Extra work on Project 2

waheeb Algabri

```
library(tidyverse)
library(readr)
library(dplyr)
```

```
songs <- read_csv("100 Songs.csv")
```

```
head(songs)
```

Load required packages

```
## # A tibble: 6 x 14
##   id    name  durat~1 energy   key loudn~2  mode speec~3 acous~4 instr~5 liven~6
##   <chr> <chr>   <dbl>  <dbl> <dbl>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 4ZtF~ Good~    2.97  0.664     9   -5.04     1  0.154  0.335   0        0.0849
## 2 5fxy~ Stay~    2.3   0.506     8  -11.3      1  0.0589 0.379   8.68e-1  0.11
## 3 5nuj~ Levi~    3.38  0.825     6   -3.79     0  0.0601 0.00883 0        0.0674
## 4 4iJy~ Peac~    3.3   0.696     0   -6.18     1  0.119  0.321   0        0.42
## 5 1SC5~ Mont~    2.3   0.503     8   -6.72     0  0.22   0.293   0        0.405
## 6 3Dar~ Kiss~    3.48  0.705     8   -3.46     1  0.0284 0.259   8.92e-5  0.12
## # ... with 3 more variables: valence <dbl>, tempo <dbl>, danceability <dbl>,
## #   and abbreviated variable names 1: duration, 2: loudness, 3: speechiness,
## #   4: acousticness, 5: instrumentalness, 6: liveness
```

```
library(dplyr)
library(tidyr)

data <- songs %>%
  rename(
    Acousticness = acousticness,
    Danceability = danceability,
    Duration = duration,
    Energy = energy,
    Id = id,
    Instrumentalness = instrumentalness,
    Key = key,
```

```
    Liveness = liveness,
    Loudness = loudness,
    Mode = mode,
    Name = name,
    Speechiness = speechiness,
    Tempo = tempo,
    Valence = valence
  )
```

```
sum(is.na(data))
```

Tiding and transforming the data

```
## [1] 0
```

```
sum(duplicated(data))
```

```
## [1] 10
```

Remove the duplicated

```
data <- unique(data)
```

change the ID value to be NA

```
data$ Id<- as.numeric(data$Id)
```

Replace the NA values in the Id column with numbers, you can use the if_else() function from dplyr package

```
library(dplyr)

data <- data %>%
  mutate(Id = if_else(is.na(Id), 1:nrow(data), Id))
```

