COMP 551 -- Assingment 1

Name:
Tristan Saumure Toupin

ID:
260688712

Question 1. Sampling

```
'''
Movies: 0.2
COMP-551: 0.4
Playing: 0.1
Studying: 0.3

or

Movies:     [0.0, 0.2)
COMP-551:   [0.2, 0.6)
Playing:    [0.6, 0.7)
Studying:   [0.7, 1.0)
'''
```

Q1.1

```
# generate random number between [0, 1)
r = random.generate(min = 0, max = 1)
activity_today = None

if r < 0.2 :
      activity_today = "Movies"
else if r < 0.6 :
      activity_today = "COMP-551"
else if r < 0.6 :
      activity_today = "Playing"
else :
      activity_today = "Studying"

return activity_today
```

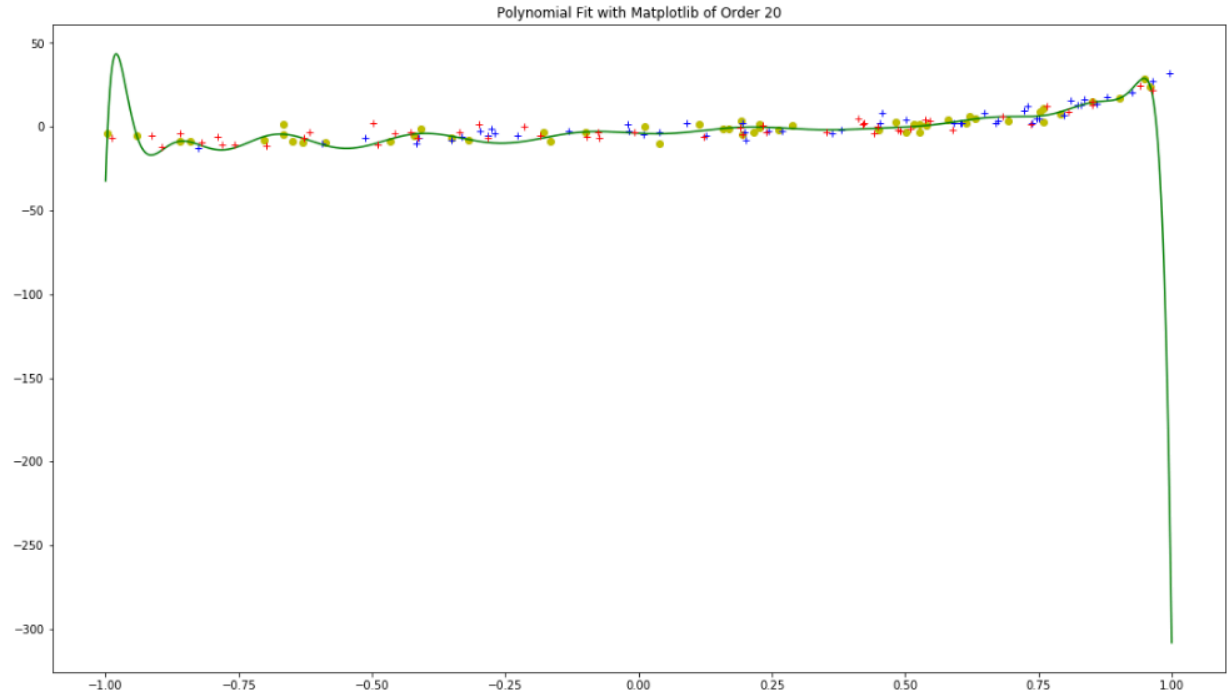
Question 2

Q2.1

a)

Mean Squared Error on the **Traning** dataset is **6.48**
Mean Squared Error on the **Validation** dataset is **1420.40**

b)



