# Lending club case study

Nikhil Khedkar

MLC33 – September 2021

# General info

- Goal
  - Analyse loan data provided
  - Understand relationships between variables
  - Find factors which influence loan defaults
- Methodology
  - Data will first be cleaned
  - EDA will be used to analyse the data
  - No machine learning methods are used

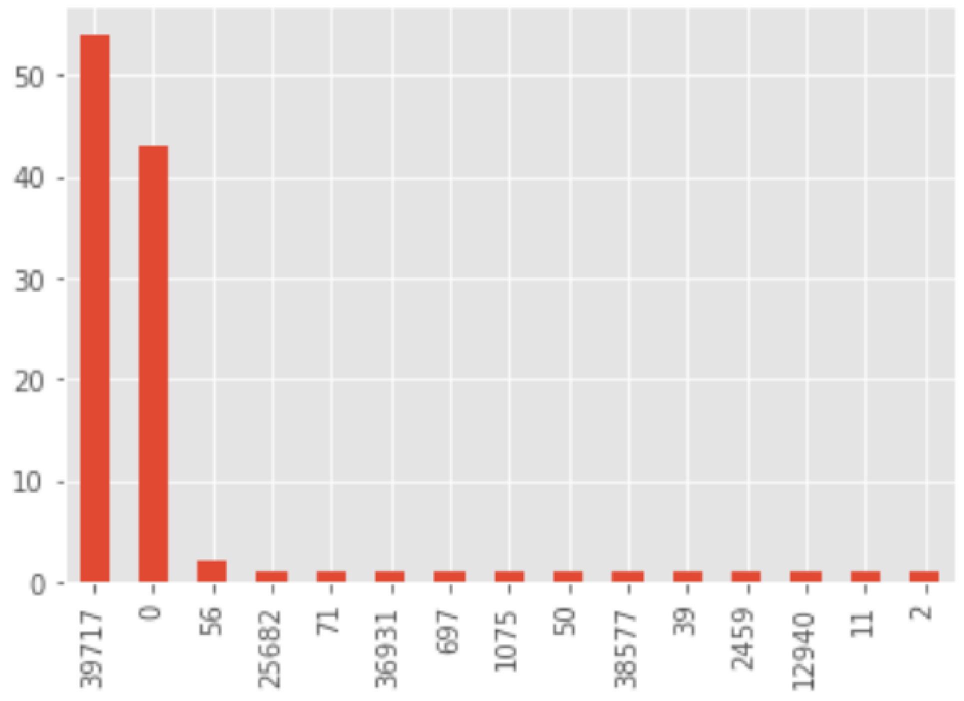
# Data info

| Item | Description |
|---|---|
| filename | Loans.csv |
| File describing data | Data_Dictionary.xlsx |
| Number of rows | 39717 |
| Number of columns | 111 |

- Data was imported into a Jupyter notebook
- All analysis was done with python

# Cleaning

- The following columns were dropped:
  - 54 columns with 39717 'nans'
  - 9 columns with all values same
  - Columns with more than 50% missing values
  - Columns not useful for prediction of default were dropped



# Missing values

- **'desc'**, **'emp_title'**, **'title'**
  - had too many unique text values. Entire columns were dropped.
- **'revol_util', 'last_pymnt_d', 'last_credit_pull_d'**
  - combined missing values were less than 0.3%
  - hence rows with missing values were dropped
- **'pub_rec_bankruptcies'**: Imputed with mode 0
- **'emp_length'**: imputed with value "unknown"

***Size of data after cleaning: (39598 rows, 37 columns)***

# Cleaning

- All date columns converted from object to datetime64
- Columns ending with % were converted to float64
- Leading and trailing spaces were removed
- Annual_inc made more readable by dividing by 1000

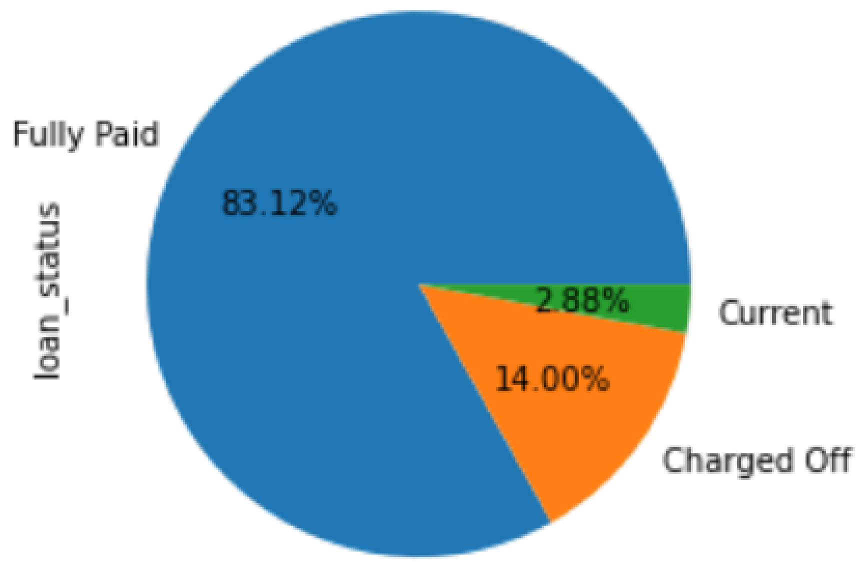
```
# Make annual_inc more readable
df1.annual_inc = df1.annual_inc / 1000
df1.annual_inc.describe()
```

```
count    39598.000000
mean        69.035085
std         63.828578
min          4.000000
25%         40.632500
50%         59.000000
75%         82.500000
max       6000.000000
Name: annual_inc, dtype: float64
```

