Lending Club
Case Study
(Upgrad assignment)

Submitted By:
Mitali Seth
Bharat Kalra

Deck Components

01 Problem Statement

02 Data Cleaning

03 Univariate Analysis

04 Segmented Analysis

05 Bivariate Analysis

06 Correlational Analysis

07 Insight Summary

08 GitHub link

Problem Statement

About Lending Company

Lending Club operates as a consumer finance marketplace specializing in providing various types of loans to urban customers. It facilitates the connection between borrowers seeking loans and investors seeking opportunities to lend their funds and earn returns..

It specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.

Problem Statement

The objective is to pinpoint these high-risk loan applicants, which would enable a reduction in such loans and subsequently minimize the overall credit loss. The primary focus of this case study is to achieve the identification of these applicants through exploratory data analysis (EDA) using the provided dataset.

In other words, the company wants 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.

Data Cleaning and Manipulation

Raw data quick summary

- No header or footer rows in the dataset
- No summary rows were there in the dataset
- No column number, indicators etc. found in the dataset
- No duplicate rows are there in the dataset

Drop unnecessary rows:

- Rows where the loan_status = CURRENT will be dropped as CURRENT loans are in progress and will not contribute in the decision making.

Drop unnecessary columns:

- Drop those columns where all values are only zero or NA.
- Drop columns where all values are constant such as application_type
- Drop unnecessary columns (id, member_id, url, emp_title, desc, purpose, title, zip_code, addr_state, next_pymnt_d, mths_since_last_record, policy_code,pymnt_plan,initial_list_status,pub_rec,delinq_2yrs,out_prncp,out_prncp_inv,total_rec_late_fee,recoveries,collection_recovery_fee,pub_rec_bankruptcies,inq_last_6mths,mths_since_last_delinq which wont add significant value in analysis

Standardizing the data:

- Removing % sign from int_rate values and converting the data type into float

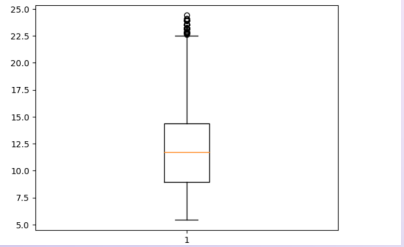
Remove Outlier data:

- Basis boxplot of annual income, we realize that that 99% of the data lies below 1,000,000. So, considering annual income above 1,000,000 as Outliers and removing them from the data

Univariate Analysis

```
plt.boxplot(df.int_rate)
```

```
{'whiskers': [<matplotlib.lines.Line2D at 0x20912009660>,
  <matplotlib.lines.Line2D at 0x20912009900>],
 'caps': [<matplotlib.lines.Line2D at 0x20912009ba0>,
  <matplotlib.lines.Line2D at 0x20912009e40>],
 'boxes': [<matplotlib.lines.Line2D at 0x209120093c0>],
 'medians': [<matplotlib.lines.Line2D at 0x2091200a0e0>],
 'fliers': [<matplotlib.lines.Line2D at 0x2091200a380>],
 'means': []}
```



Int_rate

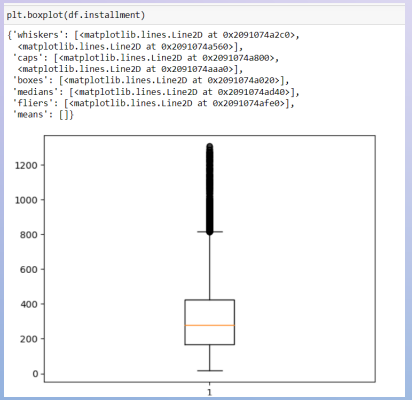
```
df['int_rate'].describe()
```

```
count    38577.000000
mean        11.932219
std          3.691327
min          5.420000
25%          8.940000
50%         11.710000
75%         14.380000
max         24.400000
Name: int_rate, dtype: float64
```

