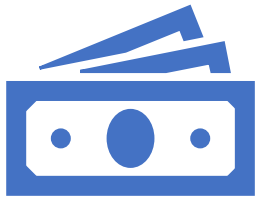


Lending Club Case Study

Usman Khan

EDA based

Approach

- Import Data
- Data Cleaning
  - Remove or delete rows with null entry
  - Remove outliers
  - Remove unwanted rows, columns
  - Check and Convert columns for correct data types
  - Generate more columns that are required
- Analysis
  - Data Understanding
  - Perform Segmented Univariate Analysis with Graphs
  - Perform Segmented Bivariate Analysiswith Graphs
  - Plots correlation graph
- Correlations
- Observations

Data cleaning

- Before Data Clean:
  - Total Data Size: (39717, 111)
- Data Cleaning

Sl. No	Columns	Remarks
1	term	Convert Term count to convert to int type {36,60}
2	int_rate	String "%" symbol to convert convert to float type
3	grade sub_grade	convert to ordered values to int type {A-G} -> {1-5}, {A1-G5 -> {1-35} Created new columns grade_int, sub_grade_int
4	emp_length	convert to integer values {0-10} convert dtype to int
5	funded_amnt	Remove outliers from funded_amnt to remove any biased analysis (Considering only 25000)
6	annual_inc	Analyze only for annual_inc < 80000 Analyze upto 80K annual income anythign > 80K are classified as high income profiles
7	Dropped	

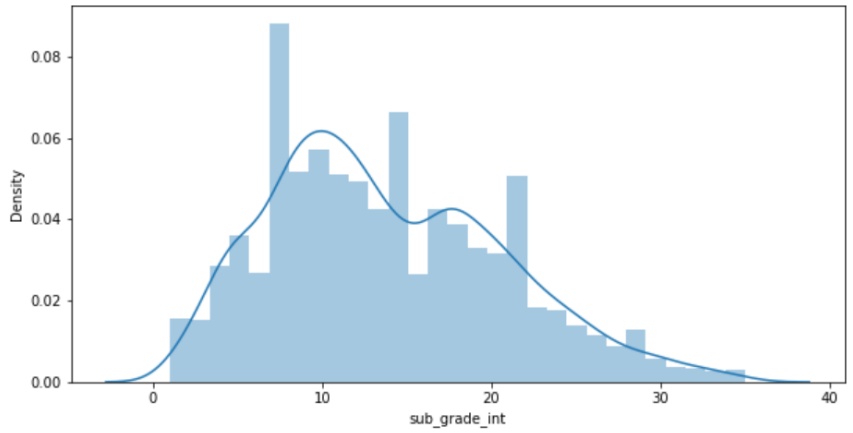
Data Cleaning

- Dropped columns with NA
  - 'mths_since_last_major_derog', 'annual_inc_joint', 'dti_joint', 'verification_status_joint', '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', 'acc_open_past_24mths', 'avg_cur_bal', 'bc_open_to_buy', 'bc_util', 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl', 'mort_acc', 'mths_since_recent_bc', 'mths_since_recent_bc_dlq', 'mths_since_recent_inq', 'mths_since_recent_revol_delinq', 'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m','num_tl_30dpd', 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m', 'pct_tl_nvr_dlq', 'percent_bc_gt_75', 'tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit'
- Drop empty columns (columns with 0) or constant columns
  - ['mths_since_last_delinq','mths_since_last_record','next_pymnt_d']
  - ['collections_12_mths_ex_med', 'acc_now_delinq', 'chargeoff_within_12_mths', 'delinq_amnt', 'tax_liens']
- New Columns
  - `'grade_int', 'sub_grade_int', 'funded_amnt_bin', 'revol_util_bucket'`
- Resized data = (36135, 57)
  - Fully Paid ~30000
  - Charged Off ~5000

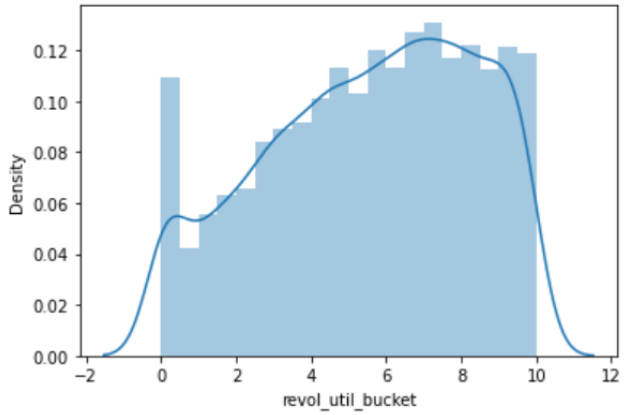
Analysis

- The number of loan account charged off has correlation as below from the data set
  - Loans with LC sub-grade 7-22 (B2 - D2) shows high charged-off with
  - Loan accounts with revol_util > 45% is likely to default
  - **Loan Status with total_acc has no significance**

```
sns.distplot(df[df['loan_status'] == "Charged Off"]['sub_grade_int'])
plt.show()
```




```
sns.distplot(df[df['loan_status'] == "Charged Off"]['revol_util_bucket'])
<AxesSubplot:xlabel='revol_util_bucket', ylabel='Density'>
```

