Final Project – Probability Course

Nur Indah Pratiwi – Skill Upgrader Batch 10

Report Highlights

Nur Indah Pratiwi | Skill Upgrader Batch 10

Introduction

The insurance industry is one of the most competitive and less predictable business spheres. It is instantly related to the profile risk of the user.
Determining the insurance charges is a challenge in itself because it really depends on several factors, especially on users side.

Through this project, the variables that have a relationship with the user's health bill will be analyzed. From personal data containing variables of gender, age, place of residence, number of children, BMI value, and whether he smokes or not, it will be seen the relationship of each variable to the value of insurance charges.

Nur Indah Pratiwi | Skill Upgrader Batch 10

Dataset

Dataset used in this project using insurance dataset which has been provided by the academy to be thoroughly explored and analyzed using the questions that have been given as a guide in compiling this report.

This dataset consists of 1338 entries (non-null entries) with a total of 7 columns.

1 Age

Characteristics of the age of the user.

2 Sex

Characteristics of the gender of the user. Marked with female or male.

3 BMI

To give an overview BMI of each user. Body mass index (BMI) is a measure of body fat based on height and weight that applies to adult men and women.

4 Children

To provide an idea of how many children the user has as a consideration for their insurance charges.

Nur Indah Pratiwi |
Probability Course Projects 2022

Dataset

Dataset used in this project using insurance dataset which has been provided by the academy to be thoroughly explored and analyzed using the questions that have been given as a guide in compiling this report.

This dataset consists of 1338 entries (non-null entries) with a total of 7 columns.

5 Smoker

An overview that identifies the user whether he or she is a smoker or not. The contents of the column are marked with categories: smokers and non-smokers.

6 Region

This is a column that identifies where the user is from. Fill in the fields marked with categories: southwest, southeast, northwest, and northeast

7 Charges

This is a column that contains the insurance charges that must be paid by the user.

Nur Indah Pratiwi |
Probability Course Projects 2022

Descriptive Statistic Analysis

Q-1: Average Age of Users

Q-2: The Average BMI of Smokers

Q-3: Charges Variance of Smoker and Non-Smokers

Q-4: Average Age of Female & Male (Smokers)

Q-5: Average Charges for Smoker/Non

Q-6: Average Charges for Smoker/Non (BMI > 25)

Q-7: The Average BMI for Female and Male

Q-8: The Average BMI for Non-Smoker

Nur Indah Pratiwi | Skill Upgrader Batch 10

Average Age of Users

18–64

The age range in the data

39,21

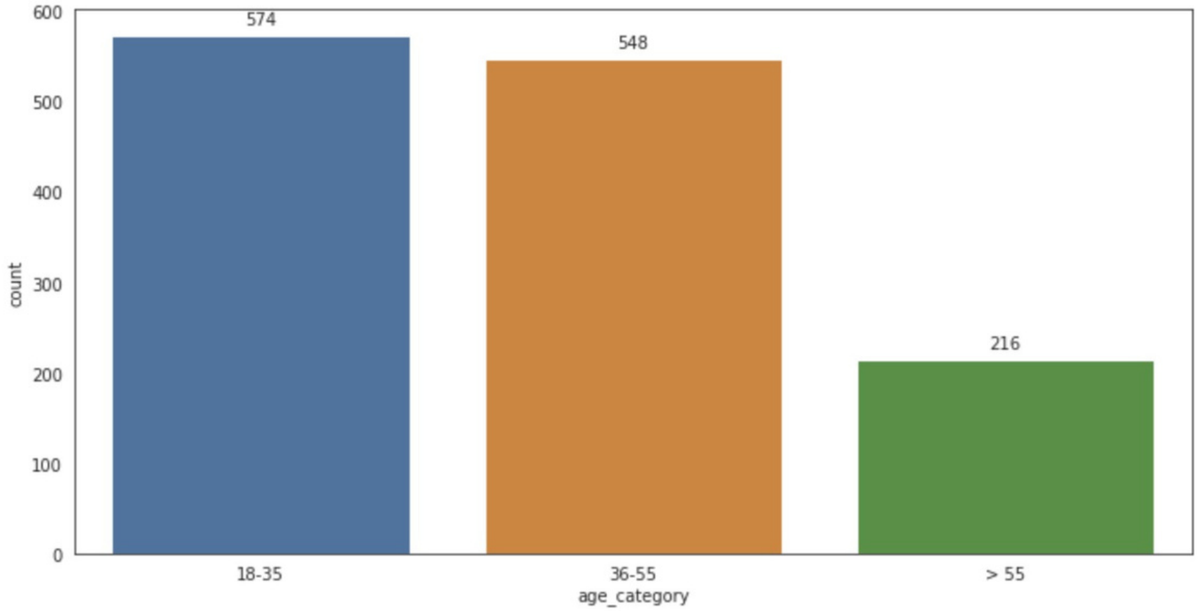
Average age from total data

The average age of the users is the result of adding up all the ages of the users divided by the total data. Age ranges are grouped into three categories. The 18–35 age category ranks first as much as 574 users.

```
print ("Average age of the data: {:.2f}" .format(df['age'].mean()))
```

Alternative

```
df.describe().transpose()
```



Nur Indah Pratiwi |
Probability Course Projects 2022

Average BMI of Smokers

30,66

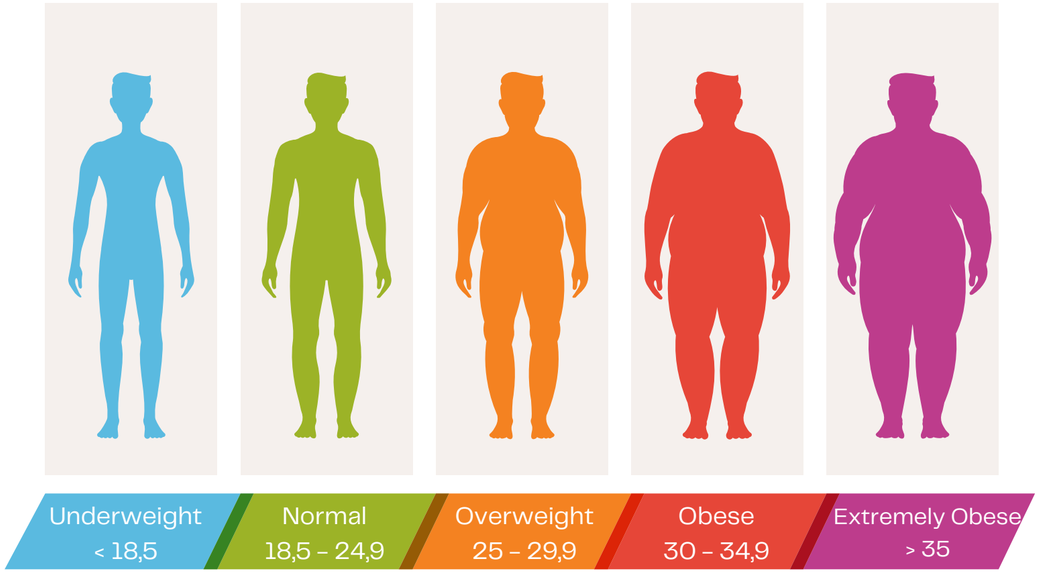
The BMI average in total

30,71

Average BMI of smokers

The average BMI of smokers is the result of adding up all the BMI of smokers divided by the total BMI of all users. The results show that the average BMI of smokers is in the obese group.

```
print ("Average BMI of smoker: {:.2f}"
.format(df.loc[df['smoker'] == 'yes']['bmi'].mean()))
```



Nur Indah Pratiwi |
Probability Course Projects 2022

Charges Variance of Smoker and Non-Smokers

Variance tells the degree of spread in the data set. The more spread the data, the larger the variance is in relation to the mean. The mean in this charges case is 13.270,42

```
sns.scatterplot(x=df['bmi'], y=df['charges'], hue=df['smoker'])
```

