Behavioral Data Analysis Report

1. Introduction

This report aims to analyze behavioral data to understand patterns and factors affecting the price of electric vehicles (EVs). The analysis involves data preprocessing, exploratory data analysis (EDA), feature engineering, and predictive modeling using machine learning techniques.

2. Data Preprocessing

2.1 Data Cleaning

- **Handling Missing Values:** The dataset is checked for missing values, and appropriate measures are taken to handle them, such as imputation or removal.
- **Categorical Encoding:** Categorical features are encoded using techniques like OneHotEncoder to convert them into numerical format.

2.2 Feature Scaling

- Numerical features are scaled using StandardScaler to ensure they have a mean of 0 and a standard deviation of 1. This step is crucial for models that rely on distance metrics.

3. Exploratory Data Analysis (EDA)

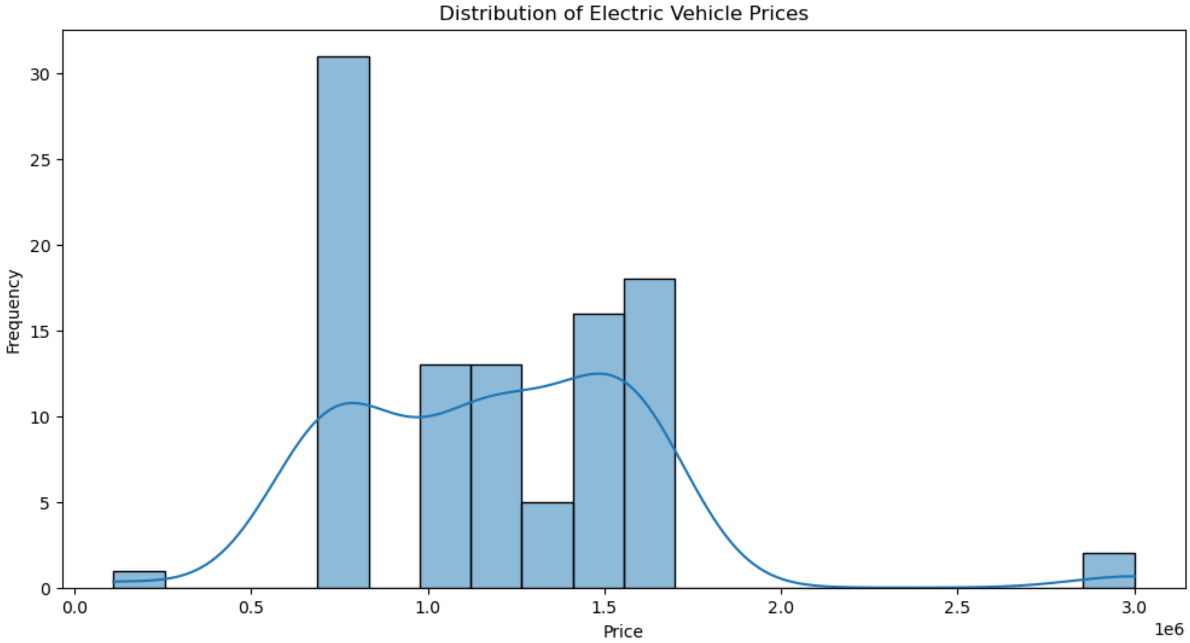
3.1 Descriptive Statistics

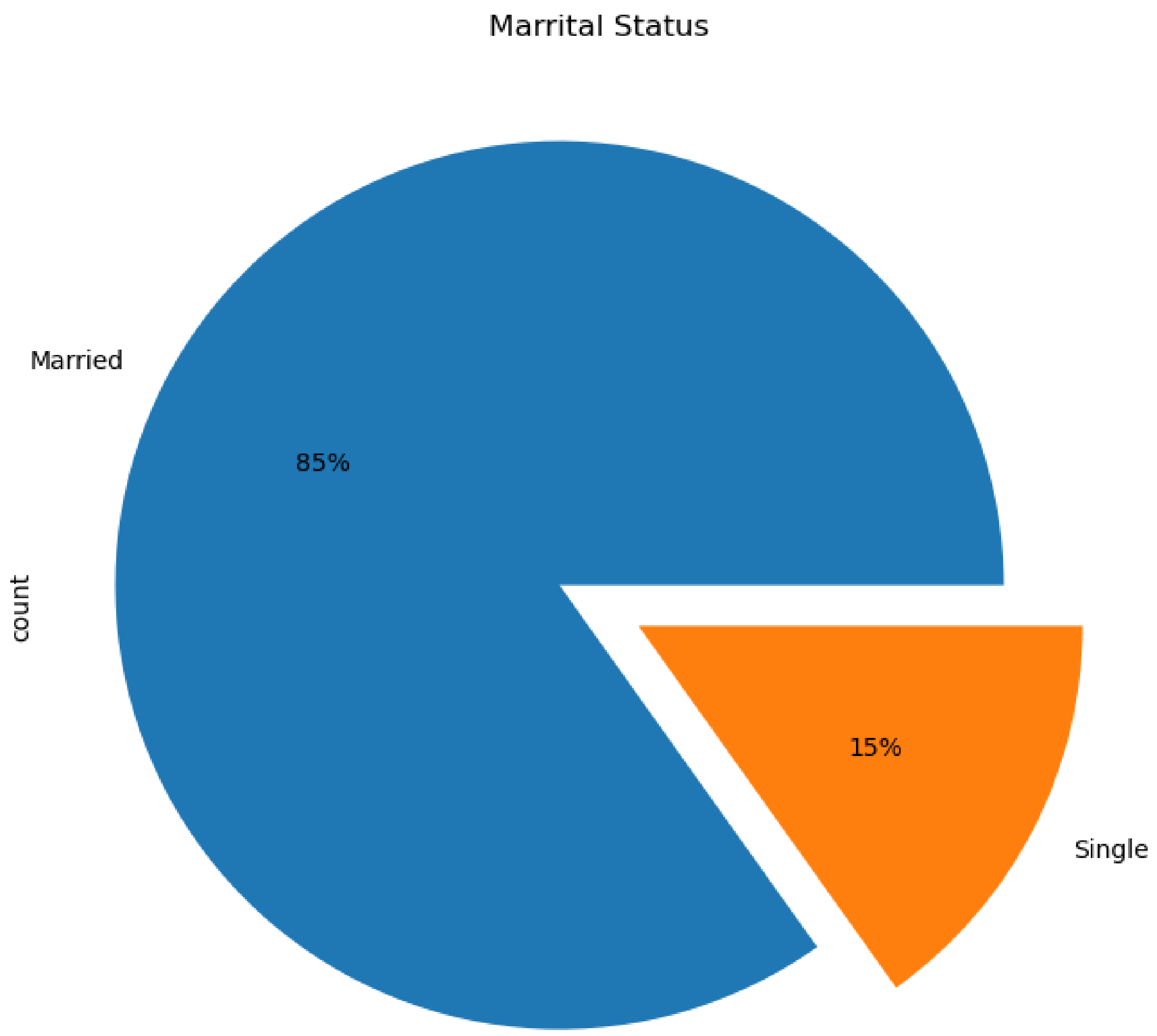
- The dataset's statistical summary is provided, including measures like mean, median, standard deviation, and interquartile ranges for numerical features.

3.2 Distribution Analysis

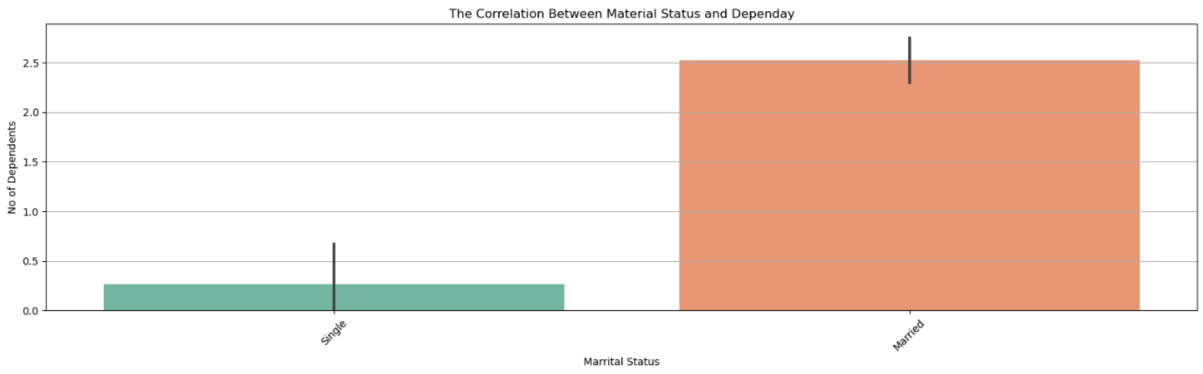
- Histograms and boxplots are used to visualize the distribution of key numerical features like age, total salary, and the target variable, Price.

- The distribution of categorical features, such as profession, marital status, education, and personal loan status, is also examined.

