2/28/2021

Multiple Linear Regression on Salary Dataset

1. ***Data Background***

The dataset contains salary figures of Assistant Professors, Associate Professors and Professors. The data was collected over a period of nine months in the years 2008 and 2009 in a college in the U.S. as part of the on-going effort of the college’s administration to monitor salary differences between male and female faculty members.

The data frame includes **397 observations** and **6 variables**.

The variables are:

"rank" "discipline" "yrs.since.phd" "yrs.service" "sex" "salary"

The data frame includes **397 observations** and **6 variables**.

**Rank (categorical)**: a factor with levels AsstProf, AssocProf and Prof

**discipline (categorical)**: a factor with levels A (“theoretical” departments) or B (“applied” departments).

**yrs.since.phd (continuous)**: number of years since PhD.

**yrs. service (continuous)**: number of years of service.

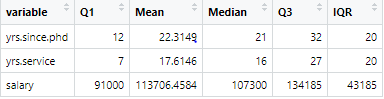
**sex (categorical)**: a factor with levels Female and Male

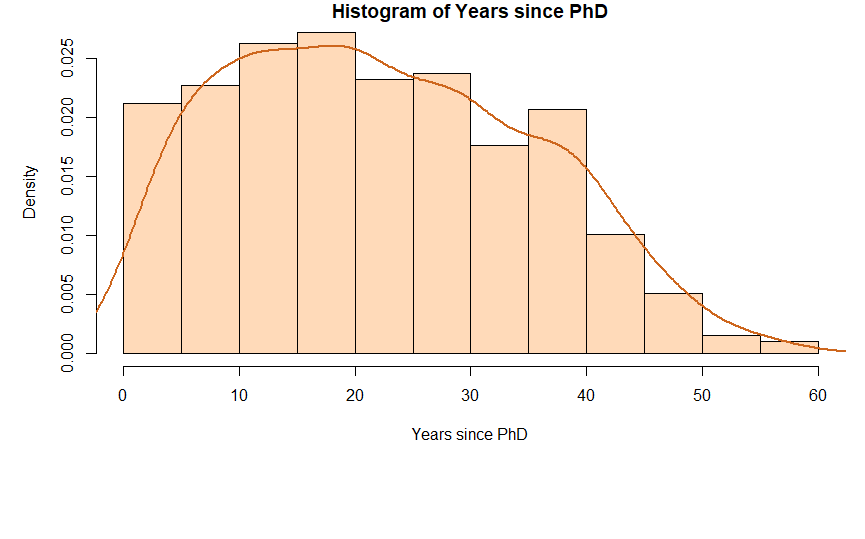
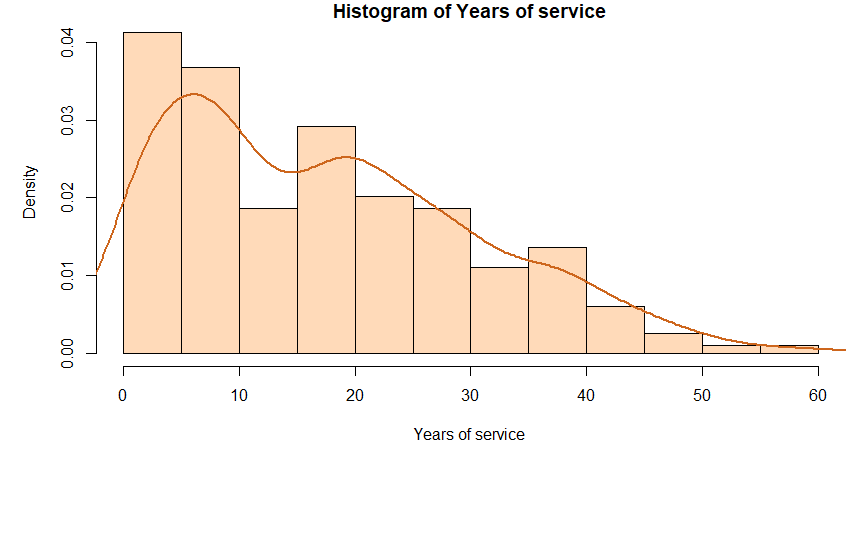
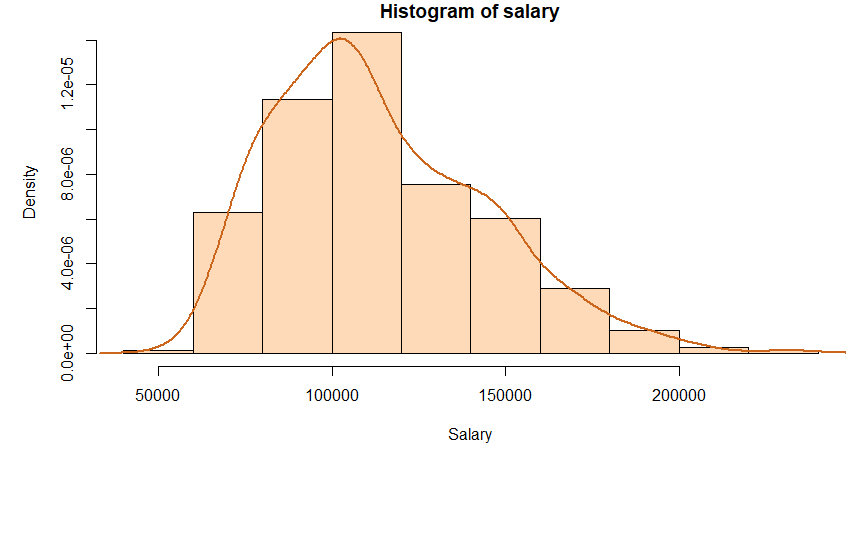
**salary (continuous)**: nine-month salary, in dollars.

