CSE 574: Introduction to Machine Learning
Spring 2022

ASSIGNMET-II

PART-I: Data Analysis

DATASET-1: Insaurance.CSV

DATA FRAME:

	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

Data Type of Dataset

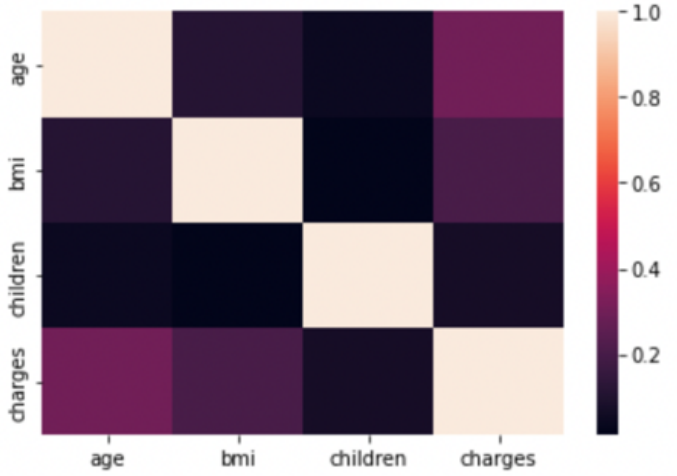
```
age          int64
sex         object
bmi        float64
children     int64
smoker      object
region      object
charges    float64
```

Information of Dataset

```
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
```

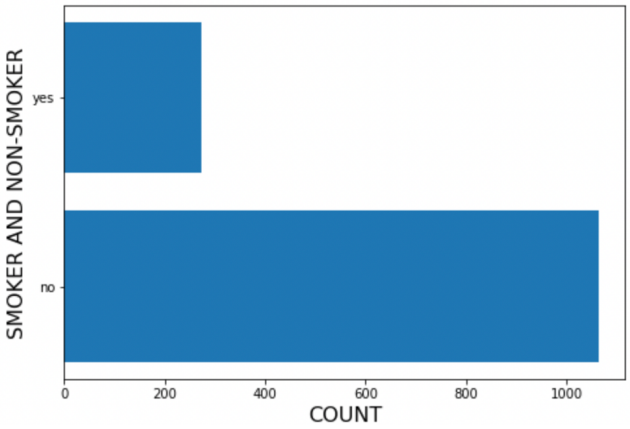
Statistics of The Dataset

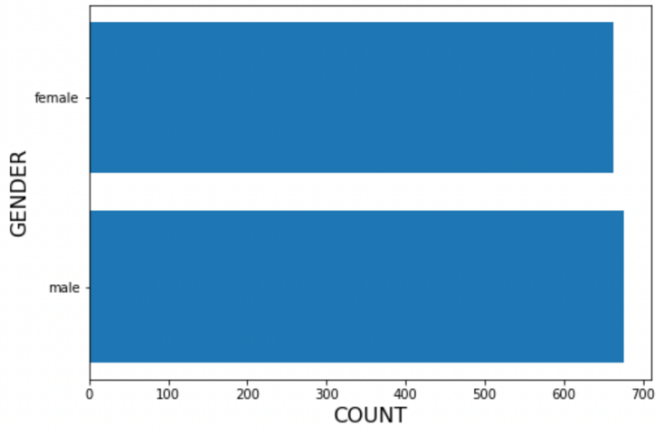
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000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