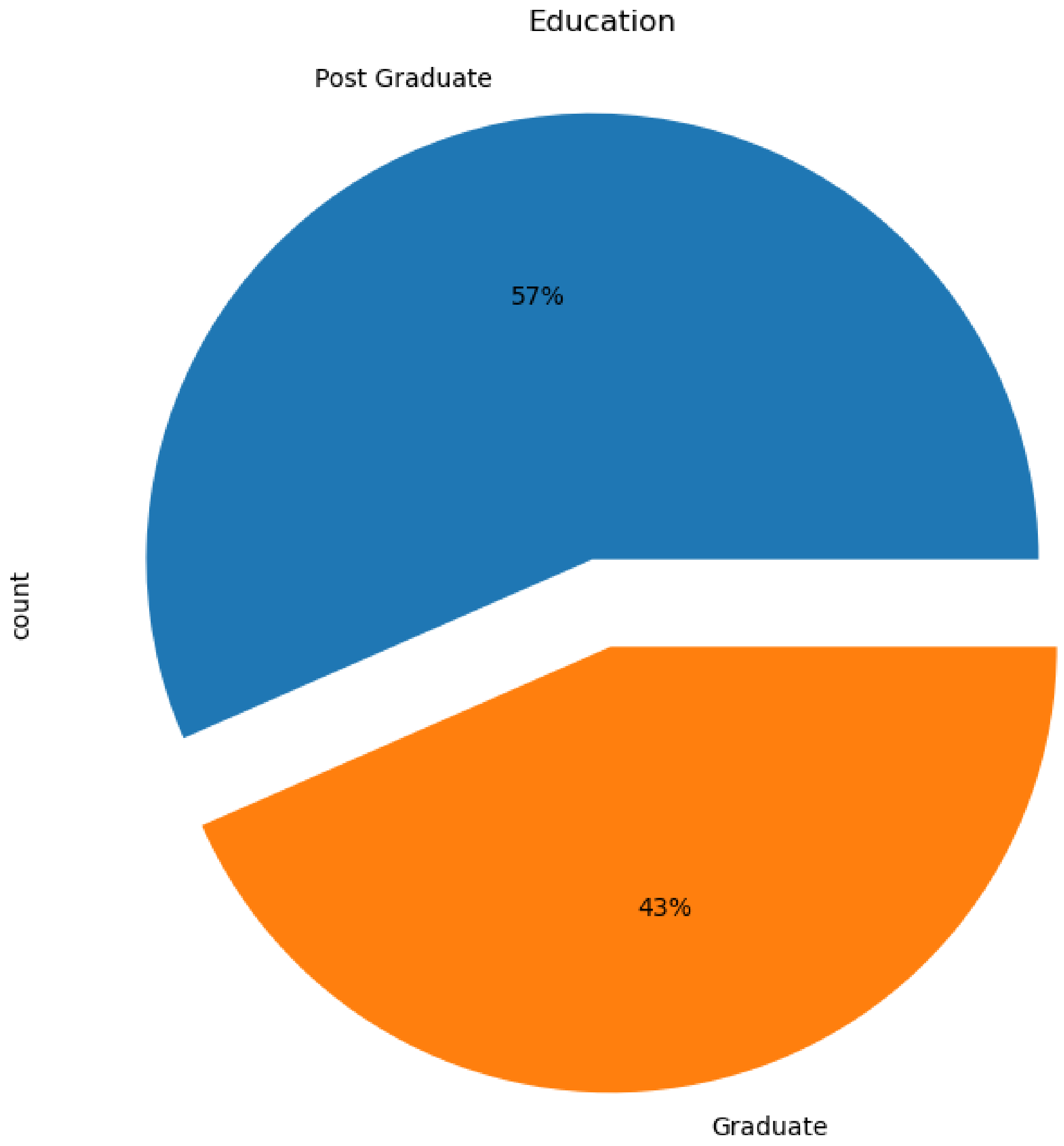

Distribution of Electric Vehicle Prices
Frequency
0
5
10
15
20
25
30
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e6
Price

