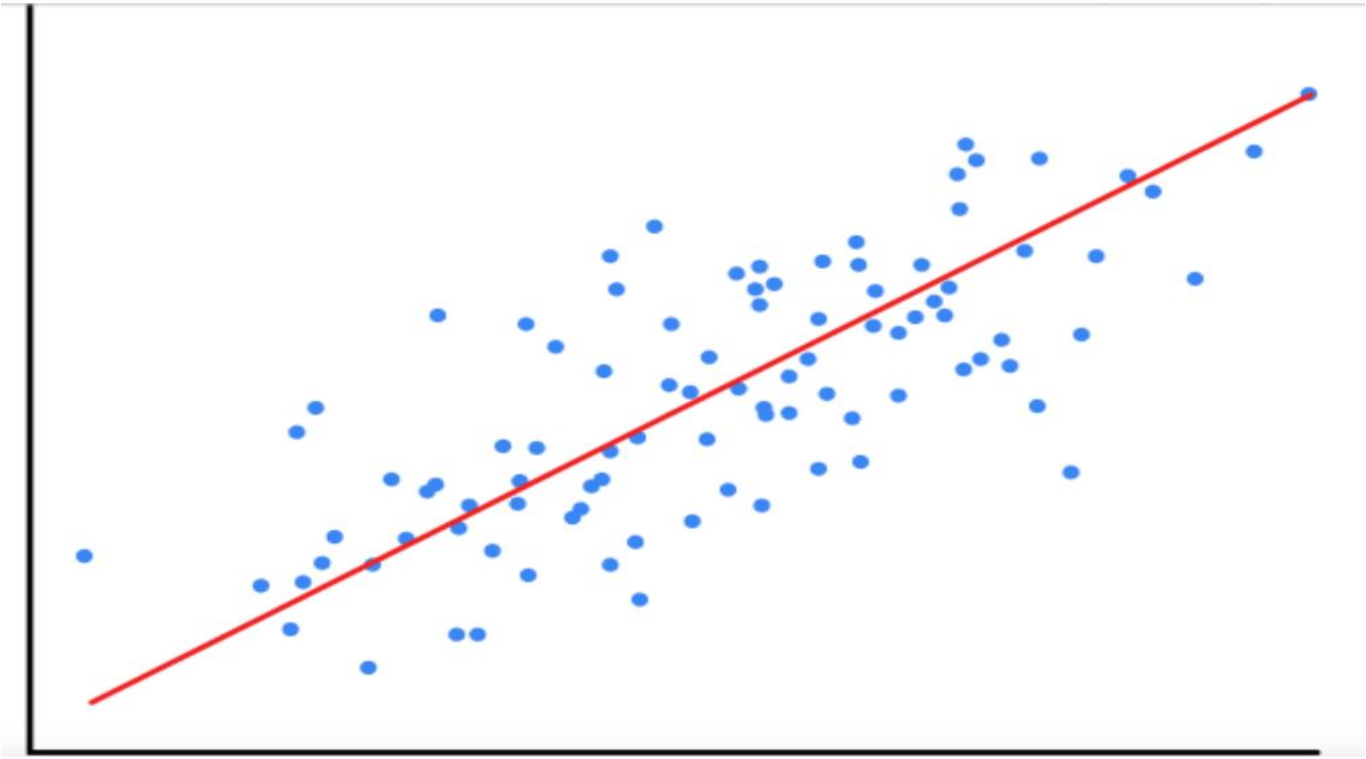
Insurance Medical Cost Analysis using Linear Regression Model

Team Members:

1. Mamatha Buddala
2. Aishwarya Jadhav
3. Lokesh Bogam
4. Karanjot Singh



Project Goal

Predicting the charges billed by the health insurance company based on current dataset

Problem Statement :

- Which attributes/factors affect the cost billed by the Insurance company?
- Can we accurately predict the insurance costs based on multiple variables?

Project Scope:

- Exploratory Data Analysis to determine the attributes that affect the cost.
- Multivariate Linear Regression to predict the cost based on multiple attributes of the dataset.

Benefits:

- Understanding the factors that affect the insurance cost
- Calculating and estimating the approx. cost based on the current dataset

Process

- Pre-processing
- EDA
- Prepare for ML training
- Training the Model
- Prediction

Attributes:

- Age, Gender, BMI, No. of Children, Smoker, Region, Charges

Population:

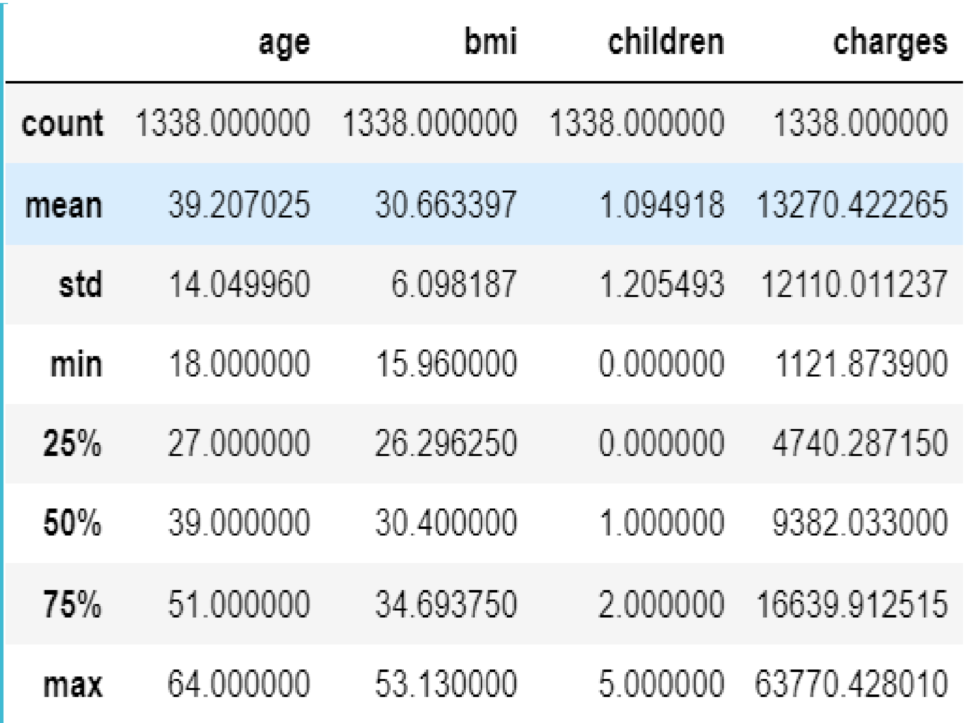
- Dataset is of age group between 18-64 years of an Insurance Firm of USA.

