Credit EDA Case Study

Group Members:
Gunavina Mehta
Kriti Rai Saini

CASE STUDY MOTIVE

1. To Understand and apply EDA to real business scenario
2. Develop a basic understanding of risk analytics in financial sector

PROBLEM STATEMENT

When a company receives a loan application, the company has to decide the loan approval on the basis of applicant's profile.

<u>**Risks involved**</u> -

1. **Credit Loss** - If the applicant is not likely to repay the loan - he/she is likely to default
2. **Interest Loss** - If the applicant is likely to repay the loan, then not approving the loan results in the loss of business.

DATASETS GIVEN

1. Application_data - Information of the client at the time of application and whether he/she has payment difficulties.
2. Previous_application - Information about the client's previous data, whether the previous loan was approved/refused/cancelled or unused offer.
3. Columns_description - Description of the attributes in the above data.

BUSINESS OBJECTIVE

To minimize the risk factor by identifying the default client patterns and taking actions accordingly -

a. Denying the loan
b. Increasing Interest Rate
c. Reducing the loan amount

Hence improving it's knowledge on **<u>portfolio and risk assessment.</u>**

WORKFLOW -

I. Importing the csv files
II. Schema Check - shape/info
III. Data Quality Check for missing values and outliers
IV. Analysis

IMPORTING THE MODULES & FILES

```python
import warnings

warnings.filterwarnings('ignore')
```

```python
#importing the required modules
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots
```

```python
#Reading the previous application dataframe
prev_data_df = pd.read_csv('previous_application.csv')
prev_data_df.head()
```

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEK
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0	17145.0	
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN	607500.0	
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN	112500.0	
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN	450000.0	
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN	337500.0	

5 rows × 37 columns

```python
#reading application_data.csv
application_data_df = pd.read_csv('application_data.csv')
application_data_df.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CRED
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000

5 rows × 122 columns

SCHEMA CHECK

Application Data

```
#Dataframe shape
application_data_df.shape
```

```
(307511, 122)
```

```
#Dataframe description
application_data_df.info(verbose = True)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 122 columns):
 #   Column                        Dtype
---  ------                        -----
 0   SK_ID_CURR                    int64
 1   TARGET                        int64
 2   NAME_CONTRACT_TYPE            object
 3   CODE_GENDER                   object
 4   FLAG_OWN_CAR                  object
 5   FLAG_OWN_REALTY               object
 6   CNT_CHILDREN                  int64
 7   AMT_INCOME_TOTAL              float64
 8   AMT_CREDIT                    float64
 9   AMT_ANNUITY                   float64
 10  AMT_GOODS_PRICE               float64
 11  NAME_TYPE_SUITE               object
 12  NAME_INCOME_TYPE              object
 13  NAME_EDUCATION_TYPE           object
```

