Pacmann

Insurance Charges Data Analysis

Probability Course - Sekolah Data Pacmann

By : Stefanus Yudi Irwan

Date : Thursday 6th October 2022

Outline

Pacmann

1. Introduction
2. Dataset
3. Descriptive Statistic Analysis
4. Categorical Variable Analysis
5. Continuous Variable Analysis
6. Correlation Variable Analysis
7. Hypothesis Testing
8. Analysis Conclusion

Introduction

Introduction

- Insurance companies need to take into account **every aspect of the client's data** to determine the **proper price** of the client's **insurance payment.**
- This insurance payment from the clients will be needed to **pay for health bills** to the **healthcare facility** (ex. hospital, clinics).
- Supervisors of a new insurance agency need help to know how is the **user's condition related** to the **insurance price.** He has some data regarding the user's condition and the insurance price in some regions.

Dataset

Dataset

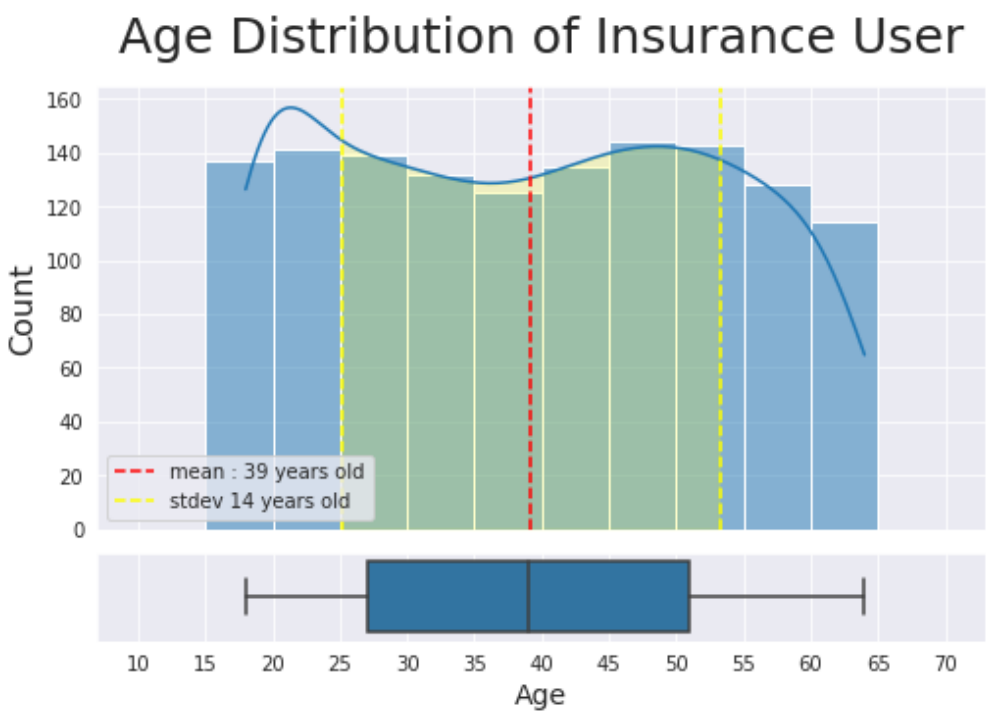
Pacmann

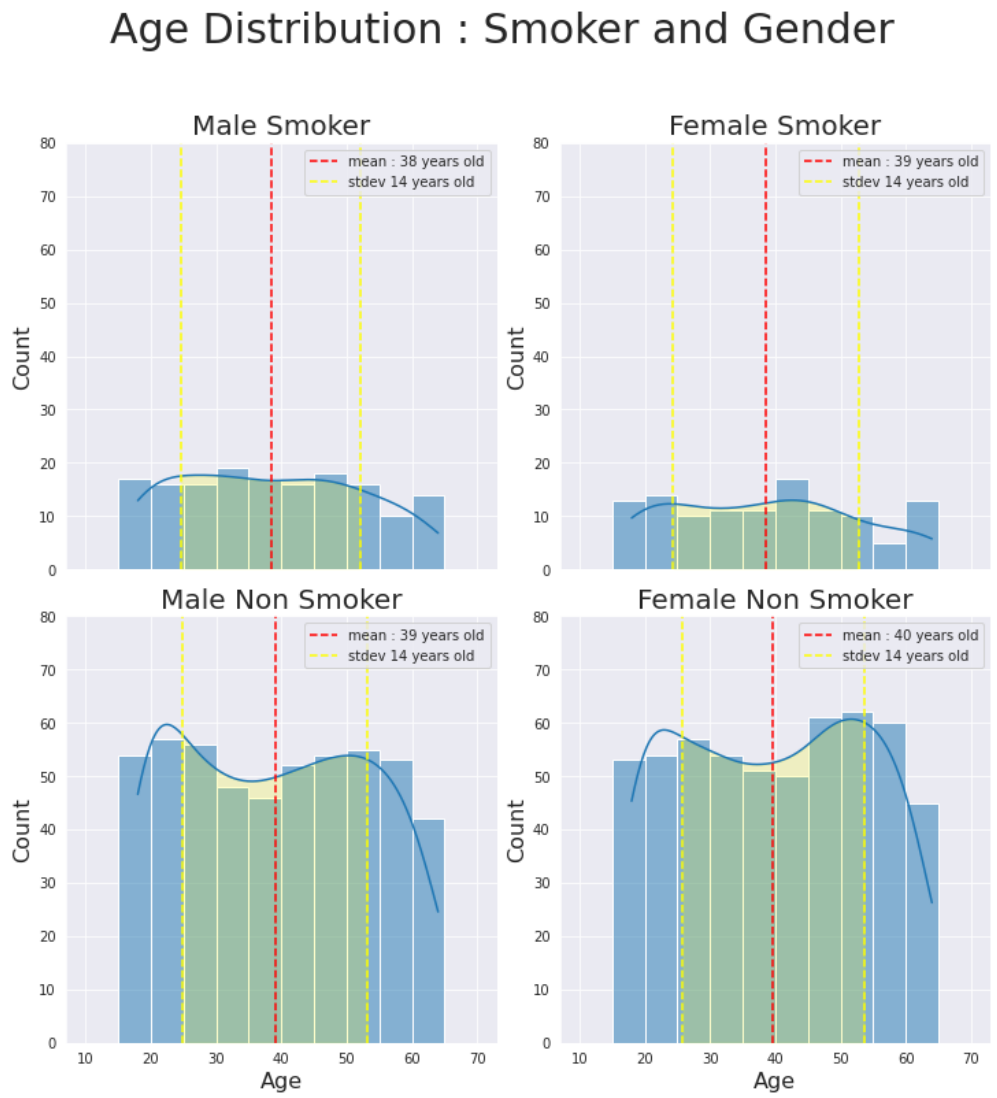
- Dataset comprises **7 columns** and **1338 rows**
- There are **3 categorical data**: Gender, Smoker, and Region
- There are **4 numerical data**: Age, BMI, Children, and Charges

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
5	31	female	25.740	0	no	southeast	3756.62160

Descriptive Statistics Analysis

Analysis on Age

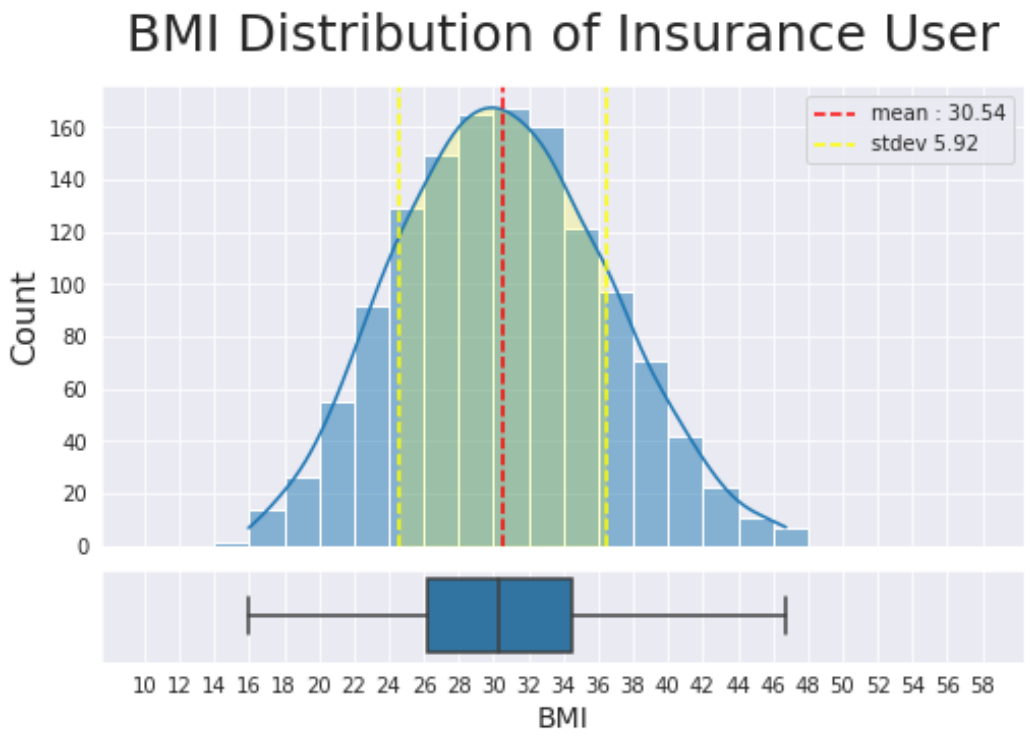


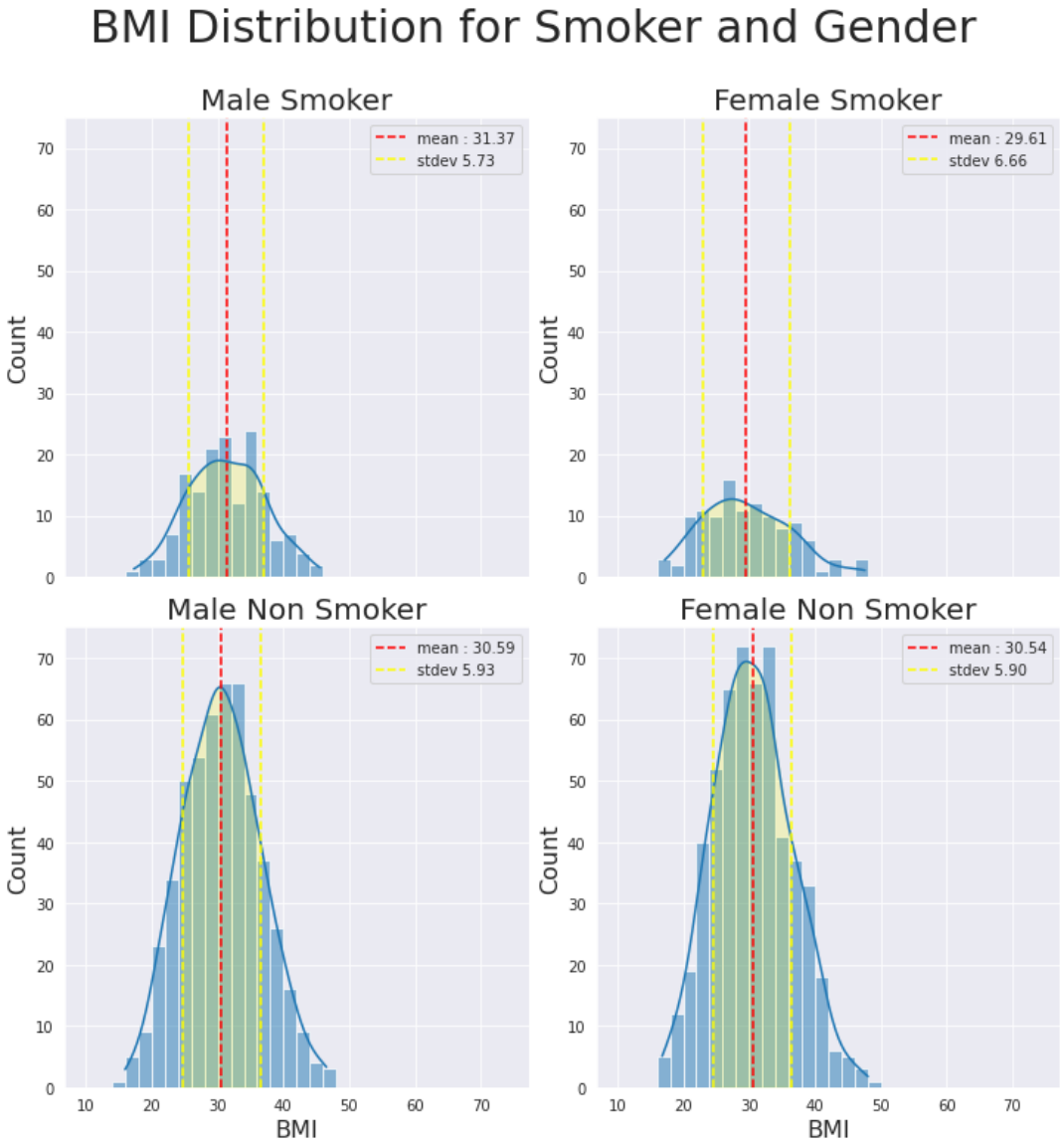


Insight.

