MOHAMMED VI POLYTECHNIC UNIVERSITY

School of Collective Intelligence

Health Insurance Analysis

DATA SCIENCE EXAM PRESENTATION

Ahlam Labiad
Oluwatobi Fakoya
Abdulwasiu Tiamiyu

SUPERVISED BY

Pr. Ikram CHAIRI
Bochra CHEMAM

Outline

- Introduction
- Dataset
    - Source
    - Dictionary
- Data Description
- Data Cleaning and Preprocessing
- Feature Engineering
- Data Exploration and Visualization
    - Univariate Analysis
    - Bivariate Analysis
    - Multivariate Analysis
- Hypothesis Testing
- Key Findings
- Future Work

Introduction

Health insurance or medical insurance is an agreement where an insurance company agrees to compensate the insured for the medical and surgical expenses incurred during the policy tenure. The medical expenses may incur if the insured falls ill, or meets an accident that leads to hospitalisation of the insured.

To be eligible to avail coverage benefits under the policy, the policyholder is required to pay a specific amount periodically, known as **premium**.

Health Insurance premium is decided by an insurance company and policyholders are required to pay the same on a monthly, quarterly, half-yearly, or yearly basis, without any lapse, to avoid losing the renewal benefits.

Introduction

Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.

PROBLEM STATEMENT: In this project, we will be using visualizations and statistical hypothesis testing to evaluate and examine a dataset for medical costs in Health Insurance in the United States, in order to draw significant insights and make some statistical-based inferences.

Dataset

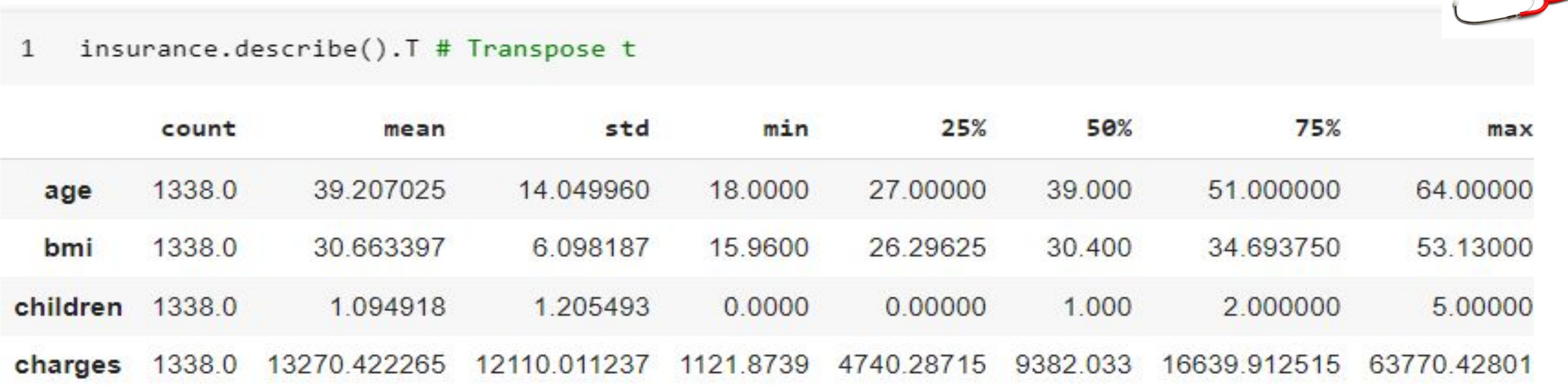
Data Source: US Health Insurance Dataset

Insurance Premium Charges in US with important details for risk underwriting.

Data Dictionary:

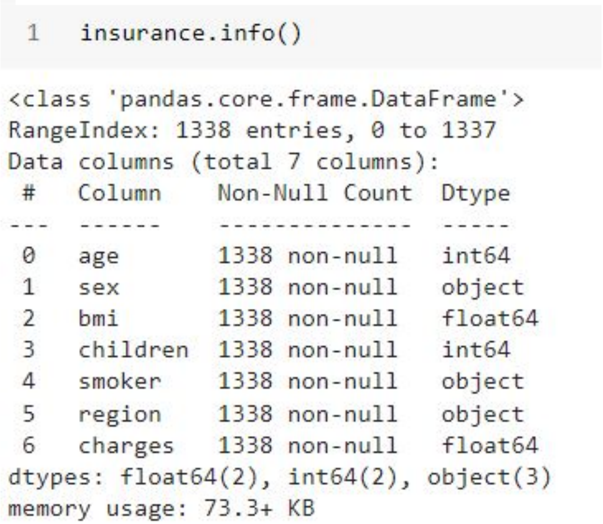
- **Age**: age of primary beneficiary
- **Sex**: insurance contractor gender, female, male
- **BMI**: Body mass index (BMI) is a value calculated by dividing a person's weight in kilograms by the square of height in meters.
- **Children**: Number of children covered by health insurance / Number of dependents
- **Smoker**: Smoking
- **Region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **Charges**: Individual medical costs billed by health insurance.

