ggplot2 lab

Lab completed by April Mariko Salazar

Lab assignment authored by Joanna Lankester

Instructions

Go through this R notebook sequentially. Questions are numbered with the letter Q. Some questions just have code you need to run. Questions requiring code or a response will have a *. Please remember to **title** plots and **label** axes (including units, where needed).

Setup

First we'll load libraries and data needed for this lab.

Q1

Run this code first.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
# in billions of dollars
df <- melt(USPersonalExpenditure,varnames=c("category","year"),value.name = "amount")

mpg %>%
    filter(manufacturer=="ford" | manufacturer=="toyota" | manufacturer=="nissan") %>%
    group_by(manufacturer,class) %>%
    tally() %>%
    group_by(manufacturer) %>%
```

```
    mutate(total = sum(n)) %>%
    group_by(manufacturer) %>%
    mutate(percent_of_brand = 100*n/total) %>%
    select(manufacturer,class,n,percent_of_brand) -> df_car
```

Time series plots

Q2

Run this command to understand what the dataset looks like.

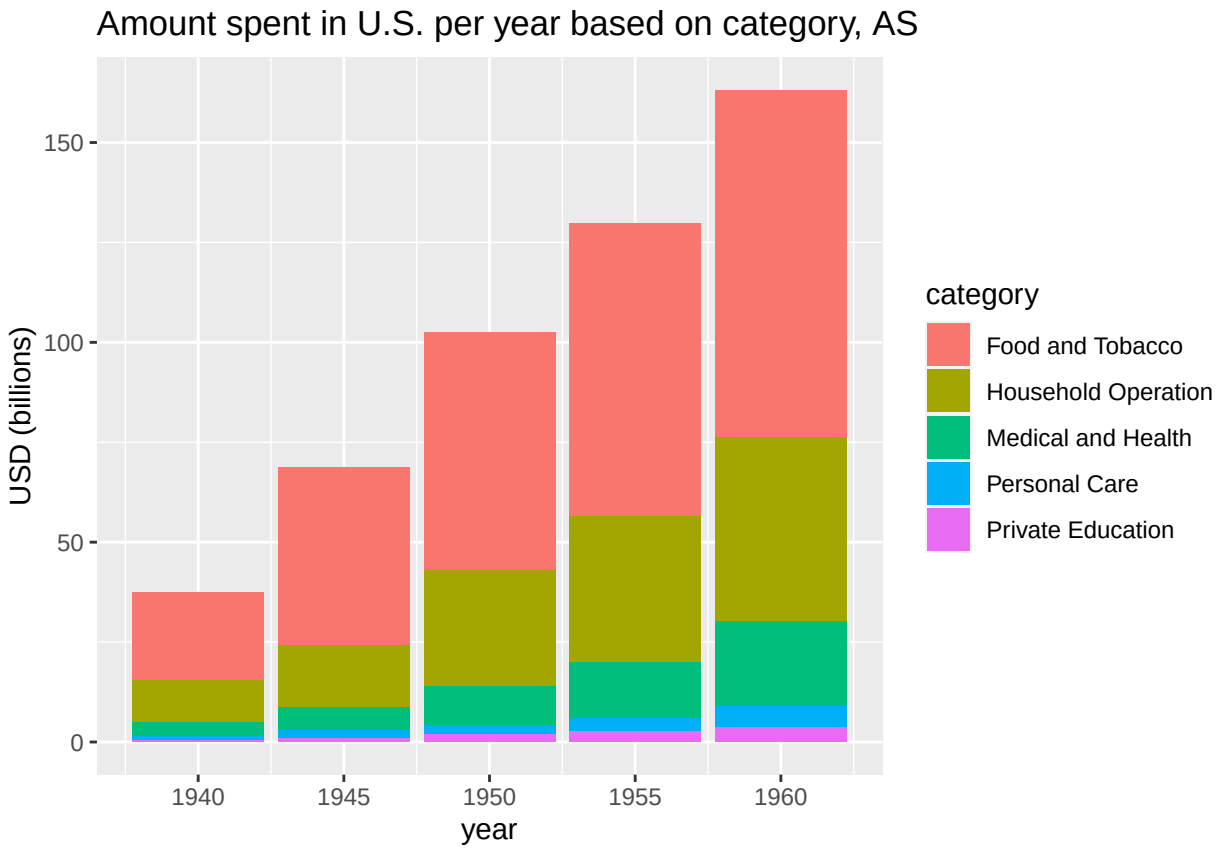
```
df %>% head(10)
```

```
##               category year amount
## 1     Food and Tobacco 1940 22.200
## 2  Household Operation 1940 10.500
## 3   Medical and Health 1940  3.530
## 4         Personal Care 1940  1.040
## 5    Private Education 1940  0.341
## 6     Food and Tobacco 1945 44.500
## 7  Household Operation 1945 15.500
## 8   Medical and Health 1945  5.760
## 9         Personal Care 1945  1.980
## 10   Private Education 1945  0.974
```

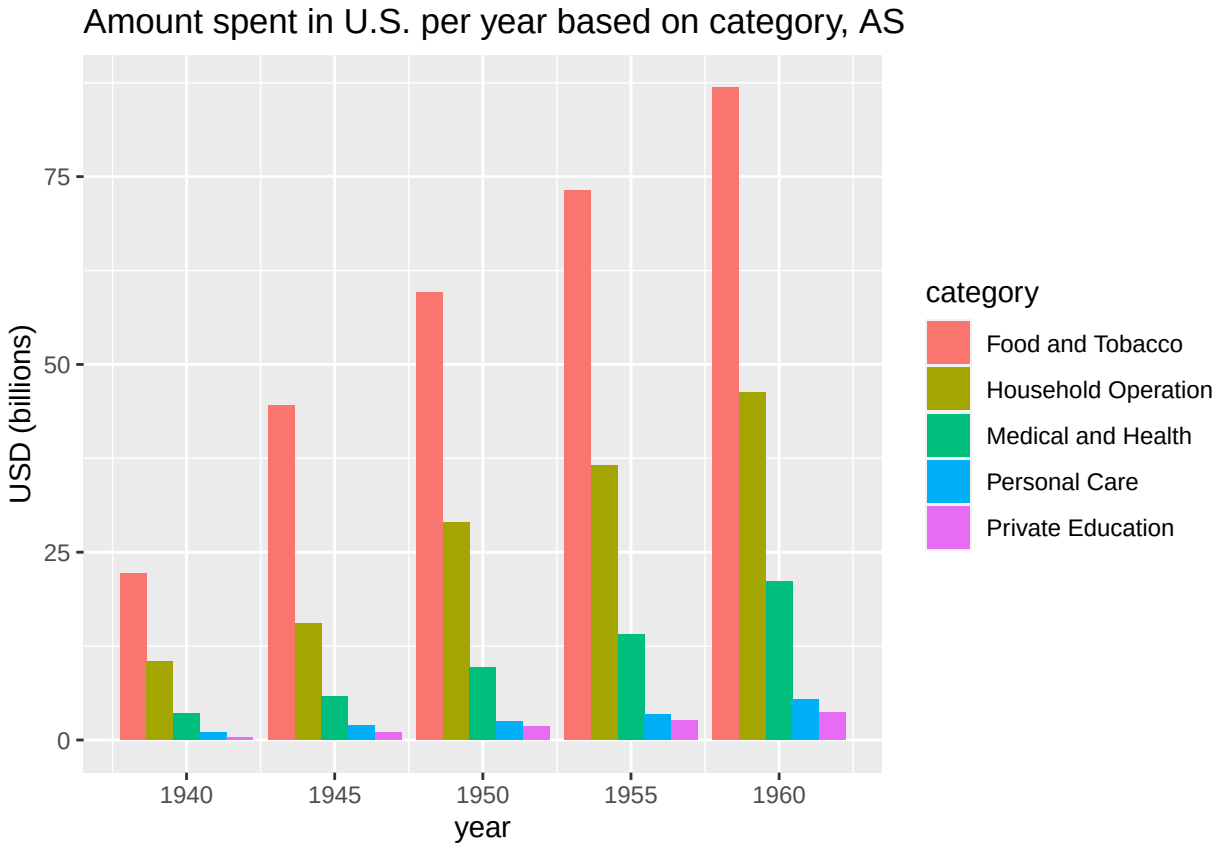
Q3*

Here are two ways of making bar plots of the amount people spent each year in the U.S. on various categories (in billions of dollars). Fill in the title (ending with your initials) and label.

