Team 98

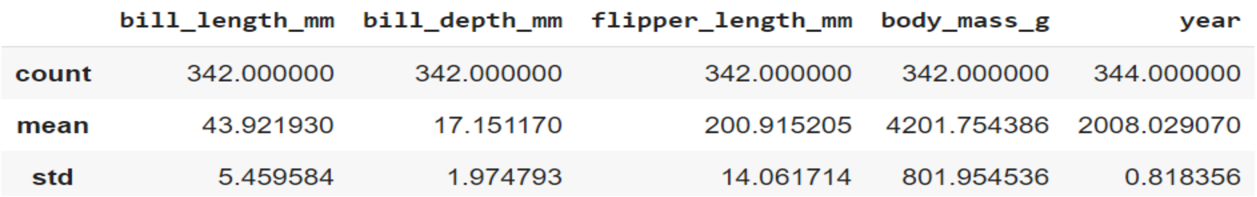
REPORT

Penguin Dataset

The penguin dataset shows the different attributes and features of the different species of penguins.

To understand more about the data, we have used the pandas built in functions.

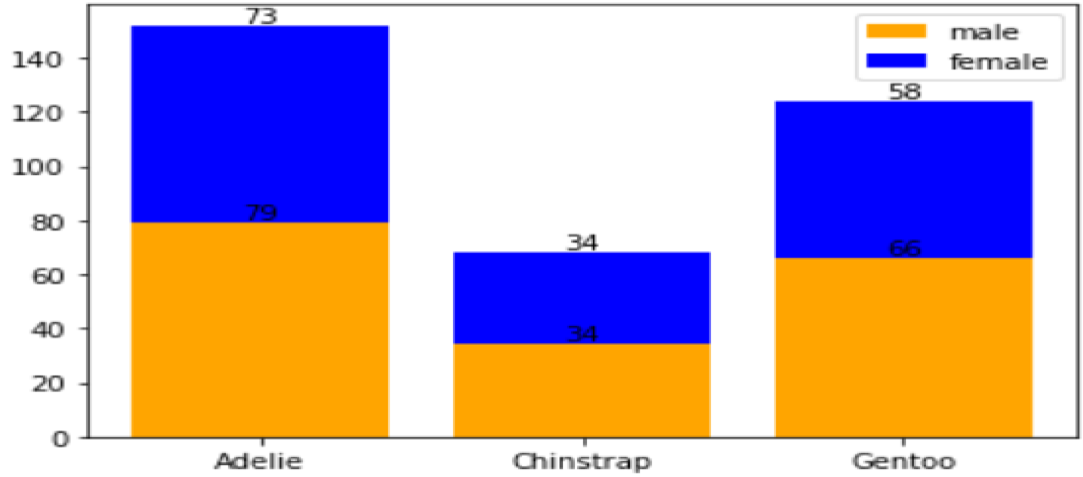
1. There are 8 columns namely 'species', 'island', 'bill_length_mm', 'bill_depth_mm', 'flipper_length_mm', 'body_mass_g', 'sex', 'year' and 344 rows
2. To clean the dataset, all NaN values have been replaced by the most frequent value of each respective column
3. There are three main species: 'Adelie', 'Chinstrap', 'Gentoo' living in three islands: 'Biscoe' , 'Dream' , 'Torgersen'
4. The data is a mix of string , float and integer datatypes.
5.



	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	year
count	342.000000	342.000000	342.000000	342.000000	344.000000
mean	43.921930	17.151170	200.915205	4201.754386	2008.029070
std	5.459584	1.974793	14.061714	801.954536	0.818356

1.

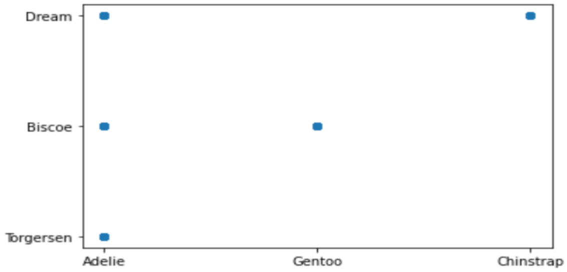
From this graph, we can see that Adelie and Gentoo has more males than females, Chinstrap has equal males and females Adelie has the most population, coming in second is Gentoo, and Chinstrap has the least amount of population



2.

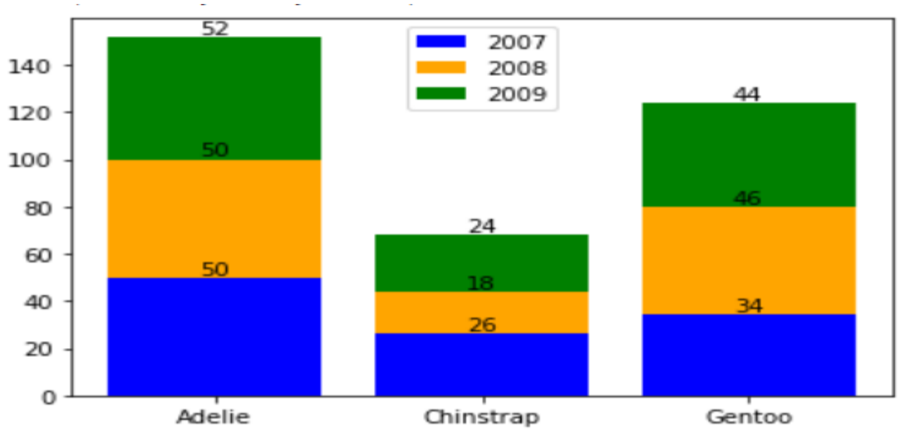
The species Adelie occupies all three islands, Gentoo only lives in Biscoe and Chinstrap only lives in Dream Island

Team 98



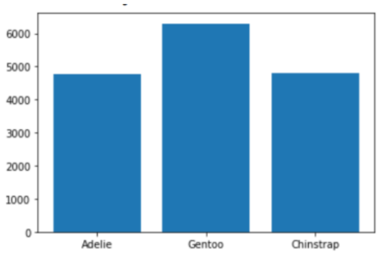
3.

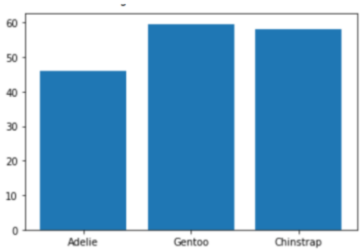
On the penguins vs year graph, we can see that in year 2009, Adelie penguins are the highest, for chinstrap, year 2007 is the highest, for Gentoo penguins, the year 2008 has the highest birth rate.

