SDS/MTH 220: Badge Challenge 1

Name: _____

Section (please circle): Kim 01 or Kinnaird 02

Instructions:

- Honor code:
  a) This is an open mind, closed stats notebook, closed textbook, and closed fellow statistician badge challenge. This badge challenge must be your own work entirely.
  b) 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.**
- What to do with these pages:
  a) Use the provided blank sheets of paper to write your answers. You may also use these pages for your scratch work.
  b) 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***.
  c) Put your name on the top right corner of each page that you submit and do not write where the staple will go.
- Taking this badge challenge:
  a) 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.
  b) Remember that you only have to do as many questions as you are ready for. Questions left unattempted will not receive a badge level. If a question has multiple parts, you must attempt all parts to earn above X (cannot be assessed).
  c) You do not have to perform any long computations. For example, if the answer is 18.5, you will receive full credit for writing $2.5 * (4 + 3.5) - (1/2)^2$.
  d) Keep your explanations contextually meaningful and concise!

Badges

This is the first of four opportunities to demonstrate your mastering of the first five badges for the course.

	Topic
1	Understand the grammar of graphics: construct graphics based on a dataset, deconstruct graphics into a data set
2	Write pseudocode for basic data wrangling & exploratory data analysis
3	Compute and interpret summary statistics: measures of centrality & spread
4	Fit & understand regression models with numerical explanatory variables
5	Fit & understand regression models with categorical explanatory variables
6	Fit & understand interaction & parallel slopes models & perform basic model selection
7	Master terminology, notation, & definitions related to sampling: All terms in 7.3
8	Understand what determines center and spread of sampling distribution: Representative sampling, the role sampling variability plays in statistical inference and the role that sample size plays in this sampling variability.
9	Highlight all differences between sampling and resampling: Why would you resample? What is difference between sampling distribution & bootstrap distribution.
10	Understand confidence intervals
11	Construct and interpret confidence intervals
12	Generalize all hypothesis tests to there is "There is only one test" framework: Fig 9.14 & infer framework
13	Master terminology & definitions related to hypothesis testing: All terms in 9.2 and 9.4, in particular correctly articulate what a p-value is and how to interpret
14	Transfer previously developed knowledge of hypothesis tests & confidence intervals to regression i.e. interpret ALL columns of a regression table
15	Verify the conditions that must be met for any inference for regression to be valid

Name: _____

Section (please circle): Kim 01 or Kinnaird 02

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

Question 1

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

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
```

Note: You do **not** need to do the arithmetic (adding, subtracting, multiplying, etc), but rather write down what you would enter into a calculator if you had one.

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?

b) 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) Namibia
(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.

Question 2

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

- the correlation coefficient is 0.25
- the fitted regression line $\widehat{y} = 45 + 0.5x$

a) Interpret the coefficients b_0 and b_1 in contextually meaningful ways.

b) Your friend Reginald Regression tells you that he made \$42,000. Reginald is now 27 and would like to know how old you think he will live to be. Show all work and interpret your result for him (recalling that he has not ever taken statistics.)

c) In reaction to your statement of his expected life expectancy, he vows to eat more spinach. Should he take age from part b) as a guarantee of how long he will live? Explain why he should **or** should not rely on this model as a guarantee.

d) 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:

- the correlation coefficient
- the fitted slope b_1 of the regression line $\widehat{y} = b_0 + b_1 x$

Question 3

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.

a) Write down the *pseudocode* of the commands that will output a table comparing survival vs death counts for the following four 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.

- by sex
- by sex and age
- by sex and class
- by sex and age and class

b) 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. On the next page, you will see 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 4

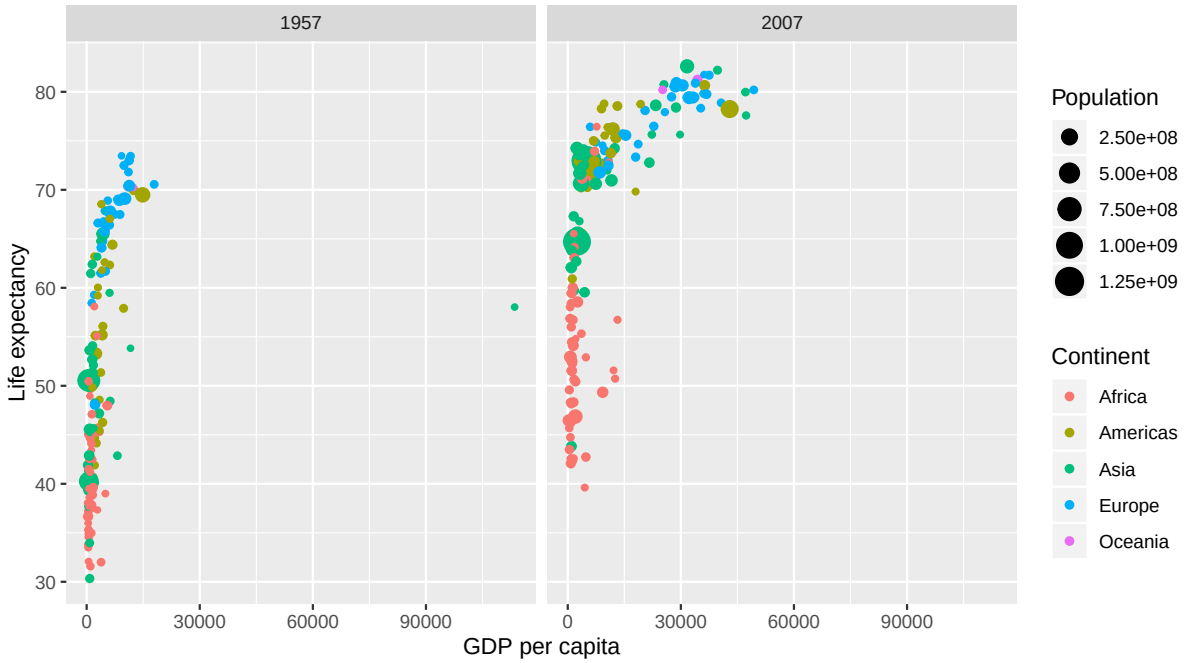
a) For which of the following pairs of variables would you visualize with a scatterplot? Circle which pairs and briefly explain your thinking

- 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"

b) 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 do **not** 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.

c) 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 5

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) You store your data as rows with `year` as the observation and the number of gold medals as a variable in a dataframe called `canada_gold`. Using pseudocode, state how you would find the Inner Quartile Range in RStudio.

(c) You find that the 25^{th}-percentile is 2 gold medals and the 75^{th}-percentile is 7 gold medals. 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.