SOC 301 Practice Problems for Exam 2

Make sure to use your cheatsheets to solve these problems as you prepare.

1 "Women and Children First"

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 `dplyr` commands that will output a table comparing survival vs death counts for the following three scenarios:

a) by sex

b) by sex and class and age

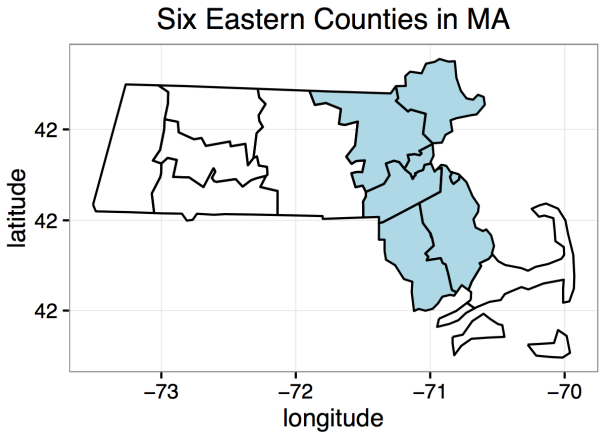
c) to answer the question if the "women and children"-first policy of the White Star Line Company (the company that ran the Titanic) held true or not.

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

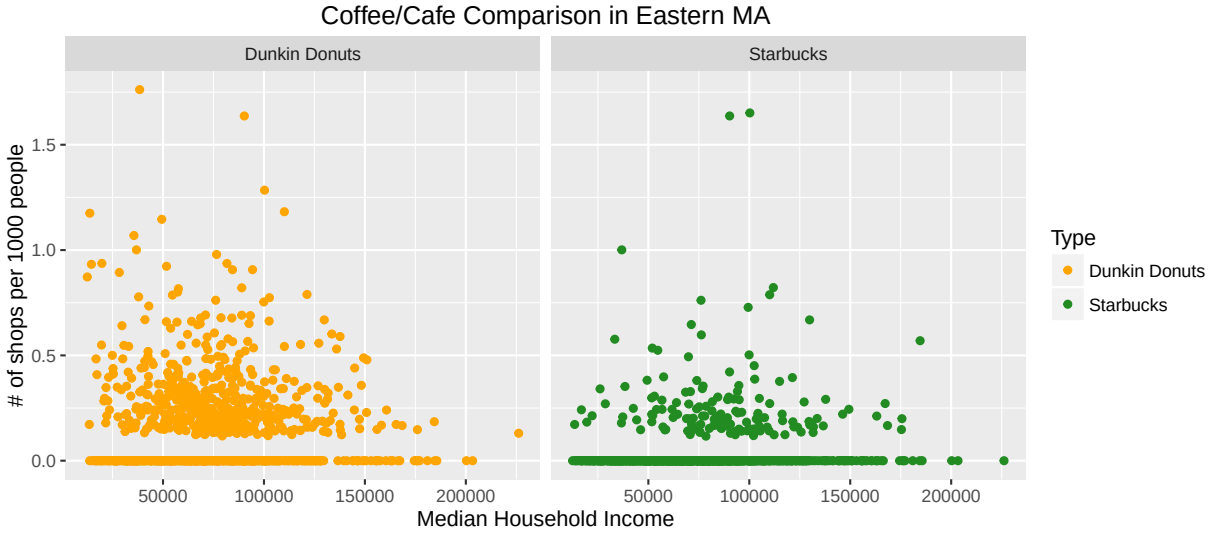
Class	Sex	Age	Survived	n
1st	Male	Child	No	0
2nd	Male	Child	No	0
3rd	Male	Child	No	35
Crew	Male	Child	No	0
1st	Female	Child	No	0
2nd	Female	Child	No	0
3rd	Female	Child	No	17
Crew	Female	Child	No	0
1st	Male	Adult	No	118
2nd	Male	Adult	No	154
3rd	Male	Adult	No	387
Crew	Male	Adult	No	670
1st	Female	Adult	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

2 America Runs on Starbucks?

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 scrape the internet for data from all 1024 census tracts (areas where decennial census data are collected) in 6 Eastern Massachusetts counties, specifically Bristol, Essex, Middlesex, Norfolk, Plymouth, and Suffolk counties:



She summarizes her results in the following graphic:



a) Sketch out (in tidy data format) the data set needed to make this graphic.