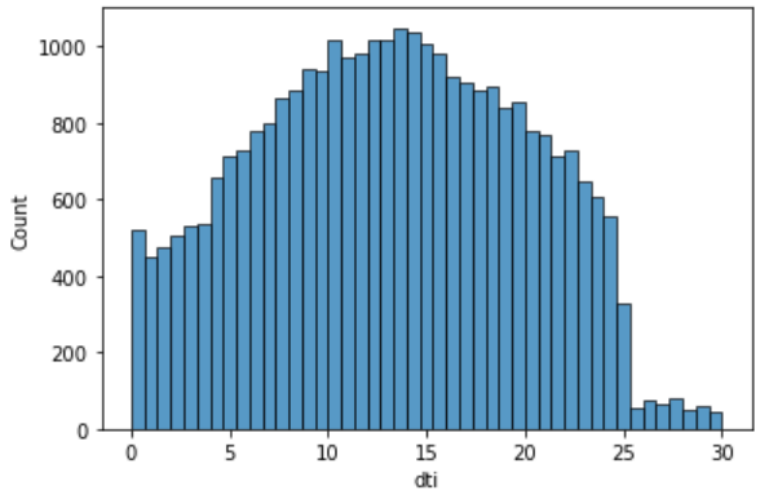


Analysis based on dti (borrower's total monthly debt payments on the total debt obligations)

There is no significant difference with loan status and dti
However between 12 - 20 shows high defaulter, hence should be accessed

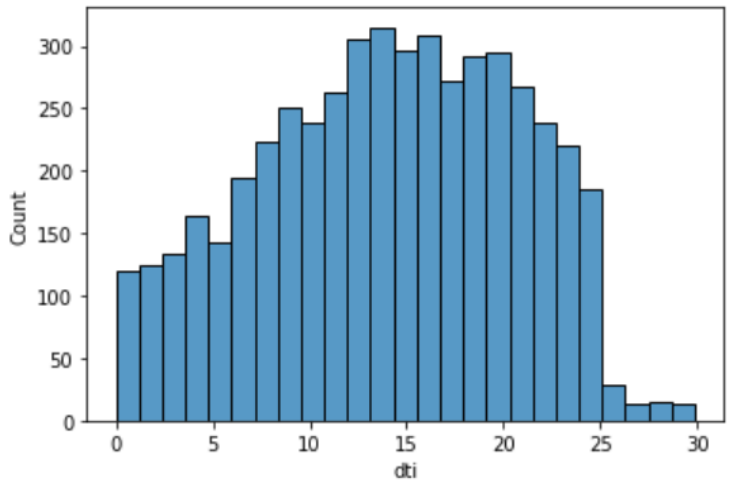
```
sns.histplot(df[df['loan_status'] == "Fully Paid"]['dti'])
```

```
<AxesSubplot:xlabel='dti', ylabel='Count'>
```




```
sns.histplot(df[df['loan_status'] == "Charged Off"]['dti'])
```

```
<AxesSubplot:xlabel='dti', ylabel='Count'>
```

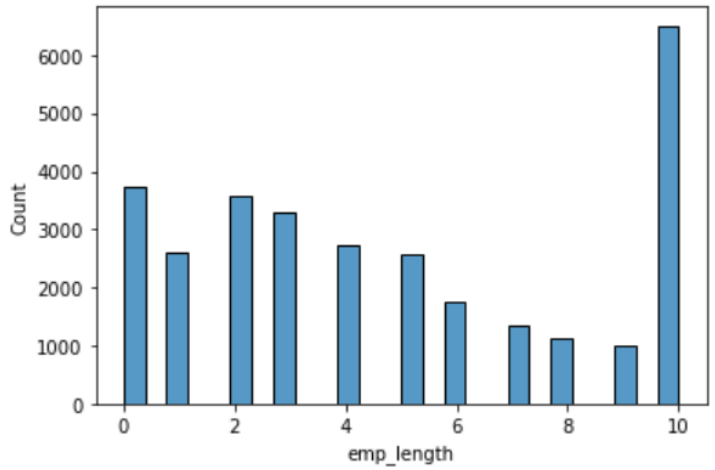


Analysis based on emp_length

- Loan application with emp_length :10+ years loan should be considered as they are likely to be the defaulter as seen below
- Full Paid: 21.493010805987904
- Charged Off: 23.22698638488112