- Insurance user age is ranged **from 25 – 51 years old**.
- **Smoker** and **gender** variable are not affecting **age central tendencies.**
- The average age of females is larger than the average age of males by **one year.**
- Smoker people **proportion** is **less than** non-smoker people proportion.

Analysis on BMI

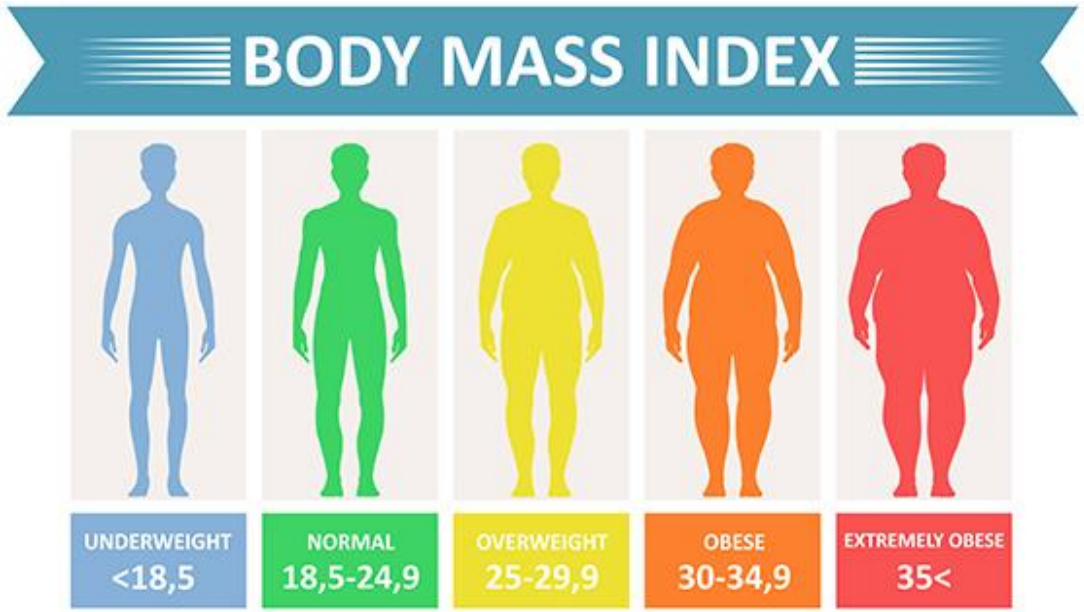




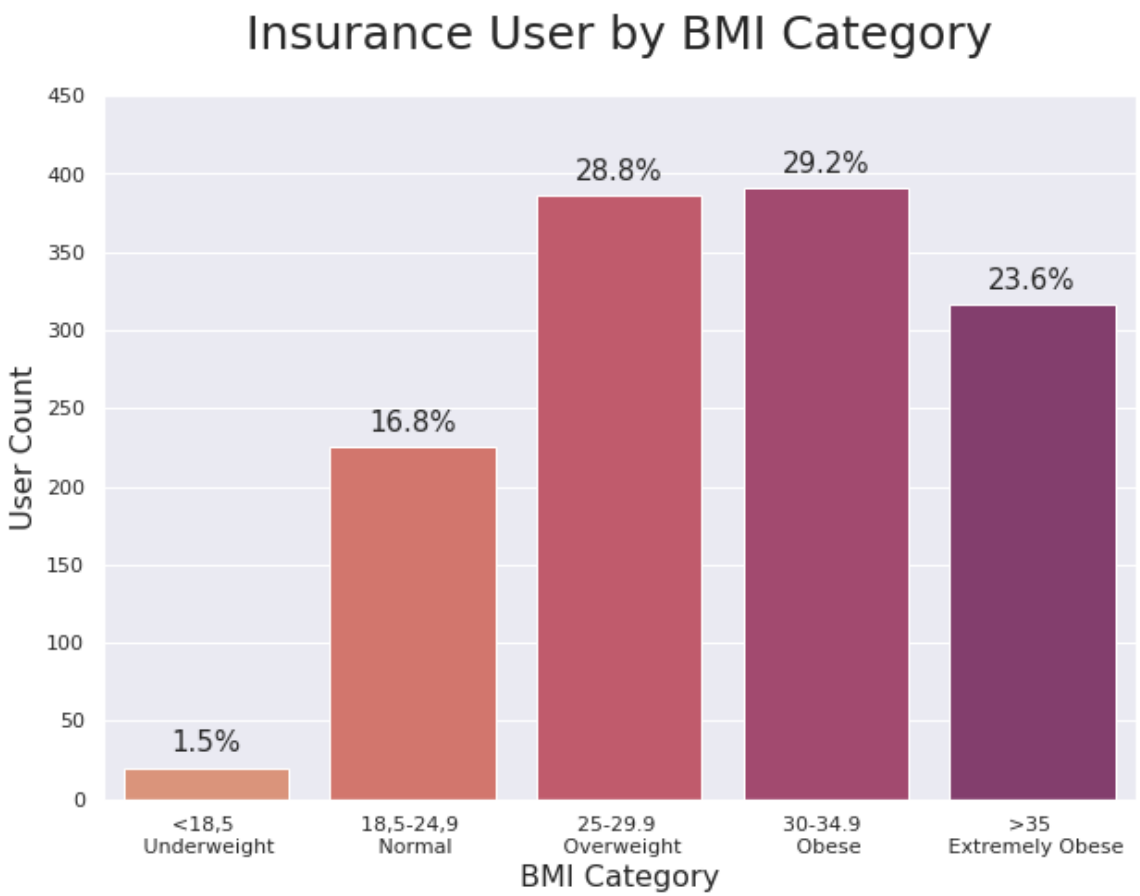
Insight.

- BMI of Insurance users is ranged **from 24.62 – 36.46**.
- **Smoker** and **gender** variable are not affecting **BMI central tendencies.**
- The mean of insurance users lies in **BMI Obese** category**.**

Analysis on BMI



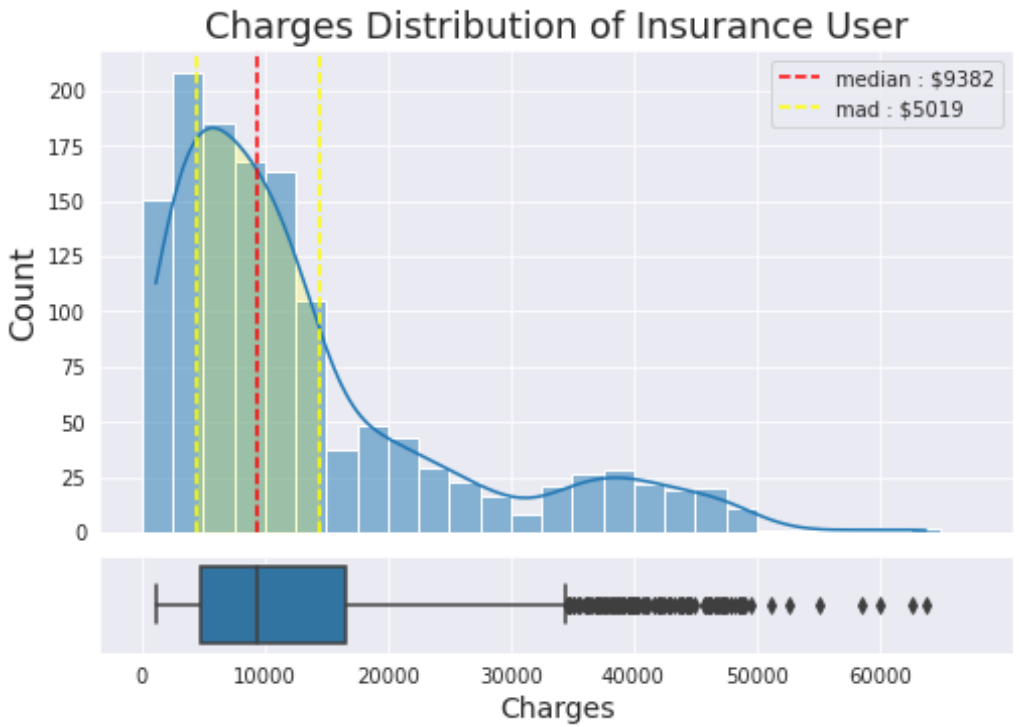
Source : www.cdc.gov

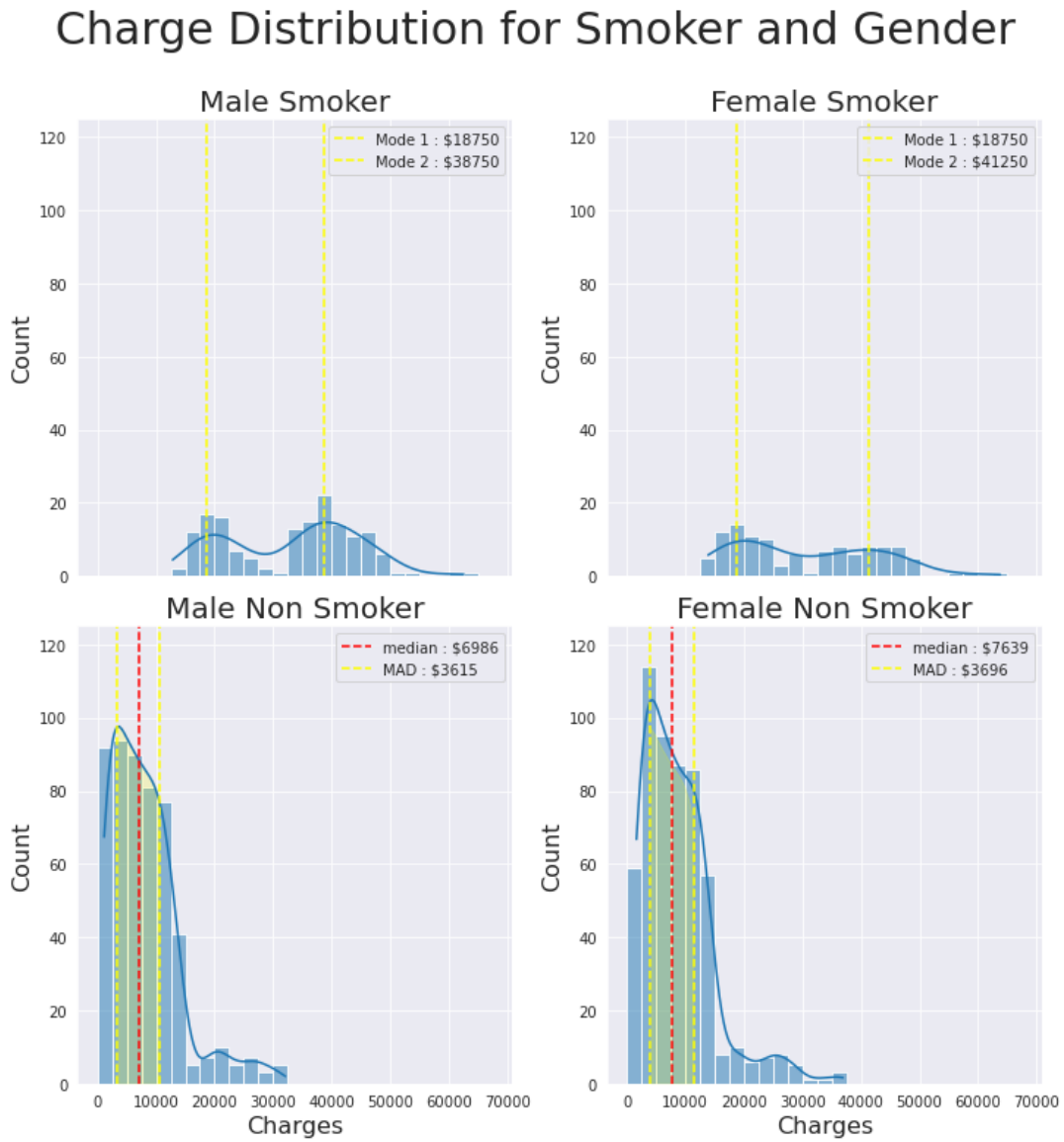


Insight.

