credit-card-loan-approval

April 17, 2024

1 Credit Card Approval Prediction

1.0.1 Project Overview:

A bank's credit card department is one of the top adopters of data science. A top focus for the bank has always been acquiring new credit card customers. Giving out credit cards without doing proper research or evaluating applicants' creditworthiness is quite risky. The credit card department has been using a data-driven system for credit assessment called Credit Scoring for many years, and the model is known as an application scorecard. A credit card application's cutoff value is determined using the application scorecard, which also aids in estimating the applicant's level of risk. This decision is made based on strategic priority at a given time Customers must fill out a form, either physically or online, to apply for a credit card. The application data is used to evaluate the applicant's creditworthiness. The decision is made using the application data in addition to the Credit Bureau Score, such as the FICO Score in the US or the CIBIL Score in India, and other internal information on the applicants. Additionally, the banks are rapidly taking a lot of outside data into account to enhance the caliber of credit judgements.

1.0.2 Project Objective:

The main objective of this assignment is to minimize the risk and maximize the profit of the bank. Bank has to make a decision based on the applicant's profile to minimize the loss from the bank's perspective. Bank considers the applicant's over their nature of work, income range and family orientaion details to take any decision to approve or reject a credit card application. The customer Credit card data contains many features and a classification approach to identify the credit worthiness of an applicant.

In this project we are utilizing the exploratory data analysis (EDA) as a data exploration technique to acquire knowledge, discover new relations, apply new methodologies and unravel patterns in data. It is important to apply the necessary rationale behind each step to address the main objective of the study.

So, The primary objective of this project is to develop a machine learning model for Credit Card Approval Prediction.

1.1 Feature Understanding

Dataset name: (Credit_Card.csv)

- Ind_ID: Client ID
- Gender: Gender information
- Car_owner: Having car or not

- Propert_owner: Having property or not
- Children: Count of children
- Annual_income: Annual income
- Type_Income: Income type
- Education: Education level
- Marital_status: Marital_status
- Housing_type: Living style
- Birthday_count: Use backward count from current day (0), -1 means yesterday.
- Employed_days: Start date of employment. Use backward count from current day (0). Positive value means, individual is currently unemployed.
- Mobile_phone: Any mobile phone
- Work_phone: Any work phone
- Phone: Any phone number
- EMAIL_ID: Any email ID
- Type_Occupation: Occupation
- Family_Members: Family size

Another data set (Credit_card_label.csv) contains two key pieces of information - ID: The joining key between application data and credit status data, same is Ind_ID - Label: 0 is application approved and 1 is application rejected.

1.1.1 Required Libraries

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.simplefilter('ignore')
pd.set_option('display.max_rows',None)
pd.set_option('display.max_columns',None)
```

1.1.2 Import Datasets

[2]:
```
credit_card=pd.read_csv('Credit_card.csv')
credit_card_label=pd.read_csv('Credit_card_label.csv')
```

1.1.3 Merging Both DataSets Using Pandas Merge Function

[3]:
```
data=pd.merge(credit_card,credit_card_label,on='Ind_ID',how='inner')
```

[4]:
```
df=data.copy()
```

[5]:
```
df.head()
```

[5]:
```
    Ind_ID GENDER Car_Owner Propert_Owner  CHILDREN  Annual_income  \
0  5008827      M         Y             Y         0       180000.0
```

```
1  5008865      F          Y             Y          2       135000.0
2  5008889      F          N             Y          0       247500.0
3  5009000      M          Y             Y          0       157500.0
4  5009023      F          N             Y          2       216000.0

            Type_Income                      EDUCATION Marital_status  \
0             Pensioner               Higher education        Married
1               Working  Secondary / secondary special        Married
2  Commercial associate               Higher education      Separated
3               Working  Secondary / secondary special        Married
4         State servant               Higher education        Married

        Housing_type  Birthday_count  Employed_days  Mobile_phone  Work_Phone  \
0  House / apartment        -18772.0         365243             1           0
1  House / apartment        -15761.0          -3173             1           0
2   Rented apartment        -17016.0          -1347             1           0
3  House / apartment         -9927.0           -828             1           0
4  House / apartment        -15444.0          -3112             1           0

   Phone  EMAIL_ID Type_Occupation  Family_Members  label
0      0         0             NaN               2      1
1      0         0        Laborers               4      0
2      0         0      Core staff               1      0
3      0         0         Drivers               2      0
4      0         1             NaN               4      0
```

1.1.4 Shape of DataFrame

[6]:
```
print(f"Total Number of Rows in Dataset={df.shape[0]}")
print(f'Total Number of Columns in Dataset={df.shape[1]}')
```

```
Total Number of Rows in Dataset=1548
Total Number of Columns in Dataset=19
```

- **We can see that, dataset contains 1548 rows and 19 colmns.**

1.2 Data Exploration

[7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1548 entries, 0 to 1547
Data columns (total 19 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Ind_ID           1548 non-null   int64
```

```
 1   GENDER           1541 non-null   object
 2   Car_Owner        1548 non-null   object
 3   Propert_Owner    1548 non-null   object
 4   CHILDREN         1548 non-null   int64
 5   Annual_income    1525 non-null   float64
 6   Type_Income      1548 non-null   object
 7   EDUCATION        1548 non-null   object
 8   Marital_status   1548 non-null   object
 9   Housing_type     1548 non-null   object
 10  Birthday_count   1526 non-null   float64
 11  Employed_days    1548 non-null   int64
 12  Mobile_phone     1548 non-null   int64
 13  Work_Phone       1548 non-null   int64
 14  Phone            1548 non-null   int64
 15  EMAIL_ID         1548 non-null   int64
 16  Type_Occupation  1060 non-null   object
 17  Family_Members   1548 non-null   int64
 18  label            1548 non-null   int64
dtypes: float64(2), int64(9), object(8)
memory usage: 241.9+ KB
```

We can gather valuable Information about the dataset.

- **Dataset contains 1548 entries(Rows) and there are 19 columns in the Dataset.**
- **Out of 19 columns, 11 columns are Numerical Columns and 8 Columns are Categorical Columns.**
- **Several columns are having missing value including GENDER, Annual_income, Birthday_count and Type_Occupation.**

1.3 Spliting columns by data types

1.3.1 Categorical Columns

[8]:
```python
categorical_columns=df.select_dtypes(include='object').columns
for i in categorical_columns:
    print(i)
```

```
GENDER
Car_Owner
Propert_Owner
Type_Income
EDUCATION
Marital_status
Housing_type
Type_Occupation
```

1.3.2 Numerical_Columns

[9]:
```
numerical_columns=df.select_dtypes(include='number').columns
for i in numerical_columns:
    print(i)
```

```
Ind_ID
CHILDREN
Annual_income
Birthday_count
Employed_days
Mobile_phone
Work_Phone
Phone
EMAIL_ID
Family_Members
label
```

1.4 Checking Null Values In Dataset

[10]:
```
df.isnull().sum()/len(df)*100
```

[10]:
```
Ind_ID              0.000000
GENDER              0.452196
Car_Owner           0.000000
Propert_Owner       0.000000
CHILDREN            0.000000
Annual_income       1.485788
Type_Income         0.000000
EDUCATION           0.000000
Marital_status      0.000000
Housing_type        0.000000
Birthday_count      1.421189
Employed_days       0.000000
Mobile_phone        0.000000
Work_Phone          0.000000
Phone               0.000000
EMAIL_ID            0.000000
Type_Occupation    31.524548
Family_Members      0.000000
label               0.000000
dtype: float64
```

Here we can see

- **Gender: 0.45% missing values**
- **Annual_income: 1.49% missing values**
- **Birthday_count: 1.42% missing values**

- **Type_Occupation: 31.52% missing values**

1.5 Drop Unnecessery Columns

- **The Features Mobile_phone, Work_Phone, Phone, EMAIL_ID are present in the dataset but these columns are unnecessary for data analysis. Drop these unnecessary columns.**
- **Type_Occupation contains 31.52% nulls values thats why we consider removing it.**

[11]:
```
df.drop(columns=["Mobile_phone", "Work_Phone", "Phone",␣
↪"EMAIL_ID",'Type_Occupation'],axis=1,inplace=True)
```

[12]:
```
df.columns
```

[12]:
```
Index(['Ind_ID', 'GENDER', 'Car_Owner', 'Propert_Owner', 'CHILDREN',
       'Annual_income', 'Type_Income', 'EDUCATION', 'Marital_status',
       'Housing_type', 'Birthday_count', 'Employed_days', 'Family_Members',
       'label'],
      dtype='object')
```

1.5.1 Feature Engineering

Calculte the approx age of customers using Birthday count:

[13]:
```
import math
Age=[]
for i in df['Birthday_count']:
    if not math.isnan(i):
        a=i/365
        Age.append(round(abs(a)))
    else:
        Age.append(np.nan)
df['Age']=Age
```

1.5.2 Creating an 'Emmployed_Status' Feature from 'Employed_days'

[14]:
```
Employed_Status=[]
for i in df['Employed_days']:
    if i<0:
        Employed_Status.append('Employed')
    else:
        Employed_Status.append('Unemployed')
df['Employed_Status']=Employed_Status
```

[15]:
```
df.drop(columns=['Birthday_count','Employed_days'],axis=1,inplace=True)
```

[16]:
```
df.rename(columns={'label':'Approved_status'},inplace=True)
```

[17]:
```
df.head()
```

[17]:

	Ind_ID	GENDER	Car_Owner	Propert_Owner	CHILDREN	Annual_income
0	5008827	M	Y	Y	0	180000.0
1	5008865	F	Y	Y	2	135000.0
2	5008889	F	N	Y	0	247500.0
3	5009000	M	Y	Y	0	157500.0
4	5009023	F	N	Y	2	216000.0

