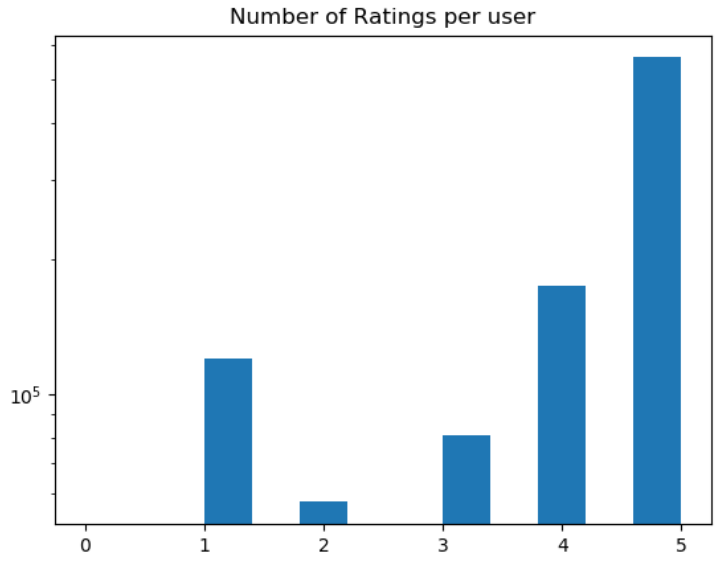
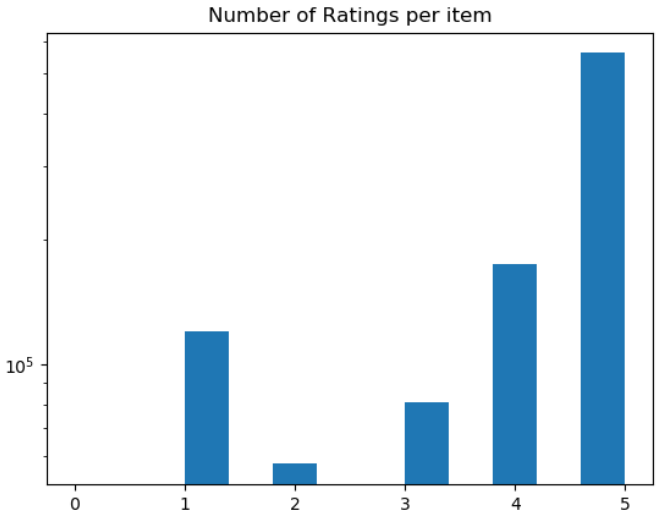
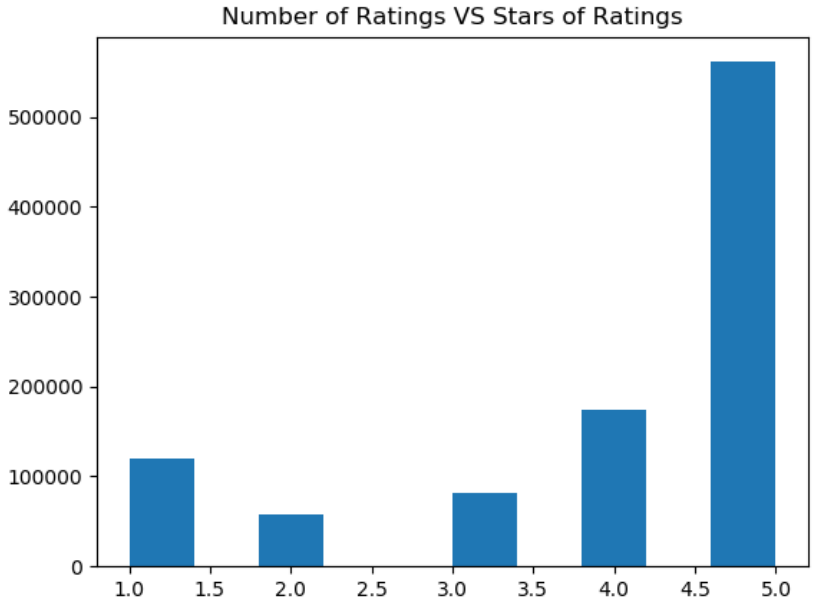
Q1
Q1.1

1. The ID is B000HCLLMM, the number of stars: 14454
2. User: A100WO06OQR8BQ, the number of Items: 161
3. The histograms generated are







