Mile Stone Report

Introduction:

Peer-to-peer lending is when borrowers take out loans from companies that pair potential borrowers with individual investors that are willing to lend them their own money. The individual investors decide after reading a profile whether or not they want to take the risk of loaning money to the potential borrower. Potential lender investors can agree to loan part – or all – of the money the borrower is asking for. Most peer-to-peer (also called P2P) loans are funded by several different investors, and as the loan payment is made each month, a portion of the payment goes back to each of the different investors involved with the loan.

Lending Club is the world's largest online marketplace connecting borrowers and investors. It enables borrowers to create unsecured personal loans and investors can search and browse the loans. Investors select loans based on the borrower's information, loan amount, loan grade, loan purpose and make money from the interest.

There is a risk to the investor is if the borrower misses the payments. After borrowers miss a loan payment, their loan will move from "current" to "late" status. When borrowers miss several payments, the loan will enter "default" status and, when there is no longer a reasonable expectation of further borrower payments, the loan will be "Charged-off."

If the risk is predicted and provided to the investor during the selection of the loans then it helps the investor to make a better decision whether to fund the loan. The risk of loan default can be predicted by building a model using the existing machine learning algorithms and Lending club historical dataset.

Problem Statement:

Create a model to predict if the loan defaults using Lending Club historical loan data.

Approach:

1.Data Collection: Extract datasets from Lending club publicly available dataset

2.Data Preprocessing: Apply data wrangling techniques to clean and organize the data.Ex: handling missing data

3.Exploratory Data Analysis: Explore data by applying visual and numerical techniques to find insights in the data, extract important features, detect outliers and anomalies

4.Modeling: Build different models by applying machine learning algorithms on the training dataset

5.Evaluation: Evaluate the performance of each model using test data.

Data Wrangling:

Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics

The original dataset contains 1321864 entries and 151 features

```
loans_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1321864 entries, 0 to 103547
Columns: 151 entries, id to settlement_term
dtypes: float64(112), object(39)
memory usage: 1.5+ GB
```

The steps to be performed in Data Cleaning can be divided into three categories

1. Columns to Retain

- Drop the columns which have more than 30% missing values.
- Drop the columns which are not applicable to the analysis (first pass of the dataset). The features like **id** and **url** are the information about LC which we don't need for the analysis are dropped from the dataset.
- Drop the columns that have only one unique value. If all the values in the column are same, then that feature would not be useful for the analysis.

2. Rows to retain

- Drop the rows which has null values in each column

3. Handling Missing Values

Missing values can lead to wrong prediction or classification and can also cause a high bias for any given model being used. The missing values are filled differently for categorical and numerical columns.

- For numeric attributes, replace the missing or null values with the mean of the respective column the missing value belongs to
- For categorical attributes, replace the missing or null values with the mode of the respective column (most frequent value) the missing value belongs to

```
In [38]: filtered_numeric_loans_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1321847 entries, 0 to 103545
Data columns (total 66 columns):
```

