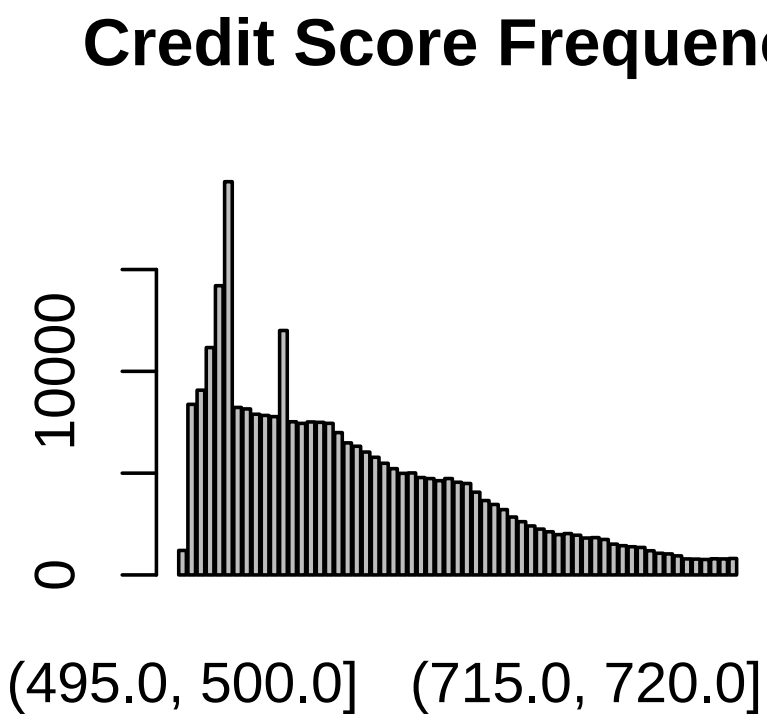
Datathon Presentation

Carolyn Chen, José San Martin, Michael Tan, Man-Lin Hsiao

11/7/2018

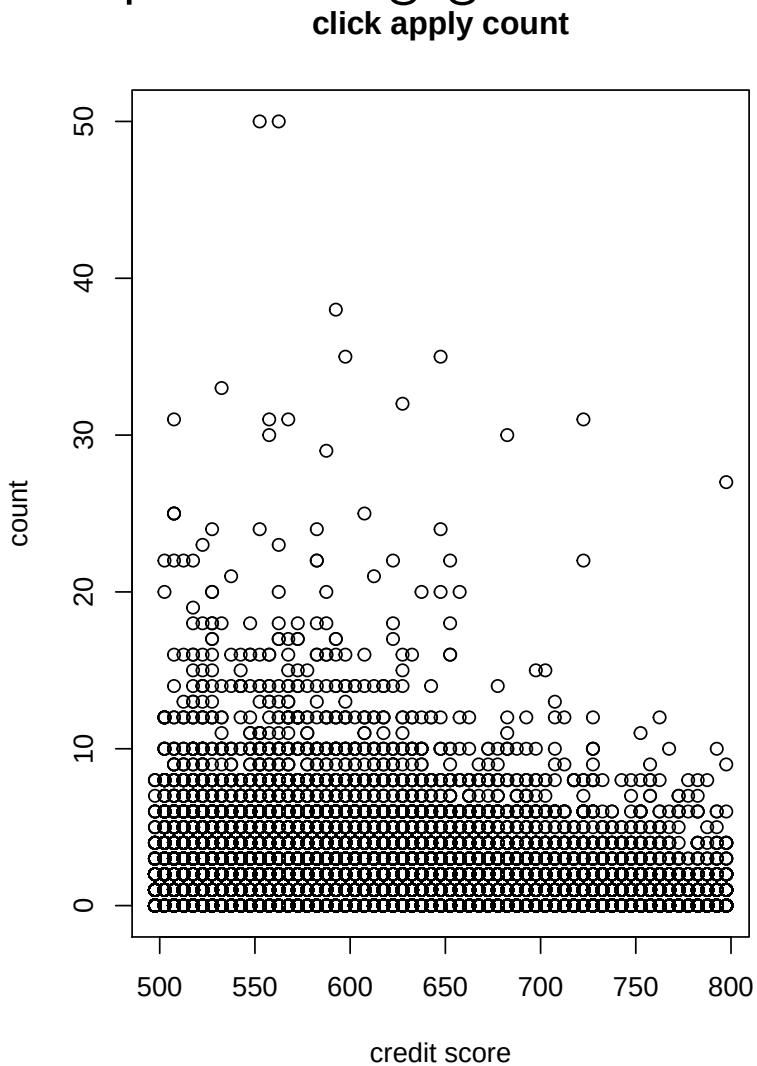
Data Set Visualizations

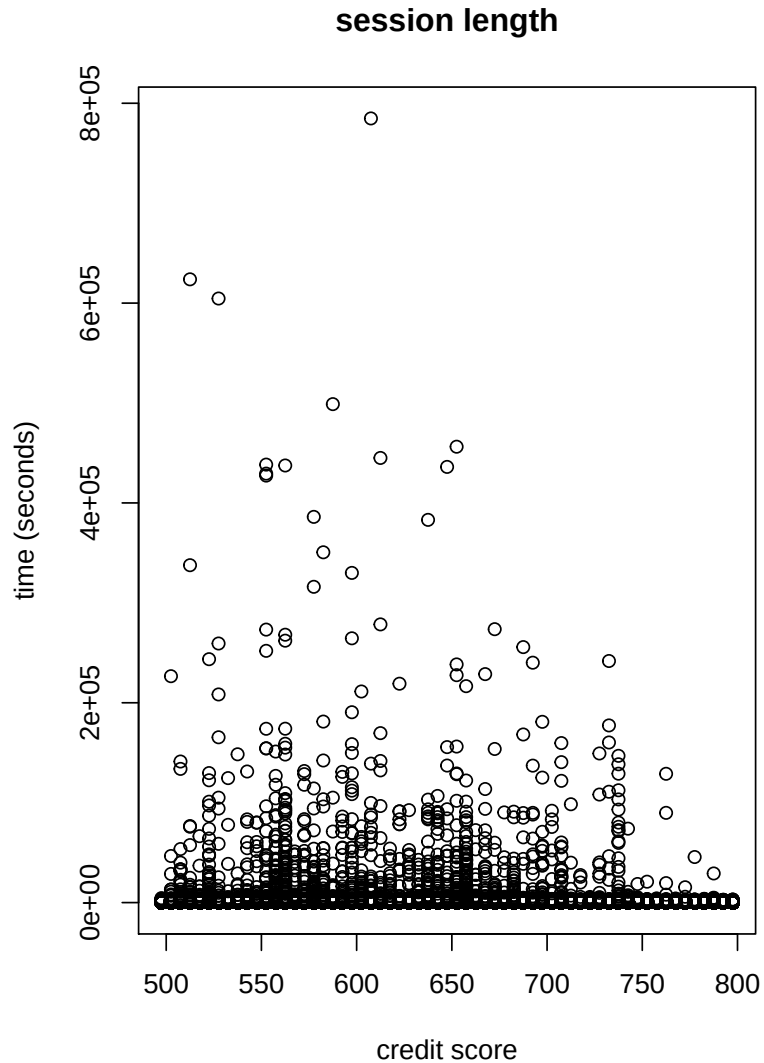
- Credit Sesame is a credit and loan-management platform
- Datasets: User Profile, First Session, 30-Day User Engagement
