Check-in 8

Jack Rellamas

Remember, **follow the instructions below and use R Markdown to create a pdf document with your code and answers to the following questions on Gradescope.** You may find a template file by clicking "Code" in the top right corner of this page.

1. Download and read the documentation for the Childcare Costs data.

```
childcare_costs <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/maste
```

```
## 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/
```

```
## 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.
```

```
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   3.4.4     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 error
```

2. Use visualization and regression to explore the following question: Is there a relationship between a county's average childcare costs and its median household income? There will be more than one way to answer this question–the important thing is to **explain** the choices you make in your analysis.

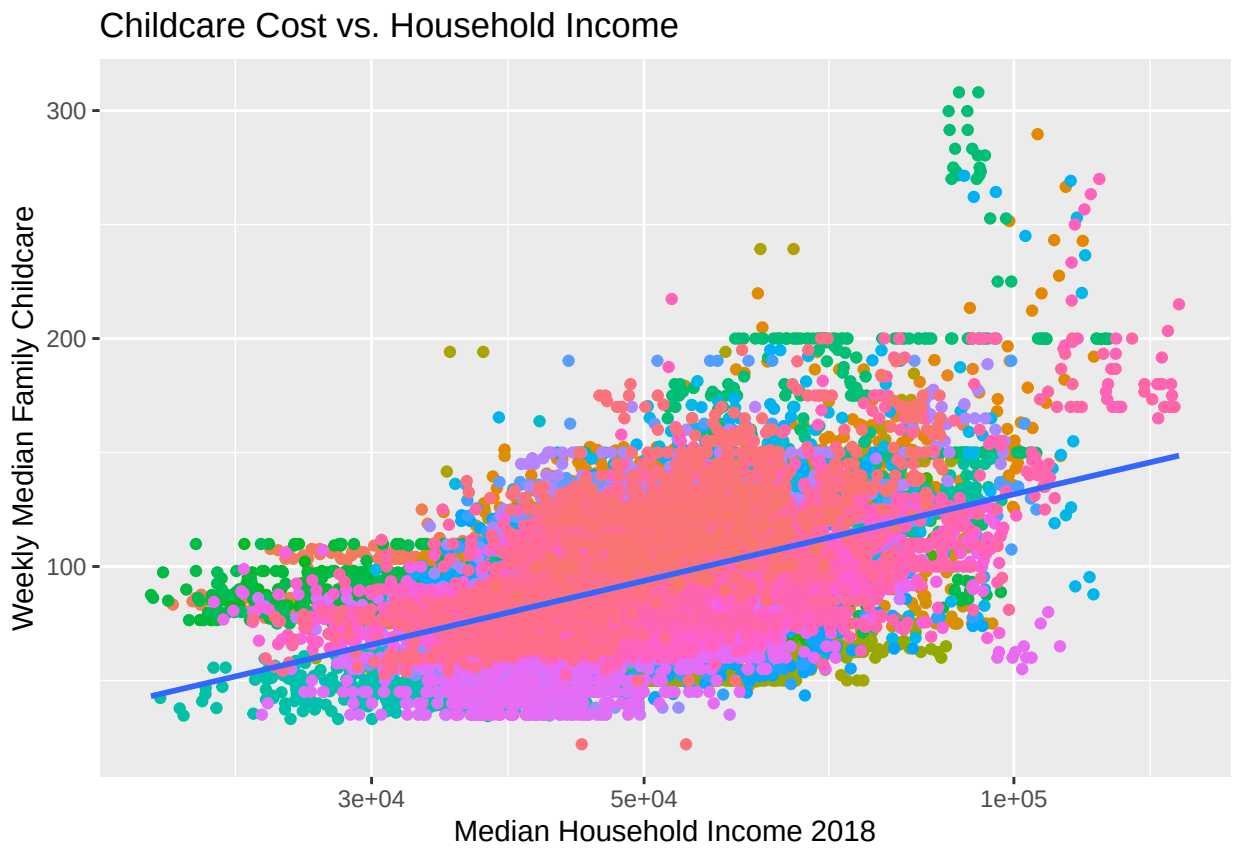
```r
# mhi_2018 : Median household income expressed in 2018 dollars.
# mcsa : Weekly, full-time median price charged for Center-Based Care for those who are school age
# mfccsa : Weekly, full-time median price charged for Family Childcare for those who are school age

childcare_costs <- childcare_costs %>%
  mutate(county_fips_code = factor(county_fips_code)) %>%
  filter_at(vars(mfccsa, mhi_2018), any_vars(complete.cases(.)))

ggplot(data = childcare_costs,
       aes(x = mhi_2018,
           y = mfccsa,
           color = county_fips_code,
           group = 1)) + #https://stackoverflow.com/questions/71761606/how-to-plot-a-single-regression
  geom_point() +
  scale_x_log10() +
  theme(legend.position = "none") +
  geom_smooth(method = 'lm') +
  xlab("Median Household Income 2018") +
  ylab("Weekly Median Family Childcare") +
  ggtitle("Childcare Cost vs. Household Income")
```

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

```
## Warning: Removed 11184 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

