U DACITY

PROJECT

## Explore and Summarize Data
A part of the Data Analyst Nanodegree Program

| PROJECT REVIEW |
| :---: |
| CODE REVIEW |
| NOTES |

## Requires Changes

**4 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.**

Please consider to include functions to simplify your code or reduce repetitions.

## 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.**

## 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.

For the bivariate since quality is the main feature of interest, please consider including more chart that explores the relationship between quality. Please consider the following.

```
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)
```
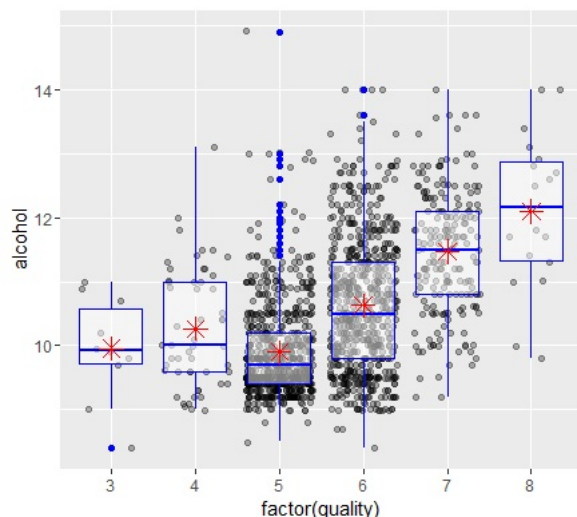


Questions and findings are placed between blocks of R code regularly so it is clear what the student was thinking throughout the analysis.

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.

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(red.wine, aes( x = 1, y = residual.sugar ) ) +
             geom_jitter(alpha = 0.1 ) +
             geom_boxplot(alpha = 0.2, color = 'red' ) ,
           ggplot(red.wine, aes( x   = residual.sugar  ) ) +
               geom_histogram(bins=30 ),ncol=2)
```
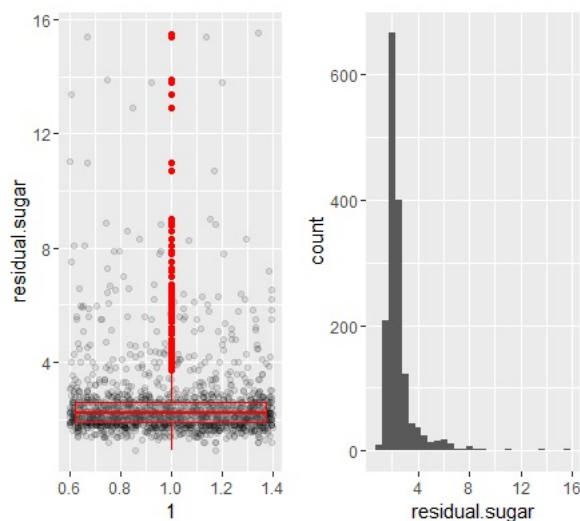
Please consider starting the bivariate section with calculating the correlation values between each couple of numerical features. That can guide and focus the following analysis.
https://briatte.github.io/ggcorr/#controlling-the-coefficient-labels

**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. Please consider to include more relevant statistics in the discussion under each chart. You can include the mean median and quartiles to quantify the distribution under the histograms. The relevant correlation value to quantify the relation in the bivariate section.

**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,

For the multivariate scatter plots, it is important to use a color table that express the fact that the feature is ordered and categorical, in addition, I strongly encourage you to add a regression line for each category to depict the separation.

```
ggplot(aes(x = alcohol,
           y = residual.sugar  , color = factor(quality)),
       data = red.wine) +
    geom_point(alpha = 0.8, size = 1) +
    geom_smooth(method = "lm", se = FALSE,size=1)  +
  scale_color_brewer(type='seq',
                   guide=guide_legend(title='Quality'))
```
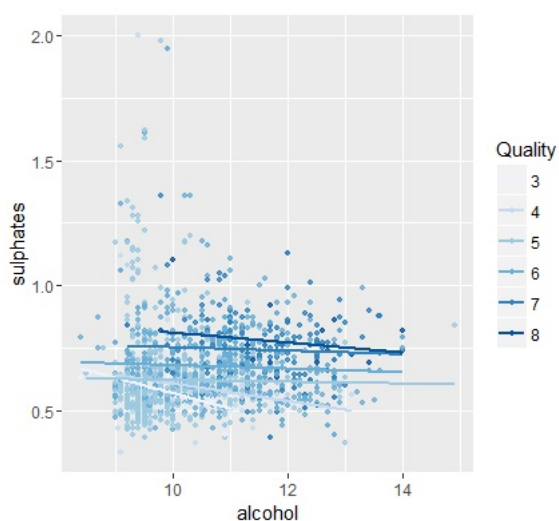
## Final Plots and Summary

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 3 please choose appropriate color map as described above.

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

For final plot 3, the y-axis is not the quality, please double check the code,

Please add the units to the axis labels, For example, " alcohol [%]".

Please include a descriptive title for each figure in the final plot section.

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 extend your discussion to all the charts in the final plot section. The discussion under each chart should include the reasoning for the choice of each chart, the insights and ideally the relevant supporting statistics. Keep in mind also that the final plot descriptions should 'stand alone'. If you didn't have the plot attached, would reading your descriptions be enough to give you a clear picture of the trends you are trying to show?

## 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.

☑ 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**