- Data cleaning for ease of visualization
- Histograms, Dot Plots, Violin Plots, Choropleth Maps
- First, we wanted to understand the demographics of Credit Sesame users



Exploratory Data Analysis (cont.)

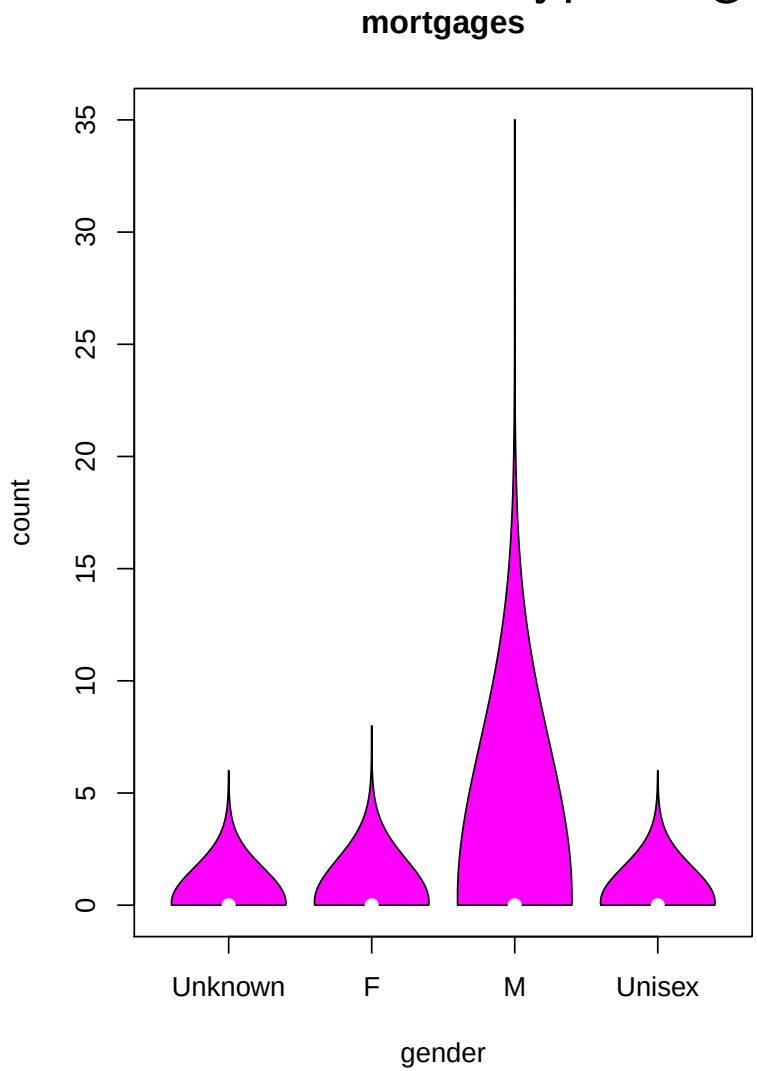
- Dot plots of engagement stats versus credit score

