IBM Machine Learning Professional Certificate

- Supervised Learning - (Regression)

Choe C.S

October 2021

Main Objective

Objective

- This report aims to analyze and predict Insurance expenses of customers using Supervised Machine Learning algorithm.

Data Set

- The data set used for this analysis is Insurance Premium Data from Kaggle.com.
- The data set includes Health Insurance Premium Charges based on important features such as Gender, BMI etc.

Steps Involved

1) Perform Data preprocessing to - Handle missing data, Data formatting, Data Binning etc.

2) EDA – To gain better understanding of each feature variables and its affects on the target data.

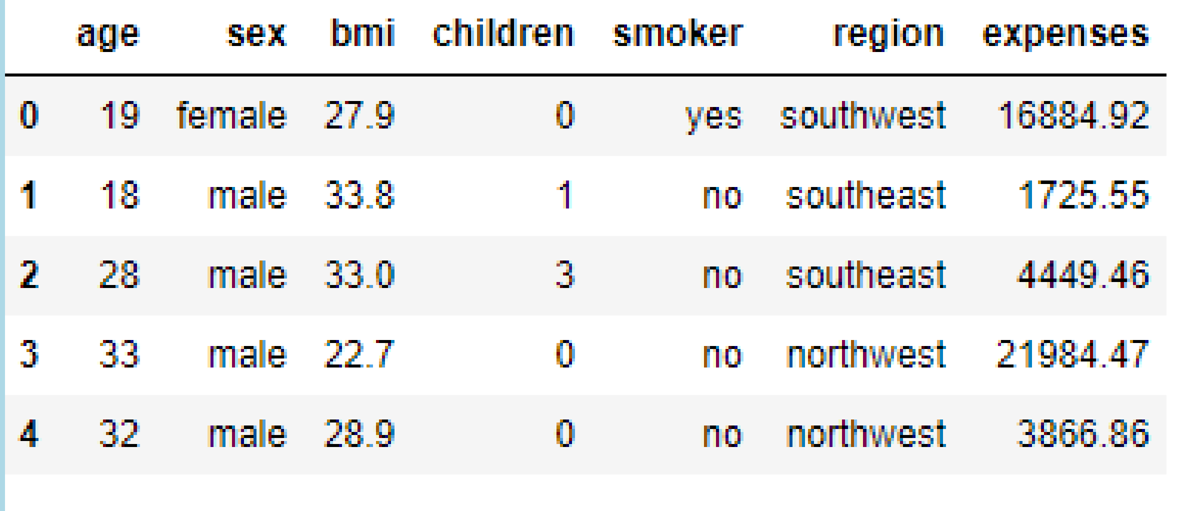
3) Perform Feature Engineering to transform categorical data into numeric format (One Hot Encoding).

4) Perform Machine Learning on the data sets using Linear regression.

5) Comparing results by including
   - Polynomial Feature transformation and various Regularization Methods
     such as (Ridge, Lasso and Elastic Net)

Import Data

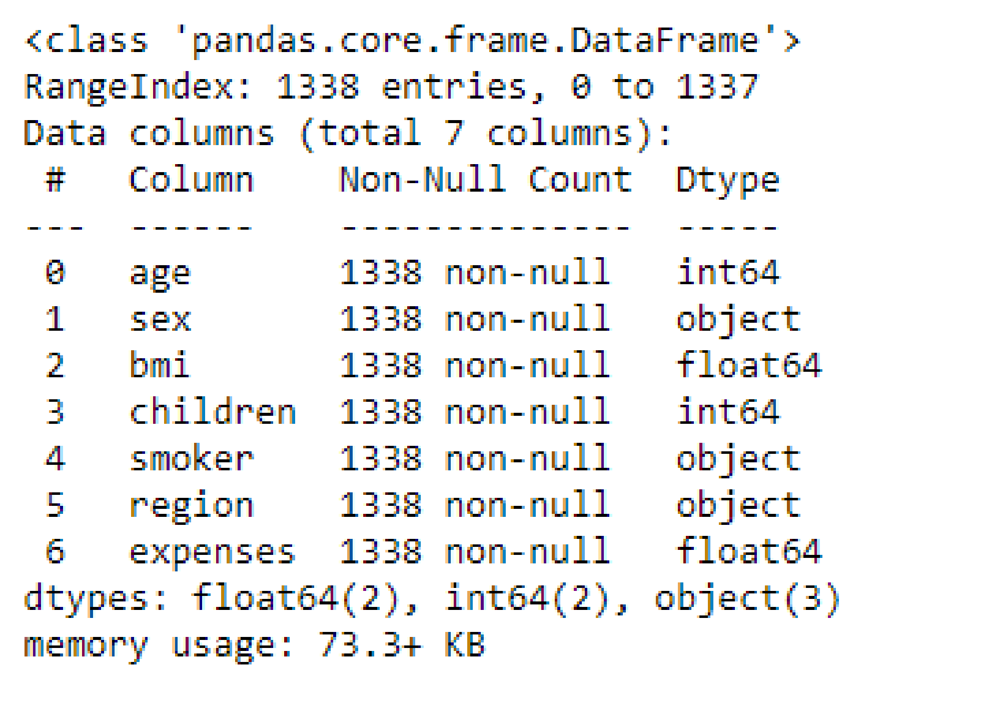
- The data set was import from Kaggle.com.
- The table on the right shows the top 5 rows for the data set



	age	sex	bmi	children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86

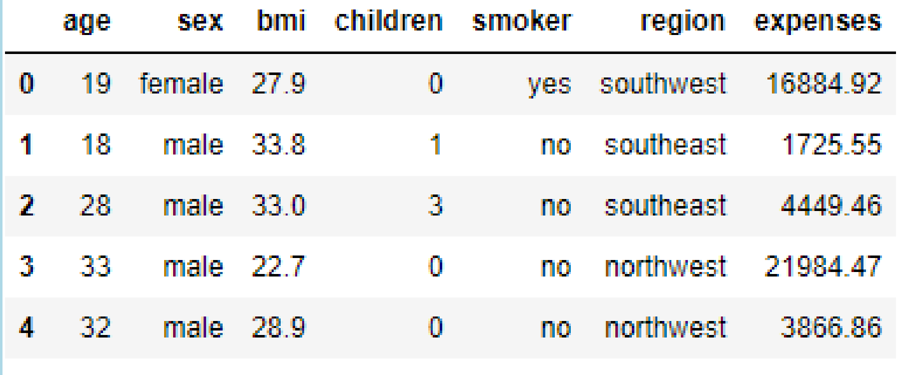
Data Preprocessing – Handling Missing Data

- There is a total of 1338 rows of data that are non-null.
- There are 2 “float64”, 2”int64” and 3 “object” columns.
- There are no missing data present in the data set.

