SDS/MTH 220: Badge Challenge 1

Name: ____

Instructions

- This badge challenge must be your own work entirely.

- This is an open mind, closed stats notebook, closed textbook, and closed fellow statistician badge challenge.

- All questions will be graded under the badge level grading system. On badge challenges, you must show ***all*** work for computational answers and justify all claims for expository questions.

- Keep your explanations contextually meaningful and concise!

- You do not have to perform any long computations. For example, if the answer is 18.5, you will receive full credit for writing $2.5 * (4 + 3.5) - (1/2)^2$.

- Use the provided blank sheets of paper to write your answers. You may also use these pages for your scratch work.

- Please write on ***one side*** *of the blank pages* and staple any pages you want graded to the badge challenge. Your *answers should appear in* ***question order***.

- Put your name on the top right corner of each page that you submit and do not write where the staple will go.

- The front page must have two timestamps on it. **Timestamps will be strictly enforced. Any badge challengs with pairs of timestamps indicating than more than 140 minutes or missing timestamps are subject to an honor board case.**

- This is the first of four opportunities to demonstrate your current understanding of the first five badges for the course. You may complete as many badges as you would like. Questions left unattempted will not receive a badge level.

Name: ________________________________

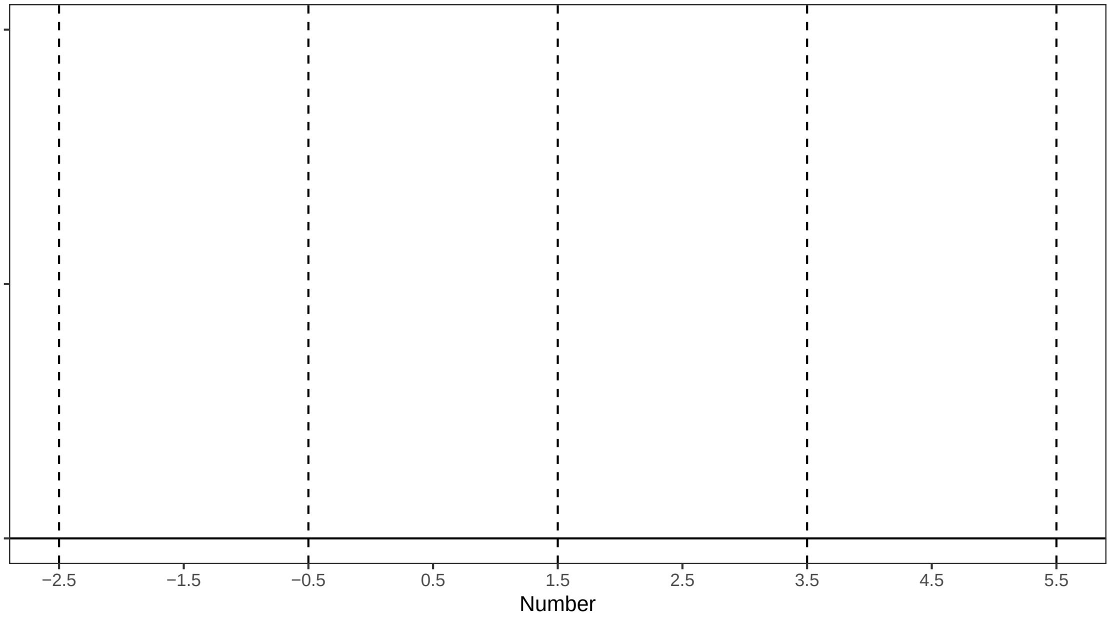
Question	Badge 1	Badge 2	Badge 3	Badge 4	Badge 5
1					
2					
3					
4					
5					
Recorded Score by Badge					

Question X

(a) Say you have a data frame named `example` that has 4 variables and 4 rows of data:

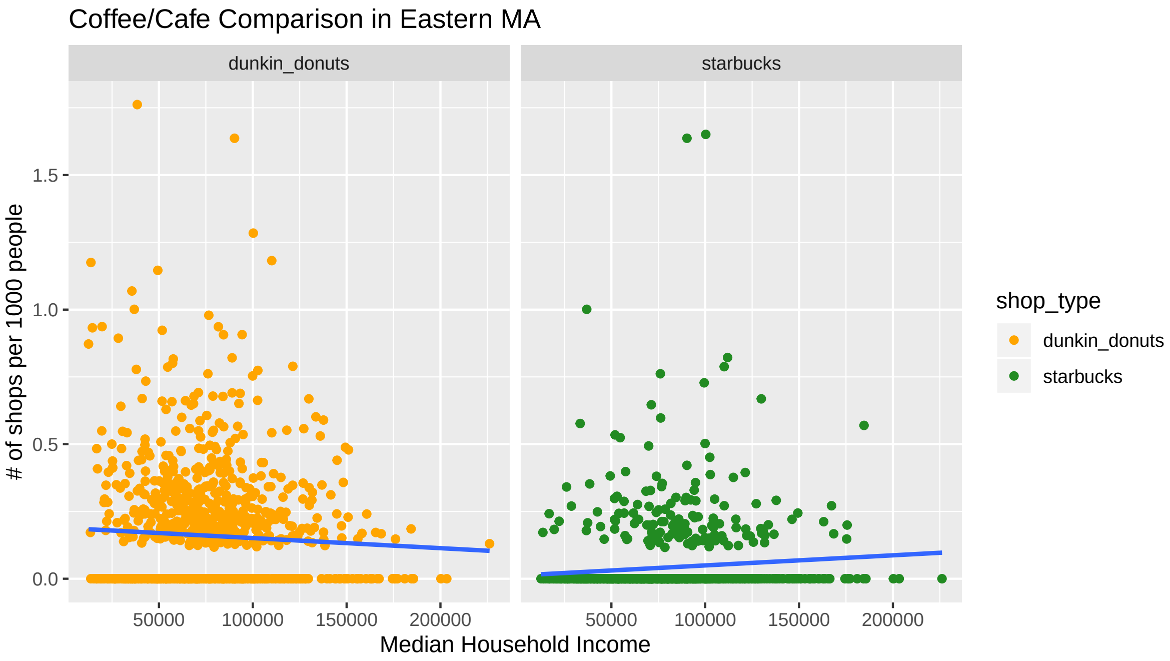
type	fruit	city	number
D	apple	Toronto	3
B	apple	Montreal	4
A	orange	Toronto	2
C	orange	Montreal	1

- Draw the histogram for the variable `number` using the bins indicated with the dashed vertical lines. Be sure to write down appropriate y-axis information.



- List the elements of the "Grammar of Graphics" for a **generic** plot. In other words, what are the ingredients of *any* visualization in `ggplot2`?
- Write out pseudocode to create the histogram you created in the first part of this question. Specify where you use each element of the "Grammar of Graphics."

(b) A researcher from eastern Massachusetts is a big Starbucks fan. She has a suspicion that Starbucks tend to locate in richer neighborhoods, while this is not the case for Dunkin Donuts. She writes code to pull data from the internet about all 1024 census tracts (areas where decennial census data are collected) in Eastern Massachusetts. She summarizes her results in the following graphic:



(a) Write out the elements of the "Grammar of Graphics" that need to be specified to create this graphic. You do not need to specify the axes labels, the plot title, nor the regression lines

