CSE 574
Introduction to Machine Learning

Instructor: Alina Vereshchaka

Assignment – 2
Classification and Regression Model

Pratik Malani (#50416266)
Babar Ahmed Shaik (#50416252)

“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 (we) did not receive any external help, coaching or contributions during the production of this work."

Part-1 Data Analysis

Dataset: Titanic

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?**

   Titanic.csv is a tabular database containing Name, Age, Sex, and a few other details of the passengers who boarded the TITANIC.

   We encounter Integer (int64), Float (float64), and Object (object) data types.

   The dataset comprises of 887 entries and has 8 variables Name, Age, Sex, Fare, Pclass (Boarding Class), Survives (Survive status), Siblings/Spouses (of passenger aboard), Parents/Children (of passenger aboard).

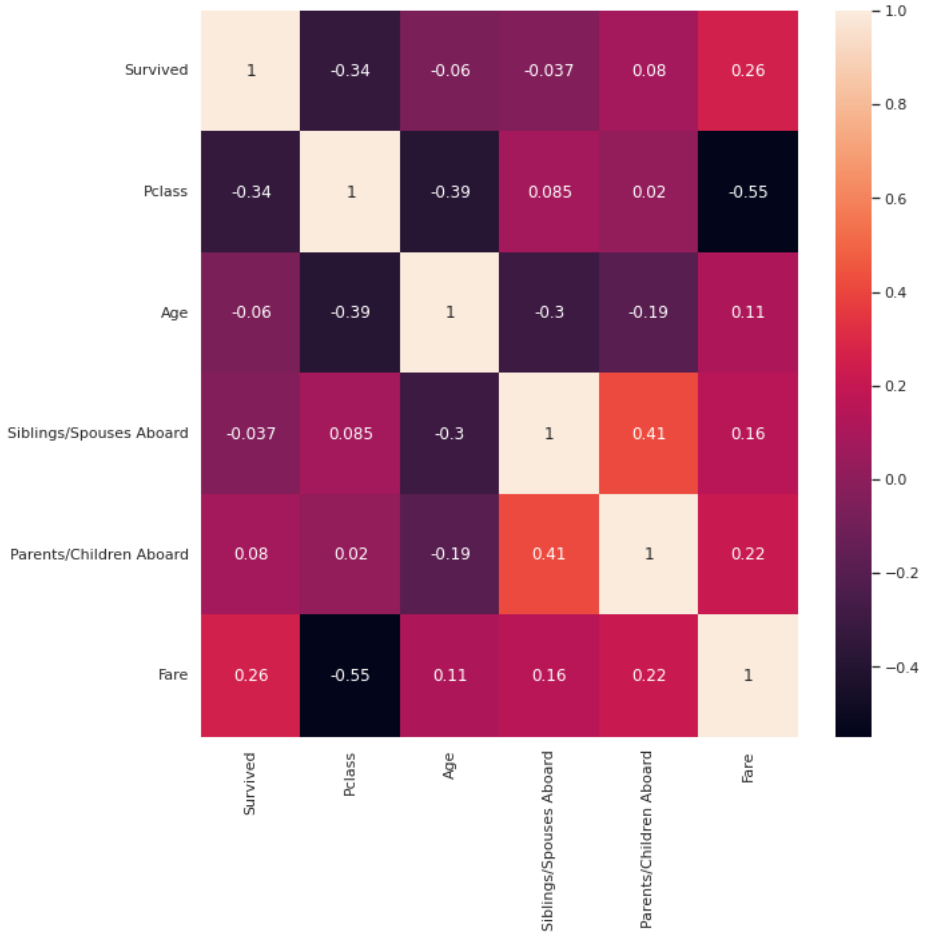
2. **Provide the main statistics about the entries of the dataset (mean, std, number of missing values, etc.)**

	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

   No Missing Values.

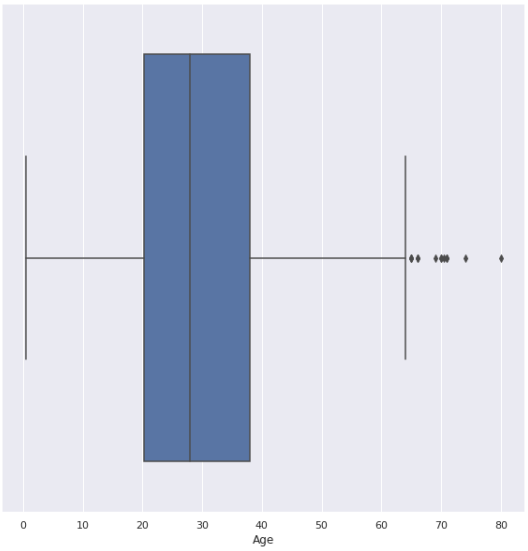
3. **Provide at least 5 visualization graphs with brief description for each graph, e.g. discuss if there are any interesting patterns or correlations.**

1. In the correlation graph we can infer the following:



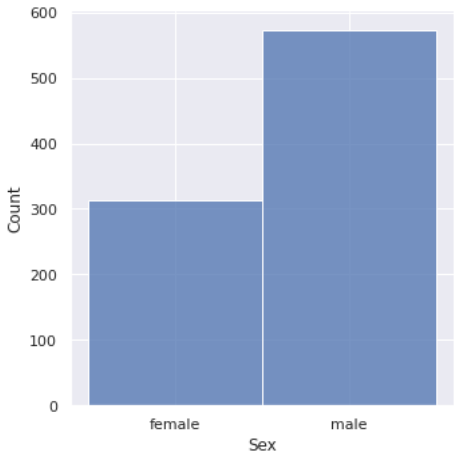
   a. P-class and Fare have a negative correlation of 0.55. Thus, we can say that they are inversely proportional to each other.
   b. The number of Parents/Children aboard is directly proportional to Number of Siblings/Spouses Aboard as it has a positive correlation of 0.41.
   c. The number of Siblings/Spouses Aboard has little to no relation to the number of survivors. We can also see the same between P-class and Parents/Children Aboard.

