Virtual Internship Experience at ID/X Partners

Loan Data Exploratory Data Analysis

Diana Eka Riyani

Tabel of Contents

Data Requirements

Data Requirements

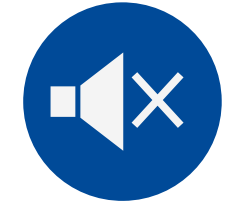Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

These files contain complete loan data for all loans issued through the 2007-2014, including the current loan status (Current, Charged-off, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter.

Data Collection

Data Collection

```
[4] df = pd.read_csv('/content/drive/My Drive/loan_data_2007_2014.csv')
    df
```

	Unnamed: 0	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	...	total_bal_il	il_util	open_rv_12m	open_rv_24m	max_bal_bc	all_util	total_rev_hi_li
0	0	1077501	1296599	5000	5000	4975.0	36 months	10.65	162.87	B	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	1	1077430	1314167	2500	2500	2500.0	60 months	15.27	59.83	C	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2	1077175	1313524	2400	2400	2400.0	36 months	15.96	84.33	C	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	3	1076863	1277178	10000	10000	10000.0	36 months	13.49	339.31	C	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	4	1075358	1311748	3000	3000	3000.0	60 months	12.69	67.79	B	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
466280	466280	8598660	1440975	18400	18400	18400.0	60 months	14.47	432.64	C	...	NaN	NaN	NaN	NaN	NaN	NaN	29900.0
466281	466281	9684700	11536848	22000	22000	22000.0	60 months	19.97	582.50	D	...	NaN	NaN	NaN	NaN	NaN	NaN	39400.0
466282	466282	9584776	11436914	20700	20700	20700.0	60 months	16.99	514.34	D	...	NaN	NaN	NaN	NaN	NaN	NaN	13100.0
466283	466283	9604874	11457002	2000	2000	2000.0	36 months	7.90	62.59	A	...	NaN	NaN	NaN	NaN	NaN	NaN	53100.0
466284	466284	9199665	11061576	10000	10000	9975.0	36 months	19.20	367.58	D	...	NaN	NaN	NaN	NaN	NaN	NaN	16000.0

466285 rows × 75 columns

Data Understanding

Data Understanding

```
[5]  # preview the dataset

     df.head()
```

	Unnamed: 0	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	...	total_bal_il	il_util	open_rv_12m	open_rv_24m
0	0	1077501	1296599	5000	5000	4975.0	36 months	10.65	162.87	B	...	NaN	NaN	NaN	NaN
1	1	1077430	1314167	2500	2500	2500.0	60 months	15.27	59.83	C	...	NaN	NaN	NaN	NaN
2	2	1077175	1313524	2400	2400	2400.0	36 months	15.96	84.33	C	...	NaN	NaN	NaN	NaN
3	3	1076863	1277178	10000	10000	10000.0	36 months	13.49	339.31	C	...	NaN	NaN	NaN	NaN
4	4	1075358	1311748	3000	3000	3000.0	60 months	12.69	67.79	B	...	NaN	NaN	NaN	NaN

5 rows × 75 columns

```
▶  # view dimensions of dataset

   df.shape
```

```
(466285, 75)
```

