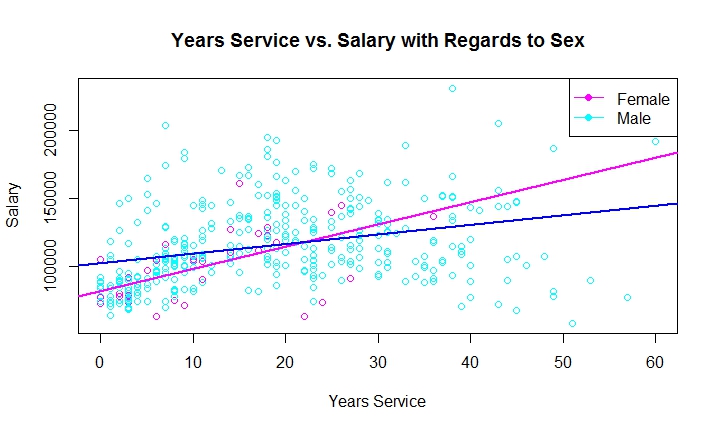
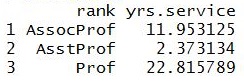
**Problem 1**

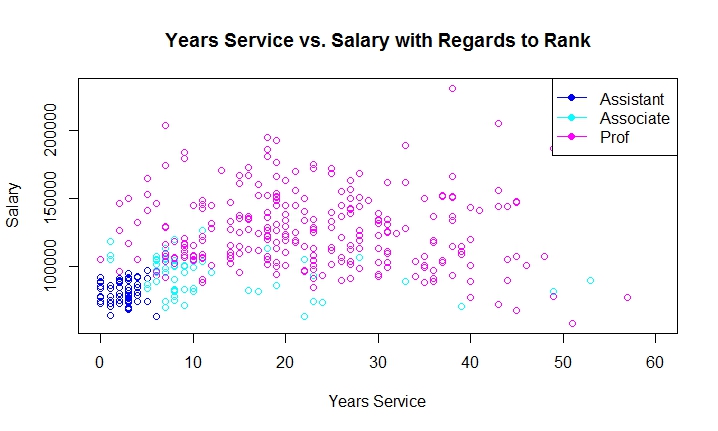
1. Do men and women make a discernable difference overall in their salaries as their time as a professor progresses?

Yes, they do. The first thing you will notice about this graph is that there are many more male professors than female. At the beginning of their career, men start out making more money than women, but as time goes on, women end up making more money more quickly than men. Men and women end up making the same amount of money around 21 to 22 years into their career.



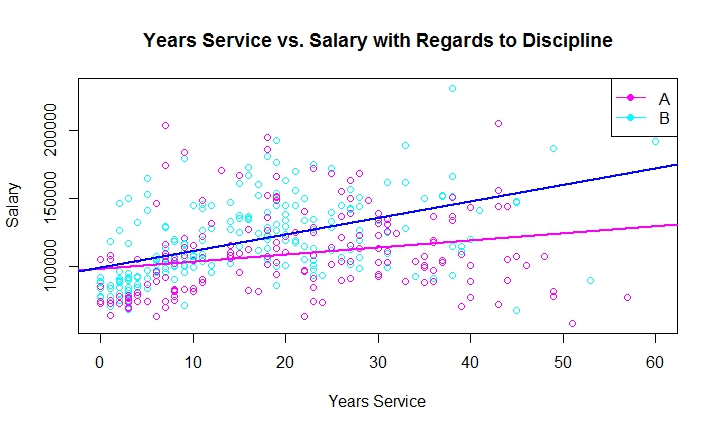
1. Relative to salary, what are typical times in terms of years of service for each rank?

To the right, I have a table with the average years of service for each rank, and below is a graph with years of service vs. salary in terms of rank. In summary, assistant professors do not make very much money and only stay assistant professors for a maximum of about 5 years; associate professors make a little more money than assistant professors and stay associate professors for 11 years on average; and full professors make much more money that assistant and associate professors and stay full professors for much longer.



1. What general relationship exists between discipline and salary as time of service progresses?

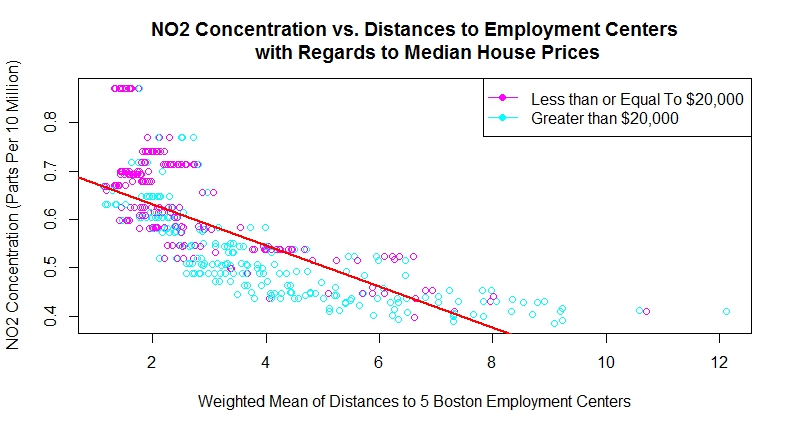
To the right, I have a table with the average years of service for each discipline, and below is a graph with years of service vs. salary in terms of discipline. Professors in each discipline start out making around the same amount of money, but as time goes on, discipline B ends up making much more money over time, despite professors in discipline A working for longer. Professors in discipline A rarely get raises.



