Building a Better Understanding of Credit Sesame's Customer Base

Julia Donheiser (leader), Naman Agarwal, Ziqi Deng, Vincent Lin

November 2018

Abstract

The 2018 Machine Learning Datathon at Duke University focused on Credit Sesame, a website that provides users with free estimates of their credit scores. Using three datasets from the company, we delved into user demographics and long-term engagement patterns to better understand Credit Sesame's customer base—and how to earn them more revenue. We particularly focused on how users engage with offers from outside companies (click-apply events) displayed on each of Credit Sesame's three platforms: mobile web, mobile app, and web. In addition to understanding user behavior, we built an analytics dashboard and API to summarise our results. The resulting webpage is an easy tool for visualizing dynamic analyses of customer usage and behavior in real time.

Keywords

EDA, Credit Sesame, mobile, app, web

Introduction

Credit Sesame is a website that provides users with free estimates of their credit score. It also provides analytics, credit monitoring and alerts, and identity theft protection. The company makes money by suggesting referral products for its users, such as new credit cards. In this paper, we analyze customer engagement and demographic data to build a better understanding of Credit Sesame's user base, and to create the most value possible for the company. We investigated how user engagement differs by platform (web, mobile web, and mobile app), with particular attention to behaviors that earn Credit Sesame revenue. Our primary motivation is to understand how users interact with Credit Sesame platforms to provide suggestions on what future features should be developed, how advertising budgets should be spent, and how to better motivate users to click outside referrals to increase revenue.

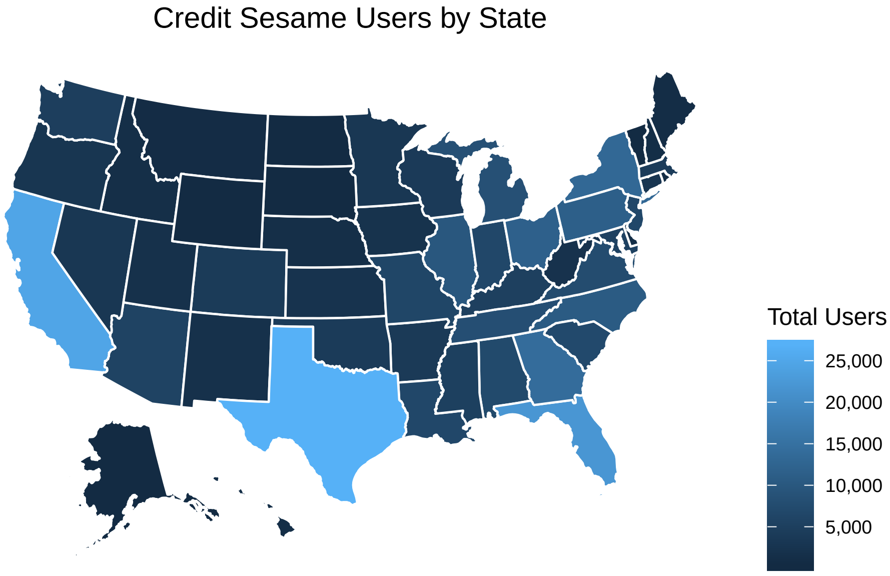
Data set

Our project utilizes three datasets provided by Credit Sesame. The first is **User Profiles**, which includes demographic and credit profile information for 285,491 people who created accounts with Credit Sesame in July 2018 — such as age group, location, credit history, whether a person is a homeowner, etc. We also used the **First Session** dataset, which details people's actions on the Credit Sesame website during their first login. Some key variables in this dataset are platform (mobile app, mobile web or computer), time spent on the website, and users' click-actions — whether they clicked on offers from Credit Sesame, how many pages of the website they visited, etc. Lastly, we used the **30-day User Engagement** dataset, which includes the same information as the **First Session** data, except for the 30-day period after a user registered with Credit Sesame.