- **81,7%** of insurance users is having a BMI of **overweight to extremely obese.** this is the cause why the BMI **average of males, females, smokers, or non-smokers has an obese value.**

Analysis on Charges

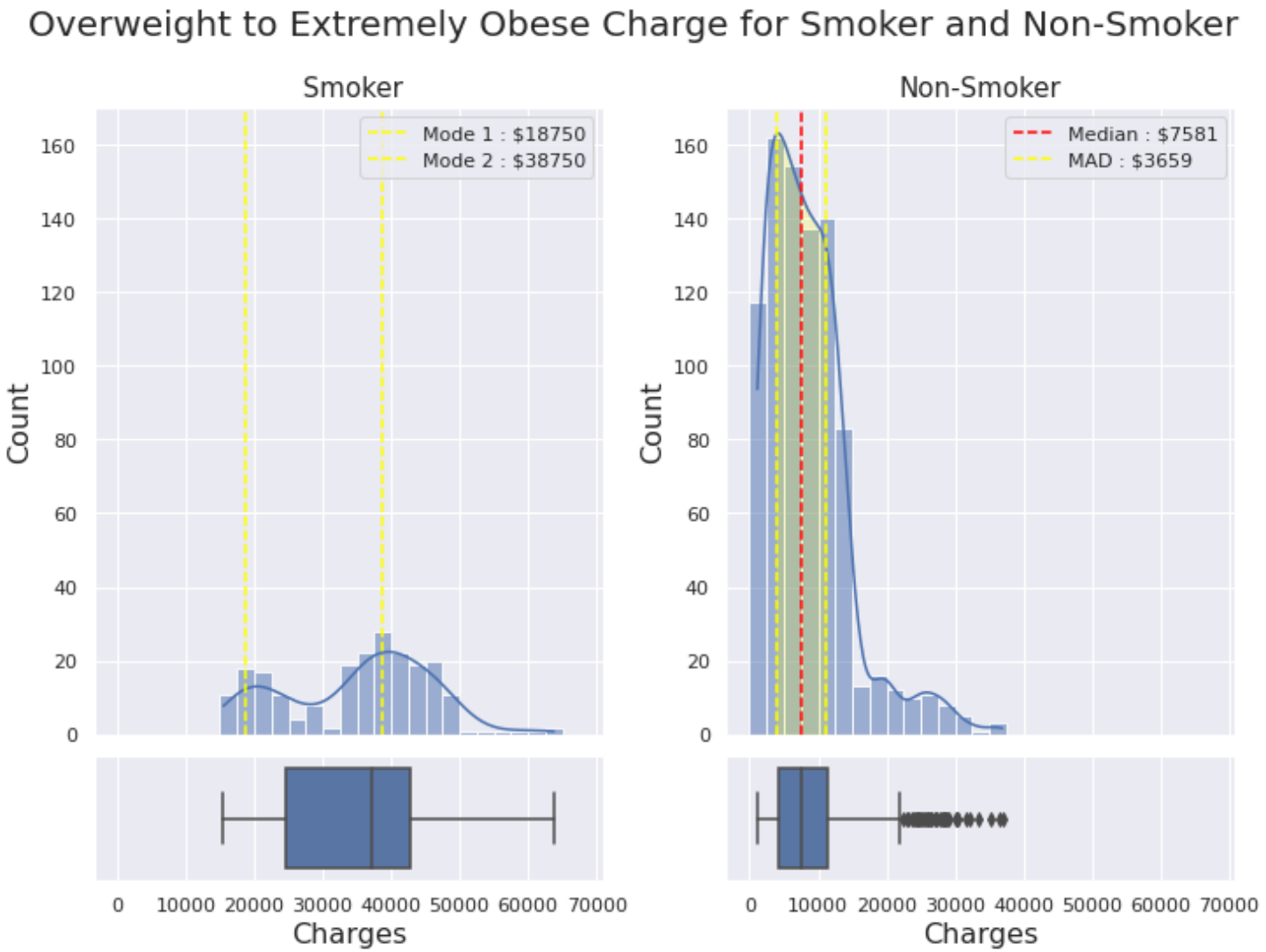




Insight.

- Overall insurance charges value is around **$4363 - $14401.**
- Smoker insurance charges are **more expensive** than non-smoker insurance charges.
- The central tendency of smoker insurance charge value is absolutely larger than **$18000**, whereas the central tendency of non-smoker insurance charge is less than **$10000**.
- For non-smoker people, Female insurance charges are more expensive than male insurance charges, but the difference is not larger than **$1000**.

Analysis on Charges



Insight.

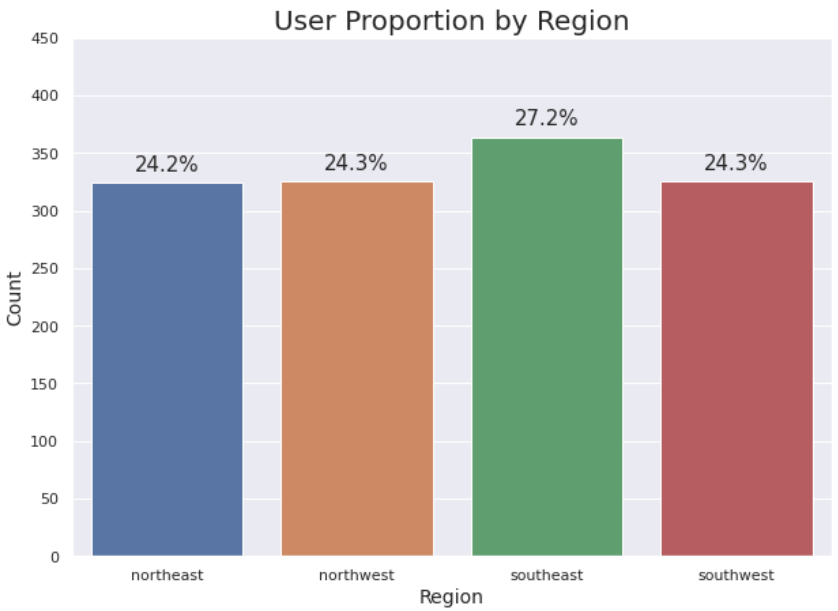
- despite of BMI value, smoker insurance charges is still **more expensive** than non-smoker insurance charges.

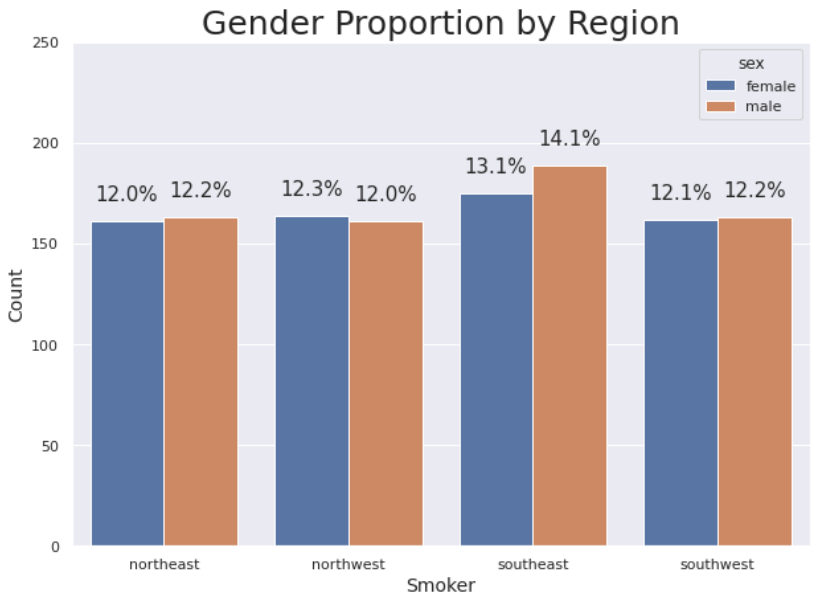
Descriptive Statistic Analysis Summary

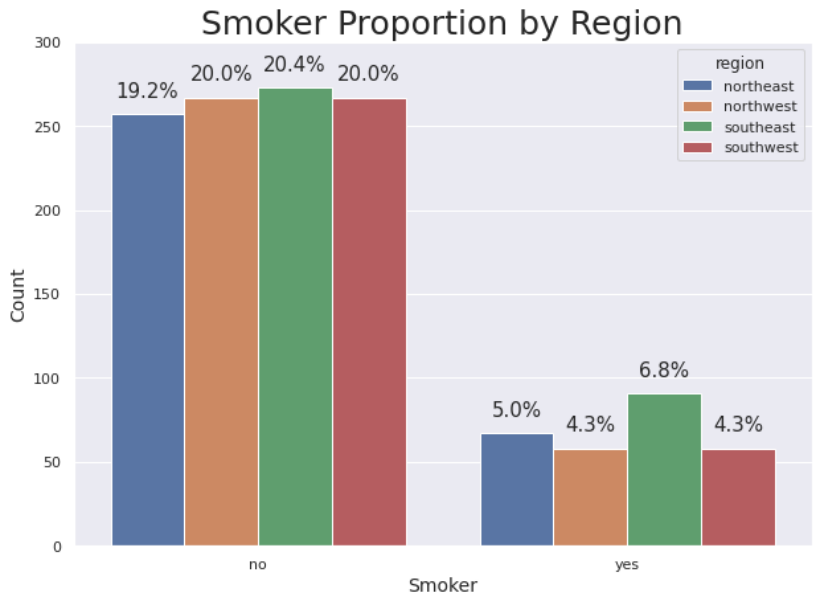
Insurance users come mainly from **productive people**, with age values ranging from **25 to 51 years old**, with an **average** value of **39-40 years old**. The unique part, is those productive people have an **average** value of **BMI** around **30** which lies in the category of **Obese**. After further analysis, I found that **87,1%** of the insurance user is having **BMI from Overweight to Extremely Obese**. This insurance user also covers people who smoke and don't smoke. Because of their habits, the insurance charge of those people will vary, the **most expensive** group of insurance user charges comes from those who **smoke**. The insurance charges for smokers are much more expensive than for non-smokers. The charges values for **smokers are larger than $18000** while charges for **non-smokers are less than $10000**.

