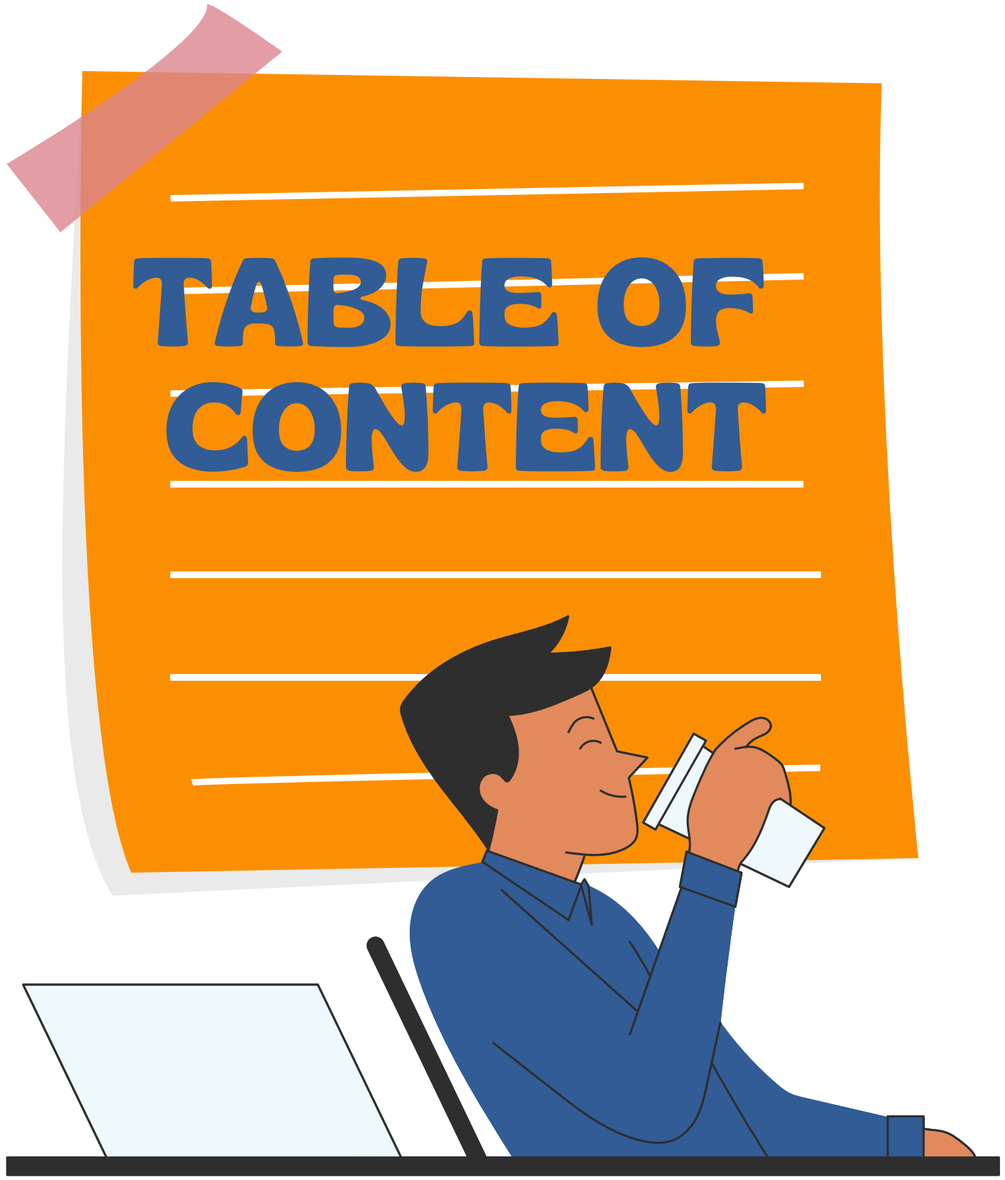
Statistics Case Study

MEDICAL COST INSURANCE

- Dimas Bagos Prasetyo -

TABLE OF CONTENT



1 Why Statistics?

2 Project Introduction

3 Descriptive Statistics

4 Hypothesis Testing

5 Regression Analysis

CASE STUDY

WHY STATISTICS

CASE STUDY

WHAT IS STATISTICS ?

Branch of mathematics dealing with the collection, analysis, interpretation and presentation of masses of numerical data.

WHY STATISTICS ?

Because statistics can assist us in making the best decisions by estimating and predicting the population based on the sampled data

CASE STUDY

STATISTICS WORKFLOW

DESCRIPTIVE STATISTICS

Understand sample data using mean, median, histogram and other descriptive statistics

PROBABILITY DISTRIBUTION

Use the sample as a model for the entire population. if it fits a probability distribution

CONFIDENCE INTERVALS

if sample doesn't fit a distribution, use central limit theorem to make estimates about population

HYPOTHESIS TESTING

Draw conclusions about what a population looks like based on a sample

REGRESSION ANALYSIS

Use relationships based on available variable to make prediction

PROJECT INTRODUCTION

CASE STUDY

OUR GOAL	Leverage customer information and statistics for the evaluation and decision-making in insurance business.
OUR OBJECTIVE	• Understand datasetwith descriptive statistics • Draw conclusions with hypothesis tests • Make predictions with regression

CASE STUDY

Type Variabel	Name Variables
Categorical varibles	sex,smoker,region,children
Quantitative Variables	age,bmi,charges

