Machine Learning Project Proposal on Credit Card Applications

By

Yogini Sanmugarajah

- **Dataset: Credit Card Approval Prediction**

- **Source: https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction?select=application_record.csv**

- There are two sets of data
  - Application record
  - Credit record

Application record

ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	DAYS_BIRTH	DAYS_EMPLOYED	FLAG_MOBIL	FLAG_WORK_PHONE	FLAG_PHONE	FLAG_EMAIL	OCCUPATION_TYPE	CNT_FAM_MEMBERS
5008804	M	Y	Y	0	427500	Working	Higher education	Civil marriage	Rented apartment	-12005	-4542	1	1	0	0		2
5008805	M	Y	Y	0	427500	Working	Higher education	Civil marriage	Rented apartment	-12005	-4542	1	1	0	0		2
5008806	M	Y	Y	0	112500	Working	Secondary / secondary specia	Married	House / apartment	-21474	-1134	1	0	0	0	Security staff	2
5008808	F	N	Y	0	270000	Commercial associate	Secondary / secondary specia	Single / not married	House / apartment	-19110	-3051	1	0	1	1	Sales staff	1
5008809	F	N	Y	0	270000	Commercial associate	Secondary / secondary specia	Single / not married	House / apartment	-19110	-3051	1	0	1	1	Sales staff	1
5008810	F	N	Y	0	270000	Commercial associate	Secondary / secondary specia	Single / not married	House / apartment	-19110	-3051	1	0	1	1	Sales staff	1
5008811	F	N	Y	0	270000	Commercial associate	Secondary / secondary specia	Single / not married	House / apartment	-19110	-3051	1	0	1	1	Sales staff	1
5008812	F	N	Y	0	283500	Pensioner	Higher education	Separated	House / apartment	-22464	365243	1	0	0	0		1
5008813	F	N	Y	0	283500	Pensioner	Higher education	Separated	House / apartment	-22464	365243	1	0	0	0		1
5008814	F	N	Y	0	283500	Pensioner	Higher education	Separated	House / apartment	-22464	365243	1	0	0	0		1
5008815	M	Y	Y	0	270000	Working	Higher education	Married	House / apartment	-16872	-769	1	1	1	1	Accountants	2
5112956	M	Y	Y	0	270000	Working	Higher education	Married	House / apartment	-16872	-769	1	1	1	1	Accountants	2
6153651	M	Y	Y	0	270000	Working	Higher education	Married	House / apartment	-16872	-769	1	1	1	1	Accountants	2
5008819	M	Y	Y	0	135000	Commercial associate	Secondary / secondary specia	Married	House / apartment	-17778	-1194	1	0	0	0	Laborers	2
5008820	M	Y	Y	0	135000	Commercial associate	Secondary / secondary specia	Married	House / apartment	-17778	-1194	1	0	0	0	Laborers	2
5008821	M	Y	Y	0	135000	Commercial associate	Secondary / secondary specia	Married	House / apartment	-17778	-1194	1	0	0	0	Laborers	2
5008822	M	Y	Y	0	135000	Commercial associate	Secondary / secondary specia	Married	House / apartment	-17778	-1194	1	0	0	0	Laborers	2
5008823	M	Y	Y	0	135000	Commercial associate	Secondary / secondary specia	Married	House / apartment	-17778	-1194	1	0	0	0	Laborers	2
5008824	M	Y	Y	0	135000	Commercial associate	Secondary / secondary specia	Married	House / apartment	-17778	-1194	1	0	0	0	Laborers	2
5008825	F	Y	N	0	130500	Working	Incomplete higher	Married	House / apartment	-10669	-1103	1	0	0	0	Accountants	2
5008826	F	Y	N	0	130500	Working	Incomplete higher	Married	House / apartment	-10669	-1103	1	0	0	0	Accountants	2
5008830	F	N	Y	0	157500	Working	Secondary / secondary specia	Married	House / apartment	-10031	-1469	1	0	1	0	Laborers	2
5008831	F	N	Y	0	157500	Working	Secondary / secondary specia	Married	House / apartment	-10031	-1469	1	0	1	0	Laborers	2
5008832	F	N	Y	0	157500	Working	Secondary / secondary specia	Married	House / apartment	-10031	-1469	1	0	1	0	Laborers	2
5008834	F	N	Y	1	112500	Working	Secondary / secondary specia	Single / not married	House / apartment	-10968	-1620	1	0	0	0		2
5008835	F	N	Y	1	112500	Working	Secondary / secondary specia	Single / not married	House / apartment	-10968	-1620	1	0	0	0		2
6153712	F	N	Y	1	112500	Working	Secondary / secondary specia	Single / not married	House / apartment	-10968	-1620	1	0	0	0		2
5008836	M	Y	Y	3	270000	Working	Secondary / secondary specia	Married	House / apartment	-12689	-1163	1	0	0	0	Laborers	5
5008837	M	Y	Y	3	270000	Working	Secondary / secondary specia	Married	House / apartment	-12689	-1163	1	0	0	0	Laborers	5
5008838	M	N	Y	1	405000	Commercial associate	Higher education	Married	House / apartment	-11842	-2016	1	0	0	0	Managers	3
5008839	M	N	Y	1	405000	Commercial associate	Higher education	Married	House / apartment	-11842	-2016	1	0	0	0	Managers	3
5008840	M	N	Y	1	405000	Commercial associate	Higher education	Married	House / apartment	-11842	-2016	1	0	0	0	Managers	3
5008841	M	N	Y	1	405000	Commercial associate	Higher education	Married	House / apartment	-11842	-2016	1	0	0	0	Managers	3
5008842	M	N	Y	1	405000	Commercial associate	Higher education	Married	House / apartment	-11842	-2016	1	0	0	0	Managers	3
5008843	M	N	Y	1	405000	Commercial associate	Higher education	Married	House / apartment	-11842	-2016	1	0	0	0	Managers	3
5008844	M	Y	Y	0	112500	Commercial associate	Secondary / secondary specia	Married	House / apartment	-20502	-4450	1	0	1	0	Drivers	2
5008846	M	Y	Y	0	112500	Commercial associate	Secondary / secondary specia	Married	House / apartment	-20502	-4450	1	0	1	0	Drivers	2
5008847	M	Y	Y	0	112500	Commercial associate	Secondary / secondary specia	Married	House / apartment	-20502	-4450	1	0	1	0	Drivers	2
5008849	M	Y	Y	0	112500	Commercial associate	Secondary / secondary specia	Married	House / apartment	-20502	-4450	1	0	1	0	Drivers	2