	Type_Income	EDUCATION	Marital_status
0	Pensioner	Higher education	Married
1	Working	Secondary / secondary special	Married
2	Commercial associate	Higher education	Separated
3	Working	Secondary / secondary special	Married
4	State servant	Higher education	Married

	Housing_type	Family_Members	Approved_status	Age	Employed_Status
0	House / apartment	2	1	51.0	Unemployed
1	House / apartment	4	0	43.0	Employed
2	Rented apartment	1	0	47.0	Employed
3	House / apartment	2	0	27.0	Employed
4	House / apartment	4	0	42.0	Employed

- **Now we have 14 Features to Analysis**

1.6 Overall Statistics about the Dataset

[18]:
```
df.describe()
```

[18]:

	Ind_ID	CHILDREN	Annual_income	Family_Members
count	1.548000e+03	1548.000000	1.525000e+03	1548.000000
mean	5.078920e+06	0.412791	1.913993e+05	2.161499
std	4.171759e+04	0.776691	1.132530e+05	0.947772
min	5.008827e+06	0.000000	3.375000e+04	1.000000
25%	5.045070e+06	0.000000	1.215000e+05	2.000000
50%	5.078842e+06	0.000000	1.665000e+05	2.000000
75%	5.115673e+06	1.000000	2.250000e+05	3.000000
max	5.150412e+06	14.000000	1.575000e+06	15.000000

	Approved_status	Age
count	1548.000000	1526.000000
mean	0.113049	43.952818
std	0.316755	11.603295
min	0.000000	21.000000
25%	0.000000	34.000000
50%	0.000000	43.000000
75%	0.000000	54.000000

```
max            1.000000    68.000000
```

1.7 Data Summary Report

- **The average income is approximetly 1,91,399.30,with a standard deviation of 1,13,253.0, suggesting a wide income distribution.**
- **The minimum and maximum Annual_income are 33,750.00 and 1,57,500.0.**
- **The Range of age is between 21 years to 68 years and average age of customers is approx 44 years.**

[19]:
```python
df.describe(include='object')
```

[19]:
```
       GENDER Car_Owner Propert_Owner Type_Income  \
count    1541      1548          1548        1548
unique      2         2             2           4
top         F         N             Y     Working
freq      973       924          1010         798

                            EDUCATION Marital_status       Housing_type  \
count                            1548           1548               1548
unique                              5              5                  6
top     Secondary / secondary special        Married  House / apartment
freq                             1031           1049               1380

       Employed_Status
count             1548
unique               2
top           Employed
freq              1287
```

1.7.1 Data Summary

- **In Dataset most of the customers are not having car. means, they are not car owners.**
- **In dataset most of the customers are married and their education is Secondary/secondary special.**
- **There is most of the customers are Employed, Working and living in House/apartment.**
- **And, Most of the customers are Property owner.***

1.7.2 Check Unique Values in Categorical Columns

[20]:
```python
categorical_column=df.select_dtypes(include='object').columns
for i in categorical_column:
    print(f"Unique values in {i} column")
    print(df[i].unique())
    print('-'*50)
```

```
Unique values in GENDER column
['M' 'F' nan]
--------------------------------------------------
Unique values in Car_Owner column
['Y' 'N']
--------------------------------------------------
Unique values in Propert_Owner column
['Y' 'N']
--------------------------------------------------
Unique values in Type_Income column
['Pensioner' 'Working' 'Commercial associate' 'State servant']
--------------------------------------------------
Unique values in EDUCATION column
['Higher education' 'Secondary / secondary special' 'Incomplete higher'
 'Lower secondary' 'Academic degree']
--------------------------------------------------
Unique values in Marital_status column
['Married' 'Separated' 'Civil marriage' 'Single / not married' 'Widow']
--------------------------------------------------
Unique values in Housing_type column
['House / apartment' 'Rented apartment' 'With parents'
 'Municipal apartment' 'Office apartment' 'Co-op apartment']
--------------------------------------------------
Unique values in Employed_Status column
['Unemployed' 'Employed']
--------------------------------------------------
```

Gender Column is having nulls values.

1.7.3 Modifying categories in categorical columns

[21]:
```
Marital_status_mapping={'Married':'Married',
                        'Separated':'Separated',
                        'Civil marriage':'Civil marriage'
                        ,'Single / not married':'not married',
                        'Widow':'Widow'}
df['Marital_status']=df['Marital_status'].map(Marital_status_mapping)

df['Marital_status'].unique()
```

[21]:
```
array(['Married', 'Separated', 'Civil marriage', 'not married', 'Widow'],
      dtype=object)
```

[22]:
```
EDUCATION_mapping={'Higher education':'Higher Education',
                   'Secondary / secondary special':'Secondary Education',
                   'Incomplete higher':'Incomplete Higher',
                   'Lower secondary':'Lower Secondary',
```

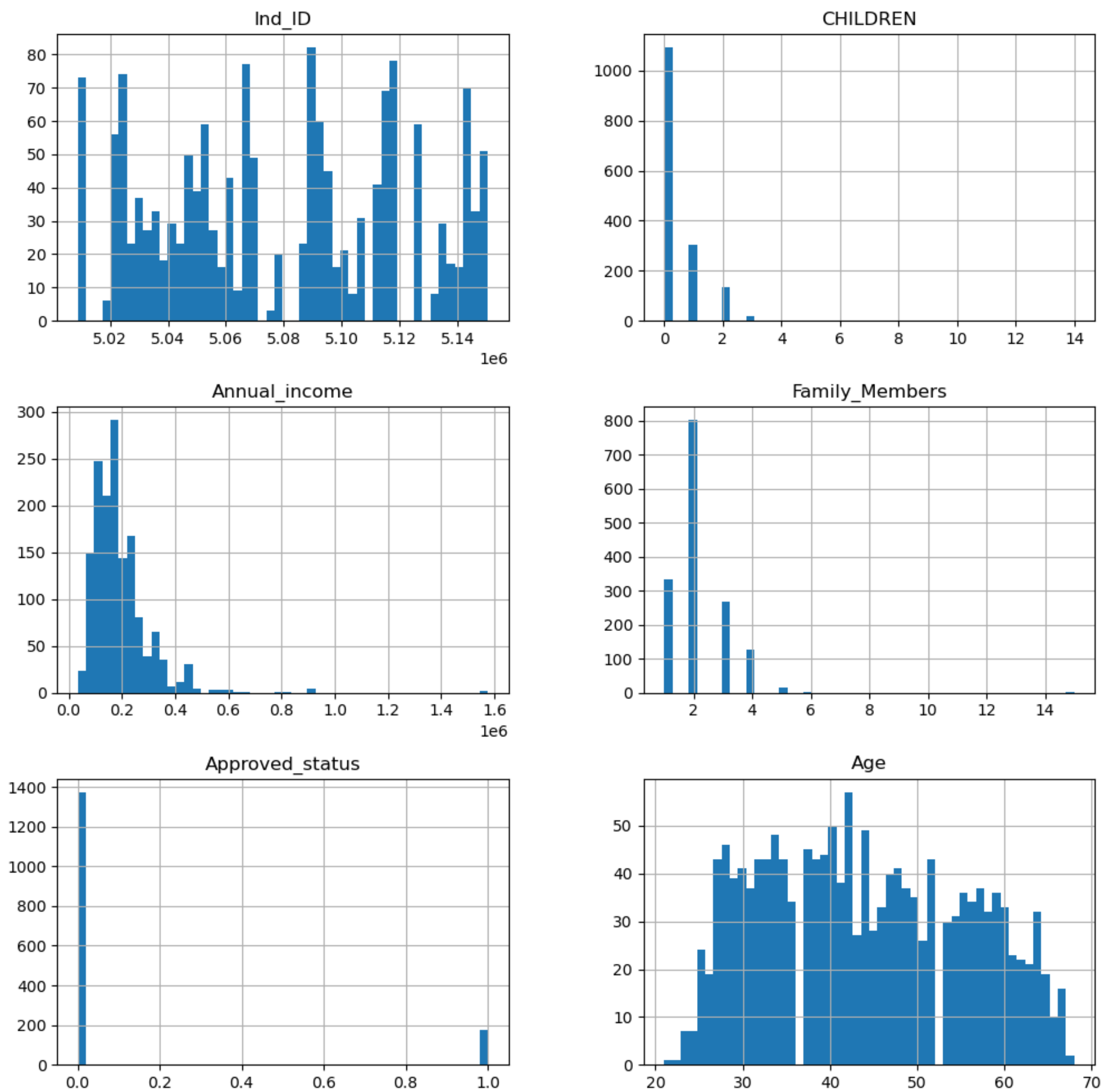
```
                        'Academic degree':'Academic Degree'}
df['EDUCATION']=df['EDUCATION'].map(EDUCATION_mapping)

df['EDUCATION'].unique()
```

[22]:
```
array(['Higher Education', 'Secondary Education', 'Incomplete Higher',
       'Lower Secondary', 'Academic Degree'], dtype=object)
```

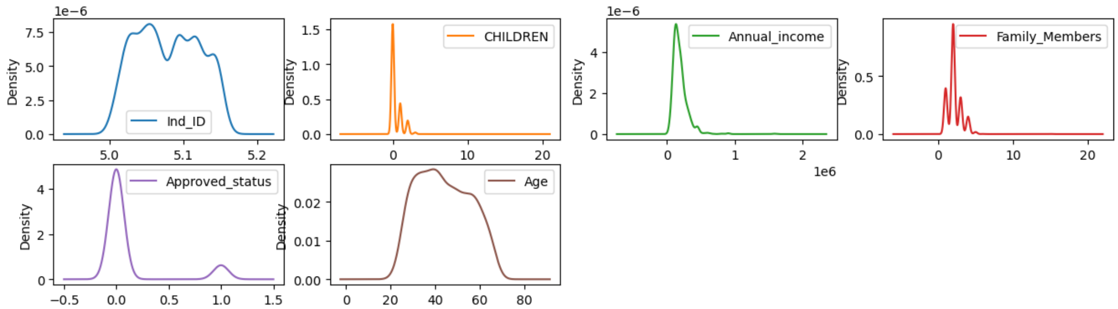
1.7.4 Visualizing the Data for Better Understanding