Categorical Variables Analysis

Region Analysis



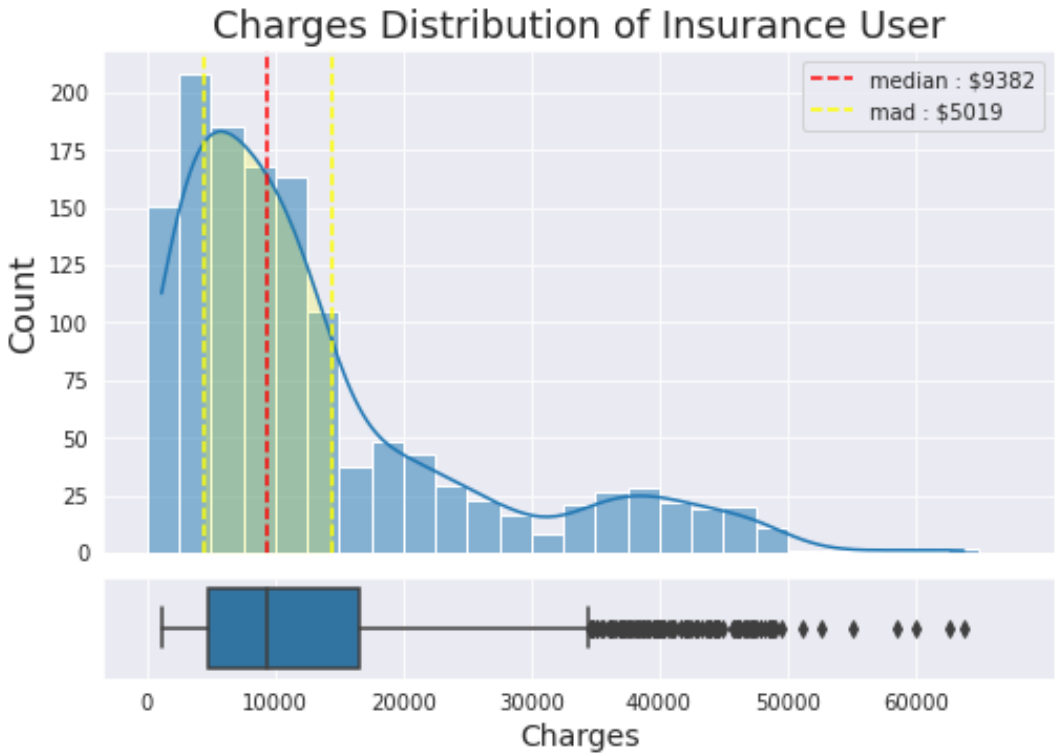


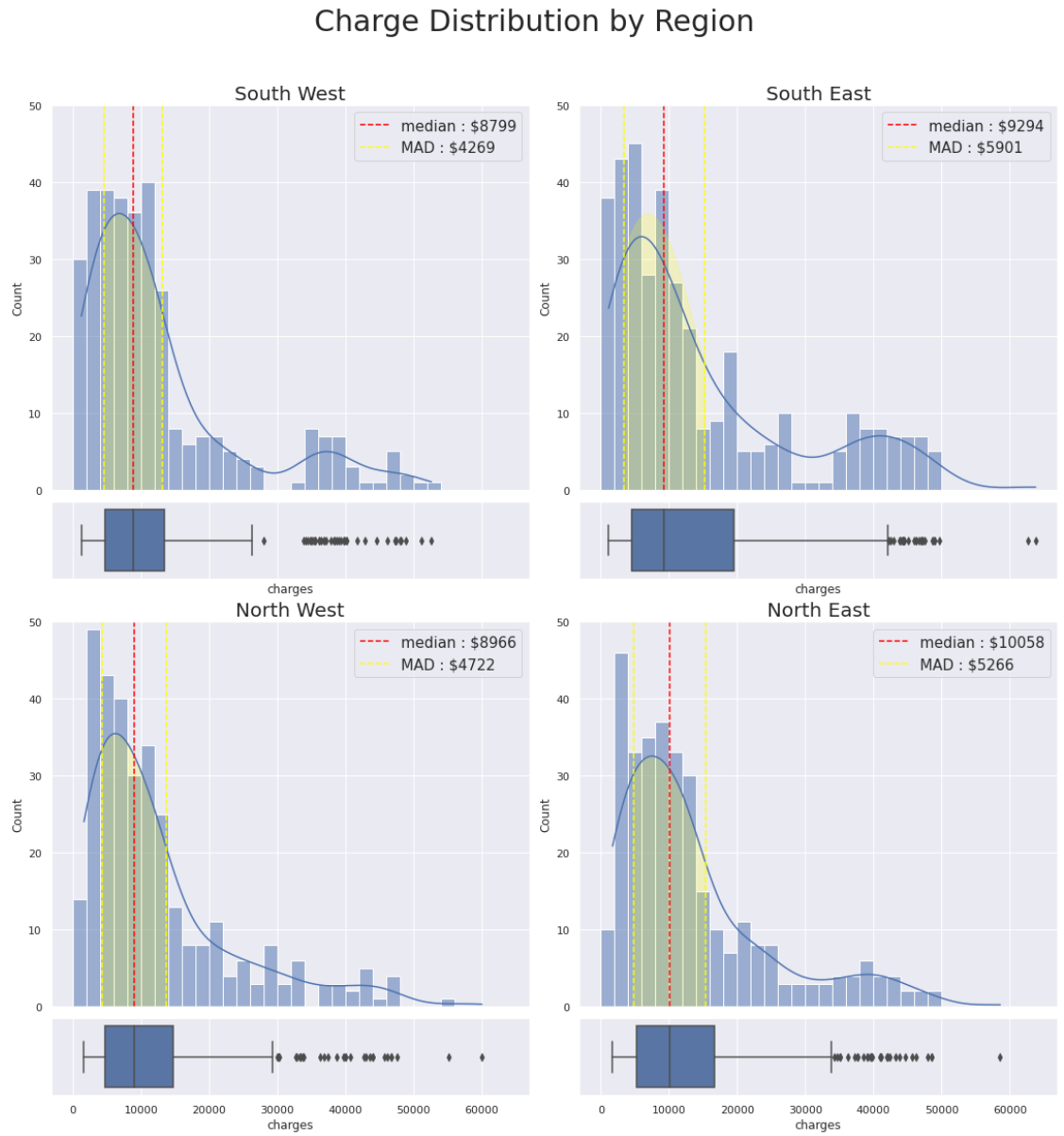


Insight.

- **Northwest**, **southwest**, and **northeast** has the same proportion of user, bar plot showed that the value is 24.2% - 24.3%, whereas the **southeast** has the largest proportion at **27.2%.**
- Every region nearly has a **50:50** female to male ratio, except the **southeast** with a difference of **1%** in gender proportion.
- In **every region** proportion of **non-smokers** is **larger** than the proportion of **smokers**. The ratio of smokers to non-smokers is around **1:4**.
- **Southeast** has the largest proportion of smoker compared to other regions, whereas **northwest and southwest** has the **same head count of smokers**.

Region Analysis



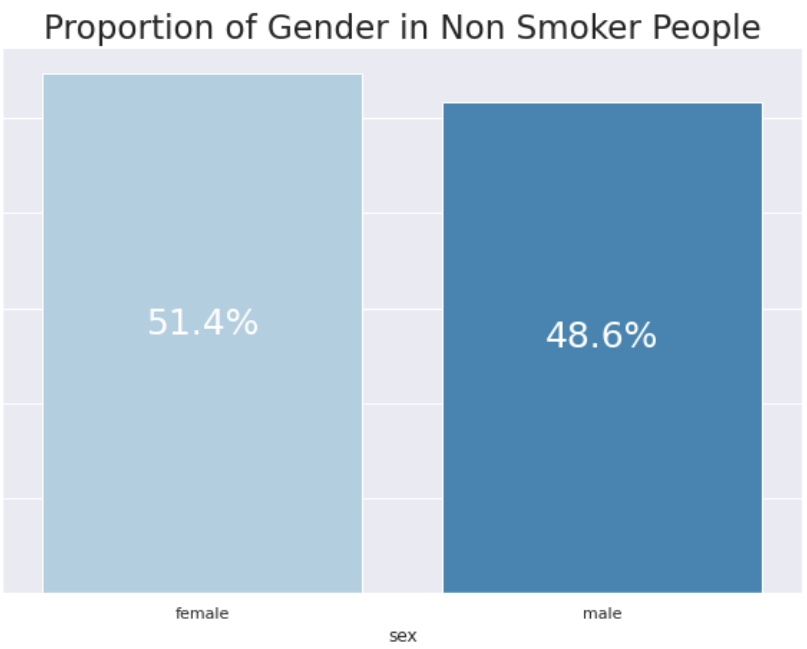


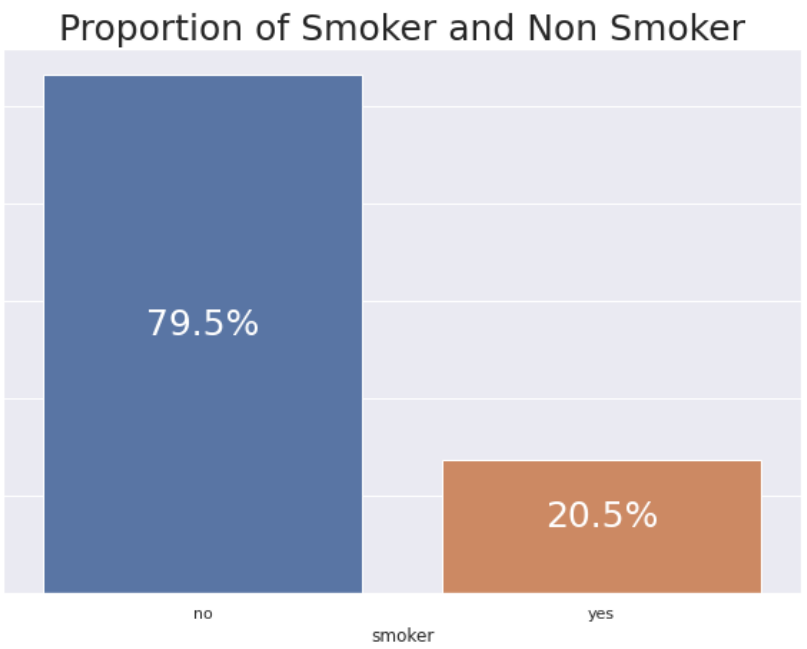
Insight.

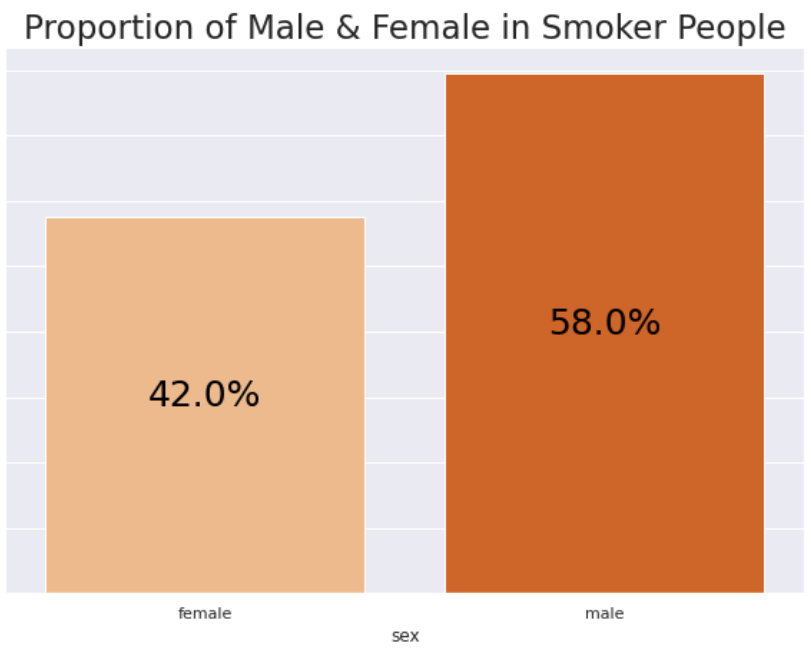
- Distribution charge from every region shaping similar shape, **positively skewed distribution**.
- North East became the most expensive charges with median **$10058**, and southwest became the cheapest charges with median **$8799**.

Smoker and Gender Analysis

Pacmann







Insight.

- Smoker proportion is much smaller than non-smoker proportion, **20.5% compare to 79.5%**.
- In the smoker group, the male has the largest proportion **58%** compared to female **42%**.
- In non-smoker group, female has the largest proportion **51.4%** compared to male **48.6%**.

