Davis and Seifarth Final Report

AUTHOR
Bryson Davis and John Seifarth

PUBLISHED
December 18, 2024

Introduction

A consumer credit score is a numerical representation of an individual's credit worthiness that is calculated through payment history, amount owed, length of credit history, and types of credit accounts. A score typically ranges between 300 to 850 with the higher score correlating to better credit. Credit score is important to lenders of money as it helps them assess risk of extending the monetary value. With a better credit score one is more likely to borrow money at a better interest rate, qualify for better credit card deals, and have better insurance rates[1]. On the other end, poor credit scores result in high insurance premiums, higher utility deposits, and difficulty in renting[2].

The following report focuses on predictors for credit score in an undisclosed multinational European bank servicing France, Spain, and Germany. The data set was collected from Kaggle [3] with a supporting iteration downloaded from Maven [4] Analytics. For the purpose of this report, we will be using the data set from Kaggle. The data includes 10,000 cases and 18 attributes per case. Not all of the attributes will be significant in our report. We will be analyzing customer geography, gender, account balance, number of products, and estimated salary as predictors of credit score to emulate an exploration taken by banks in deciding credit limits and loan decisions, for example. Note: Number of products refers to how many of the bank's services the customer is subscribed to.

Our research questions are as follows:

- Is credit score dependent on the customer's geography?
- Do gender, account balance, number of products, or estimated salary predict a customer's credit score?

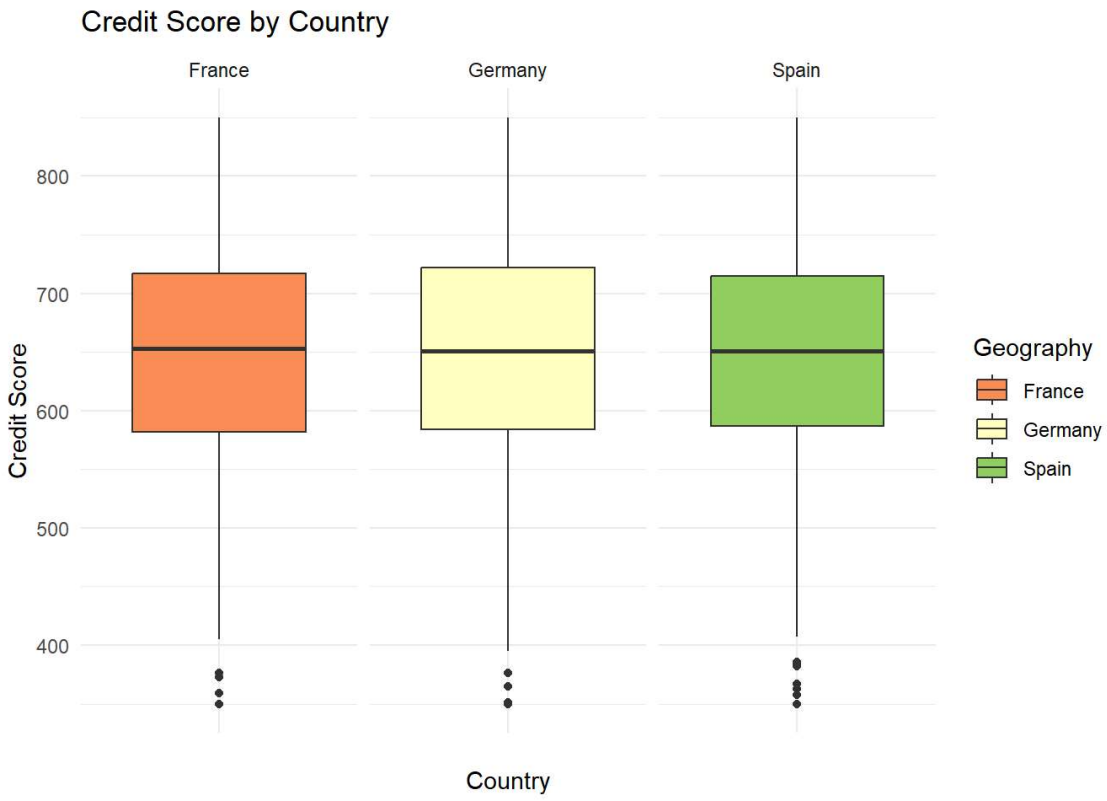
Our data meets the FAIR and CARE principle firstly because it is posted on publicly accessible sites with rich metadata and clearly defined authorization. It is stored in a CSV format and has clearly understandable data descriptions. The data collection method is straightforward and it is licensed for public use. The data does not promote prejudices and creates opportunity for clear analysis of credit data. We assume that the data is sovereign and consent was given to collect it because of its licensing with Maven Analytics. The data does not jeopardize the cases by giving up vulnerable information, and the collection method and publishing was ethical.