Distribution of Numerical Variables

[23]:
```
df.hist(bins=50,figsize=(12,12))
plt.show()
```



[24]:
```
df.plot(kind='density',subplots=True,figsize=(15,10),sharex=False,layout=(5,4))
plt.show()
```



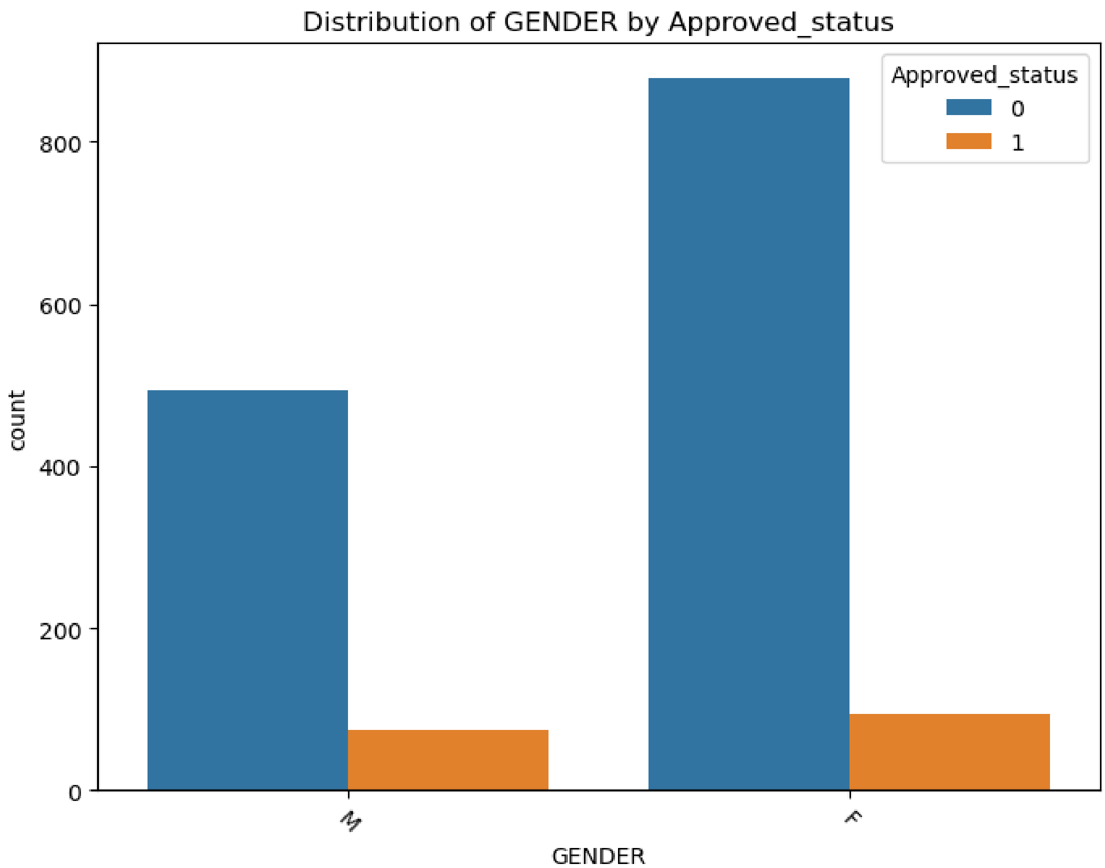
- With above visualization we can say that there is zero children customers are more and very less customers has 3 children.
- We can see that, max of customers are belongs to less than 40,000 annual income and very less customers are belongs to more than 40,000 annual income.
- we can see that,max of customer's family members are couples and very less customer,s family members are above 4.
- With above visualization, we can see most of customers are belongs to range of 21 to 68 years age.

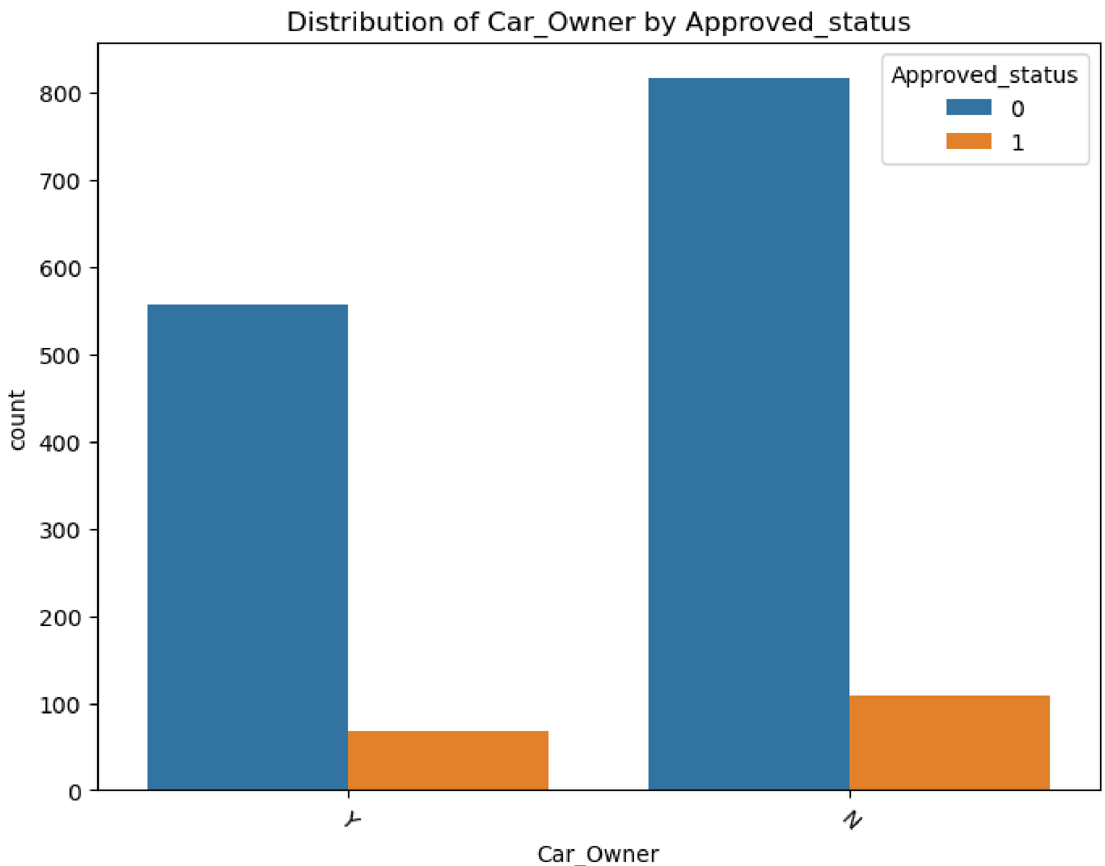
[25]:
```
numerical_columns=[]
for i in df.select_dtypes(include='number'):
    numerical_columns.append(i)
```

[26]:
```
numerical_columns
```

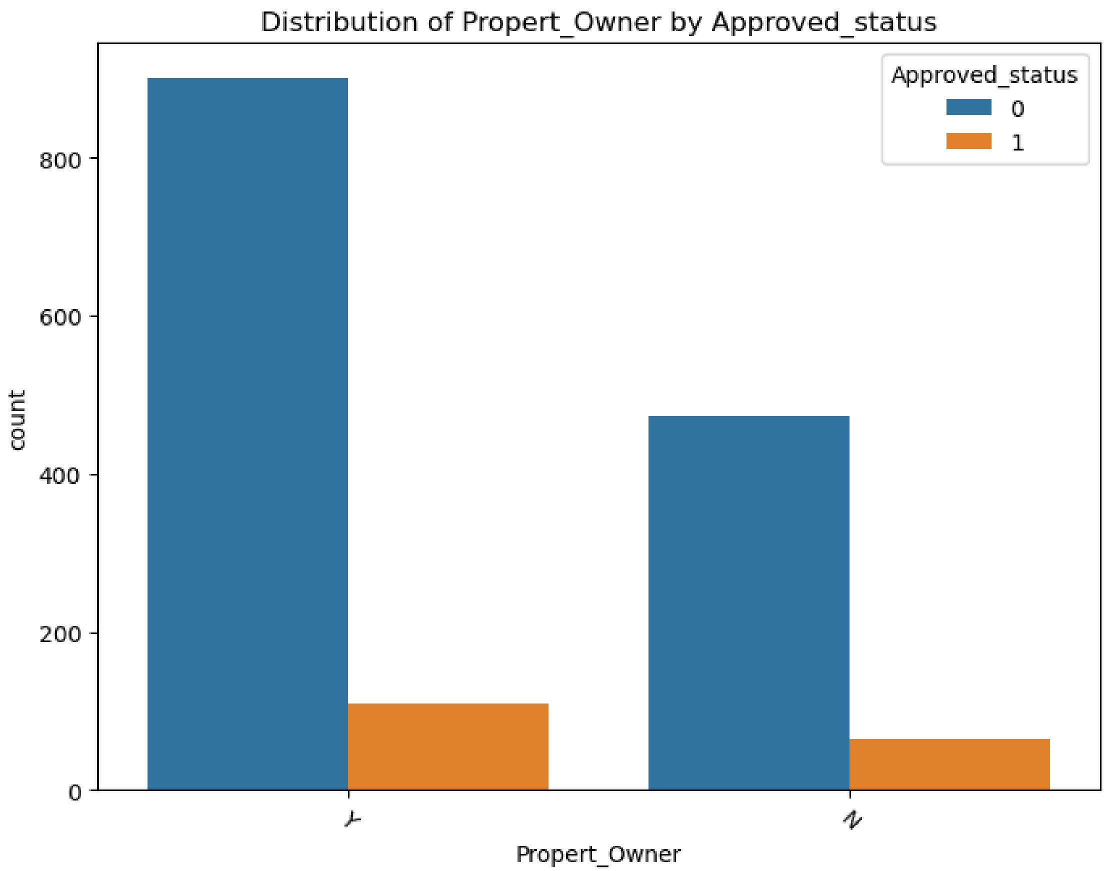
[26]:
```
['Ind_ID',
 'CHILDREN',
 'Annual_income',
 'Family_Members',
 'Approved_status',
 'Age']
```