Application dataset first 10 records

```python
df_application.head(10)
```

	ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	N
0	5008804	M	Y	Y	0	427500.0	Working	Higher education	
1	5008805	M	Y	Y	0	427500.0	Working	Higher education	
2	5008806	M	Y	Y	0	112500.0	Working	Secondary / secondary special	
3	5008808	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	
4	5008809	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	
5	5008810	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	
6	5008811	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	
7	5008812	F	N	Y	0	283500.0	Pensioner	Higher education	
8	5008813	F	N	Y	0	283500.0	Pensioner	Higher education	
9	5008814	F	N	Y	0	283500.0	Pensioner	Higher education	

No of Rows & Columns

```python
df_application.shape
```

```
(438557, 18)
```

- Column Summary

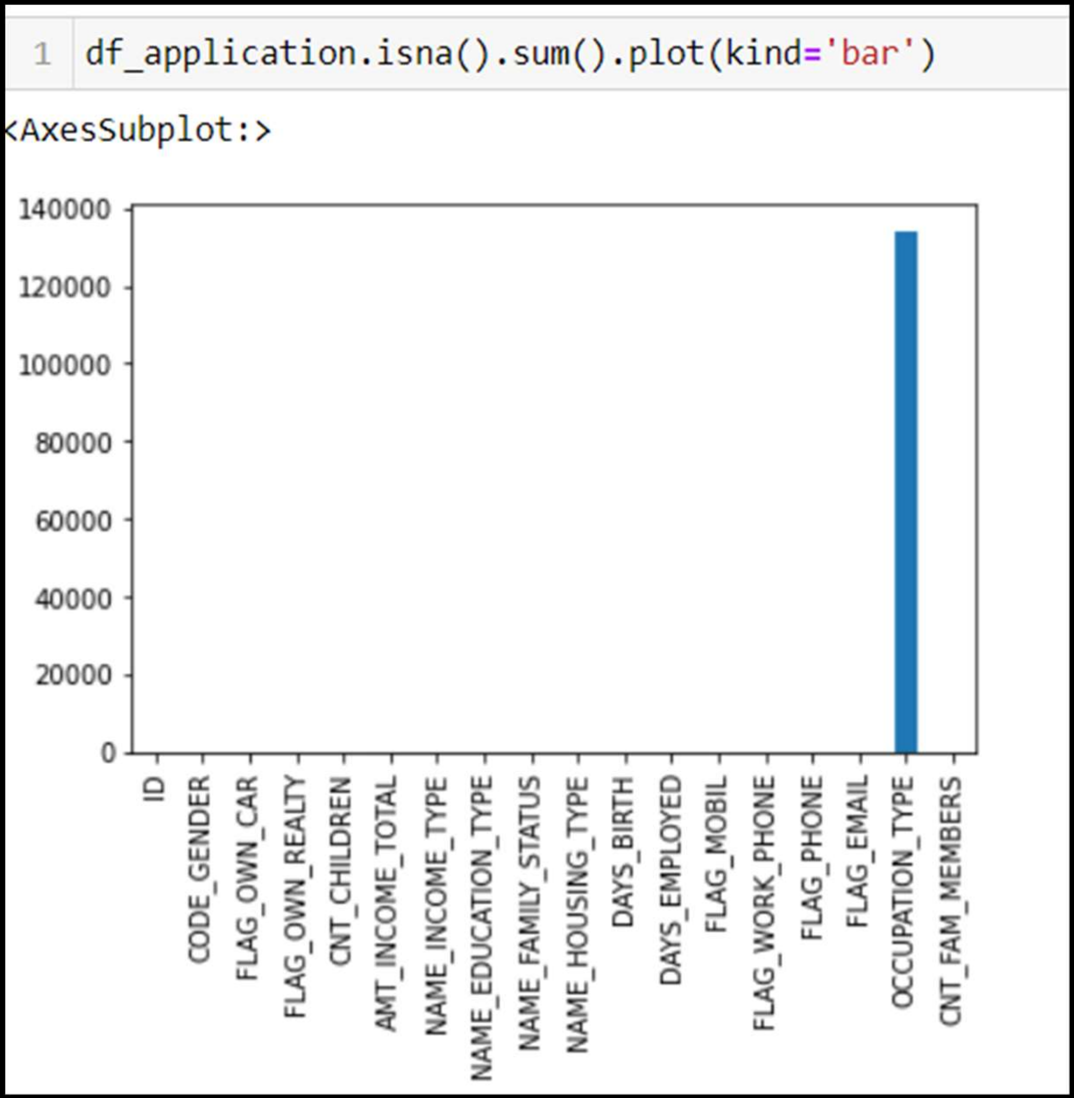
application_record.csv		
Feature name	Explanation	Remarks
ID	Client number	
CODE_GENDER	Gender	
FLAG_OWN_CAR	Is there a car	
FLAG_OWN_REALTY	Is there a property	
CNT_CHILDREN	Number of children	
AMT_INCOME_TOTAL	Annual income	
NAME_INCOME_TYPE	Income category	
NAME_EDUCATION_TYPE	Education level	
NAME_FAMILY_STATUS	Marital status	
NAME_HOUSING_TYPE	Way of living	
DAYS_BIRTH	Birthday	Count backwards from current day (0), -1 means yesterday
DAYS_EMPLOYED	Start date of employment	Count backwards from current day(0). If positive, it means the person currently unemployed.
FLAG_MOBIL	Is there a mobile phone	
FLAG_WORK_PHONE	Is there a work phone	
FLAG_PHONE	Is there a phone	
FLAG_EMAIL	Is there an email	
OCCUPATION_TYPE	Occupation	
CNT_FAM_MEMBERS	Family size	

