Assignment 5

library(ggplot2)  
library(dplyr)

Attaching package: 'dplyr'

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

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

library(vtable)

Loading required package: kableExtra

Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':  
  
 group\_rows

winners <- readr::read\_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2023/2023-04-25/winners.csv')

Rows: 163 Columns: 5

── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (3): Category, Athlete, Nationality  
dbl (1): Year  
time (1): Time  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

london\_marathon <- readr::read\_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2023/2023-04-25/london\_marathon.csv')

Rows: 42 Columns: 8  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (1): Official charity  
dbl (6): Year, Applicants, Accepted, Starters, Finishers, Raised  
date (1): Date  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

vtable(winners)

winners

| Name | Class | Values |
| --- | --- | --- |
| Category | character |  |
| Year | numeric | Num: 1981 to 2022 |
| Athlete | character |  |
| Nationality | character |  |
| Time | hms |  |

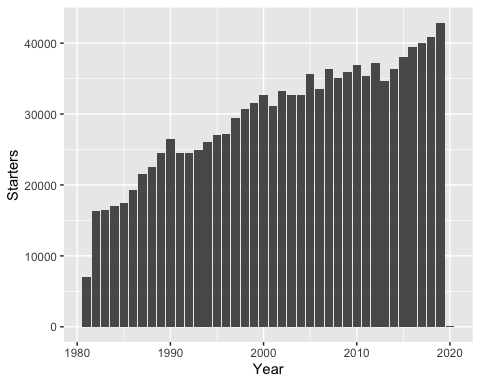
vtable(london\_marathon)

london\_marathon

| Name | Class | Values |
| --- | --- | --- |
| Date | Date | Time: 1981-03-29 to 2022-10-02 |
| Year | numeric | Num: 1981 to 2022 |
| Applicants | numeric | Num: 20000 to 457861 |
| Accepted | numeric | Num: 77 to 56398 |
| Starters | numeric | Num: 77 to 42906 |
| Finishers | numeric | Num: 61 to 42549 |
| Raised | numeric | Num: 46.5 to 66.4 |
| Official charity | character |  |

ggplot(london\_marathon, aes(x = Year, y = Starters)) +   
 geom\_col()

Warning: Removed 2 rows containing missing values (`position\_stack()`).



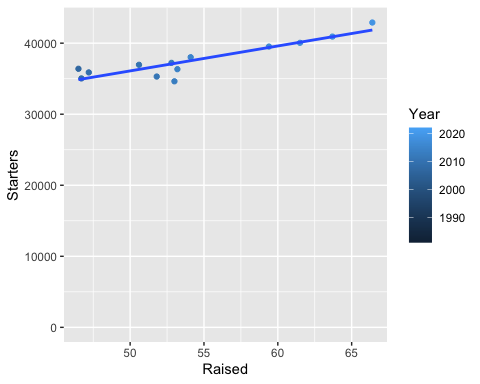
ggplot(london\_marathon, aes(x = Raised, y = Starters, color = Year)) +   
 geom\_point() + geom\_smooth(method = 'lm', se = FALSE)

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

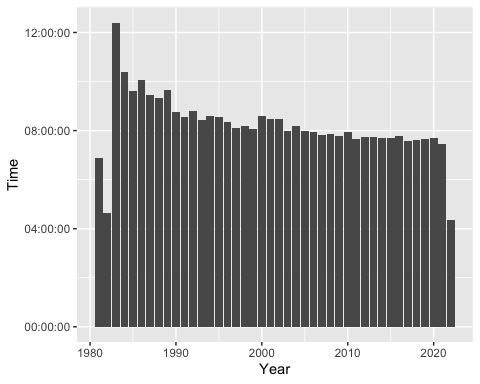
Warning: Removed 29 rows containing non-finite values (`stat\_smooth()`).

Warning: The following aesthetics were dropped during statistical transformation: colour  
ℹ This can happen when ggplot fails to infer the correct grouping structure in  
 the data.  
ℹ Did you forget to specify a `group` aesthetic or to convert a numerical  
 variable into a factor?

Warning: Removed 29 rows containing missing values (`geom\_point()`).



ggplot(winners, aes(x = Year, y = `Time`)) +   
 geom\_col()



london\_marathon <- london\_marathon %>%  
 mutate(Completed = `Finishers`/Starters)  
  
london\_marathon

# A tibble: 42 × 9  
 Date Year Applicants Accepted Starters Finishers Raised  
 <date> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
 1 1981-03-29 1981 20000 7747 7055 6255 NA  
 2 1982-05-09 1982 90000 18059 16350 15116 NA  
 3 1983-04-17 1983 60000 19735 16500 15793 NA  
 4 1984-05-13 1984 70000 21142 16992 15675 NA  
 5 1985-04-21 1985 83000 22274 17500 15873 NA  
 6 1986-04-20 1986 80000 25566 19261 18067 NA  
 7 1987-05-10 1987 80000 28364 21485 19586 NA  
 8 1988-04-17 1988 73000 29979 22469 20932 NA  
 9 1989-04-23 1989 72000 31772 24452 22701 NA  
10 1990-04-22 1990 73000 34882 26500 25013 NA  
# ℹ 32 more rows  
# ℹ 2 more variables: `Official charity` <chr>, Completed <dbl>