Data Description



```
insurance.describe().T # Transpose t
```

	count	mean	std	min	25%	50%	75%	max
age	1338.0	39.207025	14.049960	18.0000	27.00000	39.000	51.000000	64.00000
bmi	1338.0	30.663397	6.098187	15.9600	26.29625	30.400	34.693750	53.13000
children	1338.0	1.094918	1.205493	0.0000	0.00000	1.000	2.000000	5.00000
charges	1338.0	13270.422265	12110.011237	1121.8739	4740.28715	9382.033	16639.912515	63770.42801

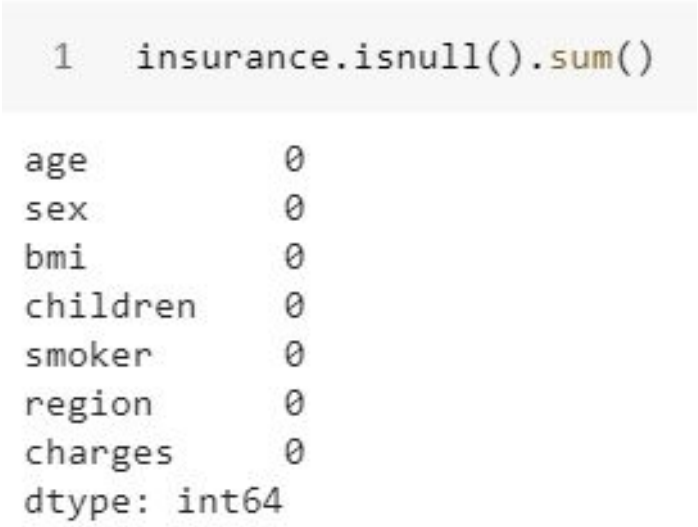


```
insurance.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

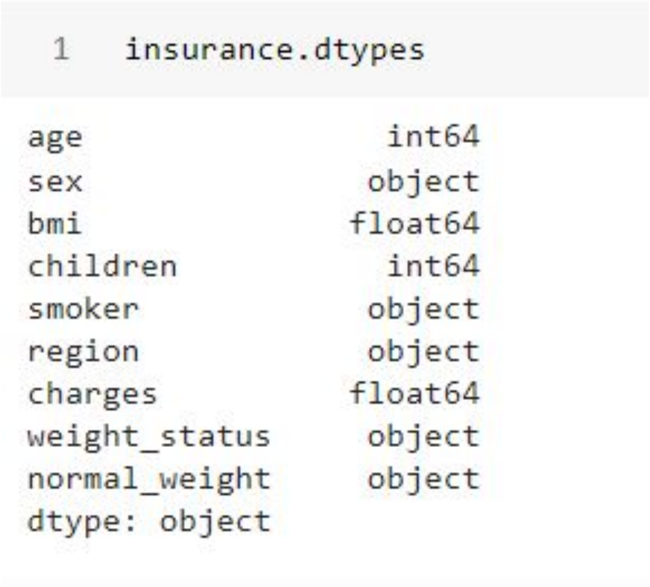
- The data set has 1338 entries with 7 attributes. 2 integer type, 2 float type and 3 object type (Strings in the column).
- There are no null values in any of the columns.
- The data statistics generally looks in good shape.
- The data in the age column represents true age distribution of the adult population.
- The charged amount is highly skewed as most people would require basic medical care and only few suffer from diseases which cost more to treat.

Data Cleaning and Preprocessing



```
insurance.isnull().sum()
```

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```



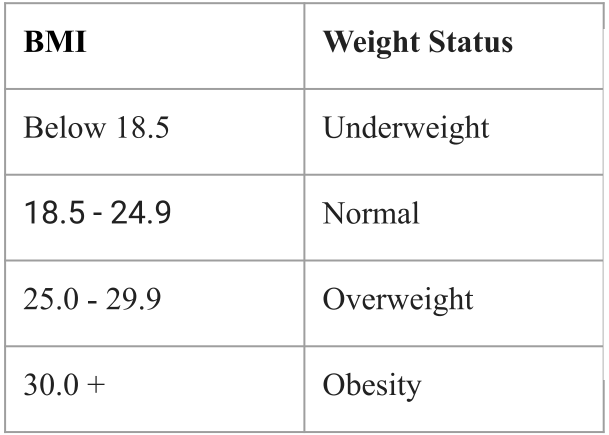
```
insurance.dtypes
```

