UDACITY

PROJECT

Explore and Summarize Data

A part of the Data Analyst Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

7 SPECIFICATIONS REQUIRE CHANGES

I find this project really well done and interesting. Through the project, you provided statistics accompanied by charts and discussions. That makes it very easy to track your line of thought. So any criticism I have is really getting down to nitpicks and shouldn't make you feel like this isn't an awesome job.

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.

Well Done for demonstrating the use of functions that reduce repetitions and simplify the code.

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 some places where you exceed the maximum line length. This seems picky but the limit is a widespread convention that ensures that future programmers can read your code easily no matter what their text editor and window size preferences are. One way to hem things in is by breaking up lists with line breaks. RStudio does the indentation automatically when you add a line break in the middle of a parameter list. RStudio also has a built in feature for finding overly long lines. In the Code Editing section of the preferences, there's an option called "Show margin" that puts a line length indicator in the code editor.

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

Unless the code includes any complicated data transformations that would give context to the plots and analysis, please avoid displaying chunks of code in your HTML document, this also implies for warnings messages. The final HTML document should include results from the analysis, figures, and discussion. You can easily do it by setting the parameter "echo=FALSE" as in this example: “{r echo=FALSE, message=FALSE, warning=FALSE, packages}”

Quality of Analysis

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

For the univariate section, it is important to include a simple histogram (continues feature) or a bar plot (categorical feature) to depict the count distribution for each feature that is included in the analysis. That will allow you to examine the distribution, identify outliers and make changes to the dataset before starting the analysis.

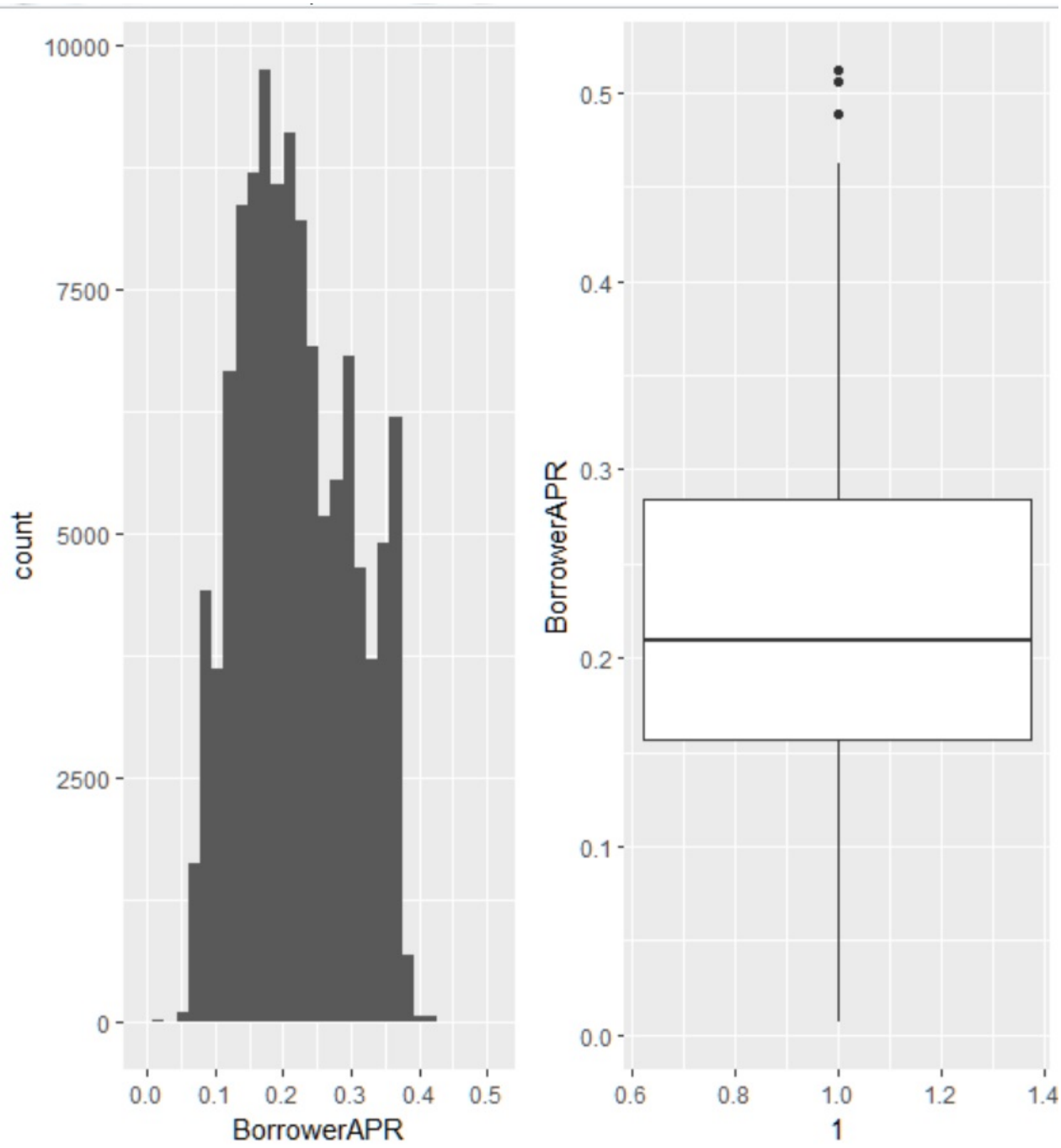
Questions and findings are placed between blocks of R code regularly so it is clear what the student was thinking throughout the analysis.

