Introduction

This repository contains the implementation of a simple linear regression model applied to the Salary Dataset from Kaggle. The primary goal of this project is to understand the relationship between years of experience and salary, and to predict the expected salary for an individual based on their years of experience.

Project Overview

Simple Linear Regression is a fundamental statistical method used to determine the relationship between two variables: an independent variable and a dependent variable. In this project:

- **Independent Variable:** Years of Experience
- **Dependent Variable:** Salary

The dataset used in this project includes 30 entries with years of experience and corresponding salaries. The objective is to fit a linear model to this data and evaluate its performance.

Dataset Description

The dataset includes the following columns:

- `Unnamed: 0`: Index column (not used in the analysis)
- `YearsExperience`: Number of years of experience
- `Salary`: Annual salary corresponding to the years of experience

Exploratory Data Analysis (EDA)

The following visualizations and analyses were conducted:

1. **Correlation Matrix and Heatmap:**
    - A heatmap of the correlation matrix was used to visualize the strength of the relationships between variables.
2. **Pair Plot:**
    - A pair plot was used to visualize the relationships between all pairs of variables in the dataset.
3. **Distribution of Years of Experience and Salary:**
    - Histograms with KDE were plotted to understand the distribution of years of experience and salary.
4. **Scatter Plot:**
    - A scatter plot was created to visualize the relationship between years of experience and salary.
5. **Box Plots:**
    - Box plots were used to detect outliers in years of experience and salary.

6. **Pearson Correlation Coefficient:**
    - The Pearson correlation coefficient was calculated to quantify the linear relationship between years of experience and salary.
7. **Regression Plot:**
    - A regression plot was used to visualize the fitted linear relationship between years of experience and salary.

Model Training and Evaluation

- **Model:** Linear Regression
- **Train-Test Split:** 80-20 split
- **Metrics:**
    - **Mean Squared Error (MSE):** 42,510,267.47
    - **R-squared (R^2):** 0.93

Results

- The linear regression model exhibits a high R^2 value of 0.93, indicating a strong fit to the data.
- The Mean Squared Error (MSE) is substantial but consistent with the scale of the data, reflecting that the model's predictions are relatively accurate.

Code Summary

The code includes:

1. **Data Loading and Preprocessing:**
    - Load the dataset and perform initial exploration.
    - Drop the `Unnamed: 0` column as it is not needed.
2. **Exploratory Data Analysis (EDA):**
    - Visualize the data using various plots and calculate correlation.
3. **Model Training:**
    - Train a simple linear regression model on the training data.
4. **Model Evaluation:**
    - Evaluate the model's performance using MSE and R^2 metrics.
5. **Visualization of Results:**
    - Plot the actual vs. predicted values to visualize model performance.