Data Scientist Modeling Exercise

Instructions: Please review the following document in its entirety. Please use any statistical program or language of your choice to address the questions below. Additionally, please prepare a presentation that addresses each of the questions below: `R` Markdown, Jupyter Notebook, PowerPoint, Shiny/GluViz, and Beamer/LaTeX are all acceptable presentation formats, and do not forget to include a copy of the code/scripts you used for this exercise. All code and related output will be reviewed by the interview panel. Be prepared to justify any assumptions that you make and present all supporting evidence of your analysis (e.g., tables, charts).

Description of Dataset

You have been provided the `Salaries` dataset from the `car` package in `R`. The dataset contains 397 rows and 6 variables, described as follows.

Variable	Description
`rank`	A factor with three levels: AsstProf, AssocProf, Prof (corresponding to assistant professor, associate professor, and full professor).
`discipline`	A factor with two levels: A ("theoretical" departments) and B ("applied" departments)
`yrs.since.phd`	Years since PhD was first received
`yrs.service`	Years of service
`sex`	A factor with two levels: Female and Male
`salary`	Nine-month salary, in dollars

Tasks to Be Completed

Analysis

Answer the following research questions to the best of your ability. Be prepared to walk the interview panel through your code and your thought process.

1) What percentage of records are Assistant Professors with less than 5 years of experience?
2) Is there a statistically significant difference between female and male salaries?
3) What is the distribution of salary by rank and discipline?
4) How would you recode `discipline` as a 0/1 binary indicator?

Model Building

Build a predictive model using salary as the response. Use any model specification and related data transformations you deem to be most appropriate for this analysis. Prepare a clear and understandable interpretation of the results.

In addition, create a 0/1 binary indicator using salary, where the indicator has a value of 0 if the salary is below the median and 1 otherwise. Build a new model using this indicator as the response, using the same set of predictors used to build the previous model.

Dataset Enhancement

Suppose you've been asked to enhance the existing dataset to enable additional rounds of analysis and facilitate additional model development.

1) State at least three research questions you would like to address and describe your thought process behind how you formulated these research questions.
2) Prepare a list of 5-7 additional attributes you would like to add to the dataset. Prepare a brief explanation for each attribute.
3) Estimate and justify the appropriate sample size (and sampling technique, if desired) that would be requires to address the research questions you defined.