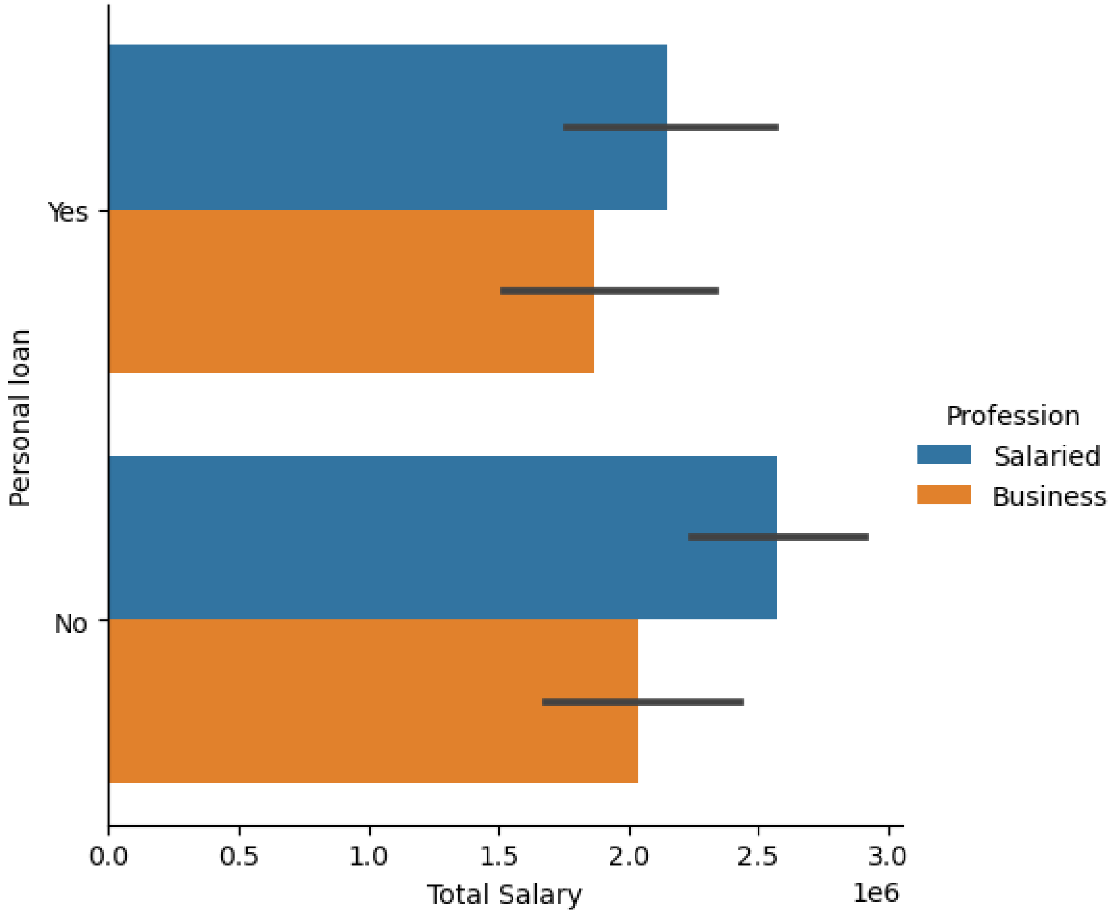


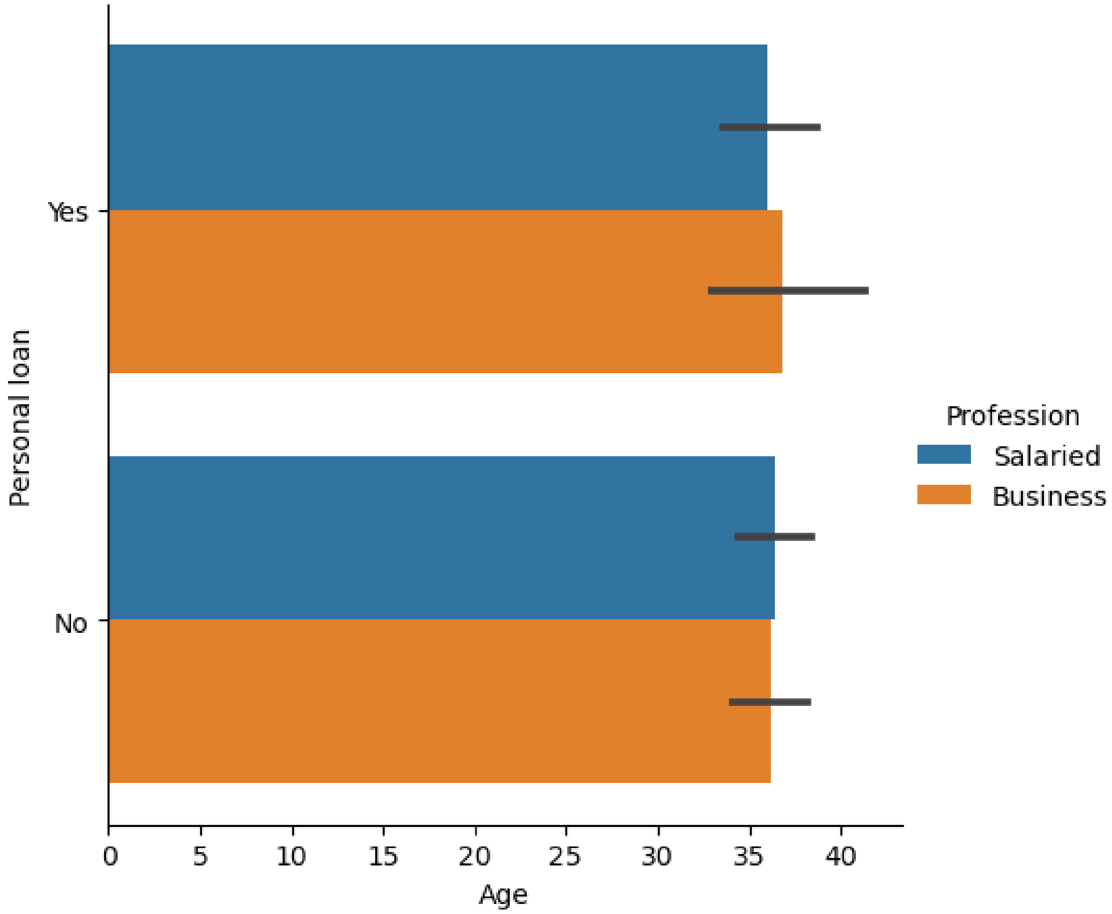
Marrital Status

