“TO LEND OR NOT TO LEND?”

Exploratory data analysis on Kaggle’s lending club loan data by Gregory Ausu

KAGGLE LENDING CLUB LOAN PROPOSAL

Problem:

Borrowers defaulting on their loans is a growing problem in the U.S. According to CNBC by 2023, nearly 40 percent of borrowers are expected to default on their student loans. Also, according to the S&P Dow Jones Indices and Experian, December of 2018 marked the first time since January of 2017 that all loan types and all metropolitan statistical areas saw an increase in sequential default rates. These issues are further compounded by auto loan defaults being at record highs. Using exploratory data analysis to accurately predict and/or prevent loan defaults or charge off's is of great importance. This will be of great help to traditional financial institutions as well as growing fintech and mobile banking companies.

Intent:

To extract and perform exploratory data analysis to assess markers that indicate the risk of a loan being charged off.

Hypothesis:

Features in this dataset such as employment title, loan amount, interest rate, and location will be used to analyze the risk of a loan being charge off.

DATA SOURCES

- https://www.kaggle.com/wendykan/lending-club-loan-data
- https://www.bls.gov/ooh/education-training-and-library/career-and-technical-education-teachers.htm
- https://www.cnbc.com/2018/08/13/twenty-two-percent-of-student-loan-borrowers-fall-into-default.html
- https://www.citylab.com/transportation/2019/02/subprime-car-loans-buy-automobile-lending-debt-trap/582652/
- https://github.com/gausu/EDA-Projects

DATA IMPORTS

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib import rcParams

loan = pd.read_csv('loan.csv')



%matplotlib inline

rcParams['figure.figsize'] = 16,6
```

```
.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 887379 entries, 0 to 887378
Data columns (total 74 columns):
id                              887379 non-null int64
member_id                       887379 non-null int64
loan_amnt                       887379 non-null float64
funded_amnt                     887379 non-null float64
funded_amnt_inv                 887379 non-null float64
term                            887379 non-null object
int_rate                        887379 non-null float64
installment                     887379 non-null float64
grade                           887379 non-null object
sub_grade                       887379 non-null object
emp_title                       835917 non-null object
emp_length                      842554 non-null object
home_ownership                  887379 non-null object
annual_inc                      887375 non-null float64
verification_status             887379 non-null object
issue_d                         887379 non-null object
loan_status                     887379 non-null object
pymnt_plan                      887379 non-null object
url                             887379 non-null object
desc                            126028 non-null object
purpose                         887379 non-null object
title                           887227 non-null object
zip_code                        887379 non-null object
addr_state                      887379 non-null object
dti                             887379 non-null float64
delinq_2yrs                     887350 non-null float64
earliest_cr_line                887350 non-null object
inq_last_6mths                  887350 non-null float64
mths_since_last_delinq          433067 non-null float64
mths_since_last_record          137053 non-null float64
open_acc                        887350 non-null float64
pub_rec                         887350 non-null float64
revol_bal                       887379 non-null float64
revol_util                      886877 non-null float64
```

