Loan-Defaulter-Dataset-Analysis(EDA)

28 March 2024 18:13 PM

I have taken this data from the Kaggle . Link is mentioned below:

https://www.kaggle.com/datasets/gauravduttakiit/loan-defaulter

Overview of the Dataset:

The objective of this case study is to illustrate the application of Exploratory Data Analysis (EDA) in a practical business context. Beyond utilizing EDA techniques, we aim to cultivate a foundational comprehension of risk analytics within banking and financial services. Through this exploration, we seek to understand how data can mitigate the risk of financial losses when extending loans to customers.

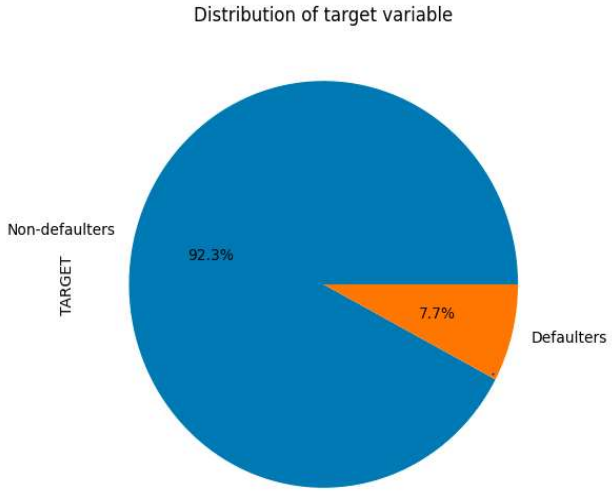
This dataset comprises 300,000 rows and 122 columns. The target variable indicates loan default status. It contains various customer details such as income, gender, occupation, and age. I will not perform univariate analysis; instead, I will address specific questions below.

Below are some questions derived from the dataset. I will address each question one by one, providing observations and graphical representations accordingly.

- What is the distribution of the target variable (loan defaulters vs. non-defaulters)?
- How does the distribution of income differ between loan defaulters and non-defaulters?
- Are there any patterns in the types of contracts (cash or revolving) for defaulters?
- What is the average age of defaulters compared to non-defaulters?
- Do clients with higher education levels tend to default less often?
- What is the relationship between family status and loan default rates?
- Are there specific housing types associated with higher rates of loan default?
- How do factors such as occupation type, organization type, and income type relate to loan default rates?

```
# 1 Question
# What is the distribution of the target variable (loan defaulters vs. non-defaulters)?
    plt.figure(figsize=(6, 6))
    data['TARGET'].value_counts().plot(kind='pie',autopct='%1.1f%%',labels=['Non-defaulters', 'Defaulters'])
    plt.title('Distribution of target variable')
    plt.show()
    print()
    print('We can observe from the plot below that 90% of customers are paying their loan amounts, while 8% are defaulters')
    print()
```