2.



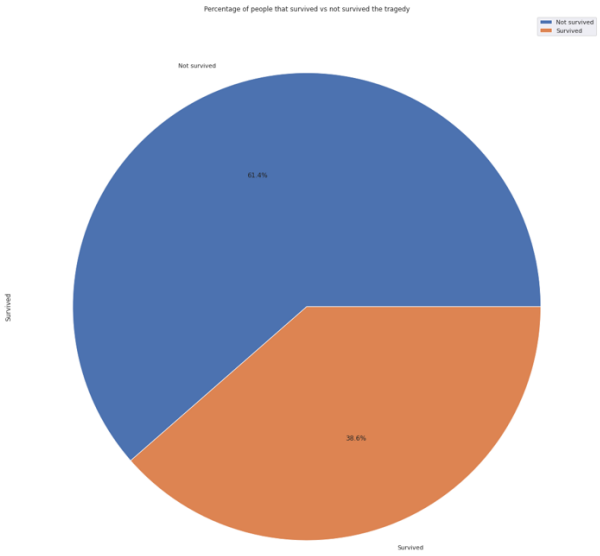
In this boxplot graph we can conclude that the maximum people aboard were from the age range between 20-40.

3.

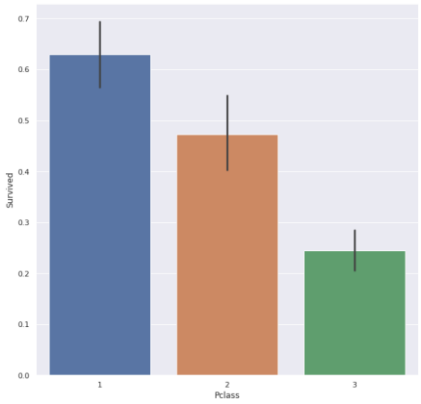


In this frequency graph of the genders of people aboard, the number of males were close to the double of female aboard.

4.



5. From the pie chart plotted, only 38.6% of the people survived and the rest were casualties.



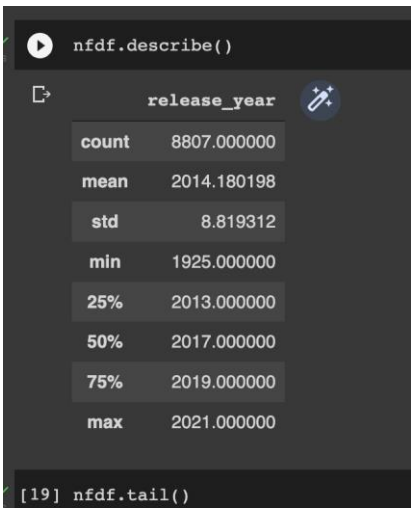
A p-class box plot was plotted to see which class had more survivors. Thus, money bought the first-class passengers life

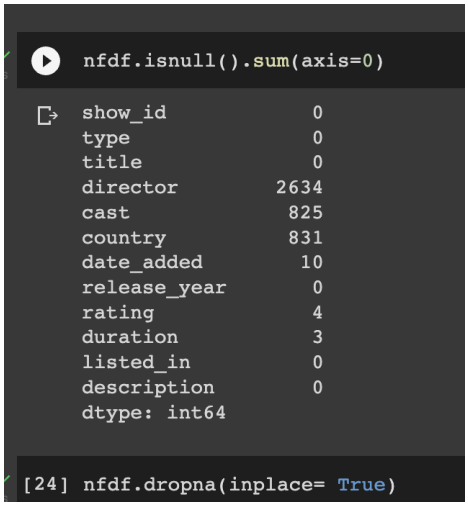
Dataset: Netflix

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?**

   Netflix.csv is a tabular database containing show id, type of show, director, cast, country, and other features about several types of entertainment shows/Movies in Netflix. Most of the features we encountered were string values. Only the release year was an integer value. The dataset comprises of 8807 rows and has 12 columns.

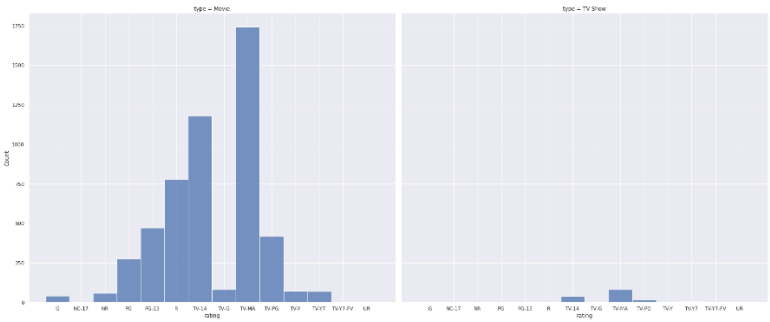
2. **Provide the main statistics about the entries of the dataset (mean, std, number of missing values, etc.)**





