Lending Club Case Study

Group Facilitator: Praveen Kumar Sharma

Team Member: Arushi Garg

Problem Statement:

Lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed.

The main objective is to be able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

Perform an analysis to understand the driving factors (or driver variables) behind loan default, i.e. The variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Steps for performing the case study:

- Step1: Data Cleaning
- Step2: Univariate Analysis
- Step3: Segmented Univariate Analysis
- Step4: Bivariate/ Multivariate Analysis
- Step5: Results

Step 1: Data Cleaning

Data cleansing or data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

jupyter praveen_kumar_sharma (autosaved)

Logout

File Edit View Insert Cell Kernel Help Not Trusted Python 3

Run Code

In the data set , there are some columns having single value. These types of columns are not contributing in analysis so removed theme

In [7]:
```
loanData.drop(['pymnt_plan', "initial_list_status",'collections_12_mths_ex_med','policy_code','acc_now_delinq', 'application_type
loanData.head()
```

Out[7]:

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	total_rec_prncp	total_rec_int	total_rec_
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	...	5000.00	863.16	
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	...	456.46	435.17	
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	...	2400.00	605.67	
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	...	10000.00	2214.92	
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	...	2475.94	1037.39	

5 rows × 48 columns

jupyter praveen_kumar_sharma (autosaved)

Logout

File Edit View Insert Cell Kernel Help

Not Trusted | Python 3

Run Code

List of post-approval features

- delinq_2yrs
- revol_bal
- out_prncp
- total_pymnt
- total_rec_prncp
- total_rec_int
- total_rec_late_fee
- recoveries
- collection_recovery_fee
- last_pymnt_d
- last_pymnt_amnt
- next_pymnt_d
- chargeoff_within_12_mths
- mths_since_last_delinq
- mths_since_last_record

In [8]:
```
#Removing the unwanted columns in dataset
loanData.drop(["id", "member_id", "url", "title", "emp_title", "zip_code", "last_credit_pull_d", "addr_state","desc","out_prncp_i
```

In [9]:
```
loanData.shape
```

Out[9]: (39717, 21)

Step 2: Univariate Analysis

Univariate analysis is perhaps the simplest form of statistical analysis. Like other forms of statistics, it can be inferential or descriptive. The key fact is that only one variable is involved. Univariate analysis can yield misleading results in cases in which multivariate analysis is more appropriate.

jupyter praveen_kumar_sharma (autosaved)

Logout

File Edit View Insert Cell Kernel Help Not Trusted Python 3

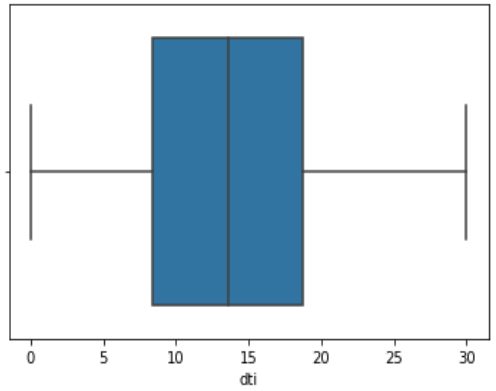
Run Code

Data Standardizing and finding outliers for following columns.

1. dti
2. loan_amnt
3. funded_amnt_inv

In [25]:
```
sns.boxplot(loanData.dti)
```

Out[25]:
```
<AxesSubplot:xlabel='dti'>
```



jupyter praveen_kumar_sharma (autosaved)

Logout

File Edit View Insert Cell Kernel Help Not Trusted Python 3

Run Code

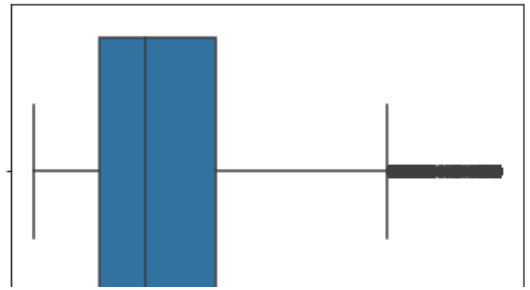
loan_amnt

In [27]:
```
loanData.loan_amnt.quantile([0.75,0.90,0.95,0.97,0.975, 0.98, 0.99, 1.0])
```

Out[27]:
```
0.750    15000.0
0.900    20000.0
0.950    25000.0
0.970    25475.0
0.975    28000.0
0.980    30000.0
0.990    35000.0
1.000    35000.0
Name: loan_amnt, dtype: float64
```