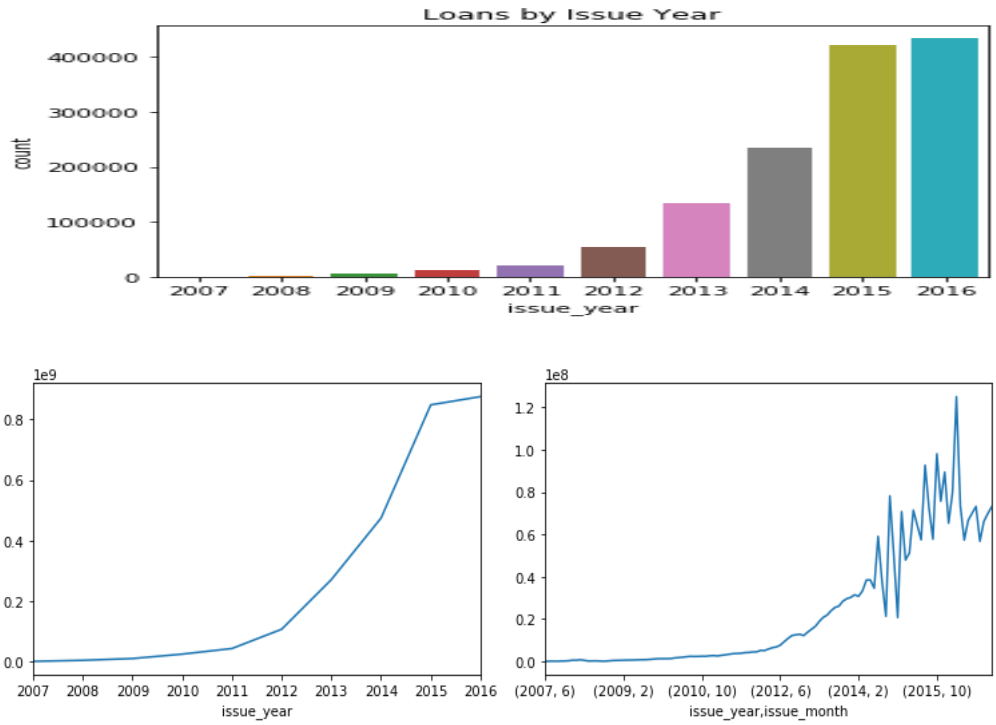
After performing data wrangling techniques, the dataset size is reduced to 1321847 and 66 columns

Exploratory Data Analysis:

Exploratory Data Analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods

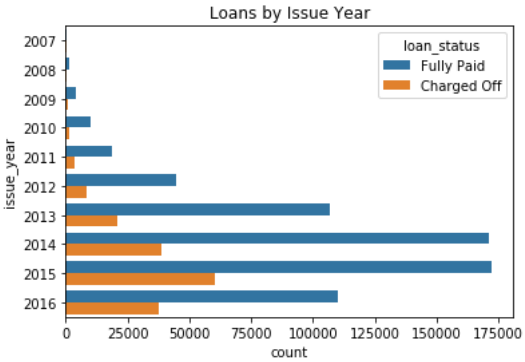
Let's answer few questions about the dataset by performing exploratory data analysis.

1. Which year has the highest number of loans issued?



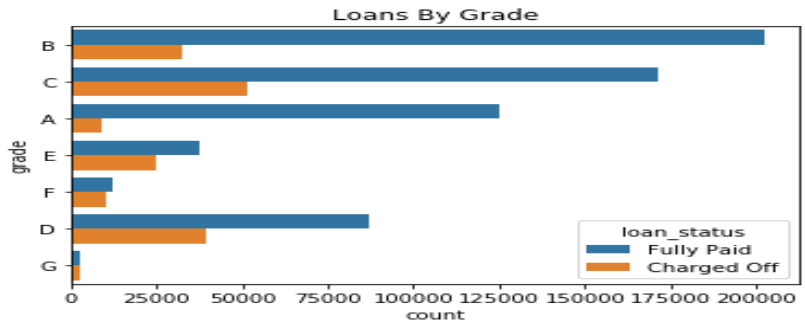
Number of Loans issued by Lending Tree have increased from 2007 to 2016. The highest number of loans are issued in the year 2015 and 2016.The number of loans are almost doubled in the year 2015 and 2016 in comparison to the loans in 2013 and 2014.

