# ASSIGNMENT-2 REPORT

**Ram Chandra B (50414405)**
**Ram Manohar Reddy C (50418469)**

We certify that the code and data in this assignment were generated independently, using only the tools and resources defined in the course and that I did not receive any external help, coaching or contributions during the production of this work.

## PART – 1 (Titanic Dataset)

**1. Provide brief details about the nature of your dataset. What is it about? What type of data are we encountering? How many entries and variables does the dataset comprise?**

A. The TITANIC dataset consists of 8 features namely, “Survived”, “Pclass”, “Name”, ”sex”, “Age”, “Siblings/spouses Aboard”, "Parents/Children Aboard”, and “Fare”. Most of these features are numerical. And the rest of the non-numerical data is converted to categorical.

**Numerical Data -** “Survived”, “Pclass”, “Age”, “Siblings/spouses Aboard”, "Parents/Children Aboard”, “Fare”.

**Categorical Data -** “Name”, “Sex”.

This dataset is all about the details of the passengers (like Name, Sex, Age...) in the Titanic who survived (1) or Deceased (0) after the disaster.

It has 887 entries and the above-mentioned 8 variables, so the shape of the data set is (887 x 8).

**2. Provide the main statistics about the entries of the dataset (mean, std, number of missing values, etc.)**

The mean, std, median for each of these variables are detailed below in a table,

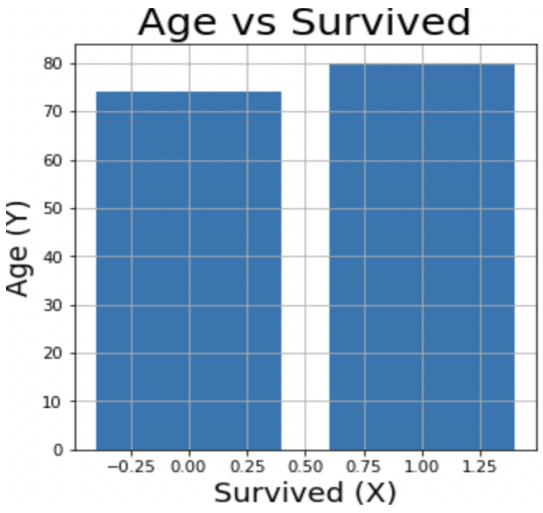
| | Survived | Pclass | Age | Siblings/Spouses Aboard | Parents/Children Aboard | Fare |
|---|---|---|---|---|---|---|
| **count** | 887.000000 | 887.000000 | 887.000000 | 887.000000 | 887.000000 | 887.00000 |
| **mean** | 0.385569 | 2.305524 | 29.471443 | 0.525366 | 0.383315 | 32.30542 |
| **std** | 0.487004 | 0.836662 | 14.121908 | 1.104669 | 0.807466 | 49.78204 |
| **min** | 0.000000 | 1.000000 | 0.420000 | 0.000000 | 0.000000 | 0.00000 |
| **25%** | 0.000000 | 2.000000 | 20.250000 | 0.000000 | 0.000000 | 7.92500 |
| **50%** | 0.000000 | 3.000000 | 28.000000 | 0.000000 | 0.000000 | 14.45420 |
| **75%** | 1.000000 | 3.000000 | 38.000000 | 1.000000 | 0.000000 | 31.13750 |
| **max** | 1.000000 | 3.000000 | 80.000000 | 8.000000 | 6.000000 | 512.32920 |

Null values are identified by using this statement isnull().sum(axis=0), which resulted in zero for all variables which states that there are no null values.

Variable “Age” has **25** floating values which are inappropriate to that column, so those values are replaced by the median of that column.

**3. Provide at least 5 visualization graphs with short description for each graph, e.g., discuss if there are any interesting patterns or correlations.**

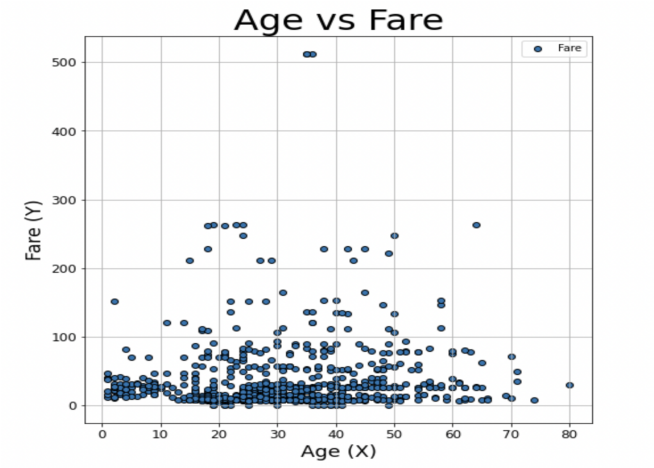
I) Age vs Survived




Irrespective of age, the death rate is more than the survival rate.  older people have higher survival rates.

