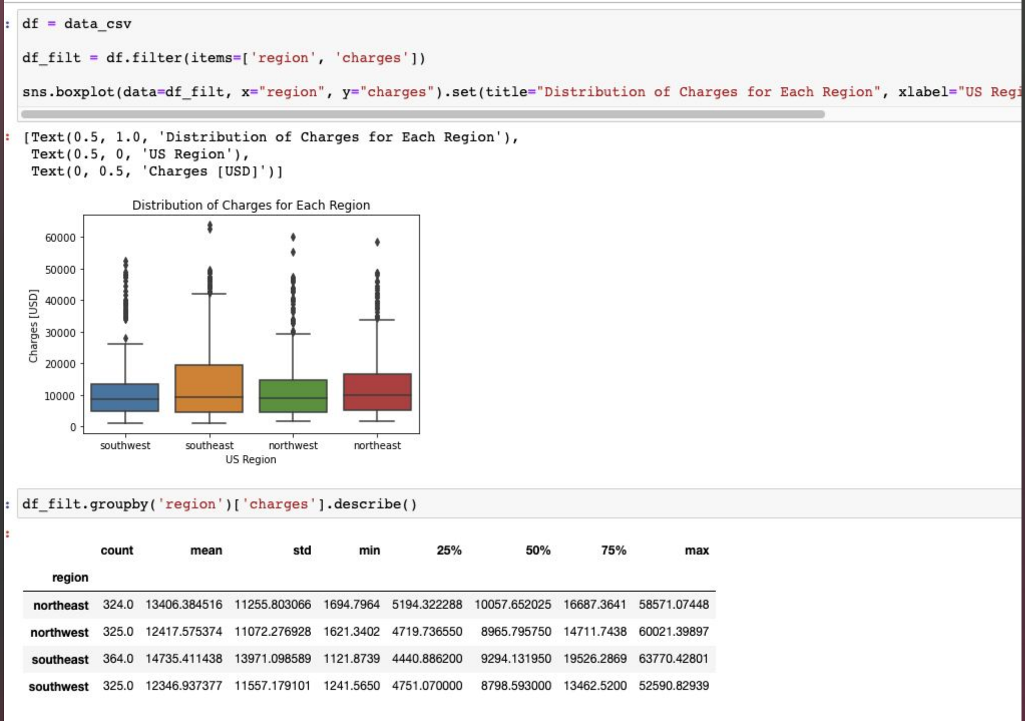
```
Probability of a person is a male given that he is a smoker is: 0.58
```

3.2 Categorical Variable Analysis

Question 5	**How is the distribution of charges for each region**
Method	1. Use box plot to analyze distribution between each regions and charges 2. Use .describe() to compare statistics of each region
Results	● Box plot that shows distribution of charges by region
Findings	● Southeast region has the widest distribution and southwest region have the lowest spread in distribution. ● Majority of data in Southeast region are in the upper quartile and this shows that majority of higher charges comes from Southeast region. ● Additionally, Southeast region also has the higher highest charges among the rest of the regions.



3.2 Categorical Variable Analysis - Method Screenshot



```
df = data_csv

df_filt = df.filter(items=['region', 'charges'])

sns.boxplot(data=df_filt, x="region", y="charges").set(title="Distribution of Charges for Each Region", xlabel="US Regi
```

```
[Text(0.5, 1.0, 'Distribution of Charges for Each Region'),
 Text(0.5, 0, 'US Region'),
 Text(0, 0.5, 'Charges [USD]')]
```

```
df_filt.groupby('region')['charges'].describe()
```

	count	mean	std	min	25%	50%	75%	max
region								
northeast	324.0	13406.384516	11255.803066	1694.7964	5194.322288	10057.652025	16687.3641	58571.07448
northwest	325.0	12417.575374	11072.276928	1621.3402	4719.736550	8965.795750	14711.7438	60021.39897
southeast	364.0	14735.411438	13971.098589	1121.8739	4440.886200	9294.131950	19526.2869	63770.42801
southwest	325.0	12346.937377	11557.179101	1241.5650	4751.070000	8798.593000	13462.5200	52590.82939

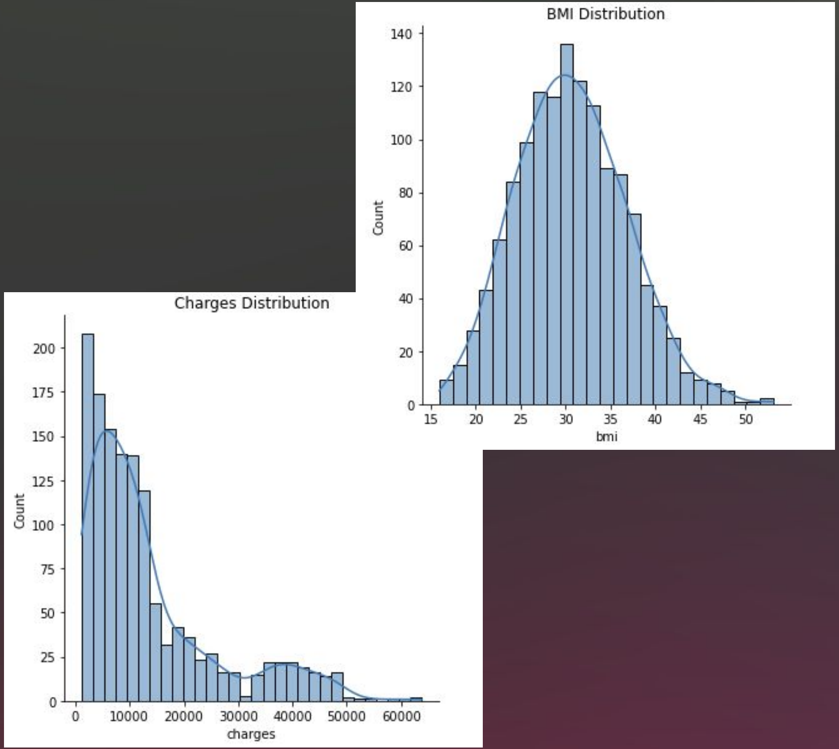
3.3 Continuous Variable Analysis

Question selected:

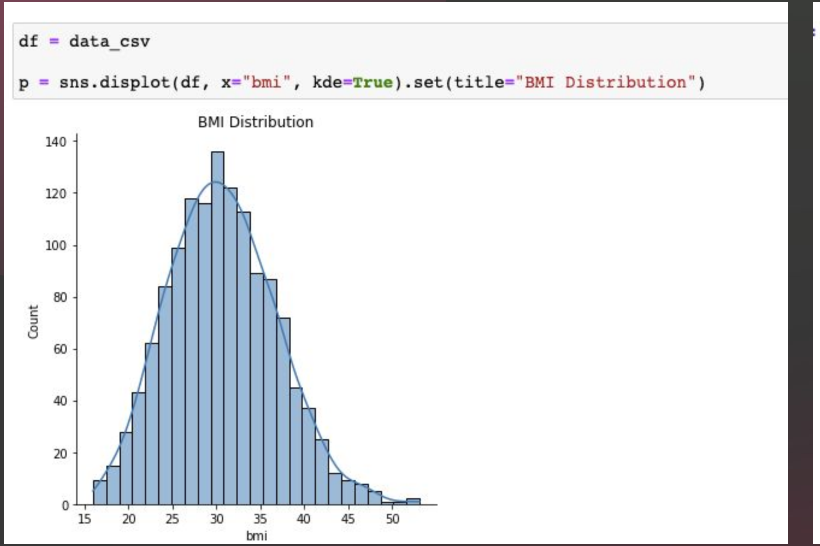
1. Distribution of BMI and Charges
2. Which is more likely to occur?
   a. An individual with BMI > 25 getting charges over 16.7k?
   b. An individual with BMI < 25 getting charges over 16.7k?
3. Which is more likely to occur?
   a. A smoker with BMI > 25 getting charges over 16.7k?
   b. A non-smoker with BMI > 25 getting charges over 16.7k?
4. Which is more likely to occur?
   a. An female individual with BMI > 25 getting charges over 16.7k?
   b. An male individual with BMI > 25 getting charges over 16.7k?
5. Which is more likely to occur?
   a. An female individual who smokes with BMI > 25 getting charges over 16.7k
   b. An male individual who smokes with BMI > 25 getting charges over 16.7k?

3.3 Continuous Variable Analysis

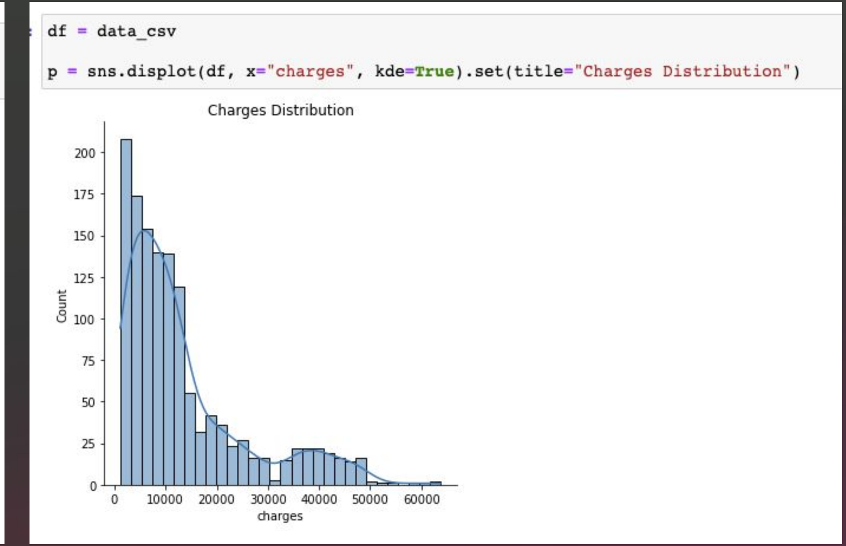
Question 1	Distribution of BMI and Charges
Method	● Plot histogram for bmi against frequency using seaborn ● Plot histogram for charges against frequency using seaborn
Results and findings	BMI Histogram: ● Normally distribution with central tendency Charges Histogram: ● Skewed distribution to the right with most data points falling in < USD 15000



3.3 Continuous Variable Analysis - Method Screenshot



```
df = data_csv

p = sns.displot(df, x="bmi", kde=True).set(title="BMI Distribution")
```

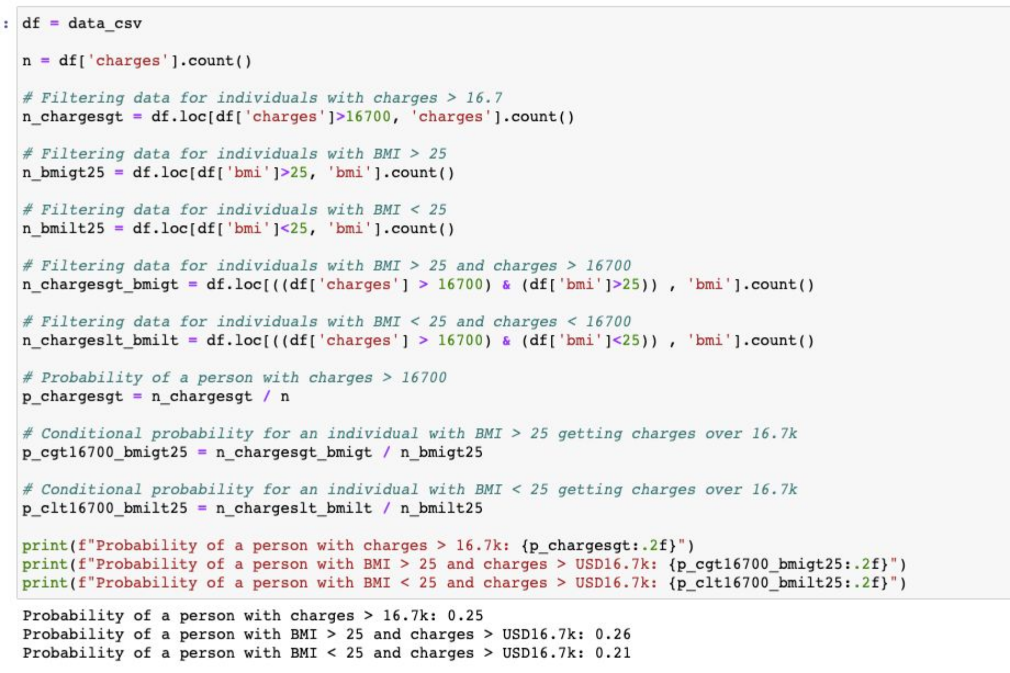


```
df = data_csv

p = sns.displot(df, x="charges", kde=True).set(title="Charges Distribution")
```