CASE STUDY

PROJECT DATASET

age	sex	bmi	children	region	smoker	charges
19	female	27,9	0	southwest	yes	16884,924
18	male	33,77	1	southeast	no	1725,5523
28	male	33	3	southeast	no	4449,462
33	male	22,705	0	northwest	no	21984,471
32	male	28,88	0	northwest	no	3866,8552
31	female	25,74	0	southeast	no	3756,6216
46	female	33,44	1	southeast	no	8240,5896
37	female	27,74	3	northwest	no	7281,5056
37	male	29,83	2	northeast	no	6406,4107
60	female	25,84	0	northwest	no	28923,137
25	male	26,22	0	northeast	no	2721,3208
62	female	26,29	0	southeast	yes	27808,725
23	male	34,4	0	southwest	no	1826,843
56	female	39,82	0	southeast	no	11090,718
27	male	42,13	0	southeast	yes	39611,758
19	male	24,6	1	southwest	no	1837,237
52	female	30,78	1	northeast	no	10797,336
23	male	23,845	0	northeast	no	2395,1716
56	male	40,3	0	southwest	no	10602,385
30	male	35,3	0	southwest	yes	36837,467
60	female	36,005	0	northeast	no	13228,847
30	female	32,4	1	southwest	no	4149,736

Field	Description
Usia	age of primary beneficiary
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 / m ^ 2)
Children	Number of children covered by health insurance / Number of dependents
Region	the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
Smoker	Smoking
Charges	Individual medical costs billed by health insurance (USD)

CASE STUDY

DESCRIPTIVE STATISTICS

DESCRIPTIVE STATISTICS

	Age	bmi	children	charges
Count	1338	1338	1338	$ 1.338,00
Mean	39,21	30,66	1,09	$ 13.270,42
Std	6,09819	14,04996	1,20549	$ 12.110,01
Min	18,00000	15,96000	0,00000	$ 1.121,87
25%	27,00000	26,29625	0,00000	$ 4.740,29
50%	39,00000	30,40000	1,00000	$ 9.382,03
75%	51,00000	34,69375	2,00000	$ 16.639,91
Max	64,00000	53,13000	5,00000	$ 63.770,43

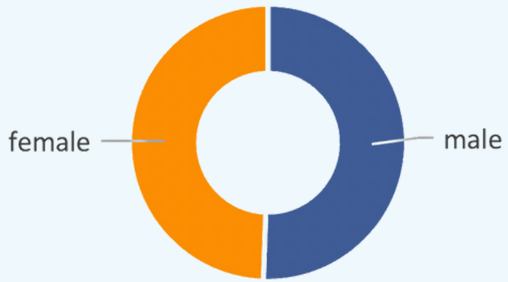
- The table shows the count, mean, standard deviation, and all quartiles of all kuantitatif column.
- Average age is 39.2 and maximum age is 64
- Average BMI is 30.66 (Obesity, Maximum BMI is 53.13
- Customer Average has 1 child and maximum 5
- average medical cost is $13.270 and maximum is $63.770

OUR SAMPLE. . .

