

# Data 607 Project 2

Jiaxin Zheng

2024-10-12

```
library(knitr)
library(stringr)
library(tidyr)
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
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v forcats   1.0.0      v purrr     1.0.2
## v ggplot2   3.5.1      v readr    2.1.5
## v lubridate 1.9.3      v tibble   3.2.1
```

```
## -- 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(ggplot2)
```

Data 1 First data is about the World Happiness Report in 2020. This data includes the Happiness Score for 153 countries. The Happiness Score is responses to the main life evaluation question asked in the Gallup World Poll(GWP). The Happiness Score is explained by the following factors: GDP per capita, Healthy Life Expectancy, Social support, Freedom to make life choices, Generosity, Corruption Perception, Residual error.

```
#view data 1
```

```
file1 <- "https://raw.githubusercontent.com/Jennyjjxxzz/Data-607_Project2/refs/heads/main/wide_data/Wor
df1 <- read.csv(file1)
head(df1)
```

##	Country.name	Regional.indicator	Ladder.score	Standard.error.of.ladder.score
## 1	Finland	Western Europe	7.8087	0.03115630
## 2	Denmark	Western Europe	7.6456	0.03349229
## 3	Switzerland	Western Europe	7.5599	0.03501417
## 4	Iceland	Western Europe	7.5045	0.05961586
## 5	Norway	Western Europe	7.4880	0.03483738
## 6	Netherlands	Western Europe	7.4489	0.02779175
##	upperwhisker	lowerwhisker	Logged.GDP.per.capita	Social.support
## 1	7.869766	7.747634	10.63927	0.9543297
## 2	7.711245	7.579955	10.77400	0.9559908
## 3	7.628528	7.491272	10.97993	0.9428466
## 4	7.621347	7.387653	10.77256	0.9746696
## 5	7.556281	7.419719	11.08780	0.9524866
## 6	7.503372	7.394428	10.81271	0.9391388
##	Healthy.life.expectancy	Freedom.to.make.life.choices	Generosity	
## 1	71.90083	0.9491722	-0.05948202	
## 2	72.40250	0.9514443	0.06620178	
## 3	74.10245	0.9213367	0.10591104	
## 4	73.00000	0.9488919	0.24694422	
## 5	73.20078	0.9557503	0.13453263	
## 6	72.30092	0.9085478	0.20761244	
##	Perceptions.of.corruption	Ladder.score.in.Dystopia		
## 1	0.1954446	1.972317		
## 2	0.1684895	1.972317		
## 3	0.3037284	1.972317		
## 4	0.7117097	1.972317		
## 5	0.2632182	1.972317		
## 6	0.3647171	1.972317		
##	Explained.by..Log.GDP.per.capita	Explained.by..Social.support		
## 1	1.285190	1.499526		
## 2	1.326949	1.503449		
## 3	1.390774	1.472403		
## 4	1.326502	1.547567		
## 5	1.424207	1.495173		
## 6	1.338946	1.463646		
##	Explained.by..Healthy.life.expectancy			
## 1	0.9612714			
## 2	0.9793326			
## 3	1.0405332			
## 4	1.0008434			
## 5	1.0080719			
## 6	0.9756753			
##	Explained.by..Freedom.to.make.life.choices	Explained.by..Generosity		
## 1	0.6623167	0.1596704		
## 2	0.6650399	0.2427934		
## 3	0.6289545	0.2690558		
## 4	0.6619807	0.3623302		
## 5	0.6702009	0.2879851		
## 6	0.6136265	0.3363176		
##	Explained.by..Perceptions.of.corruption	Dystopia...residual		
## 1	0.4778573	2.762835		
## 2	0.4952603	2.432741		
## 3	0.4079459	2.350267		
## 4	0.1445408	2.460688		