```
#Descriptive statistics for numeric columns
```

Previous Application Data

```
#Dataframe shape
prev_data_df.shape
```

```
(1670214, 37)
```

```
#Dataframe info
prev_data_df.info(verbose = True)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
 #   Column                       Non-Null Count    Dtype
---  ------                       --------------    -----
 0   SK_ID_PREV                   1670214 non-null  int64
 1   SK_ID_CURR                   1670214 non-null  int64
 2   NAME_CONTRACT_TYPE           1670214 non-null  object
 3   AMT_ANNUITY                  1297979 non-null  float64
 4   AMT_APPLICATION              1670214 non-null  float64
 5   AMT_CREDIT                   1670213 non-null  float64
 6   AMT_DOWN_PAYMENT             774370 non-null   float64
 7   AMT_GOODS_PRICE              1284699 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START   1670214 non-null  object
 9   HOUR_APPR_PROCESS_START      1670214 non-null  int64
 10  FLAG_LAST_APPL_PER_CONTRACT  1670214 non-null  object
 11  NFLAG_LAST_APPL_IN_DAY       1670214 non-null  int64
 12  RATE_DOWN_PAYMENT            774370 non-null   float64
 13  RATE_INTEREST_PRIMARY        5951 non-null     float64
 14  RATE_INTEREST_PRIVILEGED     5951 non-null     float64
 15  NAME_CASH_LOAN_PURPOSE       1670214 non-null  object
 16  NAME_CONTRACT_STATUS         1670214 non-null  object
 17  DAYS_DECISION                1670214 non-null  int64
 18  NAME_PAYMENT_TYPE            1670214 non-null  object
 19  CODE_REJECT_REASON           1670214 non-null  object
```

ANALYSIS FLOW

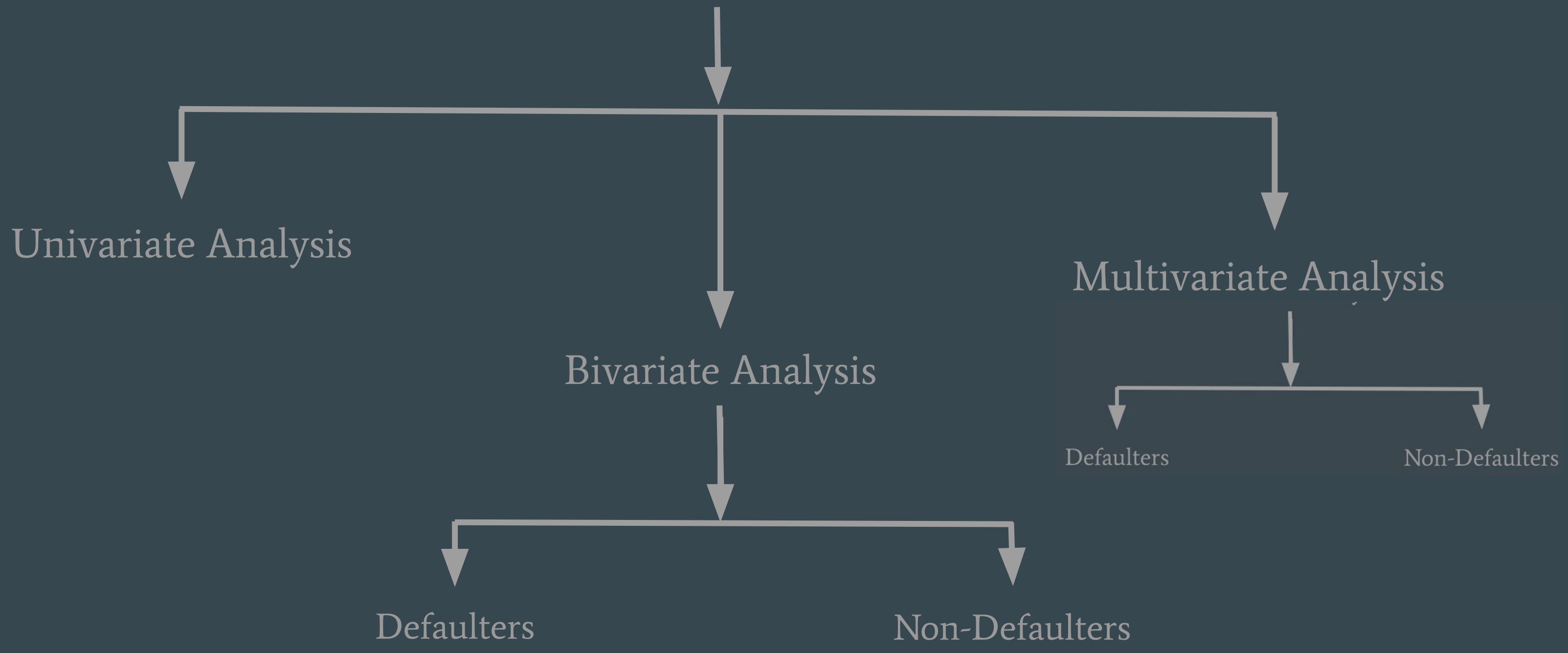
