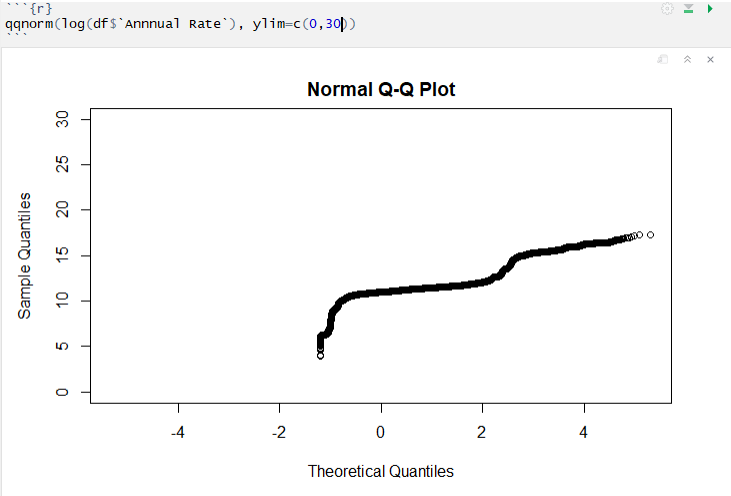
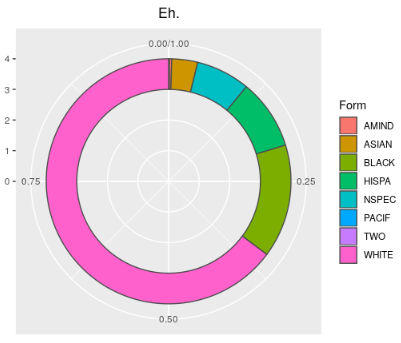


The distribution of annual rate is too strongly skewed left. So it was transformed using log, that makes it look more normal.



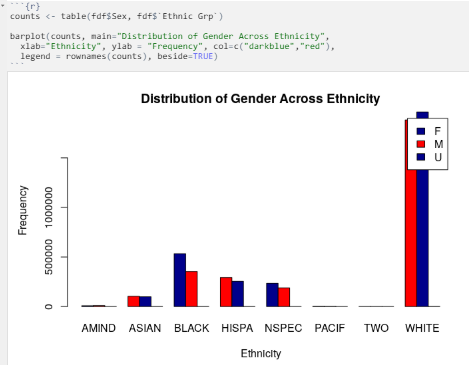
However, the qqplot shows the transformed data is not normally distributed.

After cleaning the data by removing part time and non-regular employees, we found the ethnic group distribution:



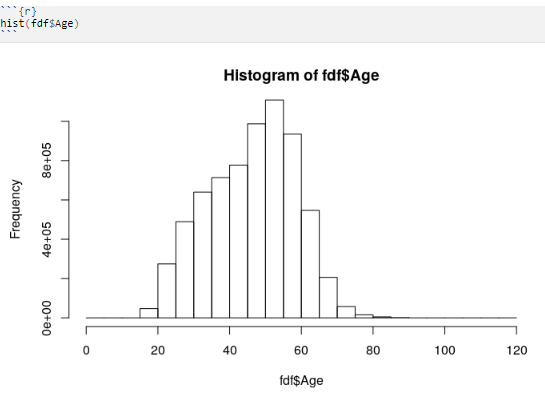
Most of the people in the data are white, followed by black, Hispanic, Pacific, and Asian. This makes sense considering the data come from Connecticut.

What about the gender distribution between these groups?

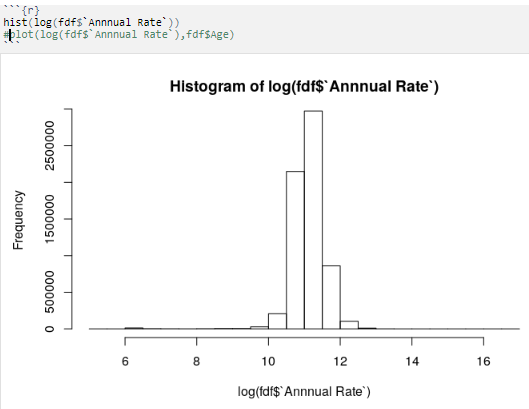


In the cleaned, subsetted data, there are substantially more black women than black men. This plot is hard to read given how sometimes men are on the right, but other times they are on the left. You also cannot see the American Indian, Pacific, and Biracial ethnic groups.

Next we’ll try to make a model from the data. Let’s look at the distribution of age and of Annual Rate.

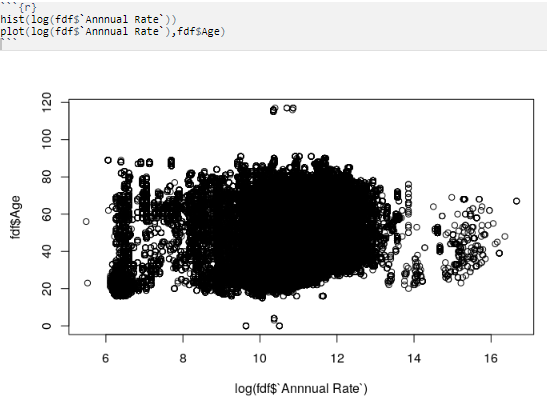


Age distributions look ok, perhaps it is skewed to the left.

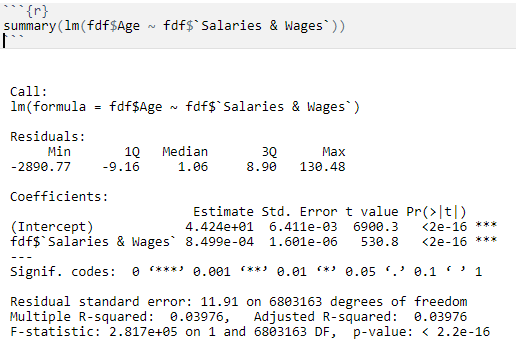


The transformed annual rate data looks normal.

Let’s plot the age against the annual rate.



From the image, age does not make a good predictor of salary. Perhaps the plot would look less clustered if we only looked at salaries for specific groups or careers. The model summary also shows little association.



The adjusted R^2 is .0398, which means salary is not a good predictor or age.

