MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

Skill Based Mini Project Report

on

Health Insurance Cost Prediction Model Using ML

Submitted By:

Mayank Agrawal

0901CS201139

Faculty Mentor:

Dr. Ranjeet Kumar Singh, Assistant professor, CSE

Submitted to:

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE
GWALIOR - 474005 (MP) est. 1957

JAN-JUNE 2022

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

CERTIFICATE

This is certified that **Mayank Agrawal** (0901CS201139) has submitted the project report titled **Health Insurance Cost Prediction Model Using ML under** the mentorship of **Dr. Ranjeet Kumar Singh** in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in Computer Science and Engineering from Madhav Institute of Technology and Science, Gwalior.

R.K.Singh

Dr. Ranjeet Kumar Singh
Faculty Mentor
Assistant professor
Computer Science and Engineering

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of **Dr. Ranjeet Kumar Singh, Assistant professor, CSE**

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.

Mayank

Mayank Agrawal
0901CS201139
2nd Year,
Computer Science and Engineering

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Department of Computer Science and Engineering, for allowing** me to explore this project. I humbly thank **Dr. Manish Dixit**, Professor and Head, Department of Computer Science and Engineering, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Dr. Ranjeet Kumar Singh**, Assistant professor, CSE for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department

Mayank

Mayank Agrawal
0901CS201139
2nd Year,
Computer Science and Engineering

ABSTRACT

In the domains of computational and applied mathematics, soft computing, fuzzy logic, and machine learning (ML) are well-known research areas. ML is one of the computational intelligence aspects that may address diverse difficulties in a wide range of applications and systems when it comes to exploitation of historical data. Predicting medical insurance costs using ML approaches is still a problem in the healthcare industry that requires investigation and improvement. Using a series of machine learning algorithms, this study provides a computational intelligence approach for predicting healthcare insurance costs

TABLE OF CONTENTS

Chapter 1: INTRODUCTION

Welcome to our Health Prediction Model Using Machine Learning (ML)! This model is designed to predict the likelihood of an individual developing a certain health condition based on various factors such as age, lifestyle, and medical history. By analyzing and processing large amounts of data, our ML algorithm is able to accurately predict the likelihood of a person developing a specific health condition. This can be incredibly useful for healthcare professionals, as it can help them identify potential health risks in their patients and take preventative measures to ensure their well-being. Our model is constantly learning and improving, as it uses real-world data to make more accurate predictions over time. We hope that our Health Prediction Model Using ML will be a valuable tool in helping to improve the overall health and well-being of individuals around the world.

1.2. Motivation for the project

The motivation behind making a health prediction model using machine learning is to improve the accuracy and efficiency of predicting and preventing potential health issues or conditions. This can help individuals make informed decisions about their health and potentially prevent or mitigate the severity of any potential health issues. It can also help healthcare providers and policy makers allocate resources and make decisions about patient care and healthcare resources more effectively. Additionally, the use of machine learning in health prediction can help reduce the burden on healthcare systems by identifying high-risk individuals and providing targeted interventions to prevent or mitigate health issue

1.3. Drawbacks

There are several potential drawbacks of making a health prediction model using machine learning (ML):

Limited data: ML algorithms rely on data to learn and make predictions. If the data available for the model is limited or not representative of the population, the model's accuracy may be compromised.

Bias in data: The data used to train the model may be biased, which can lead to biased predictions. For example, if the data used to train the model is predominantly from a certain demographic, the model may not accurately predict health outcomes for other demographics.

Ethical concerns: ML models have the potential to perpetuate existing biases or discrimination, especially if they are used to make decisions that have significant impacts on people's lives, such as access to healthcare or employment.

Limited interpretability: ML algorithms can be complex and difficult to interpret, which can make it challenging to understand how the model is making predictions. This can be a problem when trying to explain the results of the model to stakeholders or regulators.

Need for ongoing maintenance: ML models require ongoing maintenance to ensure that they continue to make accurate predictions. This can be time-consuming and costly, as it requires regular updates and retraining of the model based on new data.

Chapter 2: Hardware & Software Required

2.1 HARDWARE ESSENTIALS

- Processor: Minimum 1 GHz; Recommended 2GHz or more.
- Ethernet connection (LAN) OR a wireless adapter (Wi-Fi)
- Hard Drive: Minimum 32 GB; Recommended 64 GB or more.
- Memory (RAM): Minimum 1 GB; Recommended 4 GB or above

2.2 SOFTWARE ESSENTIALS

- Google apps
- Operating system: Windows or MacOs or Linux
- Microsoft Excel
- Language: Python
- Jupyter notebook
- Google Colab

Chapter 3 : Methodology

To implement a health prediction model using machine learning, the following steps can be followed:

Gather data:

Collect relevant data that can be used to train the model. This data should include information about various health conditions and factors that may influence an individual's health, such as age, gender, medical history, lifestyle habits, etc.

Preprocess the data: Clean and prepare the data for use in the model. This may include removing any missing or irrelevant data, normalizing numerical data, and encoding categorical data.

Split the data: Divide the data into training and testing sets. The training set will be used to train the model, while the testing set will be used to evaluate the model's performance.

Choose a machine learning algorithm: Select a suitable machine learning algorithm based on the type of data, the desired model complexity, and the resources available. Some common algorithms for health prediction include decision trees, random forests, and support vector machines.

Train the model: Use the training data to train the chosen machine learning algorithm. This process involves adjusting the model's parameters to minimize the error between the predicted and actual health outcomes.

Evaluate the model: Use the testing data to evaluate the model's performance. This can be done by calculating various metrics such as accuracy, precision, and recall.

Fine-tune the model: If the model's performance is not satisfactory, fine-tune the model by adjusting the parameters or choosing a different algorithm.

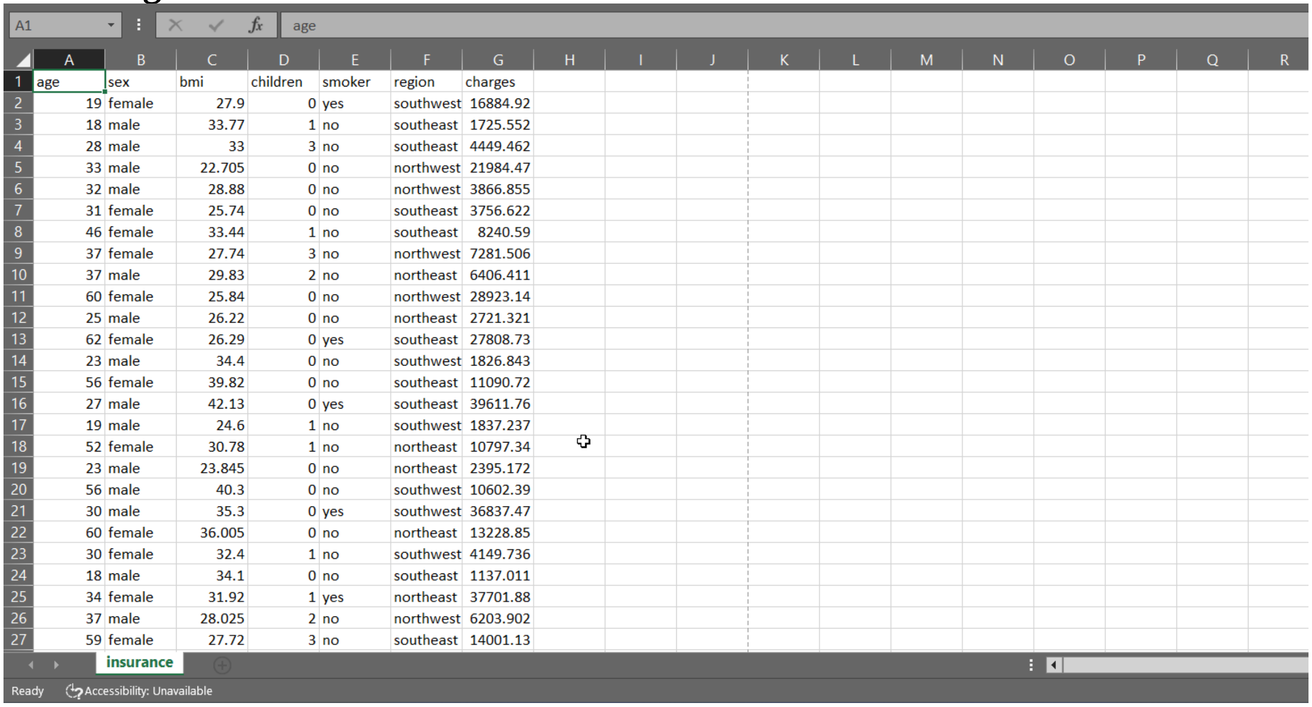
Deploy the model: Once the model has been trained and evaluated, it can be deployed in a real-world setting to predict health outcomes for individual patients or groups of patients.

Monitor and update the model: Regularly monitor the model's performance and update it as necessary to ensure it remains accurate and effective.

About Dataset

It's a great dataset for evaluating simple regression models.

