Report

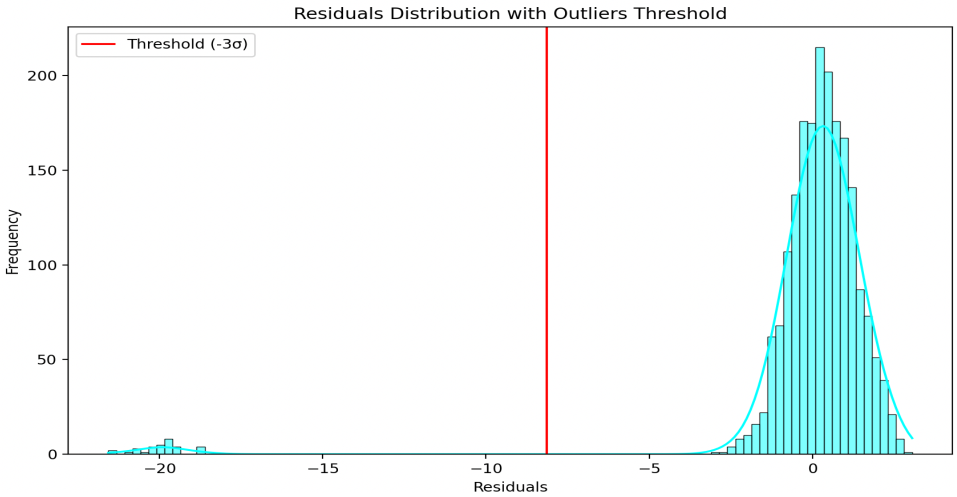
Homework – 1

Nikhil Saji
50592329

Task 1 – **Linear Regression**

1) Packages Used:
   a) Pandas
   b) Sklearn
2) Steps followed to find mislabeled data.
   a) Loaded the linear_regression CSV file.
   b) Got the description of the dataset using 'describe'.
   c) Checked for outliers.
   d) Created two data frames X and y, one is for training and other one for testing.
   e) Used Linear Regression model to fit the training dataset.
   f) Using the trained model, predicted the target value for the entire dataset.
   g) Found the **residual** between y and y_predict
   h) Checked for minimum residual value and the mean residual value
   i) The mean was around 3.44 whereas the minimum was -21 which meant there are mislabeled samples

j) After checking the graph and I created a threshold for filtering the data based on standard deviation of the residual values



k) I found 32 mislabeled samples after filtering the dataset

l) Labelled them as 1 and stored it in new column Outliers

Task 2 – **Image Editing using k-means and kNN**

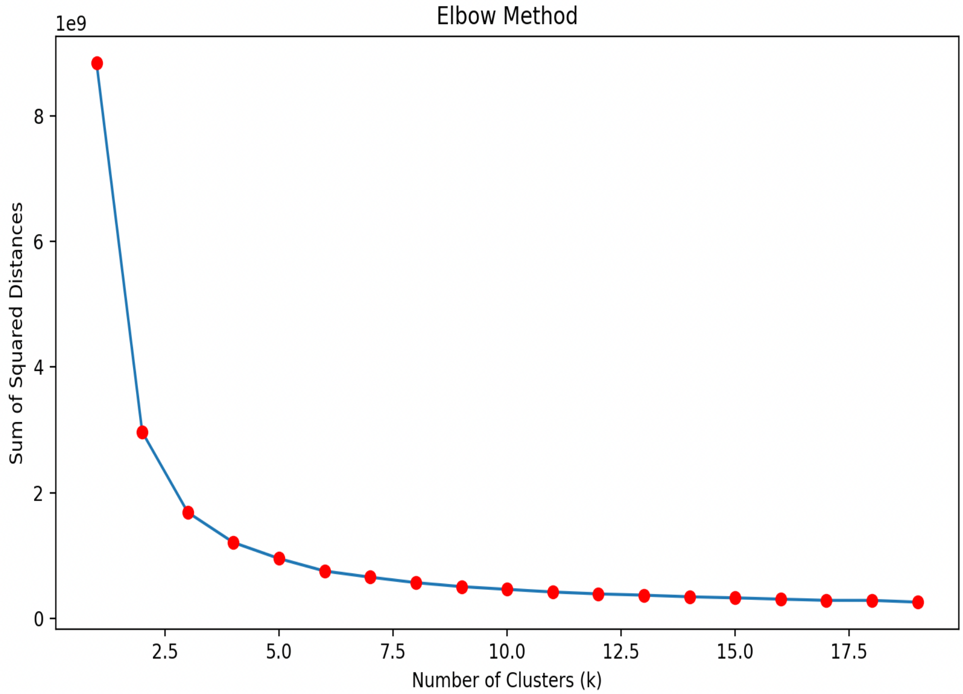
1) Packages Used –

   a) PIL
   b) Numpy
   c) Sklearn
   d) Matplotlib
   e) Tqdm

2) Steps followed to create the compressed image 1 and 2 –

   a) Used Image function from package PIL to open image
   b) Converted the image to numpy array
   c) Used k values ranging from 2 to 20

d) Using Elbow method, I found out that **k = 9** had the minimum sum of squared distance and beyond that point the difference was minor.

