Session 6: Homework 3

Leif Beckers, Dung Tran, Salman Abdullah, Andjela Bozinovic, Xiwen Wang

2020-10-10

Table of Contents

knitr::opts\_chunk$set(  
 message = FALSE,   
 warning = FALSE,   
 tidy=FALSE, # display code as typed  
 size="small") # slightly smaller font for code  
options(digits = 3)  
  
# default figure size  
knitr::opts\_chunk$set(  
 fig.width=6.75,   
 fig.height=6.75,  
 fig.align = "center"  
)

library(tidyverse) # Load ggplot2, dplyr, and all the other tidyverse packages  
library(mosaic)  
library(ggthemes)  
library(GGally)  
library(readxl)  
library(here)  
library(skimr)  
library(janitor)  
library(broom)  
library(tidyquant)  
library(infer)  
library(openintro)  
library(tidyquant)

# Youth Risk Behavior Surveillance

Every two years, the Centers for Disease Control and Prevention conduct the [Youth Risk Behavior Surveillance System (YRBSS)](https://www.cdc.gov/healthyyouth/data/yrbs/index.htm) survey, where it takes data from high schoolers (9th through 12th grade), to analyze health patterns. You will work with a selected group of variables from a random sample of observations during one of the years the YRBSS was conducted.

## Load the data

This data is part of the openintro textbook and we can load and inspect it. There are observations on 13 different variables, some categorical and some numerical. The meaning of each variable can be found by bringing up the help file:

?yrbss

data(yrbss)  
glimpse(yrbss)

## Rows: 13,583  
## Columns: 13  
## $ age <int> 14, 14, 15, 15, 15, 15, 15, 14, 15, 15, 15...  
## $ gender <chr> "female", "female", "female", "female", "f...  
## $ grade <chr> "9", "9", "9", "9", "9", "9", "9", "9", "9...  
## $ hispanic <chr> "not", "not", "hispanic", "not", "not", "n...  
## $ race <chr> "Black or African American", "Black or Afr...  
## $ height <dbl> NA, NA, 1.73, 1.60, 1.50, 1.57, 1.65, 1.88...  
## $ weight <dbl> NA, NA, 84.4, 55.8, 46.7, 67.1, 131.5, 71....  
## $ helmet\_12m <chr> "never", "never", "never", "never", "did n...  
## $ text\_while\_driving\_30d <chr> "0", NA, "30", "0", "did not drive", "did ...  
## $ physically\_active\_7d <int> 4, 2, 7, 0, 2, 1, 4, 4, 5, 0, 0, 0, 4, 7, ...  
## $ hours\_tv\_per\_school\_day <chr> "5+", "5+", "5+", "2", "3", "5+", "5+", "5...  
## $ strength\_training\_7d <int> 0, 0, 0, 0, 1, 0, 2, 0, 3, 0, 3, 0, 0, 7, ...  
## $ school\_night\_hours\_sleep <chr> "8", "6", "<5", "6", "9", "8", "9", "6", "...

Before you carry on with your analysis, it’s is always a good idea to check with skimr::skim() to get a feel for missing values, summary statistics of numerical variables, and a very rough histogram.

## Exploratory Data Analysis

You will first start with analyzing the weight of participants in kilograms. Using visualization and summary statistics, describe the distribution of weights. How many observations are we missing weights from?

weight\_distribution <- skim(yrbss$weight)  
  
# Calculate missing values in weights  
weight\_missing\_val <- yrbss %>%   
 select(weight) %>%   
 summarise(number\_of\_missing\_weight\_observations = sum(is.na(weight)))  
  
# Print tables of findings  
weight\_distribution

Data summary

|  |  |
| --- | --- |
| Name | yrbss$weight |
| Number of rows | 13583 |
| Number of columns | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

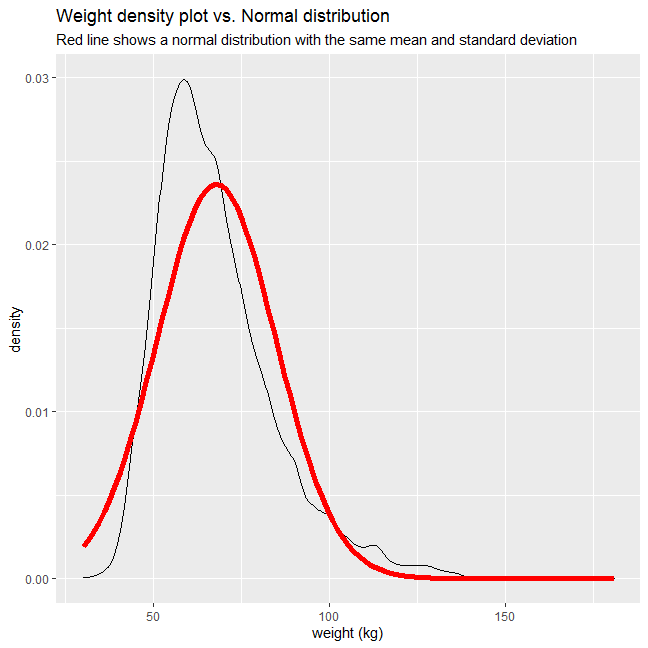
**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| data | 1004 | 0.93 | 67.9 | 16.9 | 29.9 | 56.2 | 64.4 | 76.2 | 181 | ▆▇▂▁▁ |

weight\_missing\_val

## # A tibble: 1 x 1  
## number\_of\_missing\_weight\_observations  
## <int>  
## 1 1004

# Create density Plot  
ggplot(yrbss, aes(x = weight)) +  
 geom\_density() +  
   
 # Add a red line showing normal distributrion for comparison  
 stat\_function(  
 fun = dnorm,  
 color = "red",  
 size = 2,  
 args = list(mean = mean(yrbss$weight, na.rm = TRUE),   
 sd = sd(yrbss$weight, na.rm = TRUE))) +  
   
 # Add titles  
 labs(title = "Weight density plot vs. Normal distribution",  
 subtitle ="Red line shows a normal distribution with the same mean and standard deviation",  
 x = "weight (kg)") +   
 NULL

 The data show that 1004 weight observations are missing. Also, based on the plot and the fact that mean weight is higher than median weight, we can conclude that the weight distribution is significantly right-skewed. This is most likely because there is a stricter lower bound to weights of a person.

Next, consider the possible relationship between a high schooler’s weight and their physical activity. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

Let’s create a new variable physical\_3plus, which will be yes if they are physically active for at least 3 days a week, and no otherwise.

yrbss <- yrbss %>%   
 mutate(physical\_3plus = ifelse(physically\_active\_7d >= 3, "yes", "no"))  
  
yrbss %>%   
 filter(!is.na(physical\_3plus)) %>%   
 group\_by(physical\_3plus) %>%   
 summarise(count = n()) %>%   
 mutate(prop= count/sum(count))

## # A tibble: 2 x 3  
## physical\_3plus count prop  
## <chr> <int> <dbl>  
## 1 no 4404 0.331  
## 2 yes 8906 0.669

Can you provide a 95% confidence interval for the population proportion of high schools that are *NOT* active 3 or more days per week?

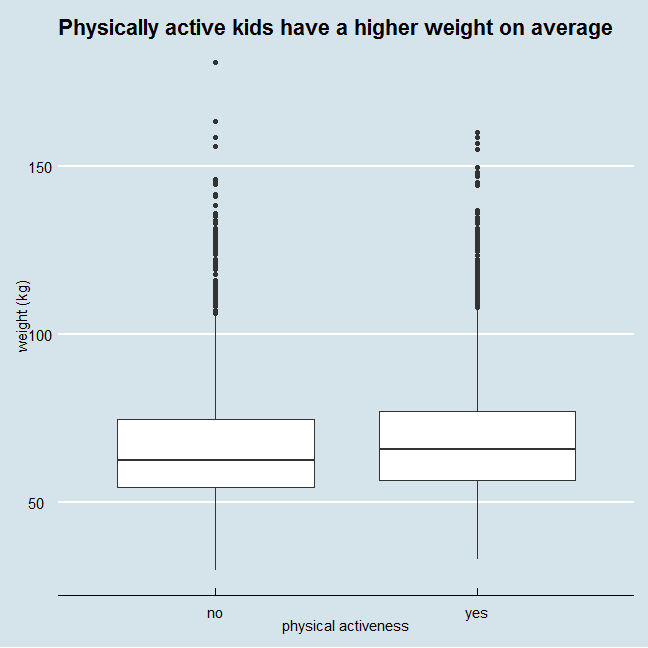
CI\_limits <- yrbss %>%   
   
 # clean NAs from weight and physical 3plus  
 drop\_na(physical\_3plus) %>%   
   
 # Seperate into 3plus yes and no category  
 group\_by(physical\_3plus) %>%   
   
 # COunt the number of active and non-active students  
 summarise(count = n()) %>%  
   
 # Calculate the proportion of each group (active and non-active), standard error, and confidence interval of each proportion  
 mutate(proportion = count/sum(count),  
 t\_critical = qt(0.975, sum(count)-1),  
 se\_proportion = sqrt(proportion \* (1 - proportion)/sum(count)),  
 margin\_of\_error = t\_critical \* se\_proportion,  
 proportion\_low = proportion - margin\_of\_error,  
 proportion\_high = proportion + margin\_of\_error) %>%  
   
 # select required columns only  
 select(physical\_3plus, proportion\_low, proportion\_high) %>%  
   
 # filter for non-active students only  
 filter(physical\_3plus == "no")  
  
CI\_limits

## # A tibble: 1 x 3  
## physical\_3plus proportion\_low proportion\_high  
## <chr> <dbl> <dbl>  
## 1 no 0.323 0.339

The 95% confidence interval for the population proportion of high school students that are *NOT* active 3 or more days per week is 0.317 - 0.334.

Make a boxplot of physical\_3plus vs. weight. Is there a relationship between these two variables? What did you expect and why?

yrbss %>%  
   
 # Drop observations where relevant entries are missing  
 drop\_na(physical\_3plus, weight) %>%  
   
 # Plot graph, stating variables for the x and y axis  
 ggplot(aes(x = physical\_3plus, y = weight)) +  
   
 # Format plot as boxplot  
 geom\_boxplot() +  
   
 # Choose theme  
 theme\_economist() +  
   
 # Add titles  
 labs(title = "Physically active kids have a higher weight on average",  
 y = "weight (kg)",  
 x = "physical activeness")

 The trend seems to be that physically active kids are likely to have higher weights than those who are not. In other words, there is a positive correlation between physical activity and weight. This is opposite to what was expected. I previously believed that since non-active kids exercise less, they burn less calories daily and hence should be heavier.

## Confidence Interval

Boxplots show how the medians of the two distributions compare, but we can also compare the means of the distributions using either a confidence interval or a hypothesis test. Note that when we calculate the mean/SD, etc weight in these groups using the mean function, we must ignore any missing values by setting the na.rm = TRUE.

yrbss %>%  
 group\_by(physical\_3plus) %>%  
 filter(!is.na(physical\_3plus)) %>%   
 summarise(mean\_weight = mean(weight, na.rm = TRUE),  
 sd\_weight = sd(weight, na.rm=TRUE),  
 count = n(),  
 se\_weight = sd\_weight/sqrt(count),  
 t\_critical = qt(0.975, count-1),   
 margin\_of\_error = t\_critical \* se\_weight,  
 lower = mean\_weight - t\_critical \* se\_weight,  
 upper = mean\_weight + t\_critical \* se\_weight  
 )

## # A tibble: 2 x 9  
## physical\_3plus mean\_weight sd\_weight count se\_weight t\_critical  
## <chr> <dbl> <dbl> <int> <dbl> <dbl>  
## 1 no 66.7 17.6 4404 0.266 1.96  
## 2 yes 68.4 16.5 8906 0.175 1.96  
## # ... with 3 more variables: margin\_of\_error <dbl>, lower <dbl>, upper <dbl>

There is an observed difference of about 1.77kg (68.44 - 66.67), and we notice that the two confidence intervals do not overlap. It seems that the difference is at least 95% statistically significant. Let us also conduct a hypothesis test.

## Hypothesis test with formula

Write the null and alternative hypotheses for testing whether mean weights are different for those who exercise at least times a week and those who don’t.

Null hypothesis: Difference in mean weights of physically active and non-active high school students is 0. Alternate hypothesis: Difference in mean weights of physically active and non-active high school students is not 0.

t.test(weight ~ physical\_3plus, data = yrbss)

##   
## Welch Two Sample t-test  
##   
## data: weight by physical\_3plus  
## t = -5, df = 7479, p-value = 9e-08  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -2.42 -1.12  
## sample estimates:  
## mean in group no mean in group yes   
## 66.7 68.4

This yields a t statistic of -5, and a p-value of 9e-08. Thus, at 95% level of confidence, we can reject the null hypothesis. In fact, based on calculated p-value, we can reject the null hypothesis even at 99.999% level of confidence. Thus, we accept the alternate hypothesis, that the mean weights are indeed different for those who exercise at least times a week and those who don’t.

## Hypothesis test with infer

Next, we will introduce a new function, hypothesize, that falls into the infer workflow. You will use this method for conducting hypothesis tests.

