UDACITY

PROJECT

Explore and Summarize Data

A part of the Data Analyst Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

8 SPECIFICATIONS REQUIRE CHANGES

There are quite a few points to improve to pass this project, but please do not be discouraged; we are here to assist you whenever you need any help. If you find it hard to complete this project, consider requesting an assistance from your assigned mentor.

Good luck with your next submission!

Code Functionality

✓

All code is functional (e.g. No Error is produced and RMD document is not prevented from being knit.)

✓

The project almost never uses repetitive code where a function would be more appropriate. The code references variables by name instead of using constants or column numbers.

Project Readability

✓

All complex code is adequately explained with comments. It is always clear what the code is doing and how and why any unusual coding decisions were made.

The code uses formatting techniques in a consistent and effective manner to improve code readability. All lines are shorter than 80 characters.

There are only a handful of code lines that are longer than 80 characters, so we may mark this as passing the specification. We suggest, however, to consistently ensure all code lines shorter than that characters length limit.

As a reference, here is a style guide to R code to follow: Hadley Wickham's R style guide.

Markdown syntax is used in the RMD file to improve readability of the knitted file.

Code blocks and irrelevant R code outputs need to be hidden

Although it is usually alright, in this case, there are too many R code outputs that are irrelevant to the analysis, like plotting code, warnings, and module notifications which made the final report less readable e.g.

```
ggplot(aes(x = Term), data = ld) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

To hide them, use knitr options `echo=FALSE`, `warning=FALSE` and `message=FALSE`, or you may add the following global options to the top of the Rmd file:

```
```{r global_options, include=FALSE}
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE)
```
```

Quality of Analysis

The project appropriately uses univariate, bivariate, and multivariate plots to explore most of the expected relationships in the data set.

Incorrect plot placements + add more multivariate plots

A couple of plots are placed in the incorrect sections. The last plot in the Univariate Plots section is supposed to a bivariate plot, and so do both of the plots in the Multivariate Plots section. Please include proper multivariate plots in your report. Remember that multivariate plots need to contain at least three variables. You may encode the third variable with colors.

Questions and findings are placed between blocks of R code regularly so it is clear what the student was thinking throughout the analysis.

Please make sure that there are sufficient comments explaining the plots and statistical calculations in the report.

Reasoning is provided for the plots made throughout the analysis. Plots made follow a logical flow. Comments following plots accurately reflect the plots' contents.

As mentioned in the specification above, most of the plots or set of plots, as well as statistics, are not accompanied by discussions. I see that Analysis part of each section does explain some plots but these sections are not the right place to describe individual plots (for readability reasons i.e. avoid readers having to scroll up to the plots and back to the summary section). Instead, please add comments right after and/or before the (important) plots as well and use Analysis sections to conclude your analysis.

There are two kinds of discussions that may be presented for each plot:

1. The reasoning behind the creation of that plot.