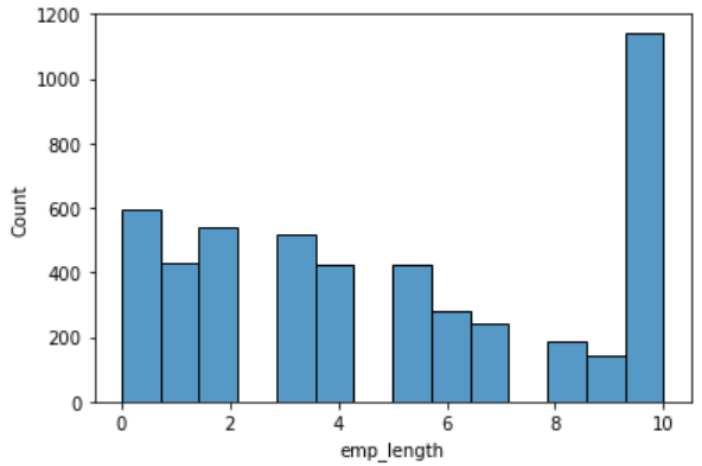
```
sns.histplot(df[df['loan_status'] == "Fully Paid"]['emp_length'])
```

```
<AxesSubplot:xlabel='emp_length', ylabel='Count'>
```




```
sns.histplot(df[df['loan_status'] == "Charged Off"]['emp_length'])
```

```
<AxesSubplot:xlabel='emp_length', ylabel='Count'>
```

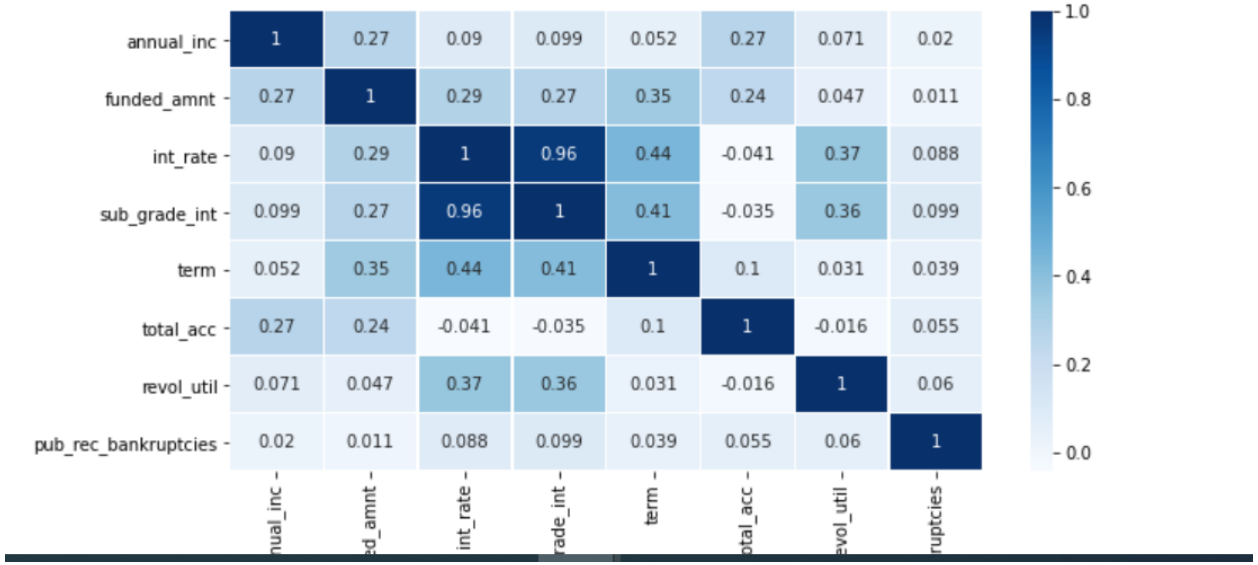


Analysis and Correlation

- Home ownership won't have significance impact on the loan_status
- Home Ownership and verification status shows no significance on loan status¶

- Correlation
  - Defaulter loans has higher correlation with LC grade (41%), loan interest rate (44%),
  - Term, interest rate shows correlation of 44%

```
corr = df[df['loan_status']=="Charged Off"][['annual_inc','funded_amnt','int_rate','sub_grade_i
fig,ax = plt.subplots(figsize=(10,5))
sns.heatmap(corr, cmap="Blues", annot=True, linewidths=.5, ax=ax)
plt.show()
```



Observation

- Loans with LC sub-grade 7-22 (B2 - D2) shows high charged-off with
- Loan accounts with revol_util > 45% is likely to default

Thank You