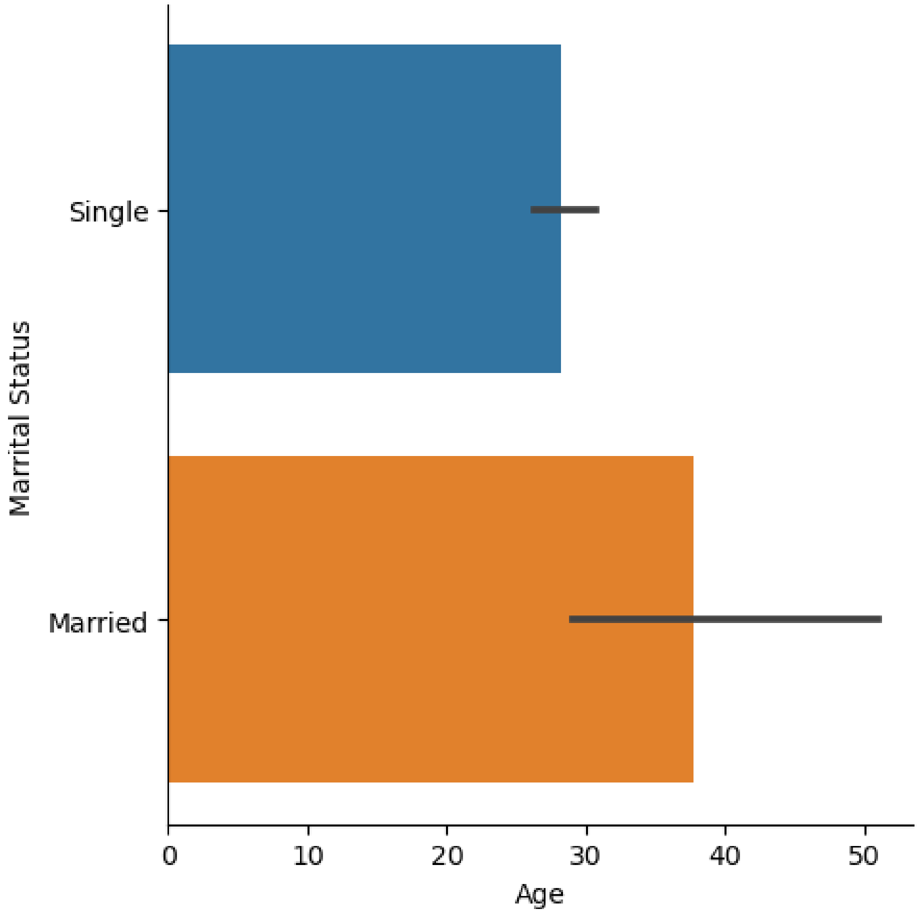


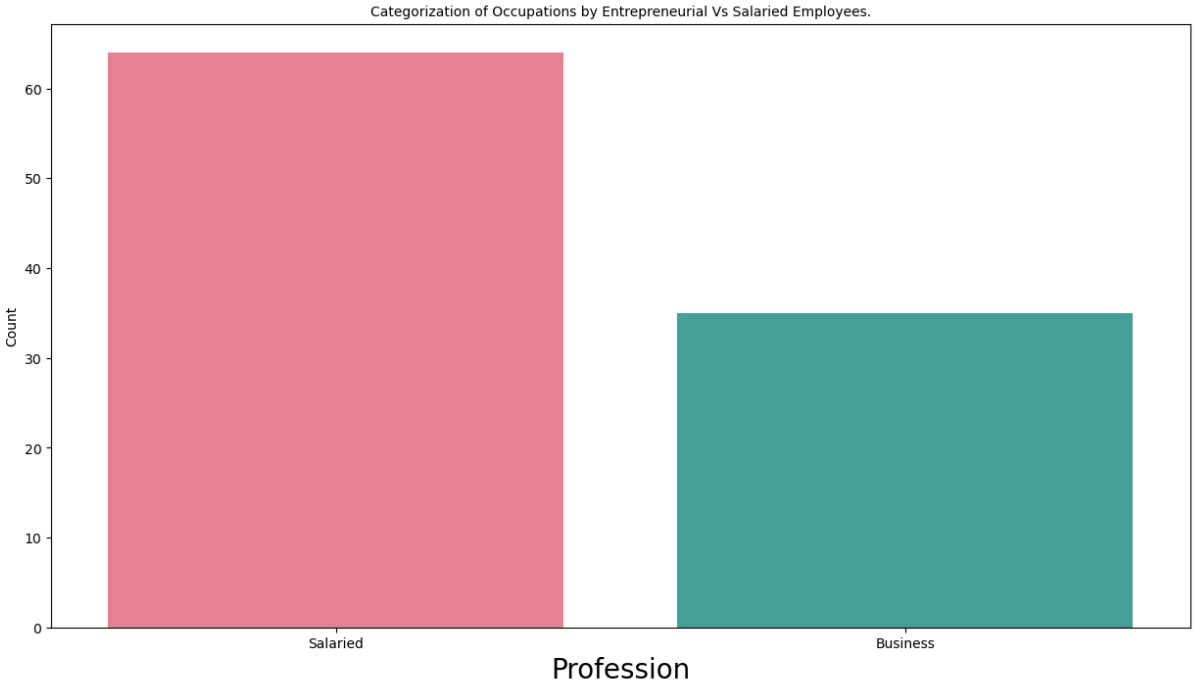
The Correlation Between Material Status and Dependay

