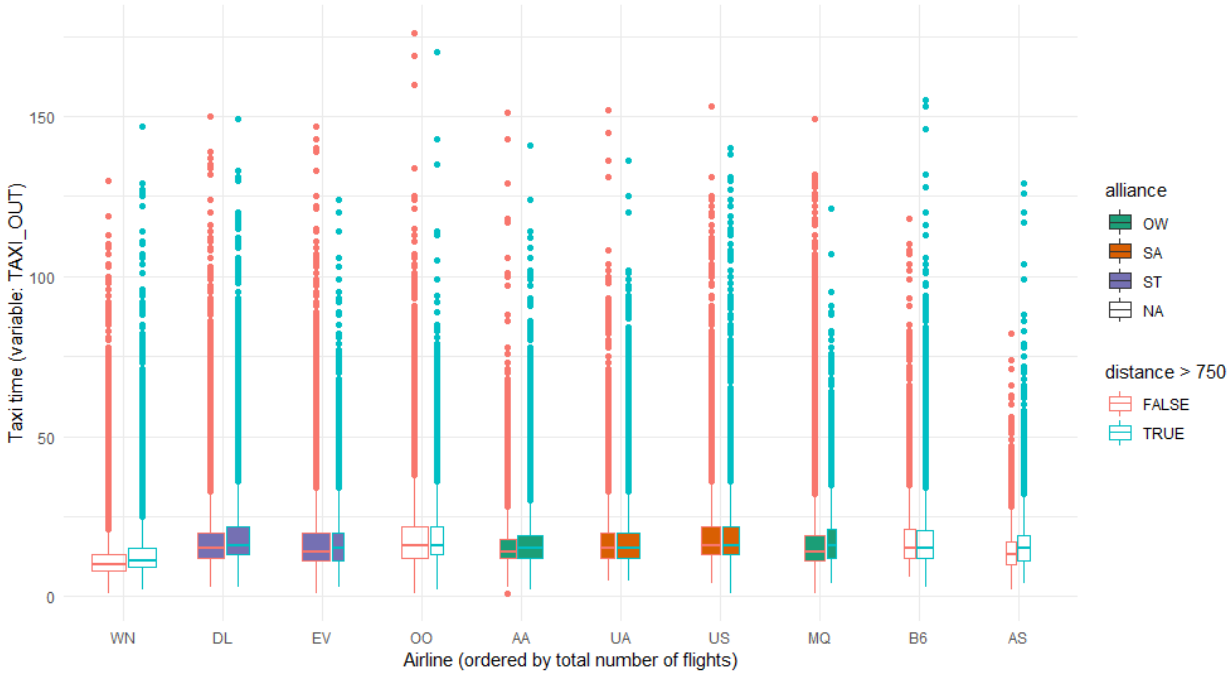
Problem 1:

Part a) Replicate the image below. You will need the "airlines.csv" dataset and the "flights_jan.csv" dataset. (15 pts)

Part b) Critique the plot. Your critique should include a description of what aesthetics are being used and anything you would like to say about their use. (5 pts)

Part c) Improve the plot (this may include dropping some information or including/modifying other information that you find to be a crucial aspect of the story). (10 pts)

