Assignment 03

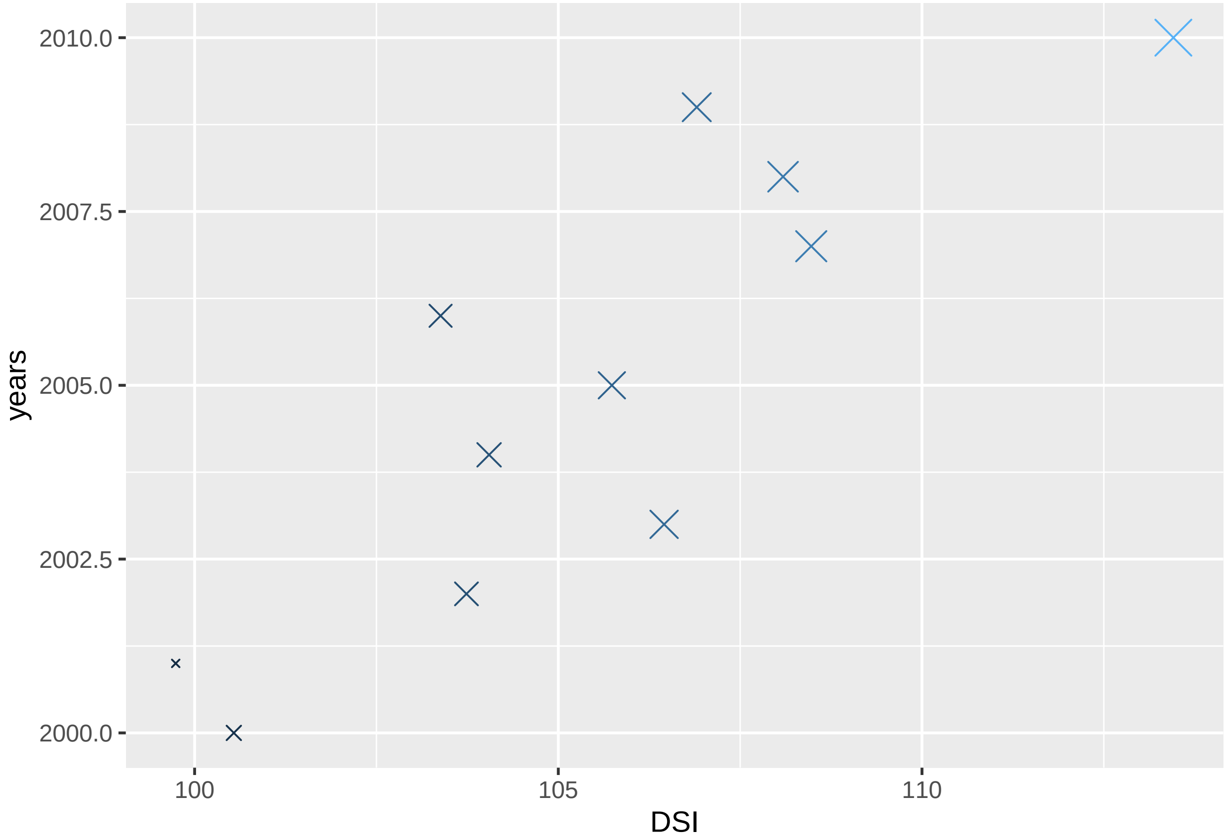
Amol Gote

2/3/2020

```
library(tidyverse)
library(scales)
```

Question 1: Using dataset 1 to replicate plot1.pdf. Here you will see that "years" is on the y-axis and "DSI" # is on the x-axis.