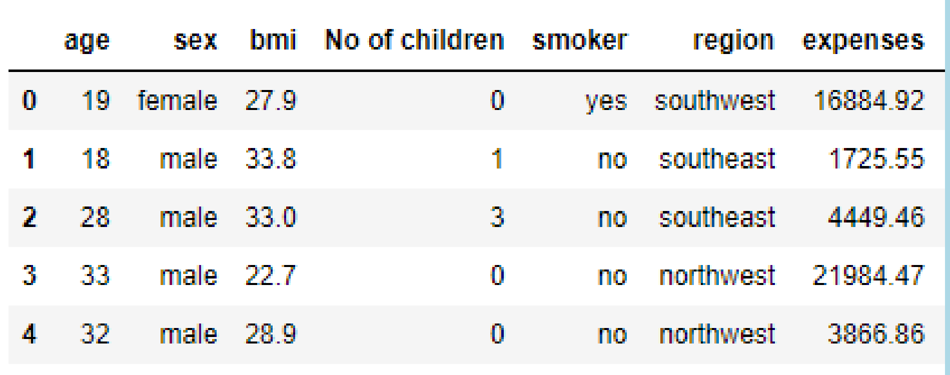


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64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

Data Preprocessing – Data Formatting (Column)



	age	sex	bmi	children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86



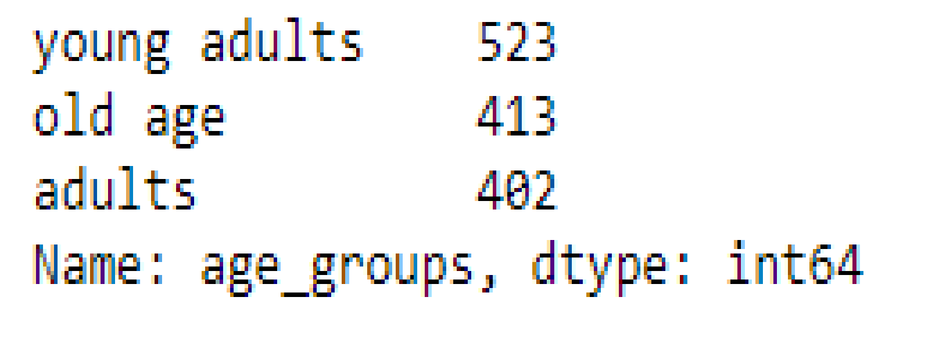
	age	sex	bmi	No of children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86

- The column name ”Children” may be difficult to interpret as to what this column represents.

- The column name “Children” will be rename to “No of Children” for better interpretation to avoid confusion.

Data Preprocessing – Data Binning

- The age of customers varies in this dataset.
- To improve visual exploration in EDA, Data Binning is perform on this column.
- The ages are binned into groups of, young adults, adults and old age.



```
young adults    523
old age         413
adults          402
Name: age_groups, dtype: int64
```

EDA – Columns & Data types

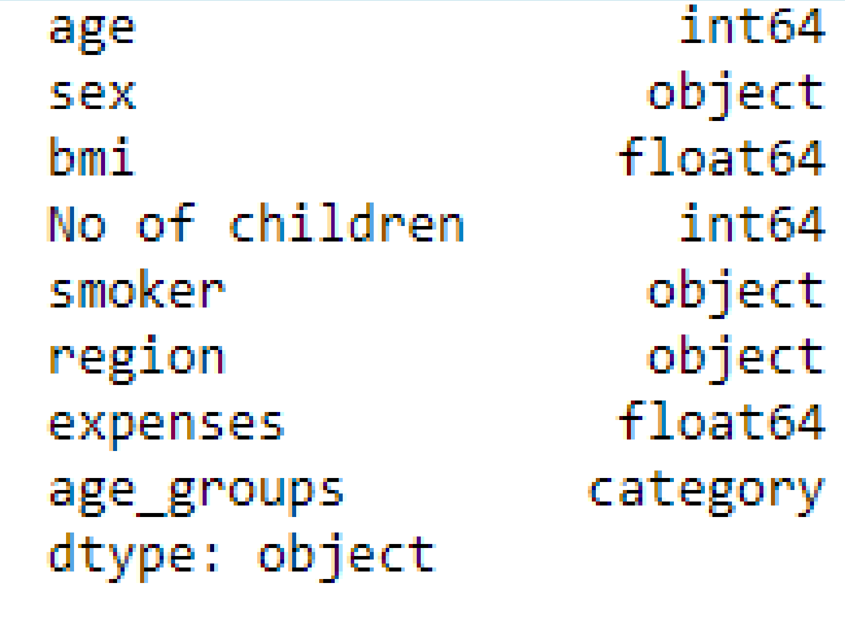
- There are 8 columns in this data set which are age, sex, bmi, No of children, smoker region, expenses age groups.

- There is 2 columns with data type “int64” which are age and no of children.

- 2 “float64” columns, bmi and expenses.

- 3 “object” type columns sex and smoker and region.

- 1 category group age_groups.



```
age                  int64
sex                 object
bmi                float64
No of children       int64
smoker              object
region              object
expenses           float64
age_groups        category
dtype: object
```