Data Understanding

```
# view columns of dataset

df.columns

Index(['Unnamed: 0', 'id', 'member_id', 'loan_amnt', 'funded_amnt',
       'funded_amnt_inv', 'term', 'int_rate', 'installment', 'grade',
       'sub_grade', 'emp_title', 'emp_length', 'home_ownership', 'annual_inc',
       'verification_status', 'issue_d', 'loan_status', 'pymnt_plan', 'url',
       'desc', 'purpose', 'title', 'zip_code', 'addr_state', 'dti',
       'delinq_2yrs', 'earliest_cr_line', 'inq_last_6mths',
       'mths_since_last_delinq', 'mths_since_last_record', 'open_acc',
       'pub_rec', 'revol_bal', 'revol_util', 'total_acc',
       'initial_list_status', 'out_prncp', 'out_prncp_inv', 'total_pymnt',
       'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int',
       'total_rec_late_fee', 'recoveries', 'collection_recovery_fee',
       'last_pymnt_d', 'last_pymnt_amnt', 'next_pymnt_d', 'last_credit_pull_d',
       'collections_12_mths_ex_med', 'mths_since_last_major_derog',
       'policy_code', 'application_type', 'annual_inc_joint', 'dti_joint',
       'verification_status_joint', 'acc_now_delinq', 'tot_coll_amt',
       'tot_cur_bal', 'open_acc_6m', 'open_il_6m', 'open_il_12m',
       'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il', 'il_util',
       'open_rv_12m', 'open_rv_24m', 'max_bal_bc', 'all_util',
       'total_rev_hi_lim', 'inq_fi', 'total_cu_tl', 'inq_last_12m'],
      dtype='object')
```

```
# view summary of dataset

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 466285 entries, 0 to 466284
Data columns (total 75 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   Unnamed: 0       466285 non-null  int64
 1   id               466285 non-null  int64
 2   member_id        466285 non-null  int64
 3   loan_amnt        466285 non-null  int64
 4   funded_amnt      466285 non-null  int64
 5   funded_amnt_inv  466285 non-null  float64
 6   term             466285 non-null  object
 7   int_rate         466285 non-null  float64
 8   installment      466285 non-null  float64
 9   grade            466285 non-null  object
 10  sub_grade        466285 non-null  object
```
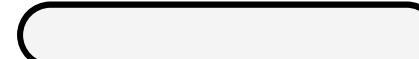
Data Preparation

Data Preparation

```
[ ] #cek data duplicated

    df.duplicated().sum()

    0


[ ] # cek missing value

    df.isnull().sum()

    Unnamed: 0            0
    id                   0
    member_id            0
    loan_amnt            0
    funded_amnt          0
                        ...
    all_util          466285
    total_rev_hi_lim   70276
    inq_fi            466285
    total_cu_tl       466285
    inq_last_12m      466285
    Length: 75, dtype: int64
```

```
[ ] # cek missing value after being columns with na's dropped

    loan_data = columns_with_na_dropped.isnull().sum()
    loan_data

    Unnamed: 0            0
    id                   0
    member_id            0
    loan_amnt            0
    funded_amnt          0
    funded_amnt_inv      0
    term                 0
    int_rate             0
    installment          0
    grade                0
    sub_grade            0
    home_ownership       0
    verification_status  0
    issue_d              0
    loan_status          0
    pymnt_plan           0
    url                  0
    purpose              0
    zip_code             0
    addr_state           0
    dti                  0
    revol_bal            0
    initial_list_status  0
    out_prncp            0
    out_prncp_inv        0
    total_pymnt          0
    total_pymnt_inv      0
    total_rec_prncp      0
    total_rec_int        0
    total_rec_late_fee   0
    recoveries           0
    collection_recovery_fee  0
    last_pymnt_amnt      0
    policy_code          0
    application_type     0
    dtype: int64
```