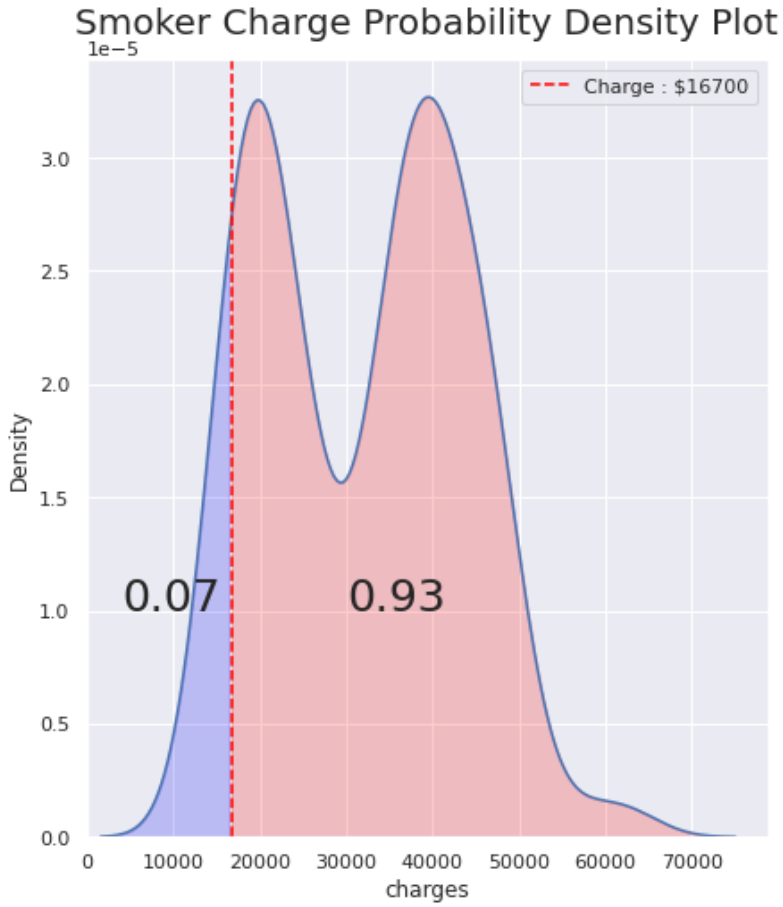
Categorical Variable Analysis Summary

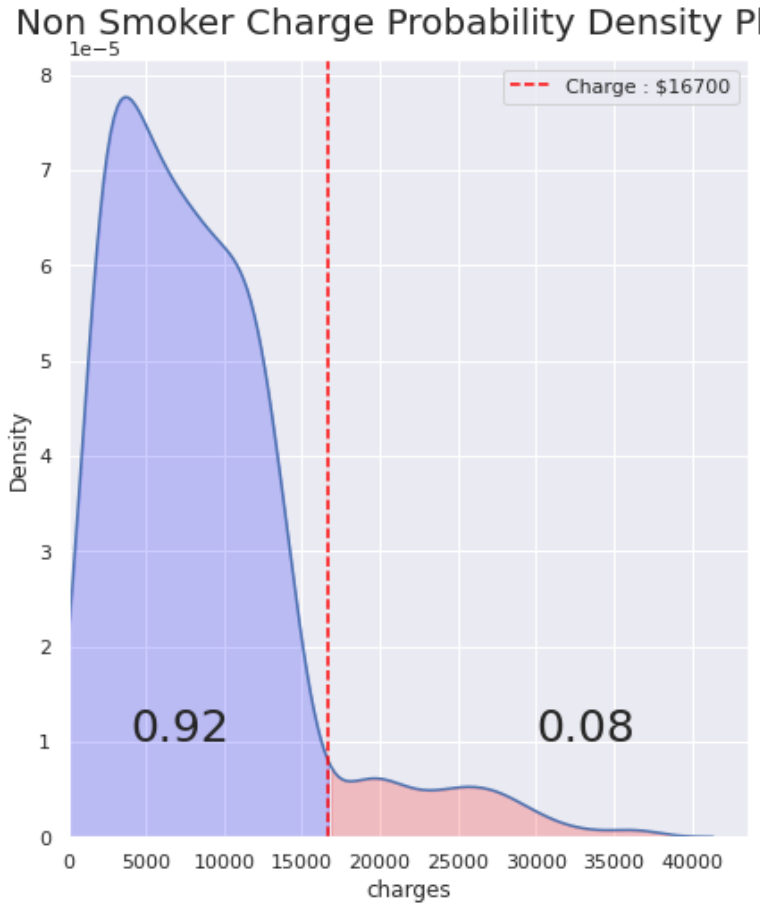
Every region has a nearly equal amount proportion of insurance users, **24.2-24.3%**, a slight difference exists in the southeast region with a proportion of **27,2%** of all insurance users. Every region also has a nearly equal amount of gender proportion between male and female, **50:50**, a slight difference also exists in the southeast region with a proportion difference of **1%** between male and female. The data confirm that the proportion of smoker people is less than non-smoker people, the ratio of the smoker and non-smokers in every region is around **1:4.** Within **s**moker people, the male proportion is larger than female, 58% compared to 42%, while in non-smoker people, the female proportion is larger than male, 51.4% compared to 48,6%**.** The northeast region became the region with the relatively expensive amount of insurance charge of around **$100508**, whereas the southwest region become the cheapest region with an insurance charge of around **$8799.**

Continuous Variables Analysis

Smoker Probability Charge Analysis

Pacmann



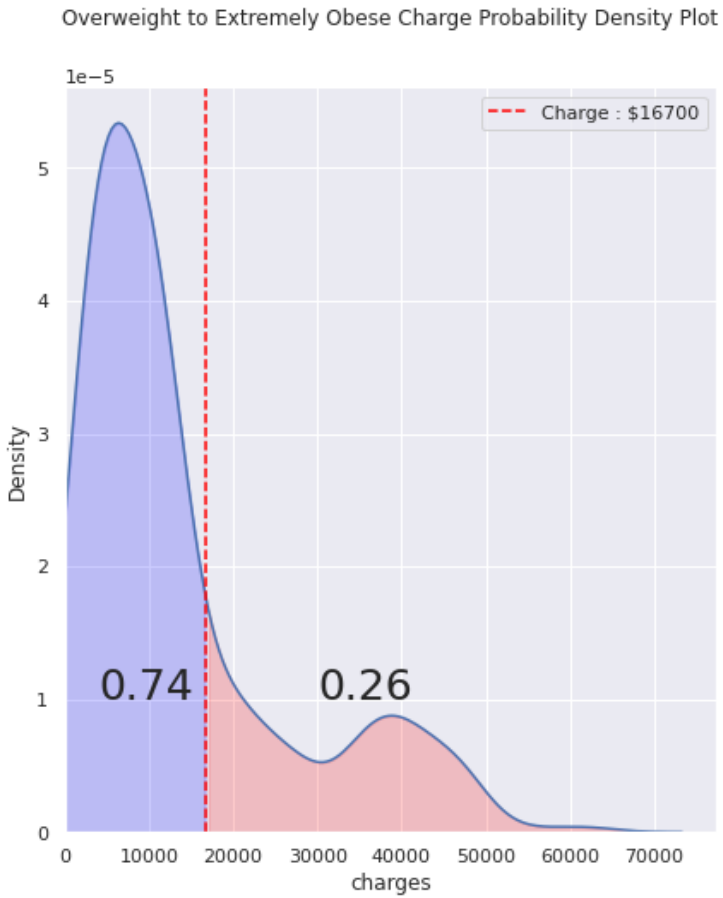


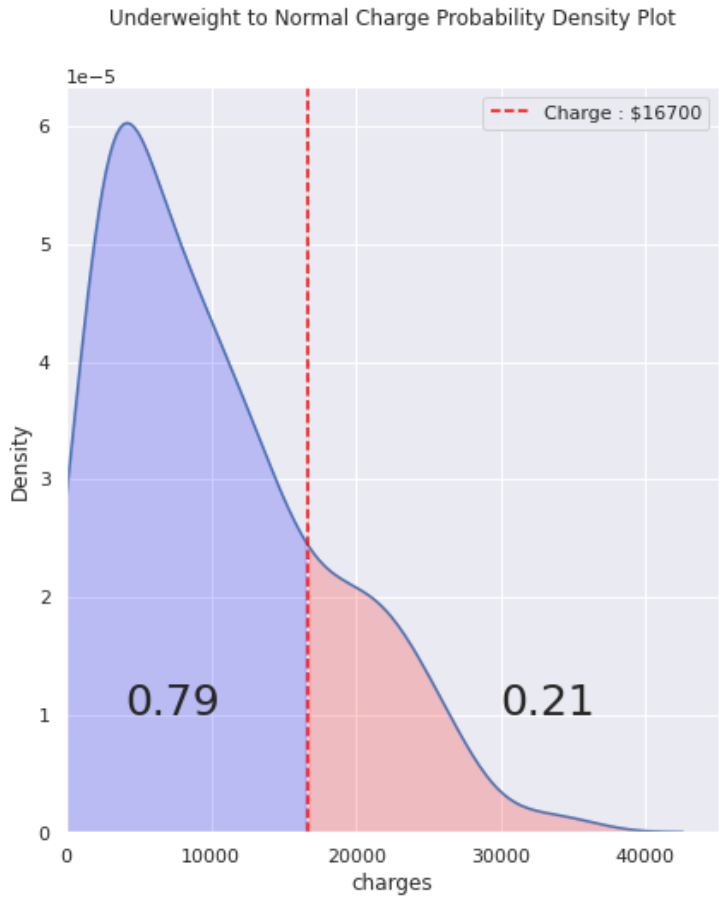
Insight.

- It is **very possible** for **smokers** to get insurance charges **more than $16700,** the probability is **0.93**. And it is **rare** that **smokers** will get insurance **charges below $16700,** the probability is just **0.07.**
- it is **rare** for **non smokers** to have an insurance charge for **more than $16700**. But it is **common** for **non-smoker** to have insurance charges below **$16700**, the probability is **0.92**.

BMI Probability Charge Analysis

Pacmann

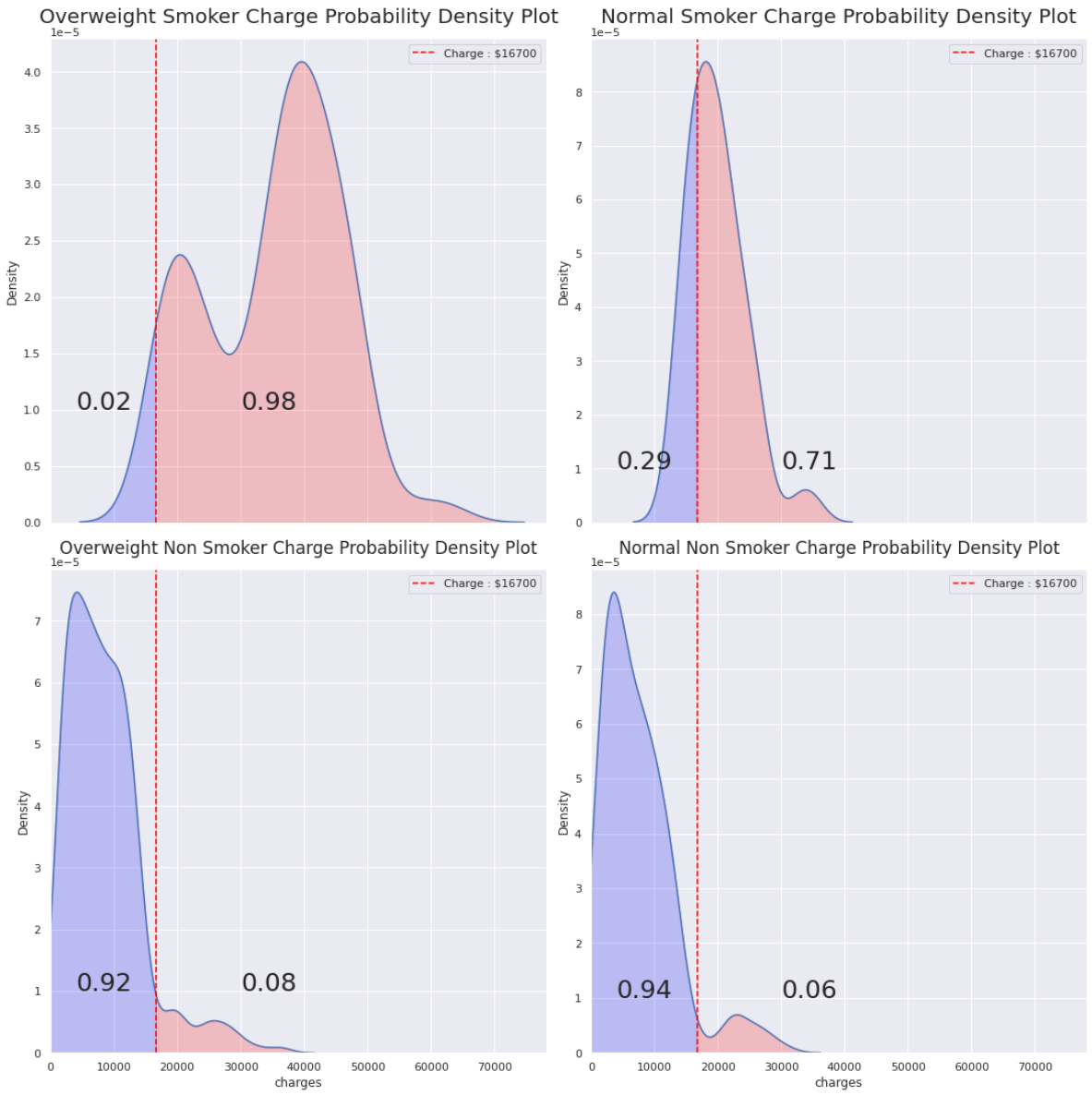




Insight.

- It is more possible for people with **BMI >= 25 (Overweight to Extremely Obese)** to get the insurance charge **below $16700** because the probability is **0.74**.
- Also It is more possible for people with **BMI < 25 (underweight to normal)** to get an insurance charge **below $16700** because the probability is **0.79**.

Smoker and BMI Probability Charge Analysis



Insight.

