Credit Sesame Data Analysis and Findings

Hari Rajan, Paul Giraud, Audreya Metz, and Calvin Ma

September 27 2018

Dataset

- The data was from Credit Sesame, a company that calculates credit scores to determine options for credit cards, mortgage rates and loans.
- This dataset included:
  - User Demographics
  - First Session Information
  - 30-day User Engagement Data

Our Research Direction

- We wanted to give Credit Sesame an insightful description of their user base
- We believe that they would have an interest in the credit score of their users, since this helps the to predict the best loans, mortgages, etc.

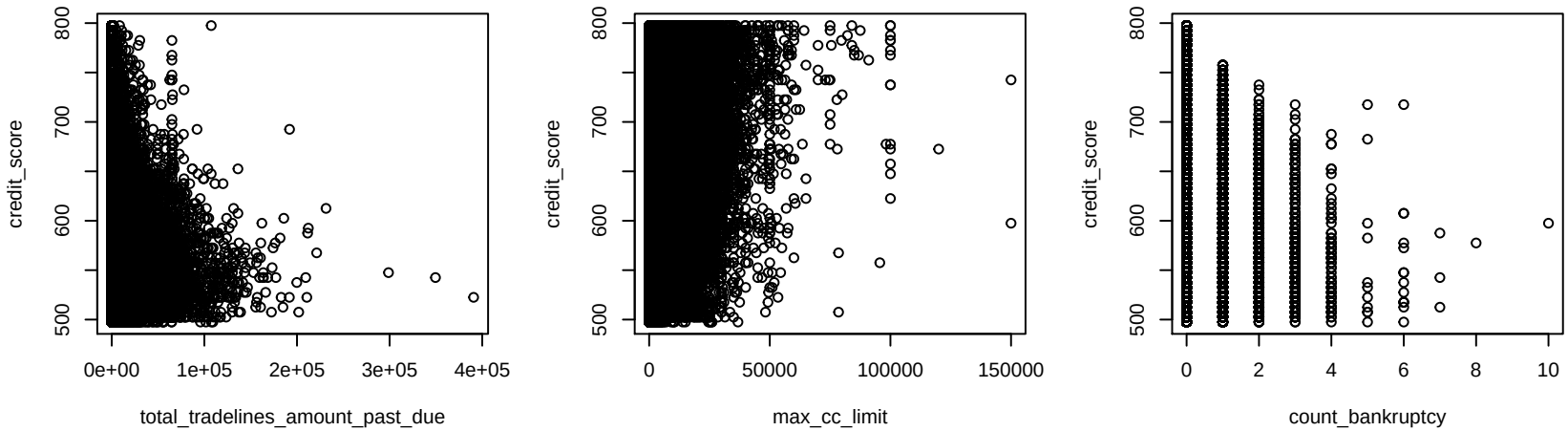
Exploring the Dataset

- We started out EDA by looking at the distribution of the variables in the datset
- We found a significant right skew in the majority of the non-categorical data

This was important to keep in mind as we proceeded in our analysis as the distribution of each variable can impact the results of our analysis methods.

