Task 1

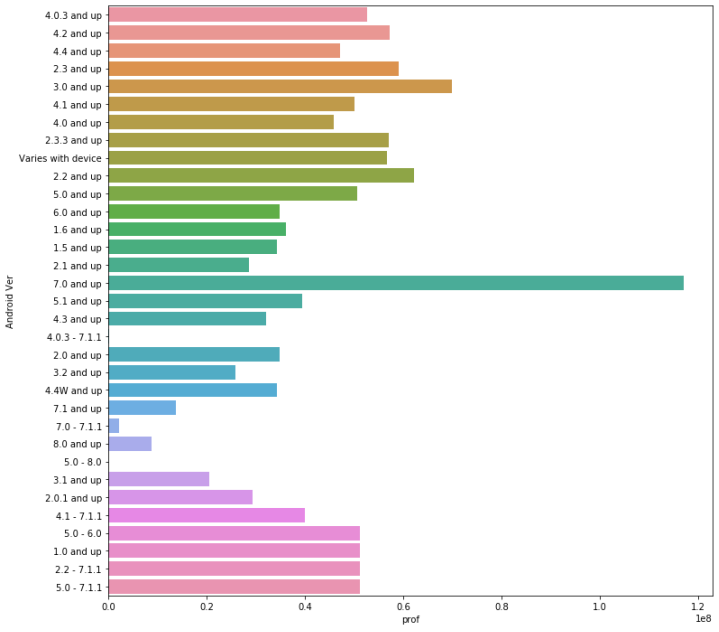
This is the dataset that was given to us. It has 13 attributes and each attribute has a type object. Which is the thing which needs to consider when plotting histogram.

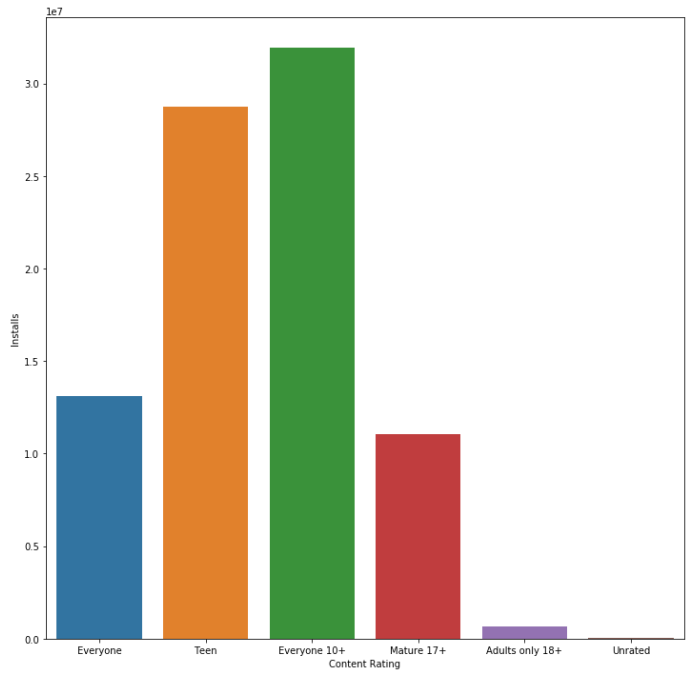
	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

