Analyzing Factors Affecting Credit Limit

Group 4: William Chettleburgh, Aaron Langtry, Charles Martel, and Jeremy Bouford

Introduction

A credit limit is the maximum amount of debt a person can accumulate on a credit card before their purchases are denied. The credit limit varies from person to person, depending on how confident the company is that they will be paid back. In this project, we explore the factors of a person that influence their credit limit. Specifically, we use the Credit dataset from "An Introduction to Statistical Learning" (available through the ISLR package in R). This dataset contains simulated data of 400 credit card users which reflect real world trends. The dataset contains a mixture of categorical and numerical variables that reflect demographic and financial information. For categorical variables, the data includes gender (either male or female), student status (student or non-student), marital status (either married or unmarried), and ethnicity (either Caucasian, African American, or Asian). For numerical variables, there is income, credit rating, average credit card balance, number of credit cards owned, age of the person, number of years of education, and (of course) credit limit.

Using this data, we specifically aim to answer the questions:

- What financial and demographic factors contribute to having a higher or lower credit limit?
- Do gender or ethnicity impact whether a person has a higher or lower credit limit?
- What are the most relevant and statistically significant relationships for predicting and describing credit limit?

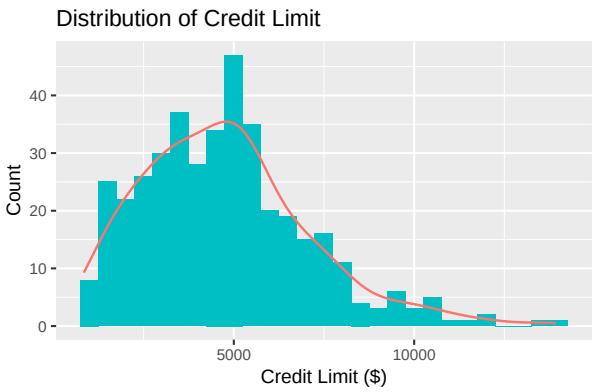
Through various forms of inference and regression, we expect to find that some of these variables are significant in predicting credit limit while some are not. It is expected that credit rating will be a very strong predictor of credit limit (the higher the credit rating, the higher the credit limit). Credit balance is also expected to have a positive relationship with credit limit, since the greater the usage of a credit card, the greater the amount of evidence that a person can meet their payments. Income is expected to have a strong positive relationship, since a person with more income is more able to make their payments (while those with lower incomes may struggle to make payments on time).

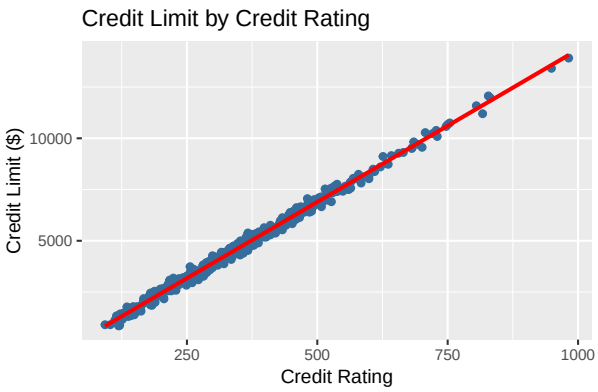
Age is expected to have a weak positive correlation at first (since younger people may be less reliable from the perspective of a credit card company), but past a certain point it is expected for age to not cause any additional change. Along this same line of reasoning, students are expected to have lower credit limits on average than non-students. As the number of years of education increases, it is expected that the credit limit will slightly increase.

We do not expect any differences in credit limit based on gender, marital status, or ethnicity, since if these trends existed, this would be a form of discrimination. The number of credit cards is expected to have a weak negative correlation, since having many credit cards suggests behaviors such as risky spending.