Exploring the Dataset

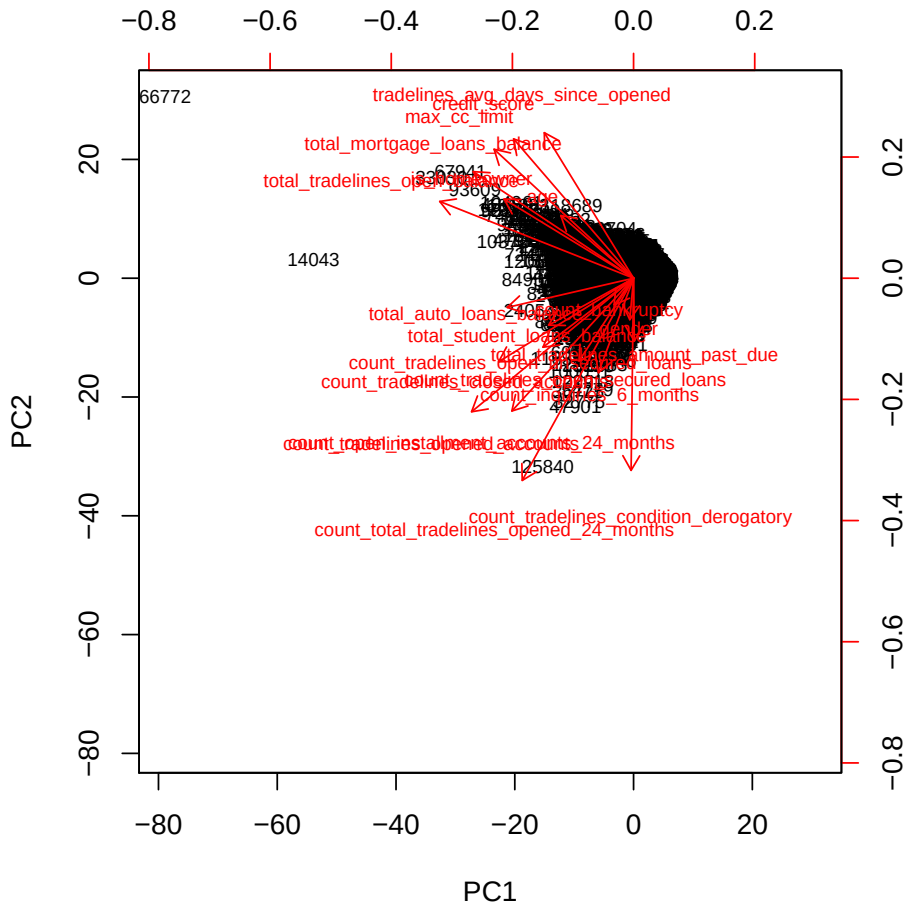
We also wanted to see if there was any obvious relationship between credit score and the others features in the user demographics.



Several of the features had a clear linear relationship with credit score, which motivated a linear regression to determine which features had the greatest impact on credit score. Other features didnt have such an obvious relationship.

Exploring the Dataset

We ended our EDA using PCA to determine any relationships between data points. Due to the high-dimensionality, the biplot is hard to interpret, but we can still obtain some valuable information from the plot.