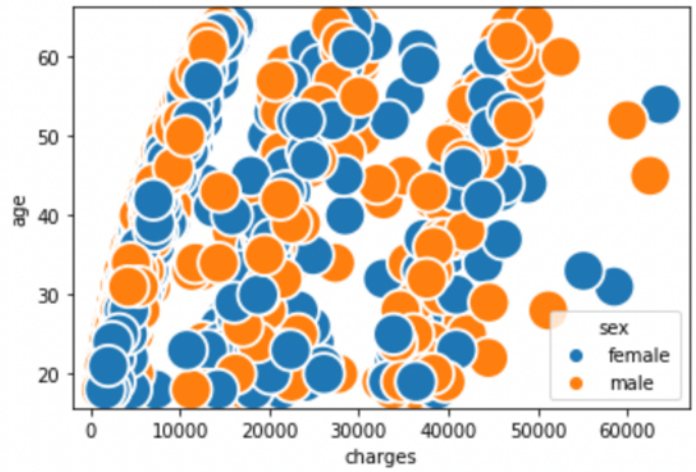
Correlation Matrix

Smoker VS Non-Smoker



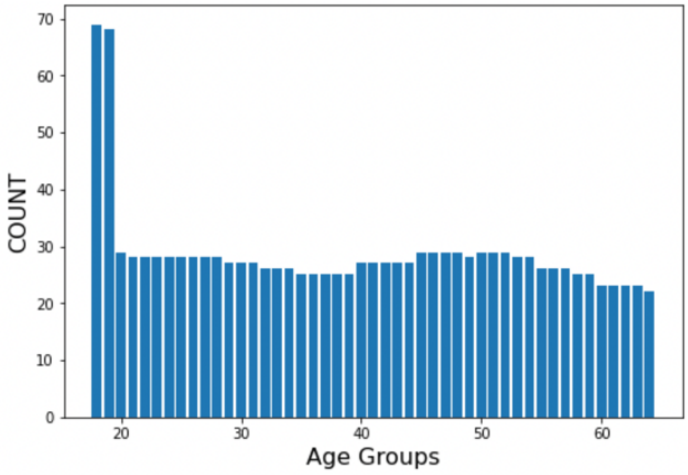


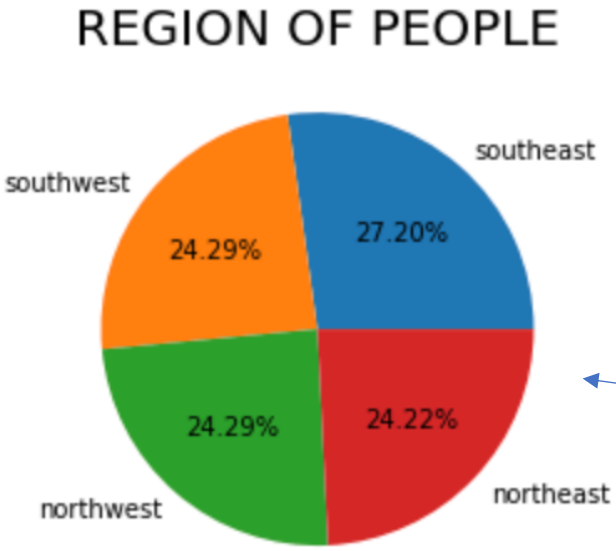
Gender Count



Scatter Plot
This plot helps us to understand the people from diverse age group and gender are paying what charges

Count VS Age Group
Count of People in different age Group





Region PIE
People belonging to different region ratio

DATASET-2: Titanic.CSV

DATA FRAME:

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	0	3	Mr. Owen Harris Braund	male	22.0	1	0	7.2500
1	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female	38.0	1	0	71.2833
2	1	3	Miss. Laina Heikkinen	female	26.0	0	0	7.9250
3	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35.0	1	0	53.1000
4	0	3	Mr. William Henry Allen	male	35.0	0	0	8.0500
..	...	...	...	...	...	...	...	...
882	0	2	Rev. Juozas Montvila	male	27.0	0	0	13.0000
883	1	1	Miss. Margaret Edith Graham	female	19.0	0	0	30.0000
884	0	3	Miss. Catherine Helen Johnston	female	7.0	1	2	23.4500
885	1	1	Mr. Karl Howell Behr	male	26.0	0	0	30.0000
886	0	3	Mr. Patrick Dooley	male	32.0	0	0	7.7500

Data Type of Dataset

```
Survived                    int64
Pclass                      int64
Name                       object
Sex                        object
Age                       float64
Siblings/Spouses Aboard     int64
Parents/Children Aboard     int64
Fare                      float64
dtype: object
```

Information of Dataset

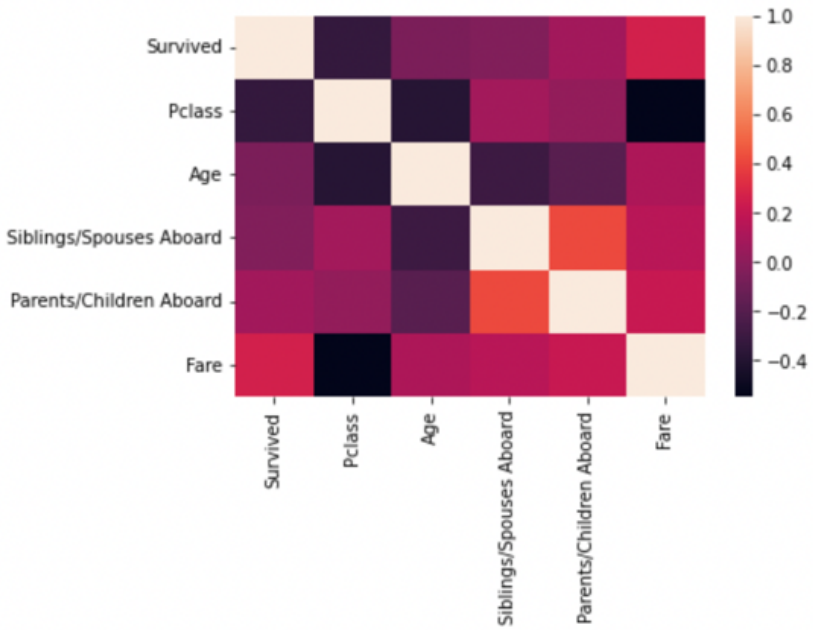
```
Data columns (total 8 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Survived                 887 non-null    int64
 1   Pclass                   887 non-null    int64
 2   Name                     887 non-null    object
 3   Sex                      887 non-null    object
 4   Age                      887 non-null    float64
 5   Siblings/Spouses Aboard  887 non-null    int64
 6   Parents/Children Aboard  887 non-null    int64
 7   Fare                     887 non-null    float64
dtypes: float64(2), int64(4), object(2)
memory usage: 55.6+ KB
```

Statistics of The Dataset

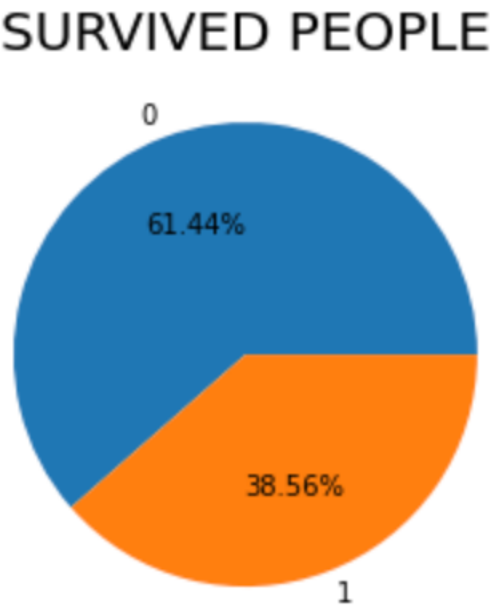
	Survived	Pclass	Age	Siblings/Spouses Aboard	\
count	887.000000	887.000000	887.000000	887.000000	
mean	0.385569	2.305524	29.471443	0.525366	
std	0.487004	0.836662	14.121908	1.104669	
min	0.000000	1.000000	0.420000	0.000000	
25%	0.000000	2.000000	20.250000	0.000000	
50%	0.000000	3.000000	28.000000	0.000000	
75%	1.000000	3.000000	38.000000	1.000000	
max	1.000000	3.000000	80.000000	8.000000	