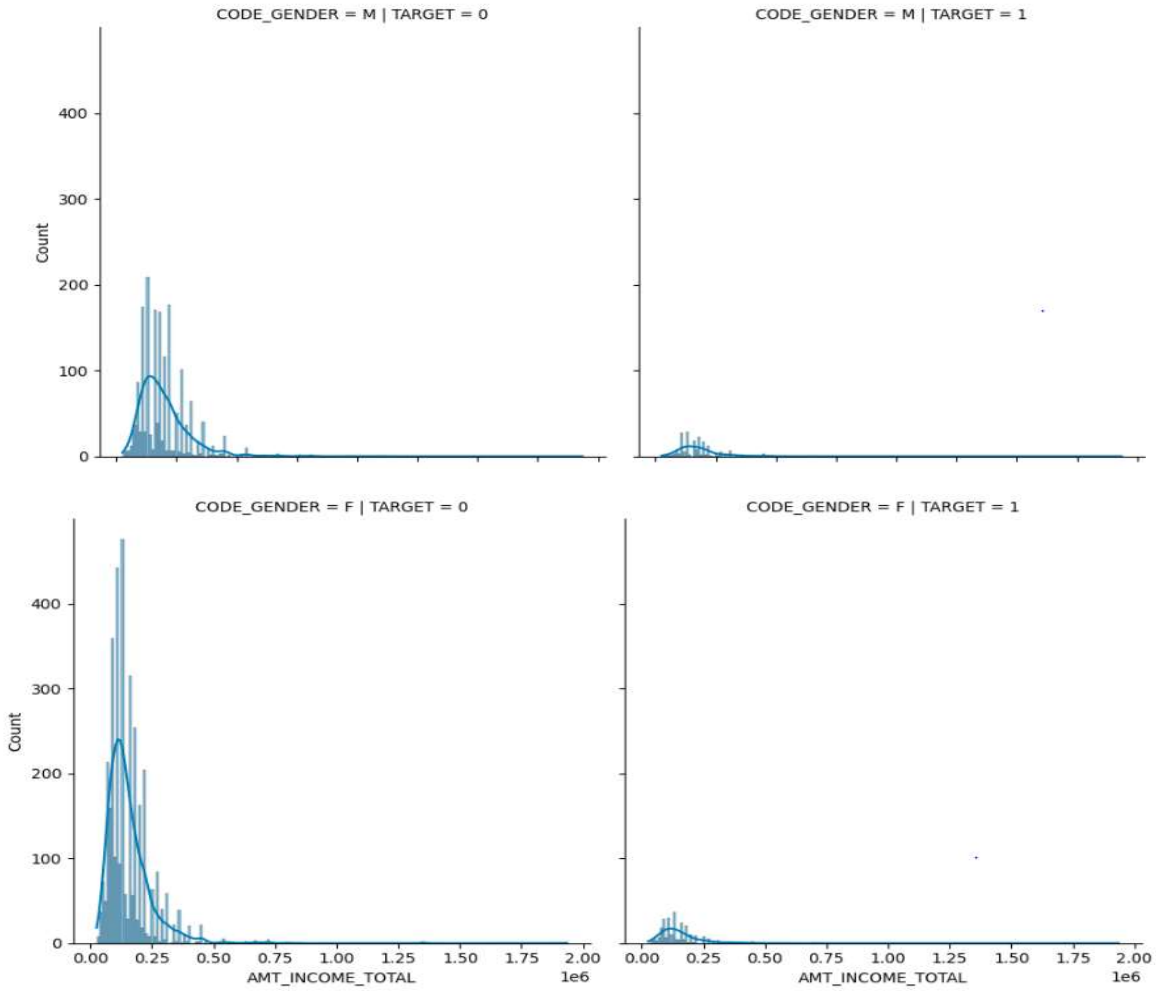
We can observe from the plot below that 90% of customers are paying their loan amounts, while 8% are defaulters




```
# 2 Question:
# 2-How does the distribution of income differ between loan defaulters and non-defaulters
    sn.displot(data, x='AMT_INCOME_TOTAL', col='TARGET',row='CODE_GENDER', kde=True,)
    plt.show()
    print()
    print("""The graph shows that most females are not paying back their loans on time, especially among the defaulters""")
    print()
```

The graph shows that most females are not paying back their loans on time, especially among the defaulters

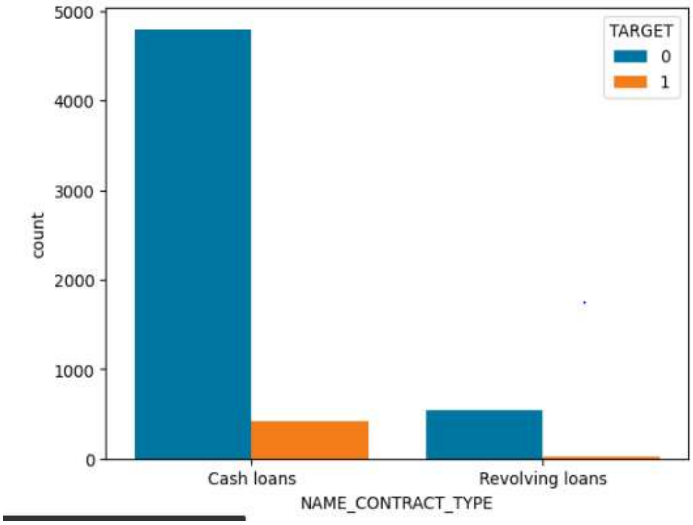
CODE_GENDER	F	M
TARGET		
0	3565	1776
1	256	192




```
# 3 Question
# 3-Are there any patterns in the types of contracts (cash or revolving) for defaulters?
    sn.countplot(data=data,x='NAME_CONTRACT_TYPE',hue='TARGET')
    print('I see from the graph and table below that 8% of defaults are associated with cash loans, while 4% are associated with
    revolving loans.')
    print()
    print()
    print(pd.crosstab(index=data['TARGET'], columns=data['NAME_CONTRACT_TYPE'], normalize='columns') * 100)
    print()
```