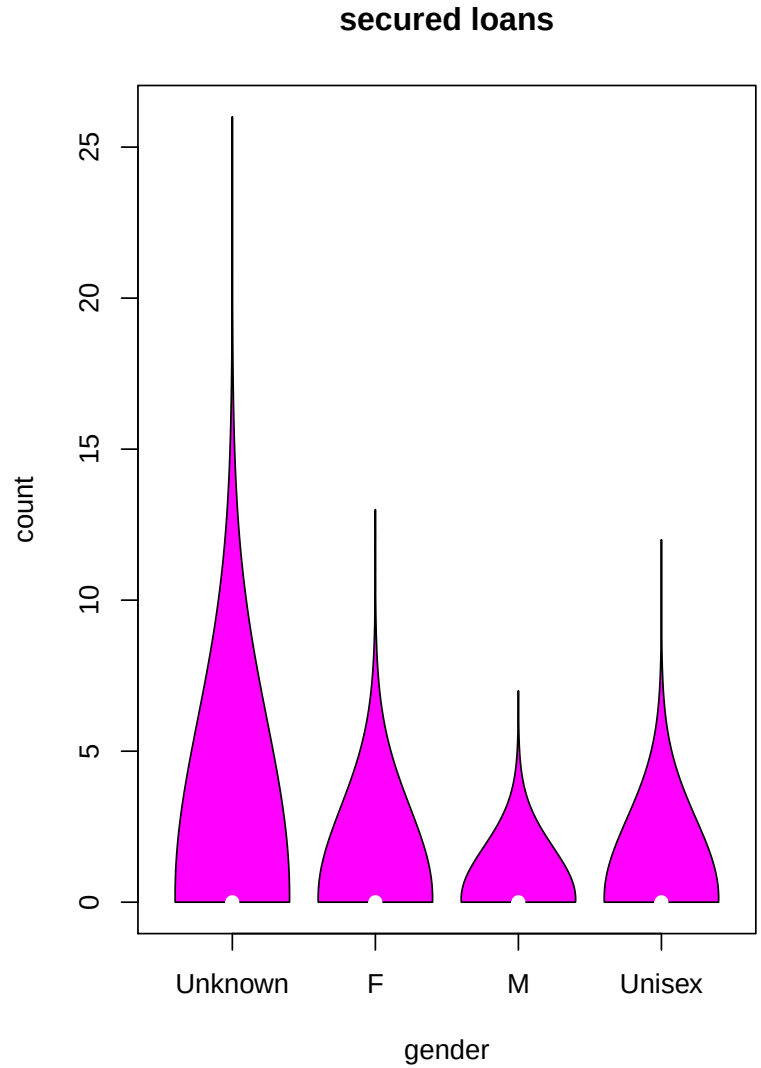




EDA Visualizations (cont.)