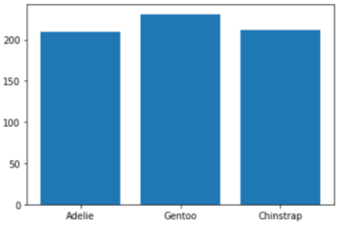


4.

Gentoo has the highest body mass, flipper length, and bill length, Adelie has the highest bill depth.

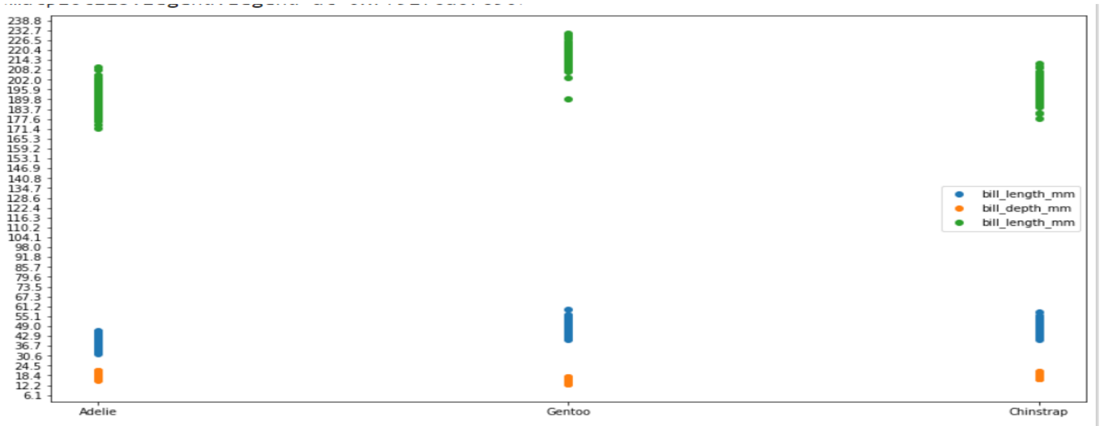






5.

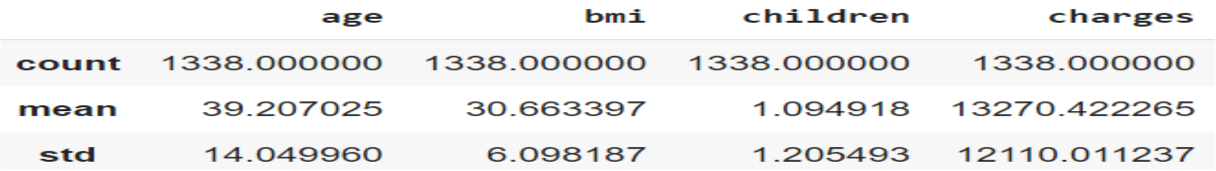
From this graph, we can see that bill depth ranges from 10mm to 25mm, bill length ranges from 30mm to 65mm, bill length ranges from 170mm to 250mm for all 3 penguin species



Report – Insurance Dataset

The Insurance dataset shows how the different factors (region, smoker/non smoker, age, sex, bmi, number of children) affect the price of the insurance.

1. There are seven columns and 1338 rows.
2.

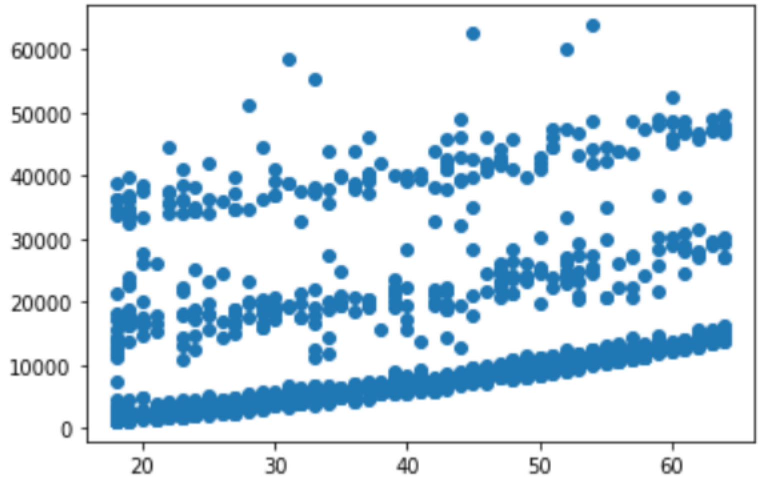


	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237

3. The data set is a combination of float, string and integer datatypes.
4. The total number of regions are 'northeast', 'northwest', 'southeast' ,'southwest'
5. The data does not have any missing values.

1.

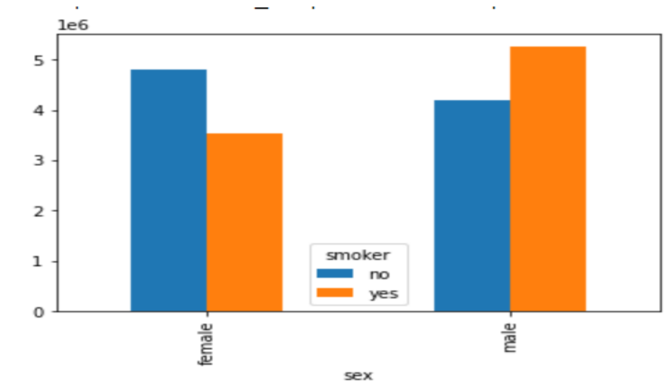
Here we can observe that as age increases(there is a very slight positive correlation), insurance charges are more.



Team 98

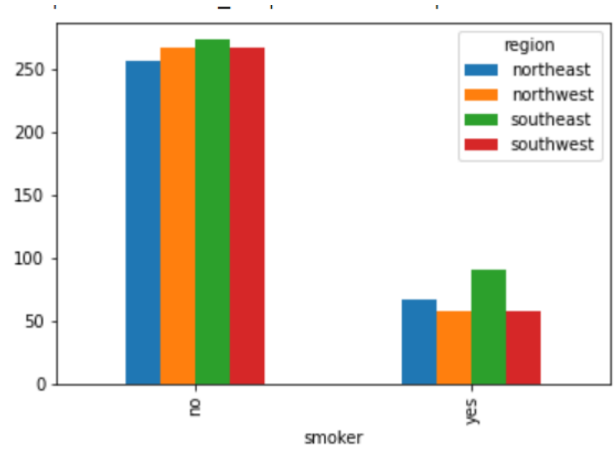
2.

