UDACITY

PROJECT

Explore and Summarize Data

A part of the Data Analyst Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Meets Specifications

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 type that explores many aspects of the data set.

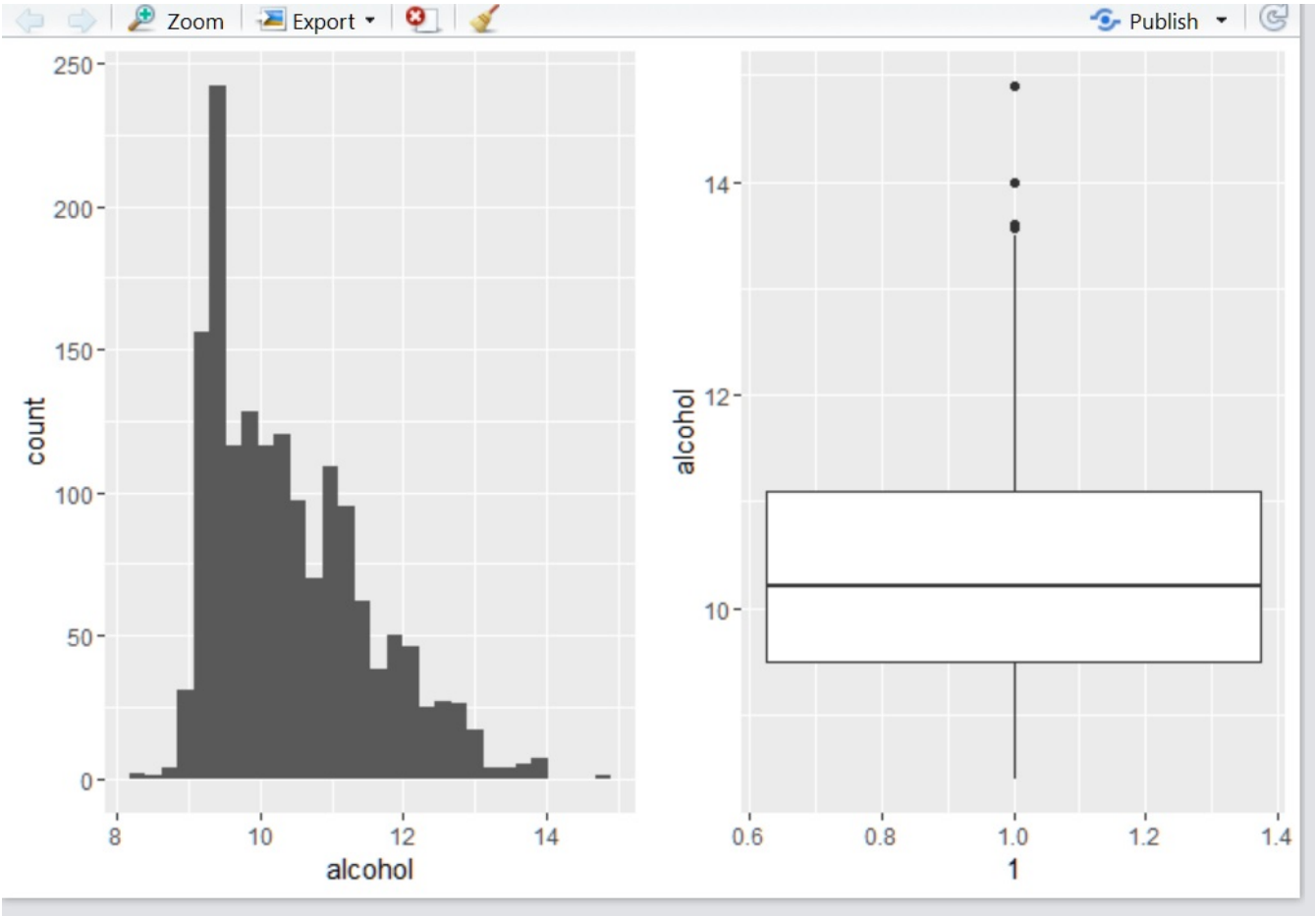
Questions and findings are placed between blocks of R code regularly so it is clear what the student was thinking throughout the analysis.

The discussion between code block includes relevant questions and interesting findings.

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, It is great that you include a 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 also use a boxplot to depict the outliers.

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



Well Done for starting the bivariate section with the correlation analysis that allow you to focus and guide the following analysis.

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.

