Assignment Brief:

You work for a consumer finance company Lending Club which specialises in lending various types of loans to urban customers. This company 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. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. When a person applies for a loan, there are two types of decisions that could be taken by the company:

1. **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:

    Fully paid: Applicant has fully paid the loan (the principal and the interest rate)

    Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

    Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

2. **Loan rejected:** The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Business Objectives:

Like most other lending companies, 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. In other words, borrowers who default cause the largest amount of loss to the lenders.

Objective is to identify the risky loan applicants at the time of loan application so that 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.

In other words, to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment. And thus minimise the risk of losing money while lending to customers.

Questions:

1. Explain the data
2. Clean the data (remove columns with null values, all random values or single categorical value + convert the data to proper int, float or date representations)
3. Perform the following analysis (show appropriate plots) and state your hypothesis:
    a. Overall loan status (study total loans per category and total loans recovered)
    b. Understand loan based on grade
    c. Defaults by interest rate
    d. Defaults by Loan purpose
    e. Defaults by borrower's income
    f. Default by ratio of amount to income
    g. Default by Revolving Line Util rate
    h. Default by prior bad record
    i. Default by Debt to income Ratio
4. Finally summarize your recommendations on better quality borrowers

Loading and cleaning:

[13]:
```
loan = pd.read_csv('../input/lending-club-loan-data/loan.csv',dtype='object')
```

[15]:
```
loan = loan[['loan_amnt', 'funded_amnt', 'term', 'int_rate', 'installment', 'grade', 'sub_grade', 'disbursement_method', 'debt_settlemer
print(loan.shape)
```

```
(2260668, 9)
```

[16]:
```
loan.head()
```

Out[16]:

	loan_amnt	funded_amnt	term	int_rate	installment	grade	sub_grade	disbursement_method	debt_settlement_flag
0	2500	2500	36 months	13.56	84.92	C	C1	Cash	N
1	30000	30000	60 months	18.94	777.23	D	D2	Cash	N
2	5000	5000	36 months	17.97	180.69	D	D1	Cash	N
3	4000	4000	36 months	18.94	146.51	D	D2	Cash	N
4	30000	30000	60 months	16.14	731.78	C	C4	Cash	N

[18]:
```
loan.describe()
```

Out[18]:

	loan_amnt	funded_amnt	term	int_rate	installment	grade	sub_grade	disbursement_method	debt_settlement_flag
count	2260668	2260668	2260668	2260668	2260668	2260668	2260668	2260668	2260668
unique	1572	1572	2	673	93296	7	35	2	2
top	10000	10000	36 months	11.99	301.15	B	C1	Cash	N
freq	187236	187146	1609754	53869	4420	663557	145903	2182546	2227612

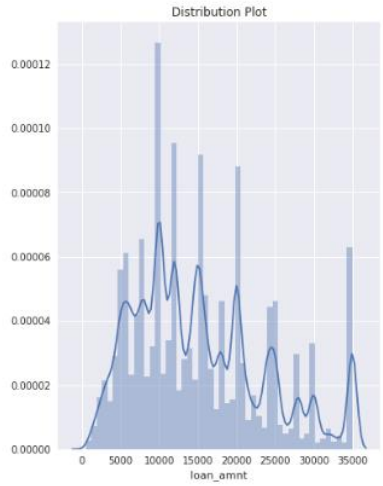
+ Code + Markdown

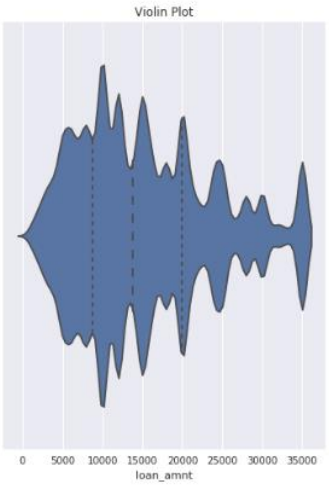
[20]:
```
loan.isnull().values.any()
```