For male, smokers have higher charges for insurance than non smokers, we can observe it is the opposite for female



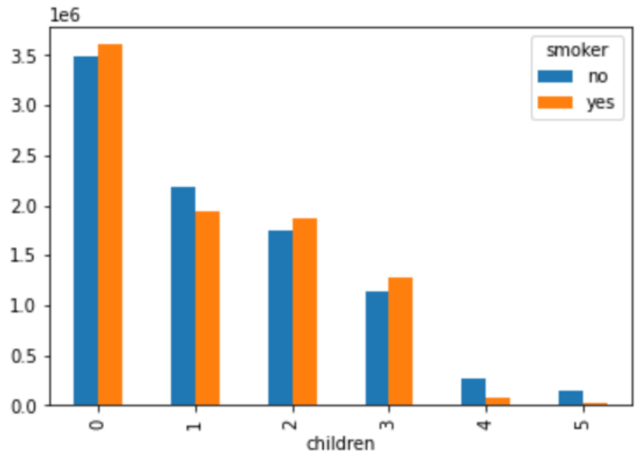
3.

From this bar graph, we can see that the majority are smokers, and the most smokers are from souhteast region.



4.

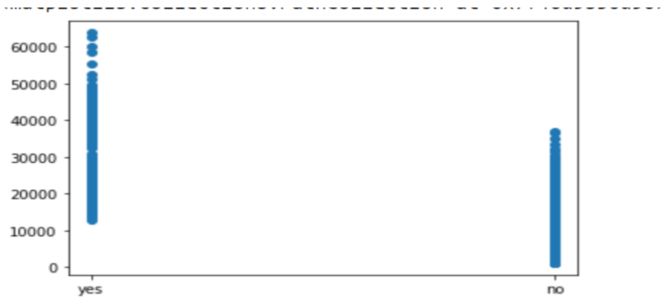
We can see a negative correlation; the highest charges are incurred for people with zero children



Team 98

5.

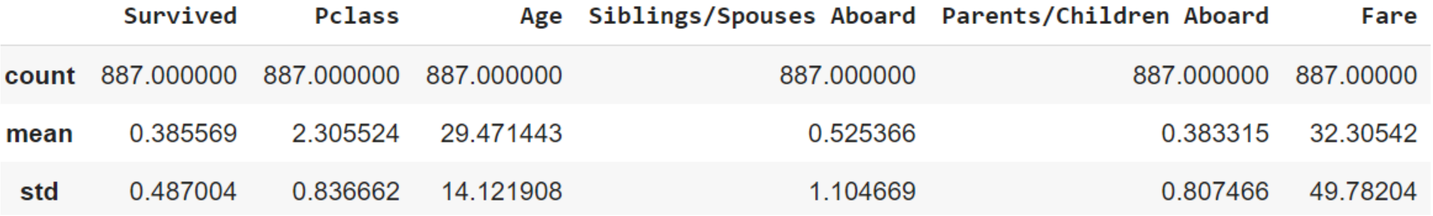
We can infer from this scatter plot that the insurance charges are significantly more for smokers when compared to non s mokers.



Report – Titanic Dataset

The Titanic dataset contains details of the different attributes of the people who have survived or not survived.

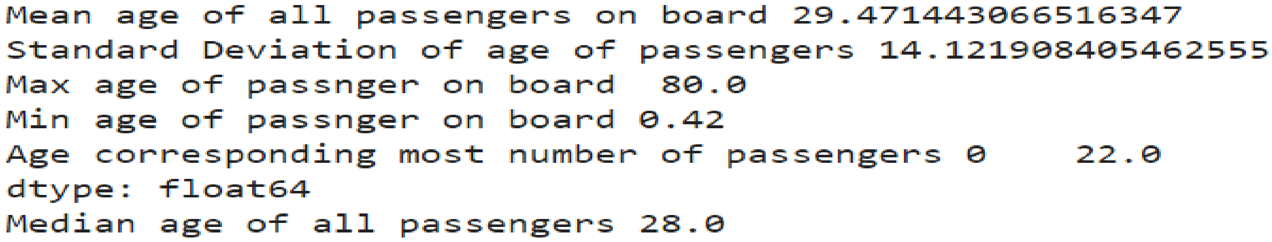
1. The data contains 887 rows and 8 columns.



	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

2.

3. There are no duplicates



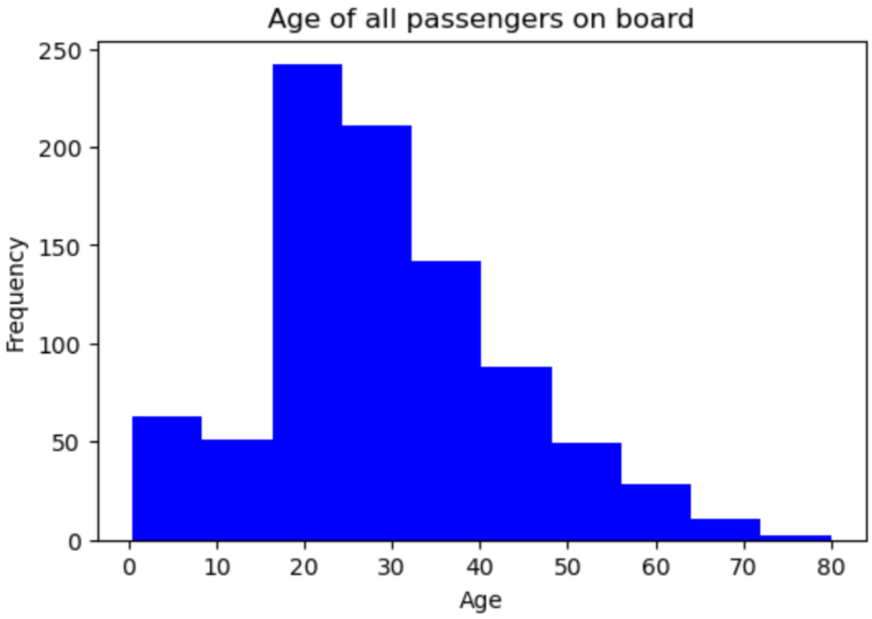
```
Mean age of all passengers on board 29.471443066516347
Standard Deviation of age of passengers 14.121908405462555
Max age of passnger on board  80.0
Min age of passnger on board 0.42
Age corresponding most number of passengers 0    22.0
dtype: float64
Median age of all passengers 28.0
```

4.