Q1.2

Q1.3

Q2

Q2.1

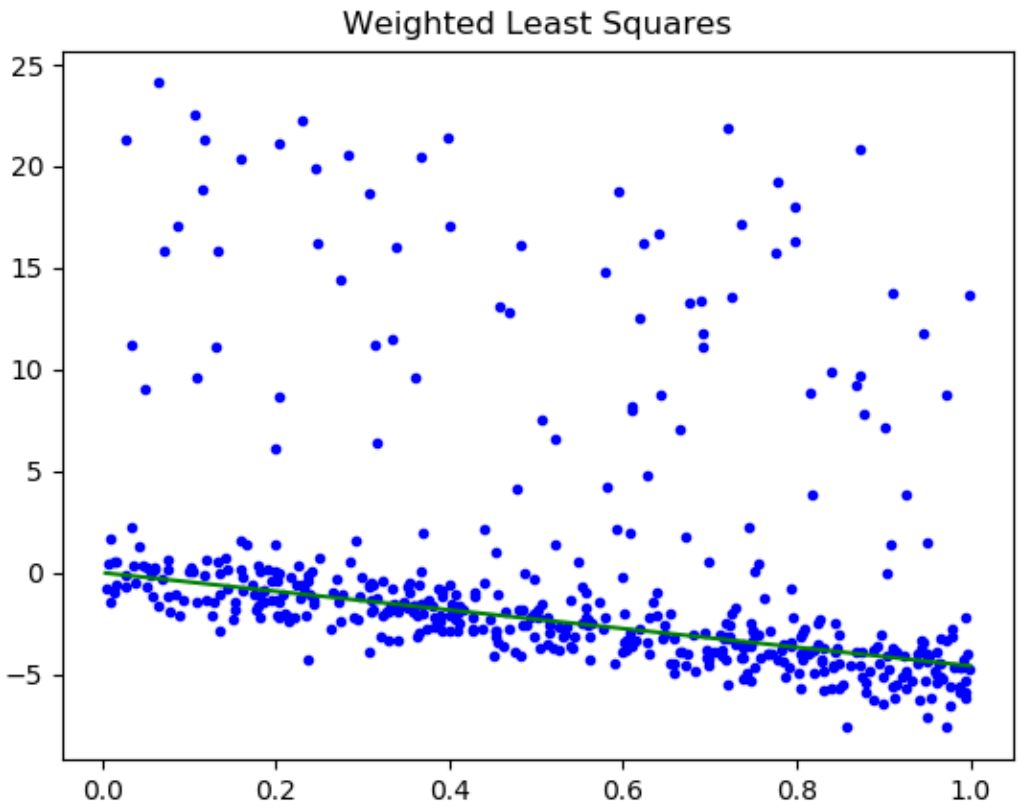
1. $||Xw - y||_{\infty}$
2. $\left||Xw - y|\right|^{T} V(Xw - y) + \frac{\lambda}{2}||w||_1$
3. $||Xw - y||_1^2 + \frac{1}{2}||\Lambda w||_1$

Q2.2

1. $f(w) = \frac{1}{2}\left||w - v|\right|^2 = \frac{1}{2}\left||w|\right|^2 - w^T v + \frac{1}{2}\left||v|\right|^2 = \nabla f(w) = w - v = 0$
2. $f(w) = \frac{1}{2}\left||Xw - y|\right|^2 + \frac{1}{2}w^T \Lambda w = X^T Xw - X^T y + \Lambda w = \nabla f(w) = (X^T X + \Lambda)w = 0$
3. $f(w) = \frac{1}{2}\sum_{i=1}^{n} v_i (w^T x_i - y_i)^2 + \frac{\lambda}{2}\left||w - w^0|\right|^2 = X^T V(Xw - y) - \lambda(w - w^0) = w(X^T VX + \Lambda) = \Lambda w^0 + X^T Vy$

Q3

Q3.1



Q3.2

Using r as an approximation

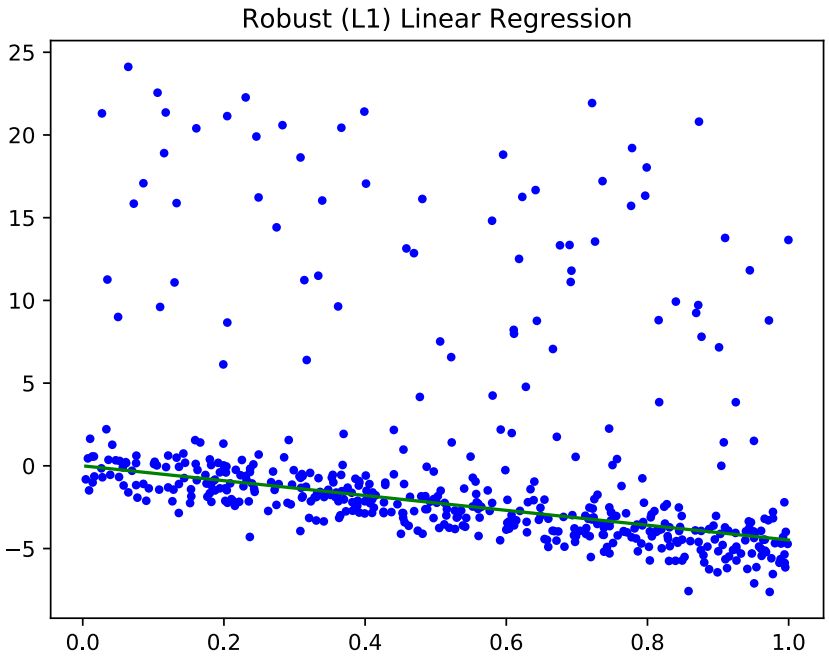
$$|r| \approx \log(\exp(r) + \exp(-r))$$

