MPA 634  
Data Science for Managers  
Final Exam: Winter 2021  
Due: April 21 by midnight

# Exam parameters:

* Closed book
* No time limit
* Complete in one sitting
* Submit before starting the second part of the exam
* Due: 21 April before midnight
* Submit a word document

# Definitions and Concepts

1. Smoothing is a very important concept for data visualization. Explain how smoothing is affected in the following cases:
   1. binwidth and number of bins on histograms and freqpoly graphs
   2. bandwidth (bw) on densities and violin plots
   3. span on loess smoothed lines
2. Explain the difference between mutating and filtering joins.
3. Explain how class and method are related to each other. Use the summary method as an example to explain and illustrate your answer.
4. First define a tibble and then explain why a tibble is a list but not all lists are tibbles.
5. The following code includes each of the seven different parts of the grammar of graphics and creates the accompanying bar chart. Use the code to define and explain each of the seven parts of the grammar of graphics.

## 

A picture containing screenshot

Description automatically generated

Data:

Aesthetics:

Geometric objects:

Stats:

Position:

Coordinate system:

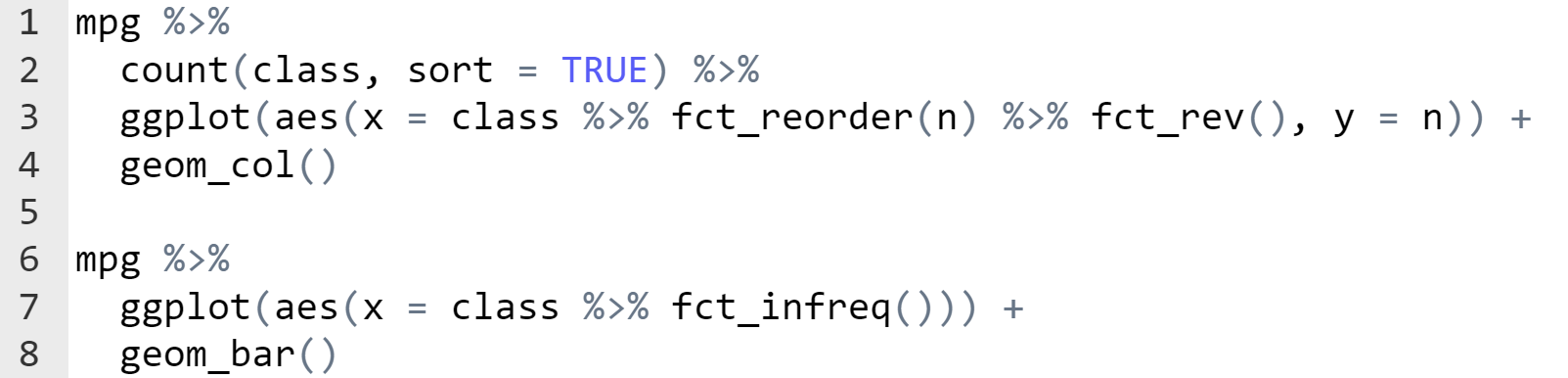
Faceting:

# Line by Line Code Interpretation

Please explain the following code chunks. By referring to line numbers, carefully explain the code and describe what it is accomplishing. Please avoid just repeating the R commands that are in the code.