I see from the graph and table below that 8% of defaults are associated with cash loans, while 4% are associated with revolving loans.

```
NAME_CONTRACT_TYPE  Cash loans  Revolving loans
TARGET
0                    91.947853        95.113438
1                     8.052147         4.886562
```

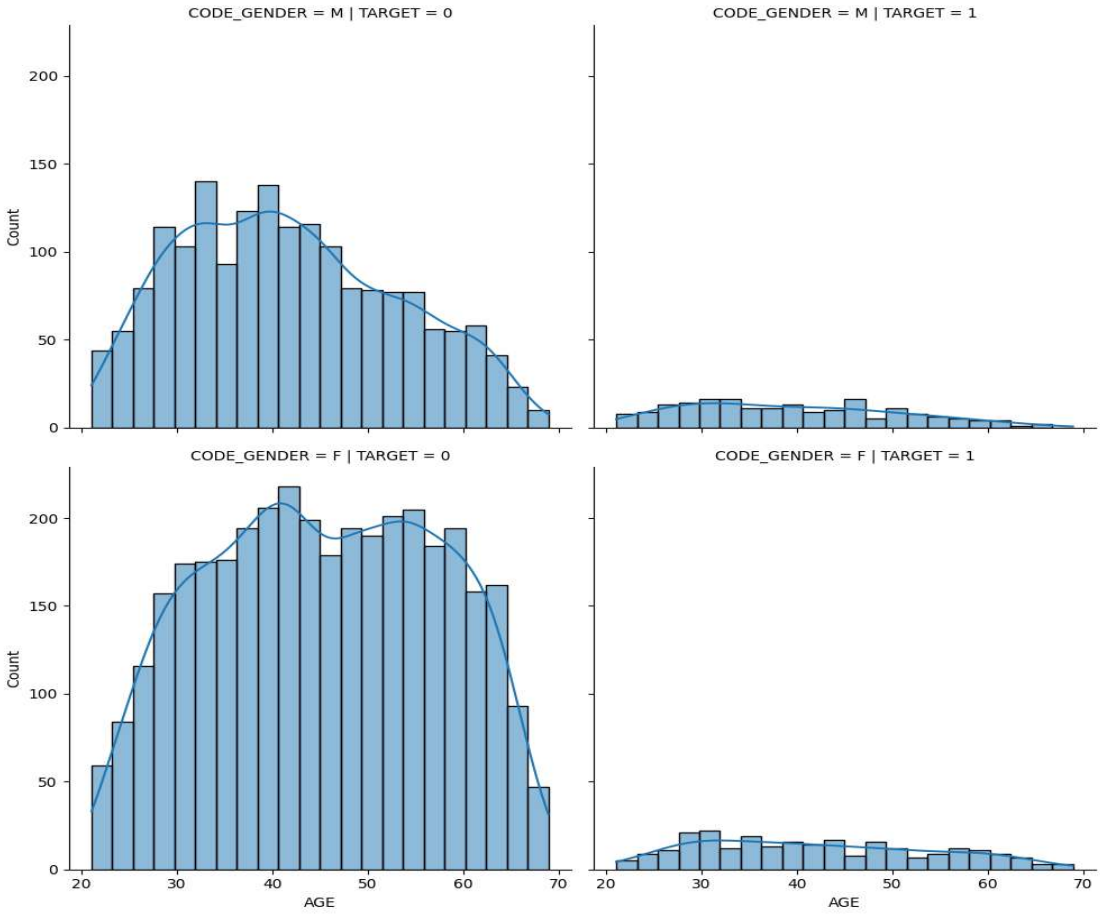



```
# 4 Question
# 4-What is the average age of defaulters compared to non-defaulters?
# Step 1: Calculate age
    data['AGE'] = round(-data['DAYS_BIRTH'] / 365, 2)
    # Step 2: Separate dataset into defaulters and non-defaulters
    defaulters = data[data['TARGET'] == 1]
    non_defaulters = data[data['TARGET'] == 0]
    # Step 3: Calculate average age for each group
    avg_age_defaulters = defaulters['AGE'].mean()
    avg_age_non_defaulters = non_defaulters['AGE'].mean()
    # Step 4: Compare average age
    print("Average age of defaulters:", avg_age_defaulters)
    print("Average age of non-defaulters:", avg_age_non_defaulters)

    # Will see the graph for Gender wise
    print()
    print('The observation is that the average age of defaulters is approximately 41.33 years, while the average age of non-
    defaulters is approximately 44.19 years.')
    print()
    sn.displot(data=data,x='AGE',col='TARGET',kde=True,row='CODE_GENDER')
    plt.show()
```

