TidyVerse Create: ggplot2

Alice Ding

2023-04-05

Introduction

For this assignment, I'll be creating a "vignette" for the tidyverse package `ggplot2`. A description of the package is as follows:

```
ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics.
You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical
primitives to use, and it takes care of the details.
```

The Data

Taken from Kaggle, the data I'll be working with shows how many internet users there are in various countries/regions between the years 1980 and 2020. A full list of fields is below:

- Entity: Contains the name of the countries and the regions.
- Code: Information about country code and where code has the value 'Region', it denotes division by grouping various countries.
- Year: Year from 1980-2020
- Cellular Subscription: Mobile phone subscriptions per 100 people. This number can get over 100 when the average person has more than one subscription to a mobile service.
- Internet Users(%): The share of the population that is accessing the internet for all countries of the world.
- No. of Internet Users: Number of people using the Internet in every country.
- Broadband Subscription: The number of fixed broadband subscriptions per 100 people. This refers to fixed subscriptions to high-speed access to the public Internet (a TCP/IP connection), at downstream speeds equal to, or greater than, 256 kbit/s.

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
internet_users <- read.csv("https://raw.githubusercontent.com/addsding/data607/main/tidyverse/internet_u

head(internet_users)
```

```
##   X      Entity Code Year Cellular.Subscription Internet.Users...
## 1 0 Afghanistan  AFG 1980                     0                 0
## 2 1 Afghanistan  AFG 1981                     0                 0
## 3 2 Afghanistan  AFG 1982                     0                 0
## 4 3 Afghanistan  AFG 1983                     0                 0
## 5 4 Afghanistan  AFG 1984                     0                 0
## 6 5 Afghanistan  AFG 1985                     0                 0
##   No..of.Internet.Users Broadband.Subscription
## 1                     0                      0
## 2                     0                      0
## 3                     0                      0
## 4                     0                      0
## 5                     0                      0
## 6                     0                      0
```

Now that we have the data, let's see what `ggplot2` can do with it.

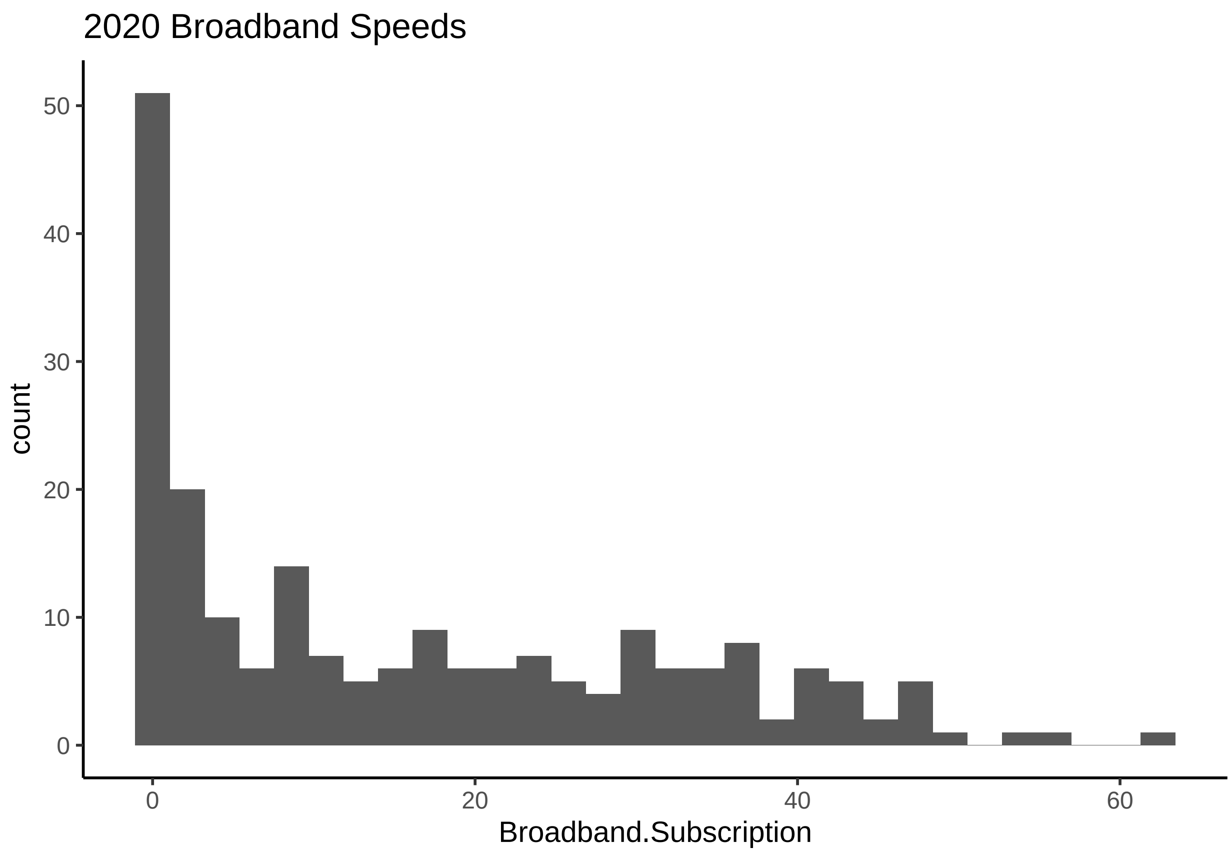
Creating the Vignette

Histograms

We can begin with histograms – how does broadband subscription speed vary for various countries overall for the year 2020?