The discussion between code block includes relevant questions and interesting findings. It is great that you summarize the results and insights after each section, that make it easier for the readers to follow the analysis.

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

The analysis follows a logical flow where the results of one analysis lead to another.
For the univariate section, please consider expanding the discussion about the outliers for each feature. You can even remove outliers if you find it appropriate, that will make the following analysis more robust.
http://www.public.iastate.edu/~maitra/stat501/lectures/Outliers.pdf
You can use a simple boxplot to depict these outliers

```
grid.arrange( ggplot(aes(x=BorrowerAPR),
        data = pl) +
  geom_histogram( bins = 30) ,
   ggplot(aes(x=1, y=BorrowerAPR),
        data = pl) +
  geom_boxplot( )  , nrow =1)
```



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.

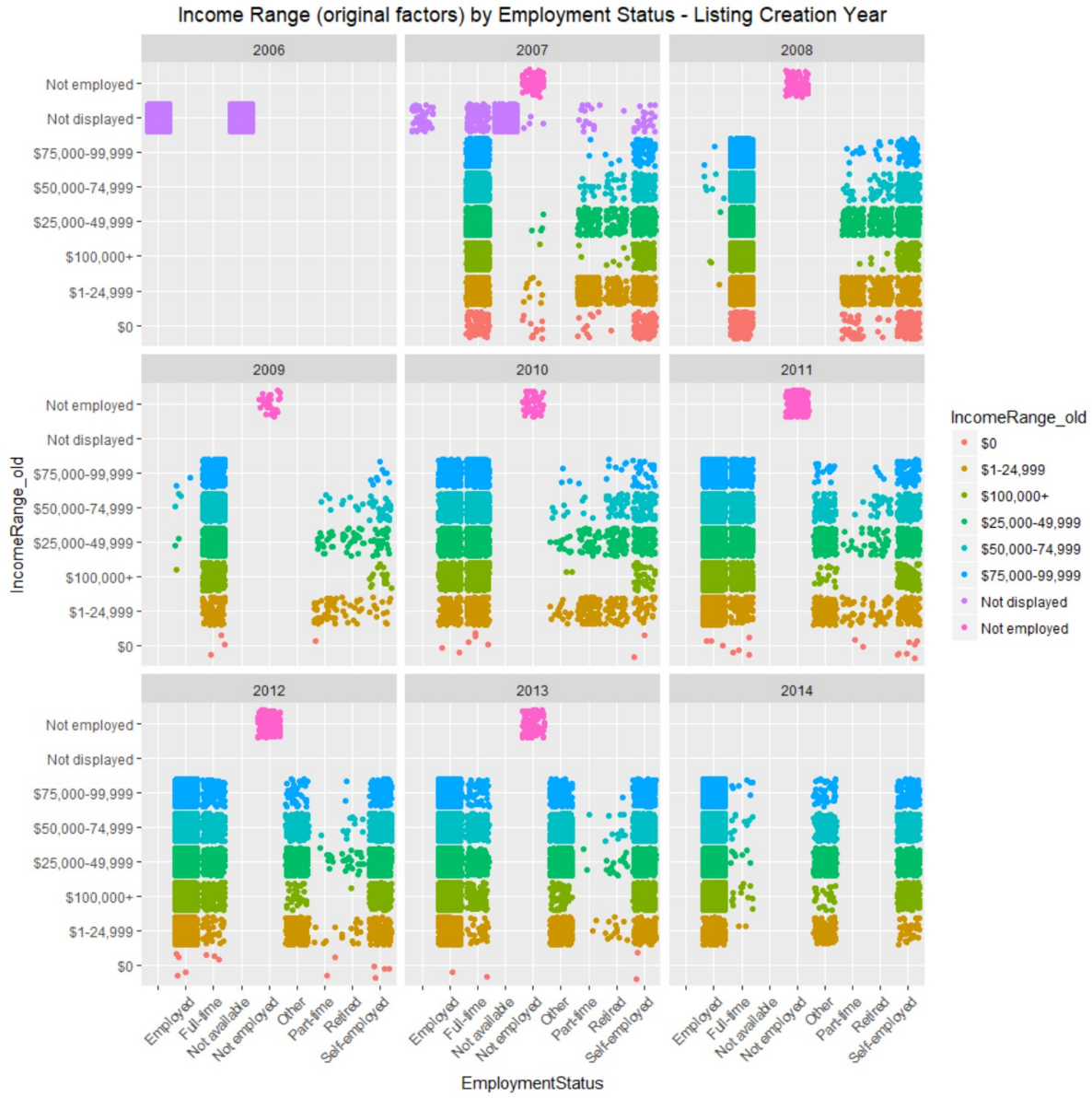
The analysis includes many figures that depict comparison, trends and relations between features. It is awesome that you include the relevant statistics in the discussion under each chart.

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.

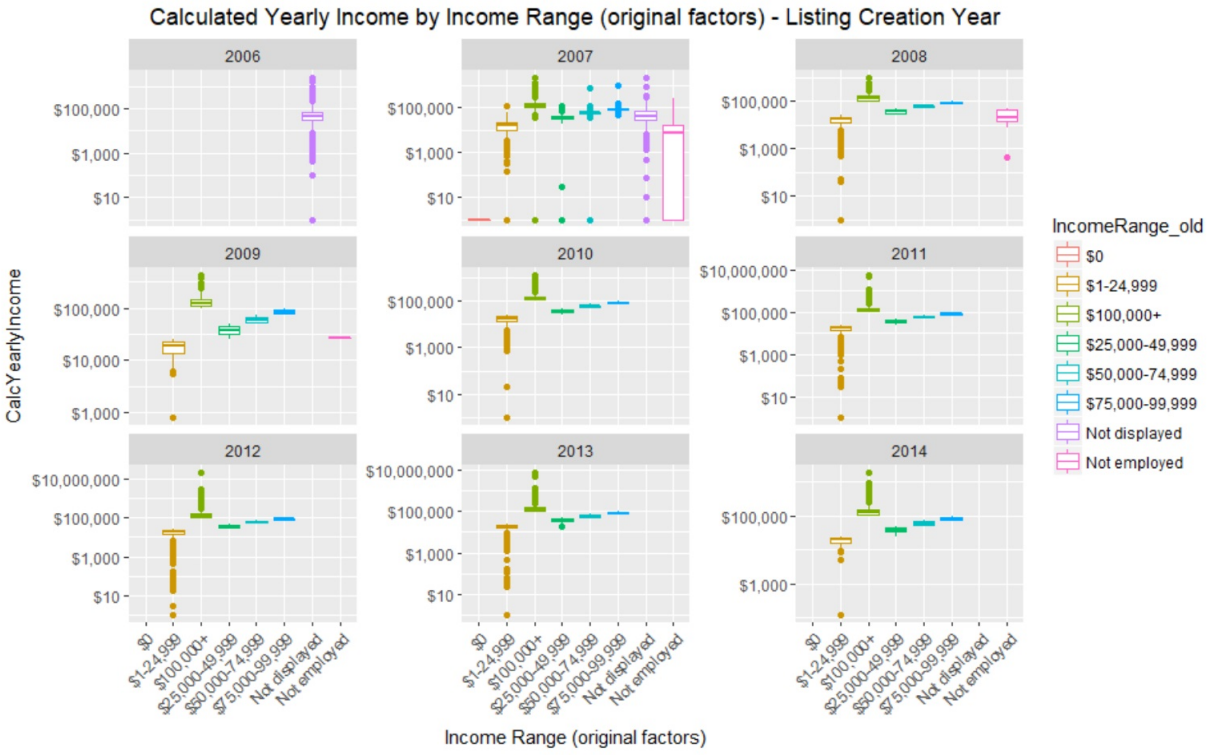
Most of the chart are well done, so I only have few comment here,

