Data visualization with ggplot

Welcome to SummeR of R at the Brandeis Library!

Who we are

Margarita Corral

mcorral@brandeis.edu

Make an appointment with Margarita!

Shannon Hagerty

shannonhagerty@brandeis.edu

Make an appointment with Shannon!

What is ggplot2?

- ggplot2 is a visualization package part of tidyverse.
- ggplot2 follows the Grammar of Graphics (GoG) [Create elegant data visualizations using the grammar of graphics, ggplot2] (https://ggplot2.tidyverse.org/)
- The idea is to build graphs from the following components:

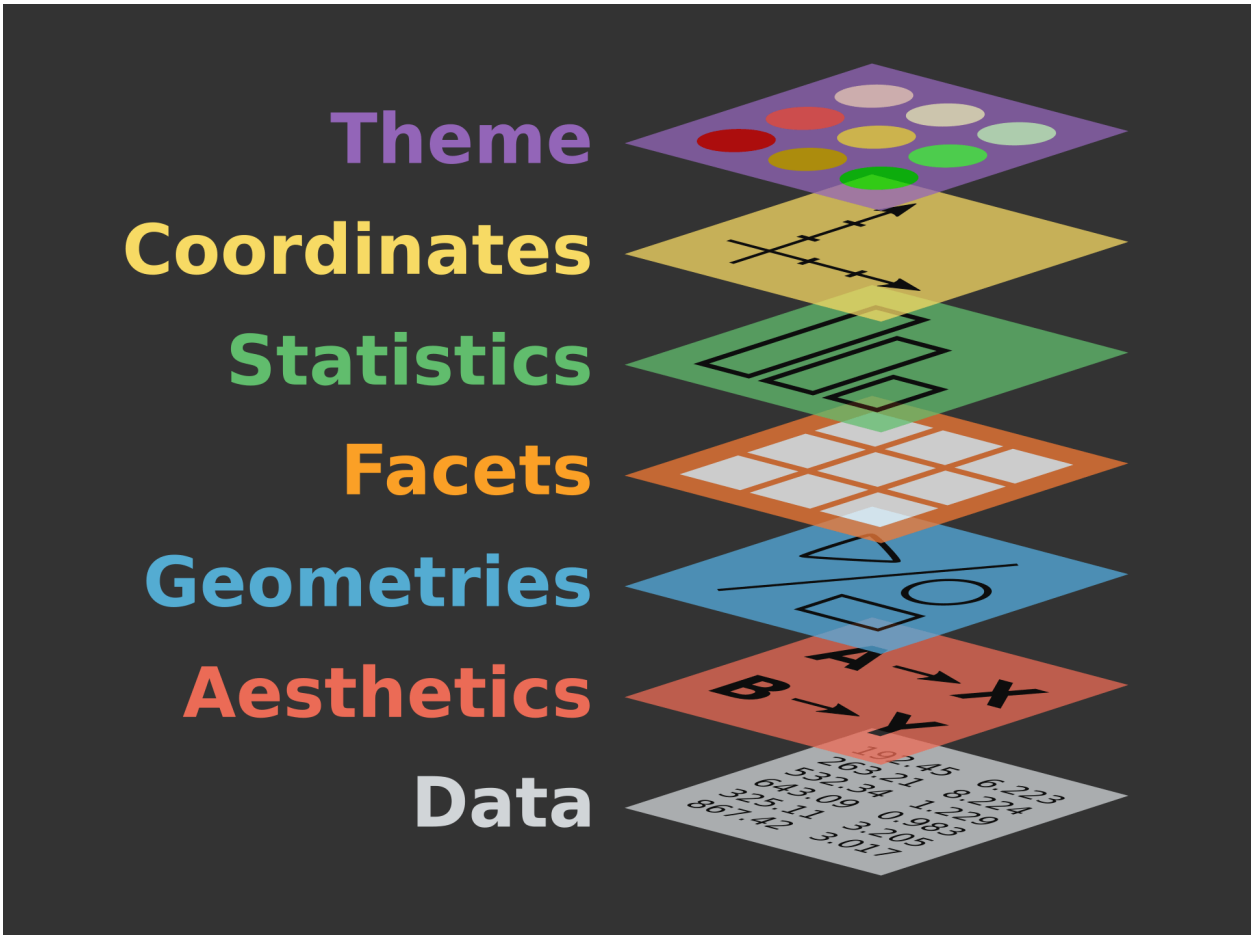


Image from *The Grammar of Graphics by Leland Wilkinson*

- Check out the https://www.rstudio.com/resources/cheatsheets/#ggplot2

Arguments for ggplot2 funtions:

- Aesthetics (Visual properties of the objects in your plot, e.g. size, shape, color, patern, fill of variables, alpha)
- Geoms (Geometric objects representing data, lines, bars, points)
- Facets (subgroups)
- Statistics (additional functions like regression lines)
- Scales (legends and labels)
- Coordinate System (Cartesian, polar..)
- Themes (Background)

Let's install/load tidyverse!

The very first time you want to use a package you first need to install it.

```
# if you have never downloaded tidyverse uncomment the line below and run to install it
#install.packages('tidyverse')
```

Load tidyverse

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.1       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Let's learn ggplot2 with some wine

- TidyTuesday:
- We will use last week's dataset. We can use this line of code to read in the data from the github page.

```
wine_ratings <- read_csv('WineRatings.csv')
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   country = col_character(),
##   description = col_character(),
##   designation = col_character(),
##   points = col_double(),
##   price = col_double(),
##   province = col_character(),
##   region_1 = col_character(),
```

```
##   region_2 = col_character(),
##   taster_name = col_character(),
##   taster_twitter_handle = col_character(),
##   title = col_character(),
##   variety = col_character(),
##   winery = col_character()
## )
```

We use the View function to look at your dataframe and check that we have tidy data (each variable is a column and each observation is a row)

```
View(wine_ratings)
```

We can delete X1.

```
wine_ratings<-select(wine_ratings, -X1)
```

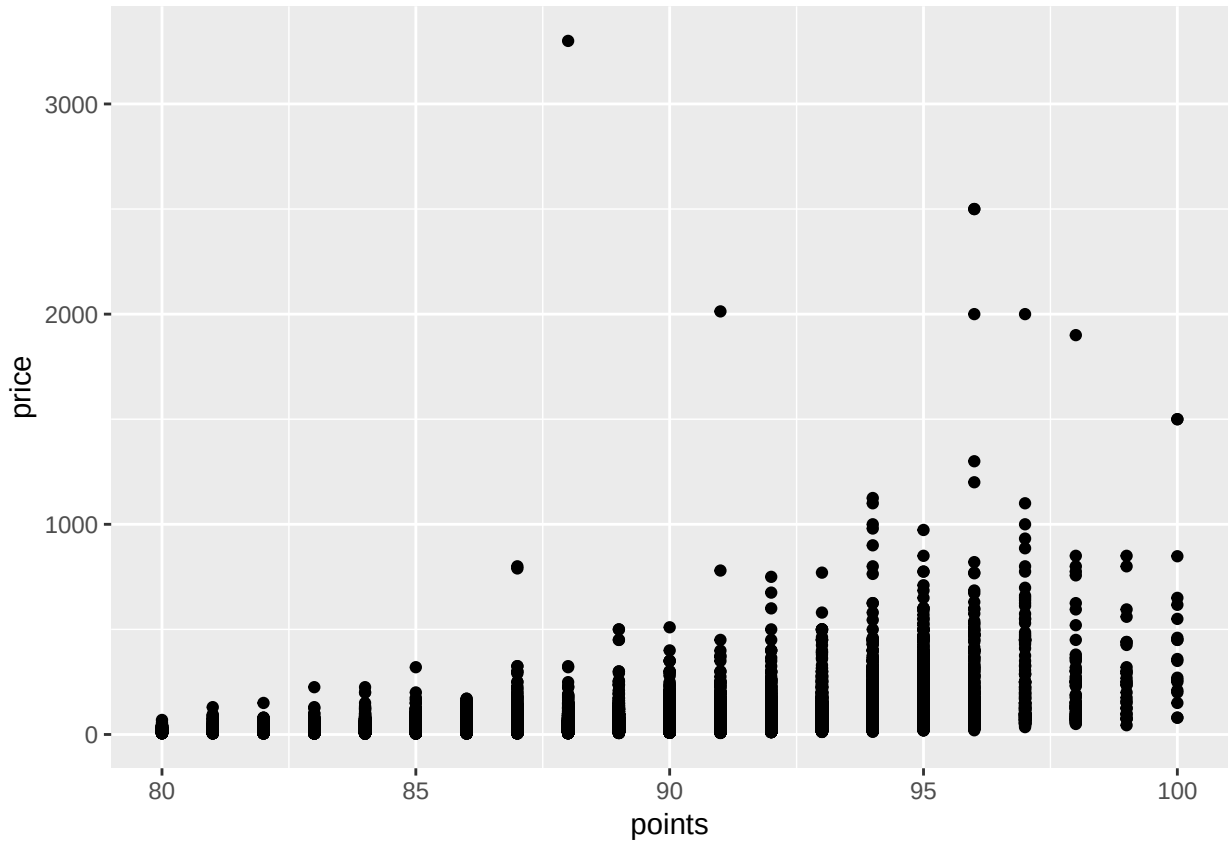
Let's create a few graphs using ggplot2.

```
ggplot(data=wine_ratings)
```


