PSTAT 10 Homework 5

Sou Hamura

2024-07-30

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

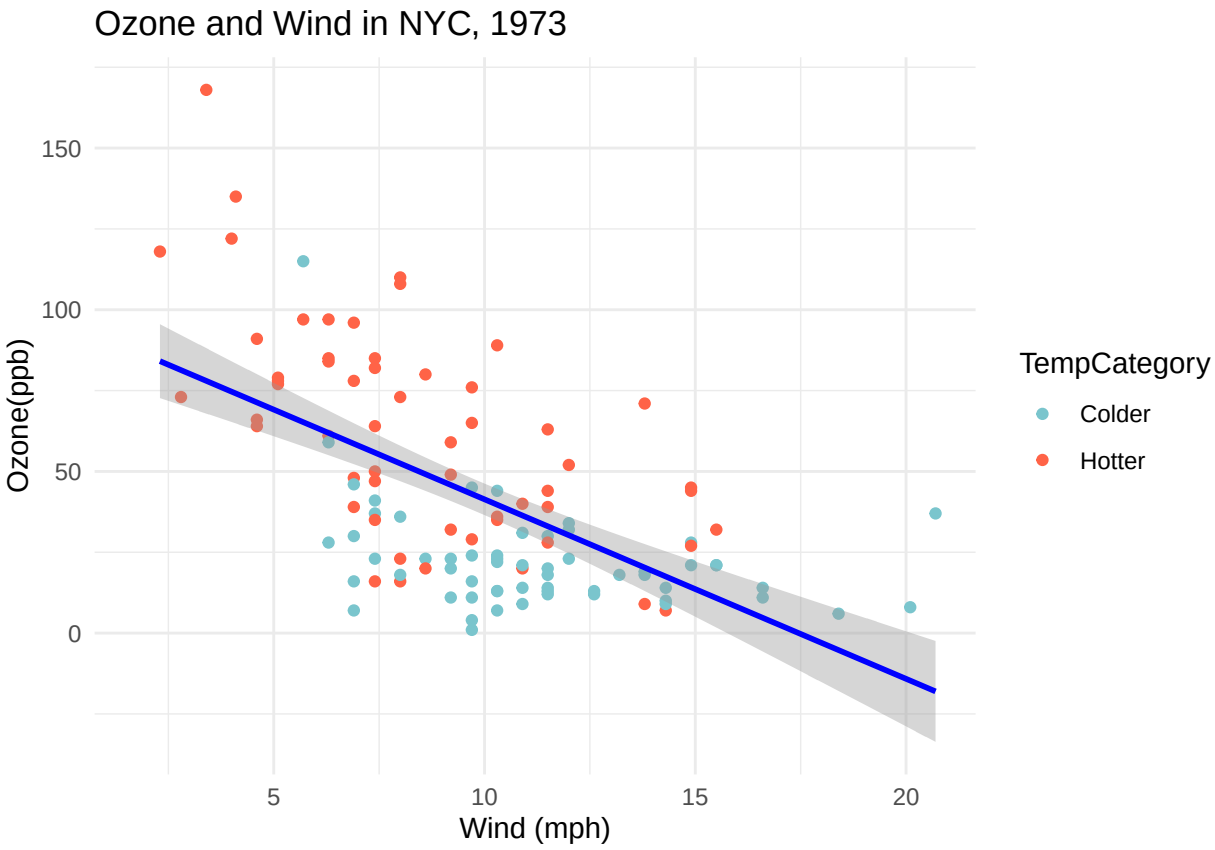
Problem 1: Airquality

```
threshold <- median(airquality$Temp, na.rm = TRUE)
airquality$TempCategory <- ifelse(airquality$Temp > threshold, "Hotter", "Colder")
airquality$TempCategory <- factor(airquality$TempCategory)
threshold
```

```
## [1] 79
```

```
ggplot(airquality, aes(x = Wind, y = Ozone, color = TempCategory)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Ozone and Wind in NYC, 1973",
       x = "Wind (mph)",
       y = "Ozone(ppb)") +
  scale_color_manual(values = c("Colder" = "cadetblue3", "Hotter" = "tomato")) +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    axis.line = element_line(color = "black"),
    panel.background = element_blank(),
    legend.title = element_blank()
  ) +
  theme_minimal()
```