```
ggplot(df,aes(x=year,y=amount,fill=category)) +
    geom_col() +
    labs(title="Amount spent in U.S. per year based on category, AS",y="USD (billions)")
```




```
ggplot(df,aes(x=year,y=amount,fill=category)) +
    geom_col(position="dodge") +
    labs(title="Amount spent in U.S. per year based on category, AS",y="USD (billions)")
```

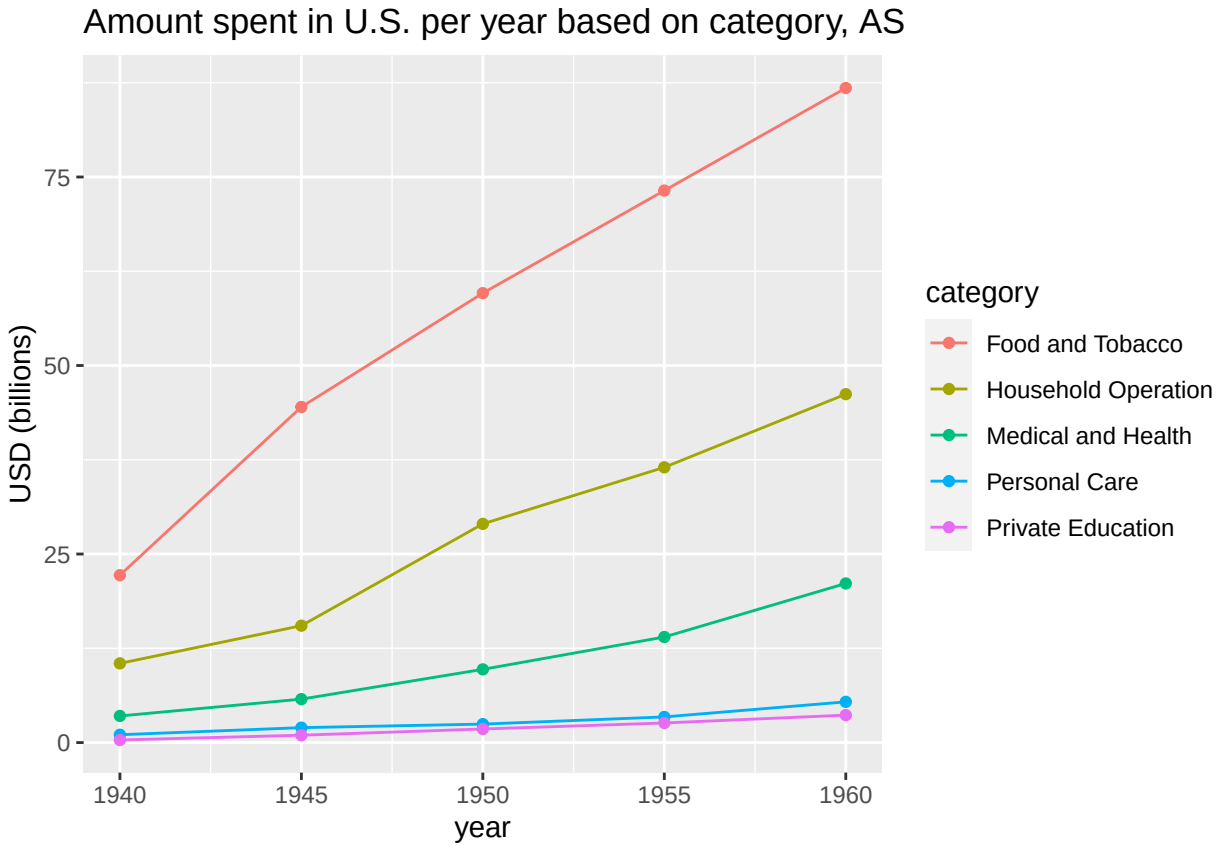


Q4*

However, there is usually a better way to represent time data. Copy and paste one of the blocks of code from above and change it to the type of plot that is generally used for time series data.

Hint: change will include a geom change, and another small item

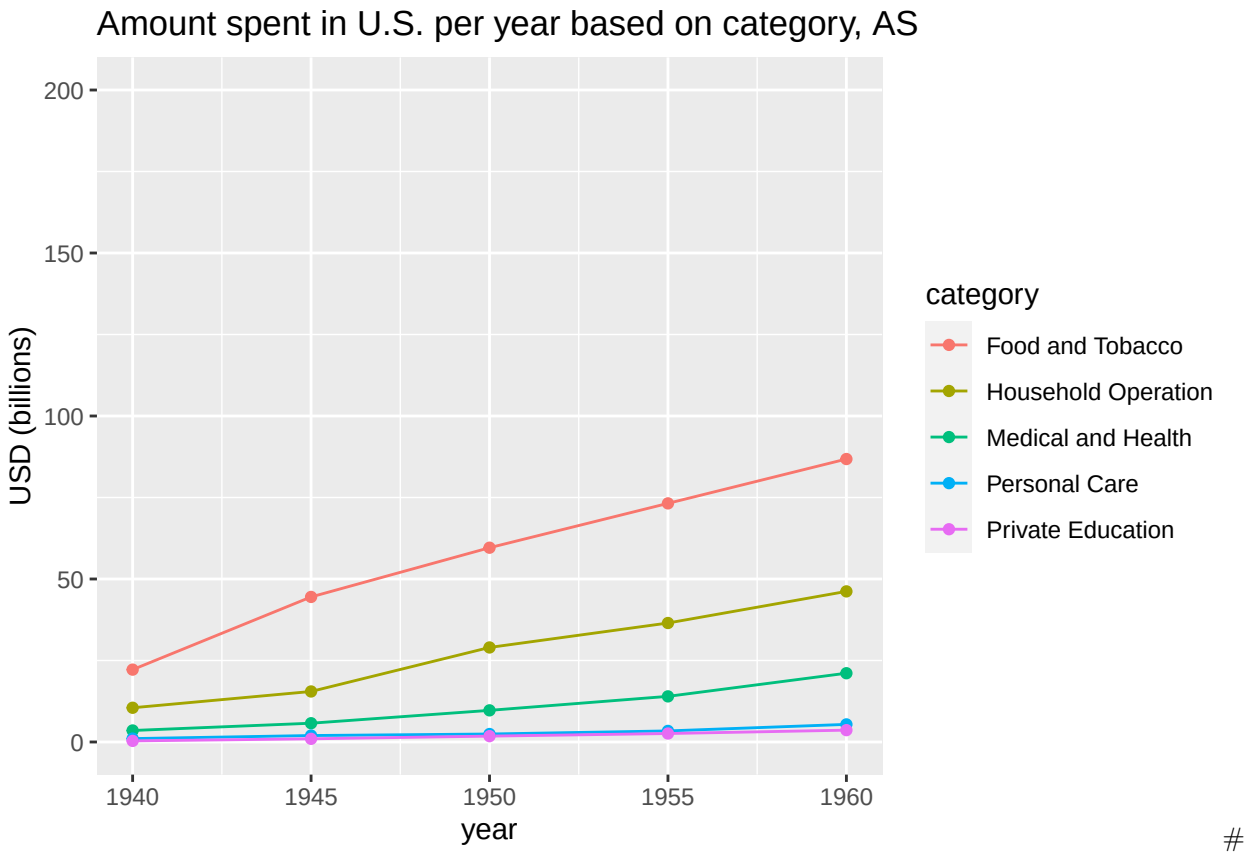
```
# put the code here...
ggplot(df,aes(x=year,y=amount,color=category)) +
    geom_line() +
    geom_point() +
    labs(title="Amount spent in U.S. per year based on category, AS",y="USD (billions)")
```



Q5 OPTIONAL

Due to the difference in order of magnitude for the expenditures, it's difficult to see how rapidly some of them grew. Add a scale change component to make it easier to see.