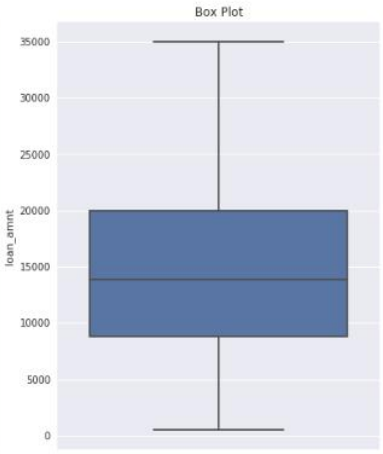
Out[20]
```
False
```

EDA:

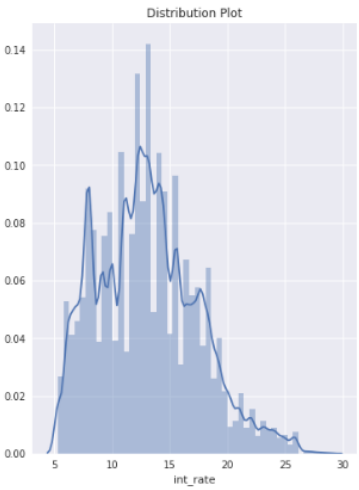
(Loan amount)

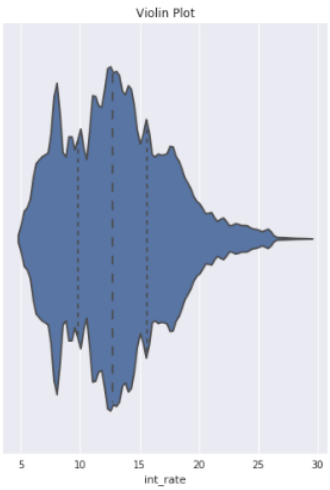


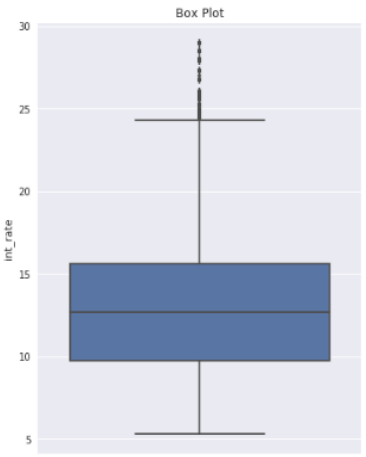




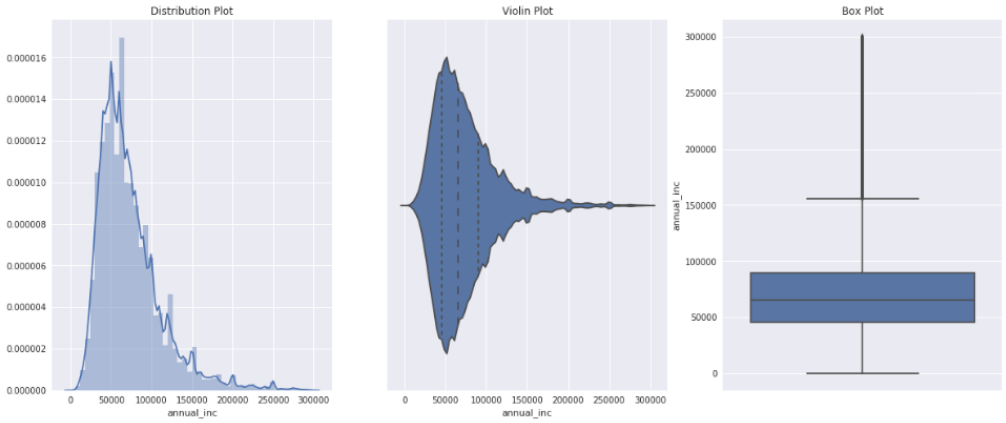
(Interest rate)



