AS3: Statistical Analysis and Hypothesis Testing

The Fifth National Bank of Springfield is facing a gender discrimination lawsuit. The charge is that its female employees receive substantially smaller salaries than its male employees. Imagine that you are an analyst for the plaintiff working on this case. The bank's employee data are stored in the file ***banksalary.csv***, which you can find in the assignment section on Canvas. For each of its 208 employees, the dataset includes the following variables:

- **EducLev:** education level, a categorical variable with categories 1 (finished high school), 2 (some college courses), 3 (obtained a B.A.), 4 (took some graduate courses), and 5 (obtained a graduate degree).
- **JobGrade:** a categorical variable indicating the current job level, the possible levels being 1 through 6 (6 is the highest).
- **YrsExper:** years of experience with the bank.
- **Age:** employee's current age.
- **Gender:** a categorical variable with values "Female" and "Male".
- **YrsPrior:** number of years of work experience at another bank prior to working at Fifth National.
- **PCJob:** a categorical yes/no variable depending on whether the employee's current job is computer related.
- **Salary:** current annual salary.

** After completing the assignment, please submit the answers along with the script file used for the assignment.* ***<u>You are highly encouraged to use R Markdown for the assignments.</u>***

<u>INSTRUCTIONS:</u>

1) Import the csv file into R and present the descriptive statistics of the numerical variables as well as the categorical variables in the dataset.
2) A plaintiff's lawyer claims that there is a significant difference in average salary between female employees and male employees. As an analyst for the plaintiff, how would you support this claim? Use a t-test and explain the results as well as your interpretation.
3) Transform **EducLev** into several dummy variables. The number of dummy variables you create will need to depend on your logical judgement. Also transform **JobGrade**, **Gender**, and **PCJob** into dummy variables.

4) The defense counsel tries to counter against the plaintiff's argument by showing that the mean difference between the two groups is biased because he or she did not control for several other factors/variables. Estimate a multiple regression model to strengthen/bolster the plaintiff's justification, then write a report explaining your results.
    - Also discuss about: what R-squared is and what it means, what the meaning of the t-values and the coefficients are (or estimates).

5) Do these data provide evidence that there is discrimination against female employees in terms of salary?

Extra Credit:

a. You may get more interesting results to talk about by including interaction terms in your regression model. Explain what an interaction term is, how we can estimate a regression model with interaction terms and how we could interpret the results.

b. How would you determine whether the interaction terms contribute in a meaningful way to the explanatory power of your estimation model?