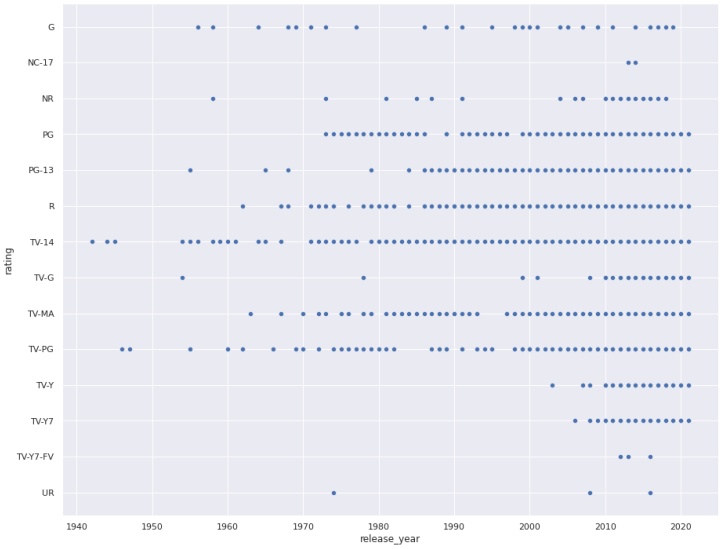
3. **Provide at least 5 visualization graphs with brief description for each graph, e.g. discuss if there are any interesting patterns or correlations.**

   1. In this frequency graph the number of movies and Tv-shows were plotted according to their rating. We can clearly state that the movies with the rating “TV-MA” and the tv shows with the same rating Is the one that most people watch.

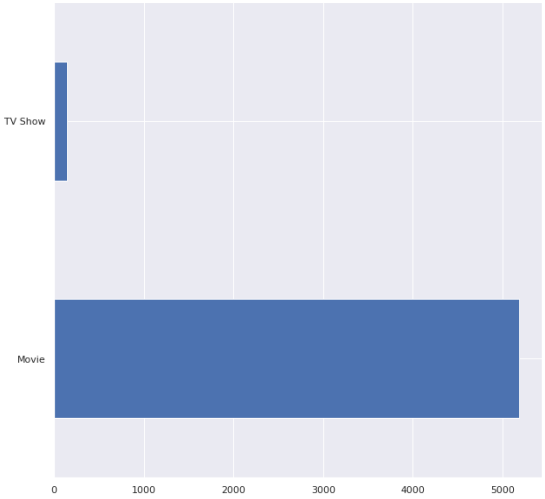


   2. In the scatter plot above we can compare movies of different ratings and their popularity over the years. The movies/shows with ratings “TV-14” and “TV-PG” are

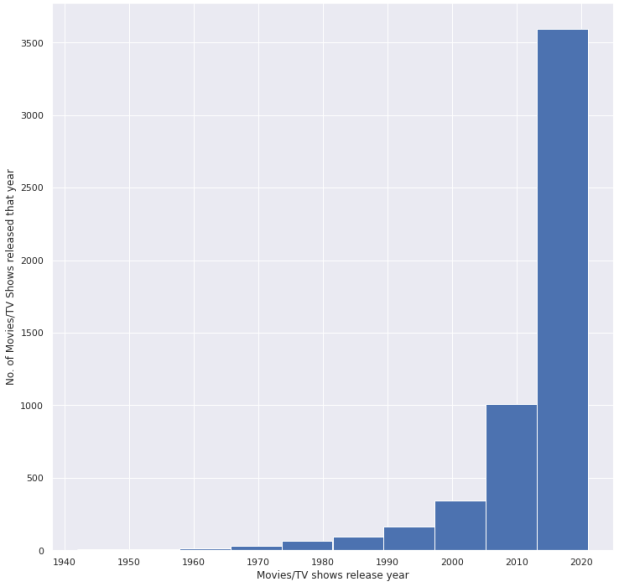
clearly the oldest and are still going strong. But movies/shows with ratings “NR” and “TV-G” started early but gained popularity only in recent years.



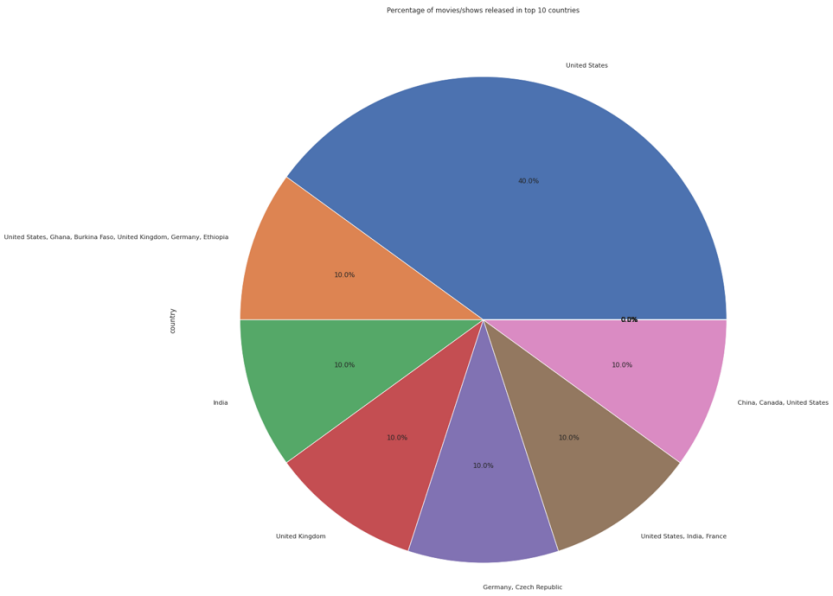
3. In this bar graph we can see the number of movies and the number of TV shows that were provided in the dataset. Clearly movies are produced more than TV-shows on Netflix in this data set.