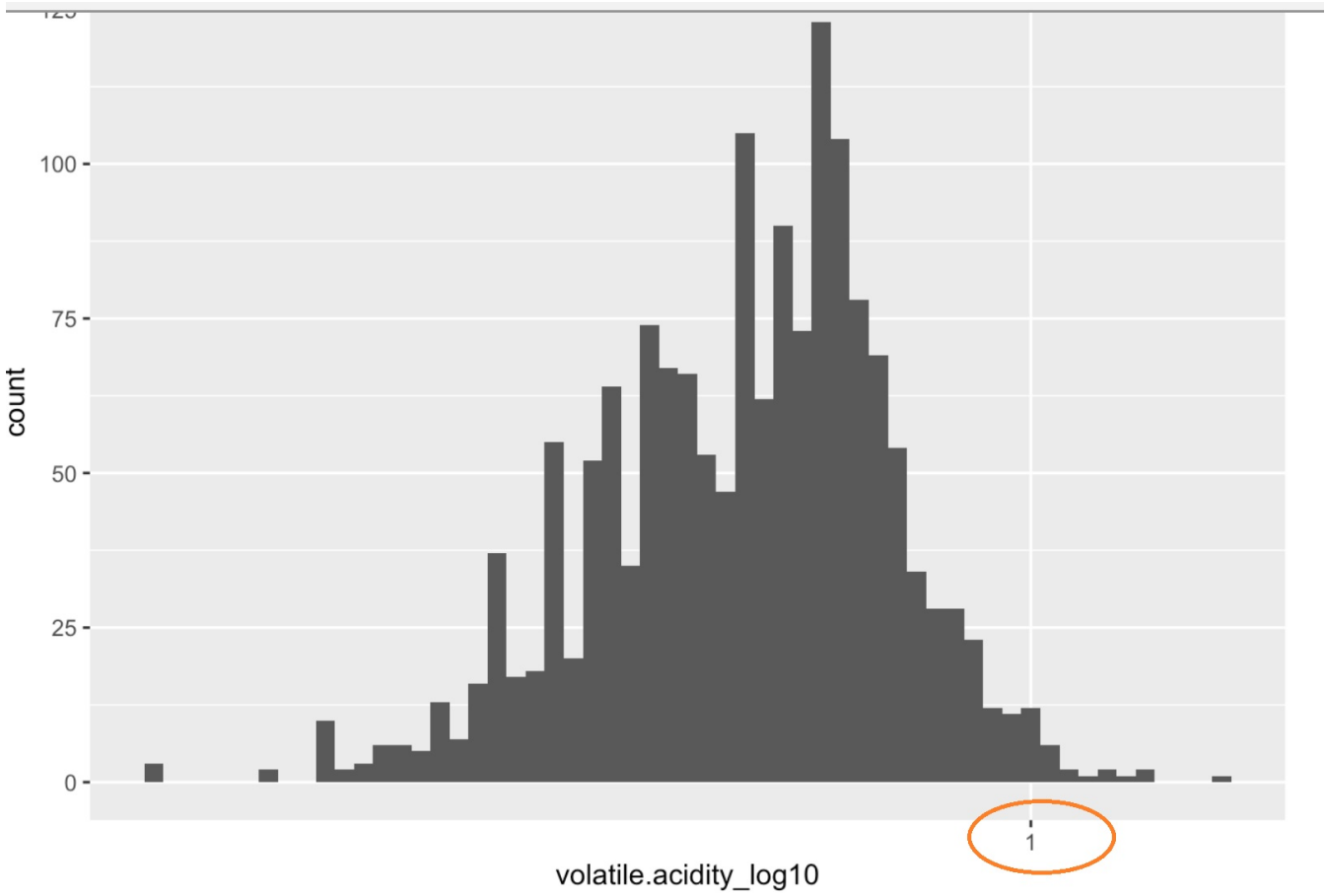
It is excellent that you add the relevant statistics (correlation values and summary statistics) next to each chart. That make the interpretation so easy and clear.

For the multivariate scatter plot, you can also calculate the correlation value for each category (color).

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.