## Code Chunk I: Pareto Charts

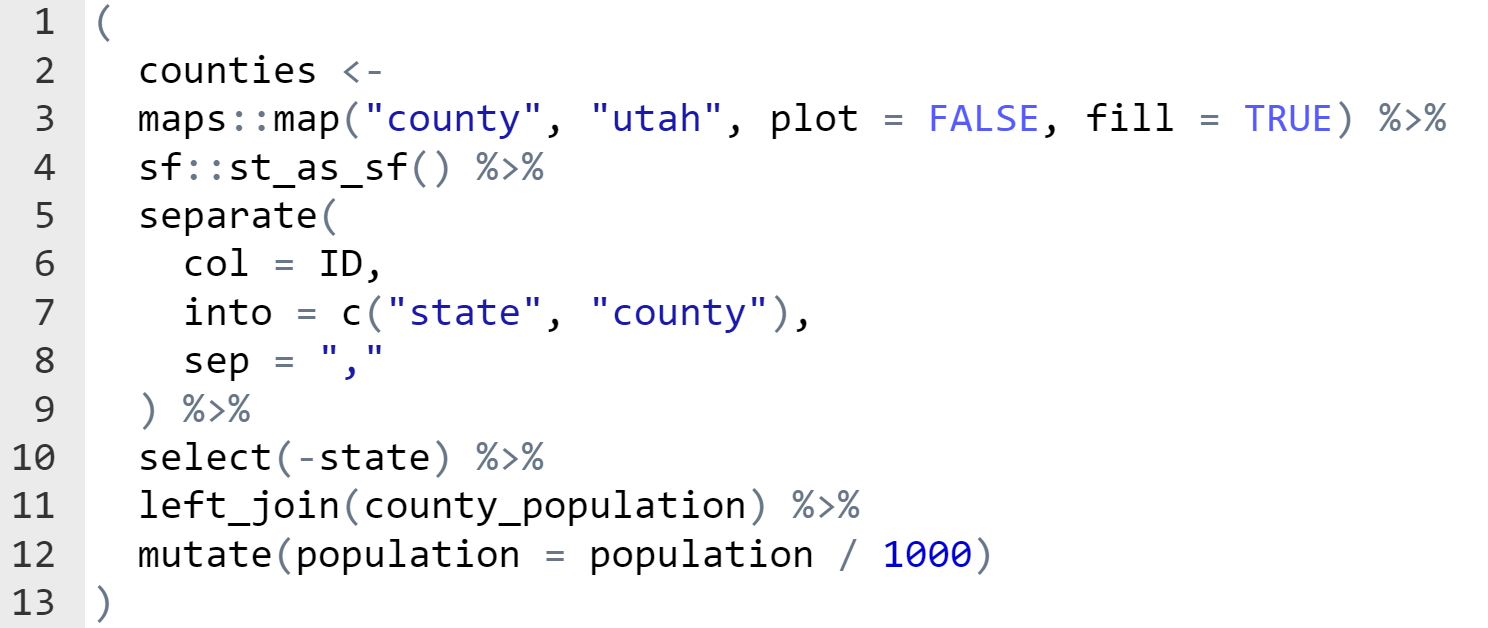
Here are two alternative ways to create Pareto charts. Please explain lines 2 – 4 and 7 - 8 and **then compare and contrast the two different approaches to creating Pareto charts**.



## Code Chunk II: Geospatial data

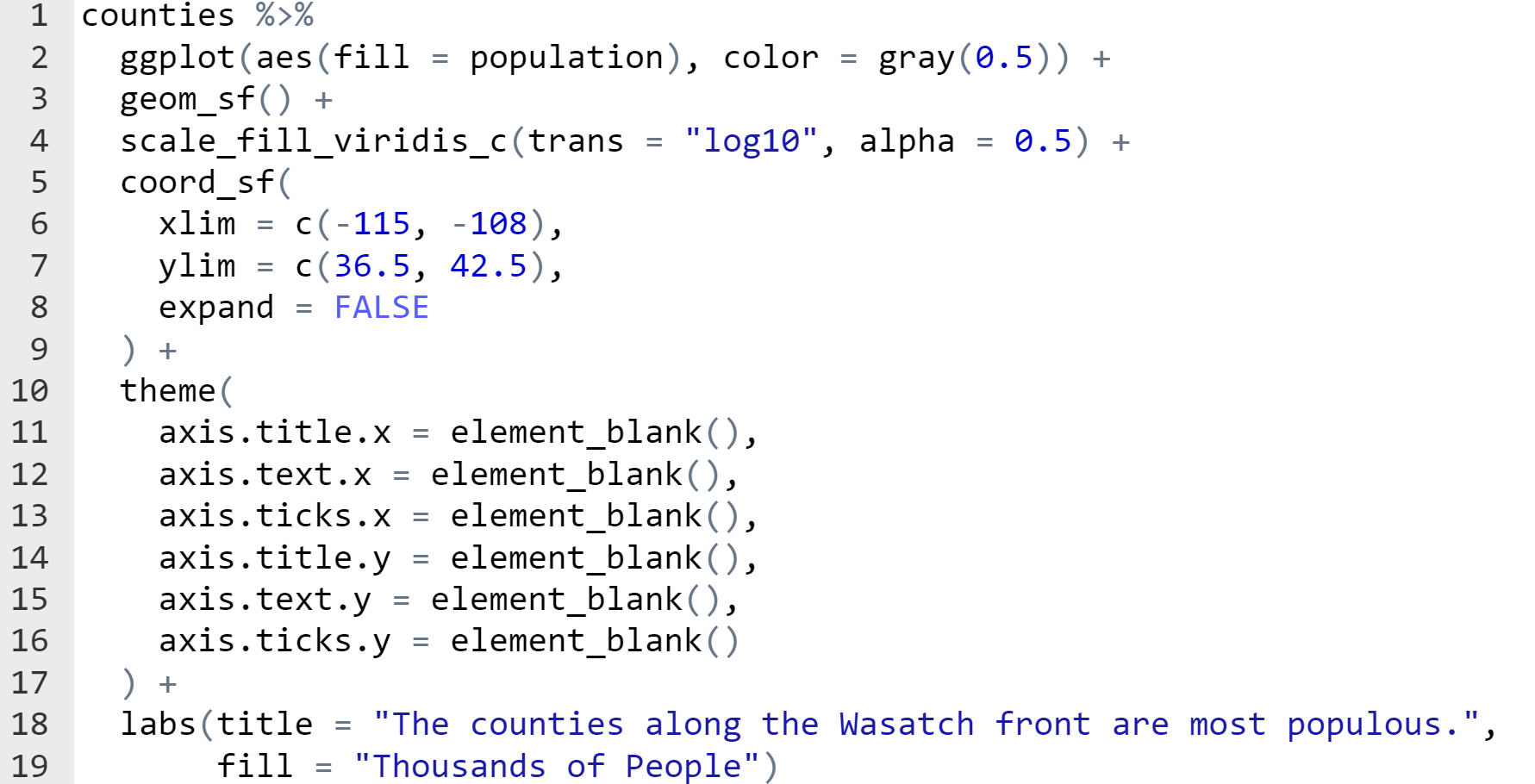
Code chunks II and III create a map of Utah with counties shaded by the size of their population.

