Stat. 651 Final

Lydia Gibson

1. Visit the Data Is Beautiful YouTube channel and watch a few of the videos. (Or find another similar channel, such as Data is Public.) These kinds of visualizations are quite popular these days. a) Describe in detail the different kinds visualizations that are present in the videos. b) Comment on what is wrong with the Most Popular Music Styles 1910 - 2019 video.

a)

The Data Is Beautiful YouTube videos are dynamic horizontal bar charts that show changing popularity of categories such as search engines, browsers, computers, websites, etc., over time. The time elapses over quarters and can span from several years to several decades. The bar charts tend to be colorful, with the exception of a handful, and the videos have background music. Often time there will be a logo used to distinguish/add context to the bars, for example the NBC peacock in the *Most Popular TV Series* videos. In the lower right hand corner of the video you will see the changing year and quarter as time in the video elapses.

b)

In the *Most Popular Music Styles 1910 - 2019* video we see that the top bar serves as a baseline for all the other bars and although the number changes, the length of the bar never changes but rather the axis values change over time. There is no mention of what the scale or measurement value is, but is presumably in percentages. There are several different colors used for the bars and while some colors or varying shades of the color appear repeatedly, there is no obvious association between the bars that may share that color. There is no obvious relationship between the choice of sound and the plot itself. There was a missed oppurtunity to change the background music to correspond with the most popular music genre at the time, but the sound remained the same throughout the video.

2. Read over Chapter 7 of the r4ds book. Do 7.5.1.1. Exercises 4 using the lv_plot and also using the violin plot. Compare your two new plots to using a boxplot.

R4DS 7.5.1.1. Exercise 4: One problem with boxplots is that they were developed in an era of much smaller datasets and tend to display a prohibitively large number of "outlying values". One approach to remedy this problem is the letter value plot. Install the `lvplot` package, and try using `geom_lv()` to display the distribution of price vs cut. What do you learn? How do you interpret the plots?

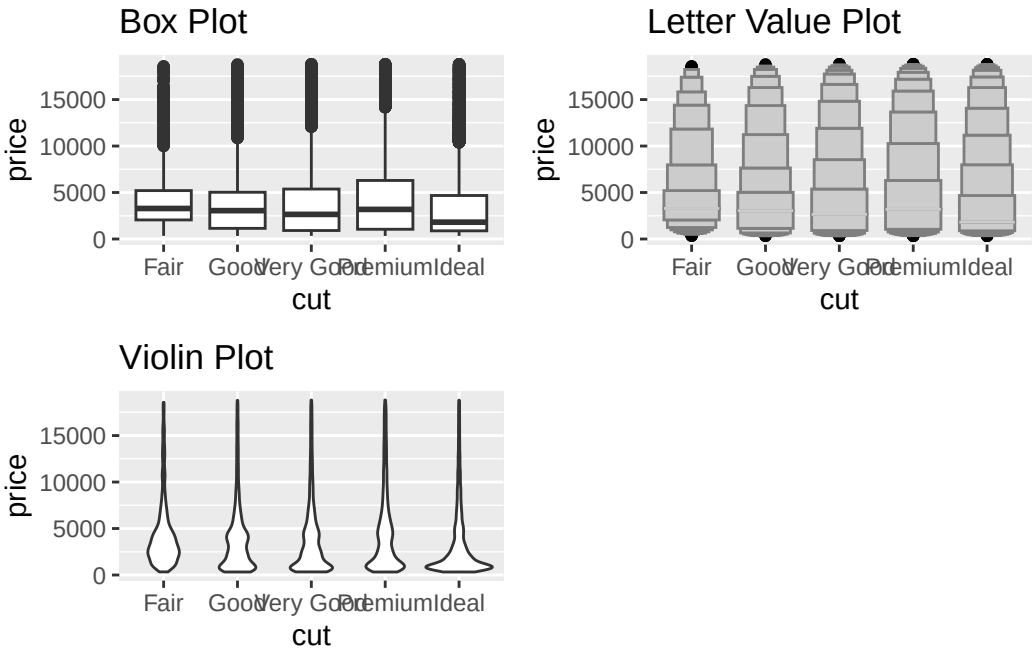
```
library(pacman)
p_load(ggplot2, lvplot, ggstatsplot)
```

```
p1 <- ggplot(data = diamonds, mapping = aes(x = cut, y = price)) +
  geom_boxplot() +
  ggtitle("Box Plot")

p2 <- ggplot(data = diamonds, mapping = aes(x = cut, y = price)) +
  geom_lv() +
  ggtitle("Letter Value Plot")

p3 <- ggplot(data = diamonds, mapping = aes(x = cut, y = price)) +
  geom_violin()  +
  ggtitle("Violin Plot")

cowplot::plot_grid(p1, p2, p3)
```



Excepted from *geom_lv: Side-by-side LV boxplots with ggplot2* RDocumentation : **Letter value plots** are "an extension of standard boxplots which **draws k letter statistics**. Conventional boxplots (Tukey 1977) are useful displays for conveying rough information about the central 50% of the data and the extent of the data. For moderate-sized data sets, detailed estimates of tail behavior beyond the *quartiles* may not be trustworthy, so the information provided by boxplots is appropriately somewhat vague beyond the *quartiles*. Large data sets afford more precise estimates of *quantiles* in the tails beyond the *quartiles* and also can be expected to present a large number of outliers. **The letter-value box plot** addresses both these shortcomings: it **conveys more detailed information in the tails** using letter values, only out to the depths where the letter values are reliable estimates of their corresponding *quantiles* (corresponding to tail areas); outliers are defined as a function of the most extreme letter value shown. All aspects shown on the letter-value boxplot are actual observations, thus remaining faithful to the principles that governed Tukey's original boxplot."

As seen above, the letter value plot gives us more details about outlier points than does a box plot, which mostly serves to inform us about the *quartiles* of data rather than their *quantiles*. Due to the size of the data set, we notice a large number of outliers in the above boxplot but have very little else information about those points. The letter value plot serves to fill in this knowledge gap between our data and the box plot, especially for larger data sets where many more observations would be considered outliers.

Much in the same way that the violin plot serves to show the density of observations at a given value, the letter value plot also does this very nicely for our outliers. In a violin plot, you will find that the widest part is the mode of the distribution and that the long tails coincide with

outlier values. Each of these plots serves to give you pieces of information, that you could not get in the other.

3. Make your own Self Evaluation checklist. Review the best_practices.html presentation and make a one page check list for Evaluating your data visualizations for use in the future.

See attached.

4. Clearly explain how latitude and longitude data can plotted on a scatterplot. Which is on the x_axis and which is on the y-axis.

According to wikipedia longitude is "a geographic coordinate that specifies the east–west position of a point on the surface of the Earth, or another celestial body" and latitude is "a coordinate that specifies the north–south position of a point on the surface of the Earth or another celestial body".

With that in mind, when plotting lat' long' data on a scatterplot, longitude goes on the x-axis and latitude goes on the y-axis.

Say we were to set a line going through x=0 as our prime meridian, anything to the right of that line would be to the east and to the left of it would be west. Say we chose y=0 as our equator, anything above it would be to the north and below it would be to the south.