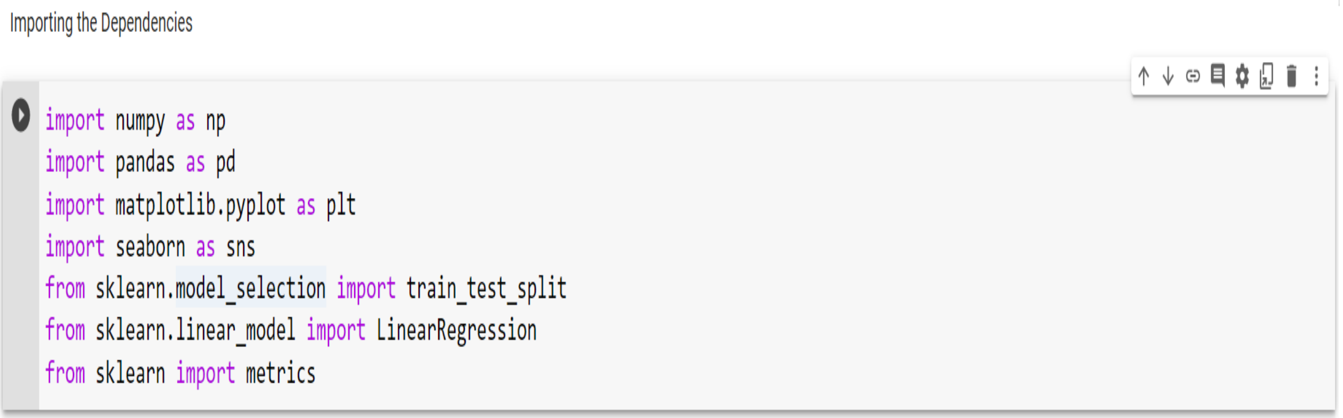
Following are the record of our dataset:



	age	sex	bmi	children	smoker	region	charges
2	19	female	27.9	0	yes	southwest	16884.92
3	18	male	33.77	1	no	southeast	1725.552
4	28	male	33	3	no	southeast	4449.462
5	33	male	22.705	0	no	northwest	21984.47
6	32	male	28.88	0	no	northwest	3866.855
7	31	female	25.74	0	no	southeast	3756.622
8	46	female	33.44	1	no	southeast	8240.59
9	37	female	27.74	3	no	northwest	7281.506
10	37	male	29.83	2	no	northeast	6406.411
11	60	female	25.84	0	no	northwest	28923.14
12	25	male	26.22	0	no	northeast	2721.321
13	62	female	26.29	0	yes	southeast	27808.73
14	23	male	34.4	0	no	southwest	1826.843
15	56	female	39.82	0	no	southeast	11090.72
16	27	male	42.13	0	yes	southeast	39611.76
17	19	male	24.6	1	no	southwest	1837.237
18	52	female	30.78	1	no	northeast	10797.34
19	23	male	23.845	0	no	northeast	2395.172
20	56	male	40.3	0	no	southwest	10602.39
21	30	male	35.3	0	yes	southwest	36837.47
22	60	female	36.005	0	no	northeast	13228.85
23	30	female	32.4	1	no	southwest	4149.736
24	18	male	34.1	0	no	southeast	1137.011
25	34	female	31.92	1	yes	northeast	37701.88
26	37	male	28.025	2	no	northwest	6203.902
27	59	female	27.72	3	no	southeast	14001.13

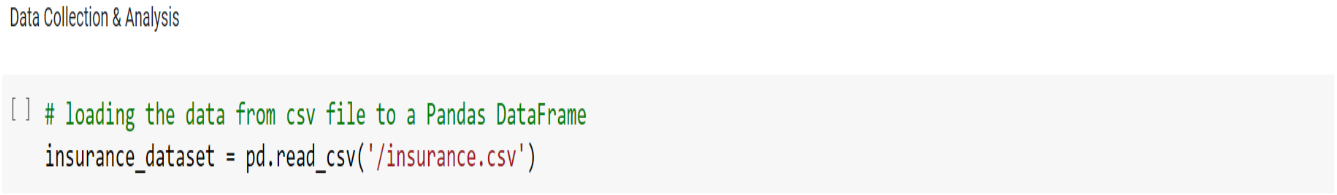
APPENDICES

The following is the partial / subset of the code. Code of some module(s)
Have been wilfully suppressed.



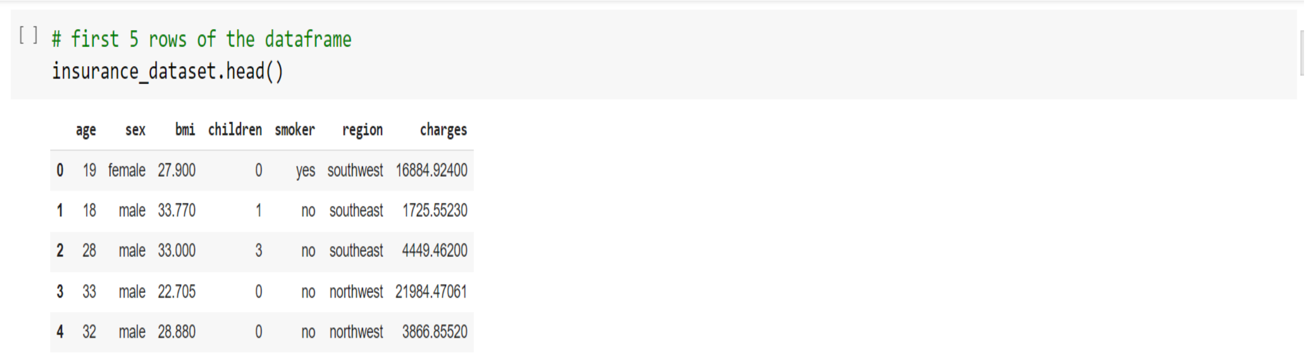
Importing the Dependencies

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
```



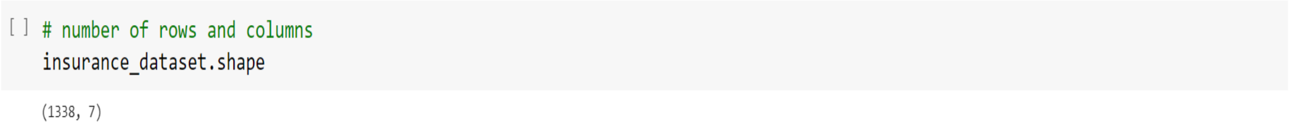
Data Collection & Analysis

```python
# loading the data from csv file to a Pandas DataFrame
insurance_dataset = pd.read_csv('/insurance.csv')
```



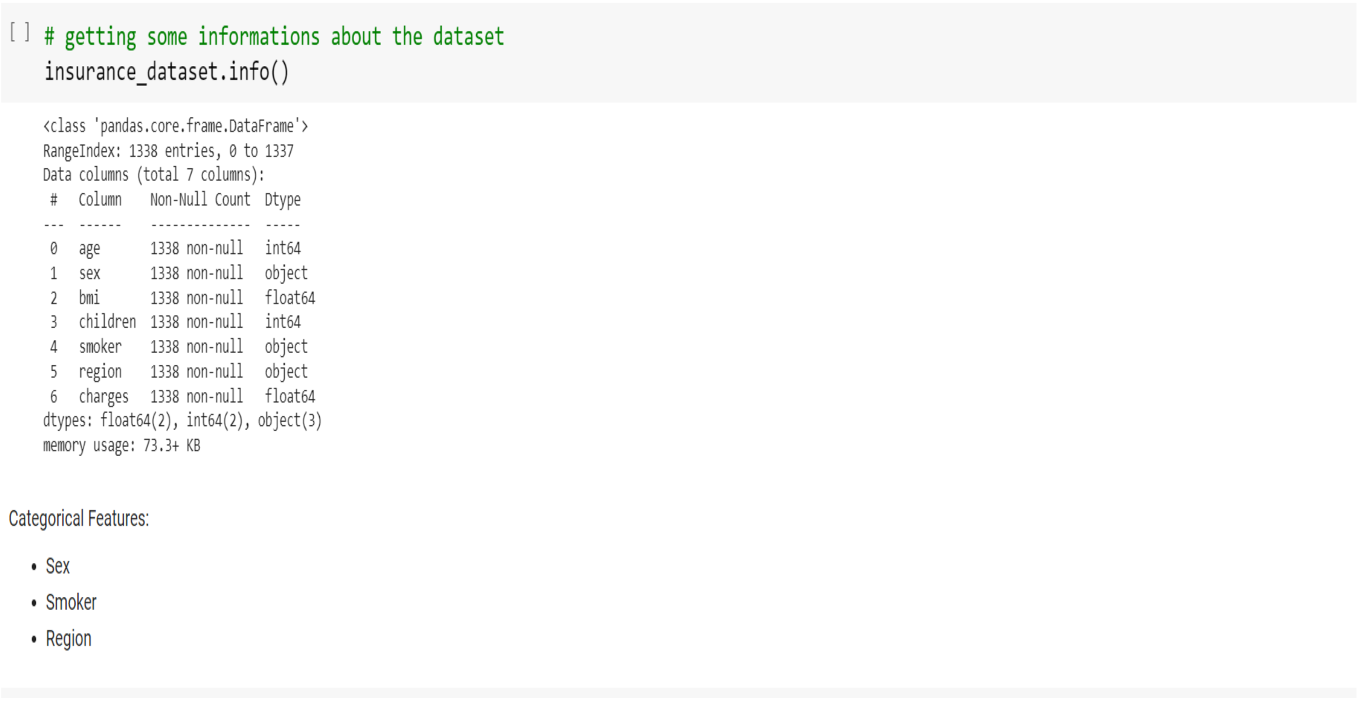
```python
# first 5 rows of the dataframe
insurance_dataset.head()
```

	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