Data Preparation

```
# cek loan_data status with value count

loan_data['loan_status'].value_counts()

Current                                                    224226
Fully Paid                                                 184739
Charged Off                                                 42475
Late (31-120 days)                                          6900
In Grace Period                                            3146
Does not meet the credit policy. Status:Fully Paid          1988
Late (16-30 days)                                           1218
Default                                                     832
Does not meet the credit policy. Status:Charged Off         761
Name: loan_status, dtype: int64
```
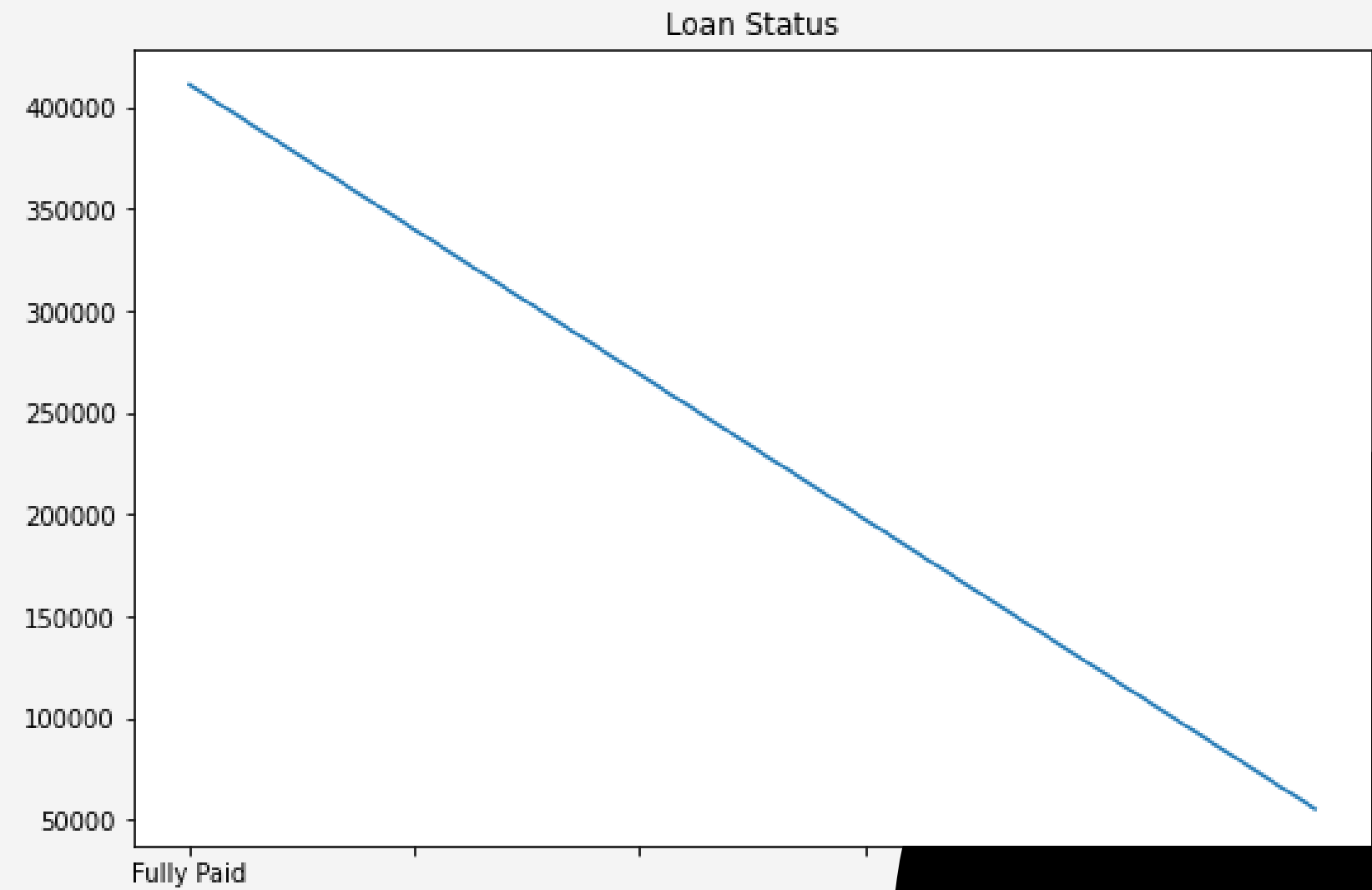
```
# cek loan_data status with value count

loan_data['loan_status'].value_counts()


Fully Paid          410953
Charged Off         55332