Now we need to add aesthetics and geometric objects. aes is what you plot (point, line, bar, boxplot), and geoms are how you plot aes (y, x, size, color, fill, shape specify aes() inside each geom_() so that we know which aes correspond to each geoms

```
ggplot(data=wine_ratings)+
  geom_point(aes(x=points,
                 y=price))
```

```
## Warning: Removed 8996 rows containing missing values (geom_point).
```

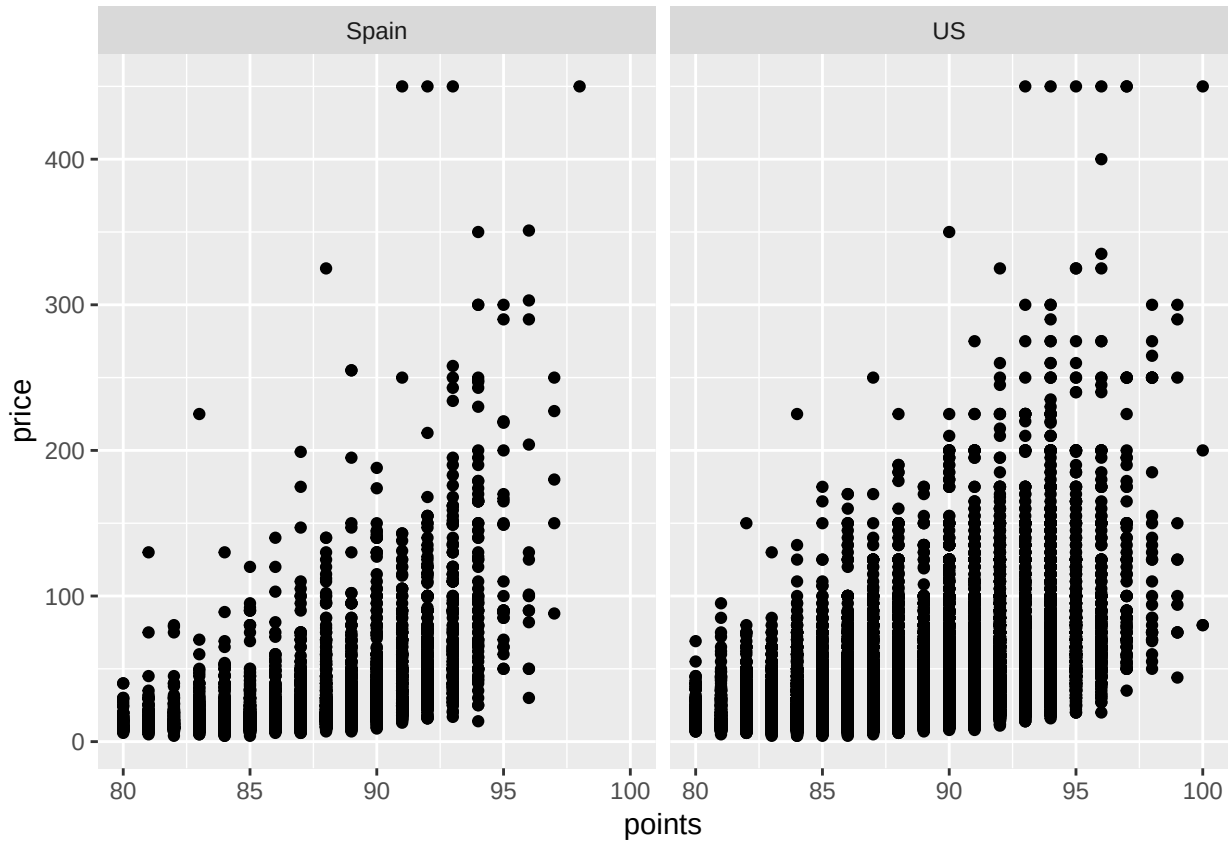


I am going to create a new data frame to compare Spain and the U.S. We will focus on cheap wine

```
Spain_and_US<- filter(wine_ratings, country %in% c("US","Spain"), price<500)
```

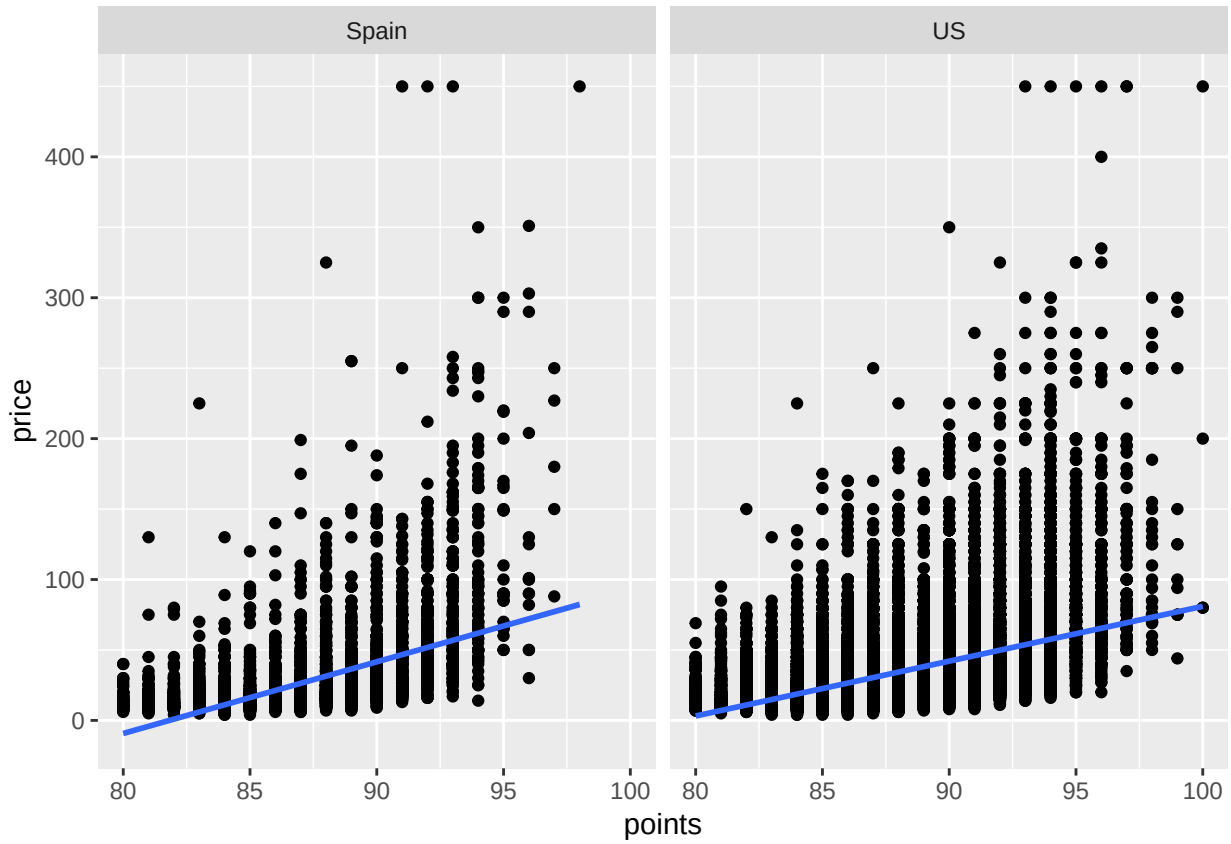
Let's add facets

```
ggplot(data=Spain_and_US)+
  geom_point(aes(x=points,
                 y=price))+
  facet_wrap(~country)
```



