Lab 2

Due: 11:59 PM on 6/15/2023

kaori hirano

2023-06-10

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- 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 becor
```

```
library(lubridate)
```

Part 1

Q1

```
date <- mdy(c(071220, 71320, 71420, 71520, 71620,71720,71820,71920, 72020,72120,72220,7232

mask <- (c(25.5, 19.5, 19.5, 20.5, 19.5, 19.5, 20.5, 20, 20.5, 21.4, 19.5, 19.5, 20.5, 19,

noMask <- (c(9.9, 9.1, 9.3, 9.9, 9.9, 9.6, 9.6, 9, 8.5, 8.6, 8.5, 9.8, 9.9, 10, 9.8, 9.8,
```

```r
(cbind(length(noMask), length(mask), length(date)))
```

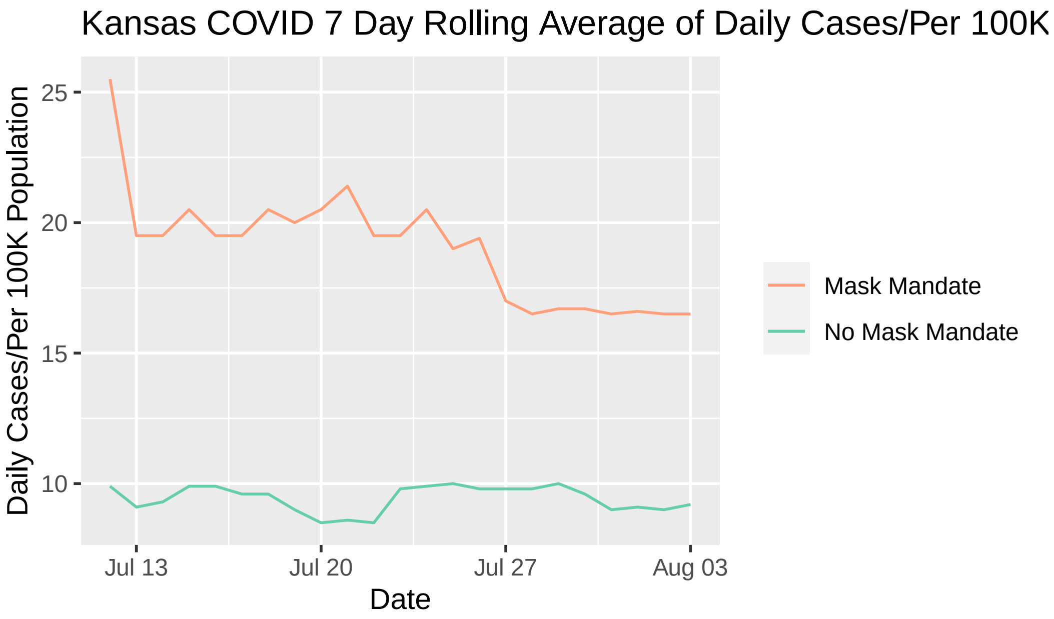
```
     [,1] [,2] [,3]
[1,]   23   23   23
```

```r
d <- data.frame(time(date), mask, noMask)
```

Q2

```r
#ggplot(d,
#       aes(x = date, y = caseNumber, color = maskStatus, group = maskStatus)) +
#  geom_line() +
#  facet_wrap(~ subgroup)

ggplot(d, aes(x = date)) +
  geom_line(aes(y = mask, color = 'Mask Mandate')) +
  geom_line(aes(y = noMask, color = 'No Mask Mandate')) +
  scale_color_manual(values = c("lightsalmon1", "aquamarine3")) +
  labs(
    x = "Date", y = "Daily Cases/Per 100K Population",
    color = NULL,
    title = "Kansas COVID 7 Day Rolling Average of Daily Cases/Per 100K Population by Mask
    caption = "Source: Kansas Department of Health and Environment"
  )
```



Source: Kansas Department of Health and Environment

Q3

In this visual it is more clear that there is a higher incidence of covid cases per 100K people in mask mandate counties than non mask mandate counties from July 11-August 3 2020 in Kansas.

Q4

It doesn't necessarily give us useful information regarding mask wearing and COVID that can be used based on this alone or immedianteley, but it gives evidence that mask mandates may not be as effective at preventing the spread of covid as expected given the lower incidence rates of non-mask mandated counties. It invites further research. One thing I would be interested to see included is the incidence rate by county when total population and/or population density is included.

...

Part 2

Q1

```
p <- read.csv('data/paygap.csv') |>
  janitor::clean_names()        # just for practice
typeof(p$due_date)
```

```
[1] "character"
```

```
p <- p |> mutate(year = year(as_datetime(due_date)) - 1)
# p$year checking it worked
```

Q2