Calculate Variance of Smoker and Non-Smokers

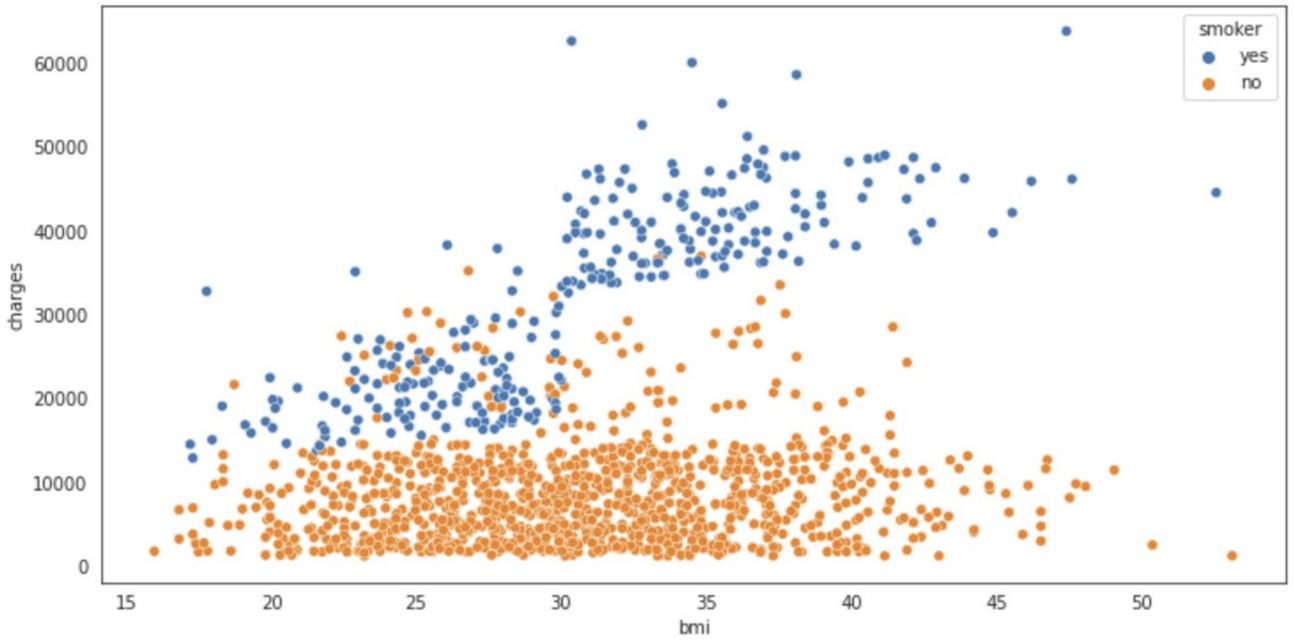
```
print (df.loc[df['smoker'] == 'yes']['charges'].var())
print (df.loc[df['smoker'] == 'no']['charges'].var())
```

133 207 311,20

Charges variance of smokers

35 925 420,49

Charges variance of non-smokers



Nur Indah Pratiwi |
Probability Course Projects 2022

Average Age of Male & Female (Smokers)

38,45

The average age of male (smokers)

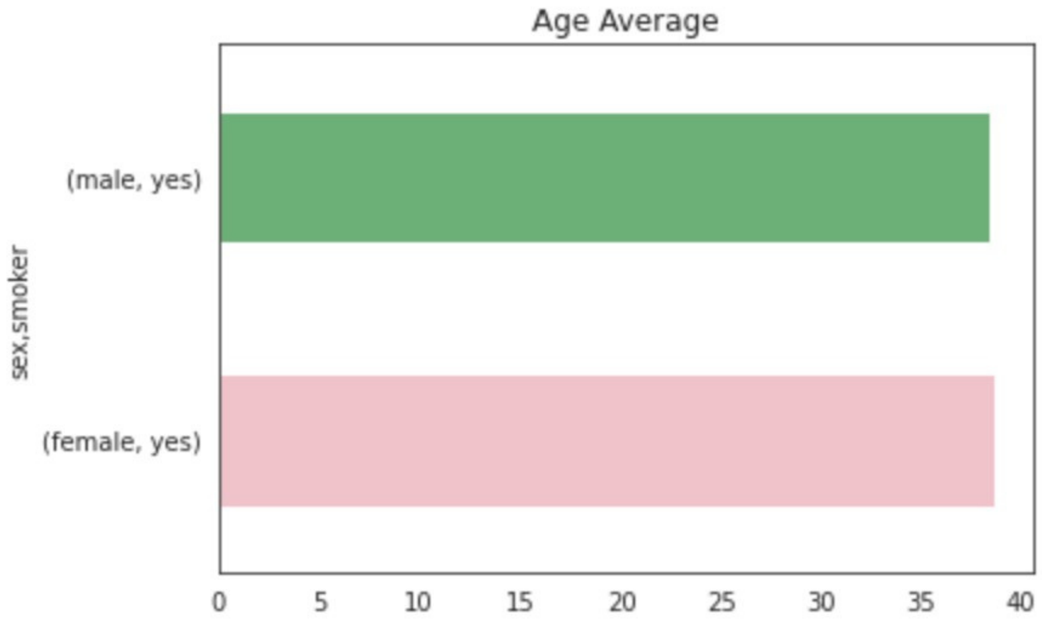
38,61

The average age of female (smokers)

After some aggregating with sex and smoker column, the result shows the age average of male and female smokers are shown the same results.

Calculate The Average Age of Male & Female (Smokers)

```
df_age_avg_smoker = (df[df['smoker'].isin(['yes'])]
                     .groupby(['sex','smoker'])
                     .agg(np.mean))
```



Nur Indah Pratiwi |
Probability Course Projects 2022

Average Charges for Smoker/Non

8.434,26

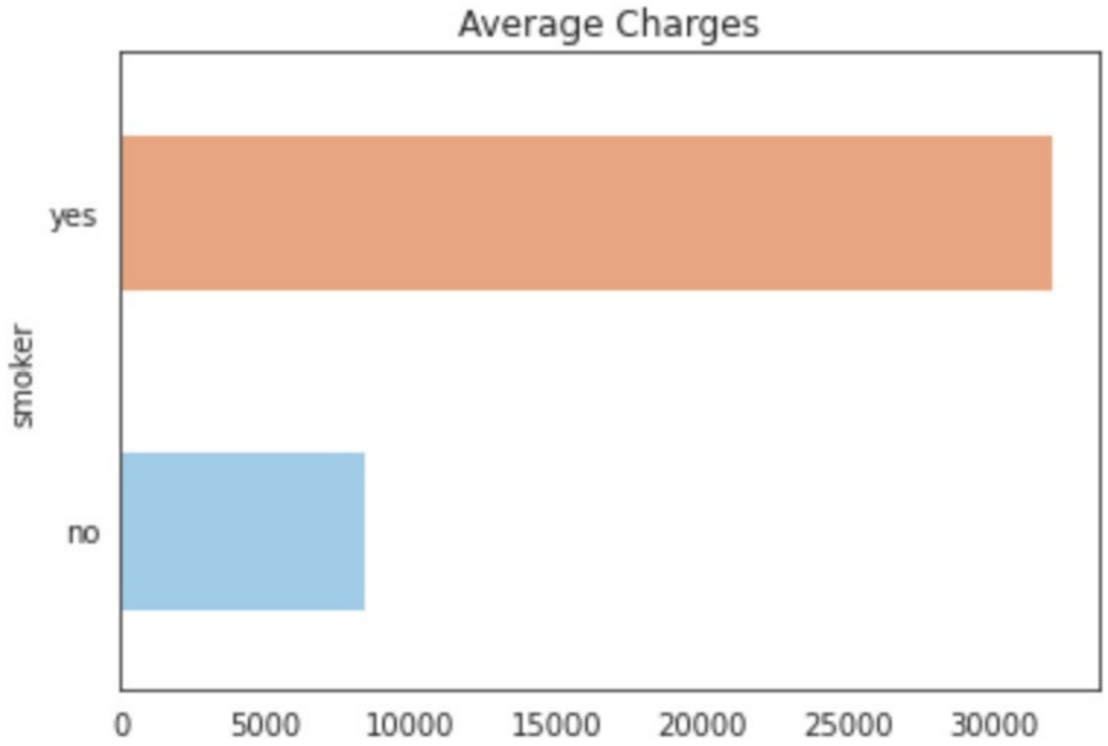
The average charges of non-smoker

32.050,23

The average charges of smoker

The result shows significant differences between charges average smokers and non-smokers. Non-smokers spend less than smokers. This makes sense.

```
df_charges_avg_smoker = (df[df['smoker'].isin(['yes', 'no'])]
                         .groupby(['smoker'])
                         .agg(np.mean))
```



Nur Indah Pratiwi |
Probability Course Projects 2022

Average Charges for Smoker/Non (BMI > 25)