Exploratory Data Analysis

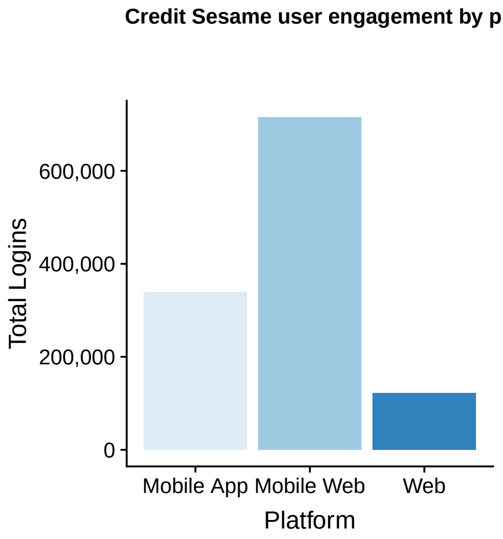
User Demographics

The was understanding who compromises Credit Sesame's userbase. Most users live in Texas, California, Florida, Georgia or New York. This isn't too surprising, as these are also some of the most populous states in America (save for Georgia). We also found that most users have credit scores ranging from 500-600 (Plot 3), and that most users are not homeowners (Table 3).

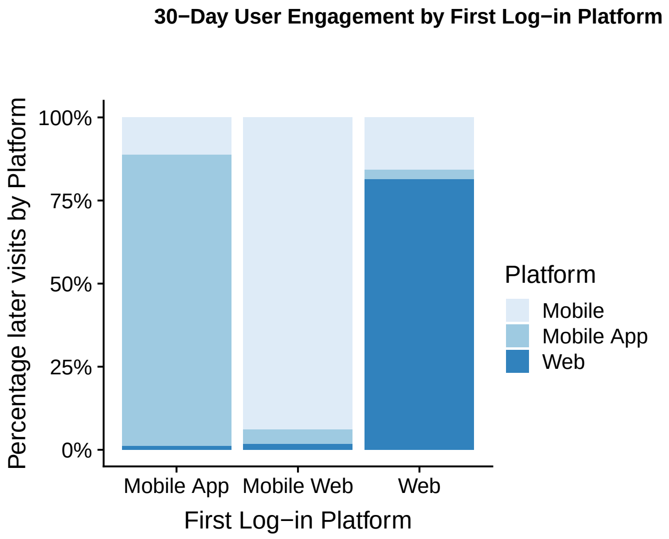


Platform

The first step of our EDA was delving into *how* users access Credit Sesame. We found that a vast majority of users accessed Credit Sesame through mobile web during their first 30-days. Comparateively, very few users logged in through a desktop computer or laptop. We also found that engagement is incredibly consistent when it comes to platform: Whatever platform someone first logged on to Credit Sesame with, they continued to use that platform as their primary means of accessing Credit Sesame services: Among users who first logged on to Credit Sesame using the mobile app, 87.7% of their following vists were made through the mobile app. For mobile web, that figure stands at 93.4%. For web on a desktop computer, 81.5%.