```
ggplot(df,aes(x=year,y=amount,color=category)) +
    geom_line() +
    geom_point() +
    xlim(1940, 1960) +
    ylim(0, 200) +
    labs(title="Amount spent in U.S. per year based on category, AS",y="USD (billions)")
```



#

Plotting a histogram, bar plot, and box plot

Q6

Run this command to see what the dataset looks like.

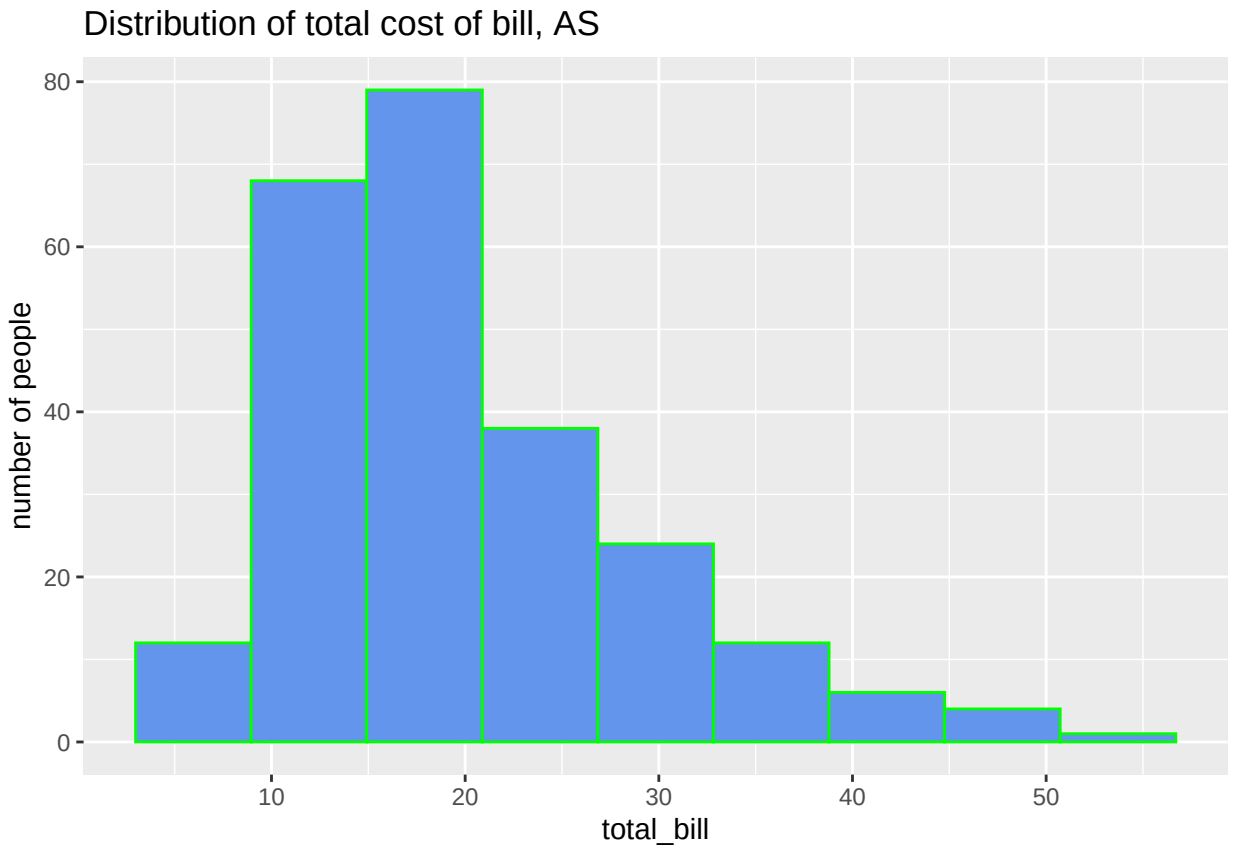
```
tips %>% head()
```