Challenges of the Dataset

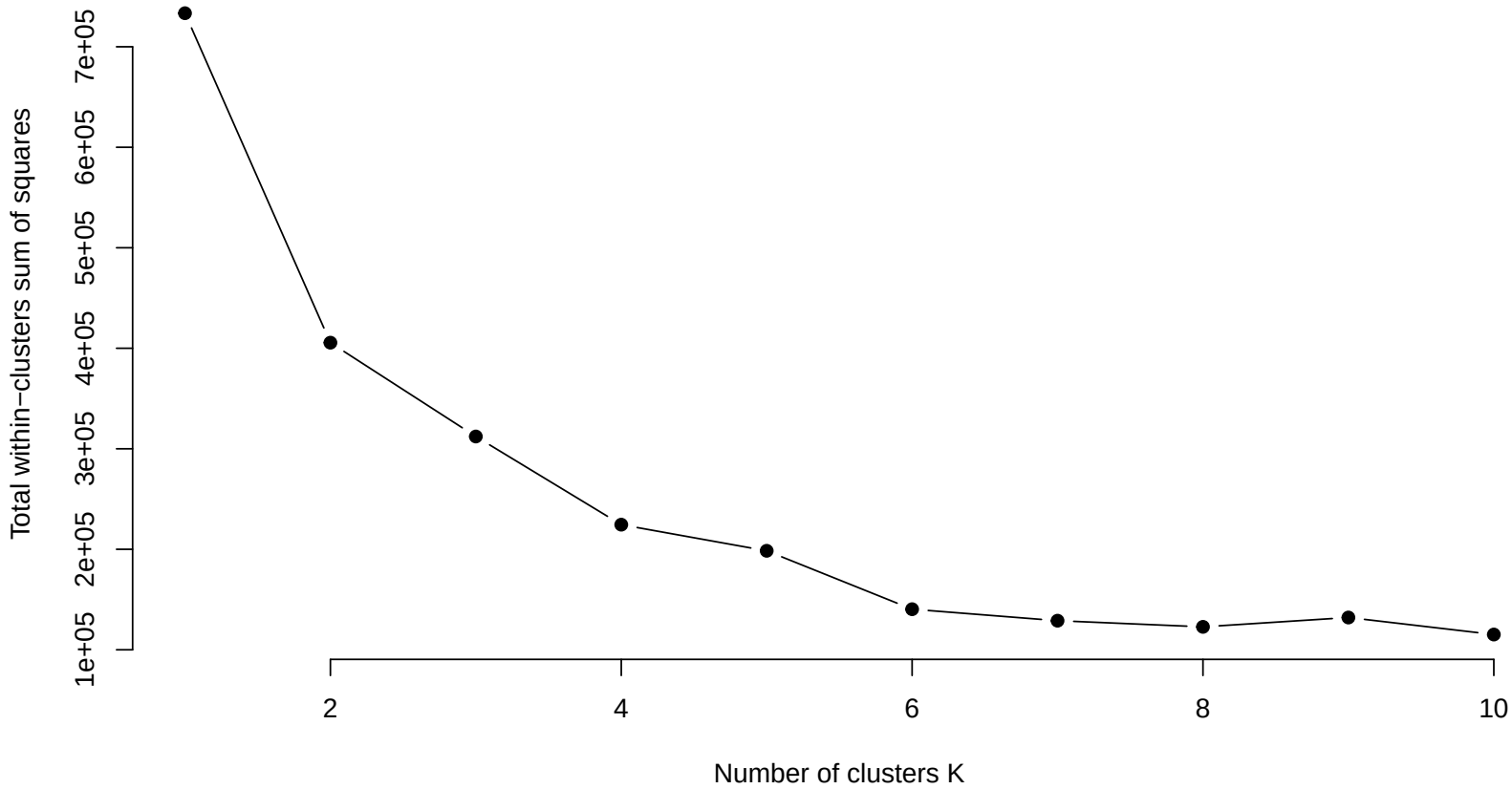
- The skewed distribution of our variables was the biggest challenge for determining how to analyze this data.
- Ultimately we did not let the skew effect our methods but were careful to consider it in our analysis.

Methods

- Linear Regression
  - We determined which features to use in the regression using the plots and PCA
  - Only achieved an R-squared of .35, which is too low to draw meanginful conclusions
  - COuld use other regression methods such as Lasso in the future
- K-means clustering
  - Cluster users based on most frequent page type
  - Compare within cluster user demographics with those of other clusters
  - Find differences between users

K-means

- We use k-means because it is the quicker than hierarchical clustering, but we first need to define k. We look at the plot of k vs. WSS.