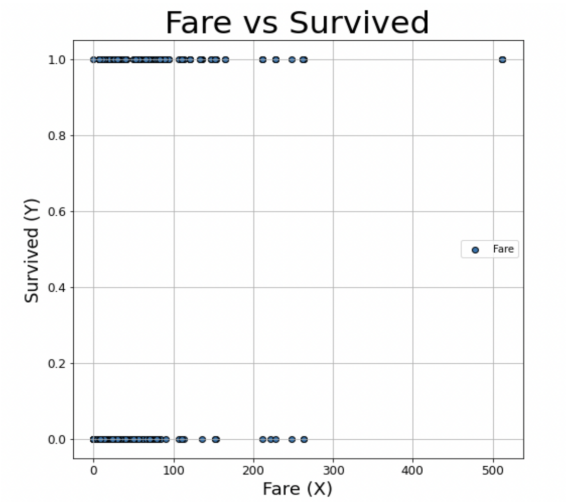
The above graph Age vs Survived is a scatter plot, which describes people who are between 65 to 80 years deceased. Overall, we can say that most of the people deceased in that incident.

II) Age vs Fare

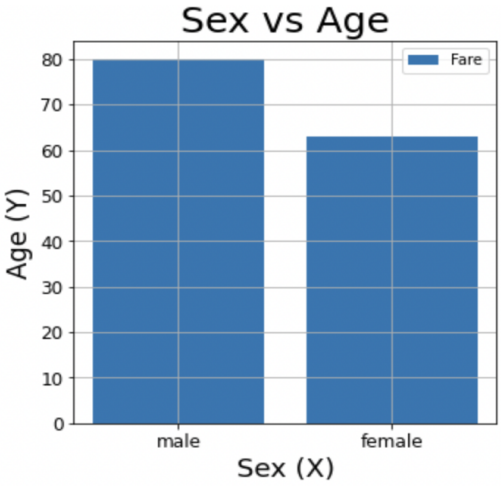



Fare is less for maximum times irrespective of age factor.

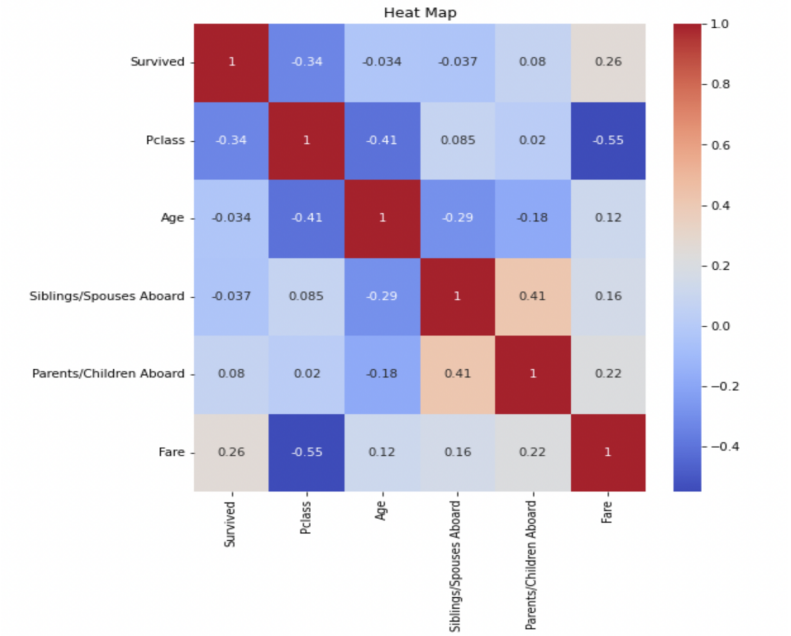
III) Fare vs Survived




IV) Sex vs Age




V) Heat map




Survival probability is high for people in ‘Pclass’, survival probability is low for people in siblings/children.

## Part – 1 (Wine Quality Dataset)

**1. Provide brief details about the nature of your dataset. What is it about? What type of data are we encountering? How many entries and variables does the dataset comprise?**

A. The WINE QUALITY dataset consists of 12 features namely, ‘fixed acidity', 'volatile acidity’, 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulfates', 'alcohol', 'quality'. As this dataset is of numerical value, this makes it easy to clean the data.