8.629,59

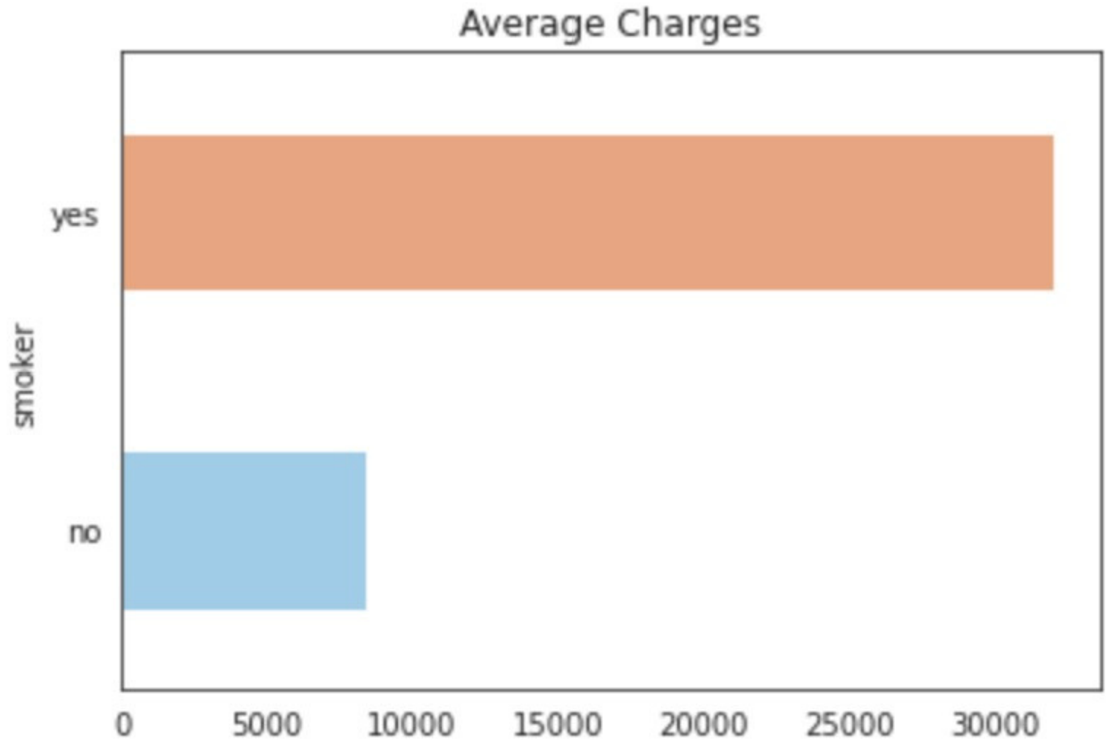
The average charges for non-smoker (BMI > 25)

35.116,91

The average charges for the smoker (BMI > 25)

This analysis wants to know specific average charges results for smokers and non with BMI more than 25. With this particular result that the average charges for the smoker with BMI > 25 show more significant charges than the others.

```
smoker_bmi = df.loc[(df['smoker'] == 'yes')&(df['bmi'] >25)]
non_smoker_bmi = df.loc[(df['smoker'] == 'no')&(df['bmi'] >25)]
smoker_bmi['charges'].mean()
non_smoker_bmi['charges'].mean()
```



Nur Indah Pratiwi |
Probability Course Projects 2022

The Average BMI for Female and Male

30,94

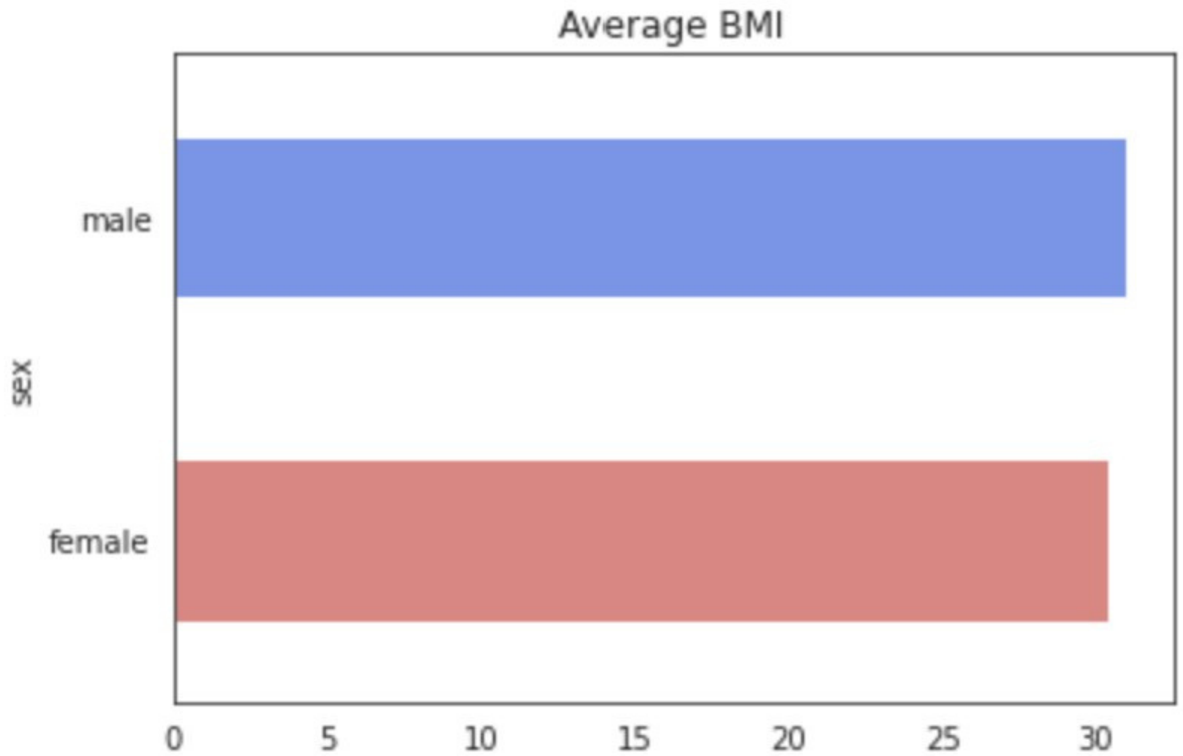
The average BMI for male

30,38

The average BMI for female

The result shows no significant differences between the average BMI for males and females. The difference is only 0.56

```
df_bmi_avg = (df[df['sex'].isin(['female', 'male'])]
              .groupby(['sex'])
              .agg(np.mean))
```



Nur Indah Pratiwi |
Probability Course Projects 2022

Average BMI of Non-Smokers

The results show that the average BMI of non-smokers is in the obese group whose results did not differ from the mean BMI of smokers. The difference is only 0.06

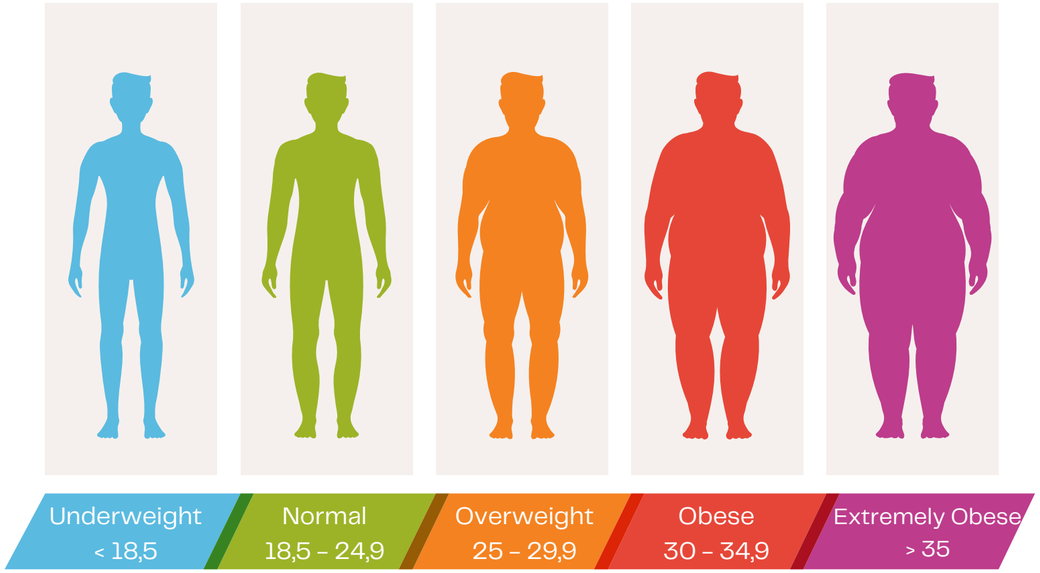
```
print ("Average BMI of smoker: {:.2f}"
.format(df.loc[df['smoker'] == 'no']['bmi'].mean()))
```

30,66

The average BMI in the data

30,65

Average BMI of non-smokers



Nur Indah Pratiwi |
Probability Course Projects 2022

Analysis

The dataset was taken from 1338 people with an average age of 39.21 where the age range in the dataset ranged from 18 – 64.

- This shows that the average person has an obese BMI where the normal BMI is in the range (of 18.5 – 24.9).
- The variance of charges is very far from the average.
- Smokers' charges are greater than non-smokers, smokers usually have other health problems involved.
- Meanwhile, if we take into account a BMI greater than 25, then the results of the health bill for smokers and non-smokers are much larger.