Exploratory Data Analysis

For categorical variables, we will be analyzing box plots of credit score faceted by the categorical variable. For continuous variables, we will be analyzing scatter plots to preliminary determine whether we wish to pursue analysis of that variable.

The first area that we want to explore is if there is a correlation between country and credit score. The three countries that this bank is affiliated with are France, Germany, and Spain. Figure 1 shows this box plot coded through R of the credit scores compared to the country. As seen, there is no significant difference between the three countries. They each have a very similar credit score with the same amount of variability.

Figure 1:



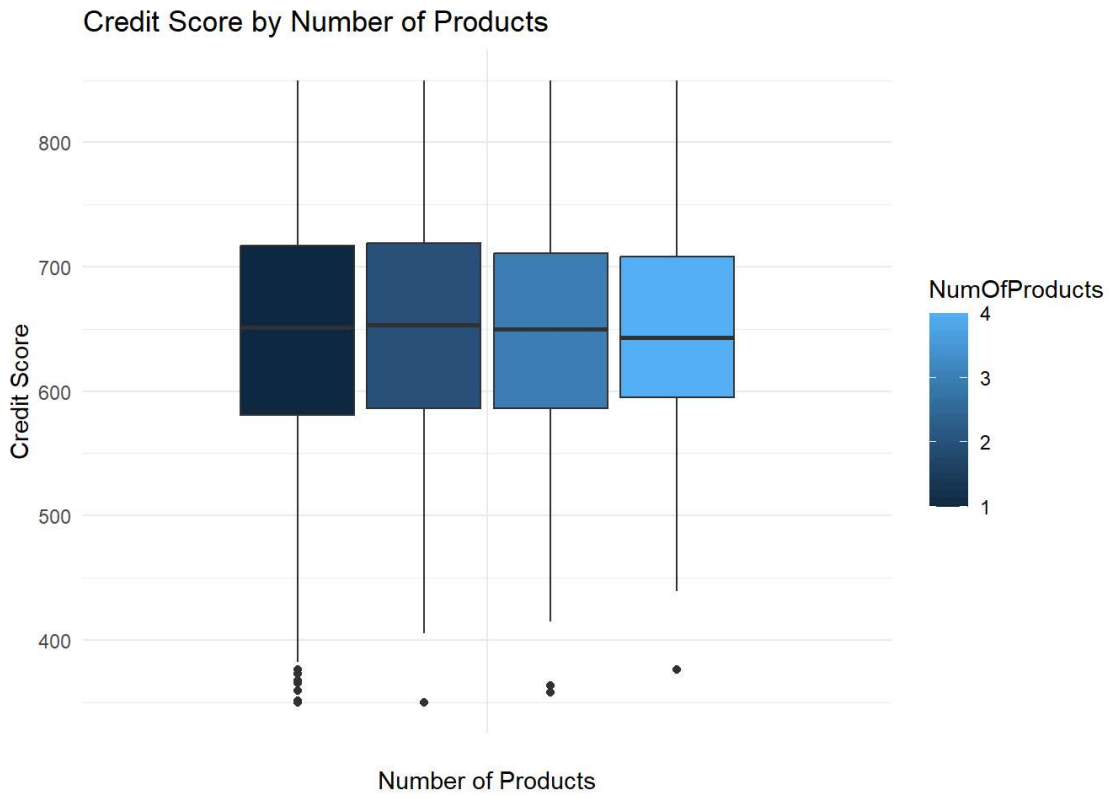
The next data visualization we conducted was also a box plot, but between Male and Female. As we can see in Figure 2, there is no significant difference between the two genders. They each have a very similar credit score with the same amount of variability.