Future Scope:

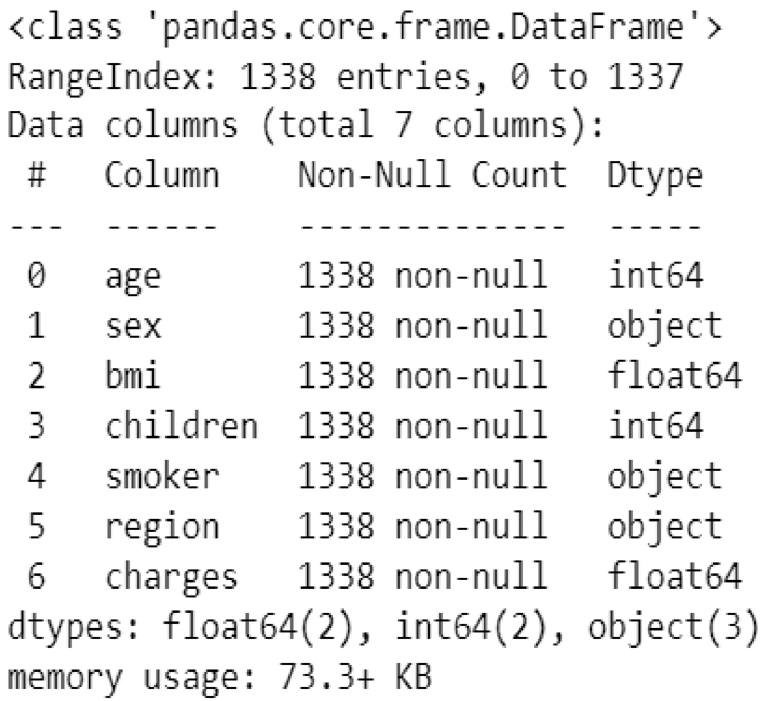
- Calculate and compare the score of different data modelling techniques against the dataset.
- Using Dimensionality reduction algorithms to preserves the salient relationships in the data

Pre-processing

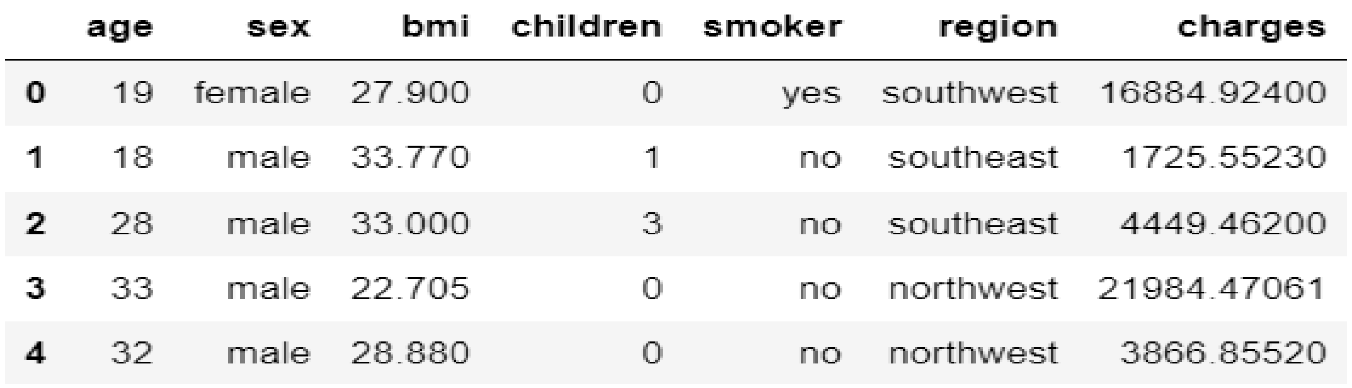
- Null Values
- Outliers
- Nan Values
- Count
- Datatypes
- Statistics



	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



```
<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
```



