Data Visualization

Setup

This lab will focus on generating plots using the `ggplot2` package, as well as practicing data subsetting with the `filter()` command from the `dplyr` package. To get started, load the following packages and data set.

```
library(dplyr)
library(ggplot2)
library(oilabs)
data(survey141)
```

This data comes from the class survey that you were asked to complete; however, the data frame we'll be using has been slightly altered to make generating plots easier. The column names of the data set have been changed as well. The help file contains the original questions.

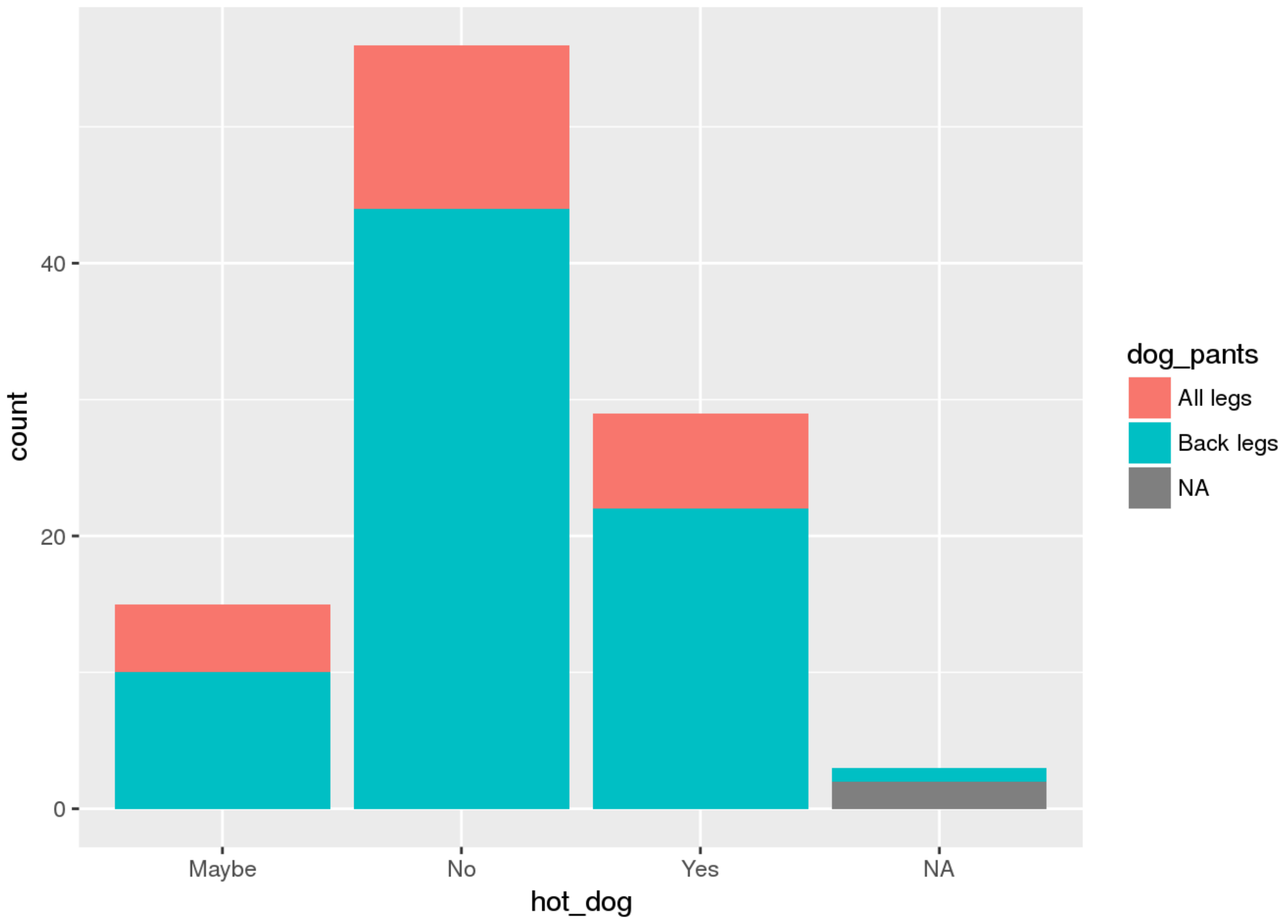
```
?survey141
```

To see a messier version of the data set, `oilabs` also contains `survey141_messy` .