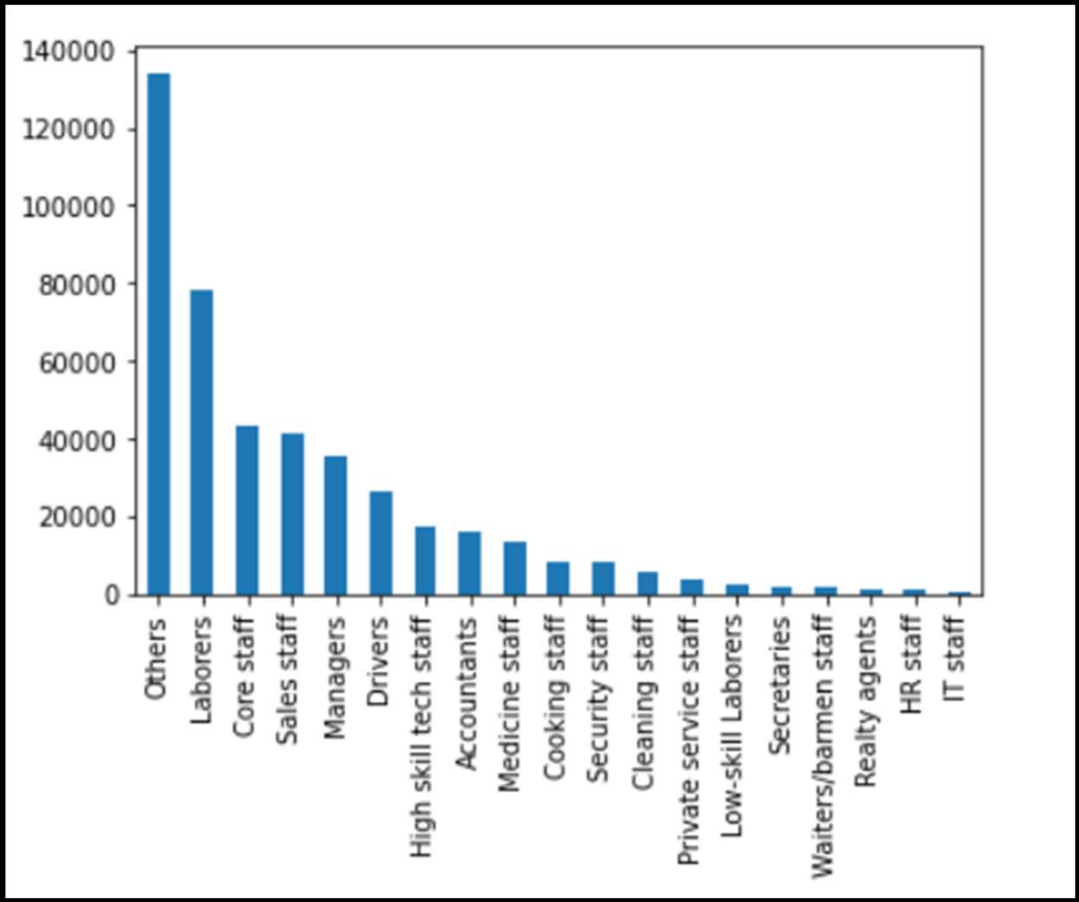
```
1 df_application.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 438557 entries, 0 to 438556
Data columns (total 18 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   ID                   438557 non-null  int64
 1   CODE_GENDER          438557 non-null  object
 2   FLAG_OWN_CAR         438557 non-null  object
 3   FLAG_OWN_REALTY      438557 non-null  object
 4   CNT_CHILDREN         438557 non-null  int64
 5   AMT_INCOME_TOTAL     438557 non-null  float64
 6   NAME_INCOME_TYPE     438557 non-null  object
 7   NAME_EDUCATION_TYPE  438557 non-null  object
 8   NAME_FAMILY_STATUS   438557 non-null  object
 9   NAME_HOUSING_TYPE    438557 non-null  object
 10  DAYS_BIRTH           438557 non-null  int64
 11  DAYS_EMPLOYED        438557 non-null  int64
 12  FLAG_MOBIL           438557 non-null  int64
 13  FLAG_WORK_PHONE      438557 non-null  int64
 14  FLAG_PHONE           438557 non-null  int64
 15  FLAG_EMAIL           438557 non-null  int64
 16  OCCUPATION_TYPE      304354 non-null  object
 17  CNT_FAM_MEMBERS      438557 non-null  float64
dtypes: float64(2), int64(8), object(8)
memory usage: 60.2+ MB
```

- Column Information

- Missing Data Columns





Handling missing data in the occupation column

- create a new type called others

Credit Record Dataset

ID	MONTHS_	STATUS
5001711	0	X
5001711	-1	0
5001711	-2	0
5001711	-3	0
5001712	0	C
5001712	-1	C
5001712	-2	C