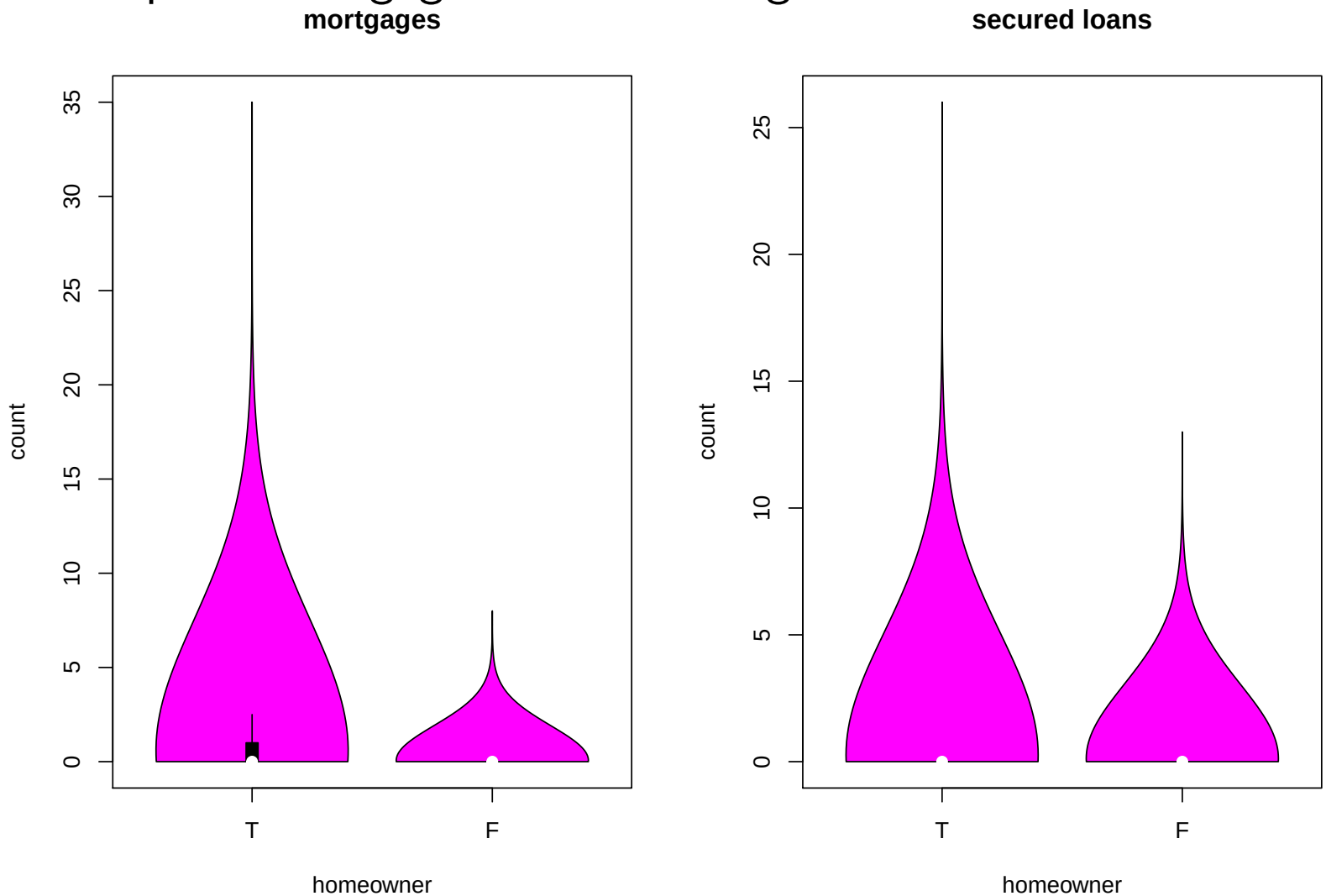
- Violin Plots of loan type vs gender





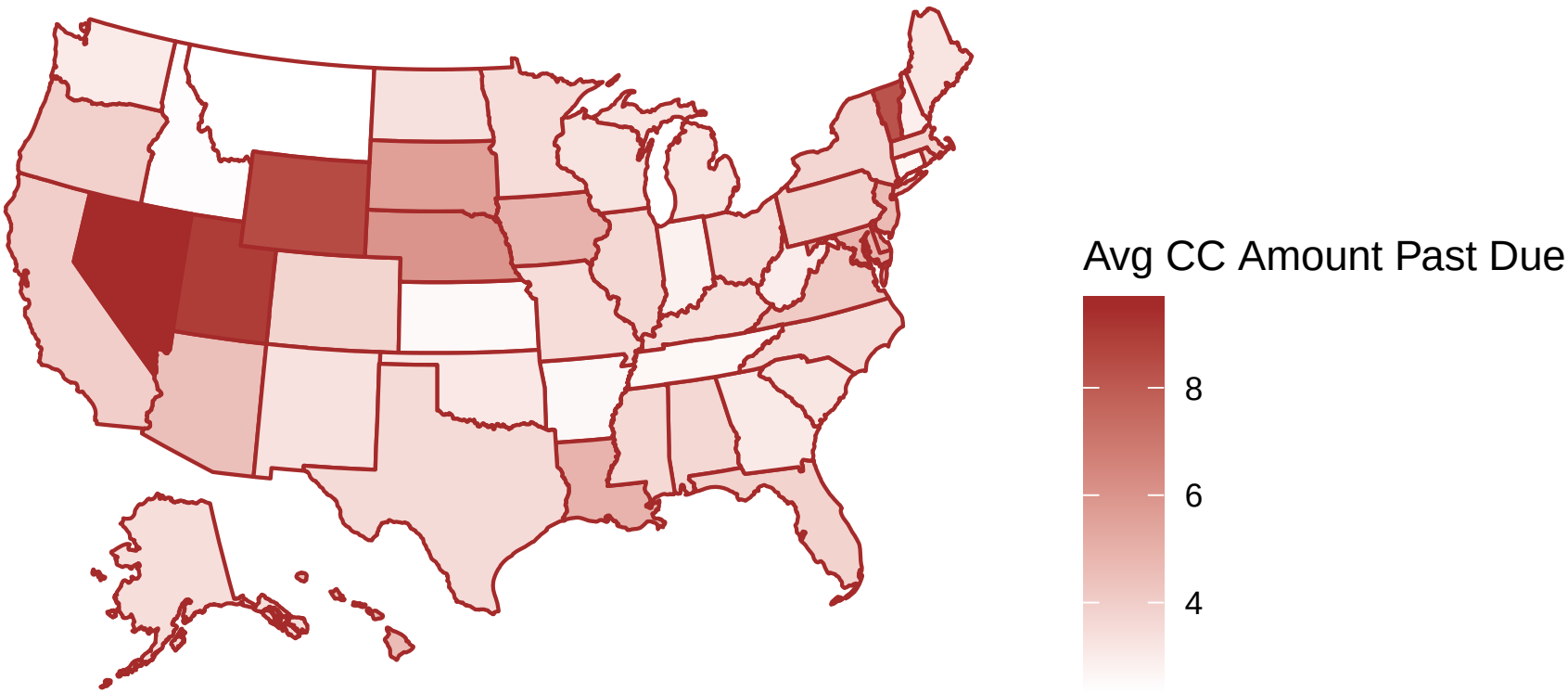
EDA Visualizations (cont.)

