MODELING EMPLOYEE SALARY DETERMINANTS USING DUMMY VARIABLES IN R

Client Overview:

A human resources consulting firm serving tech startups in California sought to build a salary benchmarking model. The client had employee data but lacked a robust way to justify pay scales across gender, education level, and job role. They needed a clean, interpretable model to present to both internal HR and external clients.

Problem Statement:

The HR team had accumulated employee data from 15 startups, but they were manually estimating salary ranges and struggling with inconsistencies. They also wanted to evaluate if any pay disparity existed across gender or education level. They asked for a transparent statistical model that could handle both numeric and categorical predictors.

Solution Approach:

I proposed a Multiple Linear Regression model using dummy variables in R. The purpose was twofold:

1. Quantify the relationship between experience, job roles, and qualifications with salary.
2. Identify if there was any statistical evidence of pay inequality between groups.

1. Data Preparation

- Imported the raw data (Excel) containing ~600 employee records.
- Cleaned the data: handled missing values, standardized currency fields, ensured consistent category labeling.
- Created dummy variables for categorical predictors:
    - **Gender** (Male = 1, Female = 0)
    - **Education_Level** (Bachelors, Masters, PhD)
    - **Job_Role** (Software Engineer, Data Analyst, Product Manager, etc.)

2. Variable Setup

- Dependent variable: Annual_Salary
- Independent variables:
    - Years_of_Experience (numerical)
    - Gender (dummy)
    - Education_Level_Masters, Education_Level_PhD (Bachelors as baseline)
    - Job_Role dummies (Product Manager as baseline)

3. Model Building

- Built model using lm() function in R.
- Used contrast coding to properly interpret coefficients relative to reference categories.
- Checked for multicollinearity using VIF and correlation matrix.
- Removed redundant dummy variables to avoid dummy variable trap.

4. Model Interpretation and Testing

- **Adjusted R^2:** 0.77
- **F-statistic:** Highly significant ($p < 0.001$)
- Coefficient interpretation included:
    - Holding everything else constant, employees with a **PhD earned ~\$18,000 more** than those with a Bachelor's.
    - Software Engineers earned ~\$10,500 less than Product Managers, on average.
    - **No statistically significant difference** found in salary based on gender ($p > 0.2$).
- Model assumptions were validated using:
    - Q-Q plot for normality
    - Residual vs Fitted plot for homoscedasticity
    - Breusch-Pagan test for variance

5. Deliverables

- Delivered a clean R Markdown report with:
    - Full model summary
    - Coefficient tables
    - Interpretations and business recommendations
    - Visual plots of salary trends by role and education
- Provided a walkthrough session to HR explaining how to interpret dummy variable coefficients.

Business Impact:

- Equipped the client with a data-driven framework for future salary decisions.
- Report served as a foundational document for their “Pay Equity Policy.”
- Helped startups under the client’s umbrella benchmark salaries against data-derived norms.

Tools Used:

- **Software:** R (RStudio)
- **Libraries:** tidyverse, broom, car, psych, stargazer
- **Reporting:** R Markdown (HTML and PDF exports)