Lab 02 - Global plastic waste

Max Asmar

Introduction

Plastic pollution is a major and growing problem, negatively affecting oceans and wildlife health. Our World in Data has a lot of great data at various levels including globally, per country, and over time. For this lab we focus on data from 2010.

Additionally, National Geographic ran a data visualization communication contest on plastic waste as seen here.

This lab builds on skills from Lab 01. You'll get more practice with ggplot2, learn about different types of visualizations, and explore a real-world dataset.

Estimated time: 90-120 minutes

Learning Goals

By the end of this lab, you will be able to:

- Create and interpret histograms and density plots
- Use faceting to compare distributions across groups
- Adjust transparency (alpha) to improve visualizations
- Create and interpret box plots and violin plots
- Make scatterplots to visualize relationships between variables
- Modify aesthetics (color, fill) to enhance plots
- Build complex multi-layered visualizations

Prerequisites

Before you begin, make sure you have:

- Completed Lab 01
- Watched the data visualization lecture
- Forked the lab-instructions repository
- Feel comfortable with the knit/commit/push workflow

Getting Started

Navigate to your forked `lab-instructions` repository in JupyterHub, open the `Lab02` folder, and open the R Markdown document `lab-02.Rmd`.

Verify Your Setup

Before proceeding:

1. Open `lab-02.Rmd` in RStudio
2. Check that the Git pane shows YOUR username (not the course organization)
3. Click **Knit** to make sure the document compiles

If you see errors or can't find the file, ask for help before continuing!

Packages

We'll use the **tidyverse** package for this analysis. Run the following code in the Console to load this package.

```
library(tidyverse)
```

Data

The dataset for this assignment can be found as a csv file in the `data` folder of your repository. You can read it in using the following.

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

The variable descriptions are as follows:

- `code`: 3 Letter country code
- `entity`: Country name
- `continent`: Continent name
- `year`: Year
- `gdp_per_cap`: GDP per capita constant 2011 international $, rate
- `plastic_waste_per_cap`: Amount of plastic waste per capita in kg/day
- `mismanaged_plastic_waste_per_cap`: Amount of mismanaged plastic waste per capita in kg/day
- `mismanaged_plastic_waste`: Tonnes of mismanaged plastic waste
- `coastal_pop`: Number of individuals living on/near coast
- `total_pop`: Total population according to Gapminder

Warm Up

Before diving into the exercises, let's make sure you're familiar with the RStudio environment and the data.

Question 1: Without looking, can you name the four panes in RStudio and briefly describe their purpose?

Your answer: There is the environment/history pane, help/packages/files/plots pane, console pane, and source pane. The environment/history pane allows you to view changes and loaded objects. The git pane lets you commit to github. The help/files/packages/plots pane allows you to

analyze the libraries that you are using, download packages, navigate directories. The console pane and terminal lets you type in code that gets directly computed. The source pane allows you to edit the YAML and other files in R Studio.

Question 2: Verify that the dataset has loaded into the Environment pane. How many observations are in the dataset?

Hint: Look in the Environment pane, or run `nrow(plastic_waste)` in the Console, or click on the dataset name to view it.

Your answer: The dataset has 240 observations. I ran `nrow(plastic_waste)` in the Console.

Question 3: Have a quick look at the data using `View(plastic_waste)` in the Console. You'll notice that some cells contain `NA`. What does `NA` mean in R?

Hint: Run `?NA` in the Console to see the documentation.

Your answer: In R, NA means that there is a missing value in the dataset.

Checkpoint: Make sure the data loaded correctly and you can see it in your Environment pane before continuing.

Knit, commit with message "Completed warm up", and push your changes.