```
##   total_bill  tip    sex smoker day   time size
## 1      16.99 1.01 Female     No Sun Dinner    2
## 2      10.34 1.66   Male     No Sun Dinner    3
## 3      21.01 3.50   Male     No Sun Dinner    3
## 4      23.68 3.31   Male     No Sun Dinner    2
## 5      24.59 3.61 Female     No Sun Dinner    4
## 6      25.29 4.71   Male     No Sun Dinner    4
```

Q7*

The geom_histogram() requires either a number of bins or a binwidth. Choose one that makes the plot look nice. Add title and label appropriately.

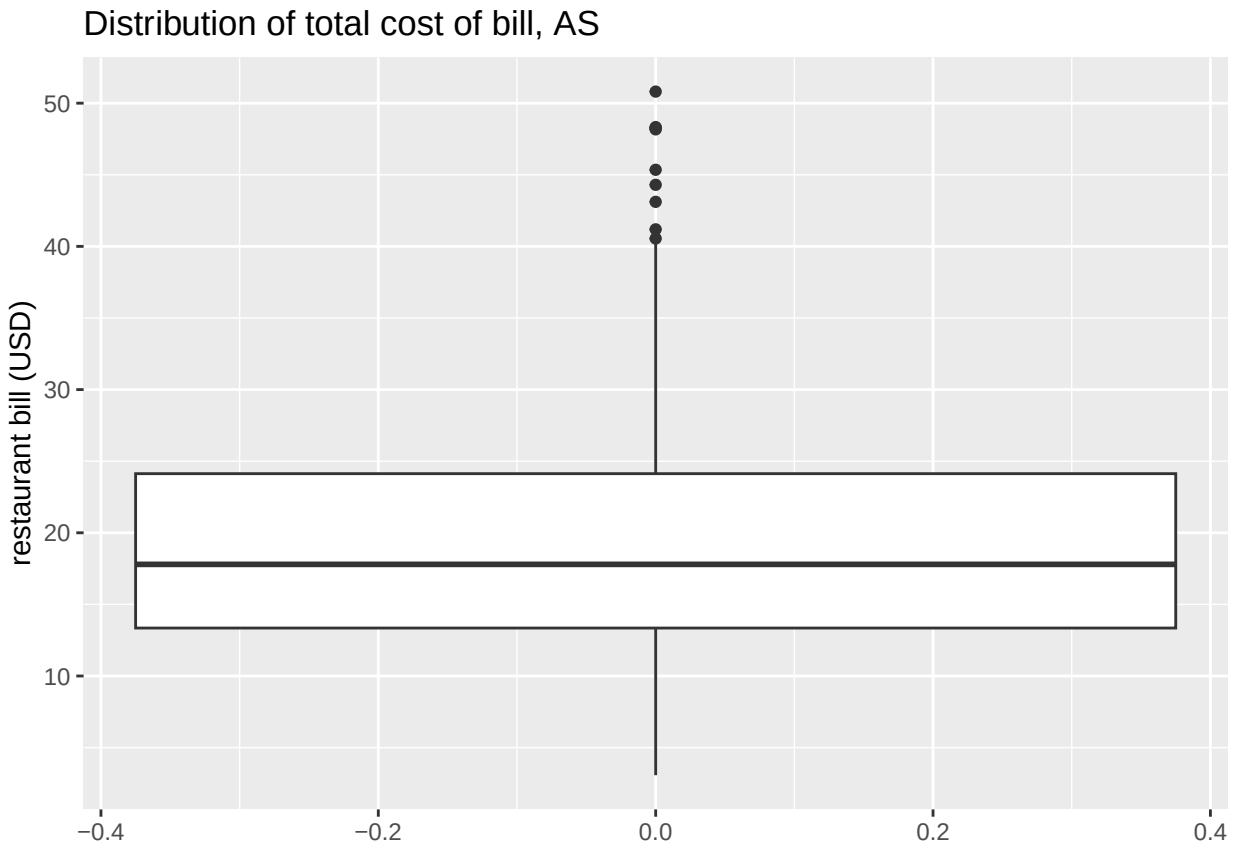
```
ggplot(tips,aes(x=total_bill)) +
    geom_histogram(bins = 9, fill = "cornflowerblue", color = "green") +
    labs(title="Distribution of total cost of bill, AS", y="number of people")
```



Q8* Now modify the histogram to become a boxplot instead. Hint: this should just involve a geom change. Also, change the "x=" to "y=". Update labels as needed.

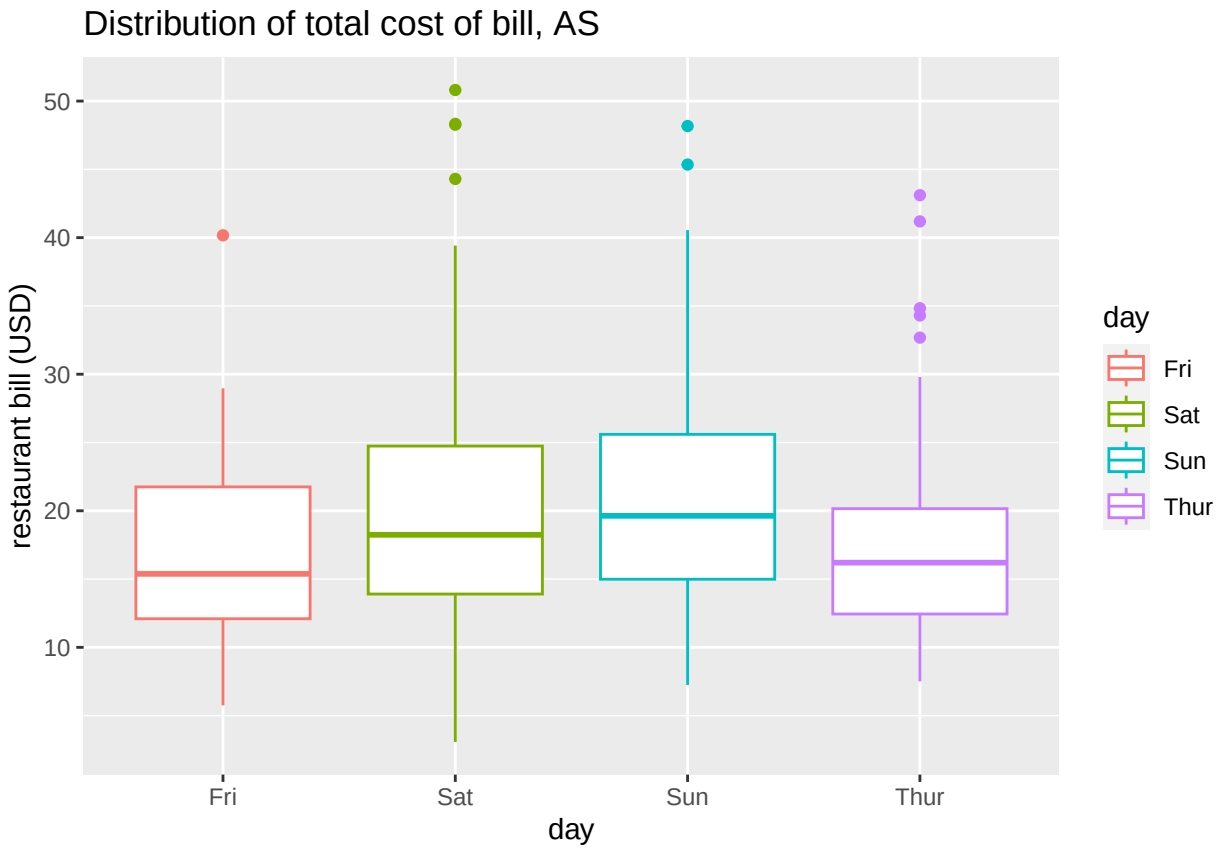
Note: this is an intermediate step (on the way to Q9) that will make a plot that doesn't yet give much information.

