**Laboratory 3: Exploratory data analysis**

**Group means**

We start by using mutate() to create a new variable that will be TRUE if the state value is either “California”, “Oregon”, or “Washington”, and FALSE otherwise. This process uses some of the tools we’ve seen before, namely the if\_else() function and the %in% operator, and puts them together.

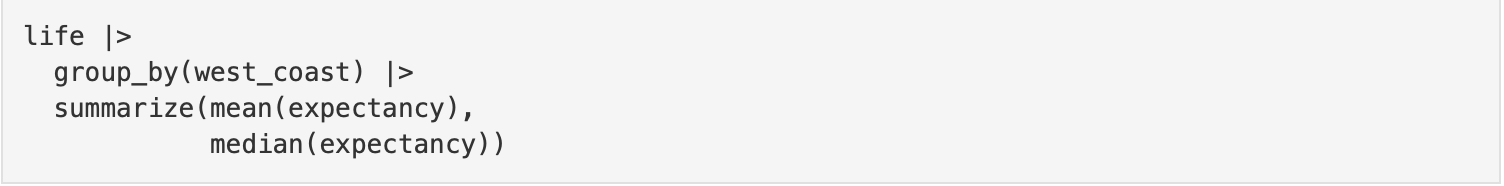
In the code below, we make a new variable named west\_coast and specify how that variable is calculated. The condition we use inside of the if\_else() function states that if the state associated with an observation is included in “California, Oregon, or Washington, the west\_coast variable should have a value of "yes". If the state of an observation is not included in this list, then the west\_coast variable should have a value of "no".

Notice the life <- life |> line in the code below. This line is telling R that we are updating the original lifedataset with a new dataset that contains the west\_coast variable.

A close-up of a computer screen

Description automatically generated

To compute means for the two groups (west coast and the rest of the country), we pipe this updated dataset into the group\_by() function, and indicate how we would like to make the groups. Then we can summarize() those groups, West Coast counties and non-West Coast counties, by taking the mean() and median() of their life expectancies.



Next, we will use data from gapminder, which tracks demographic data in countries of the world over time (before install the gapminder package). To learn more about it, you can bring up the help file with ?gapminder.

**Exercise 1.** For this exercise, focus on how the life expectancy differs from continent to continent. This requires that you conduct your analysis not at the country level, but aggregated up to the continent level. This is made possible by the one-two punch of group\_by() and summarize(), a very powerful syntax for carrying out the same analysis on different subsets of the full dataset.

* Create a dataset called gap2007 that contains only data from the year 2007.
* Using gap2007, calculate the mean and median life expectancy for each continent. Don’t worry about naming the new columns produced by summarize().
* Confirm the trends that you see in the medians by generating side-by-side boxplots of life expectancy for each continent.

More specifically:

# Create dataset of 2007 data

## Next we need to:

## 1. Use the gap2007 data

## 2. Make groups based on the continent of the country

## 3. Find the mean and median life expectancy of each continent

## Finally we need to:

## 1. Use the gap2007 data

## 2. Declare our x and y variables

## 3. Add boxplots to the plot

**Calculate spread measures**

Let’s extend the powerful group\_by() and summarize() syntax to measures of spread. If you’re unsure whether you’re working with symmetric or skewed distributions, it’s a good idea to consider a robust measure like IQR in addition to the usual measures of variance or standard deviation.

**Exercise 2.** The gap2007 dataset that you created in an earlier exercise is available in your workspace.

* For each continent in gap2007, summarize life expectancies using the sd(), the IQR(), and the count of countries, n(). No need to name the new columns produced here. Keep in mind the n() function does not take any arguments!
* Graphically compare the spread of these distributions by constructing overlaid density plots of life expectancy broken down by continent.

# Compute groupwise measures of spread

## First group the data based on the continent of the observation

# Next, specify the statistics that should be calculated

# Finally, specify what variable should be used in the calculations

## Last step -- create density plots with the different continents overlaid

## To differentiate between the continents, we fill each one with a different color

## We use an alpha of 0.3 to allow for us to see the shape of every density plot

Consider the density plots shown here. What are the most appropriate measures to describe their centers and spreads? In this exercise, you’ll select the measures and then calculate them.

**Exercise 3.** Using the shapes of the density plots you created above, calculate the most appropriate measures of center and spread for the distribution of life expectancy in the countries of the Americas. Note you’ll need to apply a filter here!

# First filter the data to only include the Americas

# Next define the summary measures that will be used

# Finally, specify what variables will be summarized

**Shape and transformations**

Here you’ll focus on the population variable, which exhibits strong right skew, and transform it with the natural logarithm function (log() in R).

**Exercise 4.** Using the gap2007 data:

* Create a density plot of the population variable (pop).
* Use mutate() to create a new variable called log\_pop that is the natural log of the population, and save this new variable back into gap2007.
* Create the same density plot as before with your transformed variable.

**Exercise 5.** Next, we’ll build a plot with that country removed. The gap2007 is still available in your workspace. To carry out this process we need to use the following steps:

1. Use filter() to retain only observations from Asia
2. Use mutate() to create a new variable called is\_outlier
3. Use if\_else() to specify how the is\_outlier variable should be made, where countries with a life expectancy less than 50 should get a value of TRUE and everyone else should get a value of FALSE
4. Create a boxplot of the life expectancies of the non-outlier countries

## First filter out observations from countries other than Asia

## Next, create a new variable named is\_outlier using the if\_else function

## Next, use the is\_outlier variable to filter out the observations flagged as outliers

# Create a boxplot of the distribution of life expectancy with the outlier removed