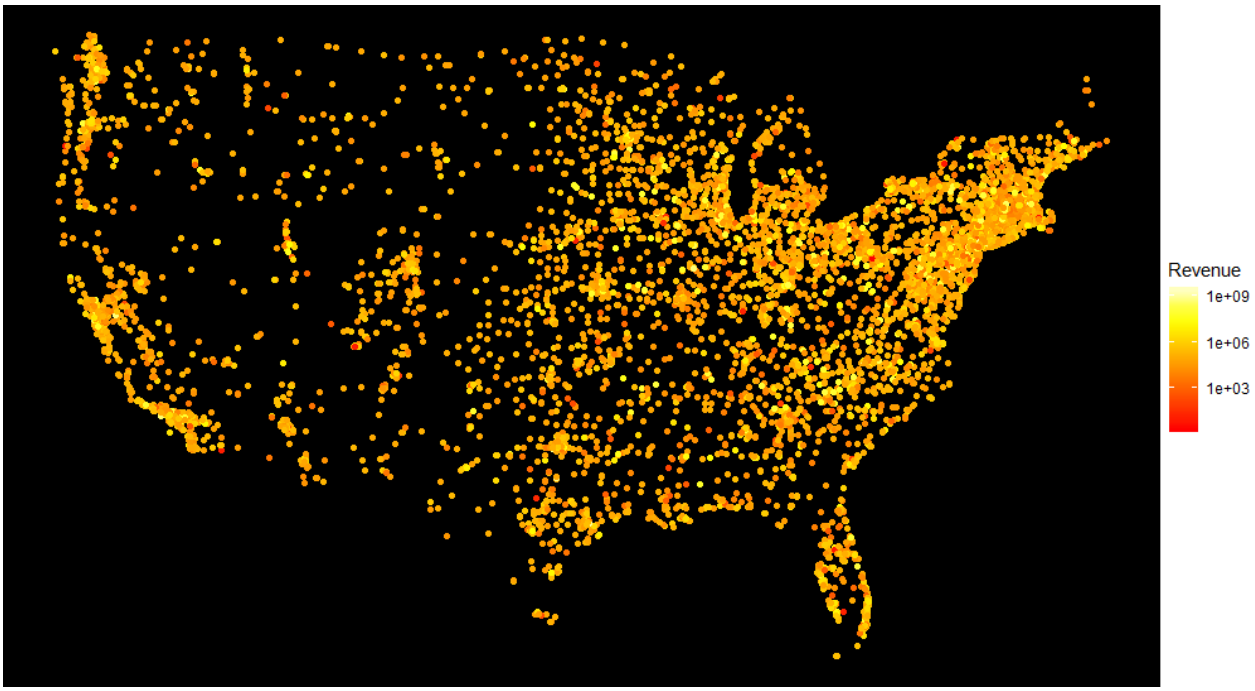


Problem 2:

Part a) Replicate the image below on museum revenue in the continental united states. You will need the "museums.csv" dataset. (15 pts)

Part b) Critique the plot. Your critique should include a description of what aesthetics are being used and anything you would like to say about their use. (5 pts)

Part c) Improve the plot. Likely, whatever story you are trying to tell with the museum dataset and revenue will not necessitate plotting the locations of the museum (10 pts)



Problem 3:

We typically think of word clouds as encoding frequency information of word content. Of course, anything that prescribes magnitude will work. In the dataset 'emnlp14age.csv' (in the emnlp2014_ageGenderLexica folder) the authors took facebook posts and used the text in the facebook posts to predict user age. The used (essentially) linear regression to do so. They then used the coefficients as indicators of words associated with age (so for instance, words with a large positive coefficient would indicate an older age, and words associated with a large negative coefficient would indicate a younger age).

There is a small peculiarity here. Let's say that the regression equation is $y = \beta_1 X_1 + \cdots + \beta_N X_K$

y is the age (centered and scaled, say) of the person who wrote the post, and the $X_1, \ldots, X_K$ are all the possible words that can go into a post. If the writer is very young, then you would want $\beta_1 X_1 + \cdots + \beta_N X_K$ to be very small (large negative number). If the writer is old, you want $\beta_1 X_1 + \cdots + \beta_N X_K$ to be large (large positive number). If the person is middle-aged, you would want $\beta_1 X_1 + \cdots + \beta_N X_K$ to be roughly, say, 0. So words associated with middle-aged writers have coefficients that are relatively small in magnitude, if we are to believe this way of categorizing words into age-groups.

More on the creation of this dataset (and more) can be found in the publication referenced in the README.txt.

Part a) Create two word clouds. One for old people and one for young people. This entails using the words with the largest positive coefficient for old people, and the largest negative coefficients for young people (though you will want to make those coefficients positive when you use them as weights for your word cloud). In this part, you should play around with the number of words to use. Play around with several options, like 50, 100, 150, 200. You may also find it helpful to set the "size" parameter to something fairly low, like 0.4. How did you decide the number of words to use? (20 pts)

Part b) The relative sizes of the words within the two wordclouds show two distinct patterns. Describe. What does this entail for extracting information from the word clouds? Is it harder in one word cloud than the other? Is there a fix for this? Or do you worry too much about distorting the information being presented? (10 pts)