# Univariate Analysis

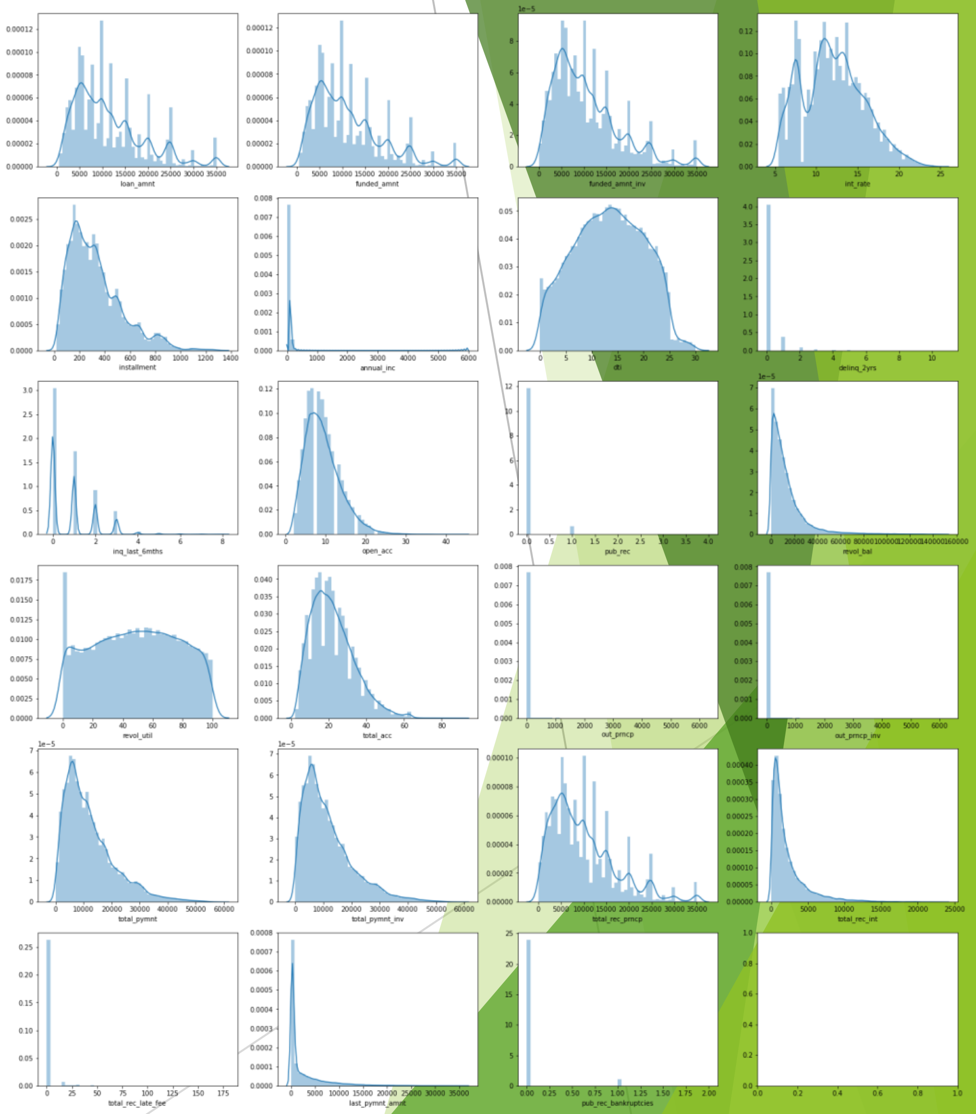
- loan_status is our target variable

```
Fully Paid      32915
Charged Off      5543
Current          1140
Name: loan_status, dtype: int64
```



# Univariate Analysis

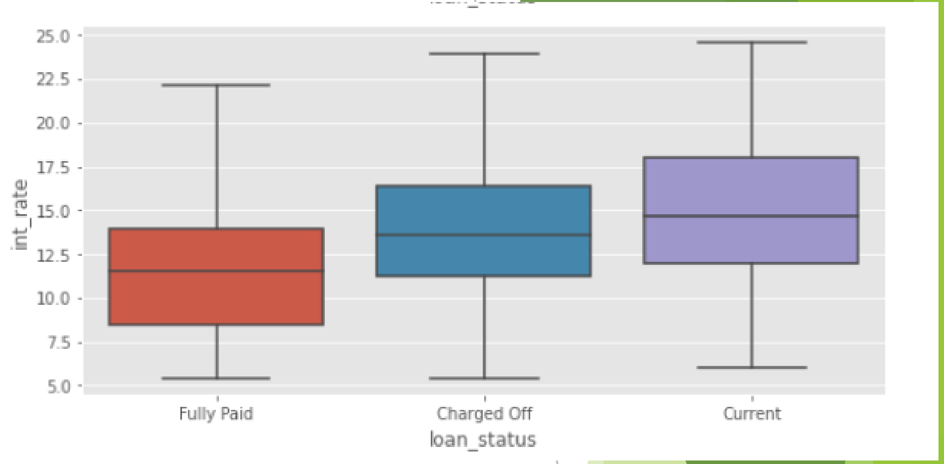
- Initial plot of continuous variables was made
  - Columns with similar distribution were dropped - *'funded_amnt'*, *'funded_amnt_inv'*
  - Columns with more than 95% same values and more than 15 unique values were dropped - *'total_rec_late_fee'*, *'out_prncp_inv'*, *'out_prncp'*
  - Total payment and total recovery cannot help in predicting default. Drop these features - *'total_pymnt'*,*'total_pymnt_inv'*,*'total_rec_int'*, *'total_rec_prncp'*



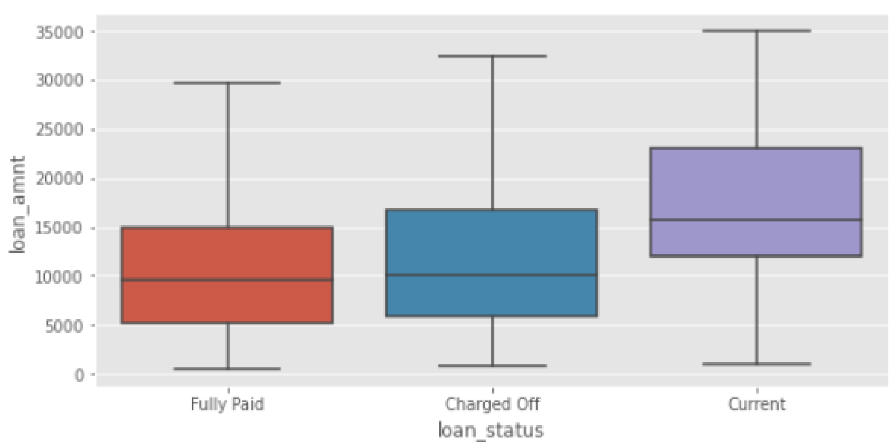
# Continuous columns – impact on default

- ['open_acc', 'last_pymnt_amnt', 'revol_bal', 'installment', 'int_rate', 'dti', 'total_acc', 'revol_util', 'loan_amnt']
- **Observations**
  - int_rate and revol_util seem to be only two variables really impacting default.
  - loan_amnt seems to have a slight impact. higher loans mean higher default.