Credit Sesame user engagement by platform



30−Day User Engagement by First Log−in Platform

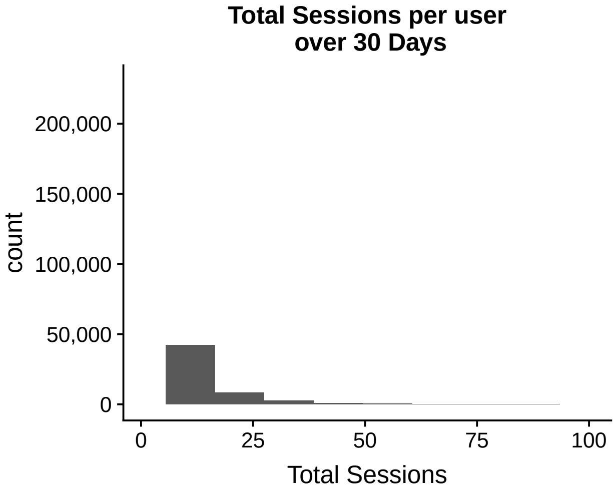
Usage

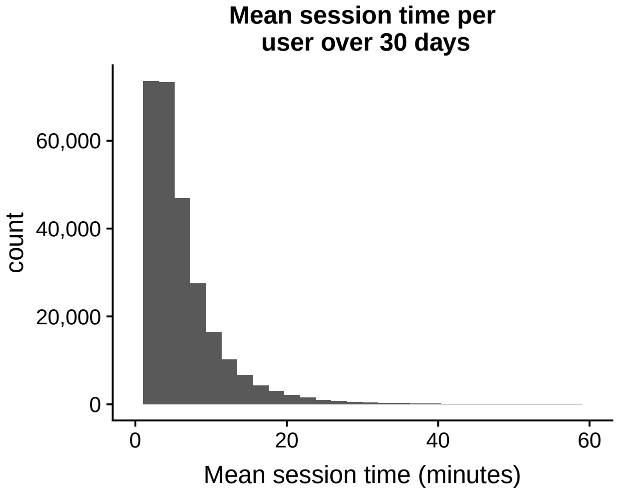
Most Credit Sesame customers are one-time users: Across all platforms, a majority of people accessed Credit Sesame services fewer than 10 times in the 30 days after registering, for less than 20 minutes per visit. When it comes to clicks— a good measure of user engagement— web and mobile app users clicked more per session, on average, than mobile web users. However, those users also had, on average, longer sessions, meaning the click/minute ratio for web and mobile app users is roughly the same as for mobile web users. Furthermore, we found that the distribution of users' age and credit scores did not differ much by platform (Plots 2 and 3, respectively).

Table 1: User engagement by platform

login_platform	mean_time	mean_click	mean_click_apply	clicks_per_minute	click_apply_per_minute
Web	5.843751	7.501176	0.3368056	1.283624	0.0576352
Mobile App	5.213268	7.440772	0.1646282	1.427276	0.0315787
Mobile Web	4.292495	5.834346	0.3473749	1.359197	0.0809261

But our main interest here are "click-apply" events, or when a user clicks on an offer displayed by Credit Sesame. This is how the company makes money, so boosting "click-applies" among users is important. Across platforms, on average, users rarely engaged with offers from Credit Sesame (fewer than 1 click-applies/visit across all platforms). If Credit Sesame is hoping to make more money, they should invest money in determining how to boost user engagement with offers presented on the relevent Credit Sesame platforms.





Methods: Linear Regression

Since Credit Sesame earns revenue from customers engaging with advertisements from outside providers, we decided to design a logistic regression model to predict a whether customer clicks on such offers— or a "click-apply" event. While our model proved unhelpful, it was designed with the goal of better understanding how Credit Sesame could boost user engagement with outside offers, thus earning the company more revenue.

Model Description

The following is our logistic regression model, which predicts the probability of a click-apply event occuring:

$$\text{total click-apply events} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i$$

Where $x_{1i} = \{1 \text{ if user is a homeowner; 0 otherwise}\}$, x_{2i} = mean credit card utilization ratio, x_{3i} = total tradeline accounts opened in last 6 months, x_{4i} = total inquiries in past 6 months, and $\epsilon_i \sim N(0, \sigma^2)$.

We included an indicator for whether a user is a homeowner in our model because it should be a useful predictor of whether a given user would engage with home loans, one of the click-events offered by Credit Sesame. We also included users' credit card utilization ratio because it may be a good predictor of future credit card use. With a similar rationale, we included the number of credit cards opened in the previous 24 months as a predictor of click-apply events because a users' credit card usage is likely tied to their need for loans and other financial services. Lastly, we included account inquiries made by a user in the past six months because that is likely tied to their need for loans and/or other financial services being advertised on Credit Sesame.

Diagnostics and Limitations of the Model

After running diagnostics on our model, we came to the conclusion that it is a poor fit for predicting click-apply events. With just over 60 variables to choose from, it was difficult to decide which to include in our model given the time constraints of the Datathon. Furthermore, reformatting the dataset via creating indicator variables and calculating summary statistics proved time consuming. If we were given more time, our group would have devoted more exploratory data analysis to understanding which variables, empircally, might be better predictors of user engagement with offers from outside companies— as well as relationships between variables, such as collinearity and interactions. We also would have devoted more time to improving our model through diagnostic measures. Ideally, a new model would include exclusively user demographics and data from their first session (i.e. total clicks, session length, platform, types of services accessed) so that Credit Sesame could identify which customers are likely to earn them the most revenue after just one login, and target advertisements accordingly. It would also be useful to bring in external knowledge of housing/auto/insurance policies by state to better contextualize users' need for services based on where they live.