	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

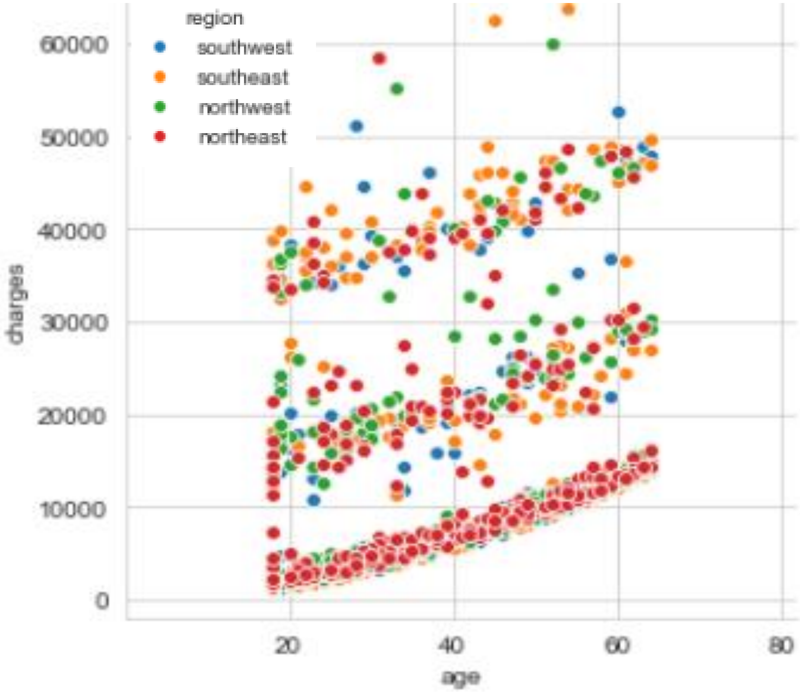
Exploratory Data Analysis

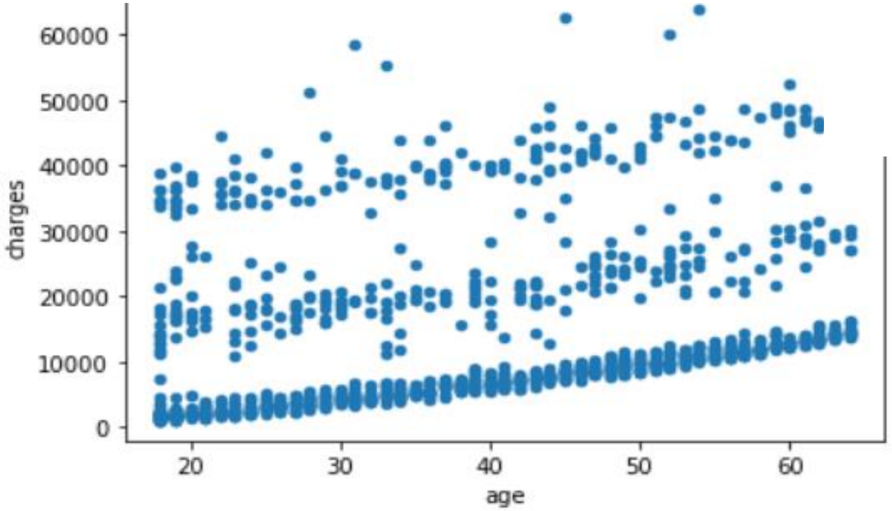
Charges against Age:

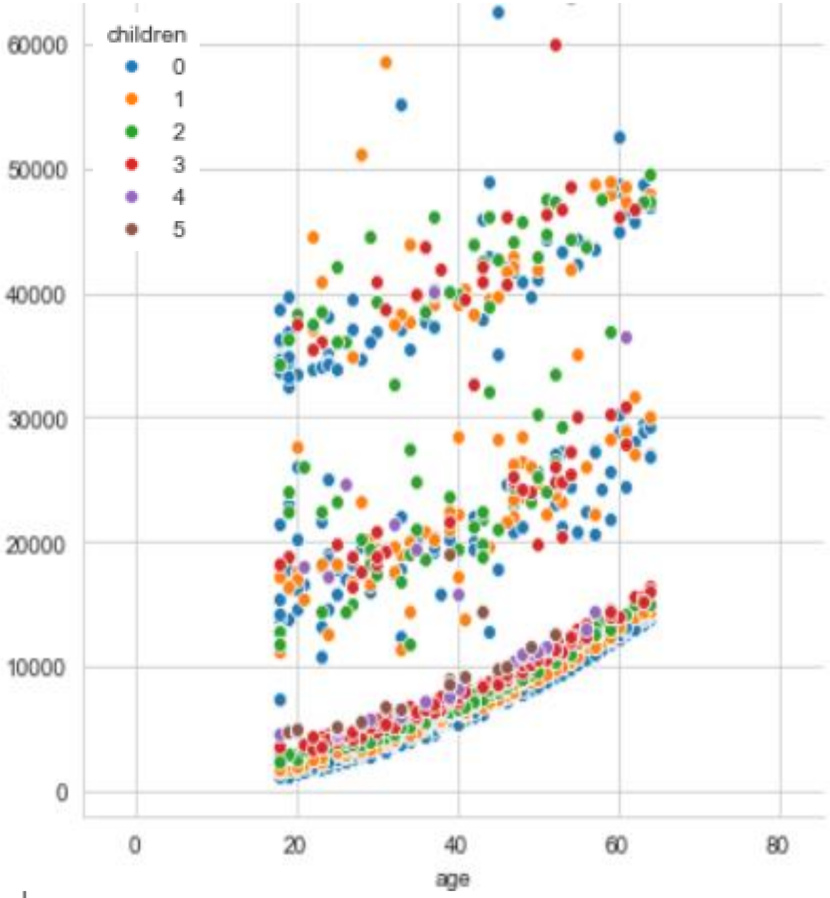
- According to region
- According to children