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

## Warning: Removed 37 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 37 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

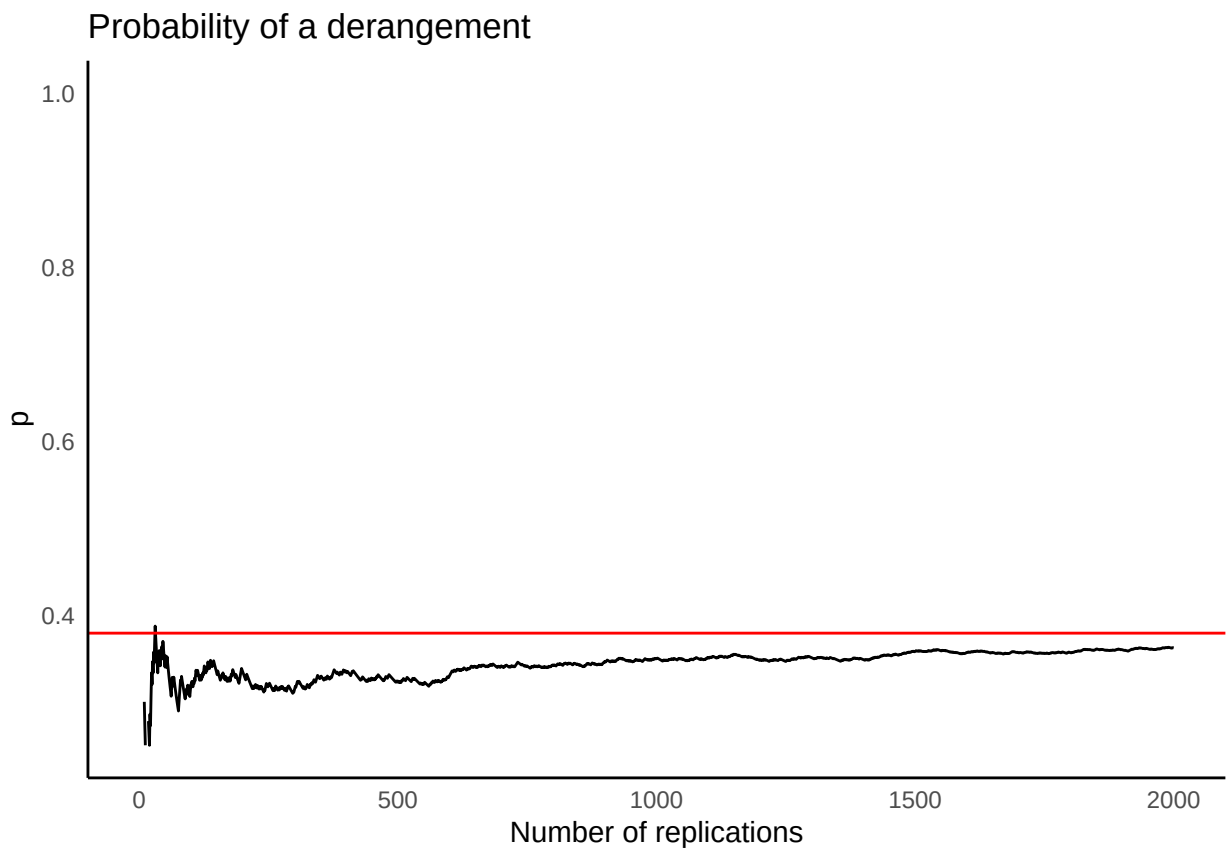


Problem 2: Derangement

```r
# from lecture 8
x <- 1:100
is_deranged <- function(){
  count <- 0
  x_val <- sample(x)
  for(i in seq_along(x)){
    if(x_val[i] == i){
      count = count + 1
    }
  }
  return(count <= 0)
}
```

```r
mean <- mean(replicate(2000, is_deranged()))
result <- replicate(2000, is_deranged())
result_avg <- cumsum(result) / 1:2000
result_data <- data.frame(x = 1:2000, y = result_avg)

ggplot(result_data, aes(x = x, y = y)) +
  geom_line() +
  labs(title = "Probability of a derangement",
       y = "p",
       x = "Number of replications") +
  ylim(0.25, 1) +
  geom_hline(yintercept = mean,color = "red") +
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    axis.line = element_line(color = "black"),
    panel.background = element_blank(),
    legend.title = element_blank())
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



