Pacmann

Final Probability Project by Sabrina

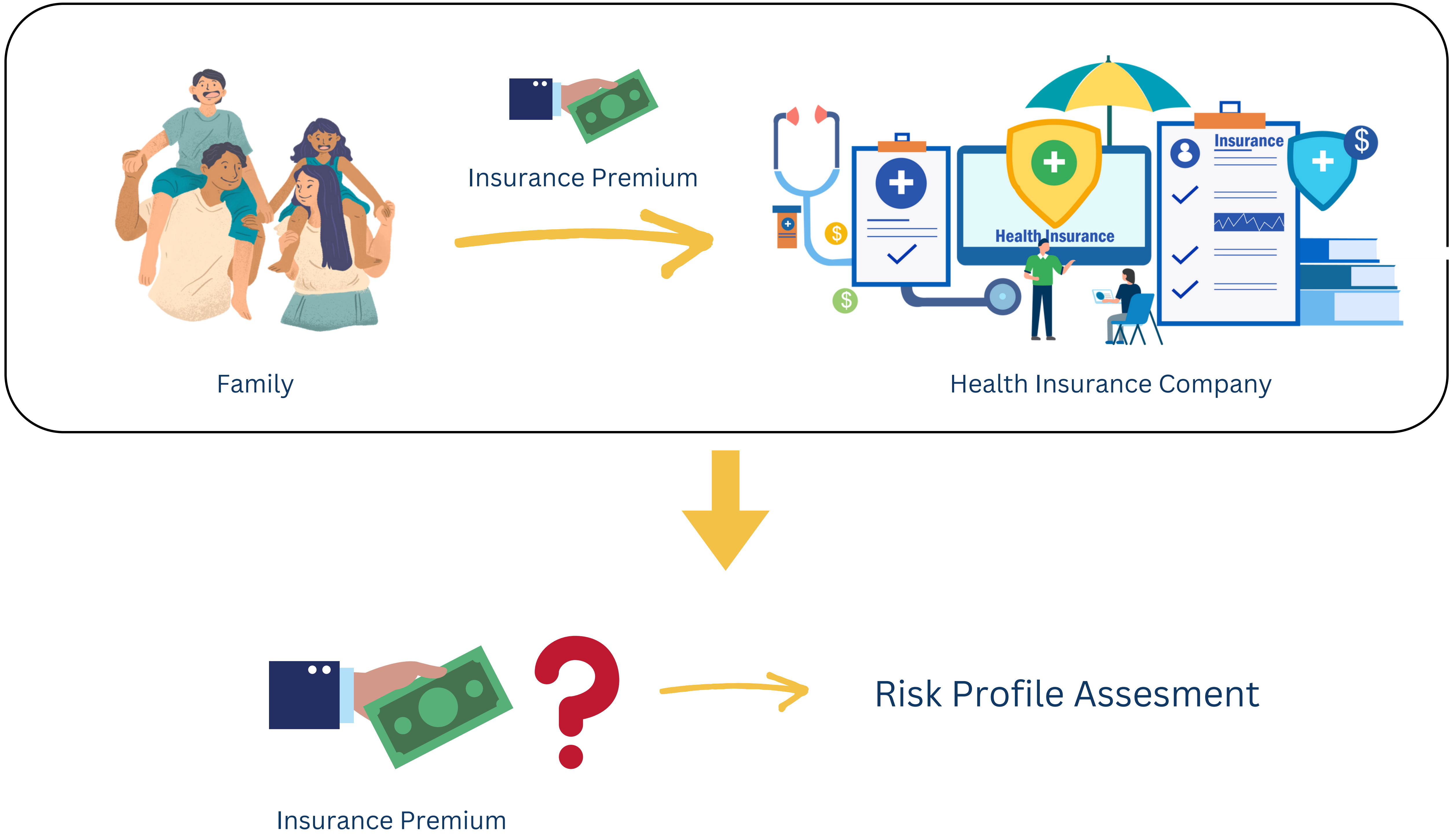
Probability Course - Sekolah Data Pacmann

Outline

- Introduction
- Dataset
- Descriptive Statistic Analysis
- Categorical Variables Analysis
- Continuous Variables Analysis
- Variables Correlation
- Hypothesis Testing
- Conclusion

Introduction

Introduction



Insurance Premium

Family

Health Insurance Company

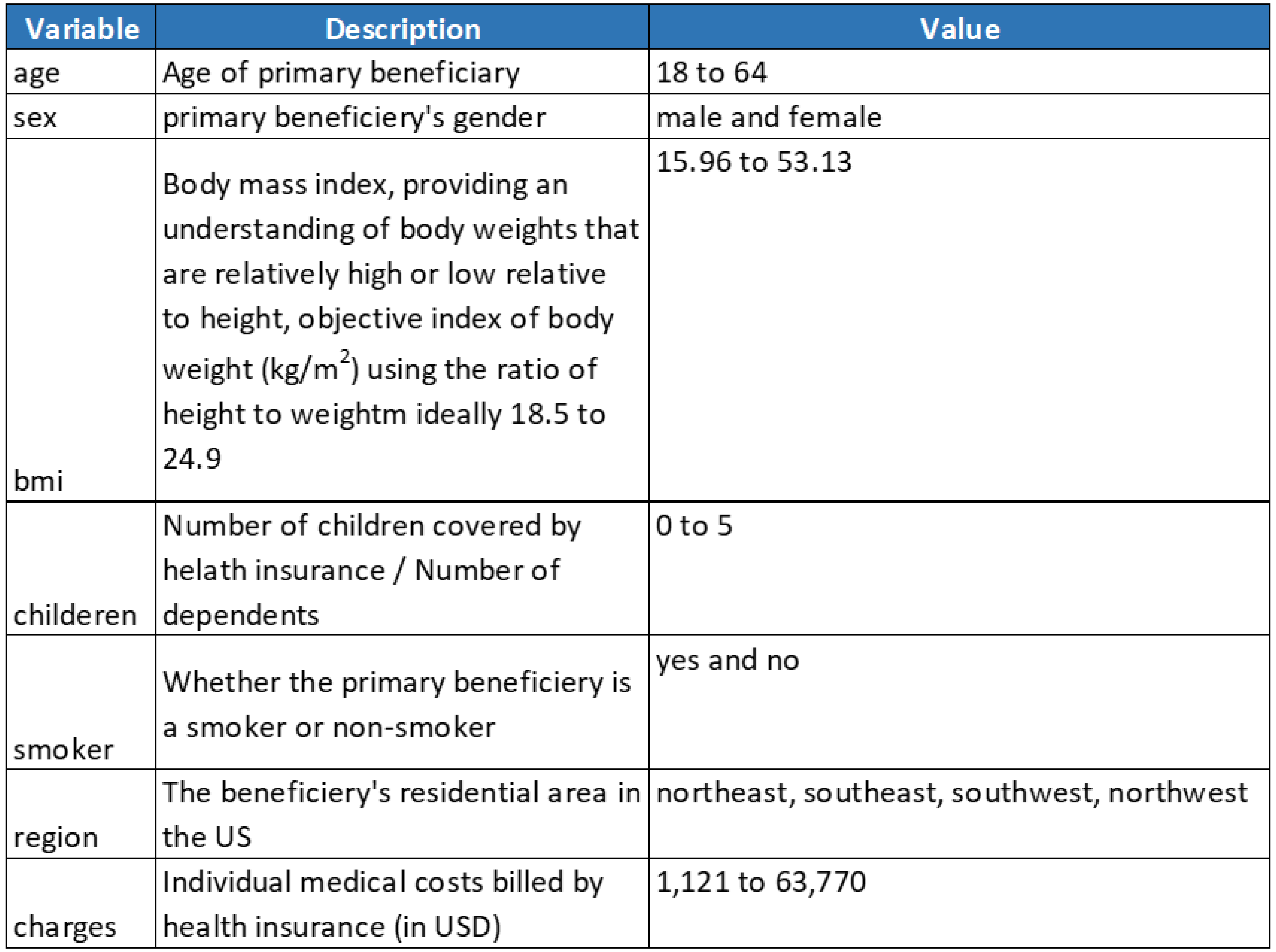
Insurance Premium

Risk Profile Assesment

Dataset

Dataset

- The dataset provided is personal health billing data.



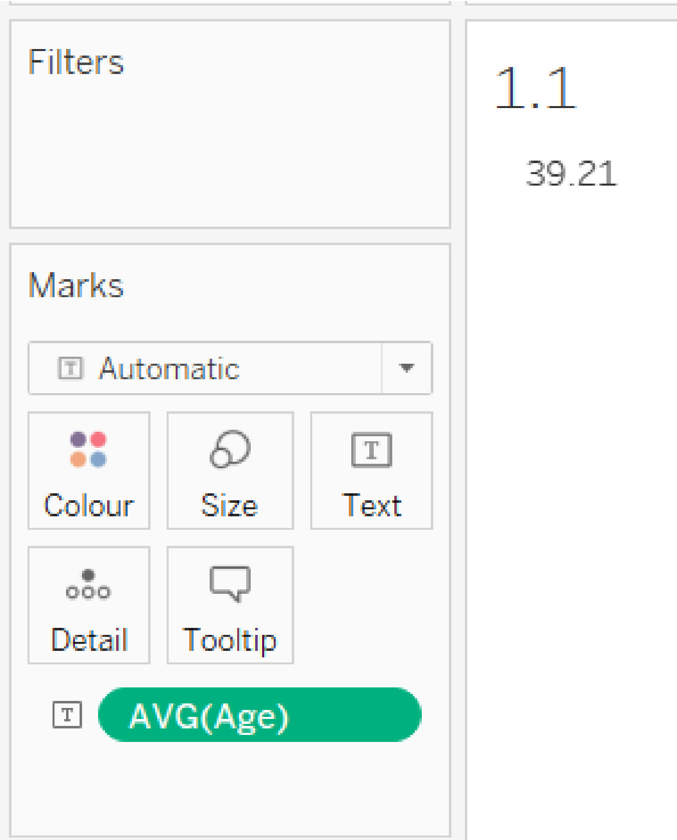
Variable	Description	Value
age	Age of primary beneficiary	18 to 64
sex	primary beneficiery's gender	male and female
bmi	Body mass index, providing an understanding of body weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weightm ideally 18.5 to 24.9	15.96 to 53.13
childeren	Number of children covered by helath insurance / Number of dependents	0 to 5
smoker	Whether the primary beneficiery is a smoker or non-smoker	yes and no
region	The beneficiery's residential area in the US	northeast, southeast, southwest, northwest
charges	Individual medical costs billed by health insurance (in USD)	1,121 to 63,770

Descriptive Statistics Analysis

Mean of Age

- Objective: The average age of the primary beneficiary.