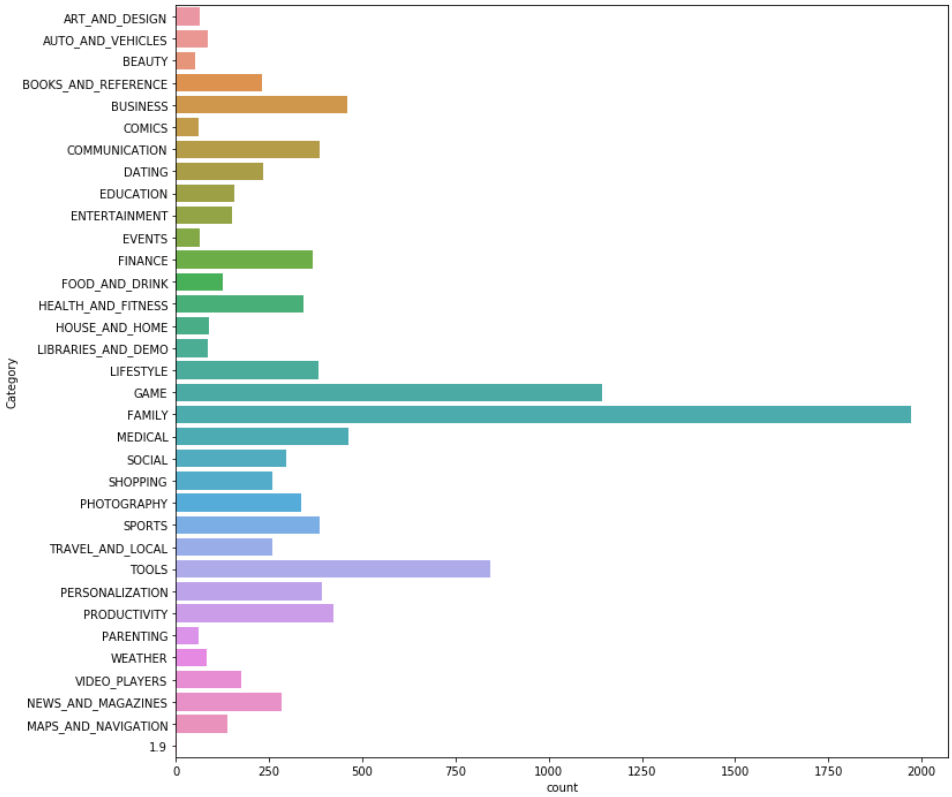
There is also an anomaly in our dataset on a row 10472 in which price and installs has strings values, therefore this also needs to be replaced by 0 in order to find profit and other graphs. As shown below

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10472	Life Made WI-Fi Touchscreen Photo Frame	1.9	19.0	3.0M	1,000+	Free	0	Everyone	NaN	February 11, 2018	1.0.19	4.0 and up	NaN

After doing these preprocessing steps, I have plotted some graphs for exploratory data analysis. Which clearly show the profit of every android version and the type of content that was installed by which category of people







Task2

Part 2_A: Linear Regression without Gradient Descent

Analysis:

In this task, we have two datasets provided and we have to solve them using gradient and without gradient.