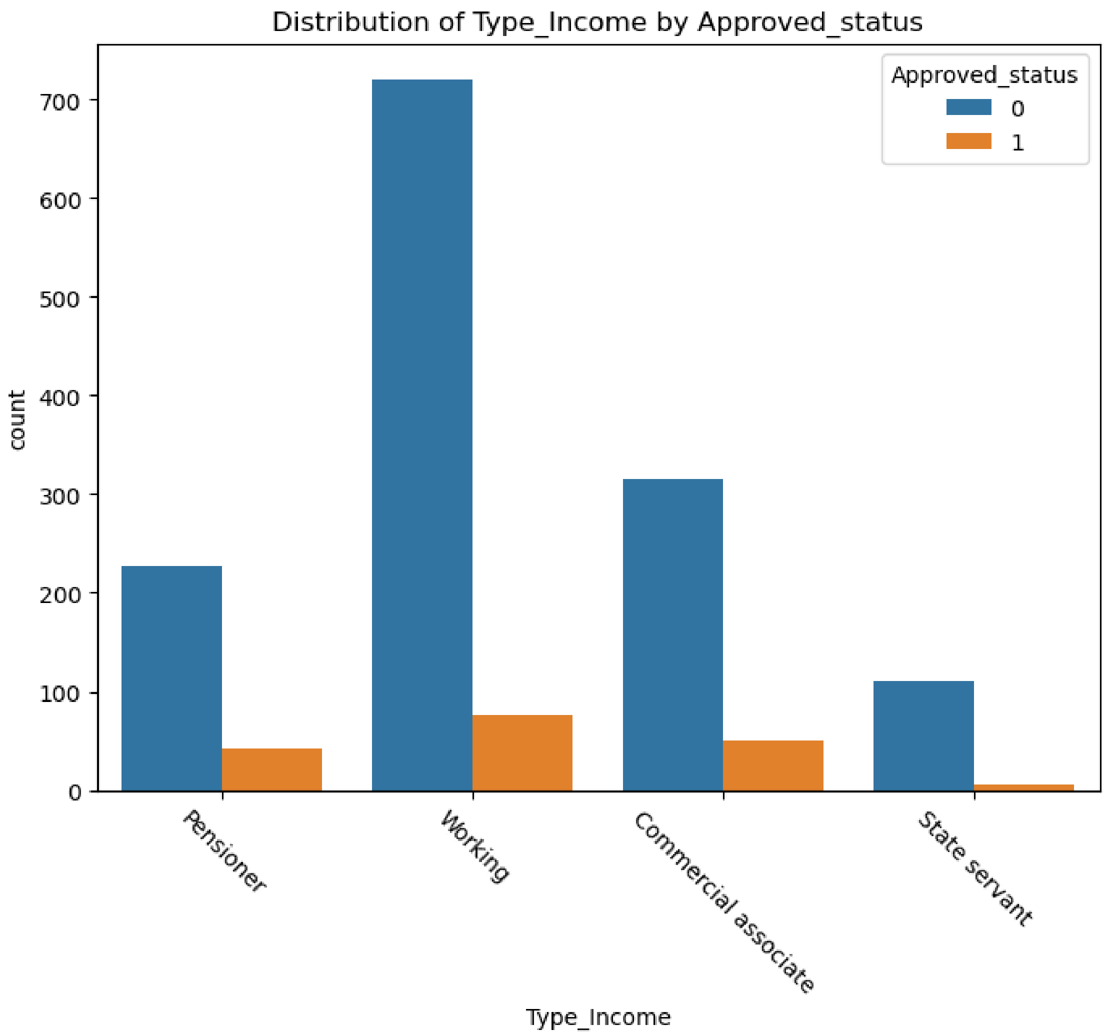
[27]:
```
categorical_features=['GENDER', 'Car_Owner', 'Propert_Owner', 'Type_Income',␣
 ↪'EDUCATION','Marital_status', 'Housing_type', 'Employed_Status']
for feature in categorical_features:
    plt.figure(figsize=(8,6))
    sns.countplot(x=feature,data=df,hue='Approved_status')
    plt.xticks(rotation=-45)
    plt.title(f'Distribution of {feature} by Approved_status')
    plt.show()
```

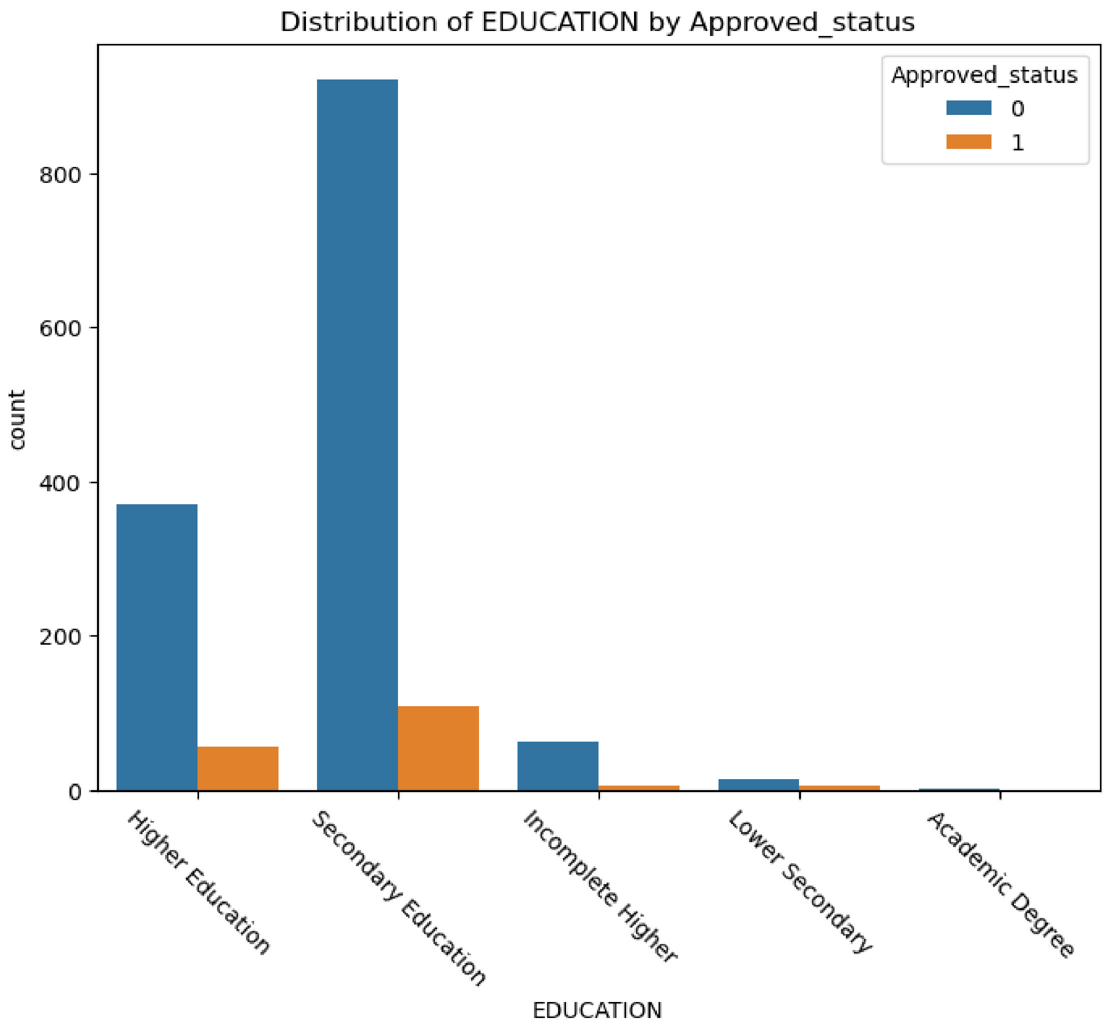

Distribution of GENDER by Approved_status
Approved_status
0
1
count
0
200
400
600
800
M
F
GENDER