RESULT: The average age is 39



AVG(Age)

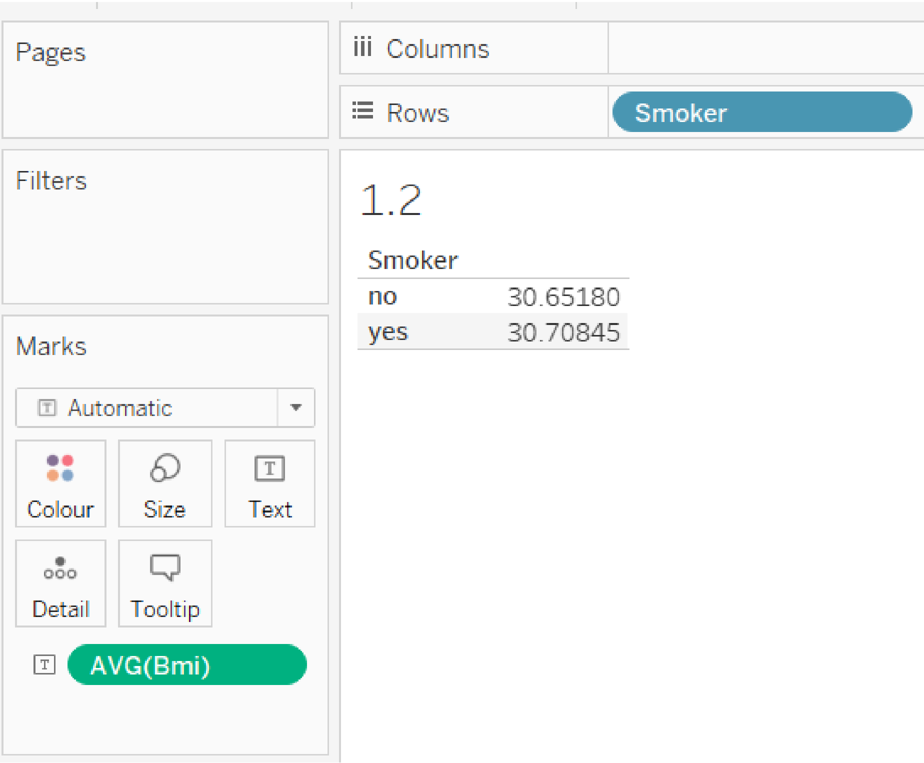
1.1

39.21

Mean of BMI who Smokes

- Objective: The average BMI of the primary beneficiary who smokes

RESULT: The average BMI of a smoker is 30.7



Rows: Smoker

AVG(Bmi)

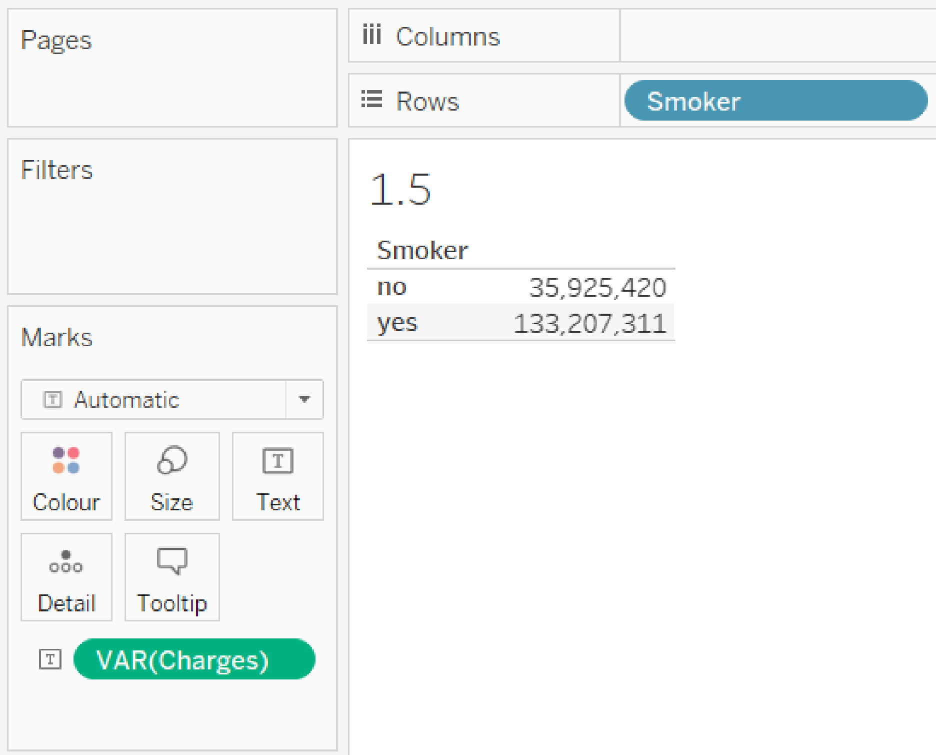
1.2

Smoker	
no	30.65180
yes	30.70845

Variance of the data charges of smokers and non-smokers

Objective: Is the variance of the data charges of smokers and non-smokers the same?

RESULT: No. The variance of a smoker is higher.



Rows: Smoker

Marks: VAR(Charges)

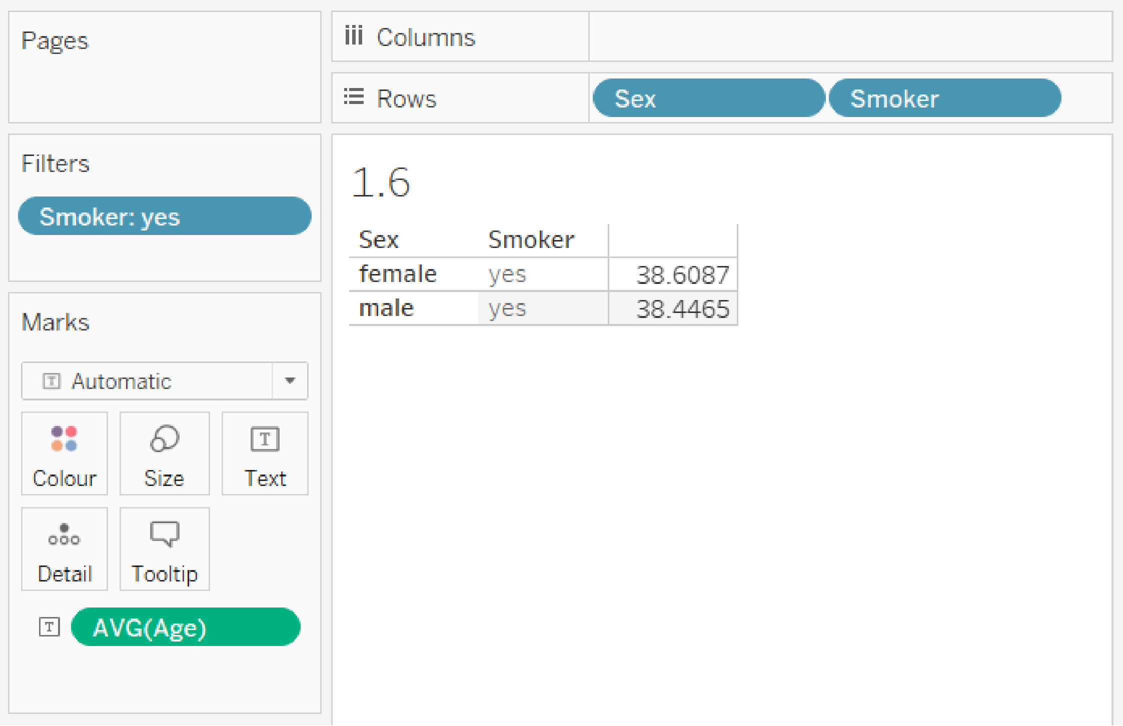
1.5

Smoker	
no	35,925,420
yes	133,207,311

Mean of Age, Gender, and a Smoker

- Objective: Is the average age of women and men who smoke the same?

RESULT: Yes, the average age of male smokers & female smokers is 38 years.



Rows: Sex, Smoker

Filters: Smoker: yes

Marks: AVG(Age)

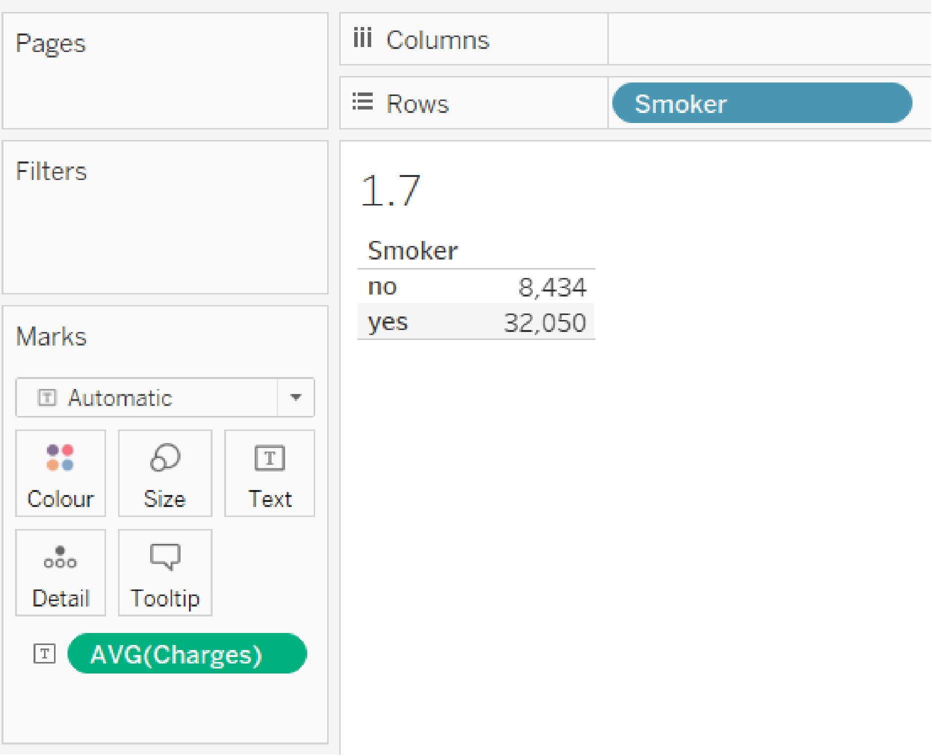
1.6

Sex	Smoker	
female	yes	38.6087
male	yes	38.4465

Charges & Smoking habit

Objective: Which is higher, the average charges of smokers or non-smokers?

RESULT: The average charge of a smoker is higher by around $23K



Rows: Smoker

Text: AVG(Charges)

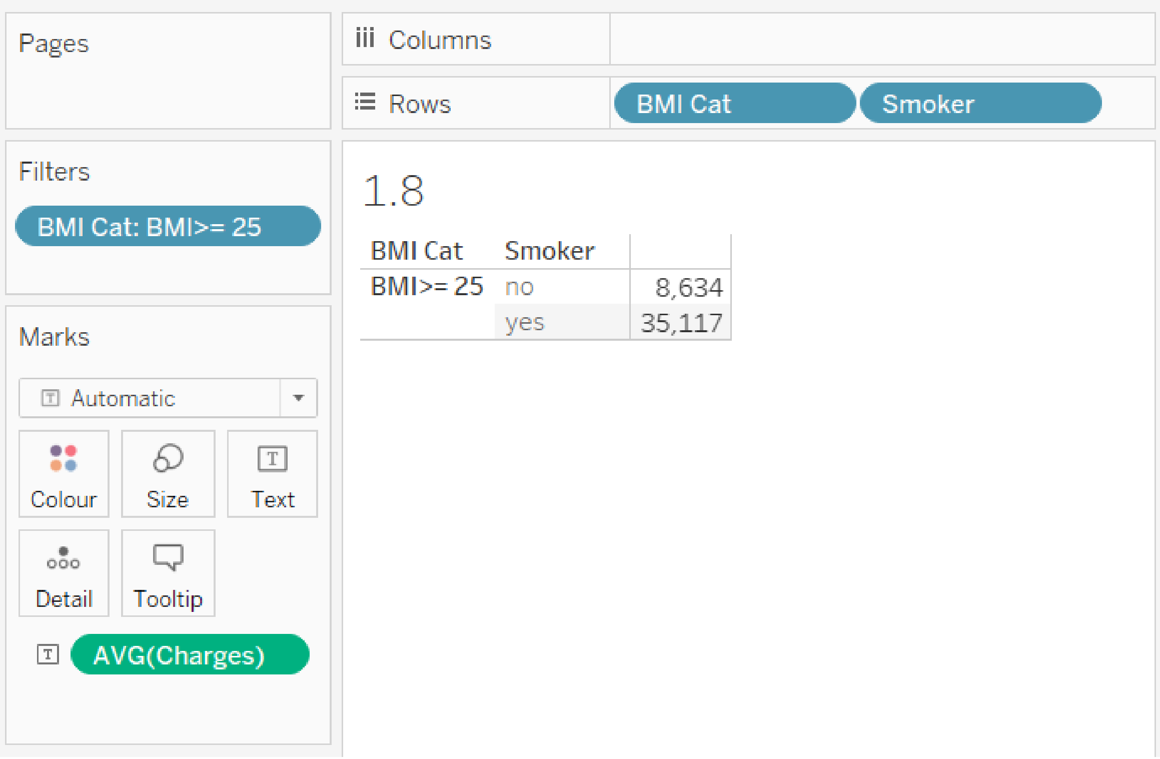
1.7

Smoker	
no	8,434
yes	32,050

Charges, Smoking Habit, and BMI

- Objective: Which is higher, the average charges of smokers whose BMI is above 25 or non-smokers whose BMI is above 25?