```
loan.shape
```

```
(887379, 74)
```

```
loan.head()
```

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	total_bal_il	il_util	open_rv_12m
0	1077501	1296599	5000.0	5000.0	4975.0	36 months	10.65	162.87	B	B2	...	NaN	NaN	NaN
1	1077430	1314167	2500.0	2500.0	2500.0	60 months	15.27	59.83	C	C4	...	NaN	NaN	NaN
2	1077175	1313524	2400.0	2400.0	2400.0	36 months	15.96	84.33	C	C5	...	NaN	NaN	NaN
3	1076863	1277178	10000.0	10000.0	10000.0	36 months	13.49	339.31	C	C1	...	NaN	NaN	NaN
4	1075358	1311748	3000.0	3000.0	3000.0	60 months	12.69	67.79	B	B5	...	NaN	NaN	NaN

5 rows × 74 columns

EXPLORATORY DATA ANALYSIS

- loan.head() reveals that this dataset has 5 rows and 74 columns.
- loan.info() reveals that there are 887,379 entries over the 74 columns. The datatypes consist of 49 floats, 2 integers, and 23 objects also revealing that there a large number of null values.
- loan.shape combines and confirms the above information

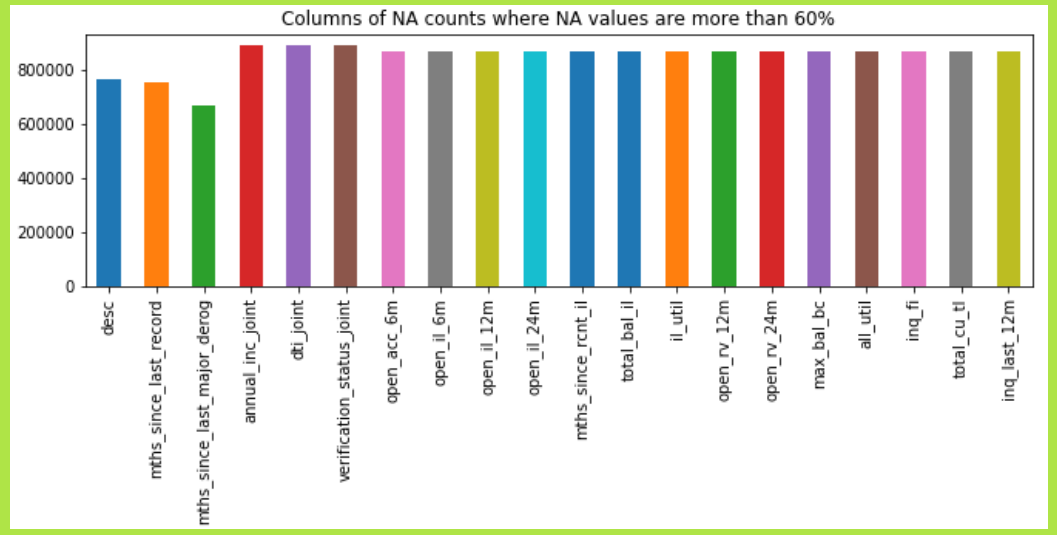
EXPLORATORY DATA ANALYSIS(NULL VALUES)

```
loan.isnull().sum()
```

id	0
member_id	0
loan_amnt	0
funded_amnt	0
funded_amnt_inv	0
term	0
int_rate	0
installment	0
grade	0
sub_grade	0
emp_title	51462
emp_length	44825
home_ownership	0
annual_inc	4
verification_status	0
issue_d	0
loan_status	0
pymnt_plan	0
url	0
desc	761351
purpose	0
title	152
zip_code	0
addr_state	0
dti	0
delinq_2yrs	29
earliest_cr_line	29
inq_last_6mths	29
mths_since_last_delinq	454312
mths_since_last_record	750326

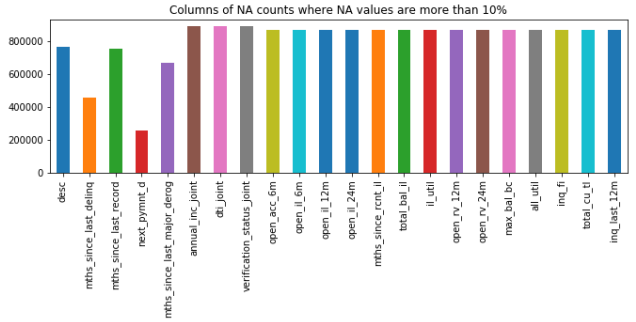
- loan.is.null.sum() REVEALS A LARGE NUMBER OF NULL VALUES.

EXPLORATORY DATA ANALYSIS(NULL VALUES)




```
nan_col = loan.isnull().sum()
nan_col = nan_col[nan_col.values >(0.6*len(loan))]
plt.figure(figsize=(11,3))
nan_col.plot(kind='bar')
plt.title('Columns of NA counts where NA values are more than 60%')
plt.show()
```

- Null values that are greater than 60% in this dataset.




```
nan_col = loan.isnull().sum()
nan_col = nan_col[nan_col.values >(0.1*len(loan))]
plt.figure(figsize=(11,3))
nan_col.plot(kind='bar')
plt.title('Columns of NA counts where NA values are more than 10%')
plt.show()
```

EXPLORATORY DATA ANALYSIS(NULL VALUES)

- Null values that are greater than 10%

EXPLORATORY DATA ANALYSIS(NULL VALUES)

- Of the selected features being used to assess the risk of a loan being “charged off”, employment title is the only feature that has null values.

