PSTAT 10 Homework 5

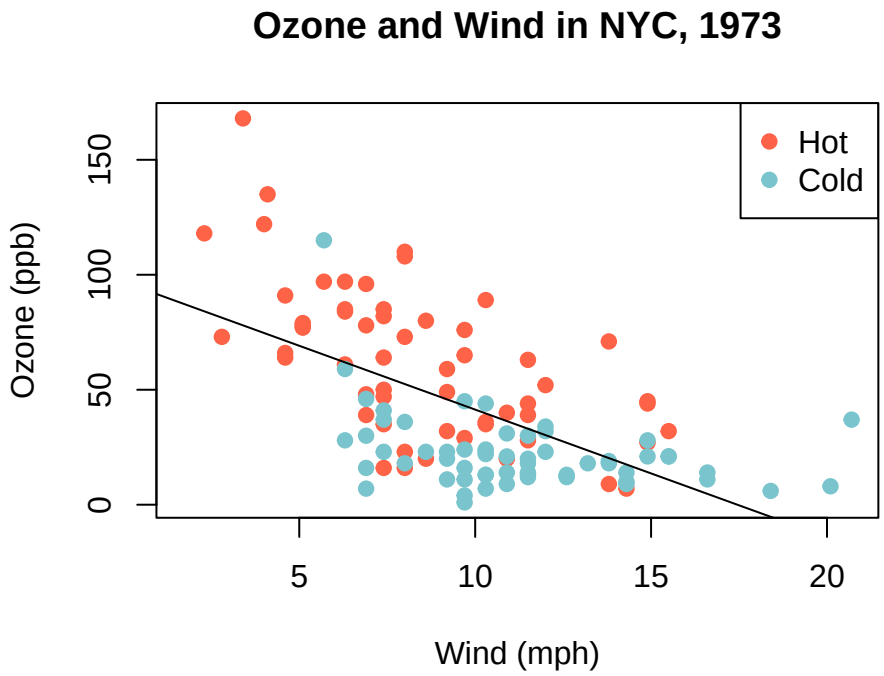
Due 7/26/22

You will need the `tidyverse` metapackage for this homework.

```
library(tidyverse)
```

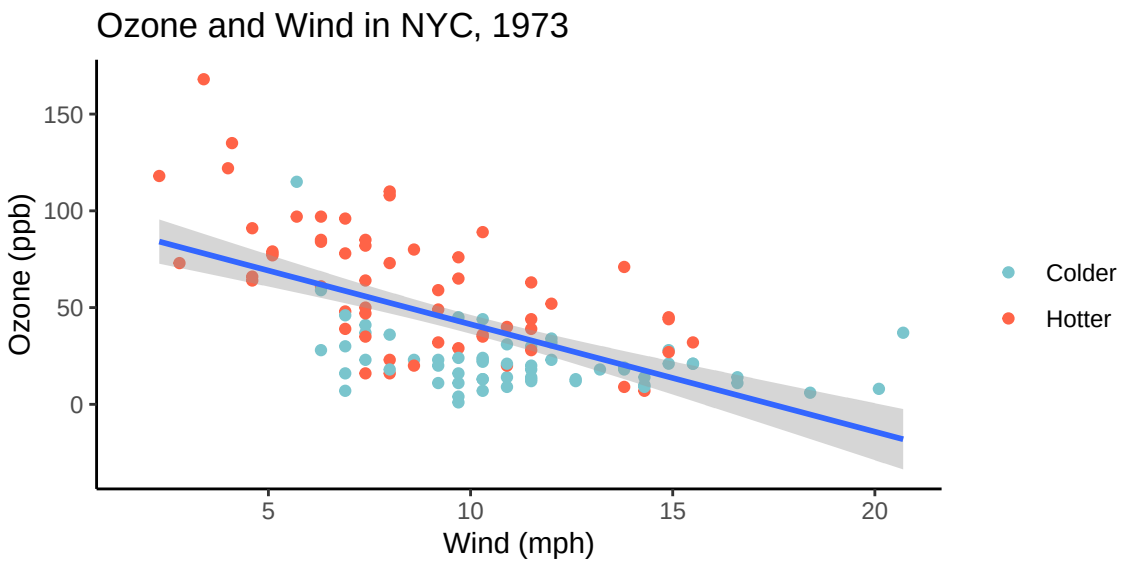
Problem 1: Airquality

In lecture 7 (slide 18), we created the following base R plot of ozone against wind and included a trend line.



Recreate this plot with `ggplot` and match my provided output.