```
age                int64
sex               object
bmi              float64
children           int64
smoker            object
region            object
charges          float64
weight_status     object
normal_weight     object
dtype: object
```

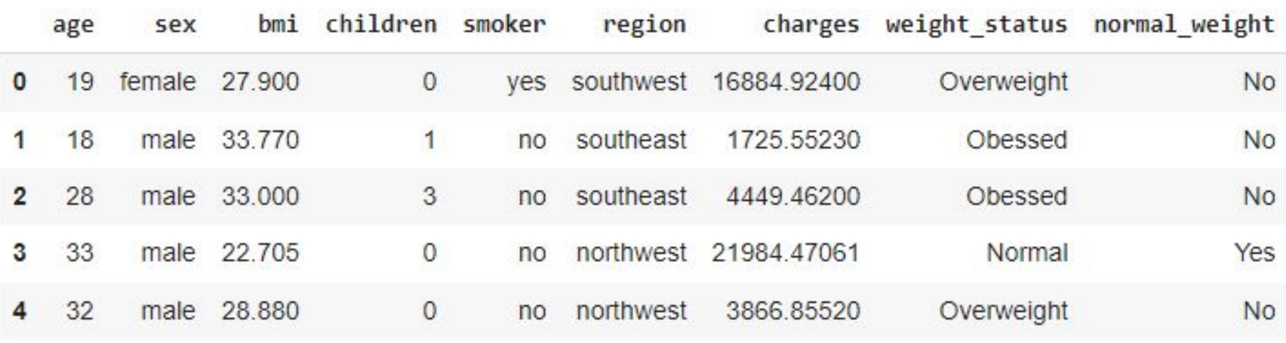
- The variable names are well written and convey actual meanings.
- There are no missing values in the data.
- The columns are in the right data types.

Feature Engineering

- Body mass index (BMI) is a value derived from the mass (weight) and height of a person. (A person's weight in kilograms divided by the square of height in meters.)
- The BMI is used by healthcare professionals to screen for overweight and obese individuals.