```
ggplot(tips,aes(y=total_bill)) +
    geom_boxplot() +
    labs(title="Distribution of total cost of bill, AS", y="restaurant bill (USD)")
```



Q9* Modify to make a boxplot by day of the week. Hint: you will need an additional specification in the "aes(...)" section.

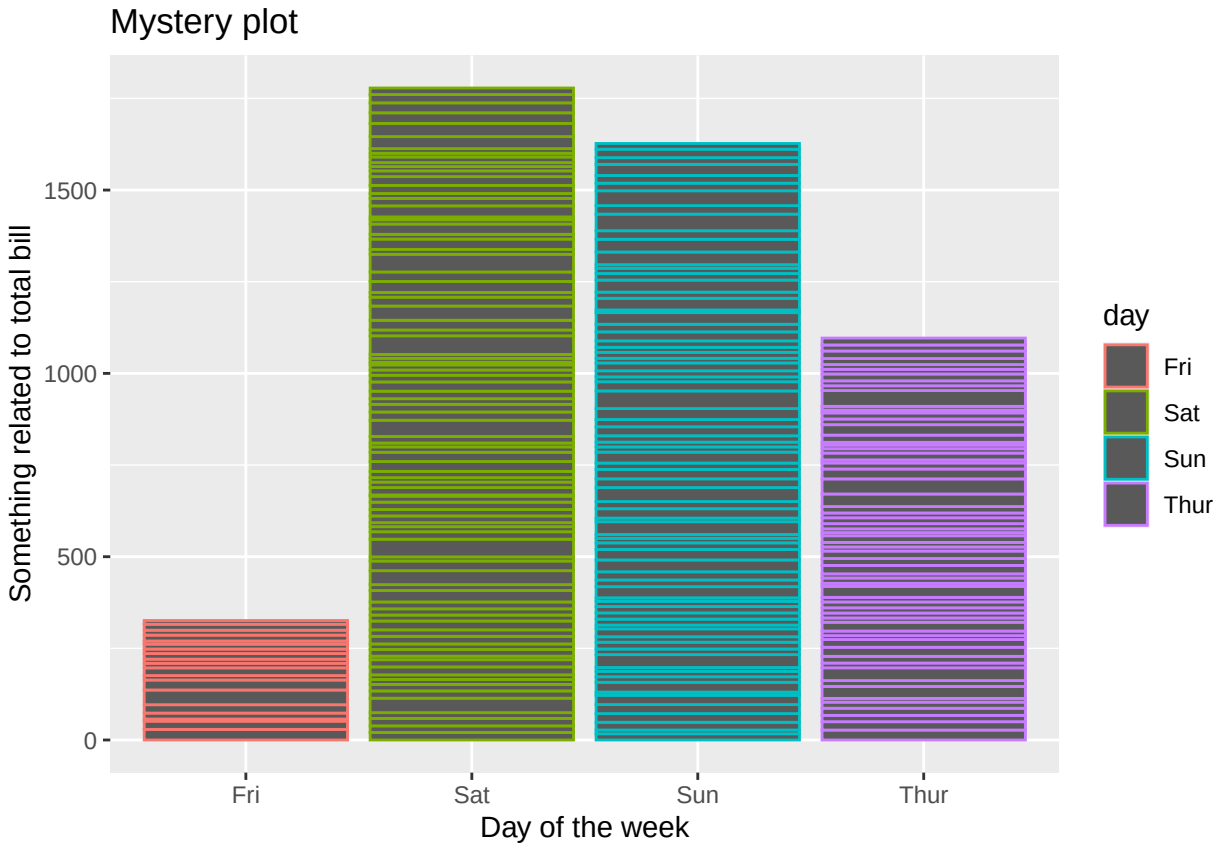
```
ggplot(tips,aes(x=day,y=total_bill,color=day)) +
    geom_boxplot() +
    labs(title="Distribution of total cost of bill, AS", y="restaurant bill (USD)")
```



Q10* (response) Suppose we want to create a bar plot of the **average** bill for dining by day of the week. You have a rough idea of how expensive the restaurant bills are based on your histogram and box plots above. What do you think the following plot might be showing? (i.e. is it showing average bill like we want, or something else, and if something else, what is it?)

It does not seem to be showing average bill spent. It is showing the sum of the total_bill per each day.

```
ggplot(tips,aes(x=day,y=total_bill,color=day)) +
    geom_col() +
    labs(title="Mystery plot",y="Something related to total bill",x="Day of the week")
```

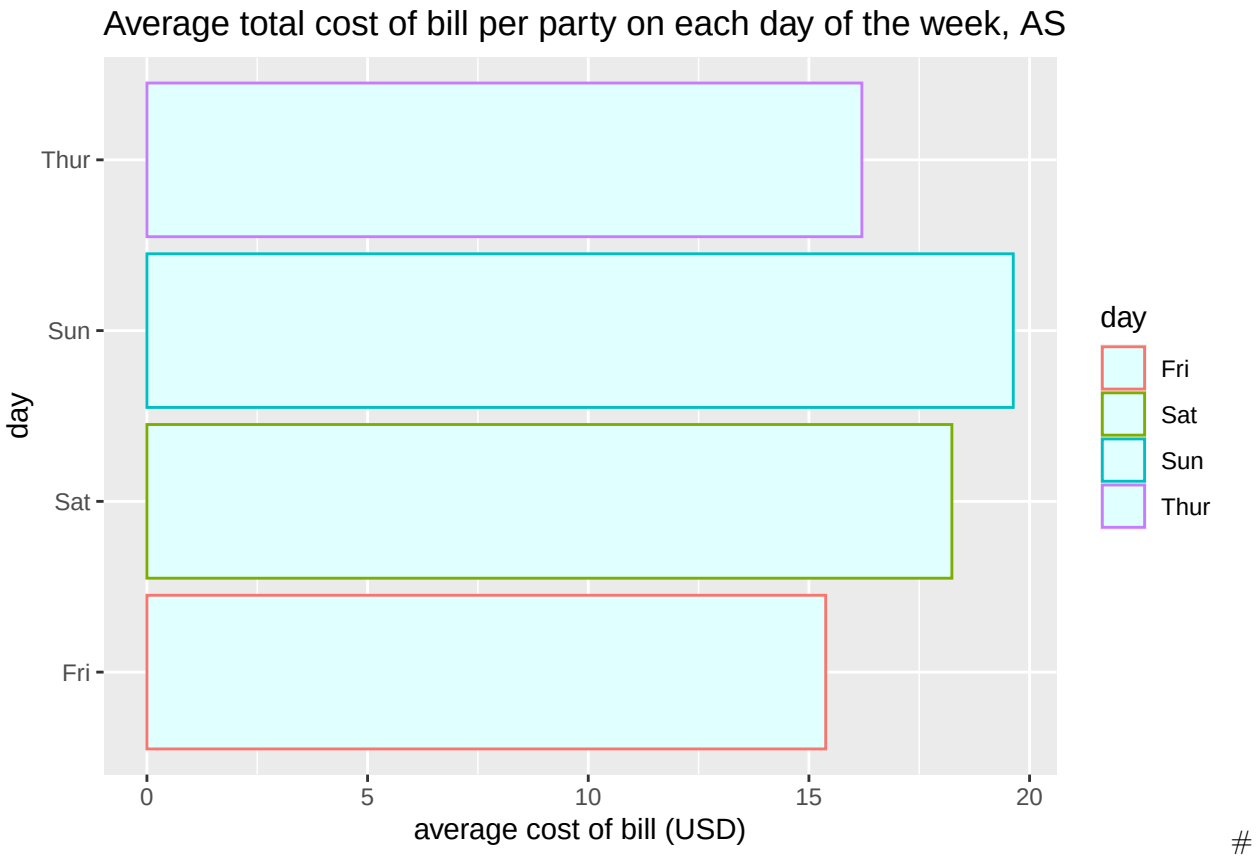


Q11* Here is the way around it. For future labs, we will learn how to do pre-processing of the data so that it is easier to make plots without this syntax.

- Customize it by flipping the coordinate to make it a horizontal plot. Do this without changing the mapping of x= and y=.

- Add a fill="colorname" option inside the stat_summary parentheses, filling in colorname whatever color you'd like. Here is a list of options: (see the assignment on Canvas)
- Title/label as needed.