RESULT: The average charge of a smoker with BMI >= 25 is higher by around $26K



Rows: BMI Cat, Smoker

Filters: BMI Cat: BMI>= 25

Text: AVG(Charges)

1.8

BMI Cat	Smoker	
BMI>= 25	no	8,634
	yes	35,117

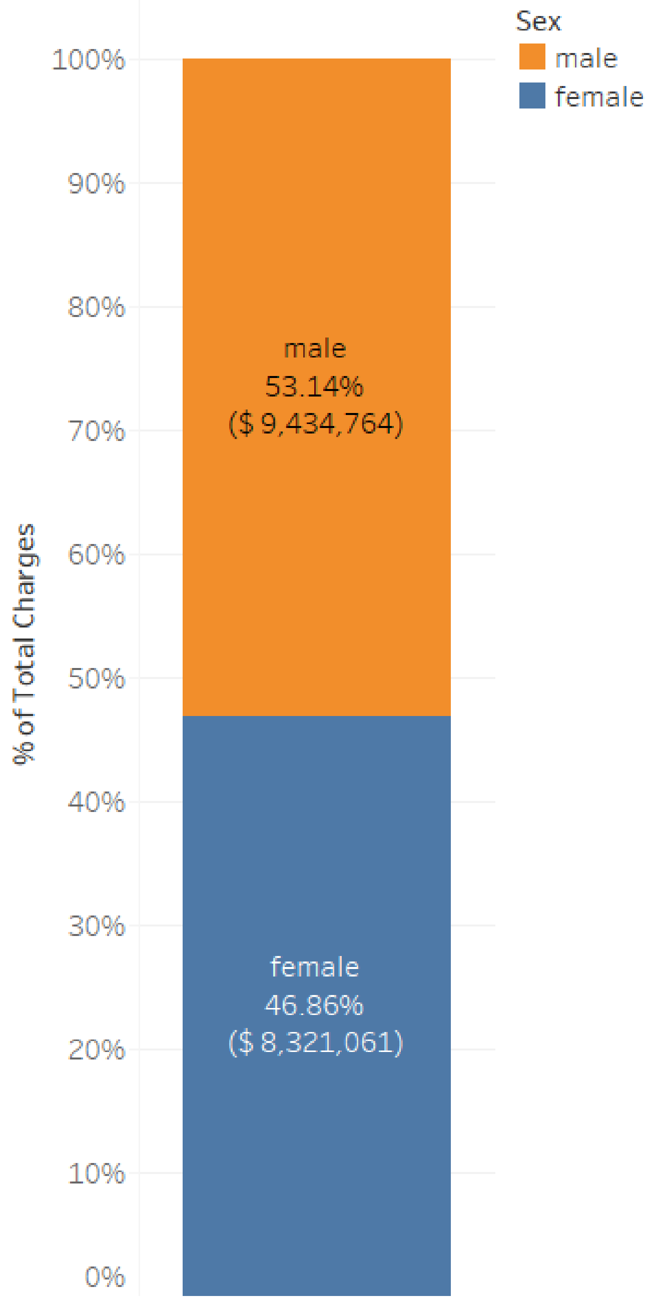
Analysis

- The average age of a smoker **is lower than** the average age of all the primary beneficiaries by one year.
- The average BMI of a smoker is 30.7. Based on the CDC, the average **BMI value of 30.7 falls within the obesity class 1 range**.
- The average charge of a smoker with BMI >= 25 is **higher by around $26K.**

Categorical Variables Analysis

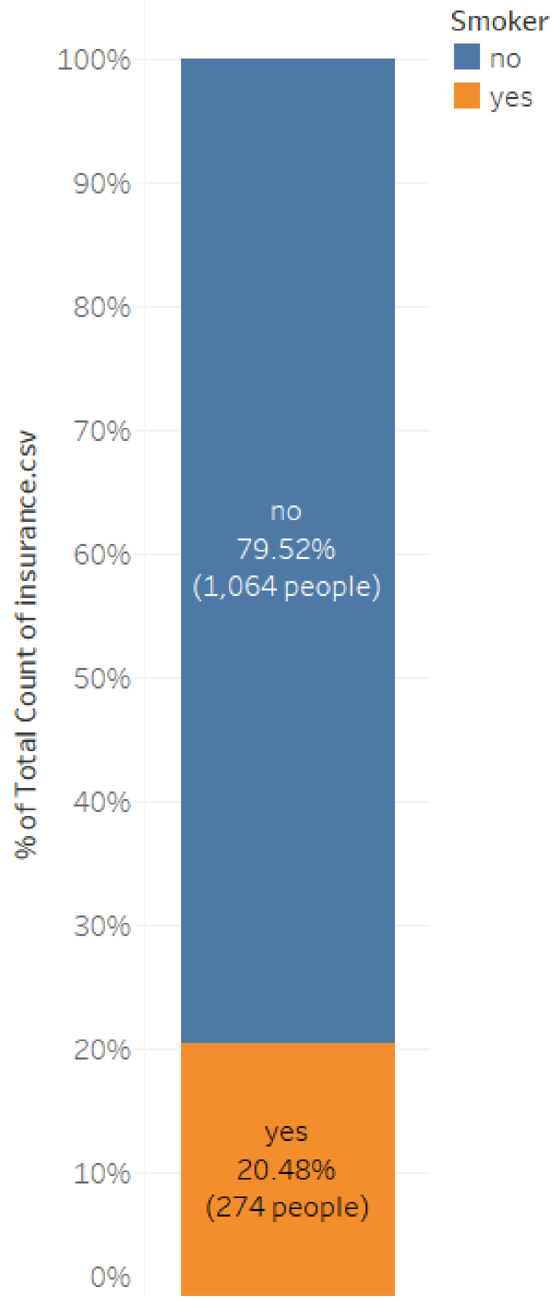
Pacmann

Gender & Charges



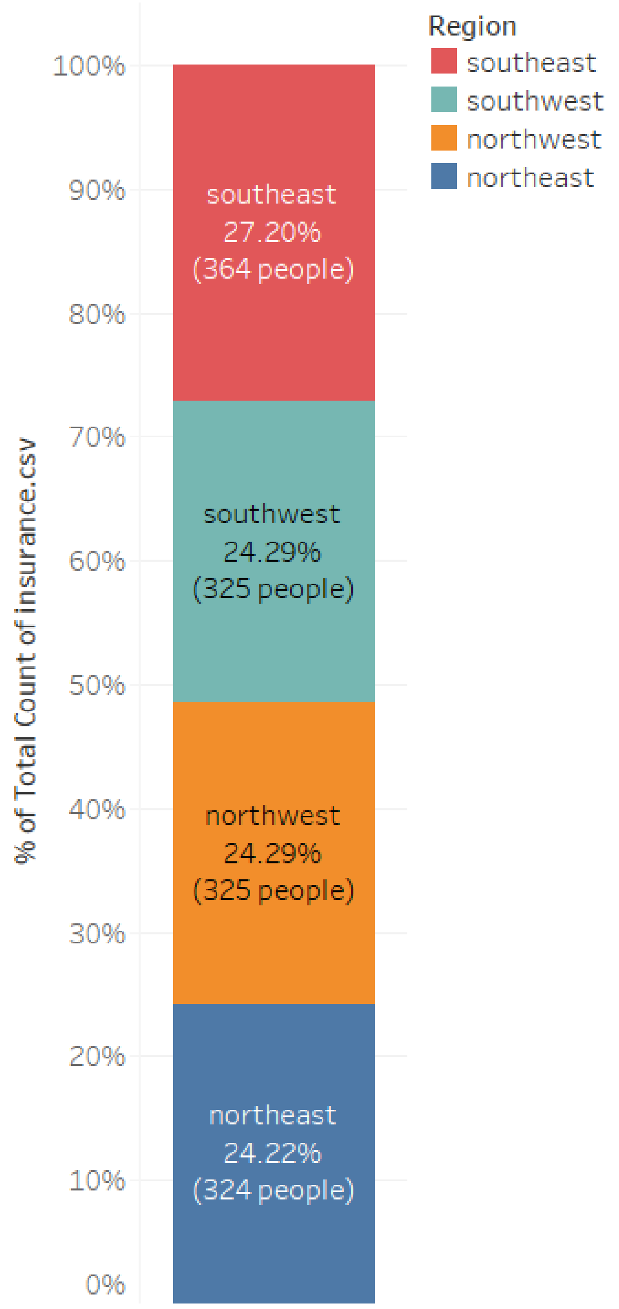
RESULT: The **male beneficiaries** have a **higher proportion** of around 6%.

Smoking Habit



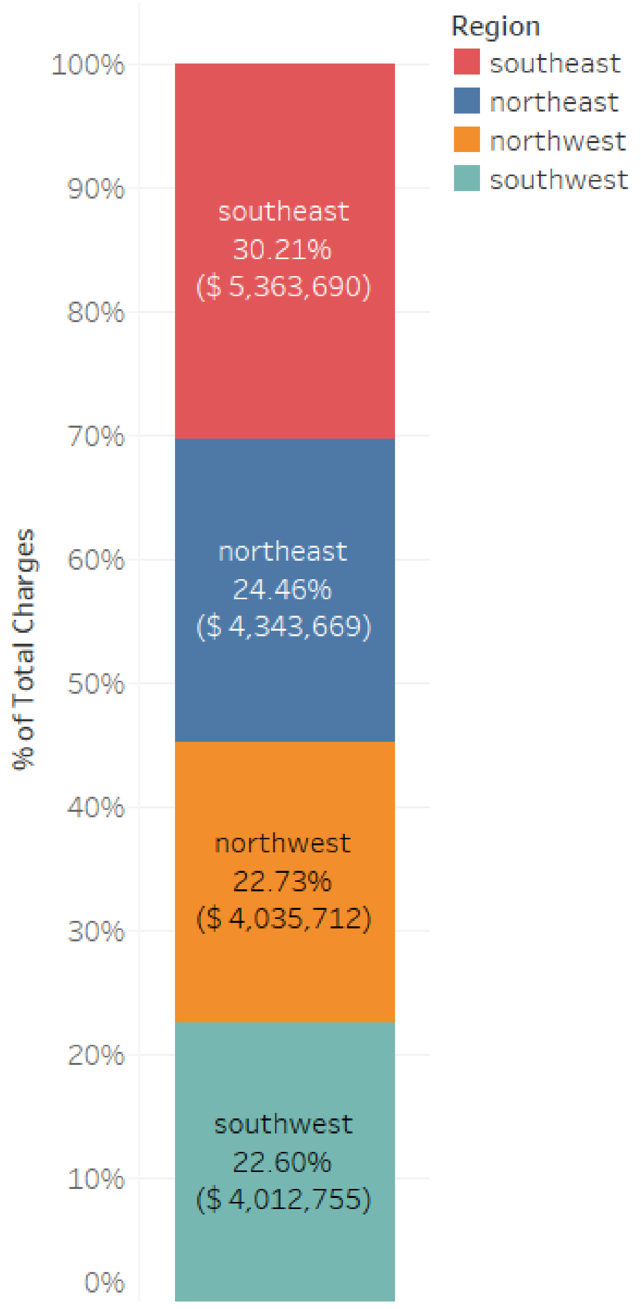
RESULT: **Non-Smokers** have a **higher proportion** of around 60%.

