# Cluster Analysis and Principal Component Analysis

Bryce Bowles
648 Business Data Analytics 901
10/09/2020

- **Due** Friday by 9am

**Submitting** a text entry box or a file upload

**Available** Aug 15 at 12am - Nov 24 at 11:59pm 3 months

Using the 2017 Q2 Lending Club data, perform a cluster analysis of the borrowers by income, loan amount, employment length, home ownership status, and debt-to-income ratio. Be sure to explain all preprocessing steps that you take.

1. Cluster the borrowers into seven groups using kmeans clustering.
2. Use principal component analysis to identify characteristics of each cluster.
3. Evaluate how your clusters compare to assigning applicants to clusters by loan grade (you do not need to run kmeans again for this step - simply change the visualization based on loan grade). Support your comparison with visualizations.
4. If you observe outliers in your analysis, conduct your analysis with the outliers removed.

Update 9/10/20: Identify and remove outliers before completing the analysis for 1, 2, and 3. Only report on 1, 2, and 3 for the data with no outliers.

Update 9/24/20: You only need to conduct an analysis with two principal components.

Create a report that contains a summary of the steps that you took and the results that you obtained. Do not include code in your report and avoid specific references to R - write the report as if you are reporting to someone unfamiliar with R (but you may assume familiarity with the analytics methods). Submit your code as a separate .R file so that the instructor can run it if necessary. Refer to the syllabus for more details.

**Data Preprocessing and Removing Outliers**

You will find an attached file, "4_Assignment-3.R", that describes each step in detail. The data was first loaded using the previous read in file lending_data_2017_Q2.rda. To narrow down the dataset to only what was needed, a dplyr function was used to select columns or filter out unneeded data. From there, it was essential to take a look at a summary of the data columns:

```
> summary(lend.df)
   annual_inc        loan_amnt          emp_length     home_ownership       dti        
 Min.   :      0   Min.   : 1000   10+ years:35438   ANY      :    5   Min.   :  0.00  
 1st Qu.:  48000   1st Qu.: 7000   2 years  : 9914   MORTGAGE:52502   1st Qu.: 12.23  
 Median :  68000   Median :12000   < 1 year : 9542   NONE     :    2   Median : 18.12  
 Mean   :  80452   Mean   :14589   3 years  : 8495   OWN      :11873   Mean   : 18.99  
 3rd Qu.:  97000   3rd Qu.:20000   1 year   : 7034   RENT     :41069   3rd Qu.: 24.58  
 Max.   :8900000   Max.   :40000   n/a      : 6697                     Max.   :999.00  
                                   (Other)  :28331                     NA's   :75
```

The highlighted values stand out, meaning all could possibly skew the data.

Data Preprocessing Issues Identified:

- At least one or a few individuals have a very high annual income in comparison to other quartiles, including the max of $8,900,000
- R thinks the employment length "n/a" is a character string – need to convert to numeric
- Home ownership has levels of very few observation (ANY and NONE)
- Debt to income ratio has 75 missing values and very high maximum, way higher than the 3rd Quartile

Steps that need to be taken to fix the issues:

- Annual Income: Identify, replace with the median or filter out outliers
- Employment Length: Either 1) Replace the "n/a" with "N/A" so R recognizes it or
  2) Convert "n/a"s to 1 numeric value
- Home Ownership: Recategorize ANY and NONE or remove from sample
- Debt to Income Ratio: Remove or recategorize NA's
- Creating dummy variables for qualitative variables

The process by which the outliers were identified and removed is: Outliers were first identified by looking at the summary (discussed above). Then, box plots, histograms, and scatter plots were used to that each Annual Income, Loan Amount, and Debt to Income Ratio had outliers to be removed. Each plot was very distorted meaning the data was very skewed. After removing the outliers, the data was replotted to determine if the amount removed/replaced was suffice.

A "which" statement narrowed down the Annual Income's outliers and removed about only %4 of the overall data. The loan amount needed no action taken and did not include outliers severe enough to remove. A R "dplyr" function was used to filter out the top 100 values of the Debt to Income Ratio, improving the dataset while keeping the dataset around %95 of its initial data.