3.3 Continuous Variable Analysis

Question 2	**Which is more likely to occur?** **An individual with BMI > 25 getting charges over 16.7k?** **An individual with BMI < 25 getting charges over 16.7k?**
Method	● Conditional probability: count of individuals with BMI > 25 & charges > 16.7k / count of individuals with BMI > 25 ● Conditional probability: count of individuals with BMI < 25 & charges > 16.7k / count of individuals with BMI < 25 ● Probability of charges > 16.7k: count of charges > 16.7k / count of sample
Results	Probability of a person with charges > 16.7k: 0.25 Probability of a person with BMI > 25 and charges > USD16.7k: 0.26 Probability of a person with BMI < 25 and charges > USD16.7k: 0.21
Findings	● Probability having charges over USD 16.7k increases slightly from 0.25 to 0.26 when BMI > 25 is known. ● Although the difference is not significant, probability of person with BMI > 25 with charges over USD 16.7k is larger than that with BMI < 25.

3.3 Continuous Variable Analysis - Method Screenshot



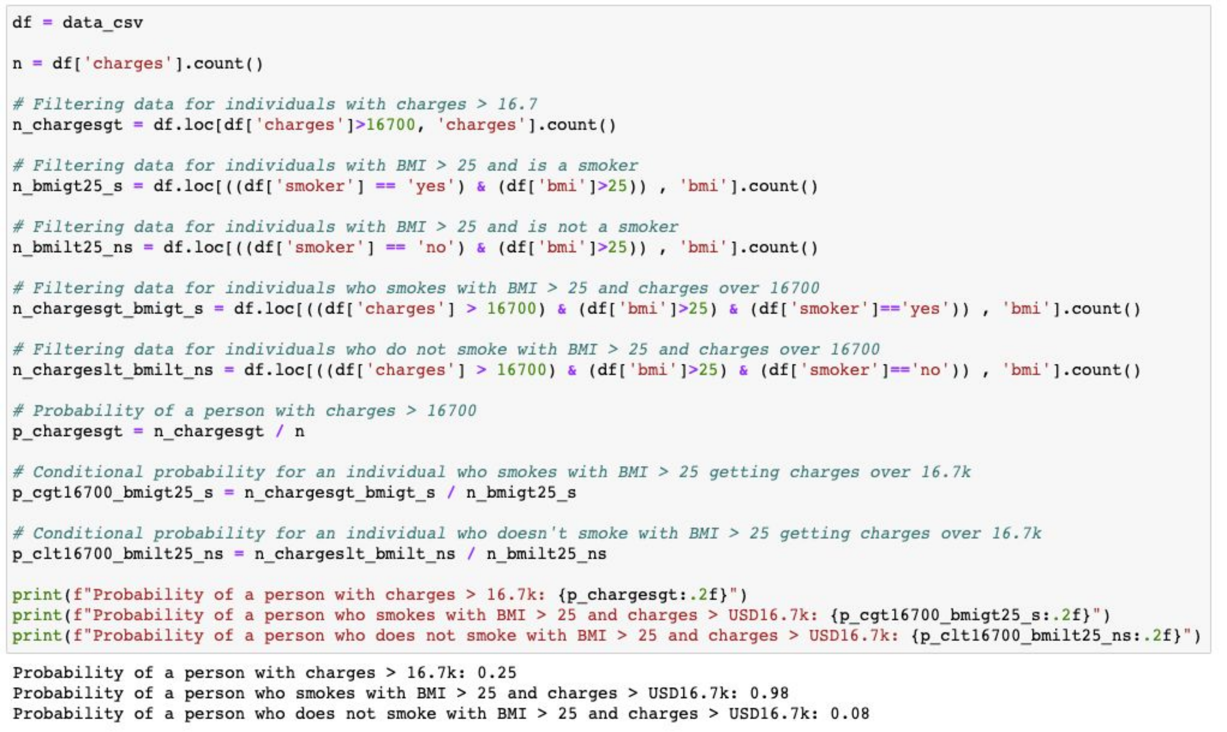
```
df = data_csv

n = df['charges'].count()

# Filtering data for individuals with charges > 16.7
n_chargesgt = df.loc[df['charges']>16700, 'charges'].count()

# Filtering data for individuals with BMI > 25
n_bmigt25 = df.loc[df['bmi']>25, 'bmi'].count()

# Filtering data for individuals with BMI < 25
n_bmilt25 = df.loc[df['bmi']<25, 'bmi'].count()

# Filtering data for individuals with BMI > 25 and charges > 16700
n_chargesgt_bmigt = df.loc[((df['charges'] > 16700) & (df['bmi']>25)) , 'bmi'].count()

# Filtering data for individuals with BMI < 25 and charges < 16700
n_chargeslt_bmilt = df.loc[((df['charges'] > 16700) & (df['bmi']<25)) , 'bmi'].count()

# Probability of a person with charges > 16700
p_chargesgt = n_chargesgt / n

# Conditional probability for an individual with BMI > 25 getting charges over 16.7k
p_cgt16700_bmigt25 = n_chargesgt_bmigt / n_bmigt25

# Conditional probability for an individual with BMI < 25 getting charges over 16.7k
p_clt16700_bmilt25 = n_chargeslt_bmilt / n_bmilt25

print(f"Probability of a person with charges > 16.7k: {p_chargesgt:.2f}")
print(f"Probability of a person with BMI > 25 and charges > USD16.7k: {p_cgt16700_bmigt25:.2f}")
print(f"Probability of a person with BMI < 25 and charges > USD16.7k: {p_clt16700_bmilt25:.2f}")
```

```
Probability of a person with charges > 16.7k: 0.25
Probability of a person with BMI > 25 and charges > USD16.7k: 0.26
Probability of a person with BMI < 25 and charges > USD16.7k: 0.21
```

3.3 Continuous Variable Analysis

Question 3	**Which is more likely to occur?** **A smoker with BMI > 25 getting charges over 16.7k?** **A non-smoker with BMI > 25 getting charges over 16.7k?**
Method	● Conditional probability: count of smokers with BMI > 25 & charges > 16.7k / count of individuals with BMI > 25 & smokes ● Conditional probability: count of non-smokers with BMI > 25 & charges > 16.7k / count of individuals with BMI > 25 & do not smoke ● Probability of charges > 16.7k: count of charges > 16.7k / count of sample
Results	Probability of a person with charges > 16.7k: 0.25 Probability of a person who smokes with BMI > 25 and charges > USD16.7k: 0.98 Probability of a person who does not smoke with BMI > 25 and charges > USD16.7k: 0.08
Findings	● Probability having charges over USD 16.7k increases significantly from 0.25 to 0.98 given an individual has BMI > 25 and is a smoker. ● A person who does not smoke has very low probability (0.08) of having charges over USD 16.7k even though she/he has a BMI > 25.

3.3 Continuous Variable Analysis - Method Screenshot



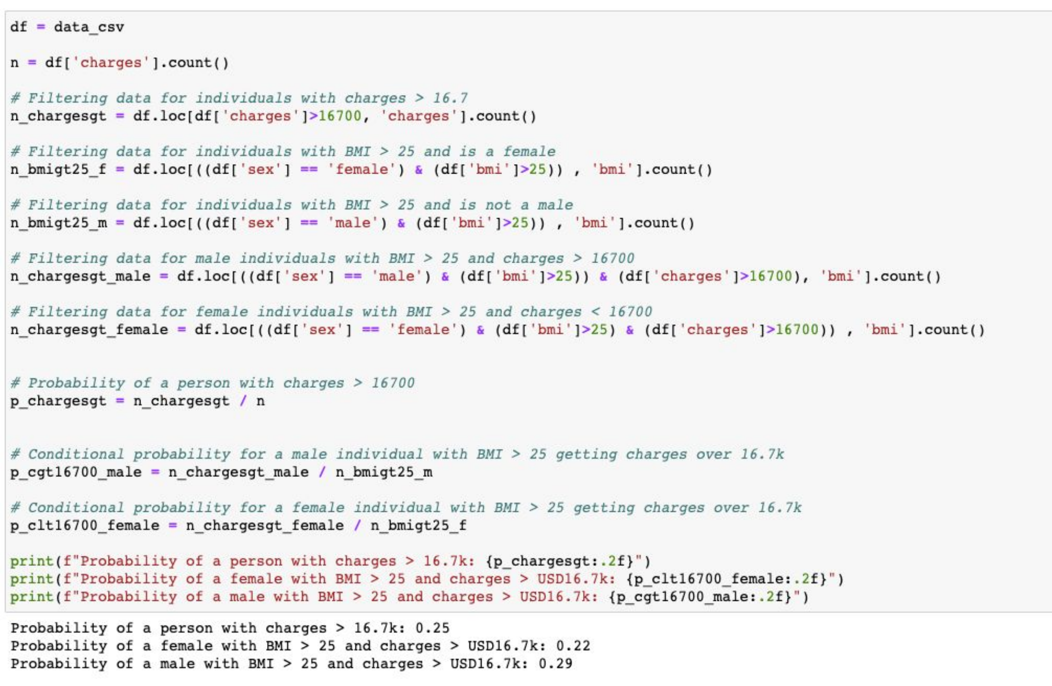
```
df = data_csv

n = df['charges'].count()

# Filtering data for individuals with charges > 16.7
n_chargesgt = df.loc[df['charges']>16700, 'charges'].count()

# Filtering data for individuals with BMI > 25 and is a smoker
n_bmigt25_s = df.loc[((df['smoker'] == 'yes') & (df['bmi']>25)) , 'bmi'].count()

# Filtering data for individuals with BMI > 25 and is not a smoker
n_bmilt25_ns = df.loc[((df['smoker'] == 'no') & (df['bmi']>25)) , 'bmi'].count()

# Filtering data for individuals who smokes with BMI > 25 and charges over 16700
n_chargesgt_bmigt_s = df.loc[((df['charges'] > 16700) & (df['bmi']>25) & (df['smoker']=='yes')) , 'bmi'].count()

# Filtering data for individuals who do not smoke with BMI > 25 and charges over 16700
n_chargeslt_bmilt_ns = df.loc[((df['charges'] > 16700) & (df['bmi']>25) & (df['smoker']=='no')) , 'bmi'].count()

# Probability of a person with charges > 16700
p_chargesgt = n_chargesgt / n

# Conditional probability for an individual who smokes with BMI > 25 getting charges over 16.7k
p_cgt16700_bmigt25_s = n_chargesgt_bmigt_s / n_bmigt25_s

# Conditional probability for an individual who doesn't smoke with BMI > 25 getting charges over 16.7k
p_clt16700_bmilt25_ns = n_chargeslt_bmilt_ns / n_bmilt25_ns

print(f"Probability of a person with charges > 16.7k: {p_chargesgt:.2f}")
print(f"Probability of a person who smokes with BMI > 25 and charges > USD16.7k: {p_cgt16700_bmigt25_s:.2f}")
print(f"Probability of a person who does not smoke with BMI > 25 and charges > USD16.7k: {p_clt16700_bmilt25_ns:.2f}")
```

```
Probability of a person with charges > 16.7k: 0.25
Probability of a person who smokes with BMI > 25 and charges > USD16.7k: 0.98
Probability of a person who does not smoke with BMI > 25 and charges > USD16.7k: 0.08
```

3.3 Continuous Variable Analysis

Question 4	**Which is more likely to occur?** **An female individual with BMI > 25 getting charges over 16.7k?** **An male individual with BMI > 25 getting charges over 16.7k?**
Method	● Conditional probability: count of female with BMI > 25 & charges > 16.7k / count of individuals with BMI > 25 & female ● Conditional probability: count of male with BMI > 25 & charges > 16.7k / count of individuals with BMI > 25 & male ● Probability of charges > 16.7k: count of charges > 16.7k / count of sample
Results	Probability of a person with charges > 16.7k: 0.25 Probability of a female with BMI > 25 and charges > USD16.7k: 0.22 Probability of a male with BMI > 25 and charges > USD16.7k: 0.29
Findings	● Probability between female and male with BMI > 25 to get charges over USD 16.7k does not differ significantly. ● Male with BMI > 25 has higher probability of 0.29 to get charge over USD 16.7k.