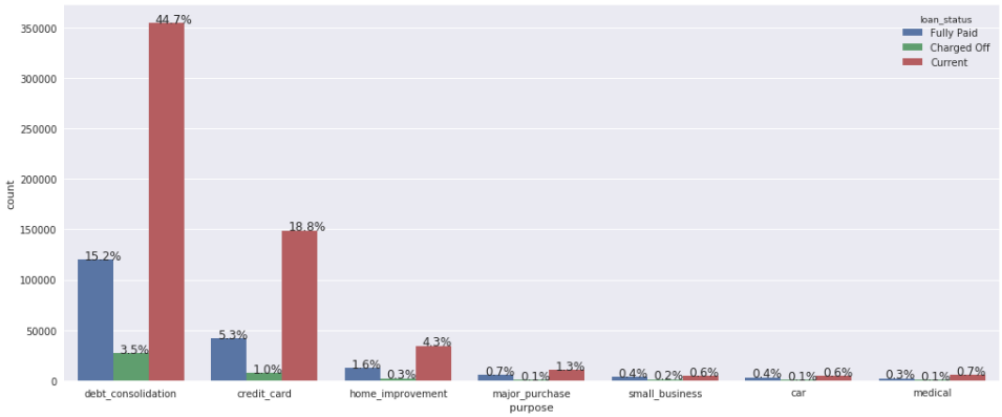




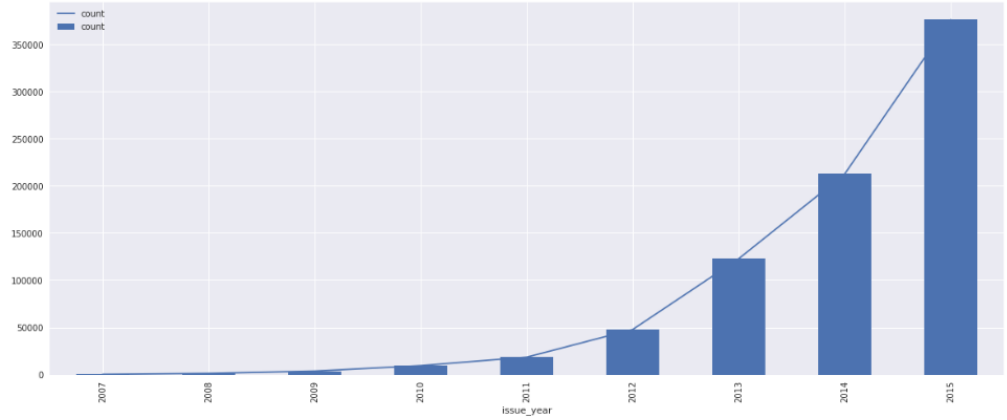
(Annual income)