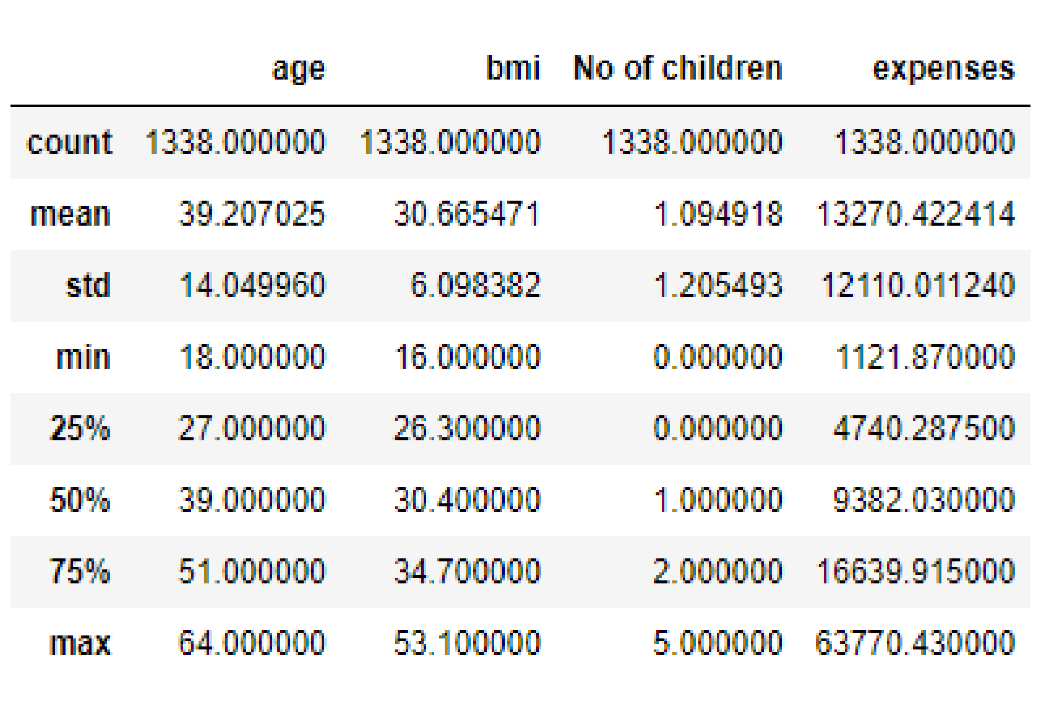
EDA - Descriptive analysis

- The table on the right shows the descriptive analysis summary for data set.

- The mean value for the age is 39, bmi is 30.67, no of children is 1 and expenses is 13270.

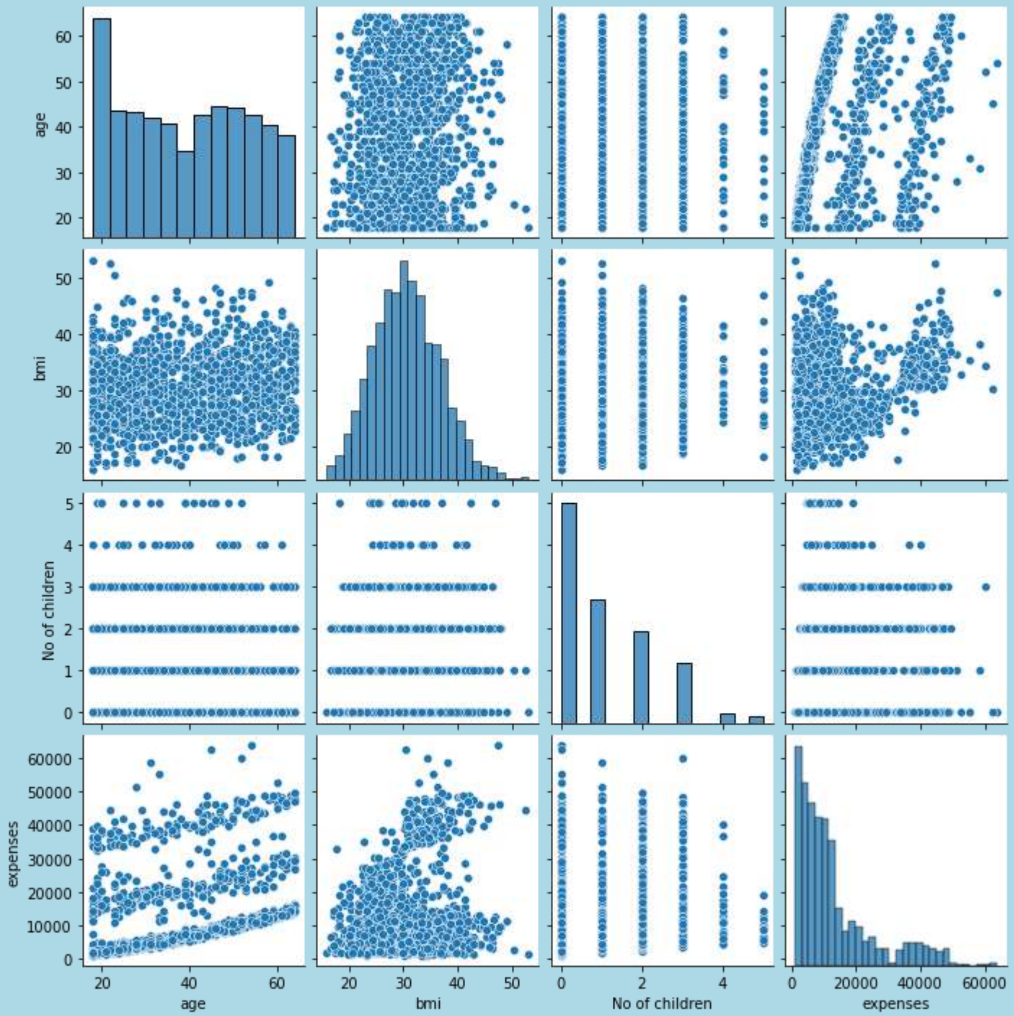
- The minimum age is 18, bmi is 16, no of children is 0 and expenses is 1121.87

- The maximum age is 64, bmi is 53.1, no of children is 5 and expenses is 63770.