Nur Indah Pratiwi |
Probability Course Projects 2022

Highlight 1

- The average BMI (total) = 30,66
- BMI for smokers = 30,71
- BMI for non-smokers = 30,65
- BMI for male = 30,94
- BMI for female = 30,38

Highlight 2

- The average charges = 13.270.42
- Smoker variance = 133.207.311,20
- Non-smokers variance = 35.925.420,49

Highlight 3

- The average charges (smokers) = 32.050,23
- The average charges (non-smokers) = 8.434,26
- Smokers with BMI>25 = 35.116,91
- Non-Smokers with BMI>25 = 8.629,59

Categoric Variable Analysis

Nur Indah Pratiwi | Skill Upgrader Batch 10

The Highest Charges by Gender

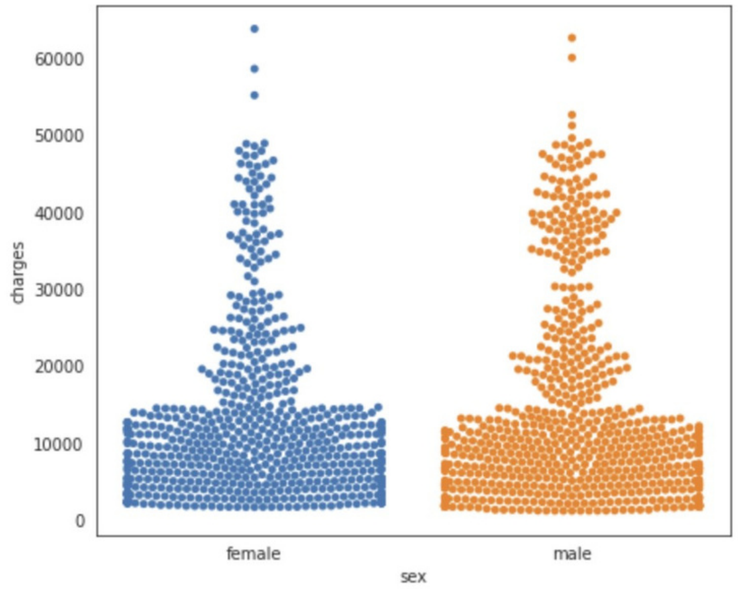
The highest charge in the female group is 63.770. this makes sense because when viewed from the data, she belongs to the elderly and smoking age group

```
sns.swarmplot(x='sex',y='charges', data=df)
print (df.loc[df['charges'].idxmax()])
```

```
age                54
sex            female
bmi             47.41
children            0
smoker            yes
region      southeast
charges     63770.42801
age_category      36-55
bmi_category     more_25
Name: 543, dtype: object
```

63.770,43

The highest charges for female

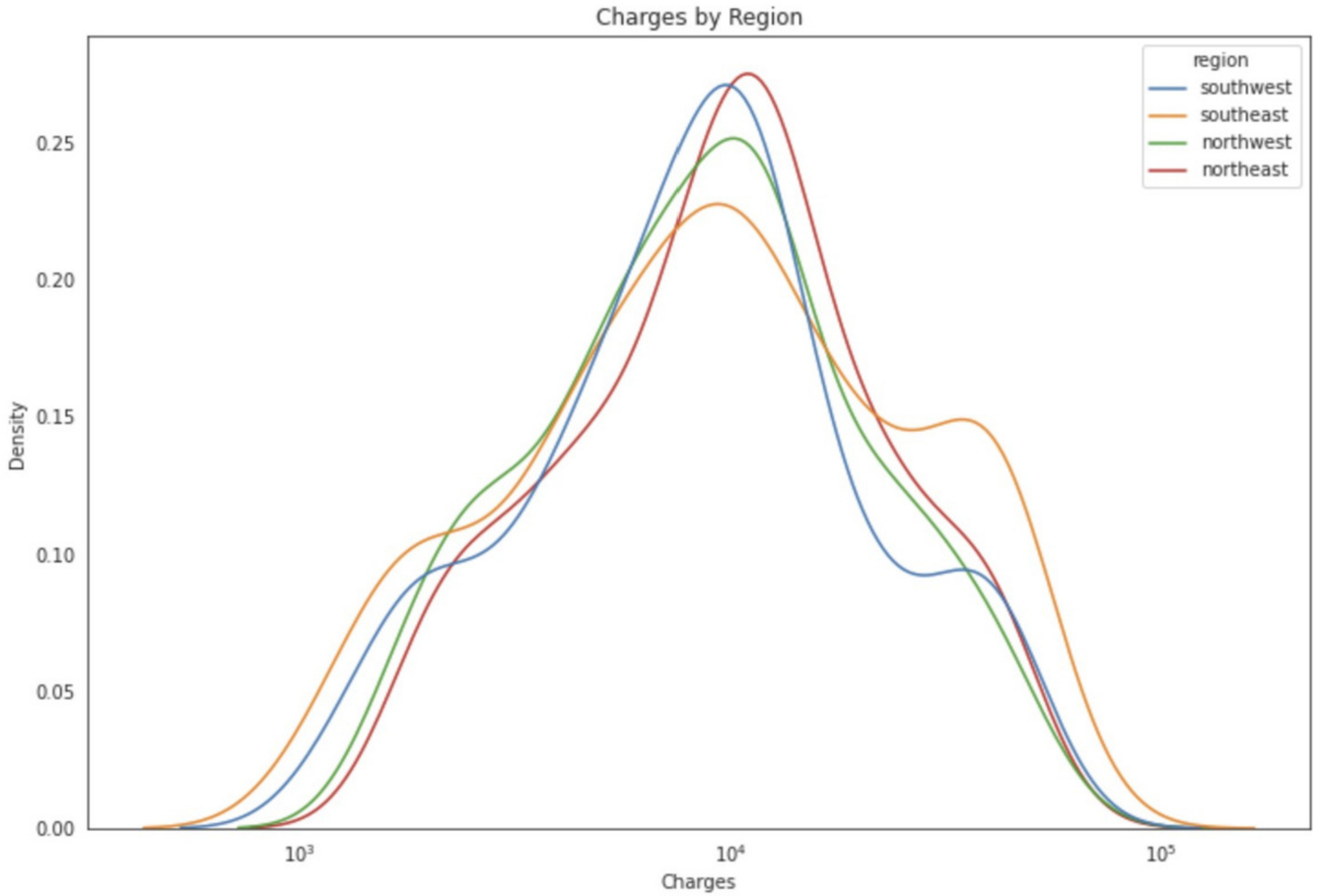


Nur Indah Pratiwi |
Probability Course Projects 2022

Charges Probability Distribution by Region

The picture shows that the highest probability of charges is in the region northeast.

```
sns.kdeplot(data=df, x="charges", hue='region')
```



Nur Indah Pratiwi |
Probability Course Projects 2022

Data Proportion by Region

From the results of the data visualization on the side, it shows that the largest proportion of data comes from southeast as many as 364 people. for southwest and northwest the proportion of data is the same as 325 people. As for northeast, it shows the least proportion of data, which is 324 people.

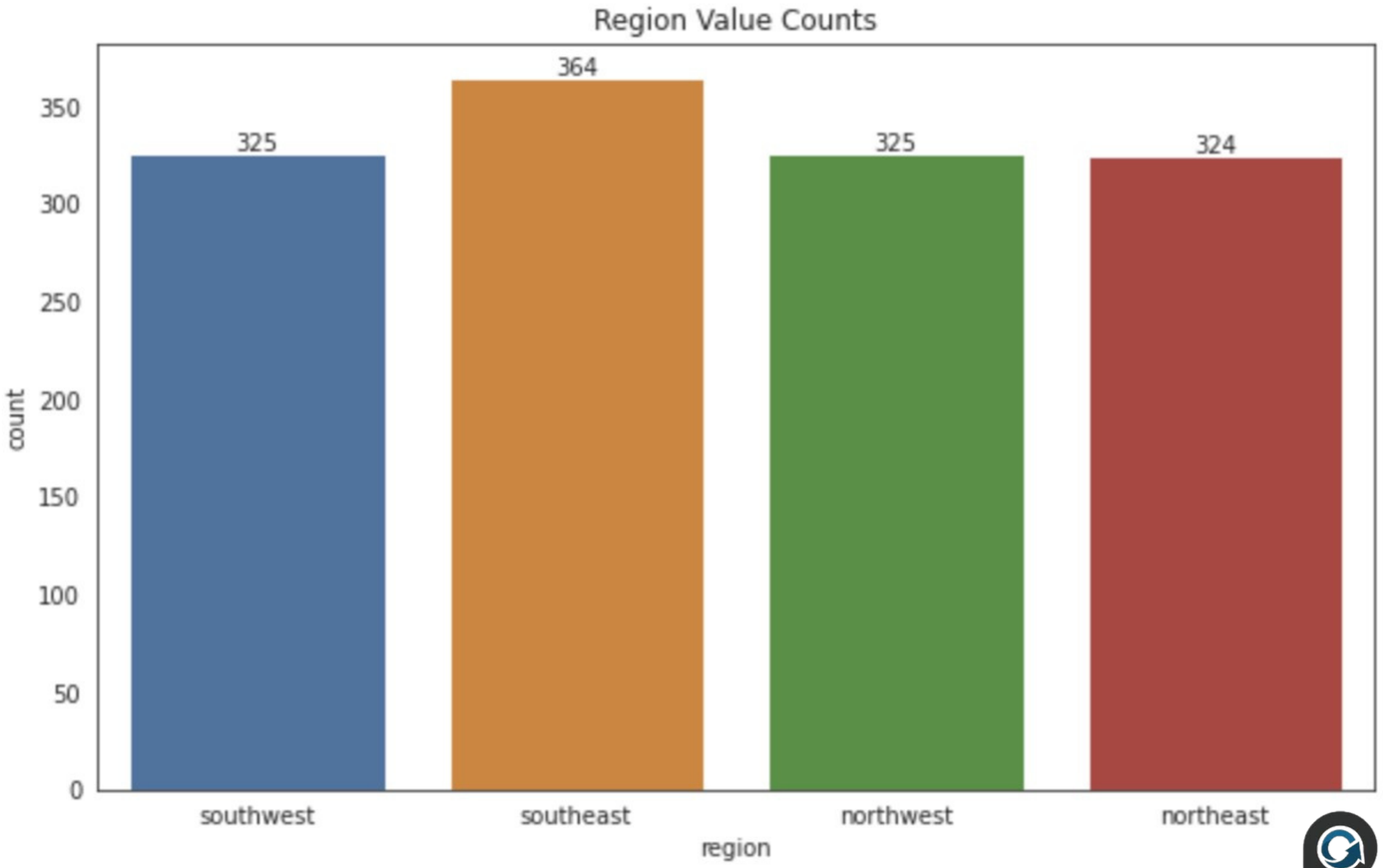
```
sns.countplot(x = 'region', data = df)
```