Name: loan_status, dtype: int64
```
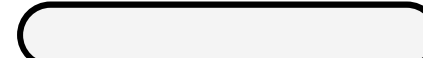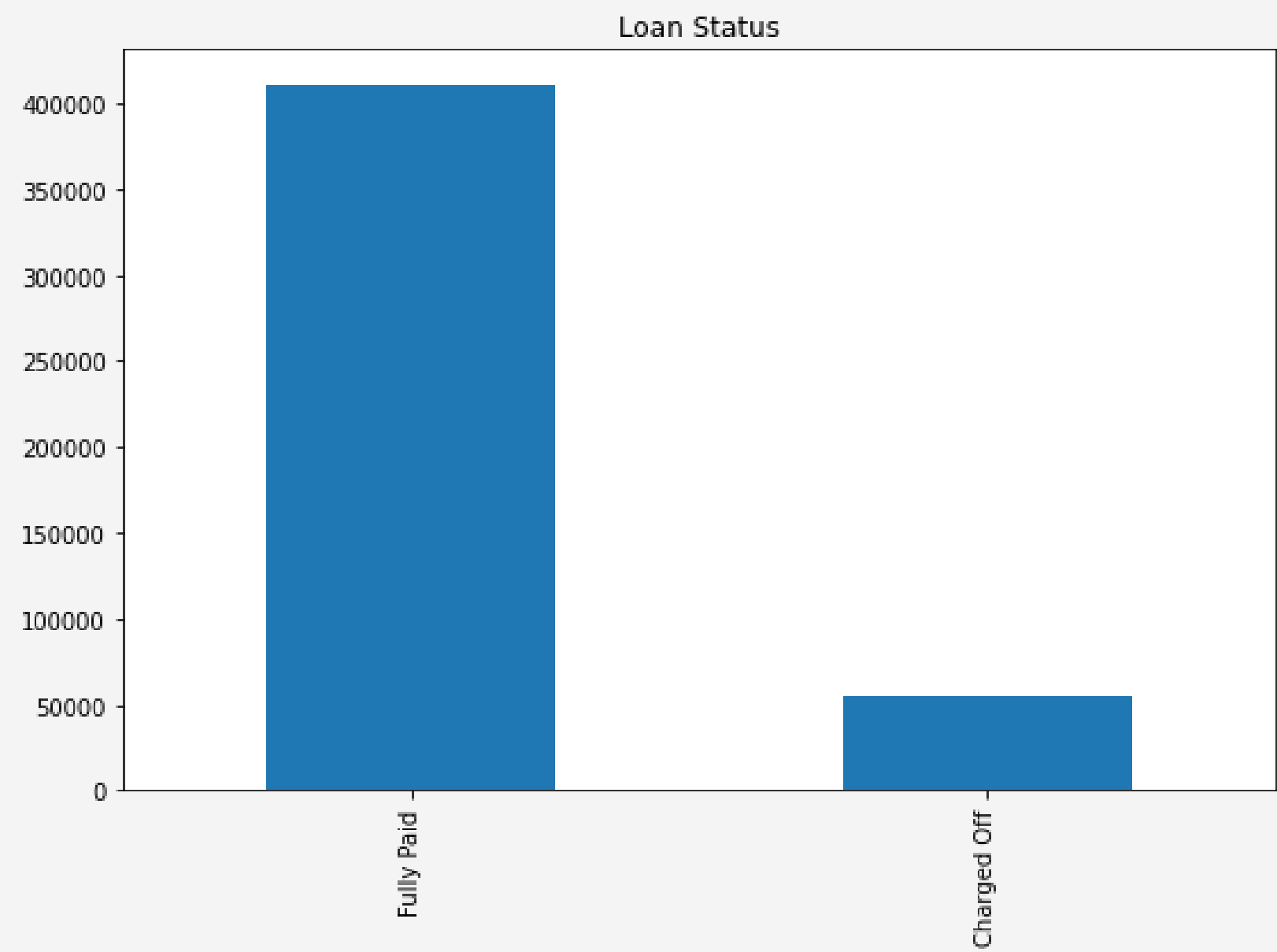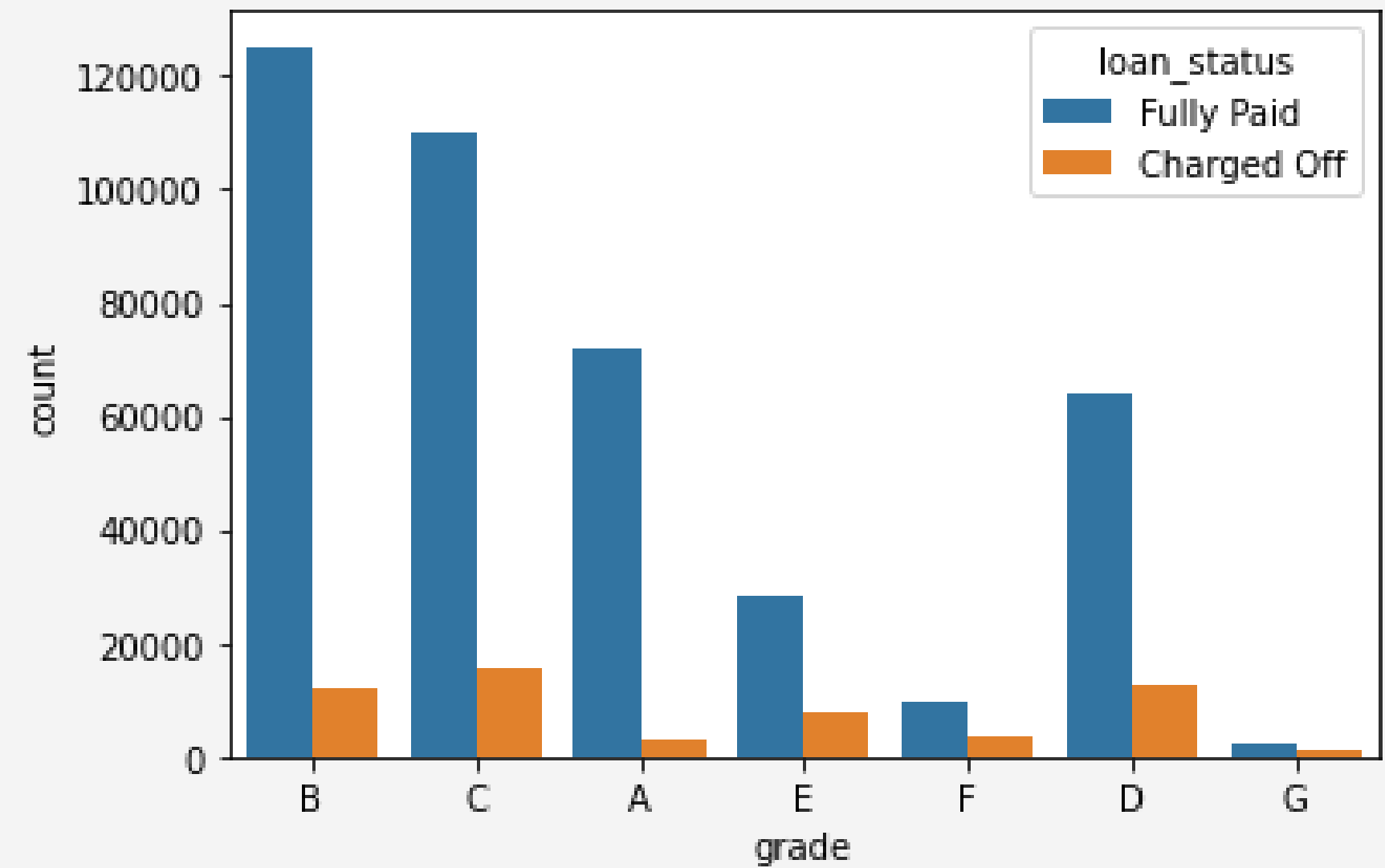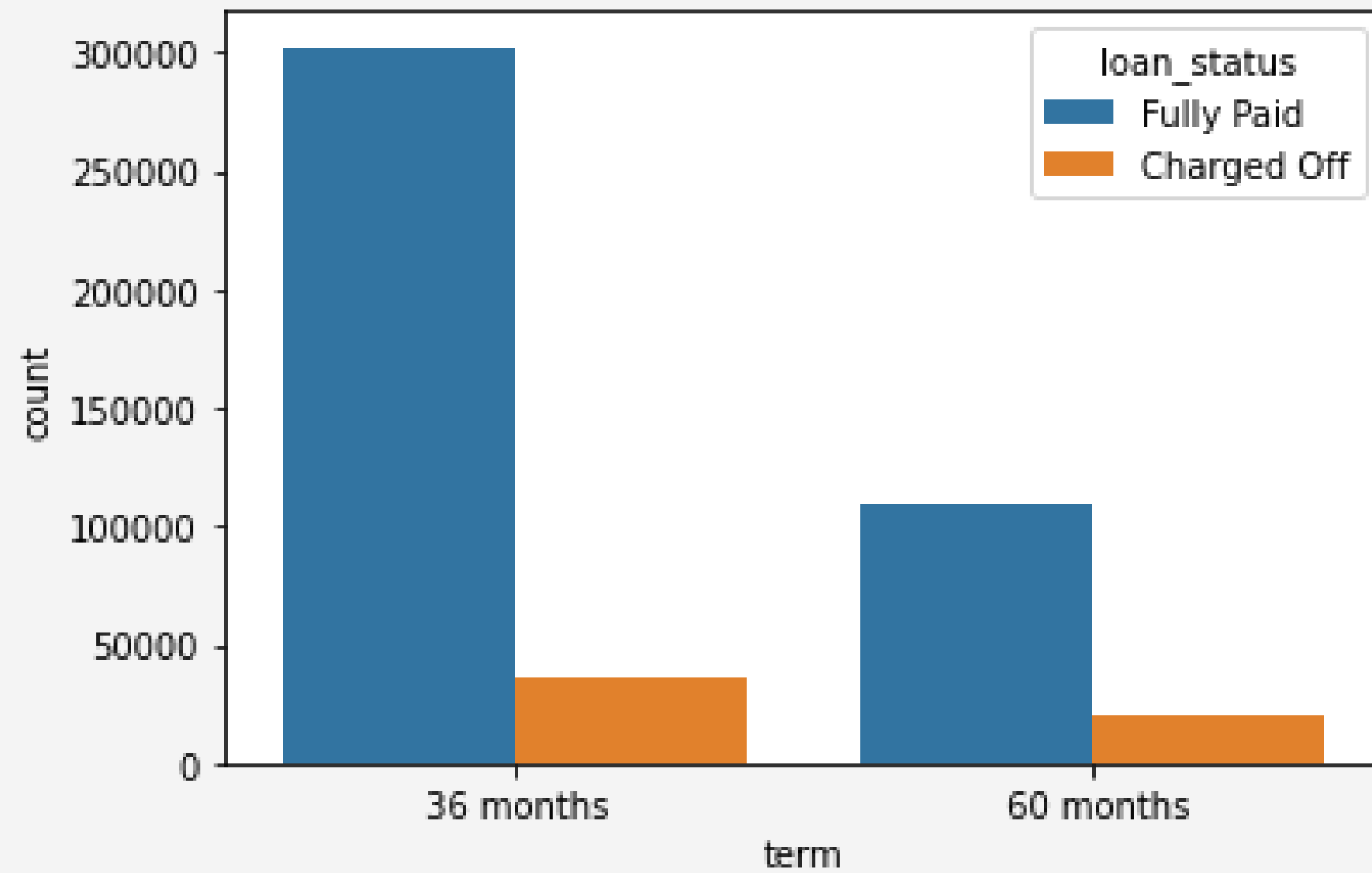
Exploratory Data Analysis

Exploratory Data Analysis

Loan Status

Loan Status

Exploratory Data Analysis

Thank You!