Pacmann

Final Project: Health Insurance Data Analysis

Probability Course - Sekolah Data Pacmann

Rahman Aziz Firmansyah
AI & ML Engineering

Outline

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

Introduction

Pacmann

Introduction

- Insurance is a **guarantee** provided for **compensation** for specified losses, damages, illness, or death **in return payment or premium** should be paid for a given time.

- In determining how much charge will be applied to a particular user, an insurance company needs to **assess the risk** regarding several factors

- A couple of analyses were performed **to answer questions** in this task and eventually **generating insight** obtained through analyses process

Dataset

Pacmann

Dataset

A data set containing charges applied to users with several factors that might influence in assessing the risk level of health insurance. Features included in the data set are:

- **age**: age of an insurance user
- **bmi**: body mass index (a measure of body fat based on height and weight)
- **children**: the number of children of an insurance user
- **smoker**: a binary feature that identifies user into smoker or nonsmoker group
- **charges**: charges applied to an insurance user
- **region**: a certain area where the user live

Descriptive Statistics Analysis

Average Insurance Users' Age

- Average users' age was calculated to provide the big picture what age of health insurance users in general

$$\bar{x}_{age} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The average users' age is approximately 39 years old

Average BMI of smokers and nonsmokers

$$\bar{x}_{(BMI|smoker)} = \frac{1}{n_{smoker}} \sum_{i=1}^{n_{smoker}} x_{i(BMI|smoker)}$$

$$\bar{x}_{(BMI|nonsmoker)} = \frac{1}{n_{nonsmoker}} \sum_{i=1}^{n_{nonsmoker}} x_{i(BMI|nonsmoker)}$$

- The average BMI of users who smoke: 30.71
- The average BMI of users who do not smoke: 30.65

Pacmann

Average charges applied to smokers an nonsmokers

$$\bar{x}_{(charges|smoker)} = \frac{1}{n_{smoker}} \sum_{i=1}^{n_{smoker}} x_{i(charges|smoker)}$$

$$\bar{x}_{(charges|nonsmoker)} = \frac{1}{n_{nonsmoker}} \sum_{i=1}^{n_{nonsmoker}} x_{i(charges|nonsmoker)}$$

- Average of charges applied to smokers: 82.05k
- Average of charges applied to nonsmokers: 8.40k

Average BMI of smoker and nonsmokers given aged over 25 y.o.

$$\bar{x}_{(BMI|smoker \cap age>25)} = \frac{1}{n_{smoker \cap age>25}} \sum_{i=1}^{n_{smoker \cap age>25}} x_{i(BMI|smoker \cap age>25)}$$

$$\bar{x}_{(BMI|nonsmoker \cap age>25)} = \frac{1}{n_{nonsmoker \cap age>25}} \sum_{i=1}^{n_{nonsmoker \cap age>25}} x_{i(BMI|nonsmoker \cap age>25)}$$

- Average BMI of smoker given whose age over 25: 30.58
- Average BMI of nonsmoker given whose age over 25: 30.91

Average BMI of male and female

$$\bar{x}_{(BMI|male)} = \frac{1}{n_{male}} \sum_{i=1}^{n_{male}} x_{i(BMI|male)}$$

$$\bar{x}_{(BMI|female)} = \frac{1}{n_{female}} \sum_{i=1}^{n_{female}} x_{i(BMI|female)}$$

- The average BMI of male users: 30.94
- The average BMI of female users: 30.38

Pacmann

Analysis

Through a number analyses performed we can claim or construct hypotheses

- The average of users age is 39 y.o.
- The charges applied to smokers are significantly higher than that of nonsmoker
- The average BMI of users are identical regardless sex, smoker or nonsmoker, over 25 y.o. or under 25 y.o.