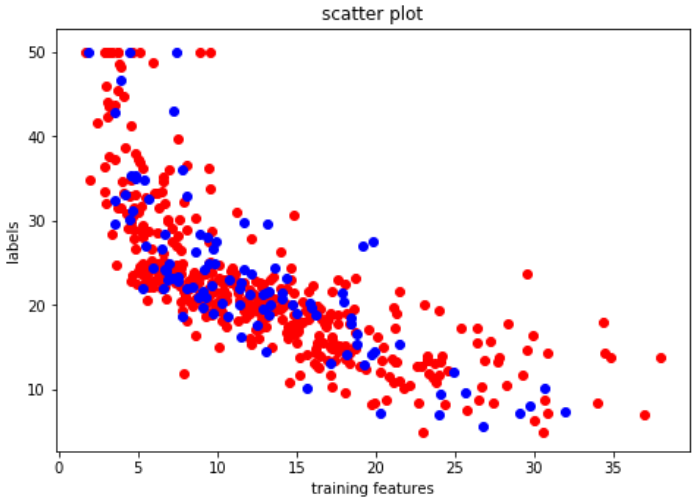


Figure 1- Scatter plot for dataset 1

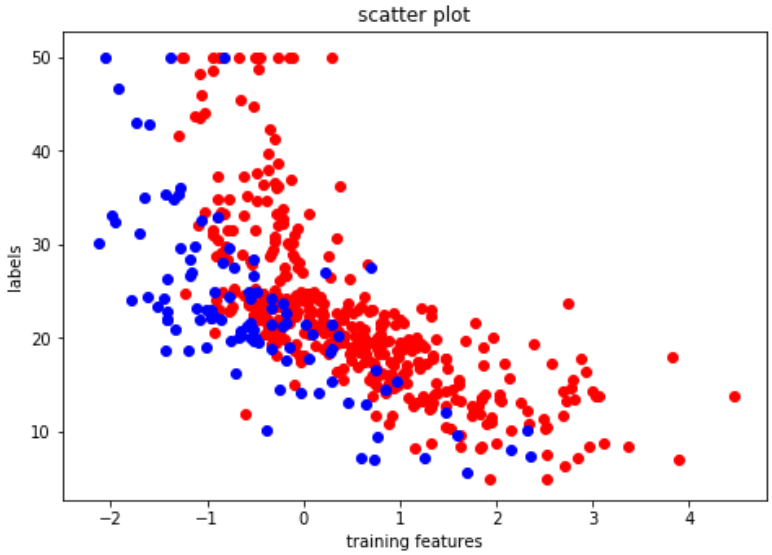


Figure 2- Scatter plot for dataset 2

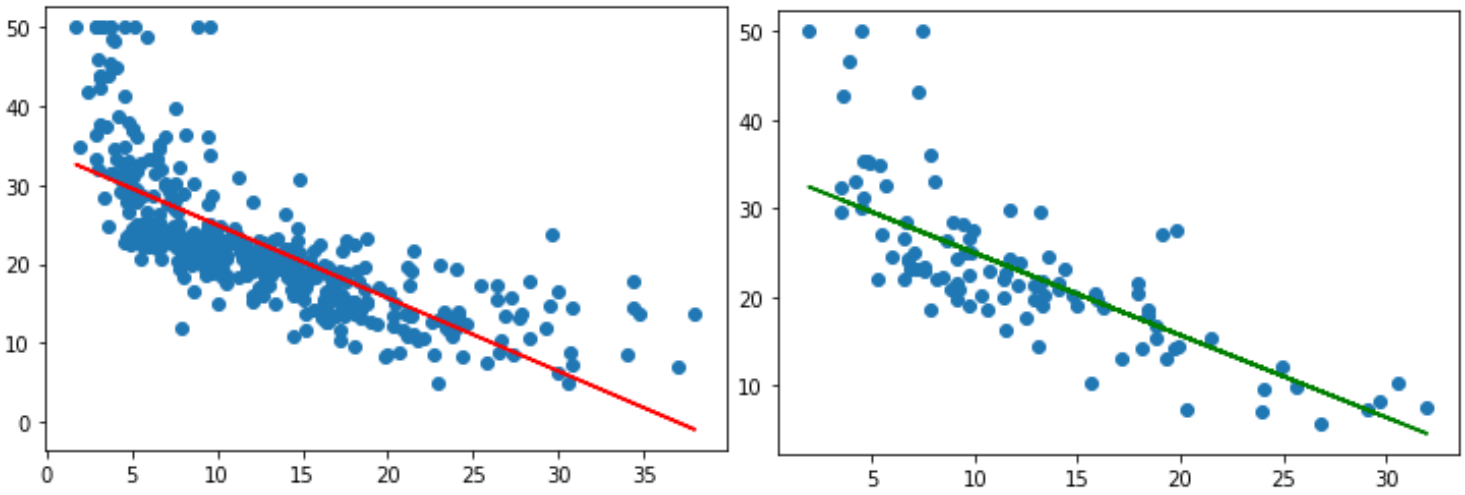
In gradient firstly visualize data points by using scatter plot. And for choosing slope and intercept I have solved an equation which gives us the best parameters where our loss will be minimum.

C = 34.21 and M = -0.92

After finding m and c I find testing and training error values for both datasets.