1.

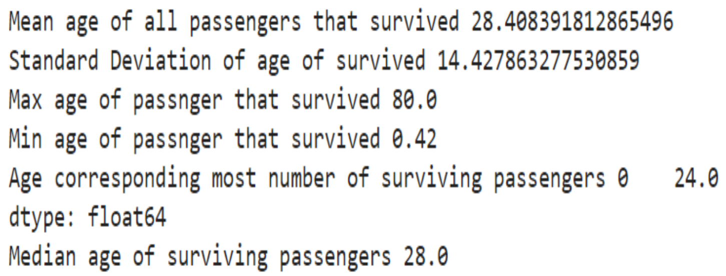
The frequency of passengers are highest between the age of 18 to 35.

Team 98

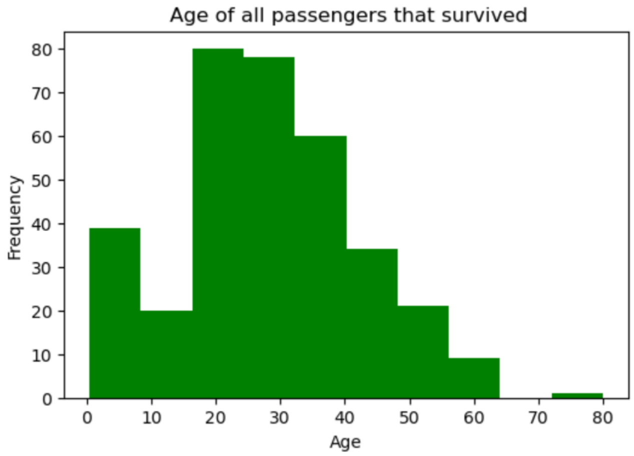


2.

The frequency of passengers that survived are highest between the age of 18 to 35.



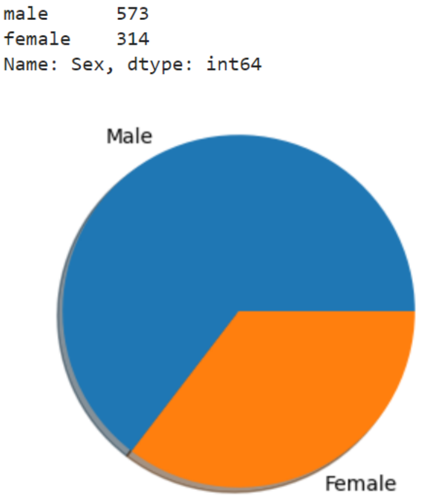
```
Mean age of all passengers that survived 28.408391812865496
Standard Deviation of age of survived 14.427863277530859
Max age of passnger that survived 80.0
Min age of passnger that survived 0.42
Age corresponding most number of surviving passengers 0    24.0
dtype: float64
Median age of surviving passengers 28.0
```



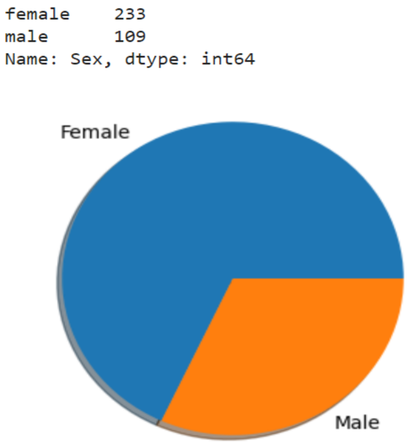
3.

The total number of males are 573, females are 314 in the ship. Out of that, 233 females and 109 males survived.

Team 98



```
male      573
female    314
Name: Sex, dtype: int64
```

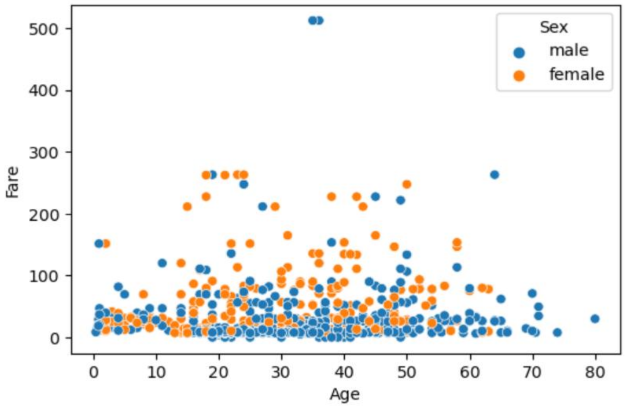


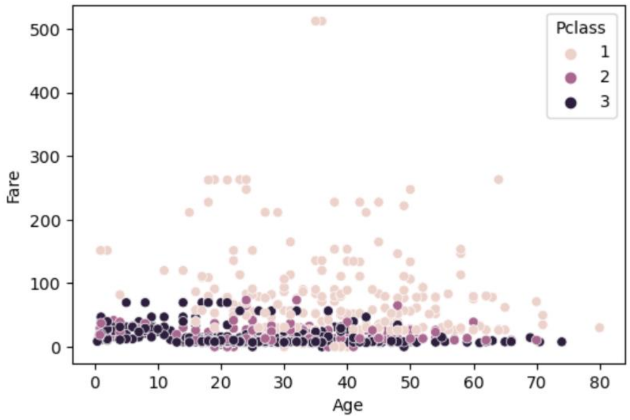
```
female    233
male      109
Name: Sex, dtype: int64
```

4.

From the age vs fare rates scatter splot, we can infer that males between the age of 30 and 40 spent the most on the fare price. The Passenger class has three classes namely class 1, class 2 and class 3. It is evident that class 1 has the highest fare rates, followed by class 2, then class 3.

It is also observed that ticket prices are lower for older age groups.