4. In the above bar graph, from the number of Movies or TV-shows that were released from the years 1956-2020, the production of the entertainment field has seen significant growth in the 21st century, especially during the pandemic.



5. This pie chart tells us about top 10 countries in which the Netflix movies/shows were released in. We can say that the maximum no. of movies/shows were released in the United States of America.



Dataset: Amazon

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?**

   Amazon_top_selling_book.csv is a tabular database containing the name of books, authors, user ratings, reviews, year, price and genre about different books in Amazon. We encounter Integer (int64), Float (float64), and Object (object) data types. The data set has 550 rows and 7 columns.

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

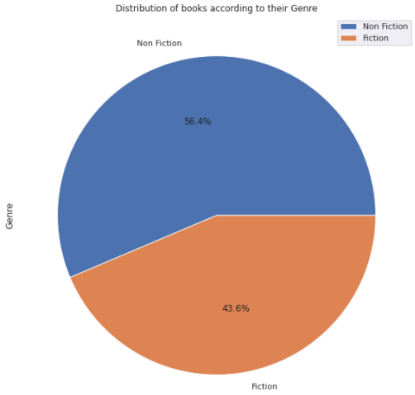
[550 rows x 7 columns]

```
amazondf.describe()
```

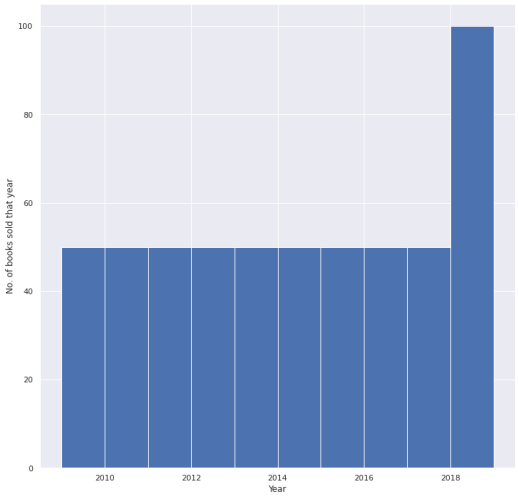
	User Rating	Reviews	Price	Year
count	550.000000	550.000000	550.000000	550.000000
mean	4.618364	11953.281818	13.100000	2014.000000
std	0.226980	11731.132017	10.842262	3.165156
min	3.300000	37.000000	0.000000	2009.000000
25%	4.500000	4058.000000	7.000000	2011.000000
50%	4.700000	8580.000000	11.000000	2014.000000
75%	4.800000	17253.250000	16.000000	2017.000000
max	4.900000	87841.000000	105.000000	2019.000000

No missing values

3. **Provide at least 5 visualization graphs with brief description for each graph, e.g. discuss if there are any interesting patterns or correlations.**
   1. In the above pie chart, we can see that the number of books sold on Amazon according to their genre is Fiction or Non-Fiction. We can infer that 56.4% of the books sold were from the Fiction category and 43.6% books sold were from the Non-Fiction category.

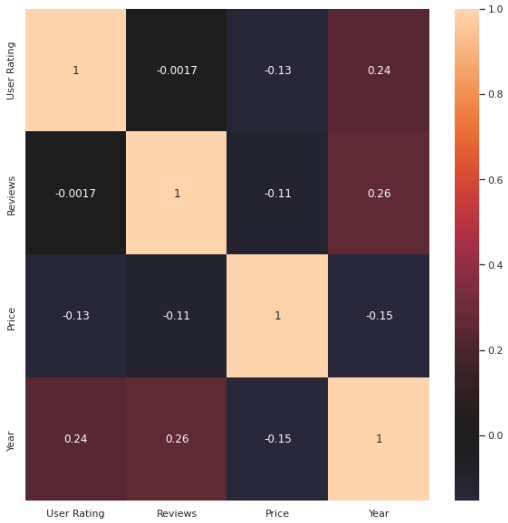


   2. In the above histogram we no. of books sold from years 2009 to 2019. The sales of books were consistent till the year 2018 and then went drastically up in the following year.

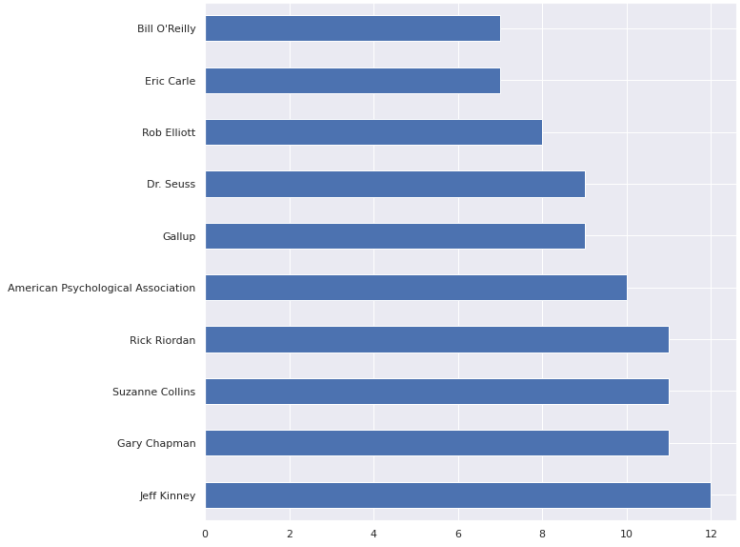


3. The books with ratings from 4.4 and above are the most selling books over the years. And there are few to none books below the rating of 4.0 over the years.