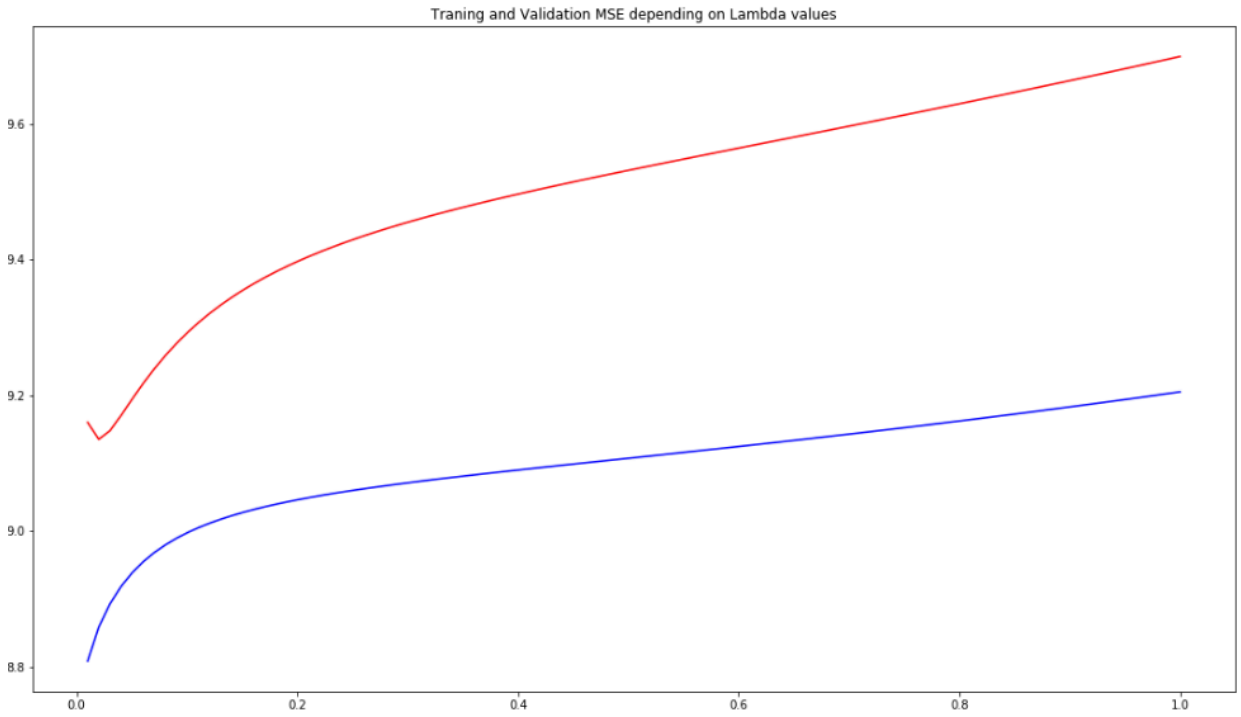
c)

Mean Squared Error on the **Testing** dataset is **50.50**
The model is overfitting the traning dataset.

1. The Mean Squared Error is much larger on the Testing dataset than on the Traning dataset.
2. The MSE of the Validation dataset is off charts due to the overfitting (not enough observation in training dataset).
3. The graph shows the regression ondulate trying to overfit datapoints from the traning dataset.

Q.2.2

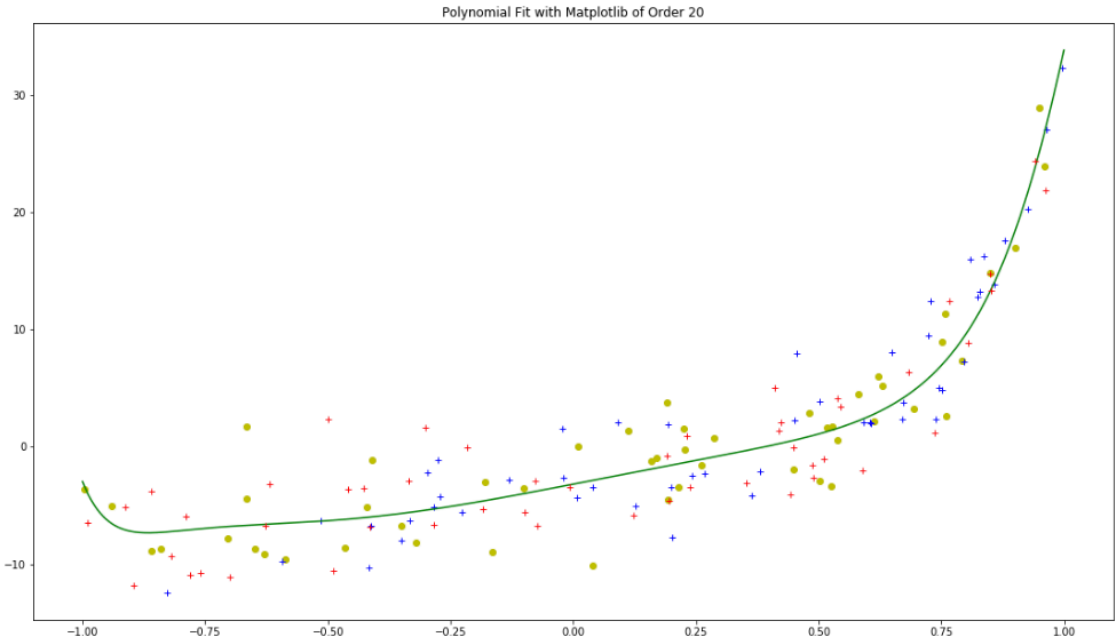
a)