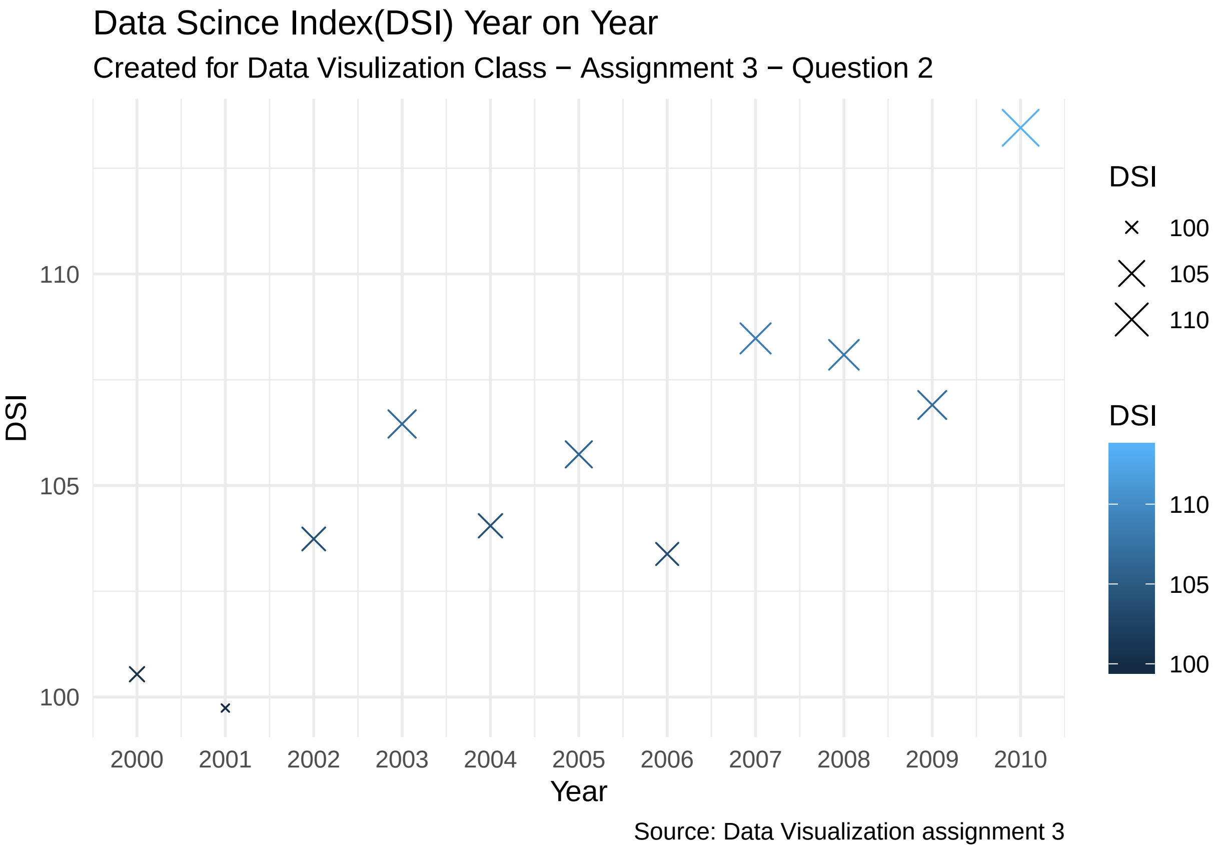
```
dataset1 <- read_csv("data/datset1.csv")
ggplot(data = dataset1, aes(x= DSI, y = years)) +
  geom_point(shape=4, aes(size = DSI, color = DSI), show.legend=FALSE)
```



Question 2: Improve the plot as you see fit (if any). Note any changes you made.

Changes Made
1. Swapped X-axis and y-axis as it is easy to visualize time on x-axis and then plot the value of DSI for each year on y-axis.
2. Added minimal theme to make it look simple and cleaner.
3. Added legend.
4. Since x-axis time range is small, plotted all year labels.
5. Added chart title, subtitle, caption