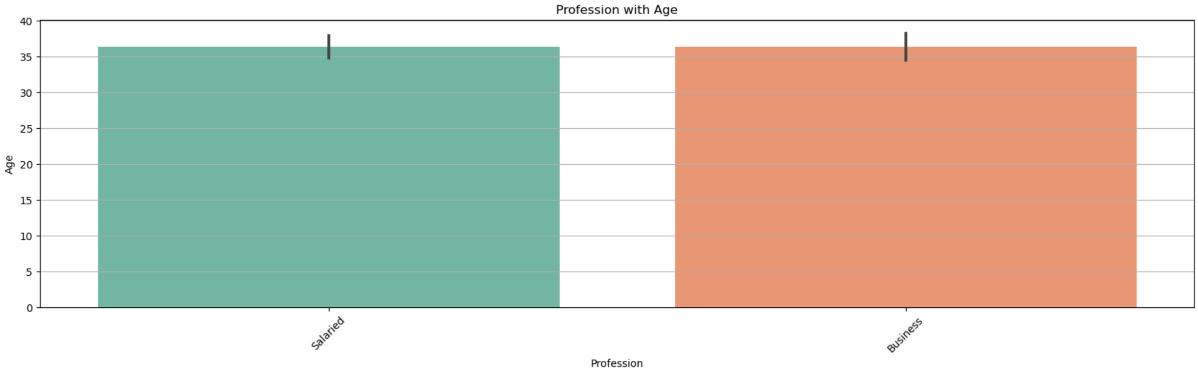

Education
Post Graduate
57%
count
43%
Graduate

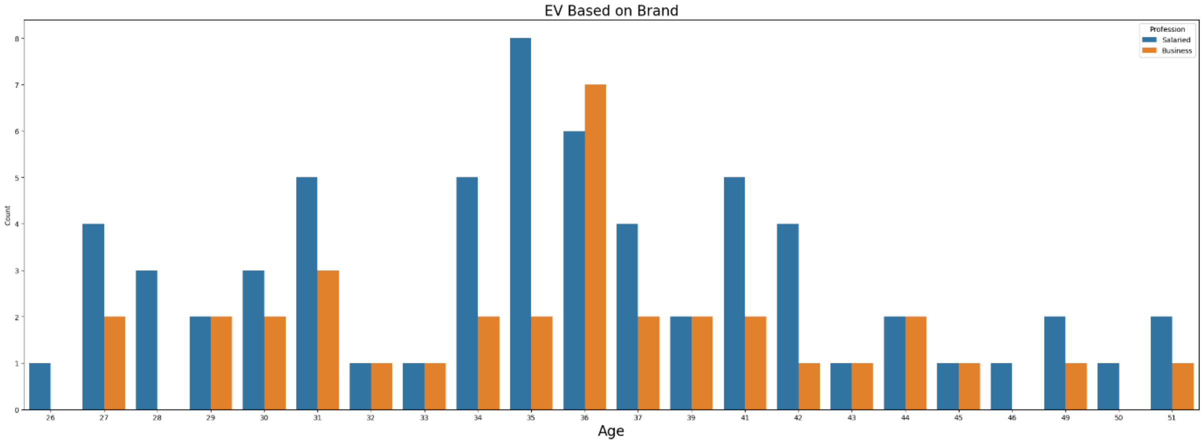

Personal loan
Yes
No
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Total Salary
1e6
Profession
Salaried
Business

