1 Short Answer

a) What's the chief difference between a barplot and a histogram?

b) What is overplotting? Name the two ways seen in class to deal with it.

c) We saw in class that there are two ways to create a barplot using `ggplot`: either using `geom_bar()` or using `geom_col()`. Illustrate the need for these two different ways using simple data frame examples.

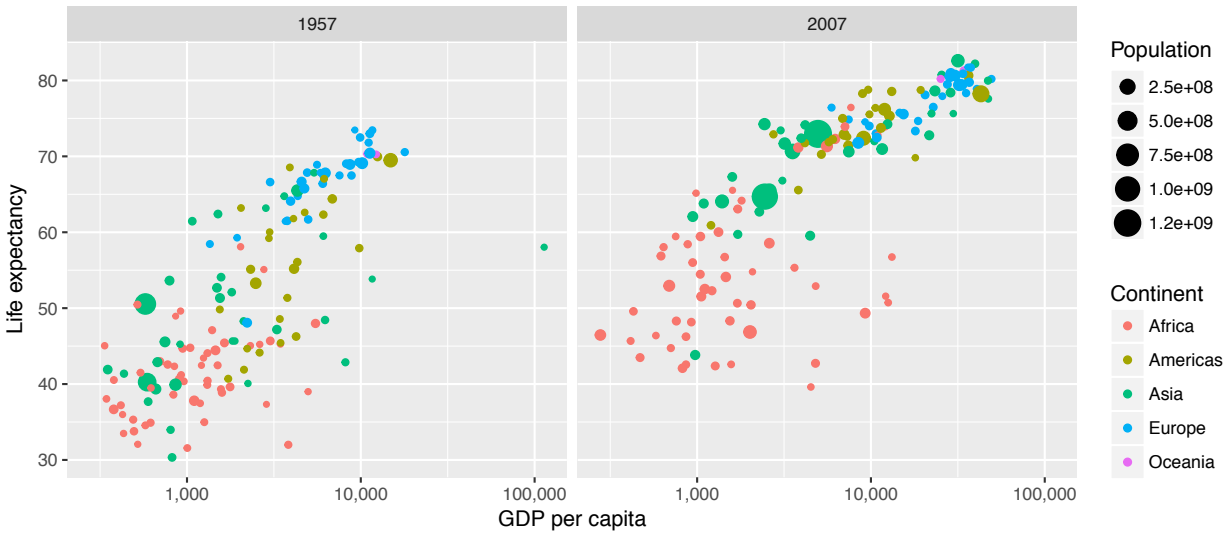
4 Gapminder

Consider a subset of the `gapminder` dataset we've seen numerous times in class:

```
gapminder_subset <- gapminder %>%
  filter(year %in% c(1957, 2007))
gapminder_subset

## # 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:



a) 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.

b) Why does the x-axis increase on a multiplicative scale (1000, 10000, 100000) instead of an additive scale (Ex: 1000, 2000, 3000)?

5 Family income by city

NPR recently posted an article titled "How Much (Or Little) The Middle Class Makes, In 30 U.S. Cities." It included the image on the following page.

a) This image most closely resembles what statistical graphic we've seen?