But first, we need to initialize the test, which we will save as obs\_diff.

obs\_diff <- yrbss %>%  
 specify(weight ~ physical\_3plus) %>%  
 calculate(stat = "diff in means", order = c("yes", "no"))

Notice how you can use the functions specify and calculate again like you did for calculating confidence intervals. Here, though, the statistic you are searching for is the difference in means, with the order being yes - no != 0.

After you have initialized the test, you need to simulate the test on the null distribution, which we will save as null.

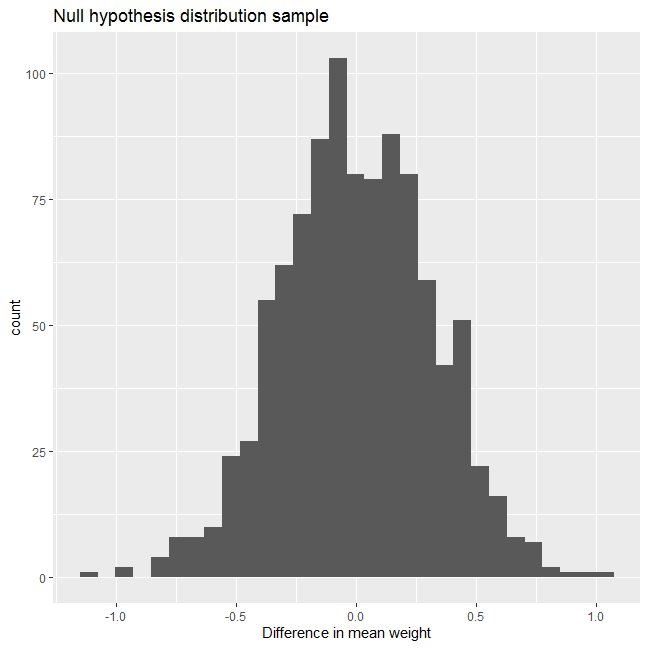
set.seed(5555)  
  
null\_dist <- yrbss %>%  
 specify(weight ~ physical\_3plus) %>%  
 hypothesize(null = "independence") %>%  
 generate(reps = 1000, type = "permute") %>%  
 calculate(stat = "diff in means", order = c("yes", "no"))

Here, hypothesize is used to set the null hypothesis as a test for independence, i.e., that there is no difference between the two population means. In one sample cases, the null argument can be set to *point* to test a hypothesis relative to a point estimate.

Also, note that the type argument within generate is set to permute, which is the argument when generating a null distribution for a hypothesis test.

We can visualize this null distribution with the following code:

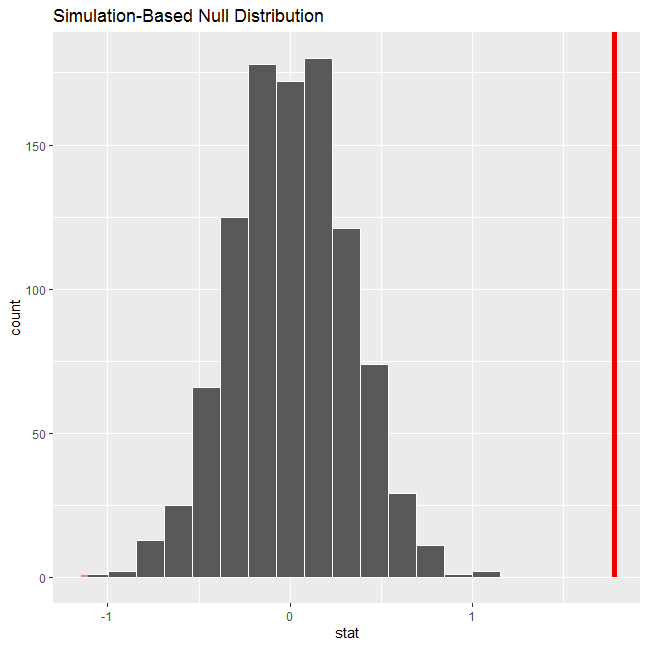
ggplot(data = null\_dist, aes(x = stat)) +  
 geom\_histogram() + # plot as histogram  
   
 # Add titles  
 labs(title = "Null hypothesis distribution sample",  
 x = "Difference in mean weight")



Now that the test is initialized and the null distribution formed, we can visualise to see how many of these null permutations have a difference of at least obs\_stat of 1.77?

We can also calculate the p-value for your hypothesis test using the function infer::get\_p\_value().

null\_dist %>% visualize() +  
 shade\_p\_value(obs\_stat = obs\_diff, direction = "two\_sided")



null\_dist %>%  
 get\_p\_value(obs\_stat = obs\_diff, direction = "two\_sided")

## # A tibble: 1 x 1  
## p\_value  
## <dbl>  
## 1 0

This the standard workflow for performing hypothesis tests.

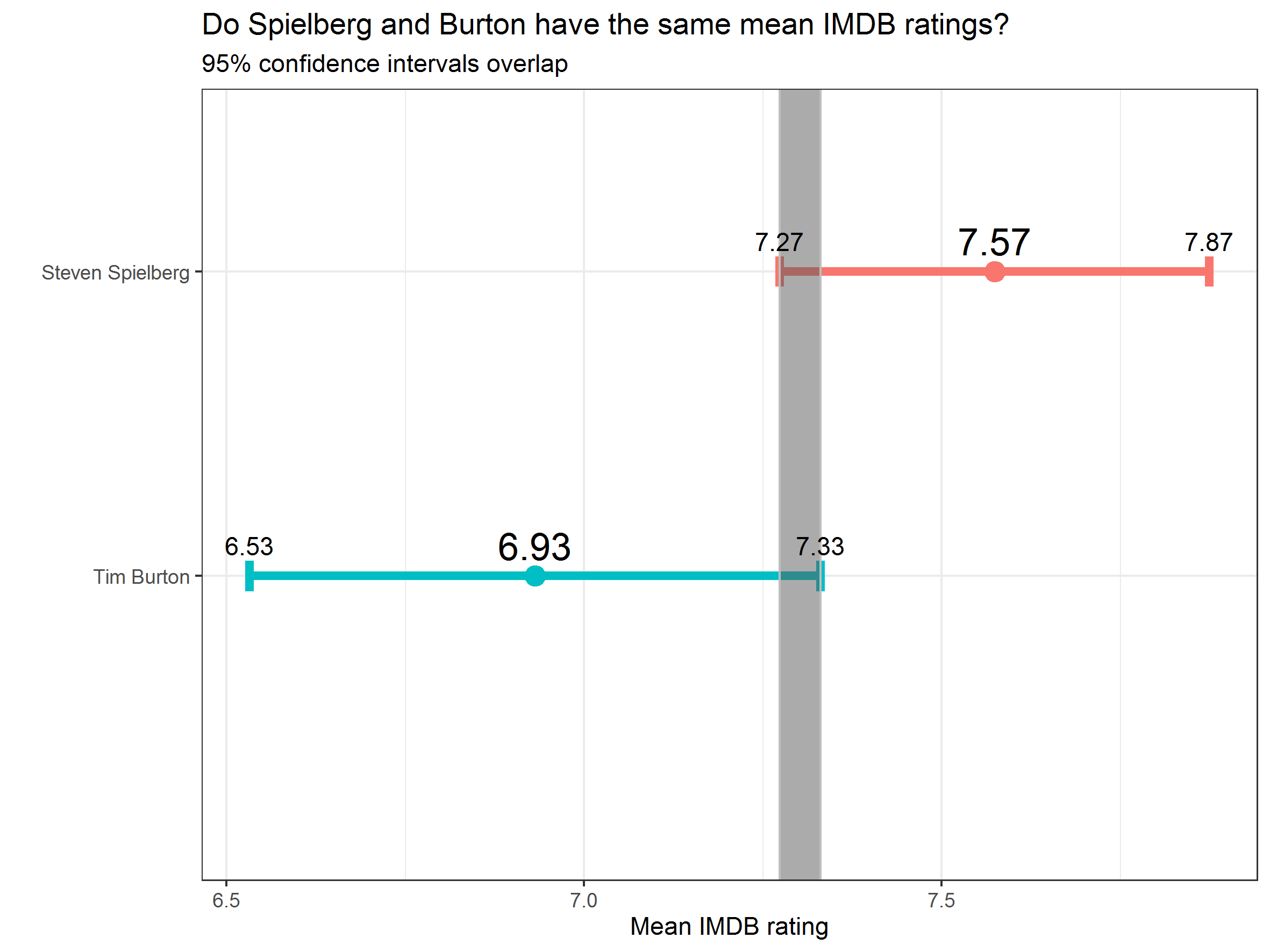
# IMDB ratings: Differences between directors

Recall the IMBD ratings data. I would like you to explore whether the mean IMDB rating for Steven Spielberg and Tim Burton are the same or not. I have already calculated the confidence intervals for the mean ratings of these two directors and as you can see they overlap. First, this graph is reproduced from the provided picture.

movies <- read\_csv(here::here("data", "movies.csv"))  
glimpse(movies)

## Rows: 2,961  
## Columns: 11  
## $ title <chr> "Avatar", "Titanic", "Jurassic World", "The Ave...  
## $ genre <chr> "Action", "Drama", "Action", "Action", "Action"...  
## $ director <chr> "James Cameron", "James Cameron", "Colin Trevor...  
## $ year <dbl> 2009, 1997, 2015, 2012, 2008, 1999, 1977, 2015,...  
## $ duration <dbl> 178, 194, 124, 173, 152, 136, 125, 141, 164, 93...  
## $ gross <dbl> 7.61e+08, 6.59e+08, 6.52e+08, 6.23e+08, 5.33e+0...  
## $ budget <dbl> 2.37e+08, 2.00e+08, 1.50e+08, 2.20e+08, 1.85e+0...  
## $ cast\_facebook\_likes <dbl> 4834, 45223, 8458, 87697, 57802, 37723, 13485, ...  
## $ votes <dbl> 886204, 793059, 418214, 995415, 1676169, 534658...  
## $ reviews <dbl> 3777, 2843, 1934, 2425, 5312, 3917, 1752, 1752,...  
## $ rating <dbl> 7.9, 7.7, 7.0, 8.1, 9.0, 6.5, 8.7, 7.5, 8.5, 7....

IMDB <- movies %>%   
   
 # filter for relevant directors  
 filter(director %in% c("Steven Spielberg", "Tim Burton")) %>%   
   
 # drop any potential NAs  
 drop\_na(rating) %>%   
   
 # group by director  
 group\_by(director) %>%   
   
 # calculate summary statistics  
 summarise(mean\_rating = mean(rating),   
 sd\_rating = sd(rating),  
 count = n(),  
 t\_critical = qt(0.975, count-1),   
 se\_rating = sd(rating)/sqrt(count),   
 margin\_of\_error = t\_critical \* se\_rating,  
 rating\_low = mean\_rating - margin\_of\_error,  
 rating\_high = mean\_rating + margin\_of\_error)  
  
# Calculate overlap points  
higher\_lower <- as.numeric(IMDB[1, 8])  
lower\_upper <- as.numeric(IMDB[2, 9])  
  
  
# Create graph  
IMDB\_plot <- ggplot(IMDB, aes(color = director)) +  
 geom\_errorbar(aes(x = reorder(director, mean\_rating),  
 ymin = rating\_low,   
 ymax = rating\_high),  
 width = 0.1,   
 size = 1.8) +  
   
 #Add grey shading for CI overlap  
 geom\_rect(aes(  
 ymin= higher\_lower,   
 ymax = lower\_upper,   
 xmin= 0,   
 xmax= Inf),  
 color = "grey",  
 alpha = 0.3) +   
   
 # add mean point  
 geom\_point(aes(x = director,  
 y = mean\_rating),  
 size = 4) +  
 # add data points to graph  
 geom\_text(aes(label = round(rating\_low,2), x = director, y = rating\_low),  
 nudge\_x = 0.1,  
 size = 4,  
 color = "Black") +  
 geom\_text(aes(label = round(rating\_high,2), x = director, y = rating\_high),  
 nudge\_x = 0.1,  
 size = 4,  
 color = "Black") +  
 geom\_text(aes(label = round(mean\_rating,2), x = director, y = mean\_rating),  
 nudge\_x = 0.1,  
 size = 6,  
 color = "Black") +  
   
 # flip graph  
 coord\_flip() +  
   
 # Choose labels  
 labs(title = "Do Spielberg and Burton have the same mean IMDB ratings?",  
 subtitle ="95% confidence intervals overlap",  
 x = "",  
 y = "Mean IMDB rating") +  
   
 # Choose theme  
 theme\_bw() +  
   
 # Hide legend  
 theme(legend.position = "none")  
  
# Save graph as custom format  
ggsave("IMDB\_graph.png",  
 plot = last\_plot(),  
 scale = 1,  
 width = 20,  
 height = 15,  
 units = "cm",  
 dpi = 300,  
 limitsize = TRUE)  
  
knitr::include\_graphics(here::here("IMDB\_graph.png"), error = FALSE)



In addition, you will run a hypothesis test. You should use both the t.test command and the infer package to simulate from a null distribution, where you assume zero difference between the two.