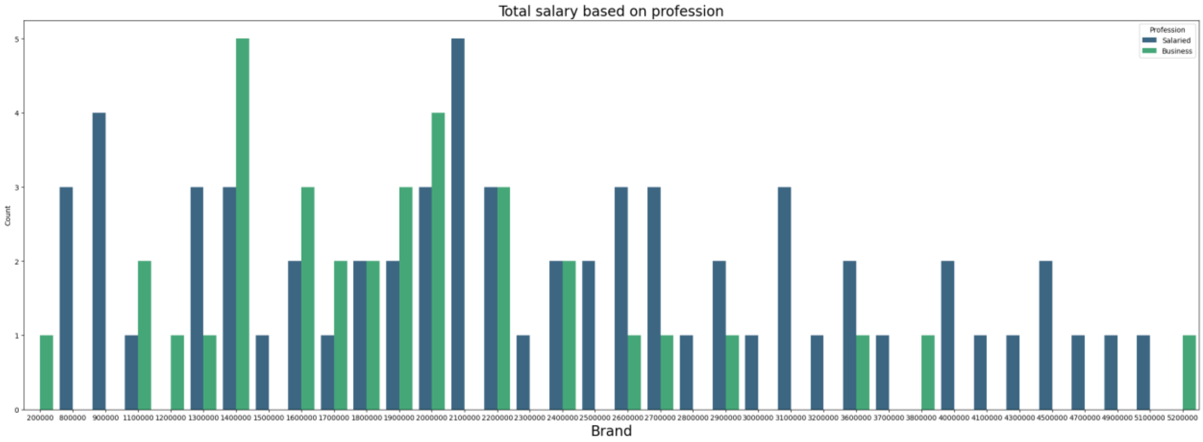

Personal loan
Yes
No
Age
0
5
10
15
20
25
30
35
40
Profession
Salaried
Business

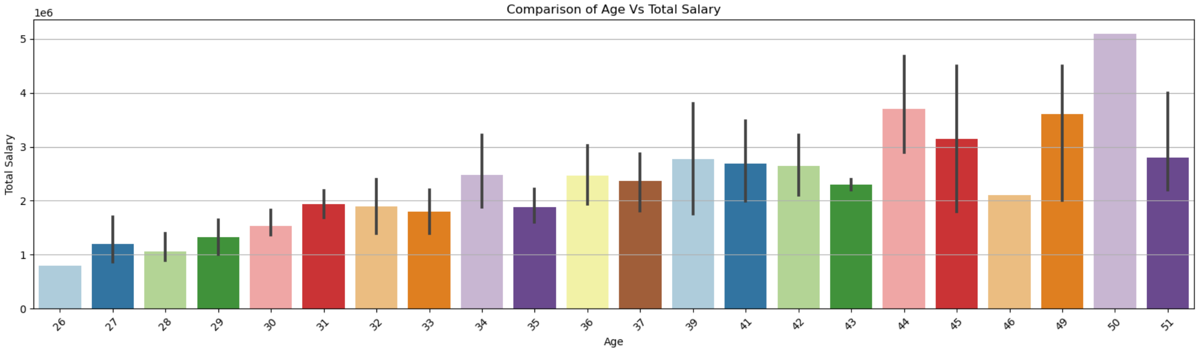

Marrital Status
Single
Married
0
10
20
30
40
50
Age

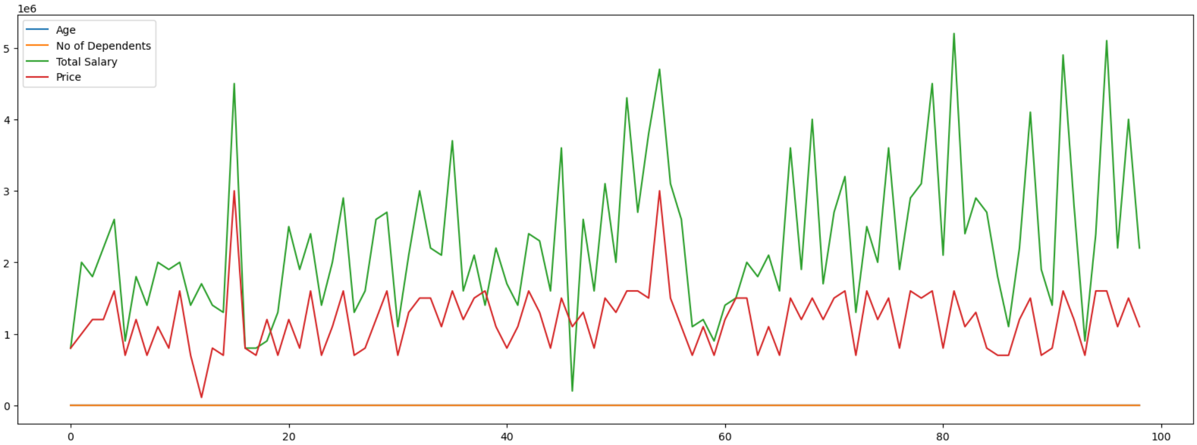

Categorization of Occupations by Entrepreneurial Vs Salaried Employees.
Count
0
10
20
30
40
50
60
Salaried
Business
Profession


Profession with Age
Age
0
5
10
15
20
25
30
35
40
Salaried
Business
Profession


EV Based on Brand
Count
0
1
2
3
4
5
6
7
8
Profession
Salaried
Business
26
27
28
29
30
31
32
33
34
35
36
37
39
41
42
43
44
45
46
49
50
51
Age