$$\frac{d}{dr}\log(exp(r) + exp(-r)) = \frac{\exp(r) - \exp(-r)}{\exp(r) + \exp(-r)}$$

The gradient of the function with respect to w_j:

$$\frac{df}{dw} = \sum_{i=1}^{n} x_i \frac{(\exp(w^T x_i - y_i) - \exp(y_i - w^T x_i))}{(\exp(w^T x_i - y_i) + \exp(y_i - w^T x_i))}$$

Q3.3

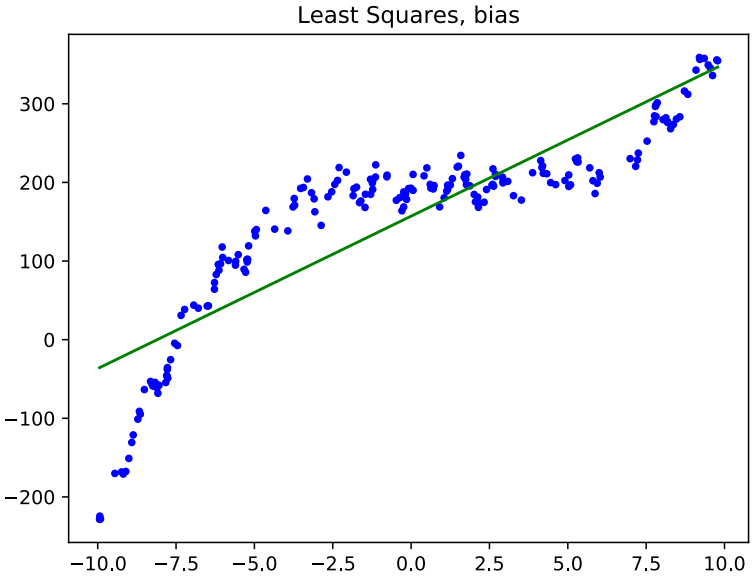


Q4

Q4.1

Training error: 3551.3

Testing error: 3393.9



Q4.2

(p=0) Training error: 15480.5	Testing error: 14390.8
(p=1) Training error: 3551.3	Testing error: 3393.9
(p=2) Training error: 2168.0	Testing error: 2480.7
(p=3) Training error: 252.0	Testing error: 242.8
(p=4) Training error: 251.5	Testing error: 242.1
(p=5) Training error: 251.5	Testing error: 239.5
(p=6) Training error: 248.6	Testing error: 246.0
(p=7) Training error: 247.0	Testing error: 242.9
(p=8) Training error: 241.3	Testing error: 246.0
(p=9) Training error: 235.8	Testing error: 259.3
(p=10) Training error: 235.1	Testing error: 256.3

Q5

1. K-means will get influenced by the outlier as it will take into the mean distance between the points of the data however, density-based clustering will not be influenced by a global outlier as there could n amount of clusters

2. We need random restarts for K-means as the clusters differ in each run and we need to generate the minimum error in distances squared however density-based clustering does not change with each iteration

3. Hierarchical clustering cannot handle non-convex clusters

4. An example of model-based outlier detection is using normal distribution and z-score, the problem with this model is that and variance which is heavily influenced by outliers.

5. Examples of graphical-based outlier detection are plots like boxplots and scatterplots however, the problem is the limitation of the variables that could be represented

6. Example of supervised-based outlier detection are decision trees; however, the limitation is that we would have to know the type of outliers as new types cannot be detected

7. Using gradient descent with 1 feature would not make sense as least square solution as gradient descent is used for higher number of features due to the runtime 0(ndt)

8. We typically use columns of 1 to add a bias and we should not do with a decision tree model

9. If the function is convex, the stationary points represent the maxima or the minima of the graph, and yes convexity implies that stationary points exists

10. We need gradient descent for robust regression because normal equations give misleading results if assumptions regarding normal equations is not true

11. The program may require more time to compute the minimum

12. The program may miss the minimum with big jumps hence, the loss could increase

13. Convex and smooth approximation to max function is the purpose of log-sum-exp function, it is related to gradient descent as it is a smooth approximation

14. We could use trigonometric functions to map the periodic function