b)

The best value for **Lambda** is **0.02020**
Mean Squared Error on the **Traning** dataset is **9.21**
Mean Squared Error on the **Validation** dataset is **9.70**
Mean Squared Error on the **Testing** dataset is **10.34**

c)



d)

The model is generalizing well the data. That being said, since the MSE is smaller for the training (9.21) set than the testing set (10.34), we could argue that the model is overfitting the training dataset.

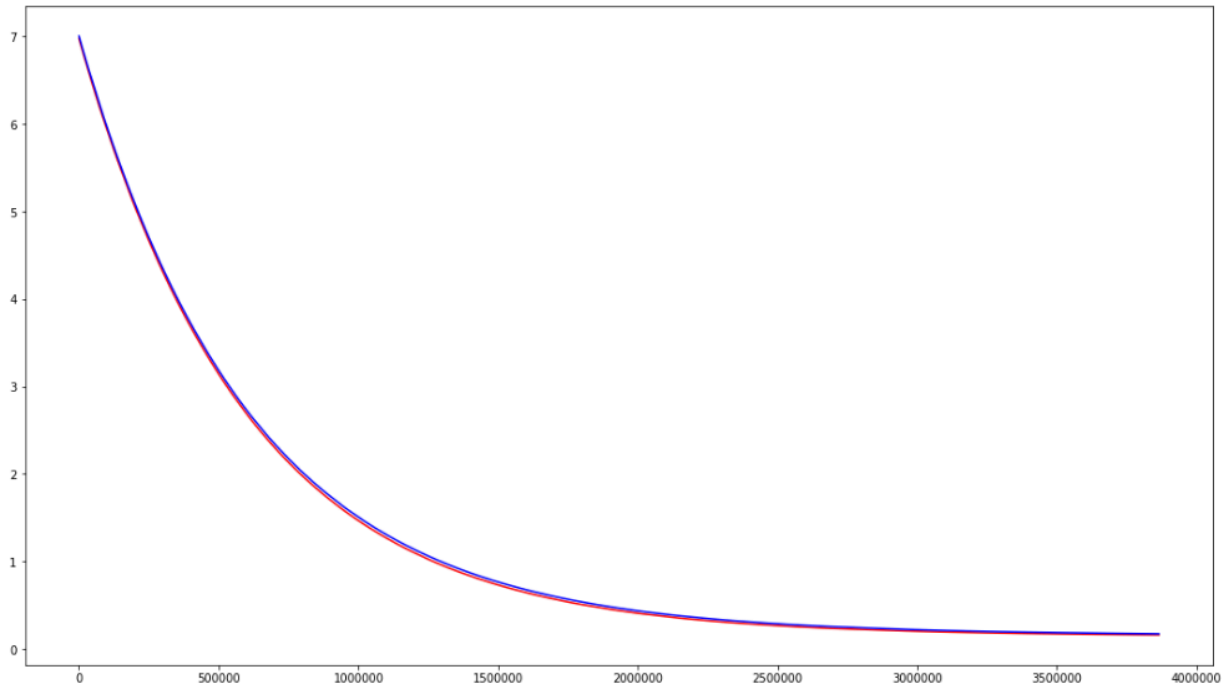
Q2.3

I believe that the degree of this polynomial is 4.
    -> There are 3 inflection points
    -> The X-axis is crossed twice.