5001712	-3	C
5001712	-4	C
5001712	-5	C
5001712	-6	C
5001712	-7	C
5001712	-8	C
5001712	-9	0
5001712	-10	0
5001712	-11	0
5001712	-12	0
5001712	-13	0
5001712	-14	0
5001712	-15	0
5001712	-16	0
5001712	-17	0
5001712	-18	0
5001713	0	X
5001713	-1	X
5001713	-2	X
5001713	-3	X

- Credit dataset first 10 record

```
1 df_credit.head(10)
```

	ID	MONTHS_BALANCE	STATUS
0	5001711	0	X
1	5001711	-1	0
2	5001711	-2	0
3	5001711	-3	0
4	5001712	0	C
5	5001712	-1	C
6	5001712	-2	C
7	5001712	-3	C
8	5001712	-4	C
9	5001712	-5	C

Column Summary

credit_record.csv		
Feature name	Explanation	Remarks
ID	Client number	
MONTHS_BALANCE	Record month	The month of the extracted data is the starting point, backwards, 0 is the current month, -1 is the previous month, and so on
STATUS	Status	0: 1-29 days past due 1: 30-59 days past due 2: 60-89 days overdue 3: 90-119 days overdue 4: 120-149 days overdue 5: Overdue or bad debts, write-offs for more than 150 days C: paid off that month X: No loan for the month

No of Rows and Columns

```
1  df_credit.shape
```

```
(1048575, 3)
```

Column Information

```
1  df_credit.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 3 columns):
 #   Column          Non-Null Count    Dtype 
---  ------          --------------    ----- 
 0   ID              1048575 non-null  int64 
 1   MONTHS_BALANCE  1048575 non-null  int64 
 2   STATUS          1048575 non-null  object
dtypes: int64(2), object(1)
memory usage: 24.0+ MB
```