3.3 Continuous Variable Analysis - Method Screenshot



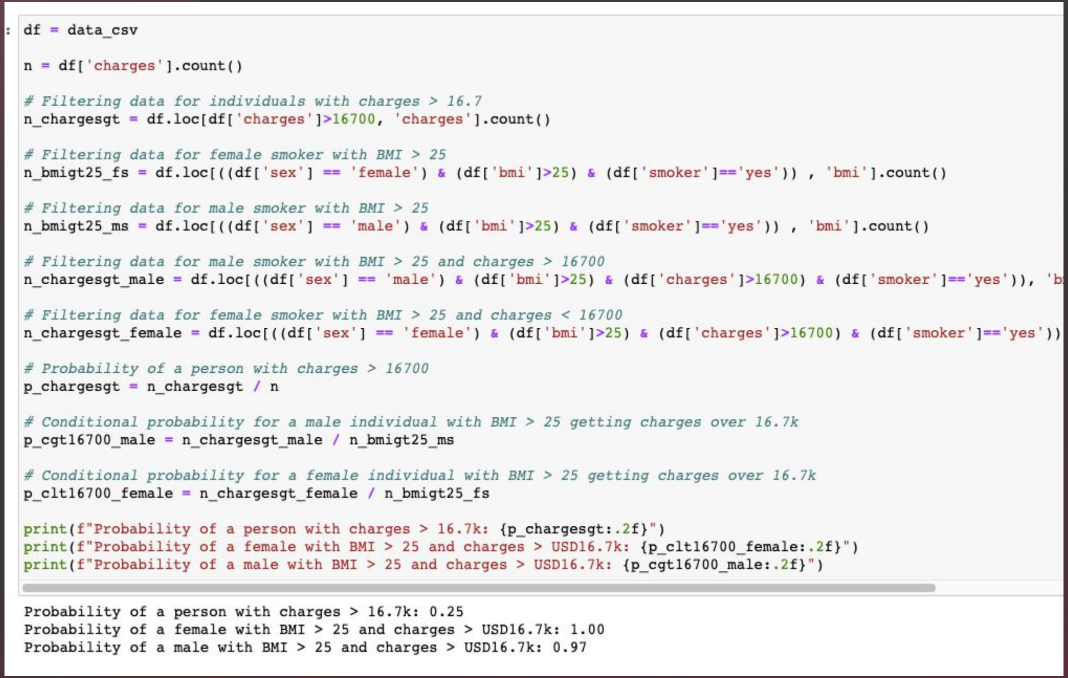
```
df = data_csv

n = df['charges'].count()

# Filtering data for individuals with charges > 16.7
n_chargesgt = df.loc[df['charges']>16700, 'charges'].count()

# Filtering data for individuals with BMI > 25 and is a female
n_bmigt25_f = df.loc[((df['sex'] == 'female') & (df['bmi']>25)) , 'bmi'].count()

# Filtering data for individuals with BMI > 25 and is not a male
n_bmigt25_m = df.loc[((df['sex'] == 'male') & (df['bmi']>25)) , 'bmi'].count()

# Filtering data for male individuals with BMI > 25 and charges > 16700
n_chargesgt_male = df.loc[((df['sex'] == 'male') & (df['bmi']>25)) & (df['charges']>16700), 'bmi'].count()

# Filtering data for female individuals with BMI > 25 and charges < 16700
n_chargesgt_female = df.loc[((df['sex'] == 'female') & (df['bmi']>25) & (df['charges']>16700)) , 'bmi'].count()


# Probability of a person with charges > 16700
p_chargesgt = n_chargesgt / n


# Conditional probability for a male individual with BMI > 25 getting charges over 16.7k
p_cgt16700_male = n_chargesgt_male / n_bmigt25_m

# Conditional probability for a female individual with BMI > 25 getting charges over 16.7k
p_clt16700_female = n_chargesgt_female / n_bmigt25_f

print(f"Probability of a person with charges > 16.7k: {p_chargesgt:.2f}")
print(f"Probability of a female with BMI > 25 and charges > USD16.7k: {p_clt16700_female:.2f}")
print(f"Probability of a male with BMI > 25 and charges > USD16.7k: {p_cgt16700_male:.2f}")
```

```
Probability of a person with charges > 16.7k: 0.25
Probability of a female with BMI > 25 and charges > USD16.7k: 0.22
Probability of a male with BMI > 25 and charges > USD16.7k: 0.29
```

3.3 Continuous Variable Analysis

Question 5	**Which is more likely to occur?** **An female smoker with BMI > 25 getting charges over 16.7k?** **An male smoker with BMI > 25 getting charges over 16.7k?**
Method	● Conditional probability: count of female smoker with BMI > 25 & charges > 16.7k / count of individuals with BMI > 25 & female smoker ● Conditional probability: count of male smoker with BMI > 25 & charges > 16.7k / count of individuals with BMI > 25 & male smoker ● Probability of charges > 16.7k: count of charges > 16.7k / count of sample
Results	Probability of a person with charges > 16.7k: 0.25 Probability of a female smoker with BMI > 25 and charges > USD16.7k: 1.00 Probability of a male smoker with BMI > 25 and charges > USD16.7k: 0.97
Findings	● 100% of female smokers with BMI > 25 have charges over USD 16.7k while 95% of male smokers with BMI > 25 have charges over USD 16.7k. ● The probability of getting charged over USD 16.7k increases greatly when someone is a smoker and have a BMI > 25. Between genders in this group, there is not much difference in probability.

3.3 Continuous Variable Analysis - Method Screenshot



```
df = data_csv

n = df['charges'].count()

# Filtering data for individuals with charges > 16.7
n_chargesgt = df.loc[df['charges']>16700, 'charges'].count()

# Filtering data for female smoker with BMI > 25
n_bmigt25_fs = df.loc[((df['sex'] == 'female') & (df['bmi']>25) & (df['smoker']=='yes')) , 'bmi'].count()

# Filtering data for male smoker with BMI > 25
n_bmigt25_ms = df.loc[((df['sex'] == 'male') & (df['bmi']>25) & (df['smoker']=='yes')) , 'bmi'].count()

# Filtering data for male smoker with BMI > 25 and charges > 16700
n_chargesgt_male = df.loc[((df['sex'] == 'male') & (df['bmi']>25) & (df['charges']>16700) & (df['smoker']=='yes')), 'b

# Filtering data for female smoker with BMI > 25 and charges < 16700
n_chargesgt_female = df.loc[((df['sex'] == 'female') & (df['bmi']>25) & (df['charges']>16700) & (df['smoker']=='yes'))

# Probability of a person with charges > 16700
p_chargesgt = n_chargesgt / n

# Conditional probability for a male individual with BMI > 25 getting charges over 16.7k
p_cgt16700_male = n_chargesgt_male / n_bmigt25_ms

# Conditional probability for a female individual with BMI > 25 getting charges over 16.7k
p_clt16700_female = n_chargesgt_female / n_bmigt25_fs

print(f"Probability of a person with charges > 16.7k: {p_chargesgt:.2f}")
print(f"Probability of a female with BMI > 25 and charges > USD16.7k: {p_clt16700_female:.2f}")
print(f"Probability of a male with BMI > 25 and charges > USD16.7k: {p_cgt16700_male:.2f}")
```

```
Probability of a person with charges > 16.7k: 0.25
Probability of a female with BMI > 25 and charges > USD16.7k: 1.00
Probability of a male with BMI > 25 and charges > USD16.7k: 0.97
```

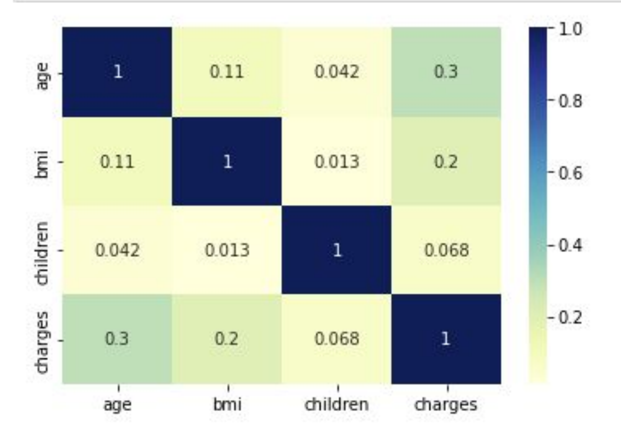
3.4 Correlation Analysis

Question: How does the correlation coefficient between each variables?

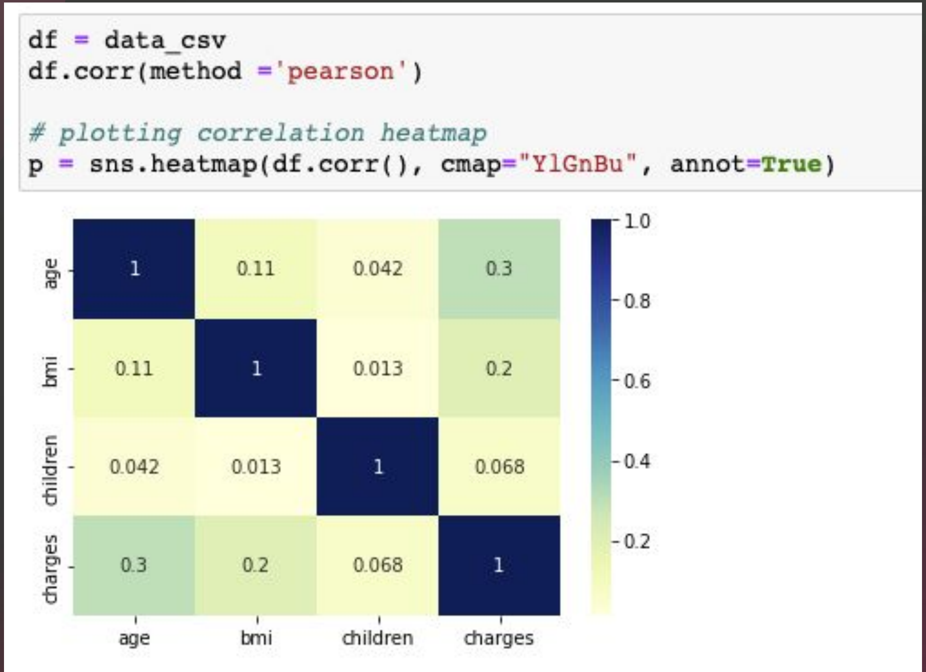
1. Correlation between charges, bmi, age and children
2. Correlation between charges and region
3. Correlation between charges and sex
4. Correlation between charges and smokers
5. Correlation between charges, smokers and sex