Results

Single-Predictor Analysis

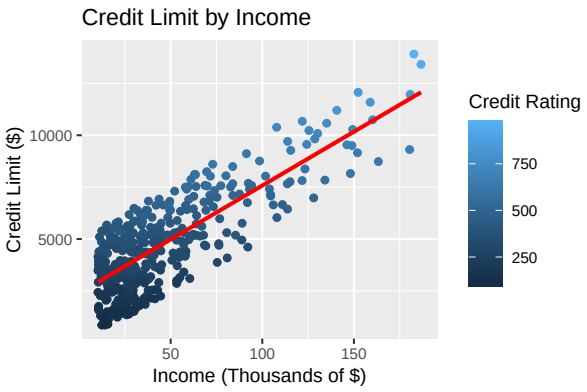


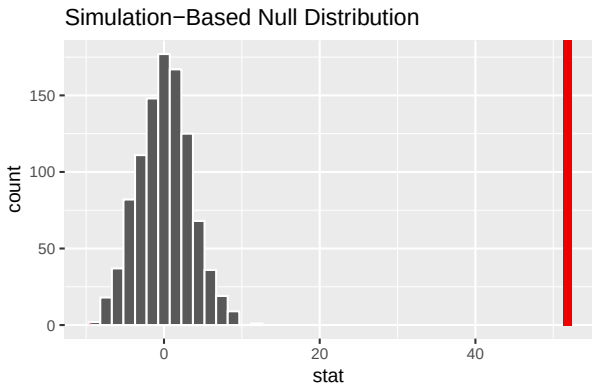


To begin, we plot the distribution of credit limits, and see that it is strongly skewed to the right. Due to this, many hypothesis tests using the `infer` package for comparing two samples failed to yield a normal null distribution, so for inference on categorical variables, we use the Mann-Whitney U Test. This is a non-parametric test with the hypotheses:

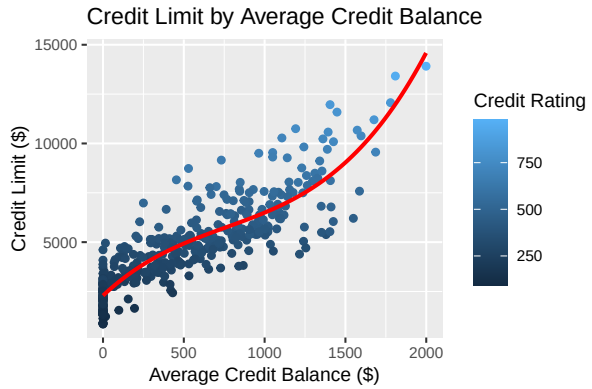
- H_0: The distributions of credit limit from which two samples are drawn are the same (they don't differ by a shift)
- H_A: The distributions of credit limit from which two samples are drawn are not the same (they do differ by a shift)

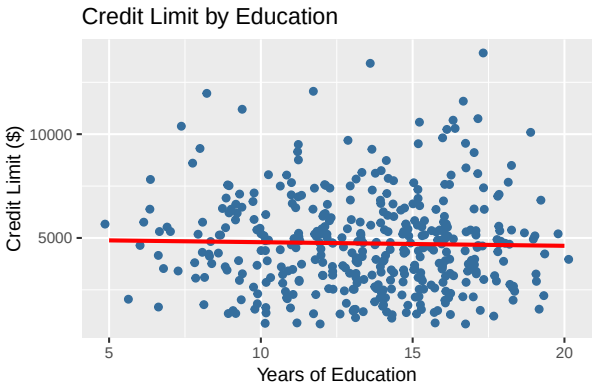
Before this, however, we first analyze some strong numerical predictors of credit limit. The strongest predictor found was credit rating. The linear model has a positive slope, an R^2 of 0.994, and a p-value less than 2.2e-16. This indicates that not only is credit rating a significant predictor, but it explains 99.4% of the variance in credit limit.