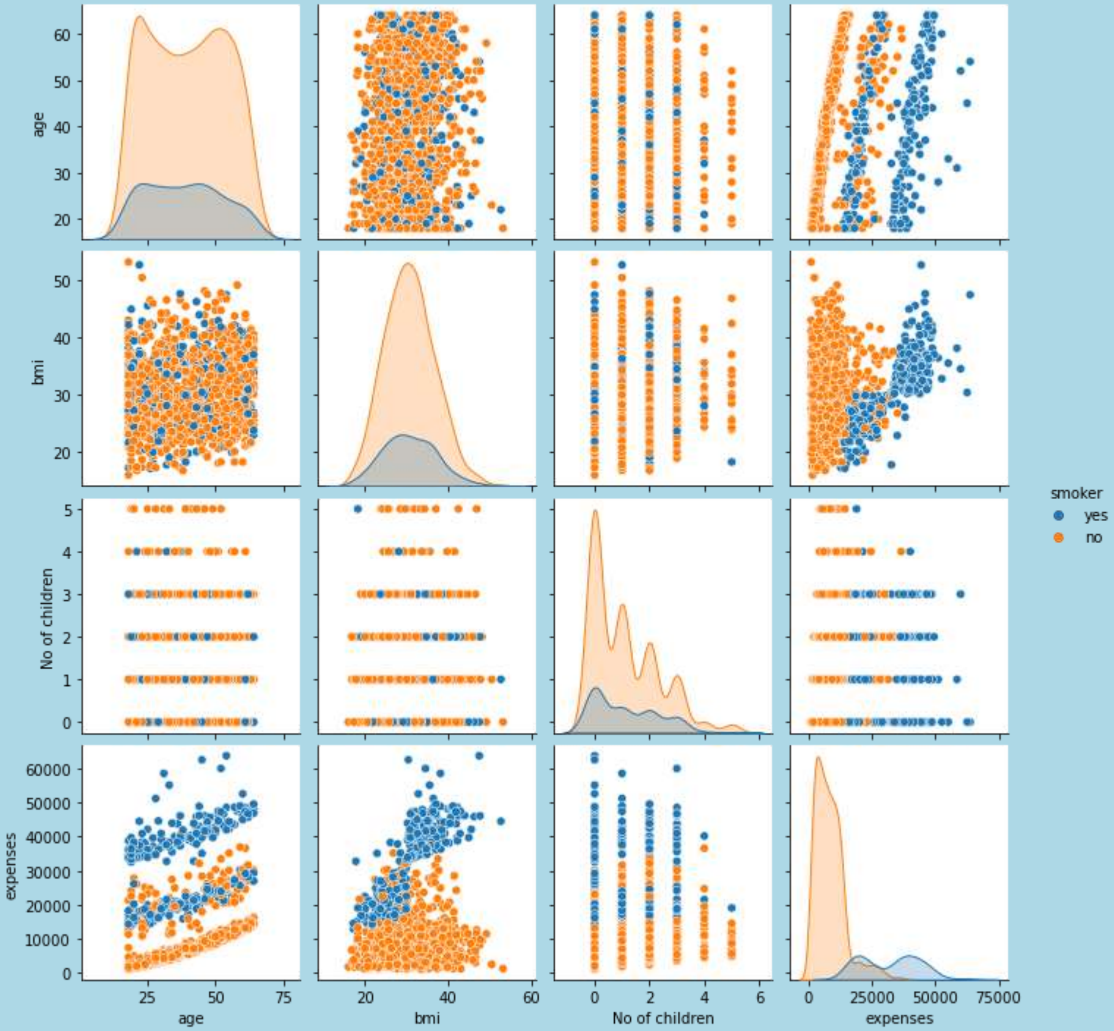
	age	bmi	No of children	expenses
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.665471	1.094918	13270.422414
std	14.049960	6.098382	1.205493	12110.011240
min	18.000000	16.000000	0.000000	1121.870000
25%	27.000000	26.300000	0.000000	4740.287500
50%	39.000000	30.400000	1.000000	9382.030000
75%	51.000000	34.700000	2.000000	16639.915000
max	64.000000	53.100000	5.000000	63770.430000

EDA – Visual exploration (pairplot)



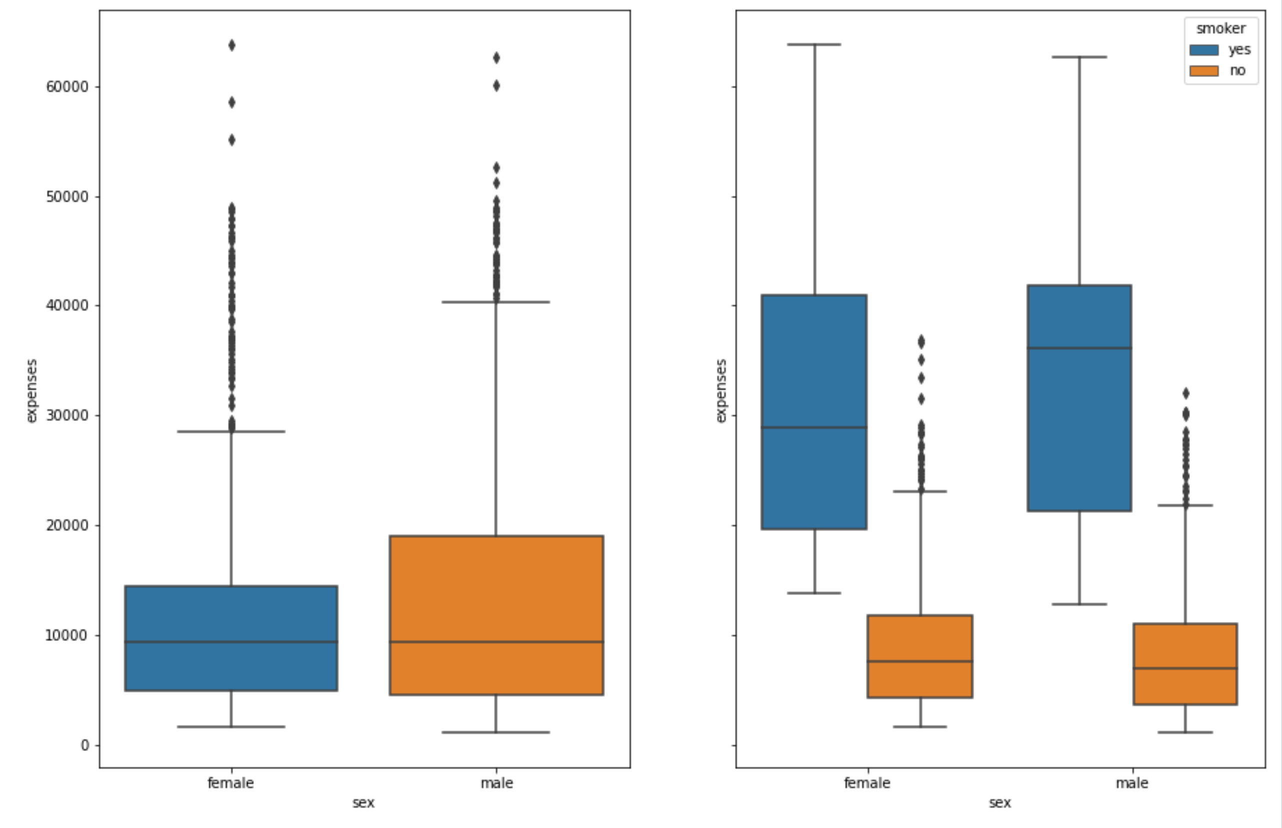
- The pair-plot on the left shows the overall relationship for all variables in data set.
- Notice that the scatter plot are not exactly distributed linearly suggesting categorical variables affecting the spread
- A hue will be included to gain better understanding of the data

EDA – Visual exploration (pairplot with hue)



- The pair-plot on the left shows the overall relationship for all variables in data set including hue for categorical variable smoker.
- The independent variables age and bmi shows significant increase in expenses between smoker and non-smoker.
- This shows that customers who smokes might have strong effect towards the expenses of insurance.