Charges increase with age
Region doesn't affect charges

No. of children don't affect charges





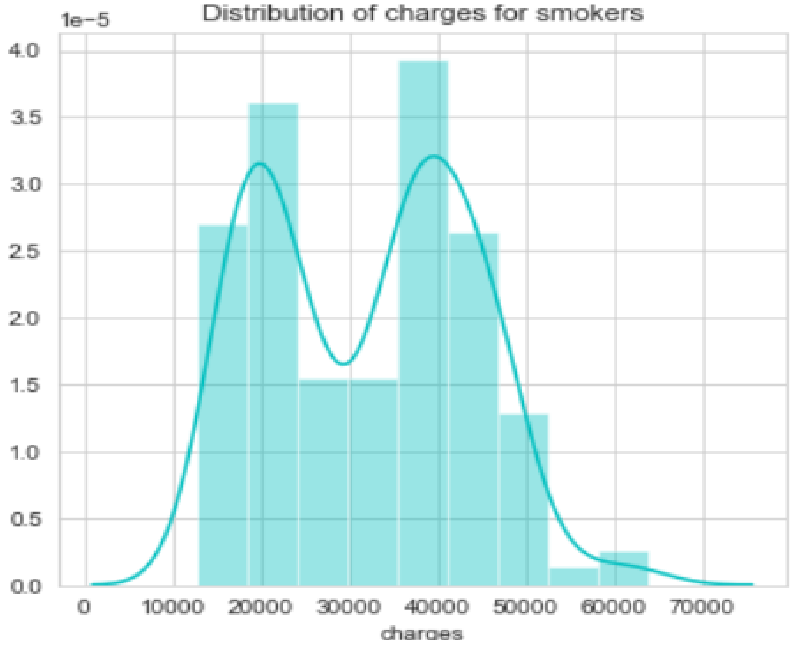


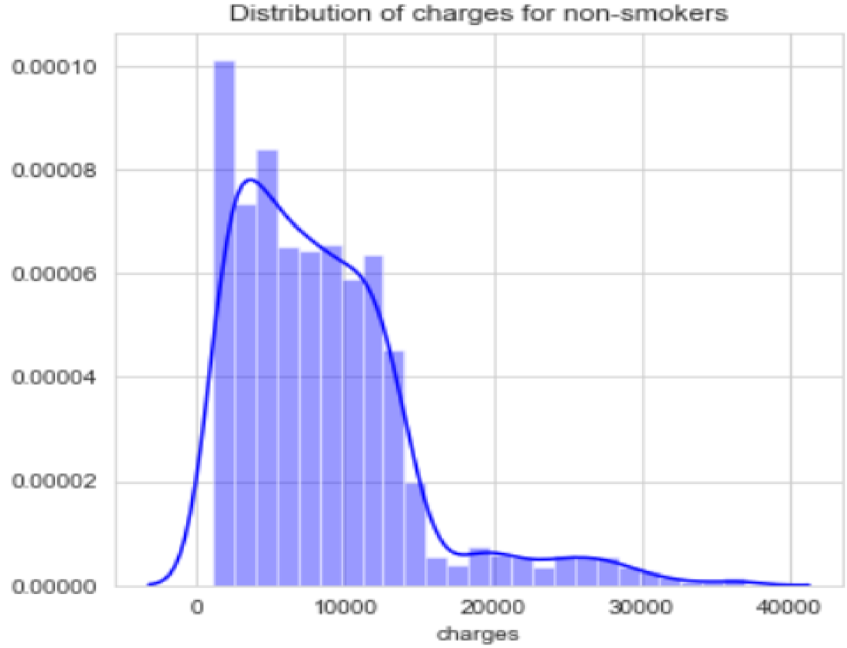
Charges against Smoker:

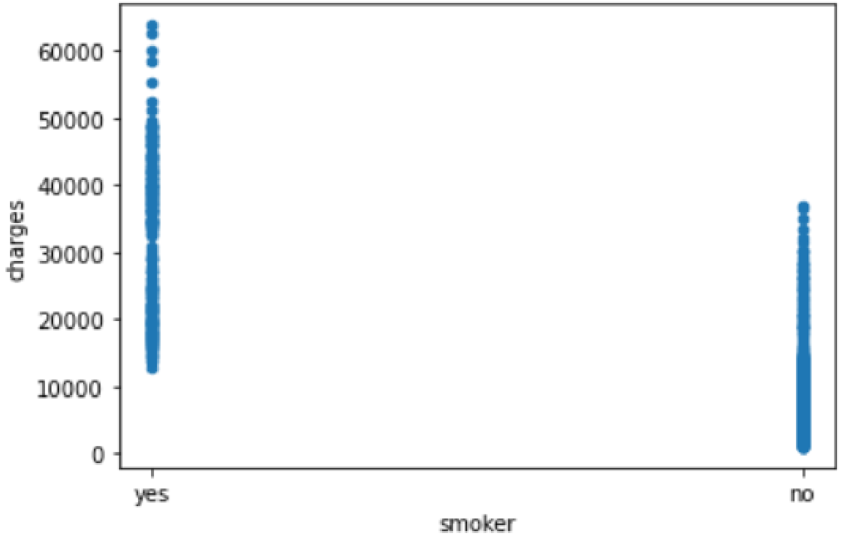
Minimum charges for smoker is more then non-smokers

Smoker population distribution is high and across the charges range

Non-smoker population distribution of charges is on the lower end of the graph.