4. In the correlation graph we can see the relation of different columns with each other. We can say that the reviews are inversely proportional to Price of the book and the user rating is directly proportional to the year the book was sold.



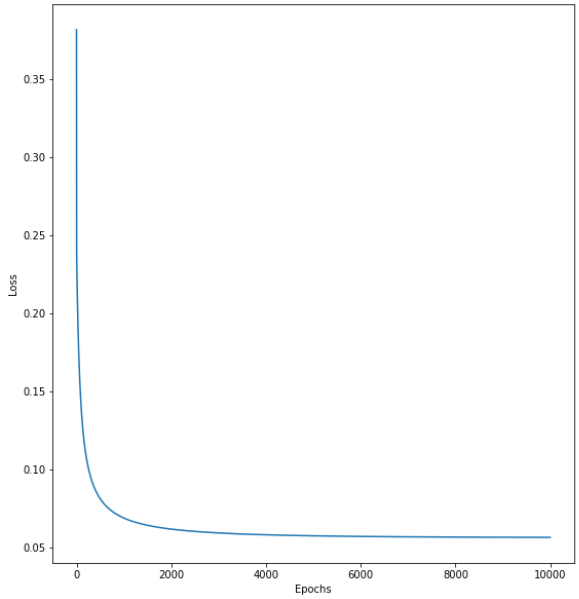
5. In this bar graph we plotted the Top 10 Authors whose books were sold the most on Amazon over the years. Author Jeff Kinney's books sold the most, that is 12 books.

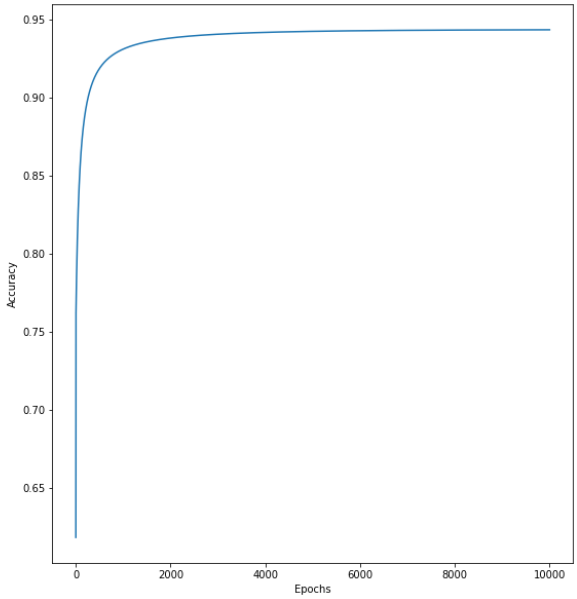


Part 2 - Logistic Regression

We analyzed the Palmer Archipelago (Antarctica) penguin dataset. The dataset was cleaned and preprocessed to start with so that good accuracy is achieved. The data was divided into two data sets: Train dataset and Test dataset. 80% of the data was in the Train dataset and 20% of the data was in test dataset.

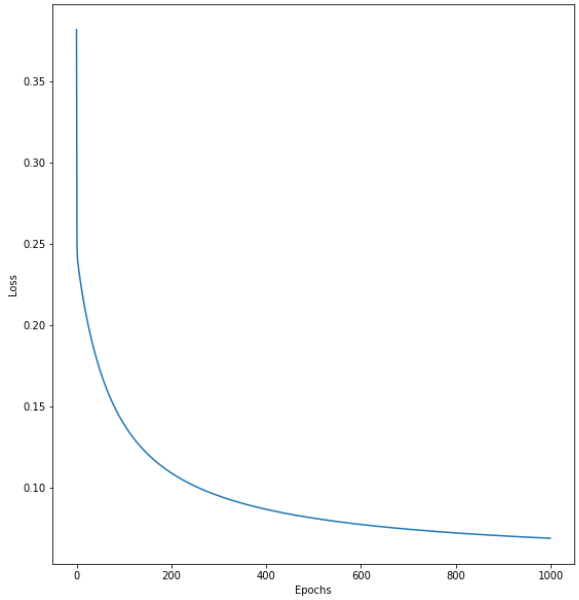
- Best Accuracy: **0.943622942995847**
- Weight Vector: **[-5.56283266, 0.80906772, -0.27478626, 0.78940926, 1.5142517, 13.33764845, 8.87860201, 3.50696969, 24.5948693, 4.23840994, 1.45322884]**