Team 98

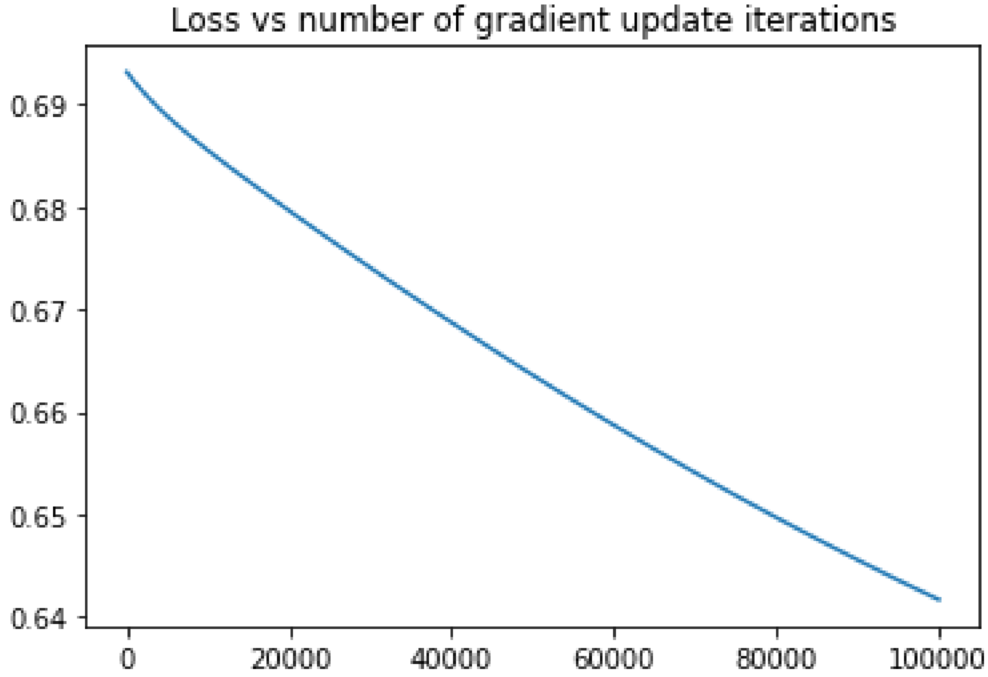
Part II: Logistic Regression

Methodology: Data preprocessing and data cleaning is similar to report of part I penguin dataset.
1. First we normalize the data to scale down the values between 0 and 1. To normalize the data without the normalize function, we use the formula: zi = (xi – min(x)) / (max(x) – min(x))
2. We take the column “sex” as the target variable Y. We form a N x d matrix for the input parameters for X. The input parameters we take into consideration are bill_length_mm, bill_depth_mm, flipper_length_mm and body_mass_g.
3. After shuffling the data, the data is split into training data (80%) and test data (20%) . Due to the shuffling and not seeding fix training and test data we get slightly different accuracy values on each run.

1. **Loss Value :**
   Weight Vector: ['bill_length_mm'],['body_mass_g'],['bill_depth_mm'],['bill_depth_mm']]

   Weights after training **[0.28498529 0.46680545 0.30475668 0.30475668]**

2.



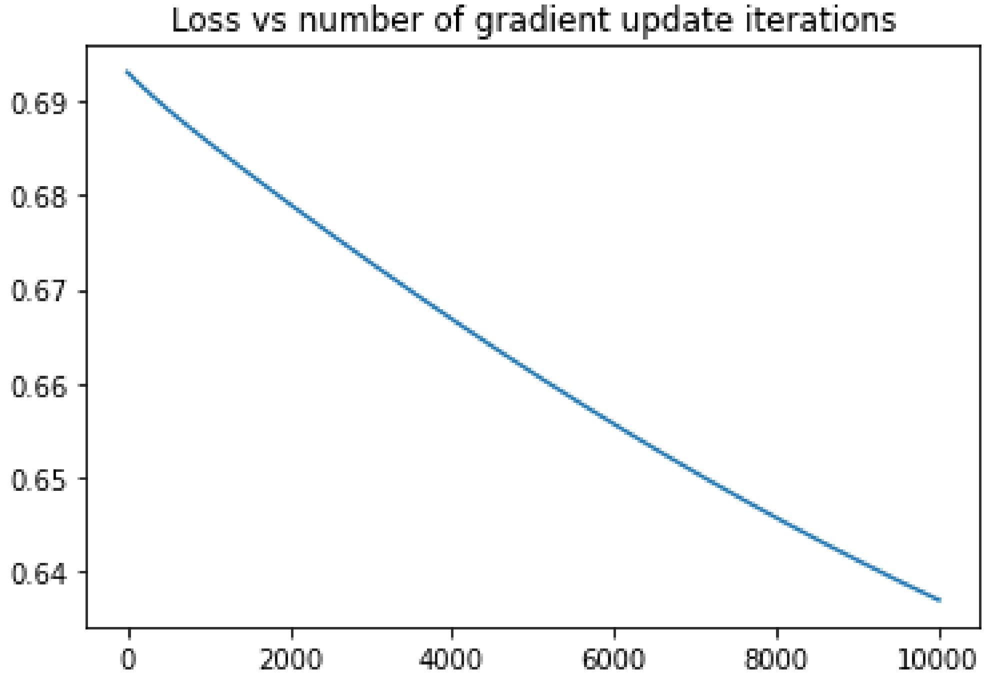
We see that the logistic log loss decreases exponentially in the beginning and then decreases approximately linearly in the asymptotic limit of number of iterations .

Accuracy is 73.13432835820896 %

3. **Different Hyperparameters tuning i.e. changing the values of the learning rate and the number of iterations.**

- **Increasing the learning rate and reducing the number of iterations**

The test accuracy increases by 3% which means that we reach to our local minima of the loss function more closely and faster this as we have increased our learning rate by a factor of 10 and at the same time we have reduced the iterations so we do not overshoot around the actual minima and keep on oscillating.