```
users_2020 <- filter(internet_users, Year==2020, Code!="Region")

ggplot(users_2020, aes(x = Broadband.Subscription)) +
  geom_histogram() +
  theme_classic() +
  ggtitle("2020 Broadband Speeds")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



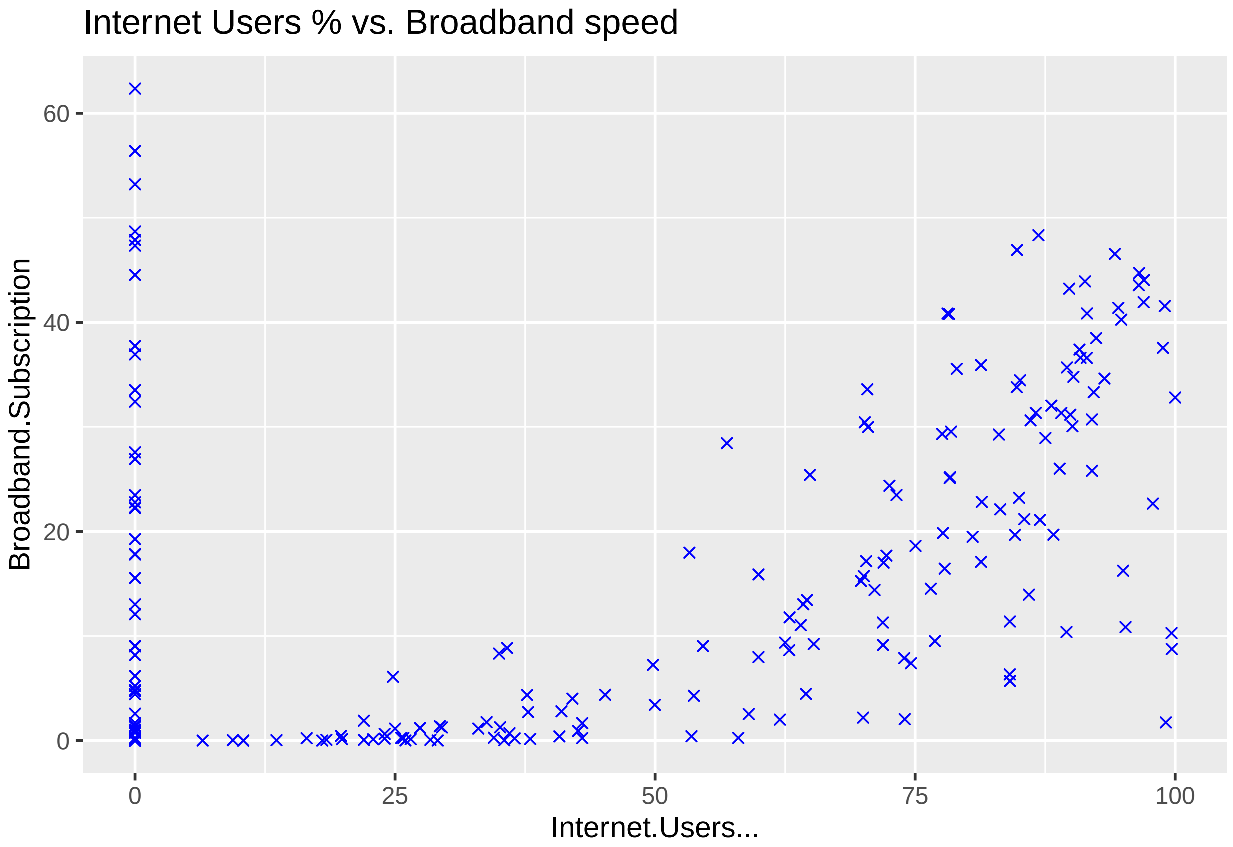
Relatively heavy for lower numbers in terms of distribution which makes sense – not every country can have fast internet.