From this analysis of int_rate, we can see that for majority of the applicants, int_rate has been between 8 to 14. However for some cases it is between 22 to 25. Probably, the higher interest rate(~say 23%) should be kept for that customer profile who fall in charged off category.
(We will see this in further slides)

```
plt.boxplot(df.installment)
```

```
{'whiskers': [<matplotlib.lines.Line2D at 0x2091074a2c0>,
  <matplotlib.lines.Line2D at 0x2091074a560>],
 'caps': [<matplotlib.lines.Line2D at 0x2091074a800>,
  <matplotlib.lines.Line2D at 0x2091074aaa0>],
 'boxes': [<matplotlib.lines.Line2D at 0x2091074a020>],
 'medians': [<matplotlib.lines.Line2D at 0x2091074ad40>],
 'fliers': [<matplotlib.lines.Line2D at 0x2091074afe0>],
 'means': []}
```



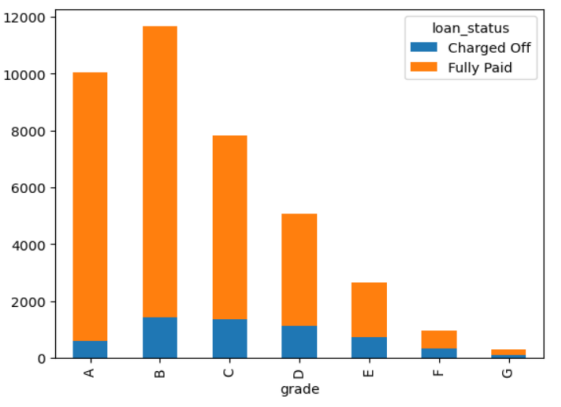
installment

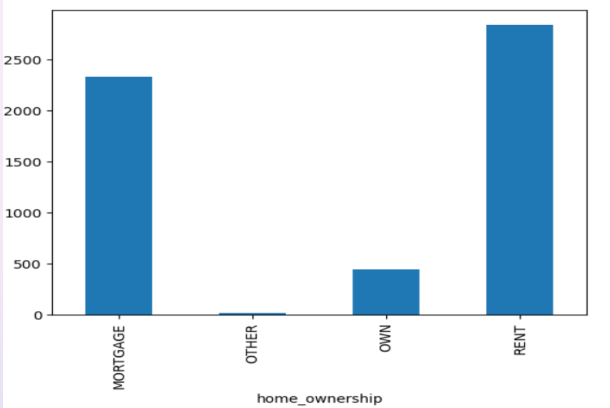
```
df['installment'].describe()
```

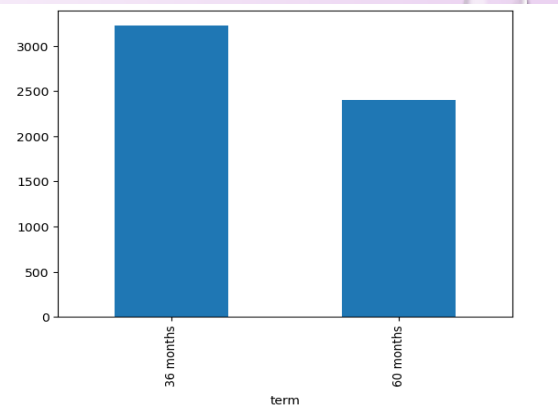
```
count    38577.000000
mean       322.466318
std        208.639215
min         15.690000
25%        165.740000
50%        277.860000
75%        425.550000
max       1305.190000
Name: installment, dtype: float64
```

From this analysis of 'installment', we can see that for majority of the applicants, installment has been between 165 to 425. However, for some cases it is more than 800 and goes till 1305. Probably, for the risky applicants with higher charged offs, loan amount should be disbursed at a higher installment amount to clear it off asap. However, this can depend on a variety of other factors

Bivariate Analysis (Categorical Variables)







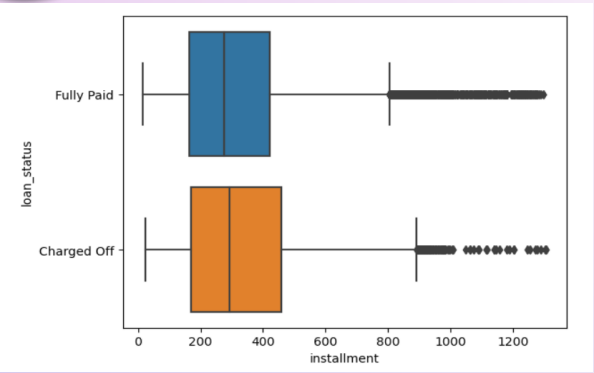
Portfolio Management:

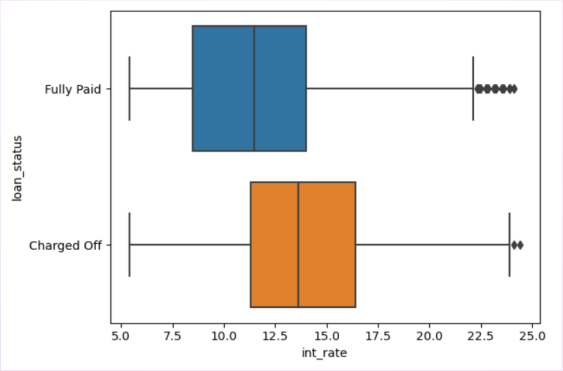
Loan Status vs Grades: Lending Club may consider adjusting its portfolio management strategies based on the analysis of loan grades. For example, it may allocate a smaller portion of funds to higher-risk grades (B and C) or more stringent approval criteria or higher interest rates may be applied to higher-risk grades (B and C) to mitigate the risk of defaults.

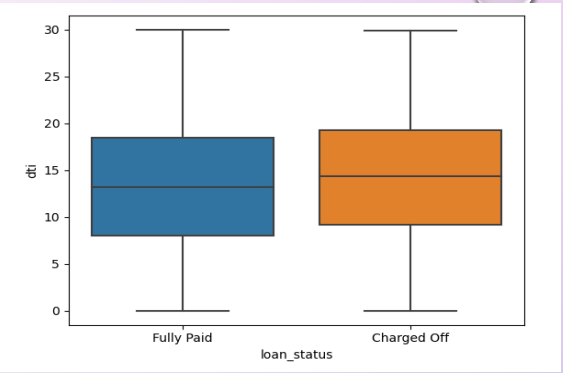
Home Ownership Impact: The data suggests that home ownership may be a factor influencing credit risk. Borrowers with "MORTGAGE" and "RENT" home ownership appear to have a higher likelihood of loan defaults (Charged Off) compared to those with "OWN," "OTHER," or "NONE.“

Loan Status vs Debt to income ratio: Charged off loans for 36 months loan term is 25% higher compared to 60 months loan term which indicates that portfolio should be managed in a way that loan disbursal is higher for 36 months loan term.

Bivariate Analysis (Numerical Variables)





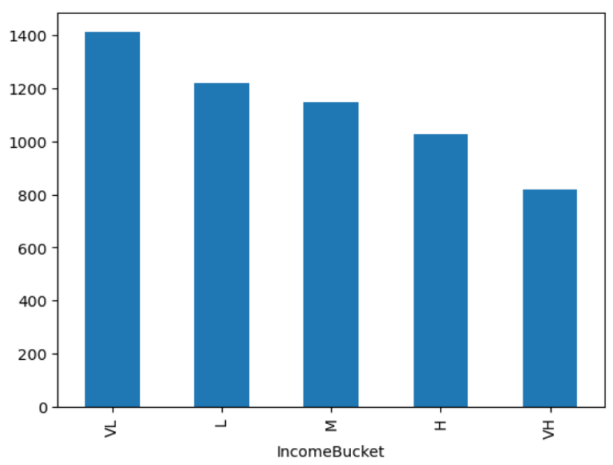


Loan Status vs Installment Amount: The fact that Charged Off loans have a higher mean and median installment amount suggests that, on average, loans that default tend to have slightly larger installment amounts

Loan Status vs Interest Rates: The higher mean, median and standard deviation in interest rates for Charged Off loans suggest that loans with higher interest rates may be associated with a higher risk of default.

Loan Status vs Debt to income ratio: With minor difference in dti median amount for charged off loans compared to Fully paid and higher variation, the result indicates that dti on a standalone basis is not a significant factor for identifying defaulters.