Average age of defaulters: 40.86229910714286
Average age of non-defaulters: 44.19487923609811

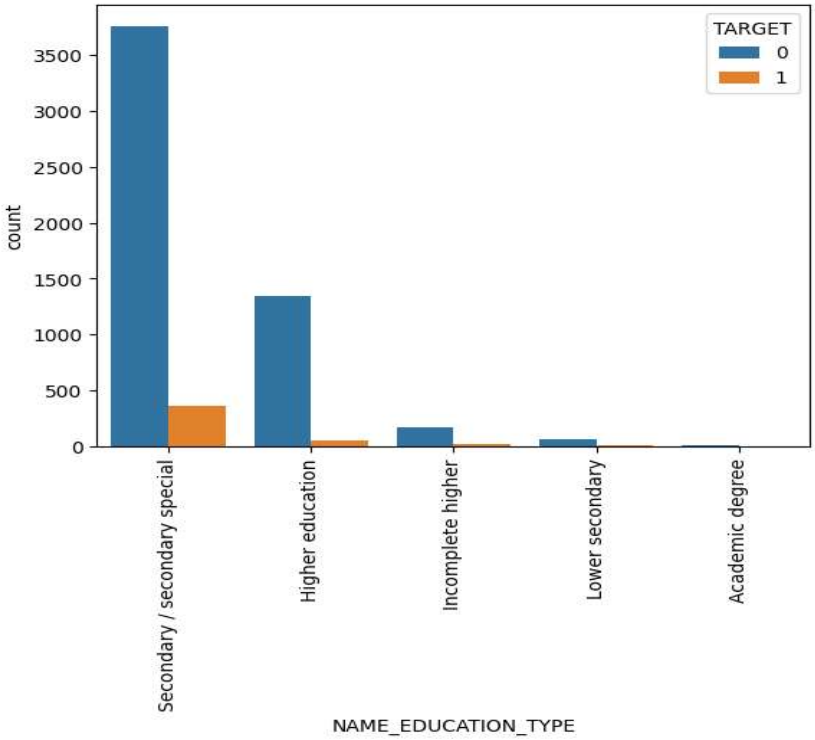
The observation is that the average age of defaulters is approximately 41.33 years, while the average age of non-defaulters is approximately 44.19 years.




```
# 5 Question :
# 5-Do clients with higher education levels tend to default less often?
    sn.countplot(x='NAME_EDUCATION_TYPE', hue='TARGET', data=data,)
    plt.xticks(rotation=90)
    plt.show()

    print('For defaulters (TARGET=1), the percentage distribution ranges from 0% to 11.63%, with the highest percentage among those
    with Lower secondary education')
    print()
    pd.crosstab(index=data['TARGET'], columns=data['NAME_EDUCATION_TYPE'], normalize='columns')*100
```

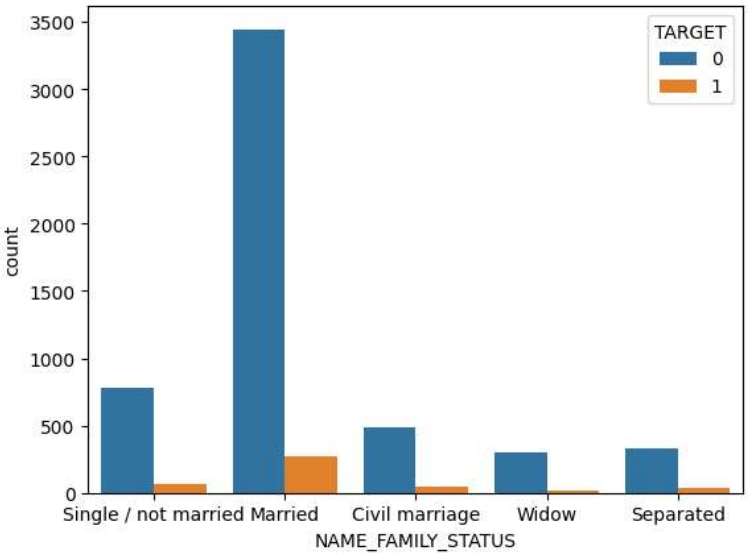
For defaulters (TARGET=1), the percentage distribution ranges from 0% to 11.63%, with the highest percentage among those with Lower secondary education