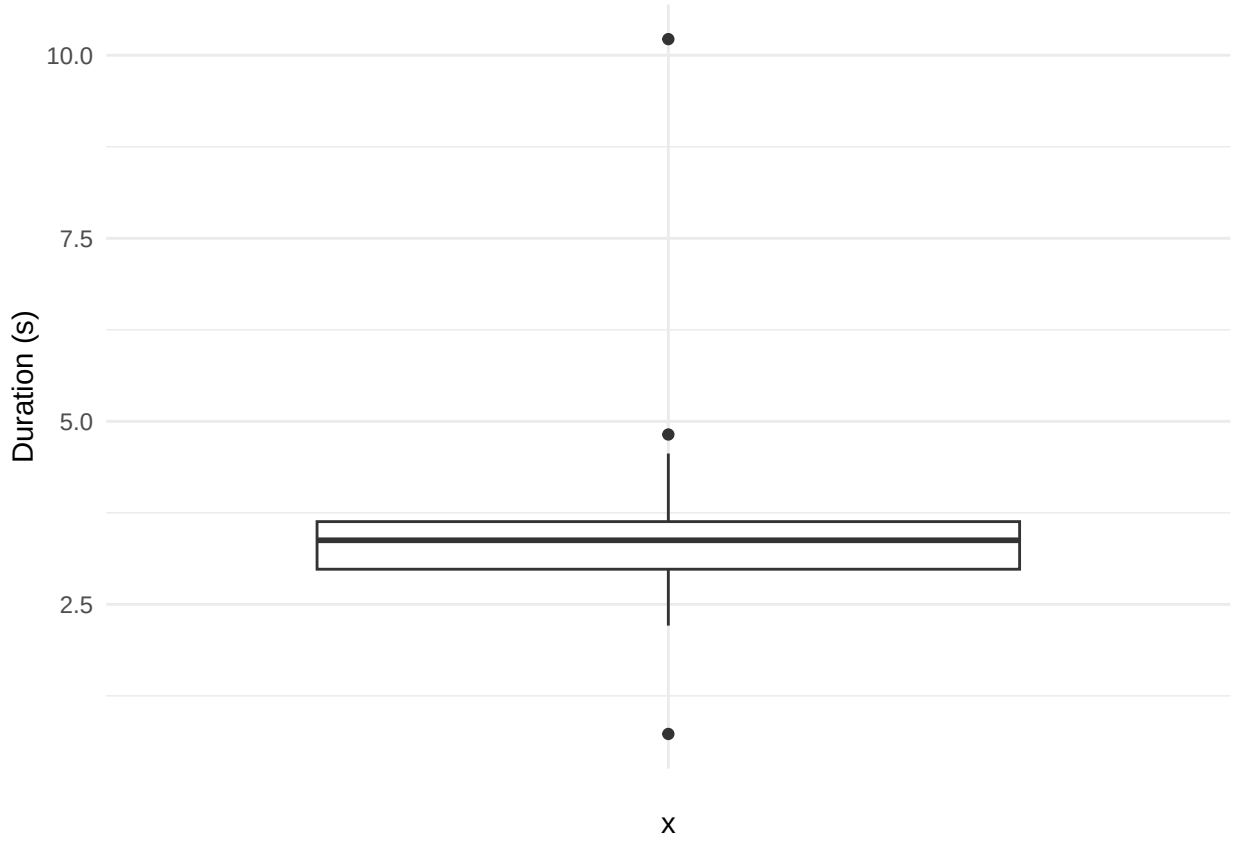
```
knitr::kable(head(data), align = "c")
```

Id	Name	Duration	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Danceability
1	Good 4 U Olivia Rodrigo	2.97	0.664	9	-5.044	1	0.1540	0.33500	0.00e+00	0.0849	0.688	166.928	0.563
2	Stay The Kid LAROI & Justin Bieber	2.30	0.506	8	-11.275	1	0.0589	0.37900	8.68e-01	0.1100	0.454	170.050	0.564
3	Levitating Dua Lipa feat. DaBaby	3.38	0.825	6	-3.787	0	0.0601	0.00883	0.00e+00	0.0674	0.915	102.977	0.702
4	Peaches Justin Bieber feat. Daniel Caesar & Giveon	3.30	0.696	0	-6.181	1	0.1190	0.32100	0.00e+00	0.4200	0.464	90.030	0.677

