Import libraries

1. Project definition
2. Unsupervised learning
   1. Sneak peek of the raw data
   2. Import the entire csv file (~61 Mega Bytes)
   3. Separate users with NO selling item listed
   4. Basic statistics for “active users”
   5. From a preview, decide what to drop and what to keep
      1. sp500 = pd.read\_csv("data/sp500.csv",index\_col='Symbol', usecols=[0, 2, 3, 7])
      2. merge ‘listing’ and ‘time\_on\_site’ -> ‘listing\_per\_day’
   6. And also this is time to think about the meaning of features (columns)
3. Dimensional Reduction

Original 11 features down to 8 features

* 1. **Feature scaling**
  2. **PCA**
     1. **Select 5 features to maintain 99% of all variances in data**

1. After this step, I tried to observe the data in various ways.
   1. Assumption
      1. Three groups (+ 1 group for new sellers)
   2. Clustering
      1. K-means clustering
         1. Silhouette score – not very promising
            1. Probably only available metric when ground-truth on clustering is unknown
         2. inspect each clusters: feature inspection
            1. positive\_rating
            2. gmv
      2. DBSCAN
         1. Takes long time and results doesn’t look good
         2. In the next subsection, I tried a different clustering algorighm of DBSCAN, which views clusters as areas of high density separated by areas of low density. It is known that the clusters found using DBSCAN can be any shape, whereas k-means assumes the clusters to be convex shapes. Having negative silhouette scores in may smaller clusters found using k-means may indicate that the ground truth clusters may not be convex shapes.
      3. Hierarchical clustering (AgglomerativeClustering)
         1. Takes long time to process
         2. Results doesn’t look appropriate
   3. Data labeling with group numbers