364

Most people come from southeast

324

people at least come from northeast



Nur Indah Pratiwi |
Probability Course Projects 2022

The Highest Data Proportion (smoker/non) by Region

273

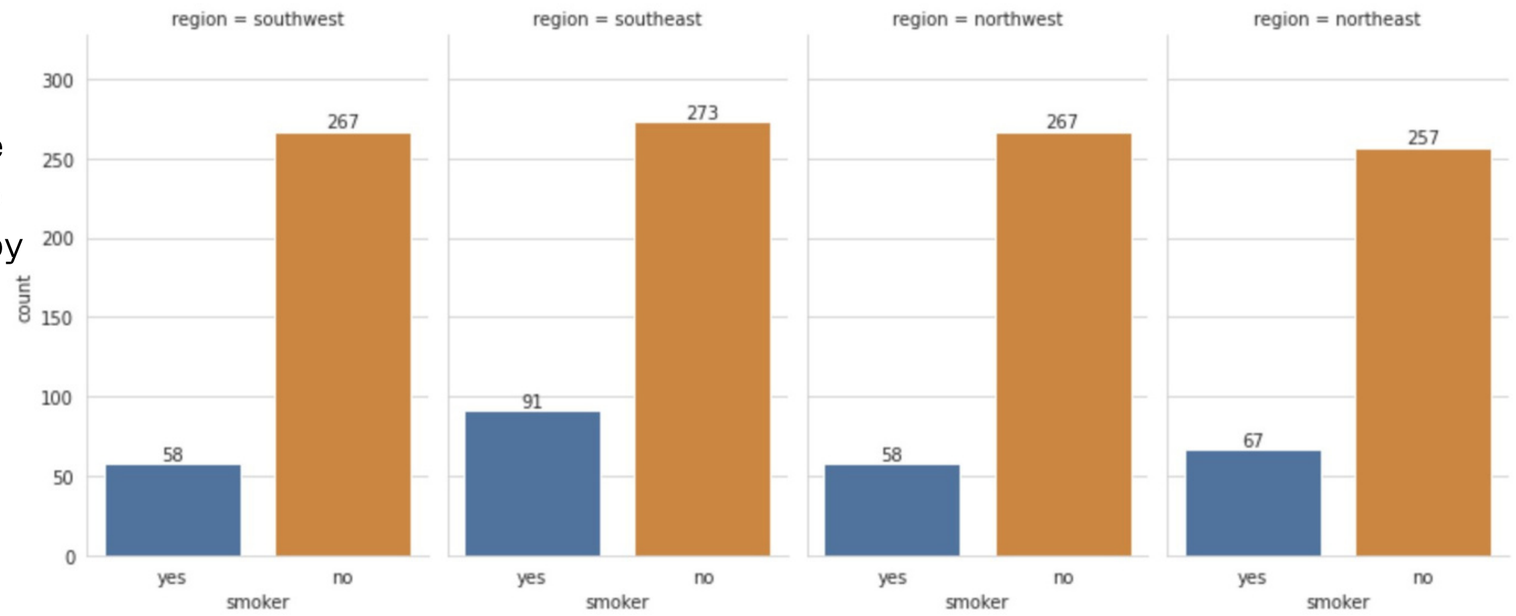
The highest data proportion for non-smokers (southeast)

91

The highest data proportion for smokers (southeast)

The highest data proportion for smokers is held by the southeast region with 91 people. At the same time, the highest data proportion for non-smokers is held also by southeast with 273 people.

```
sns.catplot(x = 'smoker', col = 'region',
            data = df, kind = 'count')
```



Nur Indah Pratiwi |
Probability Course Projects 2022

Probability of Smokers by Gender

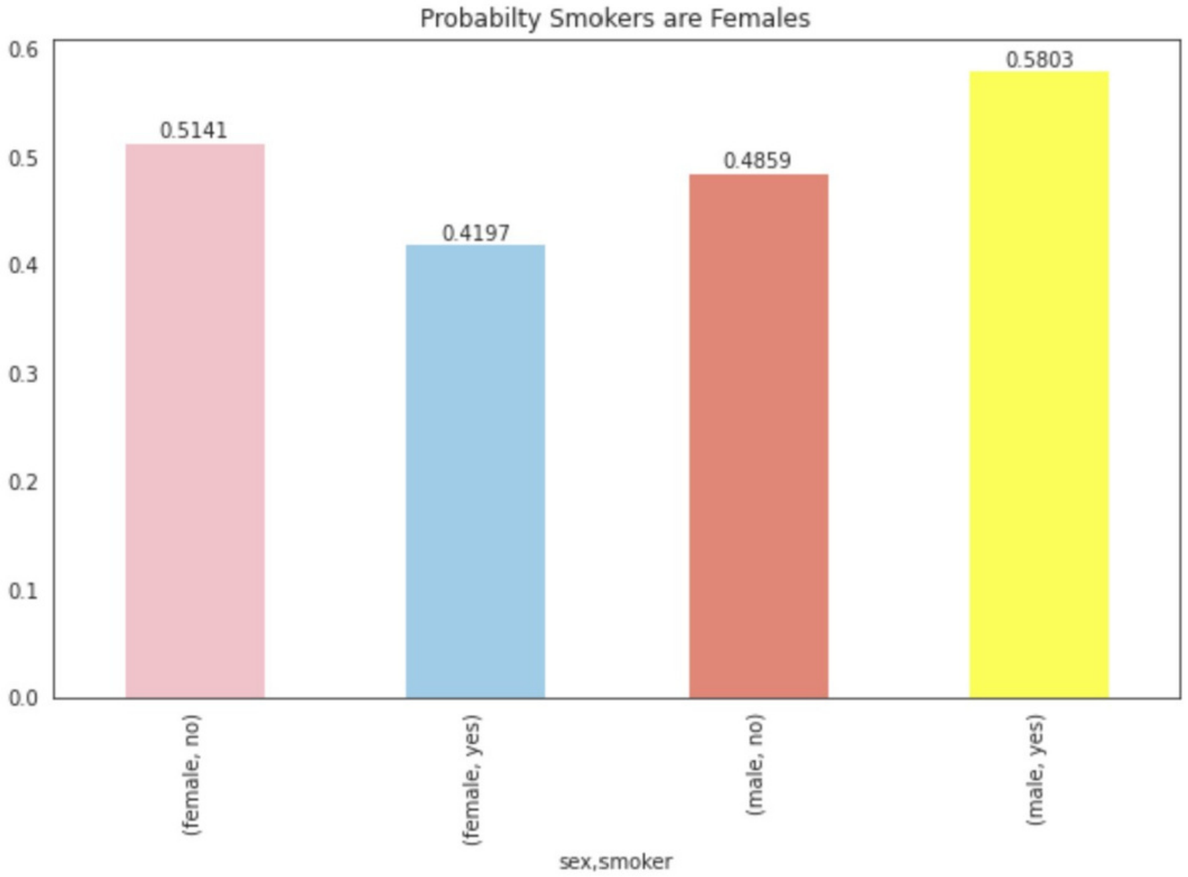
```
smoker_prob = df.groupby('smoker').size().div(len(df))

gender_prob = df.groupby(['sex', 'smoker']).size().div(len(df)).div(smoker_prob, axis=0, level='smoker')

gender_prob.plot(kind = "bar", y = "smoker", legend = False, color= ['pink', 'skyblue', 'salmon', 'yellow'], title = "Probability Smokers are Females")
```