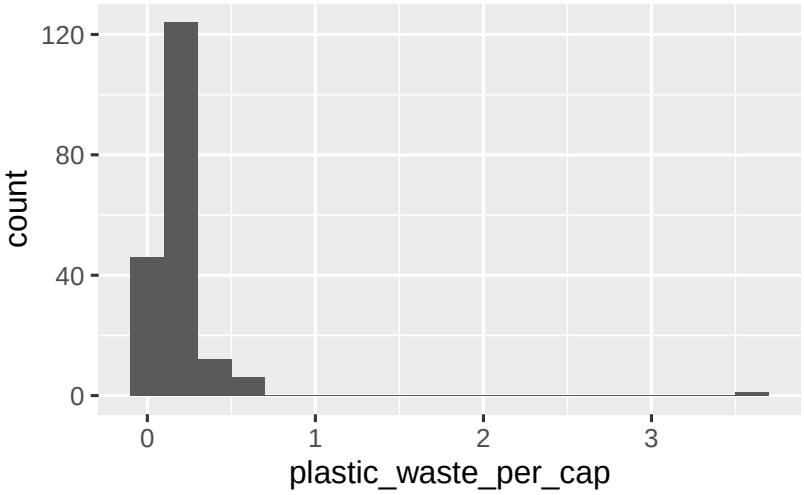
Exercises

Exploring Distributions

Let's start by taking a look at the distribution of plastic waste per capita in 2010.

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

```
## Warning: Removed 51 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



One country stands out as an unusual observation at the top of the distribution. One way of identifying this country is to filter the data for countries where plastic waste per capita is greater than 3.5 kg/person.

```
plastic_waste %>%
  filter(plastic_waste_per_cap > 3.5)
```

```
## # A tibble: 1 x 10
##   code  entity              continent     year gdp_per_cap plastic_waste_per_cap
##   <chr> <chr>               <chr>        <dbl>       <dbl>                 <dbl>
## 1 TTO   Trinidad and Tobago North Ameri~  2010      31261.                   3.6
## # i 4 more variables: mismanaged_plastic_waste_per_cap <dbl>,
## #   mismanaged_plastic_waste <dbl>, coastal_pop <dbl>, total_pop <dbl>
```

Did you expect this result? Trinidad and Tobago has unusually high plastic waste per capita. You might research why this is the case!

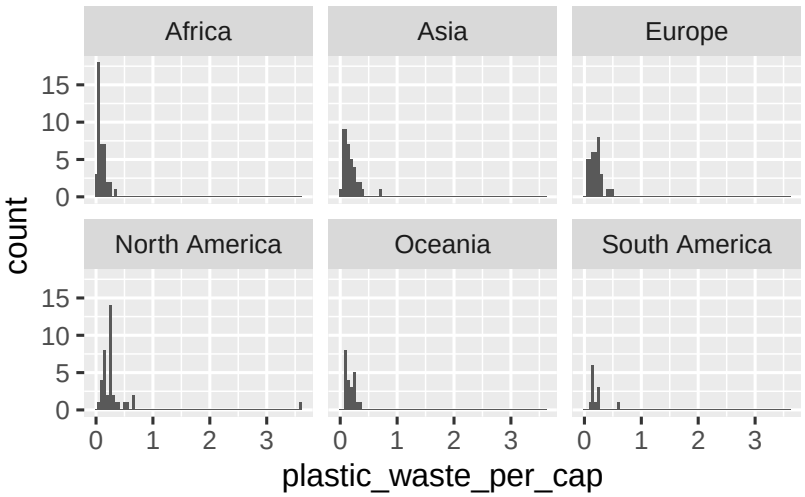
Exercise 1

Plot, using histograms, the distribution of plastic waste per capita faceted by continent. What can you say about how the continents compare to each other in terms of their plastic waste per capita?

Reminder: Faceting creates separate plots for each group. Use `facet_wrap(~variable)` to facet by a categorical variable.

```
# Remember to change eval=FALSE to eval=TRUE after filling in the blanks!
ggplot(data = plastic_waste,
       aes(x = plastic_waste_per_cap)) +
  geom_histogram(binwidth = 0.05) +
  facet_wrap(~continent)
```

```
## Warning: Removed 51 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



Interpretation: Compare the continents in terms of their plastic waste per capita. Consider:

- Which continent has the highest typical plastic waste per capita?
- Which has the lowest?
- Which continent shows the most variation?

Your answer:

