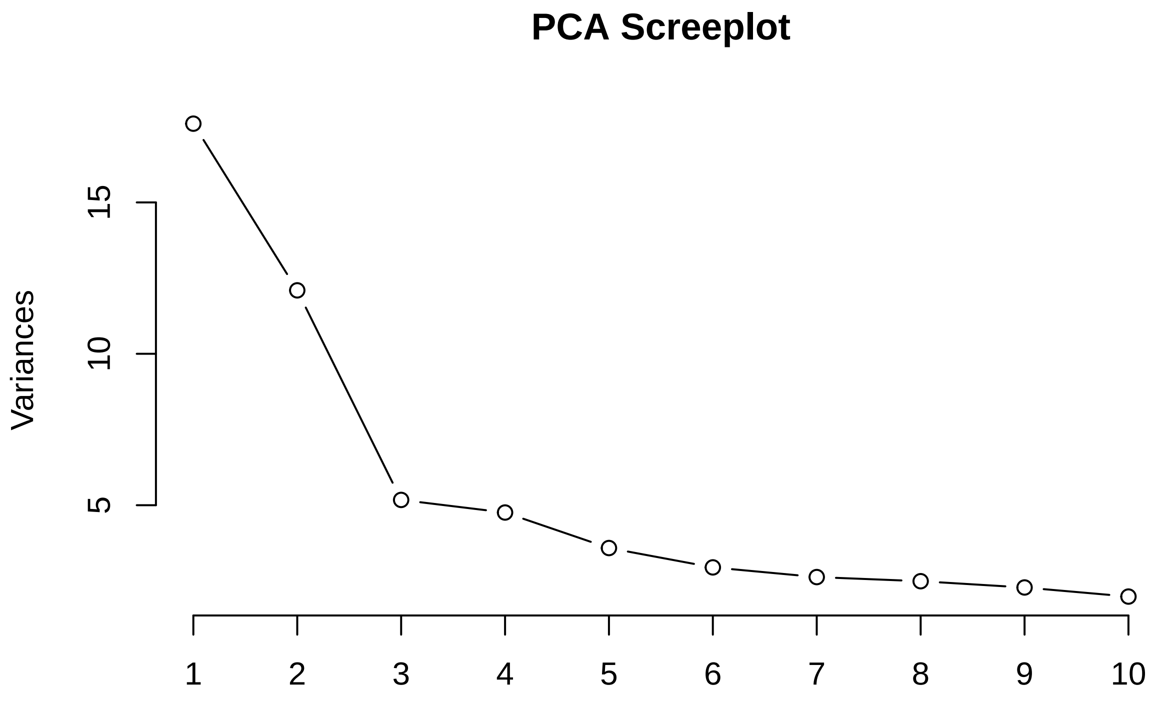
# eda_final_proj

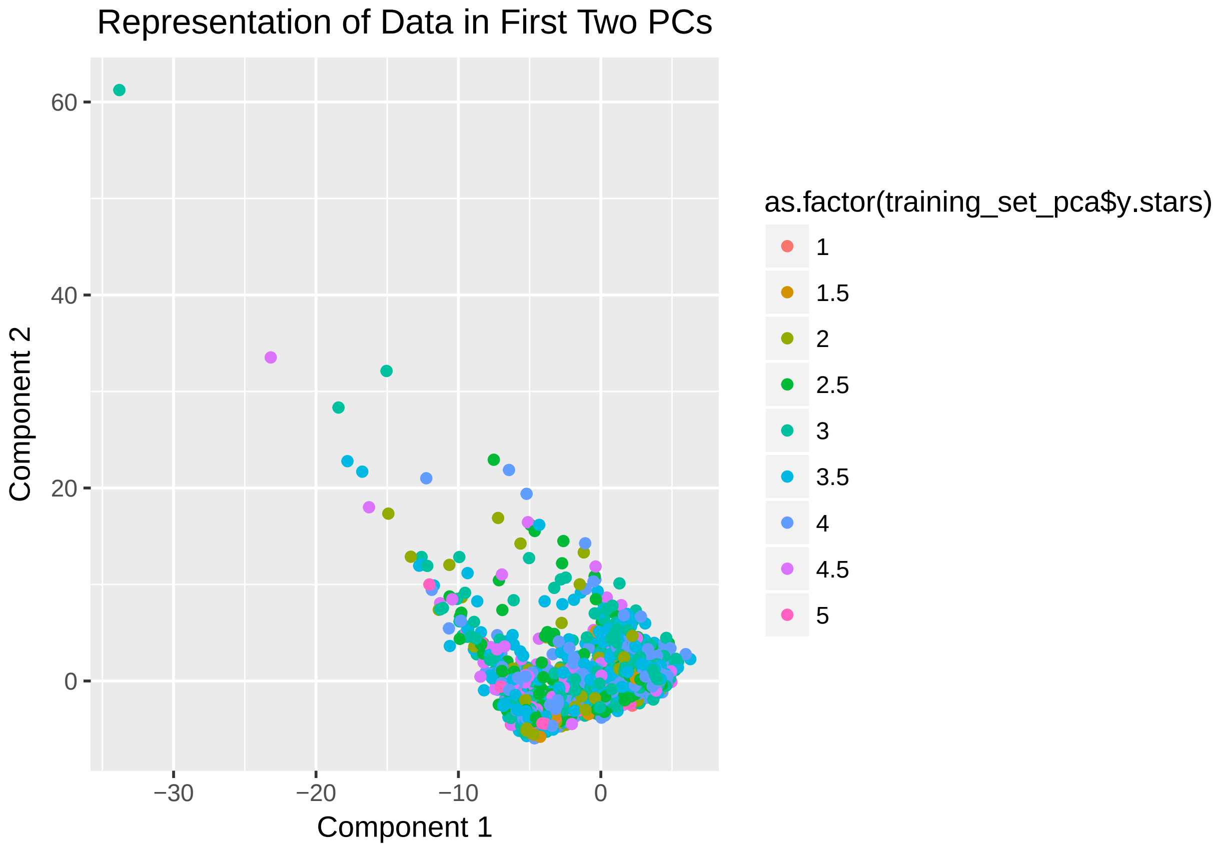
*Lydia Maher*

*May 4, 2017*

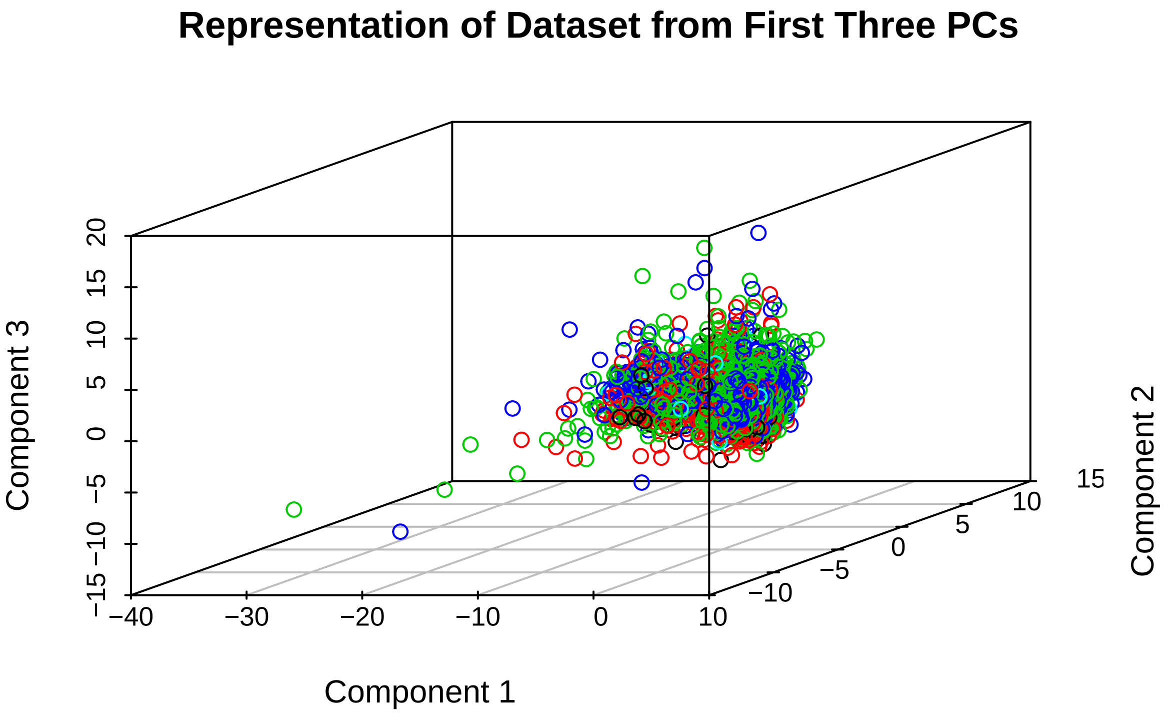
```
## Loading required package: ggplot2
```



Running PCA reveals that about 31% of the variance of this dataset can be explained with the first three principal components. Similarly, the first 10 component account for 51% of the variance. This means that not all of the variables are that predictive and our dataset is somewhat sparse. However, because 10 components still only account for about 50% of variance, this dataset might not be the best candidate for dimensionality reduction.



Plotting the two first principal components against one another reveals one concentrated cluster to the bottom right, with two sprawling streams of outliers to the right and left of this cluster. This suggests that there are a lot of fairly similar reviews, but that in the not-so-similar reviews, there is a lot of variance. Particularly for higher star ratings, there tends to be a lot of variation (we such much less 'pink' values in the center of the cluster and these colors correlate to star ratings of 4.5 and 5). This makes sense as we would expect businesses which are either really good or really bad to have quite different attributes -it wouldn't make sense if ALL restaurants described as 'fantastic' were also Mexican restaurants.