NAME_EDUCATION_TYPE	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special
TARGET					
0	100.0	95.928571	90.957447	87.323944	91.147223
1	0.0	4.071429	9.042553	12.676056	8.852777

```
# 6 Question
# 6-What is the relationship between family status and loan default rates?
    sn.countplot(data=data,x='NAME_FAMILY_STATUS',hue='TARGET')
    (pd.crosstab(index=data['TARGET'], columns=data['NAME_FAMILY_STATUS'], normalize='columns')*100)
```

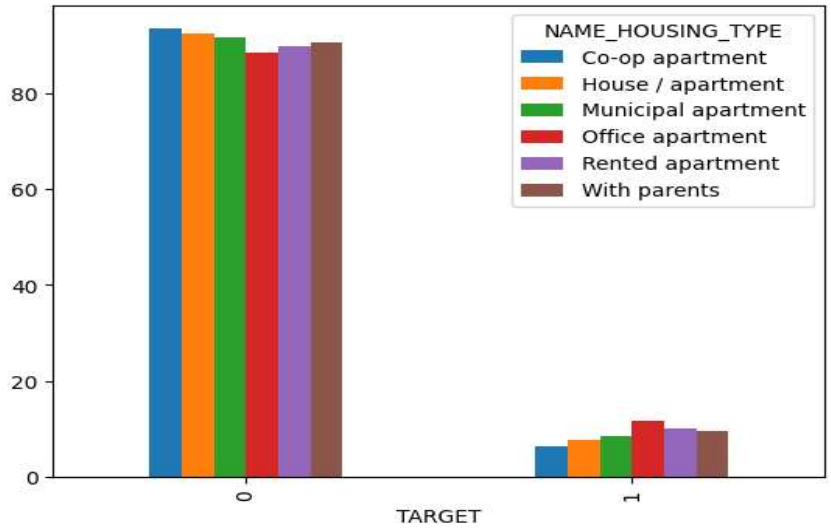
The observation indicates that the default rate among married individuals is lower compared to other categories, with the highest default rate at 10% observed in the civil marriage category.



NAME_FAMILY_STATUS	Civil marriage	Married	Separated	Single / not married	Widow
TARGET					
0	90.485075	92.632428	90.883978	91.812865	93.690852
1	9.514925	7.367572	9.116022	8.187135	6.309148

```
# 7 Questoin:
#  7-Are there specific housing types associated with higher rates of loan default?
     sn.countplot(data=data,x='NAME_HOUSING_TYPE',hue='TARGET')
     plt.xticks(rotation=90)
     plt.show()
     (pd.crosstab(index=data['TARGET'], columns=data['NAME_HOUSING_TYPE'], normalize='columns')*100).plot(kind='bar')
```

The default rate varies across different housing types.
Customers residing in a "Rented apartment" or "With parents" have a higher likelihood of default (12.3% and 11.7% respectively), while those in a "House/apartment" show the lowest default rate at 7.8%. Housing type appears to influence loan repayment behaviour.