Question 3

Q3.1

a) The model trained for 3864500 epoch. Epoch #3864500: MSE = 0.173552163312
b)
The MSE was caltulated over the complete set but trained over a single example.
In red: Training
In blue: validation



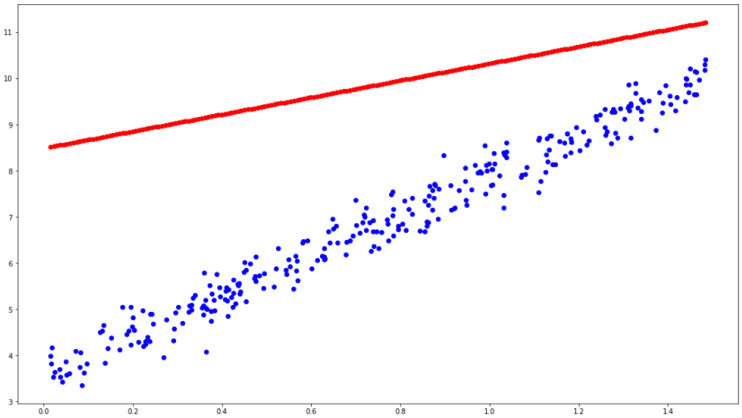
Q3.2

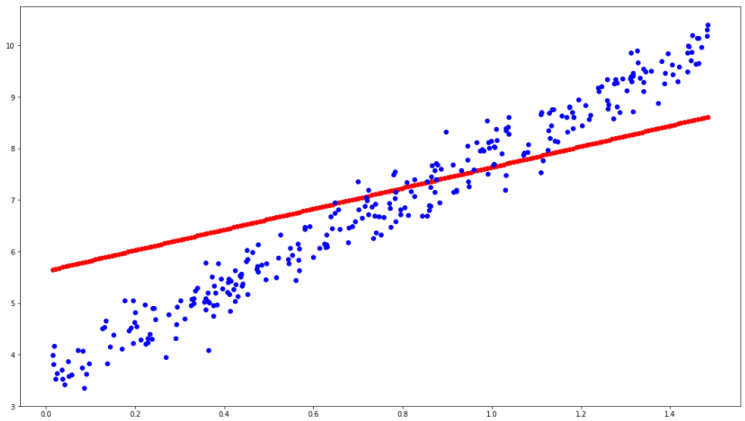
a)

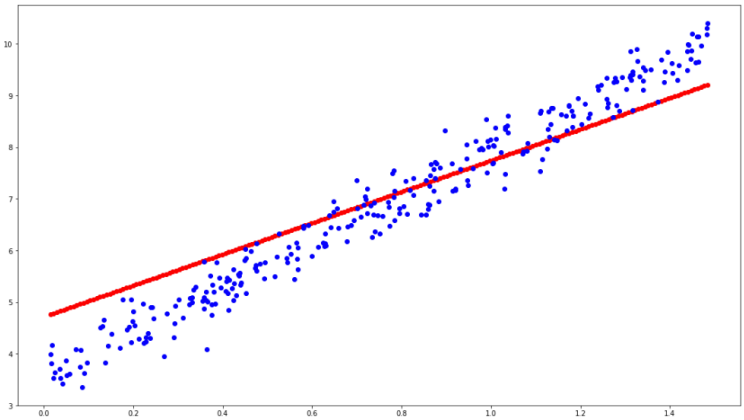
Alpha value	MSE
1e-1	0.103750538241
1e-2	0.102520499406
1e-3	0.106483725925
1e-4	0.107536859802
1e-5	0.108191426842
1e-6	0.173551892354

