Lab 01 - Plastic waste

RONIT SINGH 10/4/2020

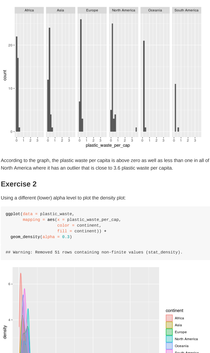
Load packages and data

```
library(tidyverse)
library(viridis)

plastic_waste <- read_csv("data/plastic-waste.csv")
```

Exercise 1

Distribution of plastic waste per capita facetted by continent using histogram:

```
ggplot(data = plastic_waste, aes(x = plastic_waste_per_cap)) +
  geom_histogram(binwidth = 0.2) + facet_grid(~continent)

## Warning: Removed 51 rows containing non-finite values (stat_bin).
```

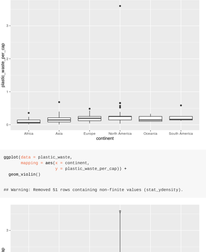


According to the graph, the plastic waste per capita is above zero as well as less than one in all of the continents, except North America where it has an outlier that is close to 3.6 plastic waste per capita.

Exercise 2

Using a different (lower) alpha level to plot the density plot:

```
ggplot(data = plastic_waste,
       mapping = aes(x = plastic_waste_per_cap,
                     color = continent,
                     fill = continent)) +
  geom_density(alpha = 0.3)

## Warning: Removed 51 rows containing non-finite values (stat_density).
```



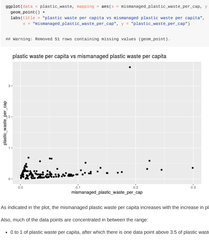
Exercise 3

We defined the color and fill of the curves by mapping aesthetics of the plot as it's controlled by a variable i.e. continent, but we defined the alpha level as a characteristic of the plotting geom as it's controlled by us and is not mapped to a variable.

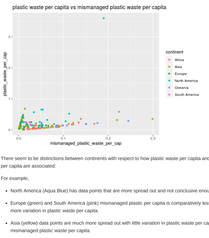
Exercise 4

Visualizing the above relationship using side-by-side box plots and violin plots:

```
ggplot(data = plastic_waste,
       mapping = aes(x = continent,
                     y = plastic_waste_per_cap)) +
  geom_boxplot()

## Warning: Removed 51 rows containing non-finite values (stat_boxplot).
```




```
ggplot(data = plastic_waste,
       mapping = aes(x = continent,
                     y = plastic_waste_per_cap)) +
  geom_violin()

## Warning: Removed 51 rows containing non-finite values (stat_ydensity).
```



Violin plots are visually attractive and more understandable.

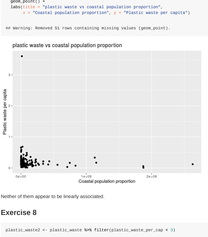
The box plots show data points outside the inter-quartile range as outliers whereas violin plots show the whole range of the data.

A violin plot is more informative than side-by-side box plots. While a box plot only shows summary statistics such as mean/median and inter-quartile ranges, the violin plot shows the full distribution of the data along with medians, ranges and variations / variabilities effectively.

Exercise 5

Relationship between plastic waste per capita and mismanaged plastic waste per capita using a scatterplot:

```
ggplot(data = plastic_waste, mapping = aes(x = mismanaged_plastic_waste_per_cap, y = plastic_waste_per_cap)) +
  geom_point() +
  labs(title = "plastic waste per capita vs mismanaged plastic waste per capita",
       x = "mismanaged_plastic_waste_per_cap", y = "plastic_waste_per_cap")

## Warning: Removed 51 rows containing missing values (geom_point).
```



As indicated in the plot, the mismanaged plastic waste per capita increases with the increase in plastic waste per capita.

Also, much of the data points are concentrated in between the range:

- 0 to 1 of plastic waste per capita, after which there is one data point above 3.5 of plastic waste per capita.
- 0 to 0.1 of mismanaged plastic waste per capita, after which there are few data points and outliers.

Exercise 6

Colored (by continent) scatterplot:

```
ggplot(data = plastic_waste, mapping = aes(x = mismanaged_plastic_waste_per_cap, y = plastic_waste_per_cap,
  geom_point() +
  labs(title = "plastic waste per capita vs mismanaged plastic waste per capita",
       x = "mismanaged_plastic_waste_per_cap", y = "plastic_waste_per_cap")

## Warning: Removed 51 rows containing missing values (geom_point).
```


There seem to be distinctions between continents with respect to how plastic waste per capita and mismanaged plastic waste per capita are associated:

For example,

- North America (Aqua Blue) has data points that are more spread out and not conclusive enough.
- Europe (green) and South America (pink) mismanaged plastic per capita is comparatively less than other continents with more variation in plastic waste per capita.
- Asia (yellow) data points are much more spread out with little variation in plastic waste per capita and more variations in mismanaged plastic waste per capita.
- Africa (orange) data points indicate more mismanaged plastic waste per capita.

Exercise 7

Relationship between plastic waste per capita and total population as well as plastic waste per capita and coastal population.

```
ggplot(data = plastic_waste, mapping = aes(x = total_pop, y = plastic_waste_per_cap)) +
  geom_point() +
  labs(title = "plastic waste vs total population",
       x = "Total population", y = "Plastic waste per capita")

## Warning: Removed 61 rows containing missing values (geom_point).
```



```
ggplot(data = plastic_waste, mapping = aes(x = coastal_pop, y = plastic_waste_per_cap)) +
  geom_point() +
  labs(title = "plastic waste vs coastal population proportion",
       x = "Coastal population proportion", y = "Plastic waste per capita")

## Warning: Removed 51 rows containing missing values (geom_point).
```


Neither of them appear to be linearly associated.

Exercise 8

```
plastic_waste2 <- plastic_waste %>% filter(plastic_waste_per_cap < 3)

ggplot(data = plastic_waste2, mapping = aes(x = coastal_pop/total_pop, y = plastic_waste_per_cap)) +
  geom_point(mapping=aes(color=continent)) + geom_smooth(color="black") + scale_color_viridis(discrete=TRU
  labs(title = "plastic waste vs coastal population proportion (by continent)",
       x = "Coastal population proportion (coastal / total population)", y = "Plastic waste per capita")

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## Warning: Removed 10 rows containing non-finite values (stat_smooth).
## Warning: Removed 10 rows containing missing values (geom_point).
```


In the first half of the plot, the regression line is curved with a positive relation between plastic waste per capita and coastal population proportion.

However, there is a slight negative relation in the second half of the plot indicating that with increase in x-axis, there is slight decrease in y-axis.

The gray area i.e. band width represents the confidence interval or zone and the data point with the highest probability of correct value lies in that zone.