2. Which year with most charged off loans?



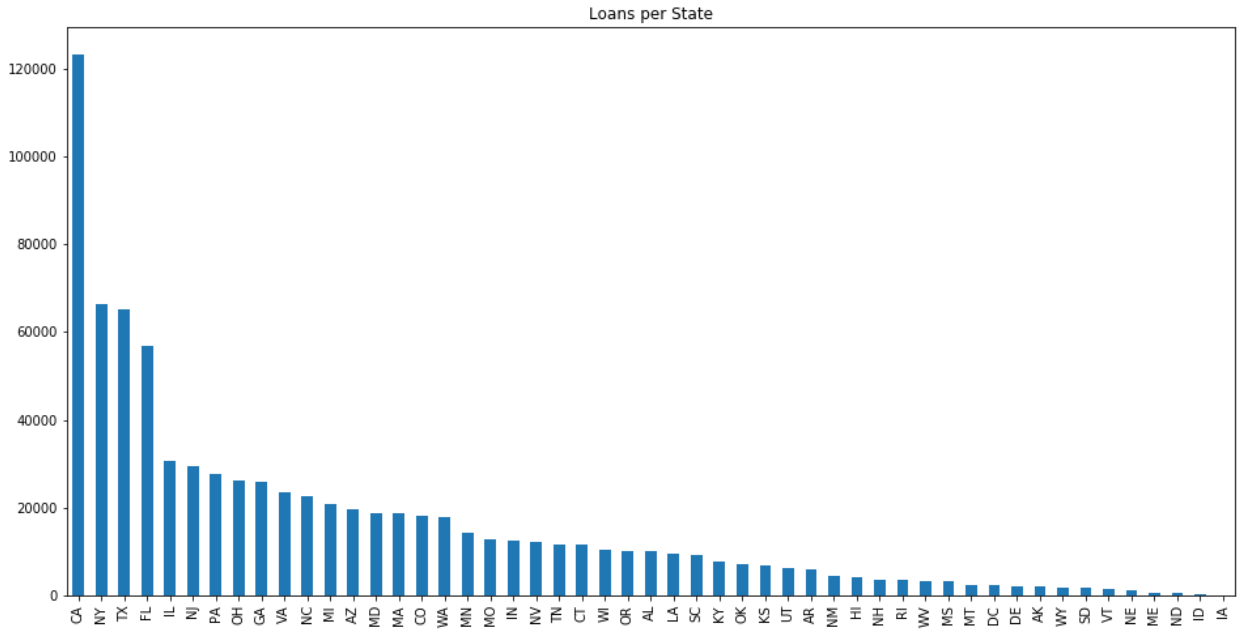
Number of maximum charged off loans are in the year 2015.Number of maximum fully paid loans are in the year 2014 and 2015

3.Which grade loans are charged off the most?



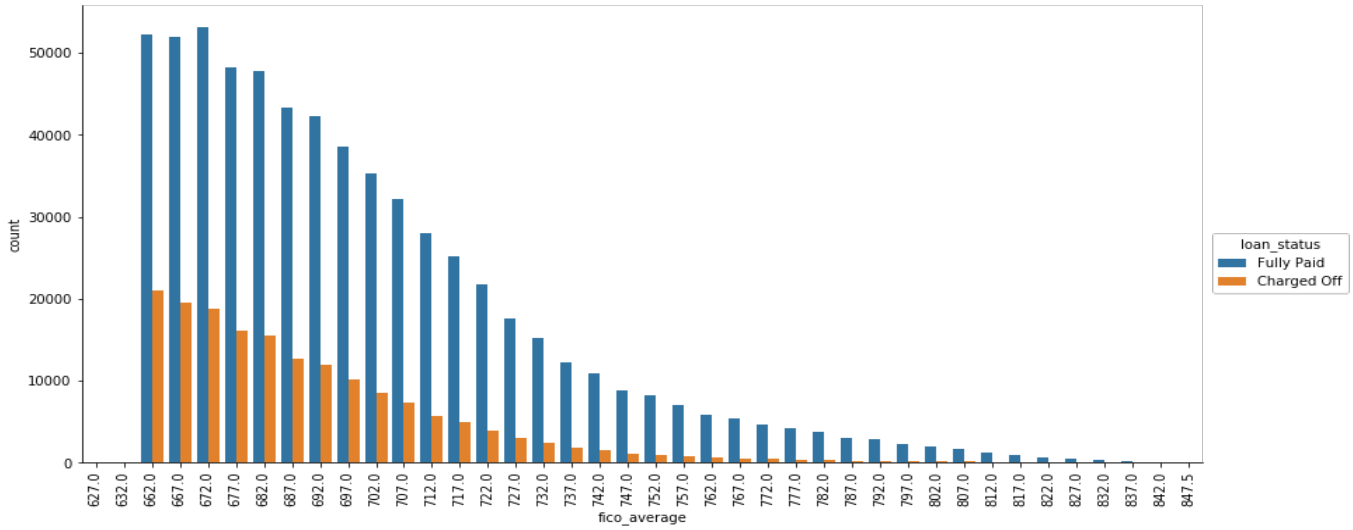
The maximum charged off loans are of grade C loans. Even the loan with good grades like A and B have significant charged off loans. Number of loans issued for F and G are very less but more than half the loans are charged off.