```
loan.notnull().sum()
```

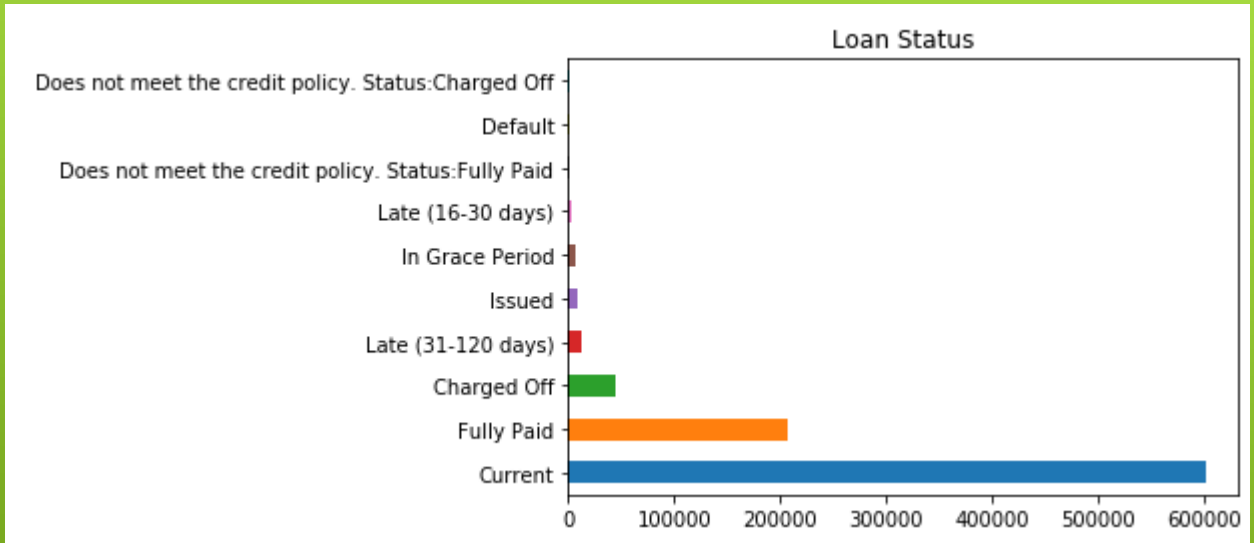
id	887379
member_id	887379
loan_amnt	887379
funded_amnt	887379
funded_amnt_inv	887379
term	887379
int_rate	887379
installment	887379
grade	887379
sub_grade	887379
emp_title	835917
emp_length	842554
home_ownership	887379
annual_inc	887375
verification_status	887379
issue_d	887379
loan_status	887379
pymnt_plan	887379
url	887379
desc	126028
purpose	887379
title	887227
zip_code	887379
addr_state	887379

LOAN STATUS

- 69% of loans are current
- 24% of loans are fully paid including 'Does not meet credit policy. Status: Fully Paid'
- 5% of loans have been charged off including 'Does not meet credit policy. Status: Charged off'
- <1% of the borrowers have defaulted
- 2% of the loans are late(16-120 days)