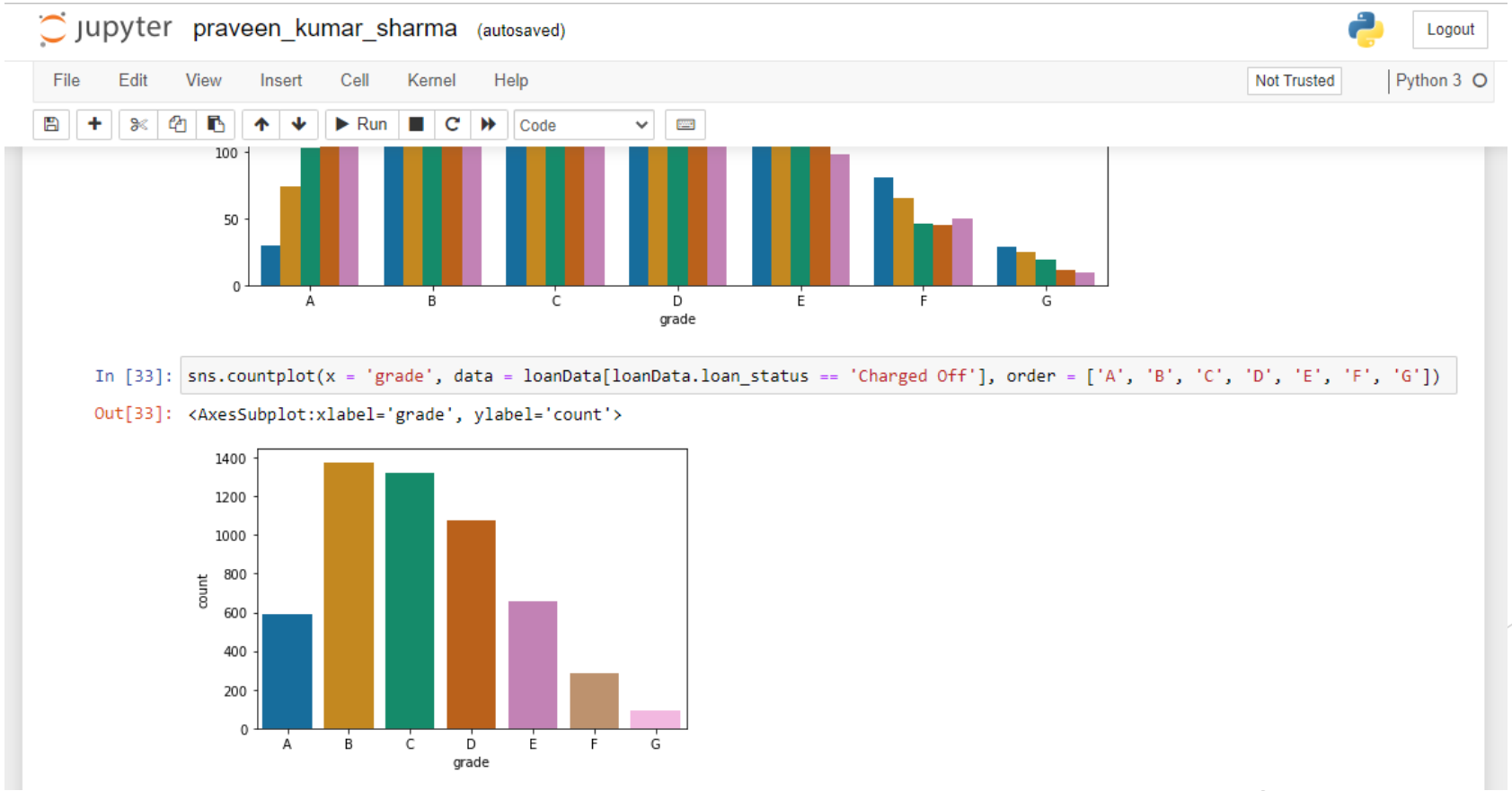
In [28]:
```
sns.boxplot(loanData.funded_amnt_inv)
```

Out[28]:
```
<AxesSubplot:xlabel='funded_amnt_inv'>
```



Step 3: Segmented Univariate Analysis

Segmented Univariate analysis can be used to find summary of a single data variable in form of segments. The dataset variable is divided into subsets and patterns can be observed across the segments. The central tendencies such as mean, mode, and median; maximum and minimum; range; variance and standard deviation are also detected.



jupyter praveen_kumar_sharma (autosaved)