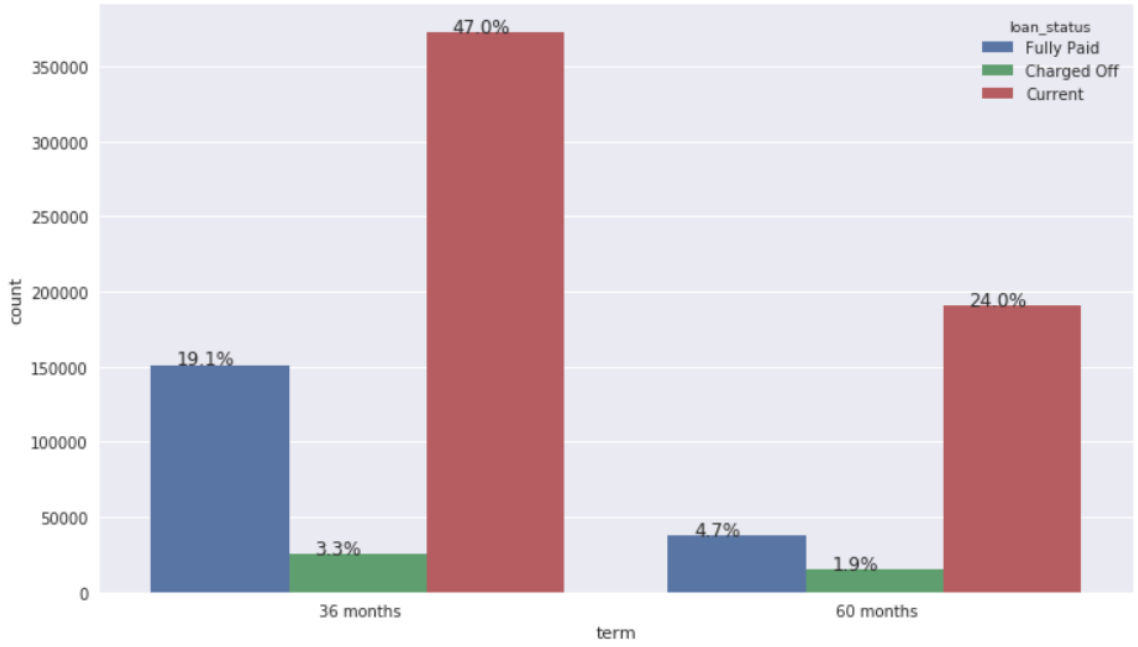
(Purpose)



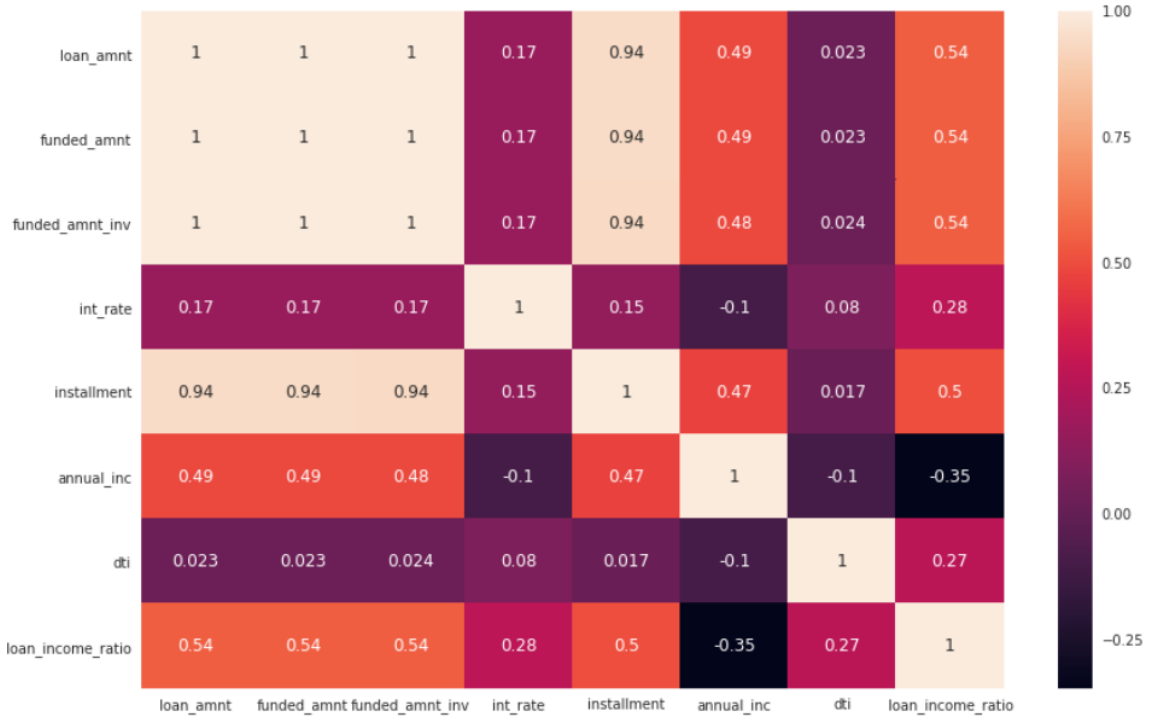
(Issue by year)