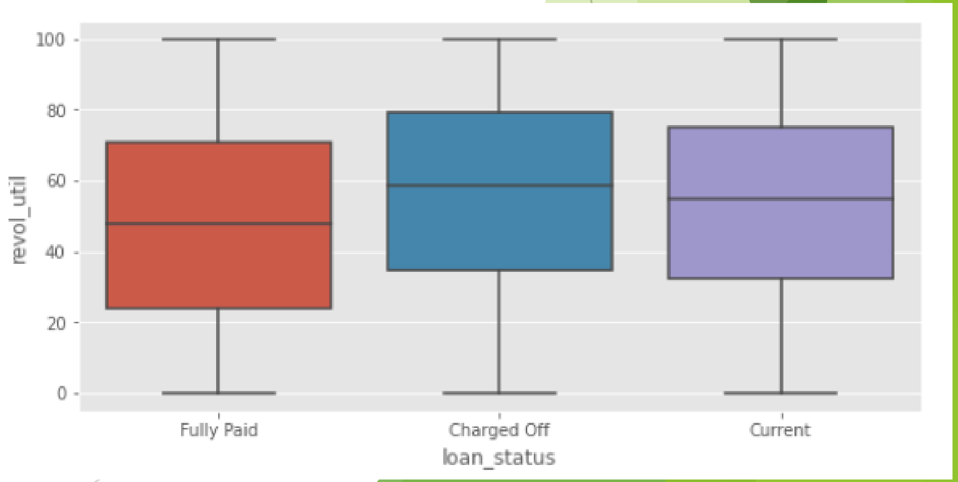
int_rate




loan_amnt




revol_bal



# Categorical columns impact

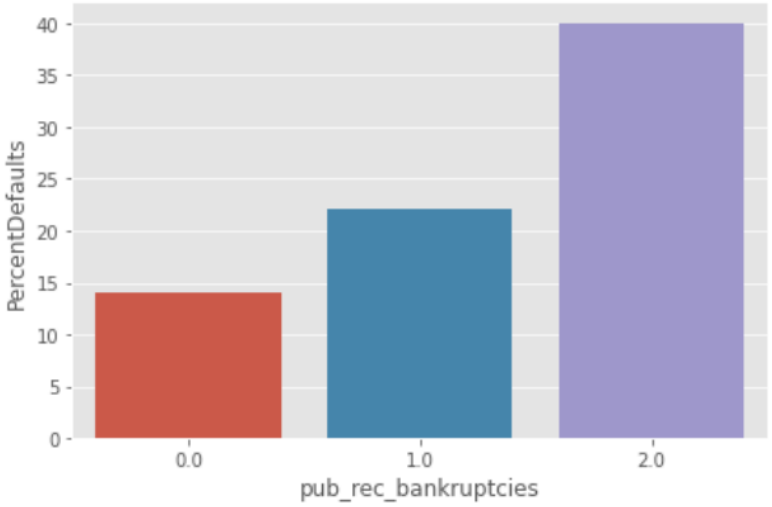
- Columns with few distinct values were treated as categorical
- `['delinq_2yrs', 'inq_last_6mths', 'pub_rec',`
- `'pub_rec_bankruptcies', 'annual_inc']`

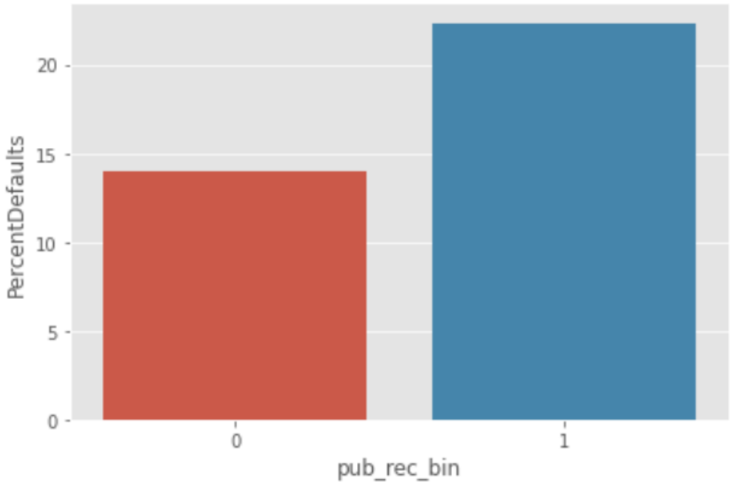
- **Observations**
- **All variables treated as categorical seem to have direct impact on defaults.**
- 'pub_rec_bankruptcies' have only 3 distinct values
- 'pub_rec_bin', 'delinq_2yrs', 'inq_last_6mths' were converted to fewer categories to make analysis easier

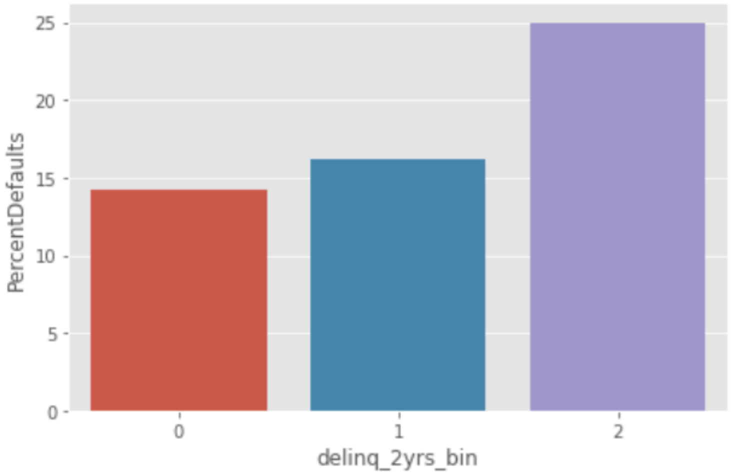
**Percent defaults** is used as the metric to analyze effect of categorical variables

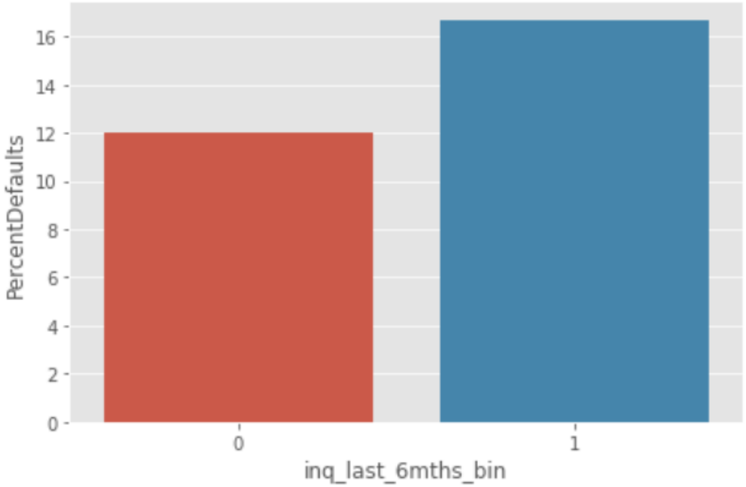
**PercentDefaults = [ 'Charged Off' / ('Charged Off' + 'Fully Paid') ] * 100**

