STAT 413/613 HW 1

Xubo Tang

Aug 30, 2020

# Instructions

Admin elements:  
1. Upload a photo (headshot) of yourself into your canvas profile  
2. Review the Syllabus and the [academic integrity code](https://www.american.edu/policies/students/upload/academic-integrity-code-2020.pdf).  
3. Fill in your information in the Student Info spreadsheet under the Canvas Collaboration site.

Analysis Elements: Rename the starter file under the analysis directory as hw\_01\_yourname.Rmd and use it for your solutions.  
1. Modify the “author” field in the YAML header.  
2. Stage and Commit R Markdown and HTML files (no PDF files).  
3. **Push both .Rmd and HTML files to GitHub**.  
- Make sure you have knitted to HTML prior to staging, committing, and pushing your final submission.  
4. **Commit each time you answer a part of question, e.g. 1.1**  
5. **Push to GitHub after each major question**, e.g., College Scorecard and World Bank Data  
- **Committing and Pushing are graded elements for this homework.**  
6. When complete, submit a response in Canvas

* Only include necessary code to answer the questions.
* Most of the functions you use should be from the tidyverse. Too much base R will result in point deductions.
* Use Pull requests and or email to ask me any questions. If you email, please ensure your most recent code is pushed to GitHub.

# Learning Outcomes:

* Operate with Git and GitHub.
* Apply concepts and methods from STAT 412/612.

# Canvas Picture, Syllabus, and Student Info

Review the Syllabus on Canvas and answer the following questions:

I, *Xubo Tang* have:

1. Added a photo of myself (headshot) to my Canvas profile
2. Reviewed the syllabus and the associated policies on the following date: Aug 29, 2020
3. Reviewed the American University policies on academic integrity, and understand how they apply to this course and agree to comply with them for this course
4. Filled in my information in the Student Info spreadsheet on Canvas collaborations

# College Scorecard

The data folder contains “college\_score\_200601.csv”, a subset of the data in the [College Scorecard](https://collegescorecard.ed.gov/data/) database as of June 1, 2020. These data contain information on colleges in the United States. The variables include:

* UNITID and OPEID: Identifiers for the colleges.
* INSTNM: Institution name
* ADM\_RATE: The Admission Rate.
* SAT\_AVE: Average SAT equivalent score of students admitted.
* UGDS: Enrollment of undergraduate certificate/degree-seeking students
* COSTT4\_A: Average cost of attendance (academic year institutions)
* AVGFACSAL: Average faculty salary
* GRAD\_DEBT\_MDN: The median debt for students who have completed
* AGE\_ENTRY: Average age of entry
* ICLEVEL: Level of institution (1 = 4-year, 2 = 2-year, 3 = less than 2-year).
* MN\_EARN\_WNE\_P10: Mean earnings of students working and not enrolled 10 years after entry.
* MD\_EARN\_WNE\_P10: Median earnings of students working and not enrolled 10 years after entry.
* FEMALE: Share of female students
* PCT\_WHITE: Percent of the population from students’ zip codes that is White, via Census data

1. Use a relative path and a readr function to load the data from data/college\_score\_200601.csv into a tibble.

library(tibble)  
library(ggplot2)  
library(ggthemes)  
library(tidyverse)

## -- Attaching packages ---------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v tidyr 1.1.1 v dplyr 1.0.2  
## v readr 1.3.1 v stringr 1.4.0  
## v purrr 0.3.4 v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

cs <- read.csv("~/hw-01-xubo1224/data/college\_score\_200601.csv", header = T)  
head(cs)

## ï..UNITID OPEID MN\_EARN\_WNE\_P10 MD\_EARN\_WNE\_P10  
## 1 100654 100200 35500 31000  
## 2 100663 105200 48400 41200  
## 3 100690 2503400 47600 39600  
## 4 100706 105500 52000 46700  
## 5 100724 100500 30600 27700  
## 6 100751 105100 51600 44500  
## INSTNM STABBR SAT\_AVG ADM\_RATE UGDS COSTT4\_A  
## 1 Alabama A & M University AL 957 0.8986 4990 22489  
## 2 University of Alabama at Birmingham AL 1220 0.9211 13186 24347  
## 3 Amridge University AL NULL NULL 351 17680  
## 4 University of Alabama in Huntsville AL 1314 0.8087 7458 23441  
## 5 Alabama State University AL 972 0.9774 3903 21476  
## 6 The University of Alabama AL 1252 0.5906 32177 29424  
## AVGFACSAL GRAD\_DEBT\_MDN AGE\_ENTRY FEMALE PCT\_WHITE ICLEVEL  
## 1 7101 34500 20.28374137 0.564030132 46.84000015 1  
## 2 10717 22500 23.60797466 0.63909074 69.01999664 1  
## 3 4292 25002 33.6722973 0.648648649 70.87999725 1  
## 4 9442 22021 22.72791963 0.476349937 76.37999725 1  
## 5 7754 32637 20.13099042 0.61341853 42.68999863 1  
## 6 10225 23250 21.1377014 0.615252417 75.34999847 1

1. If you used the default settings for reading in the data, 11 variables are probably type character when they should be numeric.