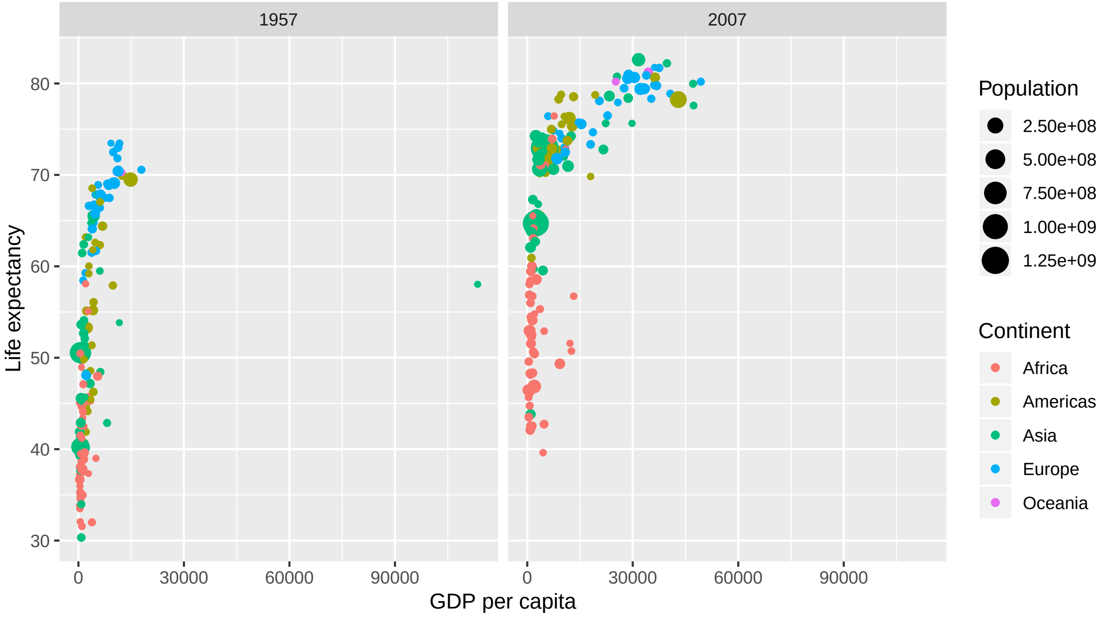
(b) Assuming there are no missing data, how many rows are in the data frame that we input into the `ggplot()` function?

(c) Does the presented visualization support or contradict the researcher's suspicion? Why?

(c)
- For which of the following pairs of variables would you visualize with a scatterplot? Circle which pairs.
  - Pair 1: "Distance from school in miles" and "mode of transportation to school (bike, walking, bus)
  - Pair 2: "Number of years at a job" and "Salary"
  - Pair 3: "Years experience playing an instrument" and "number of mistakes made playing a song"
  - Pair 4: "Number of years since a person retired" and "favorite sport"
- Consider a subset of the `gapminder` dataset we've seen numerous times in class:

```
## # A tibble: 284 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1957    30.3  9240934      821.
##  2 Afghanistan Asia       2007    43.8 31889923      975.
##  3 Albania     Europe     1957    59.3  1476505     1942.
##  4 Albania     Europe     2007    76.4  3600523     5937.
##   5 Algeria     Africa     1957     45.7 10270856     3014.
##   6 Algeria     Africa     2007     72.3 33333216     6223.
##   7 Angola      Africa     1957     32.0  4561361     3828.
##   8 Angola      Africa     2007     42.7 12420476     4797.
##   9 Argentina   Americas   1957     64.4 19610538     6857.
##  10 Argentina   Americas   2007     75.3 40301927    12779.
## # ... with 274 more rows
```

Using this data, we can create the following plot:



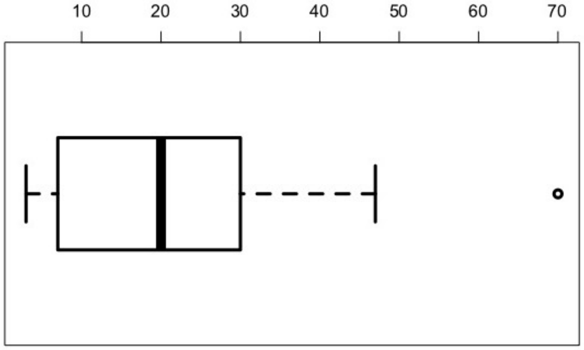
Write out **in bullet point form** all the elements of the "Grammar of Graphics" that need to be specified in a `ggplot()` function call to create this graphic. Note

- You don't need to write code, you only need to specify all components of the graphic.
- There is no need to specify the x and y axes labels.

