Machine Learning Modelling for predicting the charges of health insurance of U.S.A citizen

Author:

Muhammad Azis Rizaldi

01/20/2024

Outline

Introduction

Metodology

Explanatory Data Analysist

Machine Learning Model

Results and Discussion

INTRODUCTION

Question

ACME Insurance Inc. offers affordable health insurance to thousands of customer all over the United States. As the lead data scientist at ACME, **you're tasked with creating an automated system to estimate the annual medical expenditure for new customers**, using information such as their age, sex, BMI, children, smoking habits and region of residence, and you must be able to explain why your system outputs a certain prediction.

Description of Data

```
1 df_medical = pd.read_csv("Health_insurance.csv")
2 df_medical
```

	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
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

The data set contains 7 attributes such as age, sex, bmi, number of children, smoker (yes/no), region of domicile, and charge of health insurance for customers. In this data set, there are 1337 customers as U.S.A citizen.

METHODOLOGY

- **Data Wrangling**

  Data that we have may contains some outliers, missing/misleading value, inappropriate format. So, in the preliminary stage, we should overcome those challanges.

  There are many ways to overcome those problems:

  1. Check the data type and edit in appropriate format

  2. Check the descriptive statistics, and

  3. Consider fill or edit the missing/misleading values into zeros or mean, or just remove them.

- **Exploratory Data Analysist (EDA)**

  The data may contains thousands rows and variables. The most important thing is gaining the insights on the data. We can do exploratory data analysist and gain the correlation among variables by SQL or graph. This step is important to consider which variables must be entered in machine learning modelling

- **Machine Learning Modelling**

  Since we want to predict the annual medical charges, we choose the linear regression model. Before we start the modelling, we should determine which variables should exist in the model, and we make sure the data have been scaled or translated into numeric data type either by binary or one-hot encoding method.

  We split the data into data training and data testing for seek the accuracy of model. We can inspect the quality of model by see the r2 score and its rmse.

DATA WRANGLING

Data Wrangling

The first step to processing the data set is data cleansing or data wrangling.

```
1 df_medical.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
```

```
1 df_medical.describe()
```

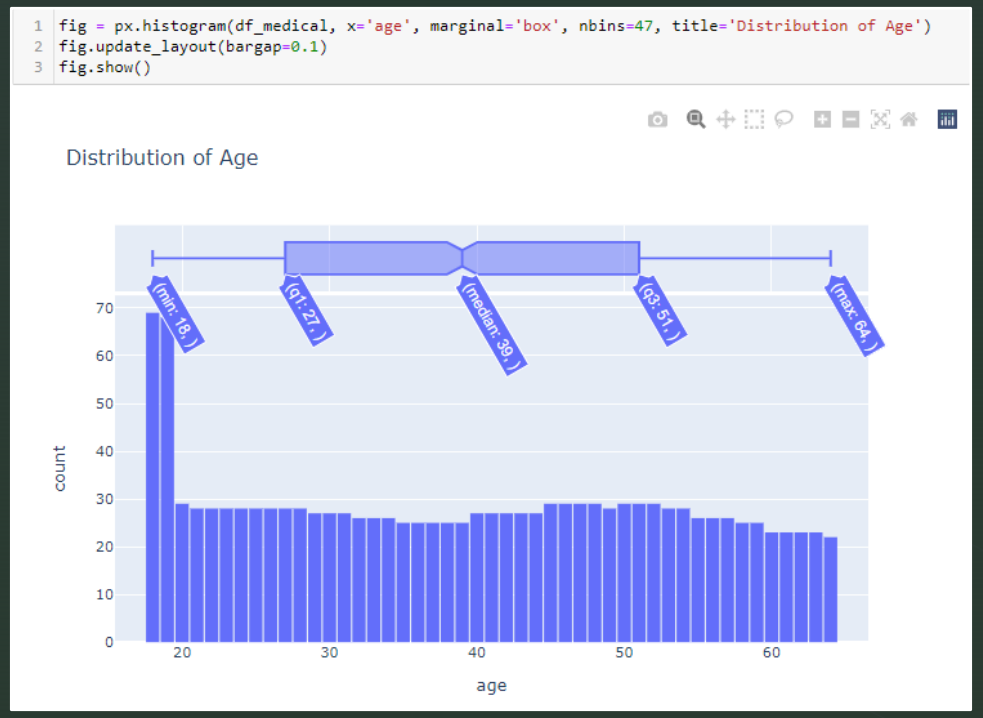
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000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

We see the data have been in appropriate format, and there are no missing values (1338 of 1338 rows). So, we are free for doing the data cleansing.