0,4197

Probability of smokers are female

0,5803

Probability of smokers are male

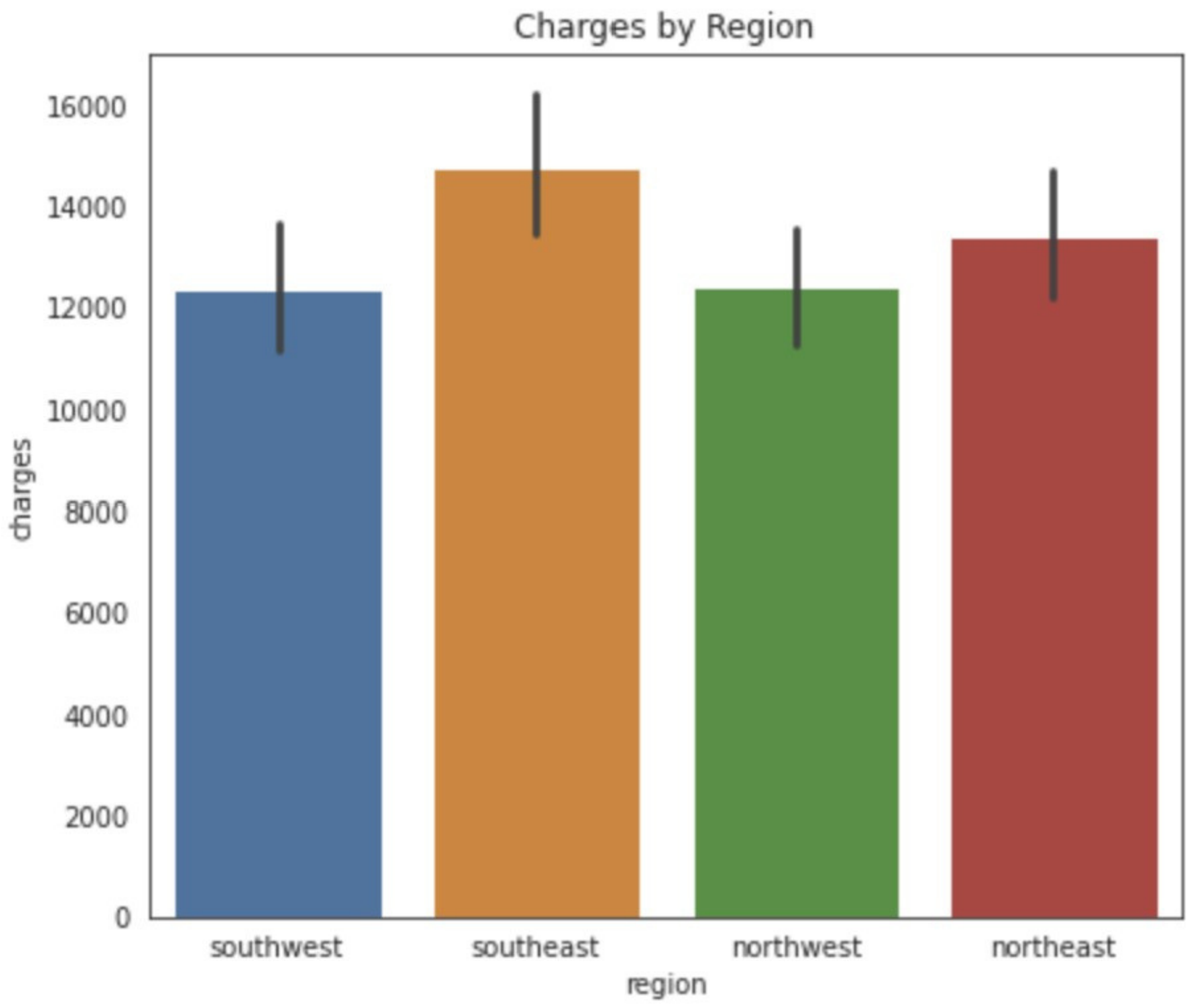


Nur Indah Pratiwi |
Probability Course Projects 2022

Distribution of Charges by Region

The Southeast region pays the highest charges. This could be due to the fact that the region has the highest number of smokers as we have seen earlier.

```
sns.barplot(x = 'region', y = 'charges', data = df)
```



Nur Indah Pratiwi |
Probability Course Projects 2022

Analysis

The highest health bill was 63,770.43 from women aged 54 years who were smokers.

- This is consistent with the analysis in the first section where smokers pay more bills than non-smokers.
- The southeast region pays the most charges, this is in line with the fact that in that region there is the most data on smokers.
- The probability that a smoker is a man is greater than a woman, because there are more male smokers than female smokers.

Highlight 1

- Probability of highest charges = Northeast
- Data proportion Southeast (most) = 364
- Data proportion Northeast (least) = 324

Highlight 2

Data proportion Southeast (most) = 364

- Smokers data = 91
- Non-smokers data = 273
- Charges distribution (most) = Southeast

Highlight 3

- Probability smokers are female = 0.4197
- Probability smokers are male = 0.5803
- Male smoker = 159
- Female smoker = 115

Nur Indah Pratiwi |
Probability Course Projects 2022

Continuous Variable Analysis

Q-1: Probability of (Charges | BMI)

Q-2: Probability of (Charges > 16.700 | BMI > 25)

Q-3: Probability of (Charges > 16.700 | Smoker)

Q-4: Probability of (Charges > 16.700 | BMI < 25)

Q-5: Probability of (Charges > 16.700 | BMI > 25 & Smoker)

Nur Indah Pratiwi | Skill Upgrader Batch 10

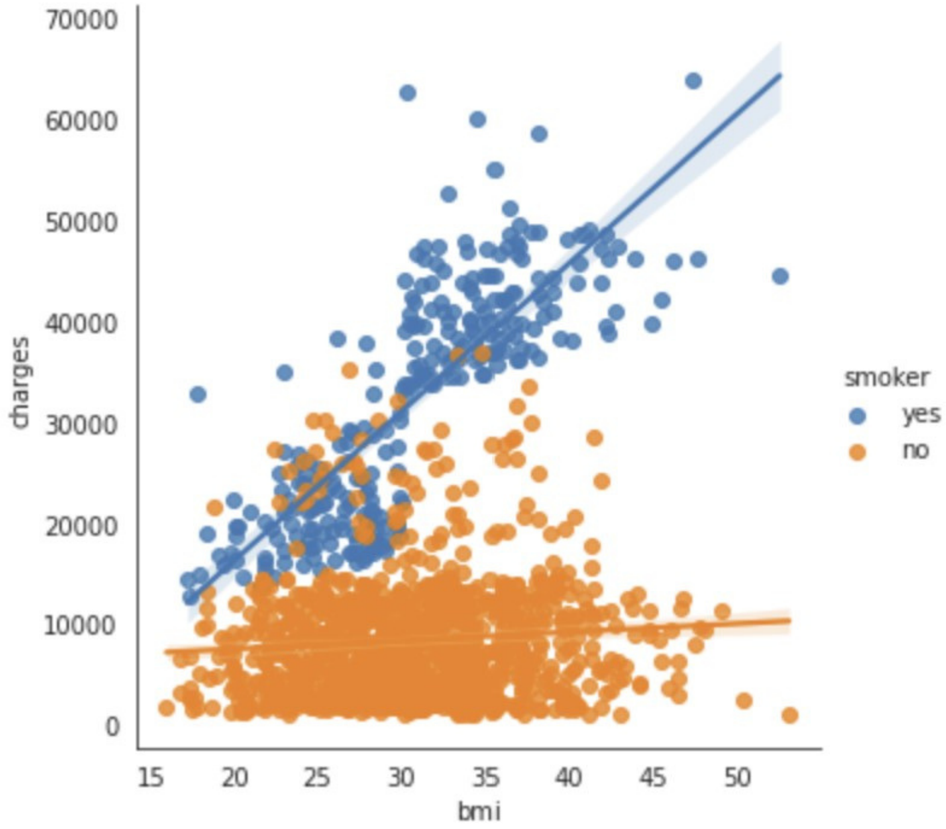
Probability of (Charges | BMI)

The results show that there is no opportunity cost if BMI is given as information.

```
print (pd.crosstab(df.charges, df.bmi, normalize='columns'))
```

0.0

Probability of charges given BMI



Nur Indah Pratiwi |
Probability Course Projects 2022

Probability of (Charges > 16.700 | Smoker)

The results show if someone is a smoker then the probability charges are most likely than 16.700.

```
charges_smoker = df.loc[(df['smoker']=='yes') &
(df['charges']>16700)]

print ('P (Charges > 16.700 | Smoker):
{:.4f}'.format(len(charges_smoker)/len(smoker)))
```

0.9270

Most people come from southeast

Nur Indah Pratiwi |
Probability Course Projects 2022

Probability of (Charges > 16.700 | BMI > 25)

the results show there is very little probability if looking for probability charges more than 16700 if given information BMI is more than 25.

```
bmi = df.loc[df['bmi']>25]
charges_bmi = df.loc[(df['bmi']>25) &
(df['charges']>16700)]

print ('P (Charges > 16.700 | BMI > 25):
{:.4f}'.format(len(charges_bmi)/len(bmi)))
```

0.2594

Probability of charges more than 16.700 given BMI more than 25 as information.

