homework5

2024-07-29

```
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
```

```
library(ggplot2)
```

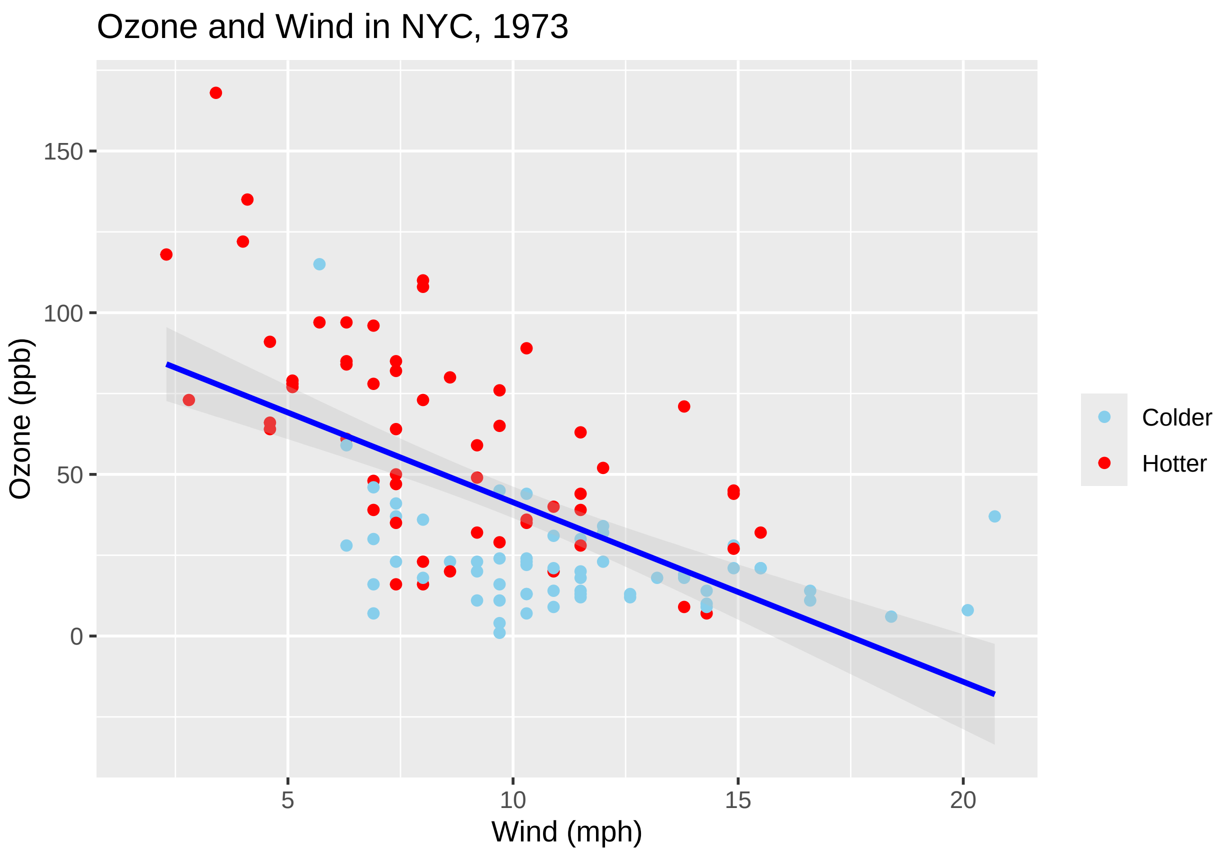
Problem 1

```
library(datasets)
airquality$Temperature <- ifelse(airquality$Temp > median(airquality$Temp, na.rm = TRUE), "Hotter", "Col
ggplot(airquality, aes(x = Wind, y = Ozone, color = Temperature)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue", fill = "gray", alpha = 0.3) +
  scale_color_manual(values = c("Hotter" = "red", "Colder" = "skyblue")) +
  labs(title = "Ozone and Wind in NYC, 1973",
       x = "Wind (mph)",
       y = "Ozone (ppb)",
       color = "")
```

```
## `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()`).
```