- The BMI is used to assess a person’s health risks associated with obesity and overweight.
- For example those with a high BMI are at risk of: high blood cholesterol or other lipid disorders, type 2 diabetes, heart disease, stroke, high blood pressure, etc...

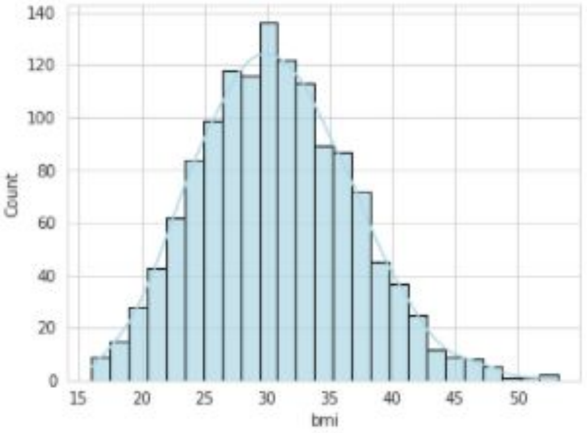


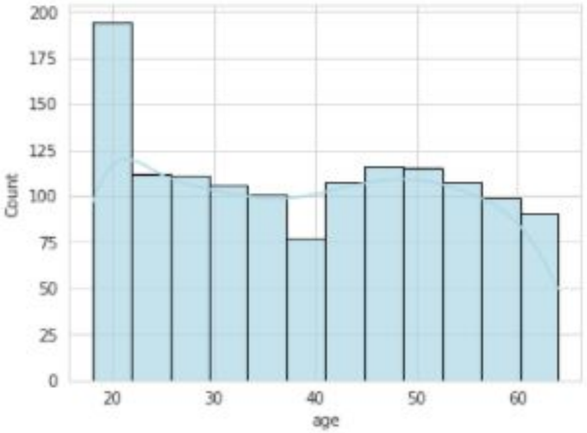
BMI	Weight Status
Below 18.5	Underweight
18.5 - 24.9	Normal
25.0 - 29.9	Overweight
30.0 +	Obesity

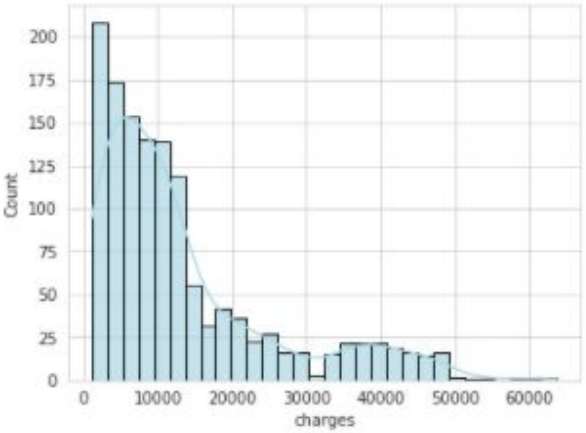


	age	sex	bmi	children	smoker	region	charges	weight_status	normal_weight
0	19	female	27.900	0	yes	southwest	16884.92400	Overweight	No
1	18	male	33.770	1	no	southeast	1725.55230	Obessed	No
2	28	male	33.000	3	no	southeast	4449.46200	Obessed	No
3	33	male	22.705	0	no	northwest	21984.47061	Normal	Yes