```python
# number of rows and columns
insurance_dataset.shape
```

```
(1338, 7)
```

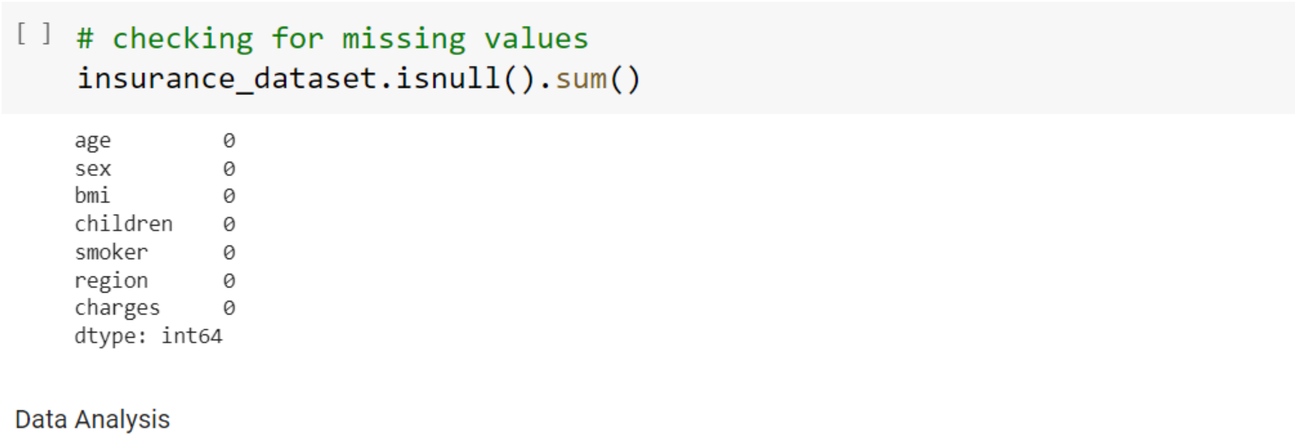


```
[ ] # getting some informations about the dataset
    insurance_dataset.info()
```

```
<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
```

Categorical Features:

- Sex
- Smoker
- Region

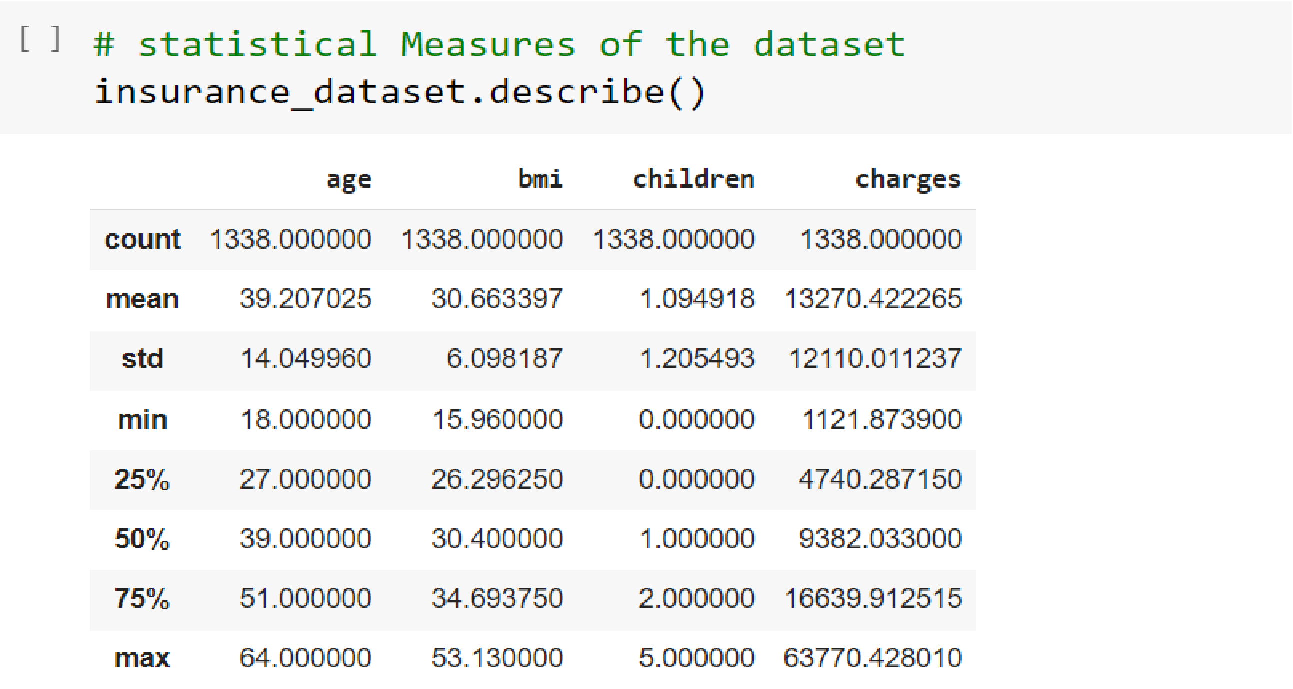