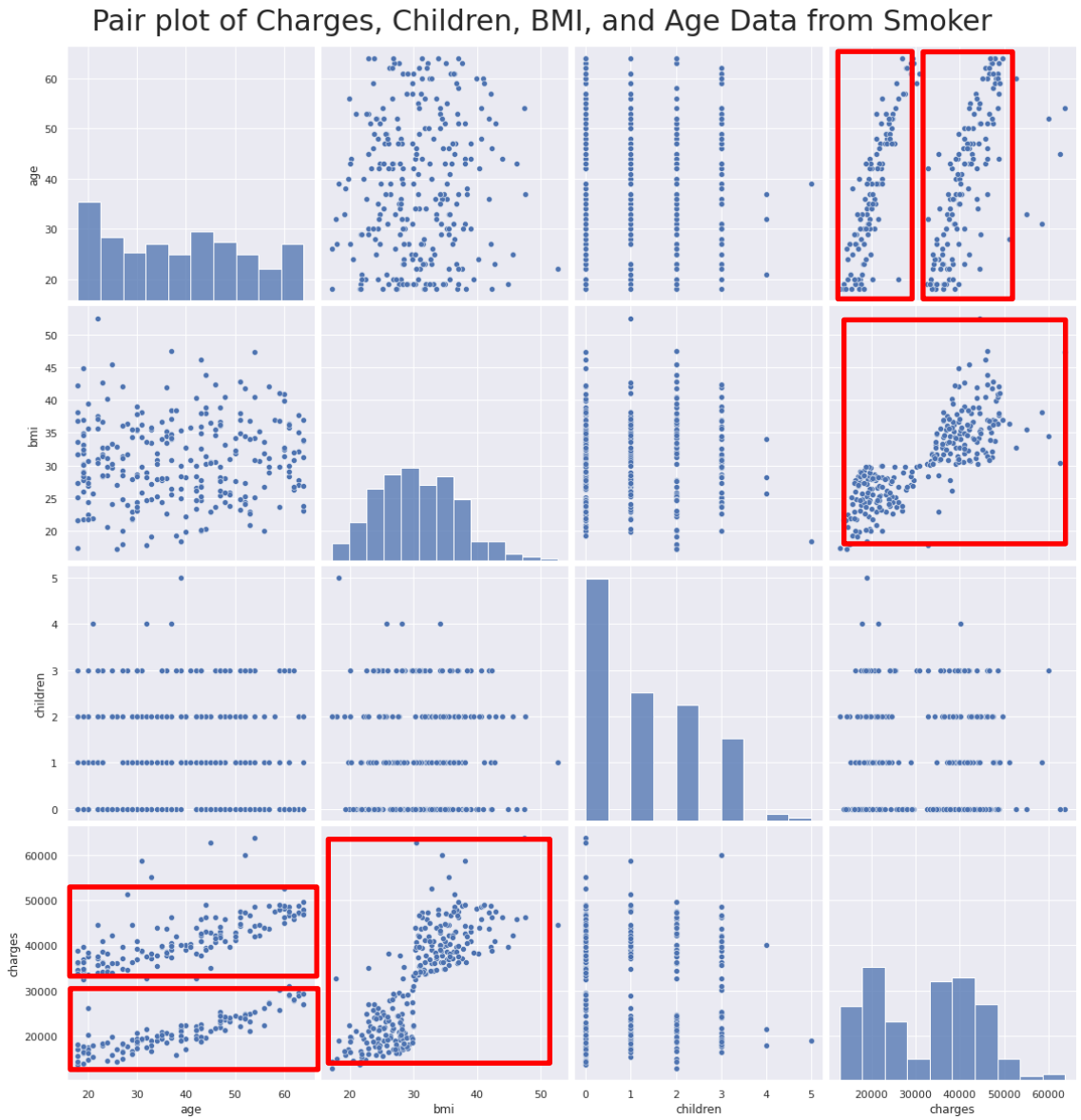
- **smoking** really increases the probability of a person getting an insurance charge for **more than $16700**.
- **overweight** also increases the probability to get an insurance charge for **more than $16700** but **not as strong as smoking**.

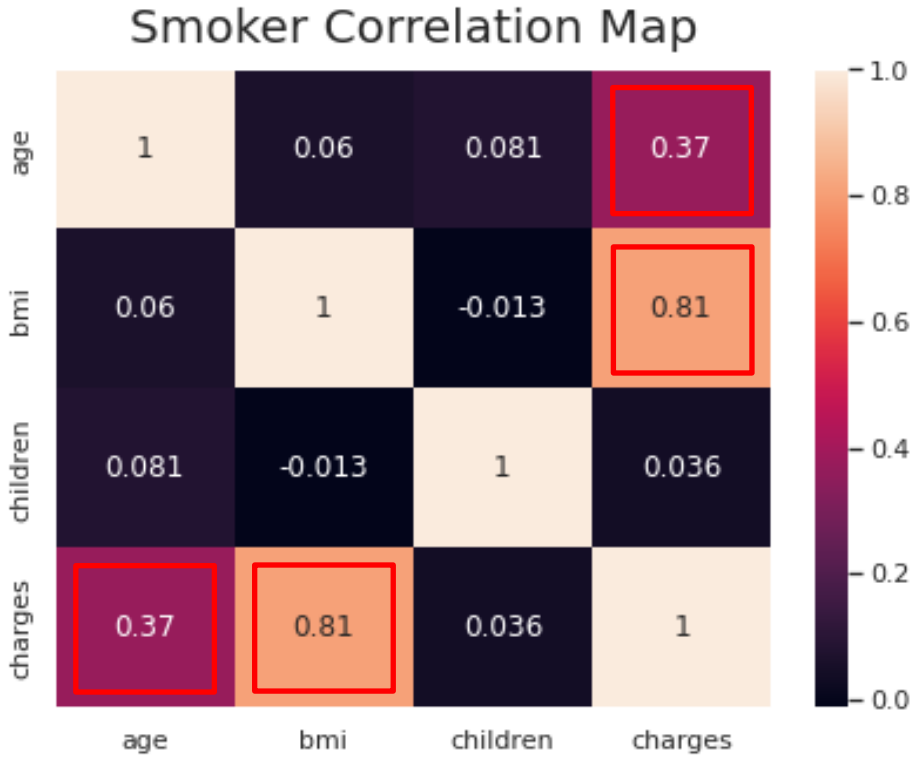
Continuous Variable Analysis Summary

The **probability of a smoker** getting an insurance charge of **more than $16700** is very large, **more than 0.7**, without taking into account his/her BMI factor. **BMI Factor** also contributes to the expensiveness of insurance charges, but the effect is **not as strong** as if he/she **smokes**. For smokers, the difference of probability between a normal BMI smoker, and with overweight BMI smoker is **0.27,** while for non-smokers the difference of probability between a normal BMI smoker, and with overweight BMI smoker is just **0.02**, this difference is very significant, thus smoking is a great factor affecting the insurance charges.

Correlation Analysis

Smoker Correlation Analysis

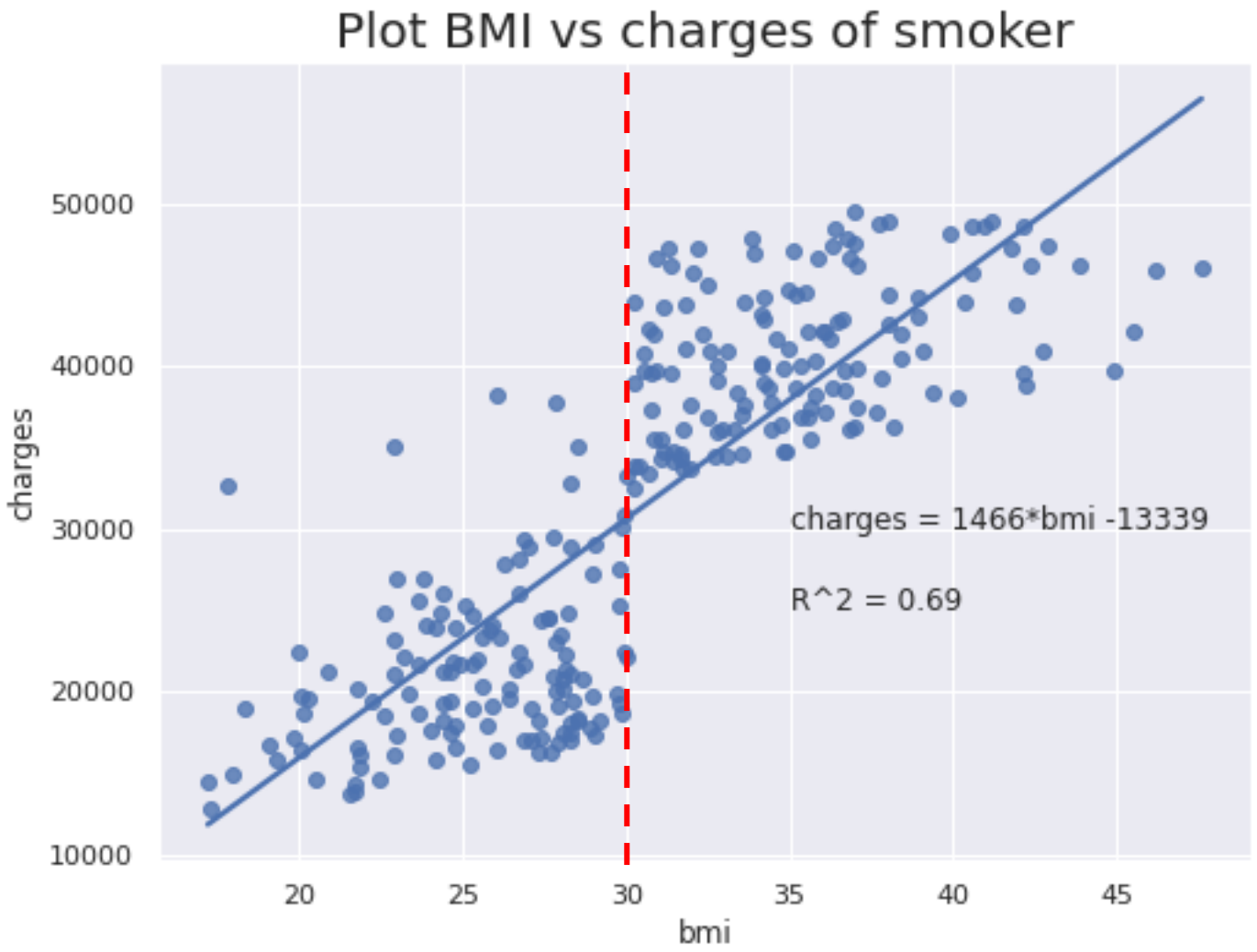




Insight.

- BMI and Charge has a strong correlation, **0.81**.
- there are positive weak correlation, **0.37**, between age and charges, but notice there are pattern in the scatter plot

Smoker Correlation Analysis

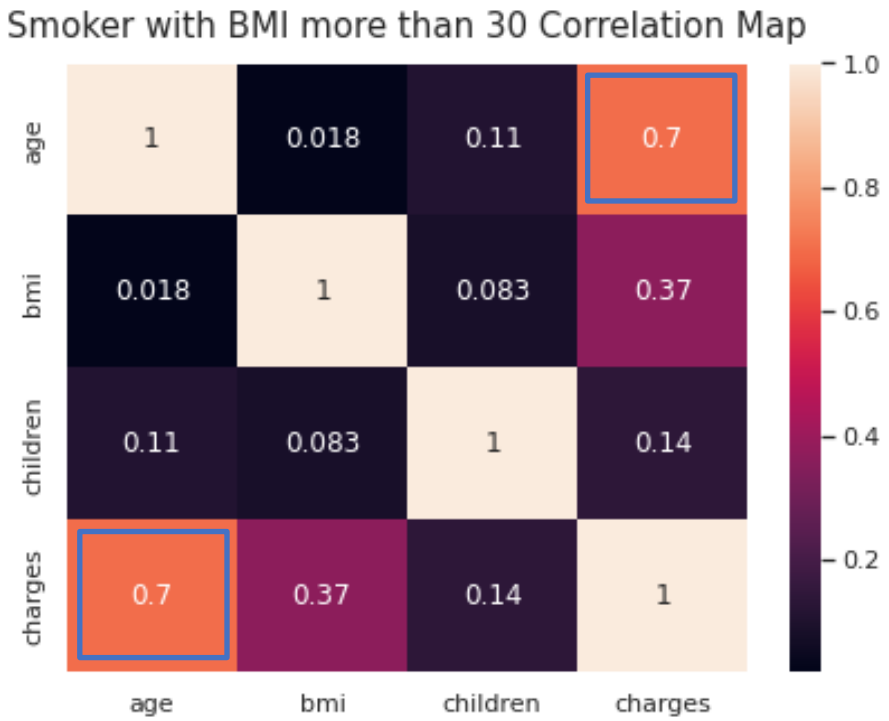


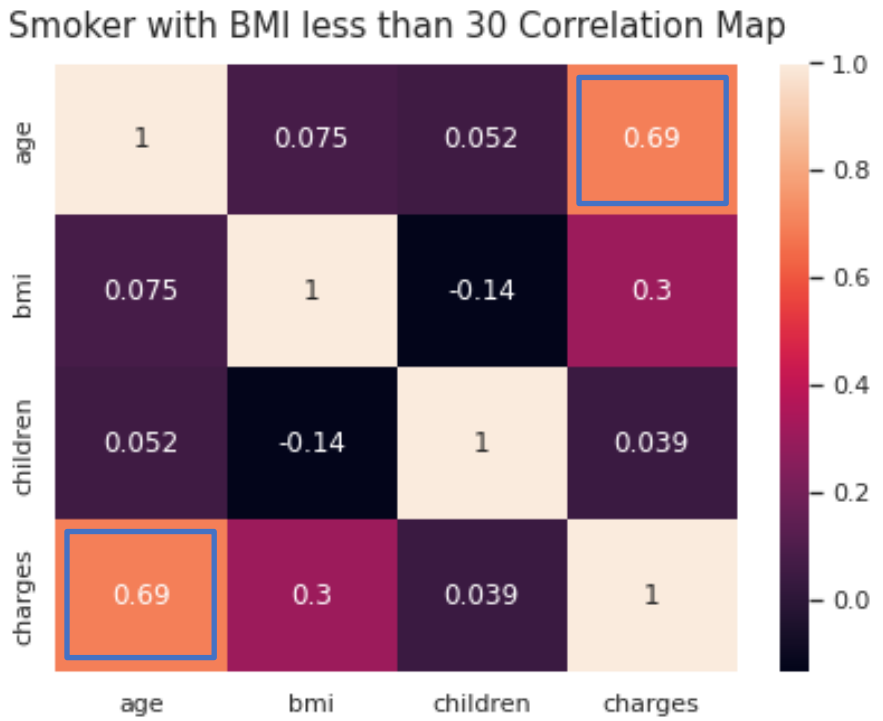
Insight.