For dataset 1 here is the curve fitting for training and testing

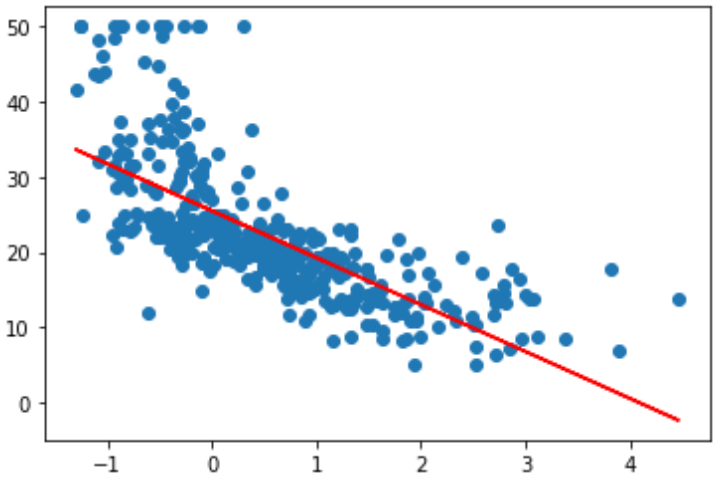
Training error = 34 and testing error = 39

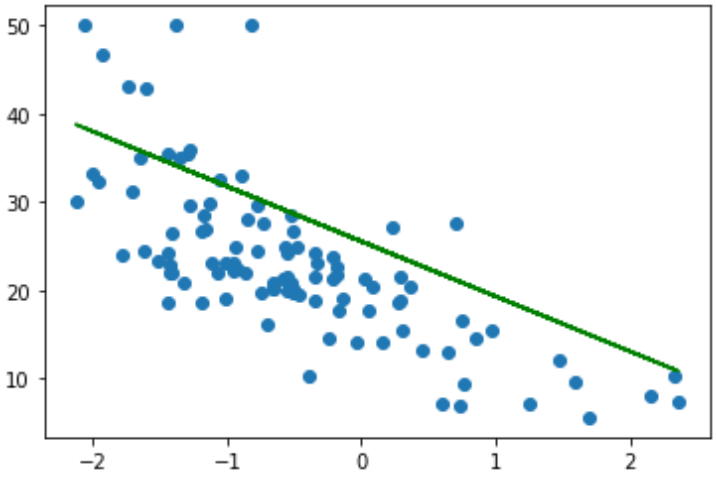


For dataset 2 here is the curve fitting for training and testing

M = -6.2381 and C = 25.5263

Training error = 42.8 and testing error = 70.8





Part 2_B: Linear Regression with Gradient Descent

Analysis:

In Part2 we have to use a stochastic gradient for finding perfect slope and intercepts. So as we know the property that stochastic update itself after every datapoint instead of updating after single epoch. For this first, we have to suppose the random value of slope and intercept. The value supposed are as follows:

So after applying that again error was calculated for both datasets and loss functions are plotted which are viewed as shown below.