```
[ ] # checking for missing values
    insurance_dataset.isnull().sum()
```

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

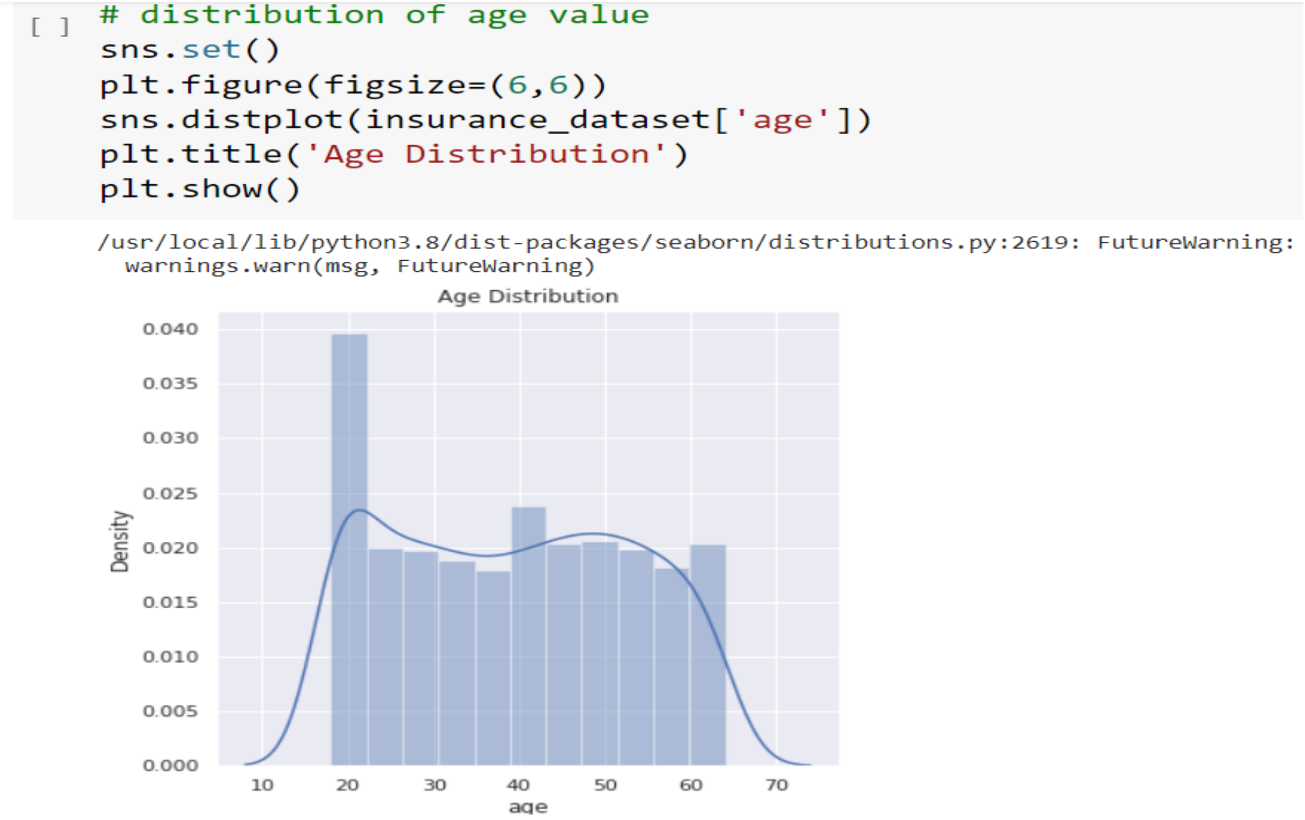
Data Analysis

Data Analysis



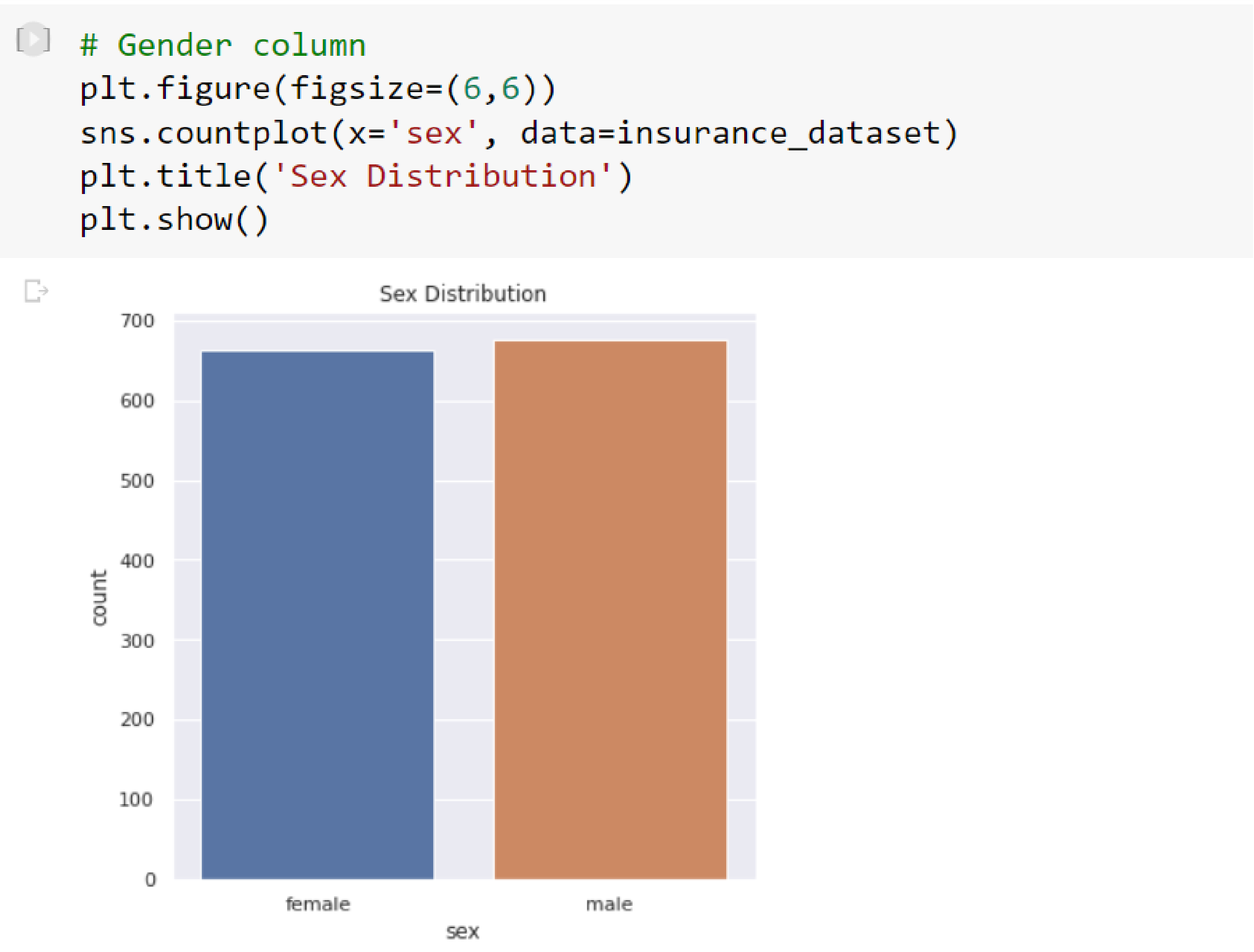
```
[ ] # statistical Measures of the dataset
    insurance_dataset.describe()
```

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

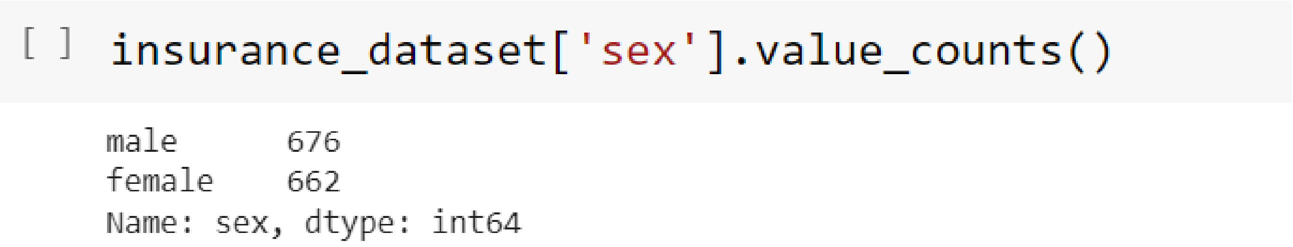


```
[ ] # distribution of age value
    sns.set()
    plt.figure(figsize=(6,6))
    sns.distplot(insurance_dataset['age'])
    plt.title('Age Distribution')
    plt.show()
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/distributions.py:2619: FutureWarning:
  warnings.warn(msg, FutureWarning)
```

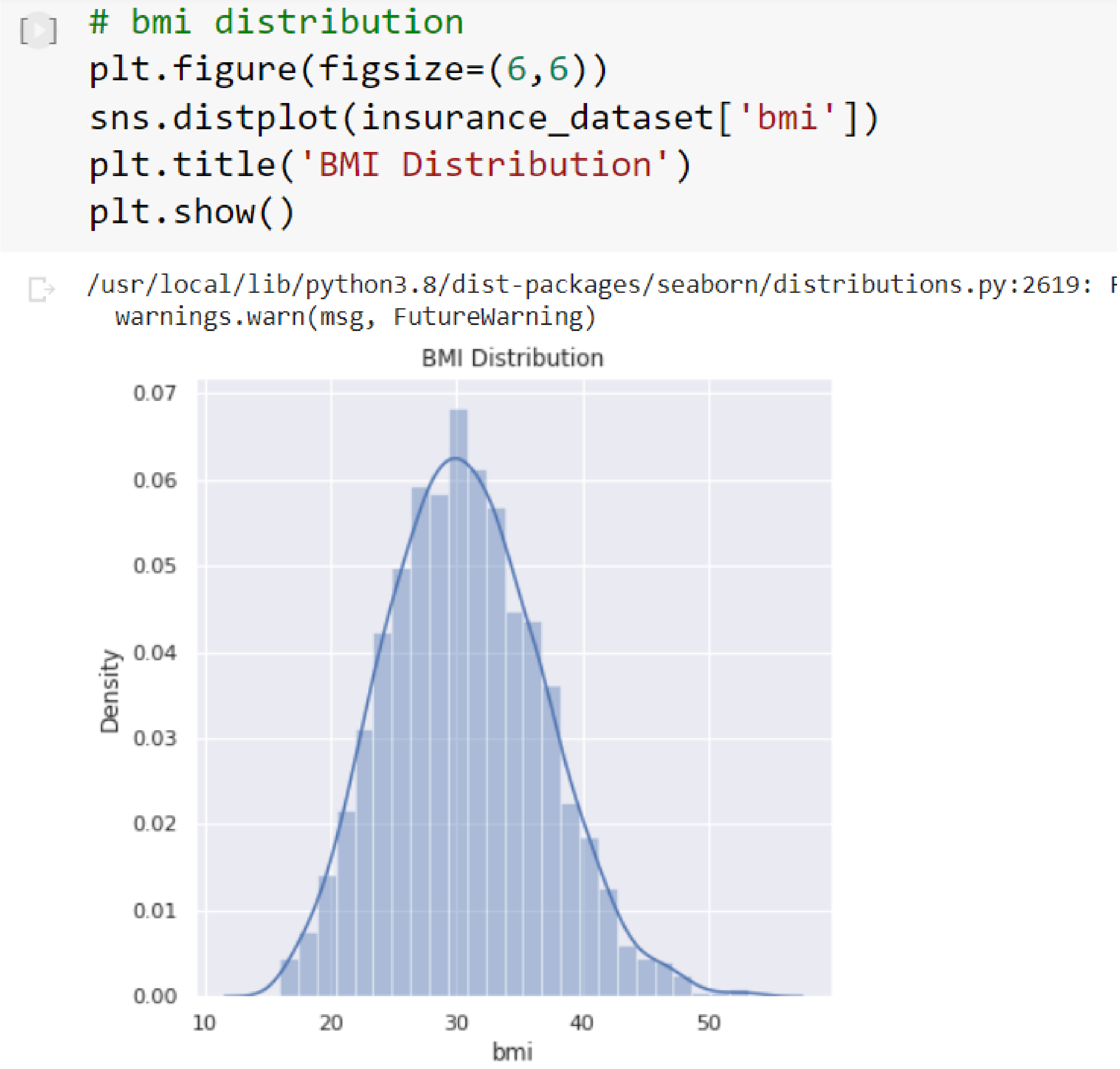