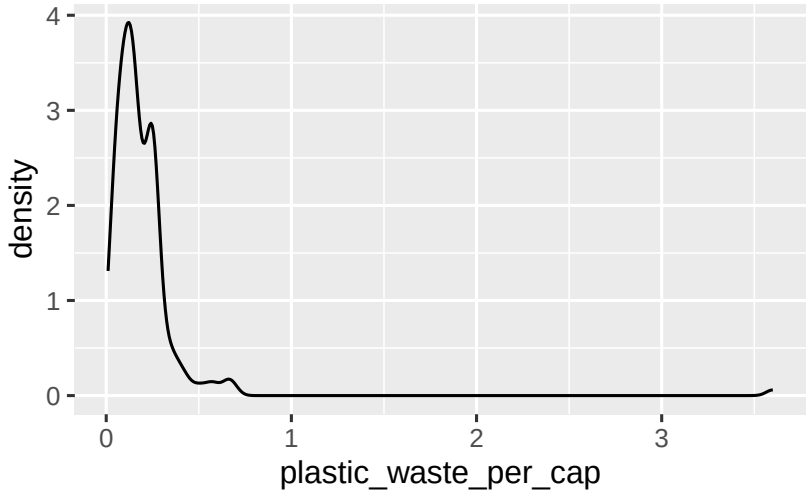
North America has the highest typical plastic waste per capita. Africa has the lowest typical plastic waste per capita. I think North America shows the most variation because its range of plastic waster per capita is the largest.

Density Plots

Another way of visualizing numerical data is using density plots. Density plots show the distribution as a smooth curve rather than bars.

```
ggplot(data = plastic_waste, aes(x = plastic_waste_per_cap)) +
  geom_density()
```

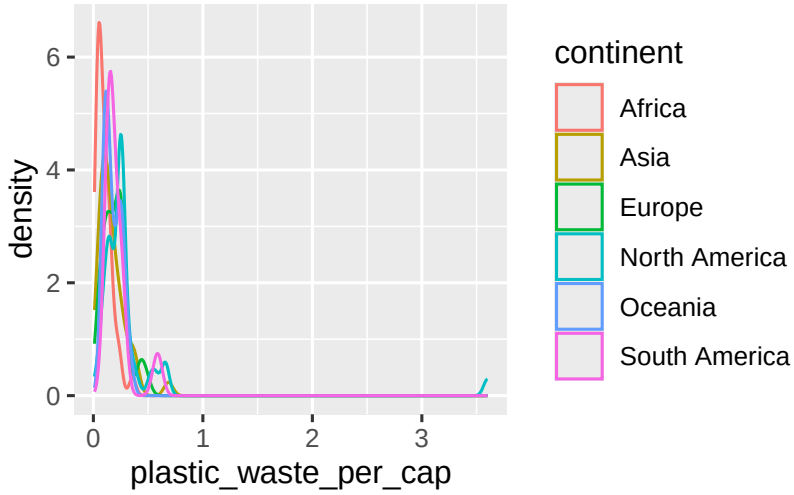
```
## Warning: Removed 51 rows containing non-finite outside the scale range
## (`stat_density()`).
```



We can compare distributions across continents by coloring density curves by continent.

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

```
## Warning: Removed 51 rows containing non-finite outside the scale range
## (`stat_density()`).
```

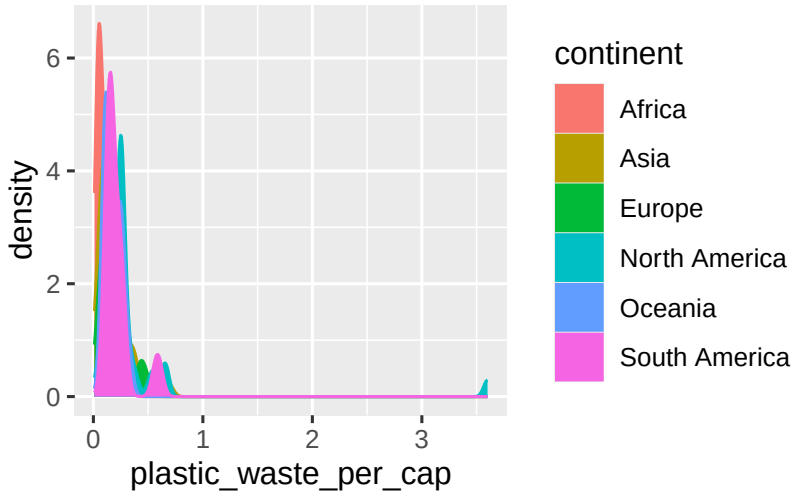


The resulting plot may be a little difficult to read, so let's also fill the curves in with colors as well.

```
ggplot(data = plastic_waste,
       mapping = aes(x = plastic_waste_per_cap,
                     color = continent,
```

```
                      fill = continent)) +
  geom_density()
```