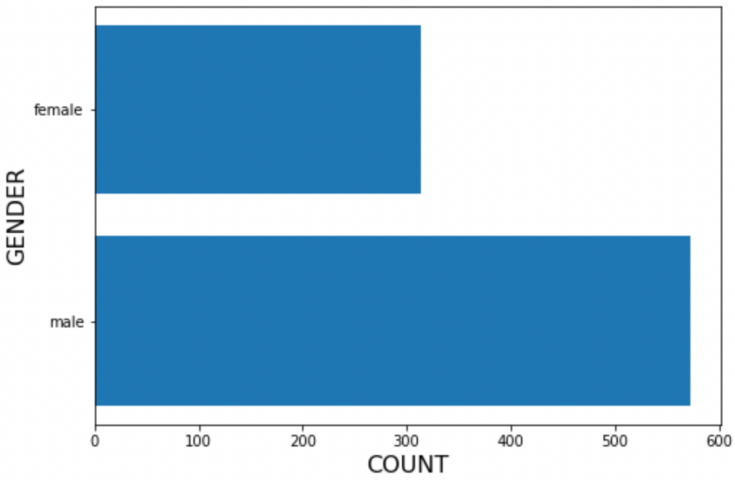
	Parents/Children Aboard	Fare
count	887.000000	887.00000
mean	0.383315	32.30542
std	0.807466	49.78204
min	0.000000	0.00000
25%	0.000000	7.92500
50%	0.000000	14.45420
75%	0.000000	31.13750
max	6.000000	512.32920

Correlation Matrix of Dataset



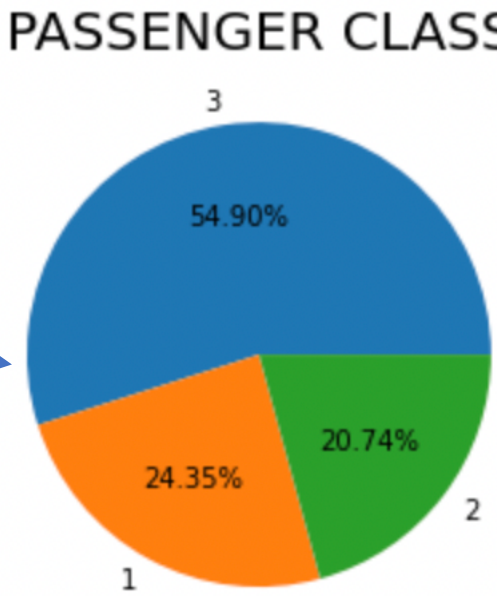
Survived People Ratio
64% people Died and
38.56% Survived