```
ggplot(tips,aes(x=day,y=total_bill,color=day)) +
    coord_flip() +
    stat_summary(aes(y=total_bill), fun="median",geom = "bar",fill="lightcyan") +
    labs(title="Average total cost of bill per party on each day of the week, AS", y="average cost of bi
```



#

Plotting proportions

Q12

Run this command to see what the following dataset looks like:

```
df_car %>% head()
```

```
## # A tibble: 6 x 4
## # Groups:   manufacturer [2]
##   manufacturer class          n percent_of_brand
##   <chr>        <chr>      <int>            <dbl>
## 1 ford         pickup         7             28
## 2 ford         subcompact     9             36
## 3 ford         suv            9             36
## 4 nissan       compact        2             15.4
## 5 nissan       midsize        7             53.8
## 6 nissan       suv            4             30.8
```

Q13

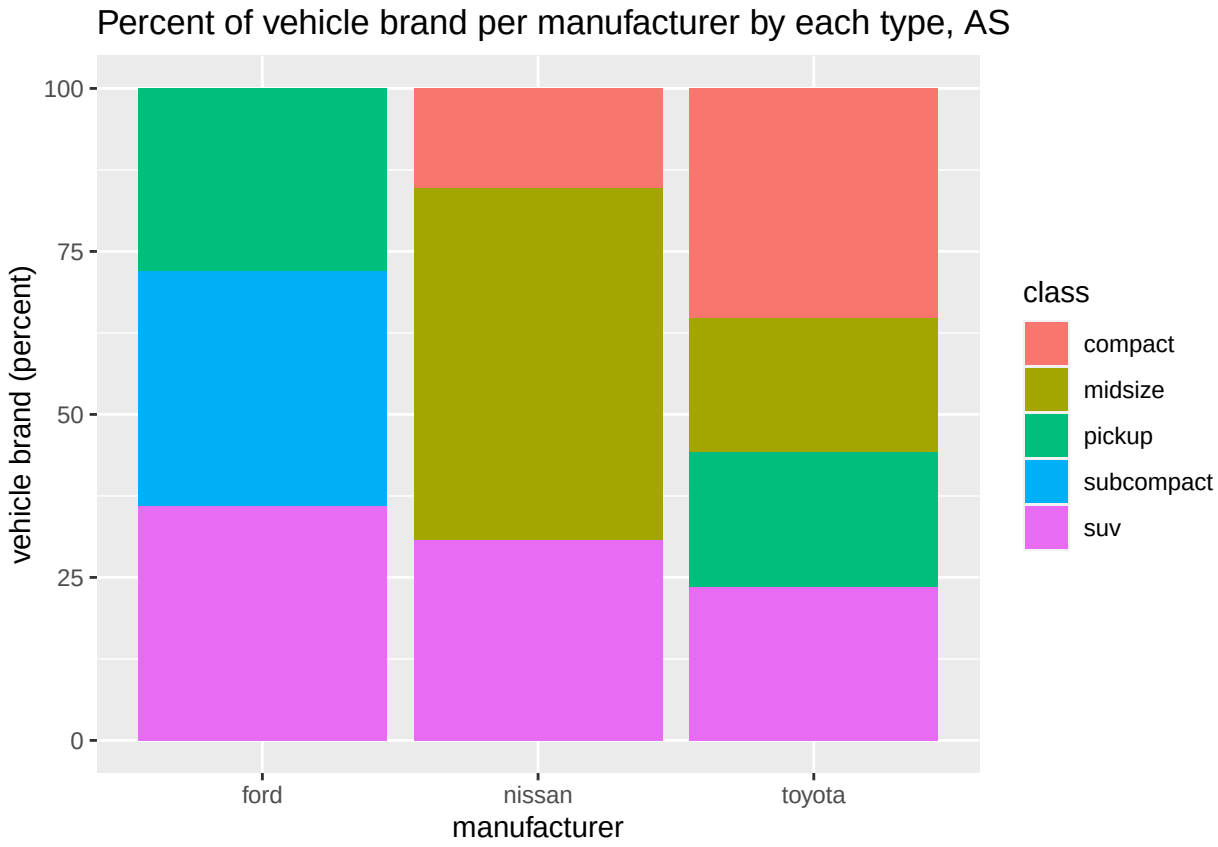
Here's a baseline plot to start with. We can add different geoms to produce plots. Run the following line of code. Note that no plot will be generated. Instead, we will obtain an object that we can add to in order to create plots.

```
car_plot <- ggplot(df_car,aes(x=manufacturer,y=percent_of_brand,fill=class))
```

Q14*

Add a geom to car_plot to make a stacked bar plot. (What else is needed to complete the plot?)

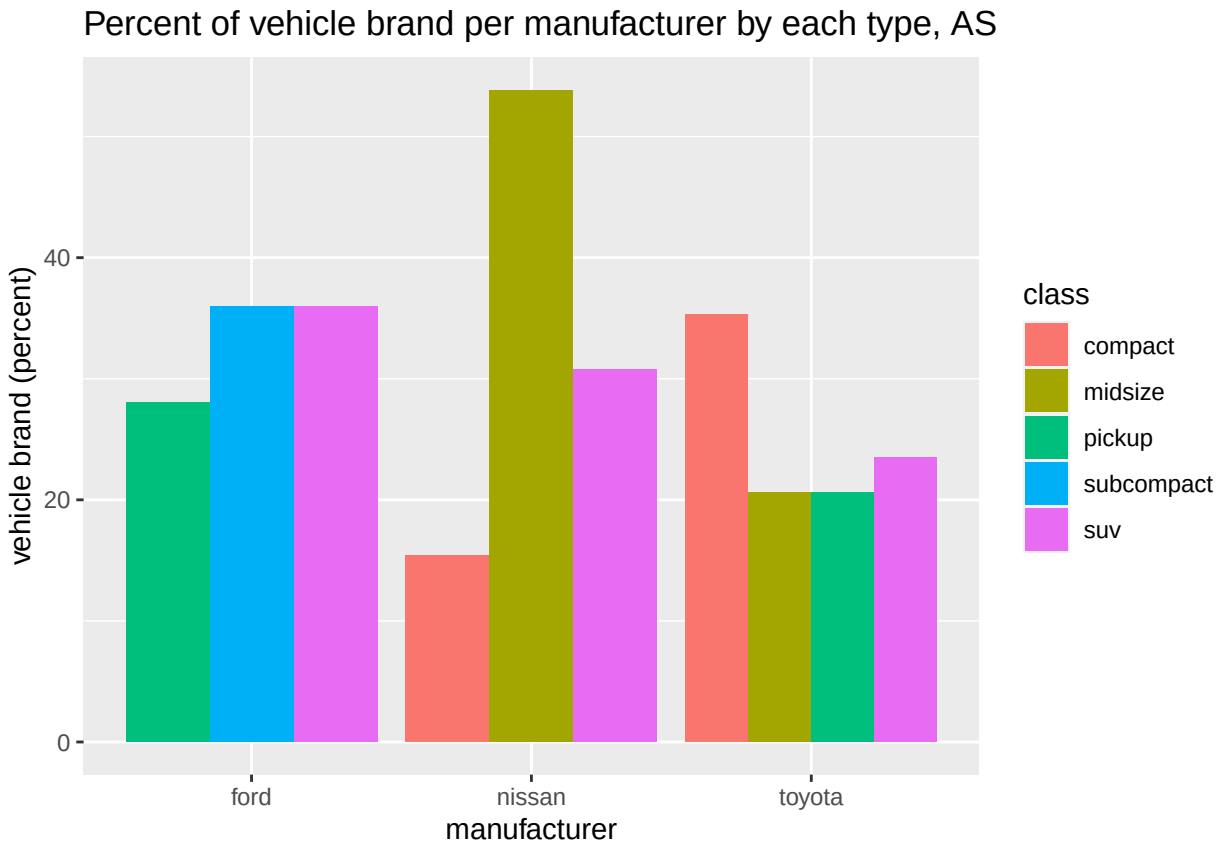
```
# Hint:
# car_plot + ...
car_plot +
    geom_col() +
    labs(title="Percent of vehicle brand per manufacturer by each type, AS",y="vehicle brand (percent)"
```



Q15*

Convert the stacked bar plot to a grouped bar chart.