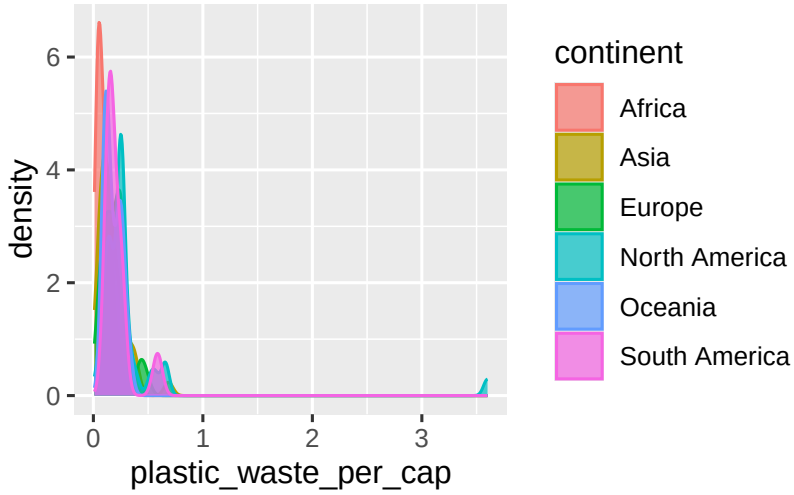
```
## Warning: Removed 51 rows containing non-finite outside the scale range
## (`stat_density()`).
```



The overlapping colors make it difficult to tell what's happening with the distributions in continents plotted first. We can change the transparency level of the fill color to help with this. The `alpha` argument takes values between 0 and 1: 0 is completely transparent and 1 is completely opaque.

```
ggplot(data = plastic_waste,
       mapping = aes(x = plastic_waste_per_cap,
                     color = continent,
                     fill = continent)) +
  geom_density(alpha = 0.7)
```

```
## Warning: Removed 51 rows containing non-finite outside the scale range
## (`stat_density()`).
```



This still doesn't look great...

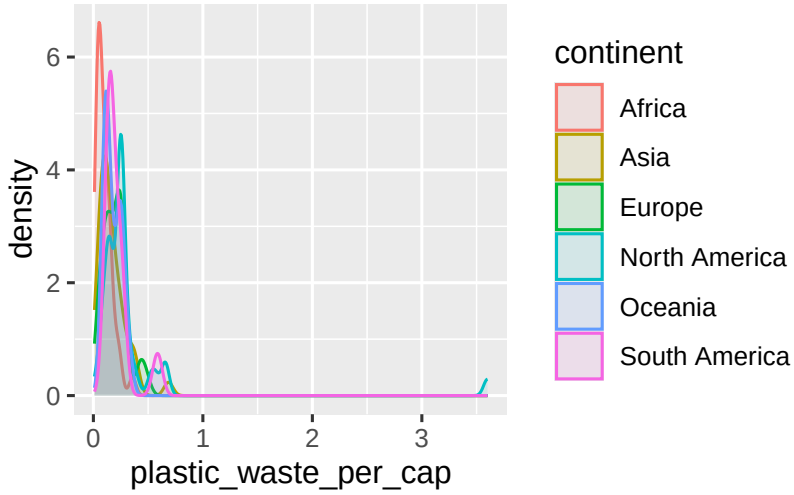
Exercise 2

Recreate the density plots above using a different (lower) alpha level that works better for displaying the density curves for all continents.

Try alpha values like 0.3, 0.4, or 0.5 to see which works best.

```
# Remember to change eval=FALSE to eval=TRUE!
ggplot(data = plastic_waste,
       mapping = aes(x = plastic_waste_per_cap,
                     color = continent,
                     fill = continent)) +
  geom_density(alpha = 0.1)
```

```
## Warning: Removed 51 rows containing non-finite outside the scale range
## (`stat_density()`).
```



Which alpha level did you choose and why?

I chose alpha = 0.1 so that I could trace every curve that is on the graph while still being able to see most of the areas under the curves.

Exercise 3

Describe why we defined the `color` and `fill` of the curves by mapping aesthetics of the plot (inside `aes()`) but we defined the `alpha` level as a characteristic of the plotting geom (inside `geom_density()`).

Hint: Think about what varies by data vs. what is constant across all data points.

Your answer: We defined the color and fill of the curves inside aes() because they vary across continent while the alpha level is constant throughout all continents because so it can be defined outside of the aes() function.