Distribution of charges for smokers


Distribution of charges for non-smokers

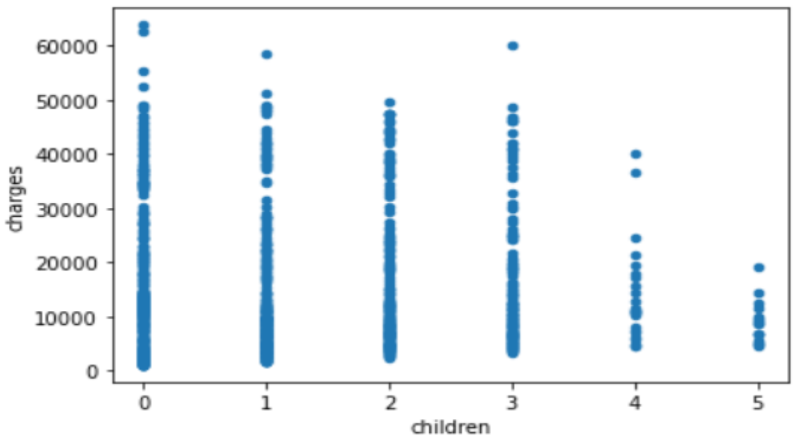


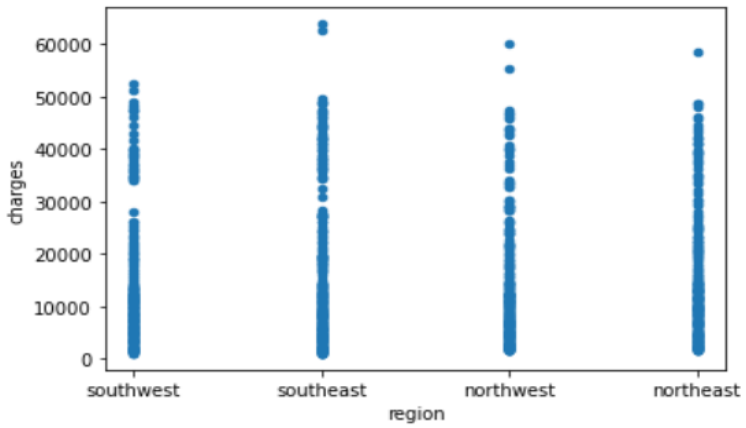
Charges against Region:

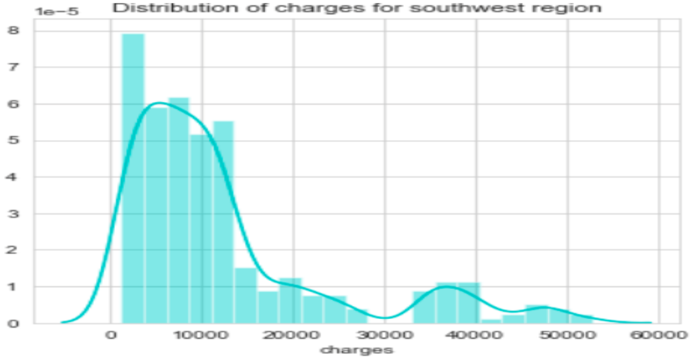
Region is not affecting the charges

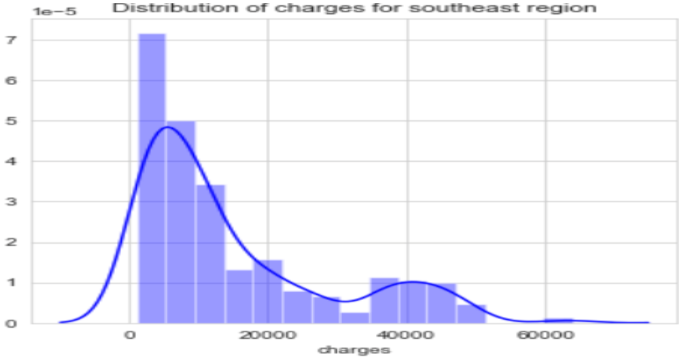
Charges against Children:

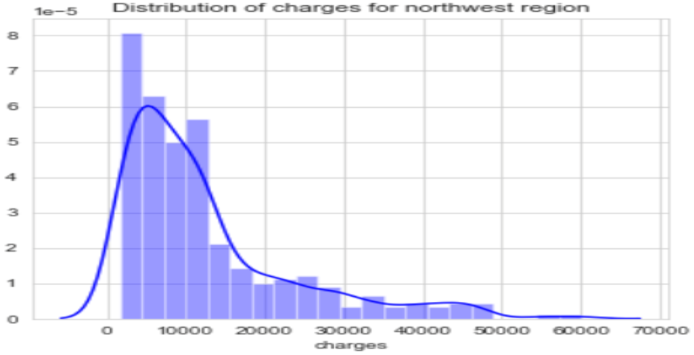
No. of children is not affecting the charges

