UDACITY

PROJECT

Explore and Summarize Data

A part of the Data Analyst Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Meets Specifications

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.

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

Quality of Analysis

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

The analysis makes use of different chart types, including univariate, bivariate and multivariate to explorers and investigates many aspects of the data set. The univariate investigation includes a simple count distribution for each feature explored in 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.

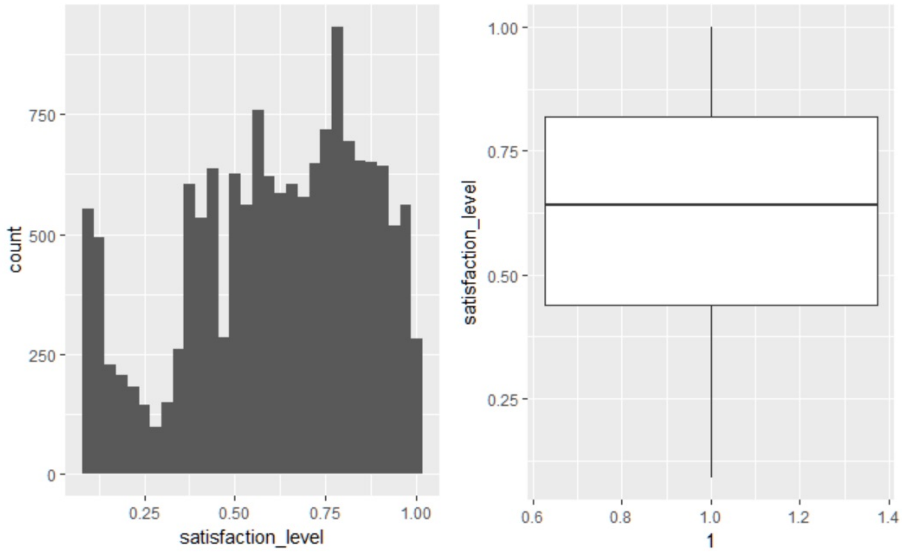
Reasoning is provided for the plots made throughout the analysis. Plots made follow a logical flow. Comments following plots accurately reflect the plots’ contents.

It is important to include a short discussion under each chart, explain what the chart depicts and what are your insights.

For the bivariate and multivariate sections, the discussion should also include reasonings, why you choose to include the analysis in the report? what did you expect to find? what are your insights?