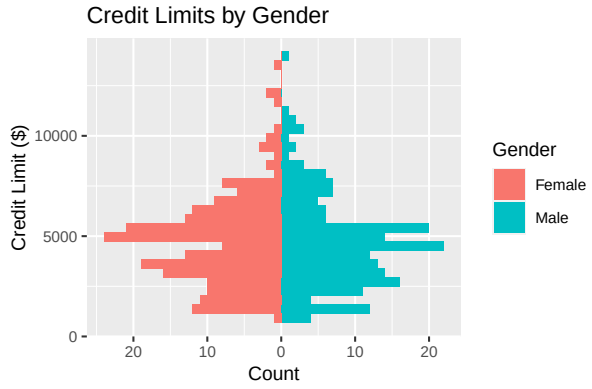
We next look at income. Again, a positive correlation is seen, this time with an R^2 of 0.627. However, the residuals are not normally distributed; rather, they appear to be uniformly distributed. Thus, simulation-based inference is needed to obtain an accurate p-value. Using the `infer` package, we obtain a p-value of 0, which is certainly significant. This is also seen in the null distribution, where the red line marks the observed slope.

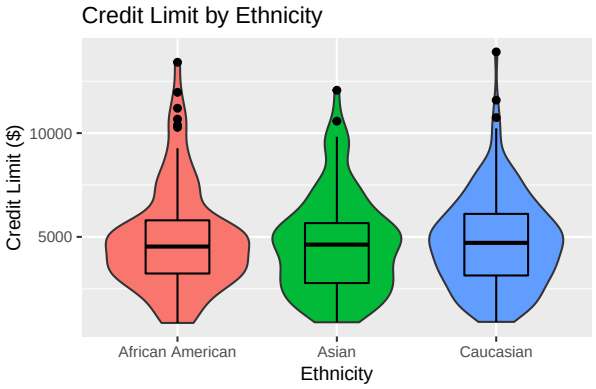




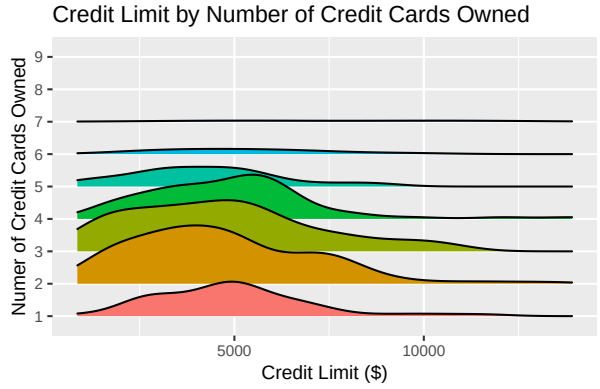
Average credit balance has a positive relationship, but it does not appear to be linear. The relationship appears concave up for high credit limits, and there are many individuals with zero average credit balance which fall lower than expected. To fit both of these features, a cubic model is chosen. This model has an adjusted R^2 of 0.76, and all of the p-values are below 1.13×10^{-7}.

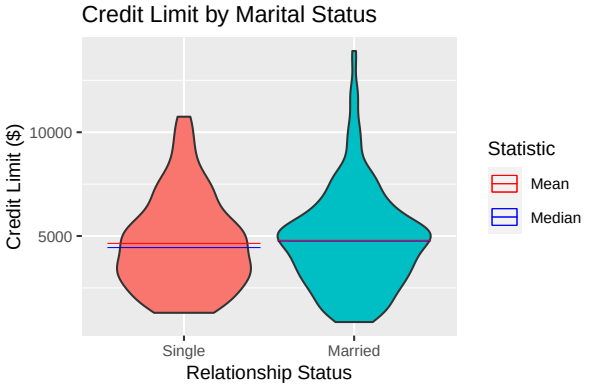
The number of years of education does not appear to have a significant relationship. Linear regression is performed, and the resulting p-value is only 0.639.