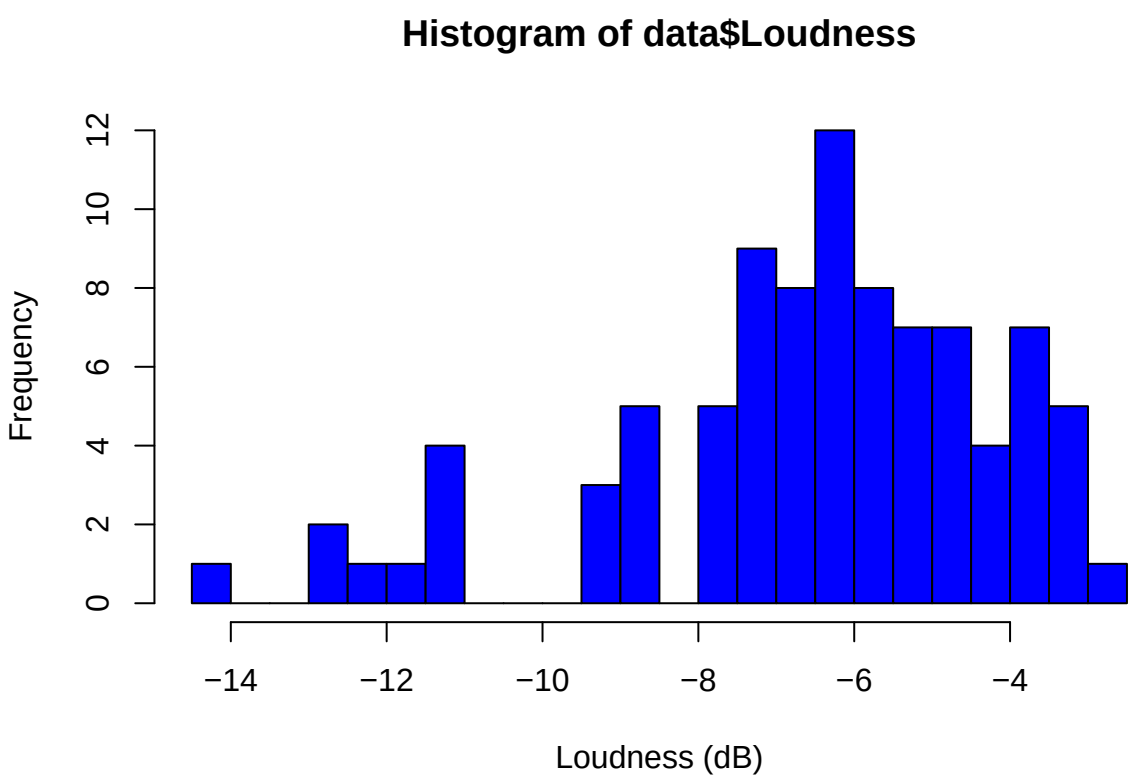
Id	Name	Duration	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Danceability
5	Montero (Call Me By Your Name) Lil Nas X	2.30	0.503	8	-6.725	0	0.2200	0.29300	0.00e+00	0.4050	0.710	178.781	0.593
6	Kiss Me More (feat. SZA) Doja Cat	3.48	0.705	8	-3.463	1	0.0284	0.25900	8.92e-05	0.1200	0.781	110.970	0.764

```
library(ggplot2)

ggplot(data, aes(x = "", y = Duration)) +
 geom_boxplot() +
  labs(y = "Duration (s)") +
  theme_minimal()
```




```
hist(data$Loudness, breaks = 20, col = "blue", xlab = "Loudness (dB)")
```