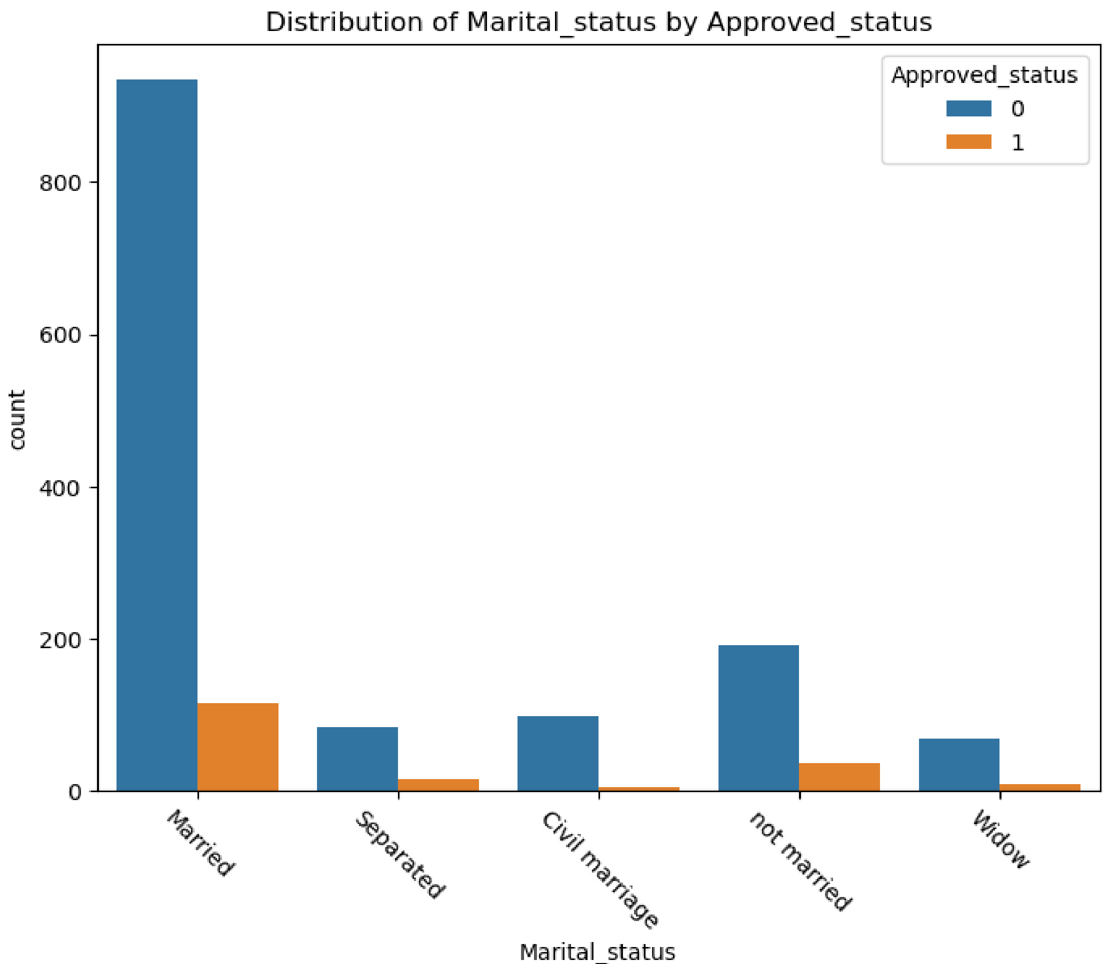


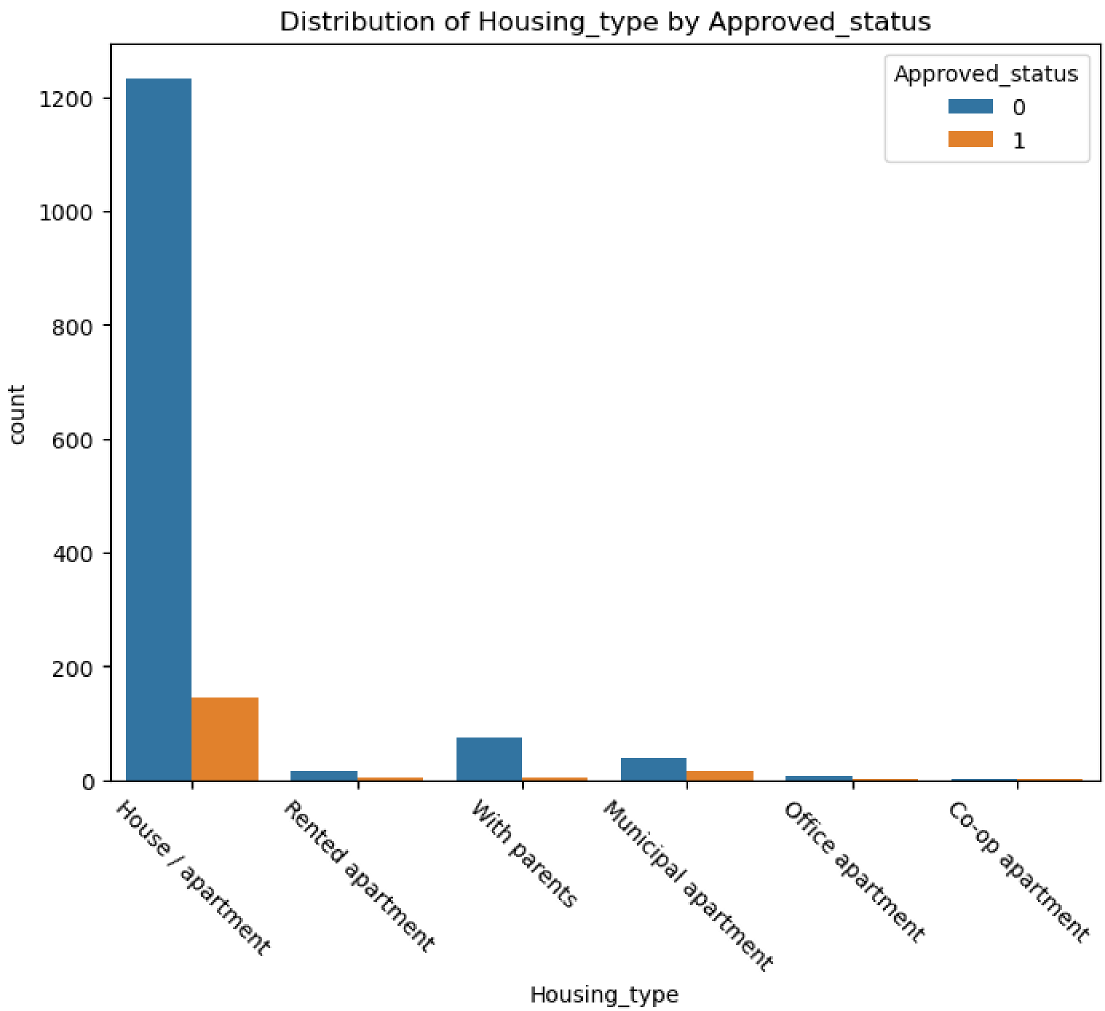
Distribution of Car_Owner by Approved_status
Approved_status
0
1
800
700
600
500
400
300
200
100
0
count
Y
N
Car_Owner

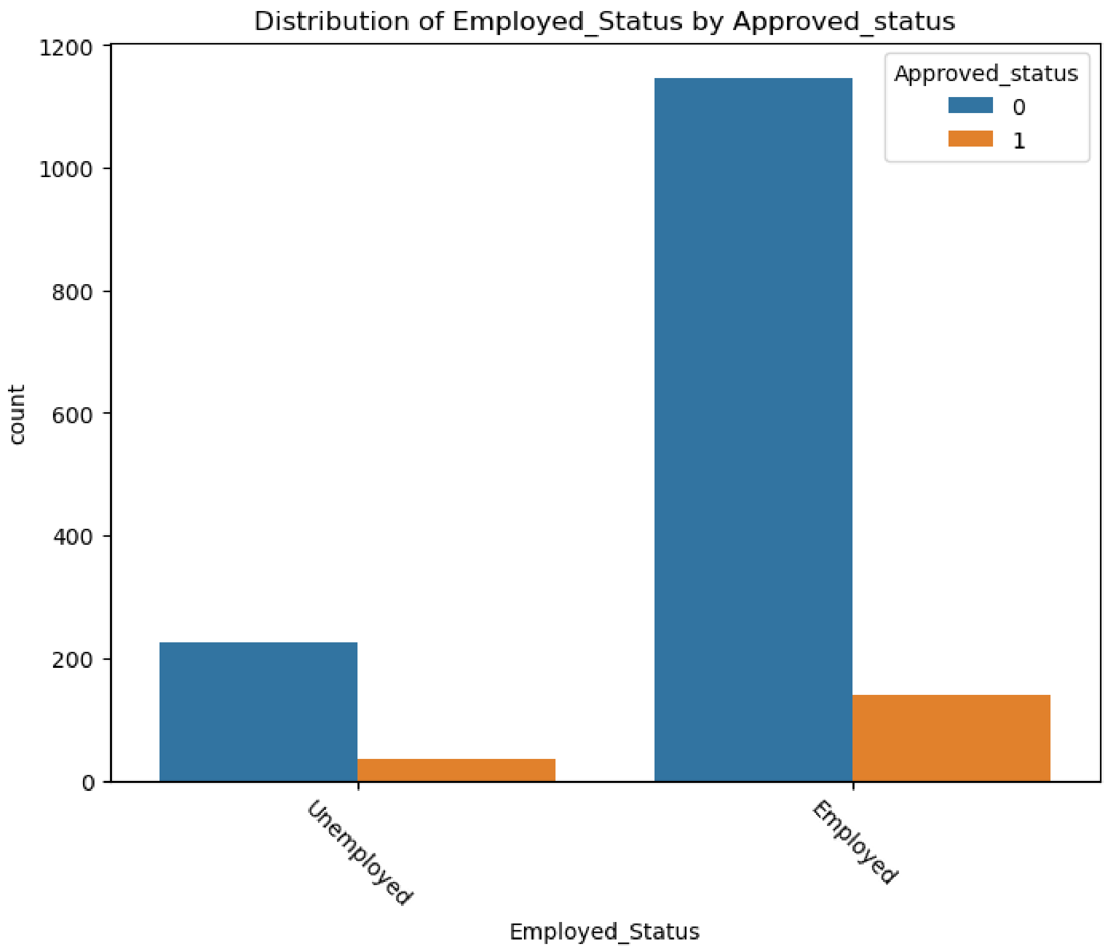

Distribution of Propert_Owner by Approved_status
Approved_status
0
1
count
800
600
400
200
0
Y
N
Propert_Owner


Distribution of Type_Income by Approved_status
Approved_status
0
1
count
700
600
500
400
300
200
100
0
Pensioner
Working
Commercial associate
State servant
Type_Income