2. Comments following that plot reflecting its contents i.e. how did the knowledge from reading that plot change the course of your analysis?

If there is nothing to be gained from reading a plot, you may explain to readers that you did not find any significant result, the point here is to communicate every step we take in our analysis to readers.

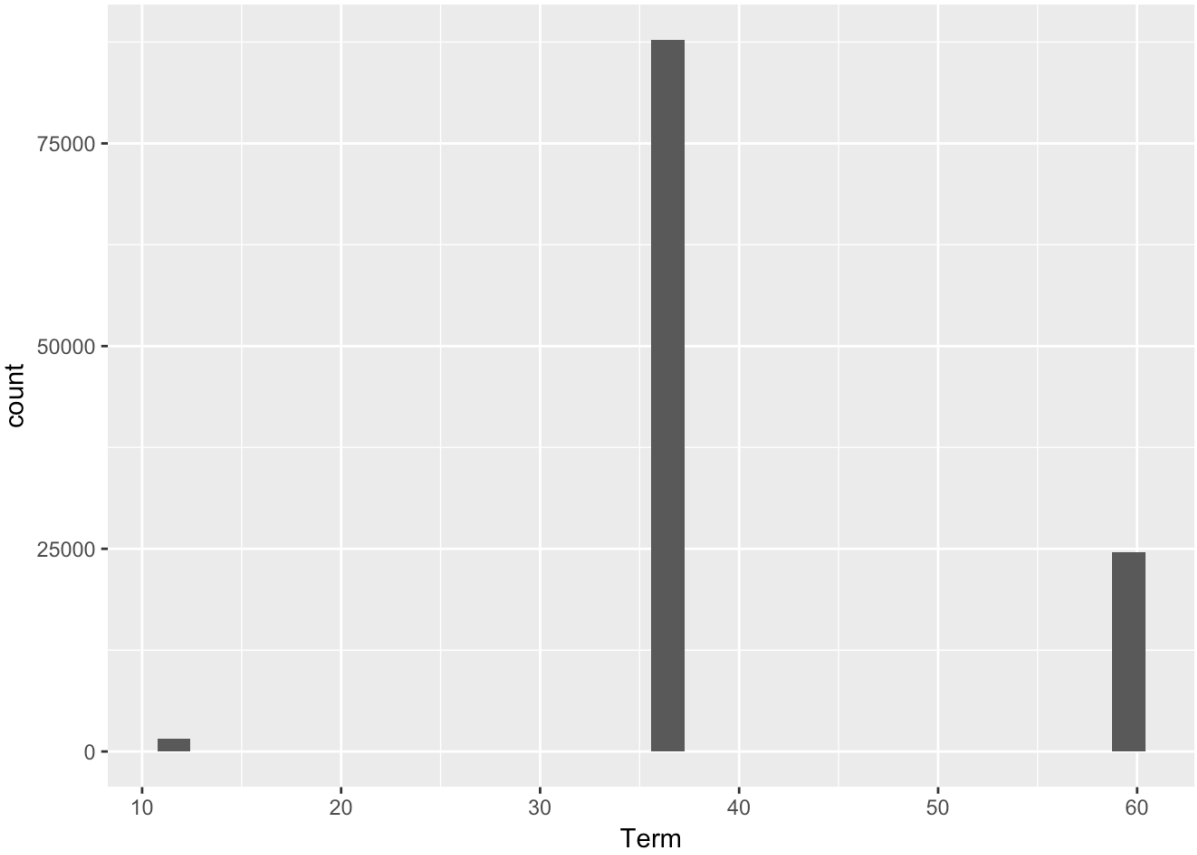
The project contains at least 20 visualizations. The visualizations are varied and show multiple comparisons and trends. Relevant statistics (e.g. mean, median, confidence intervals, correlations) are computed throughout the analysis when an inference is made about the data.

More than 20 visualizations have been provided, good work. To pass this specification, please include a summary information for each variable. This is important so readers may read a quantifiable information regarding the distribution of each variable. They can be added by using `summary()` function. Our recommendation is to include a summary statistic underneath each univariate plot for better readability. Please also include some comments explaining the finding from reading these statistics.

Visualizations made in the project depict the data in an appropriate manner that allows plots to be readily interpreted. Choice of plot type, variables, and aesthetic parameters (e.g. bin width, color, axis breaks) is appropriate.

Histogram bars for ordinal data need to be plotted as such

In this plot:



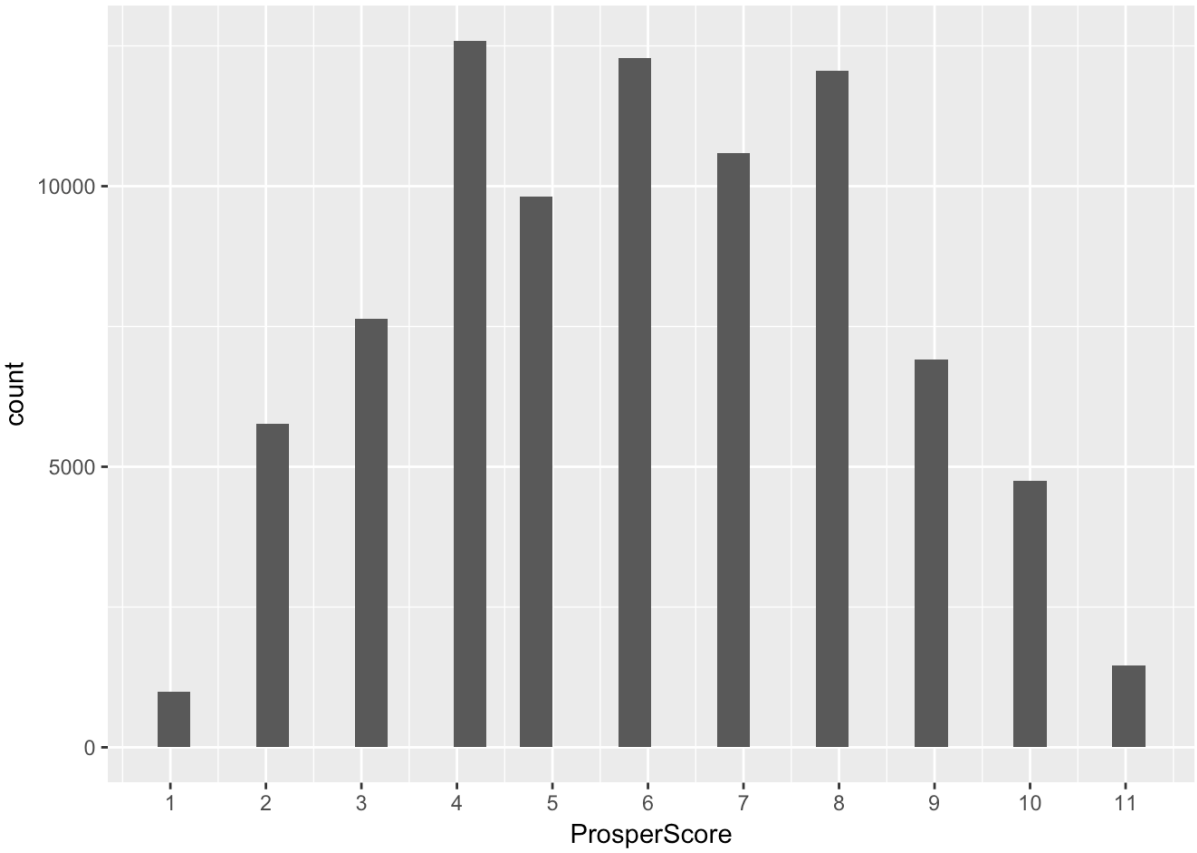
There is no other data outside 12, 36 and 60 terms. Hence, it is better to factorize the terms, so they appear categorical.

One way factorization can be done is by applying this code block before using the `Term` variable:

```
ld$Term <- factor(ld$Term, levels=c(12,36,60), ordered=TRUE)
```

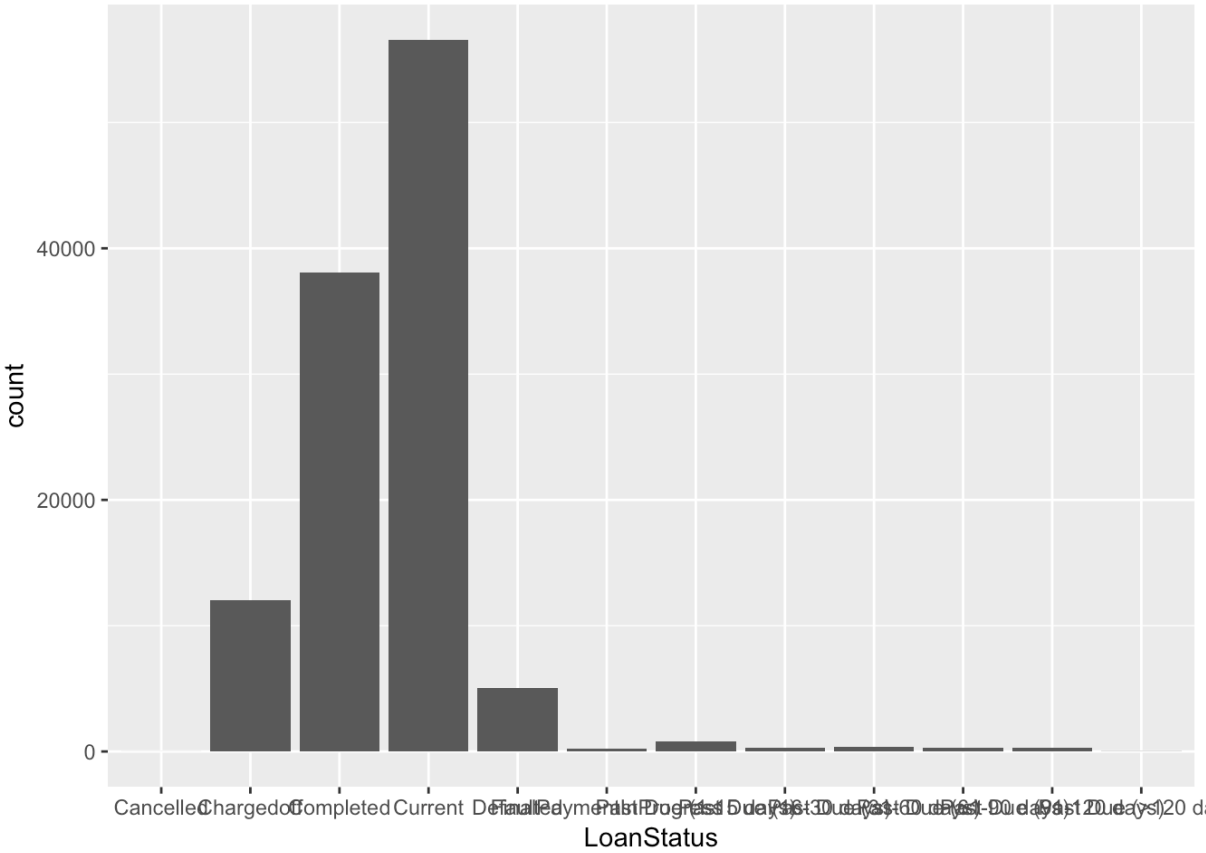
And use `geom_bar` instead of `geom_histogram`.

There is another variation of this issue in this ProsperScore histogram:



Please convert `ProsperScore` into factors as well.

Overlapping labels - general.

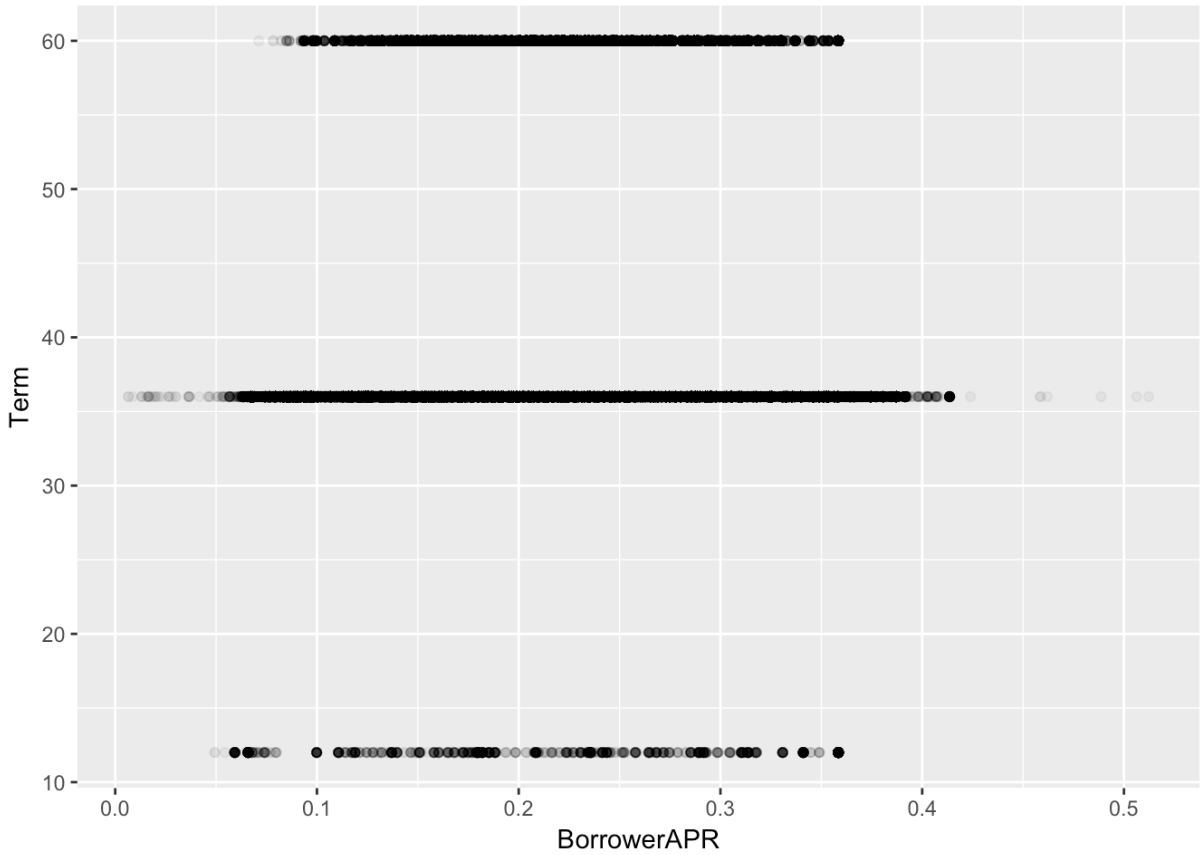


There are several plots which labels were placed too close together that they are hardly readable. They can be corrected either by adjusting the orientation of these labels with `theme` layer or resizing the plots with

`fig.width` and `fig.height` knitr options. Another option is to use a different interval of their tick-marks e.g. by adding `break` parameter (in `scale_x_continuous(break=c(...))` for example)

Discrete vs continuous variable - Needs boxplot or jitter

Several visualizations such as this plot are overplotted:



Currently, it is not possible to tell how dense each grouped data points since they are overlapping on top of each other.

To improve this plot, you may either jitter the data points or use boxplots or violin plots. Either of them will allow readers to find how dense each section of values is.

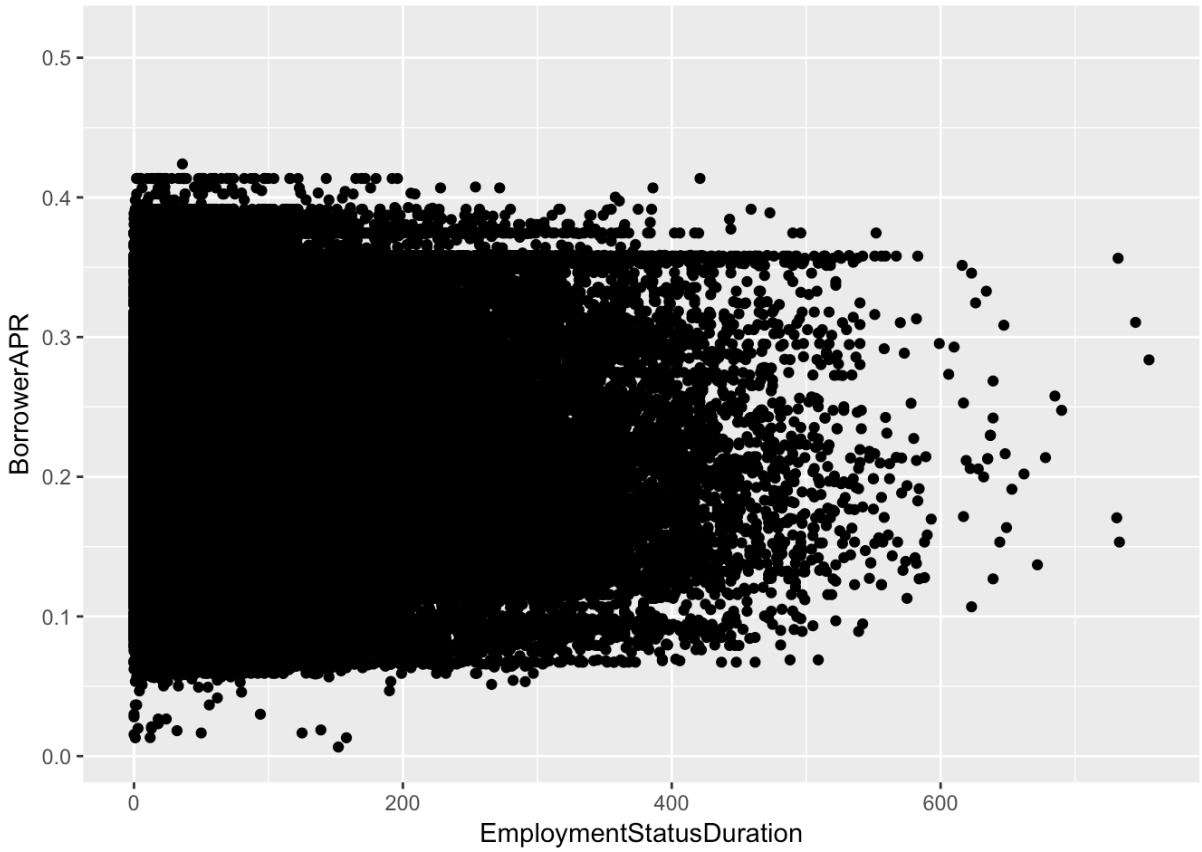