```
loan.loan_status.value_counts()
```

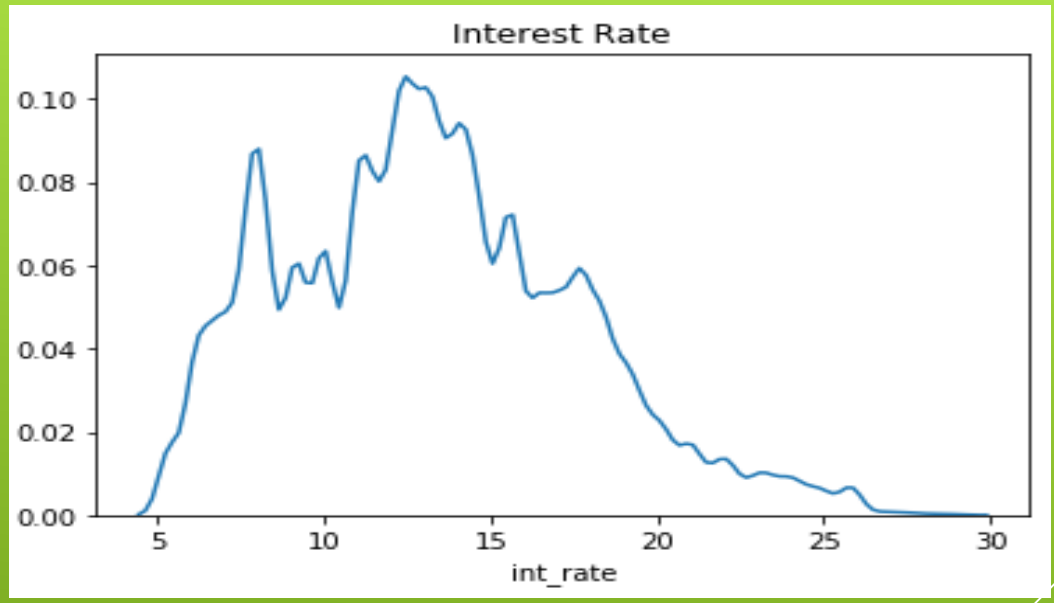
Current	601779
Fully Paid	207723
Charged Off	45248
Late (31-120 days)	11591
Issued	8460
In Grace Period	6253
Late (16-30 days)	2357
Does not meet the credit policy. Status:Fully Paid	1988
Default	1219
Does not meet the credit policy. Status:Charged Off	761


Loan Status
Does not meet the credit policy. Status:Charged Off
Default
Does not meet the credit policy. Status:Fully Paid
Late (16-30 days)
In Grace Period
Issued
Late (31-120 days)
Charged Off
Fully Paid
Current
0
100000
200000
300000
400000
500000
600000

LOAN STATUS

DESCRIPTIVE STATISTICS(INTEREST RATE)

- In this dataset, the highest interest rate is at 29%, the lowest is 5.32% and the average interest rate is 13.24%

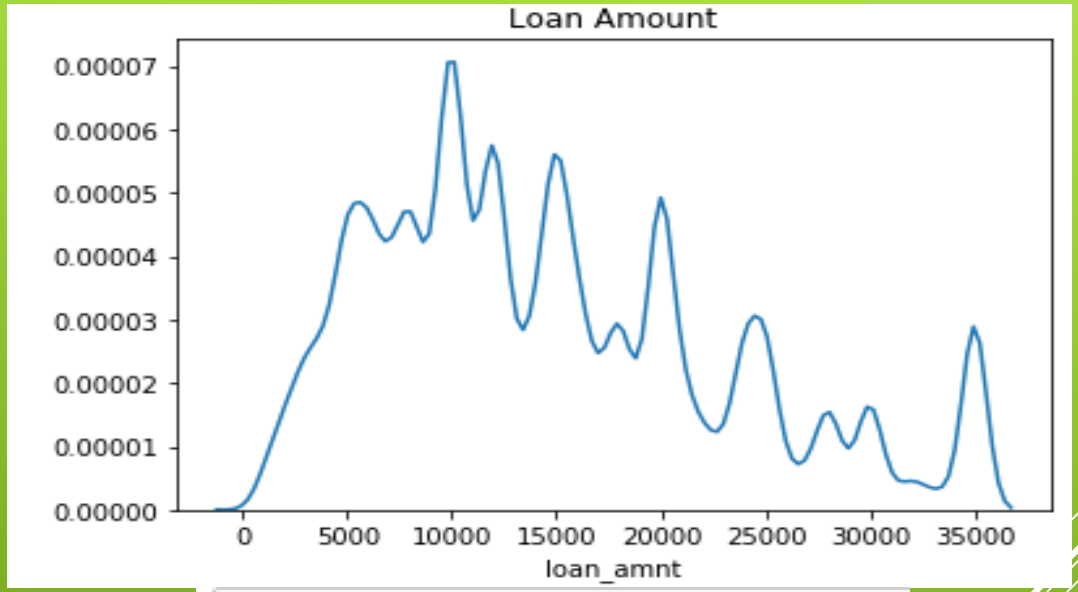