Distribution of EDUCATION by Approved_status
Approved_status
0
1
800
600
400
200
0
count
Higher Education
Secondary Education
Incomplete Higher
Lower Secondary
Academic Degree
EDUCATION

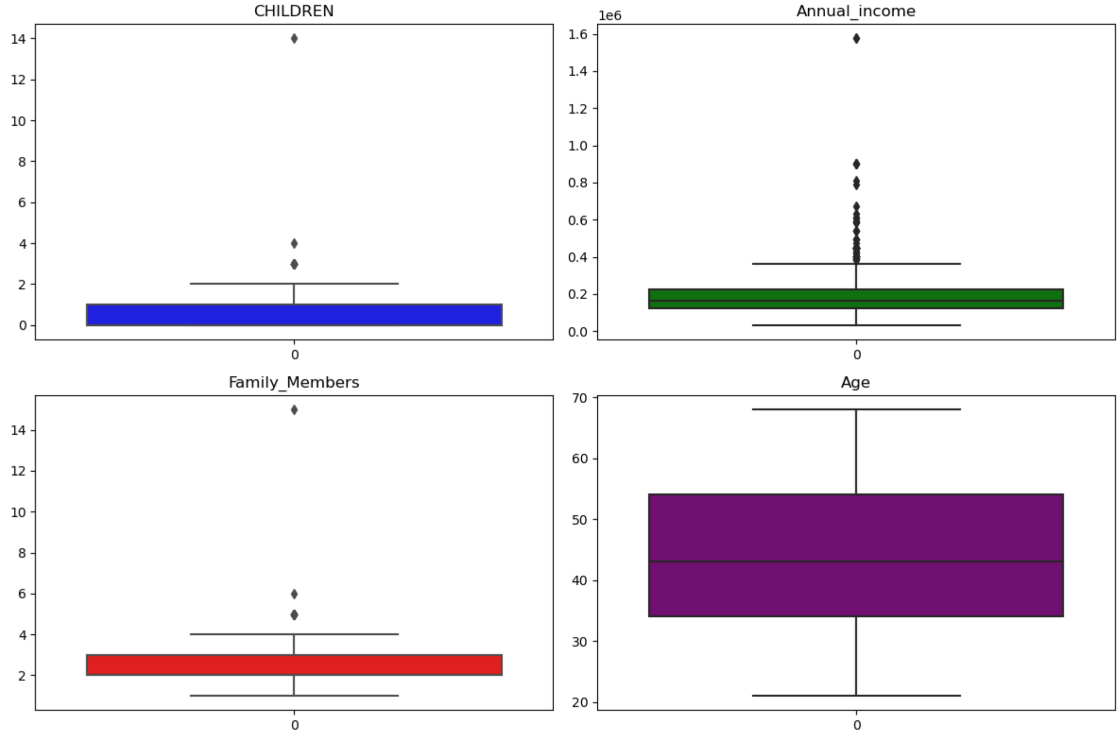

Distribution of Marital_status by Approved_status
Approved_status
0
1
800
600
400
200
0
count
Married
Separated
Civil marriage
not married
Widow
Marital_status


Distribution of Housing_type by Approved_status
Approved_status
0
1
count
1200
1000
800
600
400
200
0
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
Housing_type



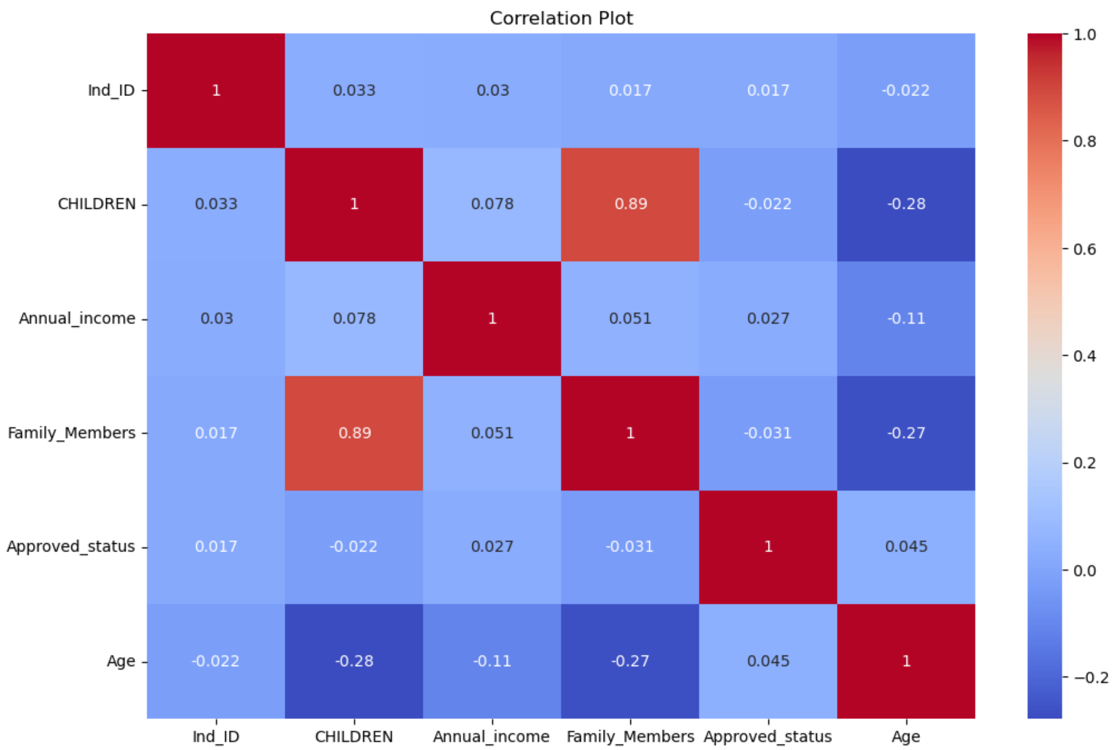
1.7.5 Boxplot to check Outliers in dataset

[29]:
```
columns_to_plot=['CHILDREN', 'Annual_income','Family_Members','Age']
colors=['blue','green','red','purple']
fig,axes=plt.subplots(nrows=2,ncols=2,figsize=(12,8))
for i,column in enumerate(columns_to_plot):
    sns.boxplot(data=df[column],ax=axes[i//2,i%2],palette=[colors[i]])
    axes[i//2,i%2].set_title(column)

plt.tight_layout()
plt.show()
```



1.7.6 So the dataset contains outliers in features like CHILDREN,Annual income and Family__Members

[30]:
```
corr_matrix=df.corr()
plt.figure(figsize=(12,8))
sns.heatmap(corr_matrix,annot=True,fmt='.
↪2g',cmap='coolwarm',linewidths=0,linecolor='black',cbar=True)
plt.title('Correlation Plot')
plt.show()
```



1.7.7 It has shown there is a strong correlation between Family_Members and Children.

1.8 Data Preprocessing

1.8.1 Handling missing values

[31]:
```
df.isnull().sum()
```

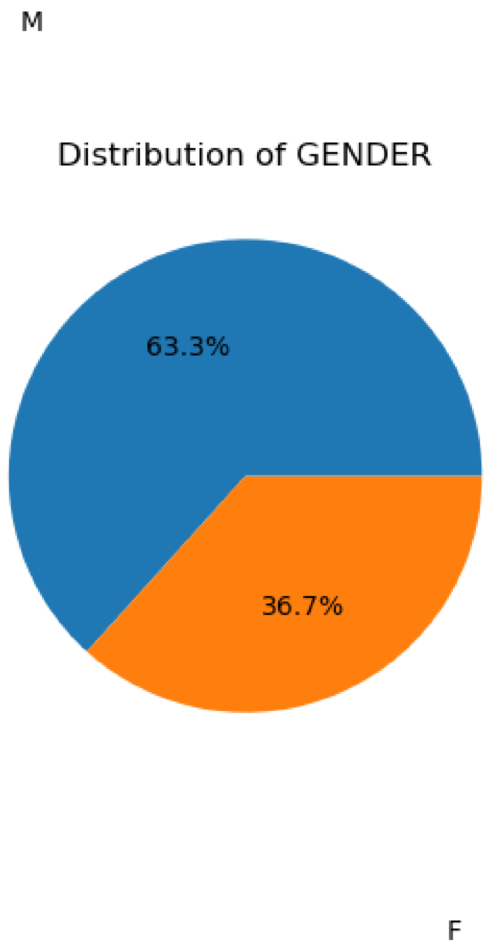
[31]:
```
Ind_ID              0
GENDER              7
Car_Owner           0
Propert_Owner       0
CHILDREN            0
Annual_income      23
Type_Income         0
EDUCATION           0
Marital_status      0
Housing_type        0
Family_Members      0
Approved_status     0
Age                22
Employed_Status     0
dtype: int64
```

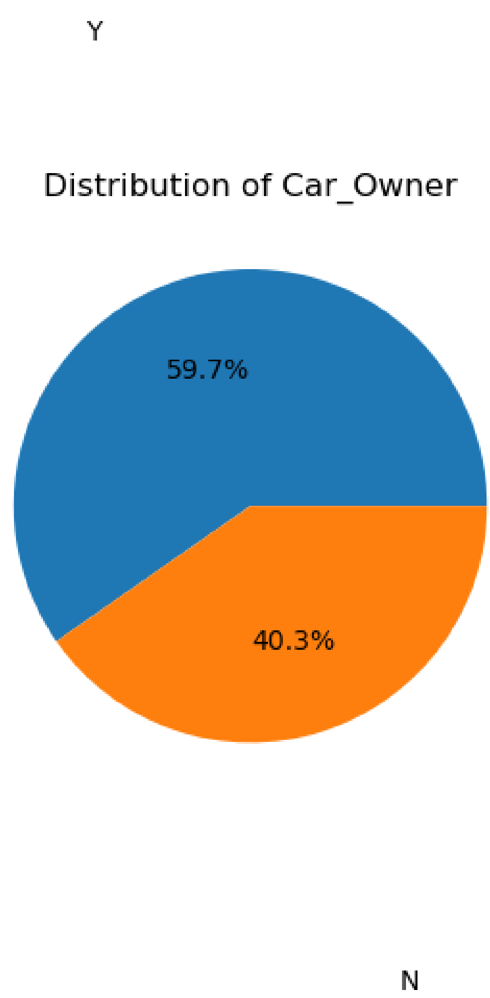
[32]:
```
df['Annual_income']=df['Annual_income'].fillna(df['Annual_income'].median())
df['Age']=df['Age'].fillna(df['Age'].mean())
df['GENDER']=df['GENDER'].fillna(df['GENDER'].mode()[0])
```