Problem 3: World Health Organization

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.

```
# from lecture 18

who1 <- who |> pivot_longer(cols = new_sp_m014:newrel_f65,
                    names_to = "key",
                    values_to = "cases",
                    values_drop_na = TRUE)
who2 <- who1 |> mutate(key = stringr::str_replace(key, "newrel", "new_rel"))
who3 <- who2 |> separate(key, c("new", "type", "sexage"), sep = "_")
who4 <- who3 |> select(-new, -iso2, -iso3)
who5 <- who4 |> separate(sexage, c("sex", "age"), 1)
who_tidy <- who |>
  pivot_longer(cols = new_sp_m014:newrel_f65,
               names_to = "key",
               values_to = "cases",
               values_drop_na = TRUE) |>
  mutate(key = stringr::str_replace(key, "newrel", "new_rel")) |>
  separate(key, c("new", "type", "sexage"), sep = "_") |>
  select(-new, -iso2, -iso3) |>
  separate(sexage, c("sex", "age"), 1)
```

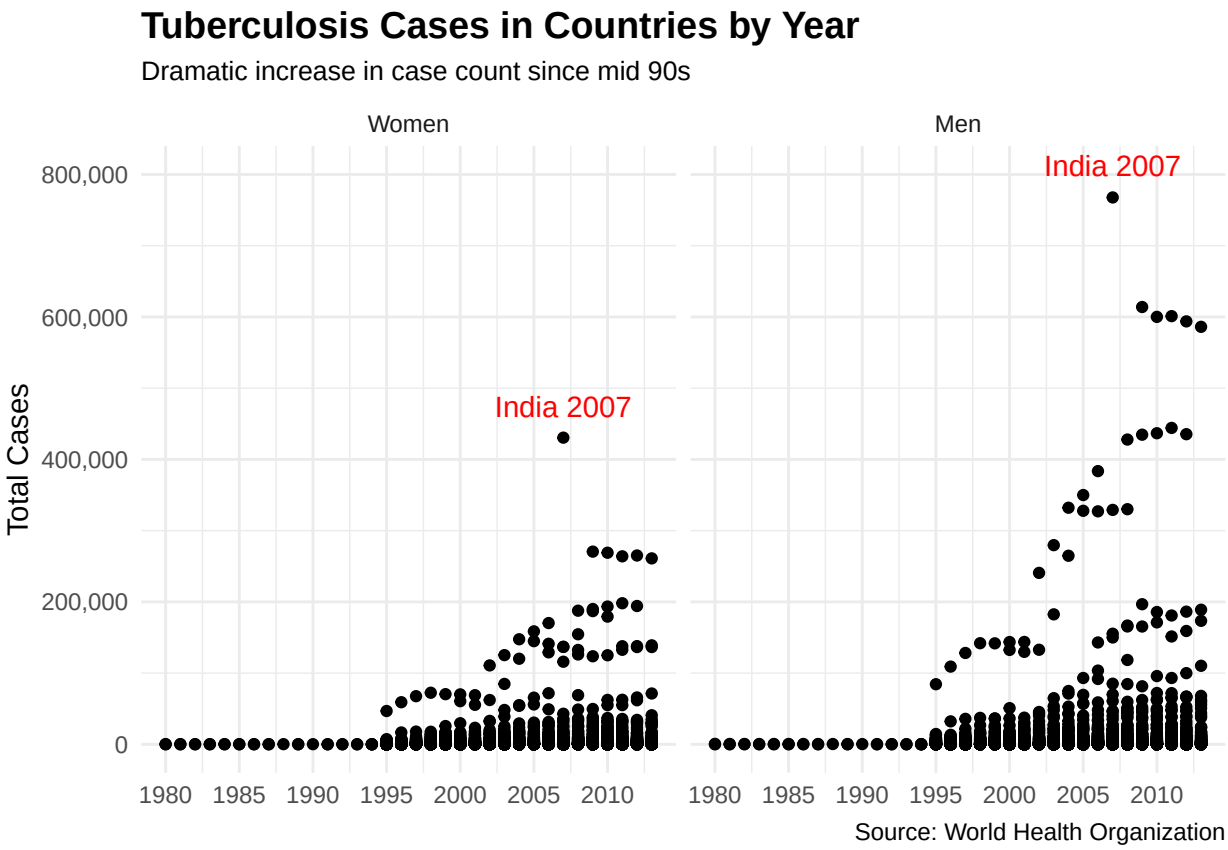
```
catplot <- who_tidy |>
  group_by(country, year, sex) |>
  summarize(cases = sum(cases))
```