* Which ones?
* Why were they read in as type character?

# They are: MN\_EARN\_WNE\_P10, MD\_EARN\_WNE\_P10, SAT\_AVE, ADM\_RATE, UGDS, COSTT4\_A, GRAD\_DEBT\_MDN, AGE\_ENTRY, FEMALE, PCT\_WHITE, ICLEVEL.  
# Because these variables contain text like "NULL" or "PrivacySuppressed" in the column.

1. Fix these variables to be numeric in the tibble.

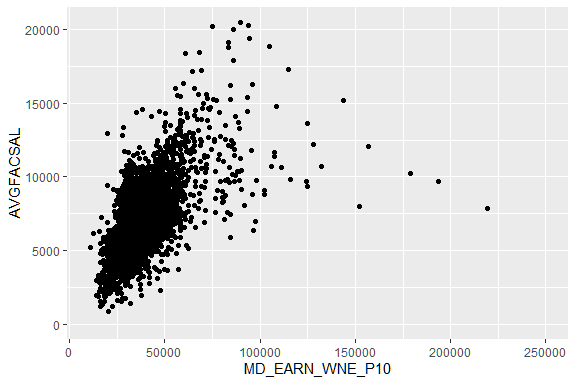
cs <- read.csv("~/hw-01-xubo1224/data/college\_score\_200601.csv", header = T, na.strings = c("NULL", "PrivacySuppressed"))  
head(cs)

## ï..UNITID OPEID MN\_EARN\_WNE\_P10 MD\_EARN\_WNE\_P10  
## 1 100654 100200 35500 31000  
## 2 100663 105200 48400 41200  
## 3 100690 2503400 47600 39600  
## 4 100706 105500 52000 46700  
## 5 100724 100500 30600 27700  
## 6 100751 105100 51600 44500  
## INSTNM STABBR SAT\_AVG ADM\_RATE UGDS COSTT4\_A  
## 1 Alabama A & M University AL 957 0.8986 4990 22489  
## 2 University of Alabama at Birmingham AL 1220 0.9211 13186 24347  
## 3 Amridge University AL NA NA 351 17680  
## 4 University of Alabama in Huntsville AL 1314 0.8087 7458 23441  
## 5 Alabama State University AL 972 0.9774 3903 21476  
## 6 The University of Alabama AL 1252 0.5906 32177 29424  
## AVGFACSAL GRAD\_DEBT\_MDN AGE\_ENTRY FEMALE PCT\_WHITE ICLEVEL  
## 1 7101 34500 20.28374 0.5640301 46.84 1  
## 2 10717 22500 23.60797 0.6390907 69.02 1  
## 3 4292 25002 33.67230 0.6486486 70.88 1  
## 4 9442 22021 22.72792 0.4763499 76.38 1  
## 5 7754 32637 20.13099 0.6134185 42.69 1  
## 6 10225 23250 21.13770 0.6152524 75.35 1

1. How is average faculty salary associated the median earnings of students ten years after initial enrollment? Create an appropriate plot and interpret the plot to justify your answer.

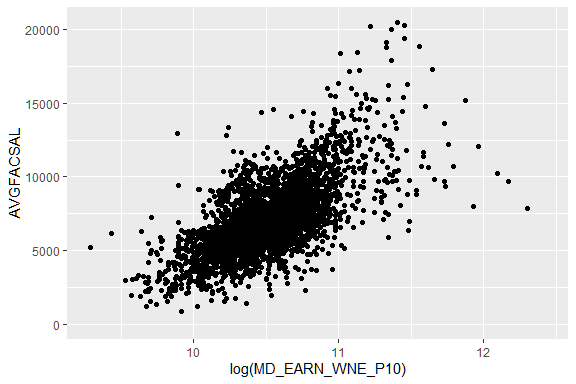
ggplot(data = cs)+  
 geom\_point(mapping = aes(x = MD\_EARN\_WNE\_P10, y = AVGFACSAL))

