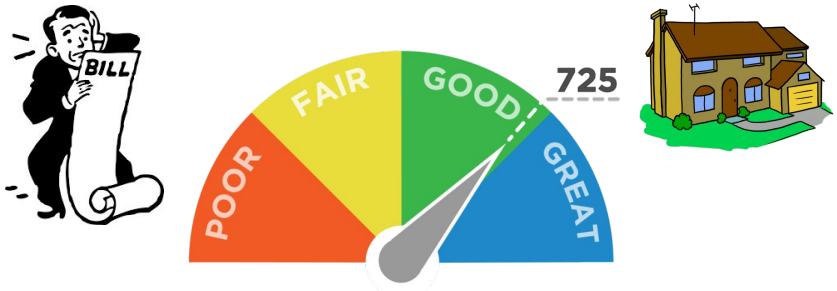
Credit Sesame Data Analysis

Team 99

Our Questions

- Who should we give loans to?
  - Can we predict a quality credit score based on other available data?

- What demographics of people displays certain engagement habits?

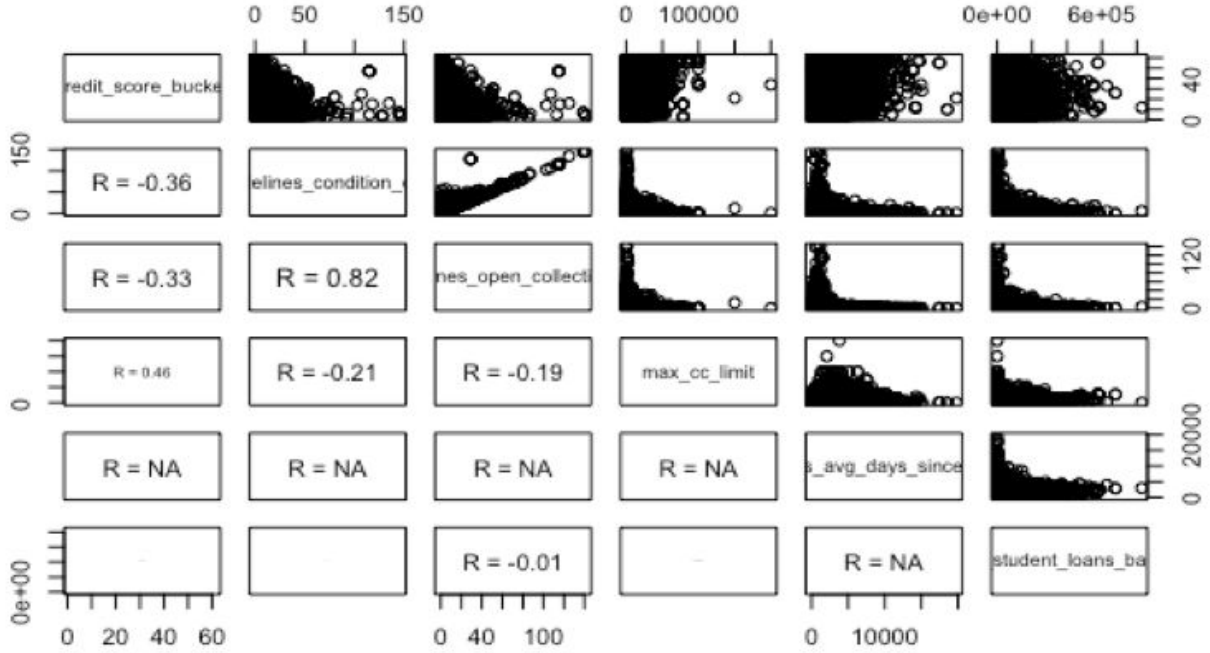


Question 1: Who should we give loans to?