1. ***Summary Statistics and Exploratory Analysis***

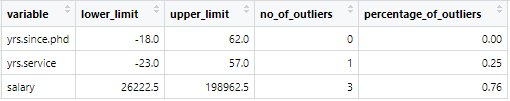
To obtain an overall idea of the variables and values present in the dataset, a 5-point summary followed by an outlier analysis was performed in R as a quality check.

**2.1. Five point summary of the continuous variables and Histograms**

For all the 3 continuous variables, it can be observed that the mean is higher than the median i.e. the distributions are right skewed (and not symmetric). This can also be observed from their respective histograms.

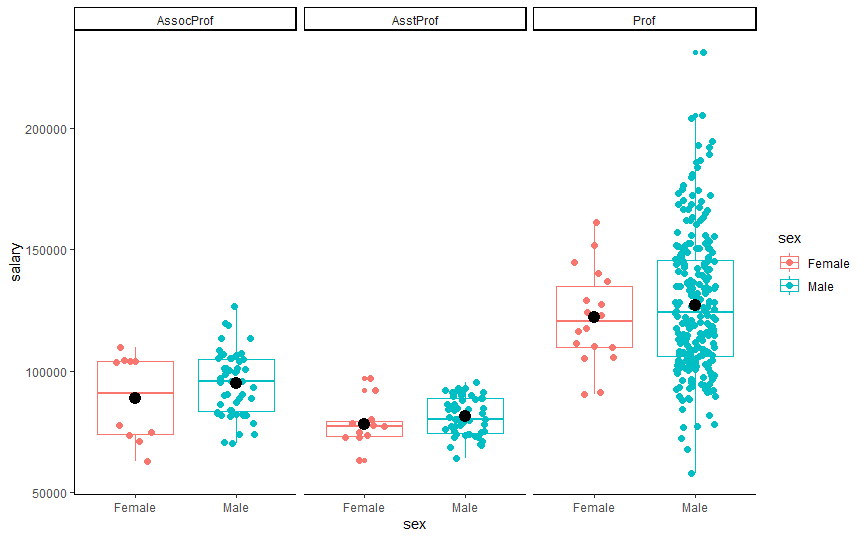
Outlier analysis was conducted to understand if any of the variables have outliers which require treatment.



From the above table, we can conclude that a very small percentage of outliers are present in our dependent variable Salary which can be overlooked.

**2.2 Boxplot**

As a part of EDA, I plotted a boxplot to see how the salary varies across different ranks and sexes present in the dataset.