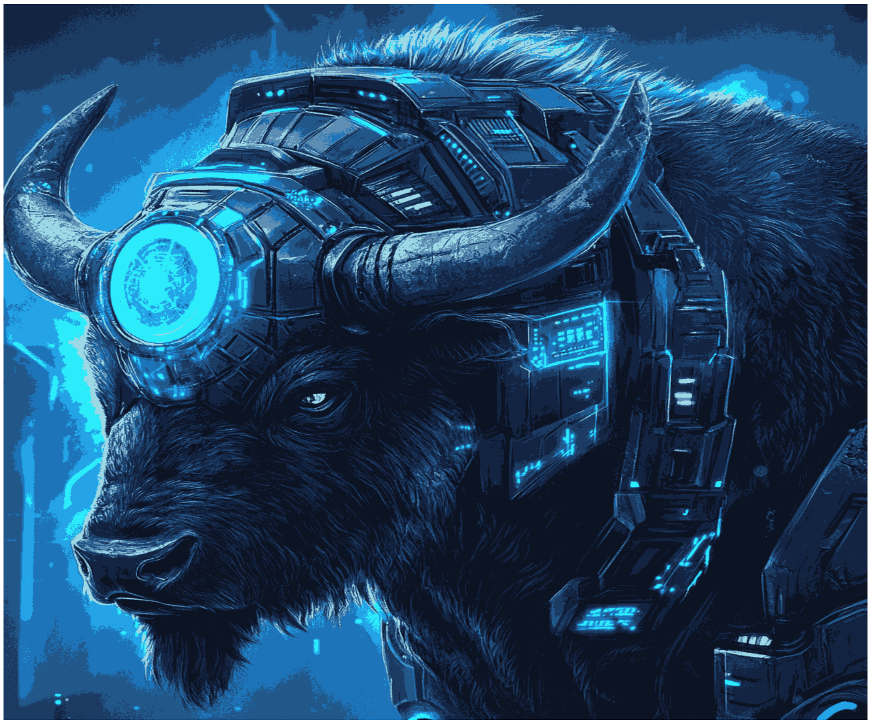


e) Used K_Means to fit the image1

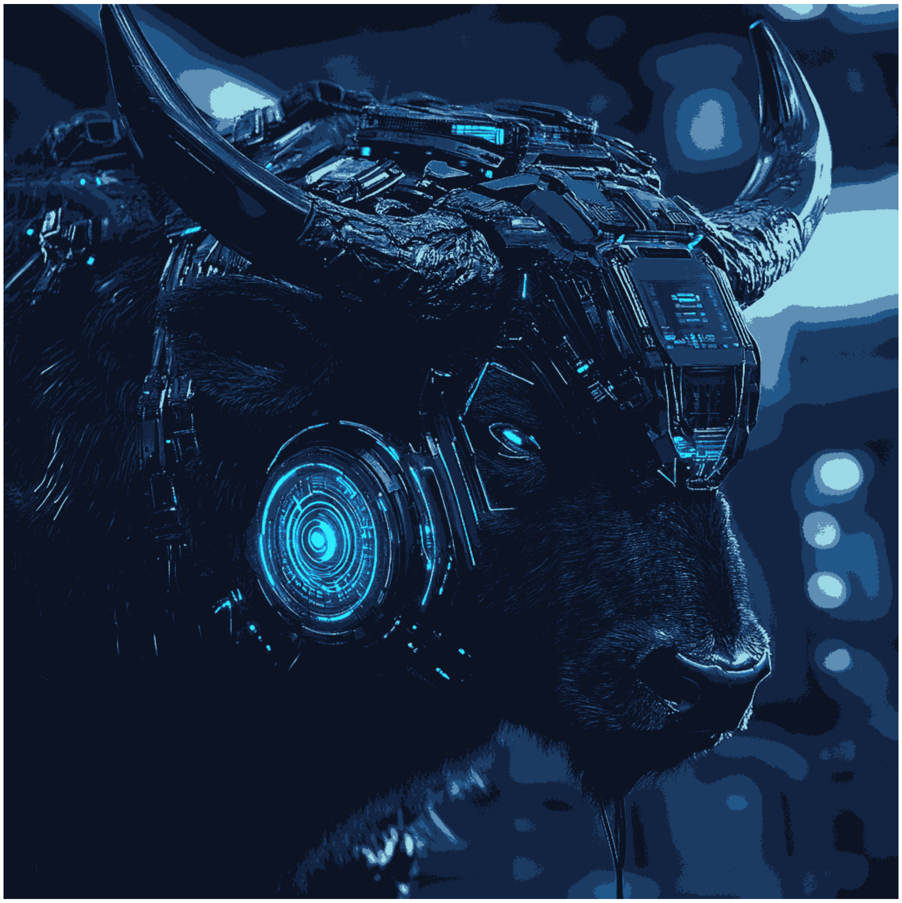
f) Using kNN , k-value and the centroid from K_Means, I fitted the model

g) Compressed the image2 and here is the output

h) Image 1 Compressed



i) Image 2 compressed using the same color palate



j) If we had used a higher k value (for instance k ≥ 60), we could have got a vibrant image

k) But using elbow method we determine that increasing k yields will have diminishing returns in terms of SSE reduction, hence adding more k (cluster value) does not significantly improve the quality of the clusters thus helping to avoid overfitting and unnecessary complexity.