# Categorical columns impact



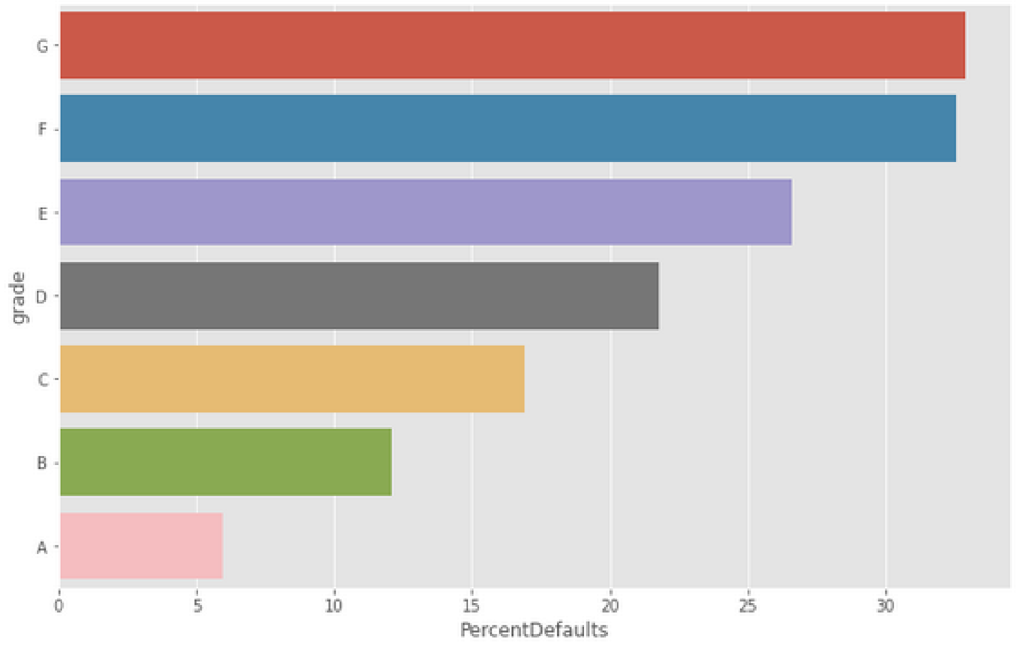



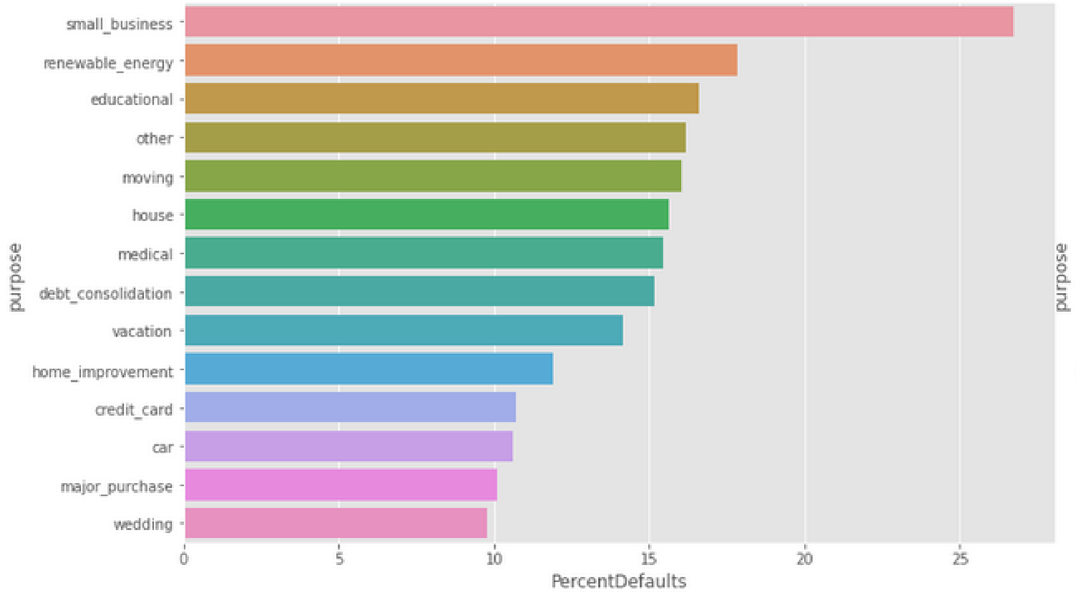
# Categorical columns impact

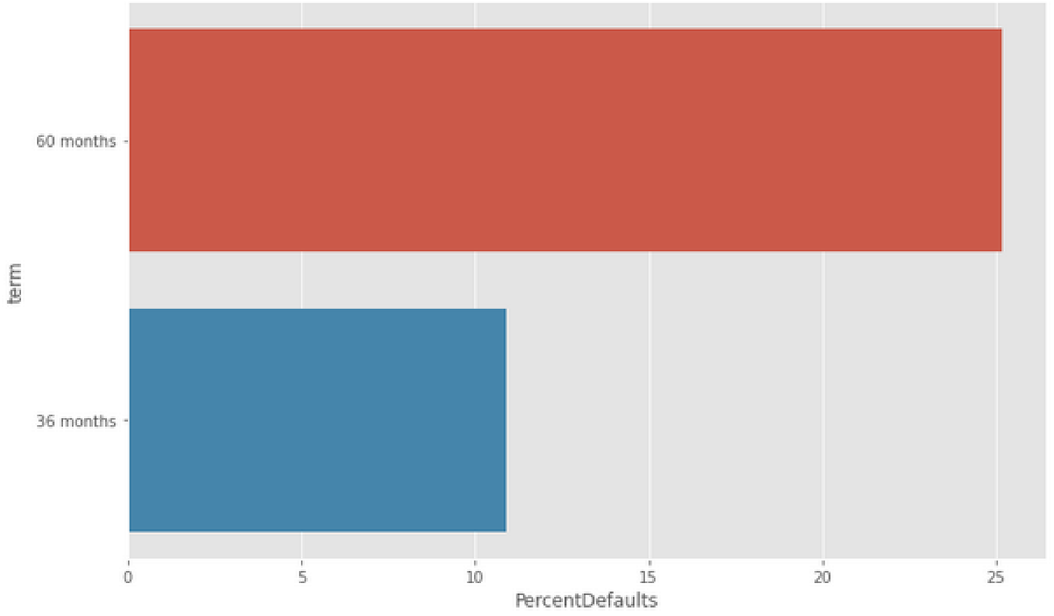
- Other Categorical variables
- ['term',
- 'grade',
- 'emp_length',
- 'verification_status',
- 'home_ownership',
- 'purpose']
- zip_code is part of state so dropped
- sub_grade is part of grade so dropped

# Categorical columns impact

- **Grade:** Better the Loan grade lower is the default percentage.
- **Term:** Shorter loans have much smaller default rates.
- **Purpose:** Small business loans have the highest rate of defaults.

