SOC 301 Name: _____

Social Statistics November 14, 2016

Professor Chester Ismay

Exam 2

- Read carefully through each problem. Take your time and relax.
- Please write your name at the top of each page of the exam.
- You are to work on this exam alone, but you can use the RStudio Cheatsheets that I provided for you. I recommend that you focus on using the **Data Wrangling with dplyr and tidyr** and the **Data Visualization with ggplot2** cheatsheets only.
- Point allocations for each part of each problem are in the box to the left of the question. Please use this as a way to help you manage your time during the exam.
- One goal of my exams (and this course) is for you to show me that you have gained the ability to think critically and statistically about problems. If you are ever in doubt, carefully write what you are struggling with and your thought processes on the back of the sheet of paper corresponding to that problem. You may receive partial credit if I can understand what you are thinking. Remember that this is largely an exam focused on checking for your understanding of writing R code though. BLANK ANSWERS I can guarantee will NOT provide you with any credit.
- You can assume that all needed R packages have been installed and that the `library` function has been run on all of these packages.
- Take time after you have completed the exam to carefully review your answers. Make sure that you have answered all parts to all questions asked.

Question:	1	2	3	4	Total
Points:	12	13	10	15	50
Score:					

1. Recall the `gap` data frame that you worked with for Exam 1. Seven rows of the `gap` data frame are below.

country	region	subRegion	year	lifeExp	pop	gdpPercap
Ecuador	Americas	South America	2007	74.994	13755680	6873.262
Jordan	Asia	Western Asia	1972	56.528	1613551	2110.856
Portugal	Europe	Southern Europe	1987	74.060	9915289	13039.309
Sao Tome and Principe	Africa	Middle Africa	2007	65.528	199579	1598.435
Serbia	Europe	Southern Europe	2007	74.002	10150265	9786.535
Trinidad and Tobago	Americas	Caribbean	1977	68.300	1039009	7899.554
Uruguay	Americas	South America	1967	68.468	2748579	5444.620

For each of the following questions, be sure to use `dplyr` and `%>%` whenever possible for full credit.

3 (a) Recall that the `gap` data frame from Exam 1 also had the `dem_rank` variable. What code did I use to remove the `dem_rank` variable here?

3 (b) What code would be required to focus on only countries in Asia or in Middle Africa?

6 (c) What code would be needed to produce the following summarized table?

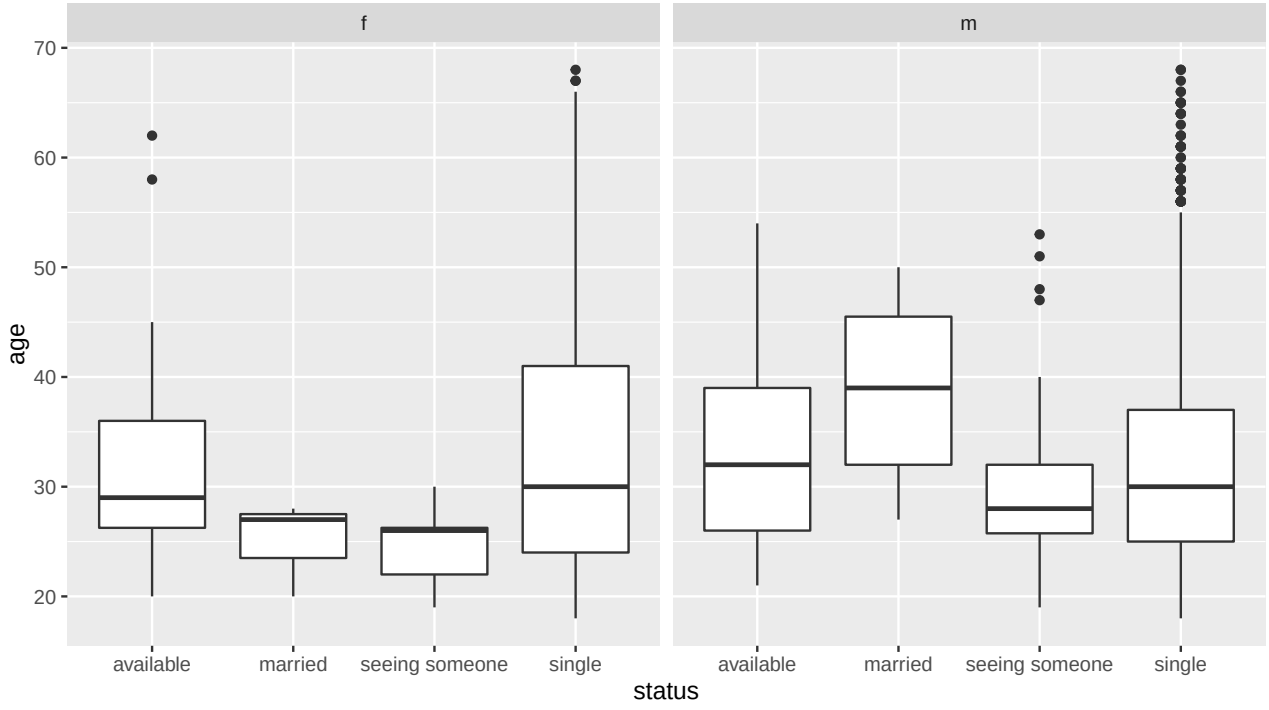
```
## Source: local data frame [10 x 3]
## Groups: region [?]
##
##      region  year mean_lifeExp
##       <chr> <dbl>        <dbl>
## 1    Africa  2002     53.52165
## 2    Africa  2007     55.00339
## 3  Americas  2002     72.48767
## 4  Americas  2007     73.66533
## 5      Asia  2002     69.23388
## 6      Asia  2007     70.72848
## 7    Europe  2002     76.74166
## 8    Europe  2007     77.68869
## 9   Oceania  2002     79.74000
## 10  Oceania  2007     80.71950
```