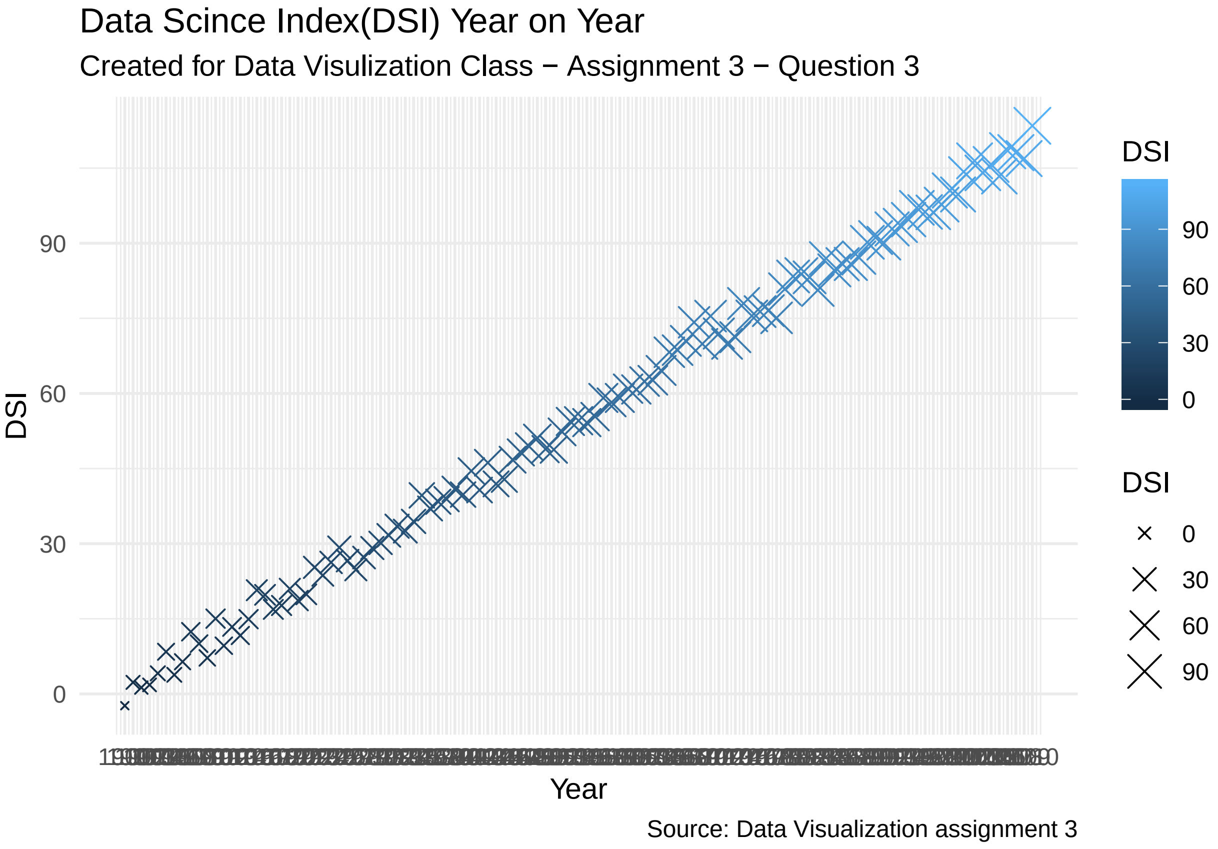
```r
dataset1 <- read_csv("data/datset1.csv")
ggplot(data = dataset1, aes(x= years, y = DSI)) +
  geom_point(shape=4, aes(size = DSI, color = DSI)) +
  theme_minimal() +
  scale_x_continuous(breaks = dataset1$years) +
  labs(x = "Year",
       y = "DSI",
       title = "Data Scince Index(DSI) Year on Year",
       subtitle = "Created for Data Visulization Class - Assignment 3 - Question 2",
       caption = "Source: Data Visualization assignment 3")
```



Question 3: Use the same plotting code from Question 2 and apply it to Dataset 2 - Describe any issues that arise from using the expanded dataset 2

Issues arising from using the expanded dataset 2
1. X-axis labels are overlapping.
2. This dataset is year on year which is timeseries, so line graph would be better fit and it would make it simple and clear.
3. Points are overlapping, so it is difficult to guage mapping between the year and DSI value.
4. Intended message of the plot is to show the year on year trend for DSI which does not seem to be conveyed with point chart.
5. There is no need of scaling the DSI point aes(size = DSI, color = DSI).

```
dataset2 <- read_csv("data/datset2.csv")
ggplot(data = dataset2, aes(x= years, y = DSI)) +
  geom_point(shape=4, aes(size = DSI, color = DSI)) +
  theme_minimal() +
  scale_x_continuous(breaks = dataset2$years) +
  labs(x = "Year",
       y = "DSI",
       title = "Data Scince Index(DSI) Year on Year",
       subtitle = "Created for Data Visulization Class - Assignment 3 - Question 3",
       caption = "Source: Data Visualization assignment 3")
```