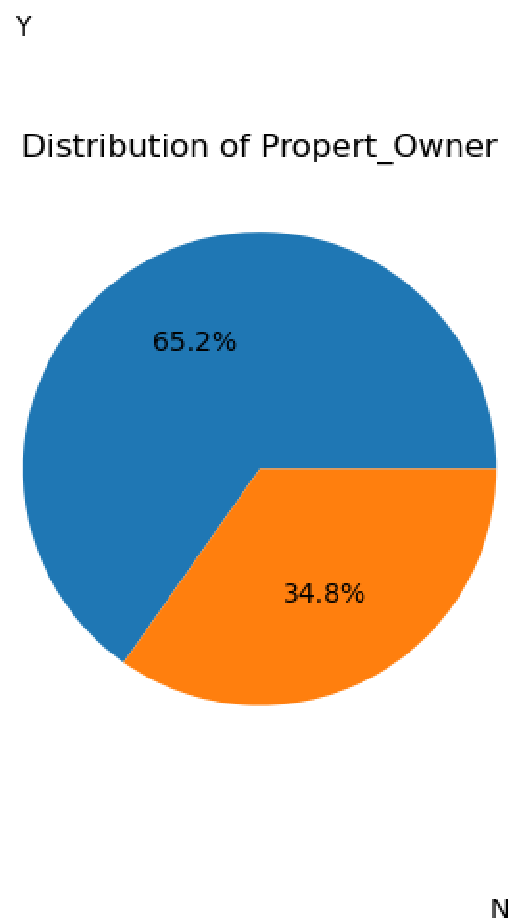
Handle Outlier in Dataset

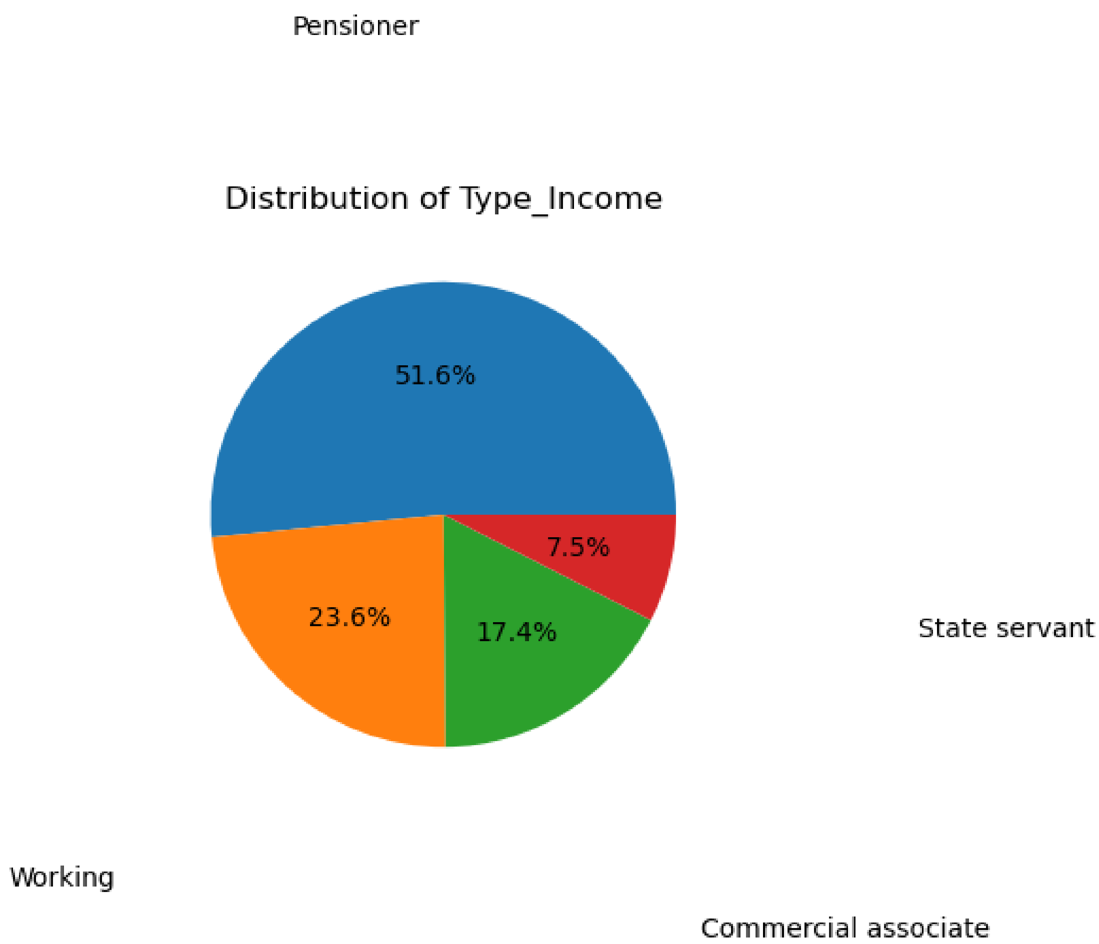
[33]:
```
Q1=df['Annual_income'].quantile(0.25)
Q3=df['Annual_income'].quantile(0.75)

IQR=Q3-Q1
Lower_limit=Q1-(1.5*IQR)
Upper_limit=Q3+(1.5*IQR)

df['Annual_income']=df['Annual_income'].clip(Lower_limit,Upper_limit)
```

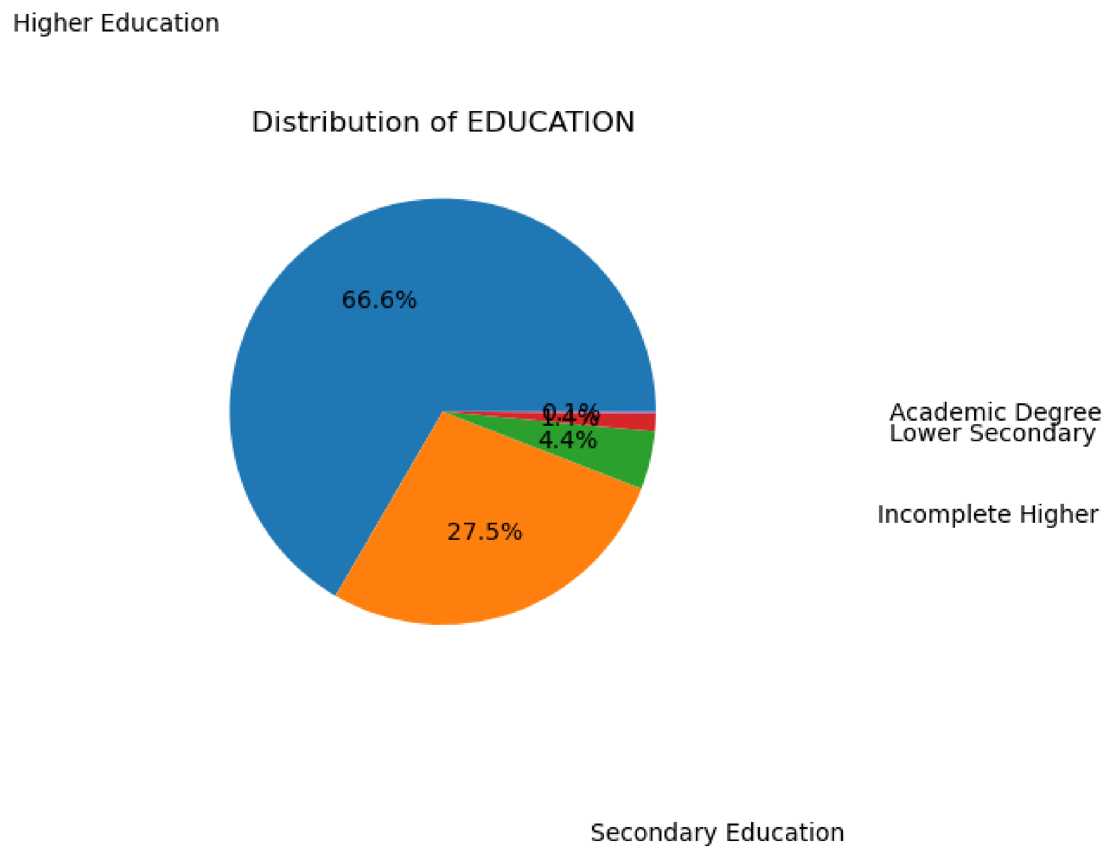
[34]:
```
categorical_features=['GENDER','Car_Owner', 'Propert_Owner', 'Type_Income',␣
 ↪'EDUCATION','Marital_status', 'Housing_type', 'Employed_Status']
for feature in categorical_features:
    plt.figure(figsize=(4,4))
    plt.pie(x=df[feature].value_counts(),normalize=True,labels=list(df[feature].
 ↪unique()),autopct='%1.1f%%',pctdistance=0.6,labeldistance=2.1)
    plt.title(f'Distribution of {feature}')
    plt.show()
```