Plot Types

The `ggplot` command is used to create a base layer for a plot, with `aes()` used to specify which visual cues to connect to variables in a data frame. After creating this base layer, a geometry can be applied which will render a specific type of plot.

```
d <- ggplot(survey141, aes(x = hot_dog, fill = dog_pants))
d + geom_bar()
```



Above a base layer for the survey data is created with `hot_dog` mapped to the x-axis and `dog_pants` mapped to fill (a color based visual cue). This base layer is assigned to `d` and then a bar plot layer is added to it via `geom_bar()` .

The following geoms are some of the most commonly used:

```
- `geom_bar()`:          for a categorical variables
- `geom_histogram()`:    for a numerical variable
- `geom_density()`:      for a numerical variable
- `geom_boxplot()`:      for a numerical variable and possibly a categorical varia
ble
- `geom_point()`:        for 2 numerical variables
- `geom_smooth()`:       for 2 numerical variables
```

Multiple geometries may be layered on top of each other, as is often the case with `geom_smooth()` and `geom_point()` to create scatter plots with trend lines.

Exercise 1

Create a density plot for the number of colleges applied to. Please describe this distribution in terms of shape, center, and spread.

Filtering Data

Often there will be entries in data sets that you'll want to exclude from a particular analysis. This is when the `filter()` command will be useful. `filter()` will take a data set and find all rows that return true under set conditions (or filters).

The following example has the `survey141` data being filtered to only include rows which have "Tea" as the response for the `coffee_tea` variable. This subset of the original survey data is then assigned to a new data frame called `tea_drinkers` .

```
tea_drinkers <- filter(survey141, coffee_tea == "Tea")
```

The following uses the piping operator to apply a filter to `survey141` , filtering out only those rows which have 0 in the `marijuana` variable, **and** a value greater than 0 in the `alcohol` variable.

```
alcohol_only <- survey141 %>% filter(alcohol > 0 & marijuana == 0)
```

Pulling up the help file will give a list of useful filter functions and more examples of the function in use.

Exercise 2

Filter out those who answered PC or Mac to the computer of choice question. Create a side-by-side box plot comparing the first kiss age between the groups. Do you notice any differences between the two groups? If so, what are they? (Note: if the box plot does not render correctly, try switching the `x` and `y` assignments.)

Settings, Labels and Limits

Beyond aesthetics and geometries, `ggplot2` also has settings and labels for customizing visualizations.

Settings are options one can apply to a plot, but unlike aesthetics, they are not directly linked to a variable.

```
ggplot(survey141, aes(x = hot_dog, fill = dog_pants)) + geom_bar(position = "fill
")
```

In the above example, the setting `position = "fill"` is used on geom_bar so that count is no longer displayed, but instead proportion.

Exercise 3

Using position = "jitter" create a scatter plot relating alcohol and marijuana.

Exercise 4

Change the axes limits so that both axis have the same range.

All that empty space makes the figure hard to read.

Exercise 5

Give more descriptive axes labels and set the axis limits back to their original values. Describe any relationship or lack thereof. How do these last two plots compare to your personal estimates?