This dataset gives us the factors like chlorides, alcohol, pH... that affect the quality of wine.

This dataset consists of 1599 entries and the above-mentioned 12 variables, so the shape of the data set is (1599 x 12).

**2. Provide the main statistics about the entries of the dataset (mean, std, number of missing values, etc.)**

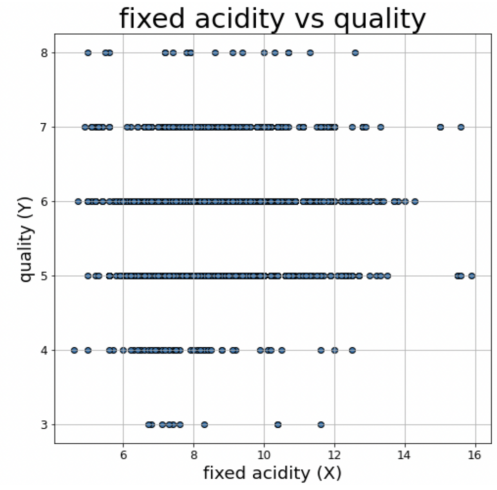
The mean, std, median for each of these variables are detailed below in a table,

| | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 |
| mean | 0.527821 | 0.270976 | 2.538806 | 0.087467 | 15.874922 | 46.467792 | 0.996747 | 3.311113 | 0.658149 | 10.422983 | 5.636023 |
| std | 0.179060 | 0.194801 | 1.409928 | 0.047065 | 10.460157 | 32.895324 | 0.001887 | 0.154386 | 0.169507 | 1.065668 | 0.807569 |
| min | 0.120000 | 0.000000 | 0.900000 | 0.012000 | 1.000000 | 6.000000 | 0.990070 | 2.740000 | 0.330000 | 8.400000 | 3.000000 |
| 25% | 0.390000 | 0.090000 | 1.900000 | 0.070000 | 7.000000 | 22.000000 | 0.995600 | 3.210000 | 0.550000 | 9.500000 | 5.000000 |
| 50% | 0.520000 | 0.260000 | 2.200000 | 0.079000 | 14.000000 | 38.000000 | 0.996750 | 3.310000 | 0.620000 | 10.200000 | 6.000000 |
| 75% | 0.640000 | 0.420000 | 2.600000 | 0.090000 | 21.000000 | 62.000000 | 0.997835 | 3.400000 | 0.730000 | 11.100000 | 6.000000 |
| max | 1.580000 | 1.000000 | 15.500000 | 0.611000 | 72.000000 | 289.000000 | 1.003690 | 4.010000 | 2.000000 | 14.900000 | 8.000000 |

NULL values are identified by using this statement isnull().sum(axis=0), which resulted in zero for all variables which states that there are no null values.

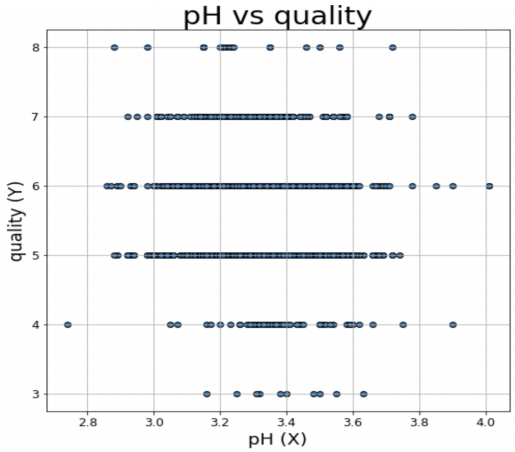
**3. Provide at least 5 visualization graphs with short description for each graph, e.g., discuss if there are any interesting patterns or correlations.**

I) Fixed acidity vs quality



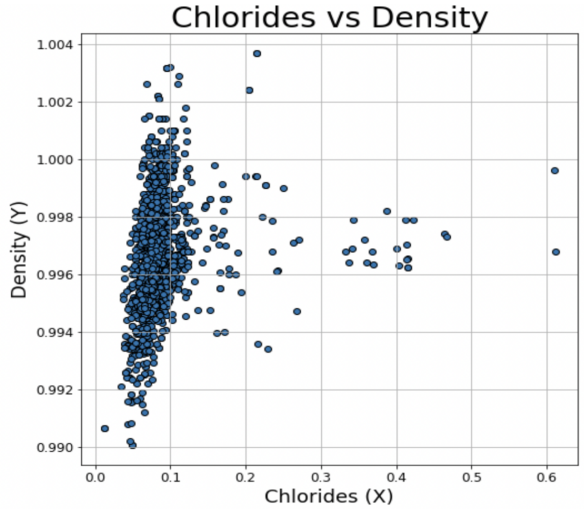
After predicting the graph, we can examine that wine quality is not dependent on acid levels of wine.