Applications

Click-through-rate and conversion-rate by login platform

In this section, we explore click through rates (CTR), conversion rates (CR) and click-apply rates (CAR) by login platform. **CTR** is the ratio of page clicks to page views. **CR** measures how often users clicked on an item being viewed that took them to an affiliate link, relative to their total clicks. This measure is particularly important as it earns Credit Sesame revenue. Lastly, we looked at **CAR**, which measures how often users clicked on an item being viewed that took them to an affiliate link, relative to their total page views. This is also an important statistic, as increasing CAR will also increase the company's revenue.

Table 2: CTR, CR and CAR by Platform

login_platform	mean_ctr	mean_cr	mean_car
Mobile App	1.81	0.03	0.03
Mobile Web	0.96	0.10	0.09
Web	1.03	0.07	0.06

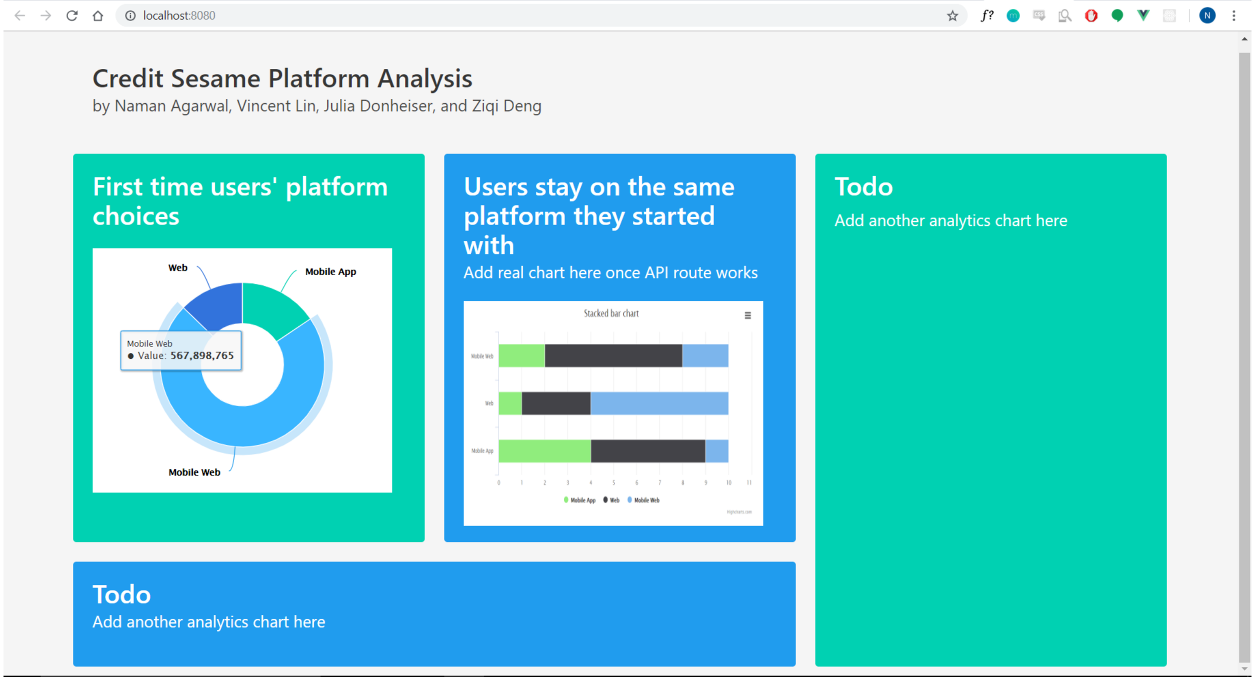
The CR and CAR are low across all platforms. However, mobile web users tend to engage with affiliate links the most. This should be kept in mind, as that also points towards mobile web users as Credit Sesame's largest revenue earners. If Credit Sesame wants to increase revenue, it should do further research on whether the difference in CRs and CARs by platform are tied to differences in interface design (e.g. usability, layout, formatting), as well as increase marketing of their mobile app.