4.Loans across different states in United States?



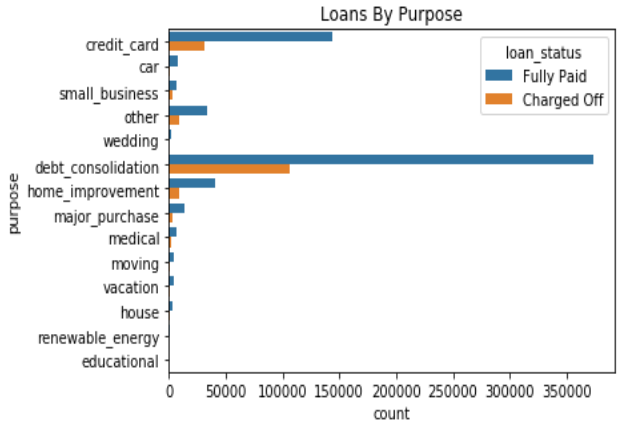
Lending Tree approves loans in most of the states in United states. The maximum number of loans are issued in California , Newyork, Texas and Florida. The lease number of loans are issed in the states Maine,North Dakota, Idaho and Iowa.

5. What is the minimum fico score to get a loan from Lending Tree?



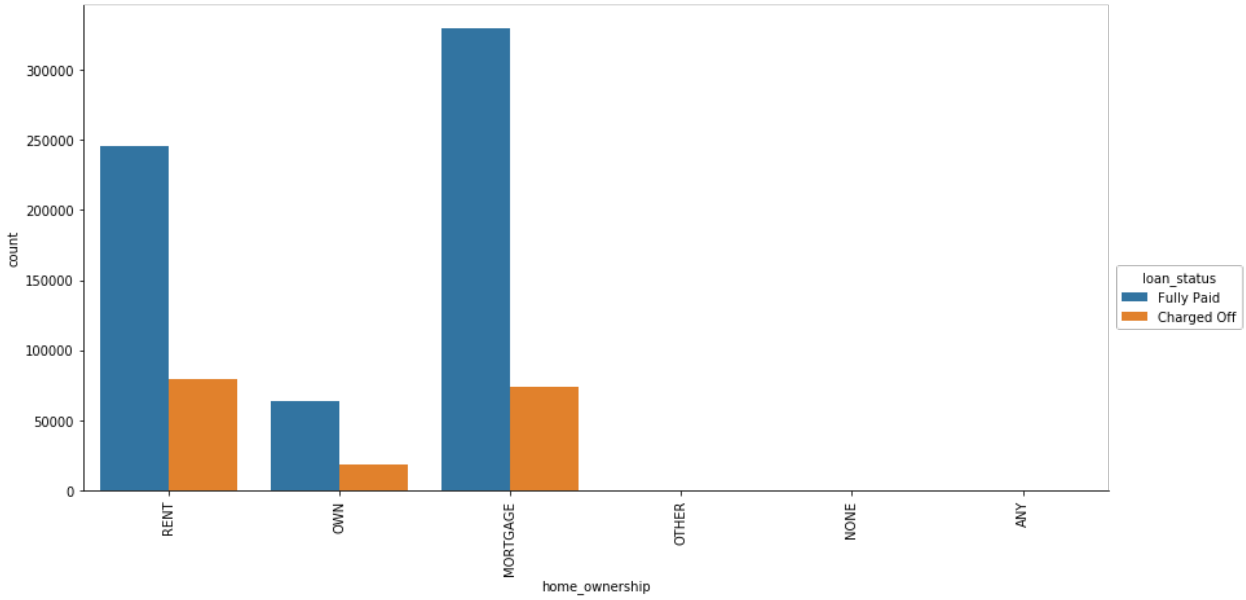
The minimum fico score is atleast 650 to get a loan from LendingTree. Higher the fico score the lesser the charged off loans. It is evident that the loans with low fico score charged off the most. The loans with high fico score also have charged off status. so there may be other features that are causing the loans to charge off.