Pacmann

Categorical Variables Analysis

Pacmann

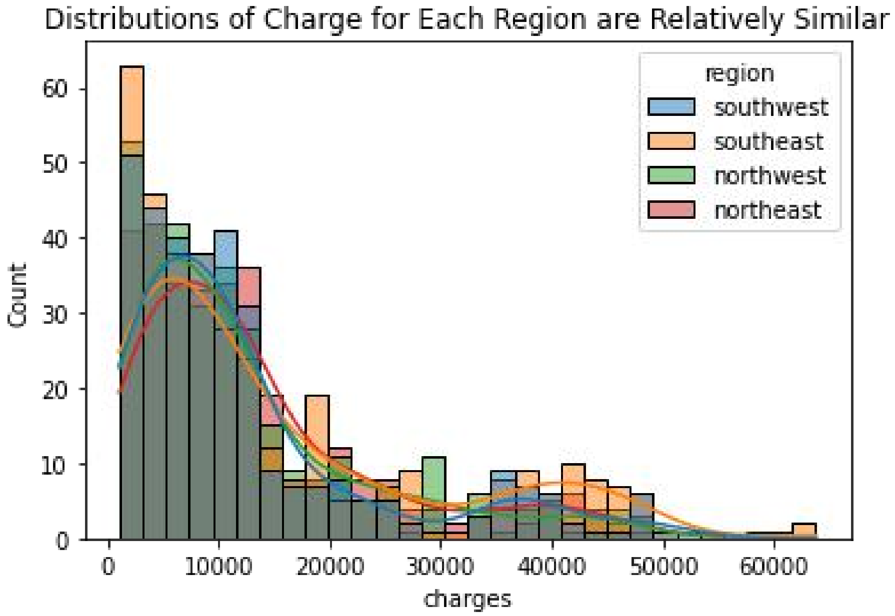
Proporsion of smokers and non smokers

$$\%smoker = \frac{n_{smoker}}{N} \times 100\% = 20.48\%$$

$$\%nonsmoker = \frac{n_{nonsmoker}}{N} \times 100\% = 79.52\%$$

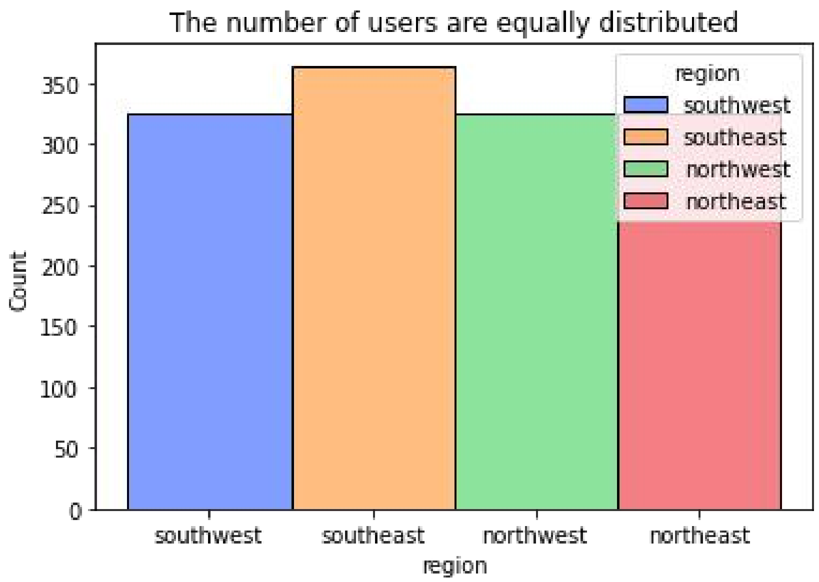
- percentage of smokers to the total is 20.48%
- percentage of nonsmokers to the total is 79.52%

Proporsion of charges in each region



- looking at the distribution pattern of each region, all region have identical proportion which is mostly concentrated at around 7.5k

Number of users distribution



- Each region has identical proportion of users even though there is slight difference in southeast region

Probability of male or female given a smoker

$$P(male|smoker) = \frac{P(male \cap smoker)}{P(smoker)} = 0.58$$

$$P(male|smoker) = \frac{P(male \cap smoker)}{P(smoker)} = 0.42$$

- Probability of a user is male given a smoker is 0.58
- Probability of a user is female given a smoker is 0.42

Analysis

- Surprisingly, nonsmokers outnumbers smokers as insurance users in this sample
- Since the distribution of charge for each region are identical, we can assume that region has no effect on risk assessment
- The number of users each region are equally distributed, thus we can claim that our data is balance in terms of region
- It is more likely a smoker to be a male than a female

Continuous Variables Analysis

Pacmann

Probability of a user whose BMI is over 25 and smoker or nonsmoker

Let be X = charges, Y = BMI, Z = smoker (takes to states, yes (1) or no (0))

$$P(X > 16700 | Y > 25 \cap Z = 1) = \frac{P(X > 16700 \cap (Y > 25 \cap Z = 1))}{P(Y > 25 \cap Z = 1)} = 0.98$$

$$P(X > 16700 | Y > 25 \cap Z = 0) = \frac{P(X > 16700 \cap (Y > 25 \cap Z = 0))}{P(Y > 25 \cap Z = 0)} = 0.08$$

- The probability of someone whose BMI is over 25 and **a smoker** being charged over 16.7k is 0.98
- The probability of someone whose BMI is over 25 and **not a smoker** being charged over 16.7 is 0.08

BMI vs Smokers

Let be X = charges, Y = BMI, Z = smoker (takes to states, yes (1) or no (0))

$$P(X > 16700|Y > 25) = \frac{P(X > 16700 \cap Y > 25)}{P(Y > 25)} = 0.26$$

$$P(X > 16700|Y < 25) = \frac{P(X > 16700 \cap Y < 25)}{P(Y < 25)} = 0.21$$

- The probability of someone whose BMI is under 25 being charged over 16.7k is 0.21
- The probability of someone whose BMI is over 25 being charged over 16.7 is 0.26