I would usually consider using a boxplot first. The y-axis of the plot shown above uses discrete values / categorical data; Scatterplots are perfect for a pair of continuous variables as the data points could be anywhere, but in this case, boxplot is a better visualization since more detailed statistics can be presented, like medians, quartiles, and outliers.

But sometimes, when there are too many categories, jittering (added with alpha transparency layer when needed) is a much better option. Note that sometimes you can also combine both jittering and boxplots to get the best of both features.

Violin plot is a better option to go with when there are multiple peaks in the distribution of each category e.g. bimodal or multimodal.

Make data points transparent

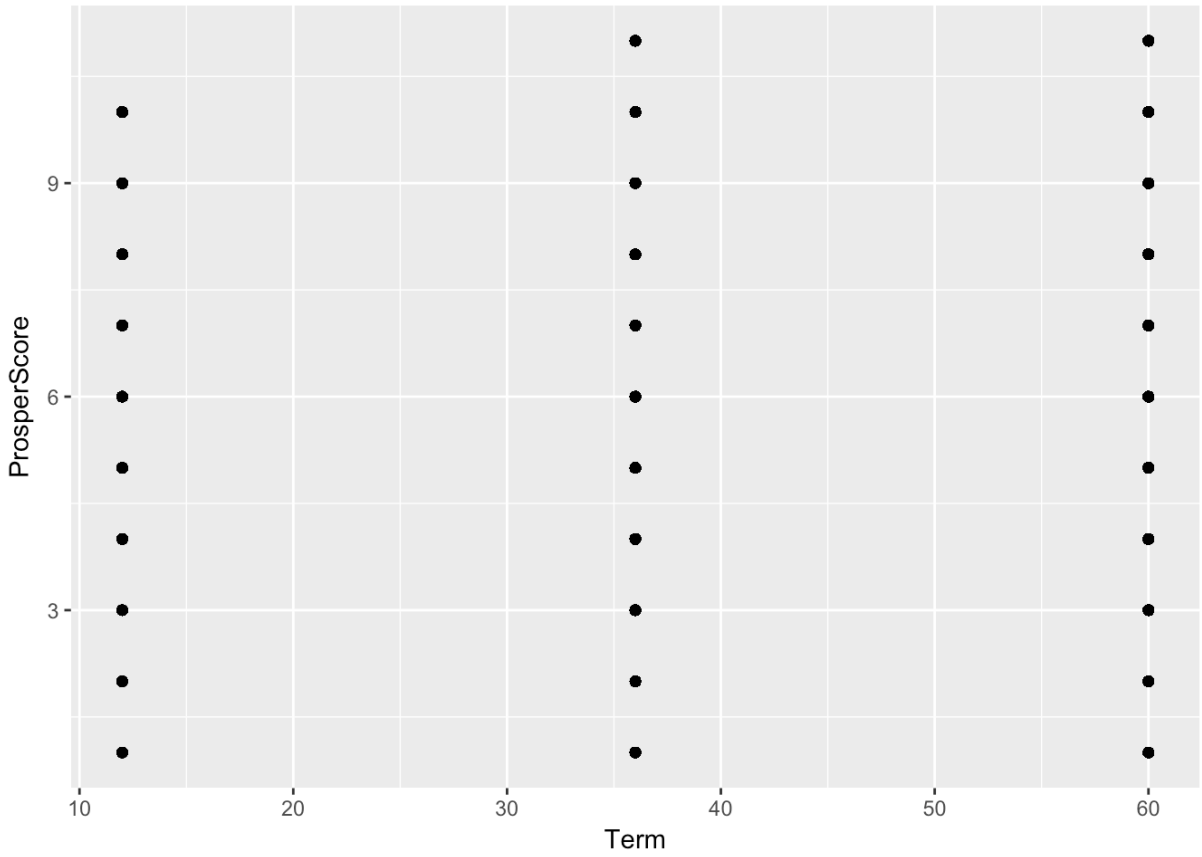
The data points in bivariate scatterplots need to be made transparent (`alpha` parameter can be added for this purpose) as they have clusters where the density of data points are not directly comparable. See this plot for example:



Please also correct other plots with the same issue.

Discrete vs discrete variables - Use heatmap or scatterplot with jitter and alpha

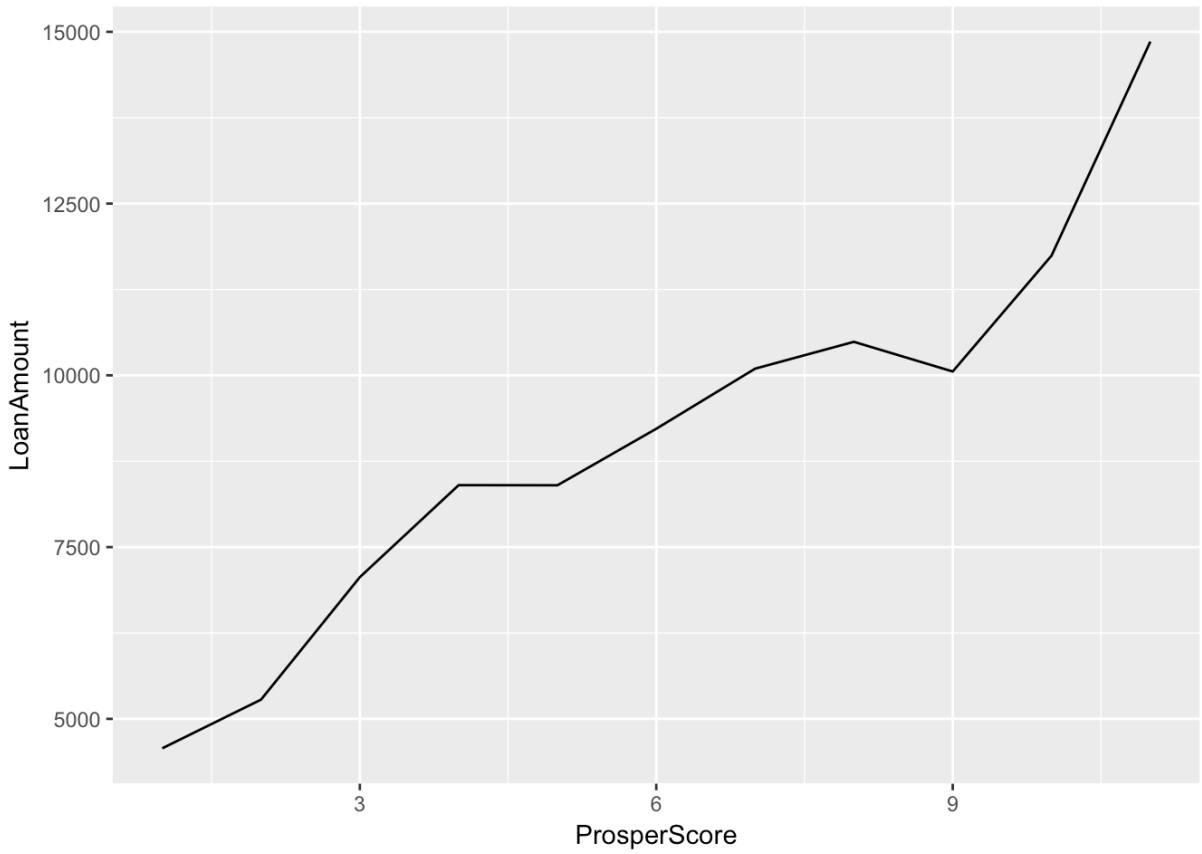
In this scatterplot:



It is not possible to appreciate the quantity of data points in each pair of variables. A better way to present this plot is by using a heatmap visualization by using modules such as geom_tile or geom_bin2d (geom_tile is usable when there are three variables, one of them being continuous, while geom_bin2d is a better option when aggregate values such as count are used in color intensity encoding).