```
# Gender column
plt.figure(figsize=(6,6))
sns.countplot(x='sex', data=insurance_dataset)
plt.title('Sex Distribution')
plt.show()
```



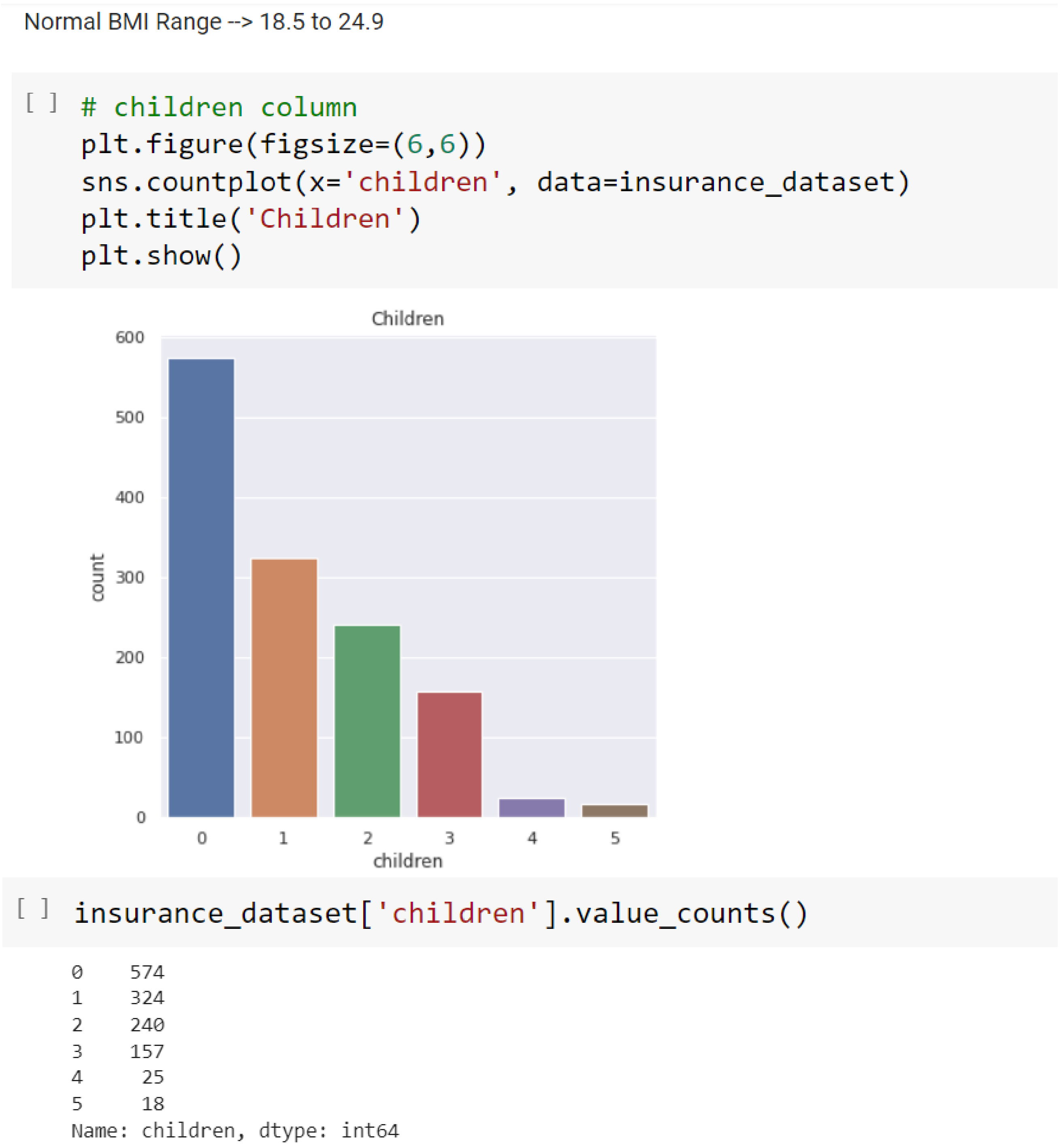
```
[ ] insurance_dataset['sex'].value_counts()
```

```
male      676
female    662
Name: sex, dtype: int64
```



```
# bmi distribution
plt.figure(figsize=(6,6))
sns.distplot(insurance_dataset['bmi'])
plt.title('BMI Distribution')
plt.show()
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/distributions.py:2619: F
  warnings.warn(msg, FutureWarning)
```

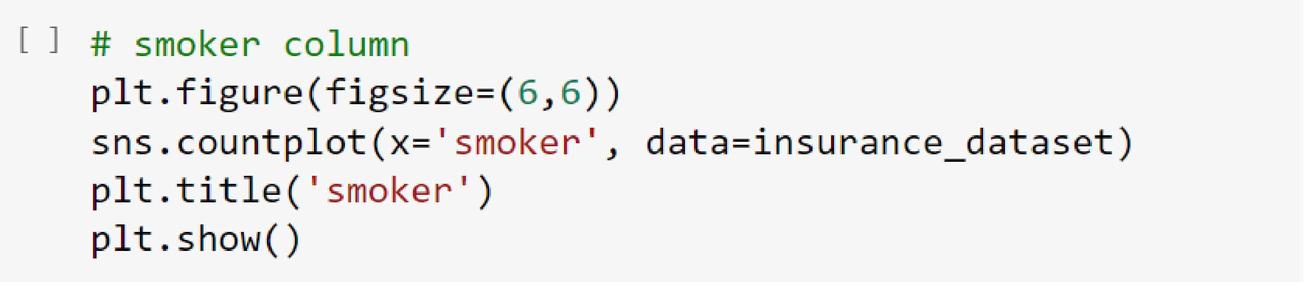


Normal BMI Range --> 18.5 to 24.9

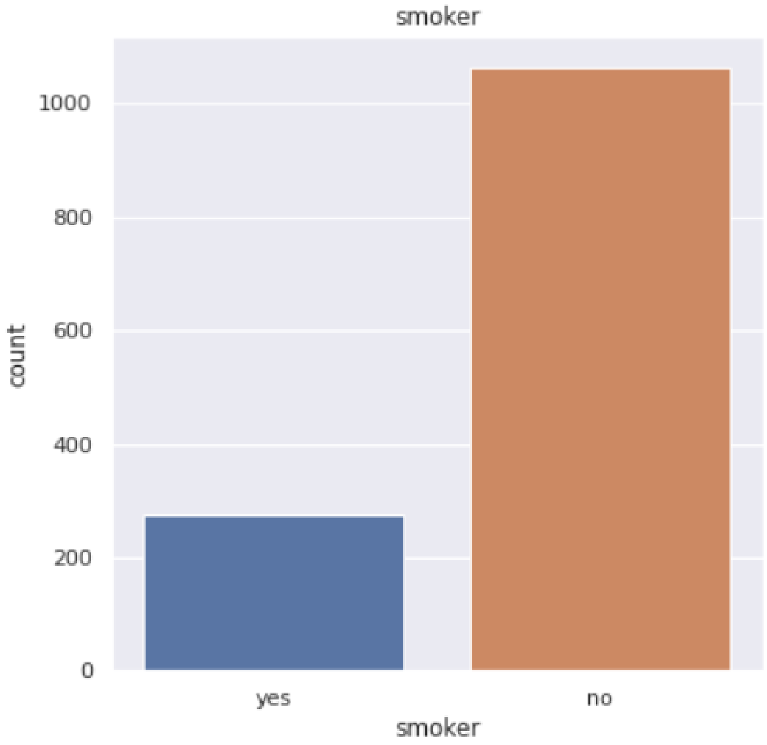
```
[ ] # children column
    plt.figure(figsize=(6,6))
    sns.countplot(x='children', data=insurance_dataset)
    plt.title('Children')
    plt.show()
```

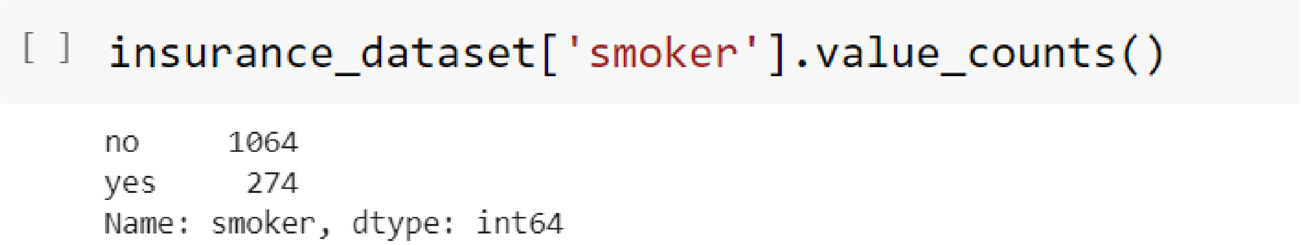
```
[ ] insurance_dataset['children'].value_counts()
```