BMI vs Smoker

Let be X = charges, Y = BMI, Z = smoker (takes to states, yes (1) or no (0))

$$P(X > 16700 | Z = 1) = \frac{P(X > 16700 \cap Z = 1)}{P(Z = 1)} = 0.93$$

$$P(X > 16700 | Z = 0) = \frac{P(X > 16700 \cap Z = 0)}{P(Z = 0)} = 0.08$$

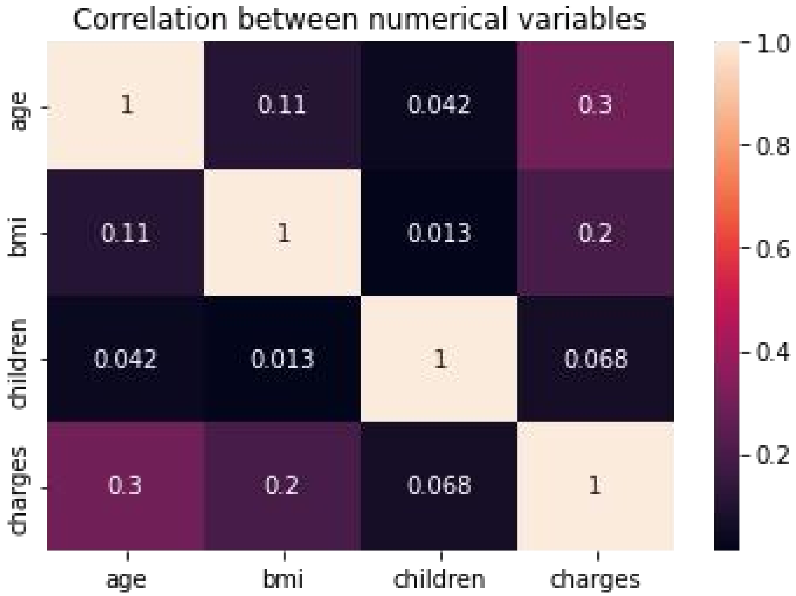
- The probability of a smoker being charged over 16.7k is 0.93
- The probability of a nonsmoker being charged over 16.71 is 0.08

Analysis

- **A smoker** with BMI over 25 is **more likely** to be charged over 16.7k than a **non smoker** with BMI over 25
- The probability of a user with BMI over 25 is being charged more than 16.7 is **slightly higher** than those with BMI under 25
- A smoker is very likely to be charged over 16.7k
- Hypothesis: Smoker has a stronger influence than BMI value to a specified charge

Variables Correlation

Correlation



- Age has strongest correlation to charges than any other numerical features

Hypothesis Testing

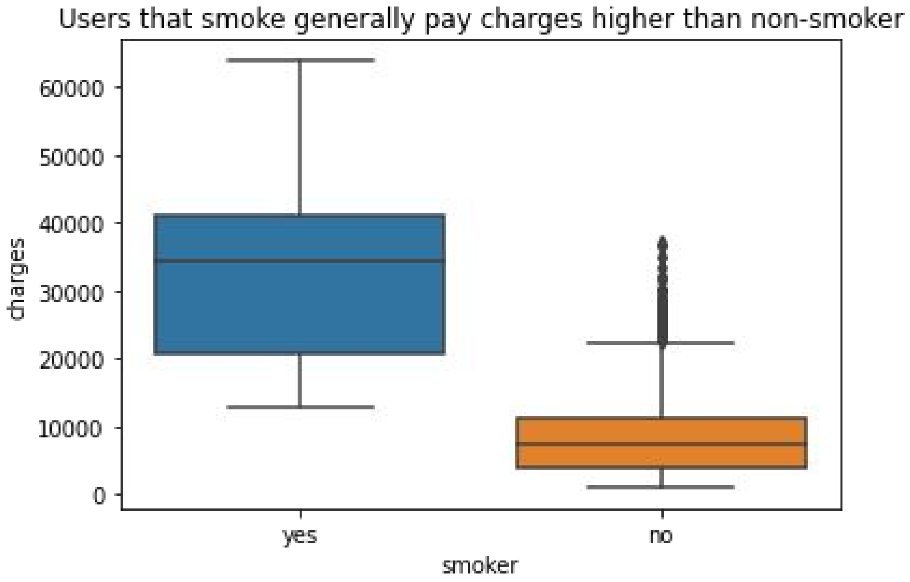
Smoker’s charges are higher than non smoker’s