EDA – Visual exploration (Categorical data)

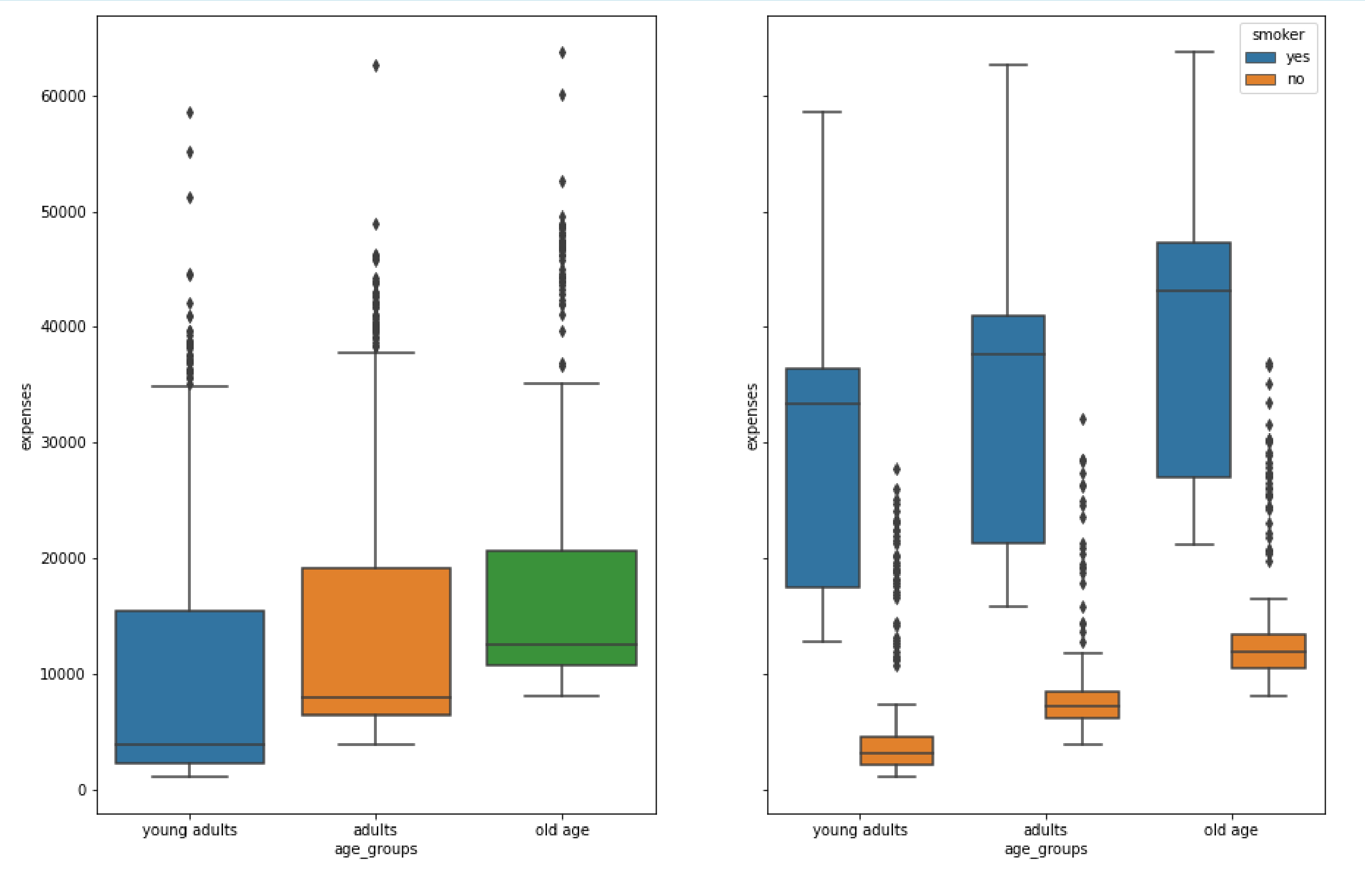


- The boxplot on the right shows the relationship between **sex** and **expenses**.

- The first plot shows that male customers tend to have higher 3rd quartile and max value compared to female while the min, average and 1st quartile are roughly equal.

- The second plot shows both category of **sex** having higher values for smokers compared to non smokers.

EDA – Visual exploration (Categorical data)

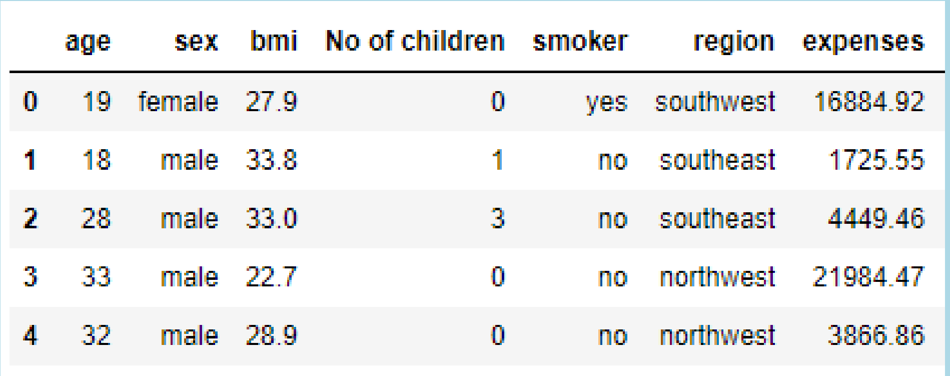


- The boxplot on the right shows the relationship between **age_groups** and **expenses**.
- The first plot shows middle age adults having the highest maximum value while young adults and old age are fairly equal.
- The min, average, 1st and 3rd quartile rises as the age group rises.
- The second plot shows a high amount of outliers for non-smokers for all **age_groups** while no outliers for smokers.

Feature Engineering – (One-Hot Encoding)

- As there are 3 important variables that are categorical (sex, smoker and region).
- Feature Encoding was performed on these 3 variables.
- One-Hot Encoding from Scikit-Learn was used to convert the variables' data type to numeric data type.

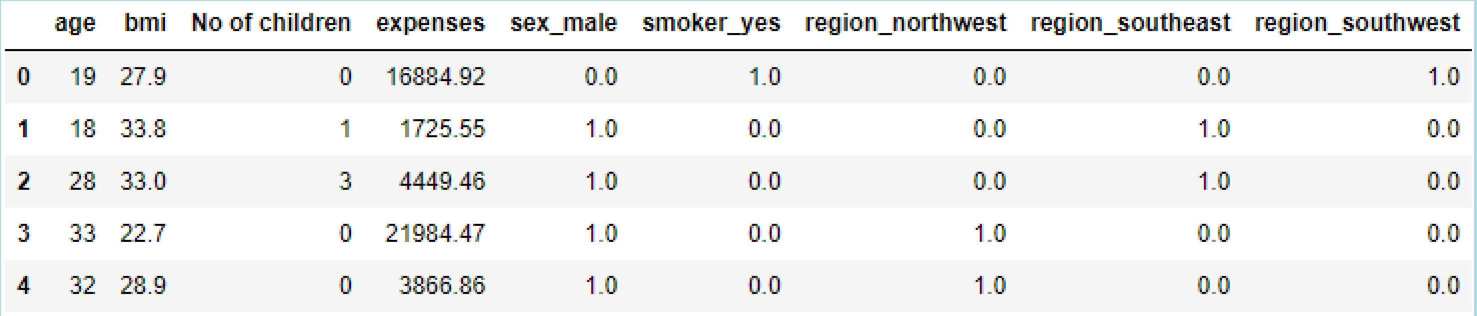
Feature Engineering – (One-Hot Encoding)



	age	sex	bmi	No of children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86