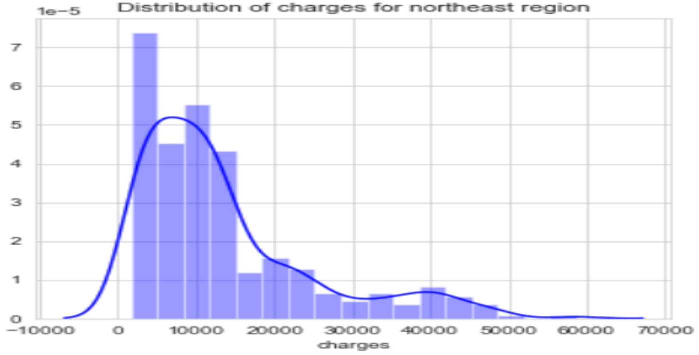












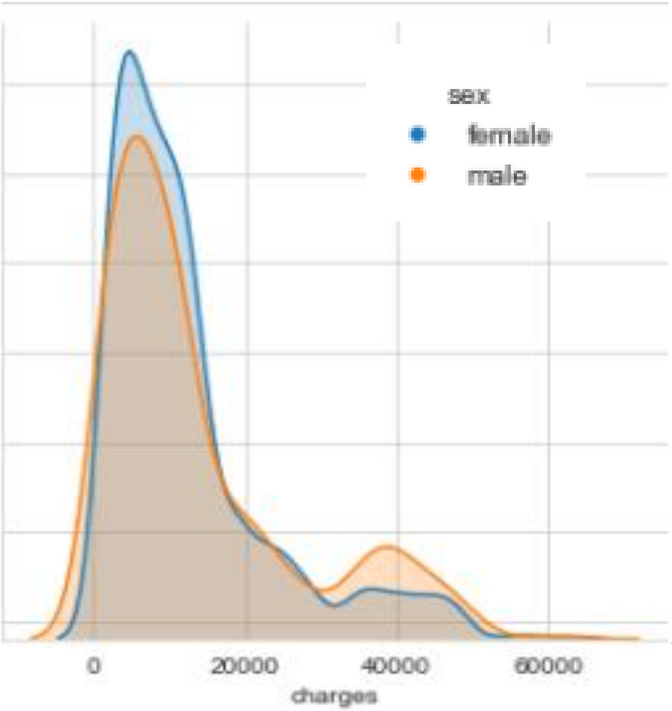
Charges against Gender:

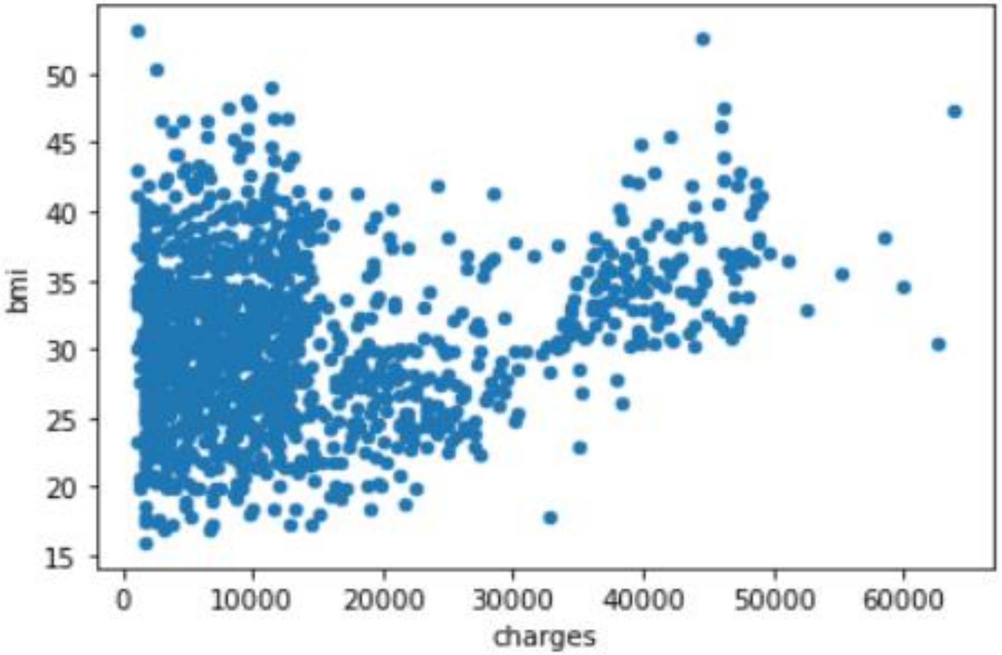
Gender data is distributed across charges

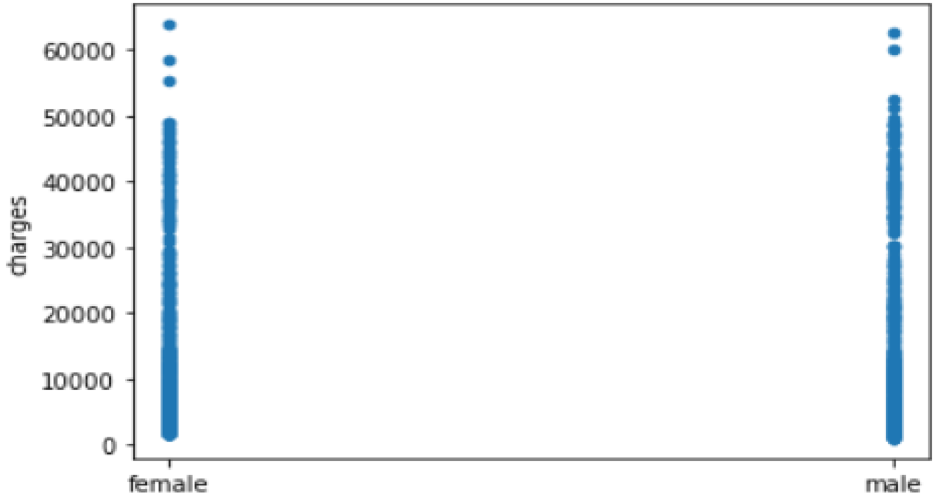
Charges against BMI:

BMI data is distributed across charges

Data is overlapping thus we will consider BMI and Gender for modelling.