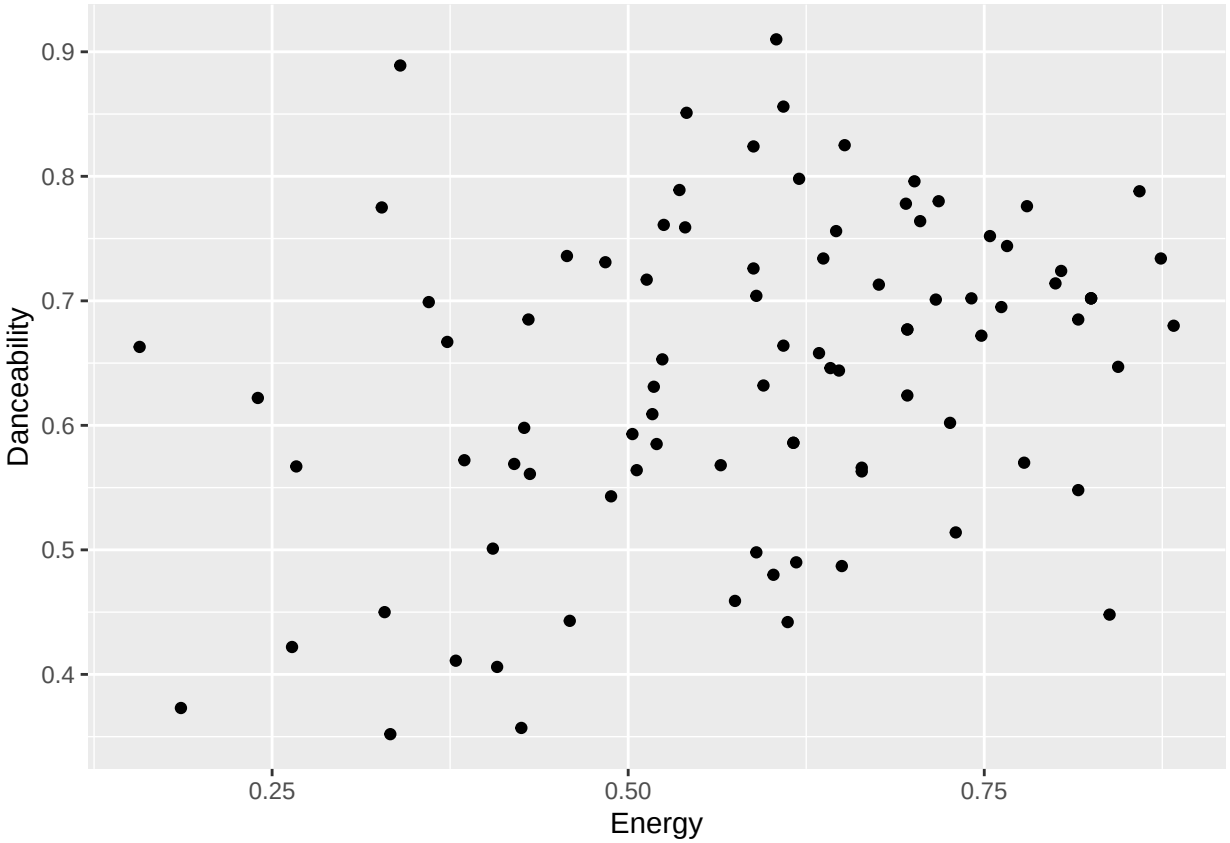



```
str(data)
```

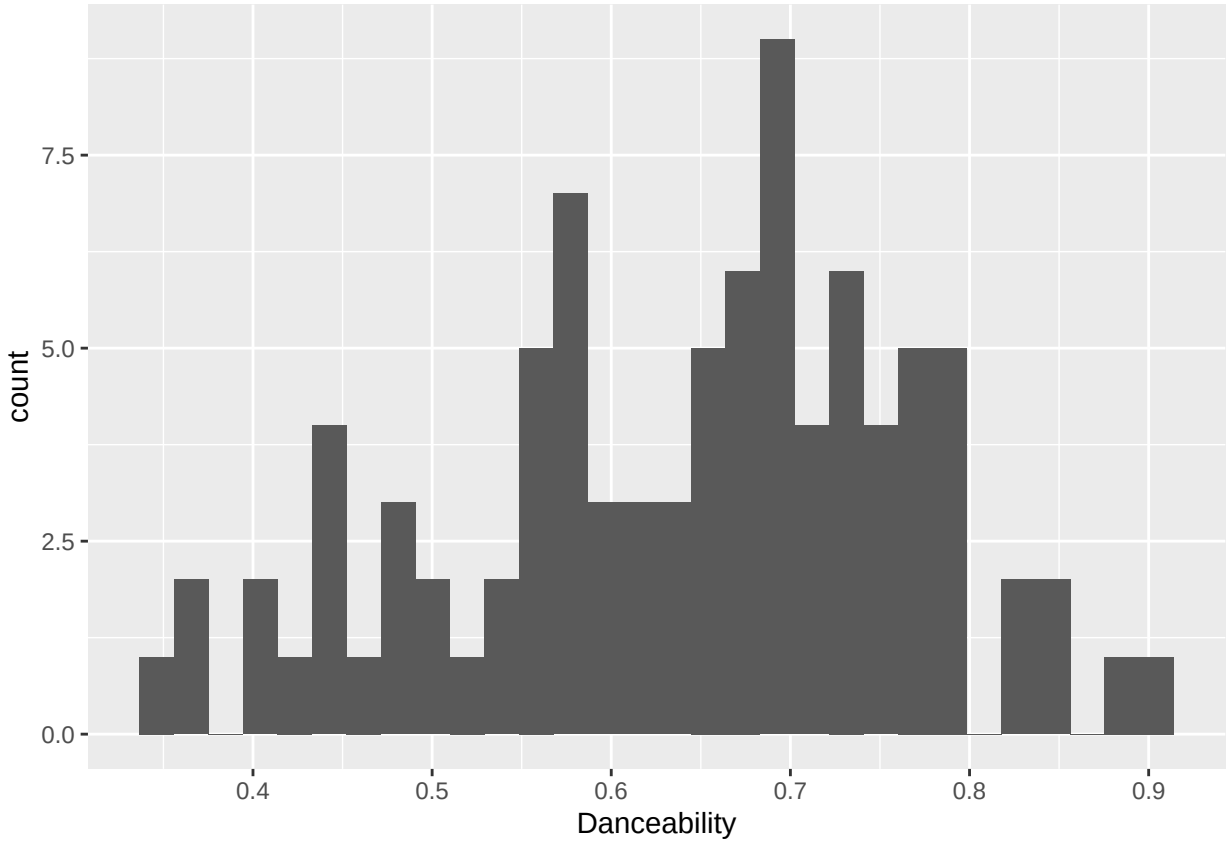
```
## tibble [90 x 14] (S3: tbl_df/tbl/data.frame)
##  $ Id              : num [1:90] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Name            : chr [1:90] "Good 4 U Olivia Rodrigo" "Stay The Kid LAROI & Justin Bieber" "Levi
##  $ Duration        : num [1:90] 2.97 2.3 3.38 3.3 2.3 3.48 4.03 4.03 3.33 3.59 ...
##  $ Energy          : num [1:90] 0.664 0.506 0.825 0.696 0.503 0.705 0.616 0.431 0.73 0.612 ...
##  $ Key             : num [1:90] 9 8 6 0 8 8 5 10 1 2 ...
##  $ Loudness        : num [1:90] -5.04 -11.28 -3.79 -6.18 -6.72 ...
##  $ Mode            : num [1:90] 1 1 0 1 0 1 1 1 1 1 ...
##  $ Speechiness     : num [1:90] 0.154 0.0589 0.0601 0.119 0.22 0.0284 0.0324 0.0578 0.0598 0.112 ...
##  $ Acousticness    : num [1:90] 0.335 0.379 0.00883 0.321 0.293 0.259 0.182 0.768 0.00146 0.584 ...
##  $ Instrumentalness: num [1:90] 0.00 8.68e-01 0.00 0.00 0.00 8.92e-05 0.00 1.42e-05 9.54e-05 5.70e-0
##  $ Liveness        : num [1:90] 0.0849 0.11 0.0674 0.42 0.405 0.12 0.0927 0.106 0.0897 0.37 ...
##  $ Valence         : num [1:90] 0.688 0.454 0.915 0.464 0.71 0.781 0.719 0.137 0.334 0.178 ...
##  $ Tempo           : num [1:90] 167 170 103 90 179 ...
##  $ Danceability    : num [1:90] 0.563 0.564 0.702 0.677 0.593 0.764 0.586 0.561 0.514 0.442 ...
```