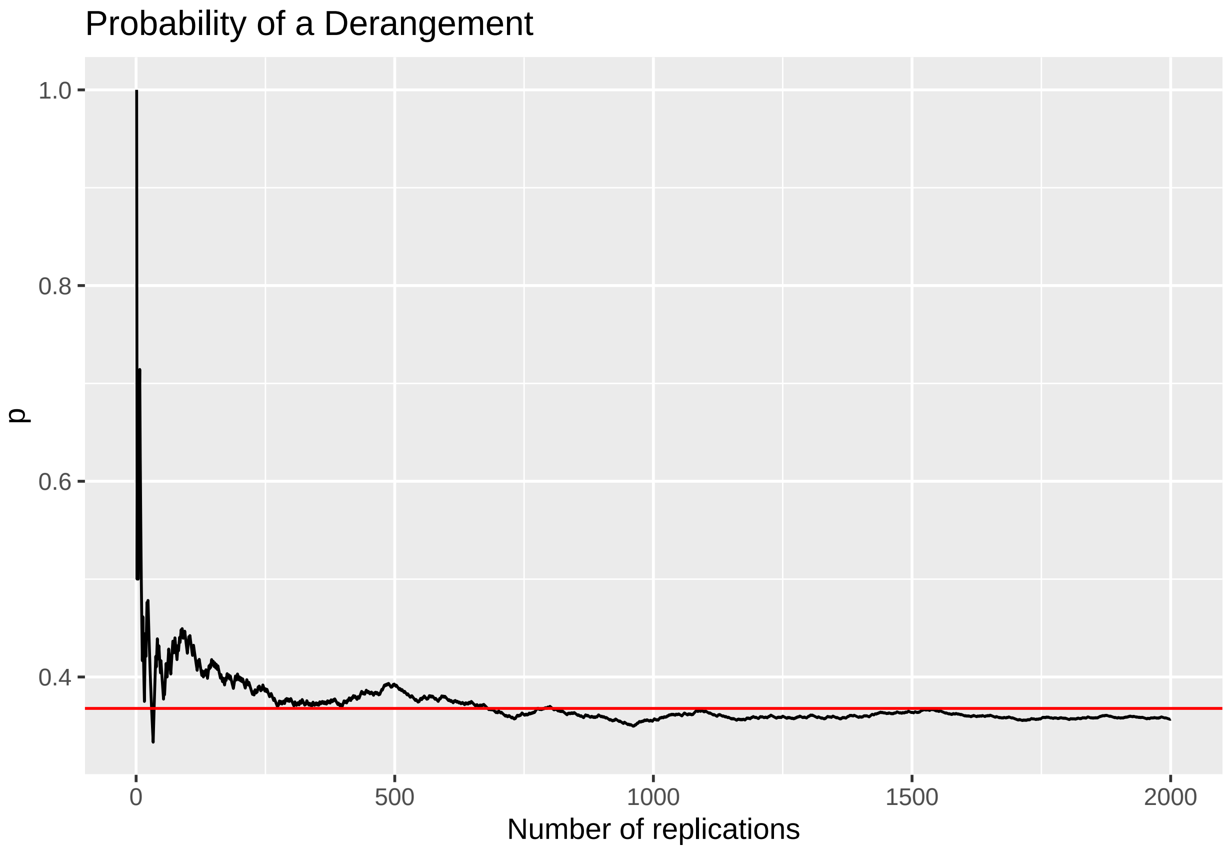
Problem 2

```
is_derangement <- function(permutation) {
  all(permutation != 1:length(permutation))
}
set.seed(123)
n <- 100
num_replications <- 2000
results <- numeric(num_replications)

for (i in 1:num_replications) {
  permutation <- sample(1:n, n)
  results[i] <- is_derangement(permutation)
}
cumulative_prob <- cumsum(results) / (1:num_replications)

df <- data.frame(
  Replications = 1:num_replications,
  Probability = cumulative_prob
)

ggplot(df, aes(x = Replications, y = Probability)) +
  geom_line() +
  geom_hline(yintercept = 1/exp(1), color = "red") +
  labs(title = "Probability of a Derangement",
       x = "Number of replications",
       y = "p")
```



Problem 3

```
library(tidyverse)
data("who")
who_tidy <- who %>%
  gather(key = "key", value = "cases", -country, -iso2, -iso3, -year) %>%
  separate(key, into = c("new", "type", "sexage"), sep = "_", extra = "merge", fill = "right") %>%
  separate(sexage, into = c("sex", "age"), sep = 1, fill = "right") %>%
  select(country, year, type, sex, age, cases) %>%
  filter(!is.na(cases))


who_tidy <- who_tidy %>%
  filter(sex %in% c("f", "m"))

tb_totals <- who_tidy %>%
  group_by(country, year, sex) %>%
  summarize(total_cases = sum(cases, na.rm = TRUE))
```