For full credit, match the output exactly (not counting the dimensions of the overall figure.)



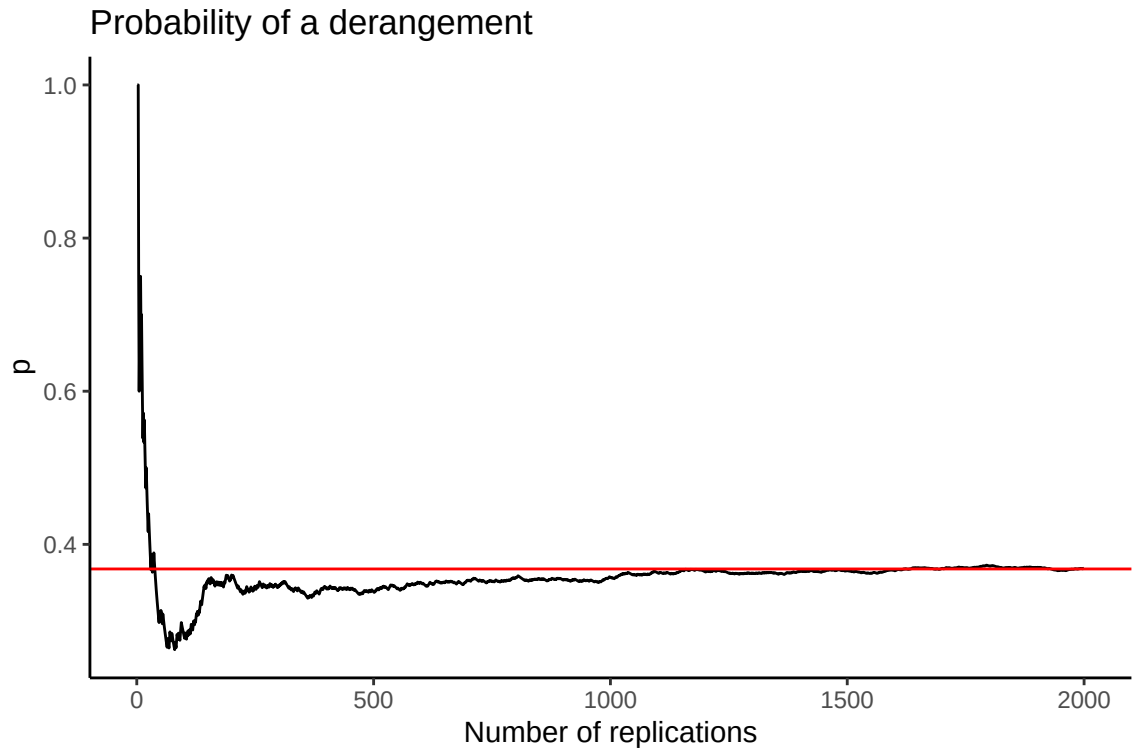
To get the right colors, you may use the following layer:

```
scale_color_manual(values = c("Colder" = "cadetblue3", "Hotter" = "tomato"))
```

Problem 2: Derangement

In lecture 8, we plotted the approximate probability that a permutation of 100 elements is a derangement.

Recreate the plot in `ggplot` (shown below), using 2000 replications. Your plot will look different due to randomness.



