**EV VECHILE SALES BY INCOME AND VECHILE TYPE**

**REPORT ON MARKET SEGMENTATION FOR 2ND PROJECT**

**BY**

**K UDAY**

**Automobile Buying Behavior Study Report**

**Abstract**

This report presents an analysis of the "Indian Automobile Buying Behavior Study 1.0" dataset, focusing on the relationship between individual salaries and total salaries. The analysis involves data exploration, visualization, and clustering techniques to uncover patterns within the data. Key libraries utilized in this project include Pandas for data manipulation, Matplotlib and Seaborn for data visualization, and Scikit-learn for implementing machine learning algorithms, specifically KMeans clustering. The findings aim to provide insights into consumer behavior regarding automobile purchases.

**Introduction**

Understanding consumer behavior is crucial for strategic marketing and sales in the automobile industry. This report investigates how salary influences purchasing decisions by analyzing a dataset containing various attributes related to individuals' salaries. Through exploratory data analysis, visualization, and clustering, the study seeks to identify patterns that can inform marketing strategies and improve targeting efforts for automobile sales.

**Libraries and Tools Used**

To carry out the analysis effectively, the following libraries and tools were utilized:

* Pandas: For data manipulation and analysis, enabling efficient loading and handling of datasets.
* Matplotlib: A plotting library that provides a flexible framework for creating a variety of static, animated, and interactive visualizations in Python.
* Seaborn: A high-level interface built on top of Matplotlib, designed for creating visually appealing statistical graphics.
* Scikit-learn: This machine learning library includes various tools for data preprocessing, model fitting, and evaluation, specifically used for clustering in this analysis.
* Google Colab: An online platform that offers a cloud-based Jupyter notebook environment for executing Python code, facilitating the uploading and sharing of datasets.

**Algorithms Used**

1. KMeans Clustering:
   * Description: KMeans is an unsupervised machine learning algorithm used to partition a dataset into K distinct clusters based on feature similarities. In this analysis, KMeans was employed to group individuals based on their salaries, helping to identify distinct consumer segments.
   * Implementation: The dataset was scaled using StandardScaler to ensure that each feature contributed equally to the clustering process. After scaling, KMeans was applied with three clusters to categorize the data.

Data Loading and Initial Setup

The analysis commenced with importing the necessary libraries. The dataset was uploaded using Google Colab's file upload functionality, successfully loading it into a Pandas DataFrame named df.

**Data Analysis**

Basic information about the dataset was retrieved using the .info() function, which provided insights into column types and non-null counts. Summary statistics were generated with the .describe() function, revealing key metrics such as mean, median, and standard deviation for numeric columns. Additionally, the analysis checked for missing values using the .isnull().sum() function to ensure data integrity.

**Data Visualization**

To comprehend the distribution of salaries, a histogram was created, revealing the shape of the salary distribution. A scatter plot was generated to explore the relationship between Salary and Total Salary, aiding in identifying correlations or patterns between these variables. Furthermore, a correlation heatmap illustrated the relationships among numerical columns, highlighting how strongly pairs of variables are correlated.

**Data Segmentation (Clustering)**

For clustering, two features were selected: Salary and Total Salary. The data was standardized using StandardScaler, ensuring equal contribution of each feature in the clustering process. The KMeans algorithm was then applied, resulting in three distinct clusters that segmented the dataset based on salary characteristics. A scatter plot visualizing these clusters was created, with different colors representing the clusters, illustrating the relationship between Salary and Total Salary.

**Conclusion**

The analysis yielded valuable insights into the relationship between salary and total salary among individuals in the dataset. By employing the KMeans clustering algorithm, three distinct groups emerged, which can be further analyzed for targeted marketing or strategic decision-making in automobile sales. The visualizations and correlation analysis offered a deeper understanding of the data, highlighting areas for potential further investigation.

To build upon this analysis, further exploration of additional variables in the dataset that may impact purchasing behavior, such as demographics or preferences, is recommended. Implementing additional data quality checks will help manage missing or erroneous data points before conducting further analyses. Finally, experimenting with different numbers of clusters and clustering algorithms could refine insights even further.