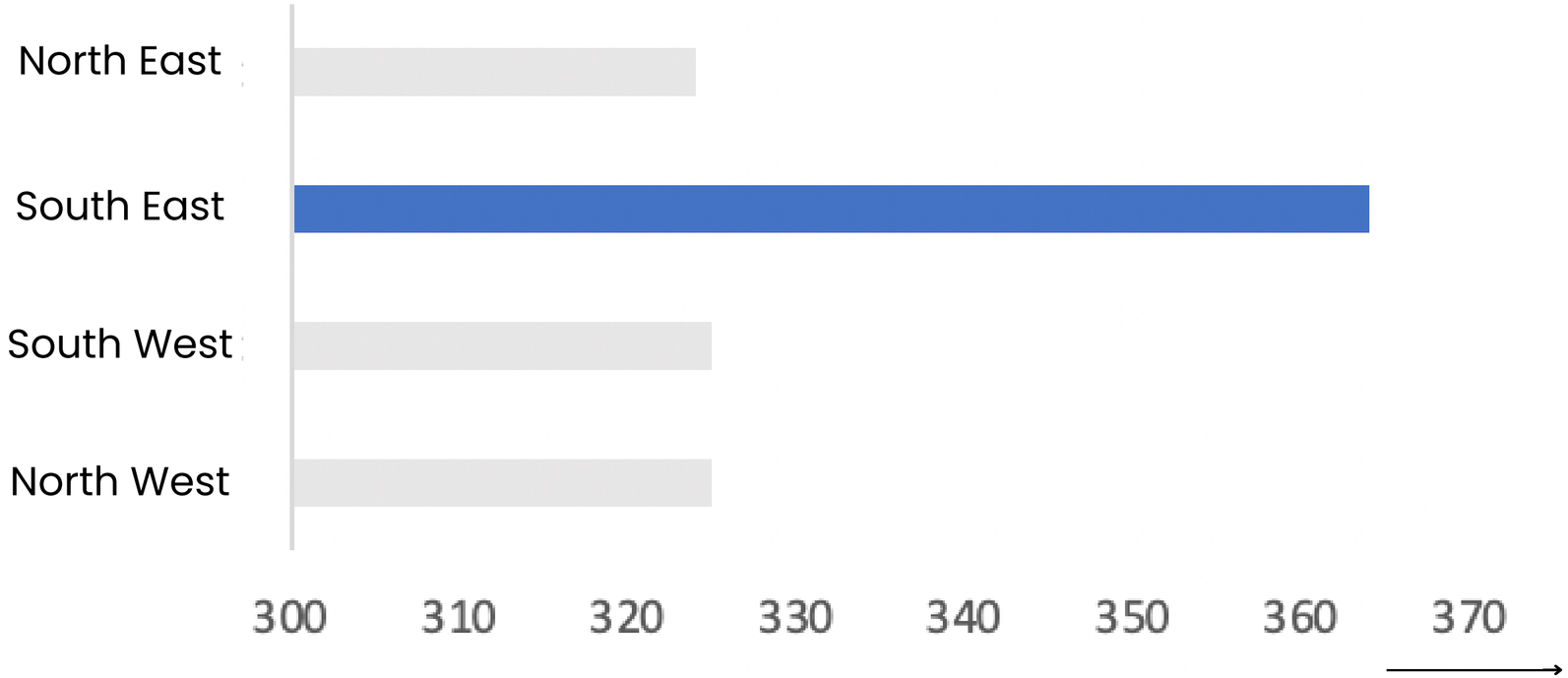
There Are

1338

Sample

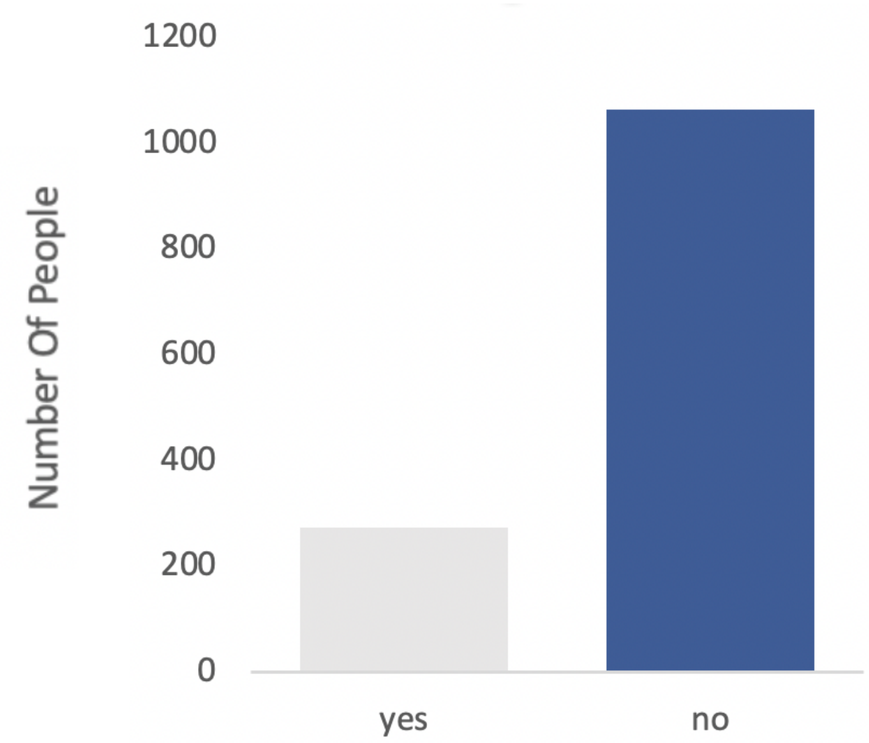


Sample Dominated from South East

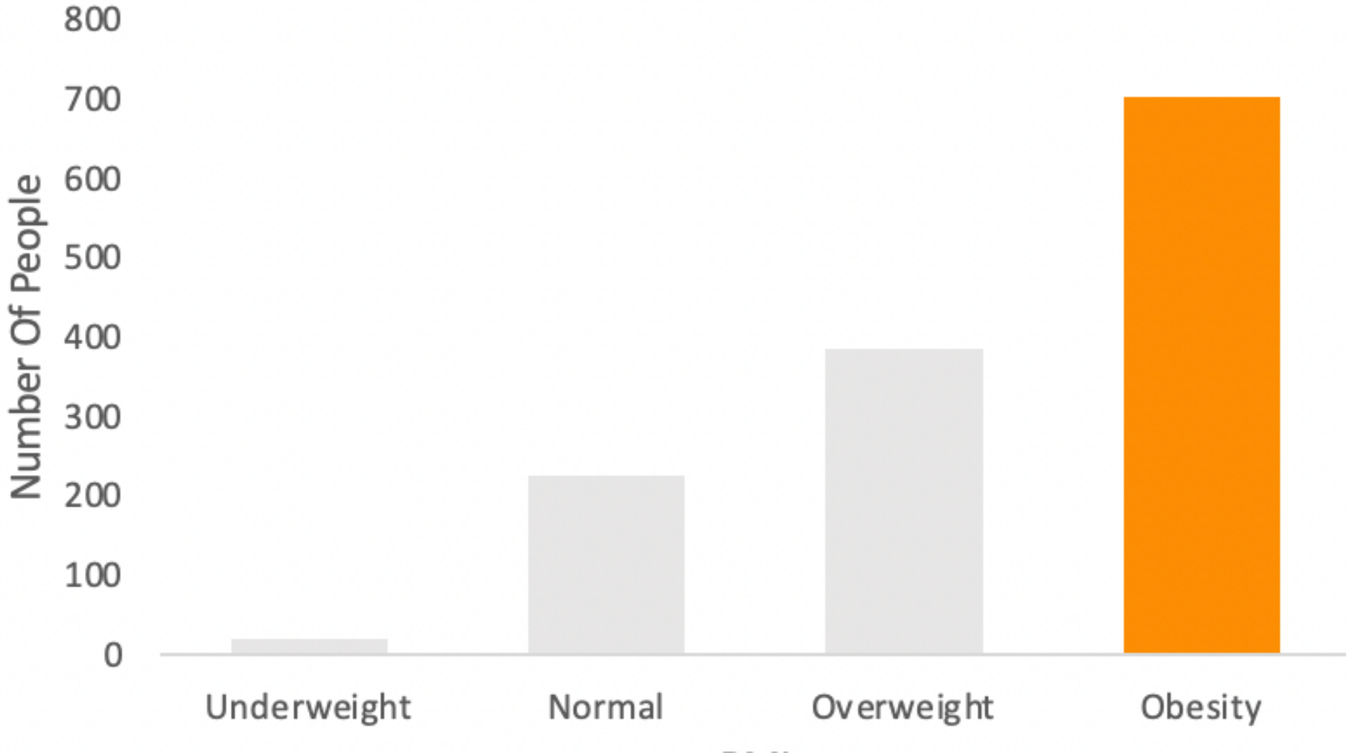


CASE STUDY

More than half of the sample are **not smoke**



More than 700 samples indicate **obesity**



SCORE HISTOGRAM

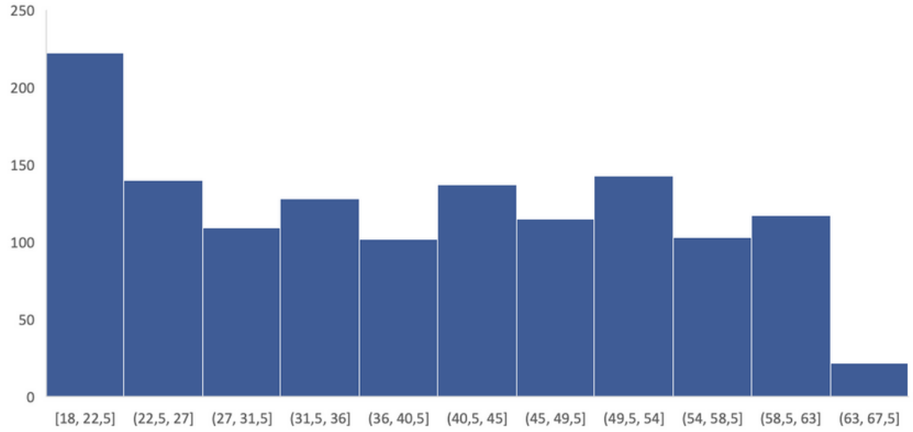