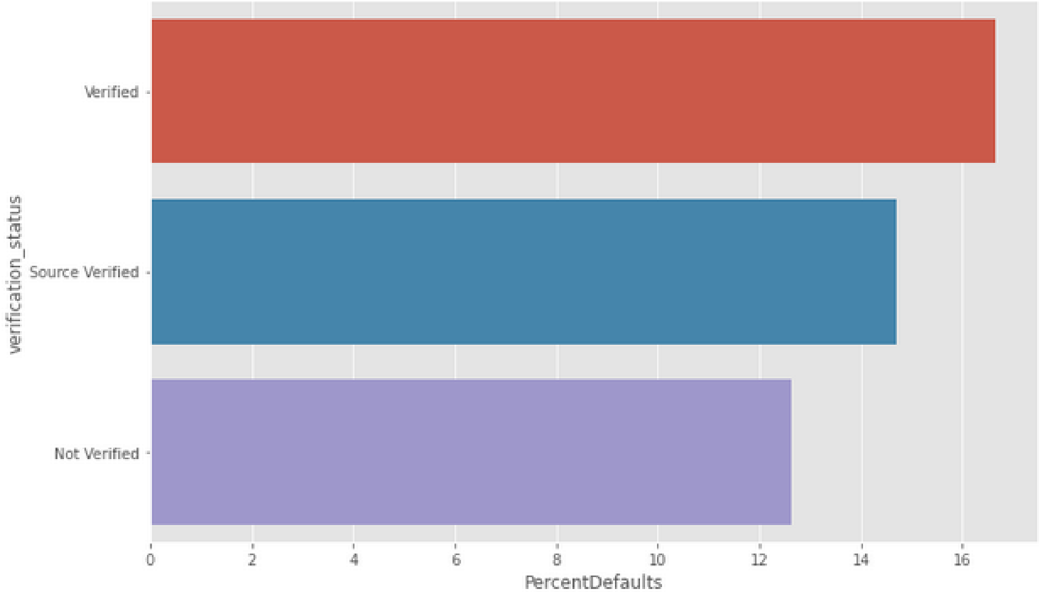


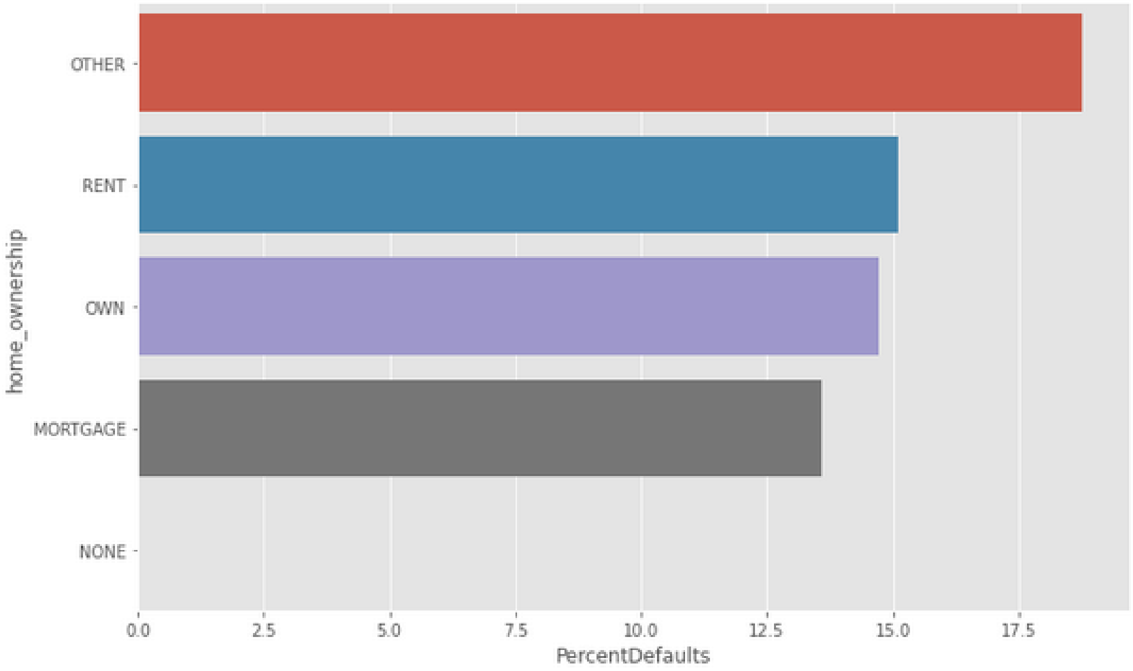



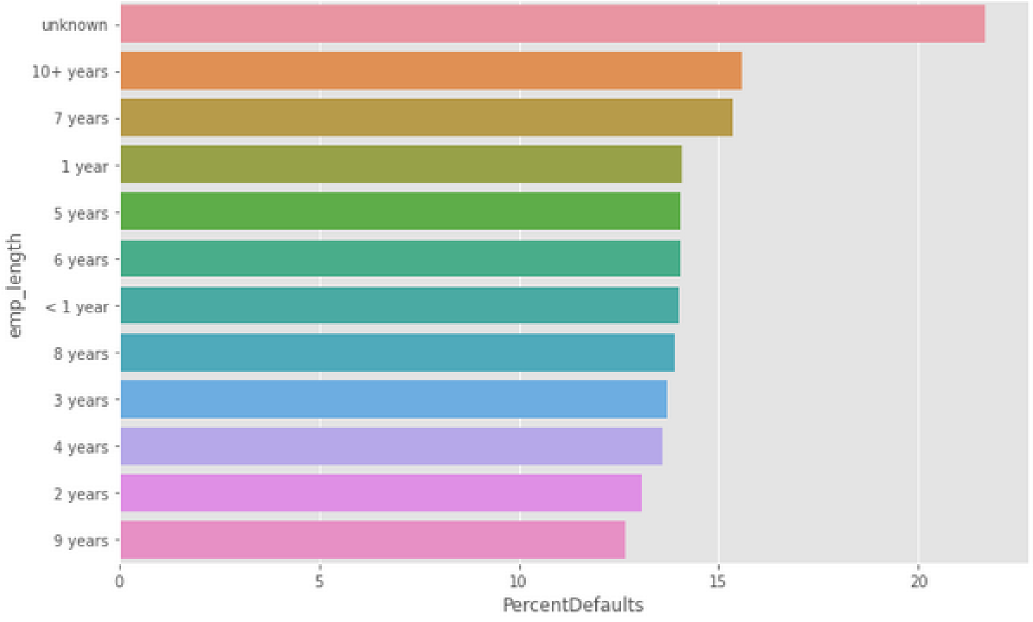



# Categorical columns impact

- **Verification:** Strangely, Verified loans have high default rates. This needs further investigation.
- **Emp_length:** If employment length is "unknown" the rate of default is high - needs further investigation.
- **Home ownership:** "Other" has high defaults, but number of loans in this are very low