```
loan.int_rate.describe()
```

count	887379.000000
mean	13.246740
std	4.381867
min	5.320000
25%	9.990000
50%	12.990000
75%	16.200000
max	28.990000

DESCRIPTIVE STATISTICS(LOAN AMOUNT)

- In this dataset, the highest amount of loan borrowed is $35,000, the lowest $500.00 and the average amount borrowed is $14,755.26.

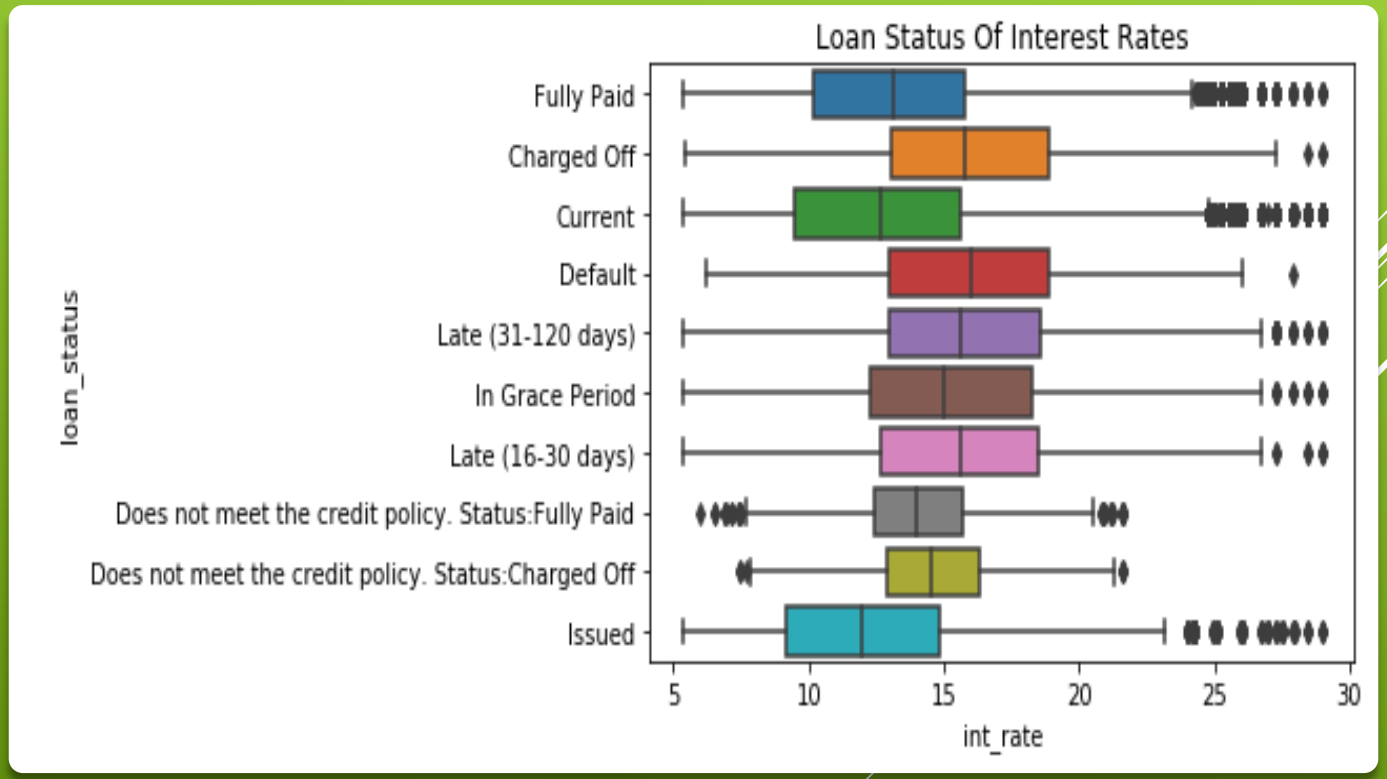