6. Which loan purpose has the highest number of charged off loans?



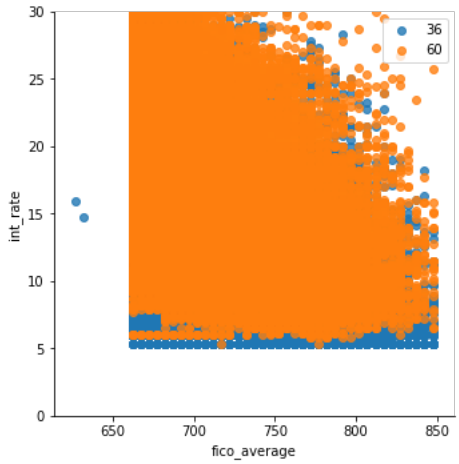
The above graph shows that the loans with purpose "Debt consolidation" and "credit card" have highest charged off rate.

7. Which homeownership status has the highest number of charged off loans



The applicants who have mortgage have been issued the most loans. The charged off rate for the loans with home ownership rent are slightly less than the loans with mortgage.

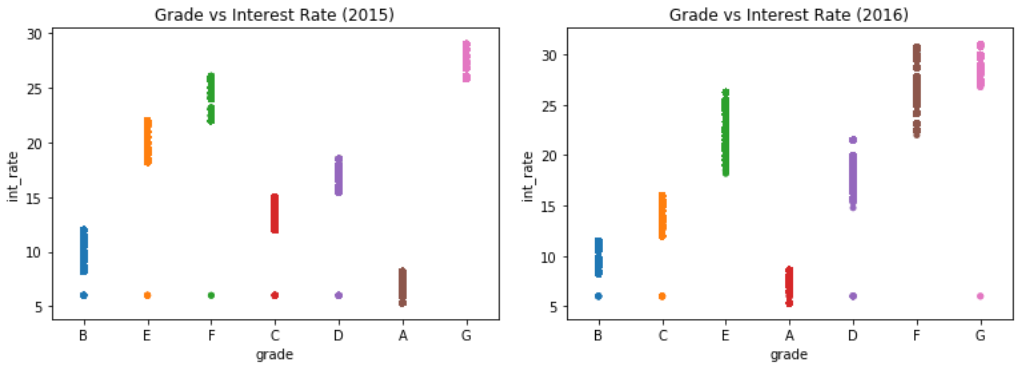
8. Fico Score Vs Interest Rate



Fico score and interest rate are negatively correlated. Lesser the fico score the higher the interest rate. The applicants with higher fico have mostly opted for 36 loan term

Inferential Statistics :