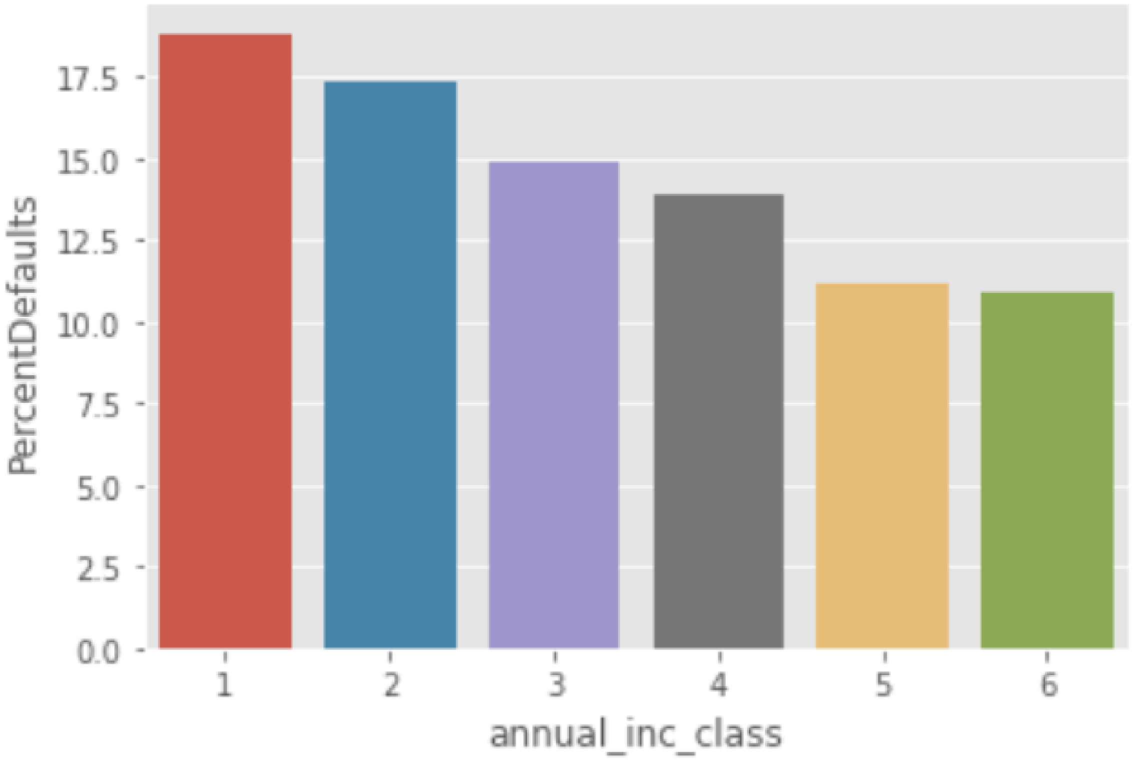



# Categorical columns impact

- Convert annual income into a categorical
  - Annual income is very skewed and hence difficult to analyse without dropping outliers.
  - We convert it to a categorical with 6 levels using quantiles [ 0, .05, .25, .5, .75, 0.95, 1. ]

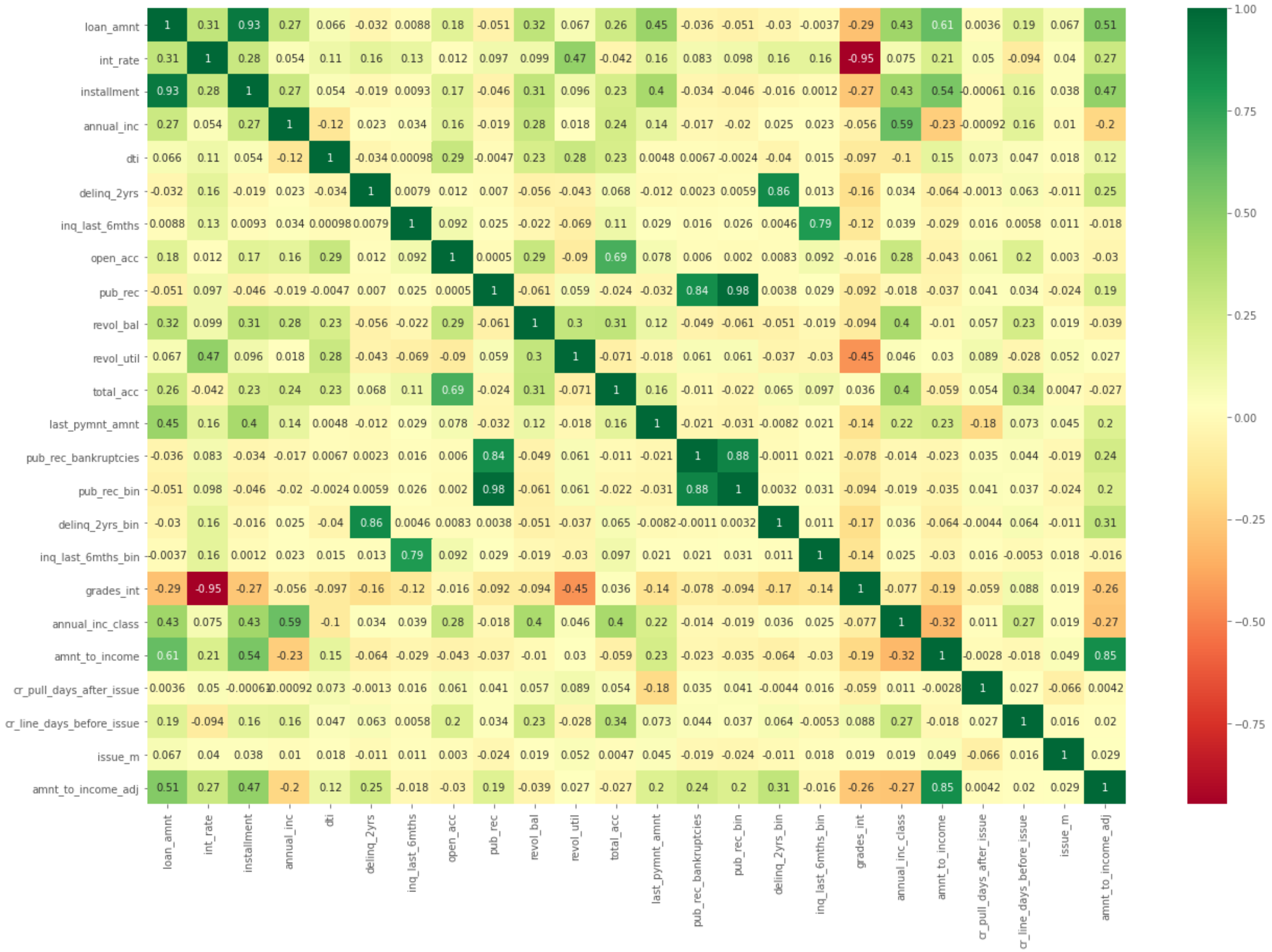
It is clearly seen that people with higher annual incomes have lower default rates.