3.4 Correlation Analysis

Question 1	**Correlation between charges, bmi, age, and children**
Method	● SNS heatmap with pearson correlation
Results	● r - charges, bmi: 0.198 ● r - charges, age: 0.299 ● r - charges, children: 0.0680
Findings	● Age, bmi and children have positive correlation coefficient with magnitude <= 0.3. This implies that there is weak positive correlation between age, bmi and children with charges. ● Hence, increase in any of these variables does not particularly lead to an increase in charges



3.4 Correlation Analysis - Method Screenshot

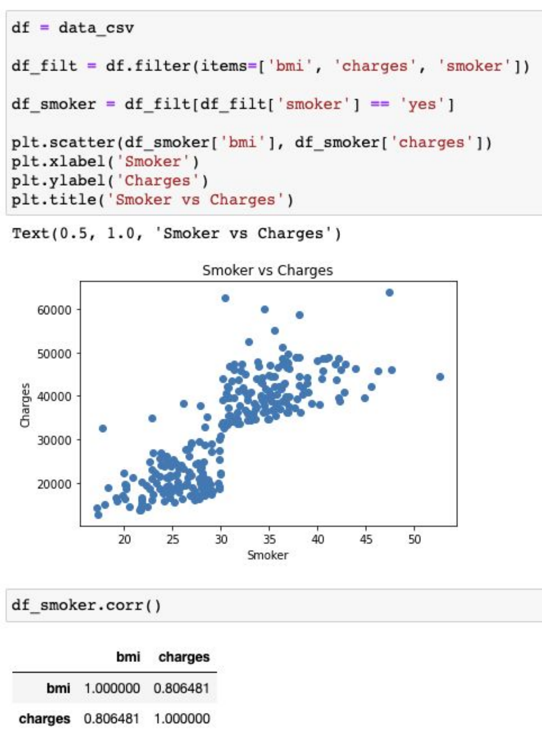


```
df = data_csv
df.corr(method ='pearson')

# plotting correlation heatmap
p = sns.heatmap(df.corr(), cmap="YlGnBu", annot=True)
```

3.4 Correlation Analysis

Question 2	**Correlation between charges, smokers and non-smokers**
Method	● Matplotlib scatter plot ● .corr() to find correlation with pearson method
Results	● r - charges, smokers: 0.81 ● r - charges, non-smokers: 0.08
Findings	● Smokers and charges are strongly correlated with r = 0.81 (r > 0.5). While non-smokers and charges have almost no correlation with r approaching 0. ● Smokers are much more likely to see a higher charges than non-smokers.

3.4 Correlation Analysis - Method Screenshot

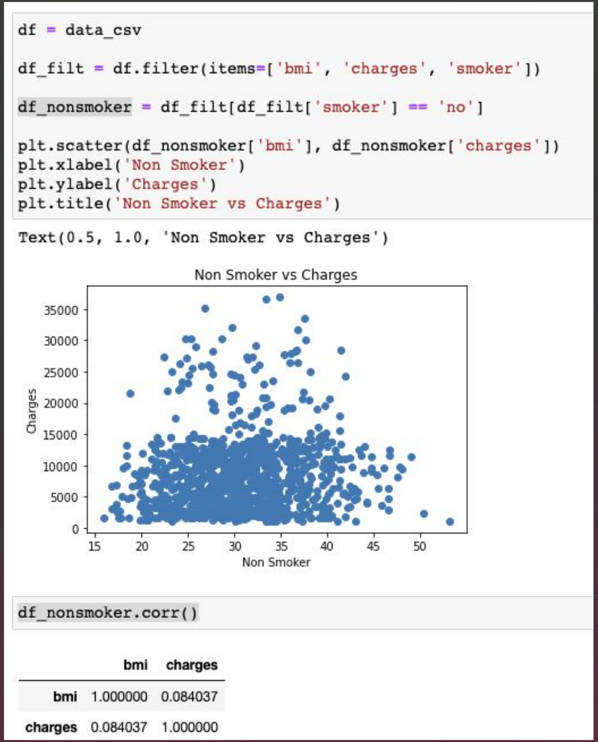


```
df = data_csv

df_filt = df.filter(items=['bmi', 'charges', 'smoker'])

df_smoker = df_filt[df_filt['smoker'] == 'yes']

plt.scatter(df_smoker['bmi'], df_smoker['charges'])
plt.xlabel('Smoker')
plt.ylabel('Charges')
plt.title('Smoker vs Charges')
```

```
Text(0.5, 1.0, 'Smoker vs Charges')
```

```
df_smoker.corr()
```

	bmi	charges
bmi	1.000000	0.806481
charges	0.806481	1.000000



```
df = data_csv

df_filt = df.filter(items=['bmi', 'charges', 'smoker'])

df_nonsmoker = df_filt[df_filt['smoker'] == 'no']

plt.scatter(df_nonsmoker['bmi'], df_nonsmoker['charges'])
plt.xlabel('Non Smoker')
plt.ylabel('Charges')
plt.title('Non Smoker vs Charges')
```

```
Text(0.5, 1.0, 'Non Smoker vs Charges')
```

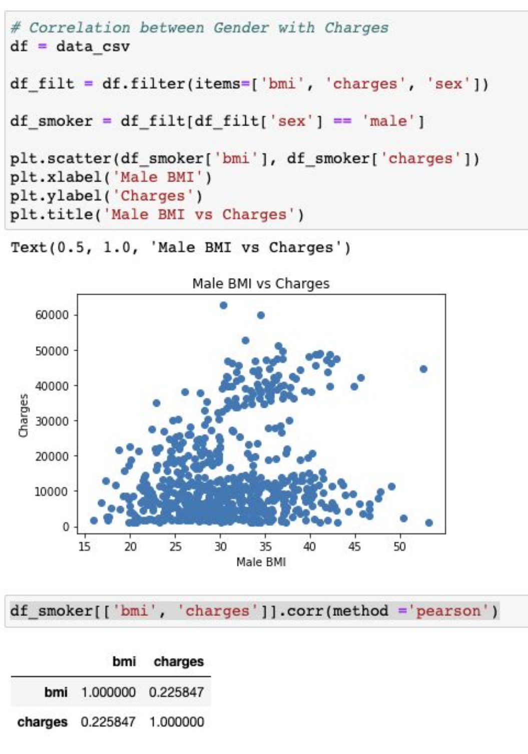
```
df_nonsmoker.corr()
```

	bmi	charges
bmi	1.000000	0.084037
charges	0.084037	1.000000

3.4 Correlation Analysis

Question 3	**Correlation between charges and sex**
Method	● Matplotlib scatter plot ● .corr() to find correlation with pearson method
Results	● r - charges, male: 0.225 ● r - charges, female: 0.161
Findings	● Both female and male have weak positive correlation with charges. With male correlation stronger than female. ● Hence, male are slightly more likely to get higher charges than female.

3.4 Correlation Analysis - Screenshot Method

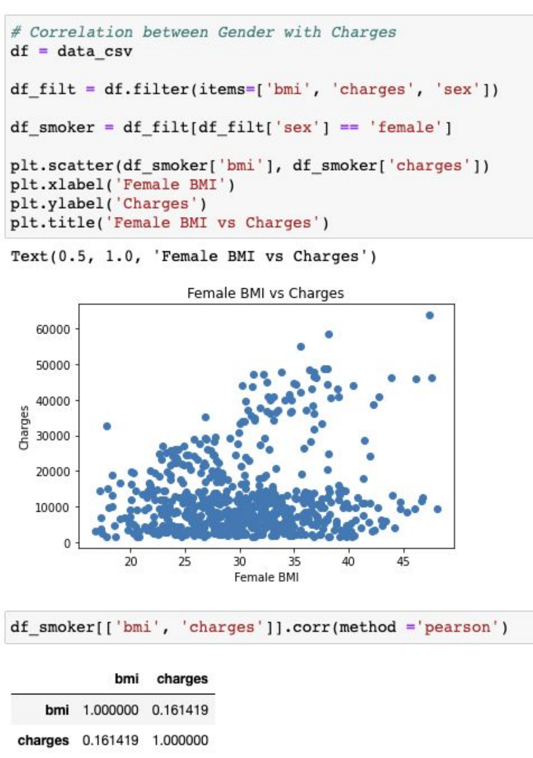


```
# Correlation between Gender with Charges
df = data_csv

df_filt = df.filter(items=['bmi', 'charges', 'sex'])

df_smoker = df_filt[df_filt['sex'] == 'male']

plt.scatter(df_smoker['bmi'], df_smoker['charges'])
plt.xlabel('Male BMI')
plt.ylabel('Charges')
plt.title('Male BMI vs Charges')
```

```
Text(0.5, 1.0, 'Male BMI vs Charges')
```

```
df_smoker[['bmi', 'charges']].corr(method ='pearson')
```

	bmi	charges
bmi	1.000000	0.225847
charges	0.225847	1.000000



```
# Correlation between Gender with Charges
df = data_csv

df_filt = df.filter(items=['bmi', 'charges', 'sex'])

df_smoker = df_filt[df_filt['sex'] == 'female']

plt.scatter(df_smoker['bmi'], df_smoker['charges'])
plt.xlabel('Female BMI')
plt.ylabel('Charges')
plt.title('Female BMI vs Charges')
```

```
Text(0.5, 1.0, 'Female BMI vs Charges')
```

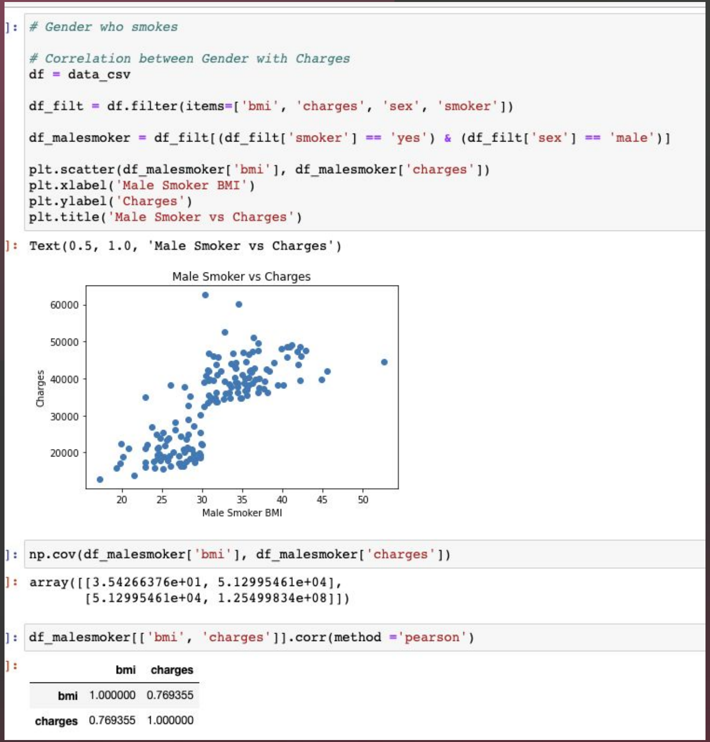
```
df_smoker[['bmi', 'charges']].corr(method ='pearson')
```

	bmi	charges
bmi	1.000000	0.161419
charges	0.161419	1.000000

3.4 Correlation Analysis

Question 4	**Correlation between charges, smokers and sex**
Method	● Matplotlib scatter plot ● .corr() to find correlation with pearson method
Results	● r - charges, male smokers: 0.789 ● r - charges, female smokers: 0.846
Findings	● Both female and male smokers have strong positive correlation with charges > 0.5. With male smokers correlation slightly smaller than female smokers .