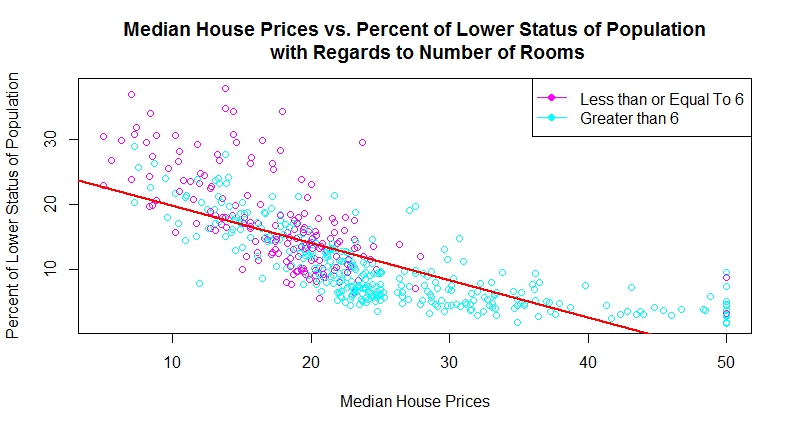
**Problem 2**

1. Explore the data using visualizations.
   1. What are some relationships that stand out as important to you and why?
   2. What hypotheses come from your data exploration?
   3. Please include at least 3 plots of interest and explain them.

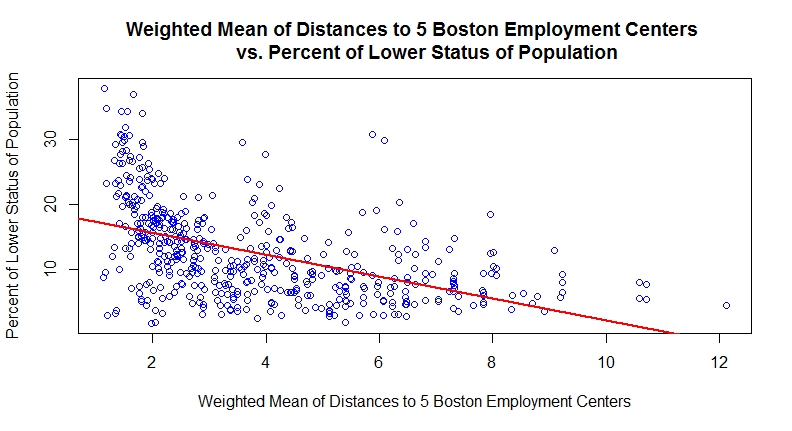
Below I have included 3 graphs which I found interesting.



This graph shows the NO2 concentration in the air near a given house vs. the mean distance the house is from 5 Boston employment centers with regards to median house prices. This graph indicates that the closer a house is to an employment center, the higher the NO2 concentration in the air. From this observation, I can hypothesize that these 5 employment centers give off a considerable amount of pollution. The pink data on the graph indicates the houses that have a value of less than or equal to $20,000. I can also hypothesize that the more NO2 in the air around the house, the less the house is worth. By using a trend line, I can also hypothesize that the safest distance from an employment center to live so that there is the least NO2 in the air possible is over 8 miles away.

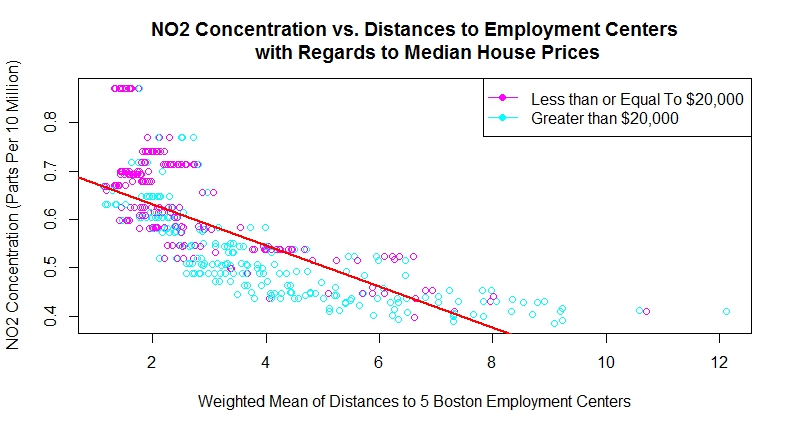


This graph shows the median house prices vs. the percent of the lower status of the population with regards to the number of rooms per house. This graph indicates that the more people of lower income status there are in a given population, the less the houses in the market are worth. I can hypothesize that the only houses the poorer people can afford are ones that aren’t worth very much, and that’s why the graph shows that trend. The graph also looks at the houses with 6 or less rooms in them. I can hypothesize from the graph that the houses with 6 or less rooms are both worth less than those with more rooms and the poorer people can mostly only afford houses with 6 or less rooms.



This graph shows the mean distance the house is from 5 Boston employment centers vs. the percent of the lower status of the population with regards to the number of rooms per house. This graph show that poorer people usually live very close to an employment center. I can hypothesize from this graph and the above graphs that the houses closest to the employment centers are the cheapest, and therefore the only ones that poorer people can afford.

1. Use regression modeling to construct a predictive model.
   1. What modeling assumptions (e.g. transformation of certain variables, inclusion/exclusion of certain variables, etc.) did you make and why?
   2. What is the final model you settled on? (i.e. the coefficients)
   3. How good is your fit?
   4. Predict prices for a random 10% of the houses contained in the data and compare to actual values. What is the mean and standard deviation of the errors?



All files and graphs can be found on my GitHub: <https://github.com/gingerchic21/DSCI401>