Before anything, write down the null and alternative hypotheses, as well as the resulting test statistic and the associated t-stat or p-value. At the end of the day, what do you conclude?

Null hypothesis: The difference in mean ratings between Spielberg and Burton is 0 Alternative hypothesis: The difference in mean ratings between Spielberg and Burton is not 0

movies2 <- movies %>%  
 filter(director %in% c("Steven Spielberg", "Tim Burton")) %>%  
 drop\_na(rating)  
  
# hypothesis testing using t.test()  
t.test(rating ~ director, data = movies2)

##   
## Welch Two Sample t-test  
##   
## data: rating by director  
## t = 3, df = 31, p-value = 0.01  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.16 1.13  
## sample estimates:  
## mean in group Steven Spielberg mean in group Tim Burton   
## 7.57 6.93

# hypothesis testing using infer package  
rating\_diff <- movies2 %>%  
 specify(rating ~ director) %>%  
 calculate(stat = "diff in means", order = c("Steven Spielberg", "Tim Burton"))  
  
set.seed(1234)  
  
 # randomise sample  
null\_dist\_imdb <- movies2 %>%  
 specify(rating ~ director) %>%  
 hypothesize(null = "independence") %>%  
 generate(reps = 1000, type = "permute") %>%  
 calculate(stat = "diff in means", order = c("Steven Spielberg", "Tim Burton"))  
  
 # calculate p-value  
null\_dist\_imdb %>%  
 get\_p\_value(obs\_stat = rating\_diff, direction = "two\_sided")

## # A tibble: 1 x 1  
## p\_value  
## <dbl>  
## 1 0.008

t.test() result: t-stat = 3 p-value = 0.01

infer package result: p-value = 0.008

Thus, for both methods, at 95% level of confidence, we reject the null hypothesis, and accept the alternative hypothesis that there is a difference in mean ratings between the 2 directors.

# Omega Group plc- Pay Discrimination

At the last board meeting of Omega Group Plc., the headquarters of a large multinational company, the issue was raised that women were being discriminated in the company, in the sense that the salaries were not the same for male and female executives. A quick analysis of a sample of 50 employees (of which 24 men and 26 women) revealed that the average salary for men was about 8,700 higher than for women. This seemed like a considerable difference, so it was decided that a further analysis of the company salaries was warranted.

You are asked to carry out the analysis. The objective is to find out whether there is indeed a significant difference between the salaries of men and women, and whether the difference is due to discrimination or whether it is based on another, possibly valid, determining factor.

## Loading the data

omega <- read\_csv(here::here("data", "omega.csv"))  
glimpse(omega) # examine the data frame

## Rows: 50  
## Columns: 3  
## $ salary <dbl> 81894, 69517, 68589, 74881, 65598, 76840, 78800, 70033, ...  
## $ gender <chr> "male", "male", "male", "male", "male", "male", "male", ...  
## $ experience <dbl> 16, 25, 15, 33, 16, 19, 32, 34, 1, 44, 7, 14, 33, 19, 24...

## Relationship Salary - Gender ?

The data frame omega contains the salaries for the sample of 50 executives in the company. Can you conclude that there is a significant difference between the salaries of the male and female executives?

Note that you can perform different types of analyses, and check whether they all lead to the same conclusion

. Confidence intervals . Hypothesis testing . Correlation analysis . Regression

Calculate summary statistics on salary by gender. Also, create and print a dataframe where, for each gender, you show the mean, SD, sample size, the t-critical, the SE, the margin of error, and the low/high endpoints of a 95% condifence interval

# Summary Statistics of salary by gender  
mosaic::favstats (salary ~ gender, data=omega)

## gender min Q1 median Q3 max mean sd n missing  
## 1 female 47033 60338 64618 70033 78800 64543 7567 26 0  
## 2 male 54768 68331 74675 78568 84576 73239 7463 24 0

# Dataframe with two rows (male-female) and having as columns gender, mean, SD, sample size,   
gender1 <- omega %>%  
 group\_by(gender) %>%  
 summarise(mean\_pay = mean(salary),  
 sd\_pay = sd(salary),  
 count = n(),  
  
# the t-critical value, the standard error, the margin of error,  
 t\_critical = qt(0.975, count-1),   
 se\_pay = sd(salary)/sqrt(count),   
 margin\_of\_error = t\_critical \* se\_pay,  
  
# and the low/high endpoints of a 95% condifence interval  
 CI\_low = mean\_pay - margin\_of\_error,  
 CI\_high = mean\_pay + margin\_of\_error)  
  
gender1

## # A tibble: 2 x 9  
## gender mean\_pay sd\_pay count t\_critical se\_pay margin\_of\_error CI\_low CI\_high  
## <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 female 64543. 7567. 26 2.06 1484. 3056. 61486. 67599.  
## 2 male 73239. 7463. 24 2.07 1523. 3151. 70088. 76390.

What can you conclude from your analysis? A couple of sentences would be enough

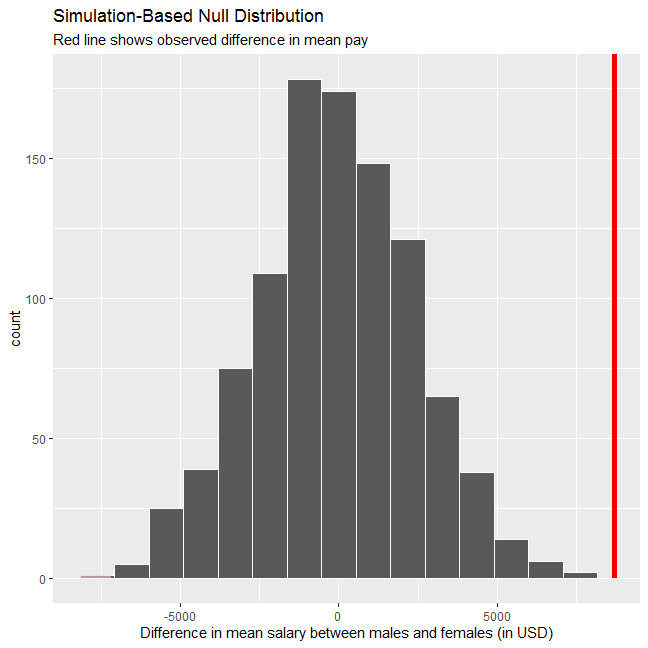
The mean pay of males is significantly higher than that of females. Furthermore, the 95% confidence intervals of their respective mean pay also do not overlap. Thus, at least at the 95% level of confidence, the data suggest that the difference in mean pay between the two genders are statistically significant. In other word, it is very likely true that males indeed do earn more than females.

You can also run a hypothesis testing, assuming as a null hypothesis that the mean difference in salaries is zero, or that, on average, men and women make the same amount of money. You should tun your hypothesis testing using t.test() and with the simulation method from the infer package.

# hypothesis testing using t.test()  
t.test(salary ~ gender, data = omega)

##   
## Welch Two Sample t-test  
##   
## data: salary by gender  
## t = -4, df = 48, p-value = 2e-04  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -12973 -4420  
## sample estimates:  
## mean in group female mean in group male   
## 64543 73239

# hypothesis testing using infer package  
pay\_diff <- omega %>%  
 specify(salary ~ gender) %>%  
 calculate(stat = "diff in means", order = c("male", "female"))  
  
set.seed(1234)  
  
 # randomise sample  
null\_dist2 <- omega %>%  
 specify(salary ~ gender) %>%  
 hypothesize(null = "independence") %>%  
 generate(reps = 1000, type = "permute") %>%  
 calculate(stat = "diff in means", order = c("male", "female"))  
  
 # visualise randomised sample   
null\_dist2 %>% visualize() +  
   
 # compare with observed difference  
 shade\_p\_value(obs\_stat = pay\_diff, direction = "two\_sided") +  
   
 # add titles  
 labs(subtitle = "Red line shows observed difference in mean pay",  
 x = "Difference in mean salary between males and females (in USD)")



# calculate p-value  
null\_dist2 %>%  
 get\_p\_value(obs\_stat = pay\_diff, direction = "two\_sided")

## # A tibble: 1 x 1  
## p\_value  
## <dbl>  
## 1 0

What can you conclude from your analysis? A couple of sentences would be enough

Null hypothesis: There is no difference in mean salary between men and women Alternate hypothesis: There is a difference in mean salary between the 2 genders

Hypothesis testing using t.test() yields t statistic = -4, p-value = 2e-4, while hypothesis testing using infer package yields a p-value of 0. The visualisation also shows that the observed difference is far outside the range of possible observations randomised under the assumption that the difference in mean pay is 0 (using the infer package). Thus, using both methods the null hypothesis must be rejected even at a 99.9% level of confidence. This further supports the previous analysis, proving that men are indeed paid higher salaries then women.

## Relationship Experience - Gender?

At the board meeting, someone raised the issue that there was indeed a substantial difference between male and female salaries, but that this was attributable to other reasons such as differences in experience. A questionnaire send out to the 50 executives in the sample reveals that the average experience of the men is approximately 21 years, whereas the women only have about 7 years experience on average (see table below).

# Summary Statistics of salary by gender  
favstats (experience ~ gender, data=omega)

## gender min Q1 median Q3 max mean sd n missing  
## 1 female 0 0.25 3.0 14.0 29 7.38 8.51 26 0  
## 2 male 1 15.75 19.5 31.2 44 21.12 10.92 24 0

Based on this evidence, can you conclude that there is a significant difference between the experience of the male and female executives? Perform similar analyses as in the previous section. Does your conclusion validate or endanger your conclusion about the difference in male and female salaries?

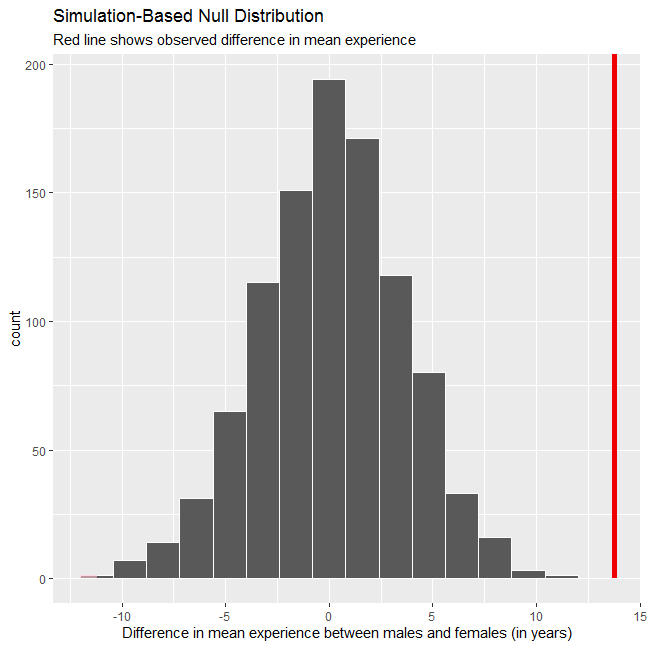
Based on the information above, we can observe that the 25-percentile, median, 75-percentile, maximum, and mean pay of men are all significantly higher than that of the women. This suggests that it is very likely that male executives have significantly more experience than female counterparts. However, to be more certain, we can perform further hypothesis testing.

Null hypothesis: There is no difference in mean experience between men and women Alternate hypothesis: There is a difference in mean experience between the 2 genders

# hypothesis testing using t.test()  
t.test(experience ~ gender, data = omega)

##   
## Welch Two Sample t-test  
##   
## data: experience by gender  
## t = -5, df = 43, p-value = 1e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -19.35 -8.13  
## sample estimates:  
## mean in group female mean in group male   
## 7.38 21.12

# hypothesis testing using infer package  
exp\_diff <- omega %>%  
 specify(experience ~ gender) %>%  
 calculate(stat = "diff in means", order = c("male", "female"))  
  
set.seed(5555)  
  
 # Randomised sample  
null\_dist3 <- omega %>%  
 specify(experience ~ gender) %>%  
 hypothesize(null = "independence") %>%  
 generate(reps = 1000, type = "permute") %>%  
 calculate(stat = "diff in means", order = c("male", "female"))  
  
 # Visualise sample normal distribution and compare with observed difference  
null\_dist3 %>% visualize() +  
 shade\_p\_value(obs\_stat = exp\_diff, direction = "two\_sided") +  
   
 # Add titles  
 labs(subtitle = "Red line shows observed difference in mean experience",  
 x = "Difference in mean experience between males and females (in years)")



# Calculate p-value  
null\_dist3 %>%  
 get\_p\_value(obs\_stat = exp\_diff, direction = "two\_sided")

## # A tibble: 1 x 1  
## p\_value  
## <dbl>  
## 1 0

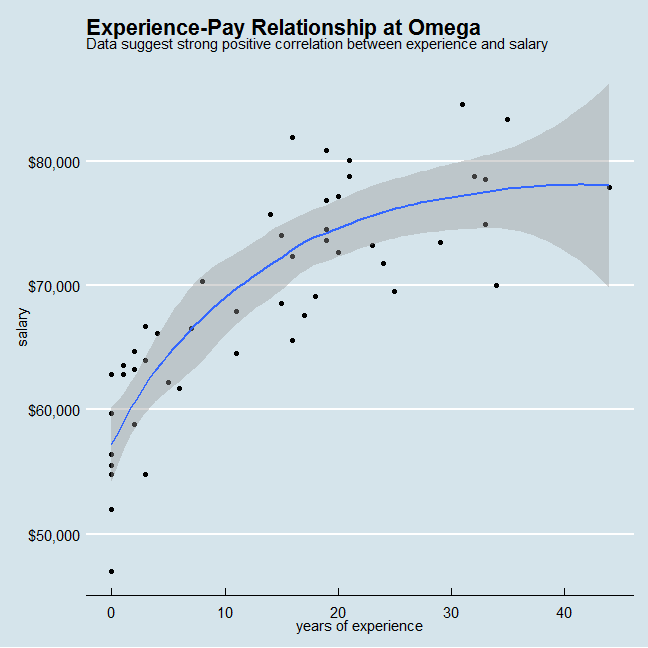
Both hypothesis testing methods yield very small p-value. Thus, we would reject the null hypothesis even at a 99.99% level of confidence. This suggests that there is a substantial difference in experience between the 2 genders. However, this would neither endanger nor validate my previous conclusion about the difference in pay between men and women - it exists regardless. What may change, however, is the assumption of what caused such difference, as now it may be due to a difference in work experience and not pure gender-based discrimination.