3.4 Correlation Analysis - Method Screenshot



```
# Gender who smokes

# Correlation between Gender with Charges
df = data_csv

df_filt = df.filter(items=['bmi', 'charges', 'sex', 'smoker'])

df_malesmoker = df_filt[(df_filt['smoker'] == 'yes') & (df_filt['sex'] == 'male')]

plt.scatter(df_malesmoker['bmi'], df_malesmoker['charges'])
plt.xlabel('Male Smoker BMI')
plt.ylabel('Charges')
plt.title('Male Smoker vs Charges')
```

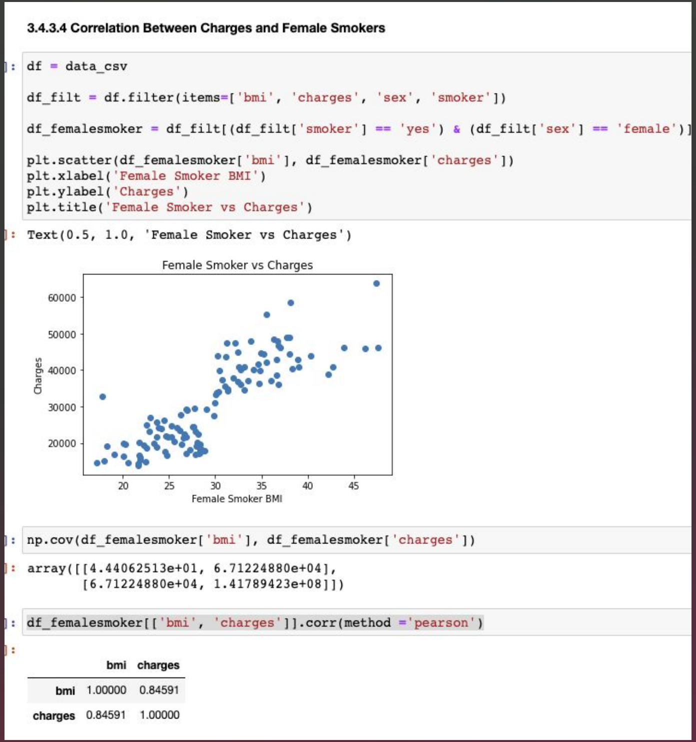
```
Text(0.5, 1.0, 'Male Smoker vs Charges')
```

```
np.cov(df_malesmoker['bmi'], df_malesmoker['charges'])
```

```
array([[3.54266376e+01, 5.12995461e+04],
       [5.12995461e+04, 1.25499834e+08]])
```

```
df_malesmoker[['bmi', 'charges']].corr(method ='pearson')
```

	bmi	charges
bmi	1.000000	0.769355
charges	0.769355	1.000000



3.4.3.4 Correlation Between Charges and Female Smokers

```
df = data_csv

df_filt = df.filter(items=['bmi', 'charges', 'sex', 'smoker'])

df_femalesmoker = df_filt[(df_filt['smoker'] == 'yes') & (df_filt['sex'] == 'female')]

plt.scatter(df_femalesmoker['bmi'], df_femalesmoker['charges'])
plt.xlabel('Female Smoker BMI')
plt.ylabel('Charges')
plt.title('Female Smoker vs Charges')
```

```
Text(0.5, 1.0, 'Female Smoker vs Charges')
```

```
np.cov(df_femalesmoker['bmi'], df_femalesmoker['charges'])
```

```
array([[4.44062513e+01, 6.71224880e+04],
       [6.71224880e+04, 1.41789423e+08]])
```

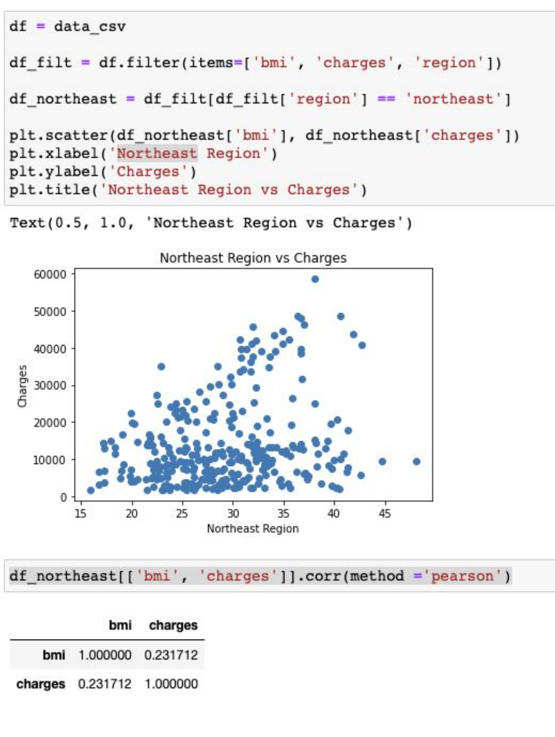
```
df_femalesmoker[['bmi', 'charges']].corr(method ='pearson')
```

	bmi	charges
bmi	1.00000	0.84591
charges	0.84591	1.00000

3.4 Correlation Analysis

Question 5	**Correlation between charges and region**
Method	● Matplotlib scatter plot ● .corr() to find correlation with pearson method
Results	● r - charges, northwest: 0.181 ● r - charges, northeast: 0.232 ● r - charges, southwest: 0.222 ● r - charges, southeast: 0.142
Findings	● Correlation coefficient between each region with charges are all less than 0.3. ● Hence, there is not much increase in charges when regions are changed.

3.4 Correlation Analysis - Method Screenshot

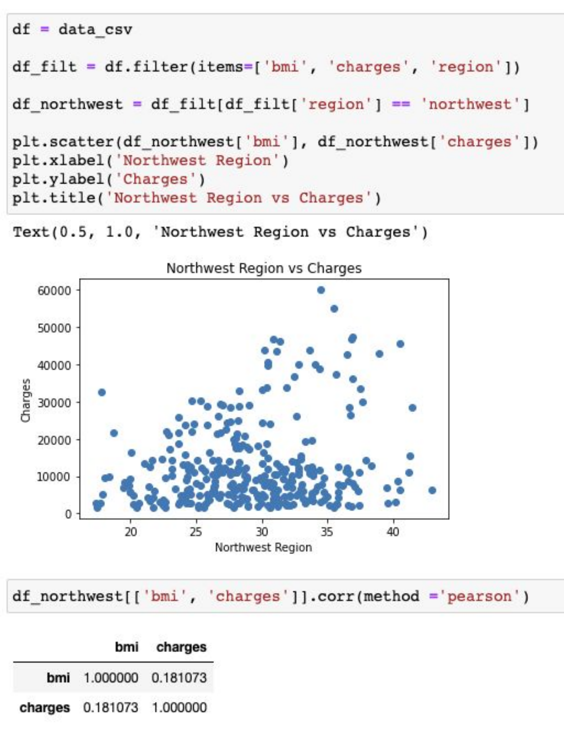


```
df = data_csv

df_filt = df.filter(items=['bmi', 'charges', 'region'])

df_northeast = df_filt[df_filt['region'] == 'northeast']

plt.scatter(df_northeast['bmi'], df_northeast['charges'])
plt.xlabel('Northeast Region')
plt.ylabel('Charges')
plt.title('Northeast Region vs Charges')
```

```
Text(0.5, 1.0, 'Northeast Region vs Charges')
```

```
df_northeast[['bmi', 'charges']].corr(method ='pearson')
```

	bmi	charges
bmi	1.000000	0.231712
charges	0.231712	1.000000



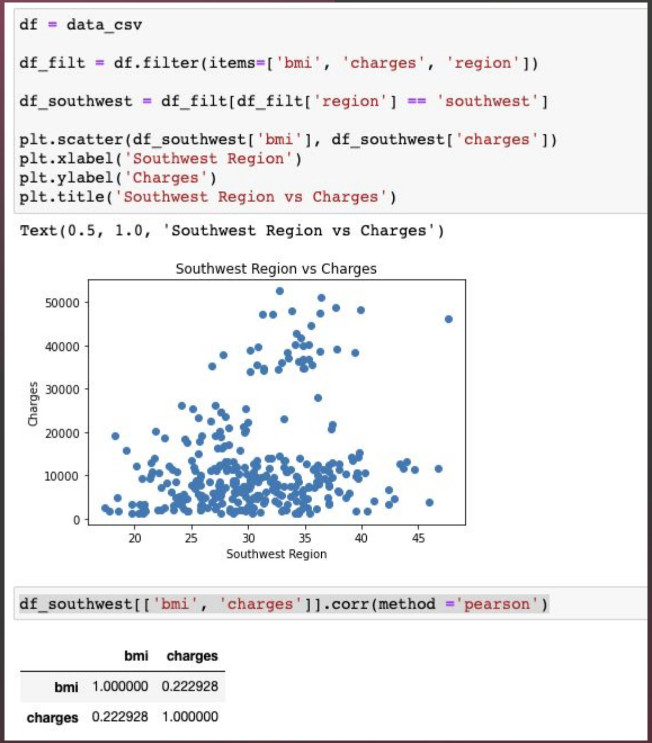
```
df = data_csv

df_filt = df.filter(items=['bmi', 'charges', 'region'])

df_northwest = df_filt[df_filt['region'] == 'northwest']

plt.scatter(df_northwest['bmi'], df_northwest['charges'])
plt.xlabel('Northwest Region')
plt.ylabel('Charges')
plt.title('Northwest Region vs Charges')
```

```
Text(0.5, 1.0, 'Northwest Region vs Charges')
```

```
df_northwest[['bmi', 'charges']].corr(method ='pearson')
```

	bmi	charges
bmi	1.000000	0.181073
charges	0.181073	1.000000

3.4 Correlation Analysis - Method Screenshot

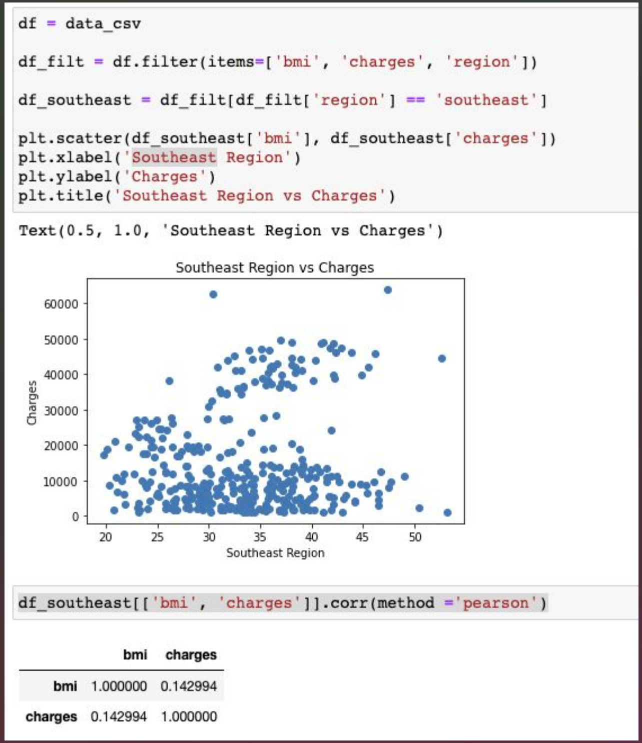


```
df = data_csv

df_filt = df.filter(items=['bmi', 'charges', 'region'])

df_southwest = df_filt[df_filt['region'] == 'southwest']

plt.scatter(df_southwest['bmi'], df_southwest['charges'])
plt.xlabel('Southwest Region')
plt.ylabel('Charges')
plt.title('Southwest Region vs Charges')
```

```
Text(0.5, 1.0, 'Southwest Region vs Charges')
```

```
df_southwest[['bmi', 'charges']].corr(method ='pearson')
```