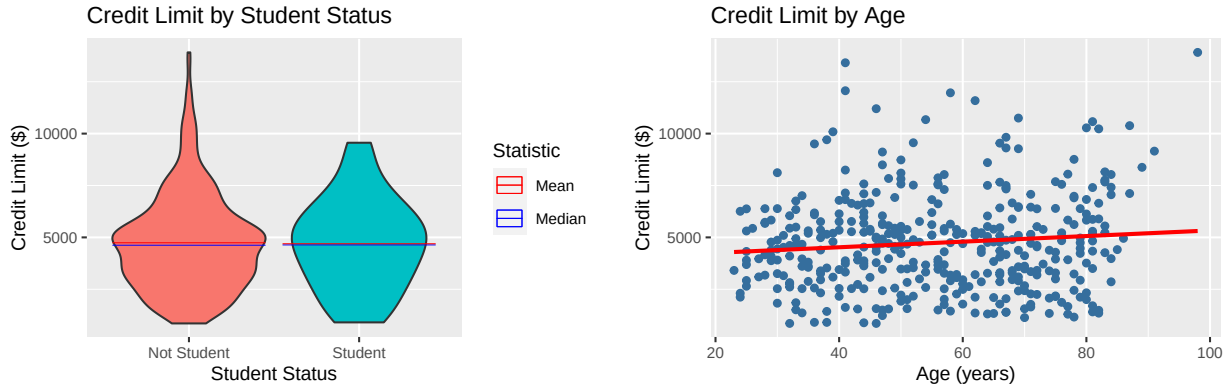
From the double histogram, no significant impact of gender on credit limit is seen. This is confirmed by a Mann-Whitney U Test, which yields a p-value of 0.63. A similar result is seen for ethnicity, where the p-values are 0.47 for comparing Caucasians and Asians, 0.995 for comparing Caucasians and African Americans, and 0.527 for comparing Asians and African Americans. Note that the first quartile for Asians are slightly lower than the others, which is contributing to lower p-values for these comparisons.





For number of cards, a discrete variable, we decided to use a ridge plot. We see that there are no major trends, since the peaks of the ridges do not seem to follow a pattern. This is confirmed by a p-value of 0.838 when conducting regression.

To analyze the effect of marital status on credit limit, two side-by-side violin plots were made. The distributions appear to have different shapes; the peak of the married plot is higher than that of the unmarried. However, due to differences in the tails of the distributions, the mean and median are quite similar for the two (with married only having slightly higher values for these quantities). A Mann-Whitney U Test fails to find any significant difference, yielding a p-value of 0.506.



A violin plot is also made to compare the credit limits for students and non-students. Since there are far fewer students in the dataset than non-students, the distribution for students appears slightly smoother. Besides this, though, no significant difference is seen; the p-value from the Mann-Whitney U Test is 0.849.

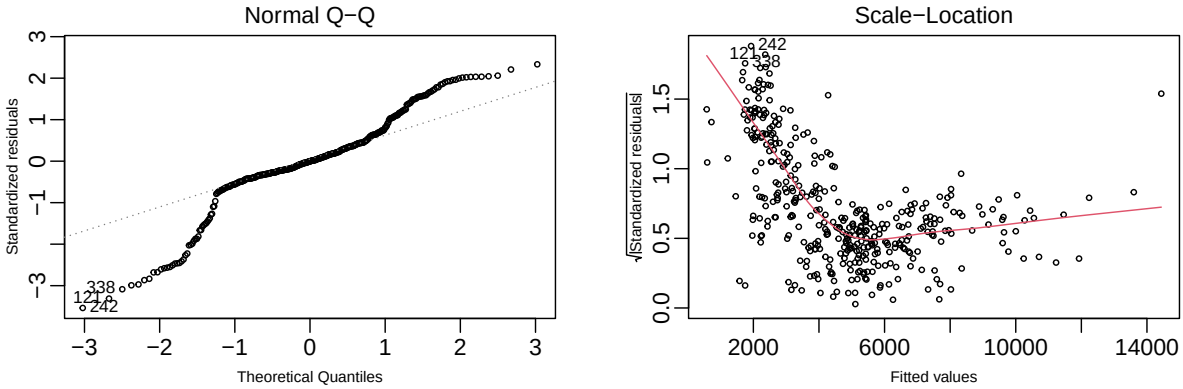
For age, a scatterplot was made and fitted with a linear model. The slope is quite small, and yet it proves to be a significant predictor (the p-value is 0.044). While it may be significant, the relationship is not strong, since the R^2 is only 0.01.