When the figure depicts 2 categorical features a scatter plot is not appropriate since you can't appreciate the number of samples in each category. Please use a heat map or other alternatives instead.

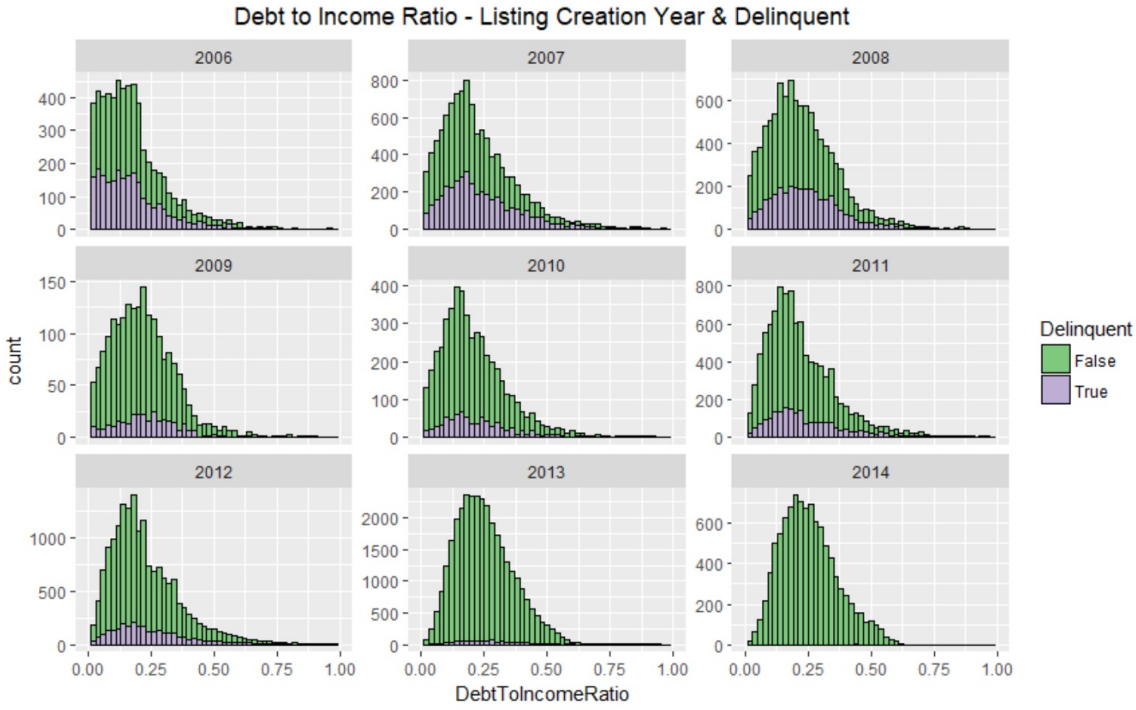


When the feature code by the color is categorical and ordered, it is important to use a sequential color map, for example. for the figure here below, one alternative is