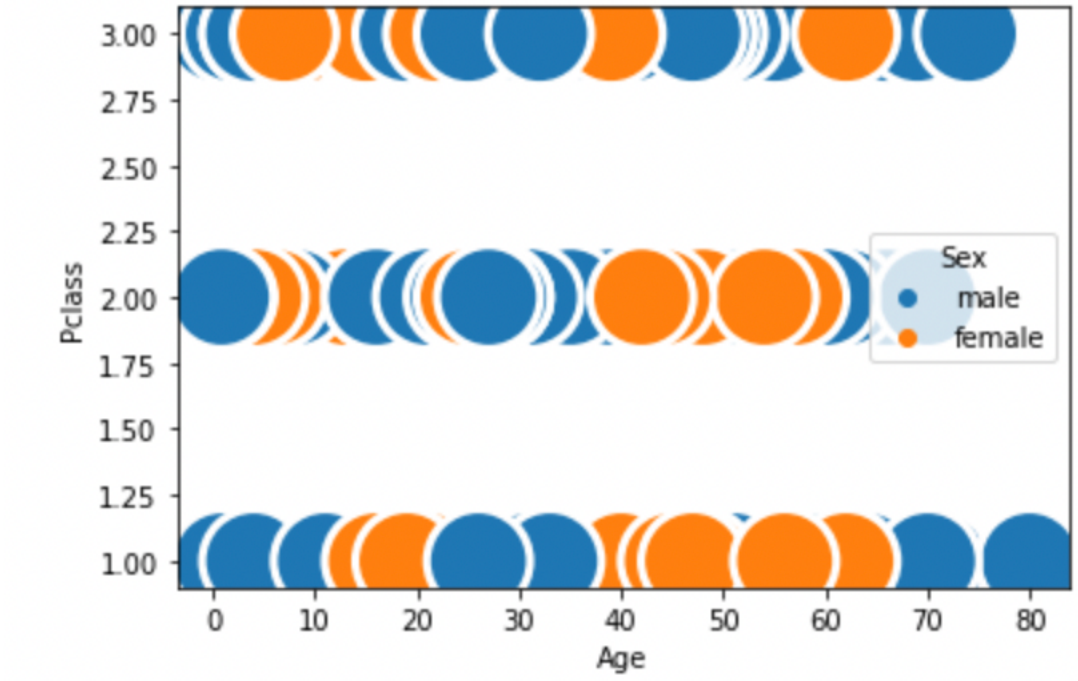
Gender VS COUNT

Passenger Class
In Titanic the percentage of people belonged to which class



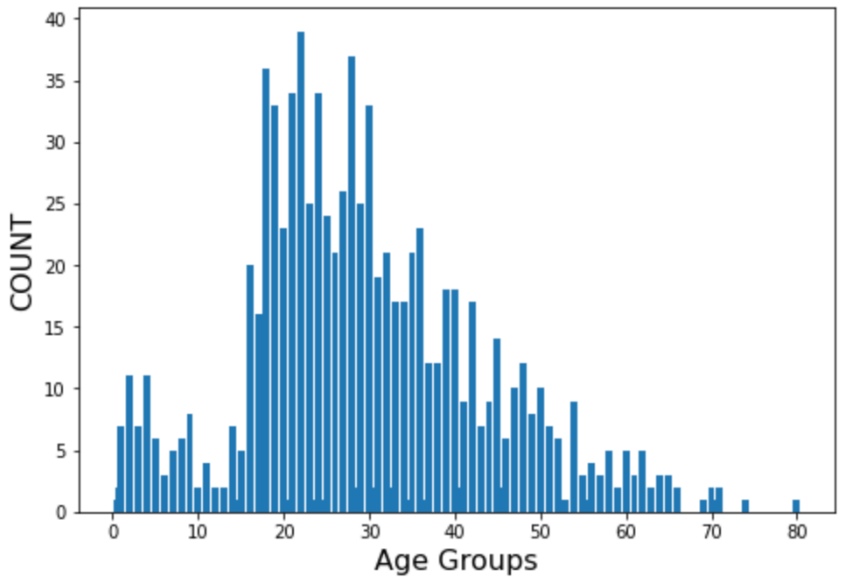
Clustering Passenger Class VS Age Group using Gender

The representation shows us the age group and gender preferred to travel using which class in the titanic



Age Group VS Count

The Graph helps us to understand the count of people in the age Group. The representation shows us the most people come under the 20-30 Age group.



DATASET-3: Amazon_top_selling_book.csv

DATA FRAME:

	Name	Author	User Rating	Reviews	Price	Year	Genre
0	10-Day Green Smoothie Cleanse	JJ Smith	4.7	17350	8	2016	Non Fiction
1	11/22/63: A Novel	Stephen King	4.6	2052	22	2011	Fiction
2	12 Rules for Life: An Antidote to Chaos	Jordan B. Peterson	4.7	18979	15	2018	Non Fiction
3	1984 (Signet Classics)	George Orwell	4.7	21424	6	2017	Fiction
4	5,000 Awesome Facts (About Everything!) (Natio...	National Geographic Kids	4.8	7665	12	2019	Non Fiction
..	...	...	...	...	...	...	...
545	Wrecking Ball (Diary of a Wimpy Kid Book 14)	Jeff Kinney	4.9	9413	8	2019	Fiction
546	You Are a Badass: How to Stop Doubting Your Gr...	Jen Sincero	4.7	14331	8	2016	Non Fiction
547	You Are a Badass: How to Stop Doubting Your Gr...	Jen Sincero	4.7	14331	8	2017	Non Fiction
548	You Are a Badass: How to Stop Doubting Your Gr...	Jen Sincero	4.7	14331	8	2018	Non Fiction
549	You Are a Badass: How to Stop Doubting Your Gr...	Jen Sincero	4.7	14331	8	2019	Non Fiction

Data Type of Dataset

```
Name           object
Author         object
User Rating   float64
Reviews         int64
Price           int64
Year            int64
Genre          object
dtype: object
```