- For smoker if the **BMI** of insurance user increase by **one value**,
  the **charges** will **increase by $1466**

Smoker Correlation Analysis

Pacmann



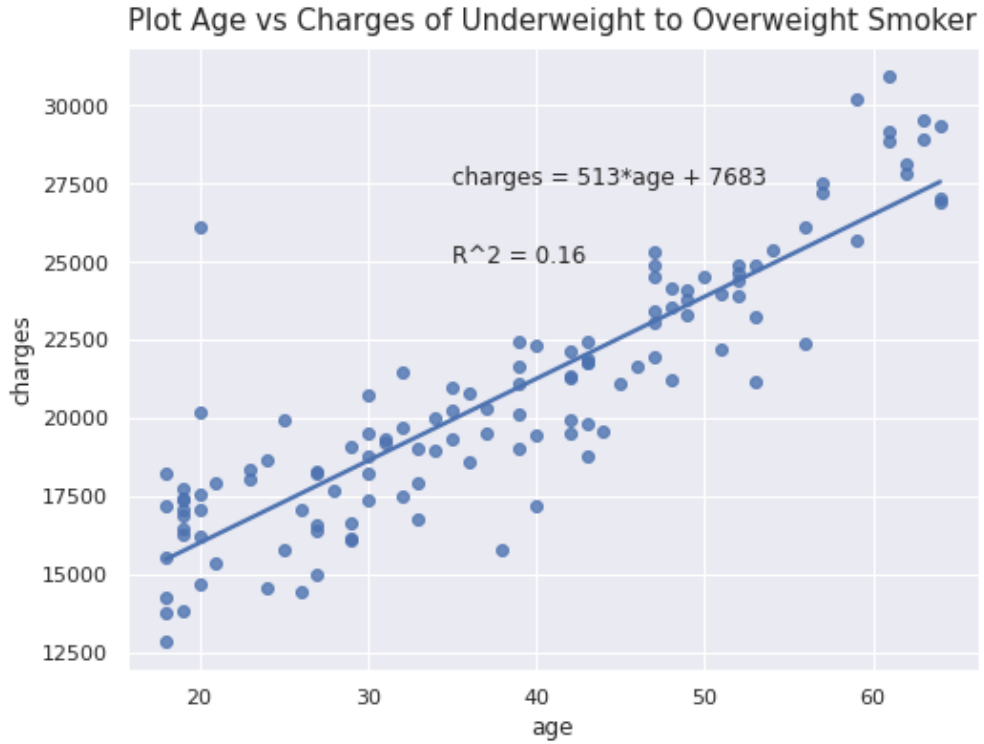


Insight.

- After separation using **BMI = 30** we could see that the **correlation value** from age and charge **increased** to **0.7**, this is a **strong correlation**. This could mean that there are indeed a relationship between **age and charges**.

Smoker Correlation Analysis

Pacmann

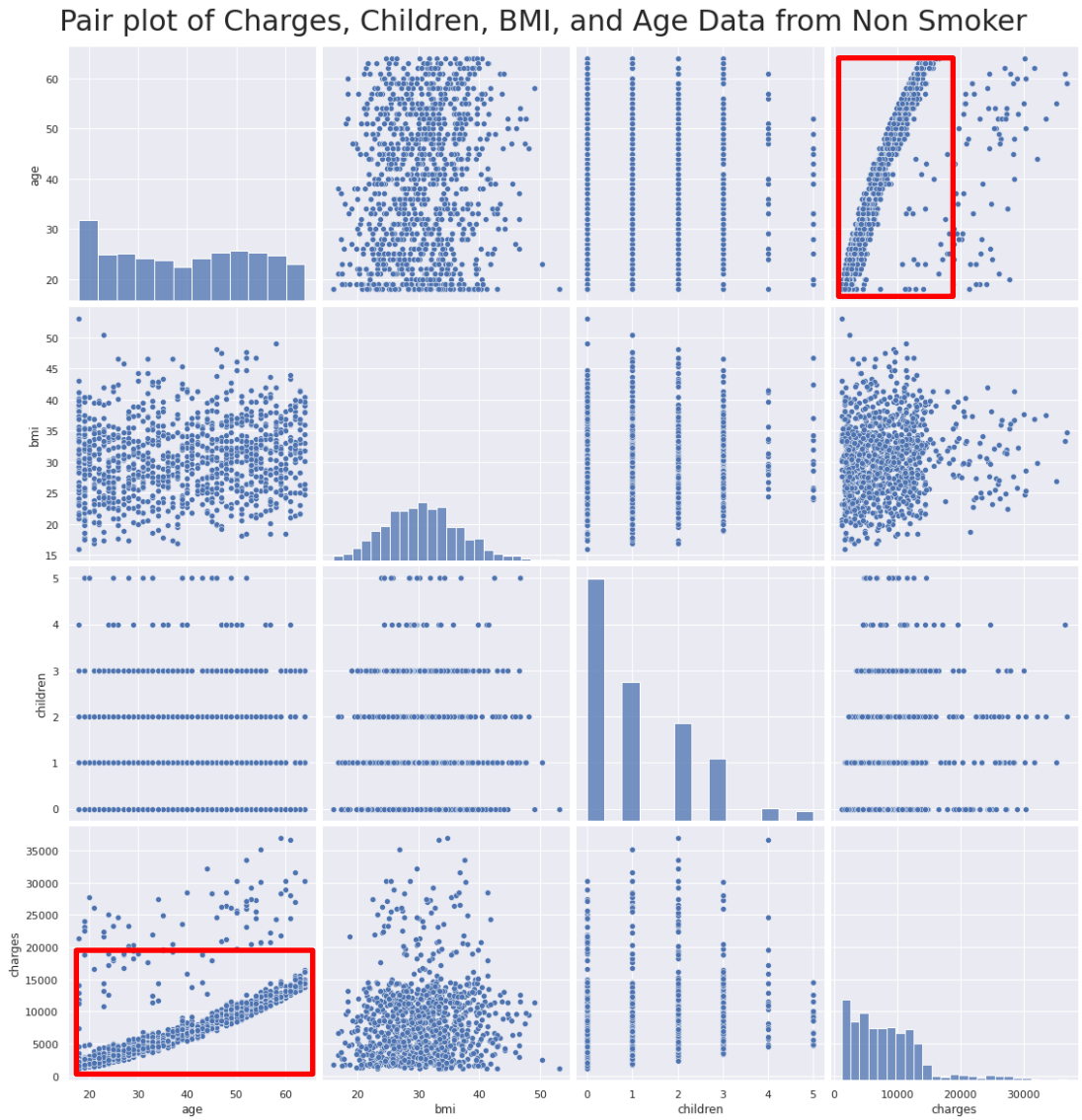


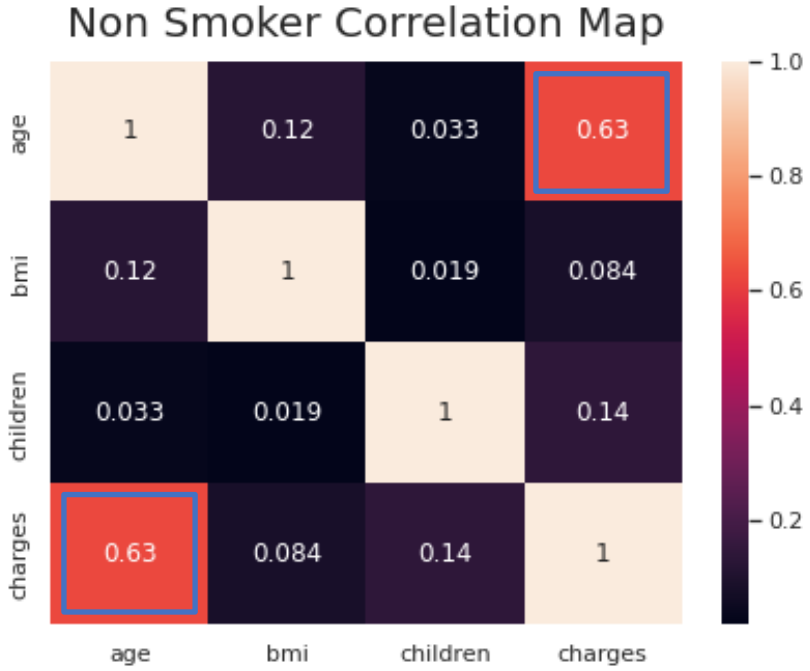
Insight.

- For smokers with BMI > 30, if the age was **increased** by **one value**, there will be an **increase in insurance charges** by **$480**. With charges ranged from **$32500 to $50000**.
- For smokers with BMI <= 30, if the age was **increased** by **one value**, there will be an **increase in insurance charges** by **$513**. With charges ranged from **$12500 to 32500**. This finding confirms analysis before, people who smoke and have a high BMI tend to have an expensive insurance charge.

Non Smoker Correlation Analysis

Pacmann





Insight.

- there are positive strong correlation between **age vs charge** for non-smoker, with a correlation value of **0.63**

Non Smoker Correlation Analysis

Insight.

- For non-smokers, if the **age** increased by **one value**, there will be an increase in charge by **$70 dollars**, this is relatively a low value to smoker data. With charges value ranged from around **$1000 to $16000.**
- Compared to the smoker linear regression minimum value, non smoker indeed has cheaper insurance charge.

Correlation Analysis Summary

There is a correlation between smokers, non-smokers, BMI values, and age of insurance users to insurance charges that they have. Correlation analysis shows that people who smoke and have a high value of BMI tend to have an expensive insurance charge, the minimum charge value for them is **$32500** and the value will increase by **$480** dollars if the insurance user gets older. Smoker with middle to low-value BMI will have relatively cheaper insurance charges than smokers with high BMI values, the minimum charge for them is **$12500** and the value will increase by **$513** dollars if the insurance user gets older. Non-smokers, compared to smokers, have really cheap charges, their minimum value is around **$1000** and the charges will increase by just **$70** dollars if they get older. Then we could know that smoking and BMI really affects the expense of insurance charge.