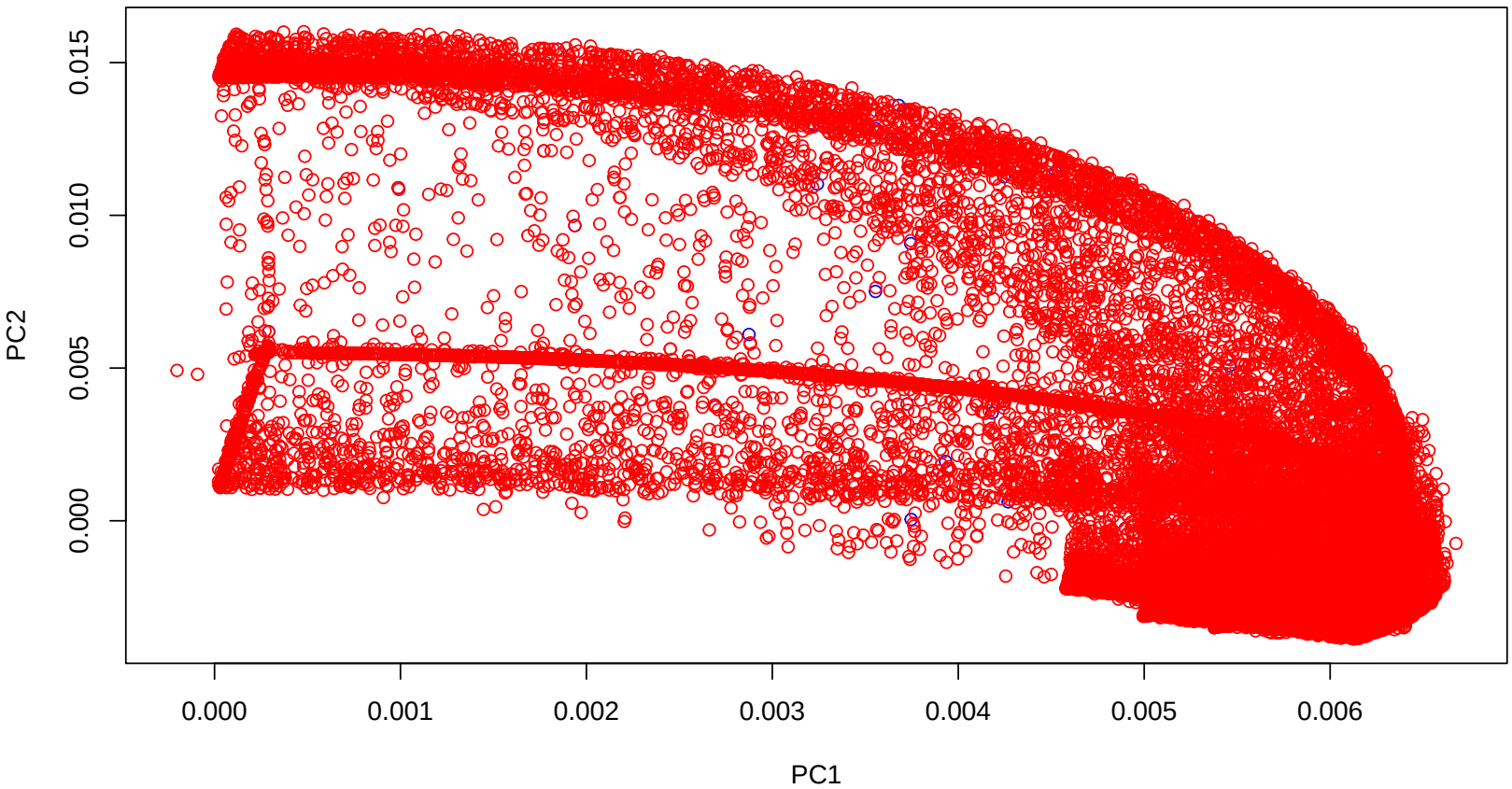
We use the Elbow method to determine the optimum number of k.

Results

- After determining the number of clusters to use and accordingly performing the k-means clustering method, we noticed that in particular, there were two clusters which had relatively high values for `click_count_credit_card` and `click_count_personal_loan`.

```
##   click_count_credit_card click_count_personal_loan
## 1               0.7028986                6.74327122
## 2               5.2224913                0.15986159
## 3              14.9978541                0.20600858
## 4               0.0000000                0.01661813
## 5               0.2266731                2.40265451
## 6               1.3833033                0.03594182
```

Results (continued)



Results (continued)

```
##                                                click_credit c
## count_open_installment_accounts_24_months     0.8966165
## count_tradelines_open_unsecured_loans      1259.0300752
## total_open_cc_amount_past_due              4040.3308271
## total_tradelines_open_balance              4173.3778195
## total_auto_loans_balance                   9858.4041353 1
## credit_score                                656.7199248
```

Conclusions

- The k-means clustering allowed us to confirm that the demographics of people who use Credit Sesame's website differently and access different pages are in fact different
- There is a distinction between Credit Sesame's website clientele that uses the credit card features and those who use loan features
  - The credit page users tend to have lower credit scores and the loan feature users tend to have higher loan amounts based on grouped findings
  - The largest difference between credit score users and loan feature users are open collection accounts, # of derogatory tradelines, max credit card limit, total credit card balance, and total mortgage loans and balance, which all happen to be strong indicators of credit score

Conclusions

- We can work to improve the current models that we used
  - Regression model can be improved using methods such as Lasso or ElasticNet to facilitate feature selection
  - Hierarchal clustering could have provided more insight into potential variable groups