Regions



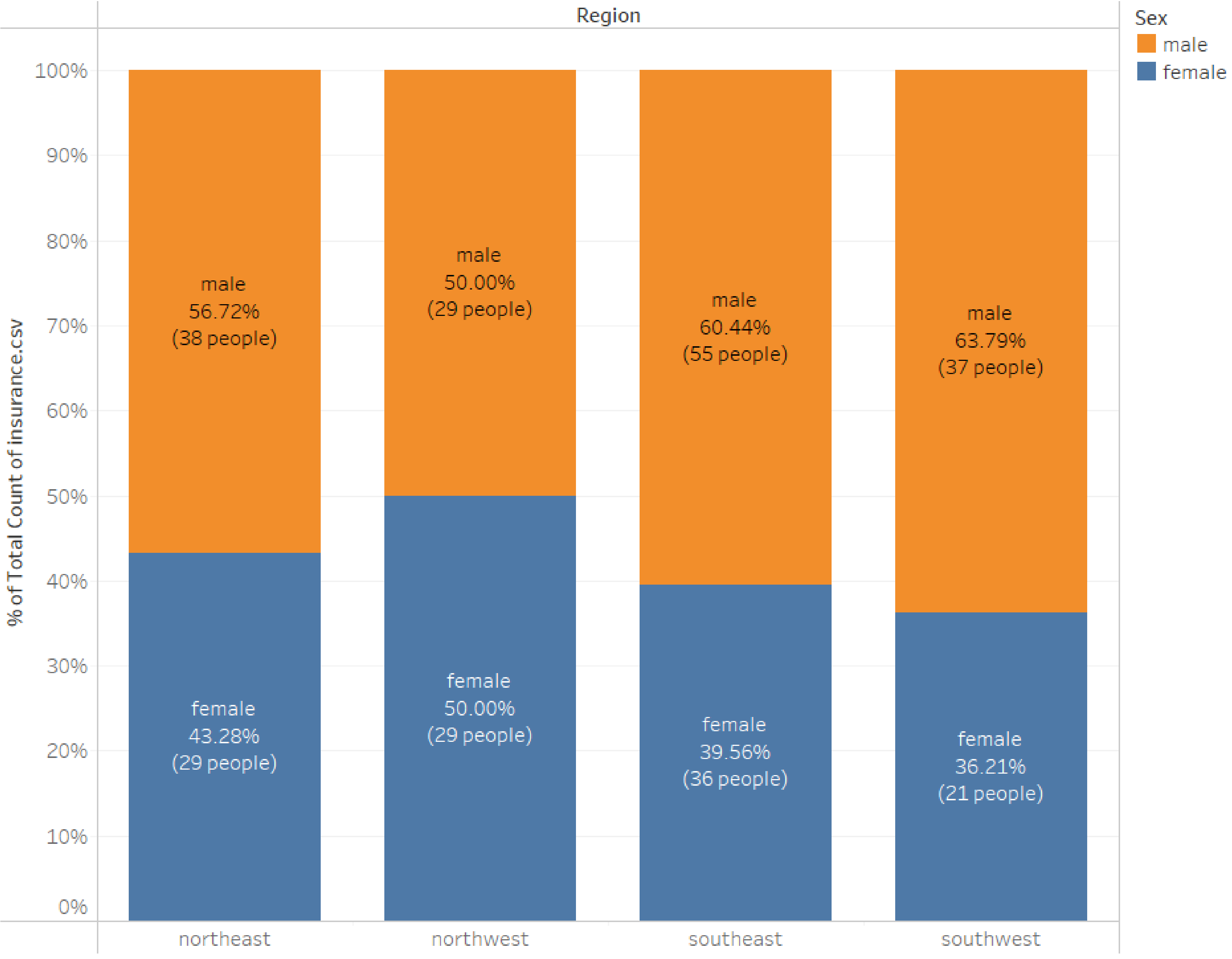
RESULT: **The southeast** has the **highest proportion of people** than other regions.

Charges & Region



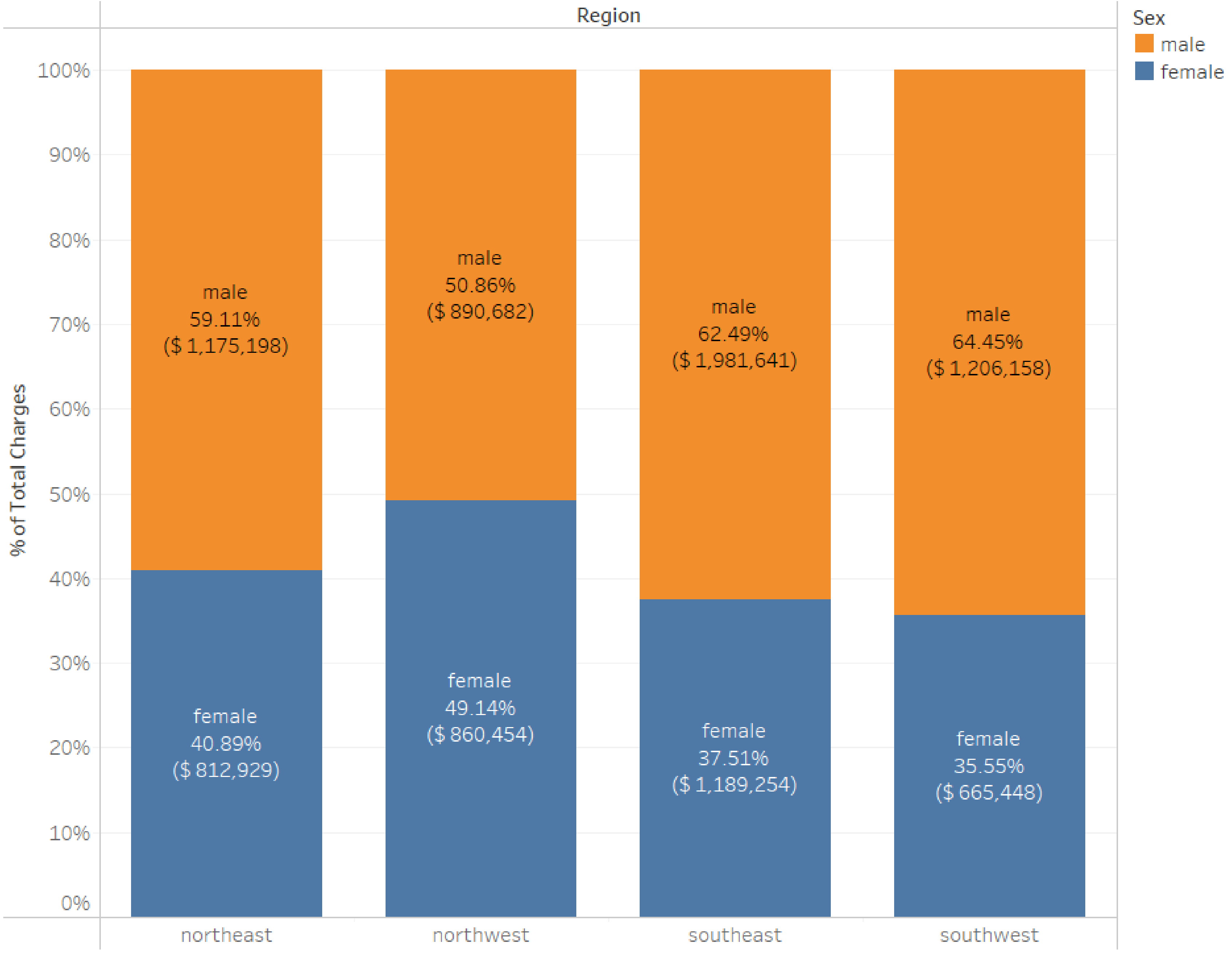
RESULT: **The southeast region** has the **highest proportions of charge** than the other regions.

Smoker and Region



RESULT: **Male Smokers** from the **southwest** have the **highest proportion of people.**

Smoker, Region, Charges



RESULT: **Male Smokers** from the **southwest** have the **highest proportion of charges.**

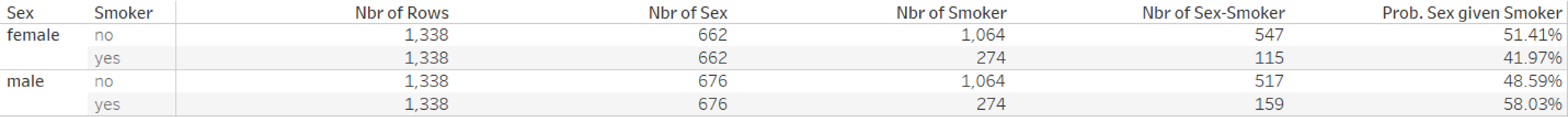
Pacmann

Female & Smoker

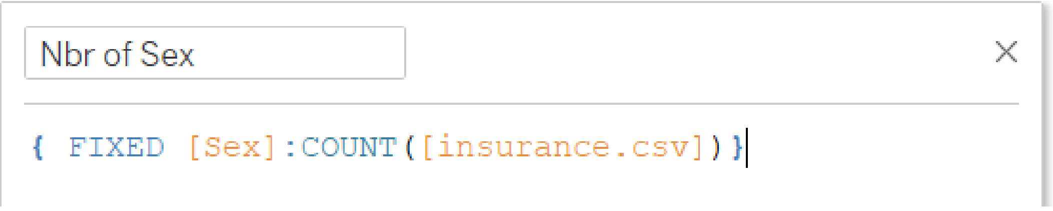
RESULT: The probability of a **female** given she's a **smoker** is **41.97%**.

Male & Smoker

RESULT: The probability of a **male** given he's a **smoker** is 58.3%.

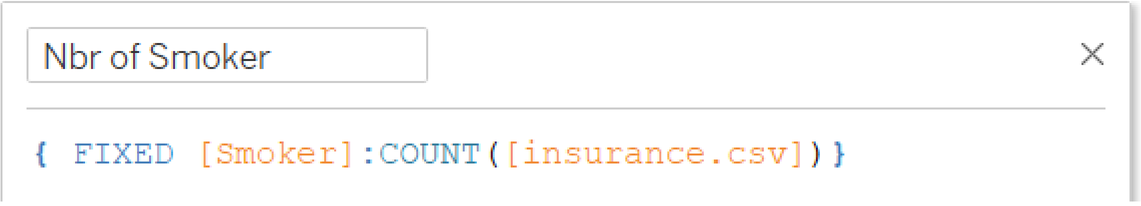


Sex	Smoker	Nbr of Rows	Nbr of Sex	Nbr of Smoker	Nbr of Sex-Smoker	Prob. Sex given Smoker
female	no	1,338	662	1,064	547	51.41%
	yes	1,338	662	274	115	41.97%
male	no	1,338	676	1,064	517	48.59%
	yes	1,338	676	274	159	58.03%



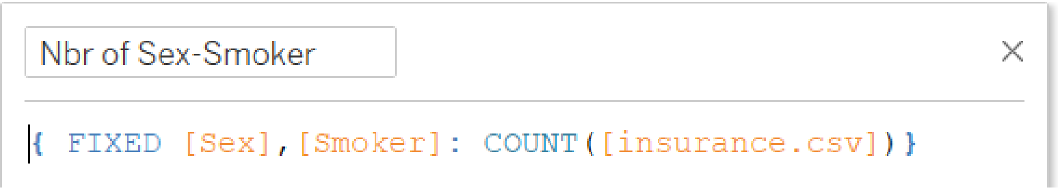
Nbr of Sex

```
{ FIXED [Sex]:COUNT([insurance.csv])}
```



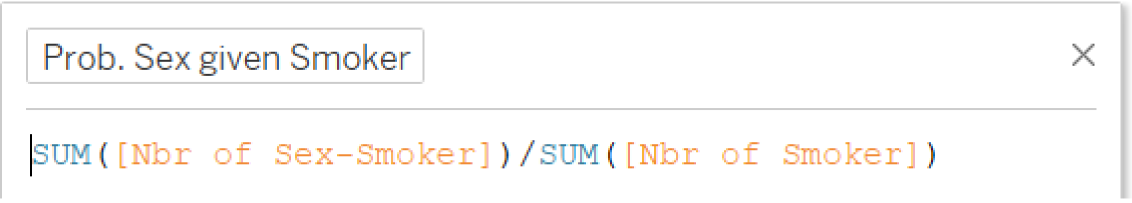
Nbr of Smoker

```
{ FIXED [Smoker]:COUNT([insurance.csv])}
```