Before One-Hot Encoding

One-Hot Encoding

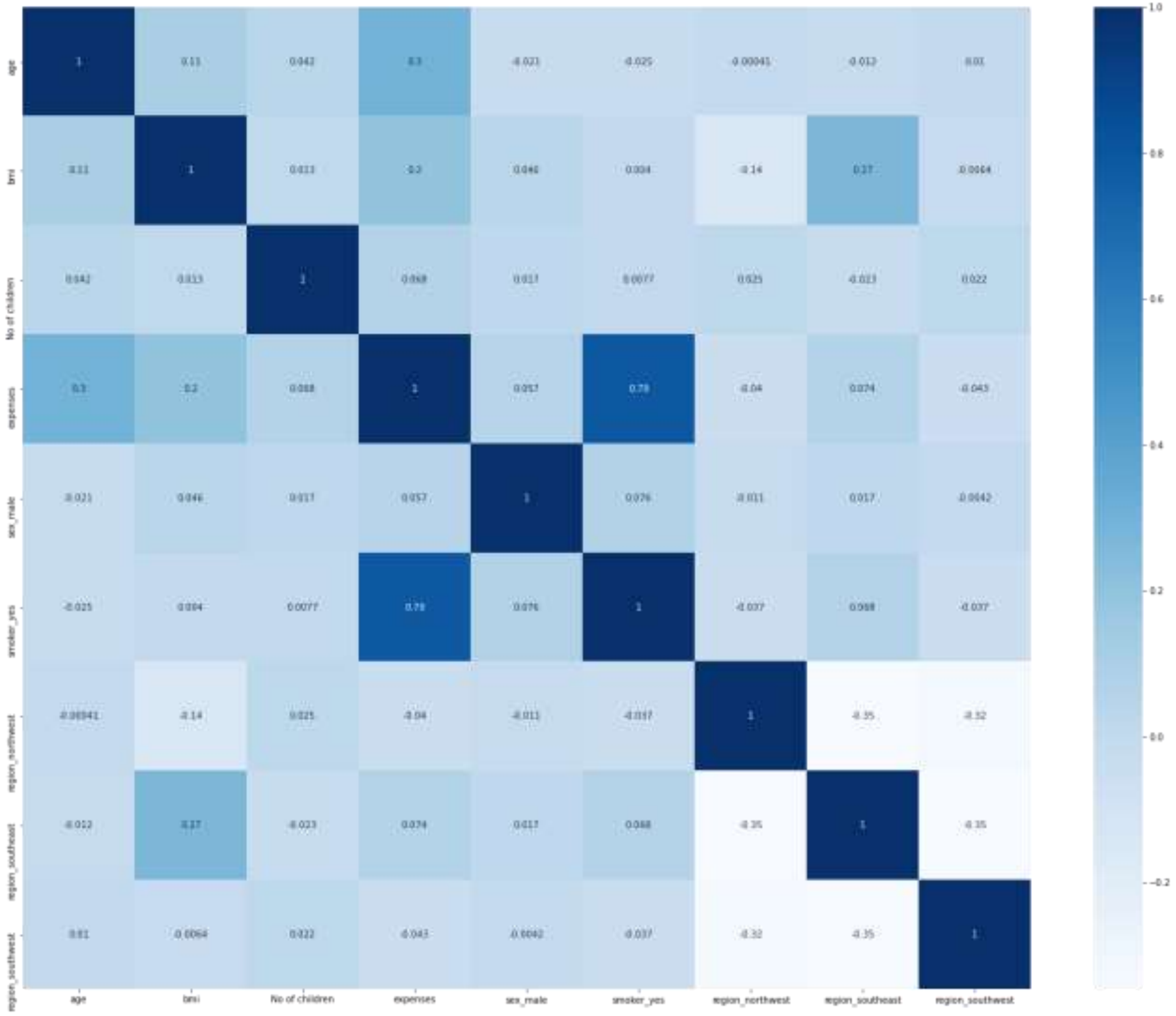


	age	bmi	No of children	expenses	sex_male	smoker_yes	region_northwest	region_southeast	region_southwest
0	19	27.9	0	16884.92	0.0	1.0	0.0	0.0	1.0
1	18	33.8	1	1725.55	1.0	0.0	0.0	1.0	0.0
2	28	33.0	3	4449.46	1.0	0.0	0.0	1.0	0.0
3	33	22.7	0	21984.47	1.0	0.0	1.0	0.0	0.0
4	32	28.9	0	3866.86	1.0	0.0	1.0	0.0	0.0

After One-Hot Encoding

All data types are in numeric form after One-Hot Encoding was performed on the data set

Heat Map



- A heat map was plotted to analyze the relationship between **feature columns** and **target column**.

- From the heat map customer who smokes has the highest correlation between expenses compared to other features.

- The correlation of smokers and expenses is at 0.79.

Machine Learning – Splitting of Data

Data Set was split into X (feature columns) and y (target column)

- Feature columns - age, bmi, No of children, smoker_yes, sex_male, region_northwest, region_southeast and region_southwest
- Target column - expenses

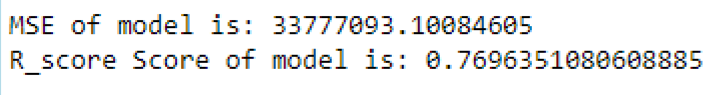
Data split (train_test_split)

- Data set was split into training (70%) and testing data (30%).
- Training set = X_train , Y_train (70%).
- Testing set = X_test , Y_test (30%).