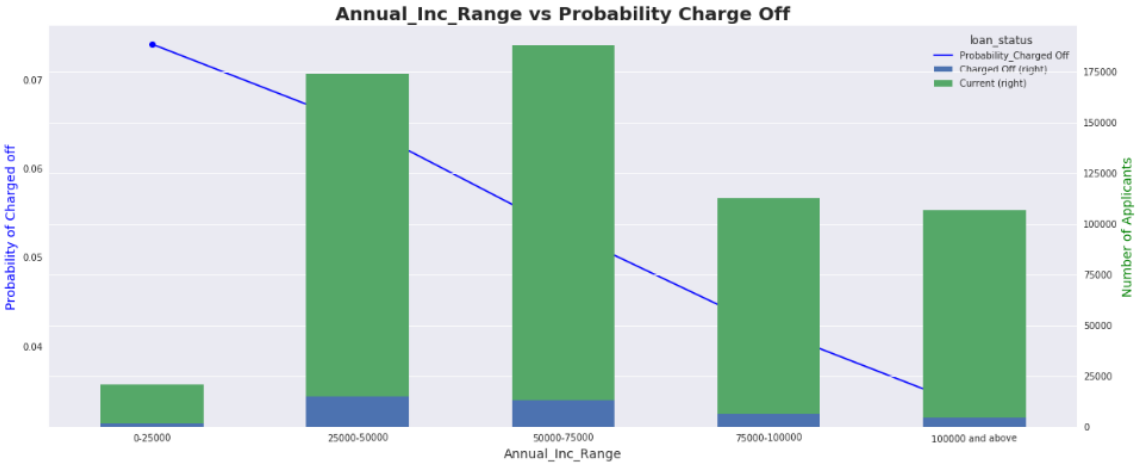
(By term)

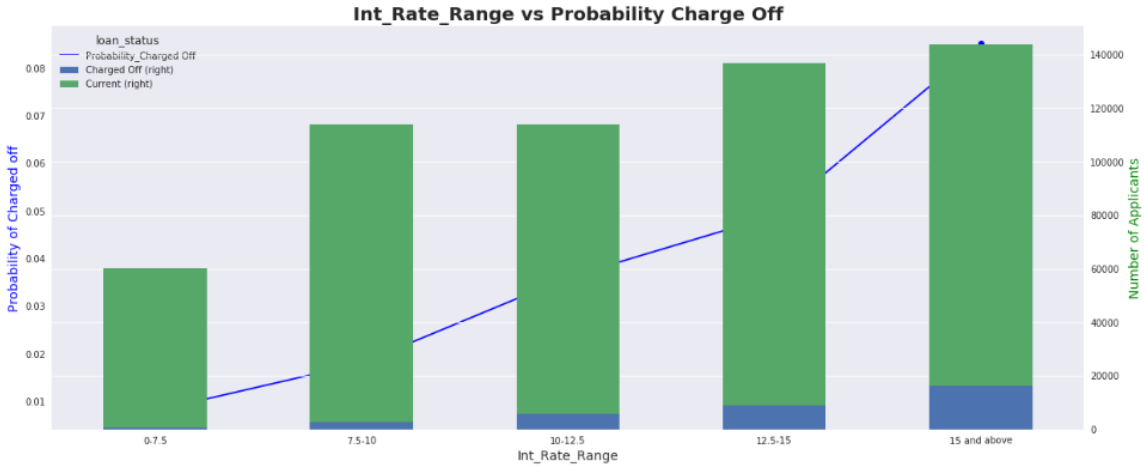


(Heatmap)



(Probability of repayment)





Summary:

- As the annual income is decreasing the probability that person will default is increasing with highest of 7% at (0 to 25000) salary bracket.
- As the interest rate is increasing the probability that person will default is increasing with highest of 9% at 15% & above bracket.
- Applicants who are self employed & less than 1 year of experience are more probable of charged off.