Figure 2:

The final box plot we created, (Figure 3) was by number of products. A number of products is defined as the number of bank products the customer uses regarding their accounts. Unfortunately we are unable to make a conclusion about their credit score

in regards to the number of products as the mean is similar, but we are able to investigate the variability differences. It appears that those with a higher number of products tend to have lower variability compared to lower numbers of products.

Figure 3:



When analyzing the scatter plot of estimated salary and credit score in Figure 4, there seems to be no trends or correlation whatsoever, as reinforced by the trend line.

Figure 4:

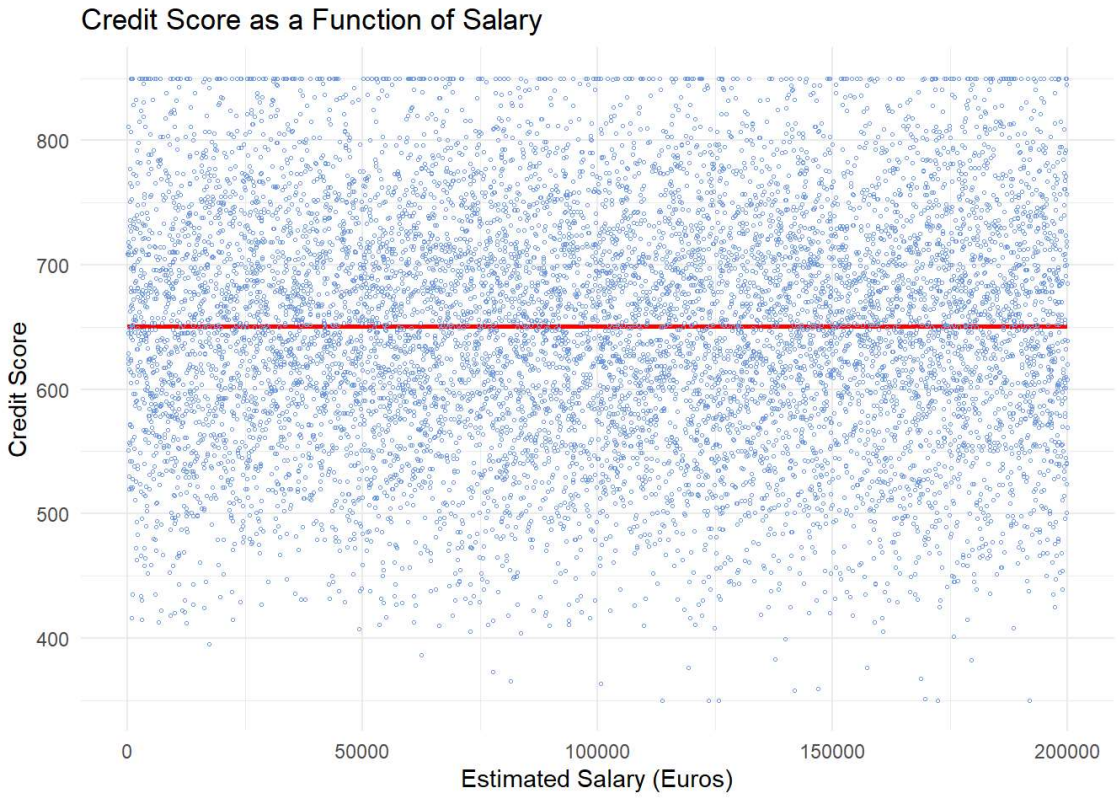
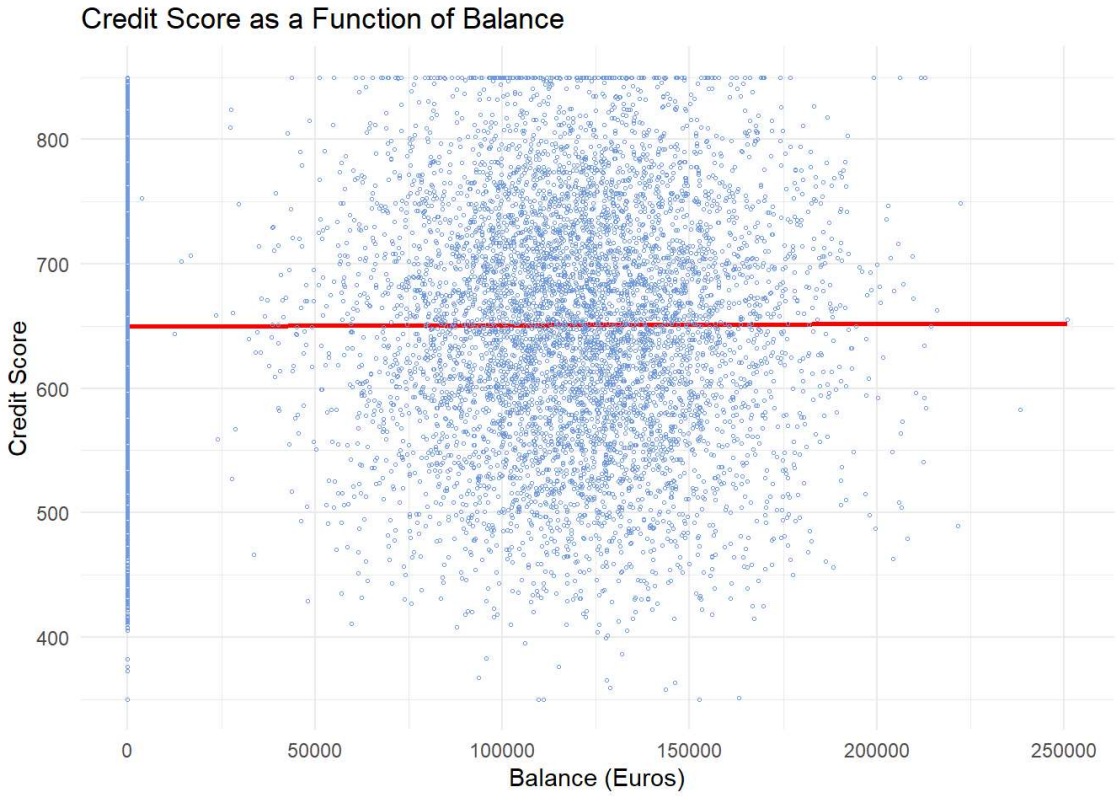