alpha = 0.001

c = 2

m = 2

epochs = 5000

m_derivative = 0

c_derivative = 0

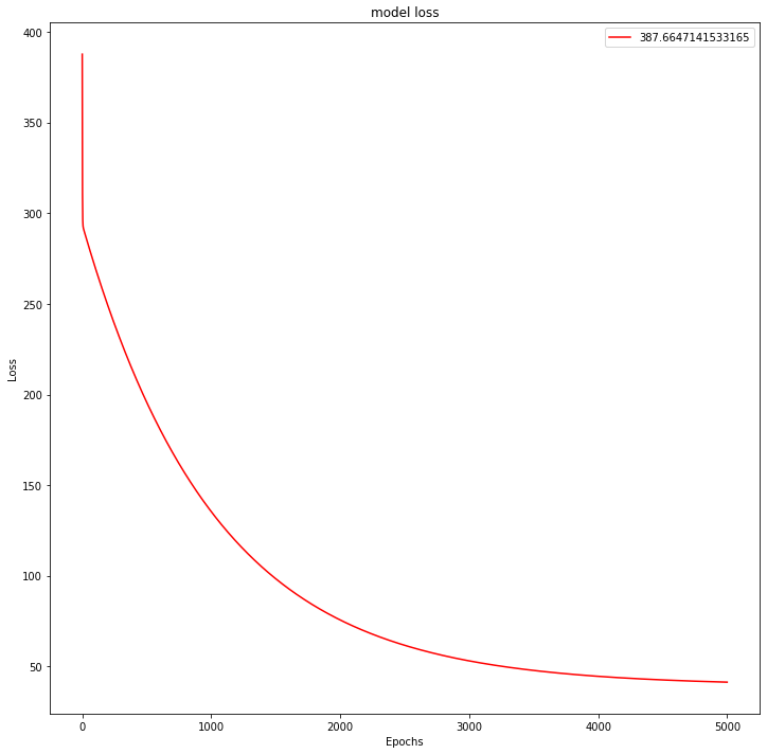
With the help of this, the best slope and intercept were found and calculate means square error for both datasets.

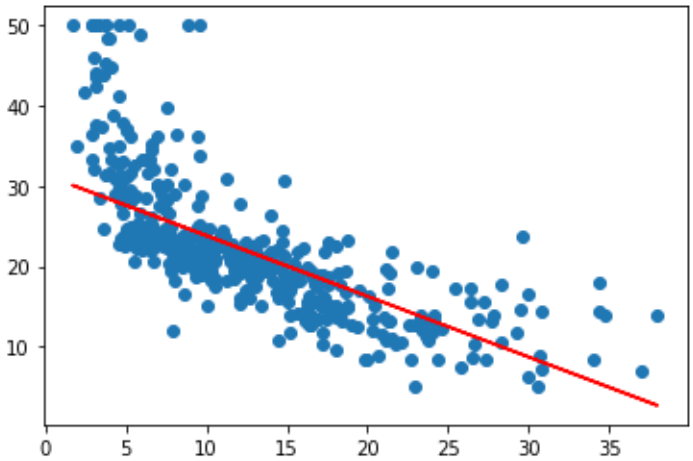
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

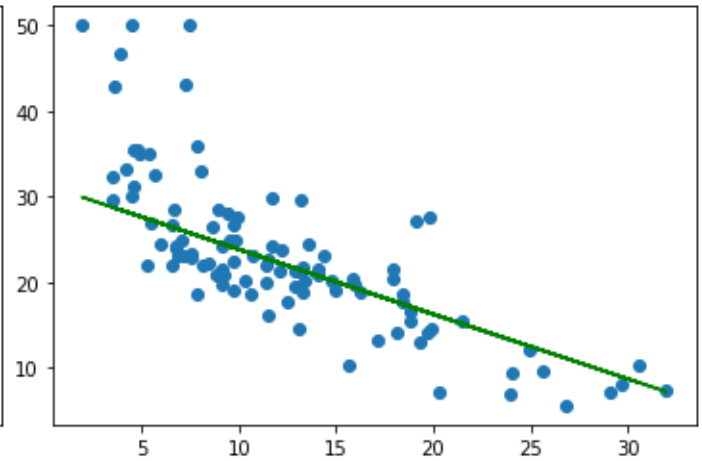
For dataset 1 here is the loss function for other results.