Another alternative is to jitter the data points (i.e. use `geom_jitter` instead of `geom_point`) and add `alpha` transparency parameter so we can see how dense each cell is. This method does not show the approximate quantity as in heatmaps, however.

(Optional) Line plot used improperly (use boxplot instead)

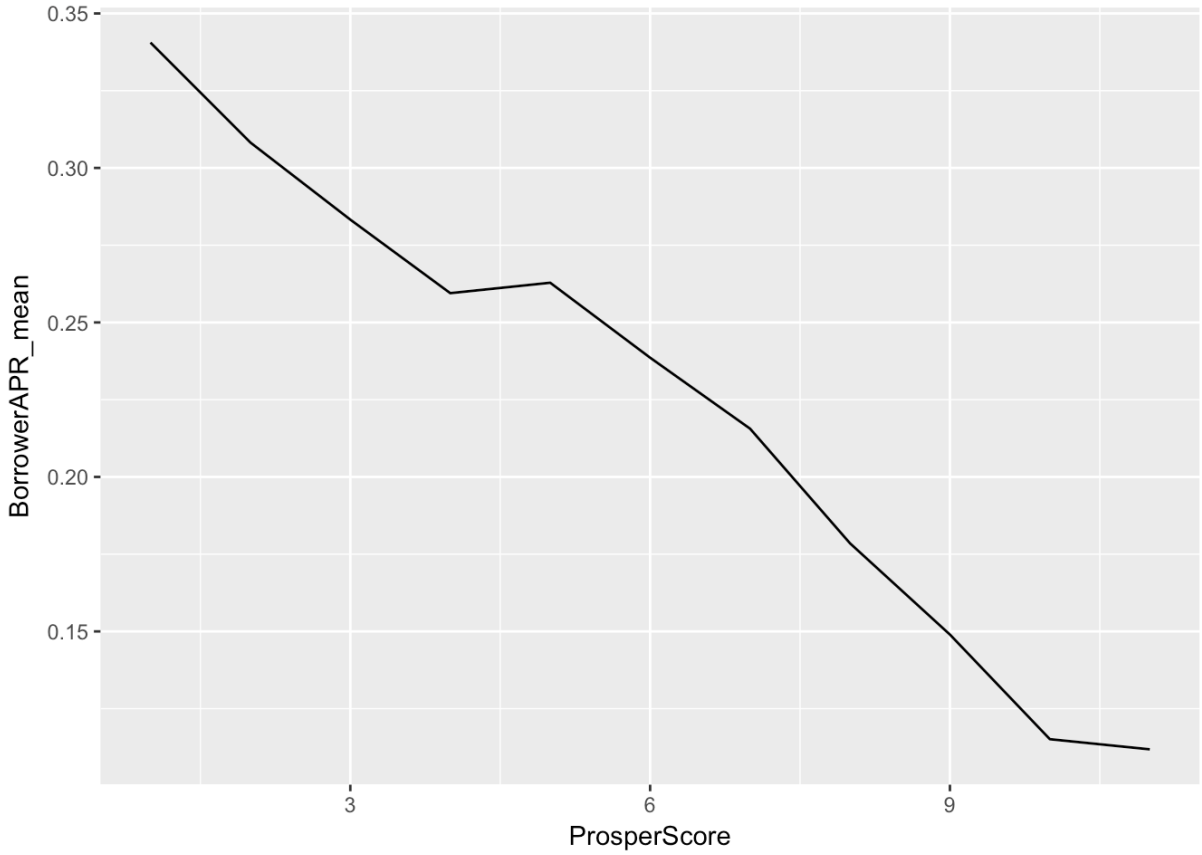


Line plot is most effective when the goal is to visualize changes in the x-axis data. In other words, one point in the x-axis must have a continuity or dependant on a point on its left and predates the point on its right i.e. periodical. That does not seem to be the case in the above visualization.

Read more here: https://infogram.com/blog/the-line-chart-how-and-when-to-use-it/

I would say a boxplot might be a better visualization type to use here, especially since readers can find other useful statistics from it, like quartiles and outliers for each category.

And for this plot:



Since the y-axis presents means, a bar chart is probably better as you can see exactly which prosper score has what value.

Final Plots and Summary

The project includes a Final Plots and Summary section containing three plots and commentary. All plots in this section reflect what has been explored in the main body of the analysis.