Application Data -

Contains info about the latest credit application

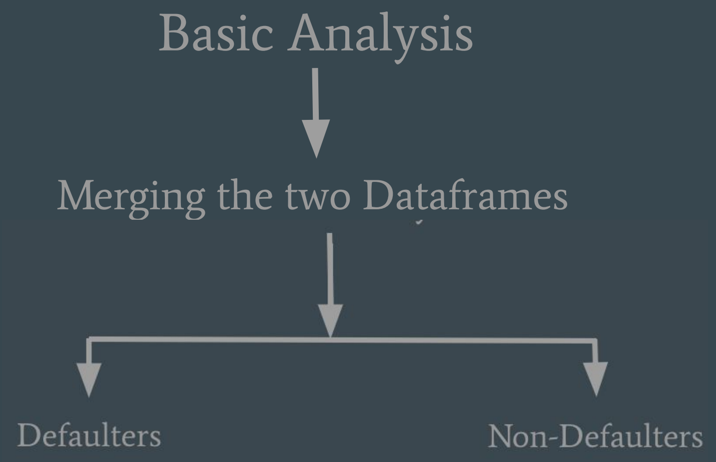
Previous Data -

Contains info about the previous credit application

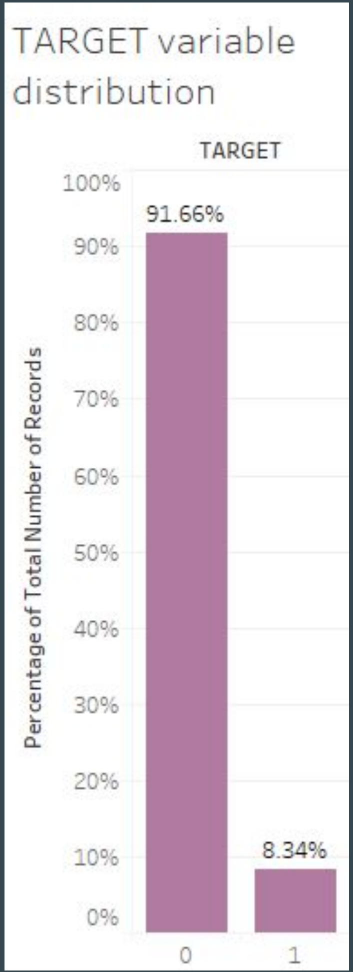
APPLICATION DATA



PREVIOUS APPLICATION DATA



THE FINAL REVELATIONS …



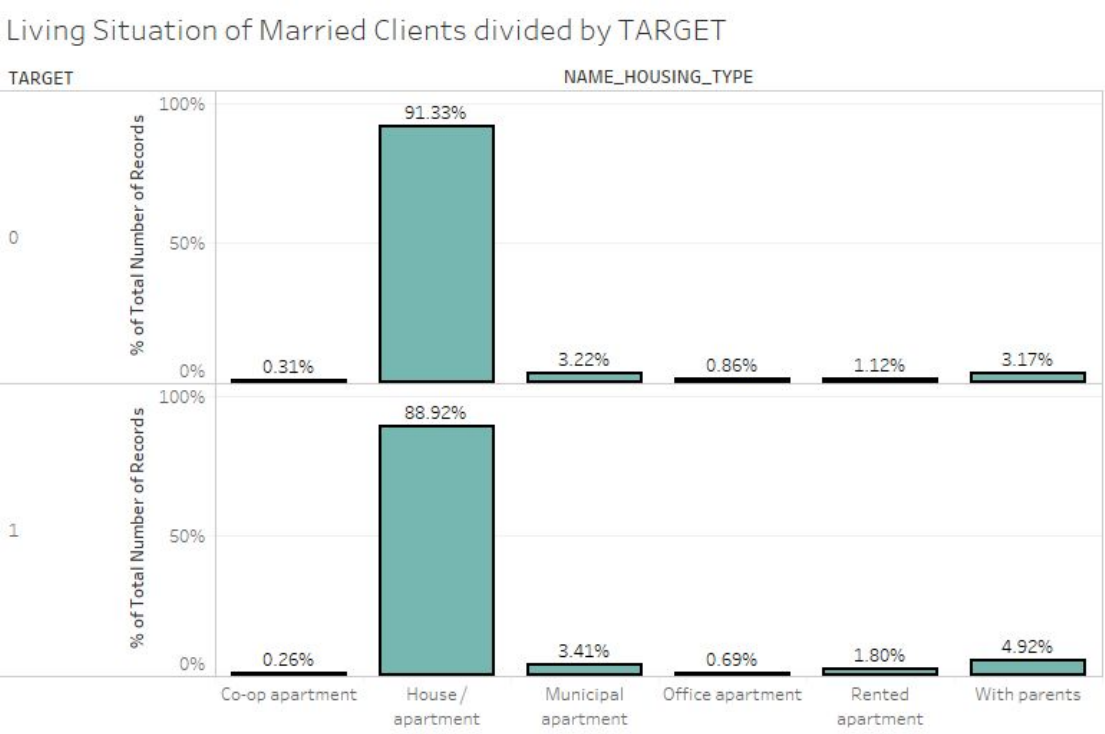
TARGET VARIABLE DISTRIBUTION

1. The data is highly imbalanced.
2. The defaulter percentage being only 8.34%.

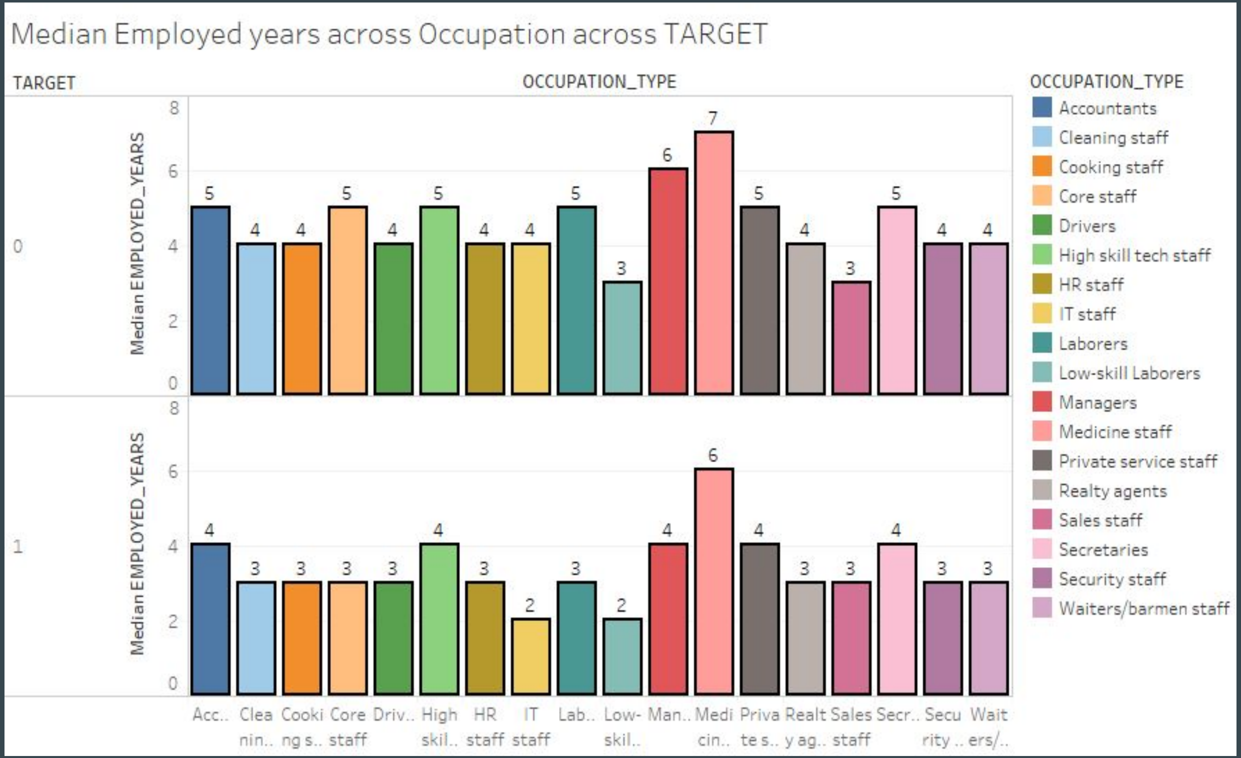
Therefore we will analyze in terms of **PERCENTAGES** as opposed to **ABSOLUTE NUMBERS**.

ANALYSIS FOR MARRIED CLIENTS

1. The dataset contains a high percentage of loan applications where the family status is married.