Information of Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550 entries, 0 to 549
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Name         550 non-null    object
 1   Author       550 non-null    object
 2   User Rating  550 non-null    float64
 3   Reviews      550 non-null    int64
 4   Price        550 non-null    int64
 5   Year         550 non-null    int64
 6   Genre        550 non-null    object
dtypes: float64(1), int64(3), object(3)
memory usage: 30.2+ KB
```

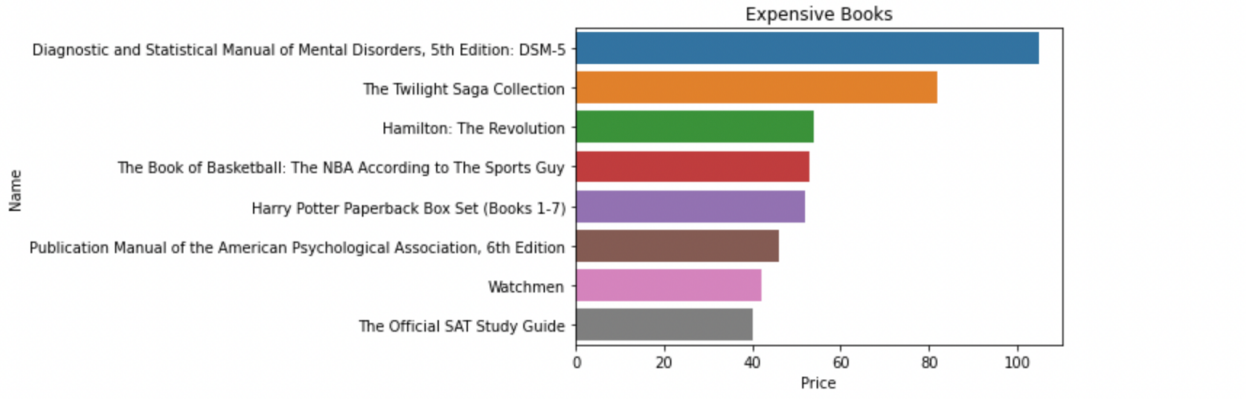
Statistics of The Dataset

	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

Correlation Matrix of Dataset

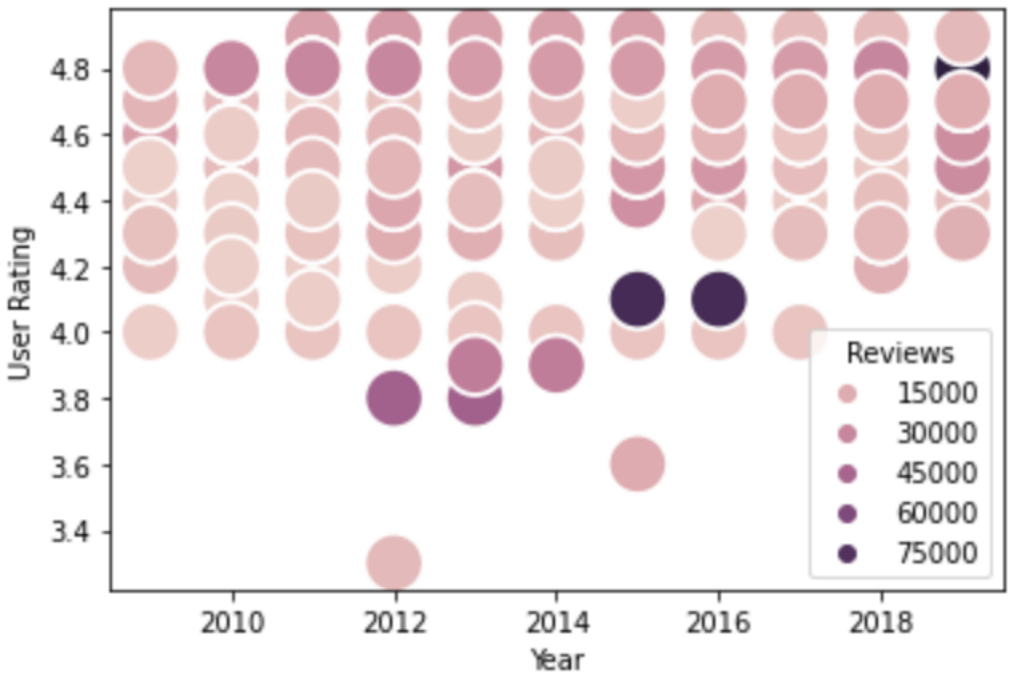
Expensive Books

The graph shows us the name of the most expensive book and coresponding cost .

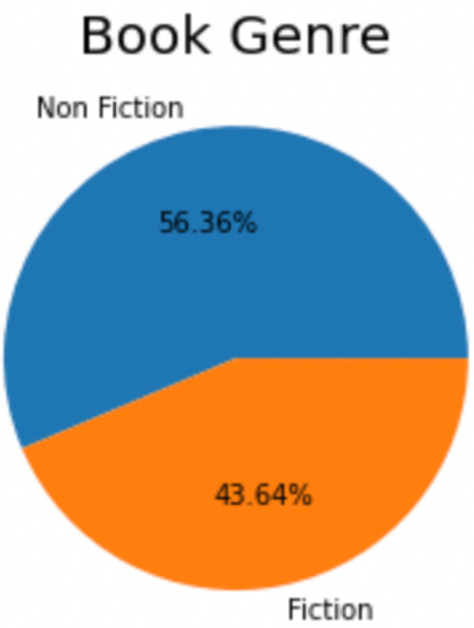


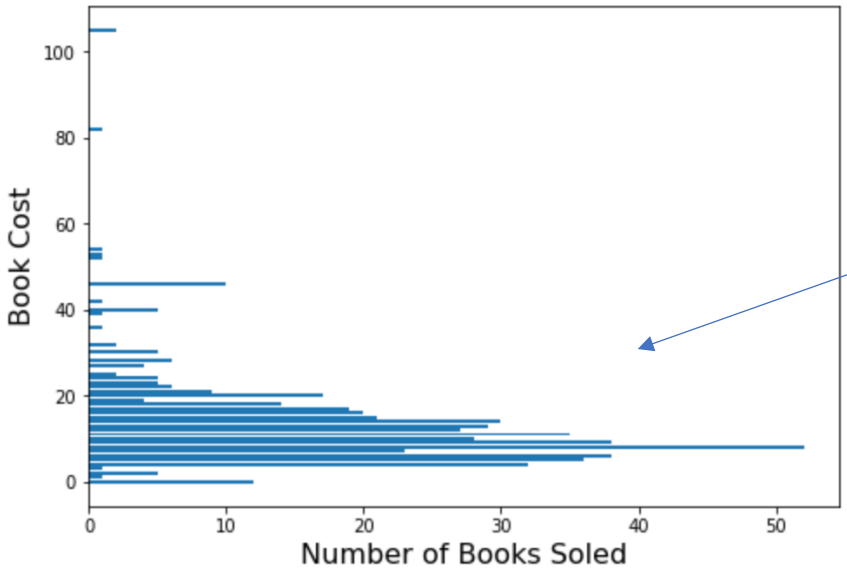
Cluster scatter Plot

The scatter plot shows the year and user rating also the number of reviews.
There were more than **75000** review in between **2014-2016** to the books and their average rating of the all the books were **4.0**



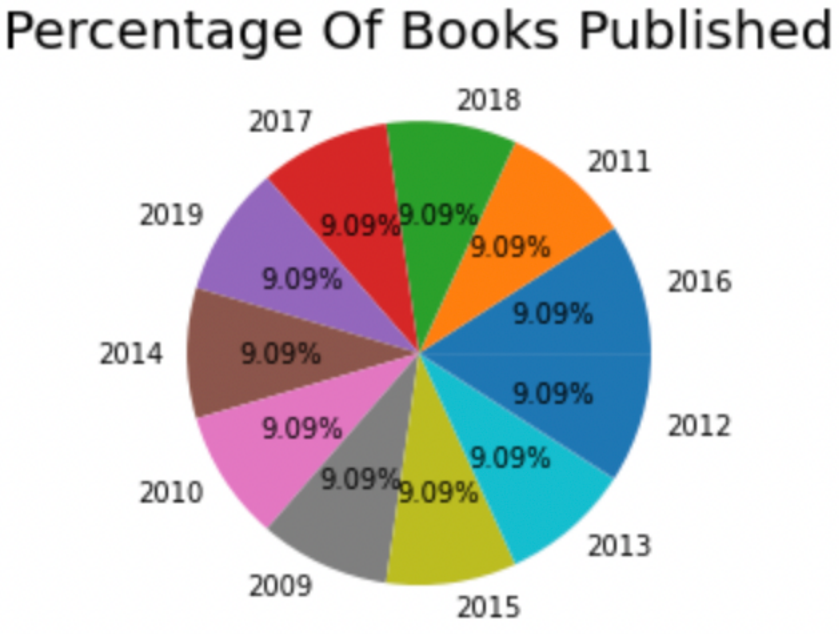
Book Genre Ratio
53.36% Non-Fiction Books and **43.64%** Fiction Books





Books soled Vs Cost
0-20 cost books **50** copy were sold the most

Survived People Ratio
64% people Died and **38.56%** Survived



Part II: Logistic Regression Report