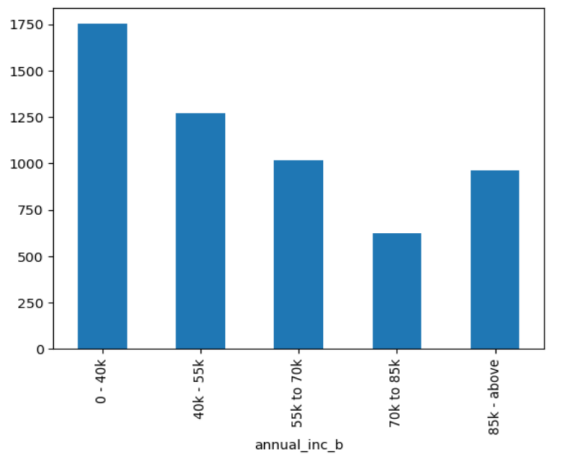
Segmented Analysis



Annual Income bucket

```
df.annual_inc.describe()
```

```
count     38563.000000
mean      68109.797129
std       47423.521494
min        4000.000000
25%       40000.000000
50%       58800.000000
75%       82000.000000
max      948000.000000
Name: annual_inc, dtype: float64
```



Using q cut and applying binning on annual income for charged off loans, we get the bucketing for Annual income. Number of elements in Annual income is binned equally in below buckets:

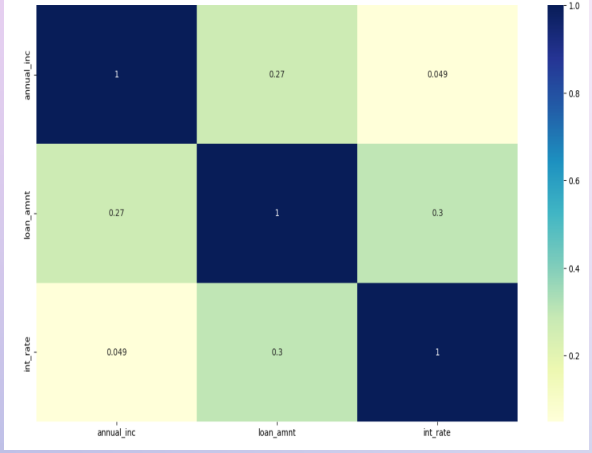
0.0 – 0.2	**:**	**VL**
0.2 – 0.4	**:**	**L**
0.4 – 0.6	**:**	**M**
0.6 – 0.8	**:**	**H**
0.8 – 1.0	**:**	**VH**

For Around 50% of the charged off loans, Annual income is below 58,800. Hence Annual income becomes an important factor while deciding loan disbursal.
Detailed scrutiny to be ensured for people in low income range to ensure their loan paying capability

Correlational Analysis

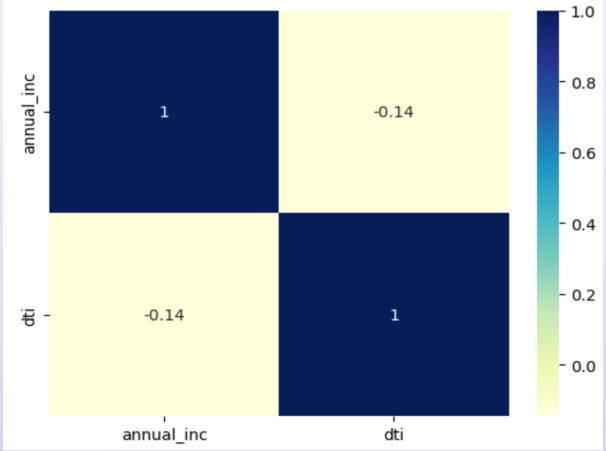
Annual_inc vs loan_amnt vs int_rate

```
# Identifying key correlations
corr = df.loc[:, ['annual_inc','loan_amnt','int_rate']].corr()
plt.figure(figsize = (15,8))
sns.heatmap(corr, annot = True, cmap='YlGnBu')
plt.show()
```



