U UDACITY

Explore and Summarize Data

REVIEW
CODE REVIEW
HISTORY

SHARE YOUR ACCOMPLISHMENT! 🐦 📘

Requires Changes

3 SPECIFICATIONS REQUIRE CHANGES

This is a very great try as your first submission, especially the reasoning part did very well, you have build a very clear path in this EDA. 👍
There are a few more points need additional attention, like code formatting and final plot's labels,
please check the comments below and try to polish the report further.

Wish good luck!

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.

Code formatting did very well, good job!

Markdown syntax is used in the RMD file to improve readability of the knitted file.
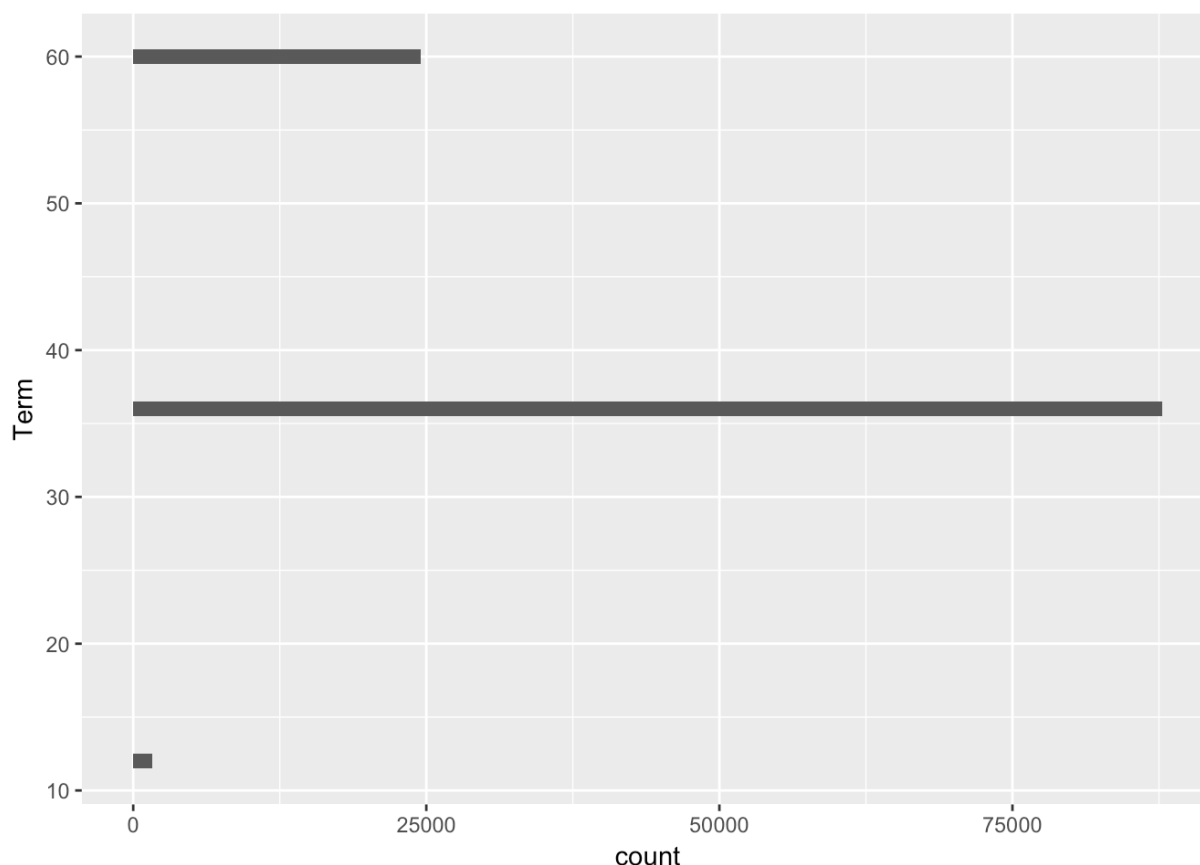
Good job on applying Markdown syntax properly.👍

Quality of Analysis

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

Variables in the dataset be explored in many ways, good job!