## Warning: Removed 3440 rows containing missing values (geom\_point).



ggplot(data = cs)+  
 geom\_point(mapping = aes(x = log(MD\_EARN\_WNE\_P10), y = AVGFACSAL))

## Warning: Removed 3440 rows containing missing values (geom\_point).

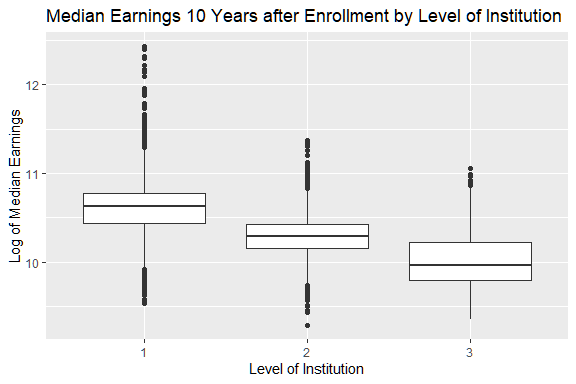


# According to the plot 1 above, we could find that there is a positive relation between average faculty salary and the median earnings of students ten years after initial enrollment. It could be more obvious if we log the median earnings of students ten years after initial enrollment.

1. Does the level of the institution seem to be associated with the median earnings of students ten years after enrollment? Reproduce this plot in R to explore this relationship and interpret the plot:

ggplot(data = cs)+  
 geom\_boxplot(mapping = aes(x = as.factor(ICLEVEL), y = log(MD\_EARN\_WNE\_P10)))+  
 labs(x = "Level of Institution", y = "Log of Median Earnings", title = "Median Earnings 10 Years after Enrollment by Level of Institution")

## Warning: Removed 1989 rows containing non-finite values (stat\_boxplot).



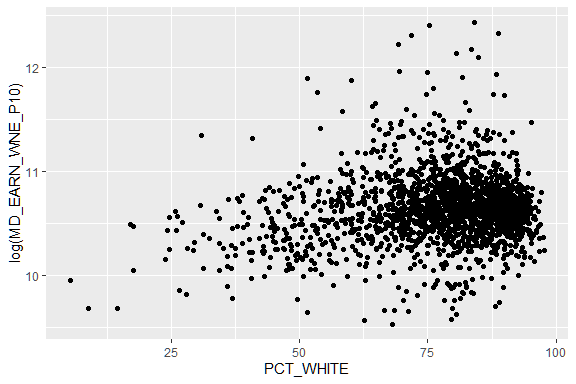
# It does show that the level of the institution is associated with the median earnings of students ten years after enrollment. Because the boxes of different levels didn't overlap a lot which means level leads to different earnings.

1. Plot the log of median earnings 10 years after enrollment for level 1 institutions as the Y axis against PCT\_WHITE and, in a second plot, against FEMALE.

* Describe the relationship if any in each of the plots.

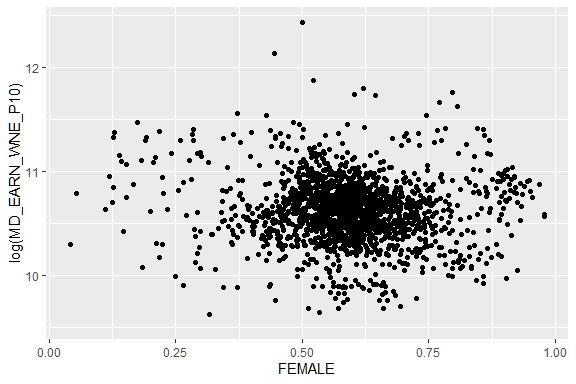
cs1 <- cs %>%  
 filter(ICLEVEL == 1)  
ggplot(data = cs1)+  
 geom\_point(mapping = aes(x = PCT\_WHITE, y = log(MD\_EARN\_WNE\_P10)))

## Warning: Removed 973 rows containing missing values (geom\_point).



ggplot(data = cs1)+  
 geom\_point(mapping = aes(x = FEMALE, y = log(MD\_EARN\_WNE\_P10)))

## Warning: Removed 798 rows containing missing values (geom\_point).



# According to the plot 1 above, we could find a weak positive realtionship between the log of median earnings 10 years after enrollment for level 1 institutions and PCT\_WHITE.  
# However, according to the plot 2 above, we didn't find obvious correlation between the log of median earnings 10 years after enrollment for level 1 institutions and FEMALE because the points scattered randomly and evenly.