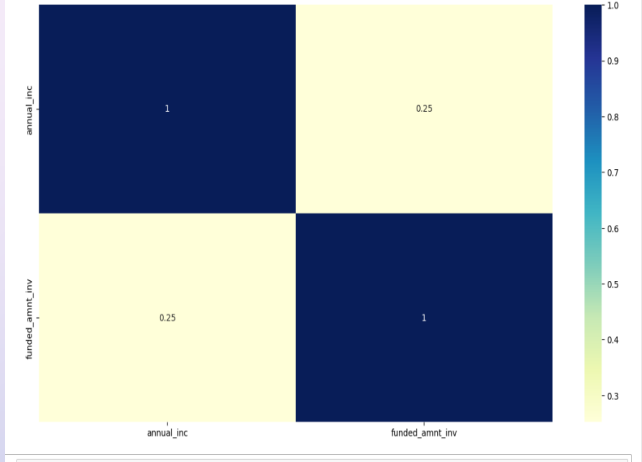
Annual_inc vs dti

```
corr = df.loc[:,["annual_inc","dti"]].corr()
sns.heatmap(corr, annot = True, cmap='YlGnBu')
plt.figure(figsize = (15,8))
plt.show()
```



Annual_inc vs funded_amnt_inv

```
# Identifying key correlations
corr = df.loc[:, ['annual_inc','funded_amnt_inv']].corr()
plt.figure(figsize = (15,8))
sns.heatmap(corr, annot = True, cmap='YlGnBu')
plt.show()
```



Here we can see that :

- annual_inc is positively correlated to loan amnt as well as int_rate
- Funded_amnt_inv is also positively correlated to loan_amnt
- dti is negatively corelated to annual income

Also, in the previous slide we saw that charged off loans are higher for lower Annual income.
Hence, we can conclude from these 2 variables that if annual income is low (or dti is high), loan amount and funded amount disbursed should be kept low so as to reduce the magnitude of loss. However, the positive correlation of loan_amnt with interest rate is something which has to be looked upon. Ideally, based on the above correlation analysis the interest rate is higher for higher loan amount and it should be that way itself businesswise. But if the customers have low annual income, then charged off probability is higher. Hence for customers with low annual income, loan amount should be evaluated in detail and interest rate can be kept slightly higher(>22%) to hedge the portfolio.

Insight Summary

Loan Term: Charged off is 25% higher for applicants with lower loan term(36 months)

Grade: Loan grades can be indicative of credit risk. Grades B and C have a relatively higher number of loans categorized as "Charged Off" compared to grades A and D. This suggests that grades B and C may represent higher credit risk.

Interest Rate : The higher mean, median and variation in interest rates for Charged Off loans suggest that loans with higher interest rates may be associated with a higher risk of default.

Home Ownership Impact: The data suggests that home ownership may be a factor influencing credit risk. Borrowers with "MORTGAGE" and "RENT" home ownership appear to have a higher likelihood of loan defaults (Charged Off) compared to those with "OWN," "OTHER," or "NONE."

Annual Income: Applicants in low incomes range(below 60k) have a greater share of defaulted loans.

Correlation Analysis: For profiles with low annual income, loan amount and funded amount disbursed should be kept low so as to reduce the magnitude of loss. Additionally, interest should be kept higher to balance out the portfolio and minimize losses.

In conclusion we can realize that factors like Loan term, Grade, Home ownership , Annual income and interest rate become the key driving factors behind loan default. Also, from the corelation analysis we can realize that dti is negatively corelated to annual income and hence dti becomes another key driving factor behind loan default. Certain other variables like issue year indicate that number of charged–off is relatively high in 2011. This indicate that the charged-off loans disbursed in year 2011 need detailed level analysis.
Although this conclusion is just limited to univariate and bivariate analysis of key variables, the above mentioned variables should be carefully analysed to ensure loan portfolio is hedged properly. Certain variables like interest rate(higher for risky applicants), loan amount (lower for risky applicants) should be adjusted properly for new applicants (basis existing charged off customer profile).

GitHub Repo Link

https://github.com/mitaliseth019/LendingClubCaseStudy

THANK YOU