Nur Indah Pratiwi |
Probability Course Projects 2022

Probability of (Charges > 16.700 | BMI < 25)

The probability that might happen is that someone whose BMI is above 25 gets a bill above 16.700.

```
bmi_under_25 = df.loc[df['bmi']<25]
charges_bmi_under_25 = df.loc[(df['bmi']<25) &
(df['charges']>16700)]

print ('P (Charges > 16.700 | BMI < 25):
{:.4f}'.format(len(charges_bmi_under_25)/len(bmi_under_25)))
```

0.2082

The highest data proportion
for non-smokers (southeast)

Nur Indah Pratiwi |
Probability Course Projects 2022

Probability of Charges > 16.700

BMI > 25 & Smoker

```
smoker_bmi_25 = df.loc[(df['charges']>16700) & (df['bmi']>25)
& (df['smoker']=='yes')]
print ('P (Charges > 16.700 | BMI > 25 & Smoker):
{:.4f}'.format(len(smoker_bmi_25)/len(df)))
```

0.1607

Probability charges given BMI more than 25 and smoker a little bigger than BMI information only and least probability than smoker information itself.

BMI > 25 & Non-Smoker

```
non_smoker_bmi_25 = df.loc[(df['charges']>16700) &
(df['bmi']>25) & (df['smoker']=='no')]
print ('P (Charges > 16.700 | BMI > 25 & Non-smoker):
{:.4f}'.format(len(non_smoker_bmi_25)/len(df)))
```

0.0508

The probability result show non-smokers with BMI more than 25 has low chance of paying insurance charges more than 16,700

Nur Indah Pratiwi |
Probability Course Projects 2022

Analysis

After analyzing the continuous variables, it was found that if a person is a smoker, he has a high chance of producing a health bill of more than 16,700.

although the odds are only 0.25, if a person has a BMI of more than 25 then he has a chance of getting a bill of more than 16,700.

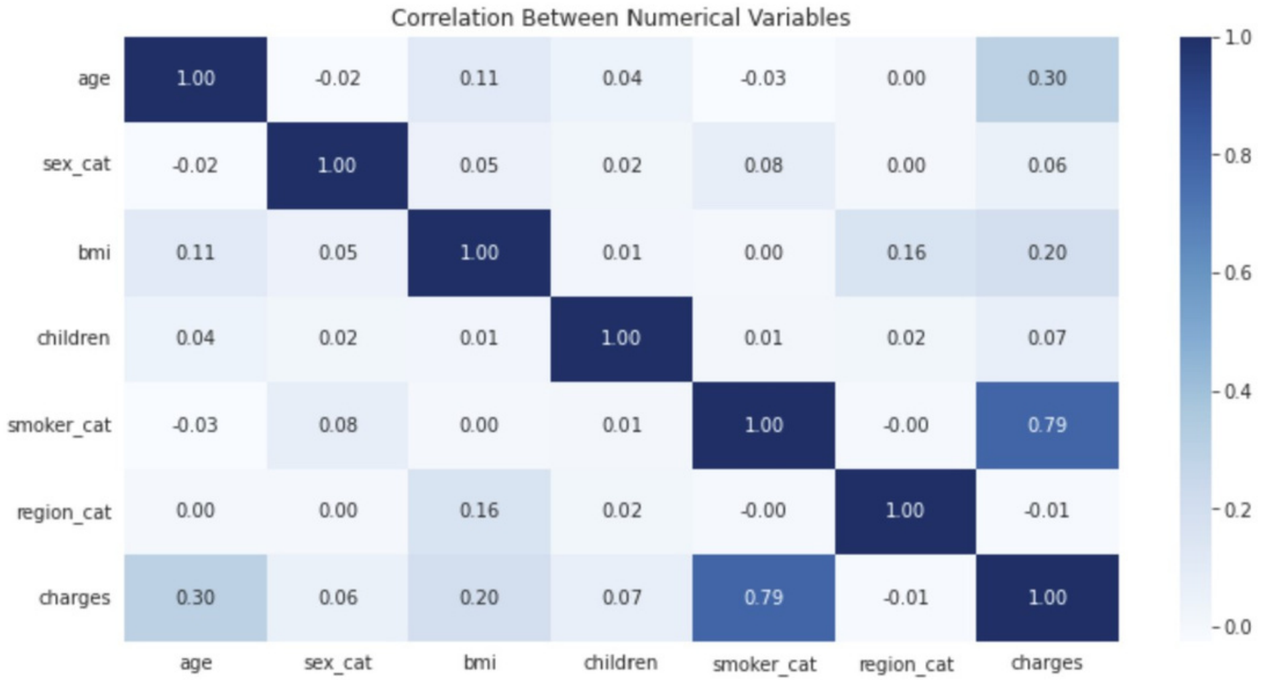
If someone is a smoker and has a BMI of more than 25, they have a greater chance of having a bill of more than 16,700 than non-smokers who have a BMI of more than 25.

Smoker	Probability of smoker get more 16.700 of charges: 0,9270
BMI	The probability that a person with a BMI of more than 25 gets a charges more than 16,700 is 0,2594
Smoker & BMI	The probability that a person with a BMI of more than 25 and he/she is a smoker gets a charges more than 16,700 is 0,1607

Correlation

In this section, we will see the correlation among features by making heatmap plot.

In this section we will look at the relationship of each feature to the other features. So categorical is changed to numeric so that correlations can be calculated between one another. can be seen in this heatmap plot, a strong correlation exists in the smoker feature or not, age and bmi.



Hypothesis Testing

Q-1: Smokers' health charges are higher than non-smokers' health charges

Q-2: Health charges with a BMI above 25 are higher than health charges with a BMI below 25

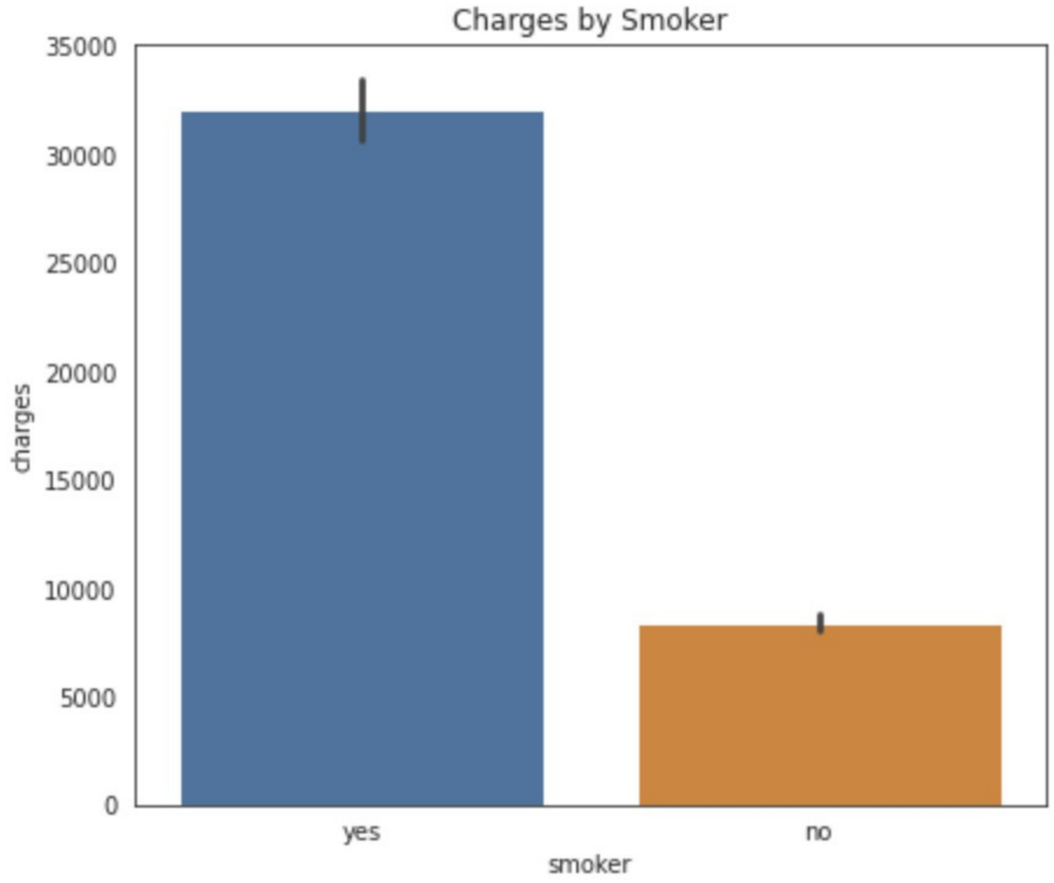
Q-3: Men's health charges are bigger than women's