Analisis Is there a relationship between certain characteristics of the songs, such as energy and danceability?

```
# Create a scatter plot of energy vs. danceability
ggplot(data, aes(x = Energy, y = Danceability)) +
  geom_point() +
  labs(x = "Energy", y = "Danceability")
```




```
# Create a histogram of danceability
ggplot(data, aes(x = Danceability)) +
  geom_histogram() +
  labs(x = "Danceability")
```




```
# Create a correlation matrix of all variables
correlation <- cor(data$Energy, data$Danceability)
print(correlation)
```

```
## [1] 0.3336112
```

A correlation coefficient of 0.3336112 suggests a weak positive correlation between the two variables being analyzed. This means that there is a tendency for higher values of one variable to be associated with higher values of the other variable, but the relationship is not very strong. In this case, it indicates that there may be some relationship between the energy and danceability of songs, but it is not a strong or definitive relationship.

```
library(tidyverse)
library(readr)
library(dplyr)
```

Load required packages

```
pharm<- read_csv("pharma spend by country.csv")
```

```
head(pharm)
```

load the data into R

```
## # A tibble: 6 x 7
##   LOCATION  TIME PC_HEALTHXP PC_GDP USD_CAP FLAG_CODES TOTAL_SPEND
##   <chr>    <dbl>       <dbl>  <dbl>   <dbl> <chr>            <dbl>
## 1 AUS       1971        16.0  0.727    35.7 <NA>              462.
## 2 AUS       1972        15.1  0.686    36.1 <NA>              475.
## 3 AUS       1973        15.1  0.681    39.9 <NA>              533.
## 4 AUS       1974        14.8  0.755    47.6 <NA>              653.
## 5 AUS       1975        11.8  0.682    47.6 <NA>              661.
## 6 AUS       1976        10.9  0.63     46.9 <NA>              658.
```