1. Is there any significant change in the interest rate for the year 2015 and 2016 for grade E loans ?

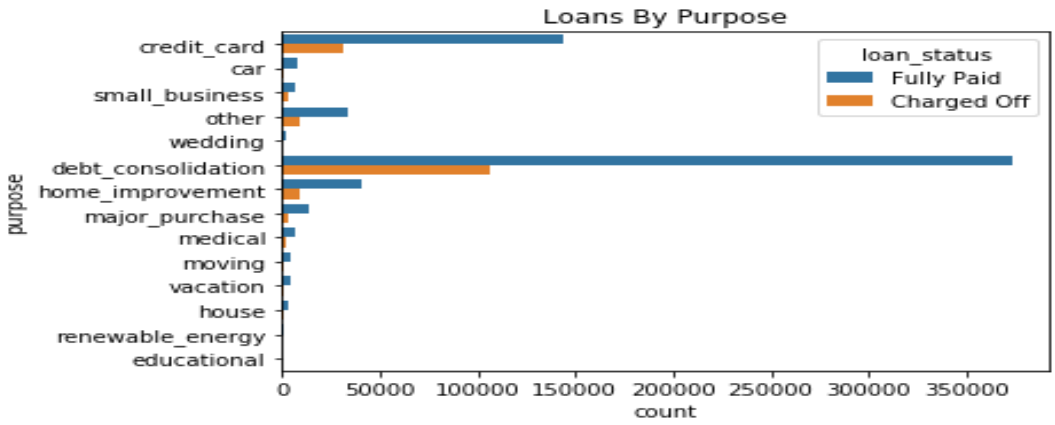


Null Hypothesis: There is no difference in the interest rates for grade E loans in 2015 and 2016.
Alternate Hypothesis : There is difference in the interest rates for grade E loans in 2015 and 2016.

```
In [18]: #Difference in the means
         print("Difference in the interest rates:", int_rate_2015.mean() - int_rate_2016.mean())

         #t-test
         stats.ttest_ind(int_rate_2015, int_rate_2016)

         Difference in the interest rates: -2.98085786549305

Out[18]: Ttest_indResult(statistic=-179.11915103758525, pvalue=0.0)
```

p-value < 0.05, so we can reject the null hypothesis in favor to alternate hypothesis. So there is a significant difference in the interest rates for the years 2015 and 2016 for grade E loans. There is an increase of 2.98 interest rate for grade E loans.

2. Does loan purpose has the significant impact on the charged off rate?



From the above, we can see that the loans with debt consolidation as purpose have highest number of loans issued and charged off loans as well. It is evident from the graph that there is relationship between the purpose and loan status.

Null Hypothesis: Loans Purpose has no significant association with the loan status
Alternate Hypothesis: Loans Purpose has no significant association with the loan status

The chi-square test of independence is a statistical test used to determine whether two categorical variables are independent of each other or not.

In [20]:
```
from scipy.stats import chi2_contingency
def chisq_of_df_cols(df, c1, c2):
    groupsizes = df.groupby([c1, c2]).size()
    ctsum = groupsizes.unstack(c1)
    print(ctsum)
    # fillna(0) is necessary to remove any NAs which will cause exceptions
    return(chi2_contingency(ctsum.fillna(0)))
print(chisq_of_df_cols(filtered_loans_df, 'loan_status' , 'purpose'))
```

```
loan_status         Charged Off  Fully Paid
purpose
car                        1320        7386
credit_card               31914      143734
debt_consolidation       106717      373614
educational                 56         270
home_improvement           9312       40252
house                       895        3015
major_purchase             3222       13809
medical                    2004        6486
moving                     1383        4163
other                      9567       33185
renewable_energy            150         445
small_business             3103        7024
vacation                   1007        3871
wedding                     277        1996
(2461.2539968957035, 0.0, 13, array([[  1.83674736e+03,   6.86925264e+03],
       [  3.70573167e+04,   1.38590683e+05],
       [  1.01337778e+05,   3.78993222e+05],
       [  6.87778127e+01,   2.57222187e+02],
       [  1.04567592e+04,   3.91072408e+04],
       [  8.24911803e+02,   3.08508820e+03],
       [  3.59311328e+03,   1.34378867e+04],
       [  1.79117678e+03,   6.69882322e+03],
       [  1.17006672e+03,   4.37593328e+03],
       [  9.01959831e+03,   3.37324017e+04],
       [  1.25530057e+02,   4.69469943e+02],
       [  2.13654267e+03,   7.99045733e+03],
       [  1.02913549e+03,   3.84886451e+03],
       [  4.79545915e+02,   1.79345408e+03]]))
```

The p-value is 0.0 , we reject the null hypothesis in the favor of alternate hypothesis. So, there is a statistical significant association between loan purpose and the loan status.