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis

```
fig = px.histogram(df_medical, x='age', marginal='box', nbins=47, title='Distribution of Age')
fig.update_layout(bargap=0.1)
fig.show()
```

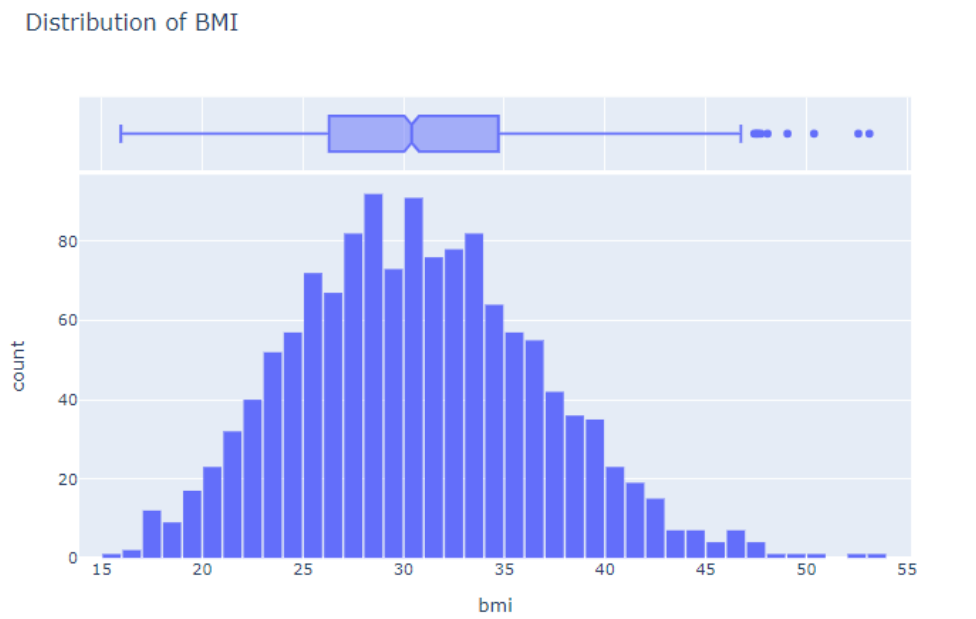


The distribution of age is normal distribution.
The number of customers are almost same for every age, except 18 and 19 years old as much twice as the other age.

The reason for this phenomena is the 18-19 years old customers are less prone to getting sick rather than the older one and thus company has to pay them less for their medical bills.

Exploratory Data Analysis

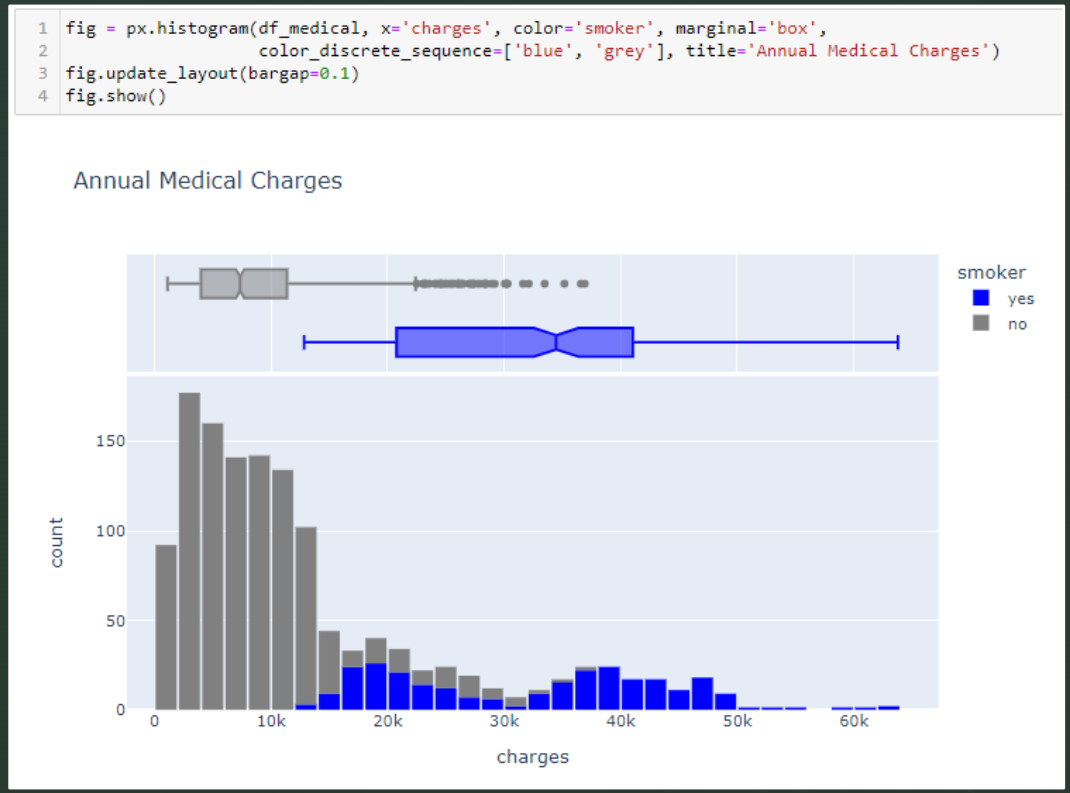
```
fig = px.histogram(df_medical, x='bmi', marginal='box', title='Distribution of BMI')
fig.update_layout(bargap=0.1)
fig.show()
```



The distribution of BMI is gaussian distribution. The bmi for healthy person is in range 20-30. If the people who bmi under or above, they are considered more prone to health issues. Hence, they will pay the health insurance at higher price because thus company will more pay the medical bill.