Nbr of Sex-Smoker

```
{ FIXED [Sex],[Smoker]: COUNT([insurance.csv])}
```



Prob. Sex given Smoker

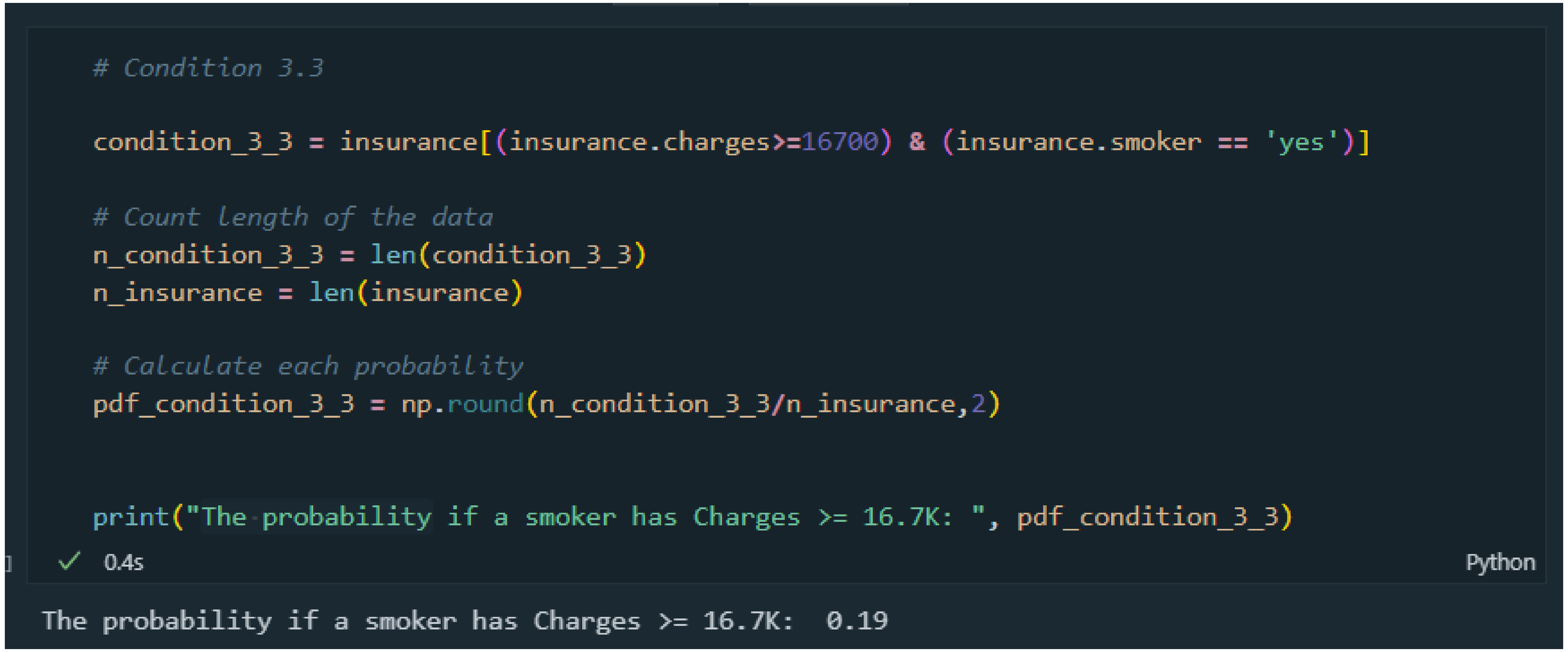
```
SUM([Nbr of Sex-Smoker])/SUM([Nbr of Smoker])
```

Analysis

- **The Southwest region** has the highest proportion of **male smokers & their charges.**
- **The Northwest region** has the highest proportion of **female smokers & their charges.**
- Although, **the Southeast region** has the highest proportion of **the number of people & their charges.**

Continuous Variables Analysis

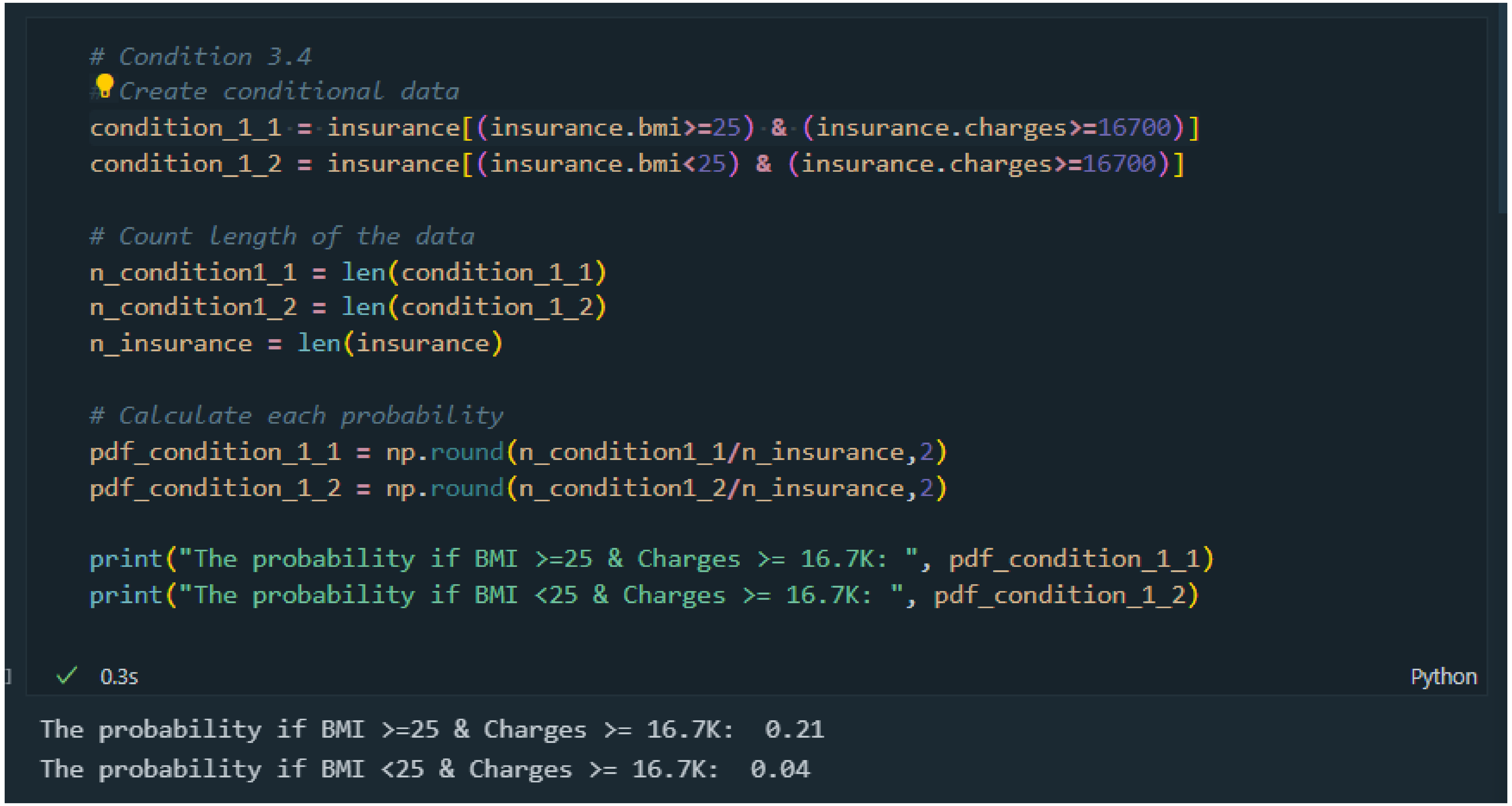
Probability of someone has high charges given he’s a smoker



```python
# Condition 3.3

condition_3_3 = insurance[(insurance.charges>=16700) & (insurance.smoker == 'yes')]

# Count length of the data
n_condition_3_3 = len(condition_3_3)
n_insurance = len(insurance)

# Calculate each probability
pdf_condition_3_3 = np.round(n_condition_3_3/n_insurance,2)


print("The probability if a smoker has Charges >= 16.7K: ", pdf_condition_3_3)
```

```
The probability if a smoker has Charges >= 16.7K:  0.19
```

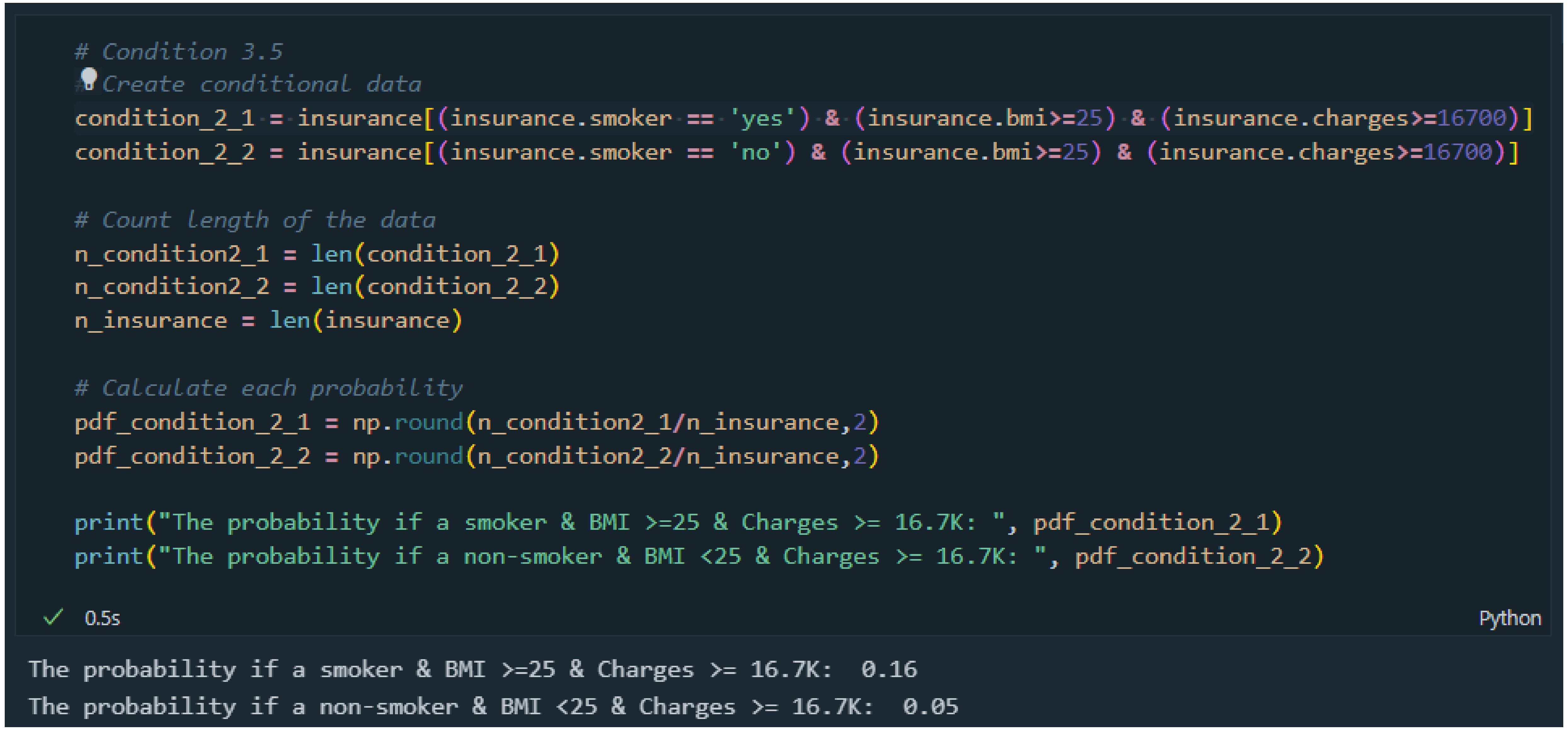
BMI vs Charges



```python
# Condition 3.4
# Create conditional data
condition_1_1 = insurance[(insurance.bmi>=25) & (insurance.charges>=16700)]
condition_1_2 = insurance[(insurance.bmi<25) & (insurance.charges>=16700)]

# Count length of the data
n_condition1_1 = len(condition_1_1)
n_condition1_2 = len(condition_1_2)
n_insurance = len(insurance)

# Calculate each probability
pdf_condition_1_1 = np.round(n_condition1_1/n_insurance,2)
pdf_condition_1_2 = np.round(n_condition1_2/n_insurance,2)

print("The probability if BMI >=25 & Charges >= 16.7K: ", pdf_condition_1_1)
print("The probability if BMI <25 & Charges >= 16.7K: ", pdf_condition_1_2)
```

```
The probability if BMI >=25 & Charges >= 16.7K:  0.21
The probability if BMI <25 & Charges >= 16.7K:  0.04
```

BMI vs Smokers



```python
# Condition 3.5
# Create conditional data
condition_2_1 = insurance[(insurance.smoker == 'yes') & (insurance.bmi>=25) & (insurance.charges>=16700)]
condition_2_2 = insurance[(insurance.smoker == 'no') & (insurance.bmi>=25) & (insurance.charges>=16700)]

# Count length of the data
n_condition2_1 = len(condition_2_1)
n_condition2_2 = len(condition_2_2)
n_insurance = len(insurance)

# Calculate each probability
pdf_condition_2_1 = np.round(n_condition2_1/n_insurance,2)
pdf_condition_2_2 = np.round(n_condition2_2/n_insurance,2)

print("The probability if a smoker & BMI >=25 & Charges >= 16.7K: ", pdf_condition_2_1)
print("The probability if a non-smoker & BMI <25 & Charges >= 16.7K: ", pdf_condition_2_2)
```