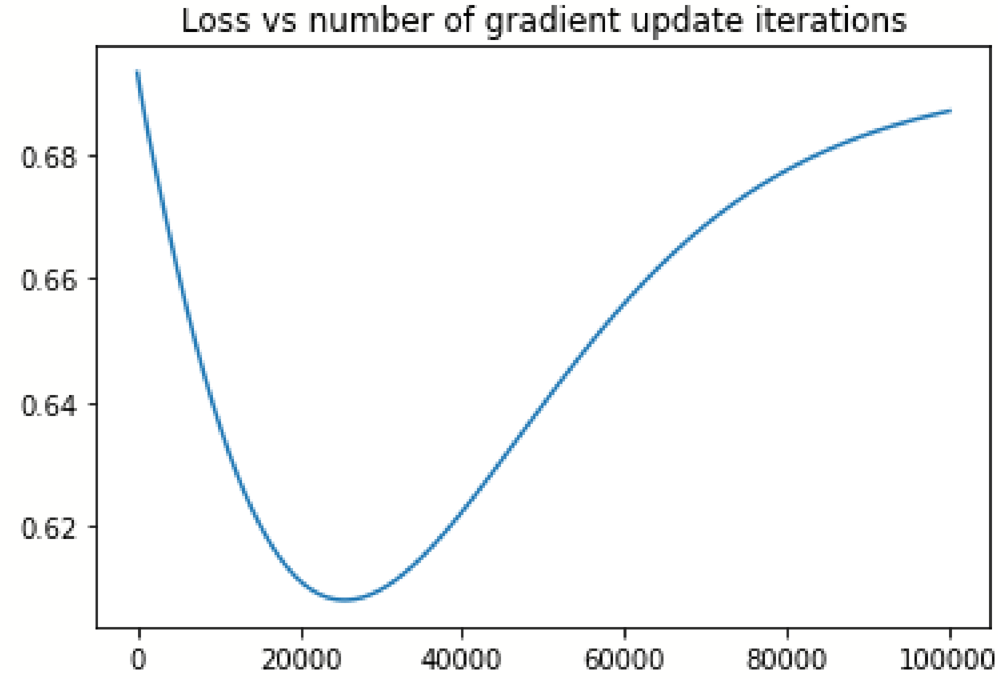
Accuracy is 76.11940298507463

Learning rate 0.001

Number of iterations 10000

- **Increasing the learning rate for same number of iterations**

We see that our accuracy has drastically reduced the reason being be have started to overshoot around the minima that is we oscillate around either side of the minima without ever truly reaching the local minima. Since our loss function, log loss is convex the local minima represents the global minima as well. The loss value actually starts to increase once we overshoot this as we have increased learning rate but we have not reduced the number of iterations nor do we have a condition for convergence.



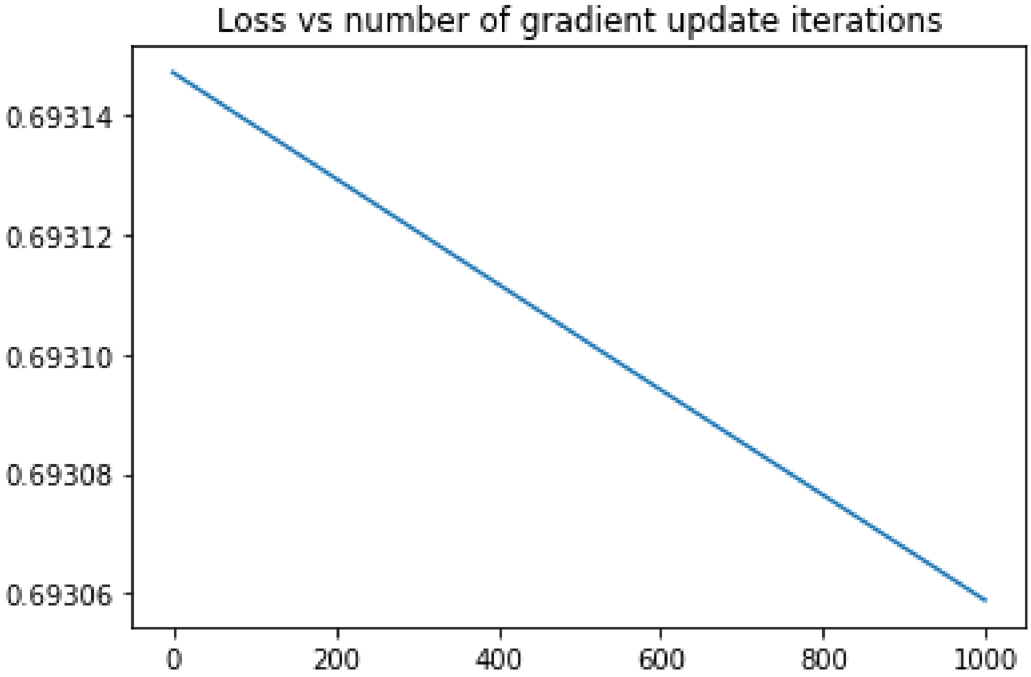
Loss vs number of gradient update iterations

Accuracy is 55.223880597014926
Learning rate 0.001

Number of iterations 100000

- **Decreasing both the learning rate and number of iterations**

The loss increases as the learning ability of our model is greatly reduced by learning rate and further compounded due to the short number of iterations we train it for, in other words we do not provide our model with enough acceleration(learning rate) nor do we give it sufficient time (number of iterations) to reach the minima. Thus even though our loss improves over the course and we move in the direction of minima but it isn’t able to reach the minima in time.

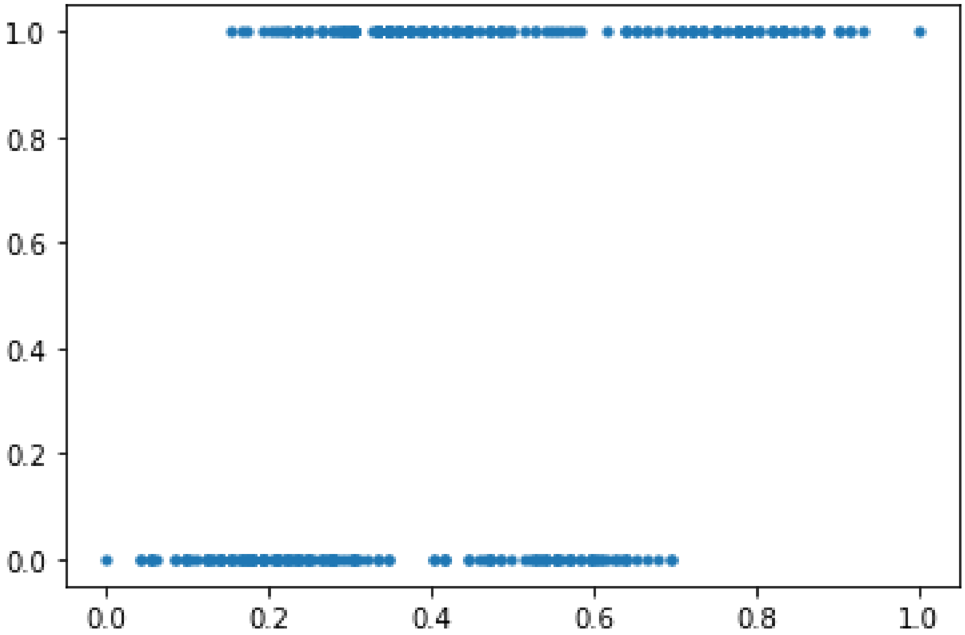


Accuracy is 50.74626865671642

Learning rate 0.00001

Number of iterations 1000

4. **Benefits of Logistic regression:**
    - **Often we cannot fit a simple straight line to the binary classification data so in that case sigmoid function which is flattened at the end helps us fit such data, for instance we have the following feature body weight based on gender it won’t be possible to fit in a line using linear regression but fitting a sigmoid function is pretty easy.**
    - **Outputs have a confidence value or probability.**
    - **We have a convex loss function, log loss.**