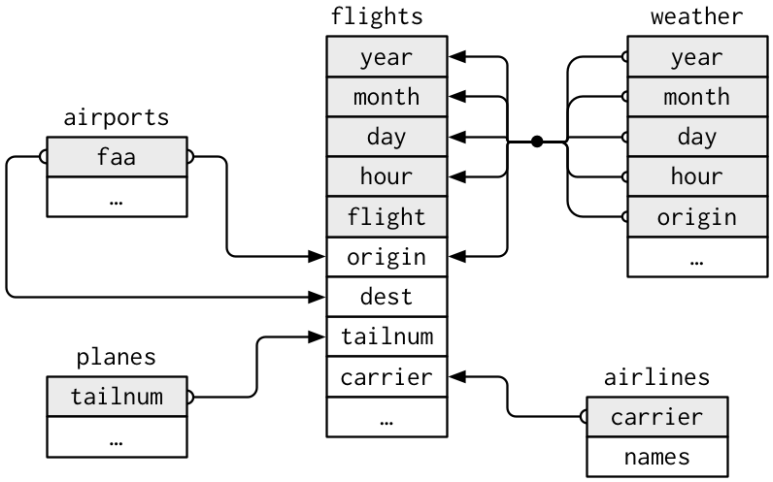
b) Write the `ggplot` code that generates this graphic. Be sure to write your code so that the various layers (the components you add to the base `ggplot()` call with + signs) are clear. (The `scale_color_manual` function can specify colors.)

c) Name two improvements that can be made to this graphic.

d) Does this evidence support or contradict the researcher's suspicion? Why?

3 NYC Flights

Recall the `airports`, `planes`, `flights`, `weather`, and `airlines` data sets in the `nycflights13` data set and that we saw the following graphic of the relationships between these data sets in the textbook:



Also, consider the following R output:

```
names(airports)

## [1] "faa"  "name" "lat"  "lon"  "alt"  "tz"   "dst"

names(planes)

## [1] "tailnum"      "year"         "type"         "manufacturer"
## [5] "model"        "engines"      "seats"        "speed"
## [9] "engine"

names(flights)

##  [1] "year"           "month"          "day"            "dep_time"
##  [5] "sched_dep_time" "dep_delay"      "arr_time"       "sched_arr_time"
##  [9] "arr_delay"      "carrier"        "flight"         "tailnum"
## [13] "origin"         "dest"           "air_time"       "distance"
## [17] "hour"           "minute"         "time_hour"

names(weather)

##  [1] "origin"     "year"       "month"      "day"        "hour"
##  [6] "temp"       "dewp"       "humid"      "wind_dir"   "wind_speed"
## [11] "wind_gust"  "precip"     "pressure"   "visib"      "time_hour"

names(airlines)

## [1] "carrier" "name"
```

a) Which data sets are you going to need to compute the distance covered by all flights leaving New York City?

b) Write the code that will output a table presenting the median departure delay of all flights for each airline leaving Newark (airport code `EWR`) .

c) Write the extra line of code that will output a table presenting the median departure delay of all flights for each airline leaving Newark, but this time in reverse alphabetical order.

d) Name a graphic that would best show all the information contained in the table in part b).

e) Write the code that will tabulate the mean humidity level recorded for all flights leaving New York City in July 2013.

4 Unisex Names... Revisited

Write the code that is going to generate an appropriate visualization to compare the trends in the "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 "Casey" and "Riley" from 1950 to 2014. As a hint, here are the first 10 rows of the `babynames` data set.

year	sex	name	n	prop
1880	F	Mary	7065	0.07
1880	F	Anna	2604	0.03
1880	F	Emma	2003	0.02
1880	F	Elizabeth	1939	0.02
1880	F	Minnie	1746	0.02
1880	F	Margaret	1578	0.02
1880	F	Ida	1472	0.02
1880	F	Alice	1414	0.01
1880	F	Bertha	1320	0.01
1880	F	Sarah	1288	0.01

5 Inference Basics

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. Describe how the process of bootstrapping could be used to create a plot and a range of possible values for the percentage of black residents, on average, by county throughout the entire US.

- Layout what the tidy data set would look like for this sample of 200 counties.
- You should carefully lay out each step of the bootstrapping process being as specific as possible. For example, you should be clear about the size of each sample and how many times you are repeating the process.
- Additionally, you should sketch a plot (free hand) of what the bootstrap distribution might look like and how one could use the distribution to help solve the problem. (Your numbers may not necessarily be correct, but it's important to get a sense of what the plot might look like.)

You should be as thorough as possible. If you can explain the bootstrapping process in this circumstance, you should be able to explain the process in any similar circumstance.