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

```
head(tb_totals)
```

```
## # A tibble: 6 x 4
## # Groups:   country, year [3]
##   country      year sex   total_cases
##   <chr>       <dbl> <chr>       <dbl>
## 1 Afghanistan  1997 f             102
## 2 Afghanistan  1997 m              26
## 3 Afghanistan  1998 f            1207
## 4 Afghanistan  1998 m             571
## 5 Afghanistan  1999 f             517
## 6 Afghanistan  1999 m             228
```

```
p <- ggplot(tb_totals, aes(x = year, y = total_cases, color = sex)) +
  geom_jitter(width = 0.3, alpha = 0.5) +
  scale_color_manual(values = c("f" = "black", "m" = "black")) +
  facet_wrap(~ sex, labeller = labeller(sex = c("f" = "Women", "m" = "Men"))) +
  labs(title = "Tuberculosis Cases in Countries by Year",
       subtitle = "Dramatic increase in case count since mid 90s",
       x = "",
       y = "Total Cases",
       color = "") +
  scale_y_continuous(labels = scales::label_comma()) +
  scale_x_continuous(breaks = seq(1980, 2015, by = 5)) +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        strip.text = element_text(face = "bold"),
        legend.position = "none")

india_2007_f <- tb_totals %>% filter(country == "India" & year == 2007 & sex == "f") %>% summarize(max_
```

```
## ‘summarise()‘ has grouped output by ’country’. You can override using the
## ‘.groups‘ argument.
```

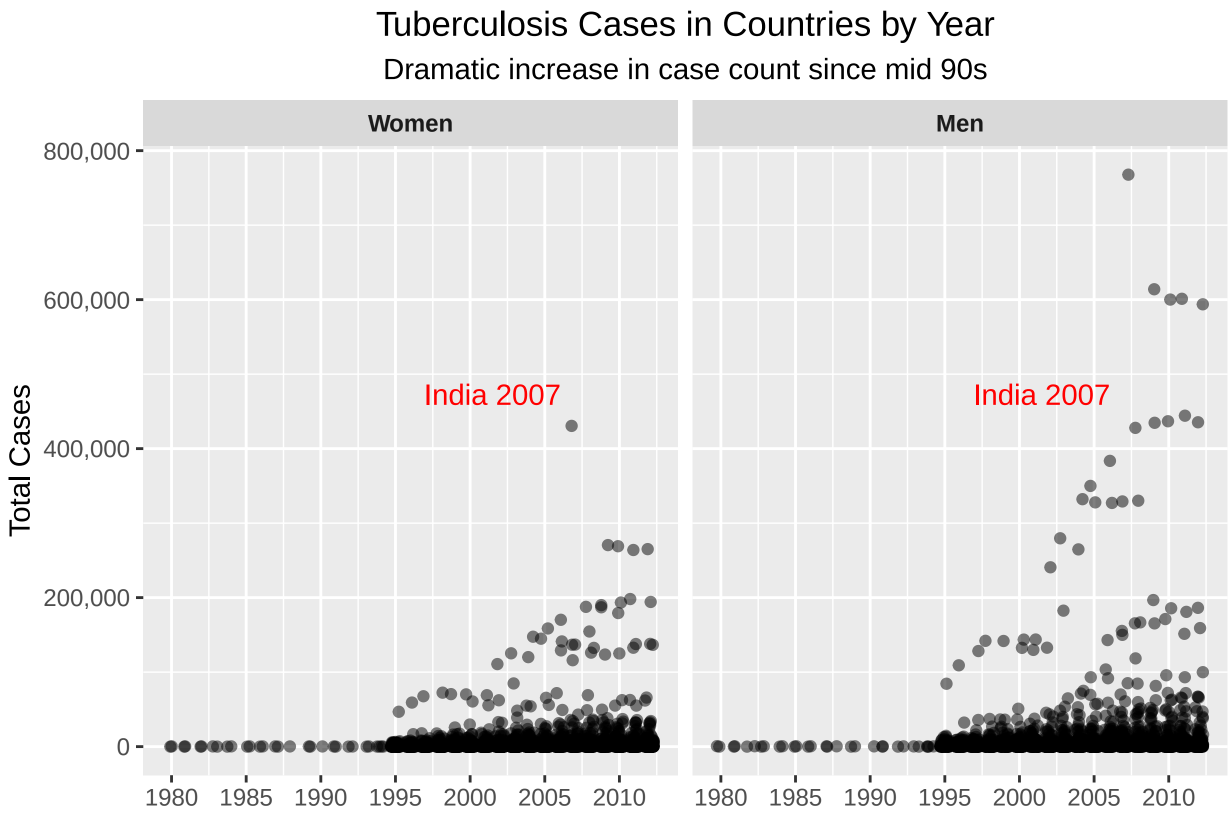
```
india_2007_m <- tb_totals %>% filter(country == "India" & year == 2007 & sex == "m") %>% summarize(max_
```

```
## ‘summarise()‘ has grouped output by ’country’. You can override using the
## ‘.groups‘ argument.
```

```
p <- p + annotate("text", x = 2007, y = india_2007_f$max_cases,
                  label = "India 2007", color = "red", vjust = -1, hjust = 1.1, size = 4)
  annotate("text", x = 2007, y = india_2007_m$max_cases,
           label = "India 2007", color = "red", vjust = -1, hjust = 1.1, size = 4)
```

```
## mapping: x = ~x, y = ~y
## geom_text: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