Multiple Regression

To draw our final conclusions, we perform multivariable regression, so that the effects of each predictor individually can be isolated. As a note, since credit rating is a near perfect predictor of credit limit on its own, further analysis including this predictor is unlikely to be fruitful. Rather, our multivariable analysis will attempt to construct a model that predicts credit limit from all of the variables except credit rating. This is a justified approach, since it can be thought of that credit limit and credit rating essentially are two measures of the same quantity: the reliability of the cardholder.

We begin by fitting a full model, including all of the possible predictors (we include average credit balance as a cubic, similar to above). Then, using the step function (offered by the stats library in R), both forward and backward selection are performed, and the best model is chosen based on the AIC (a measure that compares the quality of different models). After performing this, any variables with insignificant p-values are dropped, and the model is refit.

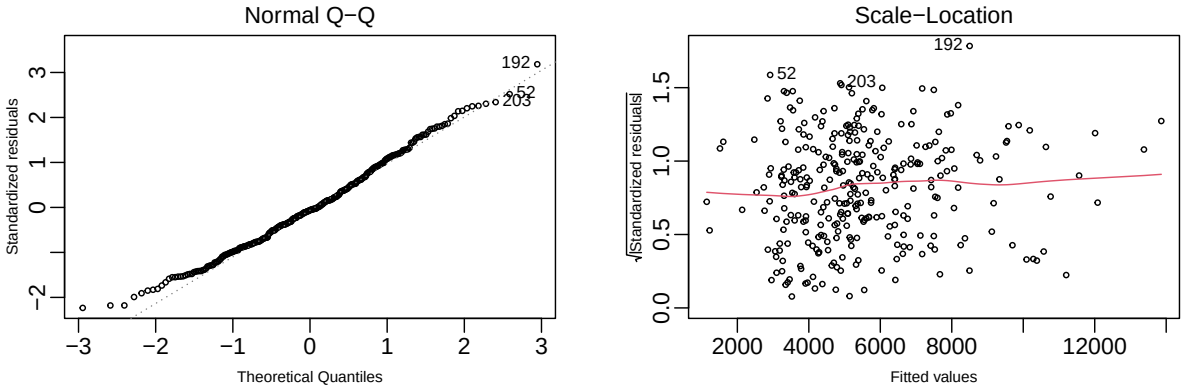
We ultimately find that gender, ethnicity, education, and marital status are not significant in predicting credit limit (while the other predictors are). However, by displaying the diagnostic plots for the model, several issues become apparent, drawing into question any conclusions made from this model.



The Normal Q-Q plot does not fall along a straight line. Since the tails twist counterclockwise, this suggests that the distribution has heavy tails (also called leptokurtosis). Also, the scale-location plot has significantly higher standardized residuals at lower fitted values of credit limit than at higher fitter values, indicating that homoscedasticity is not satisfied.