- Write out pseudocode to create a collection of boxplots of GDP per capita with one boxplot per continent. Specify where you use each element of the "Grammar of Graphics."

Question X

For this question consider the below box plot.



(a) For each of the below statements, state whether it is true or false and briefly ***explain*** your answer.

- The range of extremes is [3,47].
- According to the boxplot, the number of data points used to construct the box plot is 17.
- The 50% coverage interval is [7,30].
- The data is skew-left.
- According to the boxplot, the mean is 20.

(b) If you would like to know the above 5 summary statistics (i.e. range of extremes, skewness, mean, coverage intervals, and number of data points), is a box plot the best visual representation for your data? If not, what visualization would be better? Justify your answer.

(c) List the elements of the "Grammar of Graphics" for a **generic** plot. In other words, what are the ingredients of *any* visualization in `ggplot2`?

(d) Write out pseudocode to create this boxplot in `ggplot2`. Specify where you use each element of the "Grammar of Graphics."

Question X

(a) Recall the babynames dataset that contains all babynames used more than 5 times for any given year, split by sex, for the years 1880 through 2015. Here is a preview of the first 10 rows:

year	sex	name	n	prop
1880	F	Mary	7065	0.0723836
1880	F	Anna	2604	0.0266790
1880	F	Emma	2003	0.0205215
1880	F	Elizabeth	1939	0.0198658
1880	F	Minnie	1746	0.0178884
1880	F	Margaret	1578	0.0161672
1880	F	Ida	1472	0.0150812
1880	F	Alice	1414	0.0144870
1880	F	Bertha	1320	0.0135239
1880	F	Sarah	1288	0.0131960

a) Write the pseudocode that is going to compute the total number of babies born between 1950 and 2000 that are named “Riley.”

b) You want to compare the degree to which the names “Casey” and “Riley” have been “unisex” names for all years between 1950 to 2000, in other words the focus is on the degree to which the names have been used by both sexes

- Write the pseudocode for the data wrangling.
- Specify all the elements of the grammar of graphics that is going to generate an appropriate visualization.

(b) You are presented with data on the Titanic disaster of 1912 in a data frame Titanic, which cross-classifies survival vs death by class, sex, and age. Write down the *pseudocode* of the commands that will output a table comparing survival vs death counts for the following three scenarios. For each scenario, under your pseudocode, draw what the output table would look like, but do **not** fill in the numbers in the table.

a) by sex

b) by sex and age

c) by sex and class

(c) What would you use to address the question if the “women and children”-first policy of the White Star Line Company (the company that ran the Titanic) held true or not.

Note: you don't need to calculate the output table, just write the pseudocode that would produce it where the more concise the pseudocode the better. Here is what the `Titanic` data looks like:

Class	Sex	Age	Survived	n
1st	Male	Child	No	0
2nd	Male	Child	No	0
3rd	Male	Child	No	35
Crew	Male	Child	No	0
1st	Female	Child	No	0
2nd	Female	Child	No	0
3rd	Female	Child	No	17
Crew	Female	Child	No	0
1st	Male	Adult	No	118
2nd	Male	Adult	No	154
3rd	Male	Adult	No	387
Crew	Male	Adult	No	670
1st	Female	Adult	No	4
2nd	Female	Adult	No	13
3rd	Female	Adult	No	89
Crew	Female	Adult	No	3
1st	Male	Child	Yes	5
2nd	Male	Child	Yes	11
3rd	Male	Child	Yes	13
Crew	Male	Child	Yes	0
1st	Female	Child	Yes	1
2nd	Female	Child	Yes	13
3rd	Female	Child	Yes	14
Crew	Female	Child	Yes	0
1st	Male	Adult	Yes	57
2nd	Male	Adult	Yes	14
3rd	Male	Adult	Yes	75
Crew	Male	Adult	Yes	192
1st	Female	Adult	Yes	140
2nd	Female	Adult	Yes	80
3rd	Female	Adult	Yes	76
Crew	Female	Adult	Yes	20

