Title: Analysis and Visualization of Salary Data

1. **Introduction**

   - This script is aimed at analyzing a dataset containing salary information.
   - Libraries are installed and loaded for data manipulation, visualization, and analysis.
   - The dataset is read from a CSV file and its structure and summary are examined.
   - Various visualizations are created to understand the relationships between variables.
   - A linear regression model is fit to predict current salary based on several predictors.

2. **Library Installation and Import**

   - Libraries such as `tidyverse`, `ggplot2`, `caret`, etc. are installed for data manipulation and visualization.
   - They are then loaded into the R environment.

3. **Data Loading**

   - The dataset 'table_203b.csv' is loaded from the specified path.
   - Basic properties of the data are examined using functions like `head`, `tail`, `str`, `summary`, `dim`, and `colnames`.

4. **Data Exploration**

   - Correlation matrices are plotted to understand the linear relationships between numeric variables.
   - Histograms are plotted to visualize the distribution of starting and current salaries.

5. **Data Visualization**

   - Scatter plots are created to visualize the relationships between starting and current salaries.
   - Bar plots are created to compare the average current and starting salaries across different categories like sex and job category.
   - Scatter plots are enhanced with color to show the distribution of data points by sex, job category, and race.
   - Pairwise scatter plots and correlations are created to visualize relationships between numeric variables grouped by sex.
   - Box plots are created to compare the distribution of current salary across different categorical predictors.

6. **Model Fitting**

   - A linear regression model is fit using `lm` function to predict the logarithm of current salary based on several predictors.
   - Assumptions of the model are checked using diagnostic plots.
   - Influence measures and influence plots are created to identify leverage, influence, and outliers.

7. **Model Evaluation**

   - The data is split into training and testing sets using an 80-20 split.

  - A full model is fit on the training set, and stepwise selection is used to find the best model based on AIC.
  - Predictions are made on the test set using both the full model and the best model.
  - Root Mean Square Error (RMSE) and R-squared are calculated to compare the out-of-sample performance of the full model and the best model.

This script provides a comprehensive analysis and visualization of the salary dataset, aiding in understanding the factors affecting salary and building a predictive model to estimate the current salary based on various predictors.