The charts are well done, so there are only a few comments here,

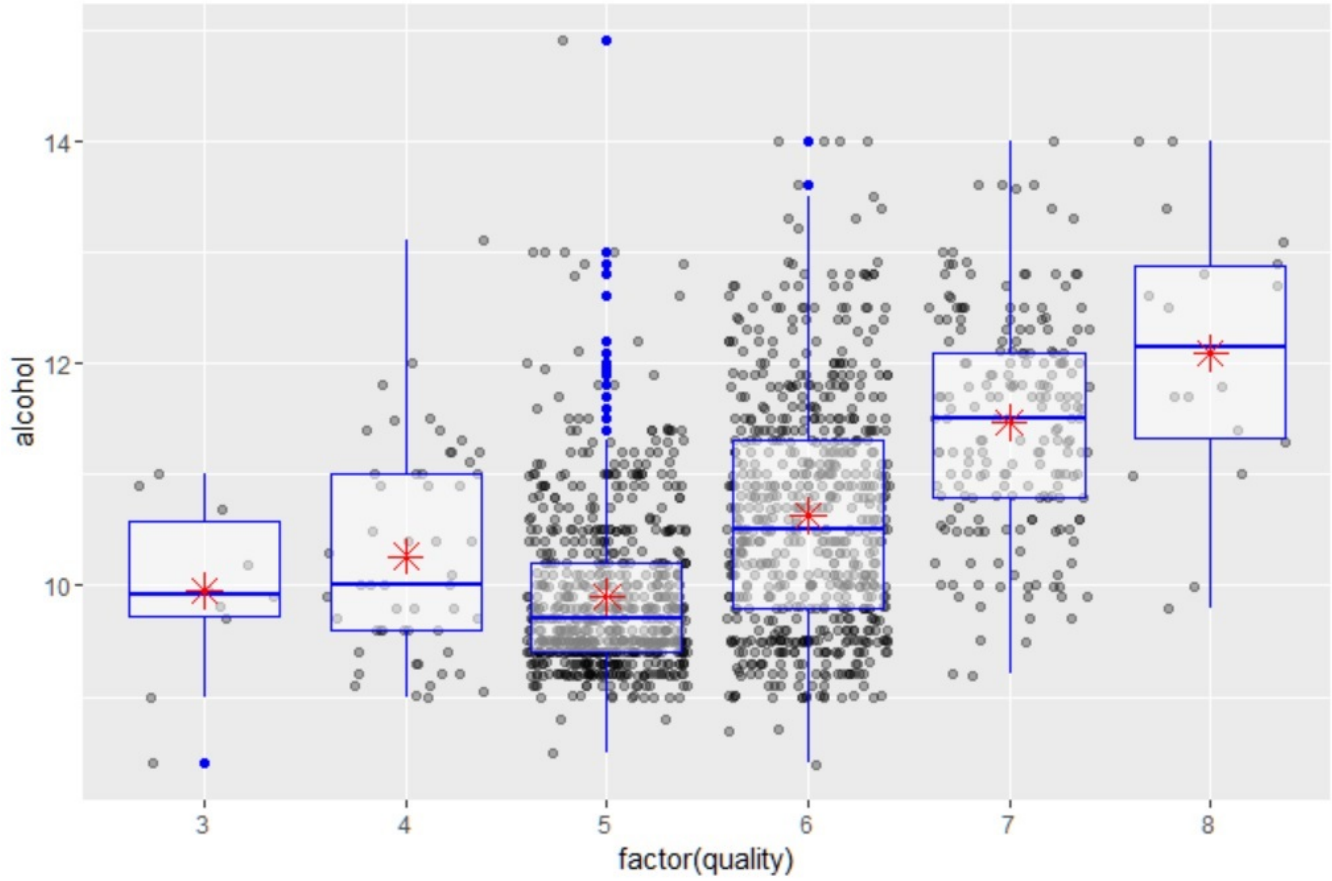
For the histogram, but also other chars, make sure that each axis includes at least 2q ticks that will allow you to appreciate the distribution.



For the correlation analysis, please consider the following, https://briatte.github.io/ggcorr/#controlling-the-coefficient-labels

For the box plot you can also include the scatter plot as another layer, for example

```
ggplot(aes(factor(quality),
           alcohol),
       data = red.wine) +
 geom_jitter( alpha = .3)  +
 geom_boxplot( alpha = .5,color = 'blue')+
 stat_summary(fun.y = "mean",
              geom = "point",
              color = "red",
              shape = 8,
              size = 4)+
 geom_smooth(aes(quality-2,
                  alcohol),
              method = "lm",
              se = FALSE,size=2)
```



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 include 3 charts that depict the analysis done in the exploration set.

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).

The image that depicts the relation between the quality and volatile.acidity appear twice.

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.

Please consider including more relevant statistics to quantifies the results and insights in the final plot section.

Reflection

The project includes a Reflection section discussing the analysis performed.