Applied Data Project

First, create an R project in an easy to access location on your computer. Then create an Rmd document that you specify to Knit to pdf.

1. Load the data from this source based on the instructions on the hosted site: https://github.com/rfordatascience/tidytuesday/blob/master/data/2023/2023-05-09/readme.md Import both **childcare_costs.csv** AND **counties.csv**. Give them names that are intuitive so that you can tell what they are.
    a. More information on the source of the data is available here: https://www.dol.gov/agencies/wb/topics/featured-childcare

2. Join these two datasets based on a join function within the tidyverse. **Hint:** call ?join in R and also look at Sarah's R code from yesterday's PM session, you will need to find a shared column between these two datasets to be able to join them together.
3. What is the unit of analysis? That is, what does each observation represent?
4. How many observations are in the dataset? How many observations per year?
5. Create a new variable called 'county_size' that categorizes counties according to the following criteria. Check your results with the 'county distribution' column. Hint: create a table of the new variable for only one year.

County size	County Distribution
Small (1-99,999)	2,548
Medium (100,000-499,999)	456
Large (500,000-999-999)	94
Very Large (1,000,000+)	44

6. The prices of childcare in the dataset are per week, but they are shown as annual costs in the report. Create a new variable that contains the annual cost (assuming that childcare is needed the 52 weeks in a year).
7. Replicate this column of the table showing the MEDIAN cost of Infant Center-Based Care for each category of counties. Note the year the data corresponds to.

Total U.S. Counties				Counties
County size	**County Distribution**	**Population Distribution**	**Share of Population**	**Infant Center-Based Care: 2018 (2022 Estimate)**
Small (1-99,999)	2,548	67,266,422	21%	$7,461 ($8,310)
Medium (100,000-499,999)	456	96,580,292	30%	$10,194 ($11,354)
Large (500,000-999-999)	94	67,437,679	21%	$13,420 ($14,947)
Very Large (1,000,000+)	44	91,618,637	28%	$15,417 ($17,171)

8. Pick **TWO** continuous variables of interest to you in the dataset. Using the function lm(y~x, data = data_name), run a regression to understand the relationship between these two variables. This is to say, if I were to pick the variables “pr_p” as my x and “mc_toddler” as my y, then I would run the function reg_poverty_toddler <- lm(mc_toddler~pr_p, data = childcare_costs).
    a. What are the results of the regression? Can you identify whether the relationship is positive or negative?
9. Create a scatter plot of the two different variables above. Try to customize some of the thematic options to develop a bit of your own aeshetic flair.
10. BONUS: Has the cost of childcare increased in tme? Plot the median cost of Center-Based Care for those who are school age (mfccsa) through time. Why is this time comparison problematic? (hint: are prices adjusted for inflation?).