Figure 5 is the same configuration of box plot as Figure 4, except replacing estimated salaray with account balance. We draw the same conclusion as Figure 4.

Figure 5:



After analyzing all five of the plots, we can see that in none of the plots are there apparent differences in credit scores across any partition of the data. The quartiles for credit score shrink around the median as the customer's number of products increases, but other than this, our data leads us to believe that finding a predictor for credit score in this data will not be as straightforward as

evaluating one variable as a predictor of credit score. We will proceed by statistically proving our hypothesis that none (or some for simplicity's sake) attributes do not predict credit score successfully. We have decided not to move forward with analysis of either estimated salary or balance.

Analysis

We will be using ANOVA to analyze whether credit score differs across the categorical variables of geography, gender, and number of products.

Figure 6 (ANOVA of Credit Score by Country):

```
              Df   Sum Sq Mean Sq F value Pr(>F)
Geography      2     7464    3732   0.399  0.671
Residuals   9997 93401796    9343
```

Our p-value of this analysis is 0.671, so we can conclude that the credit score differences between countries are not statistically significant.

Figure 7 (ANOVA of Credit Score by Gender):

```
              Df   Sum Sq Mean Sq F value Pr(>F)
Gender         1      762     762   0.082  0.775
Residuals   9998 93408497    9343
```

This ANOVA also warrants a p-value much higher than acceptable for a significan difference. It is 0.775, leading us to the conclusion that there is not a difference in credit score between gender.

Figure 8 (ANOVA of Credit Score by Number of Products):

```
                Df   Sum Sq Mean Sq F value Pr(>F)
NumOfProducts    1    13990   13990   1.498  0.221
Residuals     9998 93395270    9341
```

While this p-value is lower than both of the others, it is still not enough to prove there is a statistically strong relationship between number of products and credit score. The ANOVA warranted a p-value of 0.221.

Conclusion

In our report, we set out to find predictors of credits score in the customer data provided by a multinational European bank. We decided to analyze estimated salary, account balance, account country, gender, and number of products held by the customer. After our exploration and data analysis, we have concluded that there is not a relationship between any of these attributes and credit score. A multivariate analysis may warrant a combination of multiple attributes to be a predictor, but for the purpose of this project, none of these attributes directly affect credit score. We found this through ANOVA of the categorical attributes, and scatter plots of the continuous attributes. We had a sufficiently large sample size and generally equal variability among groups. While this may hold true for this data set, it is only representative of 10,000 customers of one bank. Elsewhere, or for another bank, one of these may be a sufficient predictor. While no statistical relationships were found, this analysis still contributes to our understanding of credit score by providing the knowledge that the attributes we analyzed are not always reliable predictors of credit score.