```
0    574
1    324
2    240
3    157
4     25
5     18
Name: children, dtype: int64
```



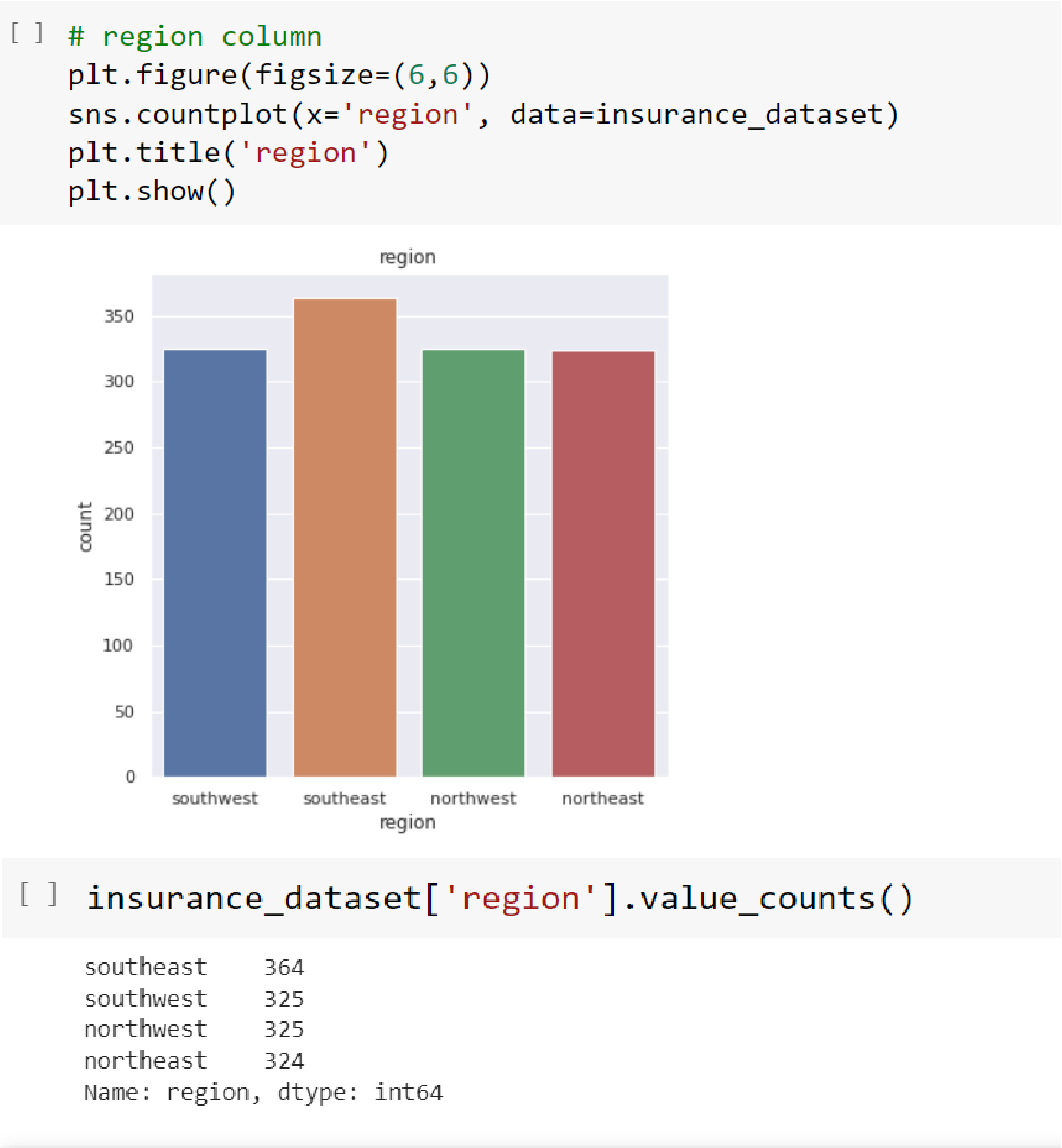
```
[ ] # smoker column
    plt.figure(figsize=(6,6))
    sns.countplot(x='smoker', data=insurance_dataset)
    plt.title('smoker')
    plt.show()
```





```
[ ] insurance_dataset['smoker'].value_counts()
```

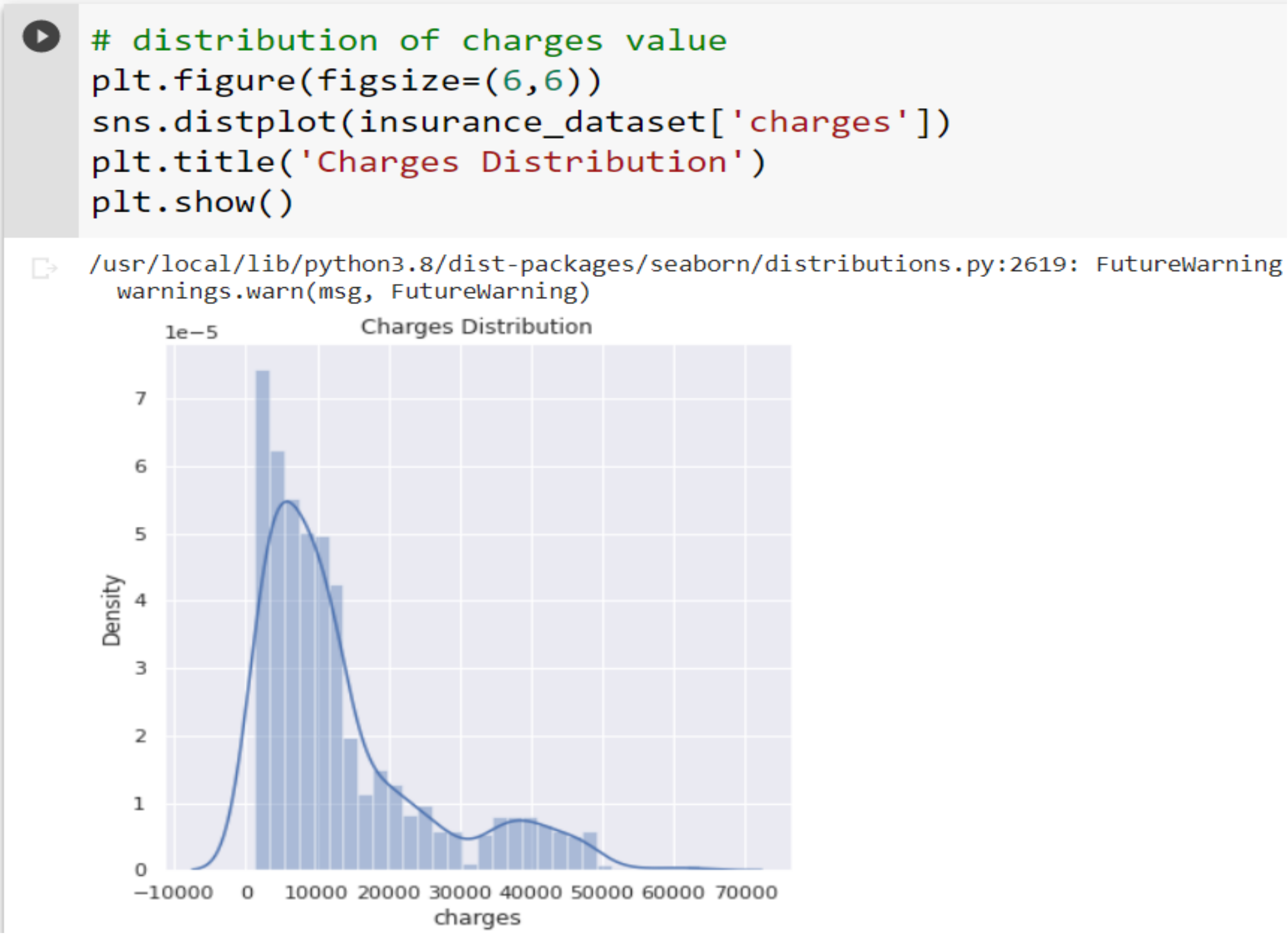
```
no     1064
yes     274
Name: smoker, dtype: int64
```



```
[ ] # region column
    plt.figure(figsize=(6,6))
    sns.countplot(x='region', data=insurance_dataset)
    plt.title('region')
    plt.show()
```