Logout

File Edit View Insert Cell Kernel Help

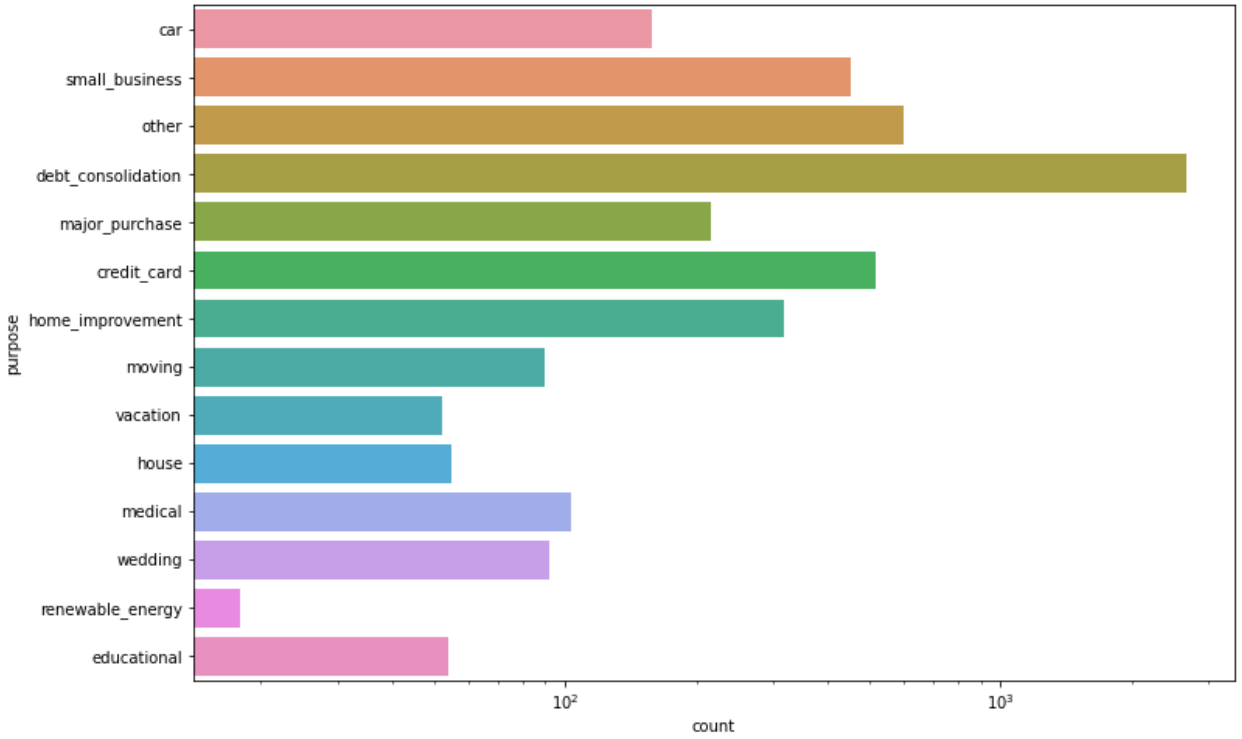
Not Trusted

Python 3

Run

Code

Out[58]: <AxesSubplot:xlabel='count', ylabel='purpose'>



Step 4: Bivariate/ Multivariate Analysis

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association.

jupyter praveen_kumar_sharma (autosaved)