The plots are well chosen and the plots fulfill at least 2 of the criteria. The plots are varied and reveal interesting trends and relationships.

All plots have appropriately selected variables and are plotted in a way that accurately conveys the data/information (i.e findings in Final Plot 1 do not depend on the findings of Final Plot 2).

I suggest improving the second and third final plots as suggested above, if you wish.

All plots are labeled appropriately (axis labels, plot titles, axis units) and can be read and interpreted easily. Plots are scaled appropriately.

Missing titles

Plot titles need to be added to all of the final plots. Title is important to keep the plot's context when it is saved as an image. You may add a `ggtitle()` layer to add a title to a plot.

Missing unit of measurements

Units of measurements need to be added in final plots, in this case the currency used for `LoanOriginalAmount`.

I understand it does sound picky, but adding a unit of measurement even though it is obvious is *super important*. NASA can tell you more about it:

> NASA lost its $125-million Mars Climate Orbiter because spacecraft engineers failed to convert from English to metric measurements when exchanging vital data before the craft was launched, space agency officials said Thursday.
>
> A navigation team at the Jet Propulsion Laboratory used the metric system of millimeters and meters in its calculations, while Lockheed Martin Astronautics in Denver, which designed and built the spacecraft, provided crucial acceleration data in the English system of inches, feet and pounds.
>
> As a result, JPL engineers mistook acceleration readings measured in English units of pound-seconds for a metric measure of force called newton-seconds.
>
> In a sense, the spacecraft was lost in translation.

The reasoning and findings from each plot are explained and the text about each plot is descriptive enough to stand alone. Comments reflect the contents of the plots that they are associated with.

Reflection

The project includes a Reflection section discussing the analysis performed.

The section reflects on how the analysis was conducted and reports on the struggles and successes throughout the analysis. The section provides at least one idea or question for future work. The section explains any important decisions in the analysis and how those decisions affected the analysis.

Well written reflection section, good work! Your reflection reports the struggles and successes you went through during the analysis process, as well as explanations behind some important decisions made there. To pass this specification, please also provide at least one idea or question for future work related to this dataset.

RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH

Student FAQ