Data Scince Index(DSI) Year on Year

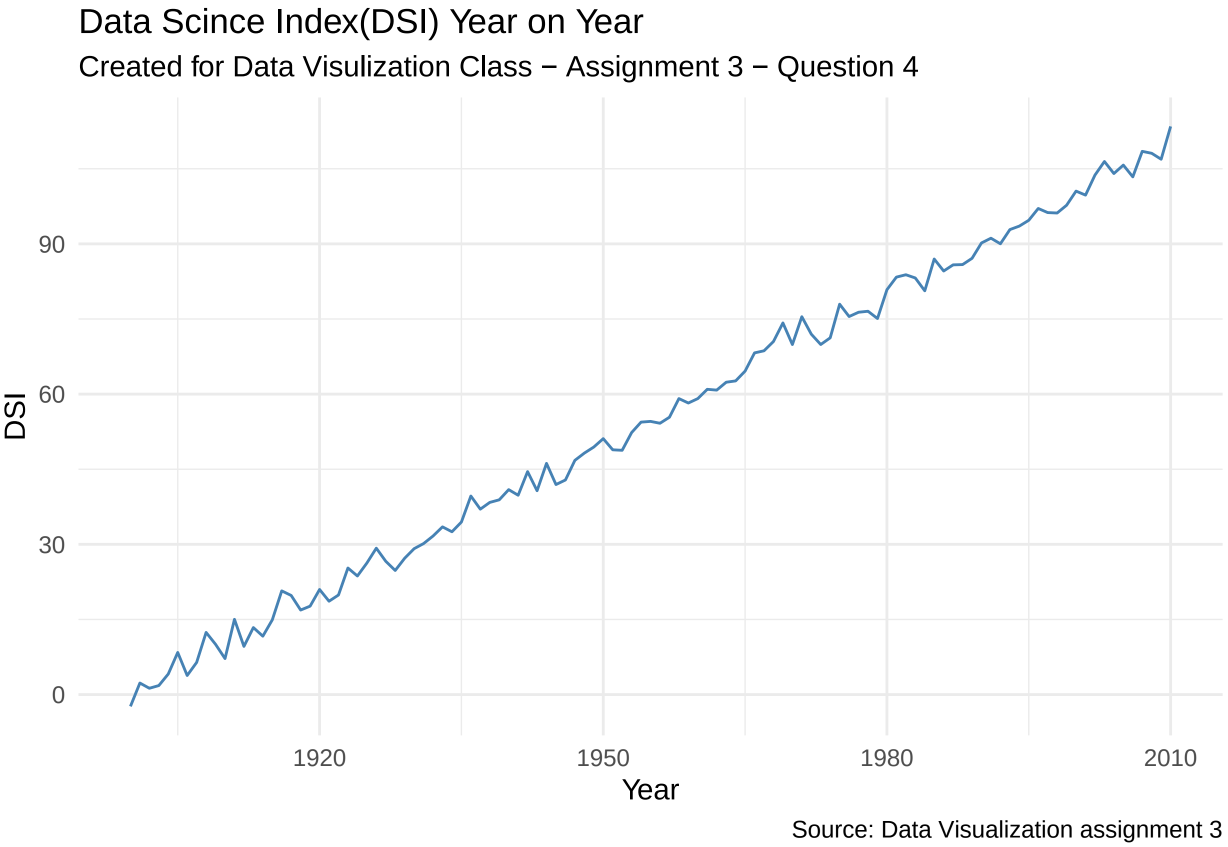
Created for Data Visulization Class − Assignment 3 − Question 3

Source: Data Visualization assignment 3

Question 4: Make any necessary improvements to your plotting code and create an improved graph (if necessary).

