**Linear Regression Case Study Report**

**Projects: Salary Prediction & Housing Price Prediction**

**Objective**

This case study focuses on applying **Linear Regression**, one of the foundational algorithms in supervised machine learning, to solve two real-world regression problems:

1. **Predict the salary** of an employee based on their years of experience.
2. **Predict the price of a house** based on various property attributes.

The study involves:

* Performing Exploratory Data Analysis (EDA) to understand patterns and relationships.
* Building regression models.
* Interpreting results using metrics like R² score and Mean Squared Error (MSE).
* Drawing business-level insights from the outcomes.

**Dataset 1: Salary Prediction Based on Experience**

**Dataset Information**

* **File:** data.csv
* **Records:** 30
* **Features:**
  + YearsExperience (numeric)
  + Salary (numeric)

**Exploratory Data Analysis (EDA)**

**1. Data Summary**

data.describe()

| **Statistic** | **YearsExperience** | **Salary** |
| --- | --- | --- |
| Count | 30 | 30 |
| Mean | 5.313 | 76003.0 |
| Std Dev | 2.837 | 27414.4 |
| Min | 1.1 | 37731.0 |
| Max | 10.5 | 122391.0 |

**2. Distribution Plots**

* **Years of Experience:**
  + Slightly right-skewed.
  + Most employees fall between 3 to 8 years of experience.
* **Salary:**
  + Right-skewed distribution.
  + Salaries cluster between ₹40,000 and ₹80,000, with a few high-earners above ₹100,000.

**3. Correlation Matrix**

data.corr()

| **Feature** | **Correlation with Salary** |
| --- | --- |
| YearsExperience | 0.978 |

**Inference:** Extremely strong positive correlation between experience and salary.

**4. Boxplot Analysis**

* A few salary values are outliers, but overall, no extreme anomalies.
* Data spread is moderate and interpretable.

**Model Development: Simple Linear Regression**

**Features:**

* **X:** Years of Experience (2D array)
* **y:** Salary

**Model Code:**

model = LinearRegression()

model.fit(X\_train, y\_train)

**Evaluation:**

* **Train-Test Split:** 80:20
* **R² Score on Test Data:** **0.71**
* **MSE on Test Data:** ~39.7 billion

**Plot:**

A regression line overlaid on the scatter plot of actual salaries shows the linear trend clearly.

**Insights:**

* As years of experience increase, salary increases linearly.
* The model explains about 71% of variance in salary for unseen data.
* The prediction is quite accurate given the small dataset size.
* Model generalizes well with a small performance drop from training to testing.

**Project 2: House Price Prediction Using Multiple Features**

**Dataset Information**

* **File:** housing.csv
* **Records:** 21,613
* **Features:** 21
  + Target: price
  + Predictors: sqft\_living, grade, bathrooms, sqft\_above, view, etc.

**Exploratory Data Analysis (EDA)**

**1. Data Cleaning**

housing.drop(['id', 'date'], axis=1, inplace=True)

* Removed uninformative columns like ID and date.

**2. Descriptive Statistics**

* Mean price: ₹540,000
* Range: ₹75,000 to ₹7,700,000
* Most houses fall within ₹200,000 to ₹500,000.

**3. Price Distribution**

* Strong **right skew**, indicating presence of **high-value outliers**.
* Not normally distributed, impacting model residuals.

**4. Correlation Matrix**

housing.corr()['price'].sort\_values(ascending=False)

| **Feature** | **Correlation** |
| --- | --- |
| sqft\_living | 0.70 |
| grade | 0.67 |
| sqft\_above | 0.61 |
| sqft\_living15 | 0.59 |
| bathrooms | 0.53 |
| view | 0.40 |
| waterfront | 0.27 |
| lat | 0.31 |

**Inference:** Price is most strongly influenced by square footage and grading.

**5. Boxplots**

* Price increases with bedroom count up to a point.
* Anomalies like 33-bedroom houses suggest data entry issues or outliers.

**6. Scatter Plots**

* **Price vs. sqft\_living:**  
  Shows a **positive, non-linear trend** with sharp increase post-3000 sqft.
* Most homes lie between 1000–2000 sqft, suitable for mid-range buyers.

**Model Development: Multiple Linear Regression**

**Features:**

* **X:** All numeric columns excluding price
* **y:** Price

**Model Code:**

house\_model = LinearRegression()

house\_model.fit(X\_train, y\_train)

**Evaluation:**

* **Train-Test Split:** 80:20
* **R² Score:** **0.71**
* **MSE:** ~39.7 billion

**Visual:**

* Actual vs Predicted Price plot.
* Most predictions are close to the ideal line, especially in the mid-price range.
* Higher spread in luxury house predictions.

**Insights:**

* The model captures **71% variance** in housing prices.
* Performs well in **predicting mid-range properties**, struggles slightly with luxury houses due to outliers and sparse representation.
* Square footage, grade, and bathrooms are the most influential features.

**Business Questions Answered**

**Salary Dataset**

1. **How many employees with >5 years experience earn >₹60,000?**  
   → **16 employees**
2. **Employees earning between ₹50,000–₹80,000?**  
   → **13 employees**

**Housing Dataset**

1. **Houses with a waterfront:** 163
2. **Houses with 2 floors:** 8,241
3. **Houses built before 1960 with waterfront:** 80
4. **Most expensive house with >4 bathrooms:** ₹7,700,000

**Conclusion**

| **Model** | **R² Score** | **MSE (Approx.)** | **Key Feature(s)** | **Suitability** |
| --- | --- | --- | --- | --- |
| Salary Prediction | 0.71 | ₹39.7 Billion | YearsExperience | Excellent for HR Planning |
| Housing Price | 0.71 | ₹39.7 Billion | sqft\_living, grade, bathrooms | Useful for real estate pricing |