2. it was found that married clients who live with ***parents contribute to around 5%*** of all the married defaulters, even though their percentage in non defaulters is less.
3. The same was observed for ***Rented apartments.***
4. The bank can therefore scrutinize such loan applications and lay strict guidelines for married -with parents and married - rented apartment categories.



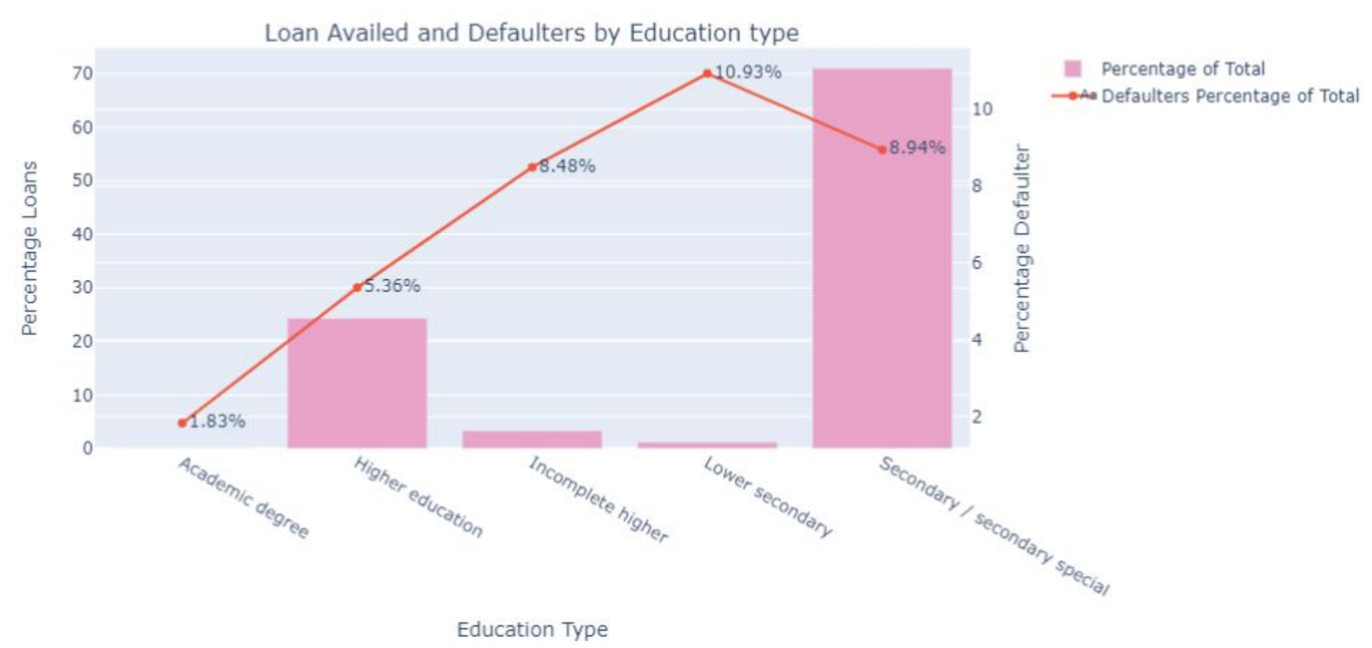
ANALYSIS BASED ON OCCUPATION/EMPLOYED YEARS



1. Across all the occupations, the **median employed years** for *defaulters* is **less than or equal** to for *non defaulters*.

2. This analysis reveals an important trend, This means that clients who have *just started out with their jobs* are **more** likely to **default.**

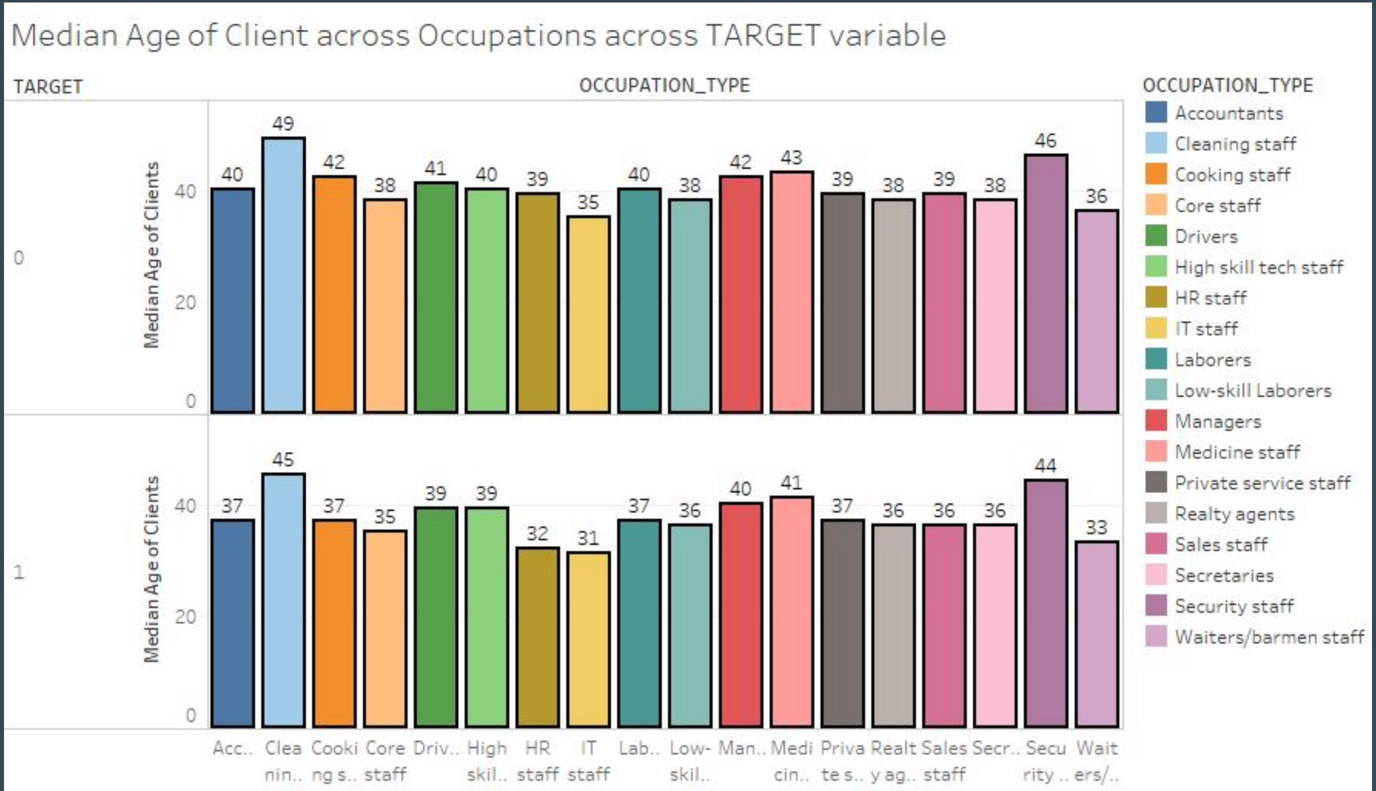
ANALYSIS BASED ON EDUCATION TYPE