```
car_plot +
    geom_col(position="dodge") +
    labs(title="Percent of vehicle brand per manufacturer by each type, AS",y="vehicle brand (percent)"
```



Q16* (response only)

Here is the code for a pie chart. Answer these questions (no coding required):

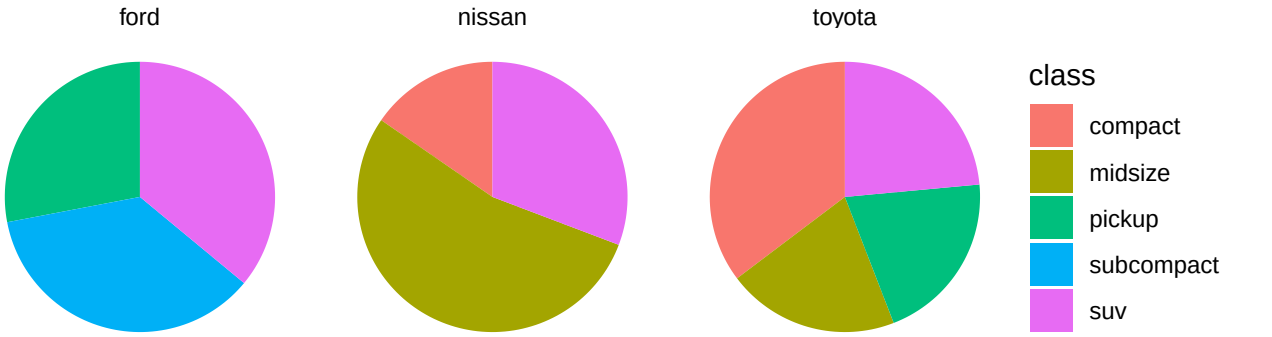
a. Why is this pie chart made with a geom_col()? In other words, what is the difference between a stacked bar and a pie chart, from the Grammar of Graphics perspective?

To directly quote the literature, 'A Layered Grammar of Graphics' (2010) by Hadley Wickham, pg. 22, it is stated that "In the grammar, a pie chart is a stacked bar geom drawn in a polar coordinate system." The pie chart is made with a geom_col() because it uses a bar chart and further maps it onto a "pie" figure to showcase its variables in this distribution.

b. What is the line "facet_grid" doing? If you're not sure, add this line on to your grouped bar plot (it won't be a correct plot, but you'll see the change that is made)

The facet_grid() function creates each individual column for the manufacturer of the car (essentially the x-axis). This aids in the audience's ease of readability.

```
ggplot(df_car,aes(x="",y=percent_of_brand,fill=class)) +
    geom_col() +
    coord_polar("y") +
    theme_void() +
    facet_grid(~manufacturer)
```

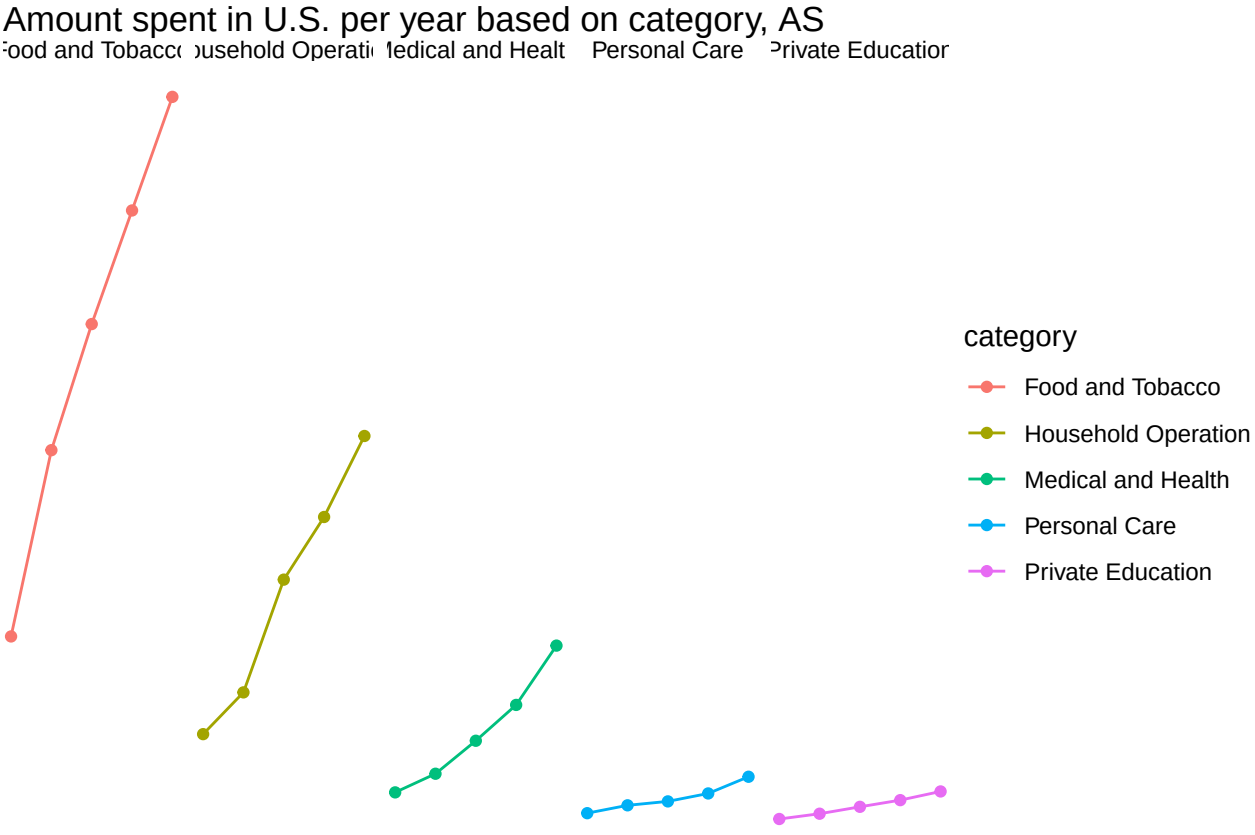



