# EDA Report - Credit Card Customer Dataset

CREDIT CARD CUSTOMER DATASET - EXPLORATORY DATA ANALYSIS (EDA)

1. Import Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style='whitegrid')
```

2. Load Dataset

```
df = pd.read_csv('CC GENERAL.csv')
df.head()
```

3. Basic Exploration

```
df.info()
df.describe()
df.isnull().sum()
```

4. Data Cleaning

```
df_clean = df.dropna()
```

# EDA Report - Credit Card Customer Dataset

df_clean.shape

5. Data Visualization

5.1 Histograms

plt.figure(figsize=(20, 15))

df_clean.hist(bins=30, figsize=(20, 15))

plt.suptitle('Histograms of Features', fontsize=20)

plt.show()

Observation:

- Many features are right-skewed, indicating presence of few large values.

5.2 Boxplots

plt.figure(figsize=(20, 10))

sns.boxplot(data=df_clean)

plt.xticks(rotation=90)

plt.title('Boxplot for All Features')

plt.show()

Observation:

- Outliers detected especially in BALANCE, PURCHASES, and CASH_ADVANCE.

5.3 Correlation Matrix (Heatmap)

plt.figure(figsize=(18, 14))

# EDA Report - Credit Card Customer Dataset

```
corr_matrix = df_clean.corr()

sns.heatmap(corr_matrix, annot=False, cmap='coolwarm')

plt.title('Correlation Heatmap', fontsize=18)

plt.show()
```

Observation:

- PURCHASES and CREDIT_LIMIT are moderately correlated.

- BALANCE also shows correlation with CREDIT_LIMIT.

5.4 Pairplot

```
sample_columns = df_clean.columns[1:6]

sns.pairplot(df_clean[sample_columns])

plt.suptitle('Pairplot of Selected Features', y=1.02)

plt.show()
```

Observation:

- Some linear relationships between BALANCE and CREDIT_LIMIT.

- Wide spread in variables like PURCHASES_FREQUENCY.

6. Summary of Findings

- Dataset mostly contains continuous numerical variables.

- Significant right skew in several features.

- Presence of notable outliers.

- Some moderate positive correlations among financial features.

# EDA Report - Credit Card Customer Dataset

Recommendations:

- Outliers may need treatment before modeling.

- Standardization/Normalization recommended.

- Dimensionality reduction (PCA) could help.

- Further unsupervised learning (e.g., clustering) can be explored.