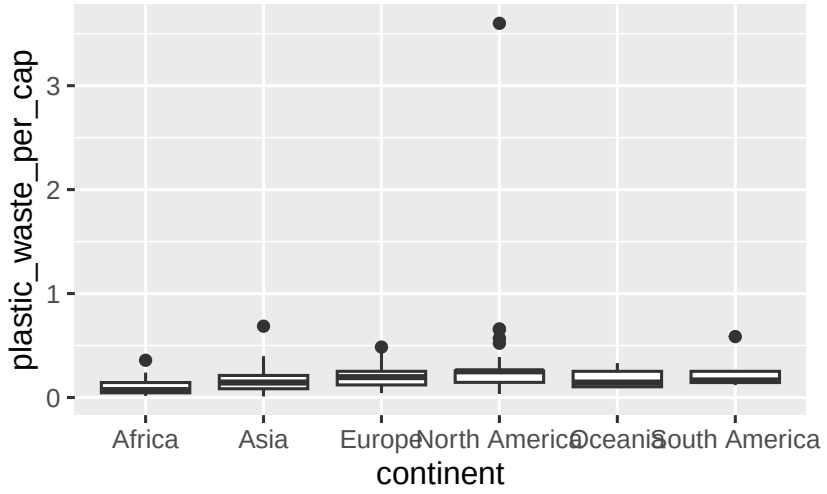
Knit, commit with message "Completed Exercises 1-3", and push your changes.

Box Plots and Violin Plots

Another way to visualize this relationship is using side-by-side box plots.

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

```
## Warning: Removed 51 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



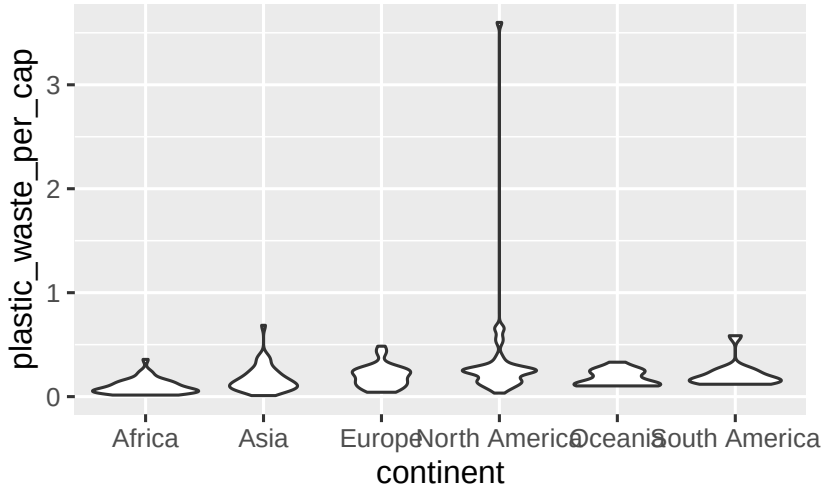
Exercise 4

Convert your side-by-side box plots from the previous task to violin plots. What do the violin plots reveal that box plots do not? What features are apparent in the box plots but not in the violin plots?

Hint: Simply replace `geom_boxplot()` with `geom_violin()`.

```
# Remember to change eval=FALSE to eval=TRUE!
ggplot(data = plastic_waste,
       mapping = aes(x = continent,
                     y = plastic_waste_per_cap)) +
  geom_violin()
```

```
## Warning: Removed 51 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```



What do violin plots reveal that box plots do not?

Violin plots better show the distribution/spread of data in the dataset. It is easier to see visually where the data is concentrated.

What features are apparent in box plots but not in violin plots?

Box plots show the median and interquartile ranges, which is harder to determine in the violin plots.

Scatterplots

Now let's explore relationships between variables using scatterplots.

