Data Cleaning and Exploratory Data Analysis Report

1. Introduction

Before making any data-driven decisions, it's important to understand and prepare the dataset properly. This report walks through the process of cleaning and exploring a dataset related to loan classification. The goal is to make sure the data is accurate, consistent, and useful by handling missing values, removing duplicates, identifying outliers, and looking at patterns through exploratory data analysis (EDA).

2. Data Cleaning

2.1 Loading the Dataset

I started by loading the dataset and checking its structure using Pandas:

```
import pandas as pd
df =
pd.read_csv("/kaggle/input/simple-loan-classification-dataset/loan.csv")
print(df.info())
print(df.head())
```

This step helped me understand the column names, data types, and any missing values that needed attention.

2.2 Handling Missing Values

Missing data can mess up the analysis, so I fixed it:

- **Checking for missing values:**

```
print(df.isnull().sum())
```

-
- **Filling numerical missing values with the mean:**

```
df.fillna(df.mean(), inplace=True)
```

-
- **Filling categorical missing values with the most common value:**

```
df.fillna(df.mode().iloc[0], inplace=True)
```

-

2.3 Removing Duplicates

To avoid unnecessary repetition, I removed duplicate entries:

```
df.drop_duplicates(inplace=True)
```

2.4 Handling Outliers

Outliers can really throw off the results, so I used the **Interquartile Range (IQR) method** to find and remove them:

```
import seaborn as sns
Q1 = df.select_dtypes(include=['number']).quantile(0.25)
Q3 = df.select_dtypes(include=['number']).quantile(0.75)
IQR = Q3 - Q1
df = df[~((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))).any(axis=1)]
```

To see the outliers, I also used a box plot:

```
sns.boxplot(data=df)
```

2.5 Standardizing Categorical Values

To keep things consistent, I standardized categorical values:

```
df['marital_status'] = df['marital_status'].str.lower().str.strip()
```

3. Exploratory Data Analysis (EDA)

3.1 Understanding Individual Variables (Univariate Analysis)

To see how each variable behaves on its own, I used summary statistics:

- print(df.describe())

And I made histograms, like this one for age:

```
import matplotlib.pyplot as plt
df['age'].hist()
plt.show()
```

3.2 Understanding Relationships Between Two Variables (Bivariate Analysis)

To see how different variables are related, I used some visual tools:

- **Correlation Matrix:**

```
 import seaborn as sns
sns.heatmap(df.corr(), annot=True, cmap="coolwarm")
plt.show()
```

-
    - **Income and credit score had a strong positive correlation**, meaning that people with higher credit scores usually had higher incomes.
    - **Age and loan approval status had weak correlation**, which suggests that age alone isn't a big factor in getting a loan approved.
- **Scatter Plot (Age vs. Income):**

```
 sns.scatterplot(x='age', y='income', data=df)
plt.show()
```

-

    - The scatter plot showed that age doesn't have a clear pattern with income. It was all over the place.
- **Box Plot (Income by Marital Status):**

```
 sns.boxplot(x='marital_status', y='income', data=df)
plt.xticks(rotation=45)
plt.show()
```

-
    - The box plot showed that **married people tended to have higher incomes than single people**, which might be because of financial stability.

3.3 Understanding Multiple Variables Together (Multivariate Analysis)

To get a bigger picture, I used:

- **Pair Plot:**

```
 sns.pairplot(df)
plt.show()
```

    - This plot confirmed that **income and credit score are strongly connected**, while other variables were more scattered.
- **Average Income by Age Group:**

```
 df.groupby('age')['income'].mean()
```

    - Incomes generally increased with age but seemed to level off after a certain point.

4. Conclusion

After cleaning the dataset, I removed missing values, duplicates, and outliers. Here's what I found:

- **People with higher credit scores tend to have higher incomes.**
- **Age doesn't really predict loan approval,** so lenders must be looking at other factors.
- **Married people generally earn more than single people.**
- **Loan approvals seem to depend on a mix of credit score, income, and other factors, not just one thing.**