CASE STUDY

Variabel	Skewness
Age	0,055
BMI	0,284
Charges	1,515

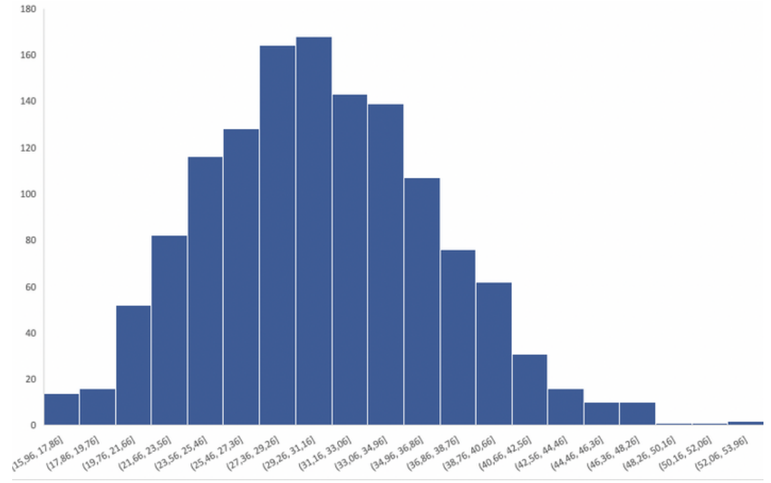
"Skewness is considered to approach symmetry/ normal when its value is less than 0.5 (either positive or negative)"

"If we want to standardize the data, it needs to go through a transformation stage (log)."