Question X

You are investigating if countries get a 'bump' in their gold medal count when they host the Winter Olympics. Recalling that the 2010 Winter Olympics were in Vancouver Canada, you decide to use Canada as a first test case. You found the following Canadian gold medal tallies for the past 9 Winter Olympics:

Year	1984	1988	1992	1994	1998	2002	2006	2010	2014
Canada's Gold Medals	2	0	2	3	6	7	7	26	10

(a) What is the median number of gold medals that Team Canada have won over the past 9 Winter Olympics? Show all work.

(b) Find the 50% coverage interval. Show all work.

(c) Based on the above table and results, do you think that Canada got a 'bump' for being the host country? Justify your answer.

(d) You double check the table's data and notice that for the 2010 games, the listed number of medals is the total number and not the just the gold medals. The correct number of gold medals is actually 13 gold medals. Does the median number of gold medals that Team Canada has won over the past 9 Winter Olympics change? Does the mean change? Justify ***both*** your answers ***without*** doing any ***additional*** computations.

(a) Consider the following hypothetical study. Say you collect two variables of information from a population of interest: y = life expectancy and x = annual income out of college measured in units of thousands of dollars. You find that

  (a) the correlation coefficient is 0.25

  (b) the fitted regression line $\widehat{y} = 45 + 0.5x$

  Write down what the following two quantities would be if x was not measured in units of thousands of dollars, but measured in units of dollars:

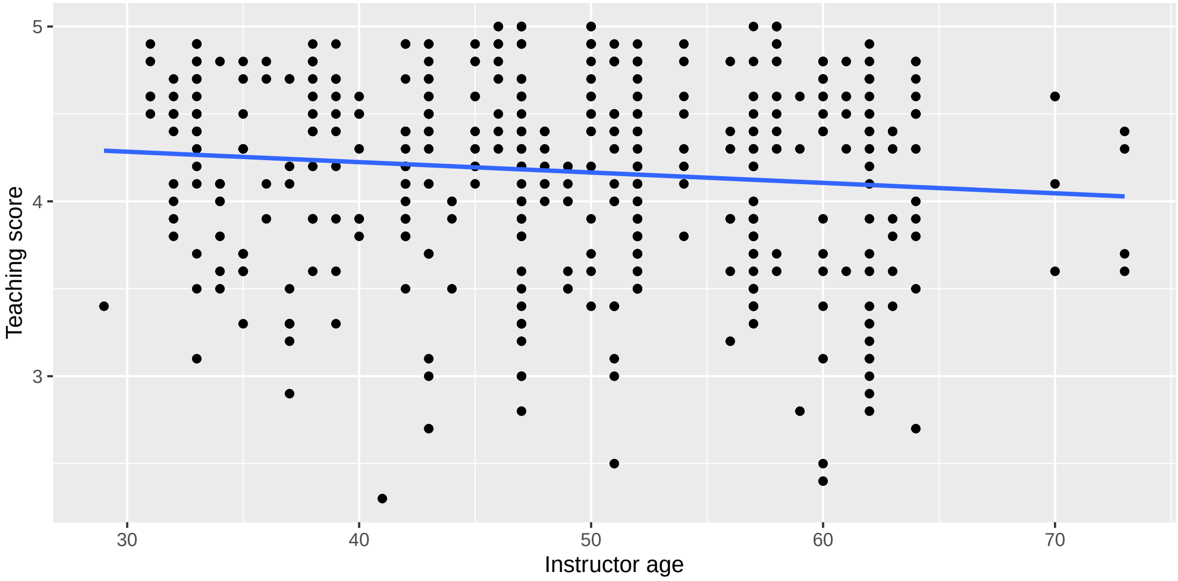
  (a) the correlation coefficient

  (b) the fitted slope b_1 of the regression line $\widehat{y} = b_0 + b_1 x$

  Interpret the coefficients b_0 and b_1 in contextually meaningful ways.

