```
# Yannique Hecht
# Harvardx: PH125.1x - (1) Data Science: R Basics
# SECTION 3: INDEXING, DATA WRANGLING, PLOTS
# ASSESSMENTS

# # # ASSESSMENT 3.1: INDEXING

# # LOGICAL VECTORS

# Store the murder rate per 100,000 for each state, in `murder_rate`
[redacted]

# Store the `murder_rate < 1` in `low`
[redacted]


# # WHICH

# Store the murder rate per 100,000 for each state, in murder_rate
[redacted]

# Store the murder_rate < 1 in low
[redacted]

# Get the indices of entries that are below 1
[redacted]


# # ORDERING VECTORS

# Store the murder rate per 100,000 for each state, in murder_rate
[redacted]

# Store the murder_rate < 1 in low
[redacted]

# Names of states with murder rates lower than 1
[redacted]


# # FILTERING

# Store the murder rate per 100,000 for each state, in `murder_rate`
[redacted]

# Store the `murder_rate < 1` in `low`
```

[redacted]

Create a vector ind for states in the Northeast and with murder rates lower than 1.
[redacted]
Names of states in `ind`
[redacted]

Store the murder rate per 100,000 for each state, in murder_rate
[redacted]0

Compute the average murder rate using `mean` and store it in object named `avg`
[redacted]

How many states have murder rates below avg ? Check using sum
[redacted]

MATCH

Store the 3 abbreviations in a vector called `abbs` (remember that they are character vectors and need quotes)
[redacted]
Match the abbs to the murders$abb and store in ind
[redacted]
Print state names from ind
[redacted]

%in%

Store the 5 abbreviations in `abbs`. (remember that they are character vectors)
[redacted]

Use the %in% command to check if the entries of abbs are abbreviations in the the murders data frame
a[redacted]

LOGICAL OPERATOR

Store the 5 abbreviations in `abbs`. (remember that they are character vectors)

[redacted]

```
# Use the %in% command to check if the entries of abbs are
 abbreviations in the the murders data frame
```
[redacted]

```
# # # ASSESSMENT 3.2: BASIC DATA WRANGLING

# # DPLYR

# Loading data
```
[redacted]
```

# Loading dplyr
```
[redacted]
```

# Redefine murders so that it includes a column named rate with the
 per 100,000 murder rates
```
[redacted]

```
# # MUTATE

# Note that if you want ranks from highest to lowest you can take the
 negative and then compute the ranks
```
[redacted]
```

# Defining rate
```
[redacted]

```
# Redefine murders to include a column named rank
# with the ranks of rate from highest to lowest

```
[redacted]

```
# # SELECT

# Load dplyr
```
[redacted]

```
# Use select to only show state names and abbreviations from murders
```

[redacted]

FILTER

Add the necessary columns
[redacted]

Filter to show the top 5 states with the highest murder rates
[redacted]

FILTER WITH !=

Use filter to create a new data frame no_south
[redacted]
Use nrow() to calculate the number of rows
[redacted]

FILTER WITH%IN%

Create a new data frame called murders_nw with only the states from the northeast and the west
[redacted]
Number of states (rows) in this category
[redacted]

FILTERING BY 2 CONDITIONS

add the rate column
[redacted]

Create a table, call it my_states, that satisfies both the conditions
[redacted]

Use select to show only the state name, the murder rate and the rank
[redacted]

```
# # USING THE PIPE %>%

# add the rate column
[redacted]

# Create a table, call it my_states, that satisfies both the
 conditions
[redacted]

# Use select to show only the state name, the murder rate and the
 rank
[redacted]


# # MUTATE, FILTER & SELECT

# Loading the libraries
[redacted]

# Create new data frame called my_states (with specifications in the
 instructions)
[redacted]



# # # ASSESSMENT 3.3: BASIC PLOTS

# # SCATTERPLOTS

# Load the datasets and define some variables
[redacted]

population_in_millions <- murders$population/10^6
[redacted]

plot(population_in_millions, total_gun_murders)

# Transform population using the log10 transformation and save to
 object log10_population
[redacted]
```

```
# Transform total gun murders using log10 transformation and save to
 object log10_total_gun_murders
[redacted]
# Create a scatterplot with the log scale transformed population and
 murders
[redacted]


# # HISTOGRAMS

# Store the population in millions and save to population_in_millions
[redacted]

# Create a histogram of this variable
[redacted]


# # BOXPLOTS

# Create a boxplot of state populations by region for the murders
 dataset
[redacted]




# # # SECTION 3 ASSESSMENT

# # Q1 First, determine the average height in this dataset. Then
 create a logical vector ind with the indices for those individuals
 who are above average height.
[redacted]
[redacted]

# # Q2 How many individuals in the dataset are above average height
 and are female?
[redacted]

# # Q3 If you use mean on a logical (TRUE/FALSE) vector, it returns
 the proportion of observations that are TRUE. What proportion of
 individuals in the dataset are female?
[redacted]

# # Q4a This question takes you through three steps to determine the
 sex of the individual with the minimum height.
[redacted]
```

Q4b Use the match() function to determine the index of the individual with the minimum height.

[redacted]

Q4c Subset the sex column of the dataset by the index in 4b to determine the individual's sex.

[redacted]

Q5a Determine the maximum height.

[redacted]

Q5b Write code to create a vector x that includes the integers between the minimum and maximum heights.

[redacted]

Q5c How many of the integers in x are NOT heights in the dataset?

[redacted]

Q6a What is the height in centimeters of the 18th individual (index 18)?

[redacted]

Q6b What is the mean height in centimeters?

[redacted]

7a How many females are in the heights2 dataset?

[redacted]

Q7b What is the mean height of the females in centimeters?

[redacted]

Q8 Plot the percent palmitic acid versus palmitoleic acid in a scatterplot. What relationship do you see?

[redacted]

Q9 Create a histogram of the percentage of eicosenoic acid in olive. Which of the following is true?

[redacted]

10 Make a boxplot of palmitic acid percentage in olive with separate distributions for each region. Which region has the most variable palmitic acid percentage? We can determine variability from

the range of values each regions's palmitic acid percentage covers:

[redacted]