	bmi	charges
bmi	1.000000	0.222928
charges	0.222928	1.000000



```
df = data_csv

df_filt = df.filter(items=['bmi', 'charges', 'region'])

df_southeast = df_filt[df_filt['region'] == 'southeast']

plt.scatter(df_southeast['bmi'], df_southeast['charges'])
plt.xlabel('Southeast Region')
plt.ylabel('Charges')
plt.title('Southeast Region vs Charges')
```

```
Text(0.5, 1.0, 'Southeast Region vs Charges')
```

```
df_southeast[['bmi', 'charges']].corr(method ='pearson')
```

	bmi	charges
bmi	1.000000	0.142994
charges	0.142994	1.000000

3.5 Hypothesis Testing

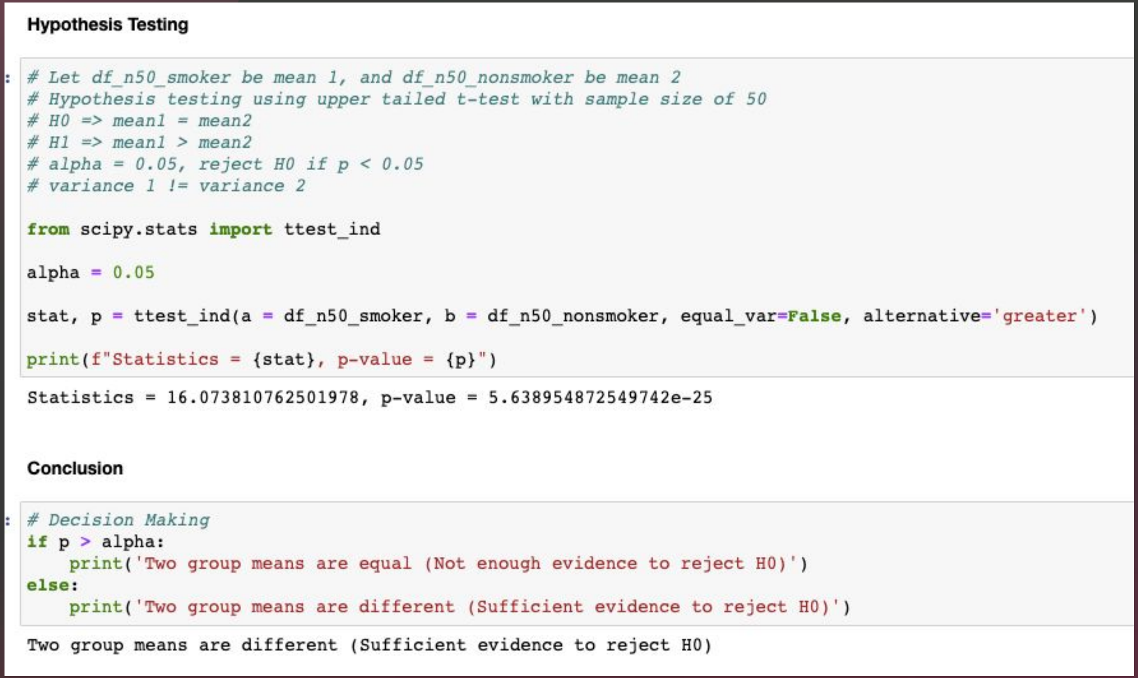
Question selected:

1. Is the mean charges for smoker greater than non smoker?
2. Is the mean charges of individuals with BMI > 25 higher than BMI < 25?
3. Is the mean charges of male greater than female?

3.5 Hypothesis Testing

Question 1	**Is the mean charges for smoker greater than non smoker?**
Assumptions and Methods	Assumption: 1. Sampling 50 data points from smoker and non-smoker 2. Data are normally distributed Method: 1. Upper tailed hypothesis testing using t-test: - H0: mean charges of smokers = mean charges of non-smokers - H1: mean charges of smokers > mean charges of non-smokers 2. Significance level = 0.05 3. Reject null hypothesis if p < alpha
Results	Statistics = 16.073810762501978, p-value = 5.638954872549742e-25
Findings	● As p-value < alpha, there is sufficient evidence to reject null hypothesis. ● Mean charges of smokers are likely greater than non-smokers

3.5 Hypothesis Testing - Screenshot Method



Hypothesis Testing

```
# Let df_n50_smoker be mean 1, and df_n50_nonsmoker be mean 2
# Hypothesis testing using upper tailed t-test with sample size of 50
# H0 => mean1 = mean2
# H1 => mean1 > mean2
# alpha = 0.05, reject H0 if p < 0.05
# variance 1 != variance 2

from scipy.stats import ttest_ind

alpha = 0.05

stat, p = ttest_ind(a = df_n50_smoker, b = df_n50_nonsmoker, equal_var=False, alternative='greater')

print(f"Statistics = {stat}, p-value = {p}")
```

```
Statistics = 16.073810762501978, p-value = 5.638954872549742e-25
```

Conclusion

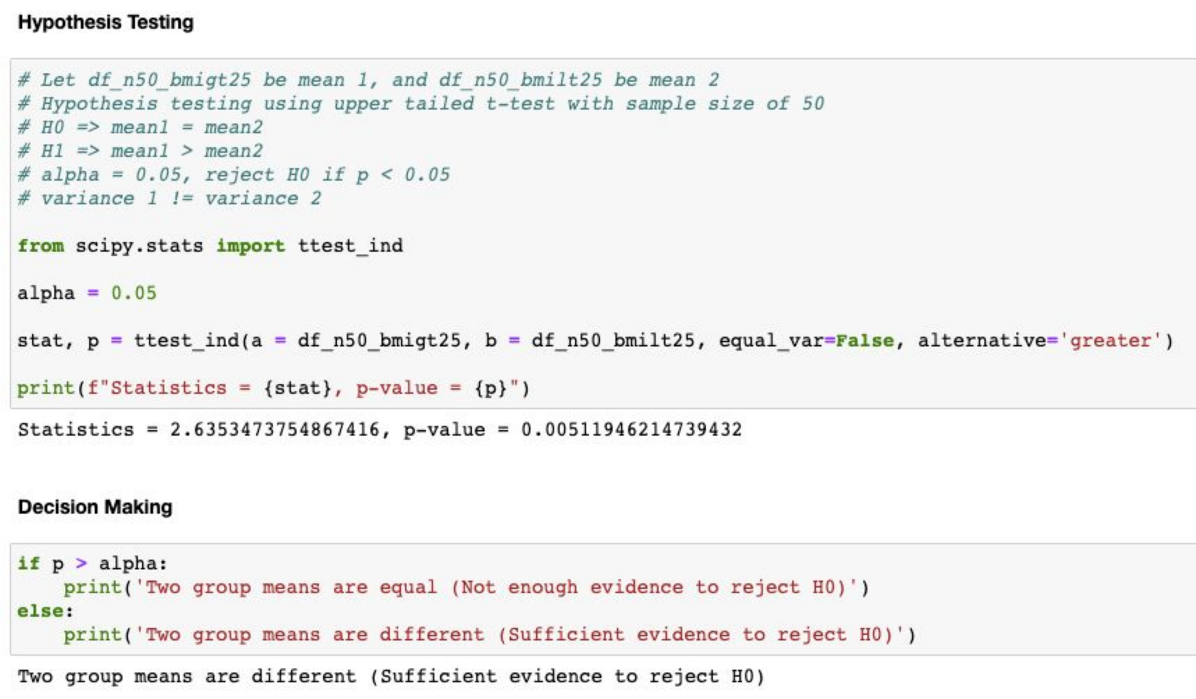
```
# Decision Making
if p > alpha:
    print('Two group means are equal (Not enough evidence to reject H0)')
else:
    print('Two group means are different (Sufficient evidence to reject H0)')
```

```
Two group means are different (Sufficient evidence to reject H0)
```

3.5 Hypothesis Testing

Question 2	**Is the mean charges of individuals with BMI > 25 higher than BMI < 25?**
Assumptions and Methods	Assumption: 1. Sampling 50 data points from BMI > 25 and BMI < 25 2. Data are normally distributed Method: 1. Upper tailed hypothesis testing using t-test: - H0: mean charges for BMI > 25 = mean charges for BMI < 25 - H1: mean charges for BMI > 25 > mean charges for BMI < 25 2. Significance level = 0.05 3. Reject null hypothesis if p < alpha
Results	Statistics = 2.6353473754867416, p-value = 0.00511946214739432
Findings	● As p-value < alpha, there is sufficient evidence to reject null hypothesis. ● Mean charges for BMI > 25 are likely greater than BMI < 25

3.5 Hypothesis Testing - Screenshot Method



Hypothesis Testing

```
# Let df_n50_bmigt25 be mean 1, and df_n50_bmilt25 be mean 2
# Hypothesis testing using upper tailed t-test with sample size of 50
# H0 => mean1 = mean2
# H1 => mean1 > mean2
# alpha = 0.05, reject H0 if p < 0.05
# variance 1 != variance 2

from scipy.stats import ttest_ind

alpha = 0.05

stat, p = ttest_ind(a = df_n50_bmigt25, b = df_n50_bmilt25, equal_var=False, alternative='greater')

print(f"Statistics = {stat}, p-value = {p}")
```

```
Statistics = 2.6353473754867416, p-value = 0.005119462147394322
```

Decision Making

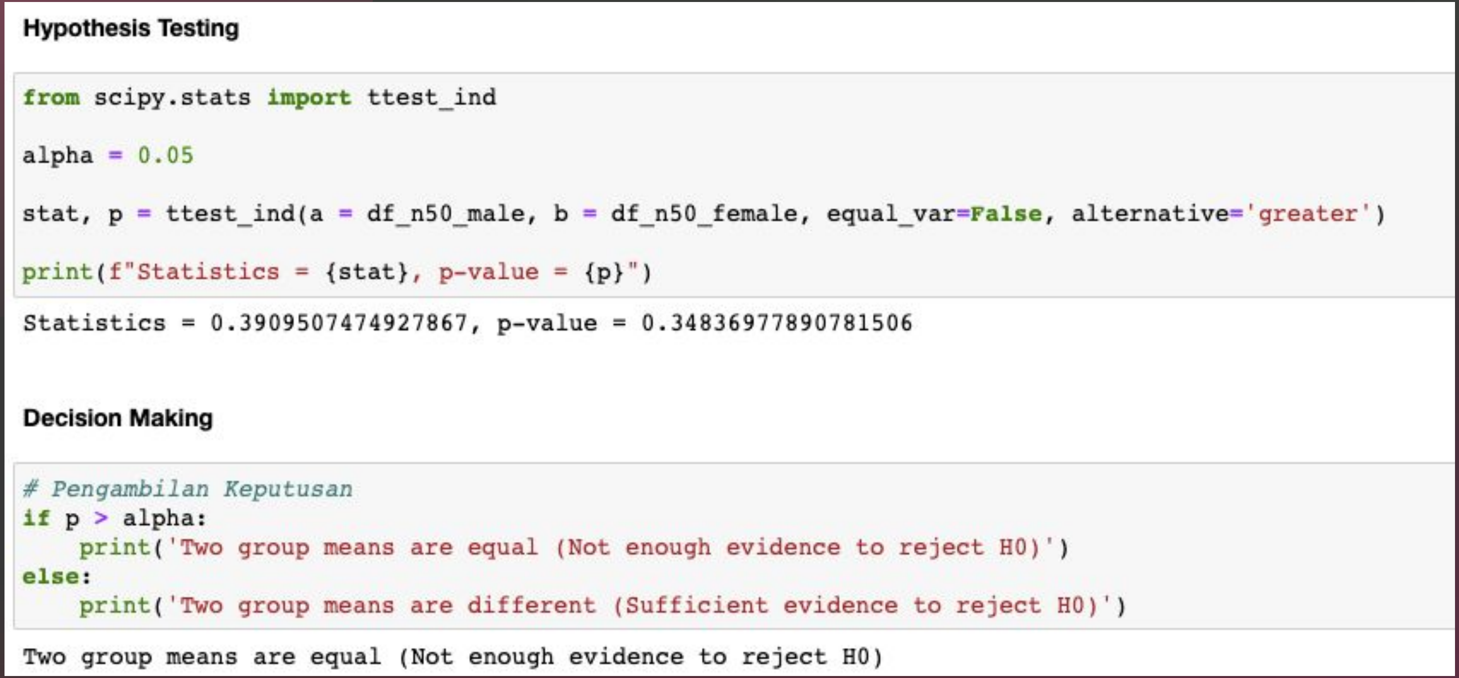
```
if p > alpha:
    print('Two group means are equal (Not enough evidence to reject H0)')
else:
    print('Two group means are different (Sufficient evidence to reject H0)')
```

```
Two group means are different (Sufficient evidence to reject H0)
```