API and Analytics Dashboard

Additionally, we created the infrastructure for an analytics dashboard and API that visualizes live data relating to our project — for example, showing the breakdown of users' platform choice on their first time using the Credit Sesame system. This allows non-technical Credit Sesame employees to get a quick snapshot of user engagement across platforms, which can influence product management, marketing, and more. This API is fully RESTful (or would be if we continued to develop it further), but as a snapshot tool it only needs to include GET requests. The project was built with Vue.js with Highcharts on the frontend with R and Plumber on the backend. This application is modular such that someone at Credit Sesame would be able to replace a few lines (specifcally, hooking the R API up to CS's actual database) and host it internally, and they could use this web interface as a method of keeping tabs on platform use and the other metrics we have discussed in this project. Other key findings from our report could easily be added to the API by taking the R code used to produce our other visualizations, serializing into JSON, and creating an API resource to make it accessible on the frontend.

Discussion

The main focus of this case study was exploratory data analysis. Through exploring the different log-in platforms. We found from the data that many people click apply on the web rather than the mobile app, which has the most frequent users. This is expected becuase we assume that people make more important



Credit Sesame Platform Analysis

by Naman Agarwal, Vincent Lin, Julia Donheiser, and Ziqi Deng

First time users' platform choices

Users stay on the same platform they started with

Add real chart here once API route works

Todo

Add another analytics chart here

Todo

Add another analytics chart here

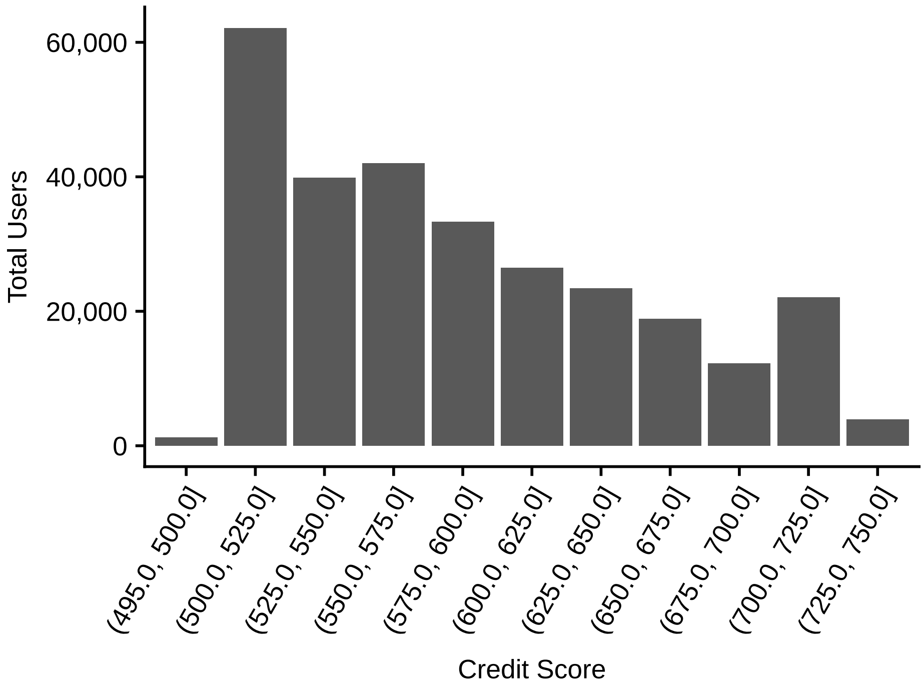
Figure 1: API screenshot

decisions like applying for a car loan or student loan through the web instead of through the mobile app. Our advice for Credit Sesame would be to make it more accessible and efficient to apply for loans or other offers through their mobile app so that they can retain their current registered users and increase the population as well. As decribed above in our model limitations, if we were granted more time to work with this dataset, we would create a model to show any statistical significant categories of clicks or the likelihood of clicking apply given a certain amount of log-in from each platform and current credit score and any kind of debt.