Exploratory Data Analysis

```
fig = px.histogram(df_medical, x='charges', color='smoker', marginal='box',
                   color_discrete_sequence=['blue', 'grey'], title='Annual Medical Charges')
fig.update_layout(bargap=0.1)
fig.show()
```

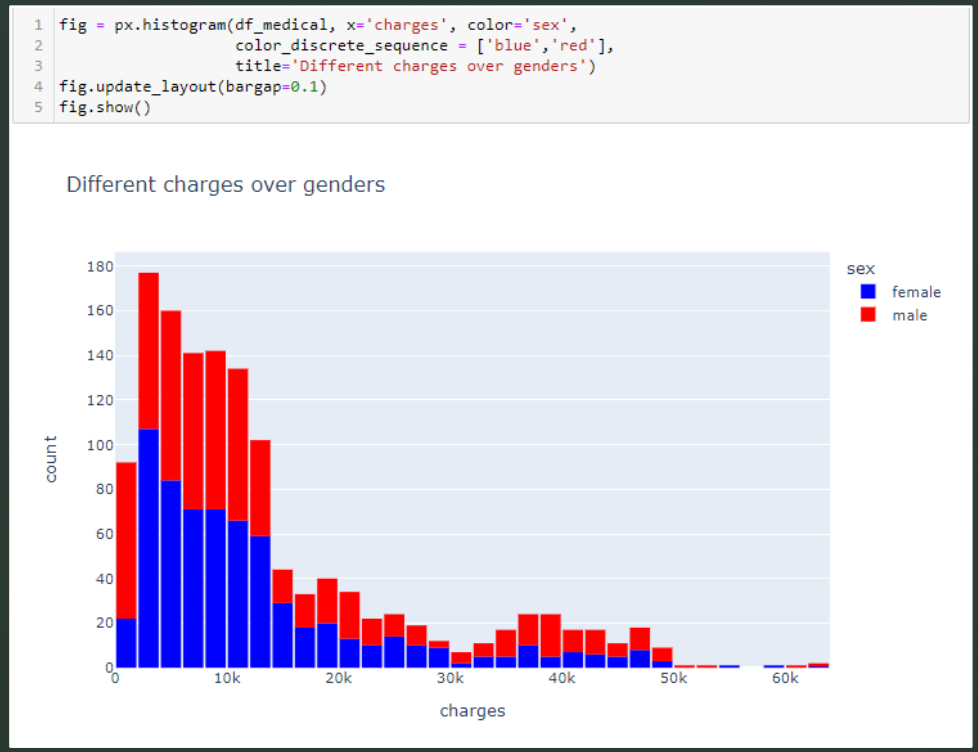


Most customers have annual medical charges under $14,000. Other customers have higher annual medical charges possibly due to accidents, major illness, and genetic diases.

There is a significant medical expenses between smokers and non-smokers. The smokers have the higher annual charges rather than the non-smokers.

Exploratory Data Analysis

```
fig = px.histogram(df_medical, x='charges', color='sex',
                   color_discrete_sequence = ['blue','red'],
                   title='Different charges over genders')
fig.update_layout(bargap=0.1)
fig.show()
```