4	32	male	28.880	0	no	northwest	3866.85520	Overweight	No

Data Exploration and Visualization - Univariate Analysis

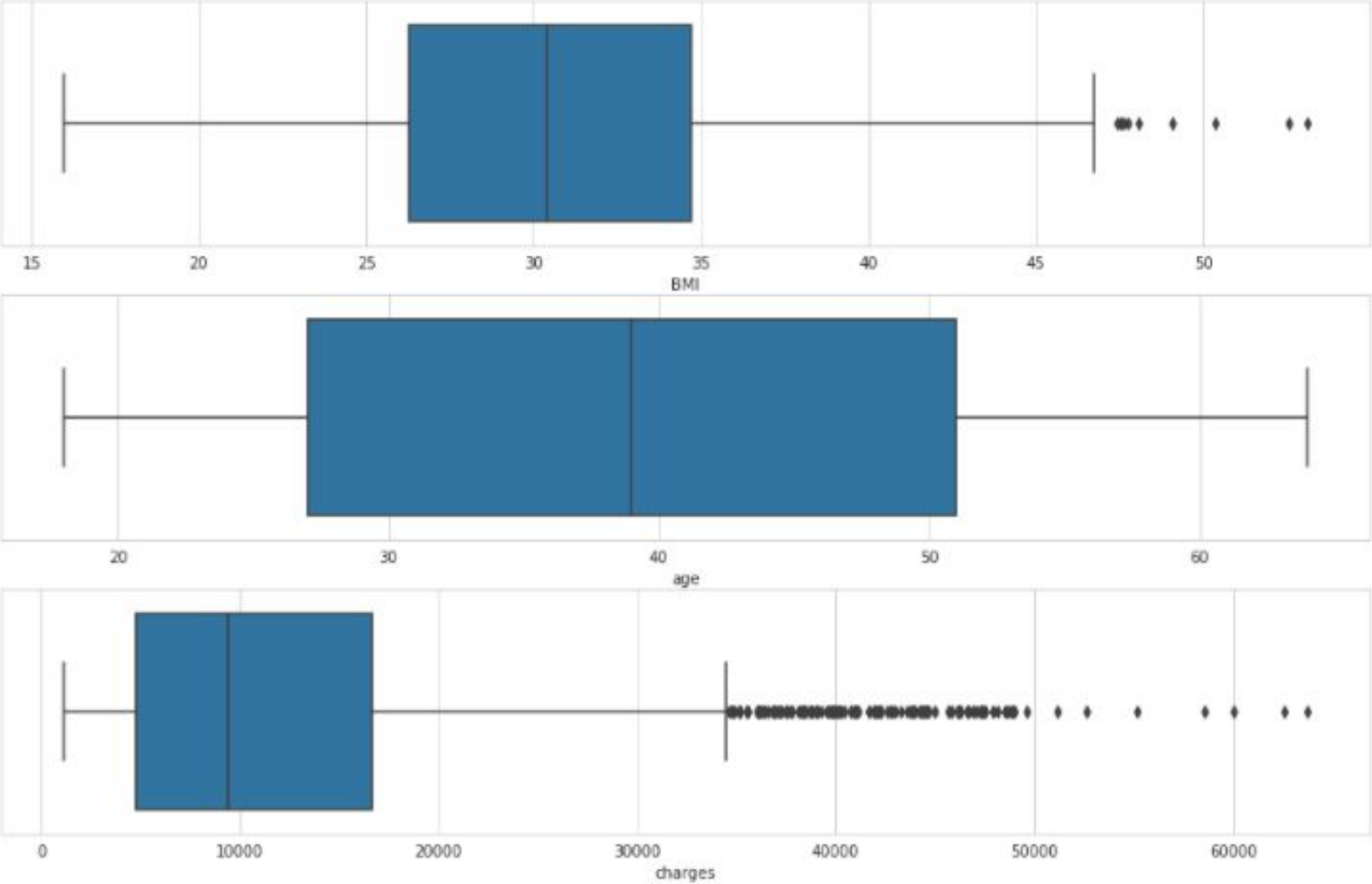






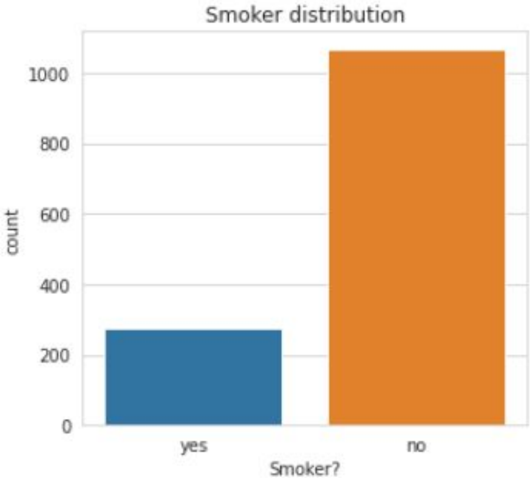
- The BMI is normally distributed.
- The age seems to assume a uniform distribution with hardly no skewness.
- Charged amount variable is rightly skewed (positive skewness).

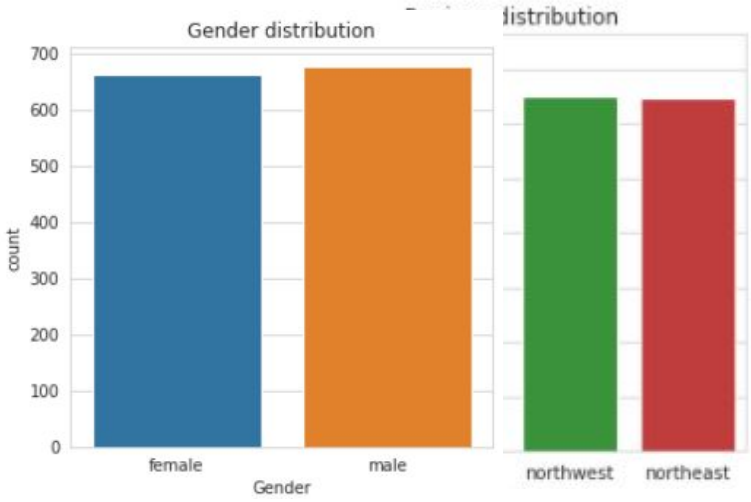
Data Exploration and Visualization - Univariate Analysis

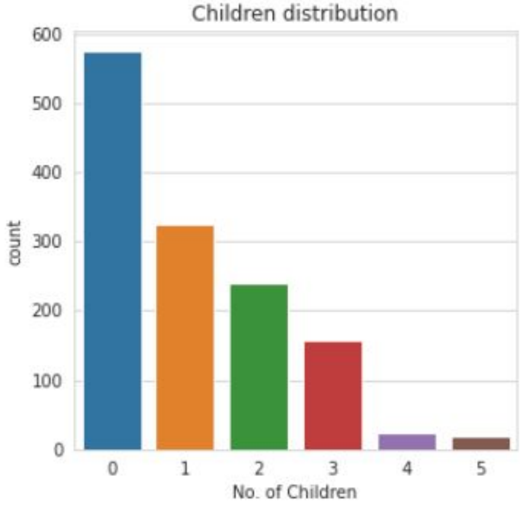


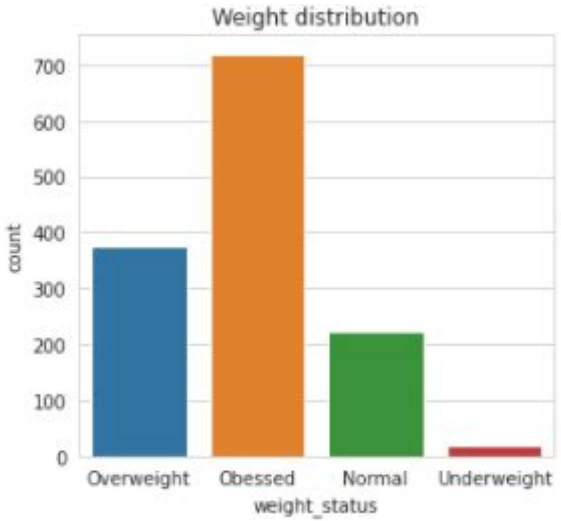
- The BMI has a few extreme values.
- The age has no extreme value.
- The charged amount is highly skewed, there are quite a lot of extreme values.