To explore this question, we…

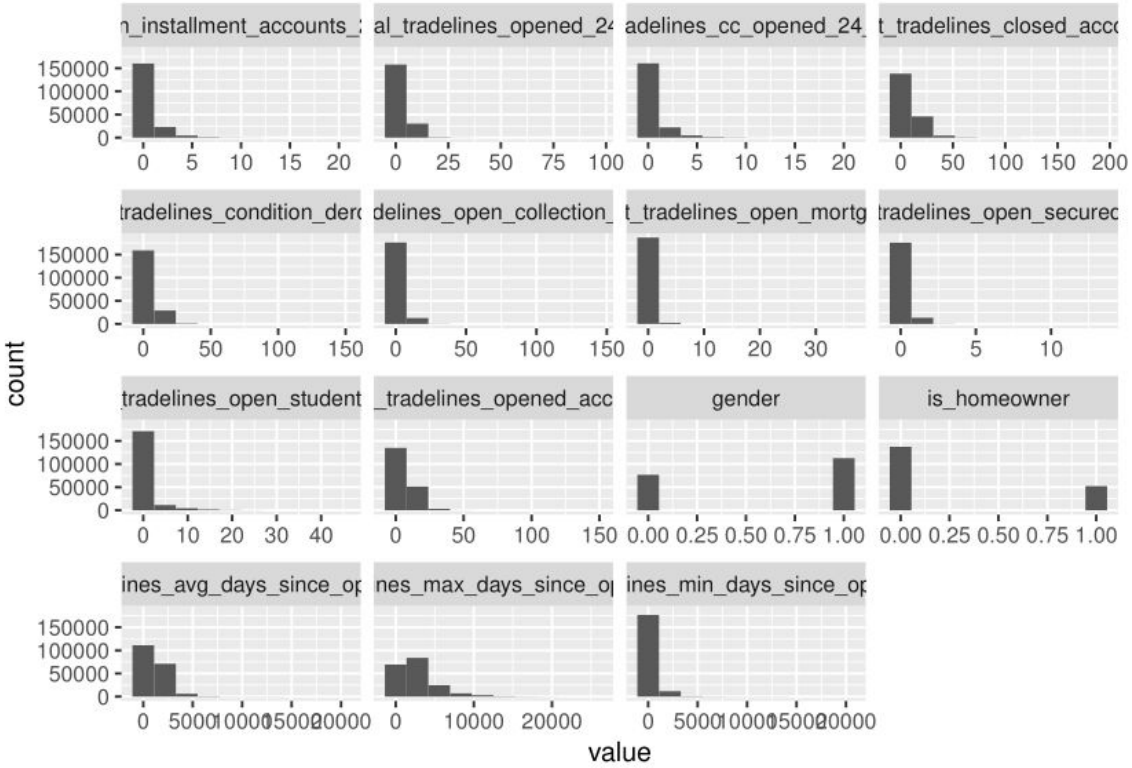
- Performed exploratory data analysis investigating correlation between key payment indicators and demographic information
- Performed PCA with the given data sets to find borrower qualities that are correlated with delinquencies
- Created a linear regression to allow us to predict quality of credit score based on demographics and these key payment indicators