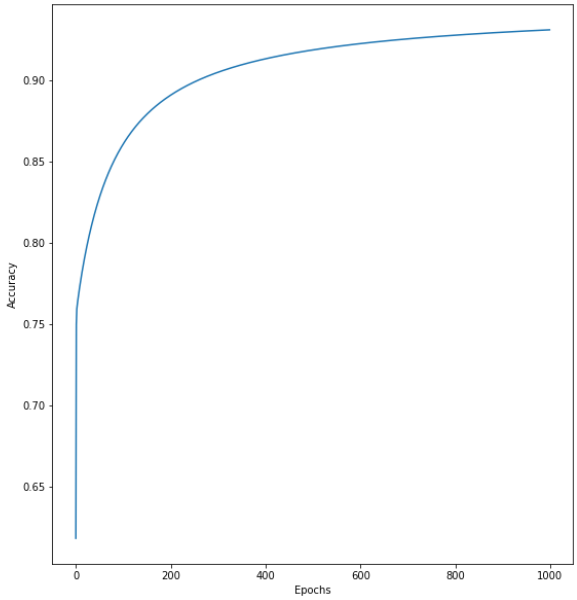




From the above Loss graph, it can be observed that as the iterations increase the weights and bias are updated and help decrease the loss.
As the loss keeps reducing it can be observed that accuracy gradually increases.

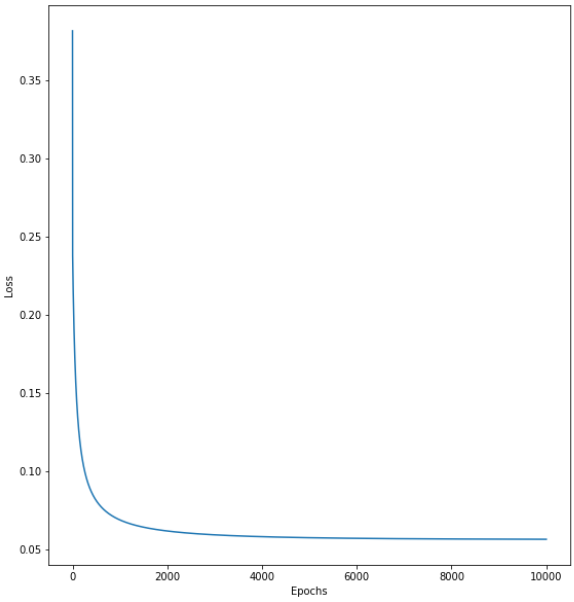
Hyperparameters
For 1000 iterations with Learning rate = 1e-9 (0.000000001)

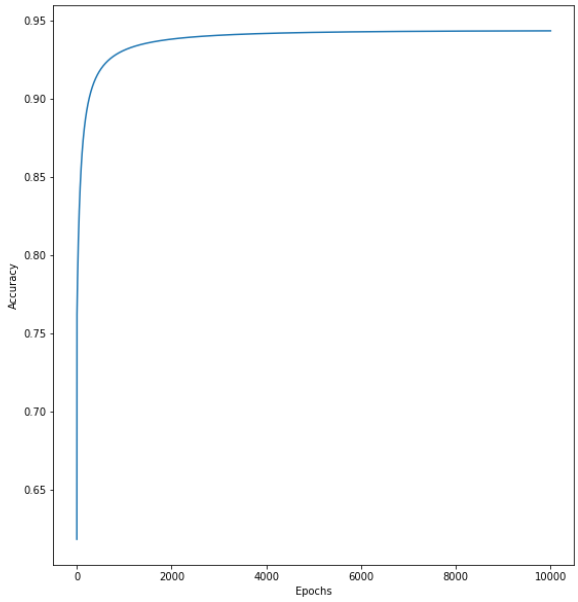




- Accuracy: **0.9312148146480524**
- Weight Vector: **[-2.89279009, 0.80906772, -0.46310694, 0.78940926, 0.0703079, 6.66899445, 5.65002224, 3.30558136, 10.91595706, 1.3072675, 1.0669664]**
  After tuning for the mentioned parameters, it can be noticed that the accuracy had reduced

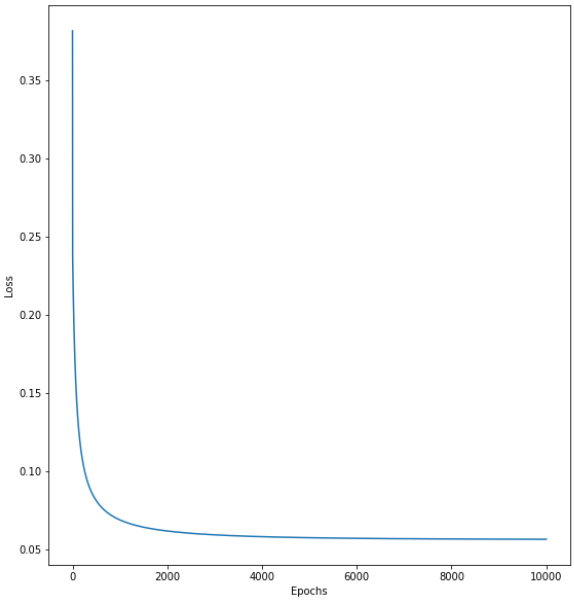
For 10,000 iterations with Learning rate = 1e-3 (0.001)