- Missing/null values

```
df_credit.isna().sum()
```

```
ID                0
MONTHS_BALANCE    0
STATUS            0
dtype: int64
```

Challenges

- Encoding categorical data
- Interpreting missing data in the occupation column
- Label column “status”- deciding whether to consider one-time nonpayment or multiple nonpayments as the criteria to define the customer is good or bad
- Imbalance dataset in term of gender in application data set

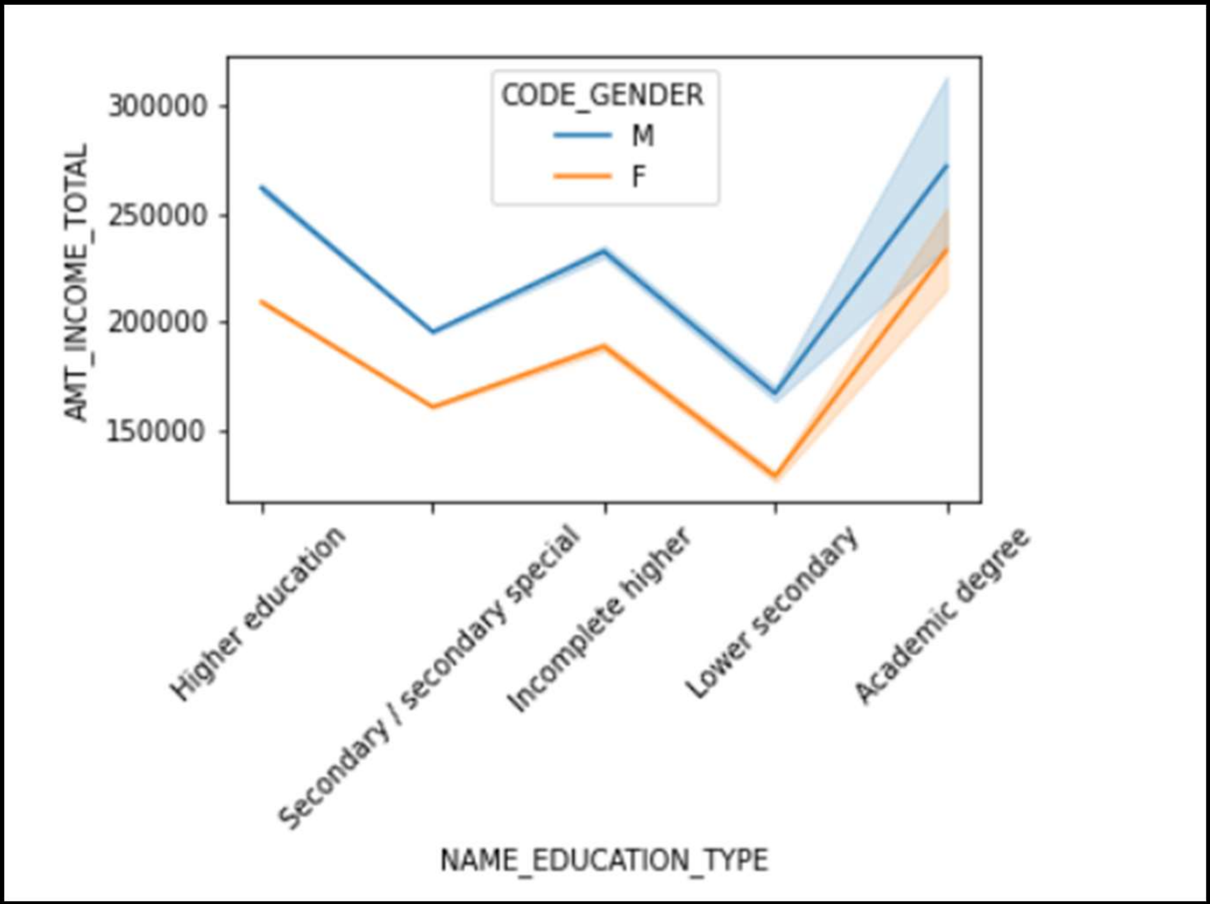
Application Record Dataset- Insides

- Drop duplicates

```
df_application.drop_duplicates(subset = 'ID')
```

	ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	NAME_INCOME_TYPE	NAME_EDUCATION_TYI
0	5008804	M	Y	Y	0	427500.0	Working	Higher educati
1	5008805	M	Y	Y	0	427500.0	Working	Higher educati
2	5008806	M	Y	Y	0	112500.0	Working	Secondary / second spec
3	5008808	F	N	Y	0	270000.0	Commercial associate	Secondary / second spec
4	5008809	F	N	Y	0	270000.0	Commercial associate	Secondary / second spec
...	...	...	...	...	...	...	...	
438552	6840104	M	N	Y	0	135000.0	Pensioner	Secondary / second spec
438553	6840222	F	N	N	0	103500.0	Working	Secondary / second spec
438554	6841878	F	N	N	0	54000.0	Commercial associate	Higher educati
438555	6842765	F	N	Y	0	72000.0	Pensioner	Secondary / second spec
438556	6842885	F	N	Y	0	121500.0	Working	Secondary / second spec