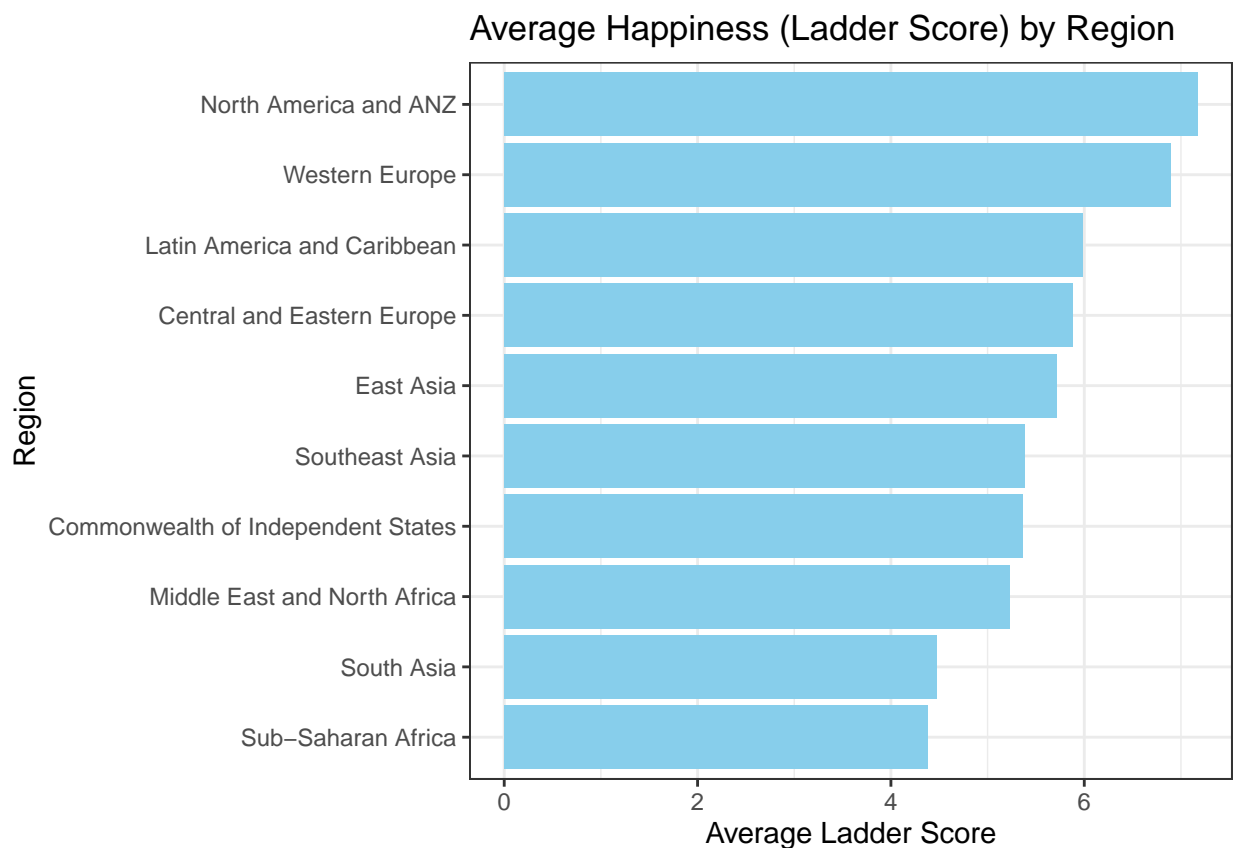
```
## 5          0.4341006          2.168266
## 6          0.3685698          2.352117
```

Question\_1: Which region has the highest average happiness (Ladder score)?

```
#group by and plot the average
```

```
region_happiness <- df1 %>%
  group_by(Regional.indicator) %>%
  summarize(avg_ladder_score = mean(Ladder.score, na.rm = TRUE)) %>%
  arrange(desc(avg_ladder_score))
```

```
ggplot(region_happiness, aes(x = reorder(Regional.indicator, avg_ladder_score), y = avg_ladder_score)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_flip() +
  labs(title = "Average Happiness (Ladder Score) by Region", x = "Region", y = "Average Ladder Score") +
  theme_bw()
```



Question\_2: Relationship between Ladder score and GDP, social support, and life expectancy

```
# Ladder score correlation analysis
```

```
ladder_score_correlations <- df1 %>%
  select(Ladder.score, Logged.GDP.per.capita, Social.support, Healthy.life.expectancy) %>%
  cor(use = "complete.obs")
```