Prepare for Modelling

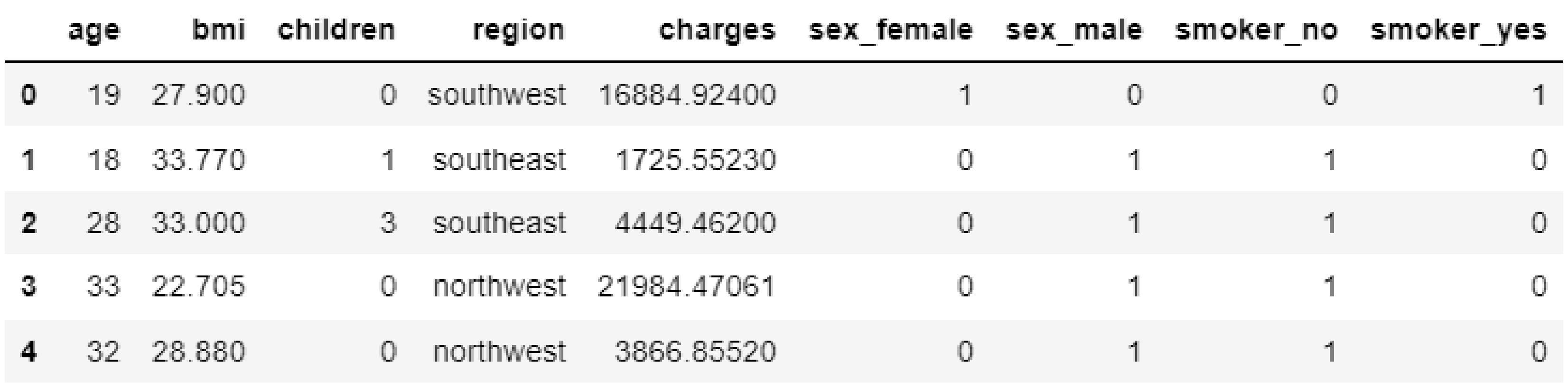
Independent Variables: Age , Smoker, Gender, BMI
Dependent Variable: Charges

Split columns to get numeric value for gender and smoker

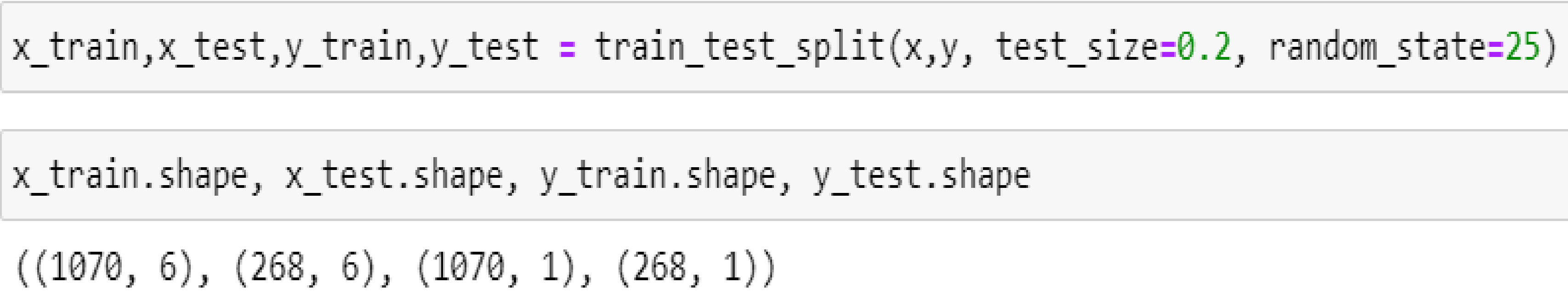
Split the data into training and testing datasets for X Independent and Y Dependent

Test set accounting for 20% of the total dataset and the training set accounting for 80%.

After Splitting dataset columns into numeric data



	age	bmi	children	region	charges	sex_female	sex_male	smoker_no	smoker_yes
0	19	27.900	0	southwest	16884.92400	1	0	0	1
1	18	33.770	1	southeast	1725.55230	0	1	1	0
2	28	33.000	3	southeast	4449.46200	0	1	1	0
3	33	22.705	0	northwest	21984.47061	0	1	1	0
4	32	28.880	0	northwest	3866.85520	0	1	1	0



```
x_train,x_test,y_train,y_test = train_test_split(x,y, test_size=0.2, random_state=25)
```

```
x_train.shape, x_test.shape, y_train.shape, y_test.shape
```

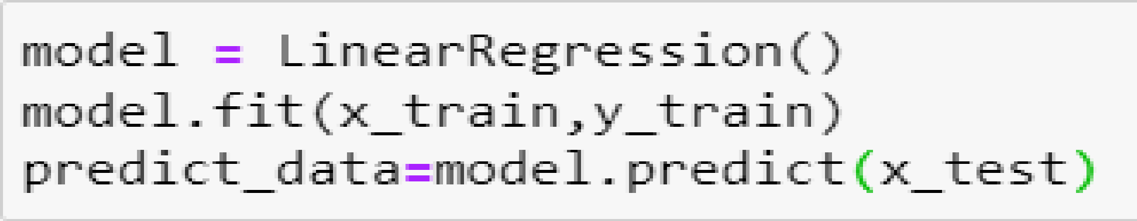
```
((1070, 6), (268, 6), (1070, 1), (268, 1))
```