- Violin plots of engagement stats vs gender



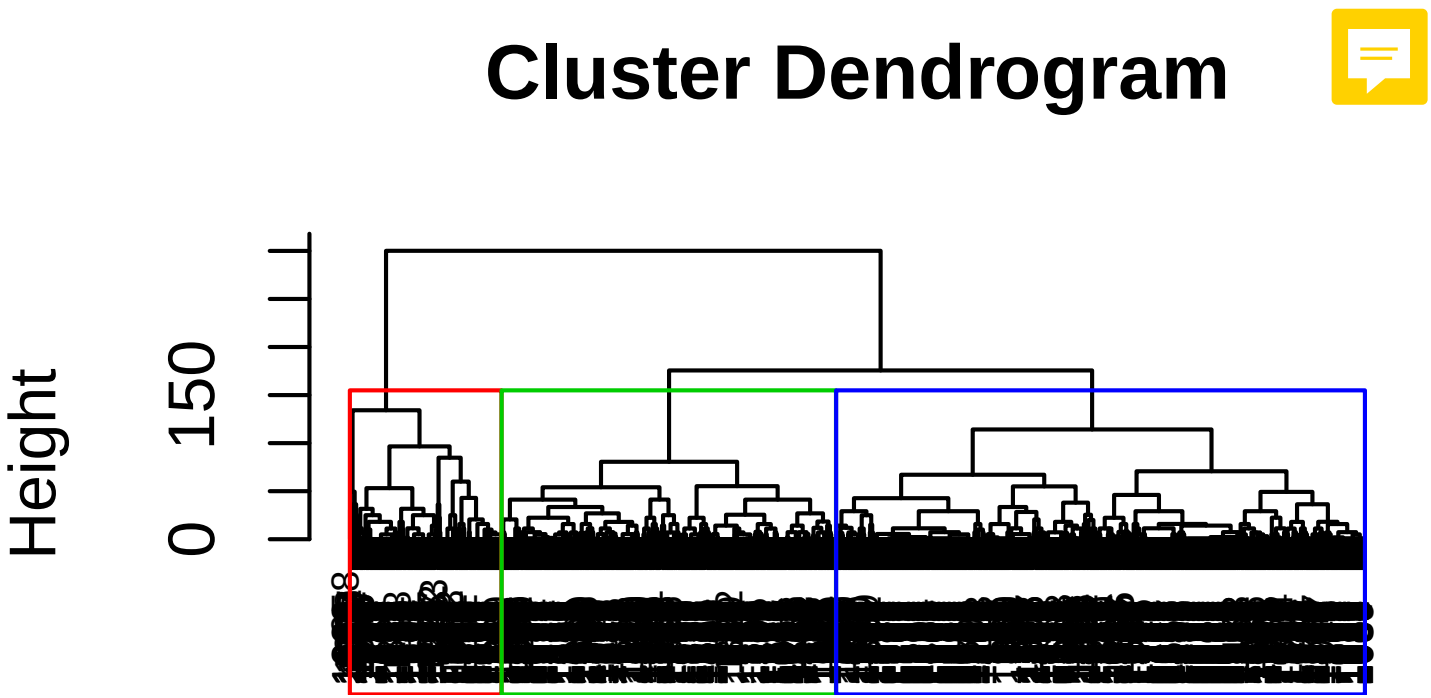
Visualizations (cont.)