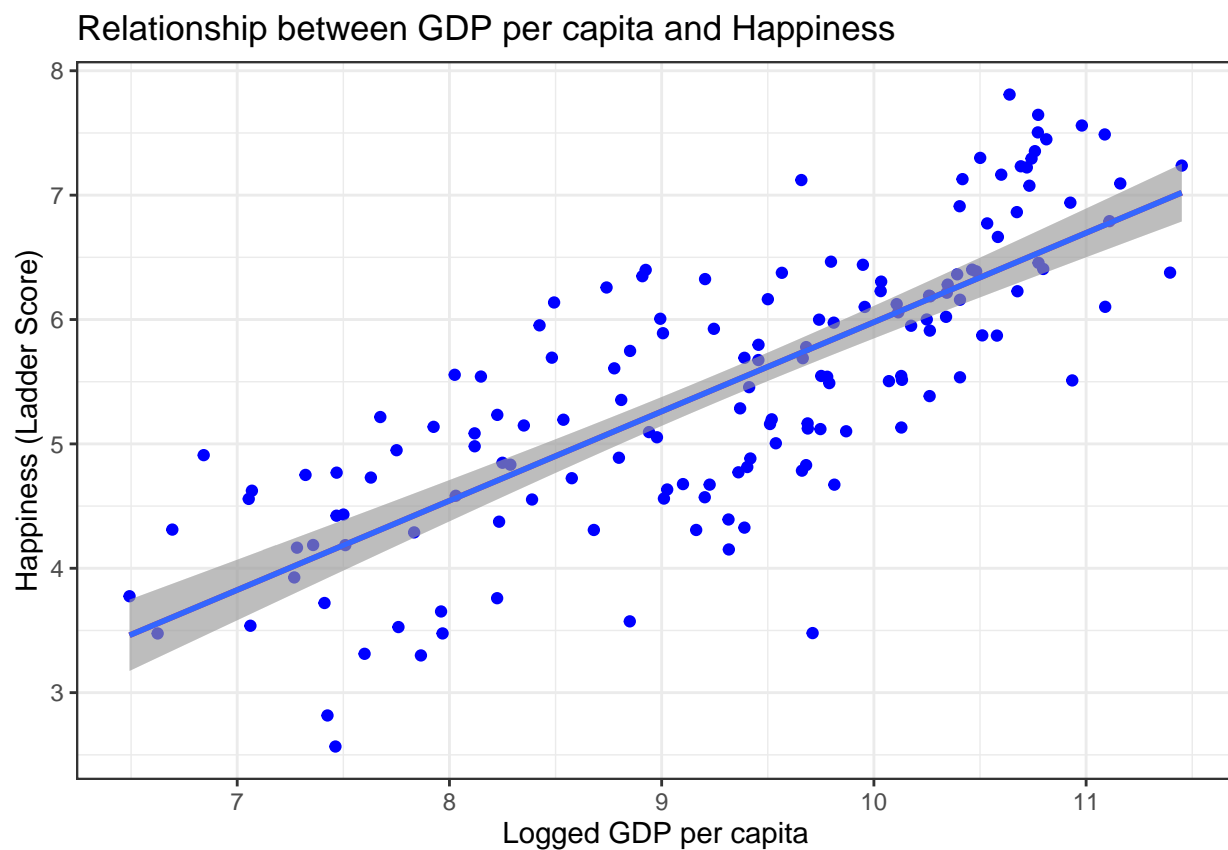
```
print(ladder_score_correlations)
```

```
##          Ladder.score Logged.GDP.per.capita Social.support
## Ladder.score          1.0000000          0.7753744          0.7650008
## Logged.GDP.per.capita  0.7753744          1.0000000          0.7818136
## Social.support         0.7650008          0.7818136          1.0000000
## Healthy.life.expectancy 0.7703163          0.8484686          0.7427441
##          Healthy.life.expectancy
## Ladder.score          0.7703163
## Logged.GDP.per.capita  0.8484686
## Social.support         0.7427441
## Healthy.life.expectancy 1.0000000
```

```
# Plot relationship between Ladder score and GDP per capita
```

```
ggplot(df1, aes(x = Logged.GDP.per.capita, y = Ladder.score)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", col = "red") +
  labs(title = "Relationship between GDP per capita and Happiness", x = "Logged GDP per capita", y = "Happiness (Ladder Score)") +
  geom_smooth(method = "lm") +
  theme_bw()
```