Scatterplots

Within this section, I'll also be playing around with the different types of colors, dots, and more.

For all the countries in 2020, what % of their population is accessing the internet vs. speed?

```
ggplot(users_2020, aes(x=Internet.Users...,
                       y=Broadband.Subscription)) +
  geom_point(color="blue", shape=4) +
  ggtitle("Internet Users % vs. Broadband speed")
```

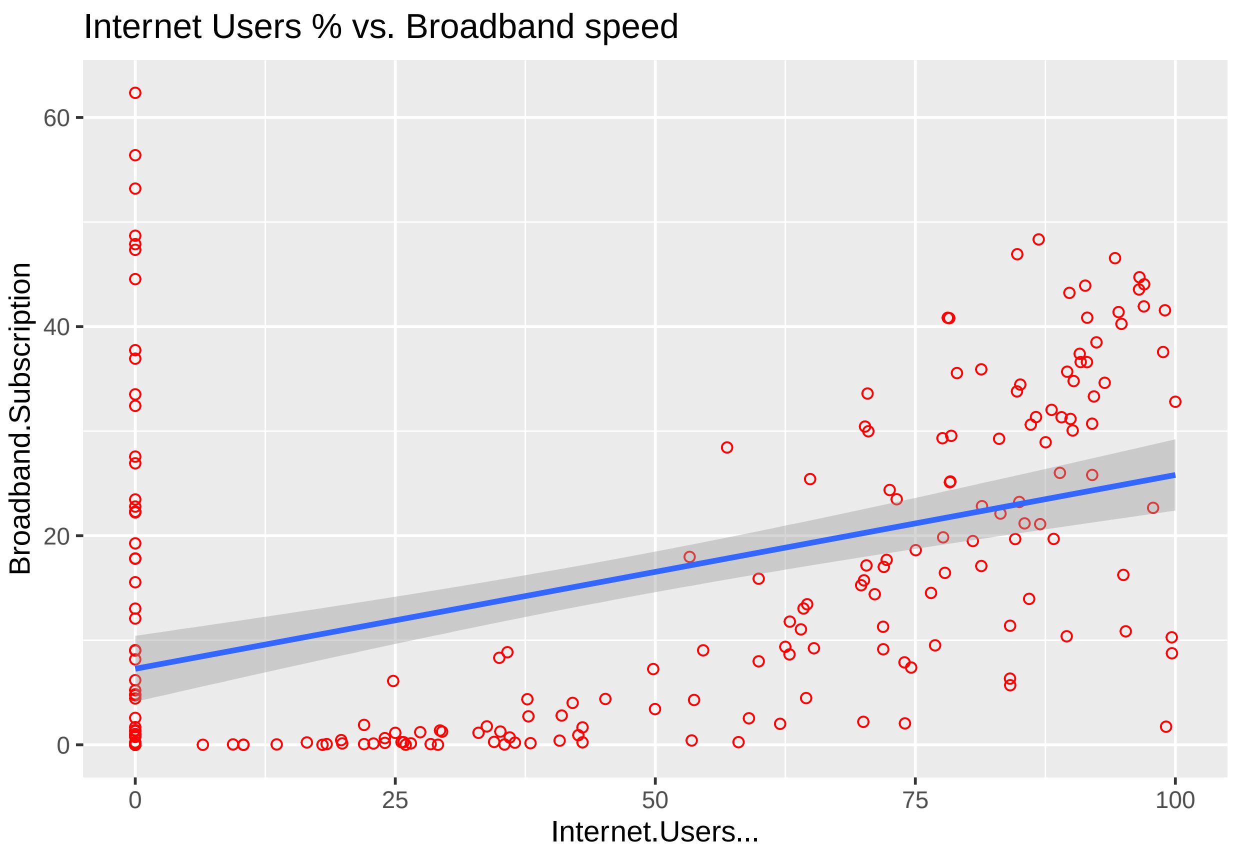


There definitely seems to be a trend here (positive slope for sure) which makes sense – the faster the internet, the more users it can support.