Nur Indah Pratiwi | Skill Upgrader Batch 10

Charges (Smoker / Non)

In this section, we will test the hypothesis that smokers' charges are higher than nonsmokers' charges. By making a barplot between smokers and non-smokers as the axis and charges as y.

```
sns.set_style('white')
plt.figure(figsize = (7, 6))
sns.barplot(x = 'smoker', y = 'charges', data = df)
plt.title('Charges by Smoker')
```

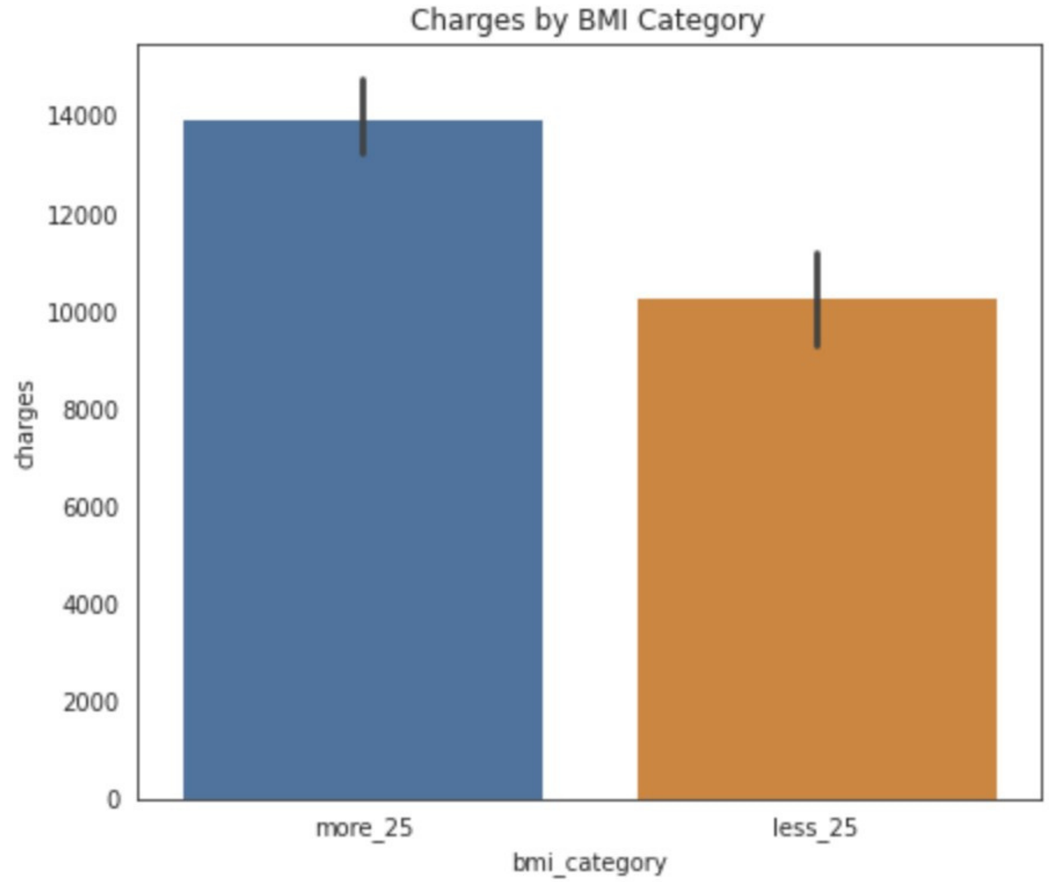


Nur Indah Pratiwi |
Probability Course Projects 2022

Charges by BMI Category

Previously, BMI had been grouped based on the two categories that we wanted to search for, so there were two categories: more than 25 and less than 25. After plotting it into bars, it was true that BMI over 25 had charges that were greater than BMI less than 25.

```
sns.set_style('white')
plt.figure(figsize = (7, 6))
sns.barplot(x = 'bmi_category', y = 'charges', data = df)
plt.title('Charges by BMI Category')
```

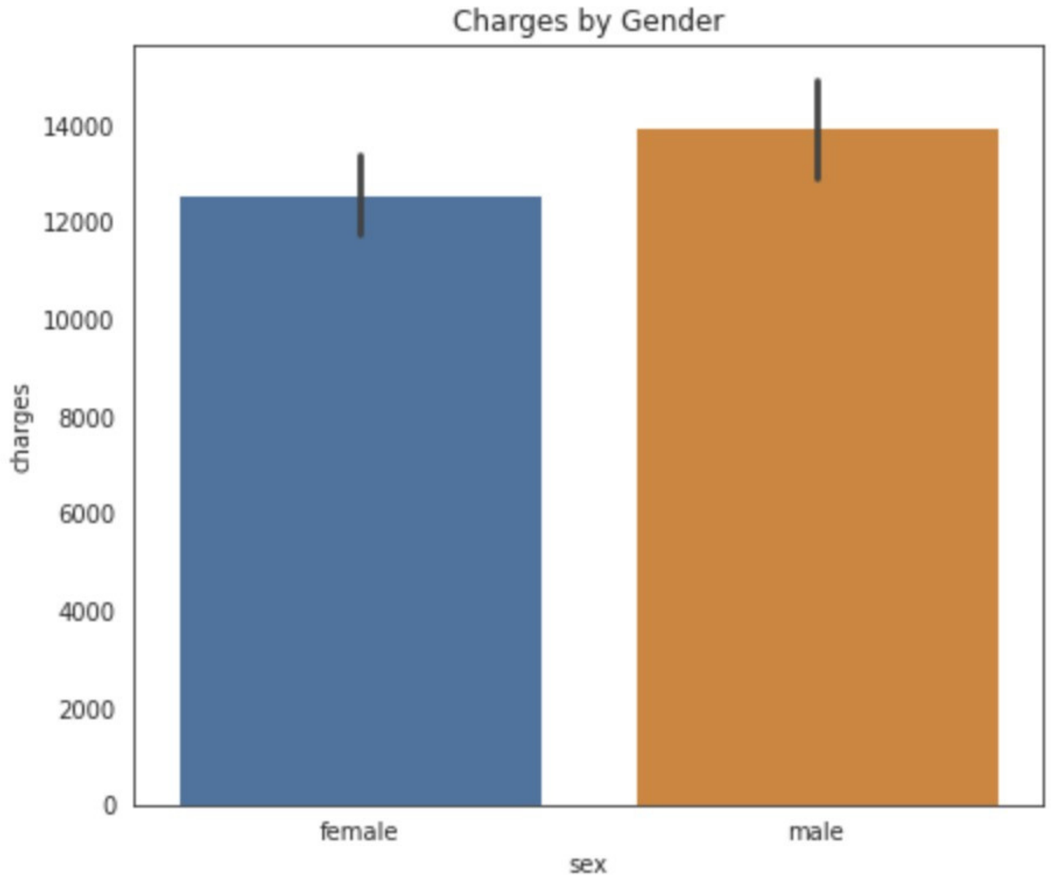


Nur Indah Pratiwi |
Probability Course Projects 2022

Charges by Gender

Variance tells the degree of spread in the data set. The more spread the data, the larger the variance is in relation to the mean

```
sns.set_style('white')
plt.figure(figsize = (7, 6))
sns.barplot(x = 'sex', y = 'charges', data = df)
plt.title('Charges by Gender')
```



Nur Indah Pratiwi |
Probability Course Projects 2022

Conclusion

After analyzing the features in the data, it can be concluded that there are 3 features that affect the size of the health charges, namely: whether he smokes or not, if he smokes then his health charges tends to be higher, the second is age. someone affects his health charges, someone will pay more when he is older, the last is the BMI feature, it turns out that a BMI that is more than 25 will affect his health charges.

Features	Describe	Result
Smoker Feature	Smoker/Non Smoker	Smokers pay a higher charges than non-smokers.
Age Feature	Range : 18-64	People pay more for health charges as they get older.
BMI Category	More than 25 Less than 25	Someone with BMI more than 25 pay a higher charges than someone who gets BMI less than 25.

Thank You!

Contact

Nur Indah Pratiwi
ID Discord: wiwaaw#4451
nurindahpratiwi@alumni.ui.ac.id