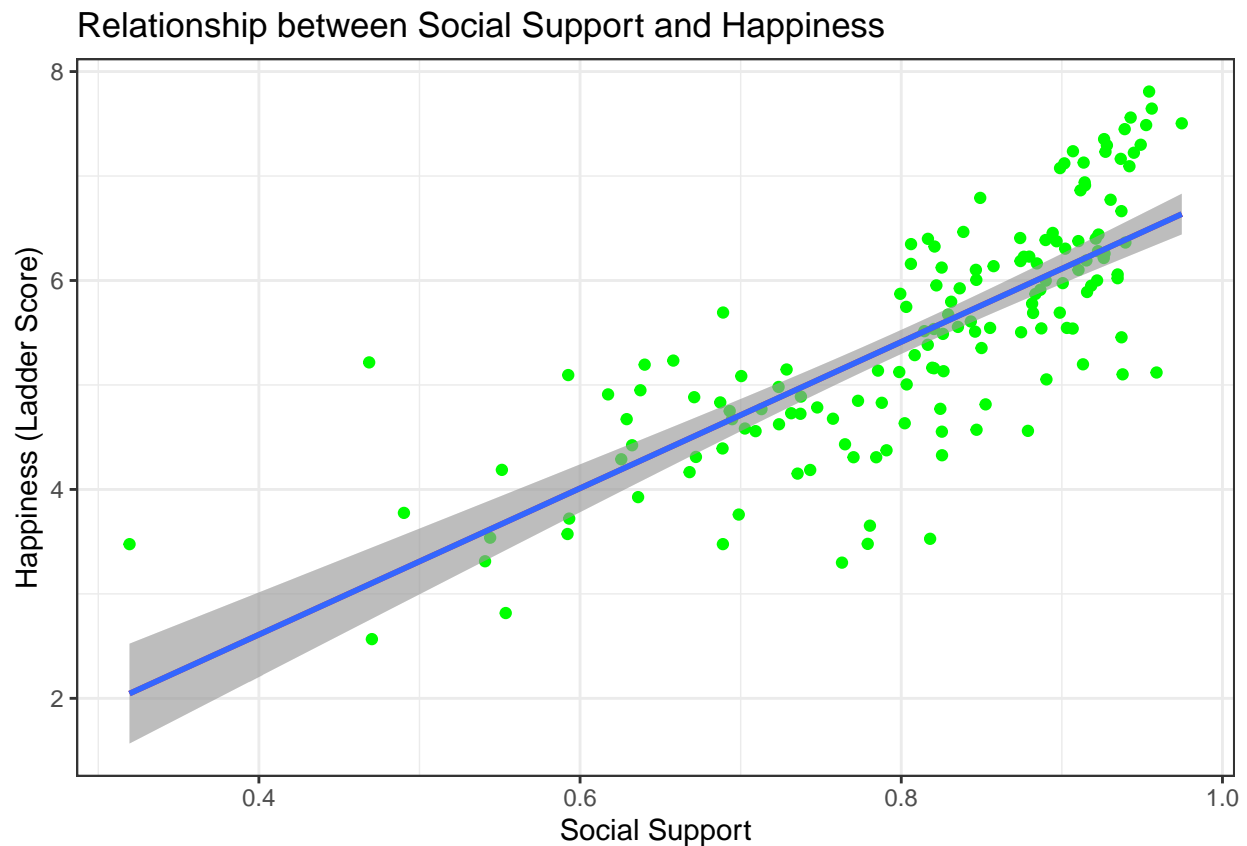
```
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
```



```
# Plot relationship between Ladder score and Social support
ggplot(df1, aes(x = Social.support, y = Ladder.score)) +
  geom_point(color = "green") +
```

```
geom_smooth(method = "lm", col = "red") +
labs(title = "Relationship between Social Support and Happiness", x = "Social Support", y = "Happiness")
geom_smooth(method = lm)+
theme_bw()
```

```
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
```



```
# Plot relationship between Ladder score and Life expectancy
ggplot(df1, aes(x = Healthy.life.expectancy, y = Ladder.score)) +
  geom_point(color = "orange") +
  geom_smooth(method = "lm", col = "red") +
  labs(title = "Relationship between Life Expectancy and Happiness", x = "Healthy Life Expectancy", y = "Happiness")
  geom_smooth(method = lm)+
  theme_bw()
```

```
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
```



Result for data 1: The North America and ANZ region ranks as the happiest region globally base on the highest average Ladder Score. Also the data shows a strong relationship between social support. Supported by their community elevated the happiness level.