```
## decided not to add title, since I'm technically not the author of this plot -- I hope that is the ri
```

Q17*

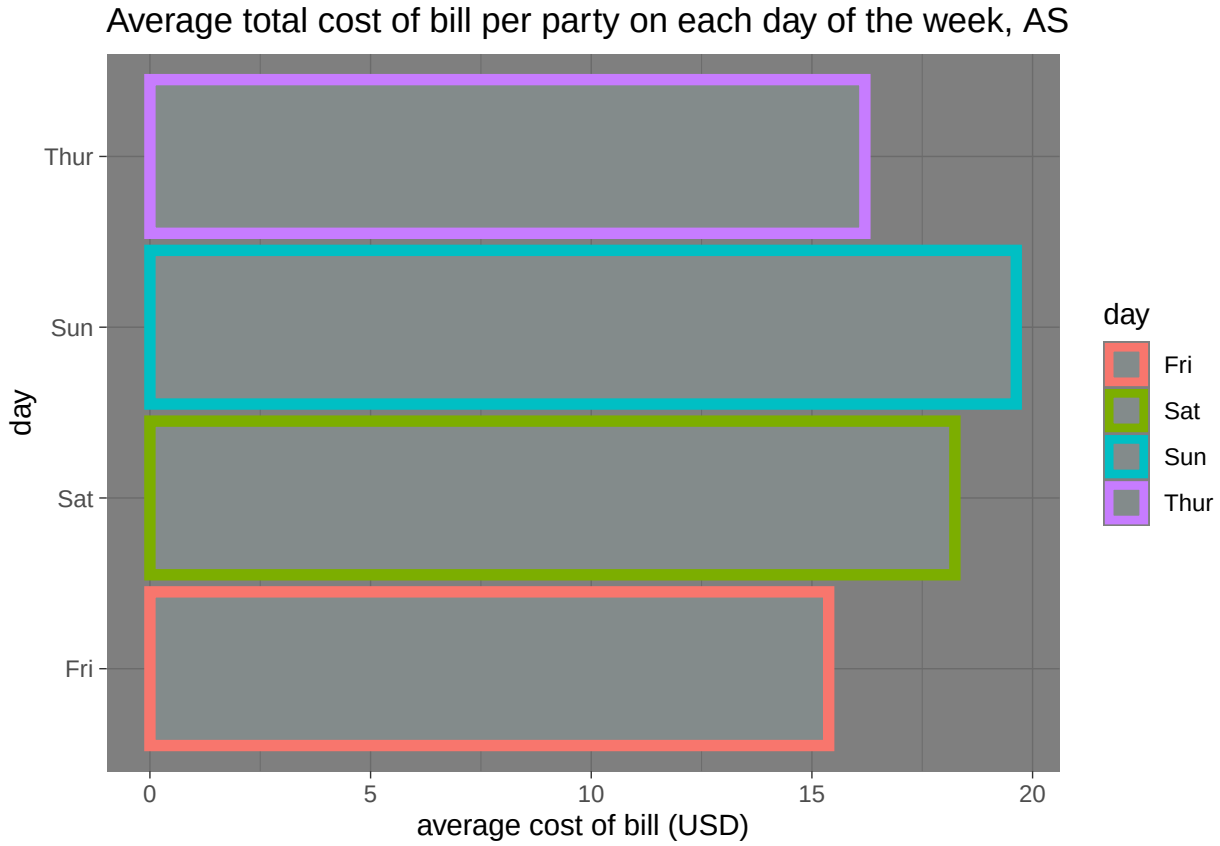
Finally, choose 3 built-in themes that you'd like to try. Choose 3 plots from this lab and paste them here, and apply a theme to each. Make these different from the default theme.

```
# theme 1:
ggplot(df,aes(x=year,y=amount,color=category)) +
    geom_line() +
    geom_point() +
    labs(title="Amount spent in U.S. per year based on category, AS",y="USD (billions)") +
    facet_grid(~category) +
    theme_void()
```

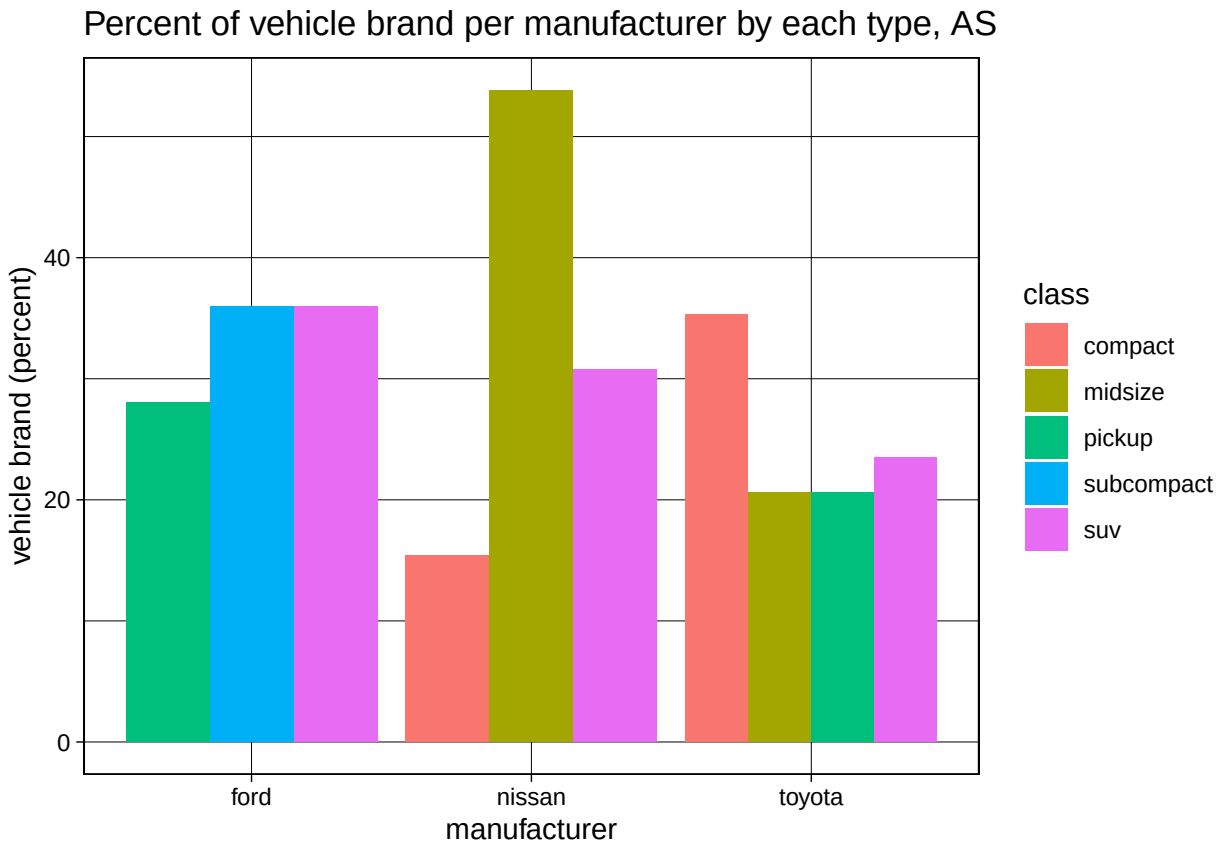



```
# theme 2:
ggplot(tips,aes(x=day,y=total_bill,color=day)) +
    coord_flip() +
```

```
    stat_summary(aes(y=total_bill), fun="median",geom ="bar",linewidth=2,fill="azure4") +
    labs(title="Average total cost of bill per party on each day of the week, AS", y="average cost of bi
    theme_dark()
```




```
# theme 3:
car_plot +
    geom_col(position="dodge") +
    labs(title="Percent of vehicle brand per manufacturer by each type, AS",y="vehicle brand (percent)"
    theme_linedraw()
```



Turning in the lab

Important:

- Don't forget to title and label all plots
- You may want to re-run everything from the top if you think you may have changed the code without running it at any point. (One easy way is to click at the top, use the "Run" dropdown, and select "Run All Chunks Below")

To turn in - 2 files:

- The .Rmd file that is the R notebook
- The .html file that is produced when you click the "Preview" button