Task 1:
Provide your best accuracy and Weight Vector.

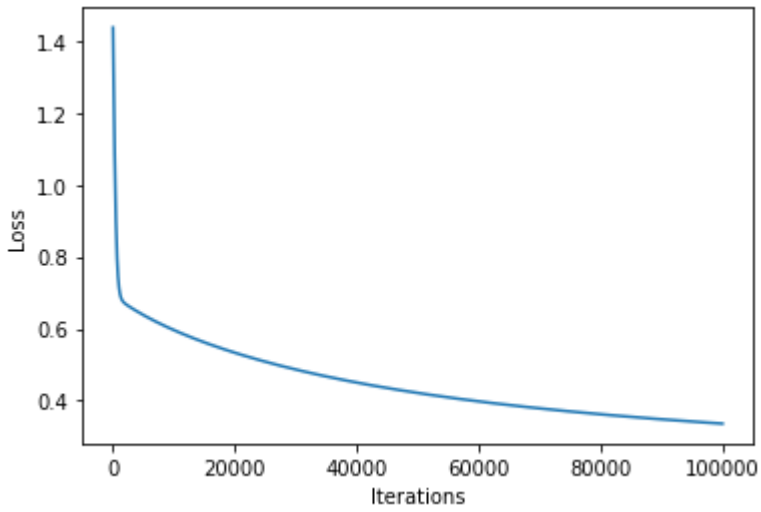
For Learning Rate = np.exp(-6), Iterations = 100000
Accuracy : 0.8955223880597015

Weight Vector :

For Learning Rate = np.exp(-6), Iterations = 100000
array([4.49147215, 1.53110854, 0.28126465, -2.99723003, -3.1685003 , -1.82750017, -5.8185686])

For Learning Rate = np.exp(-4), Iterations = 100000
array([11.38854931, 3.0876909 , 0.23140348, -8.46893043, -8.01121471, -2.17605089, -14.49528392])

Task 2:
Include Loss Graph and Short Description

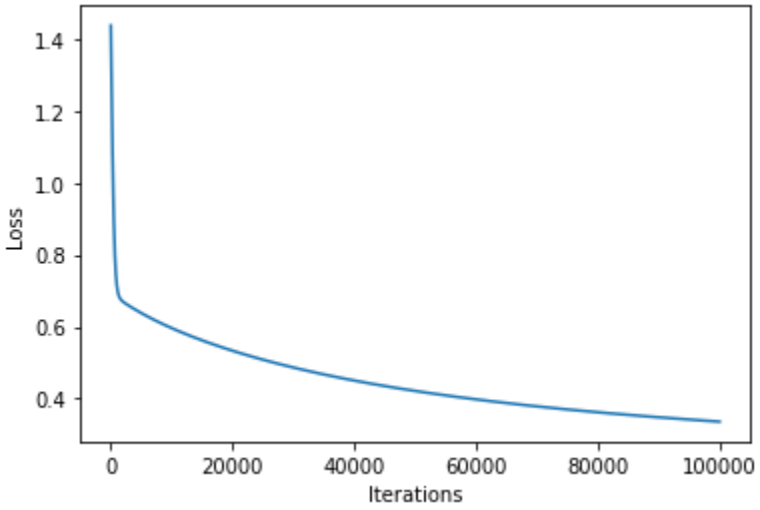


For Learning Rate = np.exp(-6), Iterations = 100000

As number of iterations increases the loss value decreases. In First few thousands(first 5 thousands) of iterations loss value is high, then it decreases gradually and for last iterations loss becomes 0.33636238.

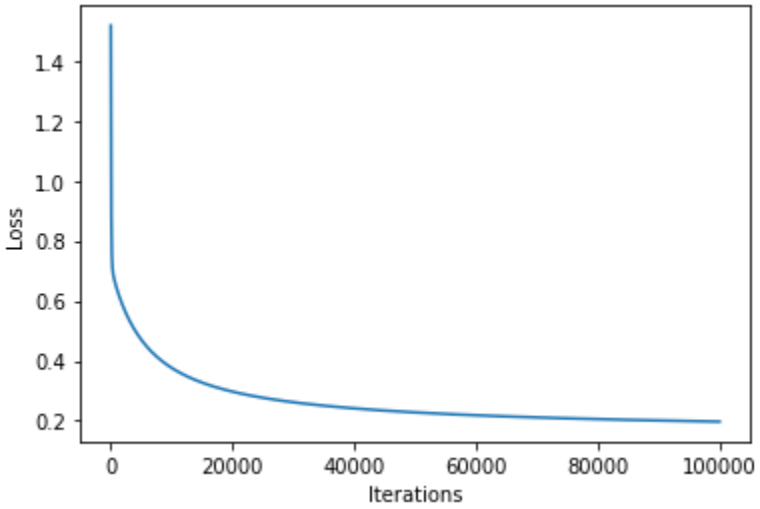
Task 3:

For Learning Rate = np.exp(-6), Iterations = 100000



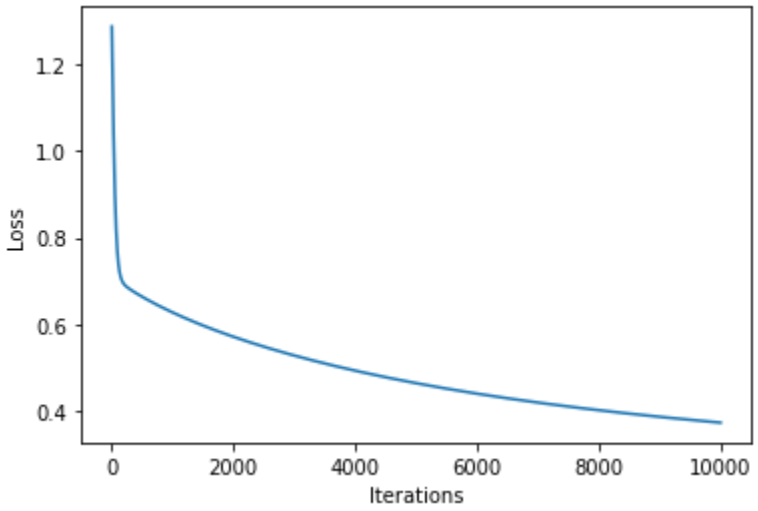
Loss for last iteration = 0.33636238

For Learning Rate = np.exp(-4), Iterations = 100000



Loss for last iteration = 0.19692161

For Learning Rate = np.exp(-4), Iterations = 10000



Loss for last iteration = 0.37301421

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

Benefits:

- Logistic regression is very straightforward to implement, interpret, and train.