Exploratory Analysis

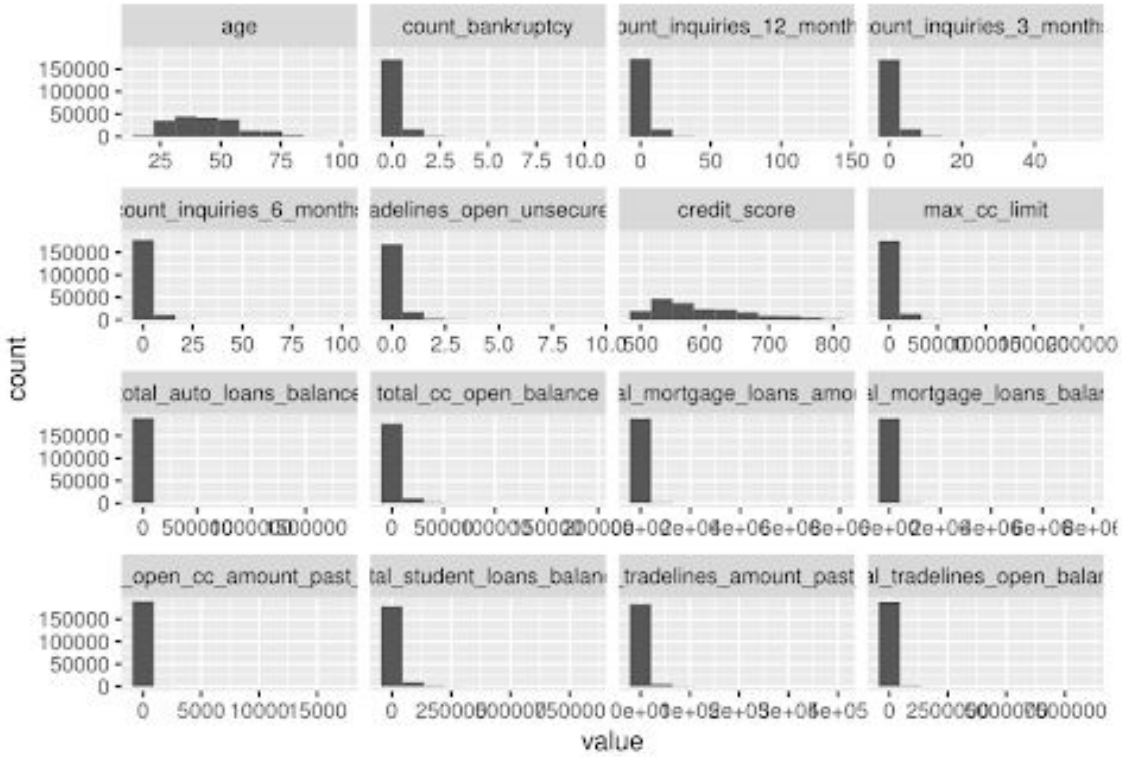


We concluded our exploratory analysis by looking at the plots of our candidate variables that we believed could be too correlated. We were ultimately able to narrow down our predictive model to four variables. A strong correlation was found here between count tradelines open collection amounts and count tradelines condition derogatory, resulting in the removal of count tradelines condition derogatory to rectify collinearity.

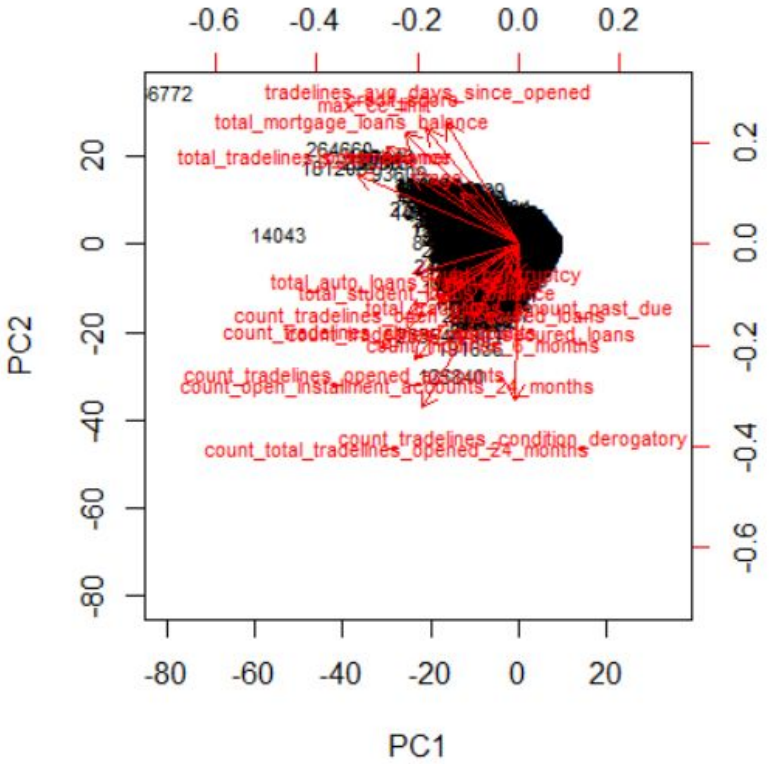
Exploratory Analysis



Exploratory Analysis



PCA Results



To the left, we have a biplot of the datapoints with loadings from Principle Component 1 and 2. We can see that all of the loadings in PC1 are negative. However, PC2 has a stark contrast in positive and negative loadings for different features. We used PC2 to determine which features were correlated with each other - we concluded that people who can be considered good clients for a loan are younger, own homes, have high mortgage loans, high credit card limits, low number of derogatory accounts and others.