II) pH vs Quality



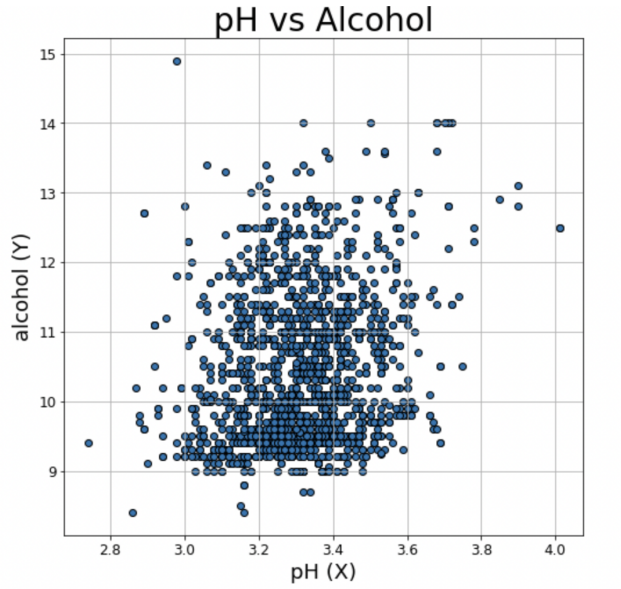

Observing the above graph, for most of the pH values wine quality is between 5 to 7. but quality doesn't seem to be dependent on pH value.

III) Chlorides vs Density



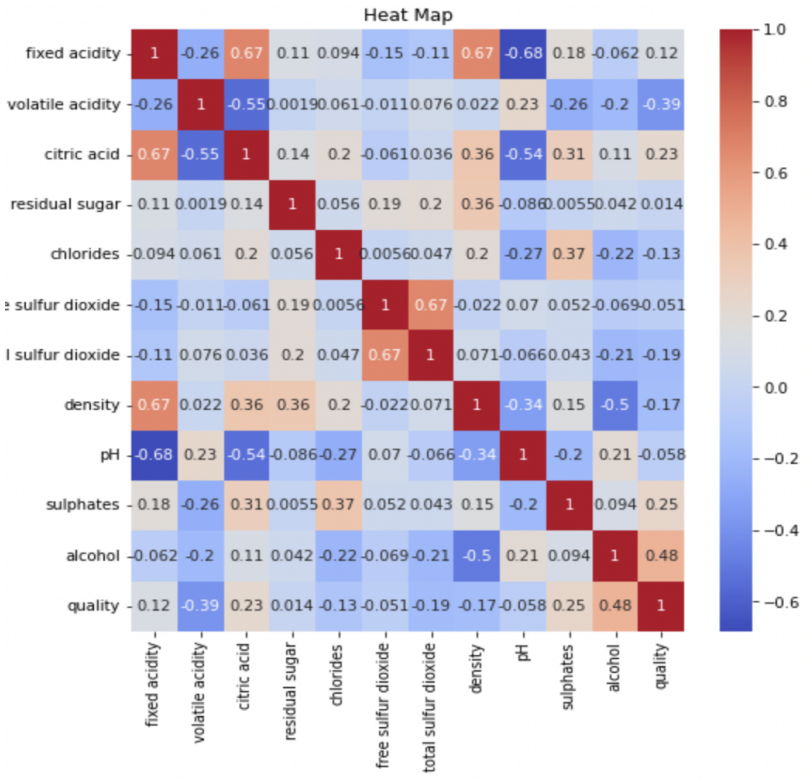

Chlorides are high at density values between 0.994 to 0.998, and chlorides levels are almost low irrespective of density level.

IV) pH vs Alcohol




Graph between pH and alcohol, for the values between 10-12 of alcohol values, pH values are between 3.0 to 3.4

V) Heat Map




Product quality is maximum when chlorides, sulfur dioxide and density levels are low.