Code Appendix

Package Loading and Data Preparation:

```r
library(dplyr)
library(ggplot2)
```

```r
bankData <- read.csv("C:\\Users\\jsei8\\Downloads\\archive\\Customer-Churn-Records.csv")

bankDataClean <- bankData %>%
  select(CustomerId, CreditScore, Geography, Gender, Balance, NumOfProducts, EstimatedSalary)
```

Figure 1:

```r
ggplot(bankDataClean) +
  aes(x = "", y = CreditScore, fill = Geography) +
  geom_boxplot() +
  scale_fill_brewer(palette = "RdYlGn", direction = 1) +
  labs(
    x = "Country",
    y = "Credit Score",
    title = "Credit Score by Country"
  ) +
  theme_minimal() +
  facet_wrap(vars(Geography))
```

Figure 2:

```r
ggplot(bankDataClean) +
  aes(x = "", y = CreditScore, fill = Gender, group = Gender) +
  geom_boxplot() +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Gender",
    y = "Credit Score",
    title = "Credit Score by Gender"
  ) +
  theme_minimal()
```

Figure 3:

```r
ggplot(bankDataClean) +
  aes(
    x = "",
    y = CreditScore,
    fill = NumOfProducts,
    group = NumOfProducts
  ) +
  geom_boxplot() +
  scale_fill_gradient() +
  labs(
    x = "Number of Products",
    y = "Credit Score",
    title = "Credit Score by Number of Products"
  ) +
  theme_minimal()
```

Figure 4:

```r
ggplot(bankDataClean) +
  aes(x = EstimatedSalary, y = CreditScore) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  geom_point(shape = "circle open", colour = "#769DE4", size = 0.5) +
  labs(
    x = "Estimated Salary (Euros)",
```

```
    y = "Credit Score",
    title = "Credit Score as a Function of Salary"
  ) +
  theme_minimal()
```

Figure 5:

```
ggplot(bankDataClean) +
  aes(x = Balance, y = CreditScore) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  geom_point(shape = "circle open", colour = "#769DE4", size = 0.5) +
  labs(
    x = "Balance (Euros)",
    y = "Credit Score",
    title = "Credit Score as a Function of Balance"
  ) +
  theme_minimal()
```

Figure 6:

```
geographyANOVA <- aov(CreditScore ~ Geography, data = bankDataClean)
summary(geographyANOVA)
```

```
             Df   Sum Sq Mean Sq F value Pr(>F)
Geography     2     7464    3732   0.399  0.671
Residuals  9997 93401796    9343
```

Figure 7:

```
genderANOVA <- aov(CreditScore ~ Gender, data = bankDataClean)
summary(genderANOVA)
```

```
             Df   Sum Sq Mean Sq F value Pr(>F)
Gender        1      762     762   0.082  0.775
Residuals  9998 93408497    9343
```

Figure 8:

```
numProductsANOVA <- aov(CreditScore ~ NumOfProducts, data = bankDataClean)
summary(numProductsANOVA)
```

```
               Df   Sum Sq Mean Sq F value Pr(>F)
NumOfProducts   1    13990   13990   1.498  0.221
Residuals    9998 93395270    9341
```

Footnotes

1. 7 benefits of good credit. NerdWallet. (n.d.). https://www.nerdwallet.com/article/finance/benefits-of-good-credit ↩
2. DeMatteo, M. (2024, October 21). These are the biggest disadvantages of having a bad credit score. CNBC. https://www.cnbc.com/select/side-effects-of-bad-credit/#:~:text=For%20a%20$300%2C000%20house%2C%20you,up%20for%20better%20career%20opportunities. ↩
3. Kollipara, R. (2023, April 28). Bank customer churn. Kaggle. https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn ↩
4. Free Data Sets & dataset samples: Maven Analytics. Free Data Sets & Dataset Samples | Maven Analytics. (n.d.). https://mavenanalytics.io/data-playground ↩