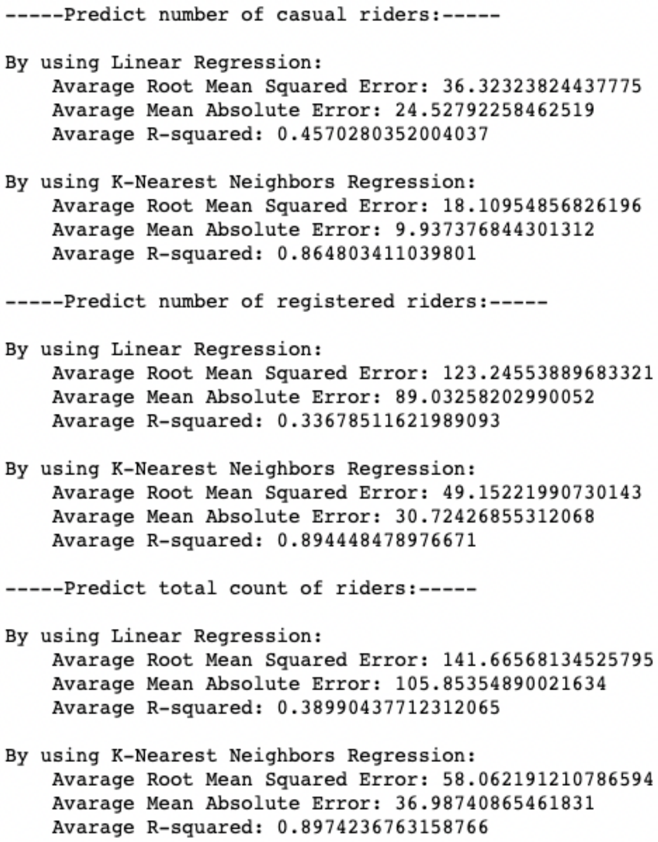
CS 5644: Assignment 3 report

Question 1:

During the preprocessing steps, I drop the useless columns ‘instant’ and ‘dteday’. The reason is that knowing ‘instance’ does not help our training, while for dropping ‘dteday’ is that we already have features such as ‘yr’, ‘mnth’, and ‘hr’ that indicates the property of date.

During the model training process, K-fold is used for cross-validation with the dataset split into 5 subsets. The number of neighbors is set to the default value for K-Nearest Neighbors, which is 5. For model evaluation, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) are applied due to they are the most common ones for regression model evaluation.