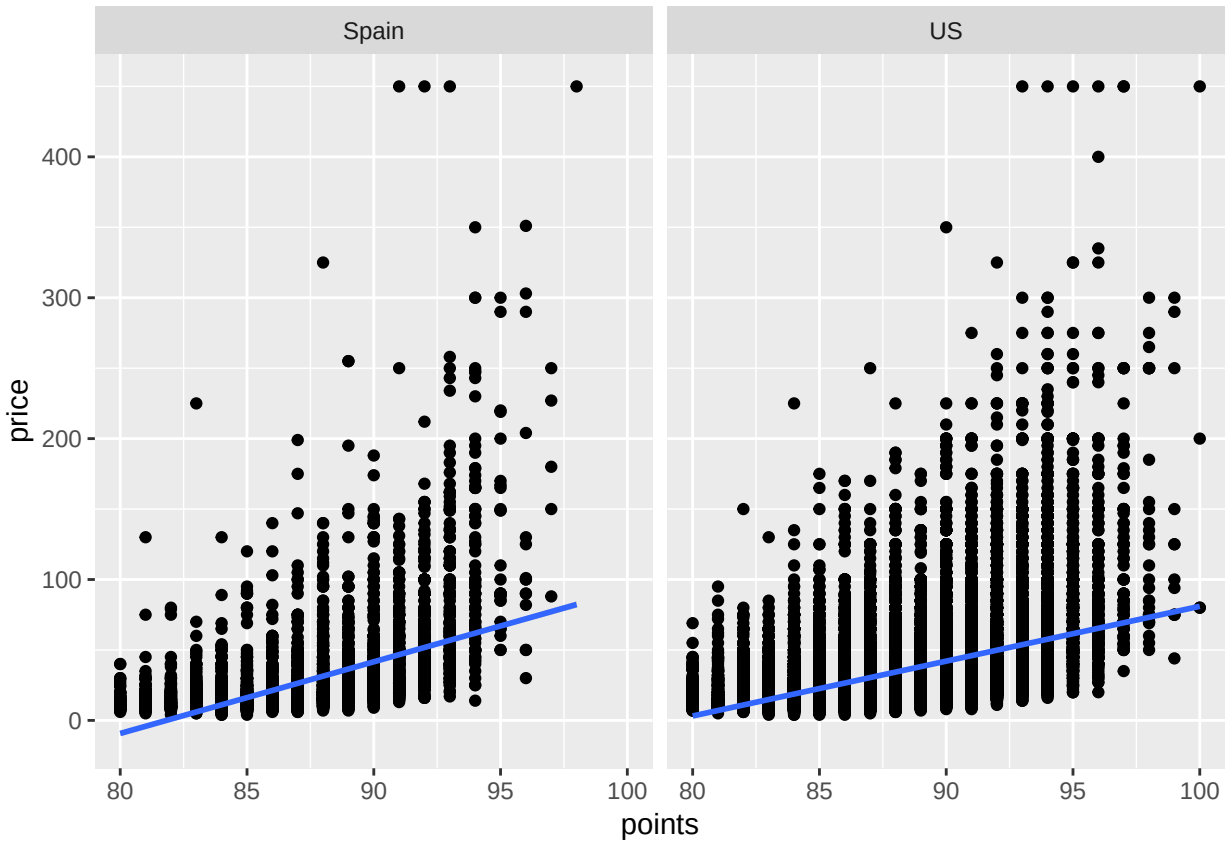
Let's add a stat layer

```
ggplot(data=Spain_and_US)+
  geom_point(aes(x=points,
                 y=price))+
  facet_wrap(~country)+
  stat_smooth(aes(x=points, y=price), method="lm", formula = y ~ x)
```