What if we added that line?

```
ggplot(users_2020, aes(x=Internet.Users...,
                       y=Broadband.Subscription)) +
  geom_point(color="red", shape=1) +
  geom_smooth(method=lm) +
  ggtitle("Internet Users % vs. Broadband speed")
```

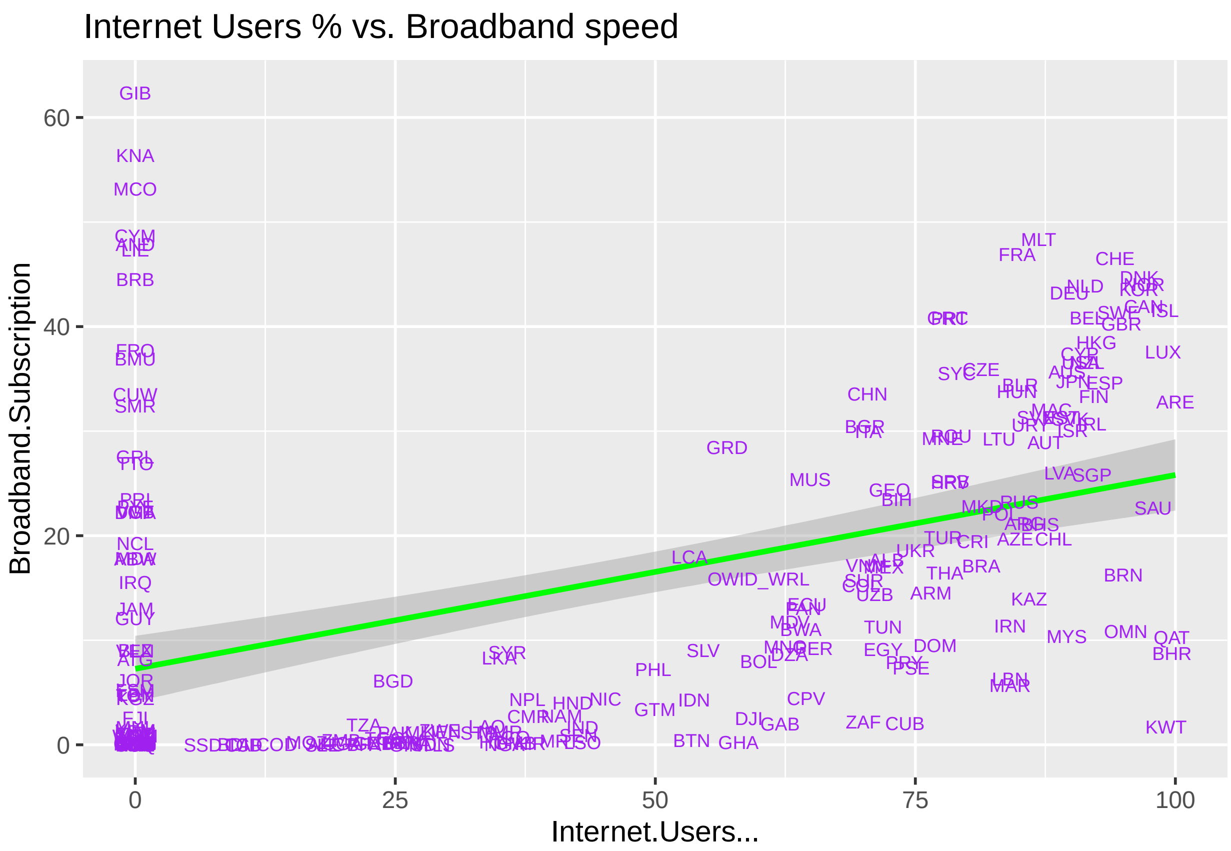
```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Looks like a positive correlation to me! We can even change the dots to actually use the country codes as labels instead.