**PART– 1 (Insurance Dataset)**

**1. Provide brief details about the nature of your dataset. What is it about? What type of data are we encountering? How many entries and variables does the dataset comprise?**

A. The Insurance QUALITY dataset consists of 7 features namely, 'age, ‘sex, ‘bmi', 'smoker', ‘children’, ‘region’, 'charges’. This dataset describes the insurance charges based on sections of category for example,

This dataset gives us factors like chlorides, alcohol, pH... that affect the quality of wine.

It has 1599 entries and the above-mentioned 12 variables, so the shape of the data set is (1599 x 12).

**2. Provide the main statistics about the entries of the dataset (mean, std, number of missing values, etc.)**

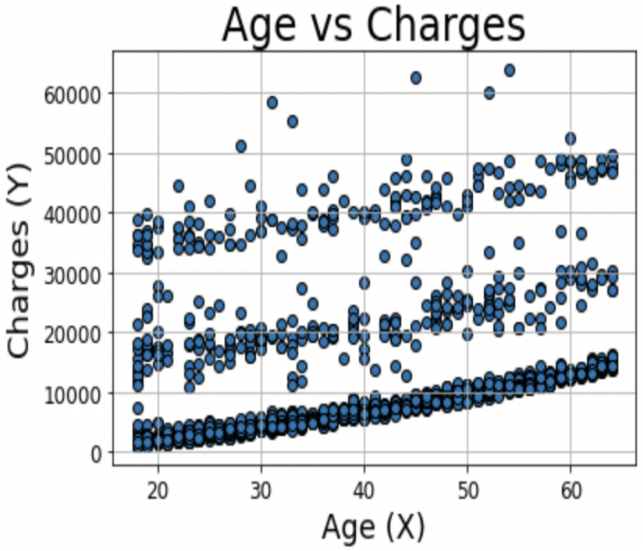
The mean, std, median for each of these variables are detailed below in a table,

| | 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 |

Null values are identified by using this statement isnull().sum(axis=0), which resulted in zero for all variables which states that there are no null values.

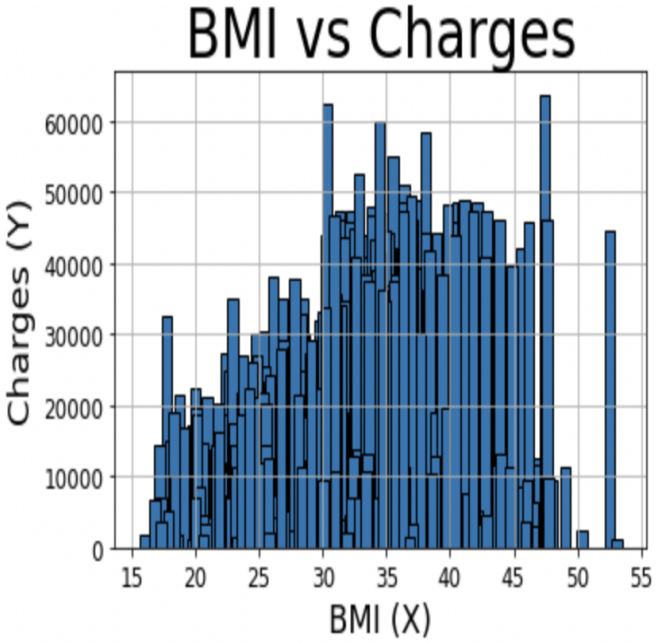
**3. Provide at least 5 visualization graphs with short description for each graph, e.g., discuss if there are any interesting patterns or correlations.**

I) Age vs charges



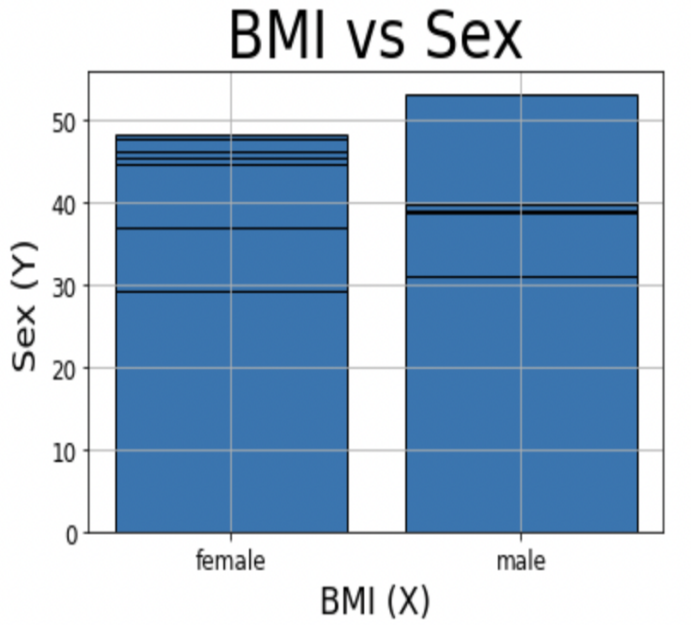

After the observation in the graph, insurance charges increase with age.