Machine Learning – Linear Regression

Linear Regression

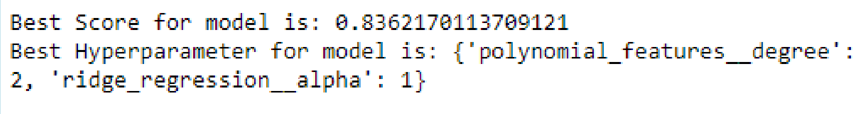
- Linear regression was performed and the results obtained was, MSE = 33777093.2 and R_score = 0.76963514



```
MSE of model is: 33777093.10084605
R_score Score of model is: 0.7696351080608885
```

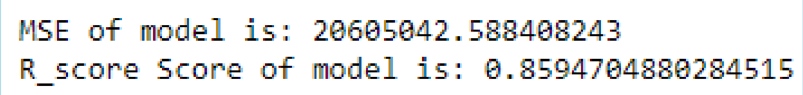
Linear Regression + Polynomial Feature transformation

- Polynomial feature was then applied.
- GridSearchCV was applied to obtained the best selection of polynomial degree.
- According to results from gridsearch the best degree = 2 and best score = 0.8166



```
Best Score for model is: 0.8362170113709121
Best Hyperparameter for model is: {'polynomial_features__degree':
2, 'ridge_regression__alpha': 1}
```

- A model was build using PolynomialFeatures(degree=2) and the results obtained was, MSE = 20605042.56 and R_score = 0.8595



```
MSE of model is: 20605042.588408243
R_score Score of model is: 0.8594704880284515
```

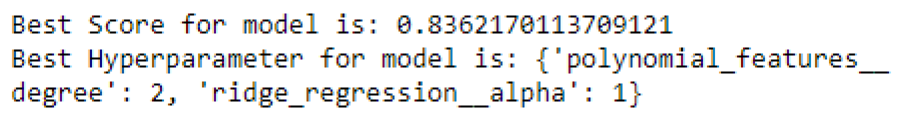
Machine Learning – Regularization (Ridge & Lasso)

Grid Search

- GridSearchCV was applied on both Ridge and Lasso Regularization to obtained best hyper-parameter selection
- The results for each model hyper-parameter were used for training of each model.

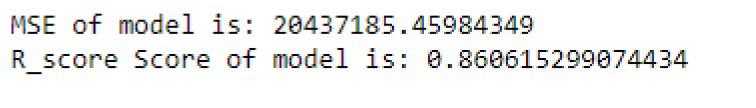
Ridge Regression

- According to results from gridsearch the best degree = 2 with alpha value = 1 and best score = 0.8362.



```
Best Score for model is: 0.8362170113709121
Best Hyperparameter for model is: {'polynomial_features__
degree': 2, 'ridge_regression__alpha': 1}
```

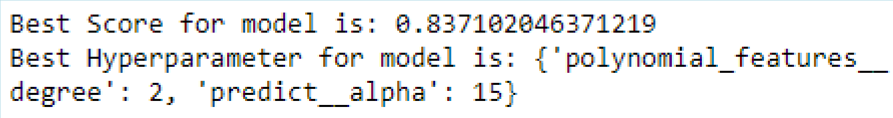
- A model was build using PolynomialFeatures(degree=2) and Ridge(alpha=1) the results obtained was, MSE = 204372 and R_score = 0.86062



```
MSE of model is: 20437185.45984349
R_score Score of model is: 0.860615299074434
```

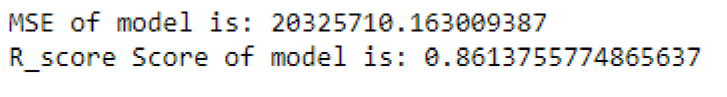
Lasso Regression

- According to results from gridsearch the best degree = 2 with alpha value = 15 and best score = 0.8371.



```
Best Score for model is: 0.837102046371219
Best Hyperparameter for model is: {'polynomial_features__
degree': 2, 'predict__alpha': 15}
```

- A model was build using Polynomial Features(degree = 2) and Lasso(alpha = 15) the results obtained was, MSE = 20325710.163 and R_score = 0.861376

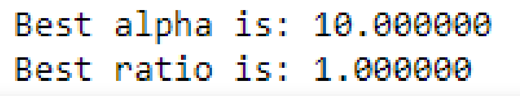


```
MSE of model is: 20325710.163009387
R_score Score of model is: 0.8613755774865637
```

Machine Learning – Regularization (Elastic Net)

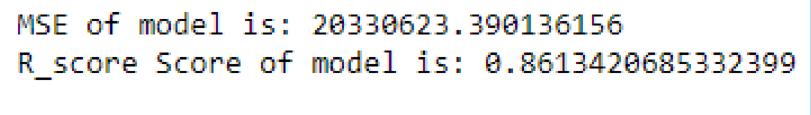
Elastic Net Regression

- ElasticNetCV was applied to obtained the best alpha and ratio value.
- According to results from ElasticNetCV the best alpha = 10 with alpha value = 1.



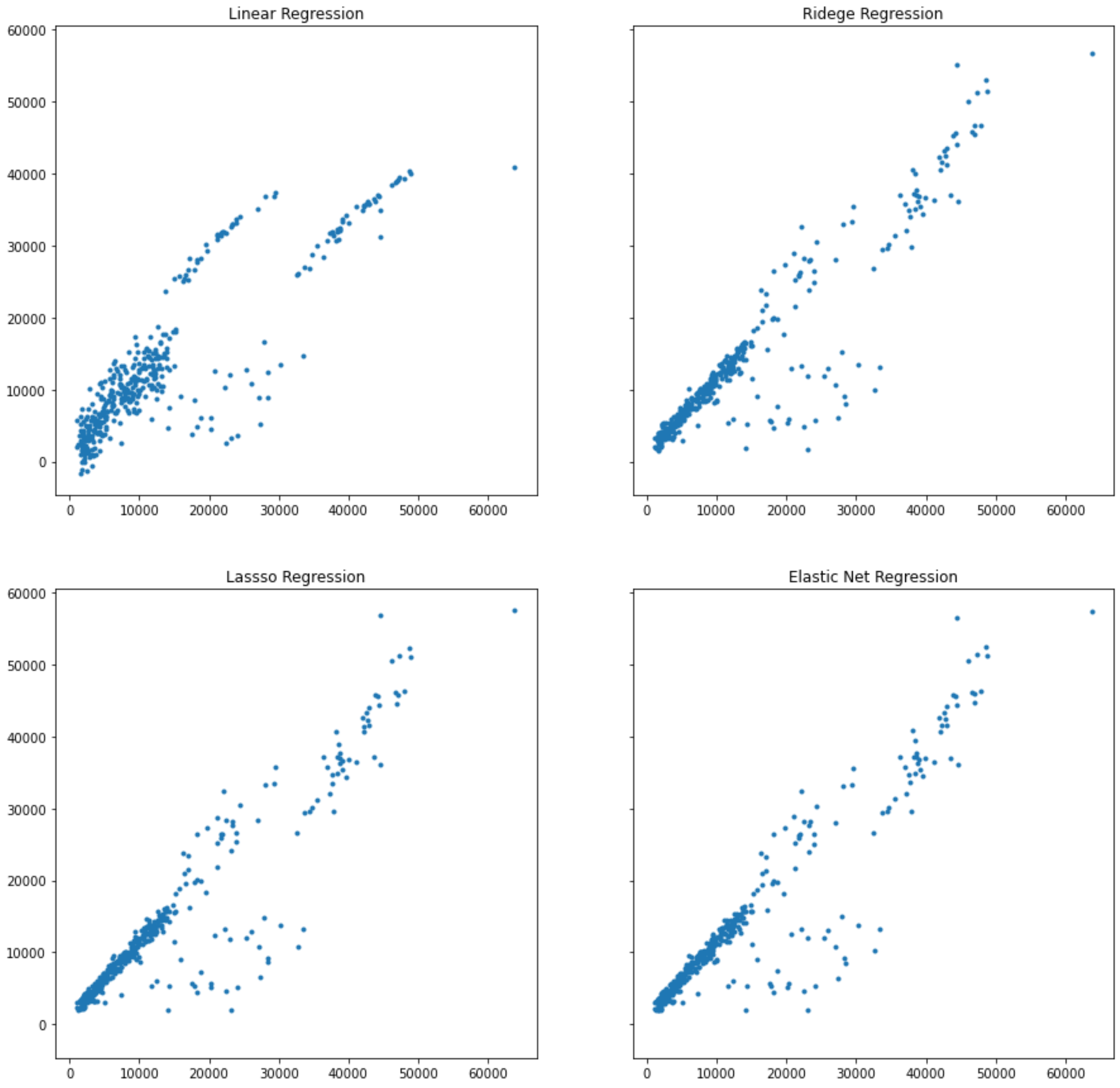
```
Best alpha is: 10.000000
Best ratio is: 1.000000
```