suggestion 1

- As mentioned in the video lessons, for univariate plots, in most cases a histogram should be used for quantitative data, while a bar chart should be used for categorical data. There are some cases where you might use one of the other plots, plots like above, it's better to try bar chat.
- It's a very good try to apply `coord_flip()` in the plot, but `coord_flip()` usually for situations that when we have a lot of category levels, or the category names are long that we might end up with overcrowding of the tick labels. Here `Term` just have 3 values, no need to flip the coordinate.
- Sample code for above plot: `ggplot(aes(x = Term), data = loanData) + geom_bar()`

suggestion 2

- In univariate plots session, we mainly want plotting be detailed enough to give a complete picture of the data, in your univariate plots session, there are several variables' explorations start with log transform, in order to make plots follow a logical flow to help the audience follow your thoughts through out the report, it's better to show the original distributions before axis transform.

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

Throughout the analysis, questions and observations be included as text regularly to help the audience follow your thoughts throughout the analysis, great job!👍

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

Reasoning is provided and comments following plots accurately reflect the plots' contents, great job!👍

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.

The visualizations are varied and show multiple comparisons and trends, good job!
To further enrich the analysis comments, you can accompany each plot with relevant statistics(e.g. mean, median, confidence intervals, correlations).

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.

Most of the plots are well chosen, there is one point need to be further improved:

1: requires changes
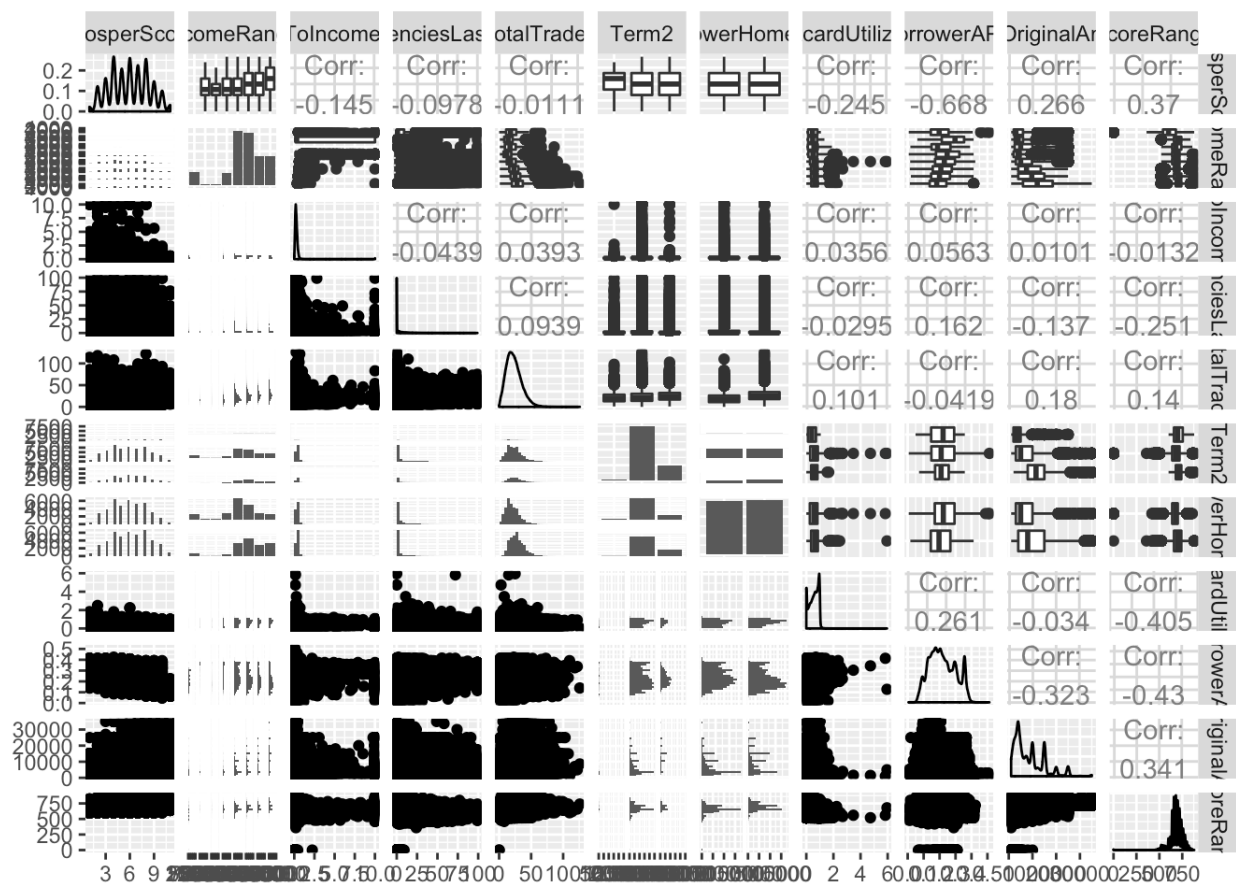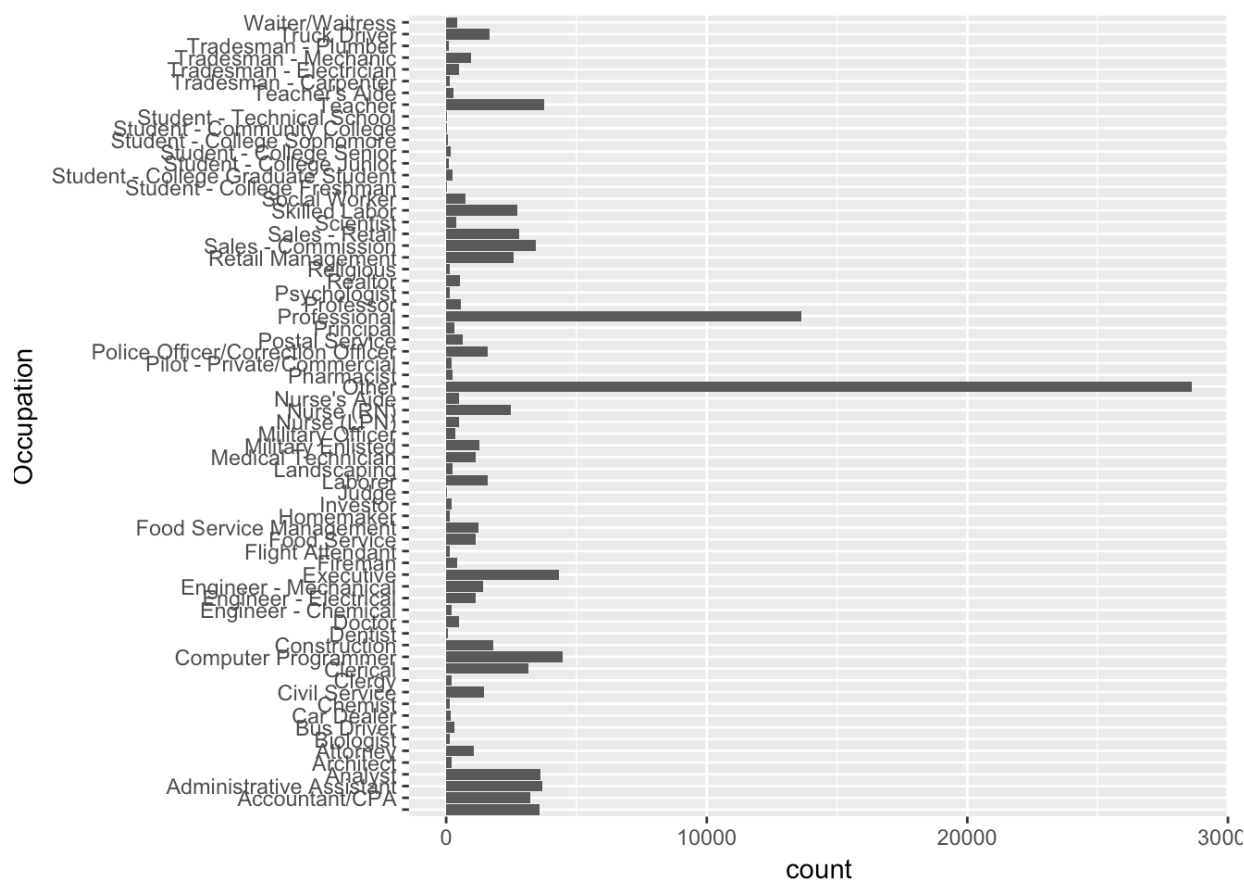
- Above plots have labels overlapping or not fully displayed, please try resize the text size or the plot size to improve report's readability.Plot can be resized by adding knitr chunk options fig.width and fig.height, eg:

```
{r Correlations, echo=FALSE, fig.height=20, fig.width=20, message=FALSE,
  warning=FALSE}
```