3.5 Hypothesis Testing

Question 3	**Is the mean charges of male greater than female?**
Assumptions and Methods	Assumption: 1. Sampling 50 data points from male and female 2. Data are normally distributed Method: 1. Upper tailed hypothesis testing using t-test: - H0: mean charges for male = mean charges for female - H1: mean charges for male > mean charges for female 2. Significance level = 0.05 3. Reject null hypothesis if p < alpha
Results	Statistics = 0.3909507474927867, p-value = 0.34836977890781506
Findings	● As p-value > alpha, there is no sufficient evidence to reject null hypothesis. ● Mean charges for female are likely to be equal to male.

3.5 Hypothesis Testing - Screenshot Method



Hypothesis Testing

```
from scipy.stats import ttest_ind

alpha = 0.05

stat, p = ttest_ind(a = df_n50_male, b = df_n50_female, equal_var=False, alternative='greater')

print(f"Statistics = {stat}, p-value = {p}")
```

```
Statistics = 0.3909507474927867, p-value = 0.34836977890781506
```

Decision Making

```
# Pengambilan Keputusan
if p > alpha:
    print('Two group means are equal (Not enough evidence to reject H0)')
else:
    print('Two group means are different (Sufficient evidence to reject H0)')
```

```
Two group means are equal (Not enough evidence to reject H0)
```

3.6 Finding a good model to predict medical charges

Background:

- Health insurance companies need to collect more premiums than they spend on medical care to make a profit.

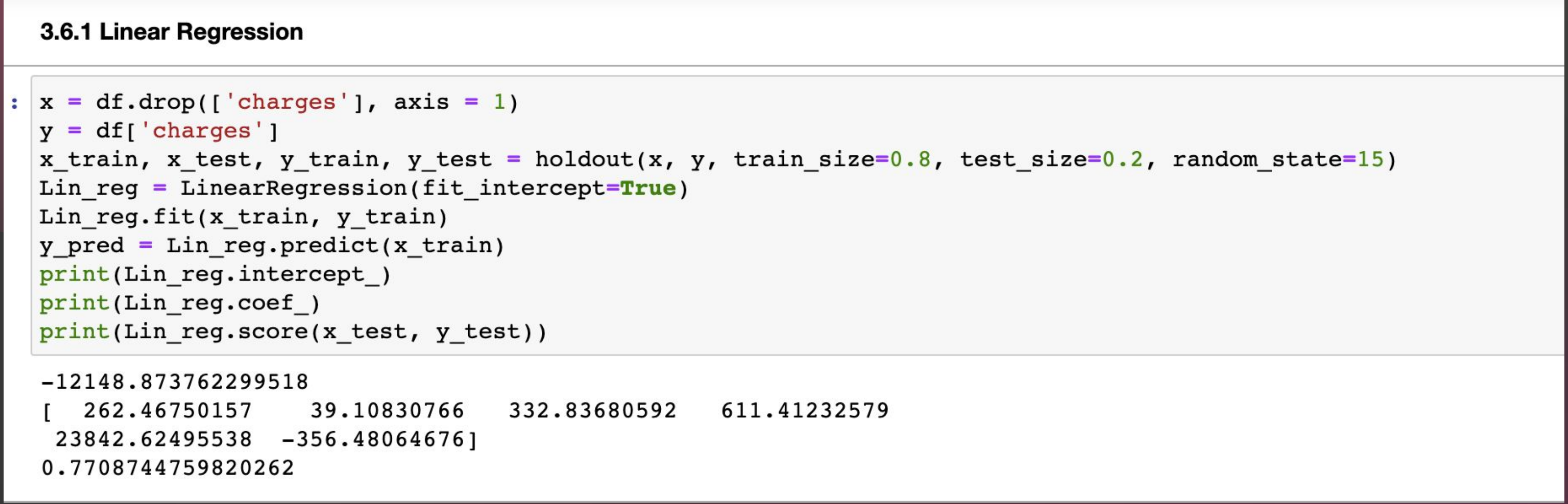
Problem:

- It is difficult to estimate medical expenses because costly conditions are rare and seemingly random. However, certain conditions are more common among certain population segments (e.g., lung cancer among smokers)

Solution:

- To compare between linear and polynomial regression and find the best models to forecast medical expenses for their insured population.
- Patient data is analyzed to estimate the average medical care expenses for these population segments and can be used to set premiums higher or lower depending on the expected treatment costs.

3.6.1 Linear Regression - Model

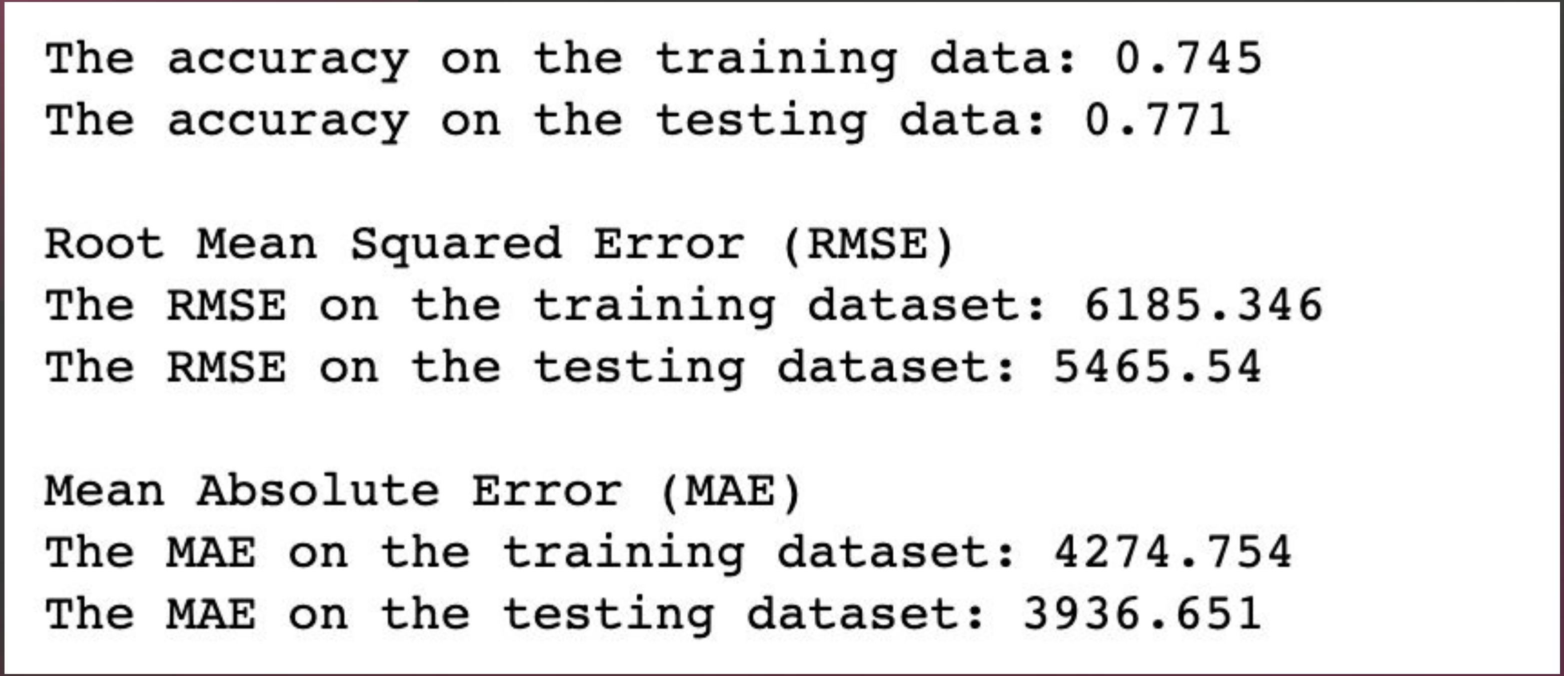


3.6.1 Linear Regression

```
x = df.drop(['charges'], axis = 1)
y = df['charges']
x_train, x_test, y_train, y_test = holdout(x, y, train_size=0.8, test_size=0.2, random_state=15)
Lin_reg = LinearRegression(fit_intercept=True)
Lin_reg.fit(x_train, y_train)
y_pred = Lin_reg.predict(x_train)
print(Lin_reg.intercept_)
print(Lin_reg.coef_)
print(Lin_reg.score(x_test, y_test))
```

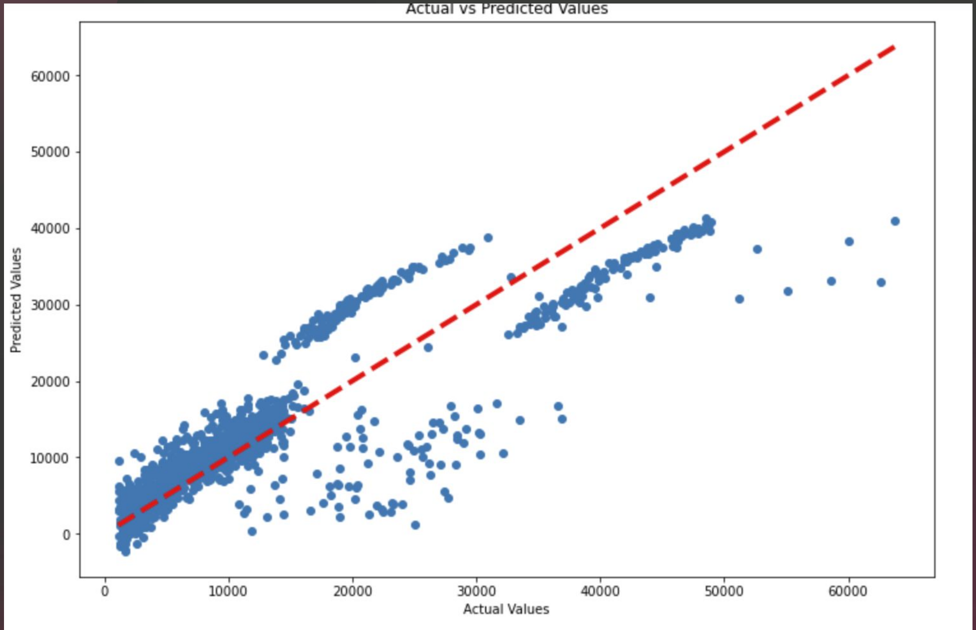
```
-12148.873762299518
[  262.46750157    39.10830766   332.83680592   611.41232579
 23842.62495538  -356.48064676]
0.7708744759820262
```