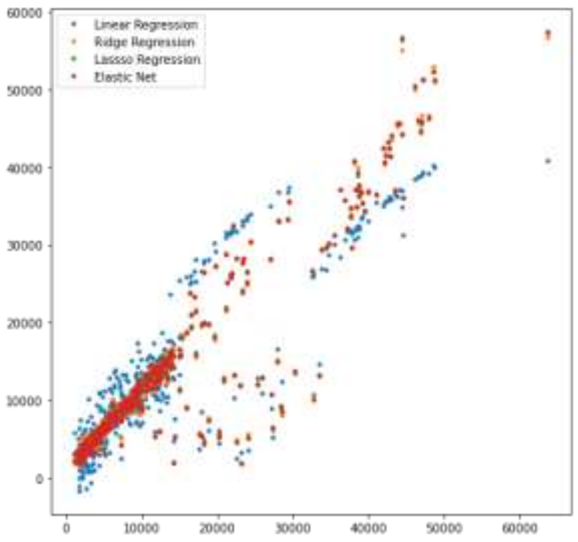
- The model was fit to the training set and predicted using the test set.
- The results obtained was MSE = 20330623.39 and R_score = 0.861342



```
MSE of model is: 20330623.390136156
R_score Score of model is: 0.8613420685332399
```

Graph (Test vs Predicted results)





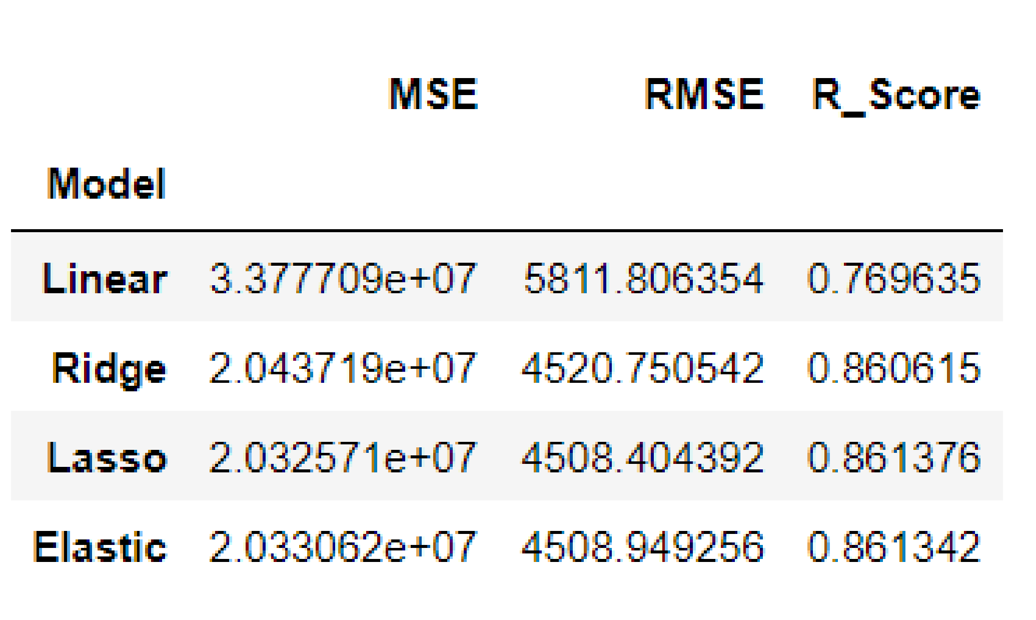
Test vs Predicted values

- The plot on the right shows the graph of **Test** vs **Predicted** values for **Linear Regression, Ridge Regression, Lasso Regression and Elastic Net**.
- The plot on the top right shows the overall results of all model.
- From the plots **Lasso Regression** appears to have the lowest difference/spread between **Test** and **Predicted** value while **Linear Regression** has the highest.

Overall Results

Overall Results

- From this analysis, the highest MSE and RMSE obtained was Linear Regression at MSE = 3.377e+07 and RMSE = 5811.806.
- The Lowest MSE and RMSE obtained was Lasso Regression at MSE = 2.032571e+07 and RMSE = 4508.404392.
- The highest R_Score is from Lasso with value 0.861376 while the lowest is Linear Regression at 0.769635
- Although Elastic Net and Lasso has close results Lasso still has slightly higher results compared to Elastic Net.



	MSE	RMSE	R_Score
Model			
Linear	3.377709e+07	5811.806354	0.769635
Ridge	2.043719e+07	4520.750542	0.860615
Lasso	2.032571e+07	4508.404392	0.861376
Elastic	2.033062e+07	4508.949256	0.861342

Conclusion & Improvements

Conclusion

- In conclusion, The best Model overall is Lasso Regression with lowest MSE & RMSE and highest R_score.

- This can be proven from the analysis done as Polynomial Feature with degree =2 and Regularization through Lasso of alpha = 15 has help reduce the overall MSE and increase the R_score for this analysis.

- In conclusion, Linear regression performs better with addition of Polynomial feature transformation to help reduce bias and Regularization to help prevent overfitting.

Improvements

- Research on other regularization methods to optimize model.

- Gather more data to improve to allow model to obtained better results.

Appendix

Link to Code

https://github.com/cs-robot-collab/IBM-ML-DL/blob/master/IBM%20Machine%20Learning%20Regression%20Report.ipynb