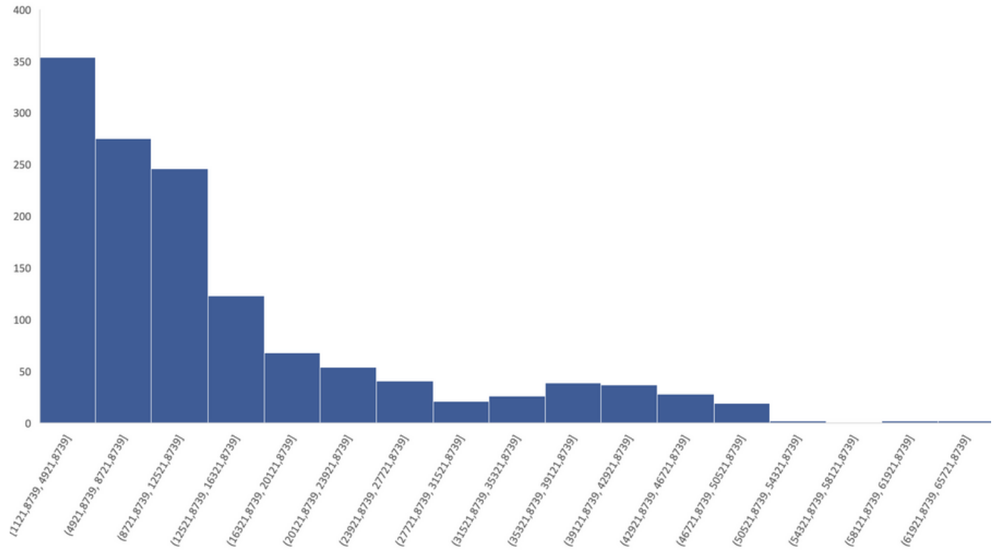
Distribution Of Age



Distribution Of BMI

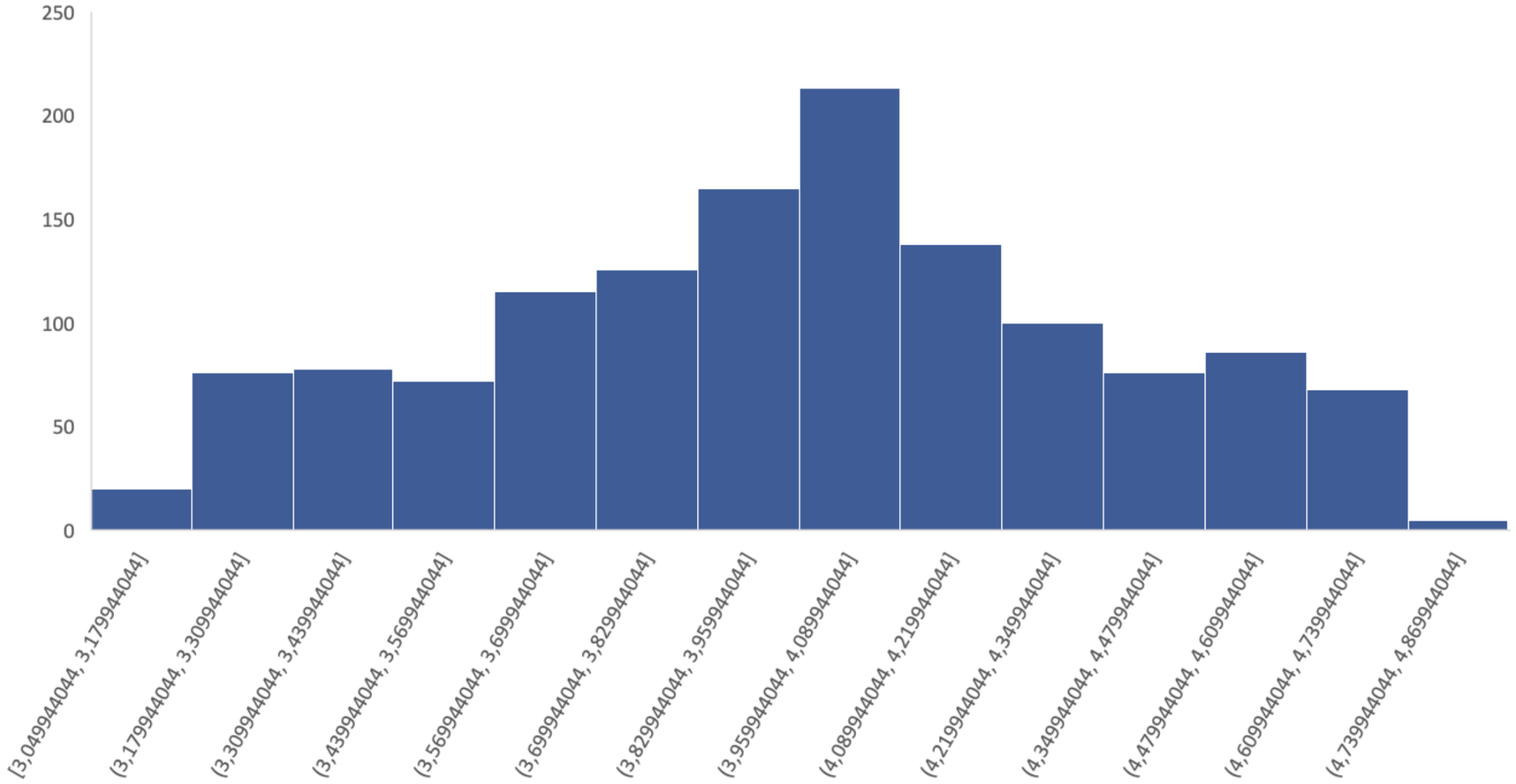


Distribution Of Charges



CASE STUDY

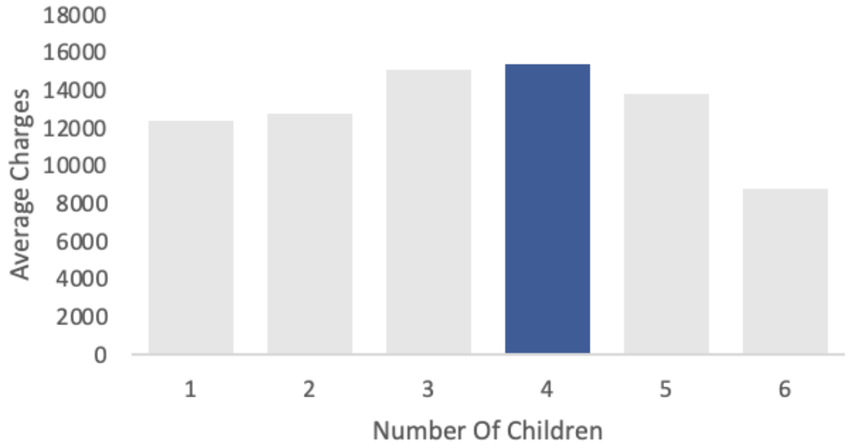
Distribution Of **Charges**



After passing through log10, Skewness Charges = -0,090, it means between -0,5 and 0,5

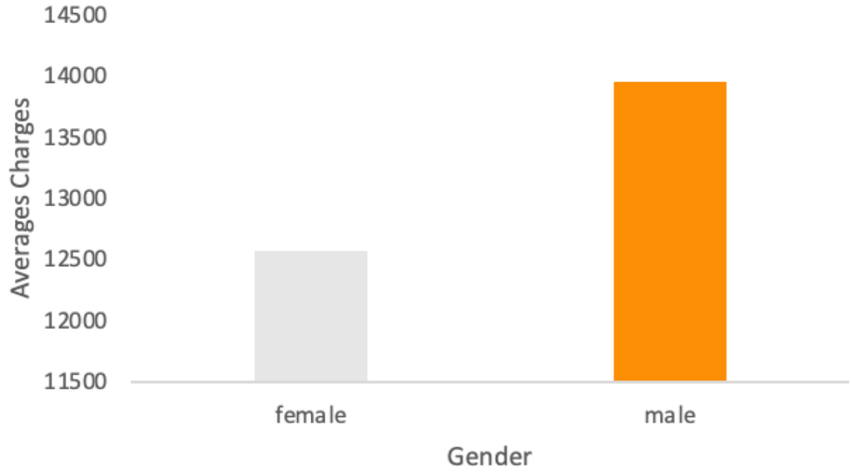
CASE STUDY

Children VS Charges



"The highest average charges are found in the sample with four children."