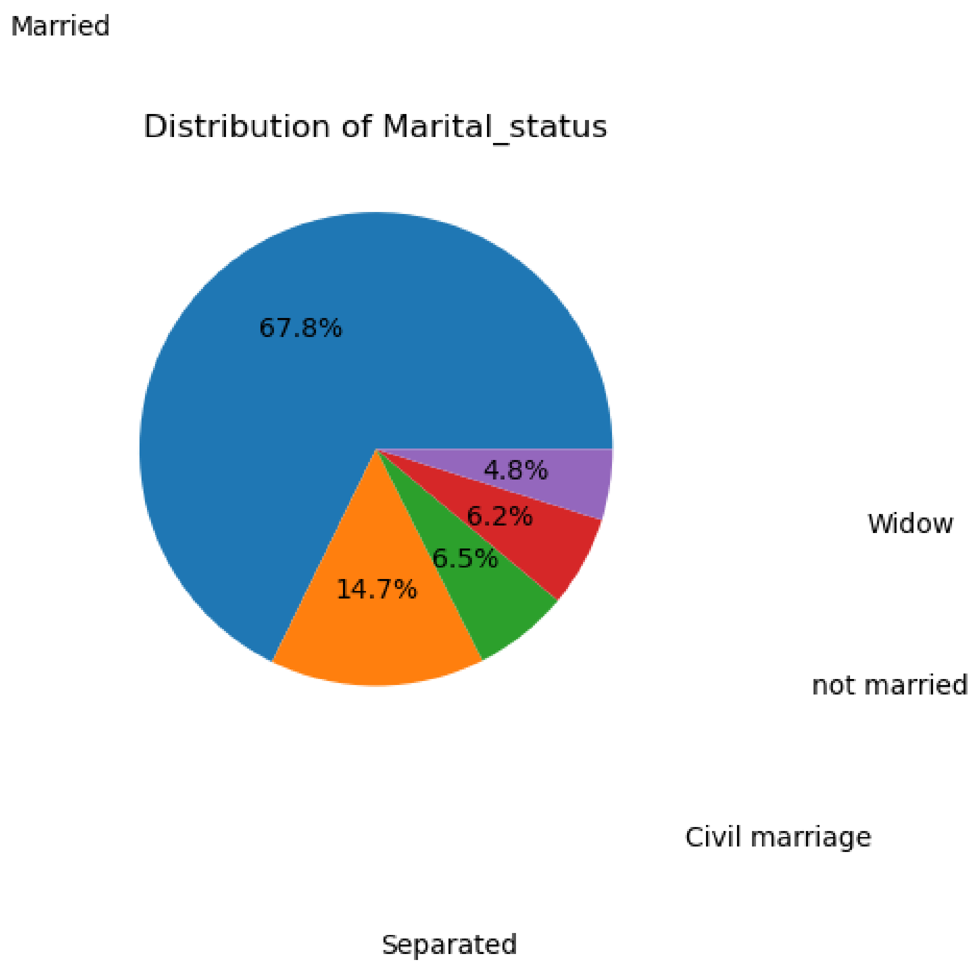

M
Distribution of GENDER
63.3%
36.7%
F

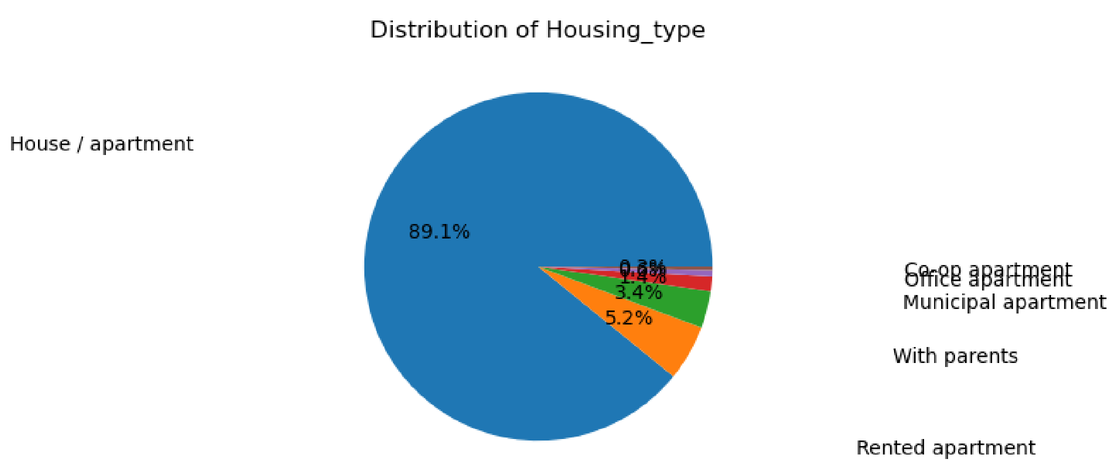

Y
Distribution of Car_Owner
59.7%
40.3%
N

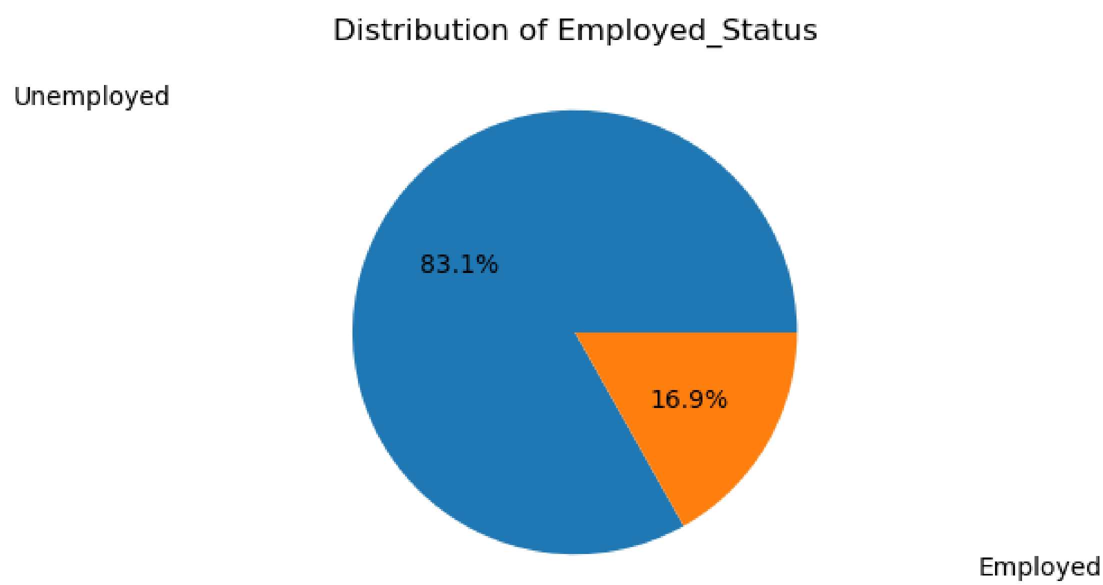

Y
Distribution of Propert_Owner
65.2%
34.8%
N


Pensioner
Distribution of Type_Income
51.6%
7.5%
23.6%
17.4%
State servant
Working
Commercial associate


Higher Education
Distribution of EDUCATION
66.6%
0.1%
1.4%
4.4%
27.5%
Academic Degree
Lower Secondary
Incomplete Higher
Secondary Education


Married
Distribution of Marital_status
67.8%
4.8%
6.2%
6.5%
14.7%
Widow
not married
Civil marriage
Separated


Distribution of Housing_type
House / apartment
89.1%
1.4%
3.4%
5.2%
Co-op apartment
Office apartment
Municipal apartment
With parents
Rented apartment



[35]:
```
df=df.drop(['Ind_ID'],axis=1)
```

[36]:
```
df.columns
```

[36]:
```
Index(['GENDER', 'Car_Owner', 'Propert_Owner', 'CHILDREN', 'Annual_income',
       'Type_Income', 'EDUCATION', 'Marital_status', 'Housing_type',
       'Family_Members', 'Approved_status', 'Age', 'Employed_Status'],
      dtype='object')
```

[37]:
```
categorical_column
```

[37]:
```
Index(['GENDER', 'Car_Owner', 'Propert_Owner', 'Type_Income', 'EDUCATION',
       'Marital_status', 'Housing_type', 'Employed_Status'],
      dtype='object')
```

1.9 Dummy Encoding

[38]:
```
categorical_column
```

[38]:
```
Index(['GENDER', 'Car_Owner', 'Propert_Owner', 'Type_Income', 'EDUCATION',
       'Marital_status', 'Housing_type', 'Employed_Status'],
      dtype='object')
```

[39]:
```
XX=pd.get_dummies(df,columns=categorical_column,drop_first=True)
```

1.10 Seperate Independent Variable X and Dependent Variable Y

[40]:
```
X=XX.drop('Approved_status',axis=1)
y=df['Approved_status']
```

1.11 Split the data in training and testing sets

[41]:
```
from sklearn.model_selection import train_test_split
```

[42]:
```
X_train,X_test,y_train,y_test=train_test_split(X,y,train_size=0.
↪8,random_state=17)
```

[43]:
```
print(len(X_train))
print(len(X_test))
```

```
1238
310
```

1.12 Feature Scaling

[44]:
```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
X_train=sc.fit_transform(X_train)
x_test=sc.transform(X_test)
```

[45]:
```
print(len(X_train))
print(len((X_test)))
```

```
1238
310
```

1.13 Key Findings

1.13.1 Income Type(Working)

- **Customers with a 'Working' Income is more than other Income types and their applications approvals are more than other type .**

1.13.2 Marital Status(Married)

- **Married customers are more than other type of Marital status and their applications approval are more than other marital status types.**

1.13.3 Housing Type(House/Apartment)

- **Most of the customers live in House/Apartment and their application approvals are more than other housing types.**

1.14 Final Conclusion

- **When considering all three types together, individual who are working, married and live in House/appartment type there is a higher probability of having their application approved.**

```
[ ]:
```