# Bivariate analysis

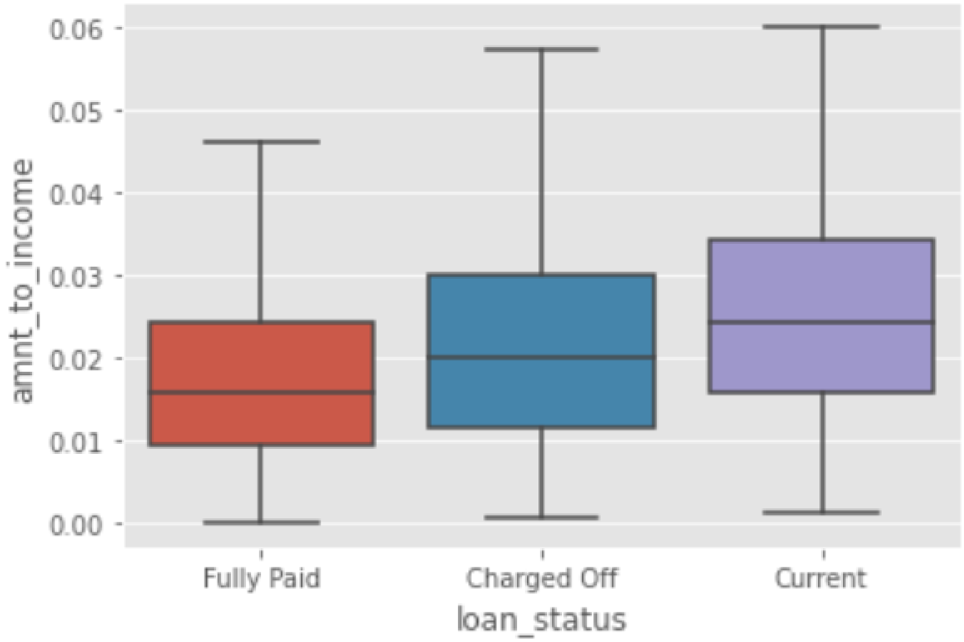
## Observations

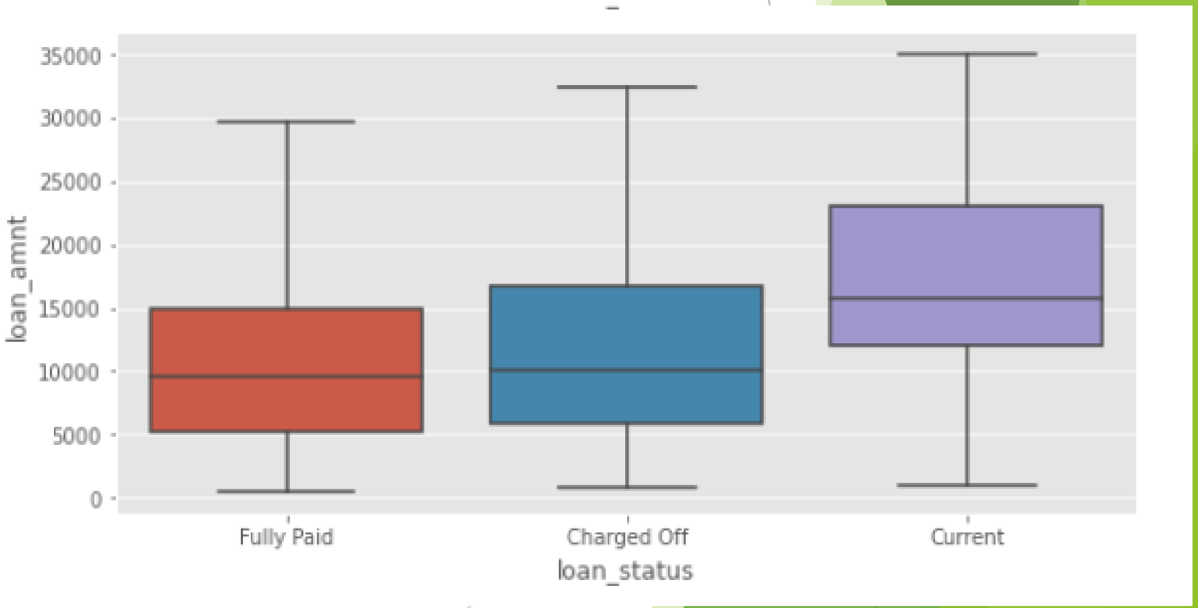
- pub_rec_bin and pub_rec_bankruptcies are highly correlated - can ignore one with lower number of non-zero values
- loan_amnt and installment are highly correlated - can ignore installment
- grades have high negative correlation with int_rate - Can ignore grade or int_rate



# Derived Metric

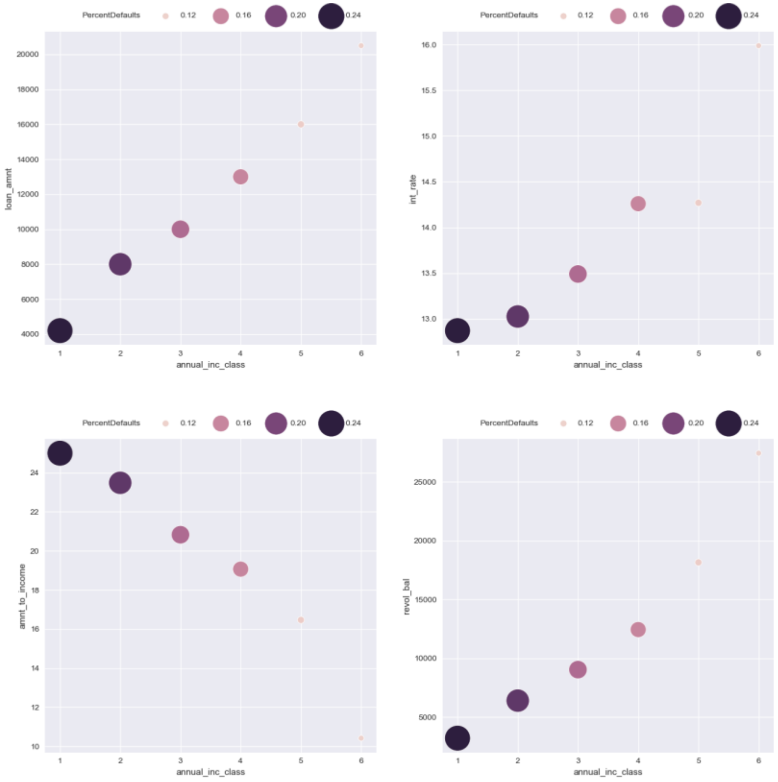
- Loan Amount as percentage of annual income
- This shows slightly better distinction than just loan_amnt
- Customers who default have higher loans as percentage of income