II) BMI vs charges



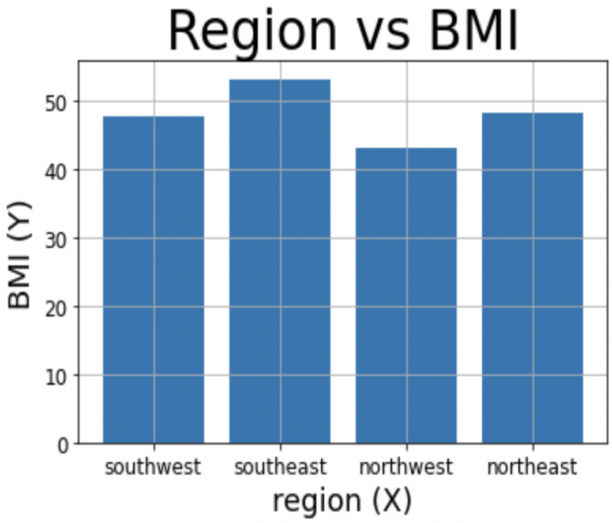

Most people are charged high, when their BMI level is 48.

III) BMI vs Sex



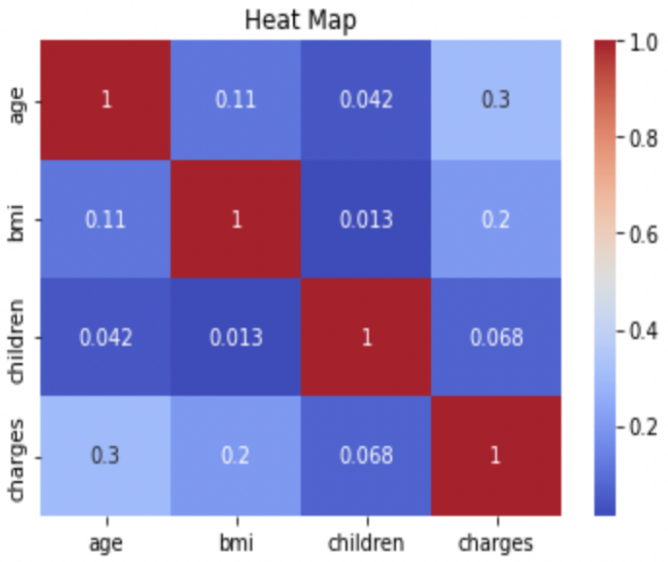

BMI value of men is higher than women.

IV) Region vs BMI




People living in the southeast region have the highest BMI values and people in the northwest have lesser BMI values.

V) HEAT MAP




From the above graph, the probability of charges and children is high, which means people with children tend to have higher charges. And people with BMI levels high and in older age, get to charge more.

## PART – 2

### Logistic Regression

**1. Provide your best accuracy and the weight vector.**

For the Penguin dataset, we got the best accuracy of 88.363636% for the training dataset whereas 82.60869% for the test dataset using 'Species', 'island', 'culmen_length_mm', 'flipper_length_mm' and 'body_mass_g' as input features and 'sex' as target variable.

And the optimized weights are -

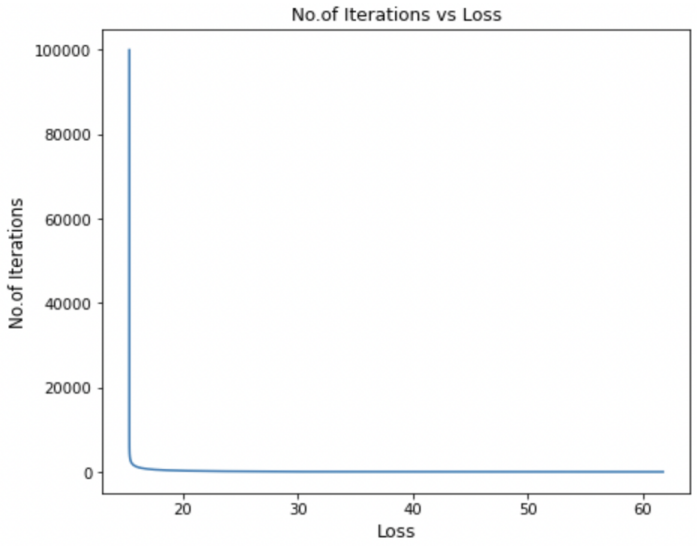
```
[[  0.41620604  -2.61871405 -12.73422648   1.11822165 -14.97090533]]
```