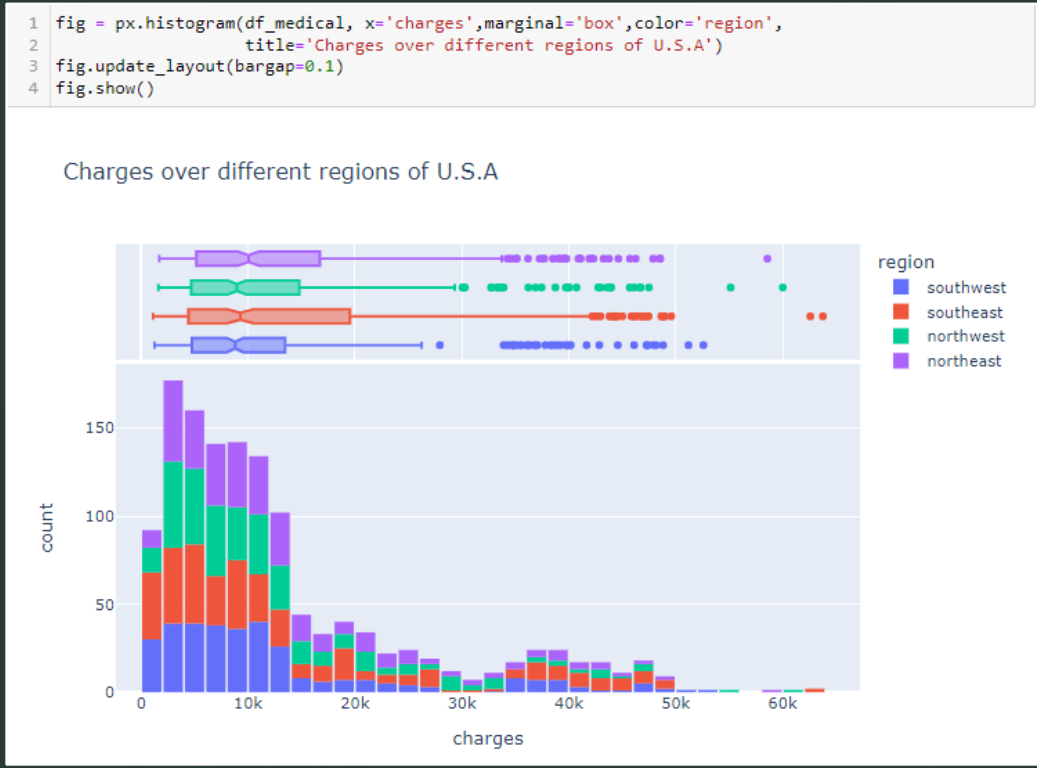


The number of male and female are equally same for wide range of annual medical charges, although men are dominant.

Males are substancially charged more because they are more likely to take risks and that keeps them in danger. Females are substancially also to be more prone to getting cancer rather than males.

Exploratory Data Analysis

```
fig = px.histogram(df_medical, x='charges',marginal='box',color='region',
                   title='Charges over different regions of U.S.A')
fig.update_layout(bargap=0.1)
fig.show()
```



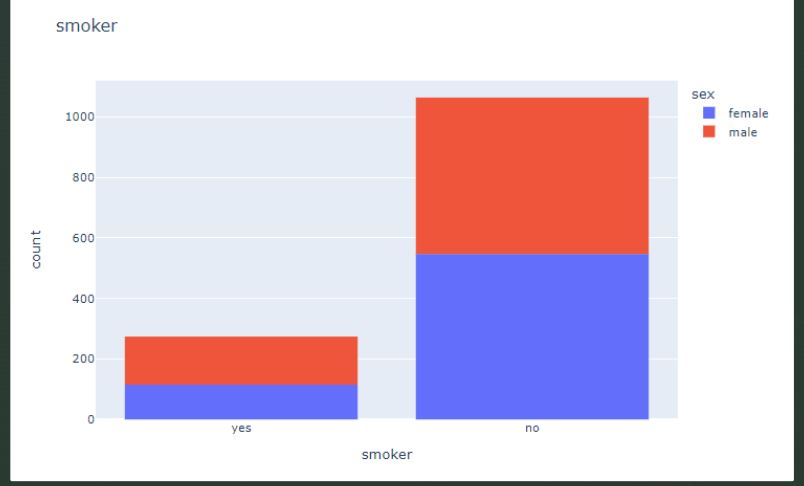
This distribution we see that southeaster part of U.S is leading in charges but majority of all customers from all parts of US are charged between 0-28k only.

Exploratory Data Analysis

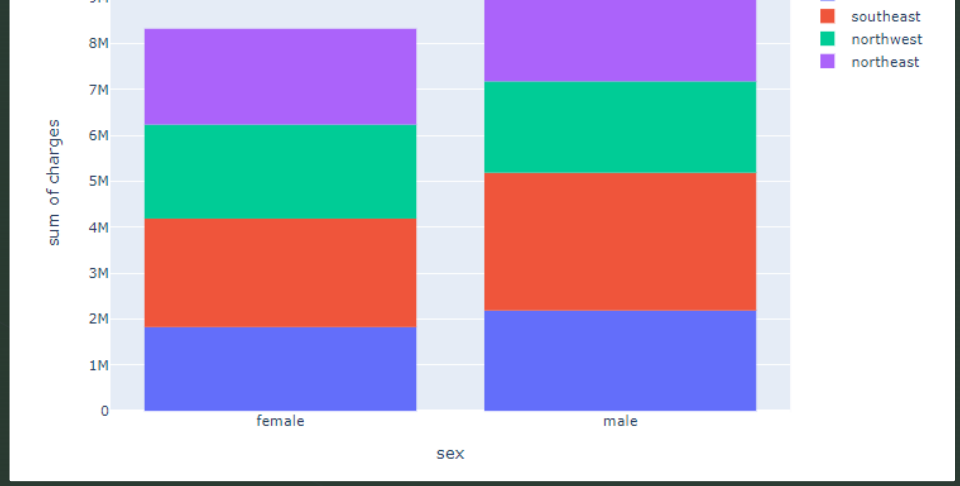
```
df_medical.smoker.value_counts()
```

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

```
px.histogram(df_medical, x='smoker', color='sex', title='smoker')
```




```
px.histogram(df_medical,x='sex',y = 'charges',color = 'region')
```