- please check GGally::ggpairs for further improvements for ggpairs.

2: requires changes
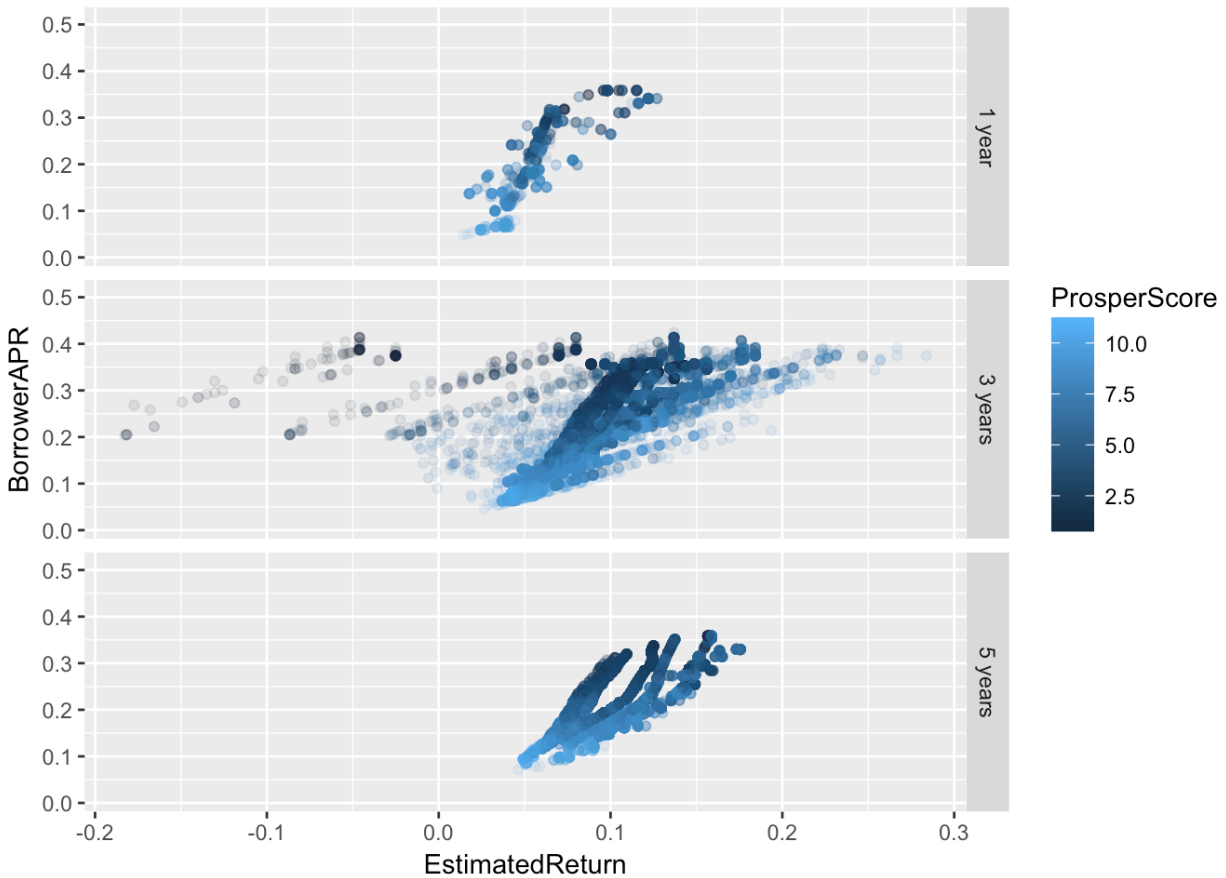
When using color as an encoding, please make sure that the palette being used is appropriate for the variable types being plotted. In the case of categorical data, the color scheme chosen should not imply a continuous relationship between levels through hue or lightness. For ordered data, the color scheme should reflect a relationship between consecutive levels of the feature being plotted.