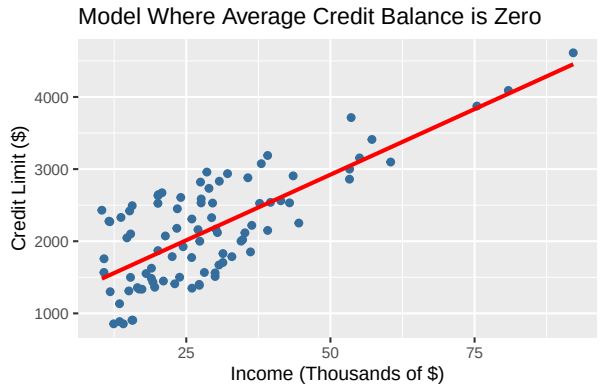
Upon trying several fixes, it was found that the model could be drastically improved by splitting it into two cases: one where the average credit balance is nonzero and one where it is zero. Recall from earlier that there is a significant proportion of individuals with zero average credit balance, who tended to not fall on a linear or quadratic regression (hence the use of a cubic). By excluding these cases from our model, the diagnostic plots drastically improve, indicating that the large residuals were in fact originating from this group of people. Below are the results and plots for the model where all individuals have a positive average credit balance.

```
##               Estimate Pr(>|t|)
## (Intercept)  2150.921   <2e-16
## Income         30.626   <2e-16
## Cards         -75.926   <2e-16
## Age             3.067   <2e-16
## StudentYes  -1532.745   <2e-16
## Balance         3.061   <2e-16
## ---
## Multiple R-squared:  0.9998
## Adjusted R-squared:  0.9998
```



For a positive average credit balance, we find that the number of credit cards, income, average credit balance, age, and student status are all statistically significant predictors of credit limit. All the numerical variables have positive correlations, except for the number of credit cards which has a negative correlation. Students on average have a lower credit limit than non-students. From the adjusted R^2, we see that the model explains over 99% of the variance in credit limit.

When the average credit balance is zero, following the above procedure and successively dropping predictors with insignificant p-values reveals that only income is significant in predicting credit limit (with a coefficient of 36.4 and a p-value of less than 2.2e-16). The model has a significantly lower adjusted R^2 of 0.562. Furthermore, this model has some noticeable limitations. For example, it appears to fail the homoscedasticity condition, and thus may be less reliable.



Answering Research Questions

We first find that credit rating is an extremely significant and strong predictor of credit limit. When attempting to predict credit limit from the other predictors, the multivariable model can be used to isolate the effects of each factor on credit limit. However, the model actually differs for two cases, depending on whether average credit balance is positive or zero. In the case that it is positive, we find the following significant predictors:

- Income: For every $1000 increase in income, there is an expected increase in credit limit by $31 on average, all else constant
- Number of Credit Cards: For each additional credit card, there is an expected decrease in credit limit by $76 on average, all else constant
- Age: For each additional year of age, there is an expected increase in credit limit by $3 on average, all else constant
- Student Status: On average, students have a credit limit of $1533 as compared to non-students, all else constant
- Average Credit Balance: For each additional dollar of average credit balance, there is an expected increase in credit limit by $3 on average, all else constant

When average credit balance is zero, only income is significant, and we find that there is an expected increase in credit limit by $36 on average.

Notably, we do not find either gender of ethnicity significant predictors in credit limit. We thus conclude that there is no significant difference in credit limit based on gender or ethnicity from this dataset.

Conclusion

From our analysis we learned that nearly all of the categorical variables do not have a significant effect on predicting credit limit. One of the only categorical variables that had noteworthy influence on credit rating was student status, where being a student would negatively affect the person's credit rating. The most influential variables were the financial factors, such as income, credit rating, and average credit balance. With credit rating being a near perfect predictor, this factor alone is the most useful information for determining a person's credit limit.

Our methodology for analysis was to first examine the results of each predictor individually, and then examine them in conjunction with each other. While the single-variable analysis occasionally provided important

information (such as the near-perfect relationship with credit rating or the large number of people with zero average credit balance), many of the conclusions drawn during this section were later superseded by the results from the multivariable regression. This suggests that our methodology was not necessarily the most efficient. This is especially true since we tended to make several different types of graphs for each predictor, only one of which was kept. This process could have been made more efficient by better planning and organization.