- Constructing hypothesis based on descriptive statistics

  $H_0 : \bar{x}_{smoker} \geq \bar{x}_{nonsmoker}$

  $H_1 : \bar{x}_{smoker} < \bar{x}_{nonsmoker}$

- performing independent t-test for two samples with different variance
- Result
  - p-value > alpha
  - We failed to reject the null hypothesis thus we claim that smoker’s charges are higher than nonsmoker’s

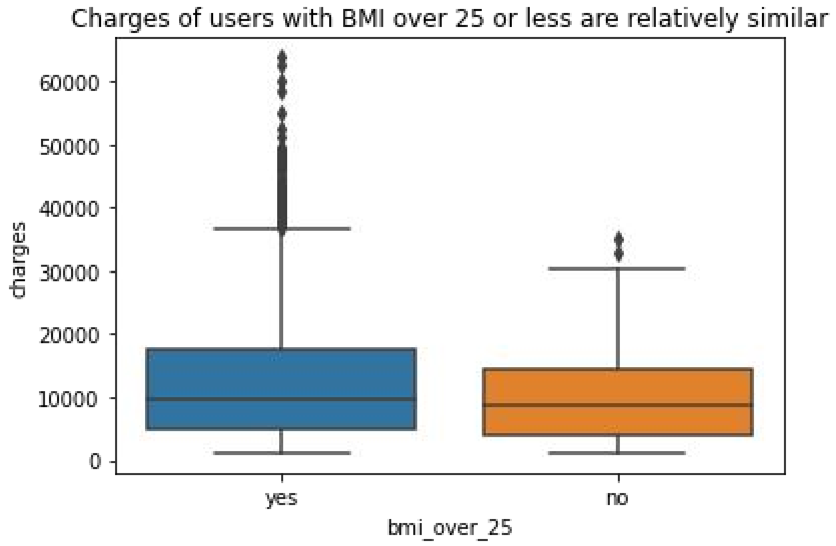


Are charges for those with BMI over 25 higher than that of under 25?

- Constructing hypothesis based on descriptive statistics

$$H_0 : \bar{x}_{(BMI>25)} \leq \bar{x}_{(BMI>25)}$$
$$H_1 : \bar{x}_{(BMI>25)} > \bar{x}_{(BMI>25)}$$

- performing independent t-test for two samples with different variance
- Result
  - p-value < alpha
  - The null hypothesis is rejected thus we should claim that charges applied for users with BMI over 25 are higher than those with BMI under 25

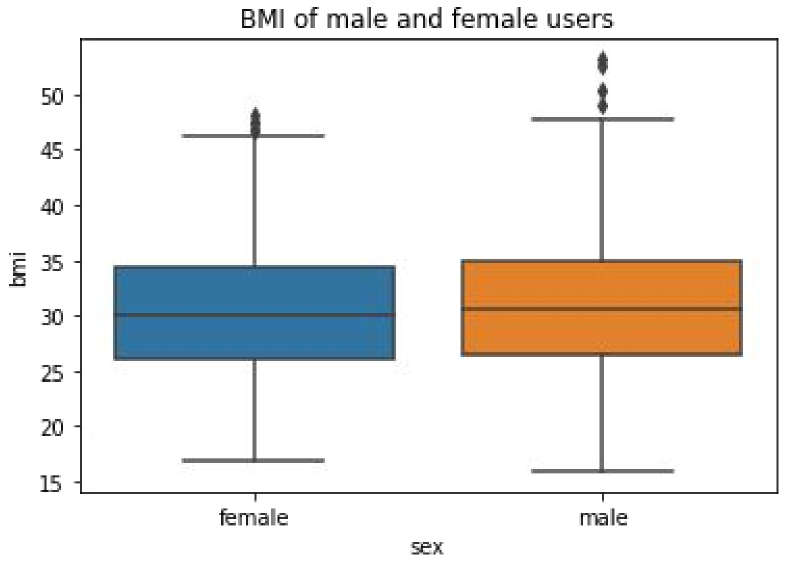


Are charges for those with BMI over 25 higher than those with BMI under 25?

- Constructing hypothesis based on descriptive statistics

$$H_0 : \bar{x}_{(BMI|male)} = \bar{x}_{(BMI|female)}$$
$$H_1 : \bar{x}_{(BMI|male)} > \bar{x}_{(BMI|female)}$$

- performing independent t-test for two samples with identical variance
- Result
  - p-value < alpha
  - The null hypothesis is rejected thus we should claim that male users have higher BMI than female users



Conclusion

Pacmann

Conclusion

- Among numerical features in the data set, users' age is the most correlated feature with the amount of charges paid by users
- Smoker users pay higher charges than that of non smoker users
- Users with BMI over 25 pay higher than that of users with BMI less than 25
- Average BMI of male users is higher than that of female

Notes

There are still analysis results need to be tested to provide sufficent statistical evidence.

Reference

Chan, Stanley. Introduction to Probability for Data Science. 2021