Problem 3: World Health Organization

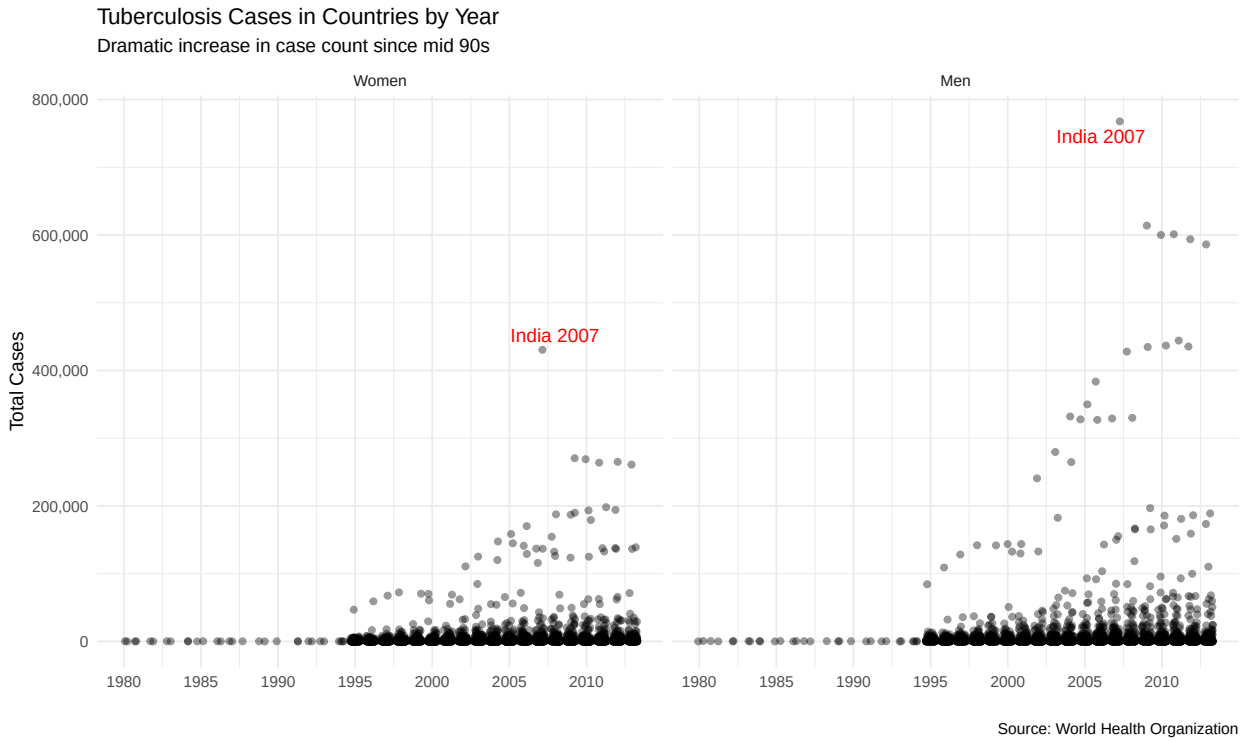
In lecture 18 we tidied the `tidyr::who` dataset. The first few rows look like this:

```
head(who_tidy, 4)
```

```
## # A tibble: 4 x 6
##   country      year var   sex   age   cases
##   <chr>       <int> <chr> <chr> <chr> <int>
## 1 Afghanistan  1997 sp    m     014       0
## 2 Afghanistan  1997 sp    m     1524     10
## 3 Afghanistan  1997 sp    m     2534      6
## 4 Afghanistan  1997 sp    m     3544      3
```

Part 1 For each country, year, and sex compute the total number of cases of TB. Put the result into a tibble with 4 columns.

Part 2 Create the following plot with `ggplot`. For full credit, match the details exactly, other than the overall dimensions of the figure and the positioning of the labels of the outlier.



Hint: To better see the overlapping points, instead of `geom_point` I used `geom_jitter` with `width = 0.3`. The following parameters in certain layers may also be helpful:

```
labeller = labeller(sex = c("f" = "Women", "m" = "Men"))
labels = scales::label_comma()
breaks = seq(1980, 2015, by = 5)
```

Problem 4: Pew Research Center

The following is data from the Pew Research Center about religion and income. It is part of the `tidyr` package which is part of the `tidyverse` metapackage.

```
relig_income
```

```
## # A tibble: 18 x 11
##    religion `<$10k` `$10-20k` `$20-30k` `$30-40k` `$40-50k` `$50-75k` `$75-100k`
##    <chr>      <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>      <dbl>
##  1 Agnostic      27        34        60        81        76       137        122
##  2 Atheist       12        27        37        52        35        70         73
##  3 Buddhist      27        21        30        34        33        58         62
##  4 Catholic     418       617       732       670       638      1116        949
##  5 Don't k~      15        14        15        11        10        35         21
##  6 Evangel~     575       869      1064       982       881      1486        949
##  7 Hindu          1         9         7         9        11        34         47
##  8 Histori~     228       244       236       238       197       223        131
##  9 Jehovah~      20        27        24        24        21        30         15
## 10 Jewish        19        19        25        25        30        95         69
## 11 Mainlin~     289       495       619       655       651      1107        939
## 12 Mormon        29        40        48        51        56       112         85
## 13 Muslim         6         7         9        10         9        23         16
## 14 Orthodox      13        17        23        32        32        47         38
## 15 Other C~       9         7        11        13        13        14         18
## 16 Other F~      20        33        40        46        49        63         46
## 17 Other W~       5         2         3         4         2         7          3
## 18 Unaffil~     217       299       374       365       341       528        407
## # ... with 3 more variables: `$100-150k` <dbl>, `>150k` <dbl>,
## #   `Don't know/refused` <dbl>
```

Part 1 In a short sentence or two, explain why this dataset is not tidy.