```
## `summarise()` has grouped output by 'country', 'year'. You can override using
## the `.groups` argument.
```

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.

```
#Part 2
ggplot(catplot,aes(x = year, y = cases)) +
  geom_point() +
  facet_wrap(~sex,labeller = labeller(sex = c("f" = "Women", "m" = "Men"))) +
  scale_x_continuous(breaks = seq(1980, 2015, by = 5)) +
  scale_y_continuous(labels = scales::label_comma(), limits = c(0, 800000)) +
  labs(
    title = "Tuberculosis Cases in Countries by Year",
    subtitle = "Dramatic increase in case count since mid 90s",
    y = "Total Cases",
    x = NULL,
```

```
  caption = "Source: World Health Organization"
) +
geom_text(data = subset(catplot, country == "India" & year == 2007),
          aes(label = paste(country, year)),
          color = "red",
          vjust = -1) +
theme_minimal() +
theme(
  legend.position = 'none',
  plot.title = element_text(face = "bold", size = 14),
  plot.subtitle = element_text(size = 10)
)
```



Problem 4: Pew Research Center

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

because each row contains multiple observations while each row should only represent a single observation in a tidy dataset

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

```
relig_income_tidy <- relig_income |>
  pivot_longer(cols = -religion, names_to = "income", values_to = "frequency")
relig_income_tidy
```