- It classifies unfamiliar records fairly quickly.
- The overfitting of result is very low for a small data
- It's simple to build to a multi-class classification system.

Drawbacks:

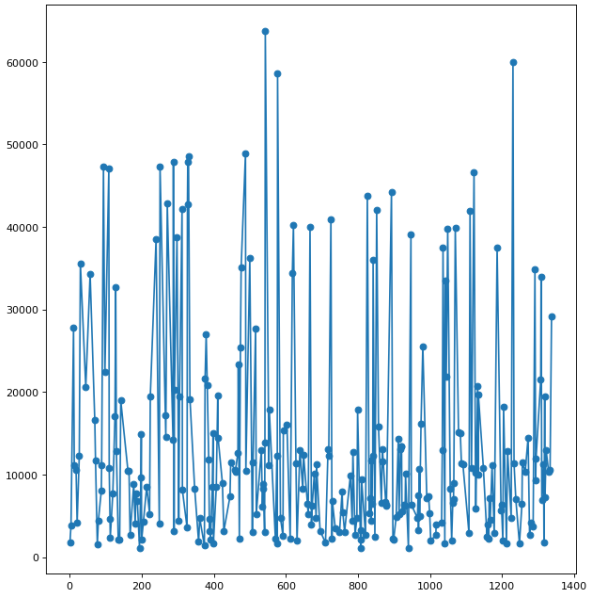
- It cannot solve Non-linear data problems
- It requires a large data with proper number of categories in feature and target set.
- It is very sensitive to the outliers
- Selection of feature is important or else the result would be affected badly.

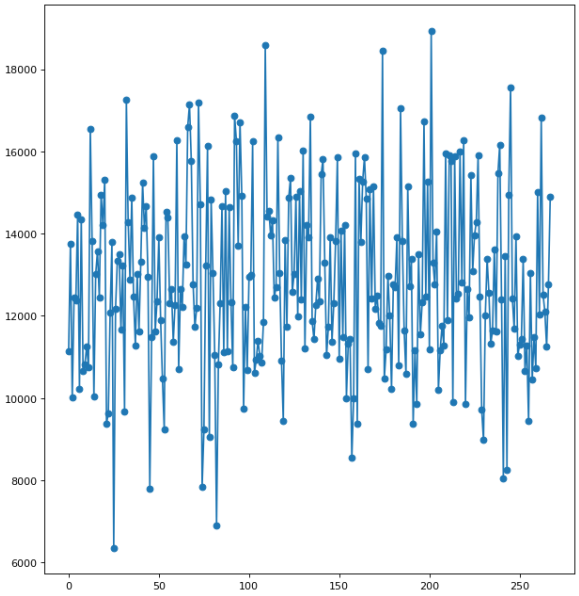
.

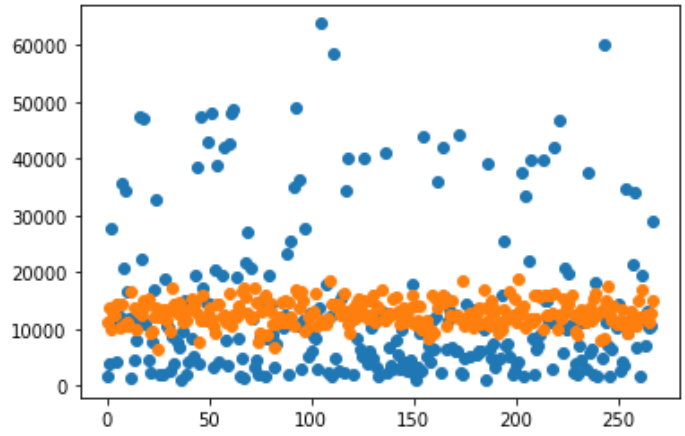
Part III: Linear Regression Report

Loss Value	1.830919e+08

Weight	5203.15482185	16285.67474828	1034.19545244







Advantages and Disadvantages of OLS(Ordinary Least Square):

Advantages:

- The OLS technique is simple to use and evaluate.
- It is simple to read due to the parametric form.
- The most important part and purpose is the reduction of errors.

Disadvantages:

- This method needs to be implemented on a generous size dataset for a better result.
- The Least squares performance suffers when outliers are present in data
- It also faces issues due to the non-linearity in the data
- In many cases over fitting of data is also seen.

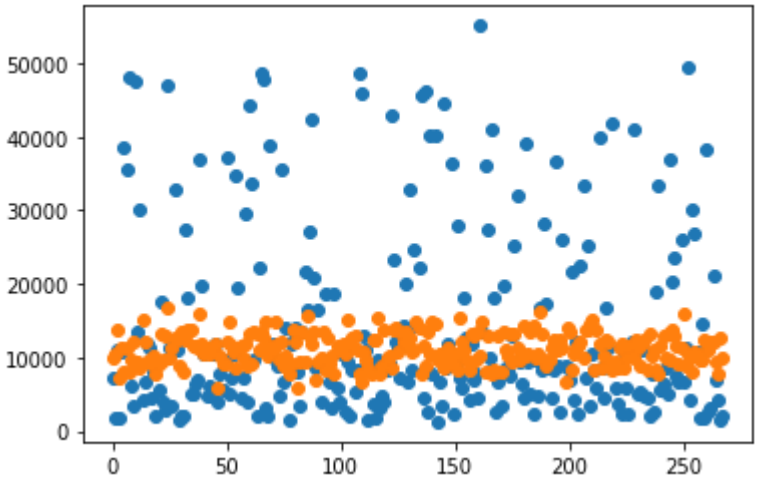
Part IV: Ridge Regression Report

1. Provide your loss value and the weight vector

Loss Value	1.2470971e+08

Weight Value	7880.85899725	10422.91112719	849.40667736