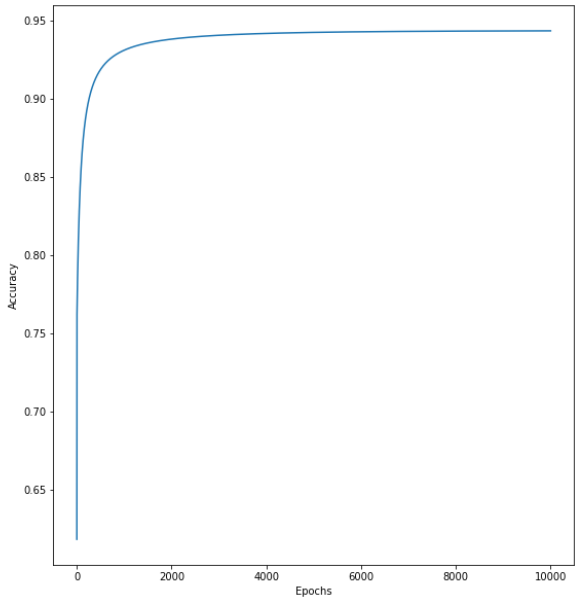




- Accuracy: **0.943622942995847**
- Weight Vector: **[-5.56283266 0.80906772 -0.27478626 0.78940926 1.5142517 13.33764845 8.87860201 3.50696969 24.5948693 4.23840994 1.45322884]**
  After tuning for the mentioned parameters, it can be noticed that the accuracy was the same as that of the initial case (1e-6)

For 10,000 iterations with Learning rate = 0.1





- Accuracy: **0.943622942995847**
- Weight Vector: **[-5.56283266 0.80906772 -0.27478626 0.78940926 1.5142517 13.33764845 8.87860201 3.50696969 24.5948693 4.23840994 1.45322884]**

After tuning for above mentioned parameters, it can be notices that the accuracy remained same.

Overall, it can be noticed that the accuracy was same for learning rate = 1+e-6, 1+e-3, 0.1 for 10,000 iterations. But when we changed the iterations to 1000 and learning rate to 1+e-9 the accuracy was decreased by a bit.

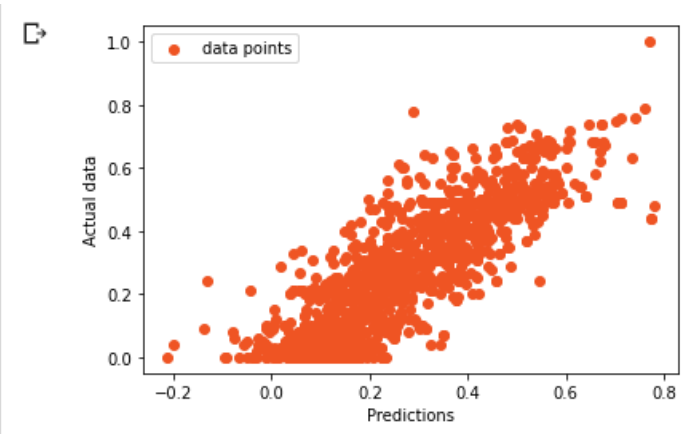
- Benefits/Drawbacks of Using Logistic Regression:

Benefits	Drawbacks
Easy to implement, interpret, and efficient to train	If the no. of observations is less than the no. of features, it leads to overfitting.
It is extremely fast at classifying unknown records.	Non-linear problems cannot be solved because Logistic Regression has a linear decision surface.
It can easily extend to multiple classes.	It constructs linear boundaries.

Part 3 - Linear Regression

For our linear regression model, we analyzed the Wine Quality – Red dataset. Initial preprocessing was done like the logistic regression model.

- Loss value - **0.012052148581248262**
- Weight Vector - **[0.71727429, -0.61855825, 0.08068472, 0.50558159, -0.1520093 0.38938336, 0.03548594, 0.02173854, 0.04710439, 0.21072576, 0.01535294]**
- Plot comparing the predictions vs the actual test data.



- The benefits/drawbacks for using OLS:
- The advantages are:

→ It is beneficial when we need to extract most extreme data from little information sets.

→ It is fast and simple to compute.
→ When the assumptions are met, it is more powerful than other regression methods.

- The disadvantages are:

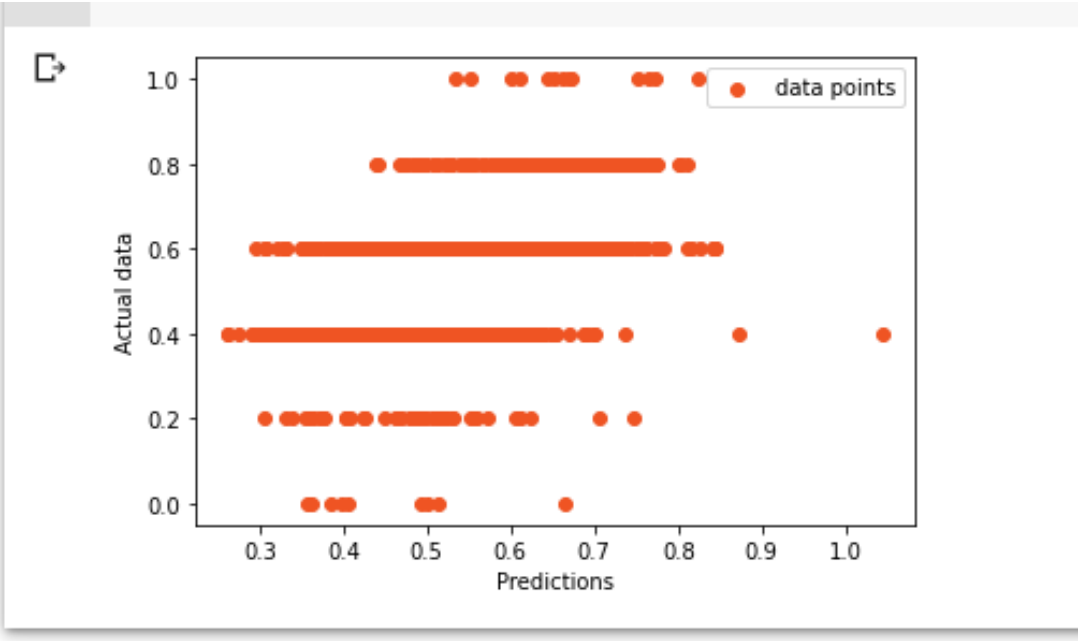
→ Exact weights are needed.
→ Estimating weights can give unpredictable results.
→ Perform badly on multivariate dataset.