- 

Drawbacks:

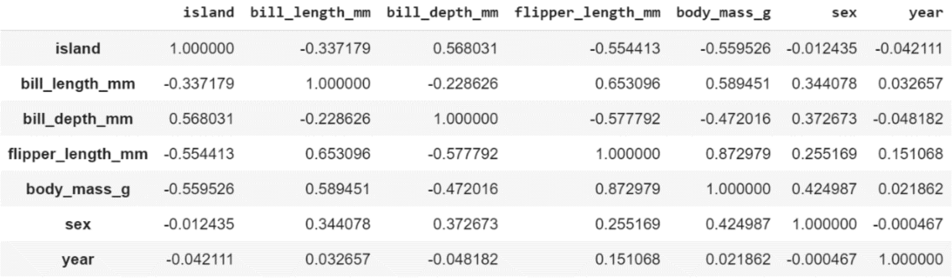
It suffers from the problem of vanishing gradient at values close to zero and one since the sigmoid function is flat towards the end regions.

Just like any other regression works poorly on discrete or non-continuous dataset.

Part III: Linear Regression

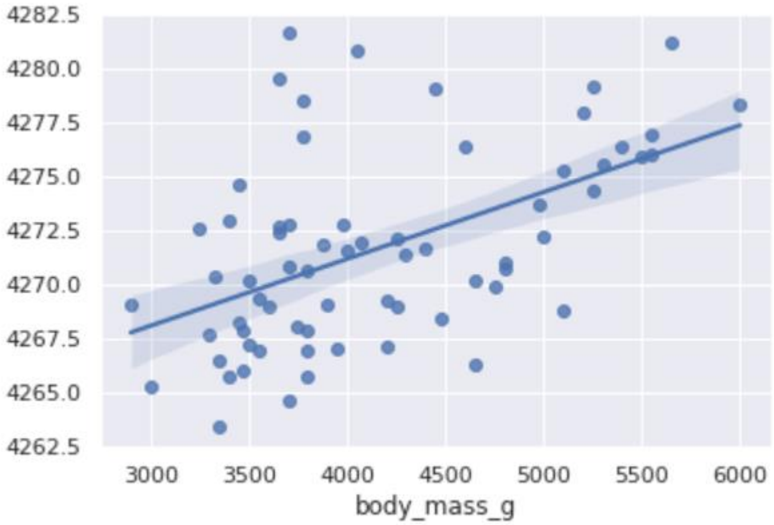
For linear regression, we use body mass as target.

From the correlation chart, we can see that the bill length, bill depth and flipper length have high correlation with the target. Hence, we choose these as our feature for the target



	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
island	1.000000	-0.337179	0.568031	-0.554413	-0.559526	-0.012435	-0.042111
bill_length_mm	-0.337179	1.000000	-0.228626	0.653096	0.589451	0.344078	0.032657
bill_depth_mm	0.568031	-0.228626	1.000000	-0.577792	-0.472016	0.372673	-0.048182
flipper_length_mm	-0.554413	0.653096	-0.577792	1.000000	0.872979	0.255169	0.151068
body_mass_g	-0.559526	0.589451	-0.472016	0.872979	1.000000	0.424987	0.021862
sex	-0.012435	0.344078	0.372673	0.255169	0.424987	1.000000	-0.000467
year	-0.042111	0.032657	-0.048182	0.151068	0.021862	-0.000467	1.000000

- The weights for linear regression using OLS is : [0.1999811 0.54198288 0.53486277]
- After calculating the weights using Ordinary least square method, the loss value is 615787.4988212374



From this graph of predicted body mass and true value, we can see that the line plotted has been plotted by calculating OLS.

Benefits:

- When using OLS to estimate weights, it is a simple algorithm which can be implemented with efficiency even with low computational power.
- OLS can be used for continuous variables.

Drawbacks:

- It can only construct linear lines.
- Prone to overfitting and underfitting
- Need continuous data cant work with discrete.

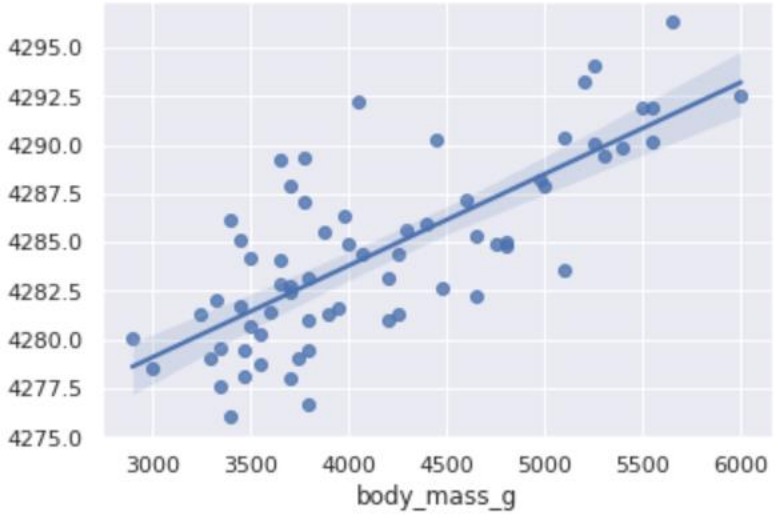
Part IV: Ridge Regression

Similar to linear regression, we take body mass as target.

For ridge regression, we keep tuning lambda to get more accurate predictions.