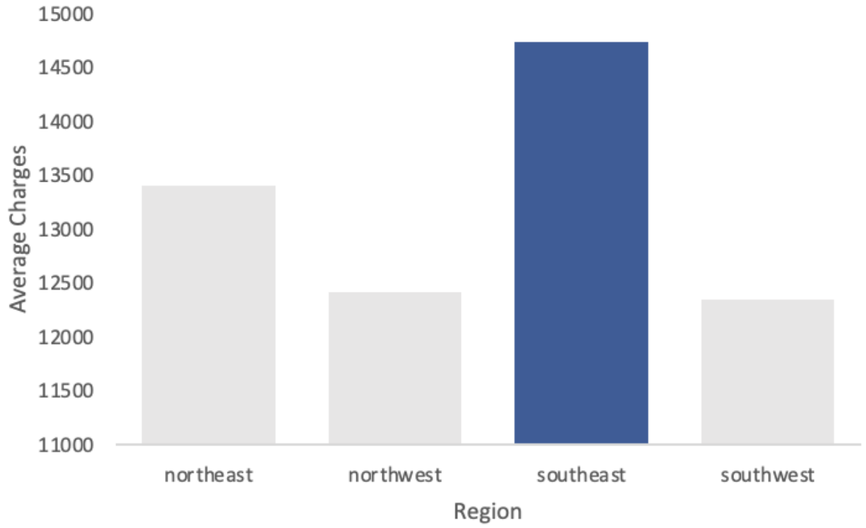
Gender VS Charges



"Male have a higher average charge compared to female."

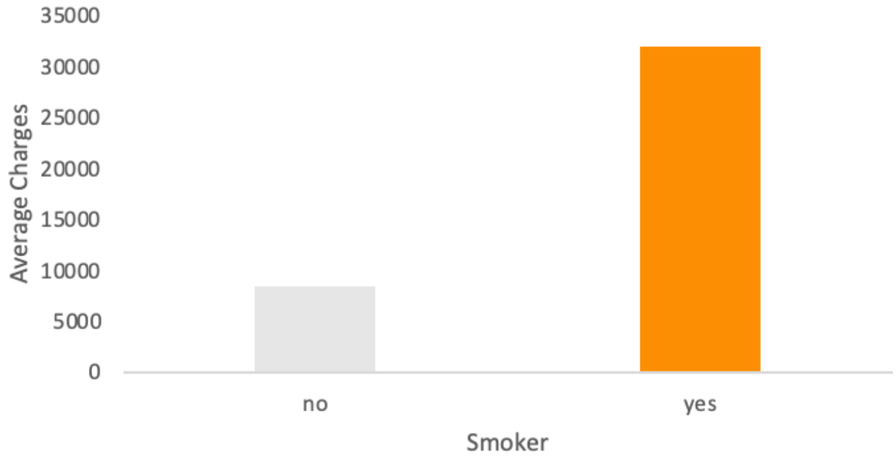
CASE STUDY

Region VS Charges



"The highest average charges are in the Southeast region."

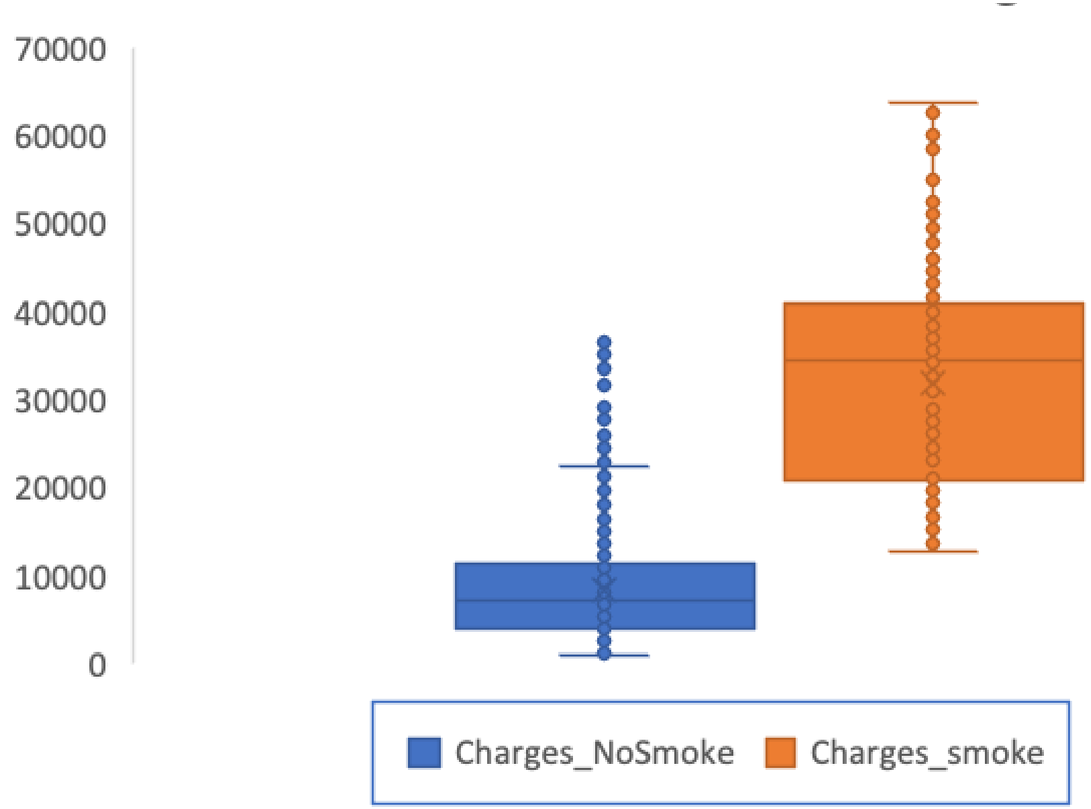
Smoker VS Charges



"Smokers have a higher average charge compared to non-smokers."