Part 2 Tidy the dataset and store the result in `relig_income_tidy`. First few rows of the result are provided.

```
head(relig_income_tidy, 4)
```

```
## # A tibble: 4 x 3
##   religion income  frequency
##   <chr>    <chr>       <dbl>
## 1 Agnostic <$10k          27
## 2 Agnostic $10-20k        34
## 3 Agnostic $20-30k        60
## 4 Agnostic $30-40k        81
```

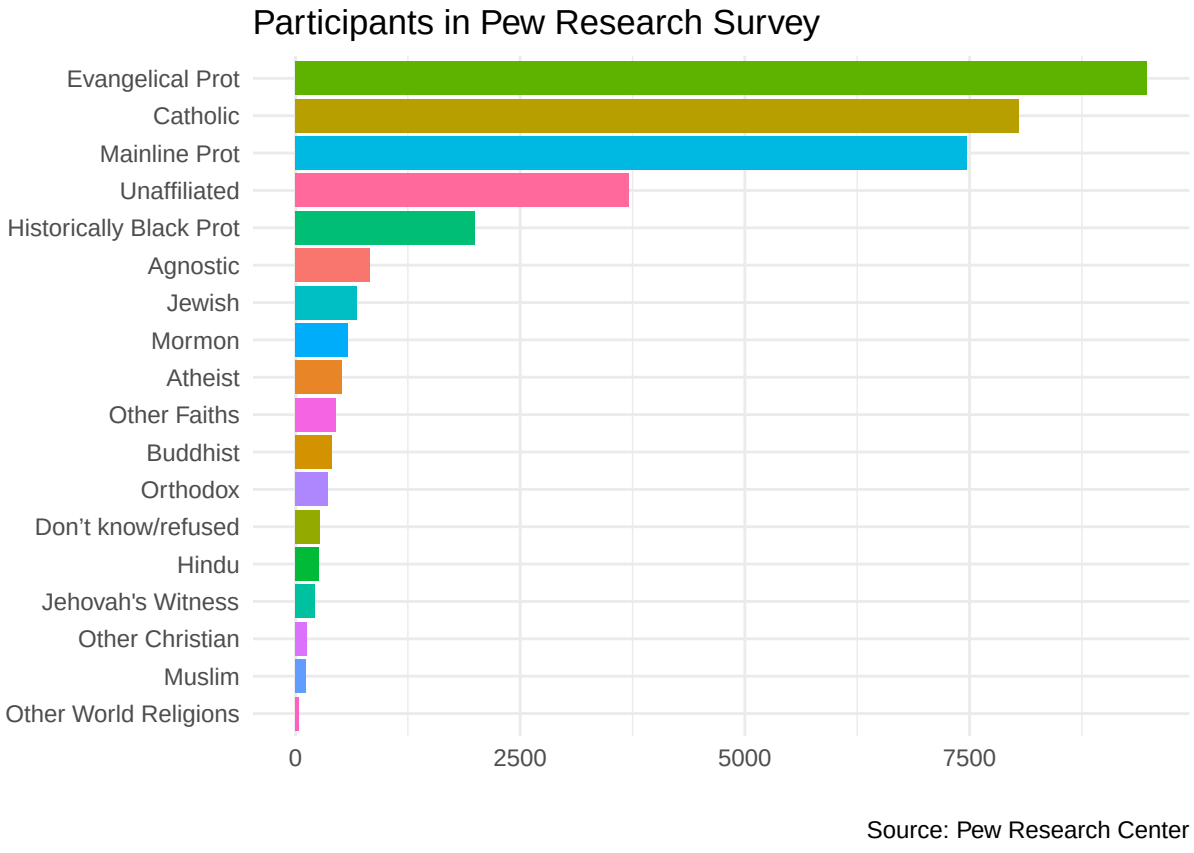
Hint: Notice that many column names are quoted, e.g. `$10-20k`. This is necessary since special characters like $ are not allowed in column names in R. You can access the column name with the quotes like any other name, for example:

```
relig_income[1:3, "$10-20k"]
```

```
## # A tibble: 3 x 1
##   `$10-20k`
##       <dbl>
## 1        34
## 2        27
## 3        21
```

Actually, the characters displayed in the above output are not quotes but *backticks*, same key as the tilde ~ on your keyboard.

Part 3 Create the following plot in `ggplot`. For full credit, match the plot exactly, not counting the overall dimensions of the figure. It is also okay if the colors are different, but the bars must have different colors.



Hint: I used the `reorder` function to order the bars.