```
ggplot(users_2020, aes(x=Internet.Users...,
                       y=Broadband.Subscription)) +
  geom_smooth(method=lm, color="green") +
  geom_text(aes(label=Code), size = 2.5, color="purple") +
  ggtitle("Internet Users % vs. Broadband speed")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

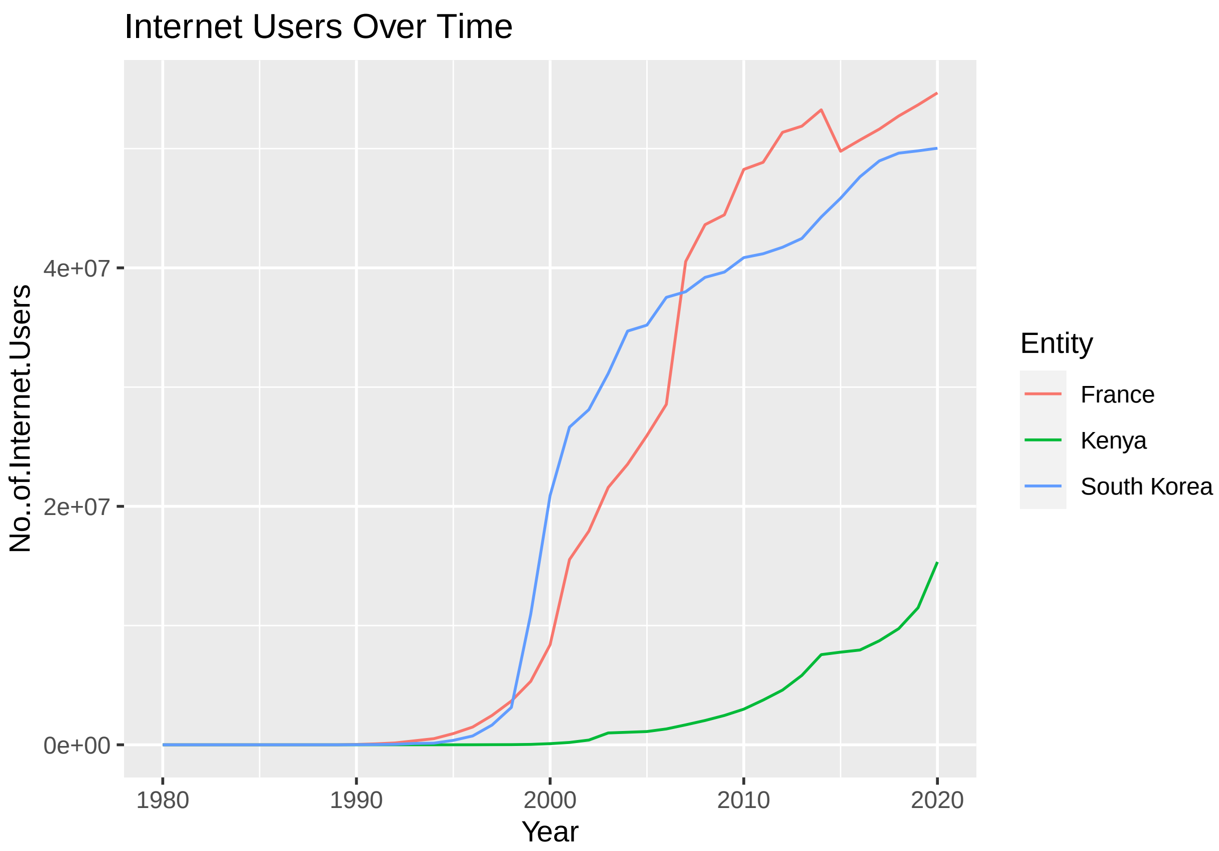


A little hard to read, but it gets the point across.