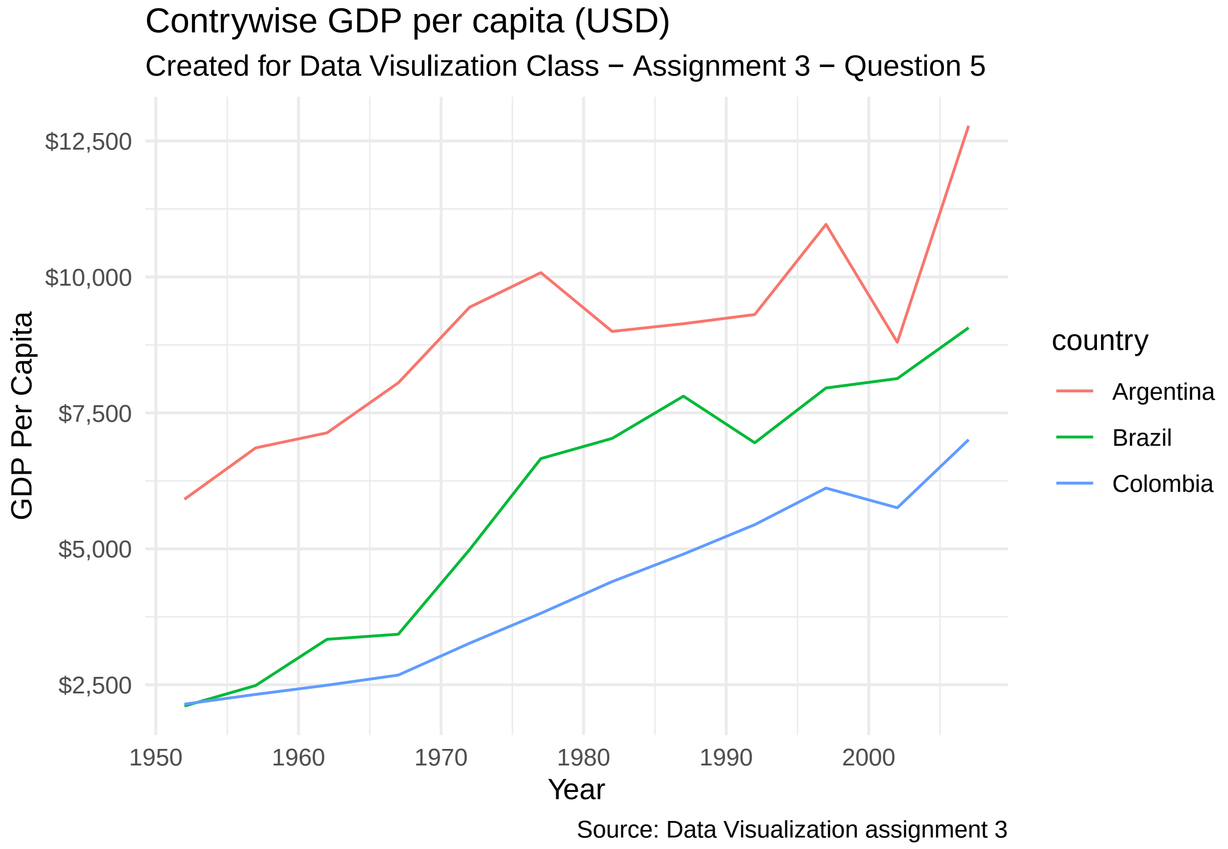
Convereted point chart to line chart this way it is easy to analyze year on year trend of the DSI value.

```
dataset2 <- read_csv("data/datset2.csv")
ggplot(data = dataset2, aes(x= years, y = DSI)) +
  geom_line(color='steelblue') +
  theme_minimal() +
  labs(x = "Year",
       y = "DSI",
       title = "Data Scince Index(DSI) Year on Year",
       subtitle = "Created for Data Visulization Class - Assignment 3 - Question 4",
       caption = "Source: Data Visualization assignment 3")
```



Question 5: Dataset 3 - Use ggplot2 to create a graphic using all the variables. Note you may want to consider re-shaping the data before plotting. The values represent GDP per Capita for each country.