- The pink bars represent what percent of the total data are contained in the respective categories.
- The red line represents what percent of application in the respective categories defaulted.

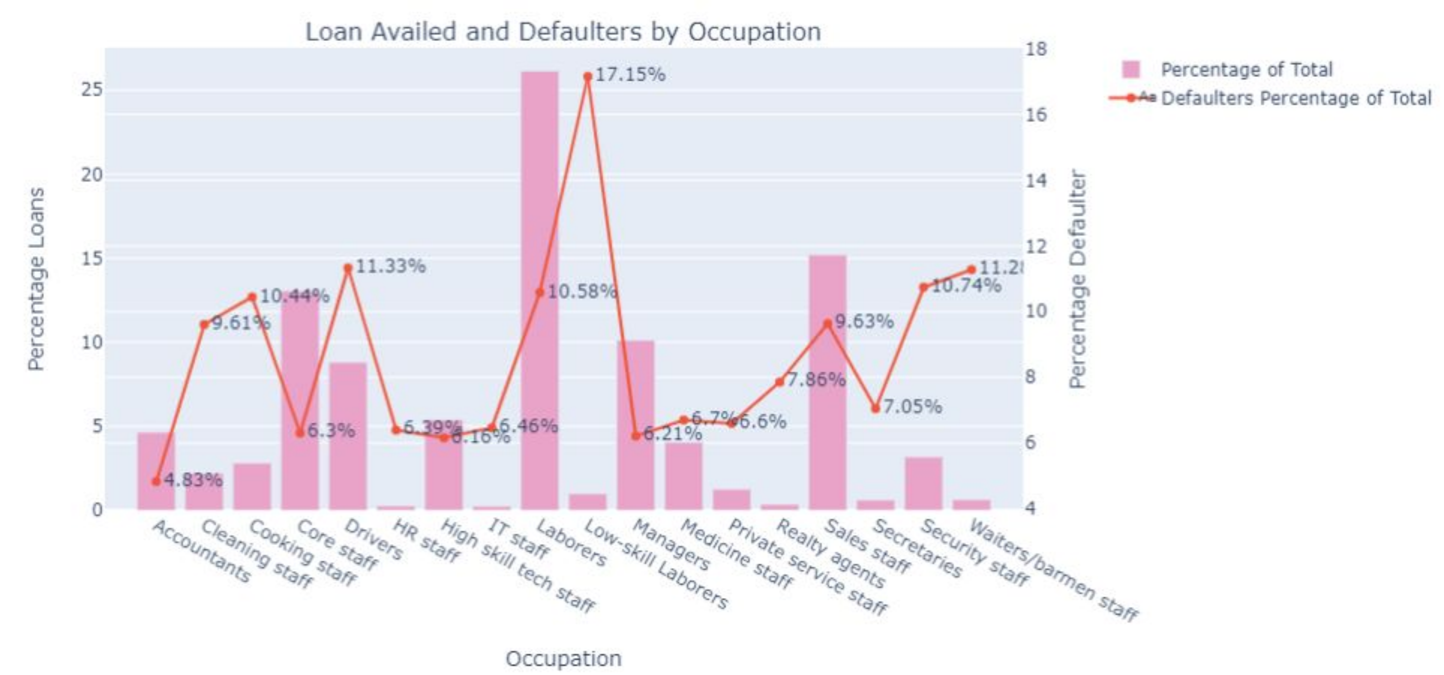
1. This analysis suggests that applications with education level ***lower secondary*** had the *highest default* percentage and their overall presence in the dataset was quite low.
2. The bank can make stringent rules for giving loans to such categories. Similar trend is observed for ‘Incomplete higher’ education category.

ANALYSIS BASED ON OCCUPATION/AGE IN YEARS



1. This analysis follows the trend in the previous analysis.
2. Across all the *occupations*, the median age in years for defaulters is ***less than or equal to for non defaulters***.
3. This means that **young clients** are more likely to default.