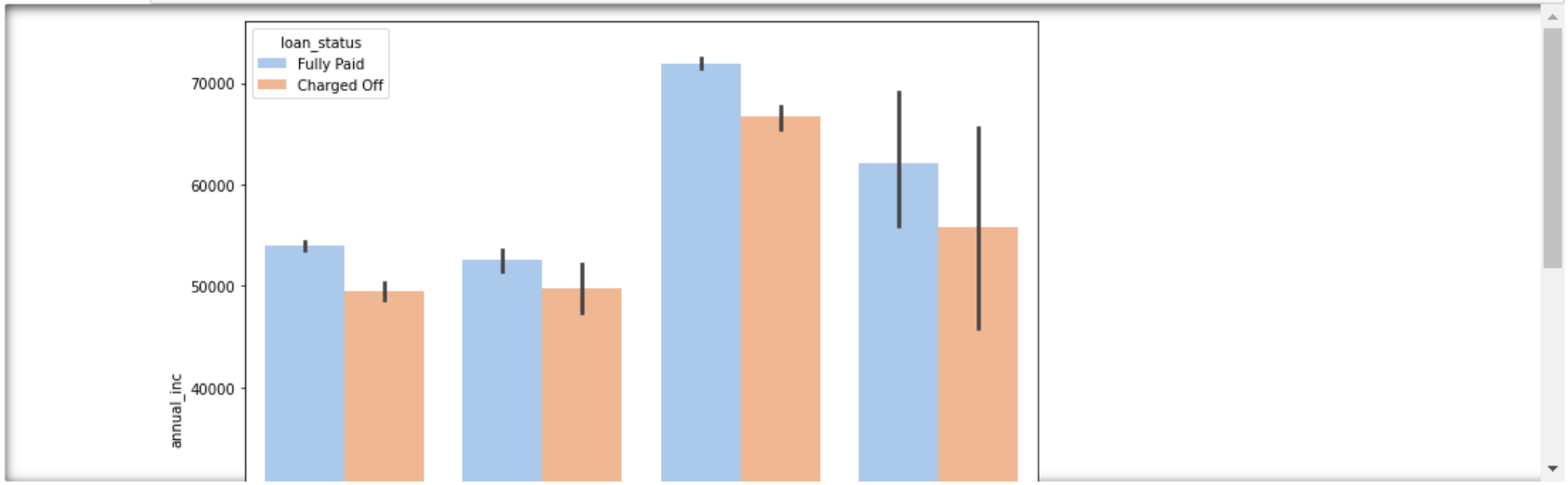
Logout

File Edit View Insert Cell Kernel Help Not Trusted Python 3

Run Code

2.Annual income vs home ownership

In [59]:
```
plt.figure(figsize=(10,10))
sns.barplot(data =loanData,x='home_ownership', y='annual_inc', hue ='loan_status',palette="pastel")
plt.show()
```



jupyter praveen_kumar_sharma (autosaved)

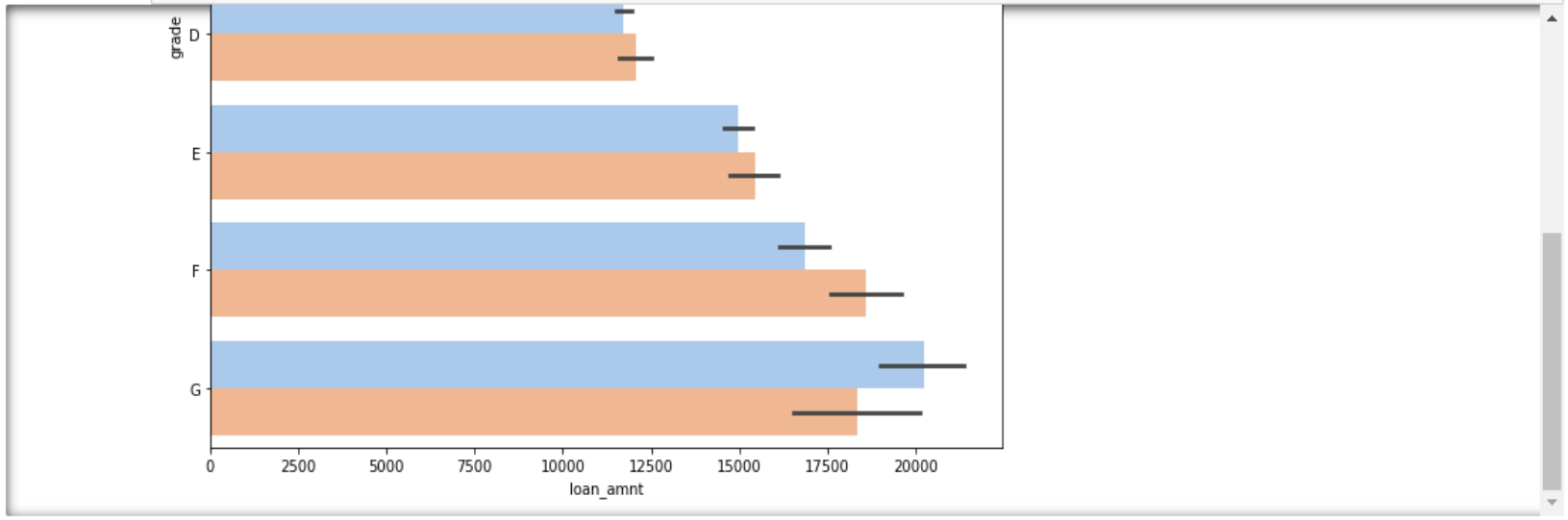
Logout

File Edit View Insert Cell Kernel Help Not Trusted Python 3

Run Code

5.Loan amount vs Grade

In [66]:
```
plt.figure(figsize=(10,10))
sns.barplot(data =loanData,x='loan_amnt', y='grade', hue ='loan_status',palette="pastel", order=['A','B','C','D','E','F','G'])
plt.show()
```



In [67]:
```
plt.figure(figsize=(20,20))
```

Observations:

- **The above analysis with respect to the charged off loans. There is a more probability of defaulting when :**
- Applicants taking loan for 'home improvement' and have income of 60k -70k
- Applicants whose home ownership is 'MORTGAGE and have income of 60-70k
- Applicants who receive interest at the rate of 21-24% and have an income of 70k-80k
- Applicants who have taken a loan in the range 30k - 35k and are charged interest rate of 15-17.5 %
- Applicants who have taken a loan for small business and the loan amount is greater than 14k
- Applicants whose home ownership is 'MORTGAGE and have loan of 14-16k
- When grade is F and loan amount is between 15k-20k
- When employment length is 10yrs and loan amount is 12k-14k
- When the loan is verified and loan amount is above 16k
- For grade G and interest rate above 20%

Thank You!