Extra credit

INFO 2950 - Spring 2023

Elisabeth Pan

5/10/23

Setup

Load packages and data:

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.2     v purrr   1.0.0
v tibble  3.2.1     v dplyr   1.1.2
v tidyr   1.2.1     v stringr 1.5.0
v readr   2.1.3     v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
library(scales)
```

```
Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```
library(tidymodels)
```

```
-- Attaching packages --------------------------------------- tidymodels 1.0.0 --
v broom        1.0.2     v rsample      1.1.1
v dials        1.1.0     v tune         1.1.1
v infer        1.0.4     v workflows    1.1.2
v modeldata    1.0.1     v workflowsets 1.0.0
v parsnip      1.0.3     v yardstick    1.1.0
v recipes      1.0.6
-- Conflicts ------------------------------------------ tidymodels_conflicts() --
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
* Use tidymodels_prefer() to resolve common conflicts.
```

```
childcare_costs <- readr::read_csv(
  'https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2023/2023-05-
```

```
Rows: 34567 Columns: 61
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (61): county_fips_code, study_year, unr_16, funr_16, munr_16, unr_20to64...

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.
```

```
counties <- readr::read_csv(
  'https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2023/2023-05-
```

```
Rows: 3144 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): county_name, state_name, state_abbreviation
dbl (1): county_fips_code
```

```
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.
```

Introduction

This dataset provides information regarding childcare prices (for both family based and center-based childcare) in different states and counties of the US from 2008 to 2018.

```r
#join data
cdata <- inner_join(x = childcare_costs, y = counties)
```

```
Joining with `by = join_by(county_fips_code)`
```

```r
#CDC defined US regions
northeast <- c("Connecticut", "Maine", "Massachusetts", "New Hampshire",
               "New Jersey", "New York", "Pennsylvania", "Rhode Island",
               "Vermont")
midwest <- c("Illinois", "Indiana", "Iowa", "Kansas", "Michigan",
             "Minnesota", "Missouri", "Nebraska", "North Dakota",
             "Ohio", "South Dakota", "Wisconsin")
south <- c("Alabama", "Arkansas", "Delaware", "District of Columbia",
           "Florida", "Georgia", "Kentucky", "Louisiana", "Maryland",
           "Mississippi", "North Carolina", "Oklahoma", "South Carolina",
           "Tennessee", "Texas", "Virginia", "West Virginia")
west <- c("Alaska", "Arizona", "California", "Colorado", "Hawaii",
          "Idaho", "Montana", "Nevada", "New Mexico", "Oregon", "Utah",
          "Washington", "Wyoming")

#add column for each state's corresponding region
cdata <- cdata |>
  mutate(region=case_when(
    state_name %in% northeast ~ "Northeast",
    state_name %in% midwest ~ "Midwest",
    state_name %in% south ~ "South",
    state_name %in% west ~ "West"
  )) |>
  mutate(avg_costs=(mcsa*mfccsa)/2)

#child care costs in 2018
```

```r
cdata_2018 <- cdata |>
  filter(study_year=="2018") |>
  select(mhi_2018, mfccsa, mcsa, county_name, state_name, region, avg_costs)
```

Research Questions

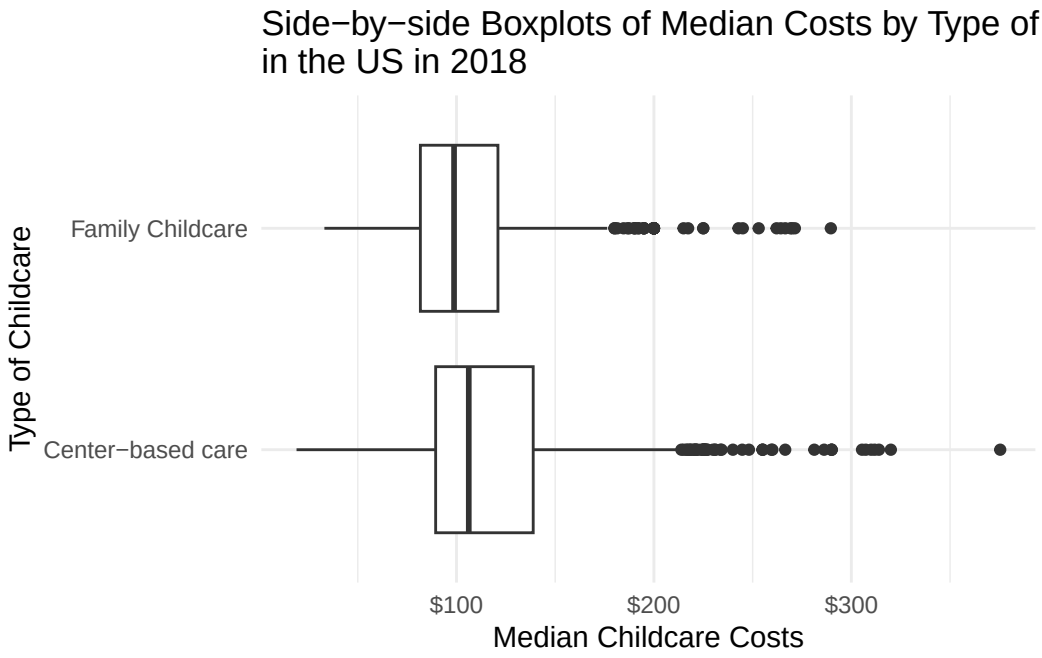
Question 1

1. What is the relationship between median family childcare costs and median center-based childcare costs in 2018 by US Region?