Data 2 Second data show the ranking of the best universities of the world make by The Times Higher Education for 2020. The data frame consists of Rank char, Score Rank, University name, Country...

```
#view data 2
```

```
file2 <- "https://raw.githubusercontent.com/Jennyjxxzz/Data-607_Project2/refs/heads/main/wide_data/Wor
df2 <- read.csv(file2)
head(df2)
```

##	Rank_Char	Score_Rank	University	Country
## 1	1	1	University of Oxford	United Kingdom
## 2	2	2	California Institute of Technology	United States
## 3	3	3	University of Cambridge	United Kingdom
## 4	4	4	Stanford University	United States
## 5	5	5	Massachusetts Institute of Technology	United States
## 6	6	6	Princeton University	United States

##	Number_students	Numb_students_per_Staff	International_Students
## 1	20,664	11.2	41%
## 2	2,240	6.4	30%
## 3	18,978	10.9	37%
## 4	16,135	7.3	23%
## 5	11,247	8.6	34%

```
## 6          7,983          8.1          25%
## Percentage_Female Percentage_Male Teaching Research Citations Industry_Income
## 1          46%          54%      90.5      99.6      98.4      65.5
## 2          34%          66%      92.1      97.2      97.9      88.0
## 3          47%          53%      91.4      98.7      95.8      59.3
## 4          43%          57%      92.8      96.4      99.9      66.2
## 5          39%          61%      90.5      92.4      99.5      86.9
## 6          45%          55%      90.3      96.3      98.8      58.6
## International_Outlook Score_Result Overall_Ranking
## 1          96.4          95.4          95.40
## 2          82.5          94.5          94.50
## 3          95.0          94.4          94.40
## 4          79.5          94.3          94.30
## 5          89.0          93.6          93.60
## 6          81.1          93.2          93.20
```

Question\_1: Which country has the most universities in the top 100?

```
#filter the universities ranked in the top 100
df_top100 <- df2 %>%
  filter(as.numeric(Rank_Char) <= 100)
```