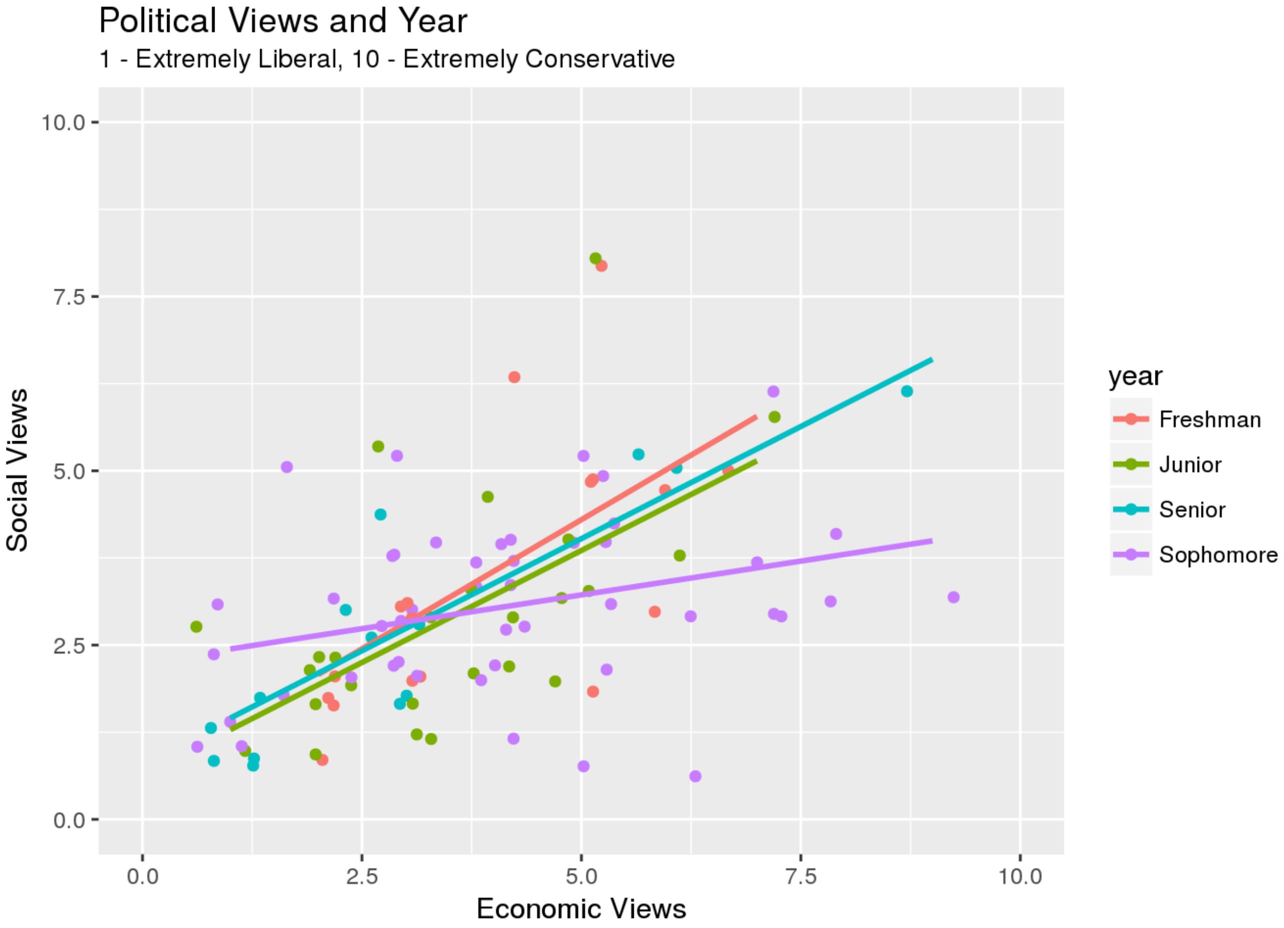
More Practice

Exercise 6

Create 2-3 more visualizations from the survey data that are of interest to you. Have at least one visualization come from a filtered subset of the data, and have at least one more map 3 or more variables. Describe any trends/patterns discovered.

Example 7

Recreate the following plot. The help file for `geom_smooth()` will be useful. (Hint: The settings used for `geom_smooth()` are `method` and `se()`)


Political Views and Year
1 - Extremely Liberal, 10 - Extremely Conservative
Social Views
Economic Views
year
Freshman
Junior
Senior
Sophomore