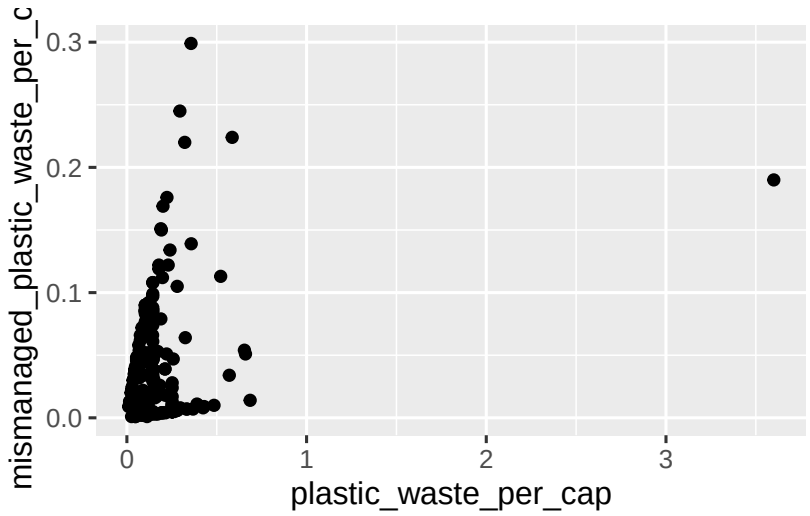
Exercise 5

Visualize the relationship between plastic waste per capita and mismanaged plastic waste per capita using a scatterplot. Describe the relationship.

Reminder: Use `geom_point()` to create scatterplots.

```
# Remember to change eval=FALSE to eval=TRUE!
ggplot(data = plastic_waste,
       mapping = aes(x = plastic_waste_per_cap,
                     y = mismanaged_plastic_waste_per_cap)) +
  geom_point()
```

```
## Warning: Removed 51 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Describe the relationship between plastic waste per capita and mismanaged plastic waste per capita:

Plastic waste per capita and mismanaged plastic waste per capita have a positive correlation. If the plastic waste per capita is greater, then the mismanaged plastic waste per capita is more likely to be greater. The scatter plot shows this relationship with a few outliers. I predict that the line of best fit is a line with positive slope, which supports this.

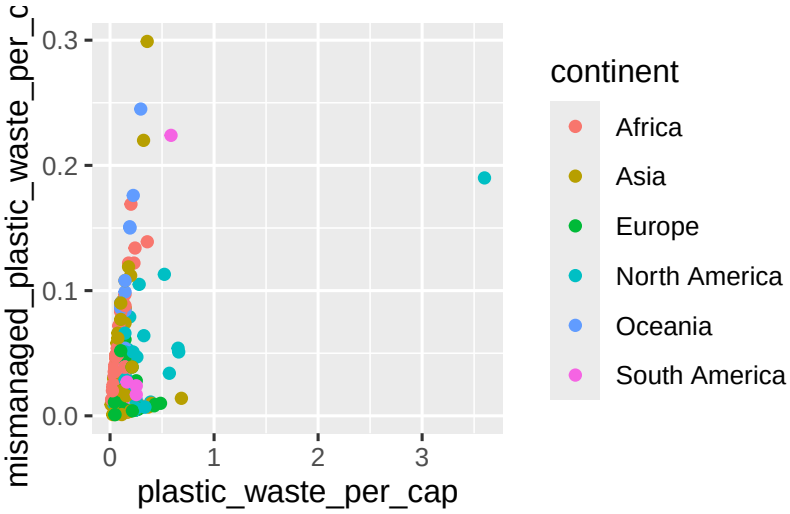
Exercise 6

Color the points in the scatterplot by continent. Does there seem to be any clear distinctions between continents with respect to how plastic waste per capita and mismanaged plastic waste per capita are associated?

Hint: Add `color = continent` to your `aes()` mapping.

```
# Remember to change eval=FALSE to eval=TRUE!
ggplot(data = plastic_waste,
       mapping = aes(x = plastic_waste_per_cap,
                     y = mismanaged_plastic_waste_per_cap,
                     color = continent)) +
  geom_point()
```

```
## Warning: Removed 51 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Are there clear distinctions between continents?

Yes, there are clear distinctions between continents. Europe tends to have the lowest mismanaged plastic waste per capita with respect to plastic waste per capita. Asia has the largest range of mismanaged plastic waste per capita. Africa qualitatively has the largest slope for plastic waste per capita to mismanaged plastic waste per capita, which highlights a problem with waste management of the waste produced.

Knit, commit with message "Completed Exercises 4-6", and push your changes.