```
count(p, vars = year) #, wt_var = NULL) # gets reports per year
```

```
  vars     n
1 2017 10230
2 2018 10471
3 2019  6924
4 2020 10536
5 2021 10425
6 2022   125
```

```
length(p$employer_id) - length(unique(p$employer_id)) # subtracts all employers
```

```
[1] 35963
```

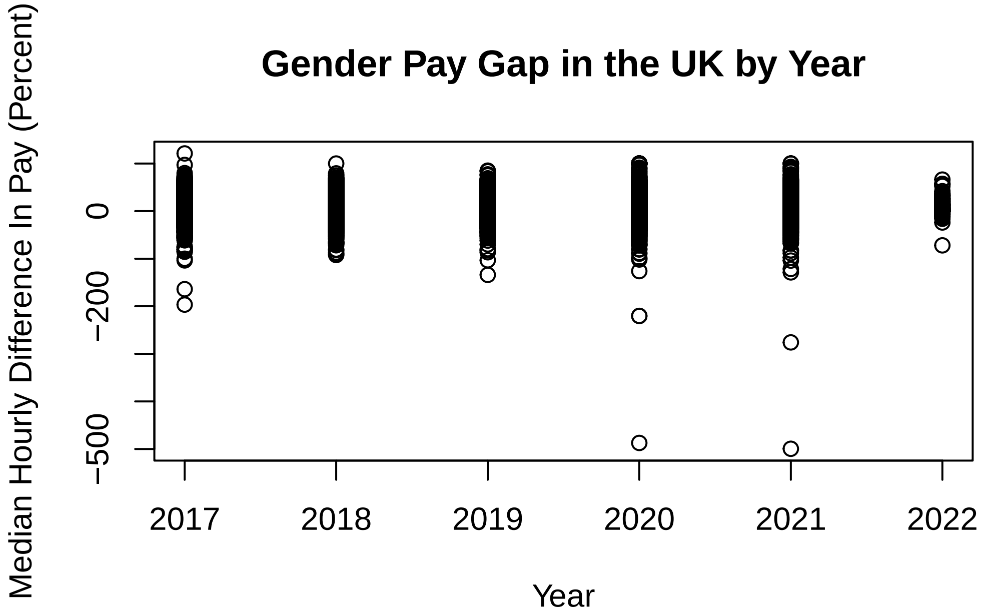
```
# from duplicates removed employers list length
```

There are far more observations in 2017-2021 than in 2022. 2019 also has fewer, by only by around 3000 rather than 9000 responses. This could be because of the year of data we are using. After looking at the codebook, it shows that the most recent one is 2023-2024. This suggests that the employers listed in 2022 may have reported earlier than the deadline so their numbers for the upcoming year are already listed.

There are 35963 unique companies reporting in the dataset.

Q3

```
plot <- plot(p$year, p$diff_median_hourly_percent,
             xlab = 'Year',
             ylab = 'Median Hourly Difference In Pay (Percent)',
             main = 'Gender Pay Gap in the UK by Year')
```



Based on the visualization, there does appear to be a (descriptive) decrease in the overall range of the gap, with more data being clustered toward more equal pay (toward 0) in later years, especially 2022. However, there is a reasonable chance that this is due to the lack of data, not necessarily a lower gap overall.

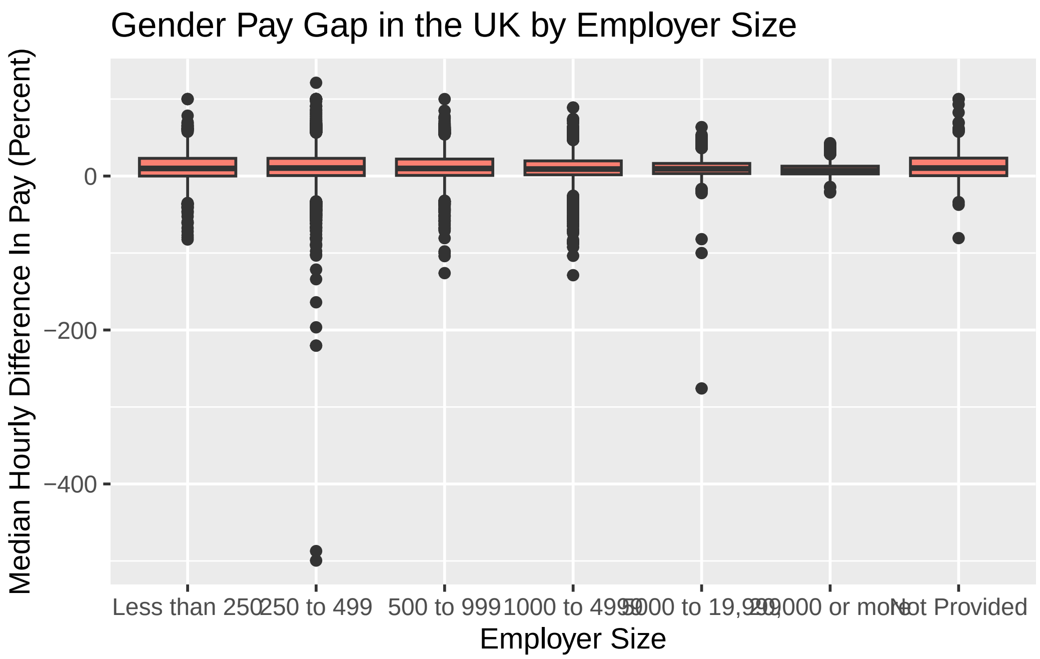
Q4