Team 98

- The means square error when using Ridge regression is 630710.5885649292
- The weights for Ridge regression is [0.84230836 0.06230377 0.2049406]
- For the plot comparing the true value vs predictions, we can see that the line is the best fit line using ridge regression formula.



The main motivation of using L2 regularization is to that it works better when we have multiple corelated features and also reduce overfitting by penalizing the cost function using an additional term which penalizes excessive variation of the coefficients, i.e reducing the model complexity.

L2 penalizes the weight^2 term.

We have the new cost function with additional term as

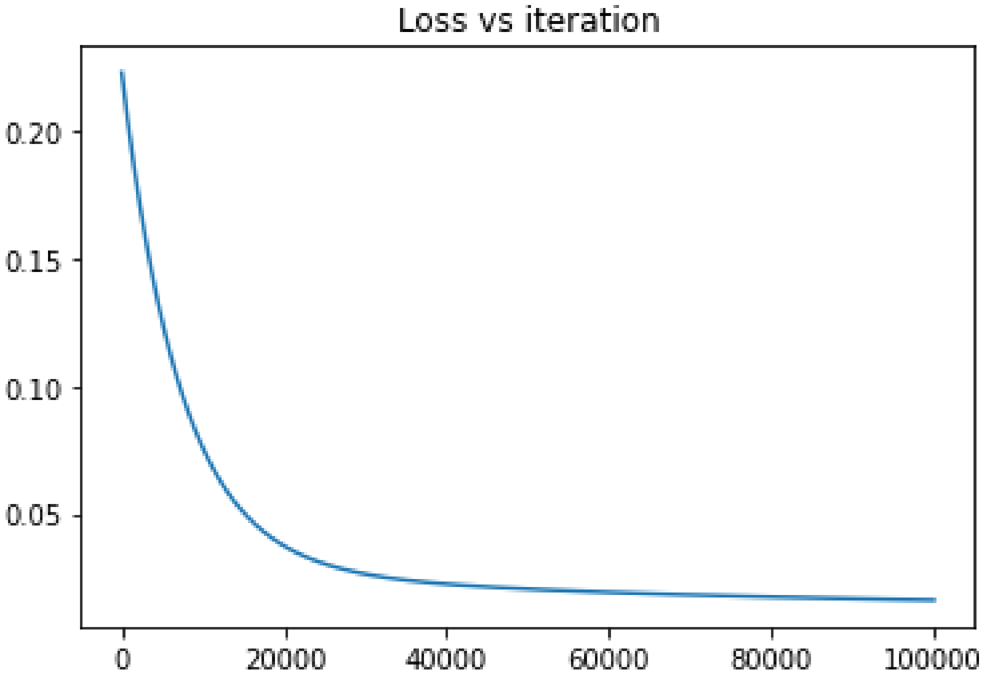
$$J_{regularized} = \underbrace{-\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\log\left(a^{[L](i)}\right) + (1-y^{(i)})\log\left(1-a^{[L](i)}\right)\right)}_{\text{cross-entropy cost}} + \underbrace{\frac{1}{m}\frac{\lambda}{2}\sum_{l}\sum_{k}\sum_{j} W_{k,j}^{[l]2}}_{\text{L2 regularization cost}}$$

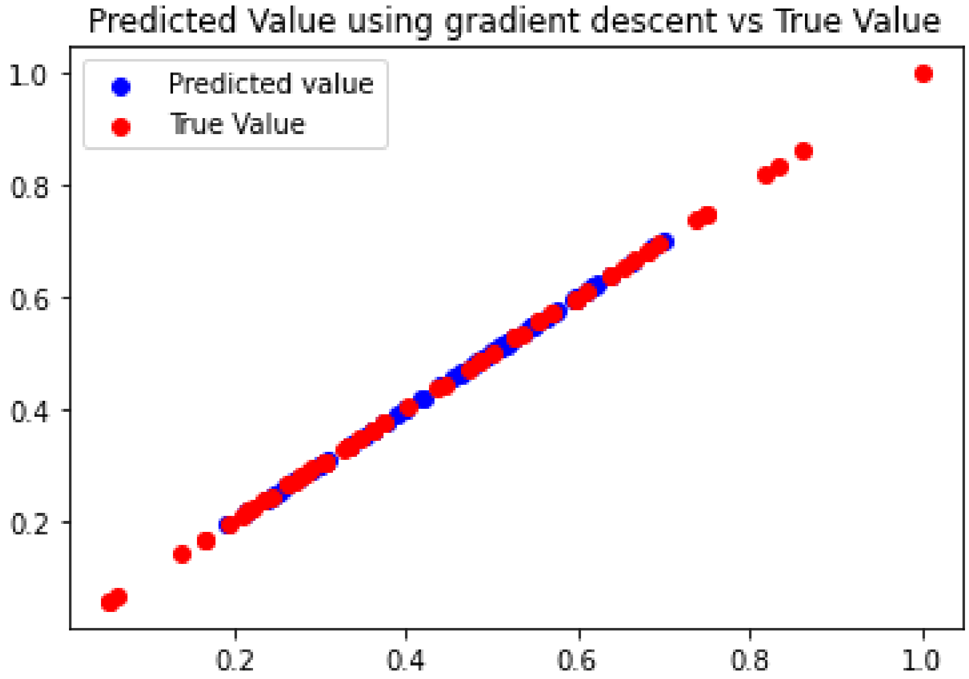
L2 regularization line can be nonlinear as well.

Bonus :

Gradient Descent

Using gradient descent we see that we have almost minimized the loss completely on the training data and even for the test data the misclassifications are only towards the extreme points.





The training time for gradient descent is smaller than linear and ridge regression although not by much for our scale we still see that the training loss decreases exponentially. We also saw linear decrease for the most part in logistic regression. Thus we can say gradient descent converges faster for same learning rate and number of iterations.

The built in gradient decent obviously still performs better due to optimizers, momentum and regularizes used in addition.

Note: Gradient Descent is combined with part 2 jupyter notebook file.

Team 98

Contribution Summary:

Team Member	**Assignment Part**	**Contribution(%)**
Vikram Segaran	Part I, II, III,IV	50
Naman Tejaswi	Part I, II, III,IV	50

References:

https://www.edureka.co/blog/linear-regression-for-machine-learning/

https://www.geeksforgeeks.org/

https://matplotlib.org/stable/gallery/lines_bars_and_markers/bar_stacked.html

https://towardsdatascience.com/