Training Model

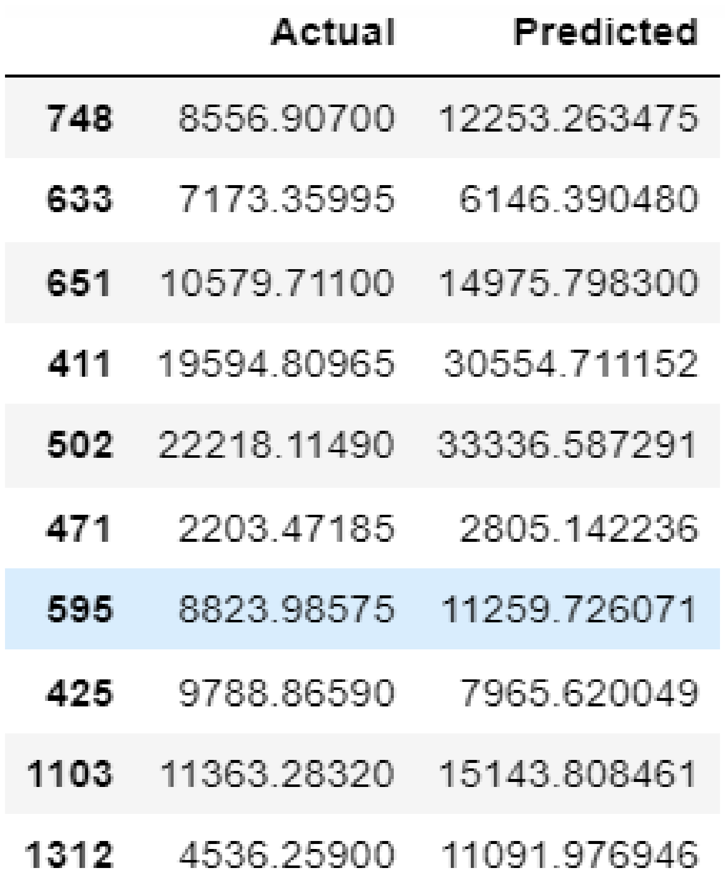
Create Linear Regression model

Fit the training dataset to the model

Predict the test dataset using the model



```
model = LinearRegression()
model.fit(x_train,y_train)
predict_data=model.predict(x_test)
```

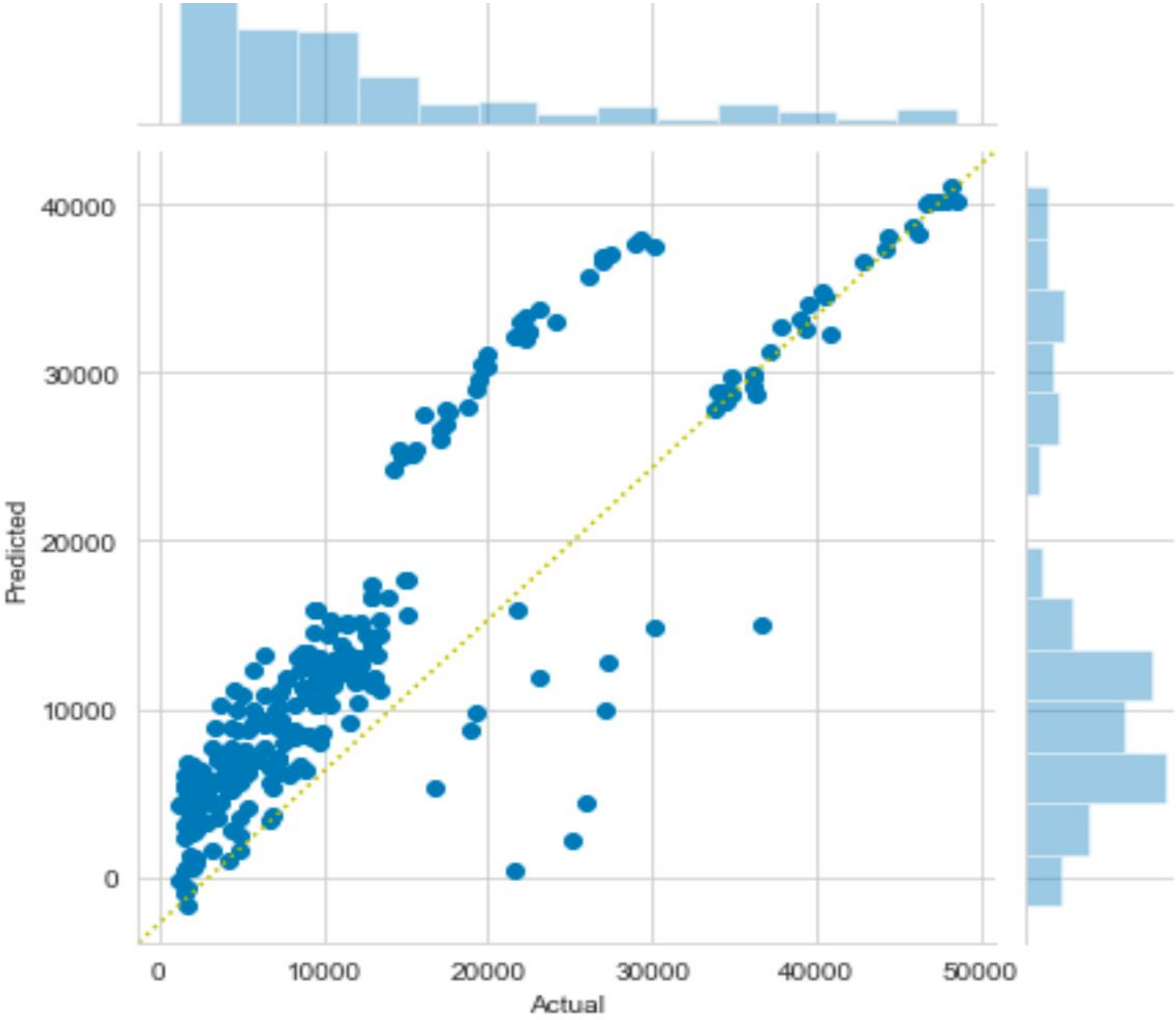


	Actual	Predicted
748	8556.90700	12253.263475
633	7173.35995	6146.390480
651	10579.71100	14975.798300
411	19594.80965	30554.711152
502	22218.11490	33336.587291
471	2203.47185	2805.142236
595	8823.98575	11259.726071
425	9788.86590	7965.620049
1103	11363.28320	15143.808461
1312	4536.25900	11091.976946

Score

R2 Score: 0.76175

Model Accuracy : 76.175%



Conclusion:

- We were able to create Multivariate Linear Regression Model with 76% accuracy prediction rate for predicting the charges of Insurance.
- Age, Smoker, BMI and Gender were the important dimensions of the dataset for modelling.

Thank you !
Any Question?