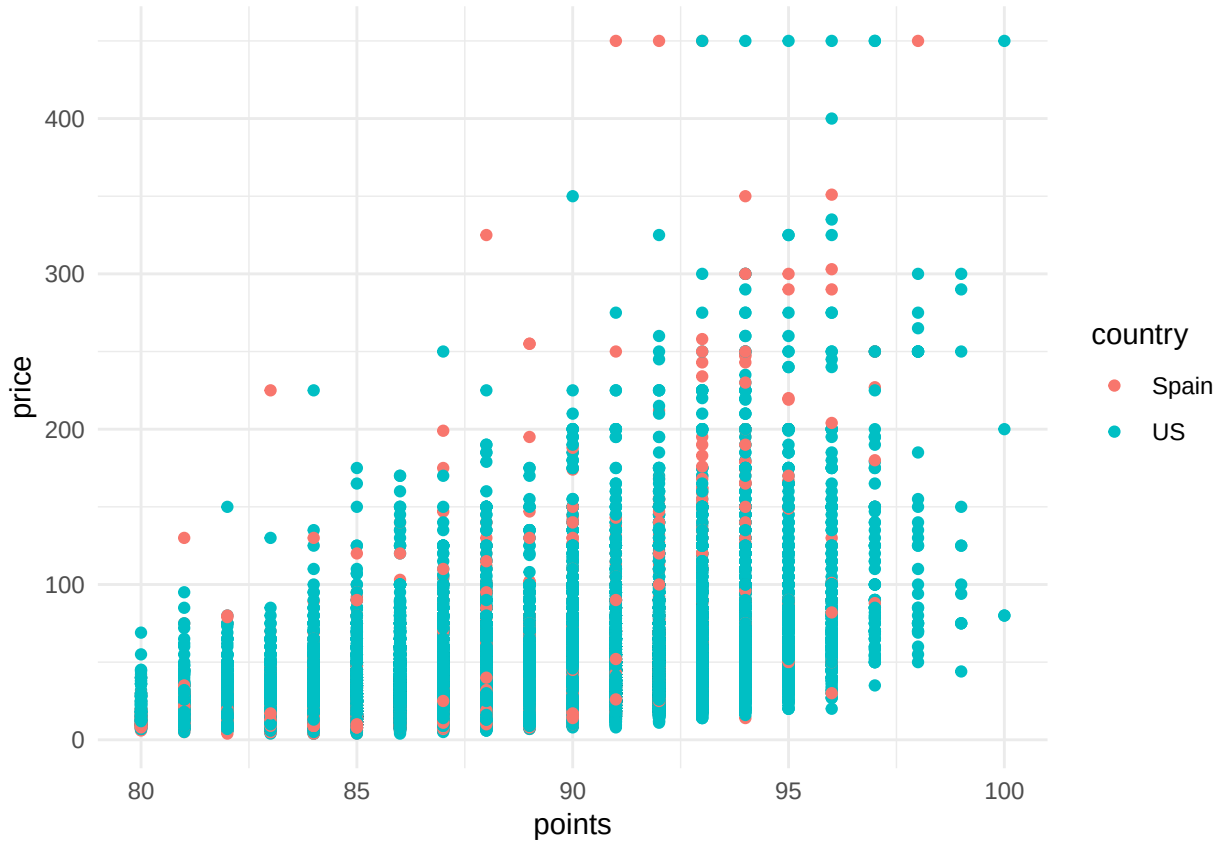

```
p<-ggplot(Spain_and_US, aes(x=points, y=price))+geom_point()+facet_grid(~country)

p+stat_smooth(method="lm", formula = y ~ x)
```



Changing the theme

```
ggplot(data=Spain_and_US)+
  geom_point(aes(x=points,
                 y=price, color=country))+
  theme_minimal()
```