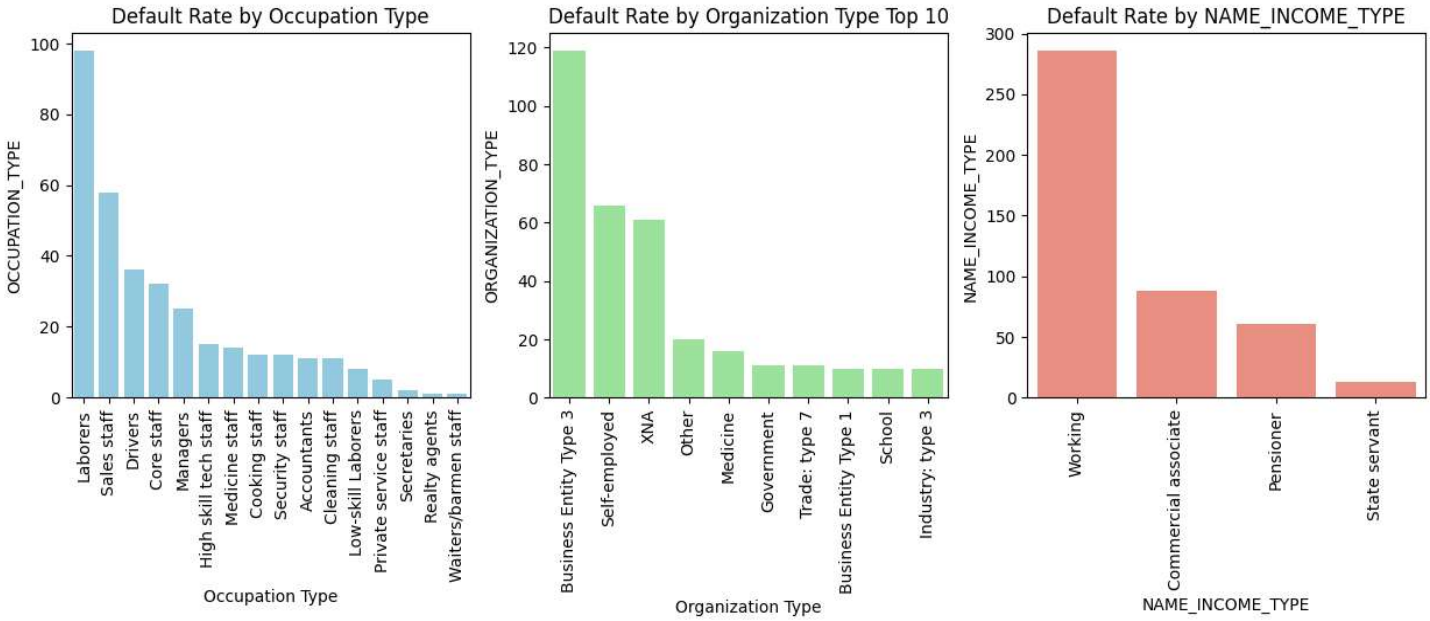



```
# 8 Question
# 8-How do factors such as occupation type, organization type, and income type relate to loan default rates?
     plt.figure(figsize=(15, 4))
     plt.subplot(1, 3, 1)
     OCCUPATION_TYPE_DATA=data[data['TARGET']==1]['OCCUPATION_TYPE'].reset_index()
     sn.barplot(OCCUPATION_TYPE_DATA['OCCUPATION_TYPE'].value_counts(),color='skyblue')
     plt.xticks(rotation=90)
     plt.title('Default Rate by Occupation Type')
     plt.xlabel('Occupation Type')
     plt.subplot(1, 3, 2)
     ORGANIZATION_TYPE_DATA=data[data['TARGET']==1]['ORGANIZATION_TYPE'].reset_index()
     ORGANIZATION_TYPE_DATA=ORGANIZATION_TYPE_DATA['ORGANIZATION_TYPE'].value_counts().head(10).reset_index()
     sn.barplot(data=ORGANIZATION_TYPE_DATA,x='index',y='ORGANIZATION_TYPE',color='lightgreen')
     plt.xticks(rotation=90)
     plt.title('Default Rate by Organization Type Top 10')
     plt.xlabel('Organization Type')
     plt.subplot(1, 3, 3)
     OCCUPATION_TYPE_DATA=data[data['TARGET']==1]['NAME_INCOME_TYPE'].reset_index()
     sn.barplot(OCCUPATION_TYPE_DATA['NAME_INCOME_TYPE'].value_counts(),color='salmon')
     plt.xticks(rotation=90)
     plt.title('Default Rate by NAME_INCOME_TYPE')
     plt.xlabel('NAME_INCOME_TYPE')
     plt.show()


     print()
     print("""
     The occupation type "Laborers" shows the highest default rate among all categories, indicating a potential risk factor for loan
     default.
     Within organization types, "Business entity type" emerges as the leading category with the highest default rate, suggesting a
     correlation between organization type and default risk.
     Individuals classified as "Working" in terms of income type exhibit a higher likelihood of defaulting on loans, indicating a
     possible association between employment status and loan repayment behavior.
```

```
""")
```

- The occupation type "Laborers" shows the highest default rate among all categories, indicating a potential risk factor for loan default.
- Within organization types, "Business entity type" emerges as the leading category with the highest default rate, suggesting a correlation between organization type and default risk.
- Individuals classified as "Working" in terms of income type exhibit a higher likelihood of defaulting on loans, indicating a possible association between employment status and loan repayment behavior.



Conclusion:

- Targeted financial education needed for females to improve repayment rates.
- Risk assessment crucial for cash loan products due to defaults.
- Younger borrowers, aged 41.33, require flexible repayment options.
- Rented apartments pose higher default risk; offer financial counseling.
- Consider occupation and organization type for tailored financial products.