```
## # A tibble: 180 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
##  5 Agnostic $40-50k                   76
##  6 Agnostic $50-75k                  137
##  7 Agnostic $75-100k                 122
##  8 Agnostic $100-150k                109
##  9 Agnostic >150k                     84
## 10 Agnostic Don't know/refused        96
## # i 170 more rows
```

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.

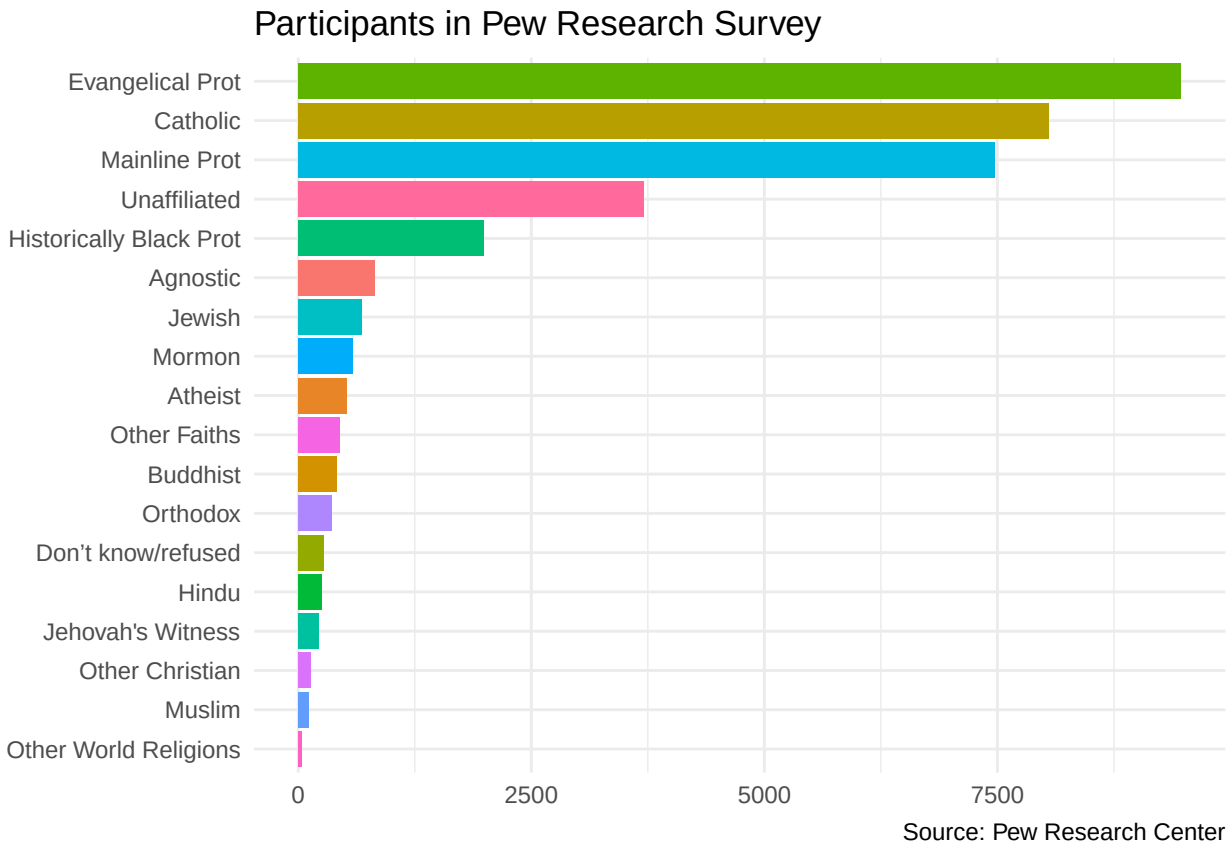
```
ri_graph <- relig_income_tidy |>
  group_by(religion) |>
  summarise(total_frequency = sum(frequency)) #, na.rm = TRUE

ri_graph <- ri_graph[order(ri_graph$total_frequency, decreasing = TRUE), ]

ri_graph
```

```
## # A tibble: 18 x 2
```

```
##    religion                total_frequency
##    <chr>                             <dbl>
##  1 Evangelical Prot                   9472
##  2 Catholic                           8054
##  3 Mainline Prot                      7470
##  4 Unaffiliated                       3707
##  5 Historically Black Prot            1995
##  6 Agnostic                            826
##  7 Jewish                              682
##  8 Mormon                              581
##  9 Atheist                             515
## 10 Other Faiths                        449
## 11 Buddhist                            411
## 12 Orthodox                            363
## 13 Don't know/refused                  272
## 14 Hindu                               257
## 15 Jehovah's Witness                   215
## 16 Other Christian                     129
## 17 Muslim                              116
## 18 Other World Religions                42
```

```
ggplot(ri_graph, mapping = aes(y = reorder(religion, total_frequency), x = total_frequency, fill = relig
  geom_col(position = "dodge") +
  labs(title = "Participants in Pew Research Survey",
       caption = "Source: Pew Research Center") +
  theme_minimal() +
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank()) +
  guides(fill = FALSE)
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```


Participants in Pew Research Survey
Evangelical Prot
Catholic
Mainline Prot
Unaffiliated
Historically Black Prot
Agnostic
Jewish
Mormon
Atheist
Other Faiths
Buddhist
Orthodox
Don’t know/refused
Hindu
Jehovah's Witness
Other Christian
Muslim
Other World Religions
0
2500
5000
7500
Source: Pew Research Center