- Cholorpleth Map shows us geographical distribution of credit card debt trends
- Deliquency: user has missed 2 consecutive payments
- What are profiles of deliquent vs. non-deliquent users and within levels of deliquency?

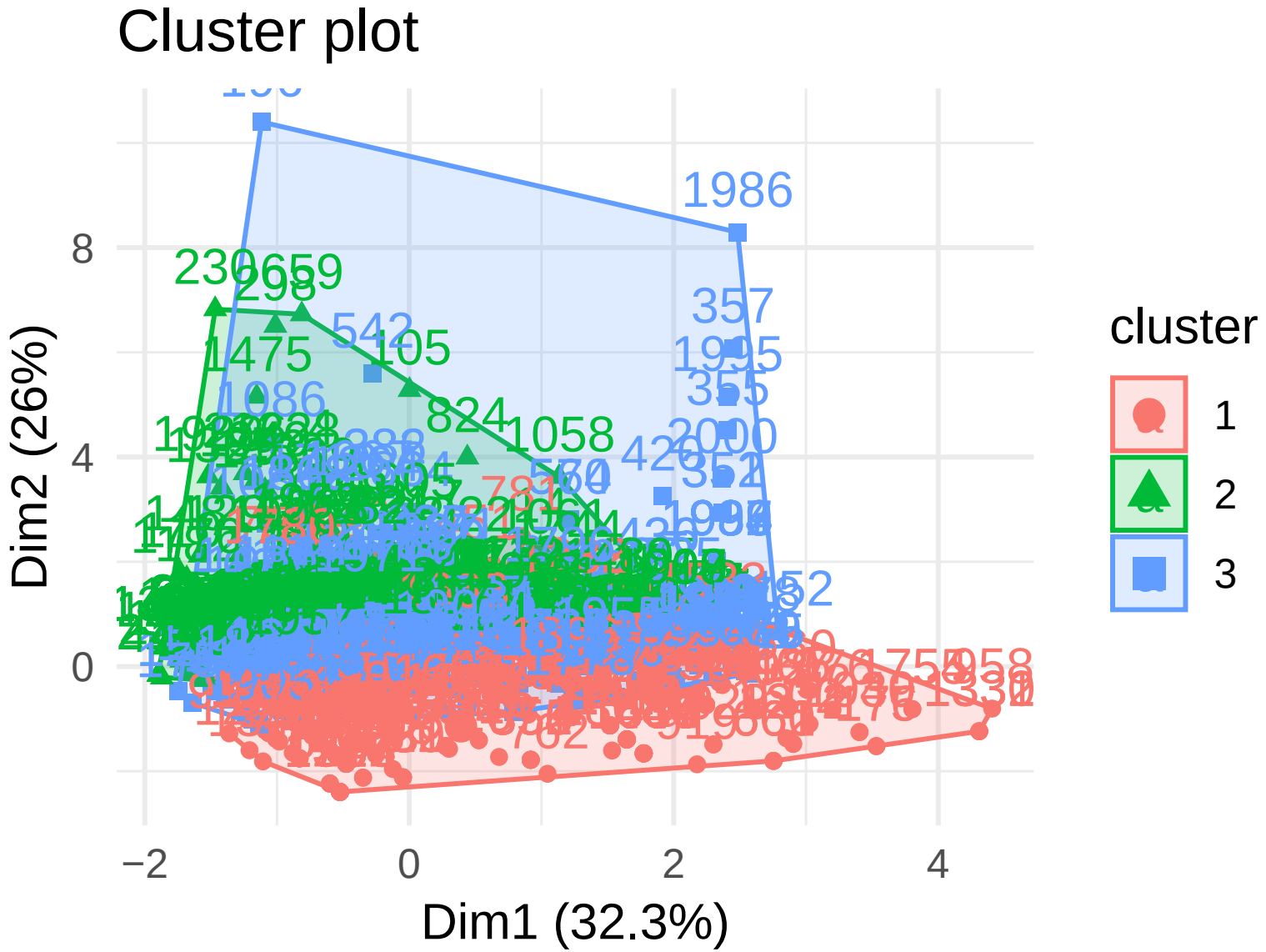


Clustering Model

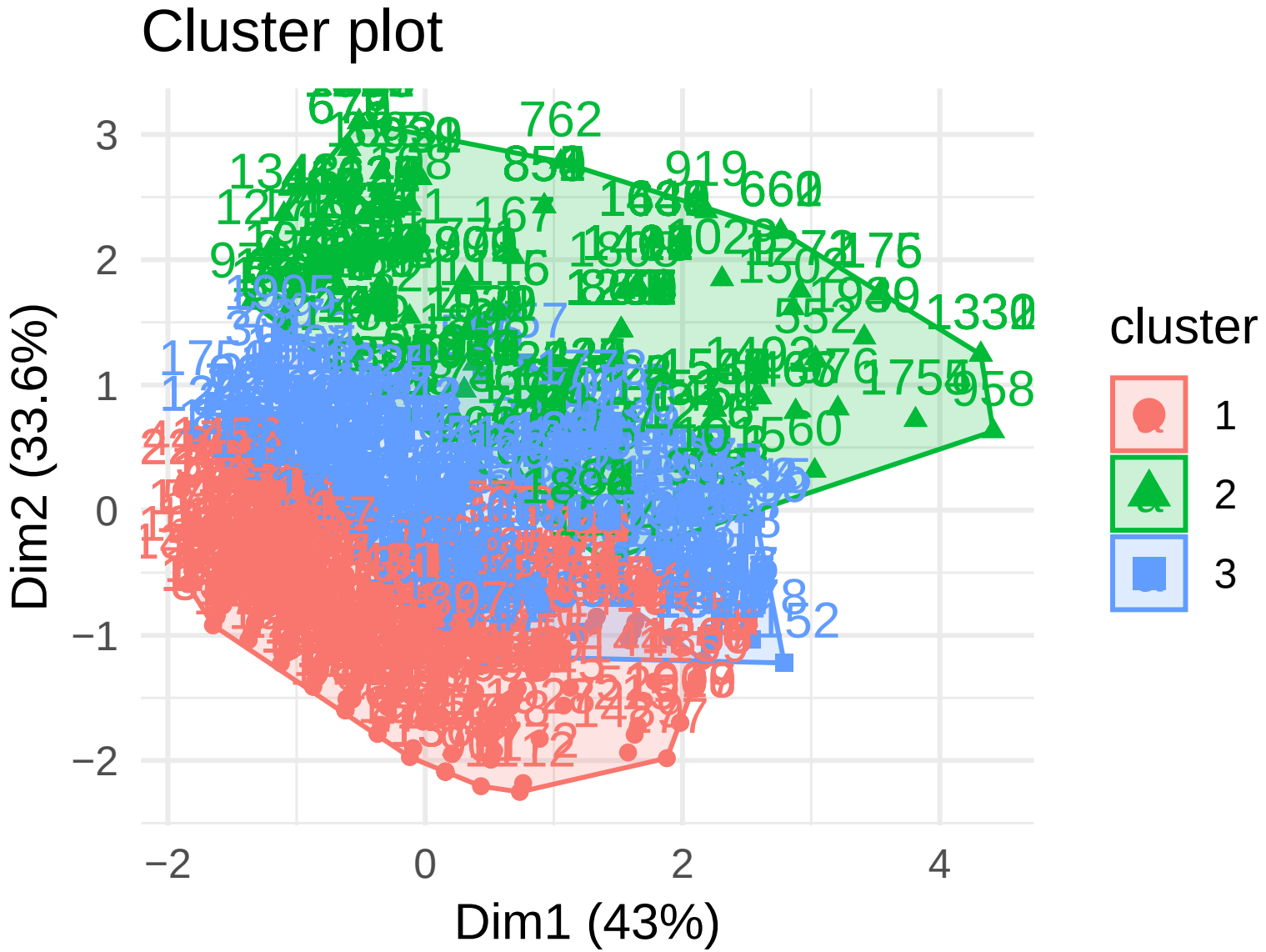
- K means clustering to identify similarity of deliquent vs non-deliquent users
- Looked at users with 'none' (0), 'some' (1-2) and 'many' (>2) deliquencies
- Users with 'some' deliquencies still quite similar to those with 'many'



K-means Visualization



K-means Visualization



Problems with Clustering Model

- The variables we chose were arbitrary and the hierarchical model was too naive.
- It wasn't a good method to use with the rest of our analysis because it didn't tell us anything substantial.
- We did not split into training/testing sets and use Cross Validation to check the model.
- We arbitrarily cut our trees at an unmotivated point.

Poisson Model

Derogratory tradelines~ age + gender + creditScore + credit card utilization ratio + auto loans balance + student loans balance + mortgage balance + auto loan x student loan + auto loan x mortgage

- Age: For every 10 years we add to a user, we expect the number of derogatory accounts to change by a multiplicative factor of e(10*9.863e-03) = 1.1. This shows age is not a significant factor.

- Credit Score: For every additional point to the credit score a user has, the expected number of derogatory accounts changes by a multiplicative of 0.99.