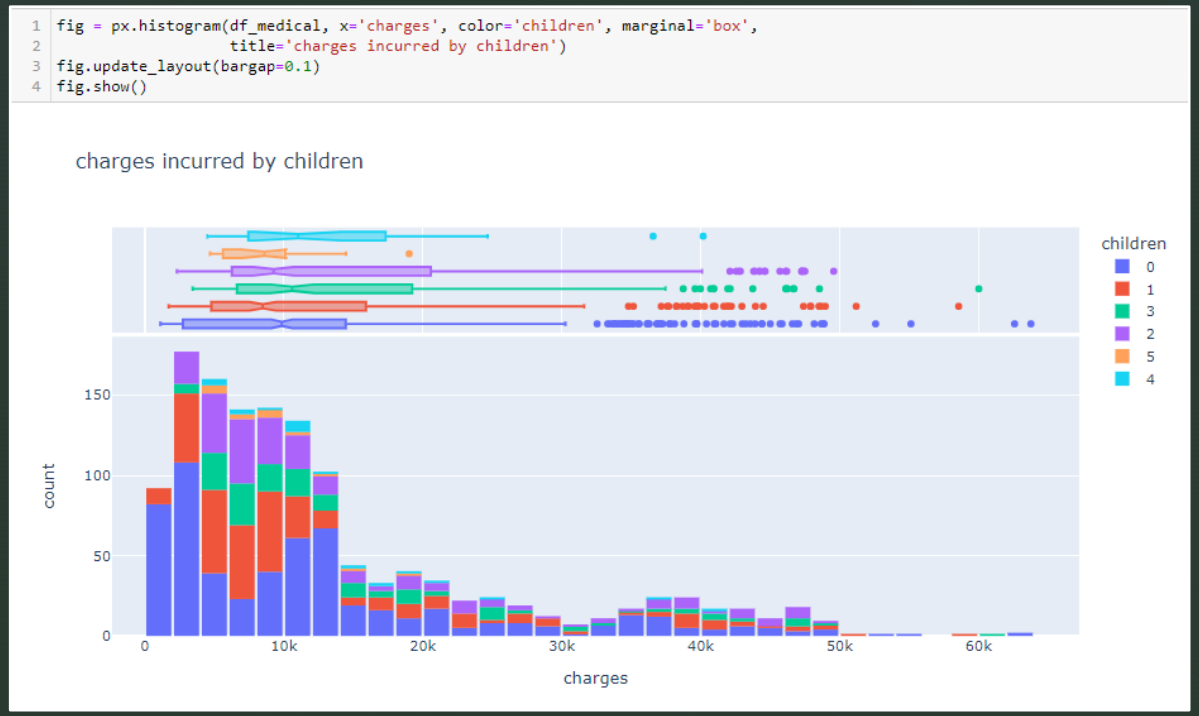


It appears that 20% of customers have reported that they smoke. We can also see that smoking appears a more common habit among males.

We see that in most regions except for northwest, the males are dominant smokers.

Exploratory Data Analysis

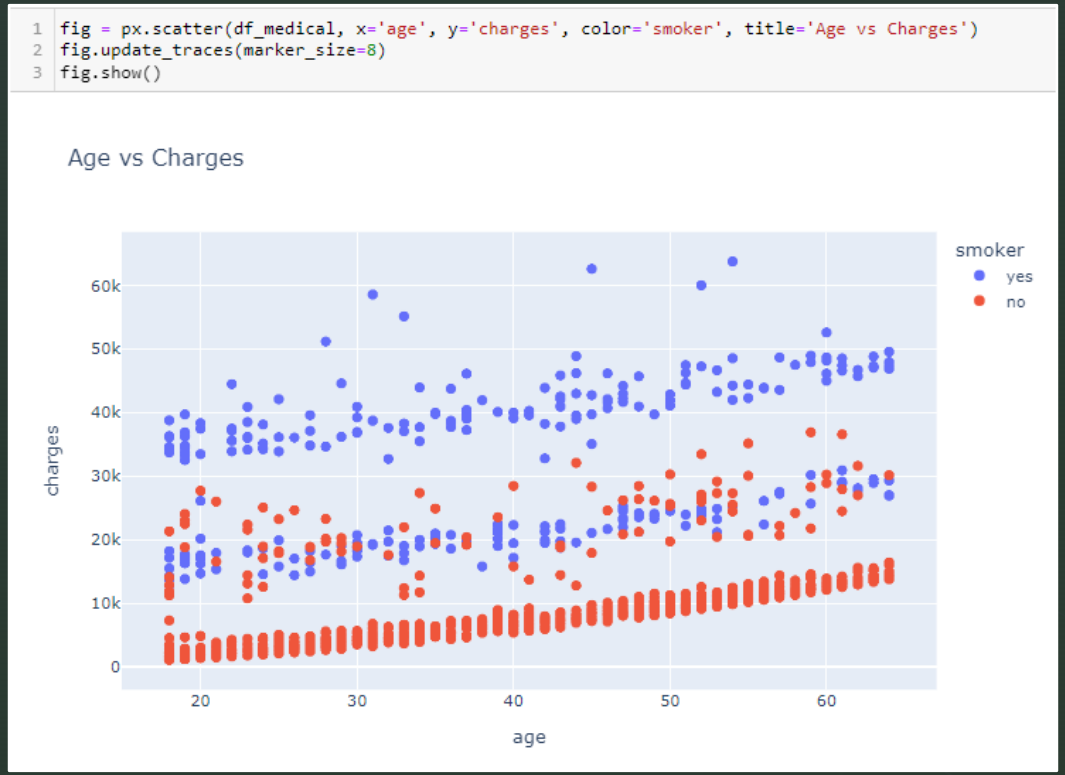
```
fig = px.histogram(df_medical, x='charges', color='children', marginal='box',
                   title='charges incurred by children')
fig.update_layout(bargap=0.1)
fig.show()
```