Time-Series

Let's look at internet users over time for the following countries: `South Korea`, `France`, `Kenya`.

```
ggplot(filter(internet_users, Entity %in% c("South Korea", "France", "Kenya")),
       aes(x=Year,
           y=No..of.Internet.Users,
           color=Entity))+
  geom_line() +
  ggtitle("Internet Users Over Time")
```



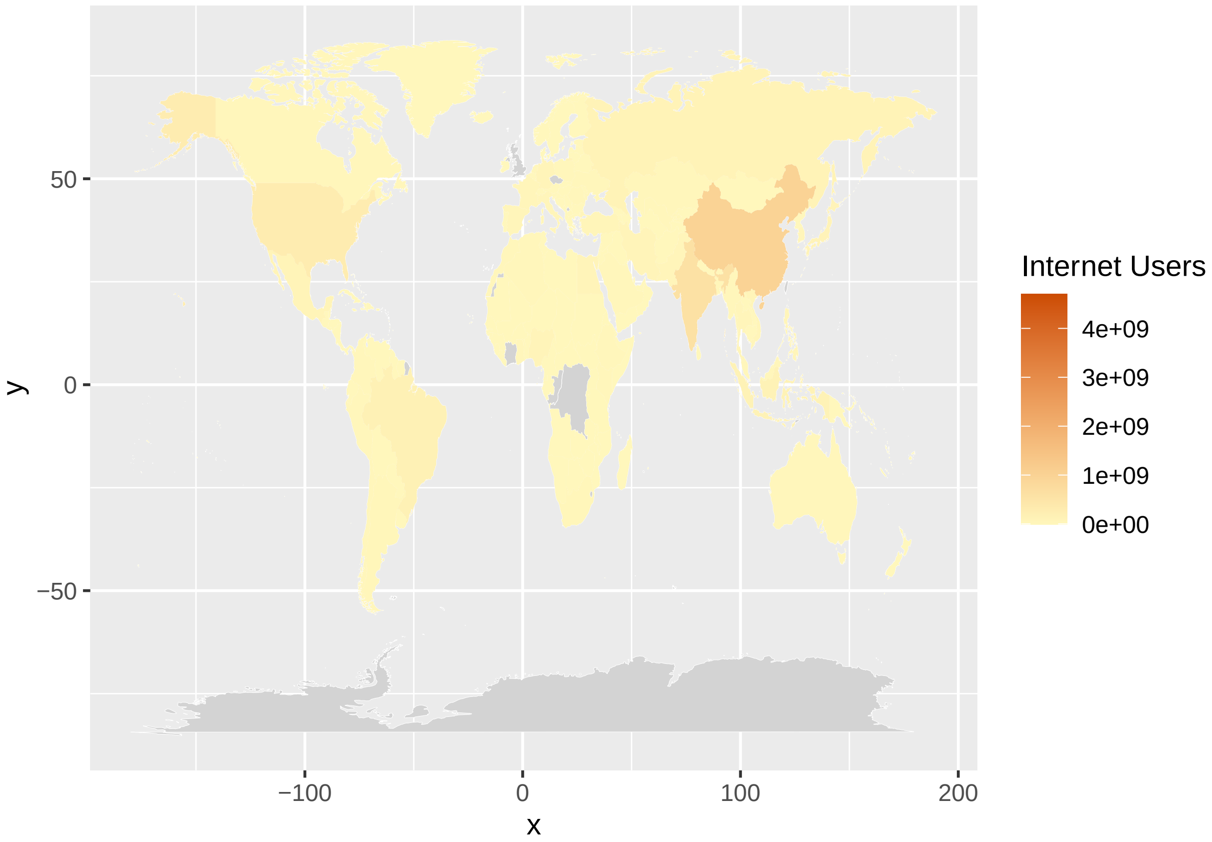
There seems to be a major spike for South Korea in the late 1990s while France seems to have a more gradual increase in the early 1990s before really jumping up in 2000. Kenya however sees their gradual increase begin in 2000 and has only recently spiked in the mid 2010s and right before 2020.