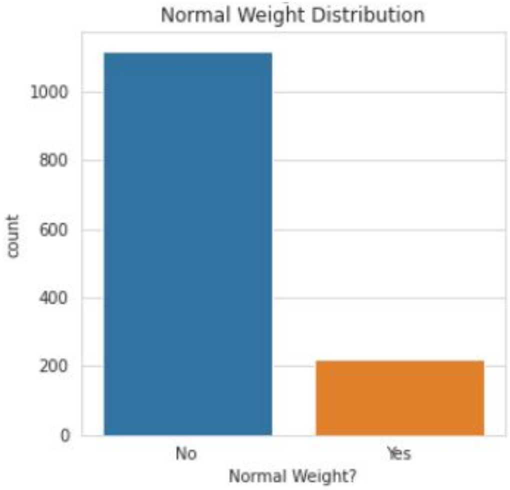
Data Exploration and Visualization - Univariate Analysis



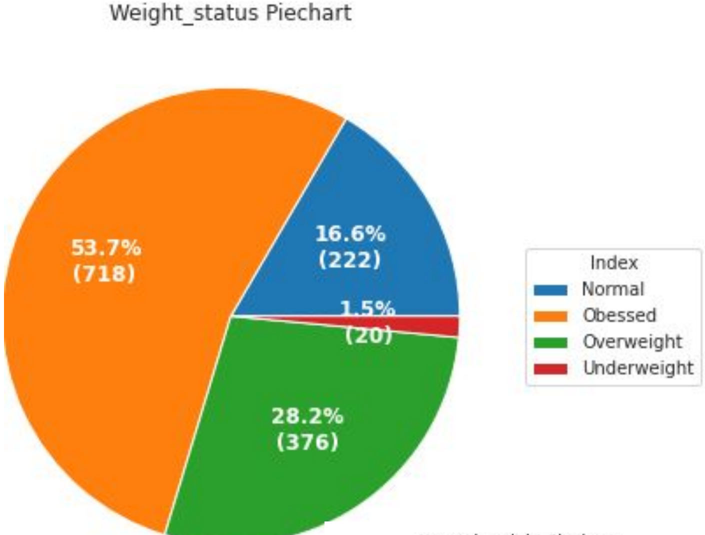


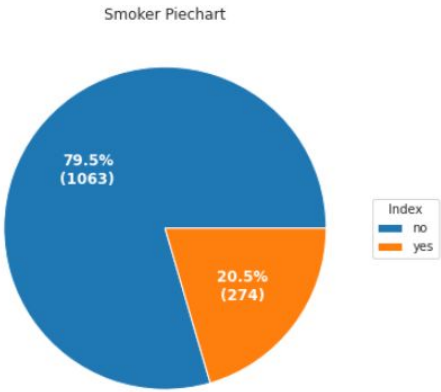


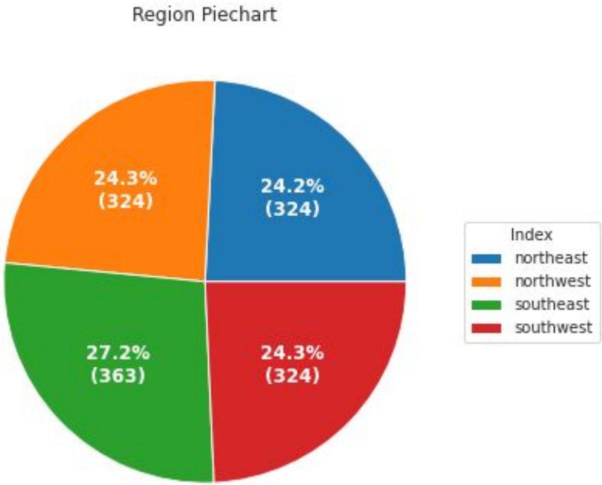


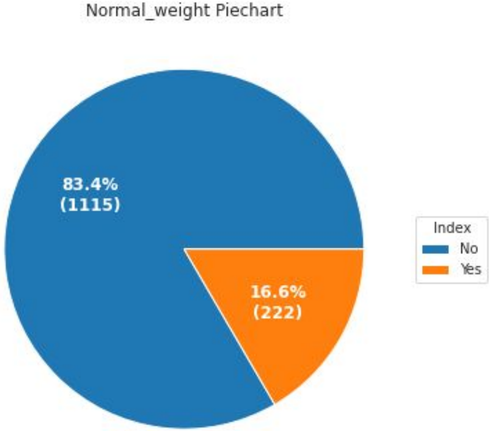


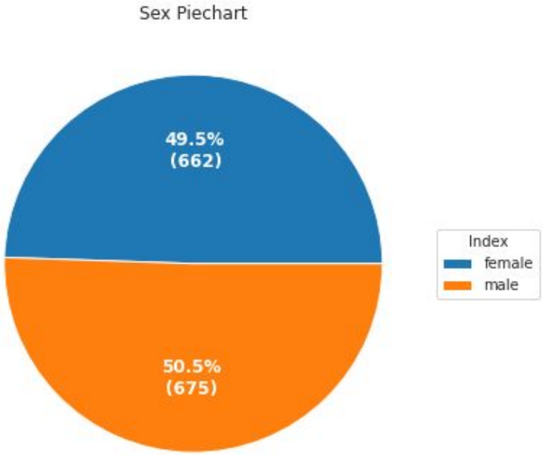
Data Exploration and Visualization - Univariate Analysis



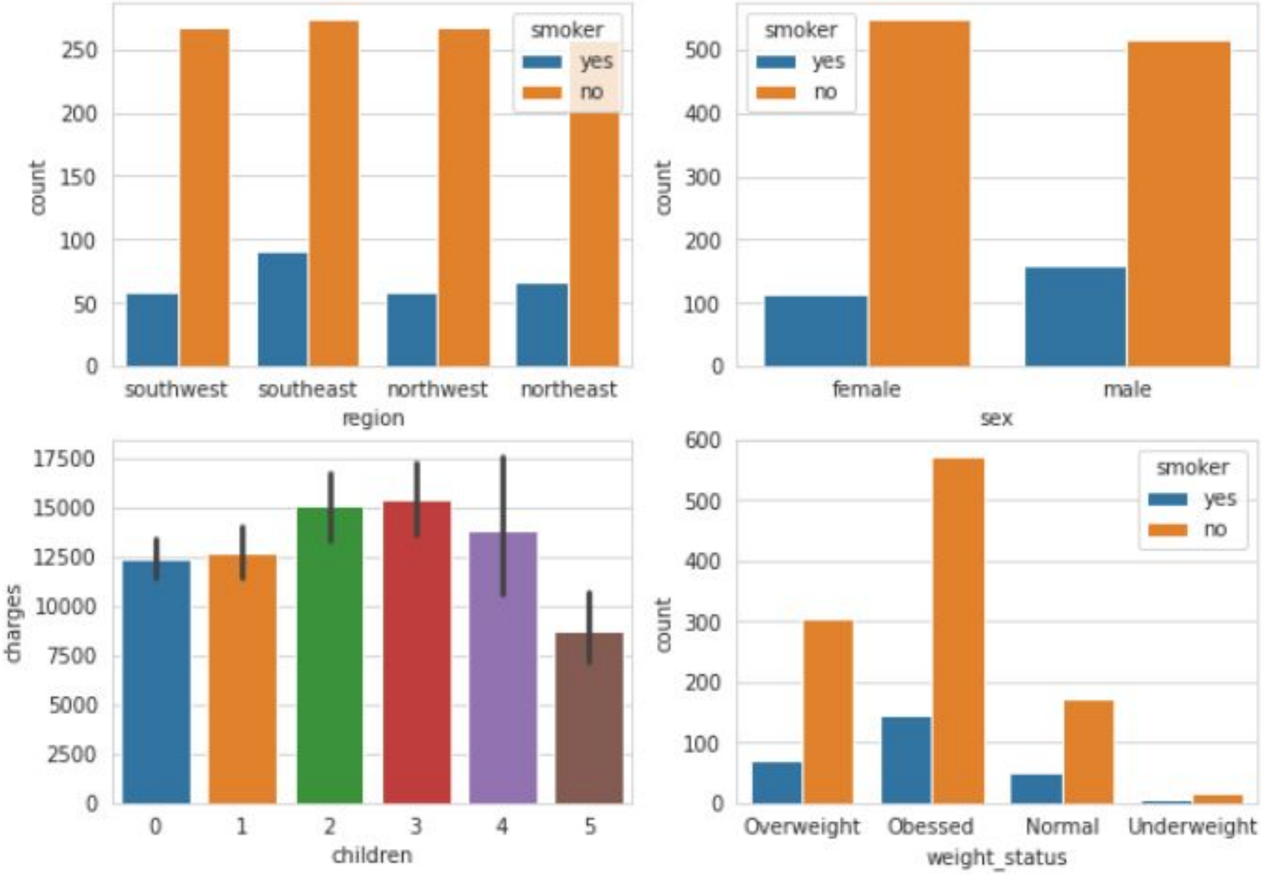






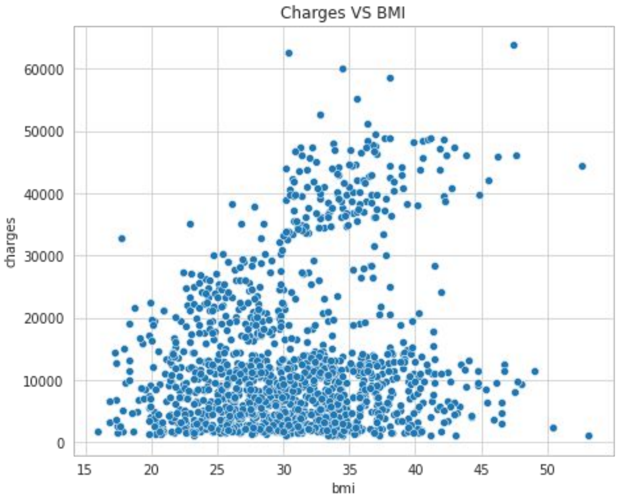


Data Exploration and Visualization - Bivariate Analysis

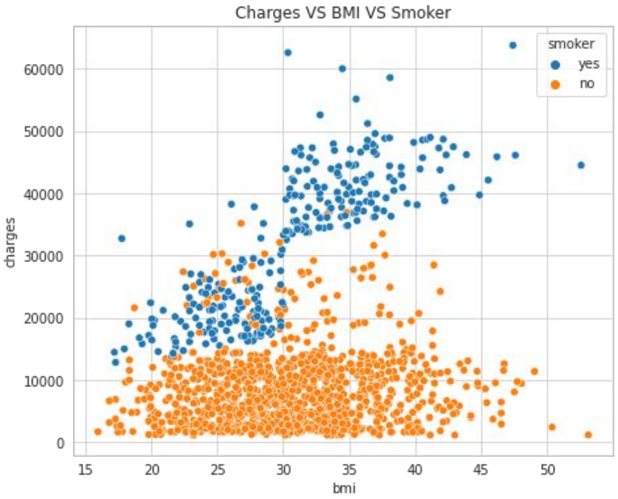


- Smoking habits of people of different regions are similar.
- There are many more male smokers than female.
- Obese individuals smoke the most, followed by the overweight.