Tidying and transforming groups the data by LOCATION using the group_by() function, and then summarizes the number of observations for each country using the summarize() function with the n() function

```
pharm %>%
  group_by(LOCATION) %>%
  summarize(n_obs = n())
```

```
## # A tibble: 32 x 2
##    LOCATION n_obs
##    <chr>    <int>
##  1 AUS         44
##  2 AUT         26
##  3 BEL         41
##  4 CAN         45
##  5 CHE         31
##  6 CZE         26
##  7 DEU         45
##  8 DNK         36
##  9 ESP         32
## 10 EST         17
## # ... with 22 more rows
```

```
pharm%>%
  group_by(TIME) %>%
  summarize(n_obs = n())
```

```
## # A tibble: 47 x 2
##     TIME n_obs
##    <dbl> <int>
##  1  1970    11
##  2  1971    11
```

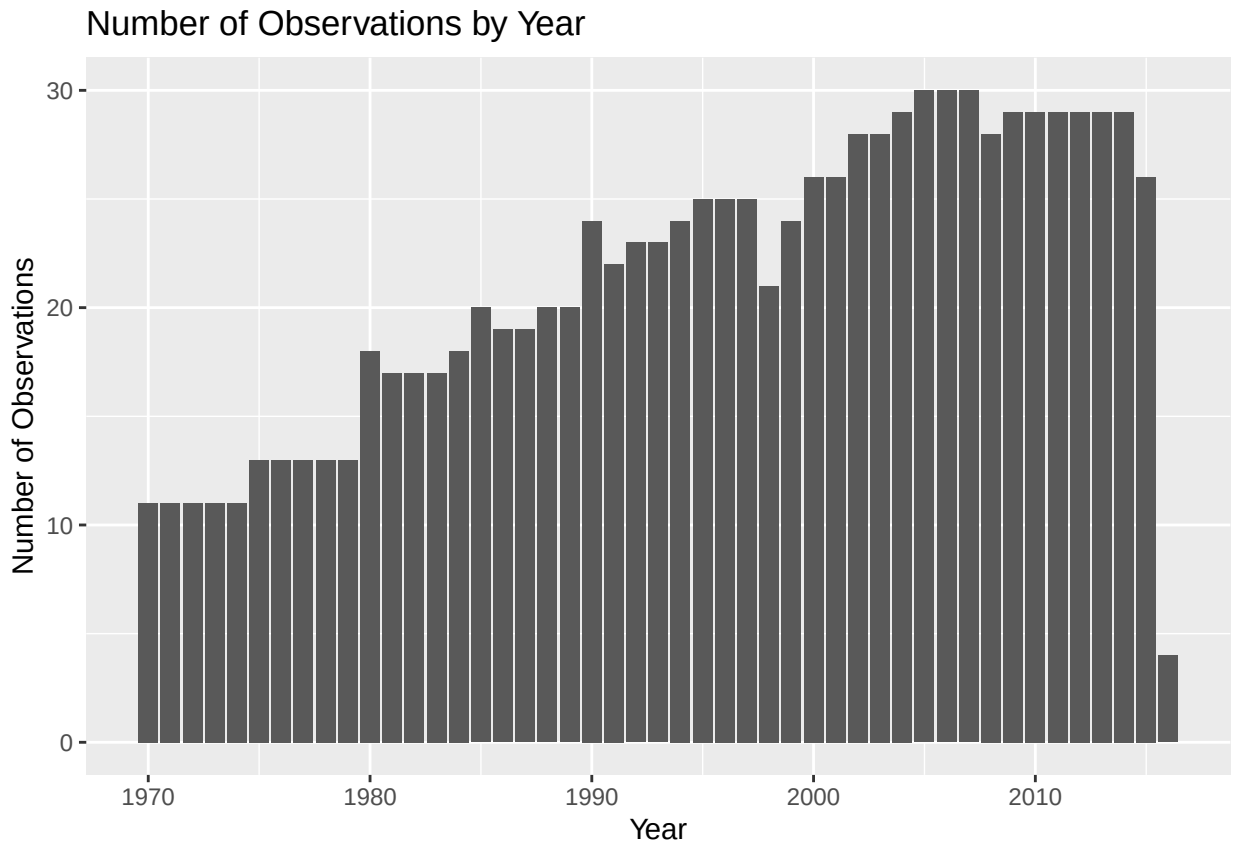
```
##  3  1972    11
##  4  1973    11
##  5  1974    11
##  6  1975    13
##  7  1976    13
##  8  1977    13
##  9  1978    13
## 10  1979    13
## # ... with 37 more rows
```

```
pharm%>%
  summarize(n_countries = n_distinct(LOCATION),
            n_years = n_distinct(TIME)) %>%
  bind_rows(pharm %>%
              group_by(LOCATION) %>%
              summarize(n_obs = n()) %>%
              summarize(n_countries = n_distinct(LOCATION),
                        n_obs = sum(n_obs)) %>%
              select(n_countries, n_obs)) %>%
  bind_rows(pharm %>%
              group_by(TIME) %>%
              summarize(n_obs = n()) %>%
              summarize(n_years = n_distinct(TIME),
                        n_obs = sum(n_obs)) %>%
              select(n_years, n_obs))
```

```
## # A tibble: 3 x 3
##   n_countries n_years n_obs
##         <int>   <int> <int>
## 1          32      47    NA
## 2          32      NA  1000
## 3          NA      47  1000
```