```r
cdata_2018_longer <- cdata_2018 |>
  pivot_longer(!c(mhi_2018, county_name, state_name, region, avg_costs),
               names_to = "childcare_type", values_to = "cost") |>
  mutate(childcare_type=ifelse(childcare_type=="mfccsa", "Family Childcare",
                               "Center-based care"))

ggplot(data=cdata_2018_longer,
         mapping=aes(x=cost, y=childcare_type)) +
  geom_boxplot() +
  scale_x_continuous(labels = label_dollar()) +
  labs(
    x="Median Childcare Costs",
    y="Type of Childcare",
    title="Side-by-side Boxplots of Median Costs by Type of Childcare
in the US in 2018",
  ) +
  theme_minimal()
```

```
Warning: Removed 1576 rows containing non-finite values (`stat_boxplot()`).
```



The purpose of this visualization is to examine the difference in costs of family childcare and center-based childcare. I chose to use a side-by-side boxplot as it is able to visually compare the difference in median childcare costs between the two types of child care. From this plot, it is evident that US families paid slightly more for center-based childcare than family childcare in 2018.

Question 2

2. What is the relationship between median household income and median childcare costs in 2018 by US Region?

```
income_cc_fit <- linear_reg() |>
  fit(avg_costs ~ mhi_2018, data = cdata_2018)

tidy(income_cc_fit)
```

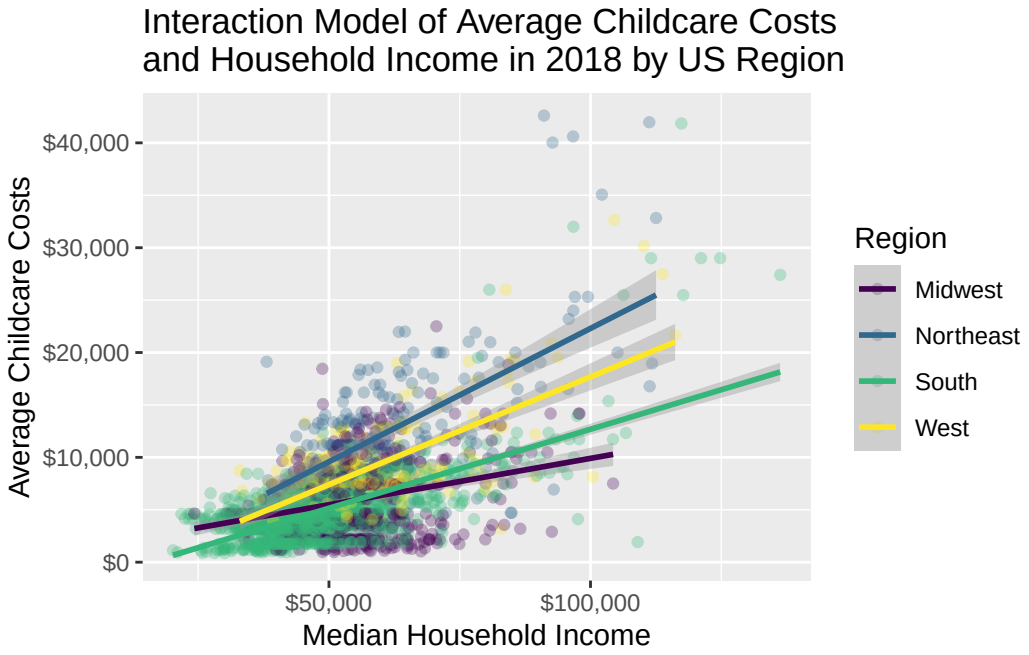
```
# A tibble: 2 x 5
  term         estimate std.error statistic   p.value
  <chr>           <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept) -3676.    280.          -13.1 4.77e- 38
2 mhi_2018        0.191   0.00519      36.9 7.68e-235
```