b) Which city has the third highest mean family income?

c) Which four cities have the highest income disparity in the US?

d) Quantify this income disparity for only one of the four chosen cities in part c) using a summary statistic of your choice.

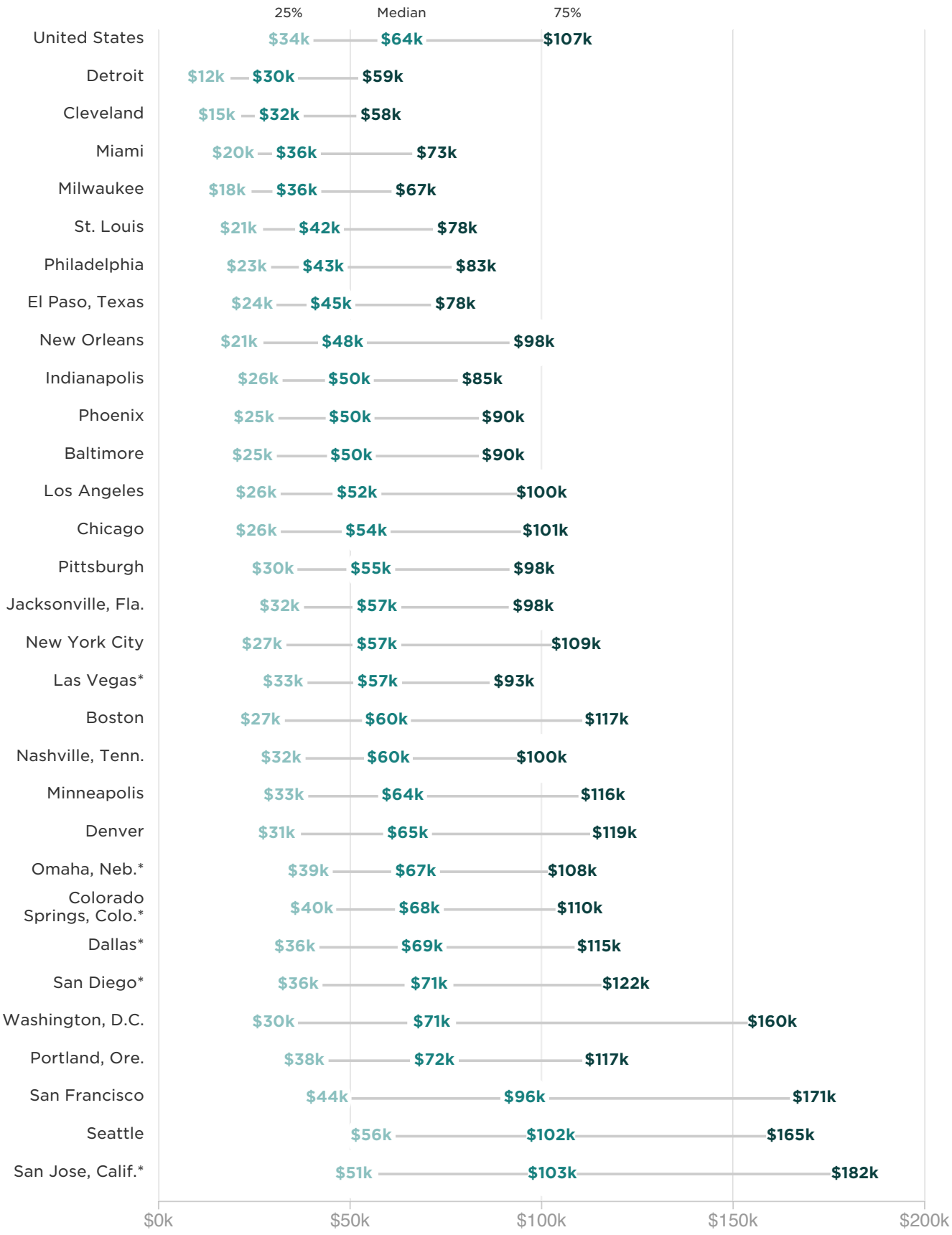
e) What proportion of Nashville families had a family income of $100K or more?

f) What proportion of Nashville families had a family income of $80K or more?

WRITE YOUR RESPONSES BELOW:

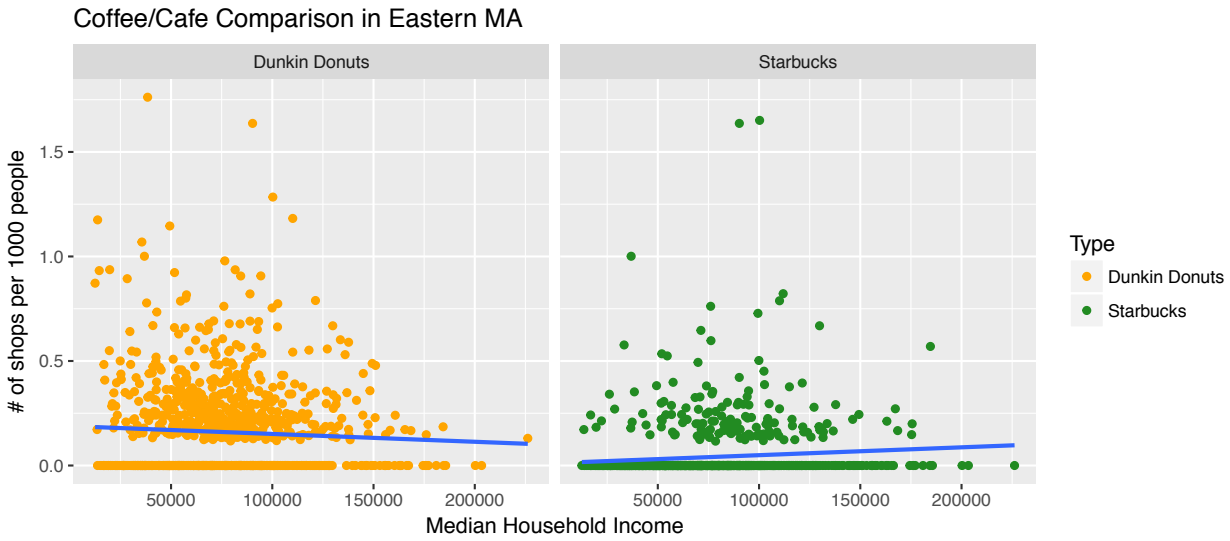
What Is Middle Class?

Family income by city, 2013



1 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 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. In particular, what `geom` is considered and what data variables correspond to what in this graphic.

b) Sketch out (in tidy data format) the dataset needed to make this graphic. Please also indicate the number of rows in this dataset.

c) Optional bonus question: Does the presented evidence support or contradict the researcher's suspicion? Why?