- Approximately 85% (1138 / 1338) of the insured have less than 3 children.

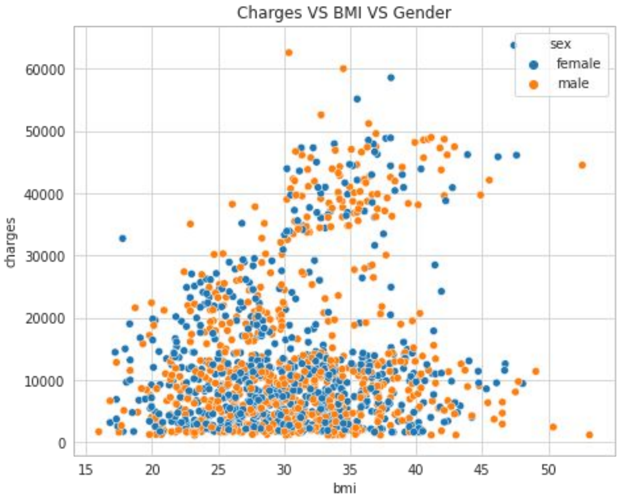
Data Exploration and Visualization - Bi/Multivariate Analysis



Charges VS BMI

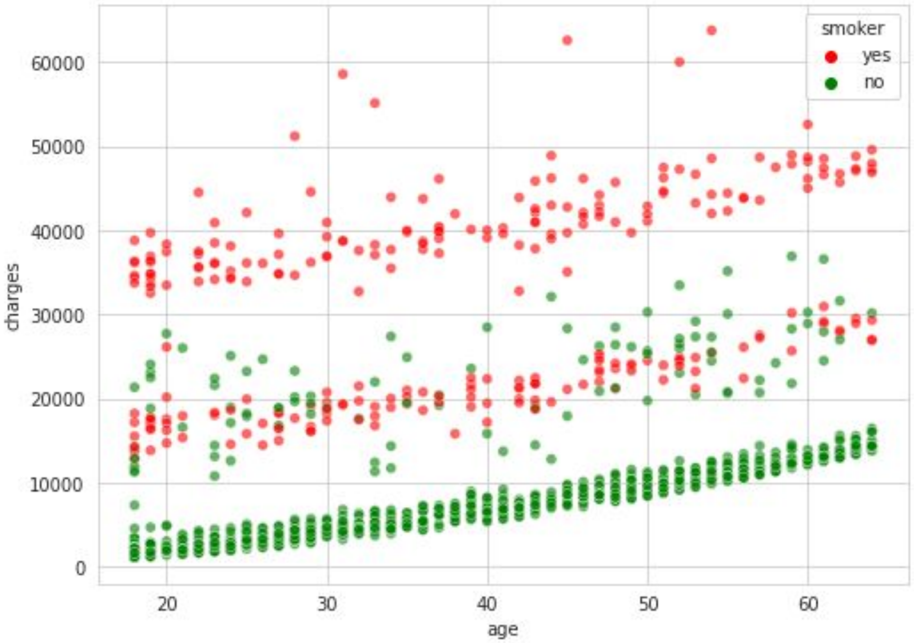


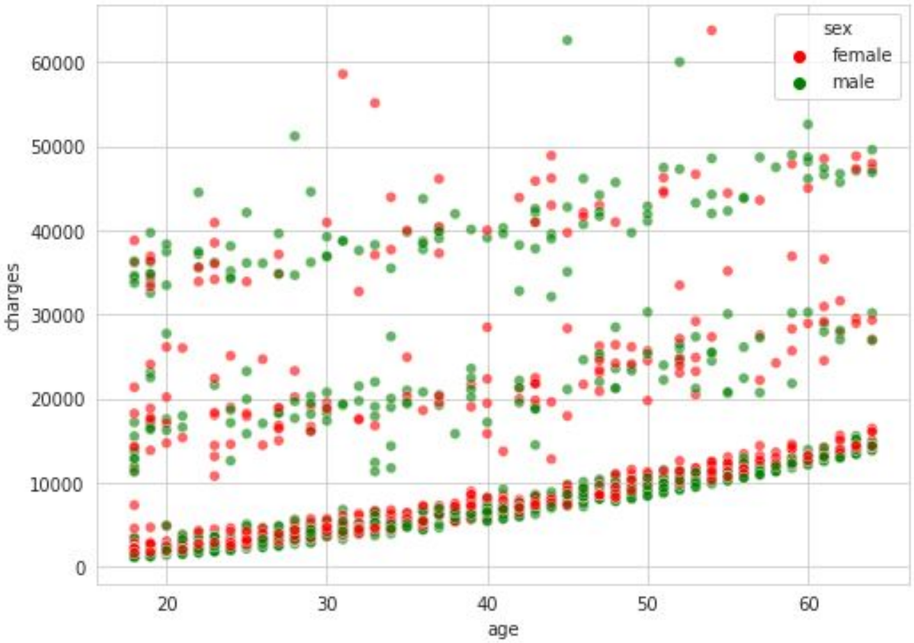
Charges VS BMI VS Smoker



Charges VS BMI VS Gender

Data Exploration and Visualization - Multivariate Analysis

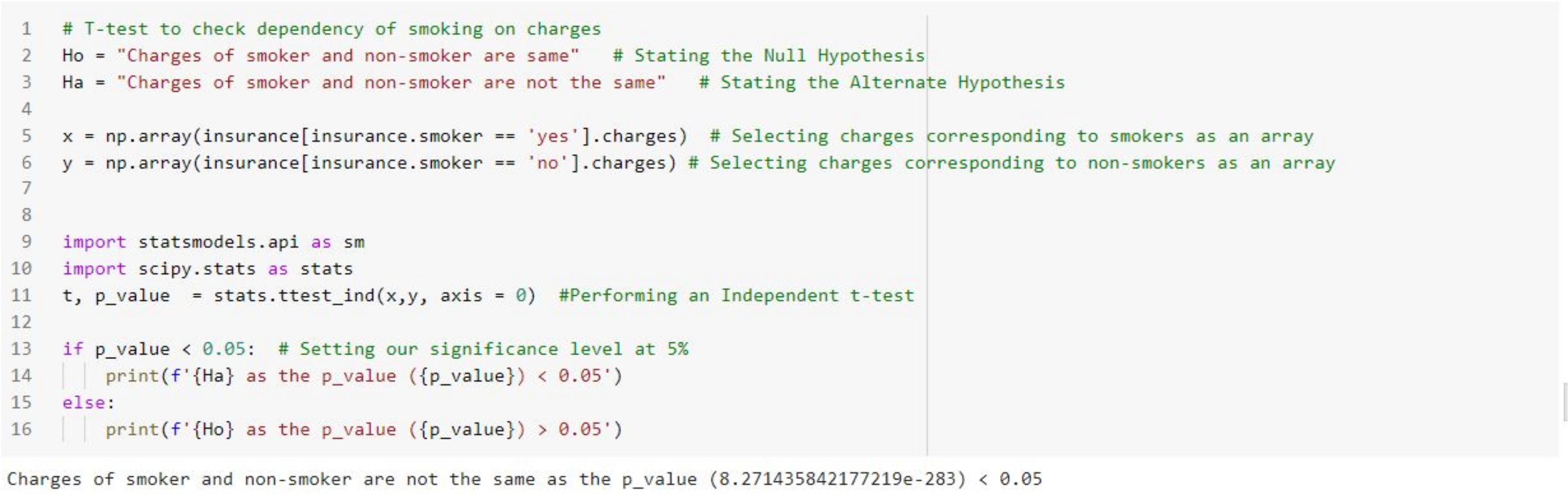




- Apparently, the non smokers’ charged amounts seem smaller than those of the smokers.
- Obviously, both genders seem to be charged even amounts across the age distribution.

Hypothesis Testing (Student T-Test)

H0: Charges of smokers and non-smokers are the same
H1: Charges of smokers and non-smokers are not the same