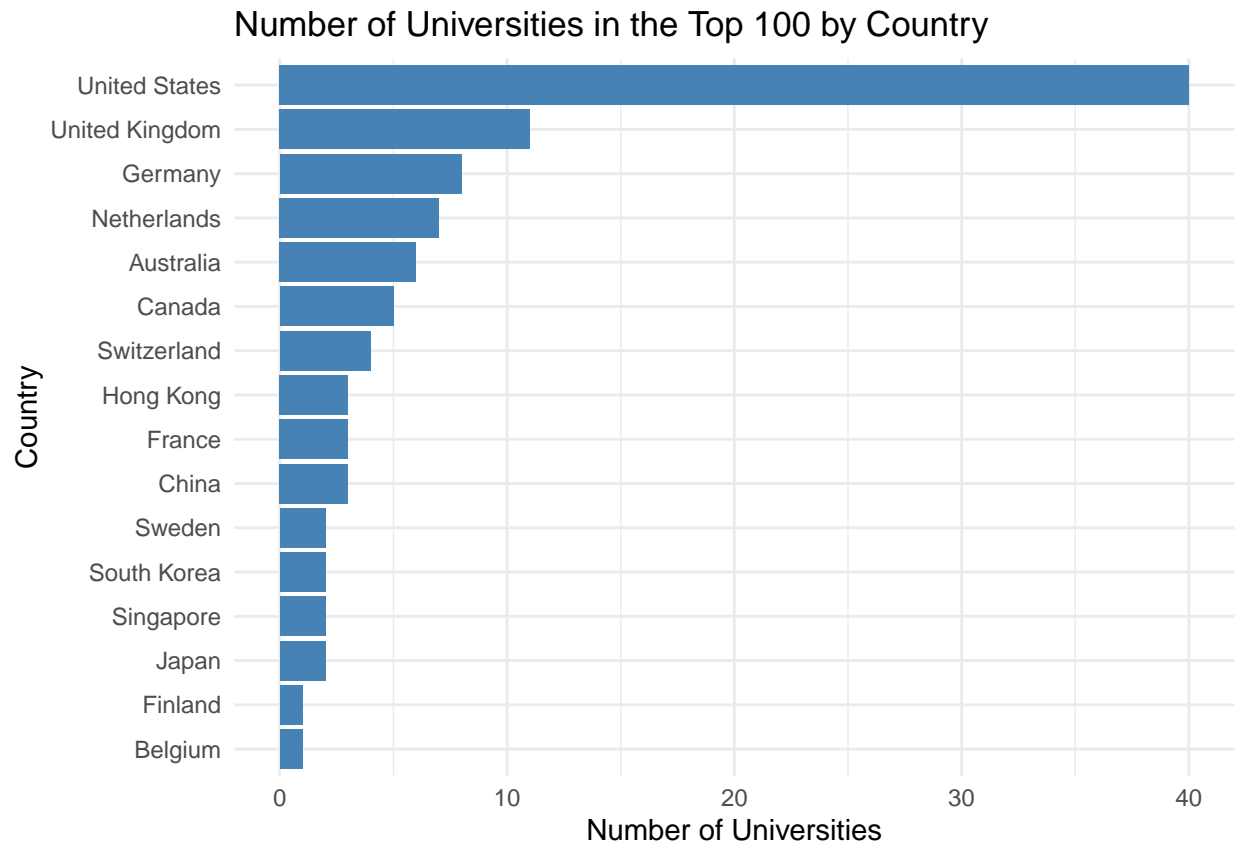
```
## Warning: There was 1 warning in 'filter()'.
## i In argument: 'as.numeric(Rank_Char) <= 100'.
## Caused by warning:
## ! NAs introduced by coercion
```

```
country_top100 <- df_top100 %>%
  group_by(Country) %>%
  summarize(university_count = n()) %>%
  arrange(desc(university_count))

print(country_top100)
```

```
## # A tibble: 16 x 2
##   Country      university_count
##   <chr>          <int>
## 1 United States      40
## 2 United Kingdom     11
## 3 Germany            8
## 4 Netherlands        7
## 5 Australia          6
## 6 Canada             5
## 7 Switzerland        4
## 8 China              3
## 9 France             3
## 10 Hong Kong         3
## 11 Japan             2
## 12 Singapore         2
## 13 South Korea       2
## 14 Sweden            2
## 15 Belgium           1
## 16 Finland           1
```

```
ggplot(country_top100, aes(x = reorder(Country, university_count), y = university_count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "Number of Universities in the Top 100 by Country",
       x = "Country",
       y = "Number of Universities") +
  theme_minimal()
```



Answer for question\_1: The United States has most number of universities in rank of top 100.

Data 3 Third data is about the population data from 2019 US Census, and also includes latitude and longitude data for each state's capital city.

```
#view data 3
```

```
file3 <- "https://raw.githubusercontent.com/Jennyjxxzz/Data-607_Project2/refs/heads/main/wide_data/2019_census_data.csv"
df3 <- read_csv(file3)
head(df3)
```

##	STATE	POPESTIMATE2019	lat	long
## 1	Alabama	4903185	32.37772	-86.30057
## 2	Alaska	731545	58.30160	-134.42021
## 3	Arizona	7278717	33.44814	-112.09696
## 4	Arkansas	3017804	34.74661	-92.28899
## 5	California	39512223	38.57667	-121.49363
## 6	Colorado	5758736	39.73923	-104.98486



Question\_1: Which states have the highest and lowest population estimates in 2019?

```
#The state with the highest population in 2019
```

```
highest_population_state <- df3 %>%  
  arrange(desc(POPESTIMATE2019)) %>%  
  slice(1)
```

```
print(highest_population_state)
```

```
##          STATE POPESTIMATE2019      lat      long  
## 1 California      39512223 38.57667 -121.4936
```

```
#The state with the lowest population in 2019
```

```
lowest_population_state <- df3 %>%  
  arrange(POPESTIMATE2019) %>%  
  slice(1)
```

```
print(lowest_population_state)
```

```
##          STATE POPESTIMATE2019      lat      long  
## 1 Wyoming      578759 41.14026 -104.8202
```