## Relationship Salary - Experience ?

Someone at the meeting argues that clearly, a more thorough analysis of the relationship between salary and experience is required before any conclusion can be drawn about whether there is any gender-based salary discrimination in the company.

Analyse the relationship between salary and experience. Draw a scatterplot to visually inspect the data

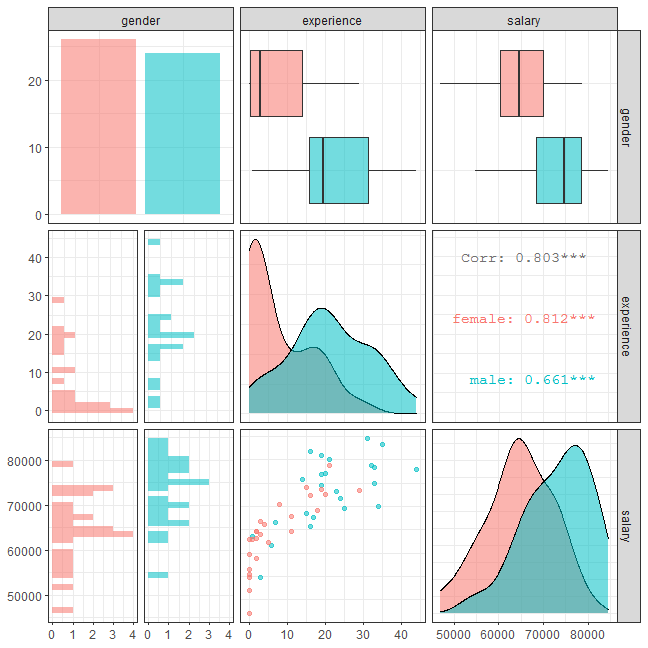
# Choose data frame, set what appears on each axis, and display different colours for different genders  
ggplot(omega, aes(x = experience, y = salary)) +  
   
 # Format plot as scatterplot with smooth line  
 geom\_point() +  
 geom\_smooth() +  
   
 # Choose theme  
 theme\_economist() +  
   
 # Set titles  
 labs(subtitle = "Data suggest strong positive correlation between experience and salary",  
 title = "Experience-Pay Relationship at Omega",  
 x = "years of experience") +  
   
 # Add dollar sign for y-axis  
 scale\_y\_continuous(labels = scales::dollar)

 The plot shows positive correlation between salary and experience. However, the correlation also seems to diminish as experience increases.

## Check correlations between the data

You can use GGally:ggpairs() to create a scatterplot and correlation matrix. Essentially, we change the order our variables will appear in and have the dependent variable (Y), salary, as last in our list. We then pipe the dataframe to ggpairs() with aes arguments to colour by gender and make ths plots somewhat transparent (alpha = 0.3).

omega %>%   
 select(gender, experience, salary) %>% #order variables they will appear in ggpairs()  
 ggpairs(aes(colour=gender, alpha = 0.3))+  
 theme\_bw()



Look at the salary vs experience scatterplot. What can you infer from this plot? Explain in a couple of sentences

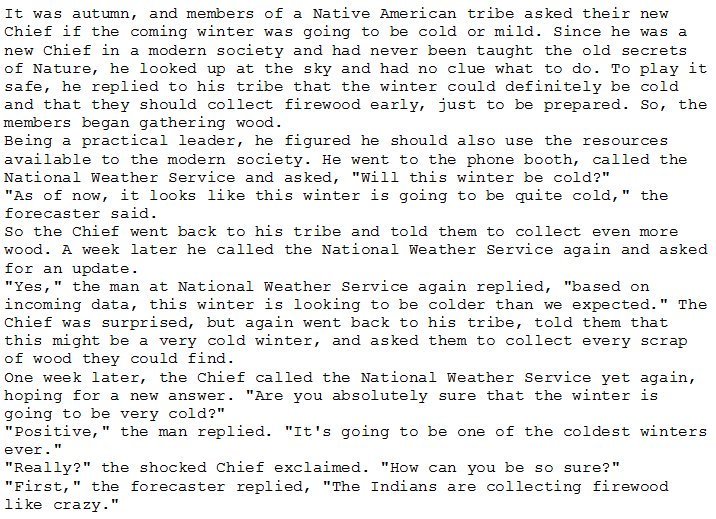
Similar to the previous plot, this salary vs experience plot also suggests strong positive correlation between and experience. In other words, the higher the experience, the higher the salary one receives. This applies for both genders. This is intuitive as higher experience is likely to expand professional skills and sharpen management aptitude, consequently allowing one to take higher positions with higher rewards.

# Challenge 1: Yield Curve inversion

Every so often, we hear warnings from commentators on the “inverted yield curve” and its predictive power with respect to recessions. An explainer what a [inverted yield curve is can be found here](https://www.reuters.com/article/us-usa-economy-yieldcurve-explainer/explainer-what-is-an-inverted-yield-curve-idUSKBN1O50GA). If you’d rather listen to something, here is a great podcast from [NPR on yield curve indicators](https://www.podbean.com/media/share/dir-4zgj9-6aefd11)

In addition, many articles and commentators think that, e.g., [*Yield curve inversion is viewed as a harbinger of recession*](https://www.bloomberg.com/news/articles/2019-08-14/u-k-yield-curve-inverts-for-first-time-since-financial-crisis). One can always doubt whether inversions are truly a harbinger of recessions, and [use the attached parable on yield curve inversions](https://twitter.com/5_min_macro/status/1161627360946511873).

knitr::include\_graphics(here::here("images", "yield\_curve\_parable.jpg"), error = FALSE)



In our case we will look at US data and use the [FRED database](https://fred.stlouisfed.org/) to download historical yield curve rates, and plot the yield curves since 1999 to see when the yield curves flatten. If you want to know more, a very nice article that explains the [yield curve is and its inversion can be found here](https://fredblog.stlouisfed.org/2018/10/the-data-behind-the-fear-of-yield-curve-inversions/).

First, we will use the tidyquant package to download monthly rates for different durations.

# Get a list of FRED codes for US rates and US yield curve; choose monthly frequency  
# to see, eg., the 3-month T-bill https://fred.stlouisfed.org/series/TB3MS  
tickers <- c('TB3MS', # 3-month Treasury bill (or T-bill)  
 'TB6MS', # 6-month  
 'GS1', # 1-year  
 'GS2', # 2-year, etc....  
 'GS3',  
 'GS5',  
 'GS7',  
 'GS10',  
 'GS20',  
 'GS30') #.... all the way to the 30-year rate  
  
# Turn FRED codes to human readable variables  
myvars <- c('3-Month Treasury Bill',  
 '6-Month Treasury Bill',  
 '1-Year Treasury Rate',  
 '2-Year Treasury Rate',  
 '3-Year Treasury Rate',  
 '5-Year Treasury Rate',  
 '7-Year Treasury Rate',  
 '10-Year Treasury Rate',  
 '20-Year Treasury Rate',  
 '30-Year Treasury Rate')  
  
maturity <- c('3m', '6m', '1y', '2y','3y','5y','7y','10y','20y','30y')  
  
# by default R will sort these maturities alphabetically; but since we want  
# to keep them in that exact order, we recast maturity as a factor   
# or categorical variable, with the levels defined as we want  
maturity <- factor(maturity, levels = maturity)  
  
# Create a lookup dataset  
mylookup<-data.frame(symbol=tickers,var=myvars, maturity=maturity)  
# Take a look:  
mylookup %>%   
 knitr::kable()

|  |  |  |
| --- | --- | --- |
| symbol | var | maturity |
| TB3MS | 3-Month Treasury Bill | 3m |
| TB6MS | 6-Month Treasury Bill | 6m |
| GS1 | 1-Year Treasury Rate | 1y |
| GS2 | 2-Year Treasury Rate | 2y |
| GS3 | 3-Year Treasury Rate | 3y |
| GS5 | 5-Year Treasury Rate | 5y |
| GS7 | 7-Year Treasury Rate | 7y |
| GS10 | 10-Year Treasury Rate | 10y |
| GS20 | 20-Year Treasury Rate | 20y |
| GS30 | 30-Year Treasury Rate | 30y |

df <- tickers %>% tidyquant::tq\_get(get="economic.data",   
 from="1960-01-01") # start from January 1960  
  
glimpse(df)

## Rows: 6,774  
## Columns: 3  
## $ symbol <chr> "TB3MS", "TB3MS", "TB3MS", "TB3MS", "TB3MS", "TB3MS", "TB3MS...  
## $ date <date> 1960-01-01, 1960-02-01, 1960-03-01, 1960-04-01, 1960-05-01,...  
## $ price <dbl> 4.35, 3.96, 3.31, 3.23, 3.29, 2.46, 2.30, 2.30, 2.48, 2.30, ...

Our dataframe df has three columns (variables):

* symbol: the FRED database ticker symbol
* date: already a date object
* price: the actual yield on that date

The first thing would be to join this dataframe df with the dataframe mylookup so we have a more readable version of maturities, durations, etc.

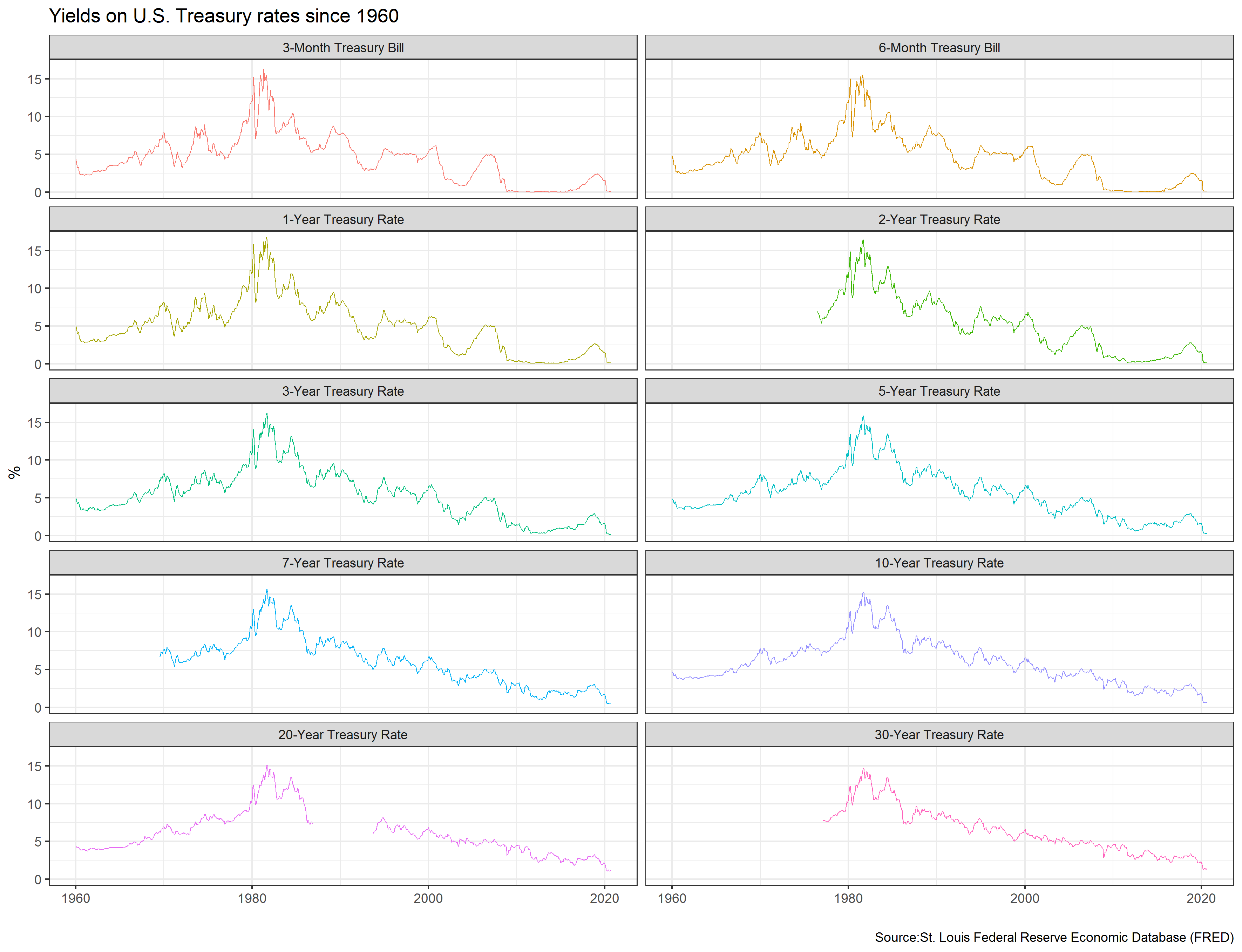
yield\_curve <-left\_join(df,mylookup,by="symbol")