Hypothesis Testing

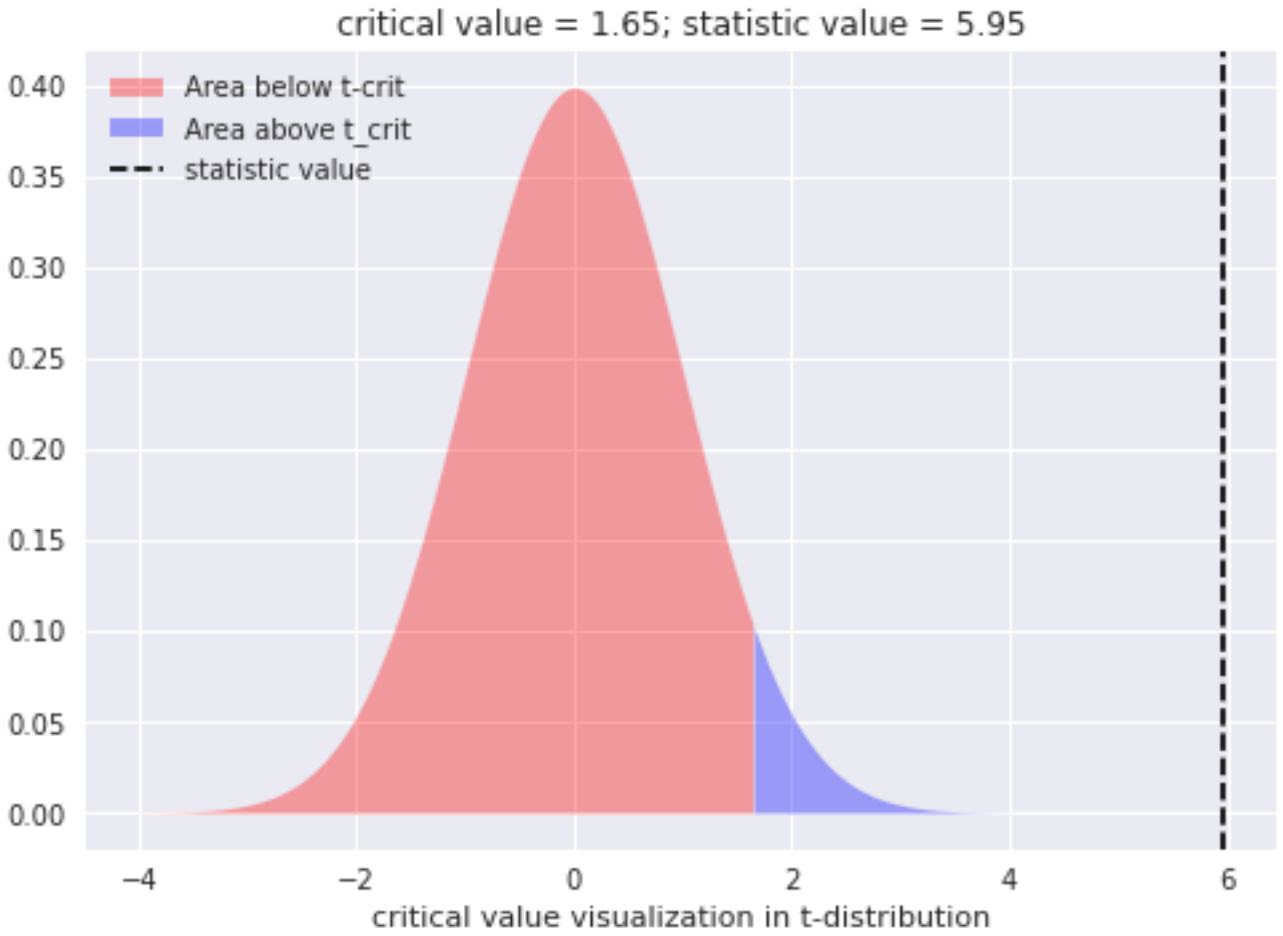
Smoker's charges are higher than non-smoker's

$$H_0 : \mu_{Charge|Smoker} <= \mu_{Charge|Non-Smoker}$$
$$H_1 : \mu_{Charge|Smoker} > \mu_{Charge|Non-Smoker}$$

Insight.

- **Reject H0**, hence in a population of insurance users, we are 95% confident that smoker's insurance charges are higher than non-smoker insurance charges

Insurance charges with BMI > 25 is higher than insurance charges with BMI < 25

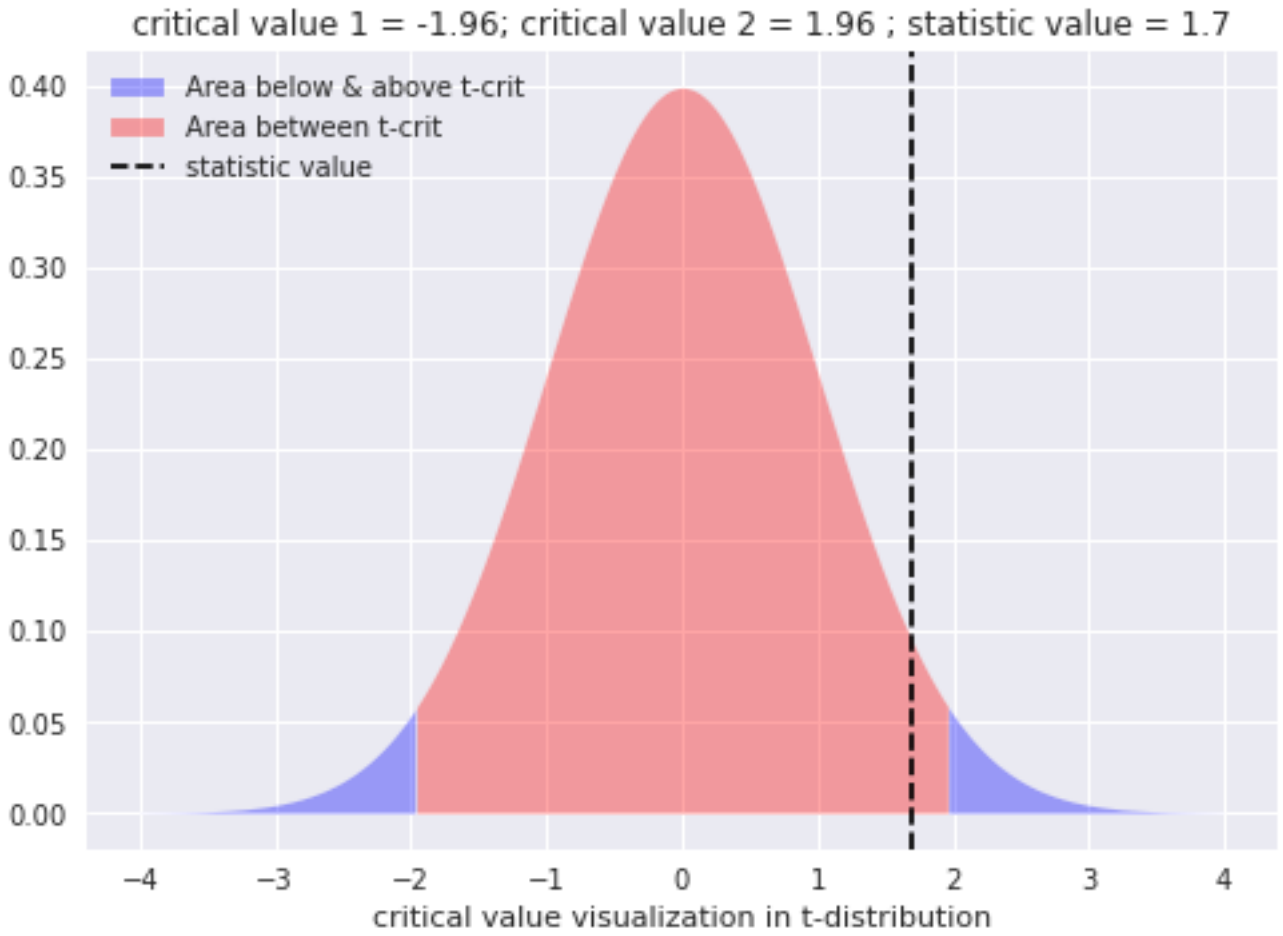


$$H_0 : \mu_{Charge|BMI>25} <= \mu_{Charge|BMI<=25}$$
$$H_1 : \mu_{Charge|BMI>25} > \mu_{Charge|BMI<=25}$$

Insight.

- **Reject H0**, hence in a population of insurance users, we are 95% confident that users with BMI more than 25 will have more expensive charges than users with BMI less than 25.

Male and Female BMI is similar

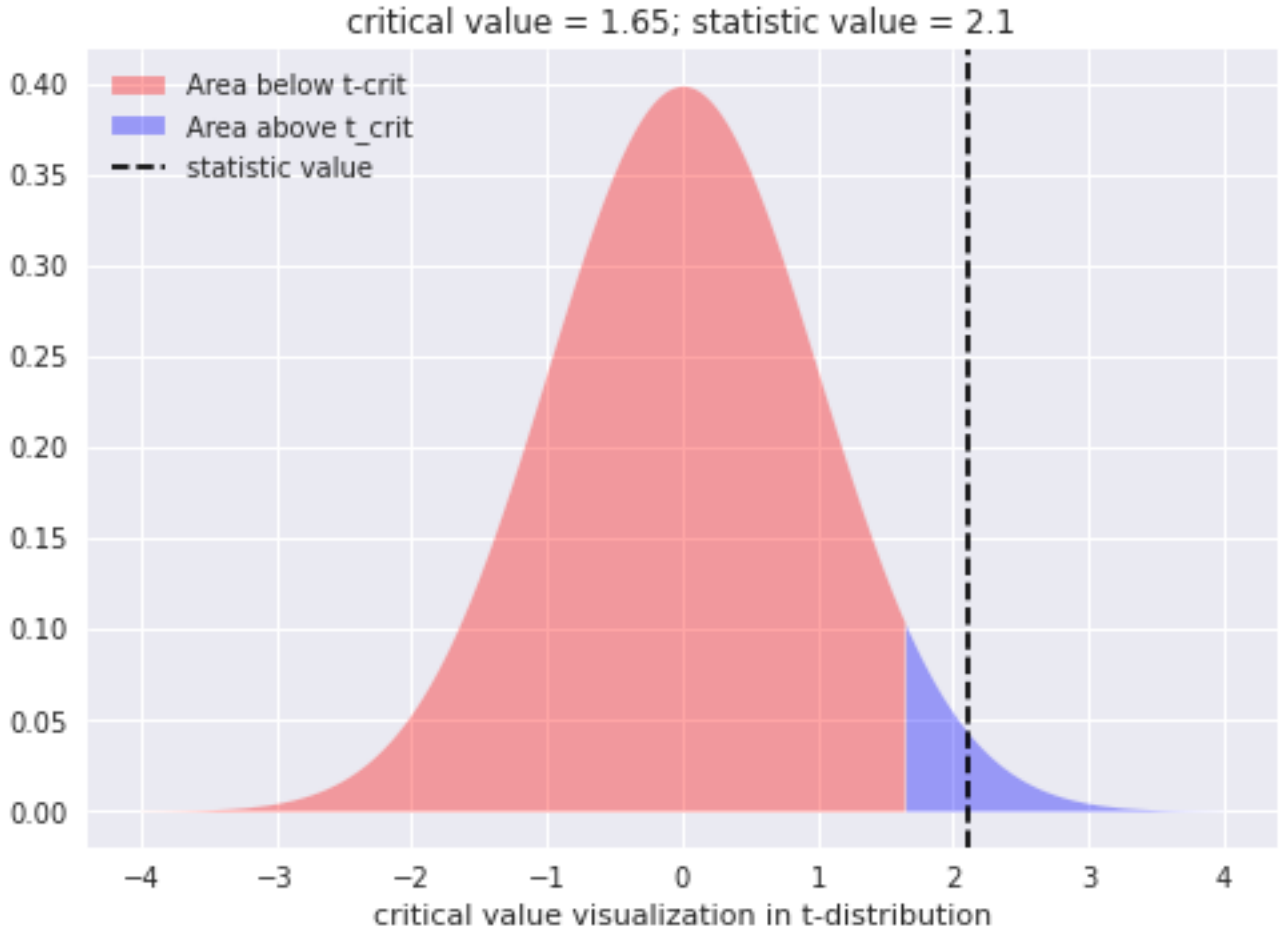


$$H_0 : \mu_{BMI|Male} \neq \mu_{BMI|Female}$$
$$H_1 : \mu_{BMI|Male} = \mu_{BMI|Female}$$

Insight.

- **Fail to Reject H0**. Hence we are 95% confident that there is not enough data/evidence to prove that in a population of insurance users Male BMI is similar to Female BMI.

Male insurance charges is higher than female insurance charges



$$H_0 : \mu_{Charge|Male} <= \mu_{Charge|Female}$$
$$H_1 : \mu_{Charge|Male} > \mu_{Charge|Female}$$

Insight.

- **Reject H0**, hence in a population of insurance users, we are 95% confident that male insurance charges are higher than female insurance charges.

Conclusion

Analysis Summary

There are **several factors** that affect insurance user charges determination like **smoking, BMI value, age, gender, region, and the number of children in a family**. Of those factors, the most **key factor** to determine **how expensive** insurance charges will users have is the **smoking habit**. Analysis shows that people who smoke will get insurance charges higher than people who don't smoke, the **charges for smokers are around $18750** for smokers with **relatively a good value of BMI,** and around **$40000** for smokers with a **bad value of BMI**. Whereas insurance charges for **non-smokers** are just around **$4000**. **BMI** also becomes the **second most contributing factor for insurance charge determination**, but the BMI contribution to charges determination is **larger in smoker people**. The analysis also shows that **as people get older**, the insurance **charge is getting expensive**, this age factor is also different in smokers and non-smokers people. The rest of the factor such as **gender, region, and the number of children** in a family does contribute to the determination of insurance charge but the value is **very small** compared to smoking, BMI, and age factor.

Reference

- Probability and Statistics for Engineers and Scientist, Ronal E. Walpole et. All
- Pacmann : Intro to Probability Learning Module

Reach me ! for further discussion

My_Resume

yudi.stefanus22@gmail.com

https://www.linkedin.com/in/stefanusyudi22