```
loan.loan_amnt.describe()
```

```
count    887379.000000
mean      14755.264605
std        8435.455601
min         500.000000
25%        8000.000000
50%       13000.000000
75%       20000.000000
max       35000.000000
```

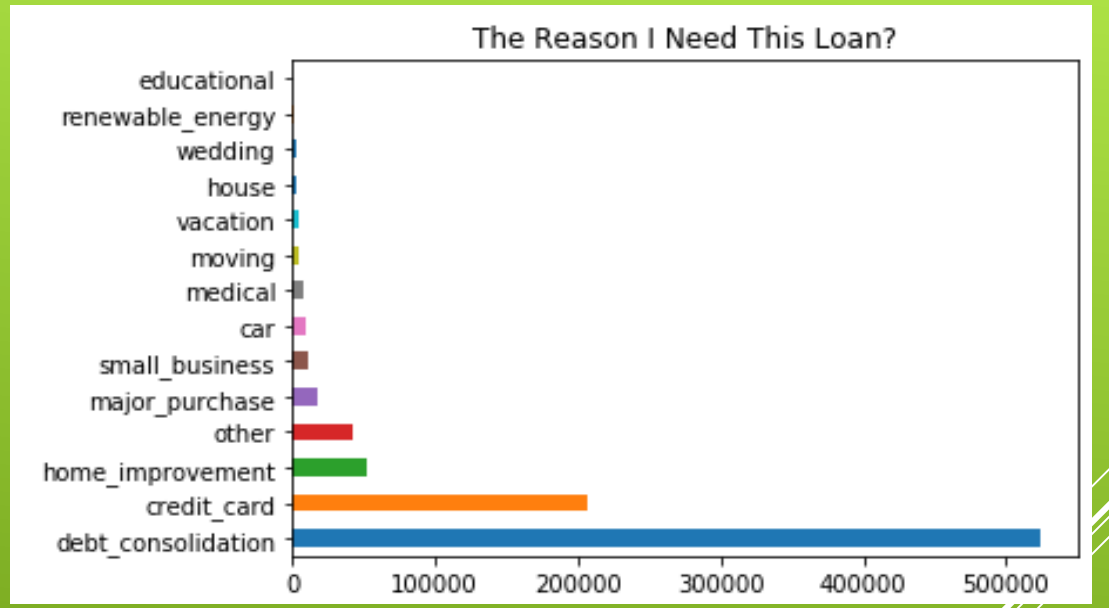
INTEREST RATE AND LOAN STATUS

- ❖ From this boxplot, we can see that the highest interest rates(25-29%) correlate with the loan status' charged off, late(16-30 days) late(31-120 days), and default. We can surmise that applications that have these interest rates are at high risk. However, this dataset has many outliers and there are individuals who are current and that have fully paid their loans with those same interest rates.
- ❖ We can also see that the average interest rate (13.24%) correlates with the median values of the loan status' fully paid and current. We can surmise that this interest rate is an optimal rate



PURPOSE OF LOAN

- Debt consolidation and credit cards account for 82% of all the purposes listed on loan applications with debt consolidation being 59% respectively.
- Car loans and student loans only account for 10% of all listed purposes. However, it is reasonable to conclude that they would be apart of debt consolidation and credit card debt considering the high amounts of debt in the U.S for student and auto loans.



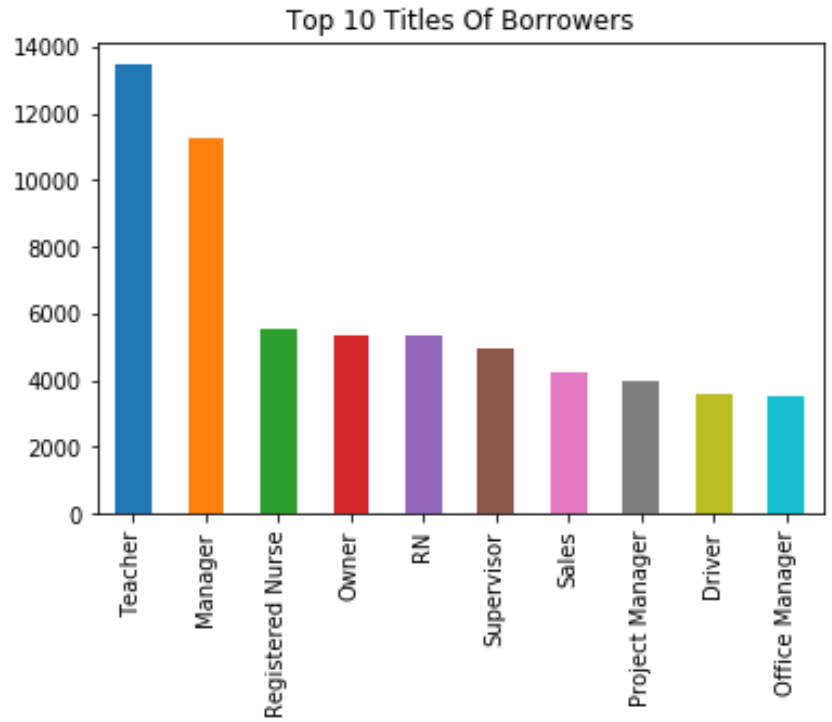

```
loan.purpose.value_counts()
```

debt_consolidation	524215
credit_card	206182
home_improvement	51829
other	42894
major_purchase	17277
small_business	10377
car	8863
medical	8540
moving	5414
vacation	4736
house	3707
wedding	2347
renewable_energy	575
educational	423

TOP 10 EMPLOYMENT TITLES

```
loan.emp_title.value_counts()[:10].plot(kind = 'bar')
plt.title("Top 10 Titles Of Borrowers")
```