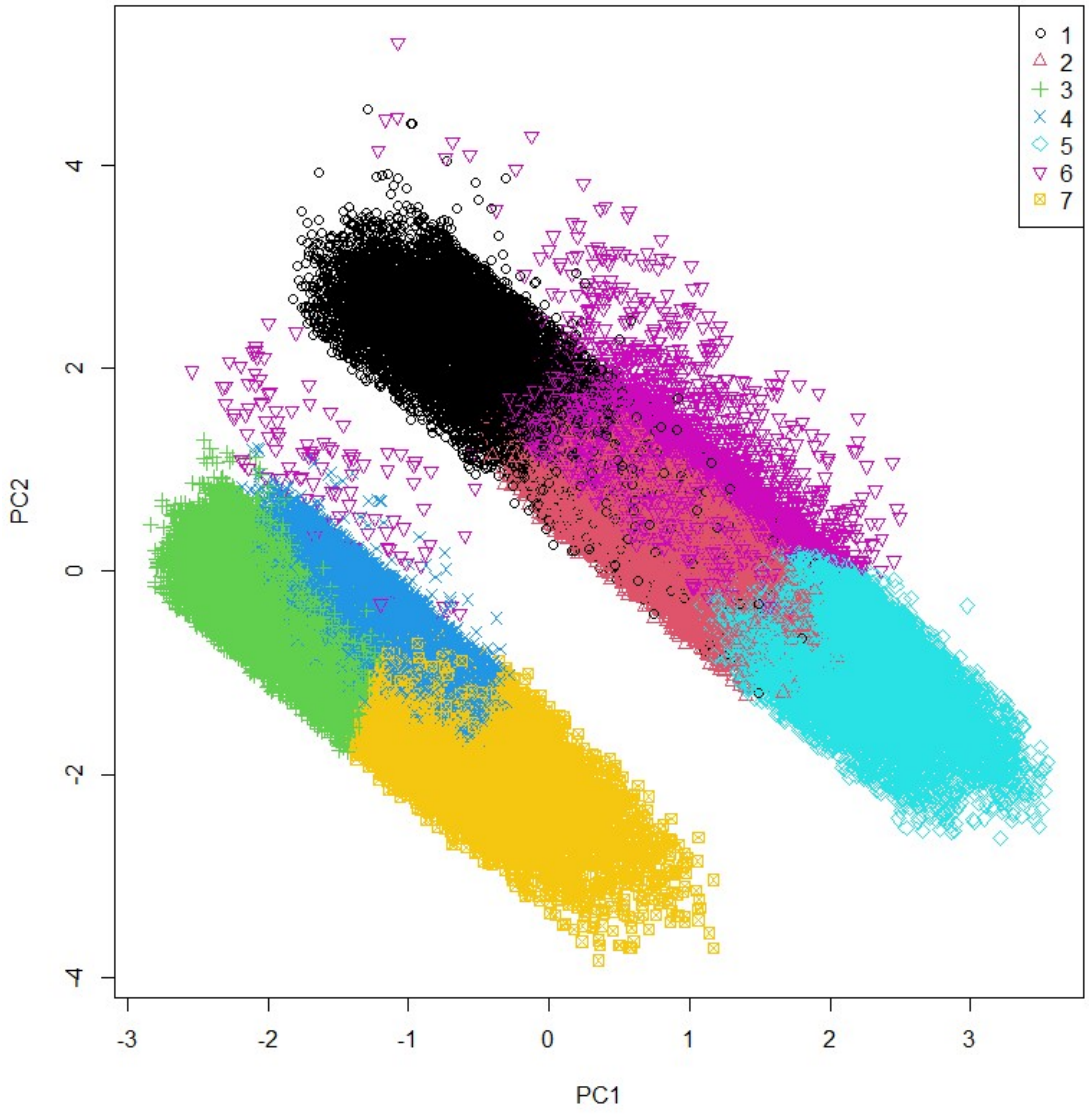
After this step, there were no longer 75 Debt to Income Ratio NA values. “n/a”s in the employment length were converted “n/a”s to 1 numeric value – using judgement to space out the years. And lastly, it was decided that ANY and NONE would be removed from the dataset.