M = -0.756 C = 31.390

Training error = 41.38 and testing error = 38.95



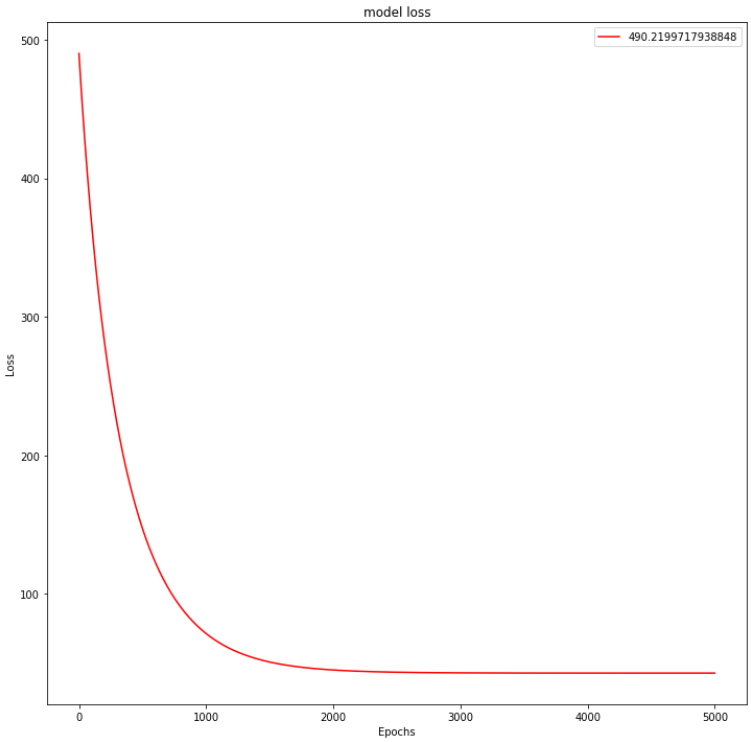




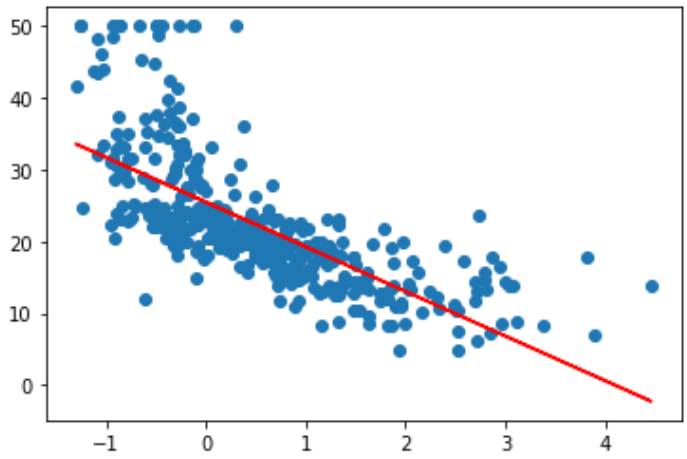
For dataset 2 here is the loss function for other results.

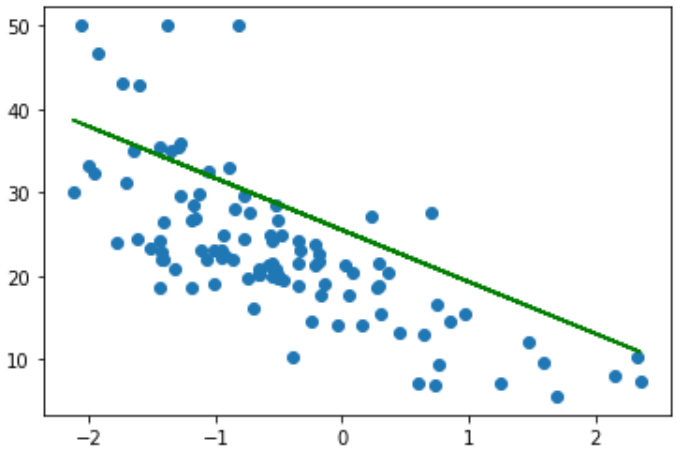
M = -6.2133 C = 25.492

Training error = 42.84 and testing error = 70.32



In this problem, the most difficult task is to tune our parameters including epochs and learning rate. By tweaking it we are able to find the best parameters. Learning rate has many effects on loss curve, if I increase it to maximum value my error did rapid increasing and decreasing in their value. Therefore, I use a minimum value of 0.001 to achieve good error value and loss curve


50
40
30
20
10
0
−1
0
1
2
3
4


50
40
30
20
10
−2
−1
0
1
2