438510 rows × 18 columns

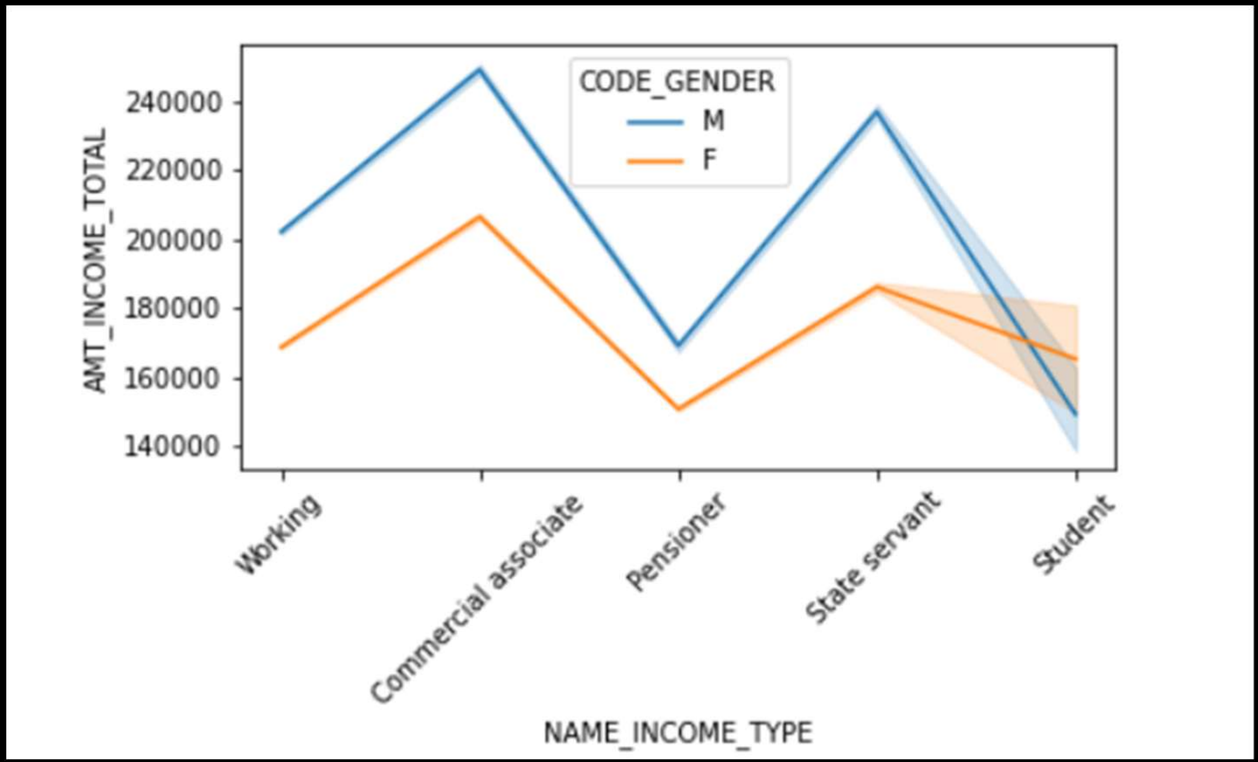


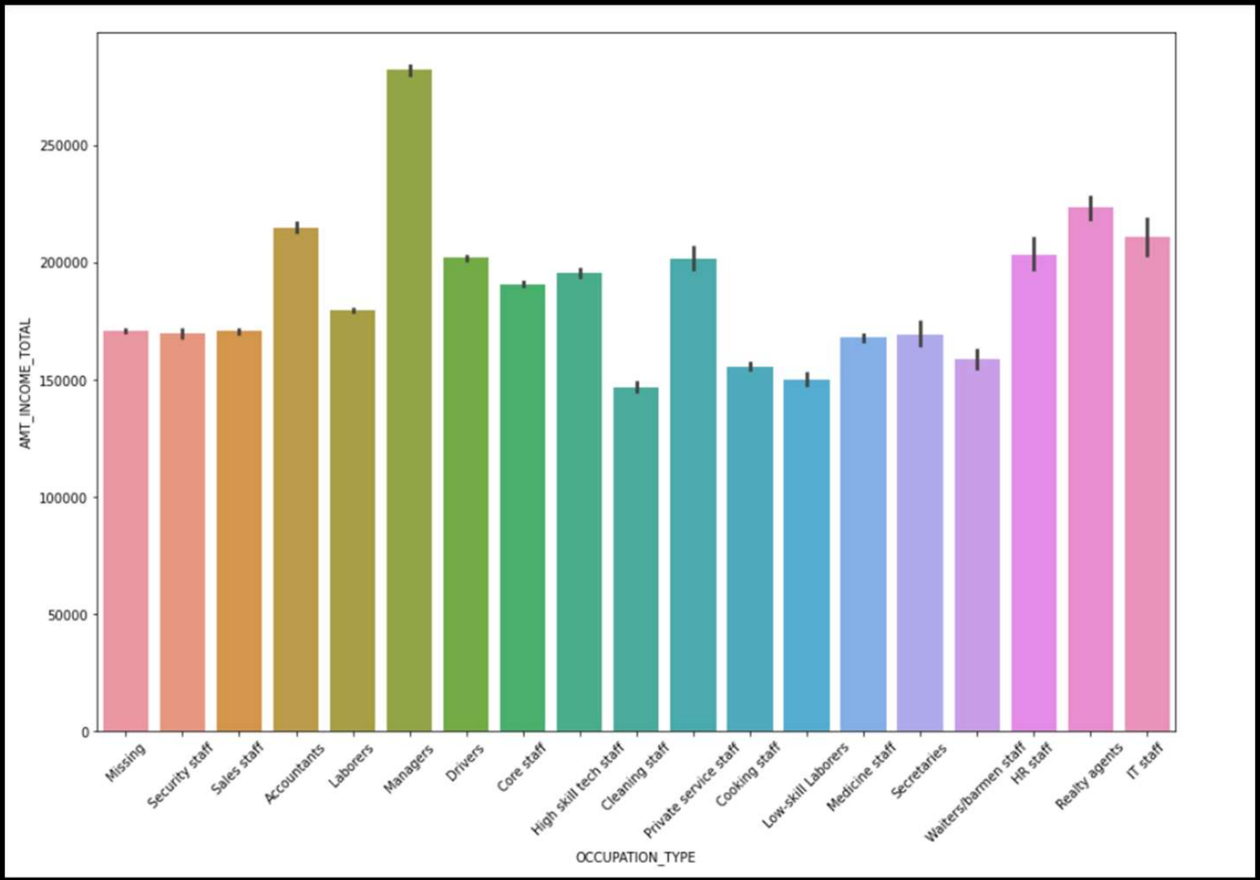
Education vs Income Observations

- Higher education and Academic degree holders earn more than the rest
- Males are earning more than females

Income Type vs Income observations

- Commercial associate income is higher than the rest
- Female income has a huge variance in State servant
- Female students are earning higher than male students. Is it due to not many male students working?