## Plotting the yield curve

This may seem long but it should be easy to produce the following three plots

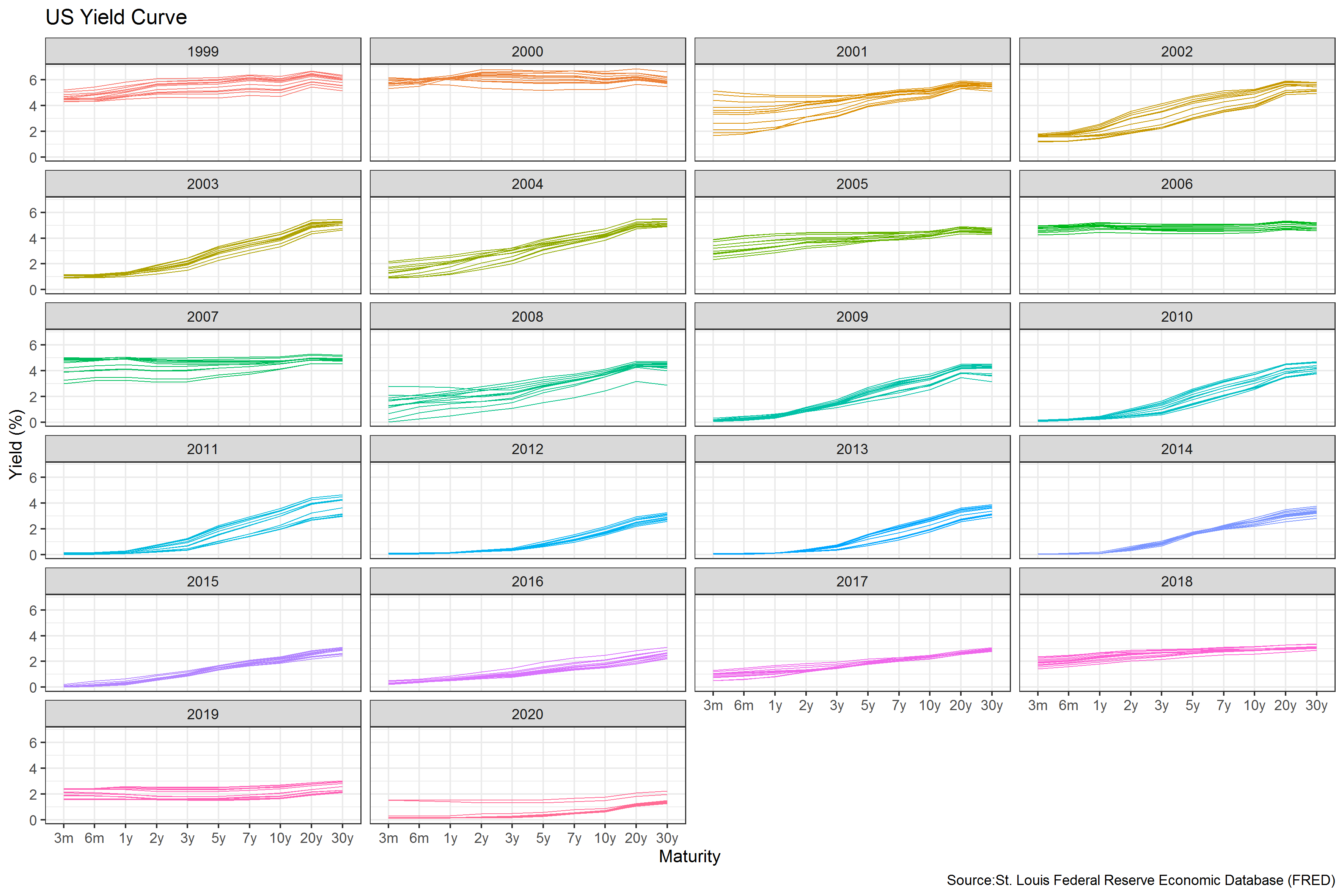
### Yields on US rates by duration since 1960

graph <- yield\_curve %>%   
   
 # Remove NA values in relevant column  
 drop\_na(var) %>%   
   
 # Reformat var column as factor  
 mutate(var = factor(var, levels = myvars)) %>%  
   
 # Create plot  
 ggplot() +  
   
 # Format plot as line plot, set variables for x and y axis  
 geom\_line(aes(x = date, y = price, color = var),   
 size = 0.25) +   
   
 # Facet by type of Treasury Bill  
 facet\_wrap(~var, ncol = 2) +  
   
 # Add titles and source caption  
 labs(title = "Yields on U.S. Treasury rates since 1960",  
 x = "",  
 y = "%",  
 caption = "Source:St. Louis Federal Reserve Economic Database (FRED)") +  
   
 # Choose theme  
 theme\_bw() +  
   
 # Hide legend  
 theme(legend.position = "none")  
  
# Save graph as custom format  
ggsave("yield\_graph1.png",  
 plot = last\_plot(),  
 scale = 1,  
 width = 30,  
 height = 23,  
 units = "cm",  
 dpi = 300,  
 limitsize = TRUE)  
  
knitr::include\_graphics(here::here("yield\_graph1.png"), error = FALSE)



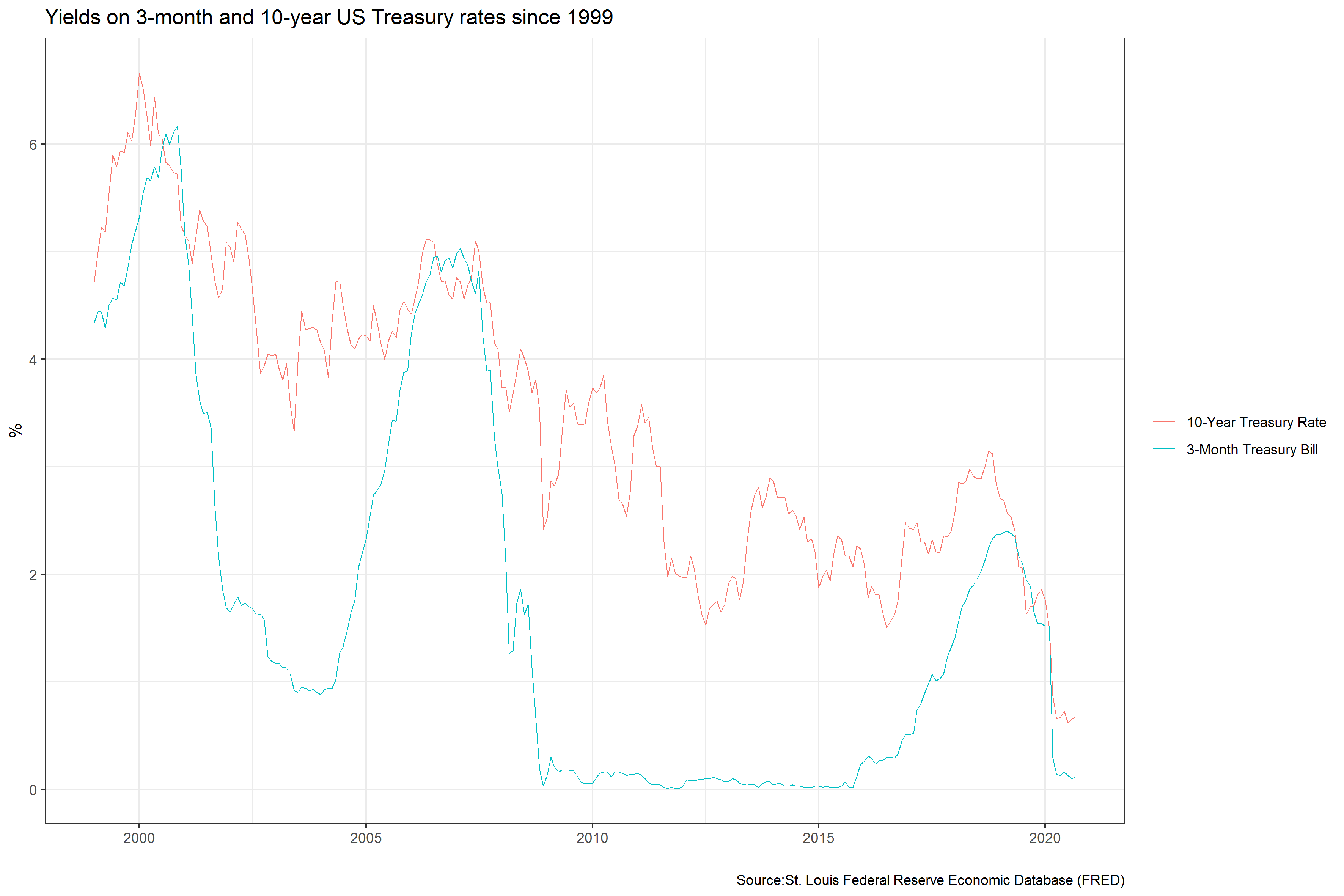
### Monthly yields on US rates by duration since 1999 on a year-by-year basis

graph2 <- yield\_curve %>%   
   
 # Remove NA values in relevant column  
 drop\_na(maturity) %>%   
   
 # Add year and month columns  
 mutate(year = format(as.Date(date, format = "%Y/%m/%d"), "%Y"),  
 month = format(as.Date(date, format = "%Y/%m/%d"), "%m")) %>%   
   
 # Filter from year 1999 onwards  
 filter(year >= "1999") %>%   
   
 # Plot data  
 ggplot() +  
   
 # Format plot as line graphs, set variables for x and y column   
 geom\_line(aes(x = maturity, y = price, color = year, group = month),   
 size = 0.25) +   
   
 # Facet by year  
 facet\_wrap(~year, ncol = 4) +  
   
 # Add titles and source caption  
 labs(title = "US Yield Curve",  
 x = "Maturity",  
 y = "Yield (%)",  
 caption = "Source:St. Louis Federal Reserve Economic Database (FRED)") +  
   
 # Choose theme  
 theme\_bw() +  
   
 # Hide legend  
 theme(legend.position = "none")  
  
# Save graph as custom format  
ggsave("yield\_graph2.png",  
 plot = last\_plot(),  
 scale = 1,  
 width = 30,  
 height = 20,  
 units = "cm",  
 dpi = 300,  
 limitsize = TRUE)  
  
knitr::include\_graphics(here::here("yield\_graph2.png"), error = FALSE)



### 3-month and 10-year yields since 1999

graph3 <- yield\_curve %>%   
   
 # Remove NA values in relevant column  
 drop\_na(maturity) %>%   
   
 # Add year column  
 mutate(year = format(as.Date(date, format = "%Y/%m/%d"), "%Y")) %>%   
   
 # Filter year from 199 onward, and maturity as 3-month or 10-year only  
 filter(year >= "1999",  
 maturity %in% c("3m","10y")) %>%   
   
 # Create plot  
 ggplot() +  
   
 # Format plot as line graph, set variables for x and y axis  
 geom\_line(aes(x = date, y = price, group = var, color = var),   
 size = 0.25) +   
   
 # Add titles  
 labs(title = "Yields on 3-month and 10-year US Treasury rates since 1999",  
 x = "",  
 y = "%",  
 caption = "Source:St. Louis Federal Reserve Economic Database (FRED)") +  
   
 # Choose theme  
 theme\_bw() +  
   
 # Hide legend title  
 theme(legend.title = element\_blank())   
   
  
# Save graph as custom format  
ggsave("yield\_graph3.png",  
 plot = last\_plot(),  
 scale = 1,  
 width = 30,  
 height = 20,  
 units = "cm",  
 dpi = 300,  
 limitsize = TRUE)  
  
knitr::include\_graphics(here::here("yield\_graph3.png"), error = FALSE)



According to [Wikipedia’s list of recession in the United States](https://en.wikipedia.org/wiki/List_of_recessions_in_the_United_States), since 1999 there have been two recession in the US: between Mar 2001–Nov 2001 and between Dec 2007–June 2009. Does the yield curve seem to flatten before these recessions? Can a yield curve flattening really mean a recession is coming in the US? Since 1999, when did short-term (3 months) yield more than longer term (10 years) debt?