It seems that majority of our customers have 0 or 1 child and median charges vary between 8.5k to 10.6k dollars

Exploratory Data Analysis

```
1 fig = px.scatter(df_medical, x='age', y='charges', color='smoker', title='Age vs Charges')
2 fig.update_traces(marker_size=8)
3 fig.show()
```



Increasing age absolutely increase the charges. There are three trend cluster

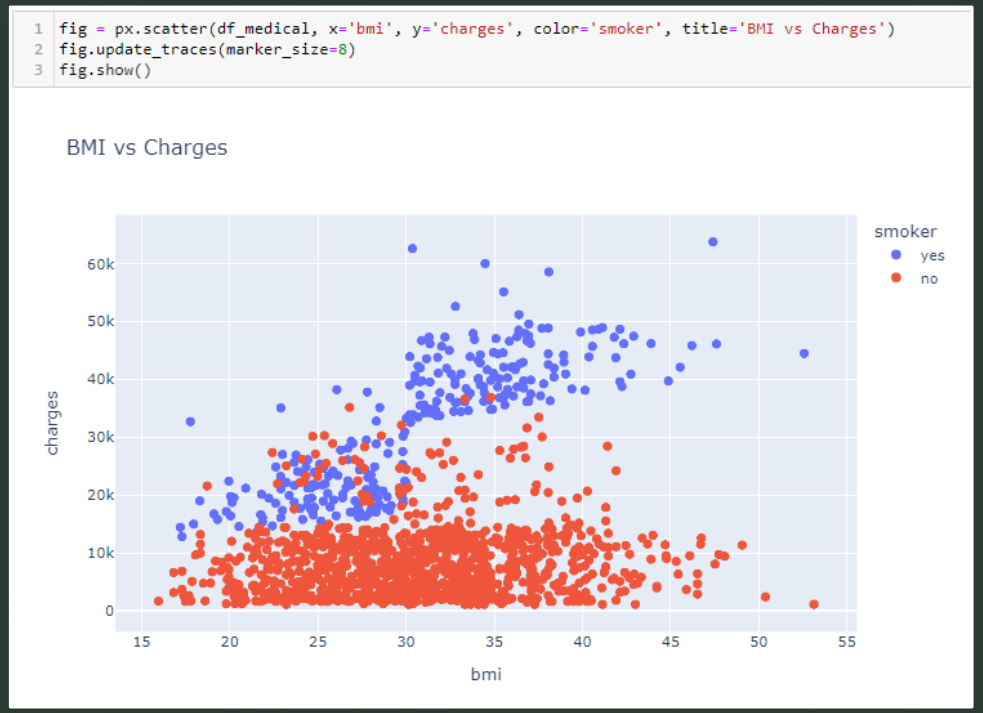
(1) The healthy non-smokers, relatively low medical charges

(2) The non-smokers with medical issues and smokers without major issues, intermediate medical charges

(3) The smokers with major medical issues, relatively high medical charges

Exploratory Data Analysis

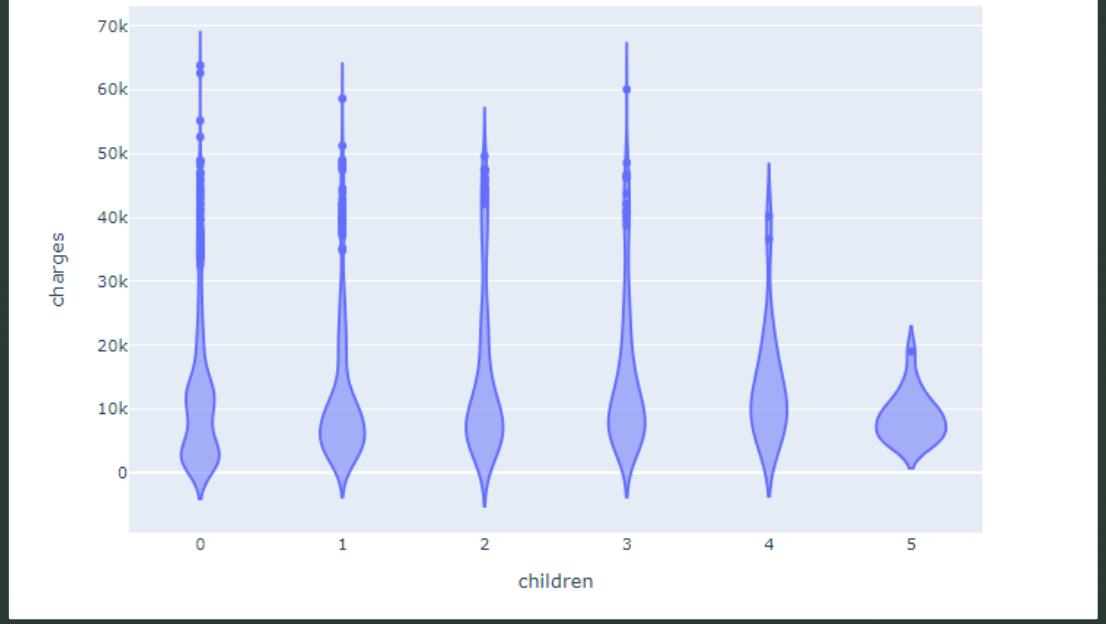
```
fig = px.scatter(df_medical, x='bmi', y='charges', color='smoker', title='BMI vs Charges')
fig.update_traces(marker_size=8)
fig.show()
```



It appears that for non-smokers, an increase in BMI doesn't seem to be related to an increase in medical charges. However, medical charges seem to be significantly higher for smokers with a BMI greater than 30.