Next, a dummy variable was created for the Home ownership variable. Since loan grade was not needing to be scaled, a new value was created for it to be used later. It was removed from the dataset before scaling. After the preprocessing steps were completed, the data was then scaled and centered.

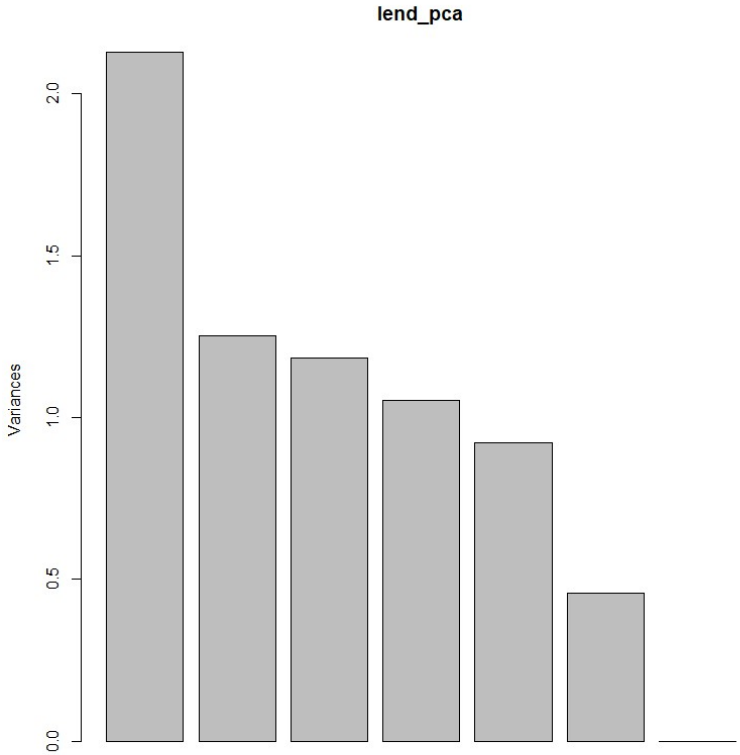
**Kmeans Clustering with Principle Component Analysis**

Kmeans was used to identify 7 groups or clusters. Many of the sample clusters can be partitioned (by color), but there are some that overlap – mainly cluster 6 with other clusters. Cluster 6 seems to overlap with cluster 2, 4 and slightly 1.

The graph is difficult to assign clusters with strict characteristics because of the diagonal shape in relation to the principle components. Principle components were looked at first to see where the load was on each. On PC1, there are larger values on home ownership mortgage (to the right of the graph) and smaller values on home ownership rent (left of the graph). Esentually, to the right of the graph, you’ll see greater values in home ownership mortgage, annual income, loan amount and a little bit of employment length. To the left, you will find larger values for home ownership rent. The values toward the top of the graph are going to include larger values in home ownership own and debt to income ratio. On the bottom of the graph, you will find larger values in annual income, home ownership rent, and loan amount. So you could make an assumption that cluster 7 has larger values in annual income, home ownership rent, and loan amount.




```
> lend_pca$rotation
                              PC1         PC2
annual_inc              0.39210394 -0.47251453
loan_amnt               0.35100896 -0.38952149
emp_length              0.24240552  0.08769482
home_ownershipMORTGAGE  0.58855695  0.10986487
home_ownershipOWN      -0.06338752  0.55558379
home_ownershipRENT     -0.55959032 -0.47169051
dti                     0.02698501  0.27217102
```


lend_pca
Variances
0.0
0.5
1.0
1.5
2.0

**How your clusters compare to assigning applicants to clusters by loan grade**

When comparing to loan grade, there is no distinction between the classifications, but you can see two different cluster shapes.