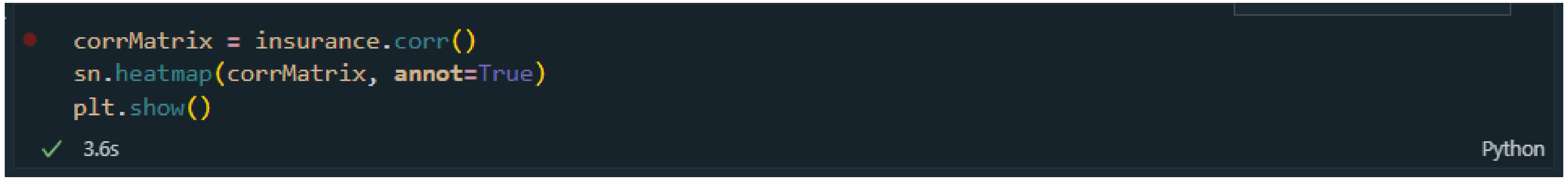
```
The probability if a smoker & BMI >=25 & Charges >= 16.7K:  0.16
The probability if a non-smoker & BMI <25 & Charges >= 16.7K:  0.05
```

Analysis

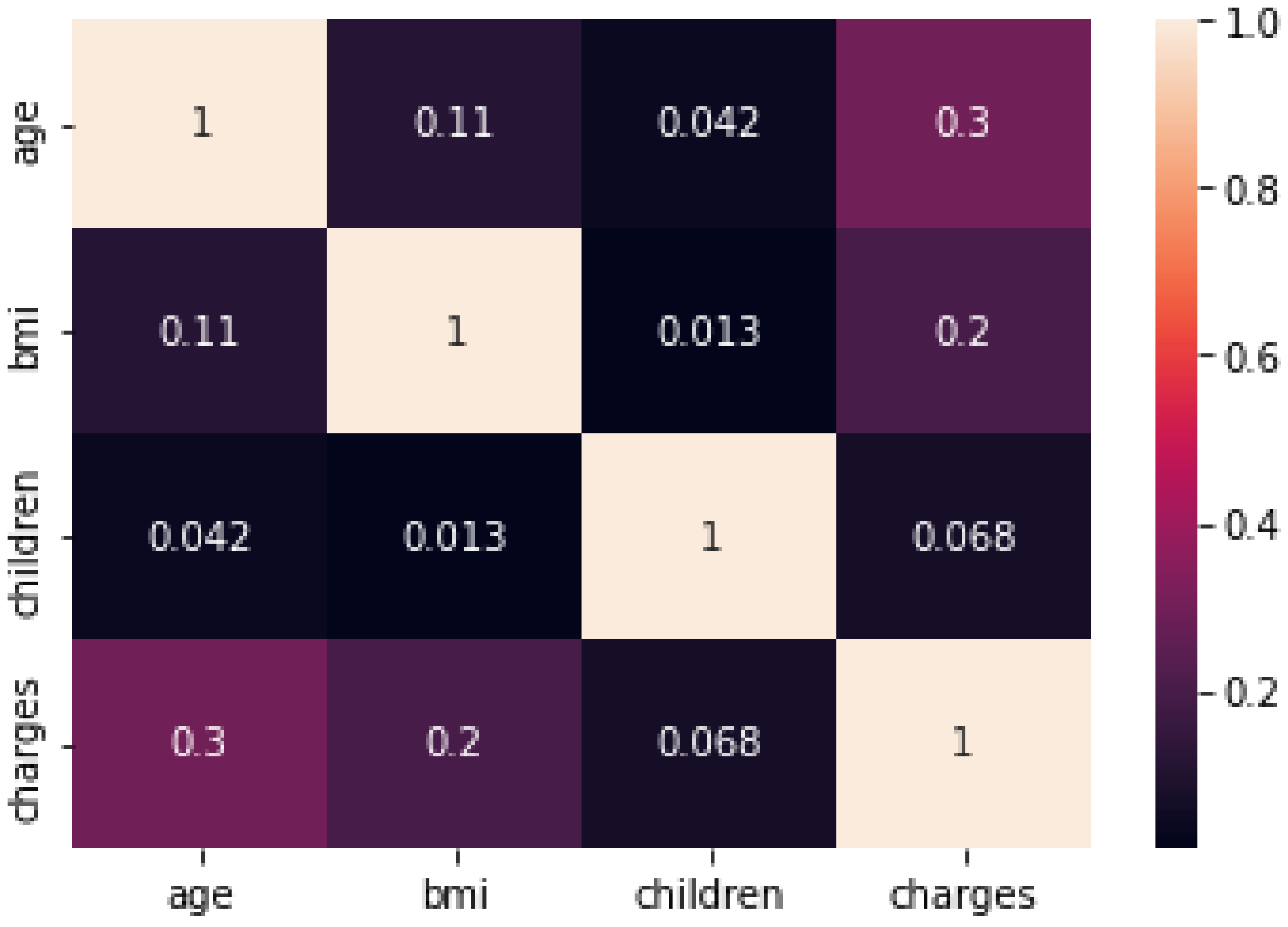
- The probability of someone who has BMI >=25 & Charges >= $16.7K is 21%
- The probability of a smoker having Charges >= $16.7K is 19%
- The probability of a smoker with BMI >=25 & Charges >= $16.7K is 16%.
- **Thus, people who have BMI over 25 & charge more than $16.7K need to pay a higher premium.**

Variables Correlation

Correlation Matrix



```
corrMatrix = insurance.corr()
sn.heatmap(corrMatrix, annot=True)
plt.show()
```



	age	bmi	children	charges
age	1	0.11	0.042	0.3
bmi	0.11	1	0.013	0.2
children	0.042	0.013	1	0.068
charges	0.3	0.2	0.068	1

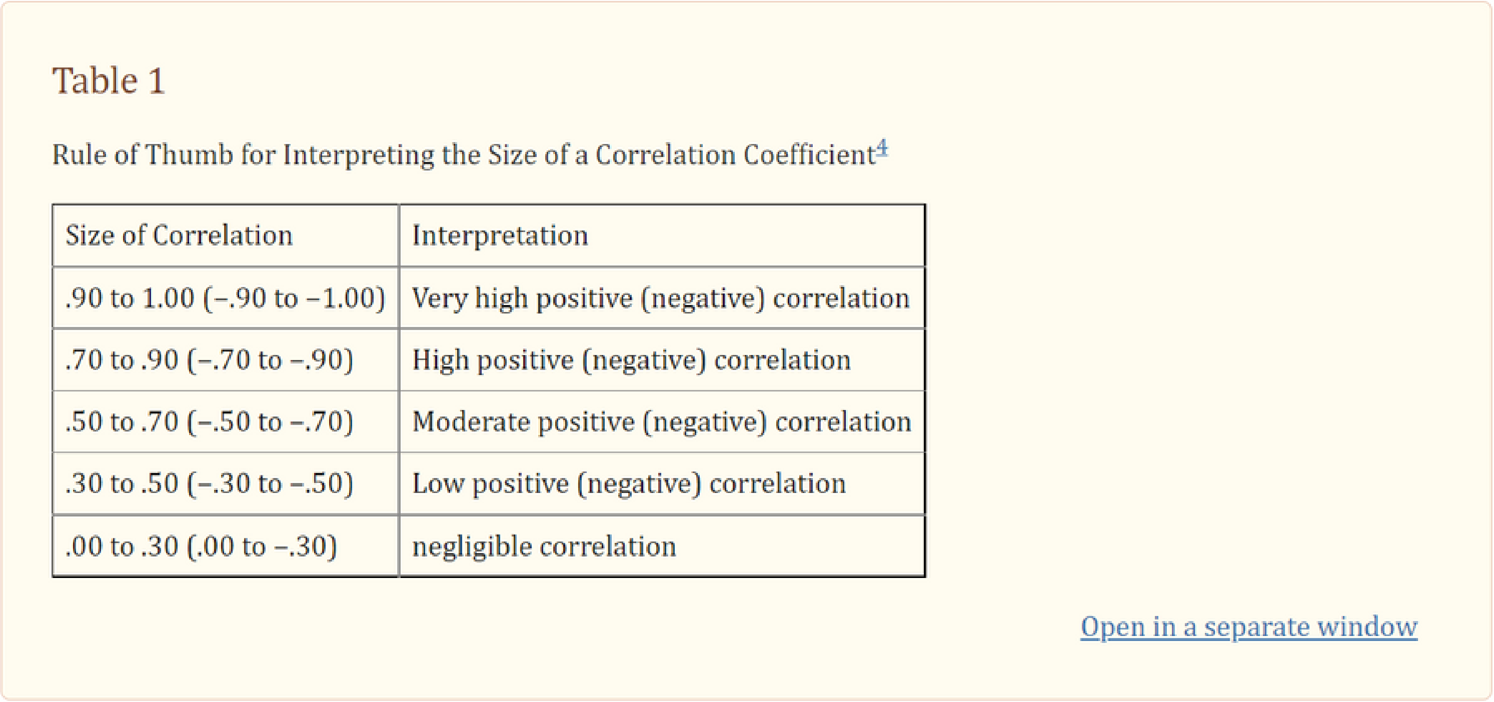


Table 1

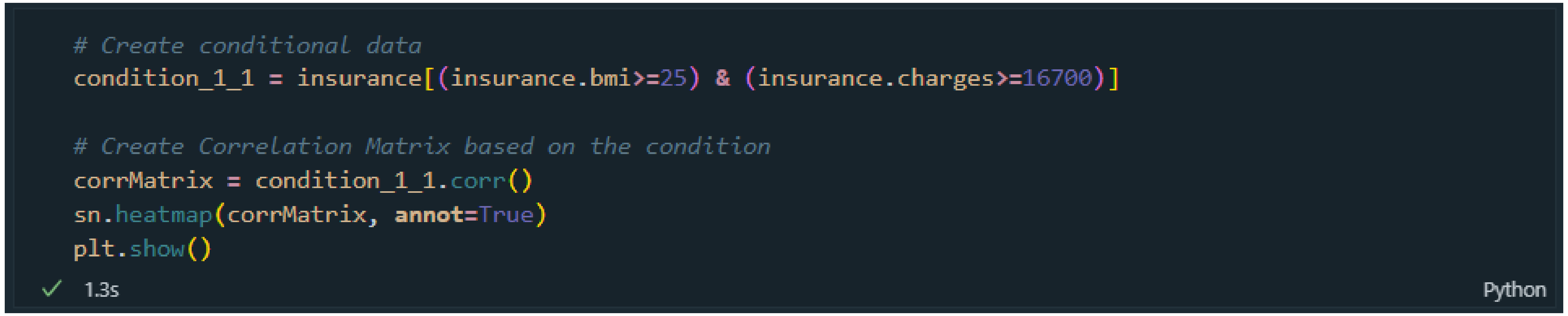
Rule of Thumb for Interpreting the Size of a Correlation Coefficient[4]