```
levels(as.factor(p$employer_size))
```

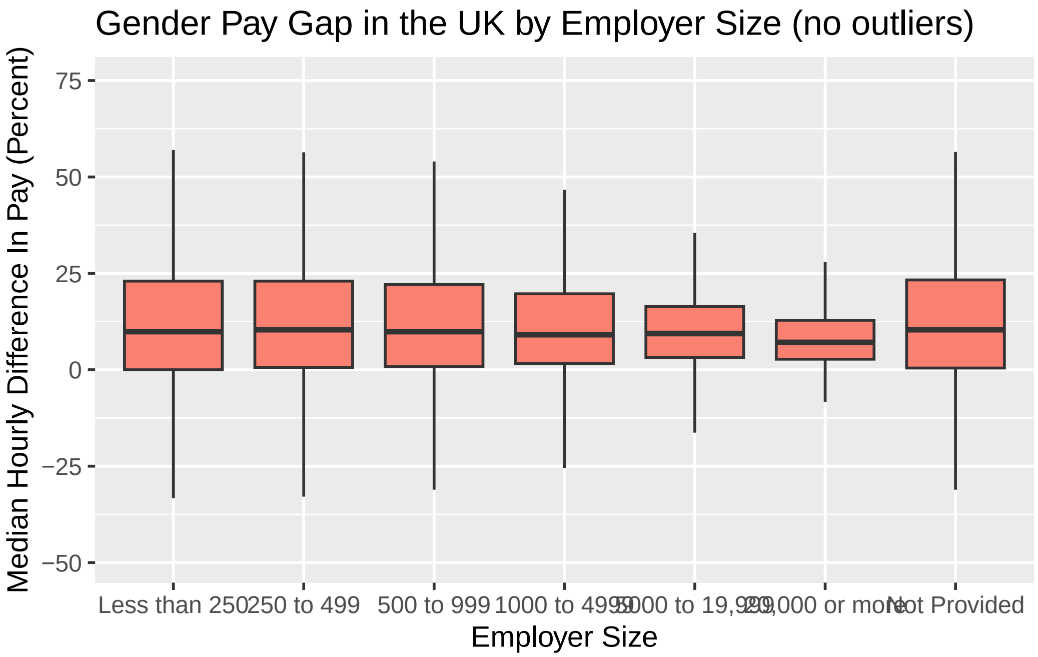
```
[1] "1000 to 4999"   "20,000 or more" "250 to 499"     "500 to 999"
[5] "5000 to 19,999" "Less than 250"  "Not Provided"
```

```
p <- p |>
  mutate(employer_size =
           fct_relevel(employer_size, c('Less than 250', '250 to 499',
                                         '500 to 999', '1000 to 4999',
                                         '5000 to 19,999', '20,000 or more',
                                         'Not Provided')))
# recoding so in order of inc size

(plot1 <- ggplot(p, aes(x = employer_size, y = diff_median_hourly_percent)) +
  geom_boxplot(fill = 'salmon') +
   labs(
    x = "Employer Size",
    y = 'Median Hourly Difference In Pay (Percent)',
    title = 'Gender Pay Gap in the UK by Employer Size'))
```



```
(plot2 <- ggplot(p, aes(x = employer_size, y = diff_median_hourly_percent)) +  #plot witho
  geom_boxplot(fill = 'salmon', outlier.shape = NA) +
  coord_cartesian(ylim=c(quantile(p$diff_median_hourly_percent, prob = .25) - 50, quantile
    labs(
     x = "Employer Size",
     y = 'Median Hourly Difference In Pay (Percent)',
     title = 'Gender Pay Gap in the UK by Employer Size (no outliers)'))
```



Descriptively speaking, there appears to be slightly less of a pay gap in employers with a size of 20,000 or more, as seen by the median which is slightly closer to zero than the other employer sizes, by around 5%.