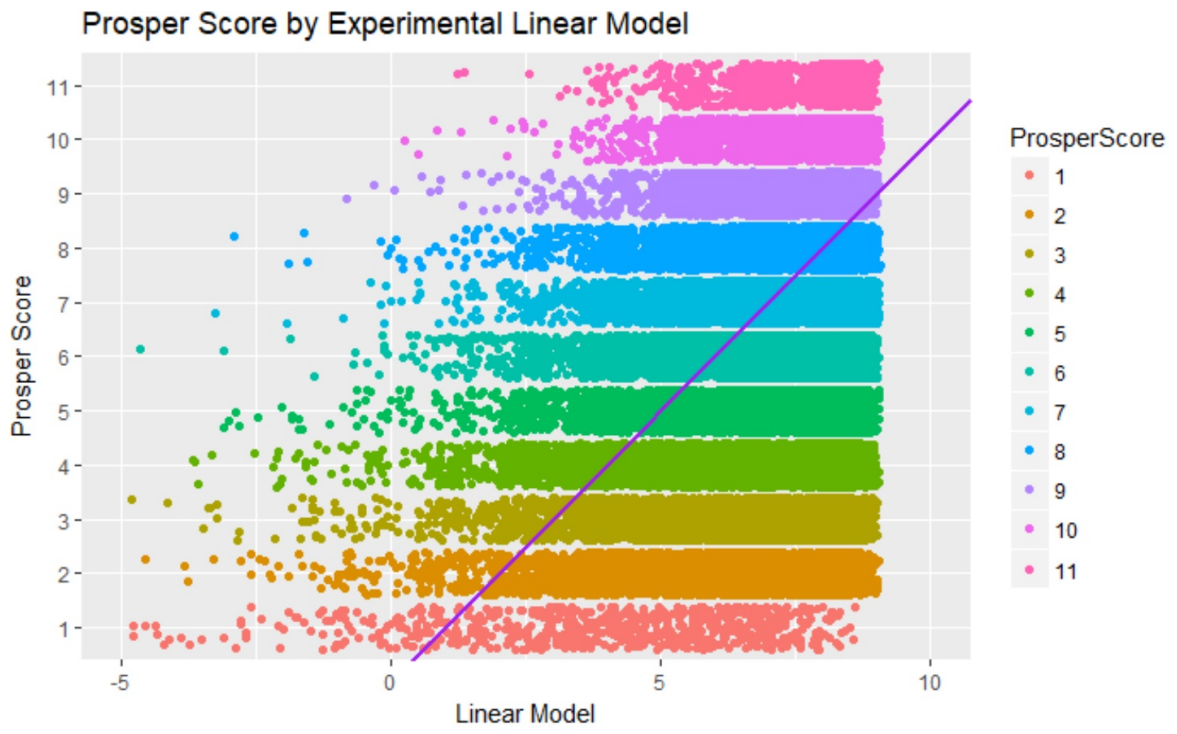
```
scale_fill_brewer(type='seq' )
```



Optional, For the figure here below, you might find that a box plot is a better option to depict the distribution of different categories.



The chart here below but also similar charts are over plotted, you can choose smaller dot size and lower alpha value. Another alternative is using a violin or a box plot.



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.

For the final plot section , please include only 3 figures. The idea is to presents the 3 most significant findings from the exploration. Please identify the 3 most significant findings and create a single figure for each such idea.

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 make sure that you are using sequential color map when appropriate, for example, final plot 1 , final plot 3 etc.

Final plot 6 is overplotted, please decrease the dot size or use a box plot instead (perhaps the color is not so significant here since the credit score seems to be highly correlated with the prosper score).

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.

Please expand the reflection section to discuss the struggles and successes throughout the analysis.

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

Student FAQ