3.6.1 Linear Regression - Accuracy

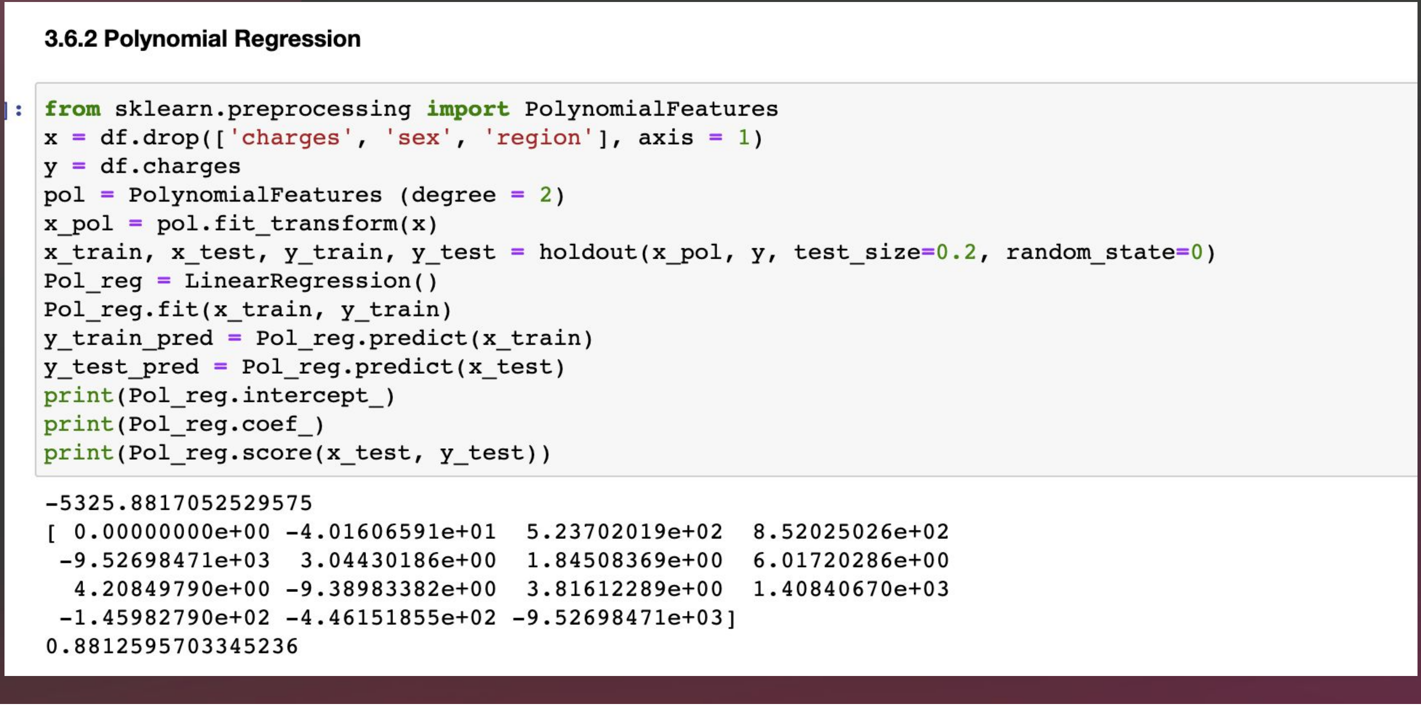


```
The accuracy on the training data: 0.745
The accuracy on the testing data: 0.771

Root Mean Squared Error (RMSE)
The RMSE on the training dataset: 6185.346
The RMSE on the testing dataset: 5465.54

Mean Absolute Error (MAE)
The MAE on the training dataset: 4274.754
The MAE on the testing dataset: 3936.651
```

3.6.1 Linear Regression - Plot



3.6.2 Polynomial Regression - Model

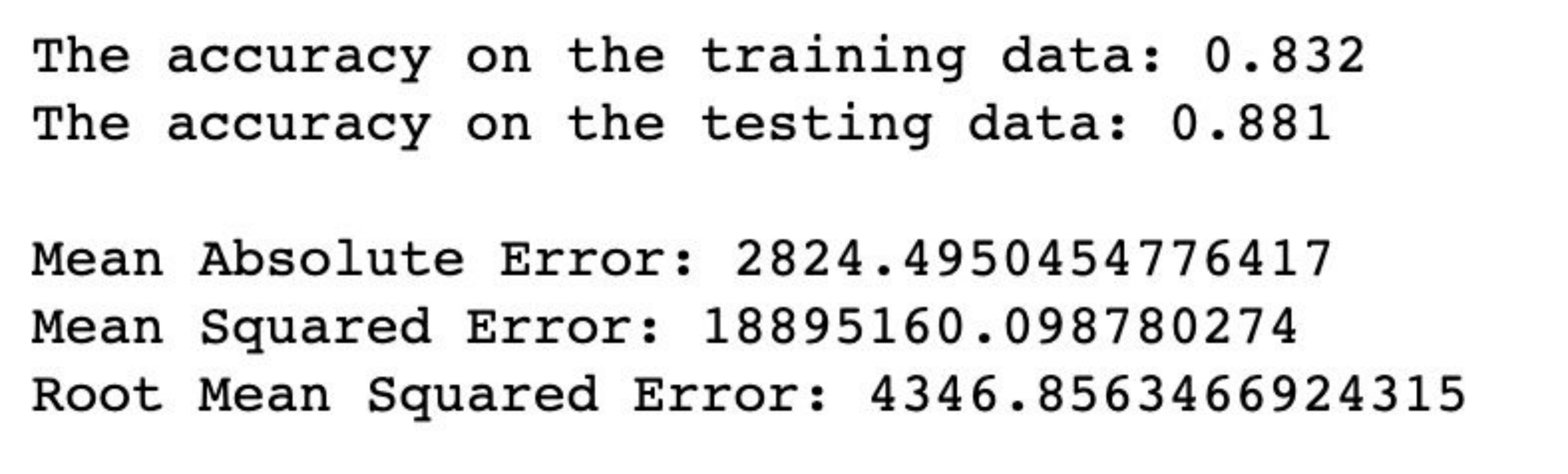


3.6.2 Polynomial Regression

```
from sklearn.preprocessing import PolynomialFeatures
x = df.drop(['charges', 'sex', 'region'], axis = 1)
y = df.charges
pol = PolynomialFeatures (degree = 2)
x_pol = pol.fit_transform(x)
x_train, x_test, y_train, y_test = holdout(x_pol, y, test_size=0.2, random_state=0)
Pol_reg = LinearRegression()
Pol_reg.fit(x_train, y_train)
y_train_pred = Pol_reg.predict(x_train)
y_test_pred = Pol_reg.predict(x_test)
print(Pol_reg.intercept_)
print(Pol_reg.coef_)
print(Pol_reg.score(x_test, y_test))
```

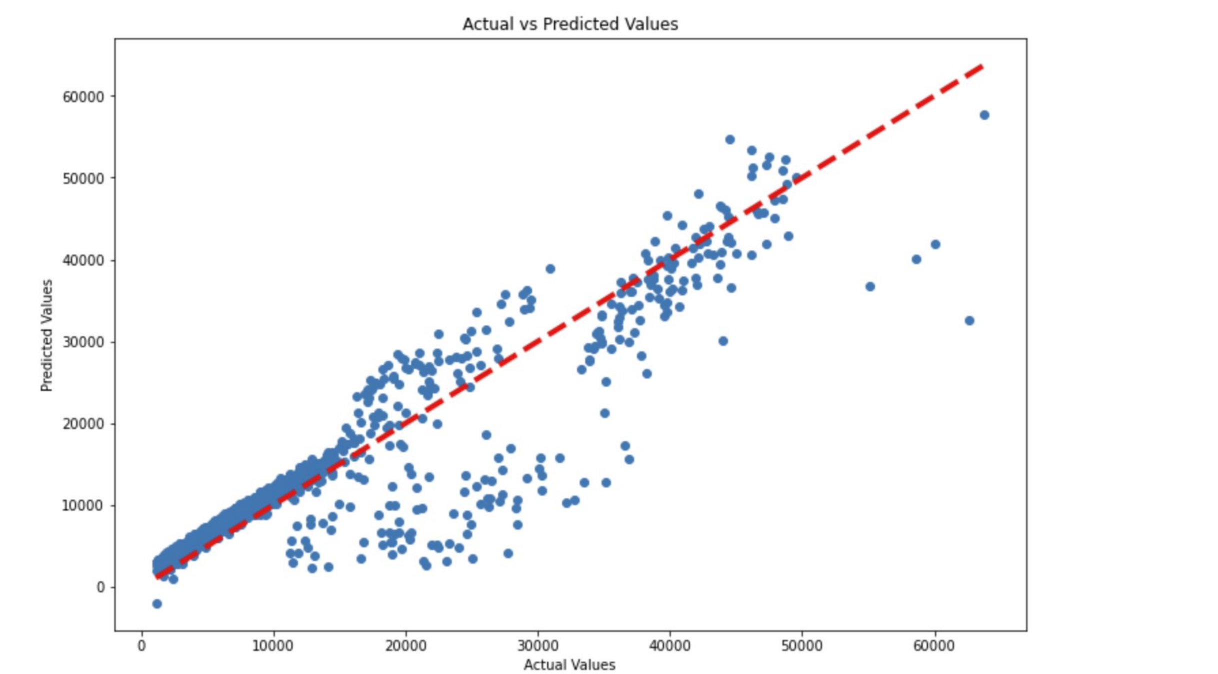
```
-5325.8817052529575
[ 0.00000000e+00 -4.01606591e+01  5.23702019e+02  8.52025026e+02
 -9.52698471e+03  3.04430186e+00  1.84508369e+00  6.01720286e+00
  4.20849790e+00 -9.38983382e+00  3.81612289e+00  1.40840670e+03
 -1.45982790e+02 -4.46151855e+02 -9.52698471e+03]
0.8812595703345236
```

3.6.2 Polynomial Regression - Accuracy



```
The accuracy on the training data: 0.832
The accuracy on the testing data: 0.881

Mean Absolute Error: 2824.4950454776417
Mean Squared Error: 18895160.098780274
Root Mean Squared Error: 4346.8563466924315
```

3.6.2 Polynomial Regression - Plot



Comparing Linear vs Polynomial Regression

Since the accuracy of Polynomial regression is higher than linear regression, polynomial turns out to be a better model to predict medical charges.

04

Conclusion

Conclusion

- Since the accuracy of Polynomial regression is higher than linear regression, polynomial turns out to be a better model to predict medical charges.
- Other variables in this dataset (age, sex, children, bmi, residential area) have weak positive correlation with charges.
- However, when we zoom into smokers data we observe that there is a strong positive correlation with charges. Smokers tend to have higher charges as compared to non-smokers.
- Hence, smoking habits of potential clients for insurance providers should be considered more closely when assessing risk and pricing premium.

Appendix

Future Enhancements and References

Future Enhancements

- Look into other variables in more details: children, age, region
- Test other hypothesis to make the analysis more all-rounded
- Add more visualizations to present better findings and analysis
- Optimize coding and practice better clean code in jupyter notebook
- Try to use machine learning to predict the medical costs based on our current understanding on how variables relate to medical costs
- Explore other models and compare accuracy for predicting medical charges

Resources

- Matplotlib documentation
- Seaborn documentation
- Scipy documentation
- Pandas documentation
- CDC data on BMI classification
- Sckit-Learn : This library includes the implementation of various machine larning algorithms. With this library, we will perform all operations from building to evaluation of regression models using functions and classes in this library.
- Pacmann lecture materials and dataset

Thank You!

Let’s connect! 🤓👇🏻