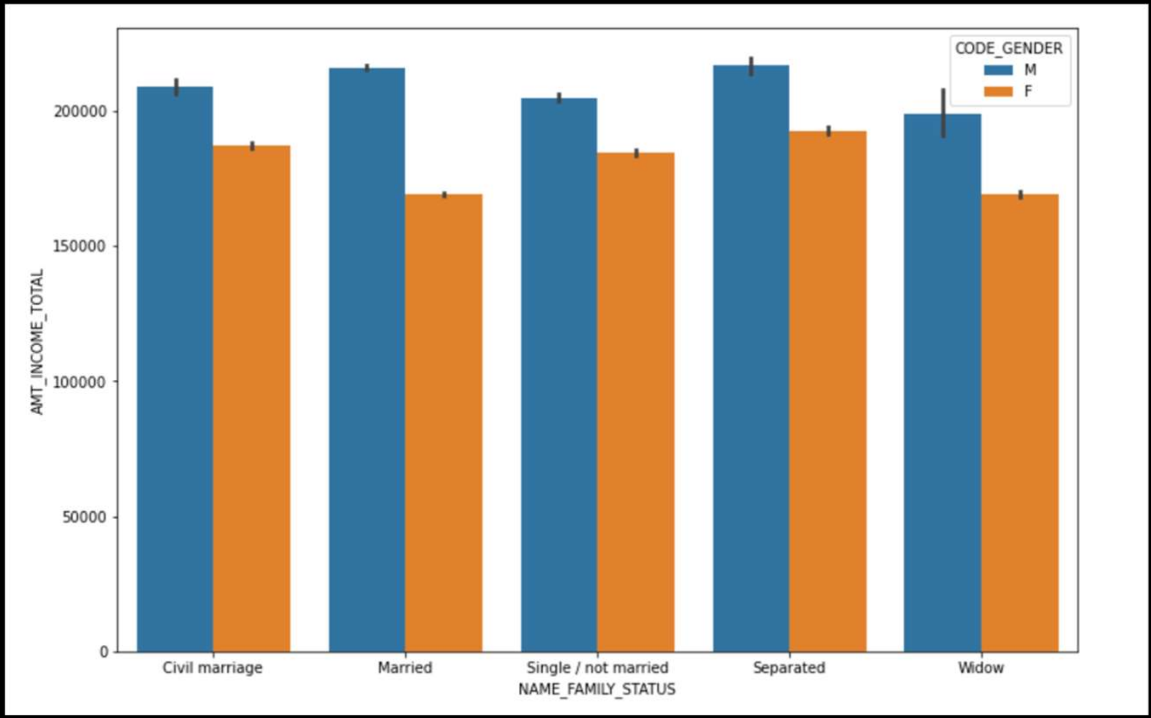


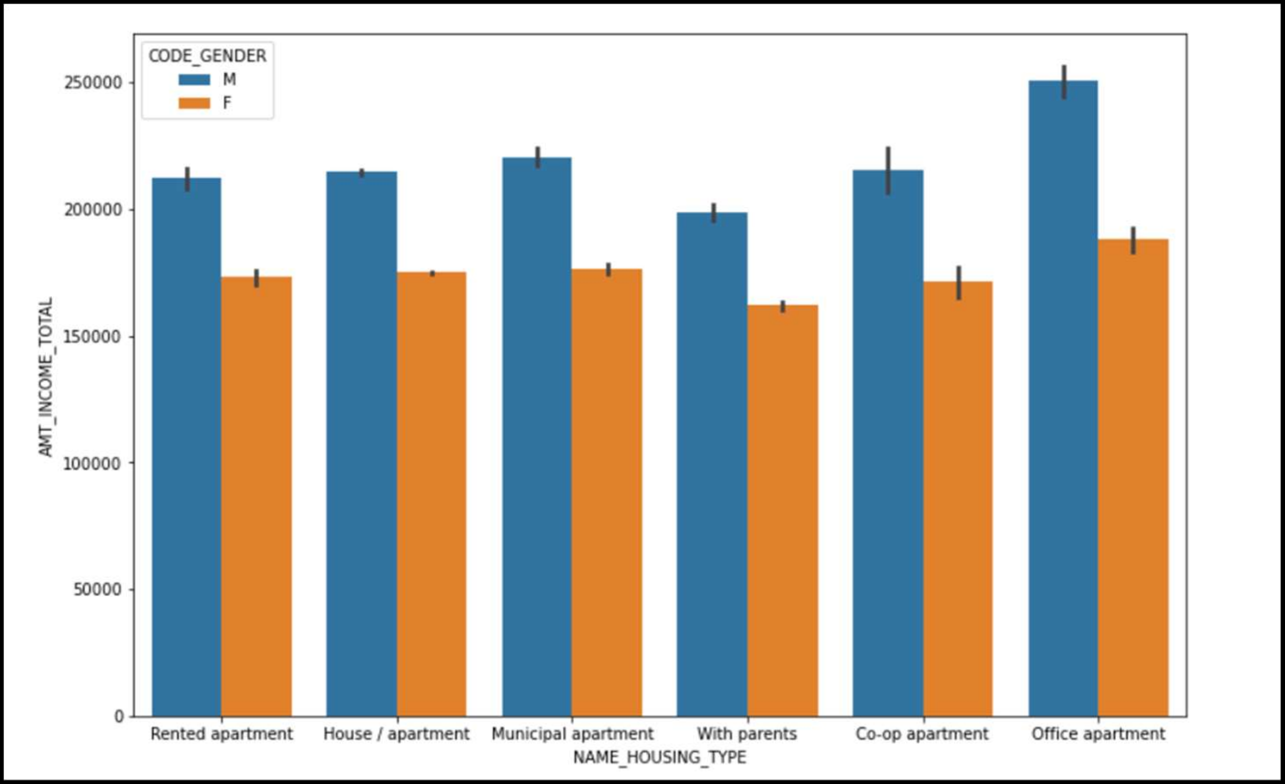
Income vs Occupation

- Managers' income level is much higher than the rest

Marital Status vs Income

- Males' income did not fluctuate much, but married females' income is much less compared to other family status





Housing Type vs Income

- Male staying in office apartment earn much higher than others

- Female and male with parent category income is lower compared to others

Model

Classification Model

- Reason – Outcome is fixed

Algorithms

- Logistic Regression R
- Naive Bayes

Problem Statements

How can we use the machine learning model to automate the credit card application rather than manual doing?

Prediction

An applicant is a good or bad customer?