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=satisfaction_level),
       data = df) +
 geom_histogram( bins = 30) ,
  ggplot(aes(x=1, y=satisfaction_level),
       data = df) +
 geom_boxplot( )  , nrow =1)
```



Please consider starting the bivariate section with calculating the correlation values between each couple of numerical features. That can guide and focus the following analysis.
https://briatte.github.io/ggcorr/#controlling-the-coefficient-labels

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.

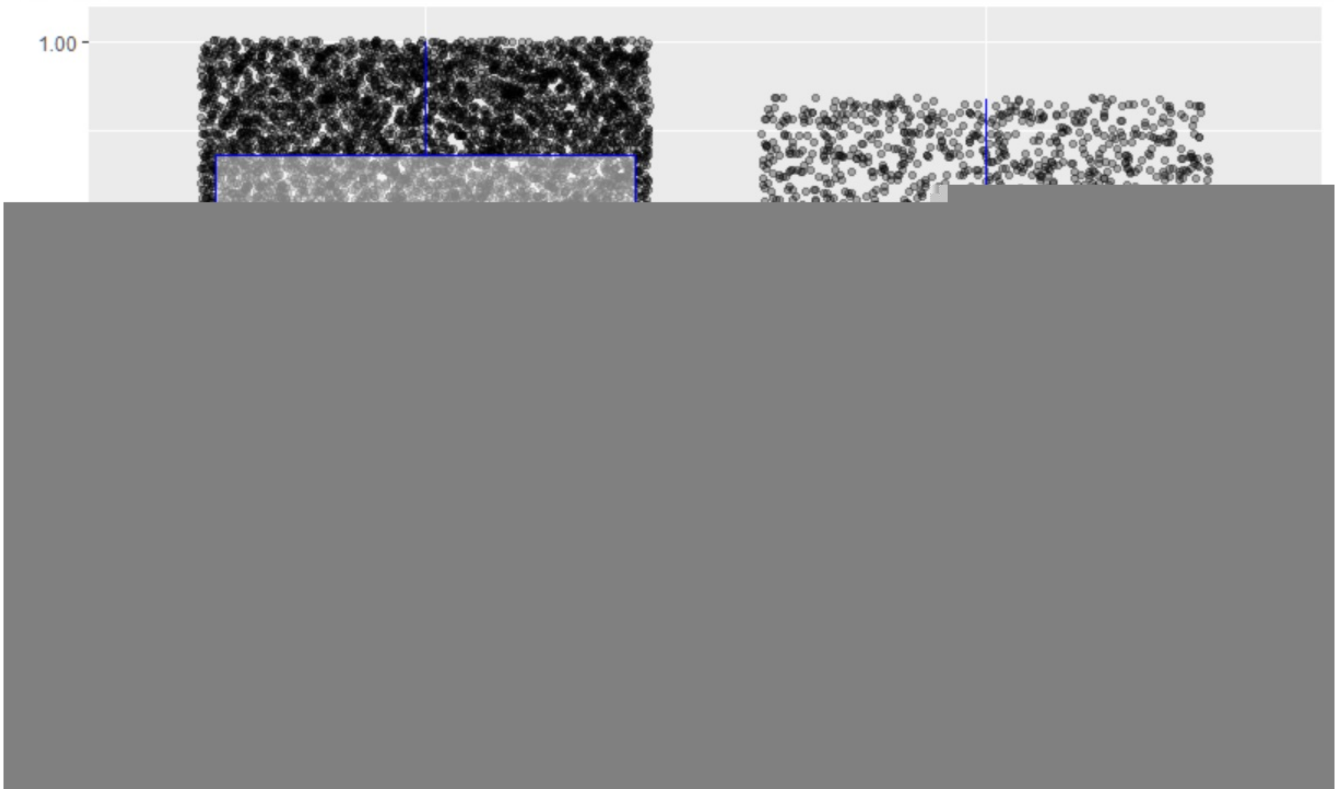
Visualizations made in the project depict the data in an appropriate manner that allows plots to be readily interpreted. Choice of plot type, variables, and aesthetic parameters (e.g. bin width, color, axis breaks) is appropriate.

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

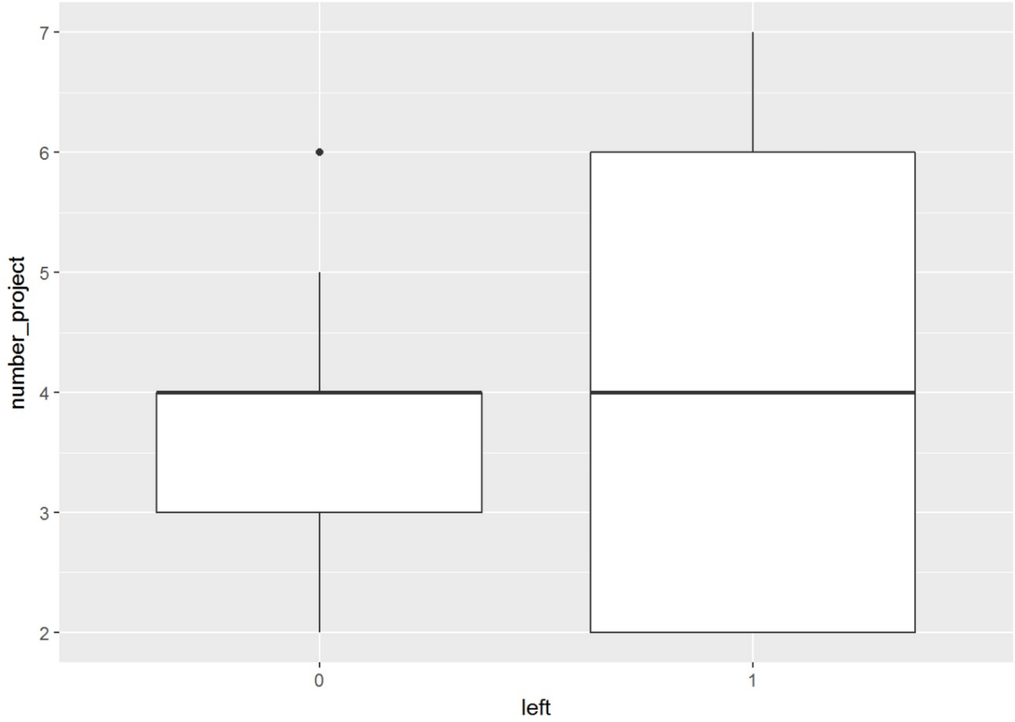
Optional, For the "number_projects", since this is a categorical feature, a bar plot is more appropriate to depict the count distribution. A histogram is optimal to depict the distribution of continuous features.

Optional, for the boxplots, you can also add markers to show the mean values and jitter plot, that will make the figure more informative,

```
ggplot(aes(factor(left),
           satisfaction_level),
       data = df) +
 geom_jitter( alpha = .3)  +
 geom_boxplot( alpha = .5,color = 'blue')+
 stat_summary(fun.y = "mean",
              geom = "point",
              color = "red",
              shape = 8,
              size = 4)
```



When the chart depicts 2 categorical features, a heat map is more appropriate,



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 final plot section includes figures that represent the analysis and demonstrate the significant findings from the exploration sections.

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).

For final plot 1, please consider combining the boxplot with the scatter plot, that will save you one figure.

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.

DOWNLOAD PROJECT

RETURN TO PATH

Student FAQ