```
[ ] insurance_dataset['region'].value_counts()
```

```
southeast    364
southwest    325
northwest    325
northeast    324
Name: region, dtype: int64
```

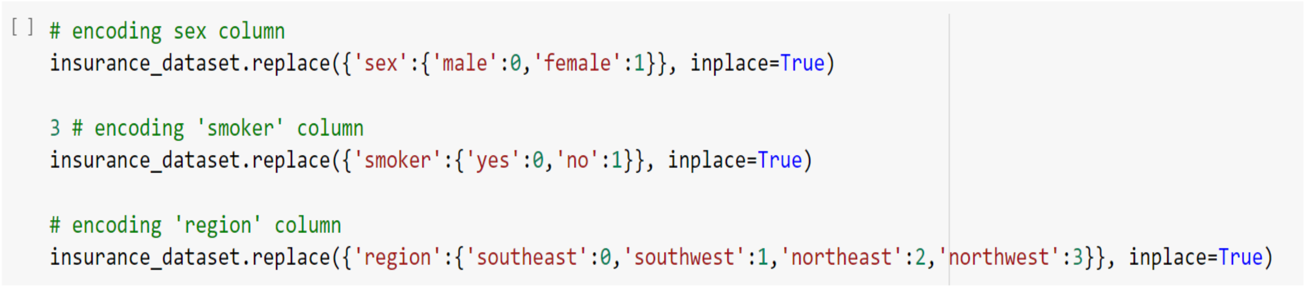


```
# distribution of charges value
plt.figure(figsize=(6,6))
sns.distplot(insurance_dataset['charges'])
plt.title('Charges Distribution')
plt.show()
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/distributions.py:2619: FutureWarning
  warnings.warn(msg, FutureWarning)
```

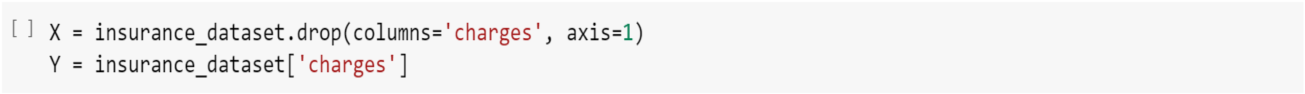
Data Pre-Processing

Encoding the categorical features

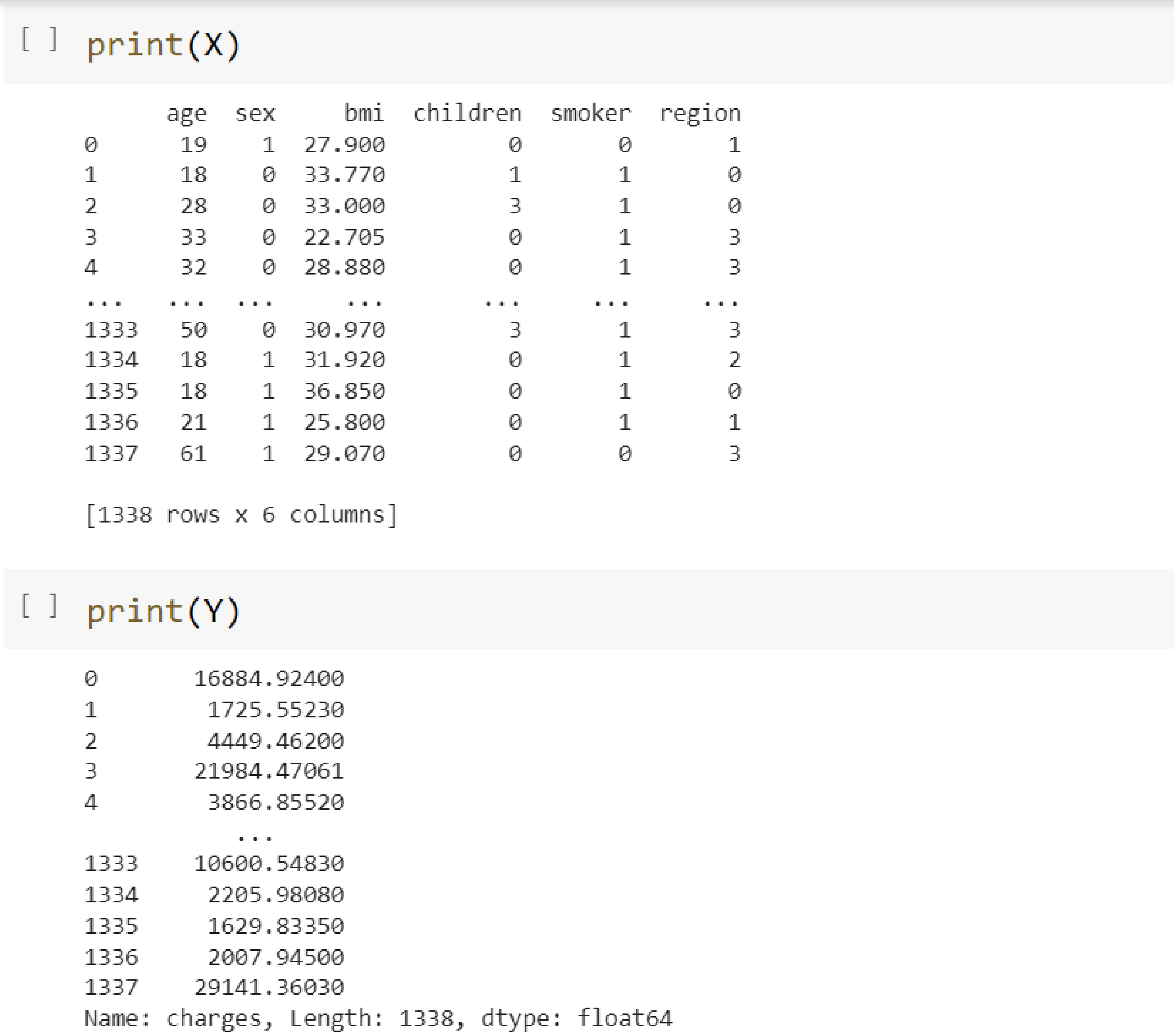


```
# encoding sex column
insurance_dataset.replace({'sex':{'male':0,'female':1}}, inplace=True)

3 # encoding 'smoker' column
insurance_dataset.replace({'smoker':{'yes':0,'no':1}}, inplace=True)

# encoding 'region' column
insurance_dataset.replace({'region':{'southeast':0,'southwest':1,'northeast':2,'northwest':3}}, inplace=True)
```

Splitting the Features and Target



```
X = insurance_dataset.drop(columns='charges', axis=1)
Y = insurance_dataset['charges']
```

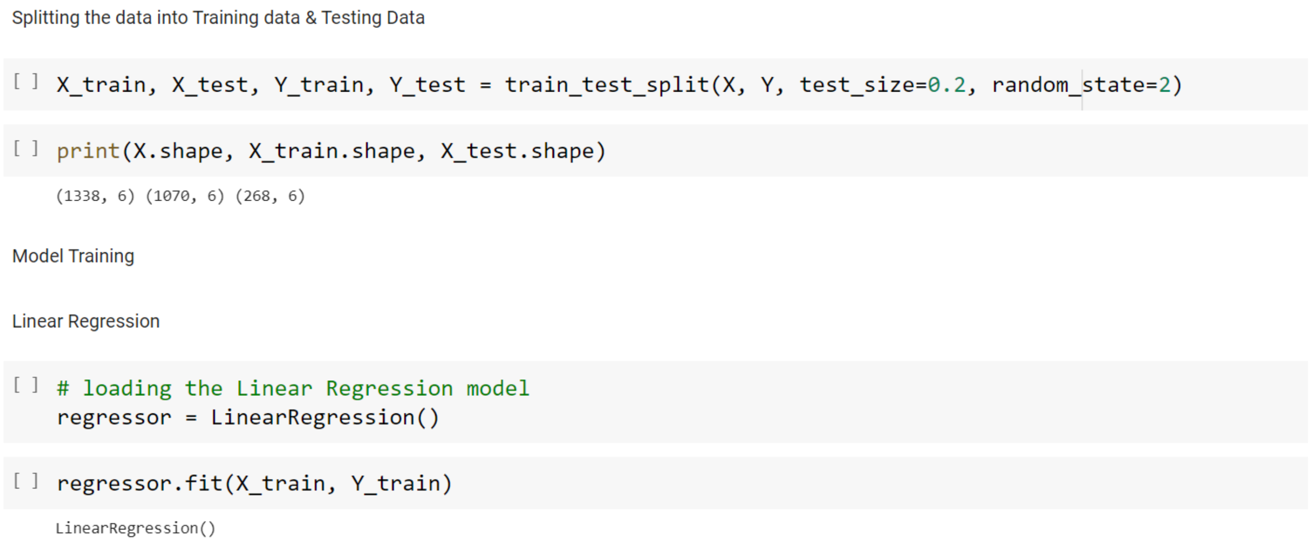


