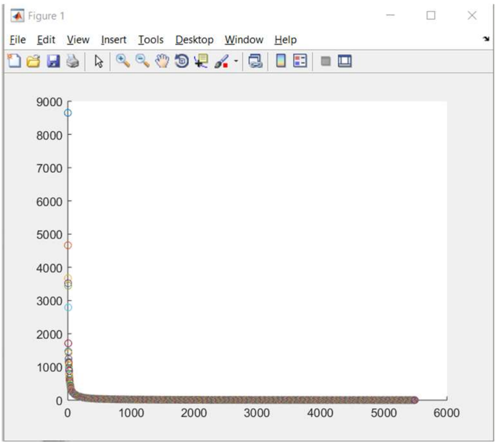
HW6

Part 1:

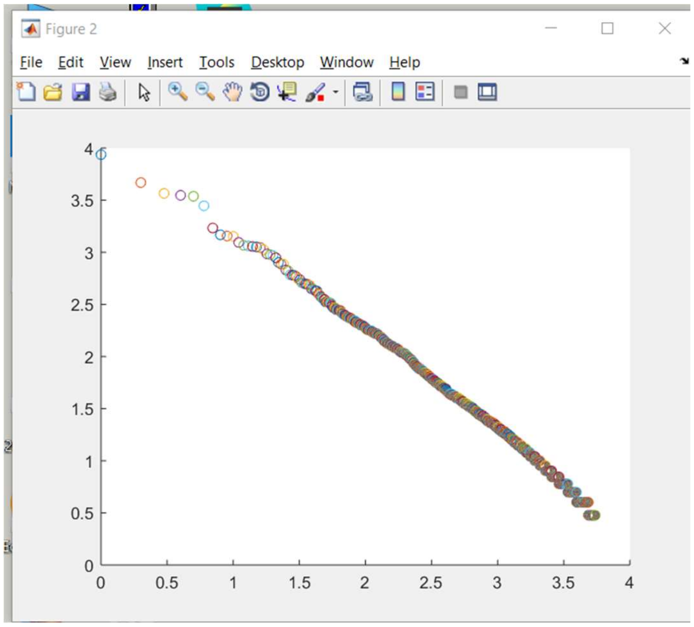
1. Words tokens: 168253, word types: 18787

2.
   1: the 8651
   2: to 4663
   3: a 3673
   4: in 3521
   5: and 3446
   6: of 2792
   7: for 1711
   8: is 1470
   9: on 1432
   10: was 1421
   11: he 1244
   12: with 1166
   13: have 1152
   14: at 1137
   15: I 1126
   16: his 1111
   17: that 1060
   18: has 965
   19: be 950
   20: but 931

3.



r on x-axis, c on y-axis



log(r) with base 10
because the log r graph is linear, it means the original data set has the rate of logarithmic function

4.

1 Ronaldo 0.732180
2 contract 0.547310
3 United 0.421776
4 Trafford. 0.386917
5 World. 0.386917
6 first-team. 0.386917
7 five-year-deal, 0.386917

```
8 knows," 0.386917
9 tomorrow. 0.386917
10 club. 0.384769
```

5. Cosine similarity of v1 and v2 using bag-of-words is

   0.4436324165581834

   Cosine similarity of v1 and v2 using tf * idf is

   0.6298192356466935. values from the two methods are not the same because bag-of-words can only extract unigram words to create unordered list of words. But tf-idf extracts words from the document and create a feature vector for the document. Thus, the two numbers are different.

6. There are two major issues: one is some unnecessary punctuations attached at the end of the words, and the other one is the differentiation of capitalization of words. Both of the features will add more inaccuracy to the word counts, which can be a potential problem for other users.

Part 2:

1. Dimension of the vector is 11, retail means = 32511.331416, horse power mean = 213.219101
2. First Eigenvector:

   [-0.27526177 -0.27353102 -0.34518922 -0.33285662 -0.3189939   0.30787157

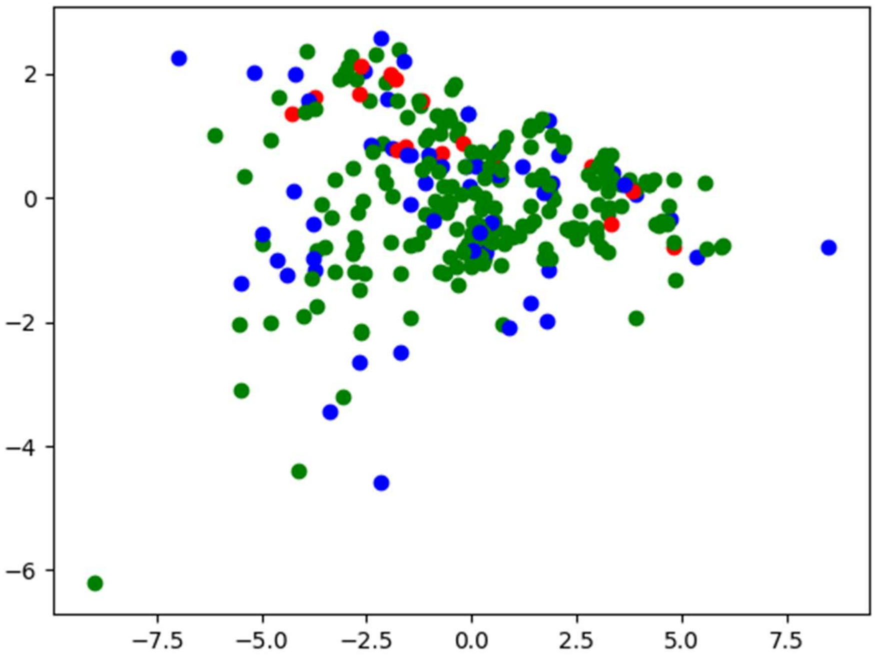
   0.30506121 -0.33520165 -0.26390318 -0.25037392 -0.29183236]

   Third Eigenvector:

   [0.25904398 0.26149657 0.06436867 0.11605896 0.09454329 0.54729793

   0.60516824 -0.11481805 0.24169352 0.31257475 0.05384033]

3. Coordinates for the first eigenvector are 6th and 7th ones. It means as PCA grows, CityMPG and HighwayMPG also grow.



4.

red is minivan, green is sedan, blue is suv

5. In the graph, which is green, sedans are clustered most strongly. Since, the first Eigenvector's greatest value is in column engine, the data shows strong correlations between the car type and engine. Because the green dots are presenting inverse correlation as it grows, it means the engine efficiency is decreasing as the PCA grows.