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

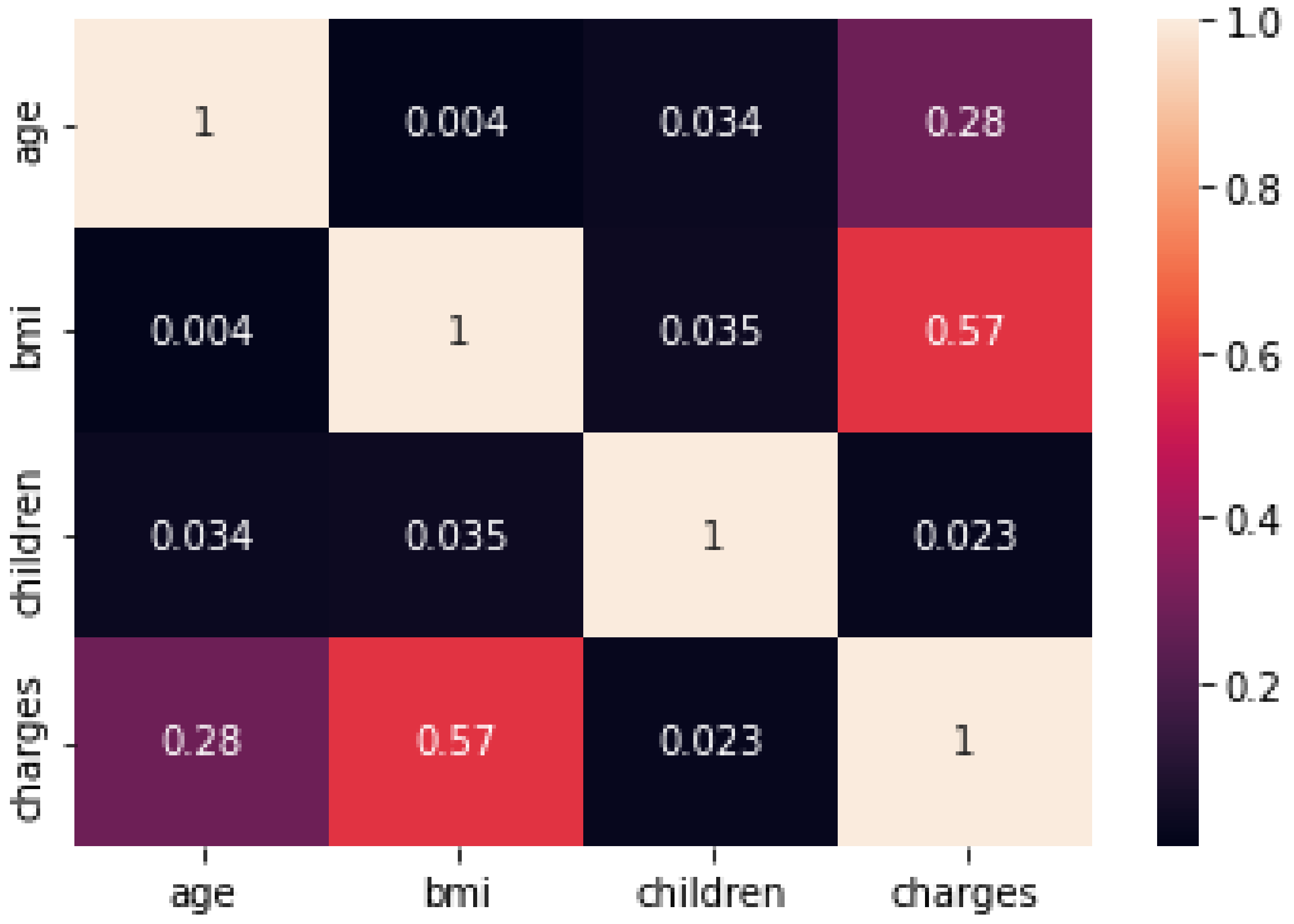
RESULT:

- Age has a low positive correlation with charges.

Correlation Matrix with High Charges & BMI



```python
# Create conditional data
condition_1_1 = insurance[(insurance.bmi>=25) & (insurance.charges>=16700)]

# Create Correlation Matrix based on the condition
corrMatrix = condition_1_1.corr()
sn.heatmap(corrMatrix, annot=True)
plt.show()
```



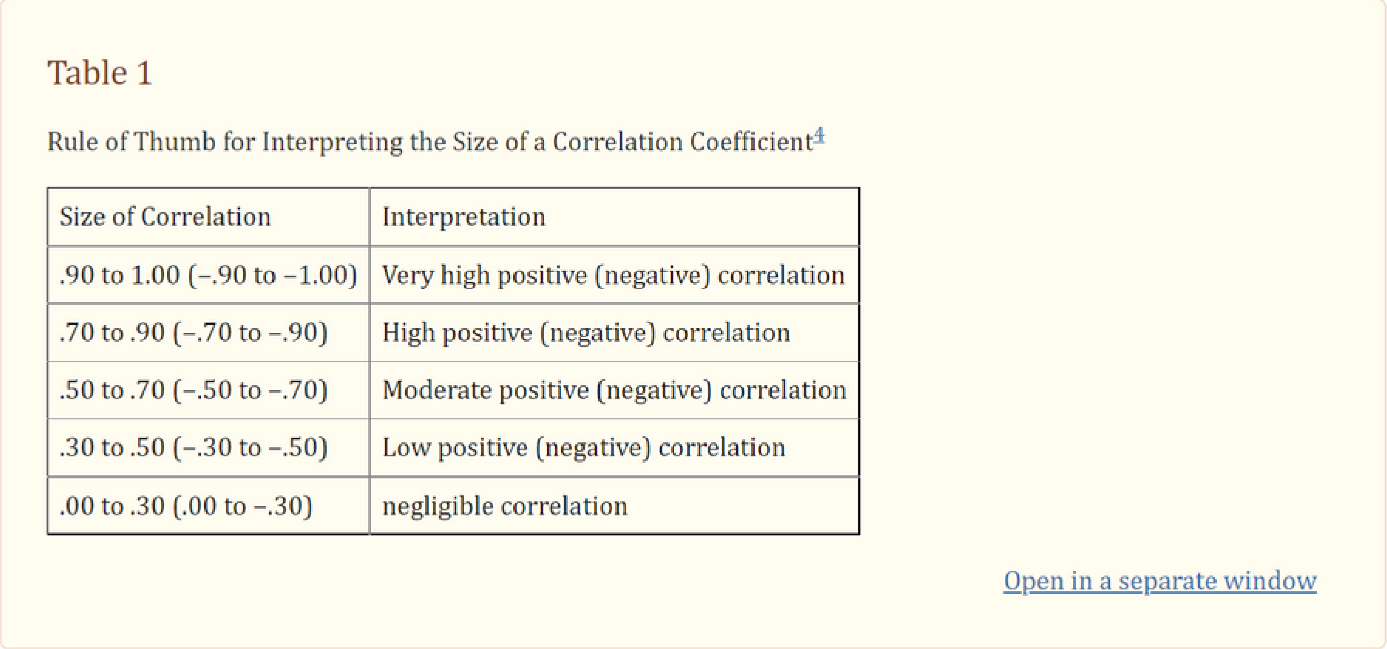


Table 1

Rule of Thumb for Interpreting the Size of a Correlation Coefficient[4]

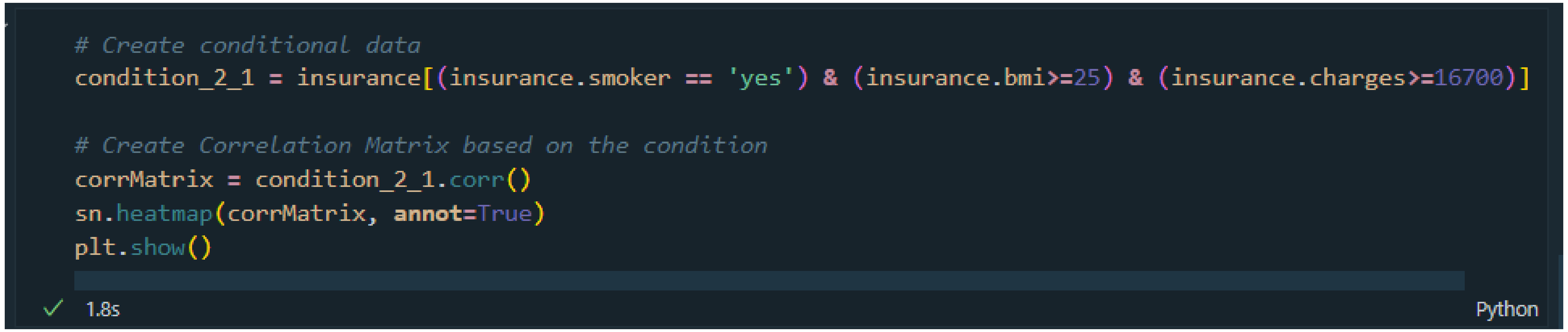
Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Open in a separate window

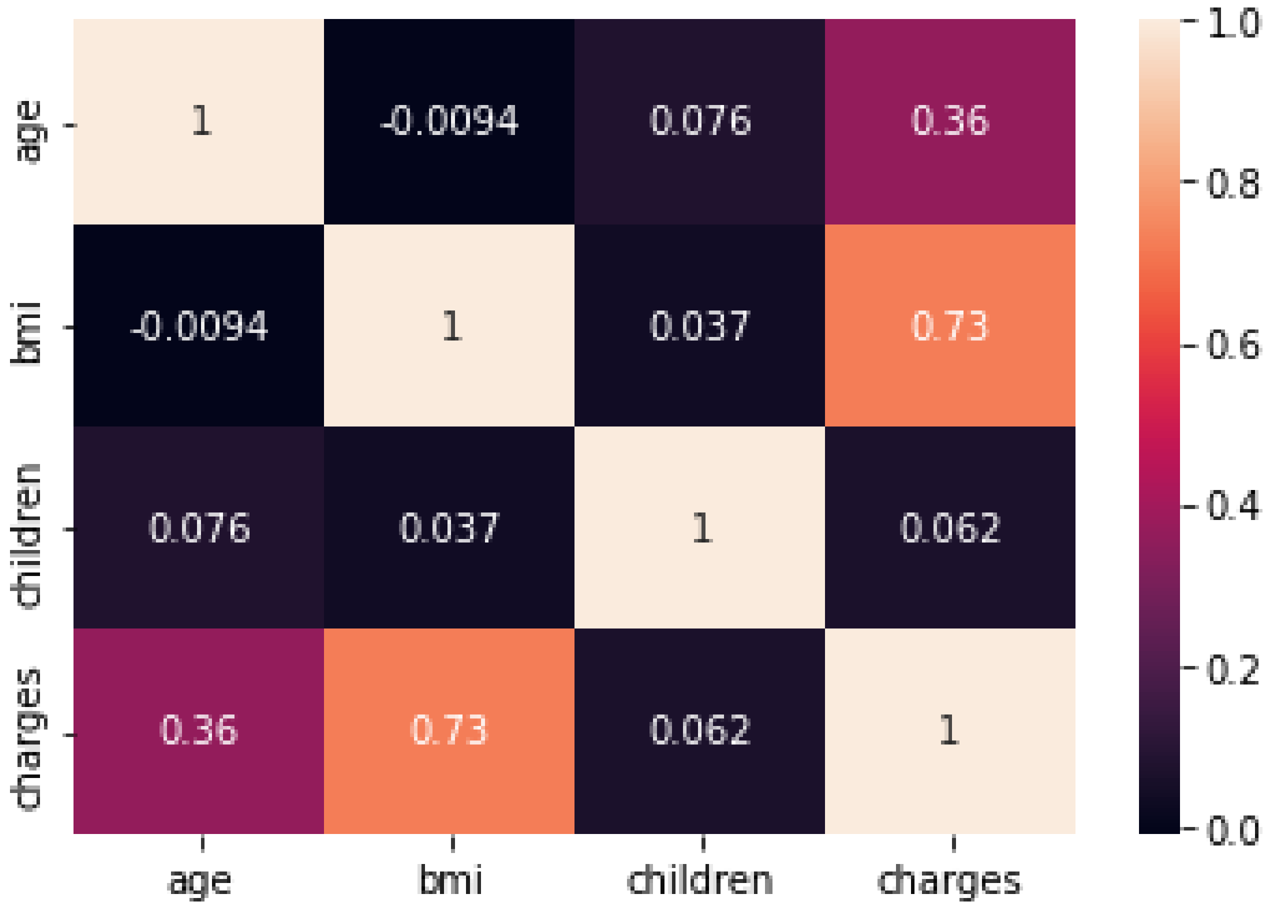
RESULT:

- BMI has a moderate positive correlation with charges.