Points earned: _____ out of a possible 12 points on this page

2. Recall the `profiles` dataset from the `okcupiddata` R package. We'll use a sample of size 2000 here from the 59,946 total number of observations and we'll call that data frame `profiles_sample`.

3 (a) What is the observational unit in this `profiles_sample` data frame? Be as specific as possible.

6 (b) What `ggplot2` code is needed to produce the following plot?



4 (c) Describe the center and spread of the `age` variable comparing those identifying as `m` for `sex` to those identifying as `f` for `sex` across each of the four levels of `status`.

Points earned: _____ out of a possible 13 points on this page

10 3. Write the code that is needed to generate an appropriate visualization to compare the trends of "unisex"iness (not a measure of gender ambiguous sexiness, but rather the degree to which a name is used by both sexes) of the names "Ashley", "Jordan", and "Taylor" from **1980 to 2010**. Use the **proportion** of names in a given year as your response variable. As a hint, here are the first 15 rows of the `babynames` data set. Note: Make sure to color based on the different names and to facet based on male versus female.

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
1880	F	Annie	1258	0.0128887
1880	F	Clara	1226	0.0125608
1880	F	Ella	1156	0.0118437
1880	F	Florence	1063	0.0108908
1880	F	Cora	1045	0.0107064

Points earned: _____ out of a possible 10 points on this page

4. This example involves thinking about county level data on the percentage of black residents. All that is collected is a random representative sample of 200 US counties.

[3] (a) Layout what three rows of the tidy data set would look like for this sample of 200 counties.

[9] (b) Place the following steps to the bootstrap process in order according to the diagram given in the textbook and the steps we followed in class. Note that some of the choices will not be used. Your final answer should look something like...1 $\rightarrow$ 4 $\rightarrow$ 6 $\rightarrow$ 9 $\rightarrow$ 15...where 1 is the entry that occurs first and then 4 and then 6, etc.

1. Collect a sample of size 200 that are the first 200 counties to appear in the listing of all counties.
2. Collect a random sample of 200 counties from the population of all US counties.
3. Determine the mean of the population of all counties.
4. Determine the mean of the original sample of 200 counties.
5. Determine the median of the original sample of 200 counties.
6. Take our original sample of 200 counties and sample without replacement to create our first bootstrap sample.
7. Take our population and shuffle them all up to create a bootstrap sample of size 300.
8. Take our original sample of 200 counties and sample with replacement to create our first bootstrap sample.
9. From our first bootstrap sample, calculate 200 different means. Each of these corresponds to a bootstrap statistic.
10. From our first bootstrap sample, calculate the mean. This corresponds to our first bootstrap statistic.
11. Resample with replacement from the original sample 10,000 times to create 10,000 new bootstrap samples and from those 10,000 bootstrap statistics/means.
12. Resample with replacement from the first bootstrap sample 10,000 times to create 10,000 new bootstrap samples and from those 10,000 bootstrap statistics/means.
13. The resulting 10,000 bootstrap means correspond to the bootstrap distribution.
14. The resulting $200 \times 10,000 = 2,000,000$ bootstrap means correspond to the bootstrap distribution.

Final answer: _____

[3] (c) Why is it so important that the original sample from the population be selected at random for the bootstrapping process to produce a "good" result?

Points earned: _____ out of a possible 15 points on this page