```
## Warning: Removed 11184 rows containing missing values (`geom_point()`).
```



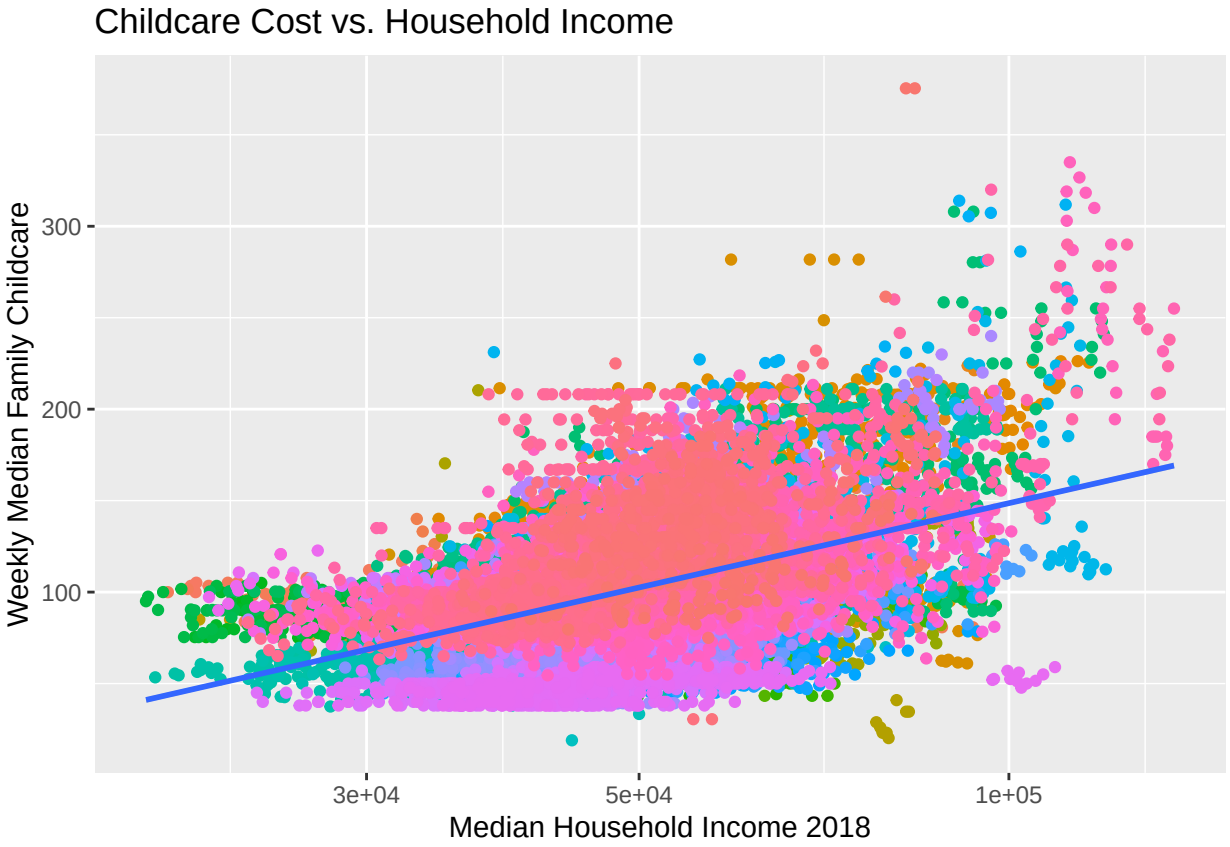

```
ggplot(data = childcare_costs,
       aes(x = mhi_2018,
           y = mcsa,
           color = county_fips_code,
           group = 1)) + #https://stackoverflow.com/questions/71761606/how-to-plot-a-single-regression
  geom_point() +
  scale_x_log10() +
  theme(legend.position = "none") +
  geom_smooth(method = 'lm') +
  xlab("Median Household Income 2018") +
  ylab("Weekly Median Family Childcare") +
  ggtitle("Childcare Cost vs. Household Income")
```

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

```
## Warning: Removed 10974 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

```
## Warning: Removed 10974 rows containing missing values (`geom_point()`).
```




```
lm_res <- lm(mfccsa ~ mhi_2018, data = childcare_costs)
summary(lm_res)
```

```
##
## Call:
## lm(formula = mfccsa ~ mhi_2018, data = childcare_costs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.064 -16.149  -0.757  14.123 173.033
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.871e+01  6.177e-01   62.66   <2e-16 ***
## mhi_2018    1.067e-03  1.185e-05   90.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.84 on 23381 degrees of freedom
##   (11184 observations deleted due to missingness)
## Multiple R-squared:  0.2575, Adjusted R-squared:  0.2575
## F-statistic:  8108 on 1 and 23381 DF,  p-value: < 2.2e-16
```

Visually, it is clear that as median household income increases, so does the cost of weekly median family childcare along with it. The relationship between income and childcare cost is very statistically significant.

There is also a resulting p-value of <2e-16 for the median household income in relation to the cost of weekly family childcare. We can reject the null hypothesis that there is no relationship between the income and the cost of childcare.