Exploratory Data Analysis

```
1 px.violin(df_medical,x = 'children',y = 'charges')
```



There doesn't seem to have a strong trend among this variables. The majority of customers having any number of children or no children altogether have charges in range of 0-20k dollars

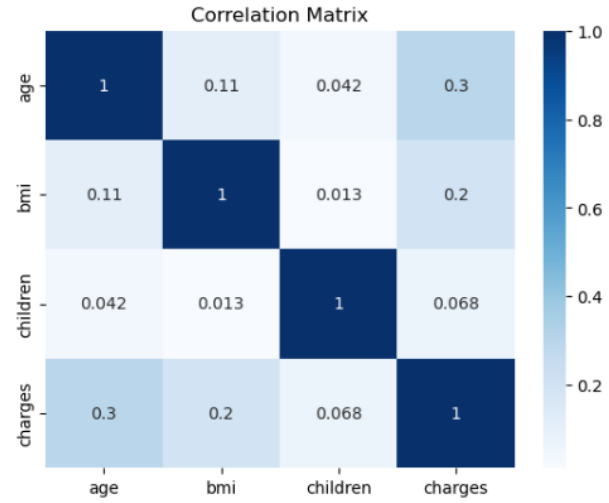
Correlation

```
1 sns.heatmap(df_medical.corr(), cmap='Blues', annot=True)
2 plt.title('Correlation Matrix')
```

```
C:\Users\ACER\AppData\Local\Temp\ipykernel_15256\2752851812.py:1: FutureWarning:

The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will defau
lt to False. Select only valid columns or specify the value of numeric_only to silence this warning.
```

```
Text(0.5, 1.0, 'Correlation Matrix')
```




```
1 map_smoker = {'no': 0, 'yes': 1}
2 smoker_binary = df_medical.smoker.map(map_smoker)
3 df_medical.charges.corr(smoker_binary)
```

```
0.7872514304984785
```

Correlation helps us to determine the weight between two variables.

In this case, we will seek the variable which having high correlation relative to the ‘charges’ variable.

In numerical variables, there are age vs. charges, and bmi vs. charges which are the highest correlation.

In categorical variable, which is ‘smoker’ has the value of correlation 0.78.

MACHINE LEARNING MODELLING

Machine Learning Modelling

```
!pip install scikit-learn --quiet
```

```
from sklearn.linear_model import LinearRegression
```

```
def rmse(targets, predictions):
    return np.sqrt(np.mean(np.square(targets - predictions)))
```

Now, we can estimate the annual medical charges by making the linear regression model. Before we do that, the variables of data should be conditioning. If the variables are numerical we should and scale the variables, whereas if the variables are categorical we should translate the variables into numeric by binary or by one-hot encoding.

RMSE (root mean square error) is the parameter which counts the difference between predicted values by model and actual values. r2 (r-square) measures how the model can explain the variability in independent variables. So, the good model is describes as having higher R2 and lower rmse.

Machine Learning Modelling

```
df_medical['smoker_code'] = smoker_binary
df_medical
```

	age	sex	bmi	children	smoker	region	charges	smoker_code
0	19	female	27.900	0	yes	southwest	16884.92400	1
1	18	male	33.770	1	no	southeast	1725.55230	0
2	28	male	33.000	3	no	southeast	4449.46200	0
3	33	male	22.705	0	no	northwest	21984.47061	0
4	32	male	28.880	0	no	northwest	3866.85520	0
...	...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830	0
1334	18	female	31.920	0	no	northeast	2205.98080	0
1335	18	female	36.850	0	no	southeast	1629.83350	0
1336	21	female	25.800	0	no	southwest	2007.94500	0
1337	61	female	29.070	0	yes	northwest	29141.36030	1

1338 rows × 8 columns

```
sex_binary = {'female': 0, 'male': 1}
df_medical['sex_code'] = df_medical.sex.map(sex_binary)
df_medical
```

	age	sex	bmi	children	smoker	region	charges	smoker_code	sex_code
0	19	female	27.900	0	yes	southwest	16884.92400	1	0
1	18	male	33.770	1	no	southeast	1725.55230	0	1
2	28	male	33.000	3	no	southeast	4449.46200	0	1
3	33	male	22.705	0	no	northwest	21984.47061	0	1
4	32	male	28.880	0	no	northwest	3866.85520	0	1
...	...	...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830	0	1
1334	18	female	31.920	0	no	northeast	2205.98080	0	0
1335	18	female	36.850	0	no	southeast	1629.83350	0	0
1336	21	female	25.800	0	no	southwest	2007.94500	0	0
1337	61	female	29.070	0	yes	northwest	29141.36030	1	0

1338 rows × 9 columns

We translate smokers into two, 1 for ‘yes’ and 0 for ‘no’ whereas sex translate into two, 1 for ‘male’ and 0 for ‘female’.