1. Create a scatter plot of the log of mean earnings 10 years after enrollment (Y) axis) compared to the log of median earnings 10 years after enrollment (X axis).

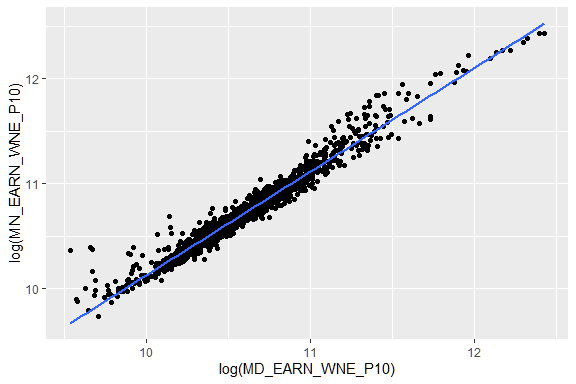
* Include an abline.
* Interpret the plot.

ggplot(data = cs1)+  
 geom\_point(mapping = aes(x = log(MD\_EARN\_WNE\_P10), y = log(MN\_EARN\_WNE\_P10)))+  
 geom\_smooth(method = lm, aes(x = log(MD\_EARN\_WNE\_P10), y = log(MN\_EARN\_WNE\_P10)))

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 715 rows containing non-finite values (stat\_smooth).

## Warning: Removed 715 rows containing missing values (geom\_point).



# According to the plot above, we could find that there is an extremely strong positive linear relationship between the log of mean earnings 10 years after enrollment and the log of median earnings 10 years after enrollment.

1. Compute a ranking of level 1 universities based on the ratio of median earnings 10 years after enrollment compared to median graduation debt.

* Identify the top 5 best and the bottom 5 worst?
* What is American University’s rank?
* Extra Credit:
  + Reproduce the following plot so the AU line adjusts as the data adjusts:
  + What is AU’s new ranking if the mean earnings are used?

rankcollege <- na.omit(cs1 %>%  
 mutate(RATIO\_EARNINGS\_DEBT = MD\_EARN\_WNE\_P10/GRAD\_DEBT\_MDN)%>%   
 select(INSTNM, RATIO\_EARNINGS\_DEBT)%>%  
 arrange(desc(RATIO\_EARNINGS\_DEBT)))  
head(rankcollege)

## INSTNM RATIO\_EARNINGS\_DEBT  
## 1 SUNY Downstate Health Sciences University 10.232000  
## 2 California Institute of Technology 9.873563  
## 3 Saint Augustine College 9.616088  
## 4 San Diego Mesa College 8.400000  
## 5 Massachusetts Institute of Technology 8.376000  
## 6 Stanford University 8.288511

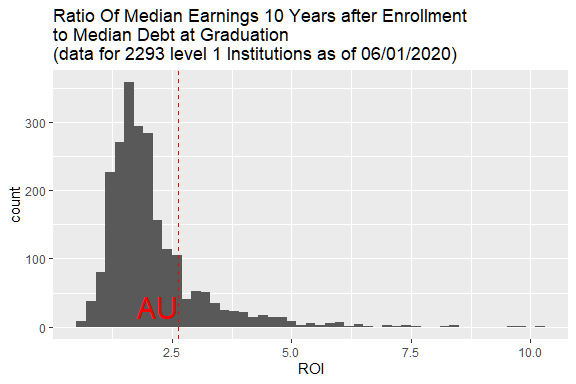
tail(rankcollege)

## INSTNM RATIO\_EARNINGS\_DEBT  
## 2288 Livingstone College 0.6685714  
## 2289 Southwest University of Visual Arts-Tucson 0.6535099  
## 2290 Southwest University of Visual Arts-Albuquerque 0.6535099  
## 2291 Benedict College 0.6350000  
## 2292 Messenger College 0.5313957  
## 2293 Martin University 0.5281276

rankcollege %>%  
 mutate(rank = order(-RATIO\_EARNINGS\_DEBT)) %>%  
 filter(INSTNM == "American University")

## INSTNM RATIO\_EARNINGS\_DEBT rank  
## 1 American University 2.619375 402

# top 5 best: SUNY Downstate Health Sciences University, California Institute of Technology, Saint Augustine College, Massachusetts Institute of Technology.  
# bottom 5 worst: Martin University, Messenger College, Benedict College, Southwest University of Visual Arts-Albuquerque, Southwest University of Visual Arts-Tucson.  
# The American University's rank is 402.  
  