Interpret lines 2 – 12 of the following script that creates the tibble needed for code chunk III.



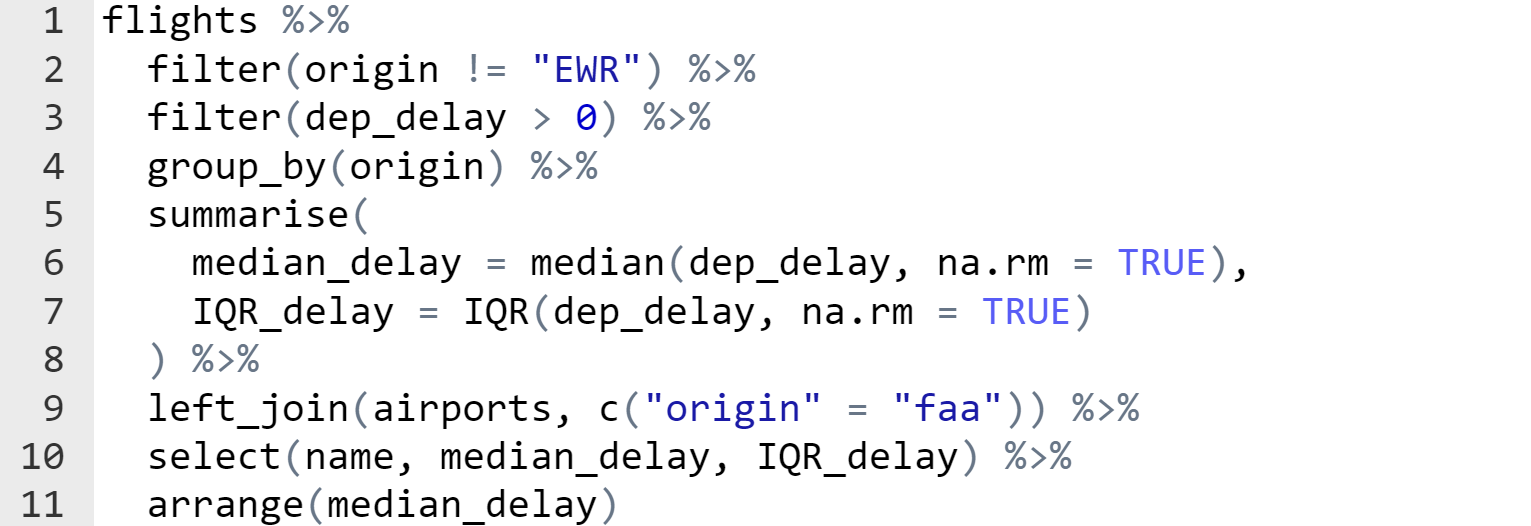
## Code Chunk III: Geospatial

The tibble county\_population contains two variables, the county name and its population. Explain all the lines of code which starts with the counties tibble created in code chunk II.



## Code Chunk IV: dplyr

Explain lines 2 - 11



## Code Chunk V: Communication Graphics

The presidential tibble has four variables initially:

* name of president
* start date of administration
* end date of administration
* political party

Interpret lines 2 - 12