Besides calculating the spread (10year - 3months), there are a few things we need to do to produce our final plot

1. Setup data for US recessions
2. Superimpose recessions as the grey areas in our plot
3. Plot the spread between 30 years and 3 months as a blue/red ribbon, based on whether the spread is positive (blue) or negative(red)

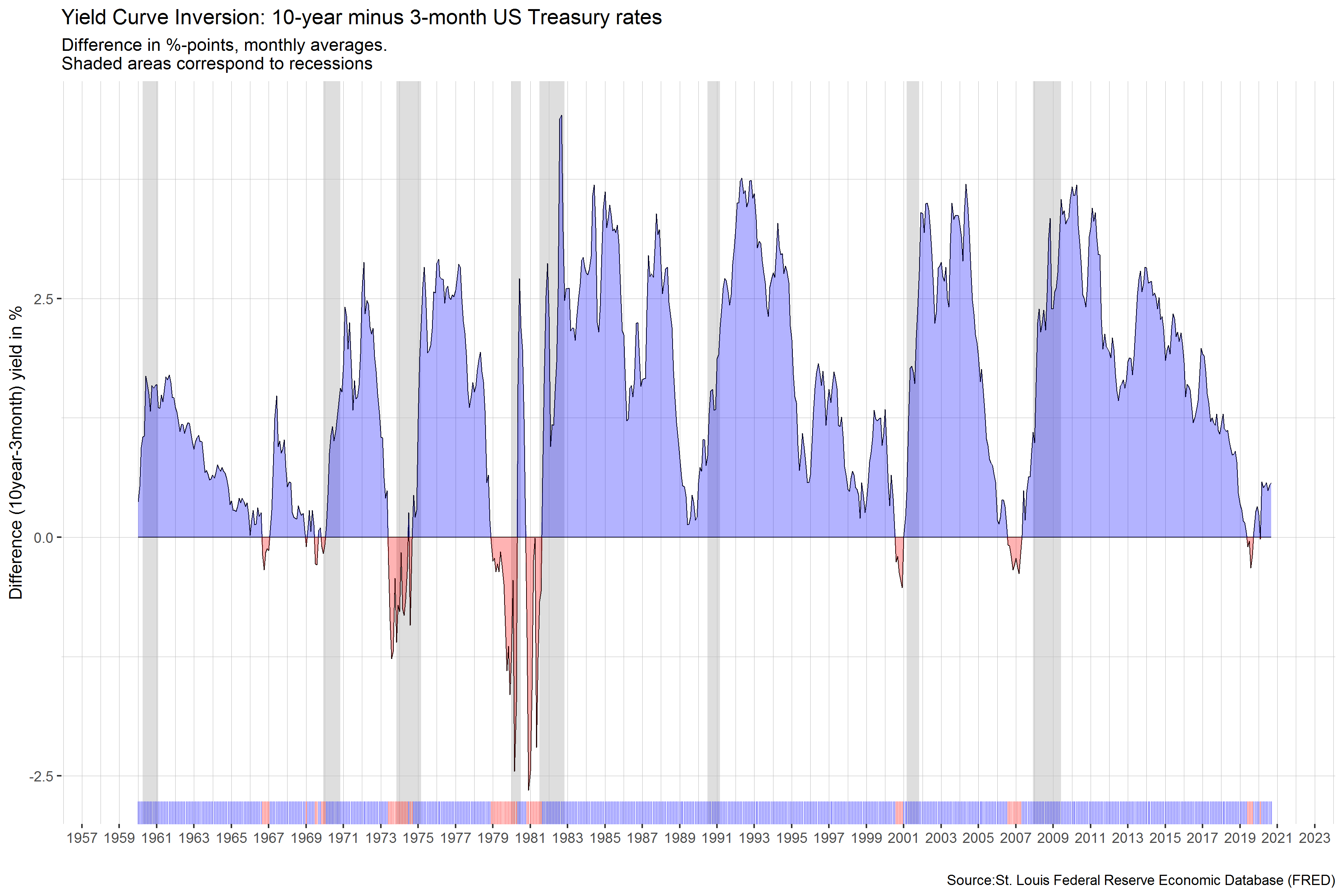
* For the first, the code below creates a dataframe with all US recessions since 1946

# get US recession dates after 1946 from Wikipedia   
# https://en.wikipedia.org/wiki/List\_of\_recessions\_in\_the\_United\_States  
  
recessions <- tibble(  
 from = c("1948-11-01", "1953-07-01", "1957-08-01", "1960-04-01", "1969-12-01", "1973-11-01", "1980-01-01","1981-07-01", "1990-07-01", "2001-03-01", "2007-12-01"),   
 to = c("1949-10-01", "1954-05-01", "1958-04-01", "1961-02-01", "1970-11-01", "1975-03-01", "1980-07-01", "1982-11-01", "1991-03-01", "2001-11-01", "2009-06-01")   
 ) %>%   
 mutate(From = ymd(from),   
 To=ymd(to),  
 duration\_days = To-From)  
  
recessions

## # A tibble: 11 x 5  
## from to From To duration\_days  
## <chr> <chr> <date> <date> <drtn>   
## 1 1948-11-01 1949-10-01 1948-11-01 1949-10-01 334 days   
## 2 1953-07-01 1954-05-01 1953-07-01 1954-05-01 304 days   
## 3 1957-08-01 1958-04-01 1957-08-01 1958-04-01 243 days   
## 4 1960-04-01 1961-02-01 1960-04-01 1961-02-01 306 days   
## 5 1969-12-01 1970-11-01 1969-12-01 1970-11-01 335 days   
## 6 1973-11-01 1975-03-01 1973-11-01 1975-03-01 485 days   
## 7 1980-01-01 1980-07-01 1980-01-01 1980-07-01 182 days   
## 8 1981-07-01 1982-11-01 1981-07-01 1982-11-01 488 days   
## 9 1990-07-01 1991-03-01 1990-07-01 1991-03-01 243 days   
## 10 2001-03-01 2001-11-01 2001-03-01 2001-11-01 245 days   
## 11 2007-12-01 2009-06-01 2007-12-01 2009-06-01 548 days

* To add the grey shaded areas corresponding to recessions, we use geom\_rect()
* to colour the ribbons blue/red we must see whether the spread is positive or negative and then use geom\_ribbon(). You should be familiar with this from last week’s homework on the excess weekly/monthly rentals of Santander Bikes in London.

graph4 <- yield\_curve %>%   
   
 # only look at 3m and 10y maturity yields  
 filter(maturity %in% c("3m","10y")) %>%   
   
 # get date values into seperate columns  
 mutate(year = format(as.Date(date, format = "%Y/%m/%d"), "%Y"),  
 month = format(as.Date(date, format = "%Y/%m/%d"), "%m")) %>%   
   
 # prepare df for tidy format  
 select(date, year, month, maturity, price) %>%   
 mutate(maturity = ifelse(maturity == "3m", "rate\_3m", "rate\_10y")) %>%   
   
 # convert df into tidy format  
 pivot\_wider(names\_from = "maturity", values\_from = "price") %>%   
 mutate(spread = rate\_10y - rate\_3m) %>%   
   
 # Create graph  
 ggplot() +  
   
 # create recession shading  
 geom\_rect(data = filter(recessions),   
 inherit.aes = F,   
 aes(xmin = From,   
 xmax = To,   
 ymin = -Inf,   
 ymax = Inf),  
 fill = "grey",alpha=0.5) +  
   
 # create deviation and forecast (y=0) lines  
 geom\_line(aes(x = date, y = spread, na.rm = TRUE),   
 color = "black", size = 0.25) +  
 geom\_line(aes(x = date, y = 0, na.rm = TRUE),   
 color = "black", size = 0.25) +   
   
 # Create ribbon to highlight normal vs inverted yield curve  
 geom\_ribbon(aes(x = date,   
 ymin = pmin(spread,0),   
 ymax = 0,   
 fill = "Normal"),   
 alpha = .3) +  
 geom\_ribbon(aes(x = date,   
 ymin = 0,   
 ymax = pmax(spread,0),   
 fill = "Inverted"), alpha = .3) +  
 scale\_fill\_manual(values=c("blue","red"),   
 name="Normal vs. Inverted") +  
   
 # Create rug at bottom of the graphs  
 geom\_rug(aes(x = date,  
 colour = ifelse(spread >= 0,">=0","<0")),   
 sides = "b",  
 alpha = 0.4,  
 size = 0.4) +  
 scale\_colour\_manual(values=c("red","blue"),  
 name="Normal vs. Inverted",   
 guide=FALSE) +  
   
 # custom x aaxis scales  
 scale\_x\_date(date\_breaks = "2 years",  
 date\_labels = "%Y",  
 limits = c(from = as.Date("1959-01-01"), to = as.Date("2020-12-31"))) +  
   
 # Create custom theme  
 theme(  
 legend.position = "none",  
 panel.background = element\_rect(color="white", fill = "white"),  
 panel.grid = element\_line(color="grey", size = 0.1)) +  
 labs(title = "Yield Curve Inversion: 10-year minus 3-month US Treasury rates",  
 subtitle = "Difference in %-points, monthly averages. \nShaded areas correspond to recessions",  
 x = "",  
 y = "Difference (10year-3month) yield in %" ,  
 caption = "Source:St. Louis Federal Reserve Economic Database (FRED)")   
  
  
# Save graph as custom format  
ggsave("yield\_challenge.png",  
 plot = last\_plot(),  
 scale = 1,  
 width = 30,  
 height = 20,  
 units = "cm",  
 dpi = 300,  
 limitsize = TRUE)  
  
knitr::include\_graphics(here::here("yield\_challenge.png"), error = FALSE)



# Challenge 2:GDP components over time and among countries

At the risk of oversimplifying things, the main components of gross domestic product, GDP are personal consumption (C), business investment (I), government spending (G) and net exports (exports - imports). You can read more about GDP and the different approaches in calculating at the [Wikipedia GDP page](https://en.wikipedia.org/wiki/Gross_domestic_product).

The GDP data we will look at is from the [United Nations’ National Accounts Main Aggregates Database](https://unstats.un.org/unsd/snaama/Downloads), which contains estimates of total GDP and its components for all countries from 1970 to today. We will look at how GDP and its components have changed over time, and compare different countries and how much each component contributes to that country’s GDP. The file we will work with is [GDP and its breakdown at constant 2010 prices in US Dollars](http://unstats.un.org/unsd/amaapi/api/file/6) and it has already been saved in the Data directory. Have a look at the Excel file to see how it is structured and organised

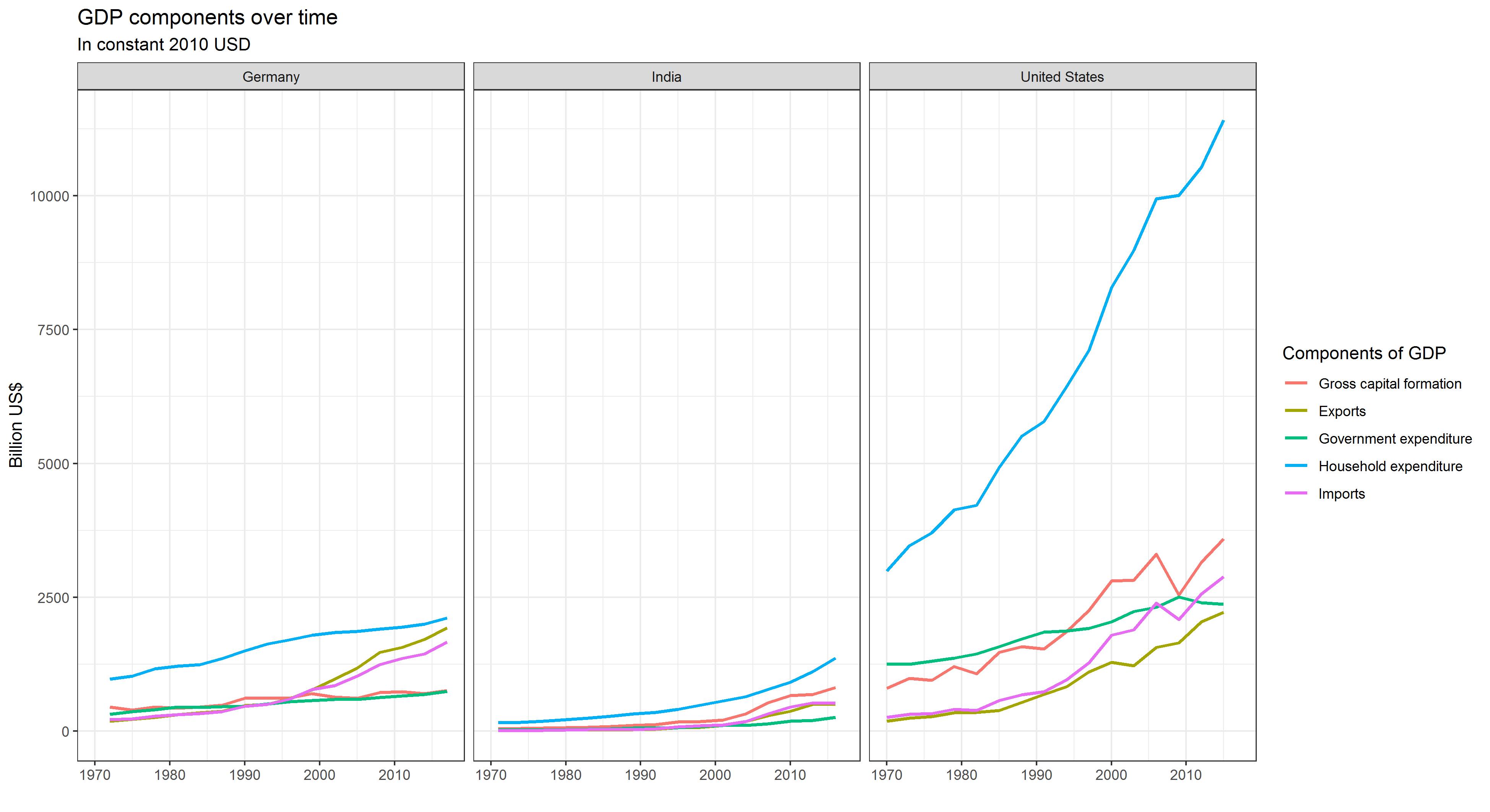
UN\_GDP\_data <- read\_excel(here::here("data", "Download-GDPconstant-USD-countries.xls"), # Excel filename  
 sheet="Download-GDPconstant-USD-countr", # Sheet name  
 skip=2) # Number of rows to skip