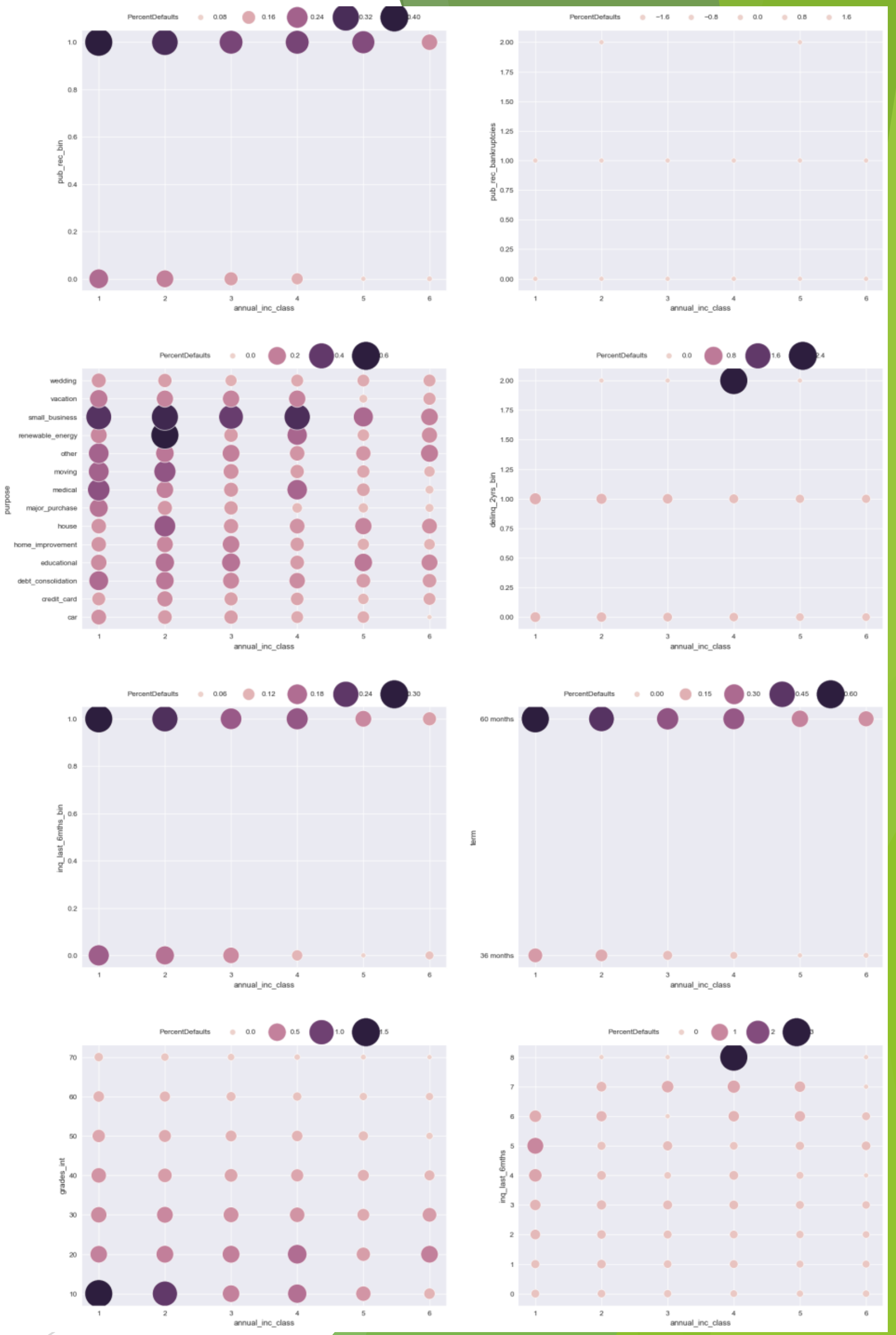


# 2-factor effect on default rate

- 2 factor effect on defauts was plotted
- Annual income class was treated as x axis



# Final list of features affecting defaults

| Feature | Description | Dangerous level - Can lead to higher default rate |
|---|---|---|
| annual_inc_class | Annual income categorical – derived from annual_inc | Under 25th percentile |
| amnt_to_income | Loan amount as % of annual income | Above median |
| loan_amnt | Amount of loan taken | Above Median |
| int_rate | Interest rate | Above Median |
| pub_rec_bin | Public derogatory records - 0 or 1 – derived from pub_rec | Above 0 |
| purpose | Why the loan was taken | Small_business, Other |
| addr_state | State | 'NE', 'NV', 'SD', 'AK', 'FL', 'MO', 'OR' |
| delinq_2yrs_bin | delinquencies in last 2 years – derived from delinq_2yrs | Above 4 |
| inq_last_6mths_bin | inquiries in last 2 years – derived from inq_last_6mths | Above 0 |
| term | Loan term | 60 months |

# High default combinations

Default rate in data is 14.4%

| Combination for default rate higher than 14.4% | Default rate |
|---|---|
| annual_inc_class < 25th percentile<br>int_rate > median<br>small_business loan | 41.66% |
| annual_inc_class < 25th percentile<br>int_rate > median<br>small_business loan<br>term = 60 months | 54.43% |
| annual_inc_class < 25th percentile<br>amnt_to_income > median<br>small_business loan<br>term = 60 months | 48.78% |
| annual_inc_class < 25th percentile<br>revol_bal > median<br>small_business loan | 43.75% |

# Conclusion

- EDA was performed to understand loan defaults
- Univariate, bivariate and multi-level interactions were studied
- Factors affecting defaults were identified
- Factor combinations that lead to higher defaults were identified
- No machine learning methods were used
- further analysis with a method such as gradient boosting my be useful