The Confusion Matrices obtained for both train and test data (without using any library) are -

```
Confusion Matrix for Train data:
[[126  15]
 [ 16 118]]
Training Accuracy for binary classification: 88.72727272727273 %
Confusion Matrix for Test data:
[[31  6]
 [ 4 28]]
Test Accuracy for binary classification: 85.5072463768116 %
```

**2. Include a loss graph and provide a short description.**

The loss values that are obtained from the cost function are stored in a list and these values are used to plot a graph against number_of_iterations. The plot is obtained below,




**3. Explain how hyperparameters influence the accuracy of the model. Provide at least 3 different setups with learning rate and #iterations and discuss the results**

The hyperparameters have a huge influence on the accuracy of the model. This can be detailed through cases,

Case-1:

When the learning rate is 0.00247 (e^ (-6)) and number of iterations is 10000, the accuracy of the train data is 88.3636% whereas test data is 82.608%.

Case-2:

When the learning rate is 0.0002 and number of iterations is 10000, the accuracy of the train data is 77.8% whereas test data is 71.3%.

Case-3:

When the learning rate is 0.000001 and number of iterations is 10000, the accuracy of the train data is 48.7% whereas test data is 46.3%.

Number of iterations: 10, 1000, 10000, 100000.....

Accuracy: 49.45%, 80.4%, 88.3%, 88.66%…...

Through these cases, we can say that, as the learning rate decreases, the accuracy also decreases. Similarly, as the number of iterations increases, accuracy also increases, and after a particular number of iterations, the accuracy stabilizes.

**4. Discuss the benefits/drawbacks of using a Logistic Regression model.**

**Benefits of using logistic regression:**

1. It is easy to implement.
2. It performs well when the dataset is linearly separable.
3. It can produce model coefficients as an indicator of feature importance.

**Drawbacks**:

1. To implement logistic regression, there should be no multicollinearity between independent variables.
2. Using logistic regression, if the number of entries is less than the number of features, is not appropriate as it may lead to overfitting.
3. The dependent variable of logistic regression is bound to a discrete number set, as it can only predict discrete functions.

## PART-3

**1. Provide your loss value and the weight vector.**

**To get the Loss Value, we need to follow the steps mentioned below,**

**Step-1: U**sing the equation below to generate a weight vector.

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

**Step-2:** Obtained weights are:

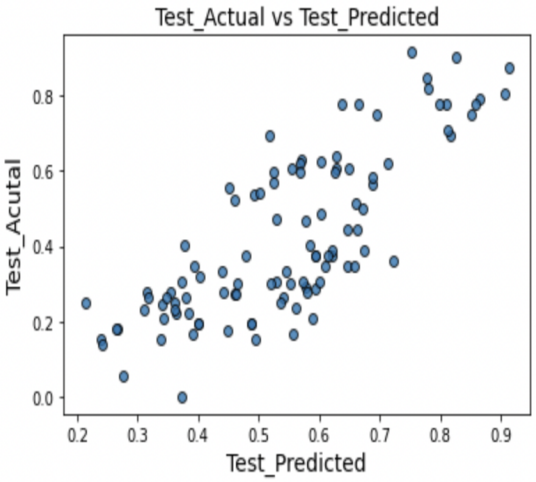
```
Weights:  [-0.00221308  0.48974073 -0.02413134  0.4066759   0.10270879]
```

Using these weights we can find the predicted values.

Mean Squares Loss is determined using predicted values.

```
Mean_Squared_Error(MSE): 3.984615031630877
```

**2. Show the plot comparing the predictions vs the actual test data**




Above plot describes the graph for predictions vs the actual data.

### 3. Discuss the benefits/drawbacks of using OLS estimate for computing weights

**Benefits:**

1. The OLS method is simple to perform.
2. Computation is made easy with the help of OLS**.**
3. This method helps to differentiate the roles of different variables to obtain an output.

**DrawBacks**:

1. We can experience deficient performance of OLS for multivariate dataset with multiple dependent variables set and single independent variables set.
2. For trustworthy results, we will need a dataset with enormous size.