When we increase the credit score of a user by 100 points, the expected number of derogatory accounts changes by a multiplicative factor of 0.36

Conclusions

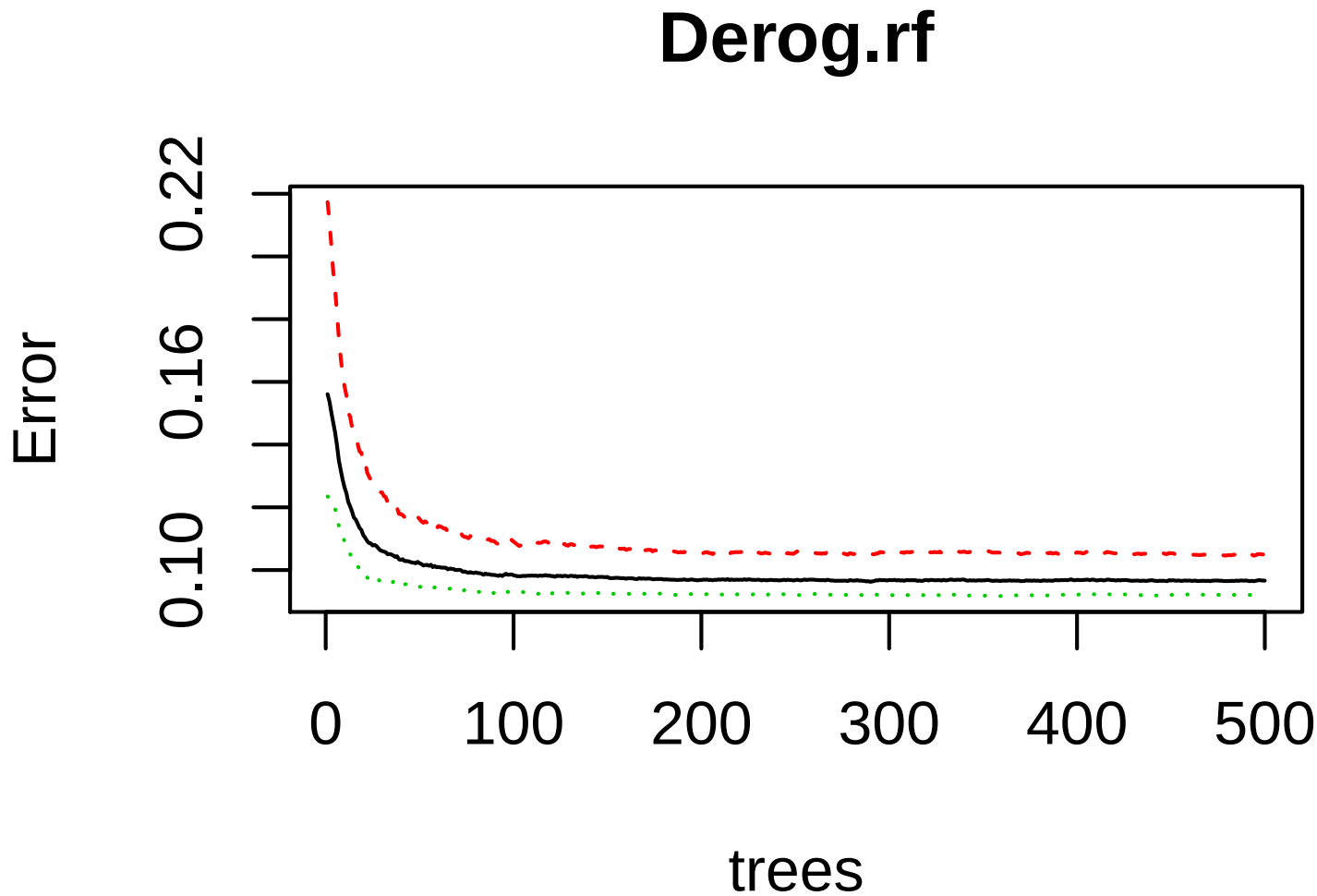
- Users with non-zero derogatory accounts tended to have similar characteristics regardless of how many of those accounts they had. Once one deliquency noted, intervention should occur to stem further ones.
- Age should not be a significant predictor of user derogatory behavior.
- Credit score is a strong predictor of derogatory behaviour, but user base of CS is also right-skewed.
- States with most past due credit card accounts are Nevada, Utah, Wyoming and Vermont. Could target reminders to people from those states.

Next Step: Random Forest Model

- From our EDA we could already tell that there was a difference in profile between people with and without derogatory accounts, and the next logical step would be to create a predictive model
- More appropriate model given we have response variable for derogratory variable

Random Forest Model

- Accuracy of about 90%, obtained a pretty small out-of-bag error rate



Random Forest Model

- Now we look at other values from our fitted Random Forest model
- Using a classifier to determine if a person has obtained a derogatory account, or not at all. We concluded that having even just one derogatory account is cause of concern for the bank.

```
##       0     1 class.error
## 0 30639  3597  0.10506484
## 1  5911 58254  0.09212187
```

```
##                                                MeanDecreaseGini
## is_homeowner                                             138.4077
## tradelines_avg_days_since_opened                        1215.7038
## tradelines_max_days_since_opened                        1284.2902
## tradelines_min_days_since_opened                        1093.7759
## count_tradelines_closed_accounts                        3212.7053
## count_total_tradelines_opened_24_months                  580.0851
```