(b) Recall our teaching evaluation dataset seen in class. We're interested in fitting a model of teaching score (evaluated by students) as a function of instructor age.

```r
ggplot(data = evals, mapping = aes(x = age, y = score)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Instructor age", y = "Teaching score")
```



```r
model_score <- lm(score ~ age, data = evals)
get_regression_table(model_score)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    4.46      0.127     35.2    0         4.21     4.71
## 2 age         -0.006     0.003     -2.31   0.021   -0.011   -0.001
```

a) Interpret the `intercept` term in the `estimate` column of the regression table.

b) Give the precise interpretation of the slope for `age` in the `estimate` column of the regression table.

c) The regression line visualized in the above figure is considered the “best fitting line” through these points. By what criteria do we mean “best”?

d) What is the correlation coefficient of `age` and `score`? Is it positive or negative?

e) Consider the first row of the following output. Write down the equation that computes the first value of `score_hat`: `4.25`.

```
model_points <- get_regression_points(model_score)
model_points

## # A tibble: 463 x 5
##       ID score   age score_hat residual
##    <int> <dbl> <int>     <dbl>    <dbl>
##  1     1   4.7    36      4.25    0.452
##  2     2   4.1    36      4.25   -0.148
##  3     3   3.9    36      4.25   -0.348
##  4     4   4.8    36      4.25    0.552
##  5     5   4.6    59      4.11    0.488
##  6     6   4.3    59      4.11    0.188
##  7     7   2.8    59      4.11   -1.31
##  8     8   4.1    51      4.16   -0.059
##  9     9   3.4    51      4.16   -0.759
## 10    10   4.5    40      4.22    0.276
## # ... with 453 more rows
```

f) Write down the equation that computes the first value of `residual`: `0.452`.

g) Write down the data wrangling pseudocode to apply to `model_points` to compute the value of the criteria described in part c).

Estimate Std. Error

Question X

(a) Note: for this question, you do not need to do the arithmetic (adding, subtracting, multiplying, dividing), but rather write down what you would enter into a calculator if you had one. Let's consider the `gapminder` development data, but only for the year 2007. Let's look at a random sample of 5 out of the 142 rows of this dataset:

country	continent	lifeExp
Namibia	Africa	52.906
Portugal	Europe	78.098
Iran	Asia	70.964
Brazil	Americas	72.390
Italy	Europe	80.546

add Nambia and Italy here

We are interested in modeling the relationship between the outcome variable y = life expectancy in years and the categorical explanatory variable x = continent. You fit a following regression and obtain the following regression table rounded to the nearest integer:

```
## # A tibble: 5 x 7
##   term              estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept             54.8      1.02      53.4       0     52.8     56.8
## 2 continentAmericas     18.8      1.8       10.4       0     15.2     22.4
## 3 continentAsia         15.9      1.65       9.68      0     12.7     19.2
## 4 continentEurope       22.8      1.70      13.5       0     19.5     26.2
## 5 continentOceania      25.9      5.33       4.86      0     15.4     36.4
```

a) What is the fitted value $\widehat{y}$ of life expectancy in years for the below countries? (Show all work to justify your steps.)

(a) Togo
(b) Lestho
(c) Bulgaria

Are you surprised by these results? Why or why not?

a) What is the fitted value $\widehat{y}$ of life expectancy in years for any given country in:

(a) Africa

(b) Europe

Show all work to justify your steps.

c) What is the residual for the following three countries?

(a) Nambia
(b) Italy

Show all work to justify your steps.

d) What is the mean life expectancy for countries in the following continents:

(a) Africa
(b) Asia
(c) Europe

Show all work to justify your steps.