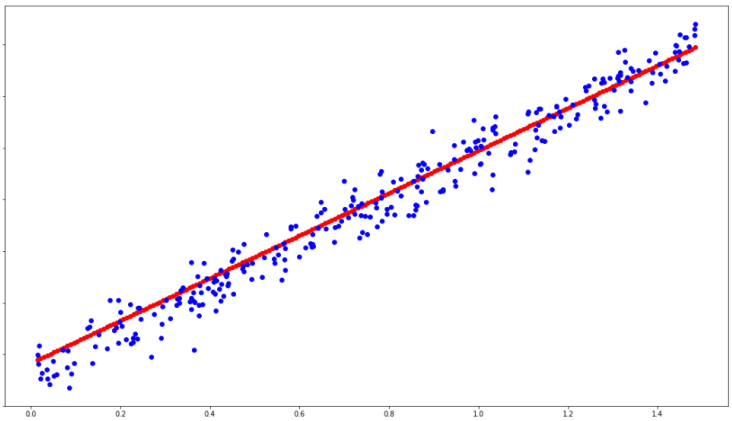
b)
Best MSE is 0.0782357984648 with alpha = 1e-05

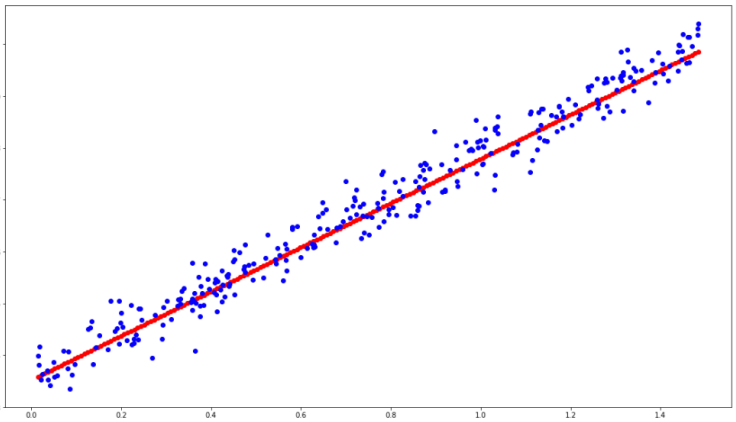
Q3.3

1. 
11
10
9
8
7
6
5
4
3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

2. 
10
9
8
7
6
5
4
3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

3. 
10
9
8
7
6
5
4
3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

4. 
10
9
8
7
6
5
4
3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

5. 
10
9
8
7
6
5
4
3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

Question 4

Q4.1

a)

It would be a terrible idea to replace all Nan cell with the mean of the columns. Here's why:

1. Some columns (e.g. county) are number representing classes.
2. Some nan can acutllay be approximated (e.g. PolicPerPop = PolicReqPerOffic / population).
3. Some columns have too many Nans (e.g. PolicBudgPerPop = 84% nan) and can be discarted.

b)

If the column represent clases...:

1. and the class ID is truly important for the task and there are not too many of them we are trying to solve, apply one hot encoding.
2. create a new class (e.g. -1).

If the column values can be calculated or approximated:

1. replace their value with the calculated/approximated value.

If the column has too many Nans:

1. Remove the column.

If the column has very little amount of Nans:

1. Remove the observation.

c)

Categorical Columns:

county
--> will replace nan with -1 (creating new category)
community
--> will be removed because according to the documentation, it is not predictive

Columns with large amount of Nan:
LemasSwornFT, LemasSwFTPerPop, LemasSwFTFieldOps, LemasSwFTFieldPerPop, LemasTotalReq, LemasTotReqPerPop, PolicReqPerOffic, PolicPerPop, RacialMatchCommPol, PctPolicWhite, PctPolicBlack, PctPolicHisp, PctPolicAsian, PctPolicMinor, OfficAssgnDrugUnits, NumKindsDrugsSeiz, PolicAveOTWorked, PolicCars, PolicOperBudg,LemasPctPolicOnPatr, LemasGangUnitDeploy, PolicBudgPerPop
--> will delete the column

Others:

OtherPerCap
--> replaced with average

d)

This is better because:

1. We only keep features that matter and have information in them.
2. We keep all the information that 'county' can provide and signal the model that there are data that wasn't labeled.
3. We add information by replacing 'OtherPerCap' nans with average.

Q4.2

Not completed.

Q4.3

a)

Not completed.

b)

Not completed.

c)

Not completed.

d)

A reduced set of feature should performs better if the selectio was done to eliminate noise in the dataset. The noice can influence (in a bad way) a model.