Maps

Taking the 2020 data, what does a heat map look like for internet users?

```
world <- map_data("world")

users_2020$Entity[users_2020$Entity == "United States"] <- "USA"

ggplot(users_2020) +
  geom_map(
    data = world, map = world,
    aes(map_id = region),
    color = "white", fill = "lightgray", size = 0.1
  ) +
  geom_map(
    map = world,
    aes(map_id=Entity, fill=No..of.Internet.Users), size=0.25) +
  scale_fill_gradient(low = "#fff7bc", high = "#cc4c02", name = "Internet Users") +
  expand_limits(x = world$long, y = world$lat)
```

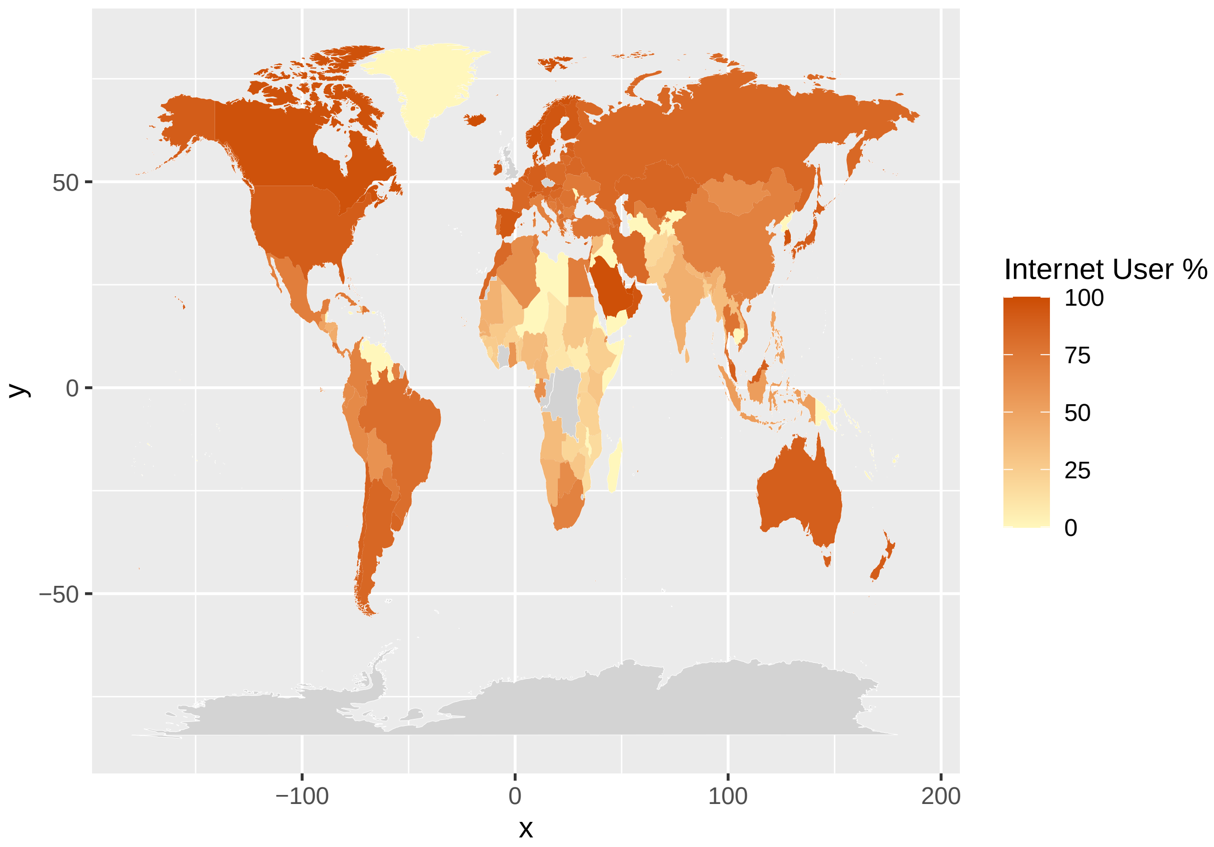
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
```



There seem to be a few missing countries here (likely just spelling/formatting – I had to adjust the United States to fit this map as well), but the basic concept of this data still holds true. It looks like that by far, China has the most internet users in the world with India, the US, and Brazil as standouts beyond that.

What if we looked at % of users for each country rather than just total count?