Plots like above, rating or score are ordinal variable, values are expressing levels here(you can use `factor()` function to transform score to be a factor in the plot), ordinal variables should be encoded by ordinal color themes, it's better to use a sequential or diverging color palette, color palette can be set by using `scale_color_brewer` or `scale_fill_brewer` function, please check here for usage details.
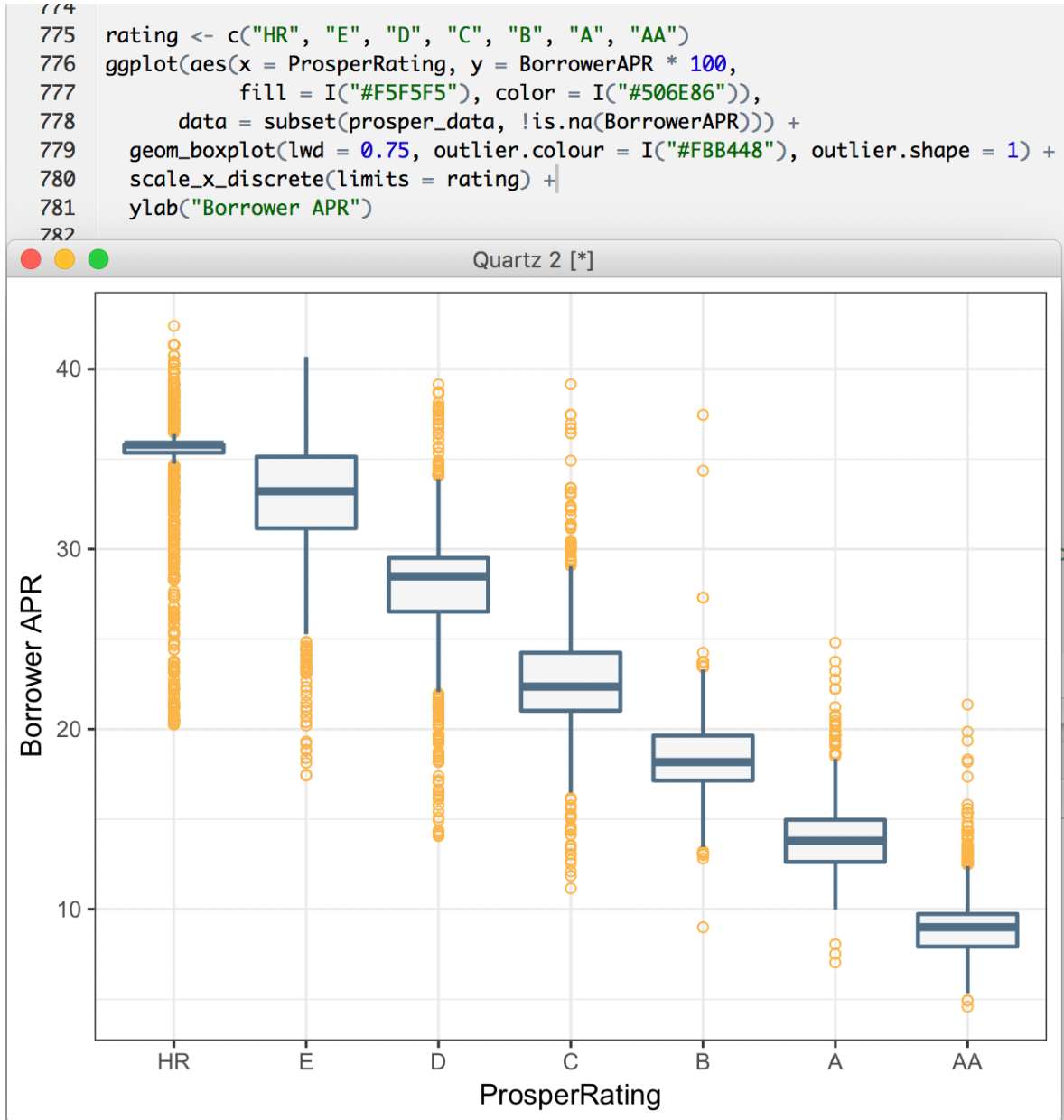
3: Suggestion for plots visualizations:

- When choosing plots type, for two numeric variables, correlation coefficient and scatterplot is a good choice, for categorical vs numerical variables, the most common exploring visualisation method is boxplot. While a line chart is common for data that you are watching over time. If you are comparing two categorical variables, the best choice is probably a side-by-side bar chart. Then in multivariate plots
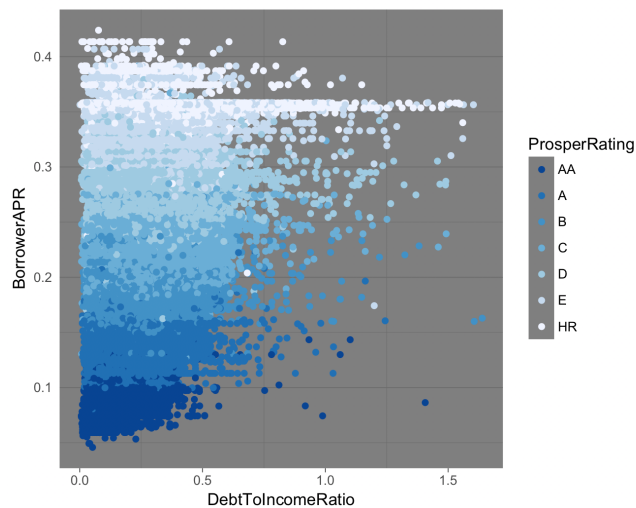
section, we can try use color to encoding or try `facet_wrap` function.

Please check here for further explanation.

- Sample plots:

```
774
775   rating <- c("HR", "E", "D", "C", "B", "A", "AA")
776   ggplot(aes(x = ProsperRating, y = BorrowerAPR * 100,
777              fill = I("#F5F5F5"), color = I("#506E86")),
778        data = subset(prosper_data, !is.na(BorrowerAPR))) +
779     geom_boxplot(lwd = 0.75, outlier.colour = I("#FBB448"), outlier.shape = 1) +
780     scale_x_discrete(limits = rating) +
781     ylab("Borrower APR")
782
```


BorrowerAPR VS DebtToIncomeRatio with colored ProsperRating

```{r}
prosper$ProsperRating <-
  factor(prosper$ProsperRating..Alpha,
         levels = c('AA', 'A', 'B', 'C', 'D', 'E', 'HR', 'NA'))
rating <- c("HR", "E", "D", "C", "B", "A", "AA")

ggplot(aes(x = DebtToIncomeRatio,
           y = BorrowerAPR,
           color =(ProsperRating)),
       data = subset(prosper,
                     !is.na(BorrowerAPR) & ProsperRating %in% rating)) +
  geom_jitter() +
  xlim(0, quantile(prosper$DebtToIncomeRatio, 0.995, na.rm = T)) +
  scale_color_brewer(name = 'ProsperRating', direction = -1) +
  theme_dark() +
  ggtitle("BorrowerAPR VS DebtToIncomeRatio with colored ProsperRating")
```

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.

Good job on choosing the final plots as a summary of your explorations above.

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).

Please refer to the visualization unit's comments to further polish final plots.

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

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, this session should also provide at least one idea or question for future work, like how could the analysis be enriched in future work (e.g. additional data and analyses), please add this part.

☑ RESUBMIT

⬇ DOWNLOAD PROJECT

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

▶ Watch Video (3:01)

RETURN TO PATH

Rate this review

Student FAQ