# Extra Credit:  
ggplot(data = rankcollege)+  
 geom\_histogram(mapping = aes(x = RATIO\_EARNINGS\_DEBT),binwidth = 0.2)+  
 geom\_vline(aes(xintercept = 2.619375), colour = 'red', linetype = 'dashed')+  
 annotate("text", x = 2.2, y = 30, label = "AU", colour = 'red', size = 8)+  
 labs(x = "ROI", title = "Ratio Of Median Earnings 10 Years after Enrollment  
to Median Debt at Graduation   
(data for 2293 level 1 Institutions as of 06/01/2020)")



# World Bank Data

The World Bank provides loans to countries with the goal of reducing poverty. The dataframes in the data folder were taken from the public data repositories of the World Bank.

* country.csv: Contains information on the countries in the data set.
  + The variables are:
    - Country\_Code: A three-letter code for the country. Note not all rows are countries; some are regions.
    - Region: The region of the country.
    - IncomeGroup: Either "High income", "Upper middle income", "Lower middle income", or "Low income".
    - TableName: The full name of the country.
* fertility.csv: Contains the fertility rate information for each country for each year.
  + For the variables 1960 to 2017, the values in the cells represent the fertility rate in total births per woman for that year.
  + Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.
* life\_exp.csv: Contains the life expectancy information for each country for each year.
  + For the variables 1960 to 2017, the values in the cells represent life expectancy at birth in years for the given year.
  + Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
* population.csv: Contains the population information for each country.
  + For the variables 1960 to 2017, the values in the cells represent the total population in number of people for the given year.
  + Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.

1. Use relative paths and a readr function to load these files into four tibbles.

country <- read.csv("~/hw-01-xubo1224/data/country.csv", header = T, check.names = F, na.strings = "")  
fertility <- read.csv("~/hw-01-xubo1224/data/fertility.csv", header = T, check.names = F)  
life\_exp <- read.csv("~/hw-01-xubo1224/data/life\_exp.csv", header = T, check.names = F)  
population <- read.csv("~/hw-01-xubo1224/data/population.csv", header = T, check.names = F)

1. These data are messy. The observational units in fert, life, and pop are locations in space-time (e.g. Aruba in 2017). Recall tidy data should have one observational unit per row.

* Tidy these three tibbles.
* Make sure the variable for year is a numeric.

fertility1 <- fertility %>%  
 gather(key = "year", value = "fertility", 3:61) %>%  
 mutate(year = as.numeric(year))  
life\_exp1 <- life\_exp %>%  
 gather(key = "year", value = "life\_exp", 3:61) %>%  
 mutate(year = as.numeric(year))  
population1 <- population %>%  
 gather(key = "year", value = "population", 3:61) %>%  
 mutate(year = as.numeric(year))

1. Combine the tibbles to create a new tibble which includes the fertility rate, population, and life expectancy in each year as well as the region for each country.

country\_loan <- full\_join(country, full\_join(full\_join(life\_exp1, fertility1, by = c("Country Name", "Country Code", "year")), population1, by = c("Country Name", "Country Code", "year")), by = "Country Code")  
country\_loan <- na.omit(country\_loan)  
head(country\_loan)

## Country Code Region IncomeGroup TableName Country Name  
## 1 ABW Latin America & Caribbean High income Aruba Aruba  
## 2 ABW Latin America & Caribbean High income Aruba Aruba  
## 3 ABW Latin America & Caribbean High income Aruba Aruba  
## 4 ABW Latin America & Caribbean High income Aruba Aruba  
## 5 ABW Latin America & Caribbean High income Aruba Aruba  
## 6 ABW Latin America & Caribbean High income Aruba Aruba  
## year life\_exp fertility population  
## 1 1960 65.662 4.820 54211  
## 2 1961 66.074 4.655 55438  
## 3 1962 66.444 4.471 56225  
## 4 1963 66.787 4.271 56695  
## 5 1964 67.113 4.059 57032  
## 6 1965 67.435 3.842 57360

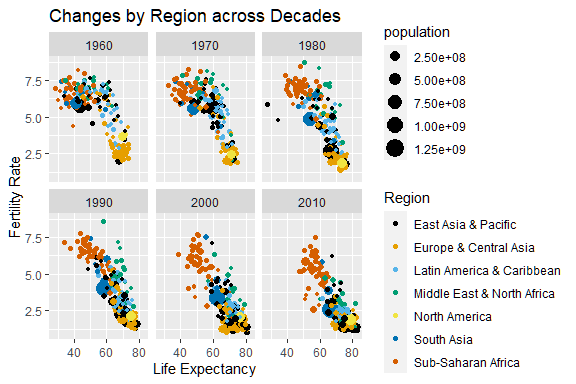
1. Make a scatterplot of fertility rate vs life expectancy, color-coding by region and annotating size by the population.