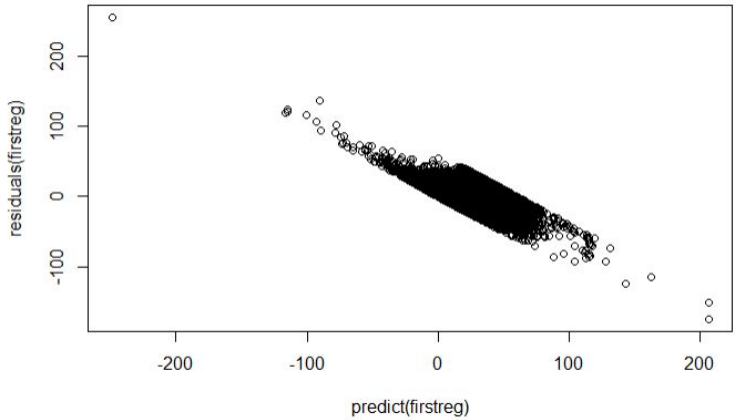
Linear Model Results

```
Call:
lm(formula = credit_cat ~ home_cat + count_tradelines_open_collection_accounts +
    max_cc_limit + tradelines_avg_days_since_opened, data = user_profile)

Residuals:
     Min       1Q   Median       3Q      Max
-173.589   -9.141   -1.967    7.520  254.818

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              1.577e+01  3.965e-02  397.79   <2e-16 ***
home_cat                                 3.962e+00  5.325e-02   74.41   <2e-16 ***
count_tradelines_open_collection_accounts -9.482e-01  6.159e-03 -153.95   <2e-16 ***
max_cc_limit                             9.073e-04  4.441e-06  204.31   <2e-16 ***
tradelines_avg_days_since_opened         1.427e-03  2.486e-05   57.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.57 on 272813 degrees of freedom
  (12673 observations deleted due to missingness)
Multiple R-squared:  0.3227,    Adjusted R-squared:  0.3227
F-statistic: 3.25e+04 on 4 and 272813 DF,  p-value: < 2.2e-16
```

Using homeowner's status, count tradelines open collections accounts, maximum credit card limit, and tradelines average days since opened, we were able to account for 32.27% of the variance in credit scores, meaning that these variables serve as a reasonable predictor of quality of credit score and therefore the reliability of the person we could potentially give a loan. The way this regression model can be interpreted can be seen in out write-up.

Linear Model Analysis and Limitations

Our linear model is reasonably good given our data and the difficulty of creating real-world model with a high R-squared value, but because of the skew of the given data, our residual plots do not have random distribution about y=0, as would be desired.



As a result, this model needs to be used with caution.

Question 2: What demographics of people display certain engagement habits?

- We used k-means clustering (with k=8) to cluster the following user_engagemenet actions
  - Click_count_credit_card, click_count_personal_loan, click_count_mortgage, click_count_credit_repair, click_count_banking, click_count_auto_products

Cluster means:

	click_count_credit_card	click_count_personal_loan	click_count_mortgage	click_count_credit_repair	click_count_banking	click_count_auto_products
1	0.0000000	0.01596800	0.002106417	0.000000000	0.0004803613	0.0003472946
2	0.7028986	6.74327122	0.050724638	0.161490683	0.0300207039	0.0372670807
3	0.1107261	0.08127264	0.017483068	1.224287289	0.0015750512	0.0053551740
4	8.6406606	0.20159453	0.042141230	0.068906606	0.0068337130	0.0034168565
5	0.2262686	2.40307348	0.021210181	0.091643989	0.0112854170	0.0116055707
6	1.2674882	0.03158102	0.010081279	0.004129523	0.0016649187	0.0016518091
7	3.8917506	0.11940997	0.025677047	0.047920081	0.0062436588	0.0031998751
8	18.2112069	0.22413793	0.030172414	0.073275862	0.0000000000	0.0215517241

Question 2: What demographics of people display certain engagement habits?

- We then identified the users from the two most substantial clusters, one with a high mean for click_count_credit_card and the other with a high mean for click_count_personal_loan, and analysed the summary statistics of the according user_profile data
  - In comparing the user_profile data for the two groups, we noticed many striking differences.
    - Ex. The mean credit score for the personal_loan cluster wads 656 whereas the mean credit score for the credi_card cluster was 595

Conclusion

Being a homeowner, your number of open collection accounts, your maximum credit card limit, and the average days since your account has been open, are the most significant indicators of credit score and future delinquent payments. We expected open collections to have a strong predictive relationship with delinquencies and credit score since open collections are caused by missed payments. Our results indicate that each account turned over to a third party for collections leads to a 5 point decrease in credit score. Homeowners tend to have a credit score that is 20 points above non-homeowners. This makes sense because homeowners tend to be more financially stable.

Every dollar increase on a credit card limit tends to indicate a 0.0045 point increase in credit score. This is probably explained by the fact that people with better credit scores who are more financially stable are given larger credit card limits. We also found that with every day an individual has an account open, his credit score increases by 0.007135 points. This last finding is also logical because accounts that stay open are accounts that have gone longer making payments.

Using our findings and these characteristics, we can fine tune credit scores as well as estimate someone’s score.