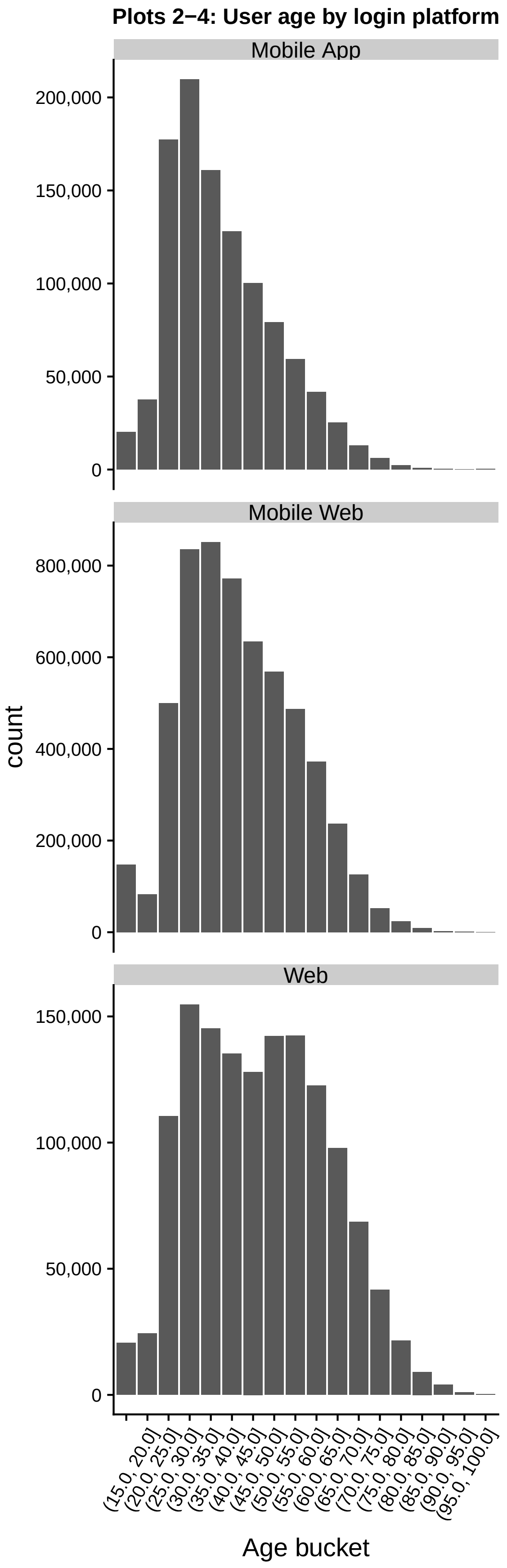
Index

Table 3: Percentage of Credit Sesame users who are homeowners

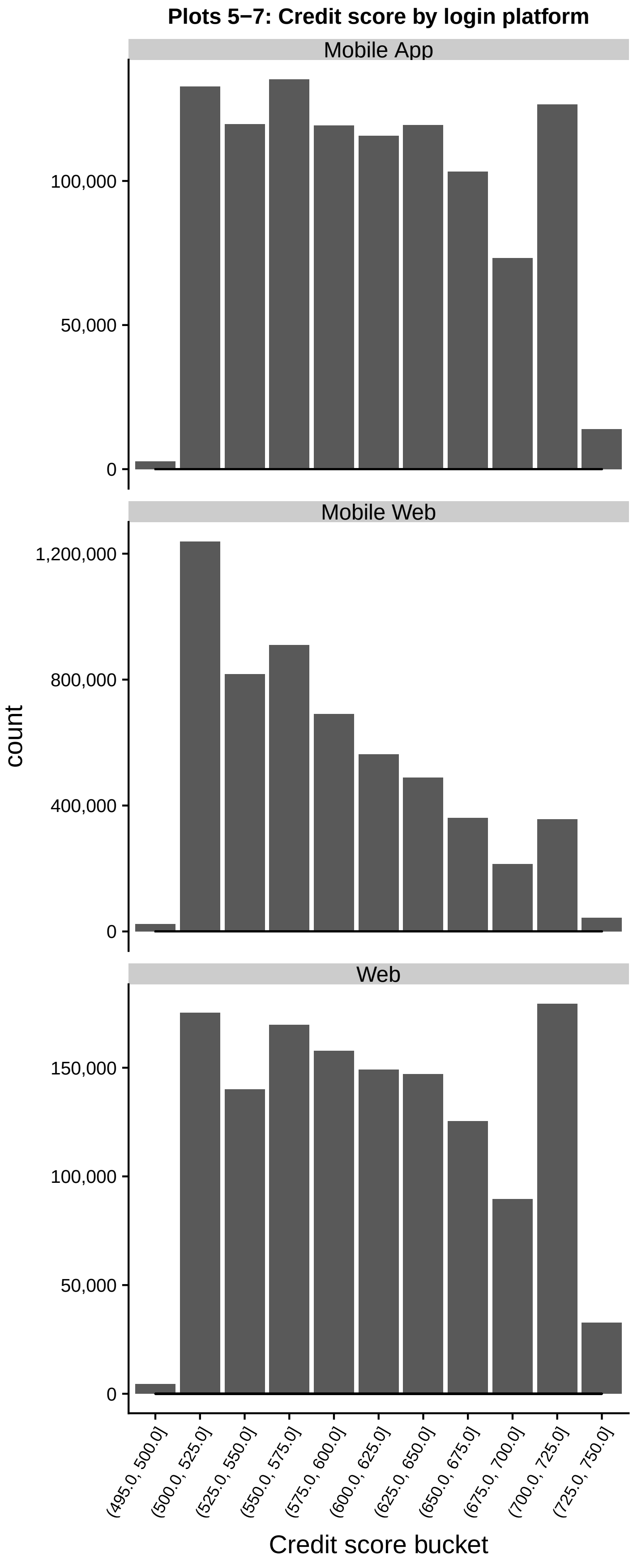
is_homeowner	percentage
FALSE	74.54
TRUE	25.46



Plot 1: Credit scores



Plots 2−4: User age by login platform


Plots 5−7: Credit score by login platform
Mobile App
100,000
50,000
0
Mobile Web
1,200,000
800,000
400,000
0
Web
150,000
100,000
50,000
0
count
(495.0, 500.0]
(500.0, 525.0]
(525.0, 550.0]
(550.0, 575.0]
(575.0, 600.0]
(600.0, 625.0]
(625.0, 650.0]
(650.0, 675.0]
(675.0, 700.0]
(700.0, 725.0]
(725.0, 750.0]
Credit score bucket

Summary of Logistic Regression Model

```
## 
## Call:
## glm(formula = apply_binary ~ is_homeowner + avg_cc_utilization_ratio + 
##     count_tradelines_cc_opened_24_months + count_inquiries_6_months, 
##     data = eng_user_join)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.6034  -0.1379  -0.1211  -0.1010   0.9153  
## 
## Coefficients:
##                                         Estimate Std. Error t value
## (Intercept)                           9.528e-02  8.786e-04 108.449
## is_homeownerTRUE                     -1.013e-02  8.616e-04 -11.761
## avg_cc_utilization_ratio              4.408e-02  1.071e-03  41.174
## count_tradelines_cc_opened_24_months -9.469e-05  2.605e-04  -0.363
## count_inquiries_6_months              5.777e-03  1.428e-04  40.449
##                                      Pr(>|t|)    
## (Intercept)                            <2e-16 ***
## is_homeownerTRUE                       <2e-16 ***
## avg_cc_utilization_ratio               <2e-16 ***
## count_tradelines_cc_opened_24_months    0.716    
## count_inquiries_6_months               <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 0.1103198)
## 
##     Null deviance: 77719  on 700806  degrees of freedom
## Residual deviance: 77312  on 700802  degrees of freedom
##   (476401 observations deleted due to missingness)
## AIC: 443972
## 
## Number of Fisher Scoring iterations: 2
```