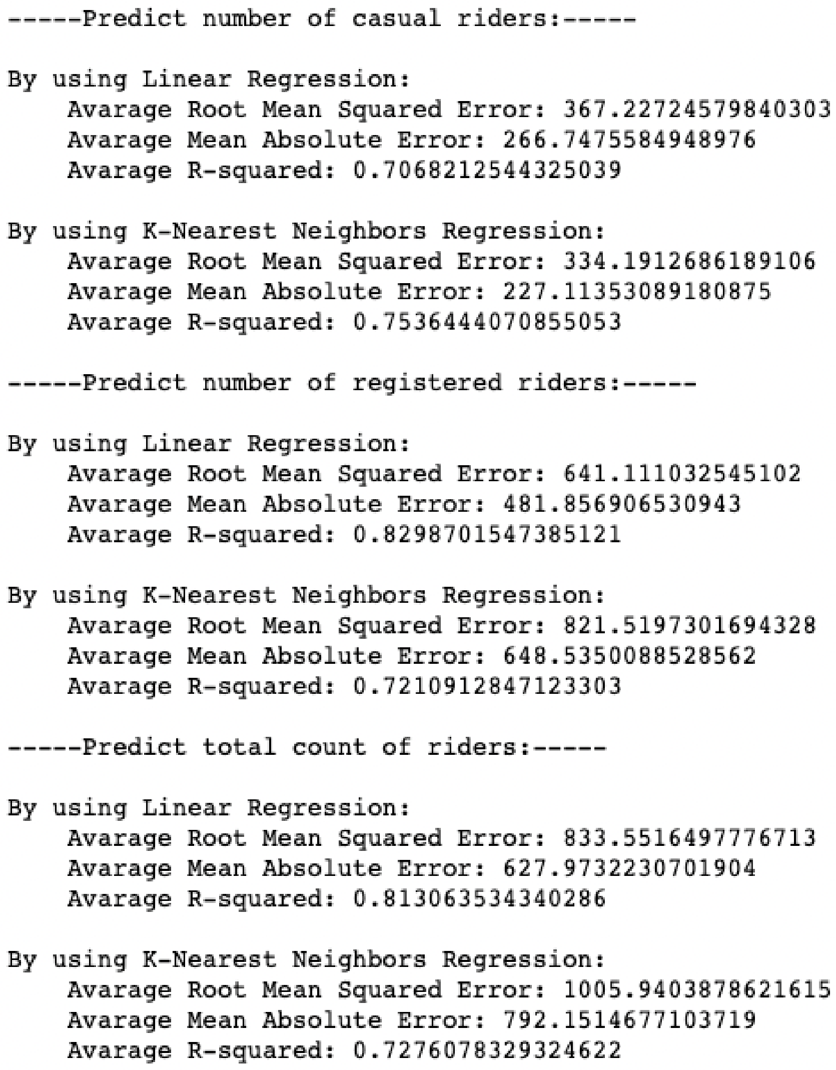
Following are the prediction results for dataset “hour.csv”.



```
-----Predict number of casual riders:-----

By using Linear Regression:
    Avarage Root Mean Squared Error: 36.32323824437775
    Avarage Mean Absolute Error: 24.52792258462519
    Avarage R-squared: 0.4570280352004037

By using K-Nearest Neighbors Regression:
    Avarage Root Mean Squared Error: 18.10954856826196
    Avarage Mean Absolute Error: 9.937376844301312
    Avarage R-squared: 0.864803411039801

-----Predict number of registered riders:-----

By using Linear Regression:
    Avarage Root Mean Squared Error: 123.24553889683321
    Avarage Mean Absolute Error: 89.03258202990052
    Avarage R-squared: 0.33678511621989093

By using K-Nearest Neighbors Regression:
    Avarage Root Mean Squared Error: 49.15221990730143
    Avarage Mean Absolute Error: 30.72426855312068
    Avarage R-squared: 0.894448478976671

-----Predict total count of riders:-----

By using Linear Regression:
    Avarage Root Mean Squared Error: 141.66568134525795
    Avarage Mean Absolute Error: 105.85354890021634
    Avarage R-squared: 0.38990437712312065

By using K-Nearest Neighbors Regression:
    Avarage Root Mean Squared Error: 58.062191210786594
    Avarage Mean Absolute Error: 36.98740865461831
    Avarage R-squared: 0.8974236763158766
```

For the dataset “hour.csv”, the K-Nearest Neighbors model shows lower RMSE and MAE and a higher R2 value than the Linear Regression model on all three targets. Since the model is said to be decent if one has a low MSE or MAE value or a high R2 value, we can conclude that when training the model with hourly data, K-Nearest Neighbors is a better machine learning method for regression model than Linear Regression.

The prediction results for dataset “day.csv” is presented as follows.



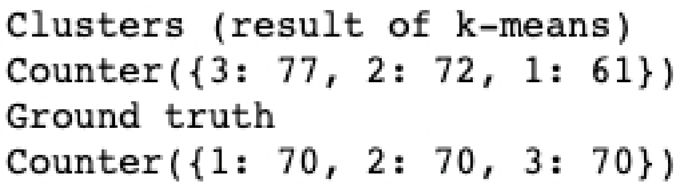
```
-----Predict number of casual riders:-----

By using Linear Regression:
    Avarage Root Mean Squared Error: 367.22724579840303
    Avarage Mean Absolute Error: 266.7475584948976
    Avarage R-squared: 0.7068212544325039

By using K-Nearest Neighbors Regression:
    Avarage Root Mean Squared Error: 334.1912686189106
    Avarage Mean Absolute Error: 227.11353089180875
    Avarage R-squared: 0.7536444070855053

-----Predict number of registered riders:-----

By using Linear Regression:
    Avarage Root Mean Squared Error: 641.111032545102
    Avarage Mean Absolute Error: 481.856906530943
    Avarage R-squared: 0.8298701547385121

By using K-Nearest Neighbors Regression:
    Avarage Root Mean Squared Error: 821.5197301694328
    Avarage Mean Absolute Error: 648.5350088528562
    Avarage R-squared: 0.7210912847123303

-----Predict total count of riders:-----

By using Linear Regression:
    Avarage Root Mean Squared Error: 833.5516497776713
    Avarage Mean Absolute Error: 627.9732230701904
    Avarage R-squared: 0.813063534340286

By using K-Nearest Neighbors Regression:
    Avarage Root Mean Squared Error: 1005.9403878621615
    Avarage Mean Absolute Error: 792.1514677103719
    Avarage R-squared: 0.7276078329324622
```