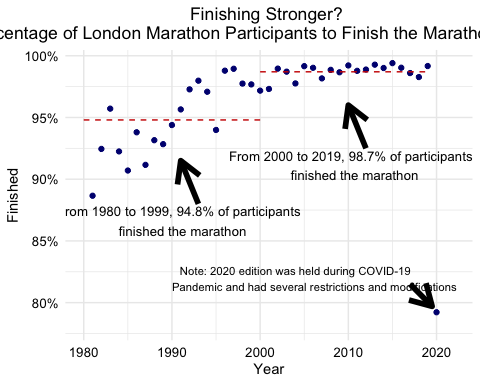
## Finishing Stronger? Percentage of London Marathon participants to **finish** the marathon over time.

ggplot(london\_marathon, aes(x = Year, y = Completed)) +   
 geom\_point(color = "navyblue") +   
 ylim(0.85, 1.0) +   
 scale\_y\_continuous(labels = scales::percent) +  
 theme\_minimal() +   
 ggtitle("Finishing Stronger? \n Percentage of London Marathon Participants to Finish the Marathon over time") +  
 labs(x = "Year", y = "Finished") +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 annotate("text", x = 1991, y = 0.82, label = "From 1980 to 1999, 94.8% of participants \n finished the marathon",   
 size = 3.5, vjust = -1.5, color = "black") +  
 annotate("text", x = 2010.5, y = 0.865, label = "From 2000 to 2019, 98.7% of participants \n finished the marathon",   
 size = 3.5, vjust = -1.5, color = "black") +  
 annotate("text", x = 2004, y = 0.78, label = "Note: 2020 edition was held during COVID-19  
 Pandemic and had several restrictions and modifications",   
 size = 3, vjust = -1.5, color = "black") +  
 annotate("segment", x = 1993, xend = 1991, y = 0.88, yend = 0.915,  
 colour = "black", size = 2, arrow = arrow()) +  
 annotate("segment", x = 2012, xend = 2010, y = 0.925, yend = 0.96,  
 colour = "black", size = 2, arrow = arrow()) +  
 annotate("segment", x = 2017, xend = 2019.5, y = 0.815, yend = 0.797,  
 colour = "black", size = 2, arrow = arrow()) +  
 theme(axis.text = element\_text(color = "black", size = 10)) +  
 annotate("segment", x = 1980, y = 0.948, xend = 2000, yend = 0.948, linetype = 2, color = "red3") +   
 annotate("segment", x = 2000, y = 0.987, xend = 2019, yend = 0.987, linetype = 2, color = "red3")

Scale for y is already present.  
Adding another scale for y, which will replace the existing scale.

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

Warning: Removed 2 rows containing missing values (`geom\_point()`).



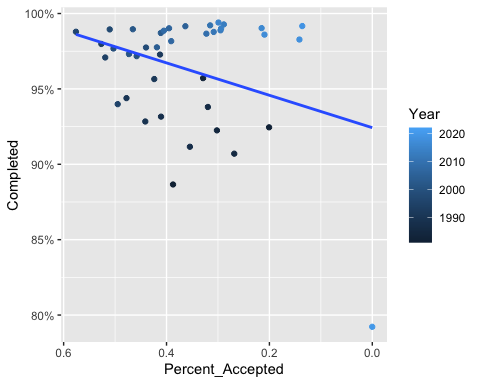
london\_marathon <- london\_marathon %>%  
 mutate(Percent\_Accepted = Accepted/Applicants)  
  
Combined <- left\_join(winners, london\_marathon, by = "Year")  
  
ggplot(london\_marathon, aes(x = Percent\_Accepted, y = Completed, color = Year)) +   
 geom\_point() +   
 ylim(0.85, 1.0) +   
 scale\_y\_continuous(labels = scales::percent) +  
 scale\_x\_continuous(labels = scales::percent) +  
 scale\_x\_reverse() +  
 geom\_smooth(method = 'lm', se = FALSE)

Scale for y is already present.  
Adding another scale for y, which will replace the existing scale.  
Scale for x is already present.  
Adding another scale for x, which will replace the existing scale.  
`geom\_smooth()` using formula = 'y ~ x'

Warning: Removed 2 rows containing non-finite values (`stat\_smooth()`).

Warning: The following aesthetics were dropped during statistical transformation: colour  
ℹ This can happen when ggplot fails to infer the correct grouping structure in  
 the data.  
ℹ Did you forget to specify a `group` aesthetic or to convert a numerical  
 variable into a factor?

Warning: Removed 2 rows containing missing values (`geom\_point()`).



averages <- london\_marathon %>%  
 filter(Year >= 1980 & Year <= 1999) %>%  
 summarize(first = mean(Completed))  
  
averages

# A tibble: 1 × 1  
 first  
 <dbl>  
1 0.948

averages <- london\_marathon %>%  
 filter(Year >= 2000 & Year <= 2019) %>%  
 summarize(second = mean(Completed))  
  
averages

# A tibble: 1 × 1  
 second  
 <dbl>  
1 0.987