Some visualizations to explore the data

```
pharm %>%
  group_by(TIME) %>%
  summarize(n_obs = n()) %>%
  ggplot(aes(x = TIME, y = n_obs)) +
  geom_bar(stat = "identity") +
  labs(x = "Year", y = "Number of Observations",
       title = "Number of Observations by Year")
```

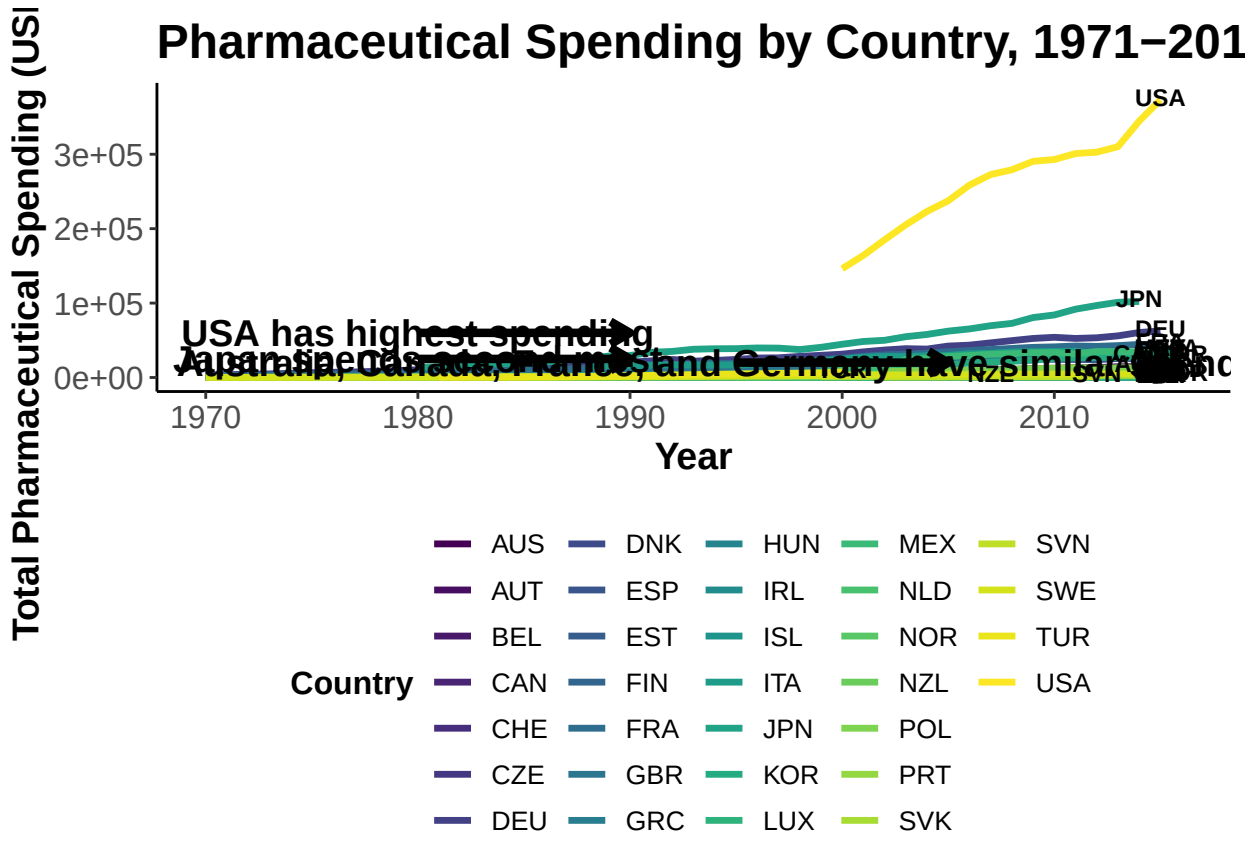



```
library(tidyverse)

for_label <- pharm %>%
  group_by(LOCATION) %>%
  summarize(year=max(TIME), totalspend=max(TOTAL_SPEND))

ggplot(data=pharm, aes(x=TIME, y=TOTAL_SPEND, group=LOCATION, color=LOCATION)) +
  geom_line(linewidth=1.2) +
  geom_text(data=for_label, aes(x=year, y=totalspend + 4000, label=LOCATION), size=3, color="black", fo
  scale_color_viridis_d() +
  labs(title="Pharmaceutical Spending by Country, 1971-2017",
       x="Year", y="Total Pharmaceutical Spending (USD billions)",
       color="Country") +
  theme_classic() +
  theme(plot.title = element_text(face="bold", size=18),
        axis.title = element_text(face="bold", size=14),
        axis.text = element_text(size=12),
        legend.position = "bottom",
        legend.title = element_text(face="bold", size=12),
        legend.text = element_text(size=10),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  annotate("text", x=1980, y=60000, label="USA has highest spending", size=5, color="black", fontface="
  annotate("segment", x=1980, xend=1990, y=60000, yend=60000, size=1.5, color="black", lineend="butt",
  annotate("text", x=1980, y=25000, label="Japan spends second most", size=5, color="black", fontface="
  annotate("segment", x=1980, xend=1990, y=25000, yend=25000, size=1.5, color="black", lineend="butt",
```

```
  annotate("text", x=1995, y=20000, label="Australia, Canada, France, and Germany have similar spending
  annotate("segment", x=1995, xend=2005, y=20000, yend=20000, size=1.5, color="black", lineend="butt",
```




```
library(dplyr)

summary_data <- pharm %>%
  group_by(LOCATION) %>%
  summarize(mean_spend = mean(TOTAL_SPEND),
            median_spend = median(TOTAL_SPEND),
            min_spend = min(TOTAL_SPEND),
            max_spend = max(TOTAL_SPEND),
            sd_spend = sd(TOTAL_SPEND))

head (summary_data)
```

```
## # A tibble: 6 x 6
##   LOCATION mean_spend median_spend min_spend max_spend sd_spend
##   <chr>         <dbl>        <dbl>     <dbl>     <dbl>    <dbl>
## 1 AUS           4786.        2765.      462.    14504.    4542.
## 2 AUT           3182.        3251.     1167.     5463.    1392.
## 3 BEL           3081.        2063.      405.     7655.    2425.
## 4 CAN           9892.        6974.      736.    27931.    9120.
## 5 CHE           3323.        2594.     1108.     8747.    2289.
## 6 CZE           3019.        3076.      994.     4659.    1159.
```

It is clear from the plots and the summary that the USA has consistently been the biggest pharmaceutical spender among the countries included in the dataset, with spending levels several times higher than other countries.