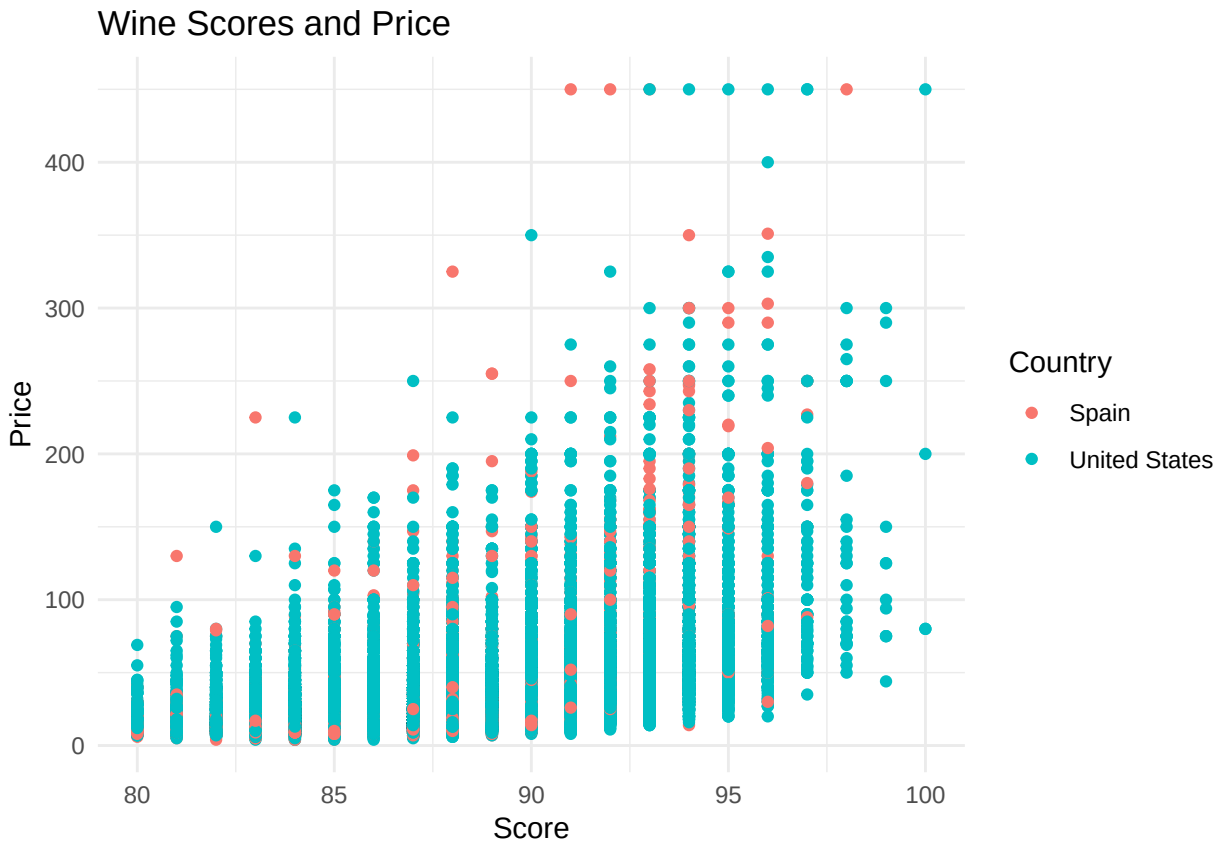


Adding Labels

```
ggplot(data=Spain_and_US)+
  geom_point(aes(x=points,
                 y=price, color=country))+
  theme_minimal()+
  labs(title = "Wine Scores and Price",
       x="Expert Scores",
       y= "Price")
```

Changing Legends

```
ggplot(data=Spain_and_US)+
  geom_point(aes(x=points,
                 y=price, color=country))+
  theme_minimal()+
  labs(title = "Wine Scores and Price",
       x="Score",
       y= "Price")+
  scale_color_discrete(name="Country", labels= c("Spain", "United States"))
```



If you would like to play with other themes, try the ggthemes package! `install.packages('ggthemes') library(ggthemes) +theme_tufte() +theme_fivethirtyeight() +theme_economist() +theme_wsj() +theme_solarized()`

- TidyTuesday:
  - This week is Ramen Ratings data. #TidyTuesday Twitter

```
ramen_ratings <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
```

```
## Parsed with column specification:
## cols(
##   review_number = col_double(),
##   brand = col_character(),
##   variety = col_character(),
##   style = col_character(),
##   country = col_character(),
##   stars = col_double()
## )
```