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



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

Ridge Regression gives more accurate result, in case of calculating loss, Ridge Regression perform better in case of decreasing the loss and increasing Accuracy.

Ridge has one extra parameter as lambda, which is added in case of calculating weight and loss. So, it gives more accurate Values.

As per my observation:

Loss in case of Ridge	1.2470971e+08
Loss in case of Linear	1.803527e+08

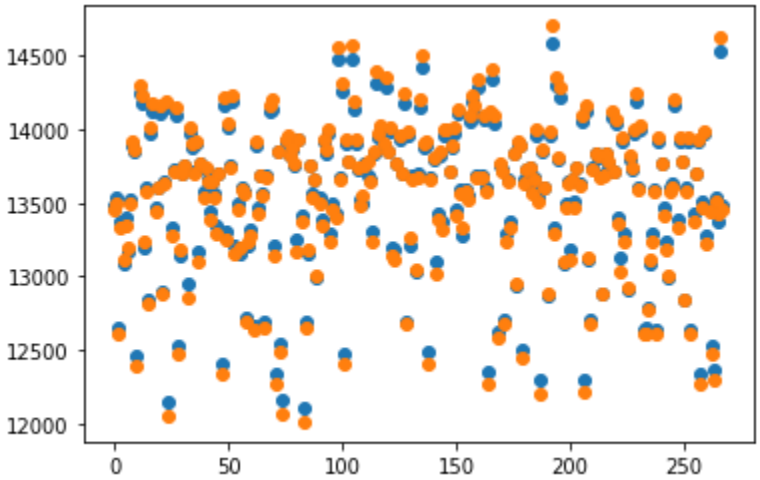
Part V: BONUS PART

From Scratch Gradient Descent Implementation:

Testing Squared Loss : 71648846.05721743
Train Squared Loss: 74374720.96260607
Time to Train Model : 6.515612602233887

From SK Learn Implementation:

Testing Squared Loss : 71747207.82940386
Train Squared Loss: 74234137.47551335
Time to Train Model : 0.010349750518798828



Graph for Predicted Output Implemented from Scratch vs SKlearn

Color Classification:
Blue dots: SKLearn Predicted Output Data
Orange: From Scratch Predicted Output

Axis Classification:
x: Number of outputs
y: Output Data

The results of the model generated from the scratch is producing outputs which are similar to the one generated using the built in sklearn method.The time required for the model implemented from scratch is taking a bit more compared to sklearn.