```
ggplot(data=cdata_2018,
         mapping=aes(x=mhi_2018, y=avg_costs, color=region)) +
  geom_point(alpha=0.3) +
  geom_smooth(method = "lm") +
  scale_x_continuous(labels = label_dollar()) +
  scale_y_continuous(labels = label_dollar()) +
  labs(
    x="Median Household Income",
    y="Average Childcare Costs",
    title="Interaction Model of Average Childcare Costs
and Household Income in 2018 by US Region",
    color="Region"
  ) +
  scale_color_viridis_d(option = "D")
```

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

```
Warning: Removed 802 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 802 rows containing missing values (`geom_point()`).
```



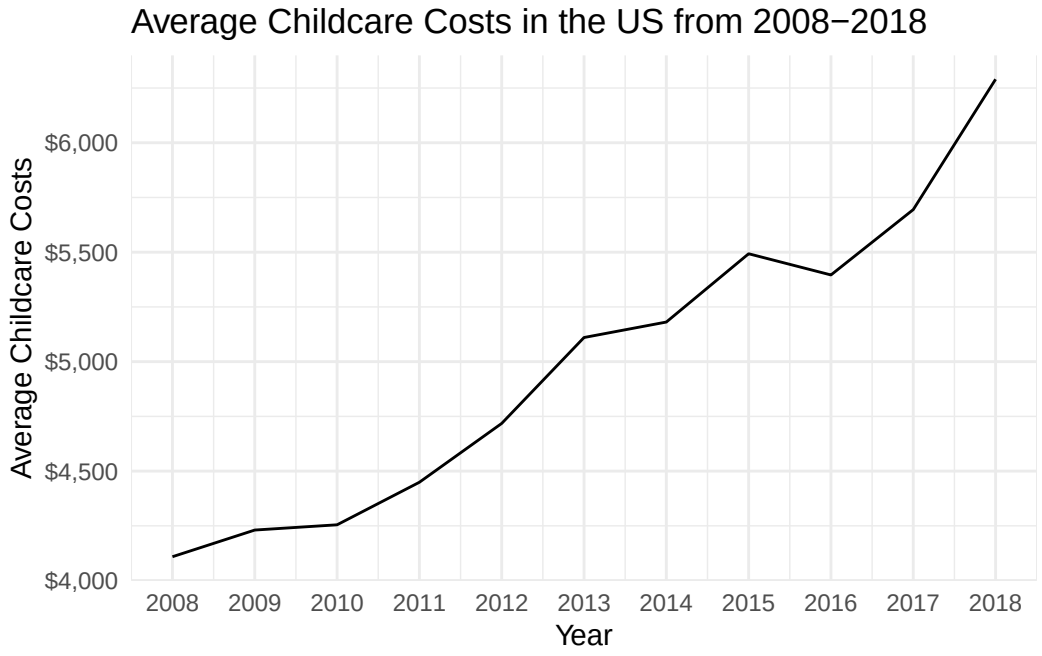
The purpose of this visualization is to explore the relationship between median household income and average childcare costs in 2018 by the four US regions. To do so, I categorized each state into one of four CDC defined US regions. I also took the average of the costs of the two types of childcare in each county to obtain the average childcare cost. I chose to use an interaction model to visualize the effect of not only income on childcare costs but also the effect of which region each family lives in on the childcare costs. From this graph, it can be concluded that the higher the median household income, the higher the average childcare costs and families paid from most to least childcare costs in the Northeast region, West region, South region, and Midwest region, respectively.

Question 3

3. How do childcare costs change from 2008-2018?

```
cdata_year <- cdata |>
  aggregate(avg_costs ~ study_year, mean)

cdata_year |>
  ggplot(mapping=aes(x=study_year, y=avg_costs)) +
  geom_line() +
  scale_x_continuous(breaks = 2008:2018) +
  scale_y_continuous(labels=label_dollar()) +
  theme_minimal() +
  labs(
    x="Year",
    y="Average Childcare Costs",
    title="Average Childcare Costs in the US from 2008-2018"
  )
```



The purpose of this visualization is to see how childcare costs have changed in the US from 2008 to 2018. I chose to use a line graph as it demonstrates how a variable changes over time. In this case, we can see that the average childcare costs in the US have steadily increased from 2008 to 2018.