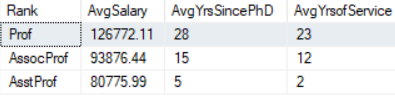
It can be observed from the plot that as the rank increases the salary also increases with most amount of variability present in the salaries of the rank Professor. The mean salary (ie the black dot) for male is comparatively higher than female for all the 3 ranks.

**2.5 Data analysis across ranks, discipline and sex**

SQL queries were used to calculate the overall average values of the 3 continuous variables present in the data. To have a better understanding of the distribution of salaries across the different categorical variables, the averages were calculated across different ranks, disciplines and sex.

|  |  |  |
| --- | --- | --- |
| **Years since PhD** | **Years of service** | **Salary** |
| 22 | 18 | 113706.46 |

**Rank wise Analysis**

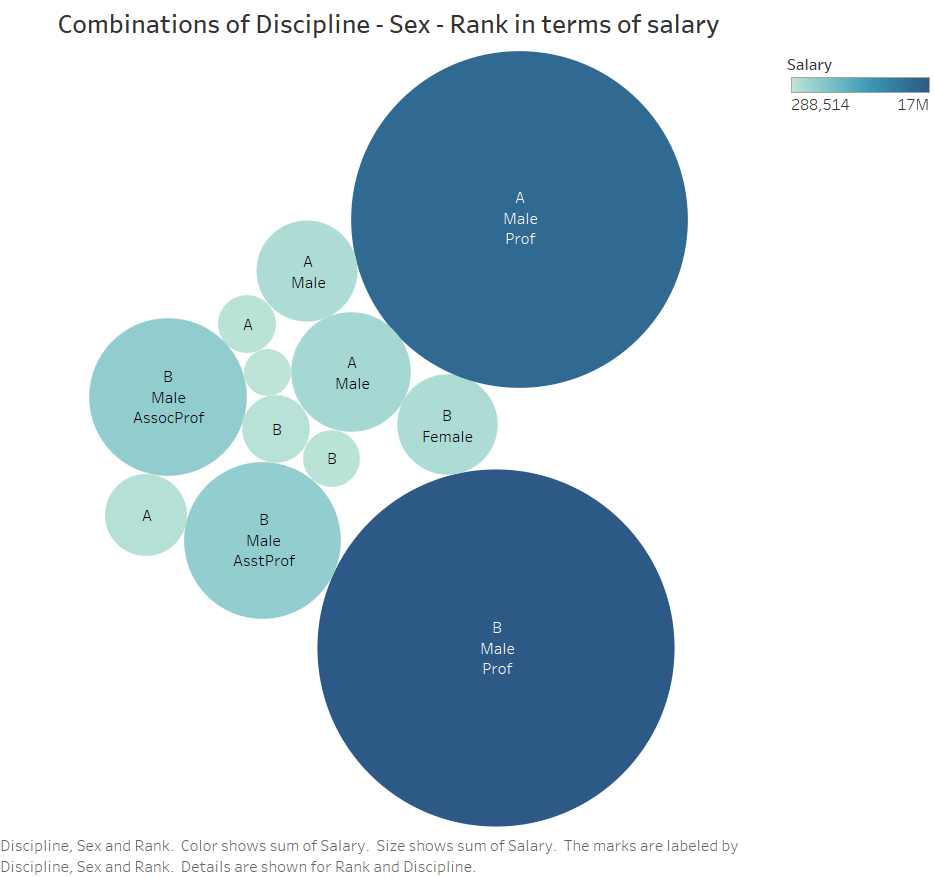


**Discipline wise Analysis**



**Gender wise Analysis**

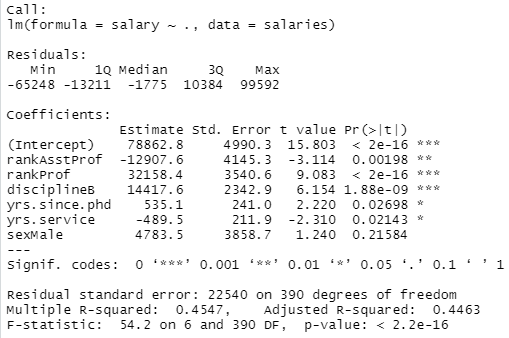
From the above 3 tables, we can observe that the rank Prof, the discipline B (applied departments) and the sex Male has a higher value of average salary in comparison to their respective counterparts.