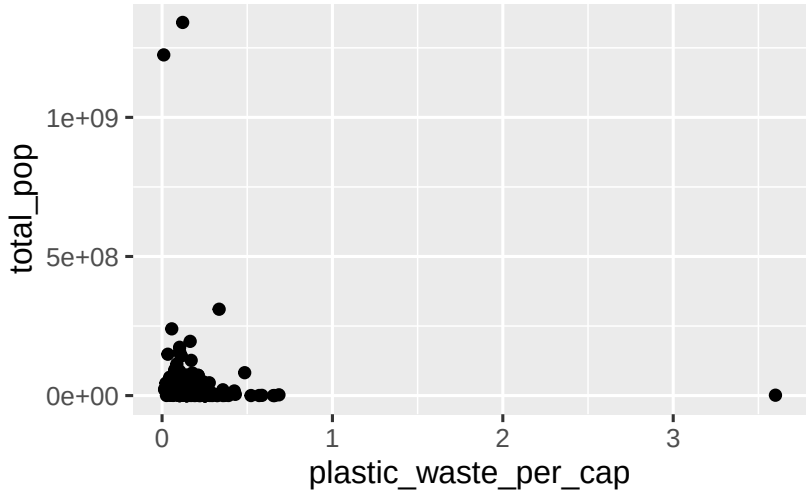
Exercise 7

Visualize the relationship between plastic waste per capita and total population as well as plastic waste per capita and coastal population. You will need to make two separate plots. Do either of these pairs of variables appear to be more strongly linearly associated?

Plot 1: Plastic waste per capita vs. total population

```
# Remember to change eval=FALSE to eval=TRUE!
ggplot(data = plastic_waste,
       mapping = aes(x = plastic_waste_per_cap,
                     y = total_pop)) +
  geom_point()
```

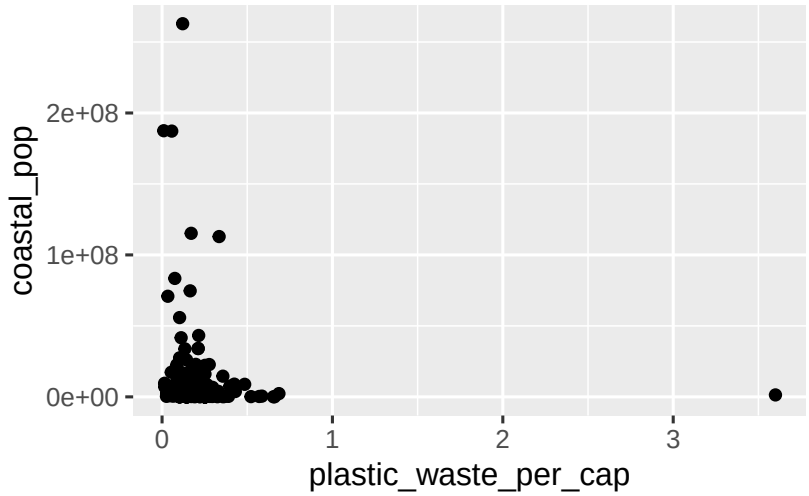
```
## Warning: Removed 61 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Plot 2: Plastic waste per capita vs. coastal population

```
# Remember to change eval=FALSE to eval=TRUE!
ggplot(data = plastic_waste,
       mapping = aes(x = plastic_waste_per_cap,
                     y = coastal_pop)) +
  geom_point()
```

```
## Warning: Removed 51 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Which pair of variables appears to be more strongly linearly associated?

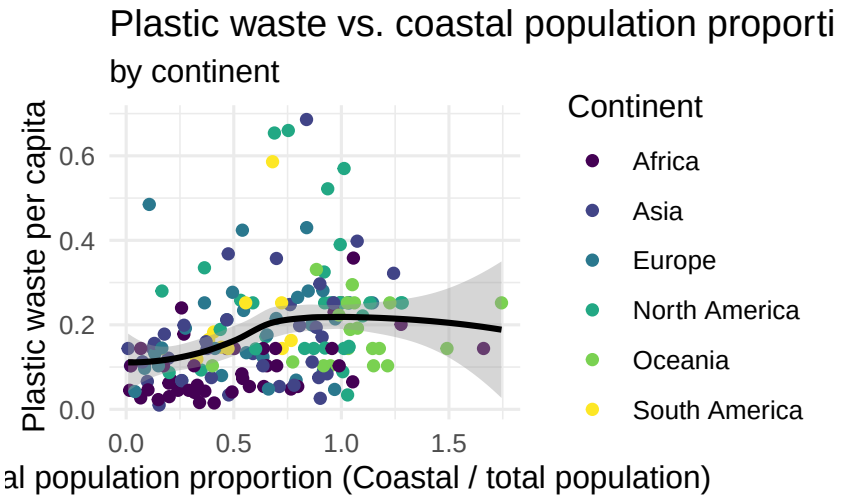
The plastic waste per capita and coastal population appear to be more strongly linear associated compared to plastic waste per capita vs. total population. The imaginary line of best fit appears to be positive for the waste per capita and coastal population comparison.