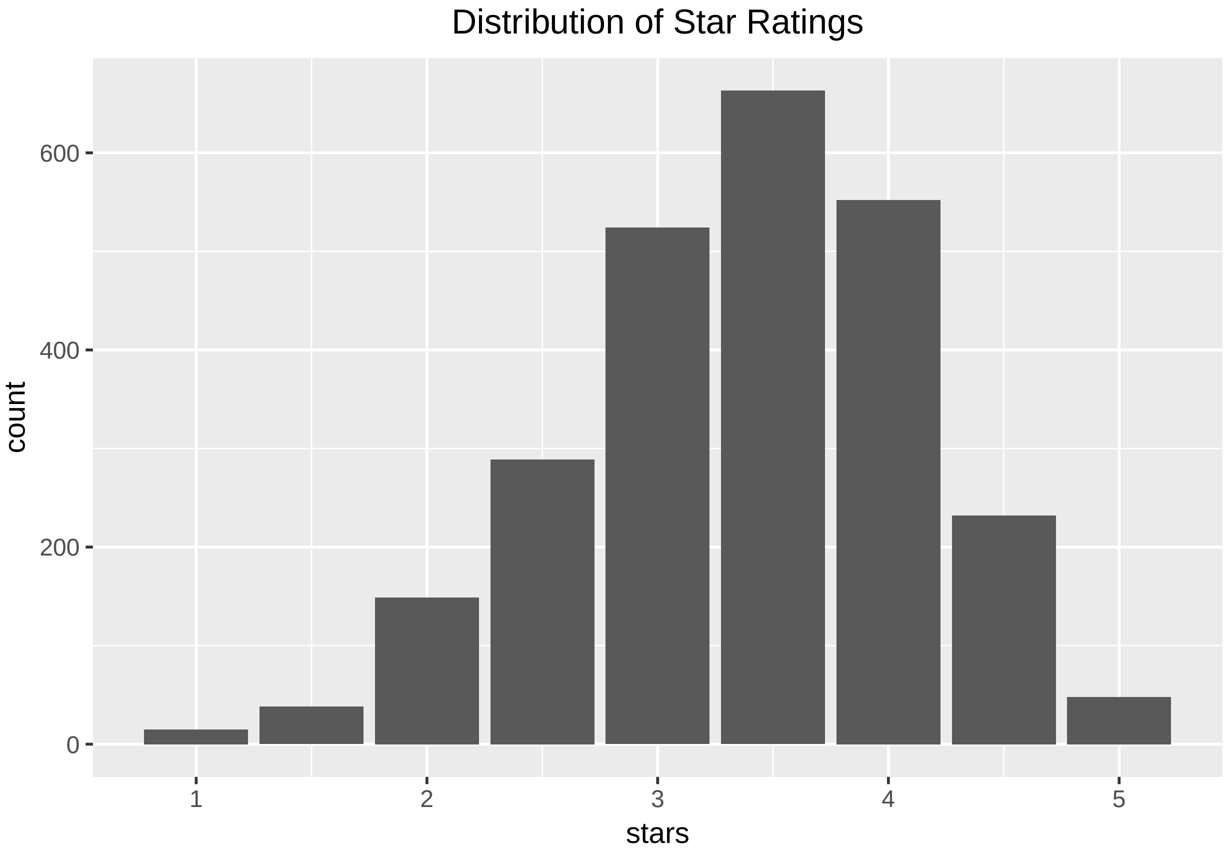
As shown above, these clusters and streams of outliers still exist in a 3D scatterplot of the first three principal components.

```
## : romantic: false   intimate: false   touristy: false    upscale: false
##          4.177273          4.175031          4.174780          4.172936
##    hipster: false     classy: false
##          4.150843          4.150153
```

Next, we looked at what attributes most contribute to the variation in these components. The table above shows that restaurants which are *not* self-labelled or labelled by Yelp as romantic, intimate, upscale, touristy, classy or hipster lead to the most variation in reviews.

```
##         out        more     y.stars      pretty      i_love    is_great
## 0.001536273 0.003054396 0.006166752 0.006270192 0.008689448 0.011323888
```

On the other hand, if restaurants are described as pretty, great, with “i love”/“more”/“out” or reviewed with similar amounts of stars, they tend to be more similar to each other (in other words, they account for less variation).



Finally, we briefly explored our response variable of star ratings across businesses. The star ratings have a mode of 3.5, with the majority of star ratings being between 3 and 4. Generally, the star ratings are skewed slightly towards the upper values of y. Looking at this same information in a boxplot, we can see that the whiskers do not extend below 1.5 stars. Moreover, reviews of 1 star are marked as outliers.