As for the dataset “day.csv”, when predicting the number of casual riders, the K-Nearest Neighbors model shows a better performance with lower RMSE and MAE and a higher R2 value. However, when it comes to predicting the number of registered riders or the total count of riders, the Linear Regression model outperforms the K-Nearest Neighbors regression with distinctly lower error scores and a higher R2 score. Therefore, K-Nearest Neighbors is a better algorithm when predicting the number of casual riders, whereas Linear Regression is a nicer one when predicting the number of registered riders or the total count of riders.

Question 2:

During the preprocessing, column names are assigned with 'area', 'perimeter', 'compactness', 'length', 'width', 'asymmetry_coef', 'groove_len', and 'class'. The last feature 'class', which is the label, will be dropped during the training process.

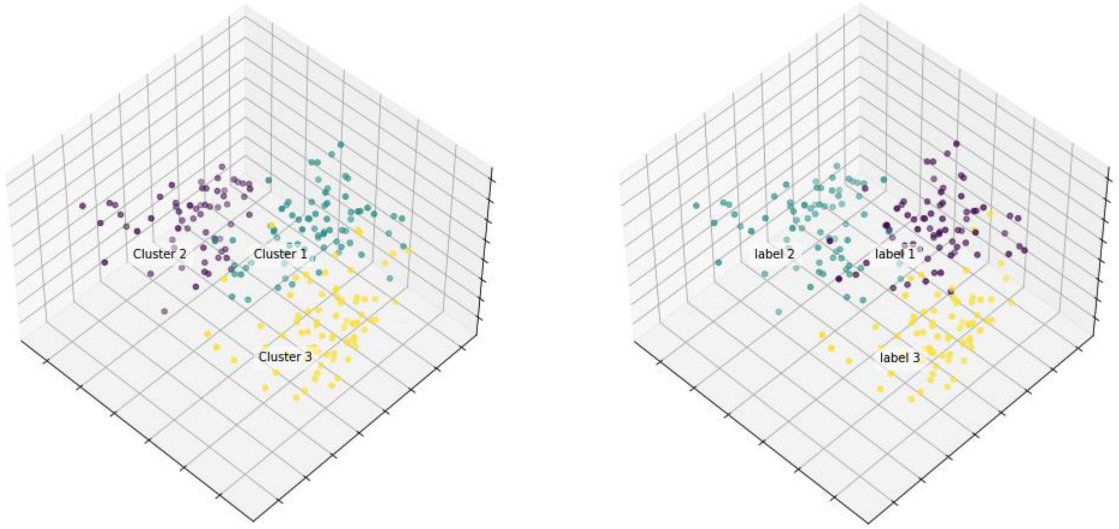
The number of clusters for the k-means model is set to 3. The result for clustering is shown below.



```
Clusters (result of k-means)
Counter({3: 77, 2: 72, 1: 61})
Ground truth
Counter({1: 70, 2: 70, 3: 70})
```

In the dataset "seeds_dataset.txt", there are 70 instances for each of the three classes. The number of instances in the three clusters is 61, 72, and 77, which seems to be a decent result.

Let's try to print it in a three-dimensional plot by using the PCA method. As presented below, on the left side is the graph colored by label class. On the right is the graph colored by the k-means clusters.



It can be clearly seen that most instances are clustered correctly. Some misclustered instances appeared close to the junction of different blocks.