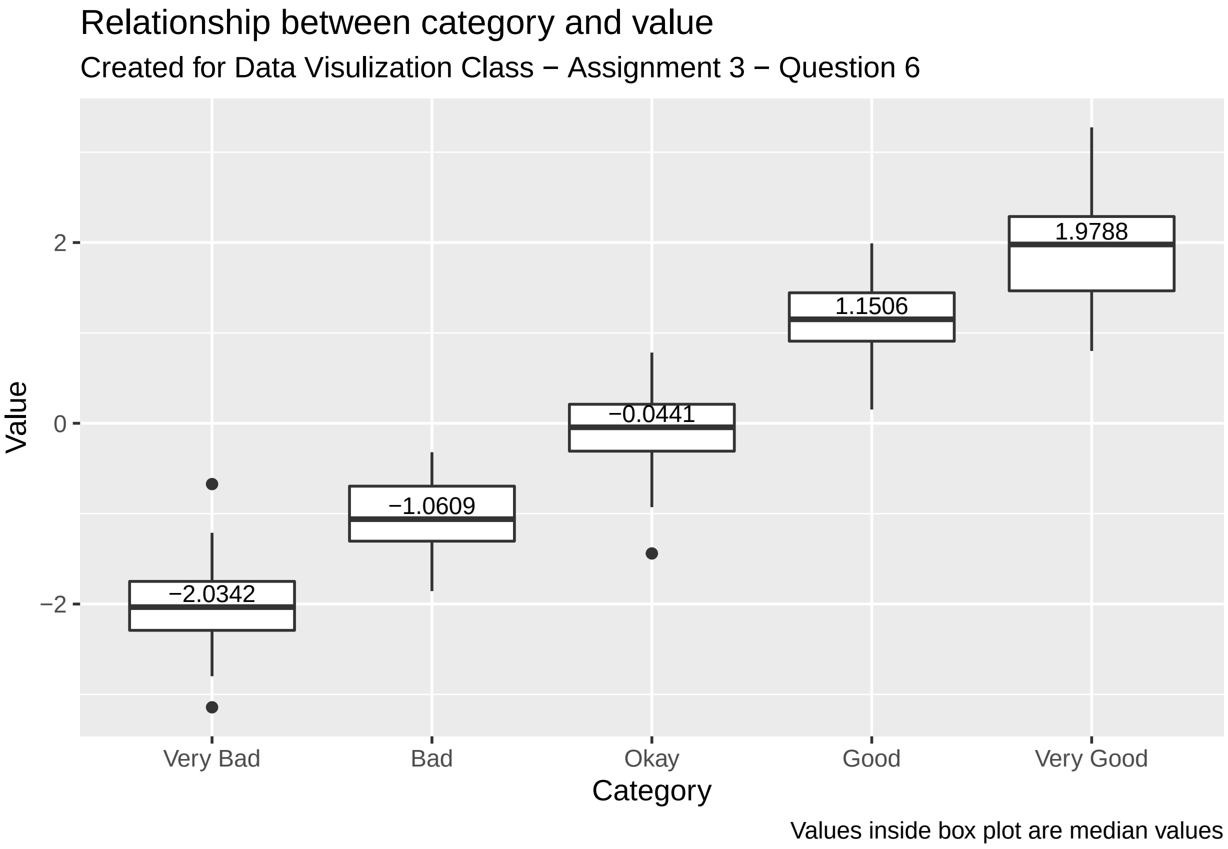
```
dataset3 <- read_csv("data/datset3.csv")
dsGdpCountrywise <- gather(dataset3, "country", "GDP", 2:4)
ggplot(data = dsGdpCountrywise, aes(x= year, y = GDP, color = country)) +
  geom_line() +
  scale_y_continuous(labels = dollar) +
  theme_minimal() +
  labs(x = "Year",
       y = "GDP Per Capita",
       title = "Contrywise GDP per capita (USD)",
       subtitle = "Created for Data Visulization Class - Assignment 3 - Question 5",
       caption = "Source: Data Visualization assignment 3")
```



Question 6: Dataset 4 - Use ggplot2 to create a graphic to explain the relationship between the two variables. You may use one of the methods described in class, in the ggplot2 book, or another method of your choosing. Explain why you chose the method you did and the pros and cons are of the method that you chose.

```
dataset4 <- read_csv("data/datset4.csv")
dataset4 <- dataset4 %>%
            mutate(category = fct_relevel(category, "Very Bad", "Bad", "Okay", "Good", "Very Good"))
dataset4_median <- summarise(group_by(dataset4, category), categoryMedian = median(value))
dataset4_median$categoryMedian <-  round(dataset4_median$categoryMedian, 4)
ggplot(dataset4, aes(x = category, y = value)) +
  geom_boxplot() +
```

```r
geom_text(data = dataset4_median, aes(category, categoryMedian, label = categoryMedian),
            position = position_dodge(width = 0.8), size = 3, vjust = -0.3) +
labs(x = "Category",
     y = "Value",
     title = "Relationship between category and value",
     subtitle = "Created for Data Visulization Class - Assignment 3 - Question 6",
     caption = "Values inside box plot are median values")
```



Method that has been chosen over here is geom_boxplot, alternative method could have been geom_jitter() or geom_violin.

Pros and Cons for geom_boxplot()

Pros

1. Simple scatter plot leads to over plotting around the categorical variable, box_plot helps in simplifying over plotting.
2. It nicely summarizes the shape of the distribution with a handful of summary statistics. geom_jitter creates unnecessary noise and while geom_violin is hard to interpret.
3. Box plot statistics includes min, max, quartile 1, quartile 2 and median, these statistics numbers are sufficient to gauge what the data says about that category.
4. Box plot highlights outliers, which can be discarded for analysis.
5. Box plot remains unaffected by outliers.
6. Box plot can handle and effectively represent summary of a large amount of data.

Cons

1. It does not show individual values. Jittered plots show every point but only work with relatively small datasets.
2. Hides the multimodality (Multiple peaks).
3. Box plot are bit technical and specific to statistics and would be difficult to be interpreted by non-technical or non-statistics background person.

3. Exact values not retained.
4. Not as visually appealing as other graphs.