```
# T-test to check dependency of smoking on charges
Ho = "Charges of smoker and non-smoker are same"   # Stating the Null Hypothesis
Ha = "Charges of smoker and non-smoker are not the same"   # Stating the Alternate Hypothesis

x = np.array(insurance[insurance.smoker == 'yes'].charges)  # Selecting charges corresponding to smokers as an array
y = np.array(insurance[insurance.smoker == 'no'].charges) # Selecting charges corresponding to non-smokers as an array


import statsmodels.api as sm
import scipy.stats as stats
t, p_value  = stats.ttest_ind(x,y, axis = 0)  #Performing an Independent t-test

if p_value < 0.05:  # Setting our significance level at 5%
    print(f'{Ha} as the p_value ({p_value}) < 0.05')
else:
    print(f'{Ho} as the p_value ({p_value}) > 0.05')
```

```
Charges of smoker and non-smoker are not the same as the p_value (8.271435842177219e-283) < 0.05
```

- Charges of people who smoke differ significantly from the people who don't. so we reject the null hypothesis.

Key Findings

- Smoking significantly influences the charged amounts: Smokers are charged more than non-smokers.
- The BMI is evenly distributed across males and females.
- The amount of male smokers is higher than female smokers.
- There is no difference in the amount charged for both genders.
- Smoking habits of people of different regions are similar.
- The BMI is directly positively correlated to the amounts charged.

Future Work/Recommendation

- Predictive Analysis to predict an individual charged amount
- Health campaign to sensitize people on the health risk factors associated with smoking and Obesity.
- Diets and Exercises that reduce weight should be recommended.