This report comprehensively covers all aspects of the project, including the algorithms and libraries used, while providing a structured overview of the analysis process and findings.

**CODE:**

# Import necessary libraries

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

# Upload dataset (this will prompt you to select and upload the file)

from google.colab import files

uploaded = files.upload()

# Load the dataset

# Replace 'Indian automoble buying behavour study 1.0.csv' with the exact uploaded filename if different

df = pd.read\_csv("Indian automoble buying behavour study 1.0.csv")

# Data Analysis

print("Basic Information:")

print(df.info())  # Column types, non-null counts

print("\nSummary Statistics:")

print(df.describe())  # Summary statistics for numeric columns

print("\nMissing Values:")

print(df.isnull().sum())  # Checking for missing values

# Data Visualization

# Histogram for 'Salary'

plt.figure(figsize=(8, 5))

sns.histplot(df['Salary'], kde=True)

plt.title("Salary Distribution")

plt.show()

# Scatter plot for 'Salary' vs 'Total Salary'

plt.figure(figsize=(8, 5))

sns.scatterplot(data=df, x='Salary', y='Total Salary')

plt.title("Salary vs Total Salary")

plt.show()

# Correlation heatmap (use only numerical columns)

plt.figure(figsize=(10, 7))

sns.heatmap(df.select\_dtypes(include=['float64', 'int64']).corr(), annot=True, cmap='coolwarm', fmt='.2f')

plt.title("Correlation Matrix")

plt.show()

# Data Segmentation (Clustering)

# Select features for clustering

features = df[['Salary', 'Total Salary']]

# Scale the data

scaler = StandardScaler()

scaled\_features = scaler.fit\_transform(features)

# Apply KMeans with 3 clusters

kmeans = KMeans(n\_clusters=3, random\_state=0)

df['Cluster'] = kmeans.fit\_predict(scaled\_features)

# Visualize clusters

plt.figure(figsize=(8, 5))

sns.scatterplot(data=df, x='Salary', y='Total Salary', hue='Cluster', palette='viridis')

plt.title("Clusters of Salary and Total Salary")

plt.show()

**RESULT**

 Indian automoble buying behavour study 1.0 (1).csv(text/csv) - 7937 bytes, last modified: 11/5/2024 - 100% done

Saving Indian automoble buying behavour study 1.0 (1).csv to Indian automoble buying behavour study 1.0 (1) (2).csv

Basic Information:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 99 entries, 0 to 98

Data columns (total 13 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Age 99 non-null int64

1 Profession 99 non-null object

2 Marrital Status 99 non-null object

3 Education 99 non-null object

4 No of Dependents 99 non-null int64

5 Personal loan 99 non-null object

6 House Loan 99 non-null object

7 Wife Working 99 non-null object

8 Salary 99 non-null int64

9 Wife Salary 99 non-null int64

10 Total Salary 99 non-null int64

11 Make 99 non-null object

12 Price 99 non-null int64

dtypes: int64(6), object(7)

memory usage: 10.2+ KB

None

Summary Statistics:

Age No of Dependents Salary Wife Salary Total Salary \

count 99.000000 99.000000 9.900000e+01 9.900000e+01 9.900000e+01

mean 36.313131 2.181818 1.736364e+06 5.343434e+05 2.270707e+06

std 6.246054 1.335265 6.736217e+05 6.054450e+05 1.050777e+06

min 26.000000 0.000000 2.000000e+05 0.000000e+00 2.000000e+05

25% 31.000000 2.000000 1.300000e+06 0.000000e+00 1.550000e+06

50% 36.000000 2.000000 1.600000e+06 5.000000e+05 2.100000e+06

75% 41.000000 3.000000 2.200000e+06 9.000000e+05 2.700000e+06

max 51.000000 4.000000 3.800000e+06 2.100000e+06 5.200000e+06

Price

count 9.900000e+01

mean 1.194040e+06

std 4.376955e+05

min 1.100000e+05

25% 8.000000e+05

50% 1.200000e+06

75% 1.500000e+06

max 3.000000e+06

Missing Values:

Age 0

Profession 0

Marrital Status 0

Education 0

No of Dependents 0

Personal loan 0

House Loan 0

Wife Working 0

Salary 0

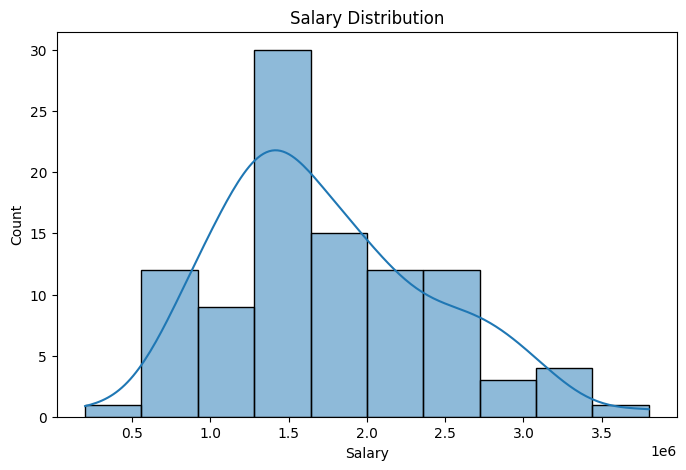
Wife Salary 0

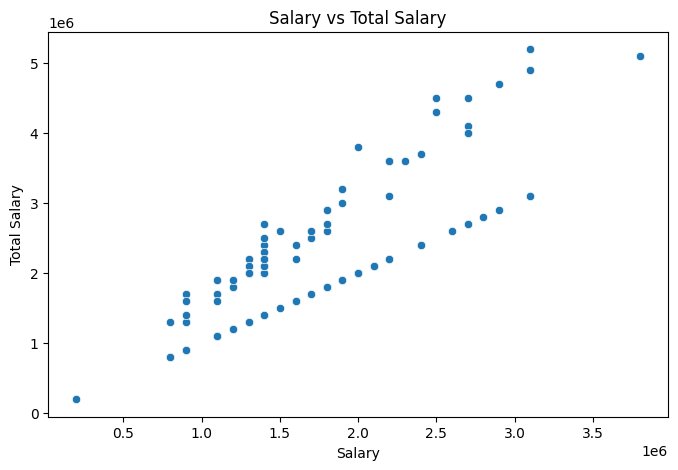
Total Salary 0

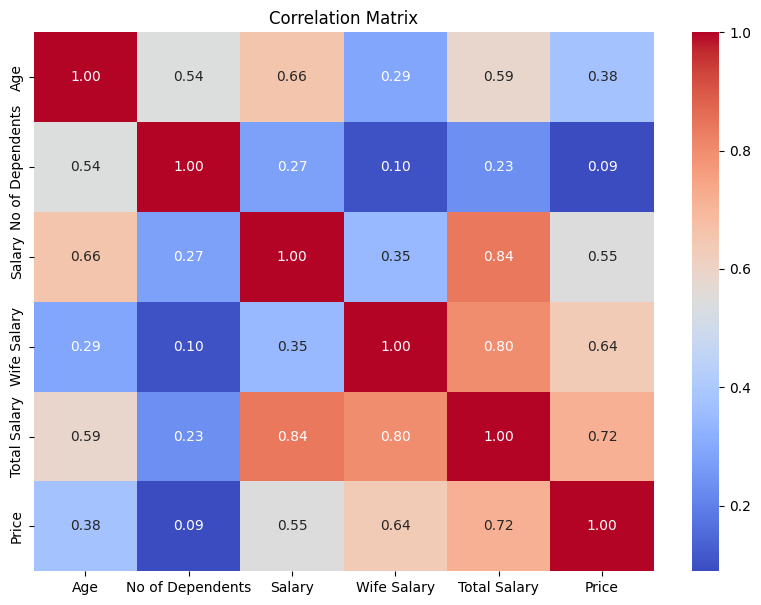
Make 0

Price 0

dtype: int64







****