```
ggplot(users_2020) +
  geom_map(
    data = world, map = world,
    aes(map_id = region),
    color = "white", fill = "lightgray", size = 0.1
  ) +
  geom_map(
    map = world,
    aes(map_id=Entity, fill=Internet.Users...), size=0.25) +
  scale_fill_gradient(low = "#fff7bc", high = "#cc4c02", name = "Internet User %") +
  expand_limits(x = world$long, y = world$lat)
```



A much more varied picture – it looks like places that are more underdeveloped such as certain countries in Africa have a lower percentage while more developed countries are closer to 100.

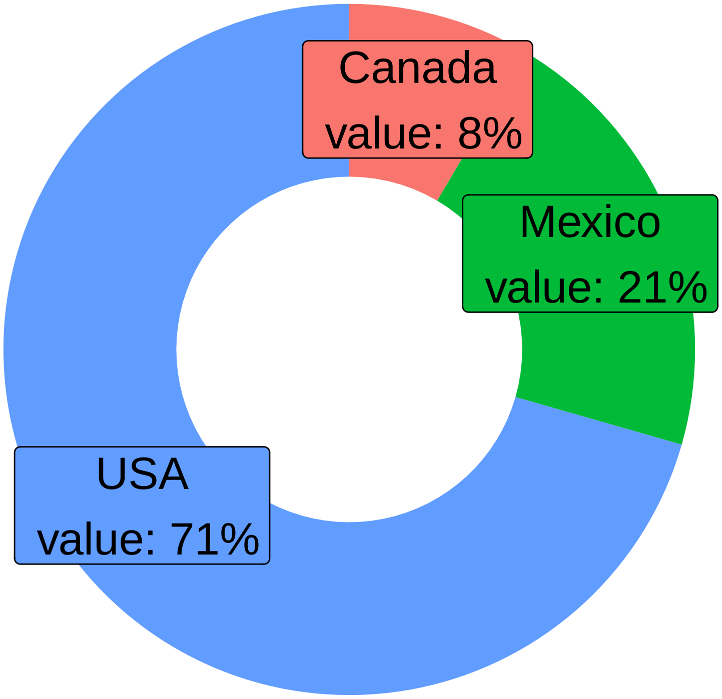
Donut Charts

While there is no specific function for this, we can hack our way into making a donut/pie chart by making a stacked rectangle chart and then making it circular.

We'll be taking a look at the 2020 numbers again to see what % lies within each country. To make things easiest, we'll just do the North American countries (Canada, USA, Mexico)

```
library(scales)

users_2020_na <- filter(users_2020, Entity %in% c("Canada", "USA", "Mexico"))

# percentages
users_2020_na$fraction <- users_2020_na$No..of.Internet.Users / sum(users_2020_na$No..of.Internet.Users
users_2020_na$fraction_label <- label_percent()(users_2020_na$fraction)

# cumulative percentages (top of each rectangle)
users_2020_na$ymax = cumsum(users_2020_na$fraction)

# bottom of each rectangle
users_2020_na$ymin = c(0, head(users_2020_na$ymax, n=-1))

# label position
```

```
users_2020_na$position <- (users_2020_na$ymax + users_2020_na$ymin) / 2

# nicer label
users_2020_na$label <- paste0(users_2020_na$Entity, "\n value: ", users_2020_na$fraction_label)

ggplot(users_2020_na, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=Entity)) +
     geom_rect() +
     coord_polar(theta="y") +
     xlim(c(2, 4)) +
     geom_label( x=3.5, aes(y=position, label=label), size=6) +
     theme_void() +
     theme(legend.position = "none")
```



Interestingly, Canada is only 8% of internet users for North America.

Conclusion

ggplot2 provides users a ton of different ways to make visualizations and customize them however they want. From the various colors, chart types, and other changeable features, ggplot2 is a great package to help tell visual stories with data and make things easily digestible with a variety of graph types to suit any use case!