CASE STUDY

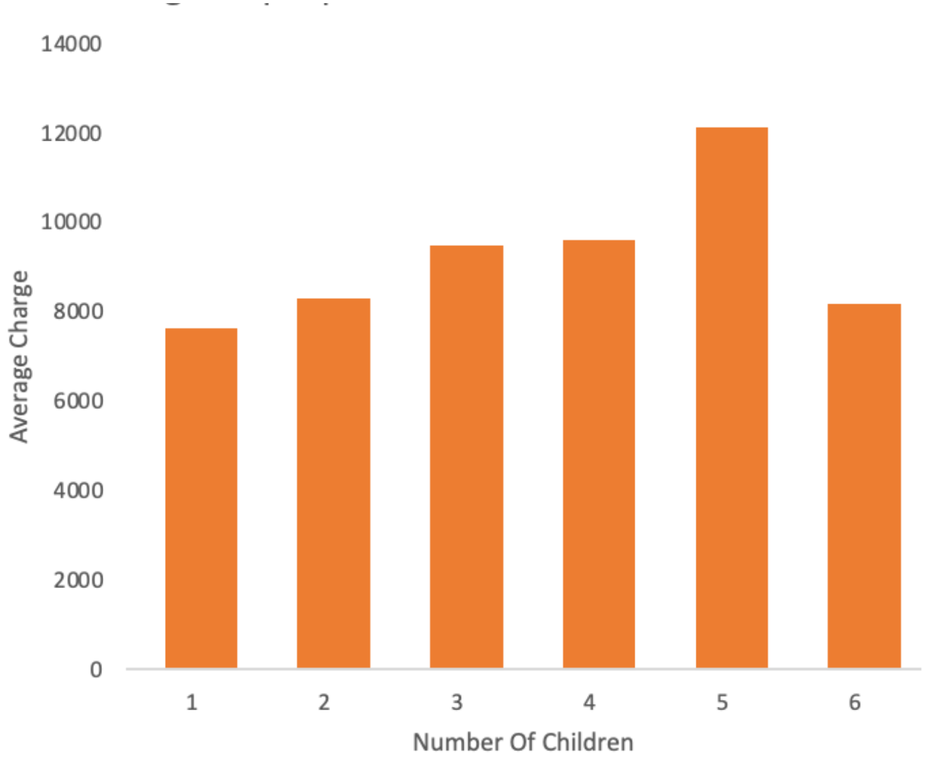
"When compared using a boxplot, the charges median for smokers indicate a **higher value** compared to those for non-smokers."



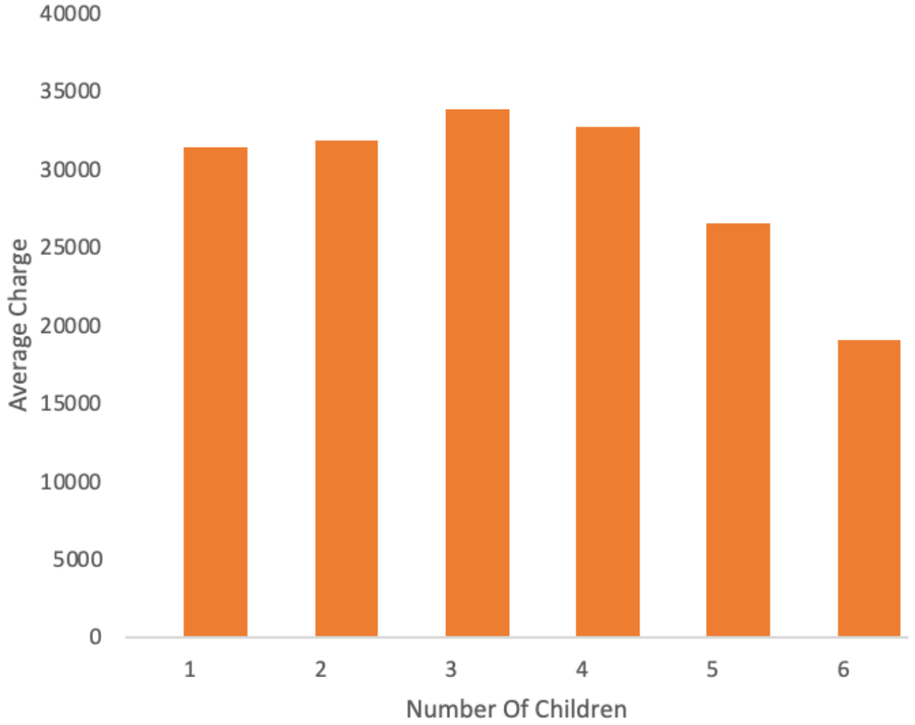
CASE STUDY

"Even though smokers have an increasing number of children, the average charges indicate a decreasing trend."

No Smoke And Have Children Charges



Smoke And Have Children Charges



CASE STUDY

HYPOTHESIS TESTING

CASE STUDY

WHAT IS HYPOTHESIS TEST

Hypothesis testing is a statistical procedure used to make decisions or draw inferences about population parameters based on observed sample data. Its goal is to evaluate the truth or falsity of a statement or hypothesis about a population based on information obtained from a sample

STEP HYPOTHESIS TEST

1. Clearly articulate the null and alternative hypotheses.
2. Set the significance level
3. Calculate the test statistics for the sample
4. Calculate the p-value
5. Draw a conclusion based on the findings of the test.

Prove that the medical claims made by the people who smoke is greater than those who don't?