## PART-4

#### 1.Provide your loss value and the weight vector

To get the Loss Value, we need to follow the steps mentioned below,

**Step-1: U**sing the equation below to generate a weight vector.

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\lambda}\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

**Step-2:** Obtained **weight vector** is:

```
Ridge Weights:  [ 0.07473064  0.43210235 -0.09375357  0.3215537   0.12693212]
```

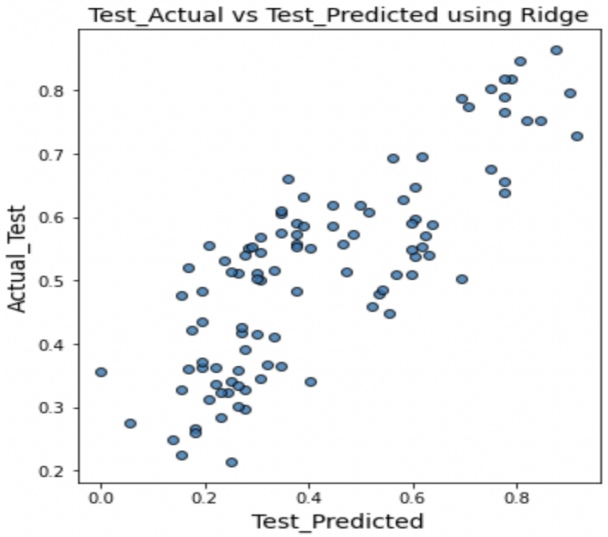
Obtained weights are optimal , produced at lambda=2.

Using these weights we can find the predicted values, which are used to find the **Loss value.**

```
Ridge_loss with optimal weights:  1.494923219297409
```

#### 2. Show the plot comparing the predictions vs the actual test data

Below plot describes the graph for predictions vs the actual data.



**3. Discuss the difference between Linear and Ridge regressions. What is the main motivation of using l2 regularization?**

| Linear Regression | Ridge Regression |
|---|---|
| This method considers the linear relationship between independent input variables and dependent output variables. | To analyze the data suffering from multicollinearity, ridge regression is used. |
| Simple linear regression contains input with single variable, Multiple linear regression contains input with multiple variables. | When count of correlated features is high, ridge regression comes into action. |
| Methods of linear regression are:<br>1. Simple linear regression<br>2. OLS (Ordinary least squares)<br>3. Gradient descent | Ridge regression is also called L2 Regression. |

L2 regularization helps in reducing the effect of correlated inputs.

**BONUS**

**Gradient Descent From Scratch:**

Gradient descent is implemented for the ridge regression by updating the weight and bias.

These are the final weights obtained.

```
weights  [ 0.04426086  0.45916822 -0.06974523  0.3539764   0.1180086 ]
Bias:  -0.016920996469272996
Mean Square Error (MSE):  8.34134538091466
```

The Training Errors are:

```
Train Loss with Gradient Descent function:  5209.900847540625
Train Loss with Inbuilt function:  1.8276016196169123
```

The Testing Errors are:

```
Test Loss with Gradient Descent function:  2.578157479193737
Test Loss with Inbuilt function:  2.8972145798913234
```

When compared to OLS, MSE is high when using gradient descent.

Here is a comparison table for MSE values of OLS , Ridge and Gradient Descent.

| | MSE |
|---|---|
| **OLS** | 3.984615 |
| **RIDGE** | 8.338488 |
| **RIDGE_GD** | 8.341345 |

Training time and predicting time of the model is :

```
training time:  1.3534331321716309
predicting time:  0.00011181831359863281
```

**References:**


1. https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b
2. https://www.geeksforgeeks.org/python-pandas-dataframe/
3. https://towardsdatascience.com/feature-transformation-for-multiple-linear-regression-in-python-8648ddf070b8
4. https://pbpython.com/dataframe-gui-overview.html
5. https://www.geeksforgeeks.org/how-to-check-the-execution-time-of-python-script/


| Team Member | Assignment Part | Contribution(%) |
|---|---|---|
| Ram Chandra Bhavirisetty | Part 1 | 50% |
| Ram Manohar Chundi | Part 1 | 50% |
| Ram Chandra Bhavirisetty | Part 2 | 50% |
| Ram Manohar Chundi | Part 2 | 50% |
| Ram Chandra Bhavirisetty | Part 3 | 50% |
| Ram Manohar Chundi | Part 3 | 50% |
| Ram Chandra Bhavirisetty | Part 4 | 50% |
| Ram Manohar Chundi | Part4 | 50% |