The first thing you need to do is to tidy the data, as it is in wide format and you must make it into long, tidy format. Please express all figures in billions (divide values by 1e9, or ), and you want to rename the indicators into something shorter.

tidy\_GDP\_data <- UN\_GDP\_data %>%  
   
 # Use pivot\_longer to move all years into a single new column  
 pivot\_longer(cols = 4:51,   
 names\_to = "year",   
 values\_to = "value") %>%  
   
 # Display figures in billions (divide by '1e9')  
 mutate(value = value/1e9,  
   
 # Rename indicators  
 IndicatorName = case\_when(  
 IndicatorName == "Final consumption expenditure" ~ "consumption\_final",  
 IndicatorName == "Household consumption expenditure (including Non-profit institutions serving households)" ~ "consumption\_household",  
 IndicatorName == "General government final consumption expenditure" ~ "gov\_spend",  
 IndicatorName == "Gross capital formation" ~ "GCF",  
 IndicatorName == "Gross fixed capital formation (including Acquisitions less disposals of valuables)" ~ "GFCF",  
 IndicatorName == "Changes in inventories" ~ "inventories\_change",  
 IndicatorName == "Exports of goods and services" ~ "exports",  
 IndicatorName == "Imports of goods and services" ~ "imports",  
 IndicatorName == "Gross Domestic Product (GDP)" ~ "GDP",  
 IndicatorName == "Agriculture, hunting, forestry, fishing (ISIC A-B)" ~ "isic\_A\_B",  
 IndicatorName == "Mining, Manufacturing, Utilities (ISIC C-E)" ~ "isic\_C\_E",  
 IndicatorName == "Manufacturing (ISIC D)" ~ "isic\_D",  
 IndicatorName == "Construction (ISIC F)" ~ "isic\_F",  
 IndicatorName == "Wholesale, retail trade, restaurants and hotels (ISIC G-H)" ~ "isic\_G\_H",  
 IndicatorName == "Transport, storage and communication (ISIC I)" ~ "isic\_I",  
 IndicatorName == "Other Activities (ISIC J-P)" ~ "isic\_J\_P",  
 IndicatorName == "Total Value Added" ~ "total\_value"),  
 # Change year to numeric  
 year = as.numeric(year))  
  
# Let us compare GDP components for these 3 countries  
country\_list <- c("United States","India", "Germany")

First, plot how GDP components change over time for Germany, India, and the US.

GDP\_plot1 <- tidy\_GDP\_data %>%  
   
 # Filter for required countries and indicators only  
 filter(Country == country\_list,  
 IndicatorName %in% c("GCF", "exports", "gov\_spend", "consumption\_household", "imports")) %>%  
 mutate(IndicatorName = factor(IndicatorName,   
 levels = c("GCF", "exports", "gov\_spend", "consumption\_household", "imports"),  
 labels = c("Gross capital formation", "Exports", "Government expenditure", "Household expenditure", "Imports"))) %>%  
   
 # Plot the graph, stating axes variables, and colour-grouping by country within each facet  
 ggplot(aes(x = year, y = value, colour = IndicatorName, na.rm = TRUE)) +  
   
 # Set plot type to geom\_line and specify size of lines  
 geom\_line(size = 1) +  
   
 # Set grid lines for both axes   
 scale\_y\_continuous(minor\_breaks = seq(from = 0, to = 12500, by = 1250),   
 breaks = seq(from = 0, to = 12500, by = 2500)) +  
 scale\_x\_continuous(breaks = seq(from = 1970, to = 2017, by = 10)) +  
   
 # Choose theme  
 theme\_bw() +  
   
 # Facet by country  
 facet\_wrap(~Country) +  
   
 # Set labels and titles  
 labs(title = "GDP components over time",  
 subtitle = "In constant 2010 USD",  
 x = "",  
 y = "Billion US$",  
 colour = "Components of GDP")   
  
# Save to resize plot  
ggsave("GDP\_plot1.jpg",   
 width = 33,   
 height = 18,   
 units = "cm")  
  
# Display plot  
knitr::include\_graphics(here::here("GDP\_plot1.jpG"), error = FALSE)



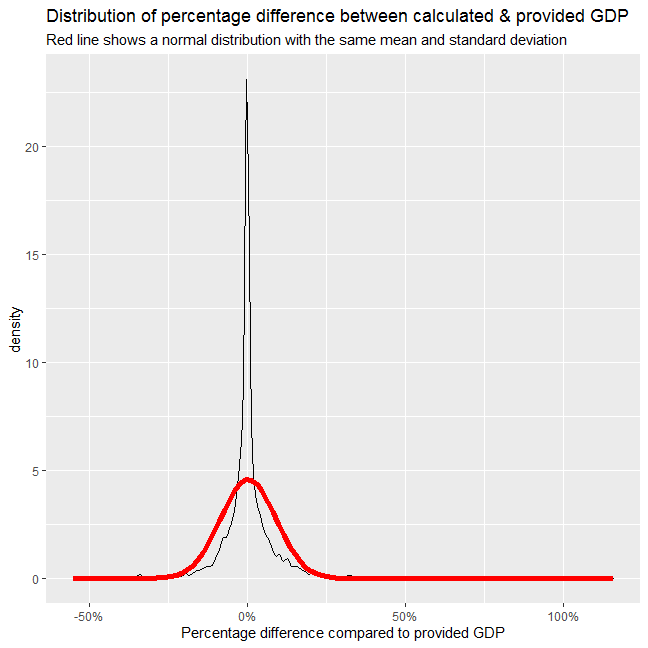
Secondly, recall that GDP is the sum of Household Expenditure (Consumption *C*), Gross Capital Formation (business investment *I*), Government Expenditure (G) and Net Exports (exports - imports). Even though there is an indicator Gross Domestic Product (GDP) in your dataframe, I would like you to calculate it given its components discussed above.

What is the % difference between what you calculated as GDP and the GDP figure included in the dataframe?

GDP\_comparison <- tidy\_GDP\_data %>%  
  
 # Use pivot\_wider to have indicators displayed as columns  
 pivot\_wider(names\_from = IndicatorName, values\_from = value) %>%  
  
 # Remove NA values in relevant columns   
 drop\_na(consumption\_household, GCF, gov\_spend, exports, imports) %>%  
   
 # Create new columns showing GDP in new method, and the corresponding percentage difference with the provided figures  
 mutate(GDP\_new\_method = consumption\_household + GCF + gov\_spend + exports - imports,  
 percentage\_difference = round((GDP\_new\_method - GDP)/GDP, digits = 3)) %>%  
   
 # Select required columns only  
 select(Country, year, percentage\_difference)  
  
GDP\_comparison

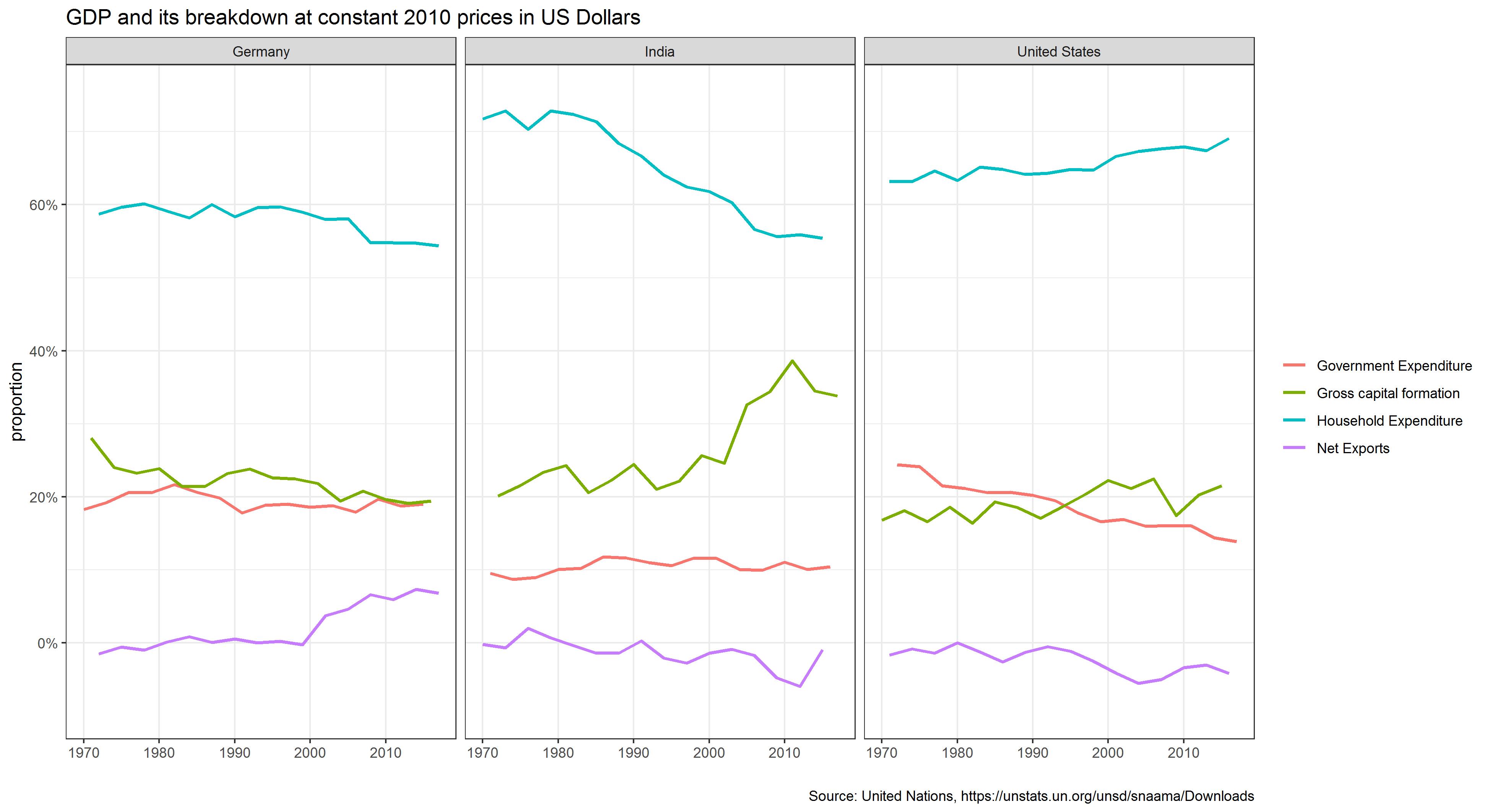
## # A tibble: 9,574 x 3  
## Country year percentage\_difference  
## <chr> <dbl> <dbl>  
## 1 Afghanistan 1970 -0.408  
## 2 Afghanistan 1971 -0.425  
## 3 Afghanistan 1972 -0.336  
## 4 Afghanistan 1973 -0.294  
## 5 Afghanistan 1974 -0.269  
## 6 Afghanistan 1975 -0.275  
## 7 Afghanistan 1976 -0.246  
## 8 Afghanistan 1977 -0.177  
## 9 Afghanistan 1978 -0.19   
## 10 Afghanistan 1979 -0.196  
## # ... with 9,564 more rows

# Draw density plot to portray difference  
ggplot(GDP\_comparison, aes(x = percentage\_difference)) +  
 geom\_density() +  
   
 # Format x-axis to show percentage  
 scale\_x\_continuous(labels = scales::percent) +  
   
 # Add normal distribution in red to compare  
 stat\_function(  
 fun = dnorm,  
 color = "red",  
 size = 2,  
 args = list(mean = mean(GDP\_comparison$percentage\_difference, na.rm = TRUE),   
 sd = sd(GDP\_comparison$percentage\_difference, na.rm = TRUE))) +  
   
 # Add titles  
 labs(title = "Distribution of percentage difference between calculated & provided GDP",  
 subtitle = "Red line shows a normal distribution with the same mean and standard deviation",  
 x = "Percentage difference compared to provided GDP")



The calculated GDP differs from the stated GDP for most countries and most years. However, the distribuion plot shows that on average the percentage difference is around 0%, with most data points highly concentrated around this average.