Part 4 - Ridge Regression

For our ridge regression model, we analyzed the Wine Quality – Red dataset. Initial preprocessing was done like the linear regression model.

- Loss Value - **0.02040755885669038**
- Weight Vector - **[0.39396151, -0.14592444, 0.01064979, 0.04066423, 0.02722234, 0.09676928, -0.06223872, -0.01094704, 0.43242394, 0.38019763, 0.46203813]**

- Plot comparing the predictions vs the actual test data.



- Difference between Linear and Ridge Regression.

Linear Regression	Ridge Regression
There is one scalar target variable “y” and one vector input variable “x.”	Ridge helps to reduce impact of correlated inputs.

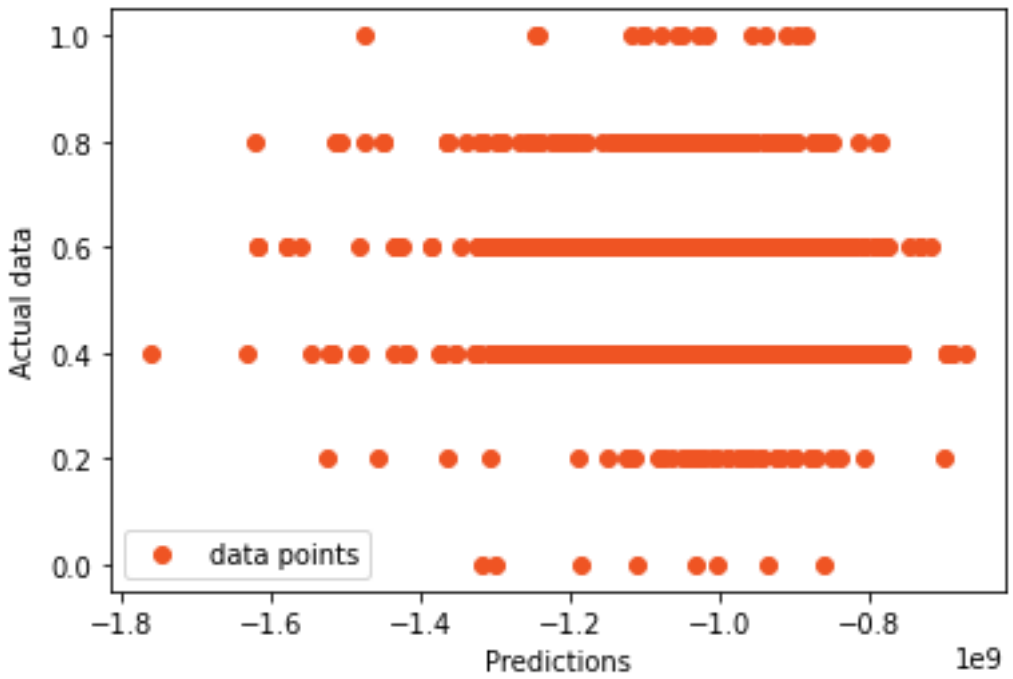
While using linear regression it tends to overfit more than ridge regression.	Ridge regression has regularization of data and prevents overfitting.
Linear regression is fast and easier to compute than Ridge regression.	In ridge regression we add a penalty term lambda in the computation of weight values.

The main motivation of using L2 regularisation is to reduce the chance of model overfitting.
It tunes the loss function by adding a penalty term, that prevents excessive fluctuation of the coefficients thereby reducing the overfitting.

Bonus Point

After implementing the gradient descent, we built from scratch, we noticed that its performance was not optimal the default case was performing better.

- Loss Value: **1.1176747831582724e+18**
- Weight vector: **[-3.76608770e+08, -3.07637411e+08, -3.15967320e+08, -1.30446945e+08, -1.42568445e+08, -2.38541090e+08, -1.62615106e+08, -5.55701397e+08, -4.92380048e+08, -2.24200452e+08, -3.45538020e+08]**
- Plot comparing the predictions vs the actual test data.



Contribution summary:

Team Member	Assignment Part	Contribution (%)
Pratik Malani	Part 1-4	50%
Babar Shaik	Part 1-4	50%

References:

- https://numpy.org/doc/
- https://datascience.stackexchange.com/questions/69661/difference-between-ridge-and-linear-regression#:~:text=Linear%20Regression%20establishes%20a%20relationship,also%20known%20as%20regression%20line).&text=Ridge%20Regression%20is%20a%20technique,independent%20variables%20are%20highly%20correlated).
- https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/
- https://en.wikipedia.org/wiki/Ridge_regression
- https://en.wikipedia.org/wiki/Logistic_regression
- https://en.wikipedia.org/wiki/Linear_regression