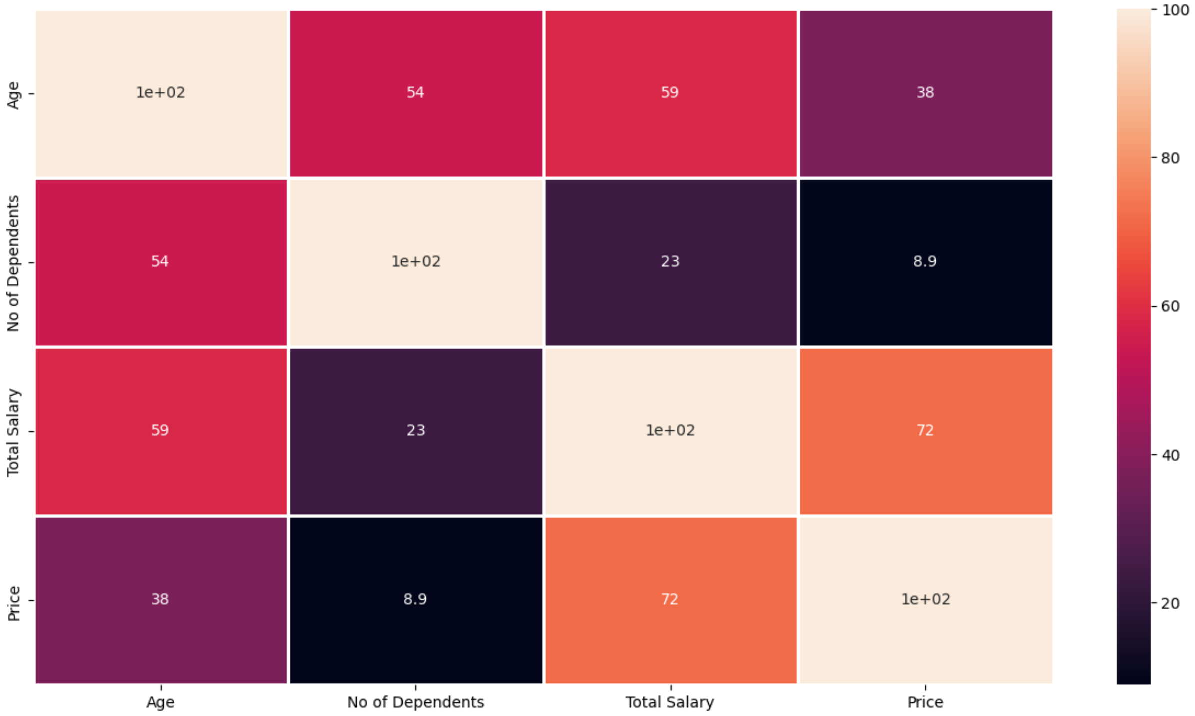

Total salary based on profession
Count
0
1
2
3
4
5
Profession
Salaried
Business
200000
800000
900000
1100000
1200000
1300000
1400000
1500000
1600000
1700000
1800000
1900000
2000000
2100000
2200000
2300000
2400000
2500000
2600000
2700000
2800000
2900000
3000000
3100000
3200000
3600000
3700000
3800000
4000000
4100000
4300000
4500000
4700000
4900000
5100000
5200000
Brand





3.3 Correlation Analysis

- The Pearson correlation coefficient is calculated to identify relationships between numerical features.
- Heatmaps are used to visualize the correlation matrix and highlight significant correlations.



3.4 Feature Importance

- Feature importance is assessed using models like Random Forest to understand the impact of each feature on the target variable.

4. Predictive Modeling

4.1 Model Selection

- **Random Forest Regressor** is chosen for its robustness and ability to handle both numerical and categorical features.

4.2 Model Training

- The data is split into training and testing sets.
- A pipeline is created to streamline preprocessing and model training.

4.3 Hyperparameter Tuning

- GridSearchCV is used for hyperparameter tuning to find the best combination of parameters for the Random Forest model.

4.4 Model Evaluation

- The model's performance is evaluated using metrics like Mean Squared Error (MSE) and R-squared (R^2) score.

Results:

- **Random Forest Mean Squared Error:** 67376721049.65011
- **Random Forest R^2 Score:** 0.3493664276321341

5. Inferences and Recommendations

5.1 Key Findings

- **Feature Impact:** Total Salary and Age are significant predictors of EV Price.
- **Model Performance:** The Random Forest model achieved a reasonable R^2 score, indicating a moderate fit to the data.

5.2 Recommendations

- **Data Quality:** Improve data quality by addressing missing values and potential outliers more rigorously.
- **Feature Engineering:** Explore additional features and interaction terms to enhance model performance.
- **Advanced Models:** Consider using more advanced models or ensemble techniques to improve prediction accuracy.

6. Conclusion

This analysis provided valuable insights into the factors influencing EV prices. While the current model offers a moderate level of predictive accuracy, further refinement and exploration of additional features could enhance its performance.