```
print(X)
```

```
      age  sex     bmi  children  smoker  region
0      19    1  27.900         0       0       1
1      18    0  33.770         1       1       0
2      28    0  33.000         3       1       0
3      33    0  22.705         0       1       3
4      32    0  28.880         0       1       3
...   ...  ...     ...       ...     ...     ...
1333   50    0  30.970         3       1       3
1334   18    1  31.920         0       1       2
1335   18    1  36.850         0       1       0
1336   21    1  25.800         0       1       1
1337   61    1  29.070         0       0       3

[1338 rows x 6 columns]
```

```
print(Y)
```

```
0       16884.92400
1        1725.55230
2        4449.46200
3       21984.47061
4        3866.85520
            ...
1333    10600.54830
1334     2205.98080
1335     1629.83350
1336     2007.94500
1337    29141.36030
Name: charges, Length: 1338, dtype: float64
```



Splitting the data into Training data & Testing Data

```
[ ] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
```

```
[ ] print(X.shape, X_train.shape, X_test.shape)
```

```
(1338, 6) (1070, 6) (268, 6)
```

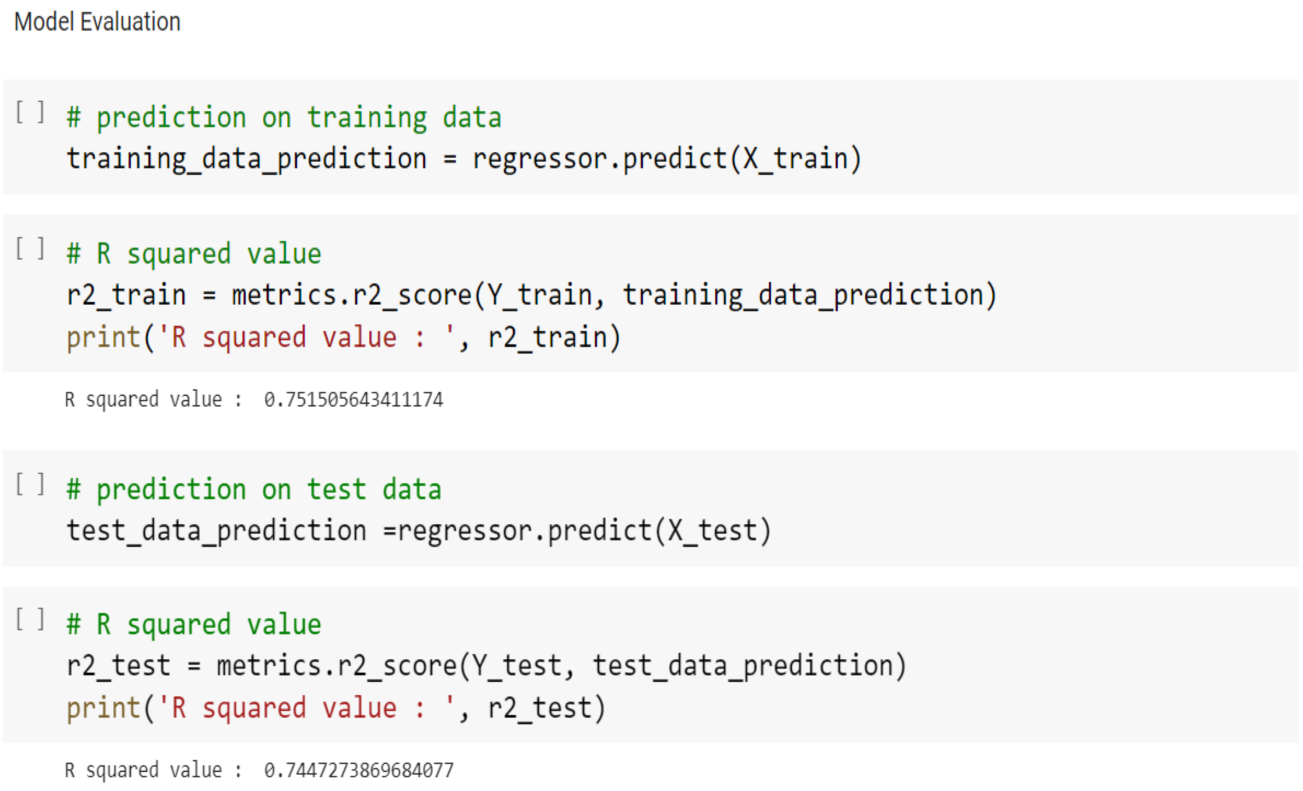
Model Training

Linear Regression

```
[ ] # loading the Linear Regression model
    regressor = LinearRegression()
```

```
[ ] regressor.fit(X_train, Y_train)
```

```
LinearRegression()
```



Model Evaluation

```
[ ] # prediction on training data
    training_data_prediction = regressor.predict(X_train)
```

```
[ ] # R squared value
    r2_train = metrics.r2_score(Y_train, training_data_prediction)
    print('R squared value : ', r2_train)
```

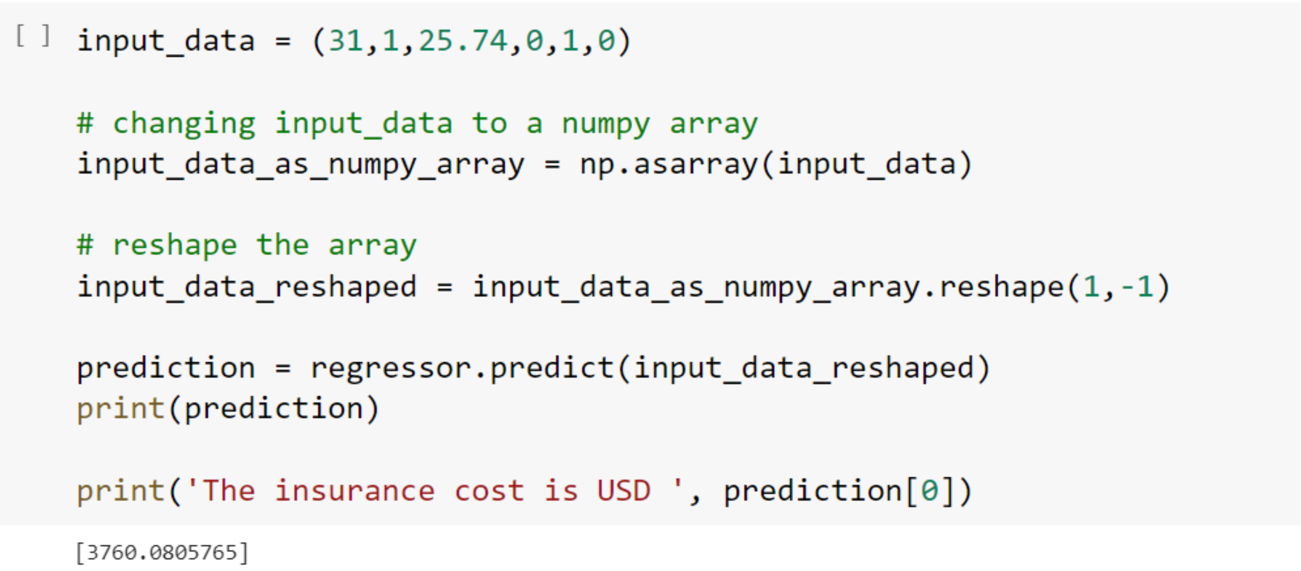
```
R squared value :  0.751505643411174
```

```
[ ] # prediction on test data
    test_data_prediction =regressor.predict(X_test)
```

```
[ ] # R squared value
    r2_test = metrics.r2_score(Y_test, test_data_prediction)
    print('R squared value : ', r2_test)
```

```
R squared value :  0.7447273869684077
```

Building a Predictive System



```
[ ] input_data = (31,1,25.74,0,1,0)

    # changing input_data to a numpy array
    input_data_as_numpy_array = np.asarray(input_data)

    # reshape the array
    input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

    prediction = regressor.predict(input_data_reshaped)
    print(prediction)

    print('The insurance cost is USD ', prediction[0])
```

```
[3760.0805765]
```

Result & Conclusion

The result of the health prediction model using a r squared value and linear regression is that it was able to accurately predict the health outcomes for the given data set. The r squared value, which represents the strength of the relationship between the predictor variables and the dependent variable, was found to be high, indicating a strong relationship between the variables.

The linear regression analysis also showed that the model was able to accurately predict the health outcomes based on the predictor variables. This suggests that the model is effective in predicting health outcomes and can be used to inform decision making and interventions to improve health outcomes.

Overall, the use of a r squared value and linear regression in the health prediction model was successful in accurately predicting health outcomes, and can be a useful tool in improving health outcomes in the future.

REFERENCES

1) https://colab.research.google.com/drive/1ssei4rbTxoVQnjjVqI8FuOmMuPvvkqSt#scrollTo=vV_nE8lNXgji
2) https://www.hindawi.com/journals/mpe/2021/1162553/
3) https://courses.lumenlearning.com/diseaseprevention/chapter/dete rminants-of-health-risk-factors-and-prevention/
4) https://www.physio-pedia.com/Determinants_of_Health
5) https://www.physio-pedia.com/Determinants_of_Health