  
  
The above packed Bubble chart from Tableau shows the combinations of Discipline – Sex and Rank that get the highest and the lowest salary.  
  
As visible from the size of the bubbles, the maximum salary is received by Male Professors in Discipline B (Applied Departments) and the minimum salary is received by Female Associate Professor in the discipline A (theoretical departments).

1. ***Modeling***

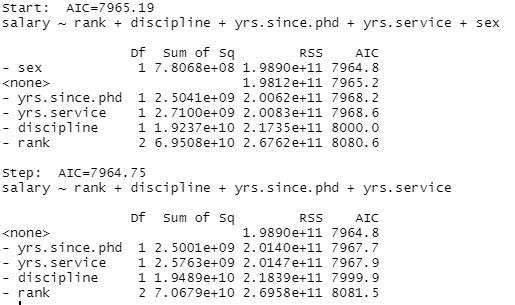
**3.1 Multiple linear regression model**

 For a multiple linear regression model, I used the variable ‘salary’ as my dependent variable and the rest as my independent variables.

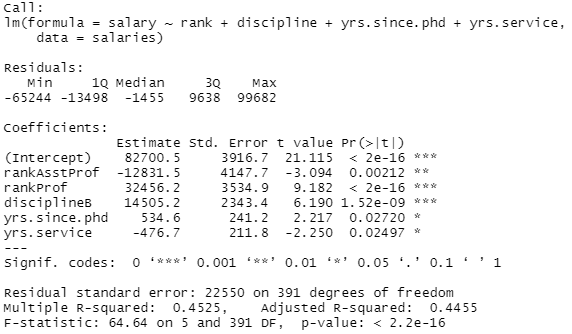


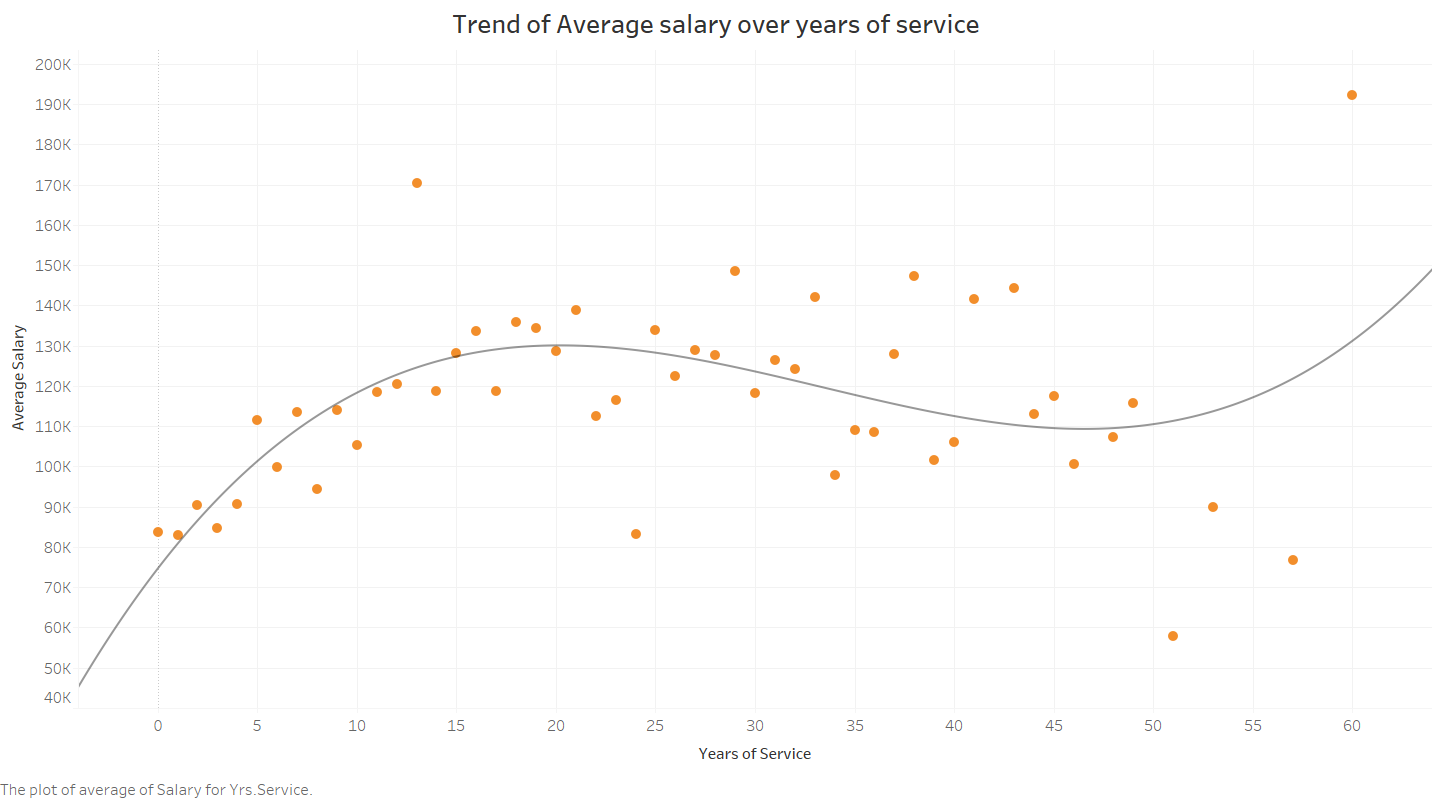
From the above table, we can observe that the intercept value implies that a person gets an average salary of $78862.8. For the variable rank, “associate professor” is set to the reference level and the average salary increases with increase in rank. Hence it is a positive coefficient for the rank Prof and a negative coefficient for the rank AsstProf.  
  