Knit, commit with message "Completed Exercise 7", and push your changes.

Challenge Exercise

Exercise 8

Recreate the following plot, and interpret what you see in context of the data.



This plot requires several steps:

1. Create a new variable `coastal_pop_prop` = `coastal_pop` / `total_pop`
2. Filter for observations where `plastic_waste_per_cap` < 3
3. Create a scatterplot with the new variable on x-axis
4. Add a smooth curve with `geom_smooth()`
5. Use `scale_color_viridis_d()` for the color scale
6. Add appropriate labels

Build it step by step:

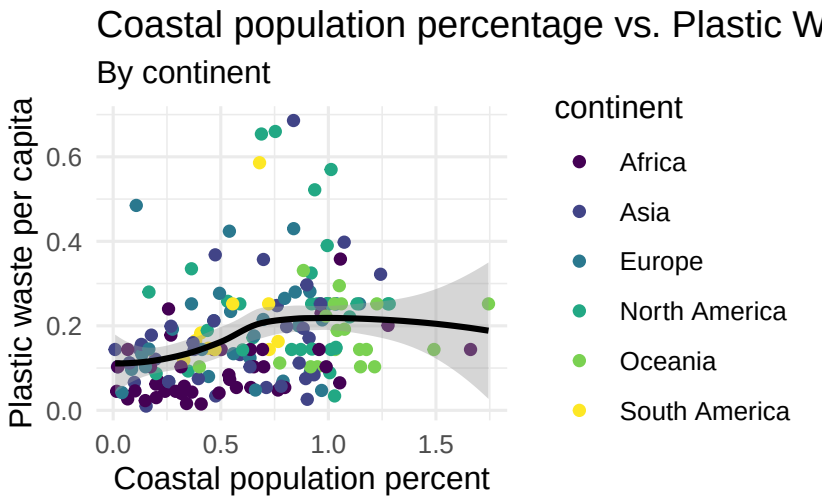
```
# Remember to change eval=FALSE to eval=TRUE!
plastic_waste %>%
  mutate(coastal_pop_prop = coastal_pop / total_pop) %>%
  filter(plastic_waste_per_cap < 3) %>%
  ggplot(aes(x = coastal_pop_prop, y = plastic_waste_per_cap, color = continent)) +
    geom_point() +
    geom_smooth(color = "black") +
    scale_color_viridis_d() +
    labs(x = "Coastal population percent",
         y = "Plastic waste per capita",
```

```
        color = "continent",
        title = "Coastal population percentage vs. Plastic Waste per Capita",
        subtitle = "By continent") +
    theme_minimal()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
## Warning: Removed 10 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 10 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Interpretation: What does this plot reveal about the relationship between coastal population proportion and plastic waste per capita? Does the relationship differ across continents?

The plot reveals that as the coastal population proportion increases, the plastic waste per capita generally increases. The relationship varies a little bit between countries, as the relationship is the least prominent in Africa but the most prominent in Europe. However, they all follow similar patterns in the positive correlation between coastal population proportion and plastic waste per capita. I think that Europe and North America's observations are closer to the trendline while Africa's observations are below the trend line generally.

Knit, commit with message "Completed Exercise 8", and push your changes.

Common Errors and Troubleshooting

Visualization errors:

- **Plot not showing** → Make sure you changed `eval=FALSE` to `eval=TRUE` in the chunk options
- **Error: "object not found"** → Check spelling of variable names (case-sensitive!)
- **Error: "could not find function 'geom_violin'"** → Make sure tidyverse is loaded
- **Weird looking plots** → Try different alpha values, binwidths, or figure sizes

Data errors:

- **Can't load data** → Make sure the `data` folder is in the same directory as your .Rmd file
- **NA values causing issues** → This is normal for this dataset; R handles them appropriately in plots

Git/GitHub errors:

- **Can't push** → Make sure you committed first
- **Changes not showing on GitHub** → Check that you pushed to YOUR fork, not the course repo

To submit to Canvas:

1. In RStudio, click the **Knit** dropdown menu (next to the Knit button)
2. Select **Knit to tufte_handout** to generate a PDF
3. Download the PDF file from the Files pane
4. Upload the PDF to Canvas

Great work! Make sure all your changes are committed and pushed to GitHub.