Correlation Matrix with Smoker, High Charges & BMI



```python
# Create conditional data
condition_2_1 = insurance[(insurance.smoker == 'yes') & (insurance.bmi>=25) & (insurance.charges>=16700)]

# Create Correlation Matrix based on the condition
corrMatrix = condition_2_1.corr()
sn.heatmap(corrMatrix, annot=True)
plt.show()
```



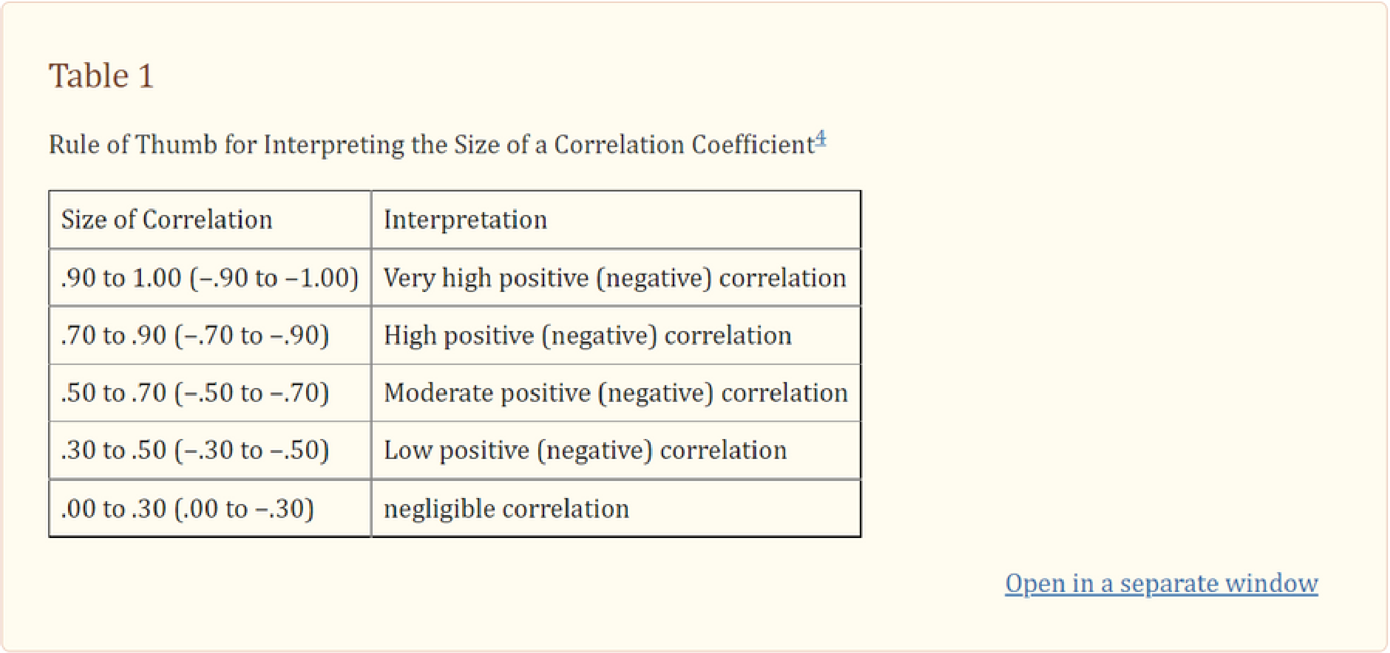


Table 1

Rule of Thumb for Interpreting the Size of a Correlation Coefficient[4]

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Open in a separate window

RESULT:

- BMI has a high positive correlation with charges.
- Age has a low positive correlation with charges

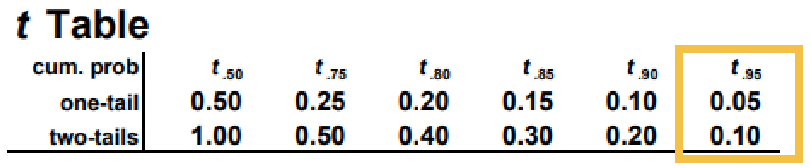
Hypothesis Testing

Smoker’s charges are higher than non smoker’s

Null Hypothesis: A smoker's charges are greater than the non-smoker's charges
Alternate Hypothesis: A smoker's charges are smaller than the non-smoker's charges

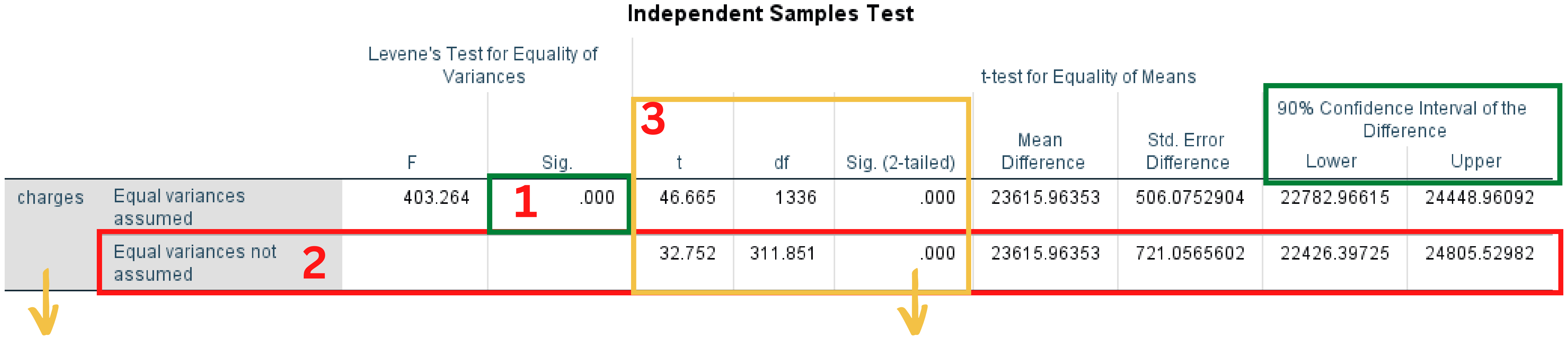
One-tailed
Independent t test

Significant level: 0.05 (One-tailed)



***t* Table**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05
two-tails	1.00	0.50	0.40	0.30	0.20	0.10



Independent Samples Test

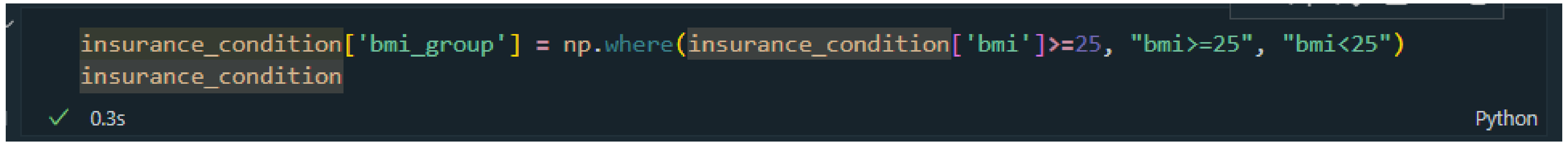
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									90% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
charges	Equal variances assumed	403.264	.000	46.665	1336	.000	23615.96353	506.0752904	22782.96615	24448.96092
	Equal variances not assumed			32.752	311.851	.000	23615.96353	721.0565602	22426.39725	24805.52982

Levene's Test for Equality of Variances

< .00001

RESULT: **The Results are statistically significant. Thus, we can reject the null hypothesis.**

People with High BMI have higher charges than people with Low BMI



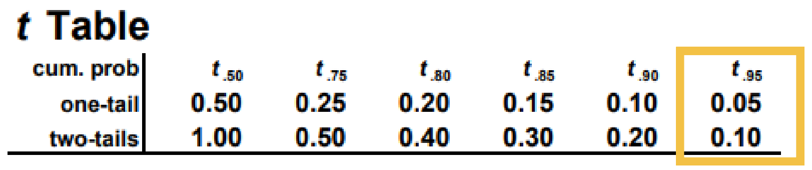
```
insurance_condition['bmi_group'] = np.where(insurance_condition['bmi']>=25, "bmi>=25", "bmi<25")
insurance_condition
```

✓ 0.3s Python

Null Hypothesis: People with High BMI have higher charges than people with Low BMI
Alternate Hypothesis: People with High BMI have lower charges than people with Low BMI

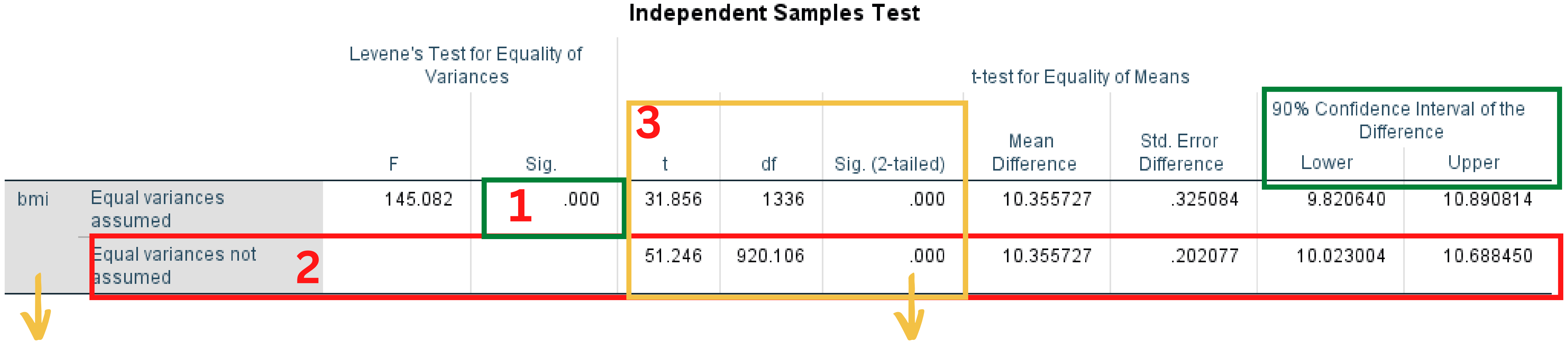
One-tailed Independent t test

Significant level: 0.05 (One-tailed)



t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05
two-tails	1.00	0.50	0.40	0.30	0.20	0.10



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									90% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
bmi	Equal variances assumed	145.082	.000	31.856	1336	.000	10.355727	.325084	9.820640	10.890814
	Equal variances not assumed			51.246	920.106	.000	10.355727	.202077	10.023004	10.688450

Levene's Test for Equality of Variances

< .00001

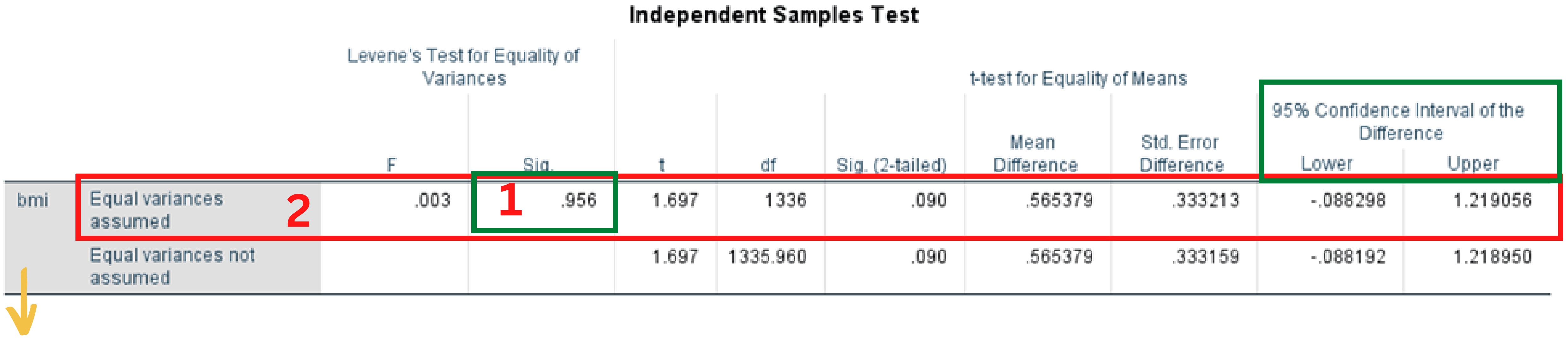
RESULT: The Results are statistically significant. Thus, we can reject the null hypothesis.

Male's BMI is equal to Female's BMI

Null Hypothesis: Male's BMI is equal to Female's BMI
Alternate Hypothesis: Male's BMI is not equal to Female's BMI

two-tailed Independent t test

Significant level: 0.05 (Two-tailed)



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
bmi	Equal variances assumed	.003	.956	1.697	1336	.090	.565379	.333213	-.088298	1.219056
	Equal variances not assumed			1.697	1335.960	.090	.565379	.333159	-.088192	1.218950

Levene's Test for Equality of Variances

RESULT: **The Results are not statistically significant. Thus, we cannot reject the null hypothesis.**

Conclusion

Conclusion

- The average BMI of a smoker is 30.7. Based on the CDC, the average BMI value of 30.7 falls within the obesity class 1 range.
- The Southwest region has the highest proportion of male smokers & their charges.
- The Northwest region has the highest proportion of female smokers & their charges.
- People who have BMI over 25 & charge more than $16.7K need to pay a higher premium.
- Based on the condition that there's a group of smokers who have a BMI over 25 & charge more than $16.7K, the BMI has a high positive correlation with charges

Thus, a group of smokers who have a BMI over 25 & charge more than $16.7K need to pay the highest premium.

Notes

- I use three different application, such as Tableau, Python, and SPSS

Reference

- A guide to appropriate use of Correlation coefficient in medical research
- Defining Adult Overweight & Obesity
- P Value from T Score Calculator