GDP\_plot2 <- tidy\_GDP\_data %>%  
 # Use pivot\_wider to have indicators displayed as columns  
 pivot\_wider(names\_from = IndicatorName, values\_from = value) %>%  
   
 # Create new columns showing Net Exports  
 mutate(net\_exp = exports - imports) %>%  
   
 # Use pivot\_longer to change data back to long format  
 pivot\_longer(!c(1,2,3,11), names\_to = "IndicatorName", values\_to = "value") %>%  
   
 # Calculate variables as percentage of total GDP  
 mutate(proportion = (value)/(GDP)) %>%  
   
 # Filter for required countries and indicators only  
 filter(Country == country\_list,  
 IndicatorName %in% c("gov\_spend", "GCF", "consumption\_household", "net\_exp")) %>%  
 mutate(IndicatorName = factor(IndicatorName,   
 levels = c("gov\_spend", "GCF", "consumption\_household", "net\_exp"),  
 labels = c("Government Expenditure", "Gross capital formation", "Household Expenditure", "Net Exports"))) %>%  
   
 # Plot the graph, stating axes variables, and colour-grouping by country within each facet  
 ggplot(aes(x = year, y = proportion, colour = IndicatorName, na.rm = TRUE)) +  
   
 # Set plot type to geom\_line and specify size of lines  
 geom\_line(size = 1) +  
   
 # Set grid lines for both axes   
 scale\_y\_continuous(limits = c(-0.09, 0.75),   
 minor\_breaks = seq(from = (0), to = 0.75, by = 0.1),   
 breaks = seq(from = (0), to = 0.75, by = 0.2),  
 labels = scales::percent) +  
 scale\_x\_continuous(breaks = seq(from = 1970, to = 2017, by = 10),   
 minor\_breaks = NULL) +  
   
 # Choose theme  
 theme\_bw() +  
   
 # Remove legend title  
 theme(legend.title = element\_blank()) +  
   
 # Facet by country  
 facet\_wrap(~Country) +  
   
 # Set labels, source caption, and axes titles  
 labs(title = "GDP and its breakdown at constant 2010 prices in US Dollars",  
 x = "",  
 y = "proportion",  
 caption = "Source: United Nations, https://unstats.un.org/unsd/snaama/Downloads")   
  
# Save to resize plot  
ggsave("GDP\_plot2.jpg",   
 width = 33,   
 height = 18,   
 units = "cm")  
  
# Display plot  
knitr::include\_graphics(here::here("GDP\_plot2.jpG"), error = FALSE)



What is this last chart telling you? Can you explain in a couple of paragraphs the different dynamic among these three countries?

The chart reflects different economic trends between Germany, India, and the United States from 1970 until 2017. Household expenditure as proportion of GDP had been increasing in the US, but falling in Germany and India, with the latter experiencing a steeper decline. This may signal changes in consumption behaviour between the 3 countries. As consumption is income less tax and savings, increased spending in the US must be due to decreased tax rates and/or decreased average savings propensity. Similarly, the opposite must have happened in Germany and India.

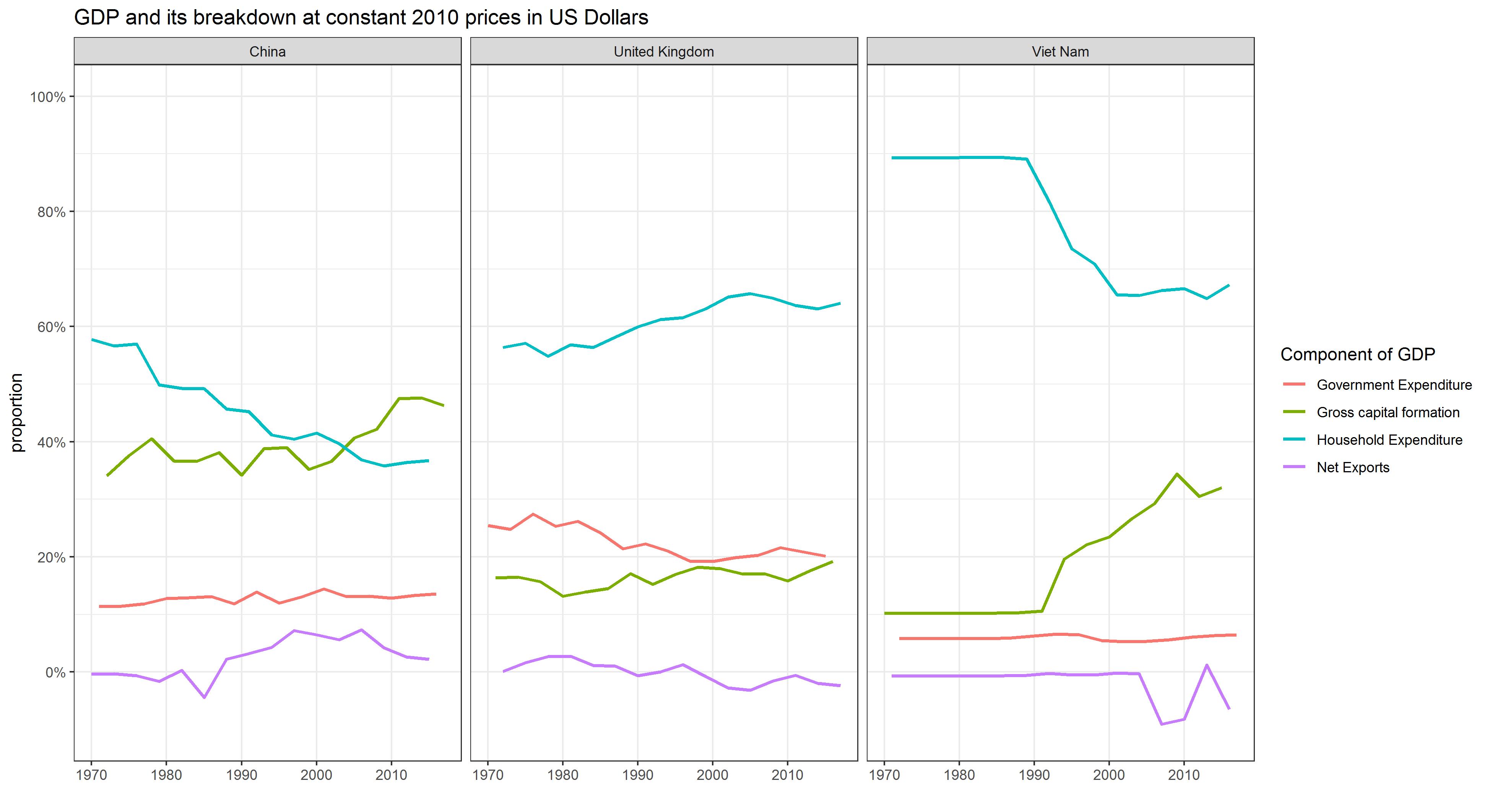
During the same period, business investment (gross capital formation) as percentage of GDP had remarkably surged in India, slightly increased in the US, and decreased in Germany. This trend can be explained by investor’s confidence and potential business returns. As India is still a developing country, it must have higher potential returns than the other 2 developed countries, where markets have matured and business opportunities are swindling.

Meanwhile, there had not been much changes in government expenditure relative to income in Germany and India, whereas the same figure significantly declined in the US. It is more difficult to explain these trends, but fiscal policy would be a noteworthy explanation. While fiscal policy had not changed as much in the first 2 countries, in the latter contractionary fiscal policy may have been employed. As a more ambitious elaboration of this idea, it may be the case that the US had been decreasing tax rates to attract and maintain business investments. This would also help to explain why gross capital formation is increasing relative to GDP in the US, but falling in most other developed nations.

Finally, net exports had proportionally expanded in Germany, and slightly contracted in the other 2 countries. This can be propelled by different factors in different countries. Higher net exports in Germany is most likely a result of its strength in producing high quality exports, especially automobiles, machinery, and pharmaceuticals. On the other hand, rising income and consequently demand for luxury imports in India, as well as the outward migration of manufacturing facilities by US corporations, may have contributed to each respective country’s negative net exports.

If you want to, please change country\_list <- c("United States","India", "Germany") to include your own country and compare it with any two other countries you like

country\_list2 <- c("Viet Nam","China", "United Kingdom")  
  
GDP\_plot3 <- tidy\_GDP\_data %>%  
 # Use pivot\_wider to have indicators displayed as columns  
 pivot\_wider(names\_from = IndicatorName, values\_from = value) %>%  
   
 # Create new columns showing Net Exports  
 mutate(net\_exp = exports - imports) %>%  
   
 # Use pivot\_longer to change data back to long format  
 pivot\_longer(!c(1,2,3,11), names\_to = "IndicatorName", values\_to = "value") %>%  
   
 # Calculate variables as percentage of total GDP  
 mutate(proportion = (value)/(GDP)) %>%  
   
 # Filter for required countries and indicators only  
 filter(Country == country\_list2,  
 IndicatorName %in% c("gov\_spend", "GCF", "consumption\_household", "net\_exp")) %>%  
 mutate(IndicatorName = factor(IndicatorName,   
 levels = c("gov\_spend", "GCF", "consumption\_household", "net\_exp"),  
 labels = c("Government Expenditure", "Gross capital formation", "Household Expenditure", "Net Exports"))) %>%  
   
 # Plot the graph, stating axes variables, and colour-grouping by country within each facet  
 ggplot(aes(x = year, y = proportion, colour = IndicatorName, na.rm = TRUE)) +  
   
 # Set plot type to geom\_line and specify size of lines  
 geom\_line(size = 1) +  
   
 # Set grid lines for both axes   
 scale\_y\_continuous(limits = c(-0.1, 1),   
 minor\_breaks = seq(from = (0), to = 1, by = 0.1),   
 breaks = seq(from = (0), to = 1, by = 0.2),  
 labels = scales::percent) +  
 scale\_x\_continuous(breaks = seq(from = 1970, to = 2017, by = 10),   
 minor\_breaks = NULL) +  
   
 # Choose theme  
 theme\_bw() +  
   
 # Facet by country  
 facet\_wrap(~Country) +  
   
 # Set labels and titles  
 labs(title = "GDP and its breakdown at constant 2010 prices in US Dollars",  
 x = "",  
 y = "proportion",  
 colour = "Component of GDP")   
  
# Save to resize plot  
ggsave("GDP\_plot3.jpg",   
 width = 33,   
 height = 18,   
 units = "cm")  
  
# Display plot  
knitr::include\_graphics(here::here("GDP\_plot3.jpG"), error = FALSE)

 Chosen countries and rationale: Vietnam - my country. China - similar economic trends in the last 5 decades. United Kingdom - more developed, different and more stable economic trend.

Analysis

As Vietnam and China both experienced rapid growth at around the same time, their economic trends have been similar: fluctuating net exports, stable government spending, falling household expenditure, and growing business investment (all relative to GDP). The UK, on the other hand, experienced industrialisation much earlier and had been a developed country since the commencement of the period. Thus, its components of GDP moved in different directions compared to Vietnam and China. Relative to GDP, there had been a more gradual increase in gross capital formation and consumption, as well as a slight decrease in net exports and government spending.

There is a lot of explanatory text, comments, etc. You do not need these, so delete them and produce a stand-alone document that you could share with someone. Knit the edited and completed R Markdown file as an HTML document (use the “Knit” button at the top of the script editor window) and upload it to Canvas.

# Details

* Who did you collaborate with: Leif Beckers, Dung Tran, Salman Abdullah, Andjela Bozinovic, Xiwen Wang
* Approximately how much time did you spend on this problem set: 10h
* What, if anything, gave you the most trouble: NA

**Please seek out help when you need it,** and remember the [15-minute rule](https://mfa2021.netlify.app/syllabus/#the-15-minute-rule). You know enough R (and have enough examples of code from class, your previous homeworks, and your readings) to be able to do this. If you get stuck, ask for help from others, post a question on Slack– and remember that I am here to help too!

As a true test to yourself, do you understand the code you submitted and are you able to explain it to someone else?

Yes

# Rubric

Check minus (1/5): Displays minimal effort. Doesn’t complete all components. Code is poorly written and not documented. Uses the same type of plot for each graph, or doesn’t use plots appropriate for the variables being analyzed.

Check (3/5): Solid effort. Hits all the elements. No clear mistakes. Easy to follow (both the code and the output).

Check plus (5/5): Finished all components of the assignment correctly and addressed both challenges. Code is well-documented (both self-documented and with additional comments as necessary). Used tidyverse, instead of base R. Graphs and tables are properly labelled. Analysis is clear and easy to follow, either because graphs are labeled clearly or you’ve written additional text to describe how you interpret the output.