```
print(p)
```



Problem 4

1. Because they are so mu ch different number and it is not ordered. also the symbols are mess.

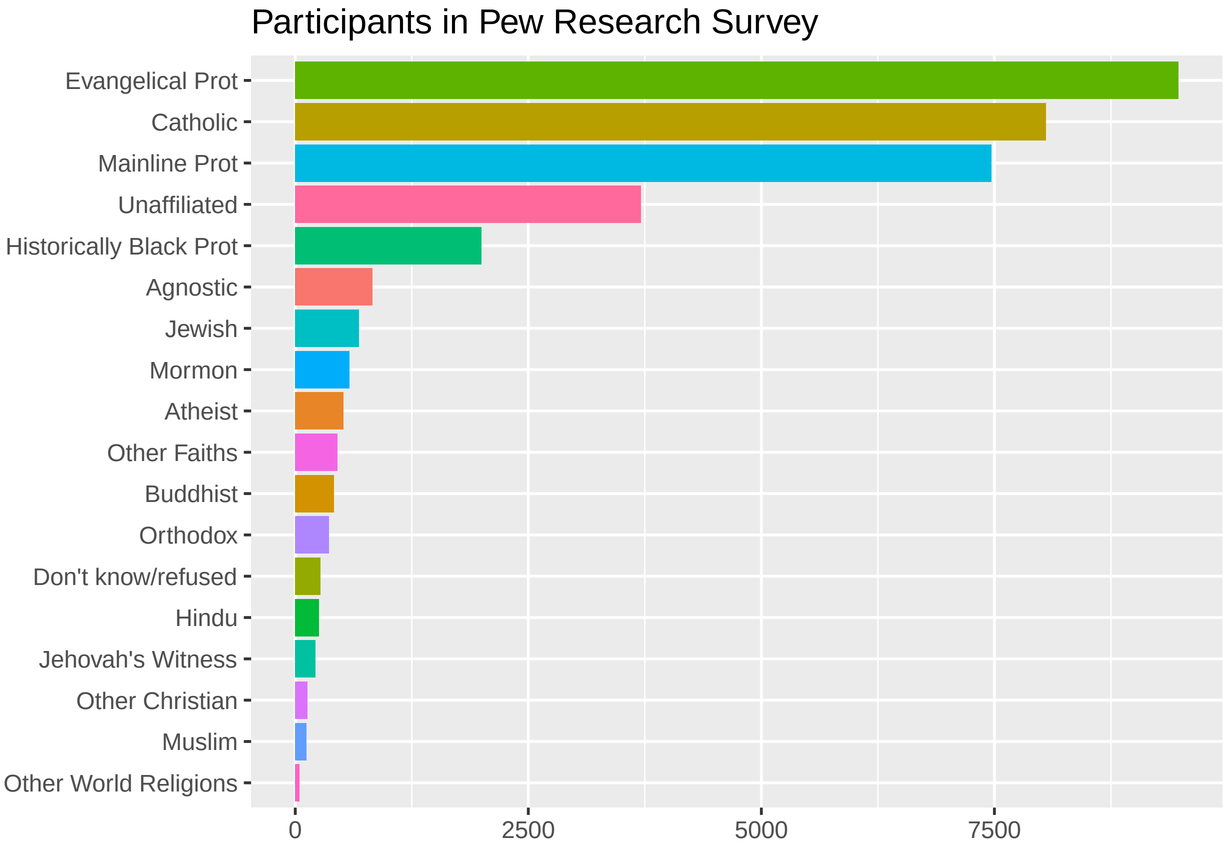
```
relig_income_tidy <- relig_income %>%
  pivot_longer(cols = -religion, names_to = "income_range", values_to = "count")
head(relig_income_tidy,4)
```

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

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

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

```
ggplot(relig_income_tidy, mapping=aes(x = count, y = reorder(religion, count, FUN = sum), fill = religi
  geom_col(show.legend = FALSE) +
  labs(title = "Participants in Pew Research Survey",
       x = NULL,
       y = NULL,)
```


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