HYPOTHESES			
Ho:	μ1	<=	μ2
Ha:	μ1	>	μ2

Define null and alternative hypothesis

The average charges of smokers is less than or equal to nonsmokers

The average charges of smokers is less than or equal to nonsmokers

} Right-tail

Dan karena ini merupakan uji satu arah maka kita melewati step 4

SIGNIFICANCE LEVEL	
Alpha:	5%

Decide the significance level : 0,05

SAMPLE DATA	
Sample Size:	275

HYPOTHESIS TEST	
Tstatistics-onetail	1,0595E-117

Reject H0. 1,0595E-117 < Alpha

Based on the one-tailed t-statistic test, the t-statistic value < alpha. As a result, the null hypothesis (H0) is rejected, and the alternative hypothesis (Ha) is accepted. Therefore, it can be concluded that smokers incur higher charges compared to non-smokers.

CASE STUDY

REGRESSION ANALYSIS

WHAT IS REGRESSION ANALYSIS

Linear regression is a type of statistical analysis used to predict the relationship between two variables. It assumes a linear relationship between the independent variable and the dependent variable, and aims to find the best-fitting line that describes the relationship.

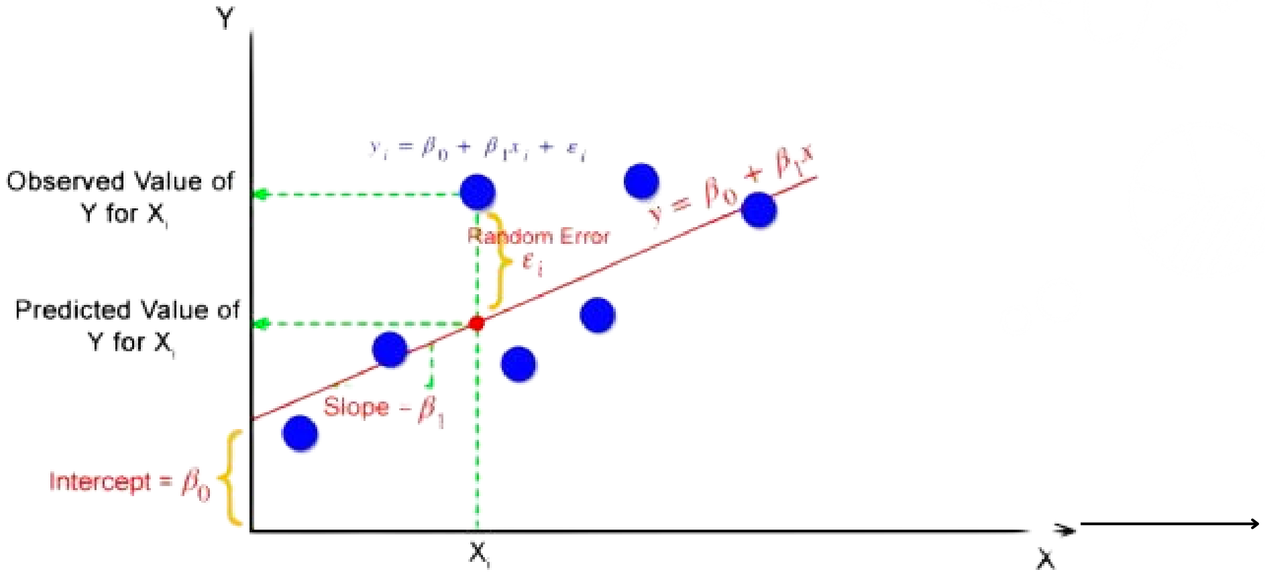
Linear Regression: Single Variable

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

Predicted output | Coefficients | Input | Error

Linear Regression: Multiple Variables

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$



Source : analyticsvidhya

CASE STUDY

SCORE CORRELATION TO CHARGES

Variabel	Score
Age	0,527834
Gender	0,00563186
BMI	0,13266941
Children	0,0679982
Region	0,0242018
Smoker	0,6655057

Use Function =CORREL()

"Because smokers and age variabele have a high correlation, I want to use them as independent variables."

Class of Pearson Correlation

Class	Range
Not Correlated	< 0,1
Weak	0,1 to 0,2
Moderate	0,2 to 0,5
Strong	> 0,5

CASE STUDY

Regression Statistics	
Multiple R	0,84935318
R Square	0,72140083
Adjusted R Square	0,72098345
Standard Error	6396,75223
Observations	1338

- The Multiple R value of 0.8493 indicates a strong relationship between the independent variables (smoker and age) and the dependent variable (Charges).
- The table shows an R-squared value of 0.7214, meaning that the independent variables (smoker and age) collectively account for 72.14% of the variation in the dependent variable (Charges). The remaining 28.86% is influenced by other factors

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	1,4145E+11	7,0724E+10	1728,41522	0
Residual	1335	5,4626E+10	40918439,1		
Total	1337	1,9607E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	-2391,6264	528,3021105	-4,5270051	6,5155E-06	-3428,01909	-1355,233629	-3428,01909	-1355,233629
age	274,871186	12,45531425	22,068587	2,9098E-92	250,4370659	299,3053058	250,4370659	299,3053058
smoker	23855,3048	433,4883733	55,0310142	0	23004,91223	24705,6974	23004,91223	24705,6974

P-value ≤ 5 %, it can be stated that these variables individually have a significant impact.

Regression model

Regression model = -2391,62 + 274,87 x Age + 23855,30 *x Smoker +Error Term

CASE STUDY

Using the model to predict charges by utilizing the variables age and smoker.

Performing a prediction of charges for a smoker who is 68 years old.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

Y = -2391,62 + 274,87* 68 + 23855,30 * 1

Y = $40.154,91

THANK YOU


Budget
Actual
Difference
£500.00
£375.00
SALE REPORT
£125.00
£0.00
Auto
Food
Medical
Travel
Other