Machine Learning Modelling

```
In [30]:
from sklearn import preprocessing
enc = preprocessing.OneHotEncoder()
```

```
In [32]:
enc.fit(df_medical[['region']])
one_hot_enc = enc.transform(df_medical[['region']]).toarray()
df_medical[['northeast','northwest','southeast','southwest']] = one_hot_enc
df_medical
```

	age	sex	bmi	children	smoker	region	charges	smoker_code	sex_code	northeast	northwest	southeast	southwest
0	19	female	27.900	0	yes	southwest	16884.92400	1	0	0.0	0.0	0.0	1.0
1	18	male	33.770	1	no	southeast	1725.55230	0	1	0.0	0.0	1.0	0.0
2	28	male	33.000	3	no	southeast	4449.46200	0	1	0.0	0.0	1.0	0.0
3	33	male	22.705	0	no	northwest	21984.47061	0	1	0.0	1.0	0.0	0.0
4	32	male	28.880	0	no	northwest	3866.85520	0	1	0.0	1.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830	0	1	0.0	1.0	0.0	0.0
1334	18	female	31.920	0	no	northeast	2205.98080	0	0	1.0	0.0	0.0	0.0
1335	18	female	36.850	0	no	southeast	1629.83350	0	0	0.0	0.0	1.0	0.0
1336	21	female	25.800	0	no	southwest	2007.94500	0	0	0.0	0.0	0.0	1.0
1337	61	female	29.070	0	yes	northwest	29141.36030	1	0	0.0	1.0	0.0	0.0

Since the ‘region’ variable contains more than two values, we translate this variable into numeric by one-hot encoding.

Number 1 is defined as True, and number 0 is defined as False.

Machine Learning Modelling

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
```

```
numeric_cols = ['age', 'bmi', 'children']
scaler.fit(df_medical[numeric_cols])
scaled_num_cols = scaler.transform(df_medical[numeric_cols])
df_medical
```

	age	sex	bmi	children	smoker	region	charges	smoker_code	sex_code	northeast	northwest	southeast
0	19	female	27.900	0	yes	southwest	16884.92400	1	0	0.0	0.0	0.0
1	18	male	33.770	1	no	southeast	1725.55230	0	1	0.0	0.0	1.0
2	28	male	33.000	3	no	southeast	4449.46200	0	1	0.0	0.0	1.0
3	33	male	22.705	0	no	northwest	21984.47061	0	1	0.0	1.0	0.0
4	32	male	28.880	0	no	northwest	3866.85520	0	1	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830	0	1	0.0	1.0	0.0
1334	18	female	31.920	0	no	northeast	2205.98080	0	0	1.0	0.0	0.0
1335	18	female	36.850	0	no	southeast	1629.83350	0	0	0.0	0.0	1.0
1336	21	female	25.800	0	no	southwest	2007.94500	0	0	0.0	0.0	0.0
1337	61	female	29.070	0	yes	northwest	29141.36030	1	0	0.0	1.0	0.0

1338 rows × 13 columns

```
cat_cols = ['smoker_code', 'sex_code', 'northeast', 'northwest', 'southeast', 'southwest']
cat_data = df_medical[cat_cols].values
```

```
X = np.concatenate((scaled_num_cols, cat_data), axis=1)
Y = df_medical.charges
```

After we translate the categorical variables into numeric, we can scale the numerical variables due to the wide range of value.

Now, we can concate the numerical and categorical variables (include age, bmi, children, smoker_code, sex_code, northeast, northwest, southeast, southwest) as inputs and charges as output

Machine Learning Modelling

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.1, random_state=3)

# Create and train the model
model = LinearRegression().fit(x_train, y_train)
r2 = model.score(x_train, y_train)
print('model score: ', r2)

# Generate predictions
pred_test = model.predict(x_test)

# Compute loss to evalute the model
loss_test = rmse(y_test, pred_test)
print('Loss of testing data: ', loss_test)
```

```
model score:  0.7505721678490005
Loss of testing data:  6287.332697217845
```

Before we build the model, we split the data into training data (90%) and testing data (10%).

We do linear regression modelling to predict the annual medical charge. Unfortunately as we see, the model have score 0.75 and rmse of charges $6287 which is bad model.

Result and Discussion

After all procedure have been conducted, we found that:

1. The higher annual medical charge, fewer the customers.
2. The annual medical charge is getting higher along with the high health costs that need to be incurred by the company.
3. The high annual medical charge incurred by the Company are due to customers who smoke, excess or deficit of BMI (less than 22 or more than 30), and increasing age.
4. Customers who smoke are predominantly male.

There are strong correlation between charges and age, charges and bmi, charges and smoker. Machine learning modelling including all variables relative to the charges variable, has accuracy 0.75 and rmse $6287 which is low. So, other machine learning model may result the better accuracy for prediction, classification, or clustering.