ANALYSIS BASED ON OCCUPATION

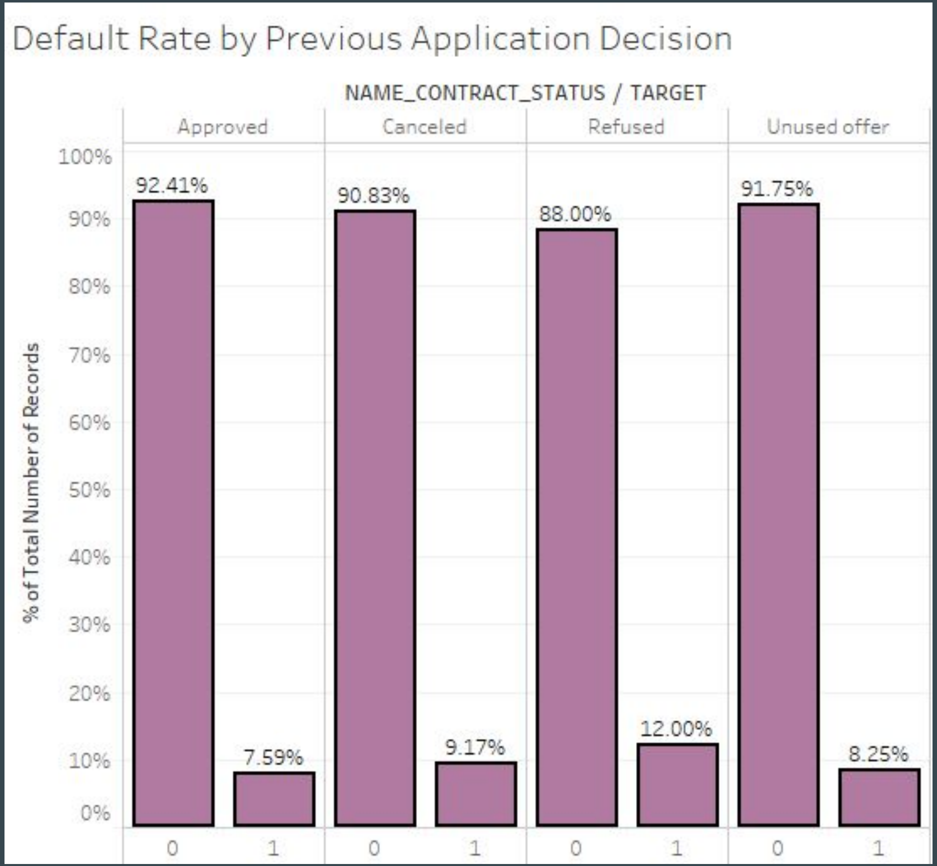


1. Laborers and ***Low Skill Laborers*** have a ***huge default percentage***.

2. The bank should be selective in giving loans clients with these occupations.

- The pink bars represent what percent of the total data are contained in the respective categories.
- The red line represents what percent of application in the respective categories defaulted.

ANALYSIS BASED ON PREVIOUS APPLICATION DATA



1. This analysis suggests that the ***applications that were rejected*** **the previous time** had the most ***default percentage***.
2. Therefore the bank should avoid giving loans for applications that have been rejected.
3. Instead they should either reduce the loan amount or increase the interest rate if they have to give out loans.

ANALYSIS BASED ON SELLER INDUSTRY(PREVIOUS DATA)

1. Loans in ***Auto technology*** sector had the highest default percentage followed by Connectivity industry.

2. The bank should be selective in lending in such sectors.

- The pink bars represent what percent of the total data are contained in the respective categories.
- The red line represents what percent of application in the respective categories defaulted.

SUMMARY -

1. Applications is scenarios like
   a. Married with Parents
   b. Rented Apartments
   c. Less employment years
   d. Education less that lower secondary
   e. Low Skill Labourers
   f. Previous application is rejected
   g. From Auto Technology industry

Need to be scrutinized properly as their default percentages are high through analysis.