* Include only the years 1960, 1970, 1980, 1990, 2000, and 2010.
* Facet by these years.
* Your final plot should look like this (Each element of the formatting is graded):
* Hint: use ggthemes
* **Interpret the plot in one sentence**.

country\_loan1 <- country\_loan %>%  
 filter(year == 1960 | year==1970 | year == 1980 | year == 1990 | year == 2000 | year == 2010)  
head(country\_loan1)

## Country Code Region IncomeGroup TableName Country Name  
## 1 ABW Latin America & Caribbean High income Aruba Aruba  
## 2 ABW Latin America & Caribbean High income Aruba Aruba  
## 3 ABW Latin America & Caribbean High income Aruba Aruba  
## 4 ABW Latin America & Caribbean High income Aruba Aruba  
## 5 ABW Latin America & Caribbean High income Aruba Aruba  
## 6 ABW Latin America & Caribbean High income Aruba Aruba  
## year life\_exp fertility population  
## 1 1960 65.662 4.820 54211  
## 2 1970 69.140 2.908 59063  
## 3 1980 72.293 2.392 60096  
## 4 1990 73.468 2.249 62149  
## 5 2000 73.787 1.872 90853  
## 6 2010 75.016 1.776 101669

ggplot(data = country\_loan1)+  
 geom\_point(mapping = aes(x = life\_exp, y = fertility, color = Region, size = population))+  
 facet\_wrap(~year)+  
 labs(x = "Life Expectancy", y = "Fertility Rate", title = "Changes by Region across Decades")+  
 scale\_colour\_colorblind()



# There is a negative linear relationship between Fertility Rate and Life Expentancy and there is no big difference year by year but fertility decrease and life expectancy increase in general.   
# Representative region of high fertility and low life expectancy: Sub-Saharan Africa.  
# Representative region of low fertility and high life expectancy: Europe & Central Asia.

1. Calculate the total population for each region for each year.

* Exclude 2018.
* Make a line plot of year versus log of total population, color-coding by region.
* Your final plot should look like this:
* **Interpret the plot in one sentence**.

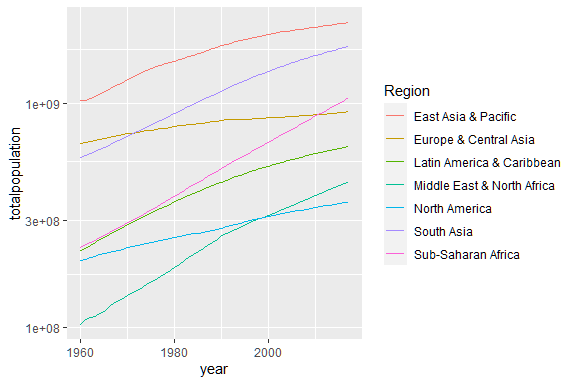
RegionPopulation <- country\_loan %>%  
 group\_by(Region, year) %>%  
 summarise(totalpopulation = sum(population))

## `summarise()` regrouping output by 'Region' (override with `.groups` argument)

head(RegionPopulation)

## # A tibble: 6 x 3  
## # Groups: Region [1]  
## Region year totalpopulation  
## <chr> <dbl> <dbl>  
## 1 East Asia & Pacific 1960 1029267702  
## 2 East Asia & Pacific 1961 1032458486  
## 3 East Asia & Pacific 1962 1046524748  
## 4 East Asia & Pacific 1963 1071884492  
## 5 East Asia & Pacific 1964 1096877540  
## 6 East Asia & Pacific 1965 1122925649

ggplot(data = RegionPopulation) +  
 geom\_line(mapping = aes(x = year, y = totalpopulation, group = Region, color = Region))+  
 scale\_y\_log10()

 6. Make a bar plot of population vs region for the year 2017.  
+ Order the bars on the -axis in **decreasing** order of population. + Your final plot should look like this:

population2017 <- RegionPopulation %>%  
 filter(year == 2017)  
ggplot(data = population2017) +  
 geom\_bar(mapping = aes(x = reorder(Region, -totalpopulation), y = totalpopulation), stat = "identity")+  
 coord\_flip()+  
 labs(x = "Region", y = "Total Population", title = "2017 Population by Region")