 We also observe that with an increment of a year of service, the average salary drops by $489.5 holding all other variables constant.

Category A is set to the reference level for the variable discipline. Category B (applied departments) results in an average increase of $14417.6 in salary compared to discipline A (theoretical departments) keeping other variables at constant.  
  
Since there is an underlying correlation between the independent variables, it may result into a bias in the model. To get rid of such a situation (also known as multicollinearity), we perform stepwise regression to filter out variables that may have a direct relation with any other independent variable. In a backward stepwise regression, the process starts with initially fitting all the variables and after that, with each iteration, it starts eliminating variables one by one if the variable does not improve the model fit. The AIC metric is used for checking model fit improvement.



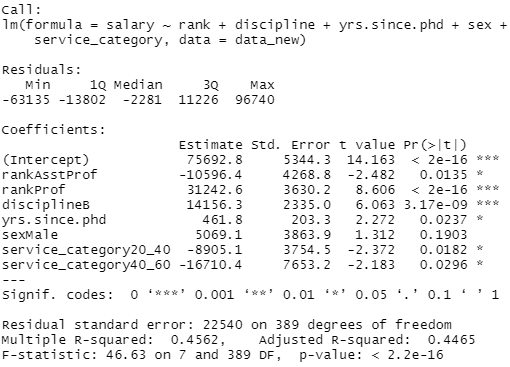
As is evident from the 2nd table, the variable sex was eliminated to improve the AIC value. Refitting the model without the variable ‘sex’.





On plotting the salary vs years of service plot, I noticed that there is a steep increasing trend in salary amount for the first 20 years of service. To understand the effect of the variable ‘Years of service’ better, the values of the variable was bucketed into 3 categories namely, ‘under 20’, ‘20-40’, ‘40-60’ as the range of this variable is from 0 to 60 years. The distribution of records in the 3 categories are as follows:  
  


Refitting the model using this newly created categorical variable,



The model coefficient table showed that as the service time increases the salary decreases (negative coefficients) when compared to the 0–20 years of service. Compared to 0–20 years of service years category, a person is in 20–40 years of service gets on average 8905.1$ less salary, similarly, a person is in 40–60 years of service earns 16710.4$ less salary.

1. ***Conclusion***

The multiple linear regression model shows the varying impact of different ranks and disciplines on the salary of different faculty members of a college in the US. The analysis and the weights of the different categories can help monitor and reduce the salary difference in future.