We have two specific ideas/suggestions for how to improve our analysis. First, when finding the relationship between credit balance and credit limit, we developed a cubic model. While this model met the criteria for regression, we later learned that the people with zero average credit balance have fundamentally different relationships. Therefore, this model may not be accurate for low (but nonzero) credit balances. Other models may do a better job, such as an exponential model fit only to those with a positive average credit balance.

Second, it would have been helpful to make a correlation matrix near the beginning of the project. This would have revealed several trends near the beginning of the project, such as the strong correlation between credit rating and credit limit.

The conclusions drawn from this project are reliable. This is because the factors in the dataset are well-defined (they are not subjective) and there are many observations. If similar datasets were obtained and the same analyses were conducted, it would thus be expected for the same (or similar) results to be found. However, the conclusions from this project are not necessarily valid. Since the dataset is simulated, there is no guarantee that it will reflect real-world trends. Furthermore, it may contain bias based on the opinions and perceptions of the person who created it. Therefore, while the results we found make sense, we must be cautious in generalizing them.

The use of the Mann-Whitney U Test was appropriate, since the null distributions were not normally distributed (and thus a non-parametric test was required). For the multivariable regression where average credit balance is positive, this method is also justified, as evidenced by the diagnostic plots. However, for the regression where average credit balance is zero, homoscedasticity is not satisfied, so it may not be the most appropriate. Further research could attempt to identify other factors not present in this dataset that (when included in the model) may fix this issue.

Aside from the limitations discussed above (such as the difficulty in generalizing our results), one of the main limitations was the unpredictable trends associated with people with zero average credit balance. This resulted in two models needing to be made, one of which is questionable in whether it meets the assumptions and has a low R^2.

After all of our analysis was done, there are still further questions that can be explored about the dataset 'Credit' or similar datasets. Some of these questions include:

- Can we determine if someone is reliable to pay off their credit card balance without knowing their credit rating?
- What relationships exist between the financial/demographic factors and the utilization ratio (the ratio of credit balance to credit limit)?
- Can we predict a person's monthly or yearly spending based on the utilization ratio?
- What other predictors could be found/included to better predict the credit limit of individuals with zero average credit balance?

References

Emulate ggplot2 default color palette. (2016). In *Stack Overflow*. Stack Exchange Inc. https://stackoverflow.com/a/40181166

Gupta, S. (2016). R-source/lm.r. In *GitHub*. GitHub Inc. https://github.com/SurajGupta/r-source/blob/master/src/library/stats/R/lm.R

Gupta, S. (2015). R-source/anova.r. In *GitHub*. GitHub Inc. https://github.com/SurajGupta/r-source/blob/a28e609e72ed7c47f6ddfbb86c85279a0750f0b7/src/library/stats/R/anova.R

Henrik. (2017). Center plot title in ggplot2. In *Stack Overflow*. Stack Exchange Inc. https://stackoverflow.com/a/41487043

Kabacoff, R. I. (2017). Graphical parameters. In *Quick-R*. https://www.statmethods.net/advgraphs/parameters.html

Manian, S. S. (2016). How to reduce space between axis ticks and axis labels in r. In *Stack Overflow*. Stack Exchange Inc. https://stackoverflow.com/a/38935883

Narasimhan, R. (2014). Number format in ggplot: No sign on y-axis-labels. In *Stack Overflow*. Stack Exchange Inc. https://stackoverflow.com/a/21212990

strictlystat. (2014). Making back-to-back histograms. In *R-bloggers*. https://www.r-bloggers.com/2014/06/making-back-to-back-histograms/

user12117520. (2019). How to change font size for all text in a ggplot object relative to current value? In *Stack Overflow*. Stack Exchange Inc. https://stackoverflow.com/a/58093667

Yu, G. (2021). Convert plot to grob and ggplot object. In *The Comprehensive R Archive Network*. https://cran.r-project.org/web/packages/ggplotify/vignettes/ggplotify.html