```
Text(0.5,1,'Top 10 Titles Of Borrowers')
```



- The top employment title is a teacher. However, this section of the data could be more helpful if employment metrics were more specific such as field or industry. Titles such as “owner”, “office manager”, and “supervisor” are undetailed and do not give us an accurate indicator of the possibility of a loan being charged off. The employment titles that are specific in the data do not have great numerical representation for analysis for ex., “airport operations supervisor” accounts for only 1 title out of 835,917 entries.
- Employment titles such as teachers and registered nurses are more useful for analysis because we can locate their median incomes and job growth expectancies at the bureau of labor statistics.

TOP EMPLOYMENT TITLES(CONT.)

- This handicap in the employment titles column of the data is further elucidated in the fact that the graphical and numerical representations of this feature show an overrepresentation of the title teacher which could lead to an assumption or bias towards lending to teachers. However, upon further inspection teachers account for only 2% of the total employment titles and there are 2 different titles for the title teacher in the dataset.

```
loan.emp_title.value_counts()
```

Teacher	13469
Manager	11240
Registered Nurse	5525
Owner	5376
RN	5355
Supervisor	4983
Sales	4212
Project Manager	3988
Driver	3569
Office Manager	3510
General Manager	3178
Director	3156
manager	3138
teacher	2925

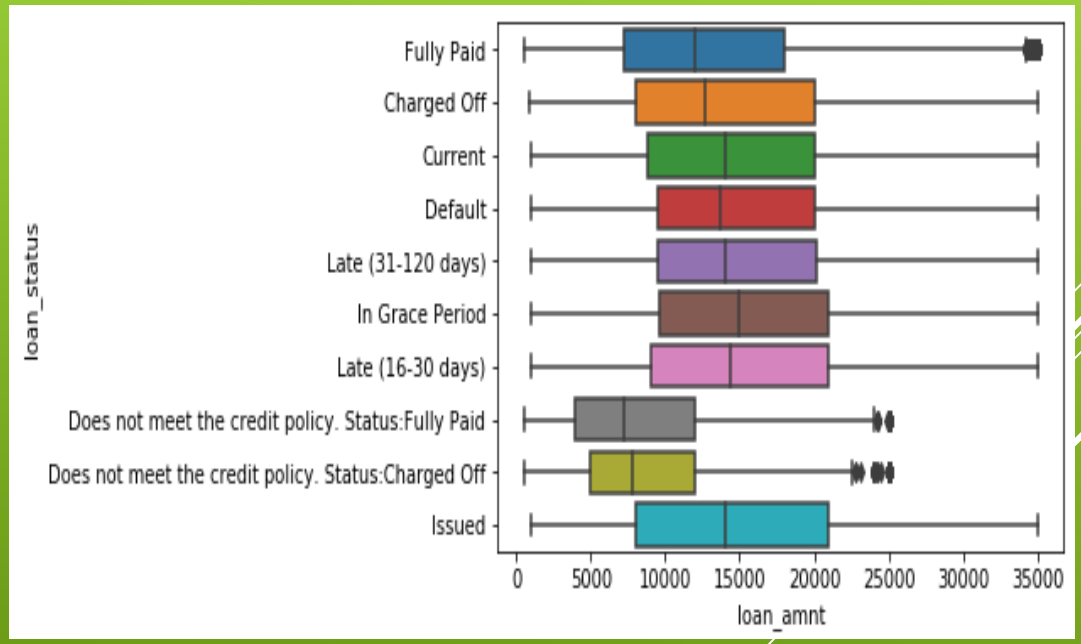
```
loan.emp_title.value_counts("teacher")
```

Teacher	0.016113
Manager	0.013446
Registered Nurse	0.006610
Owner	0.006431
RN	0.006406
Supervisor	0.005961
Sales	0.005039
Project Manager	0.004771
Driver	0.004270
Office Manager	0.004199
General Manager	0.003802
Director	0.003775
manager	0.003754
teacher	0.003499
owner	0.003408

LOAN STATUS AND LOAN AMOUNT RELATIONSHIP

- As we saw earlier, there is a relationship between the loan status "charged off" and interest rates. Here we see the relationships with the loan amount.
- The highest loan borrowed is between $33,000 and $35,000 dollars and it appears across all loan status'(Excluding the "Does not meet the credit policy" categories).
- The median value of all the loan status' and loan amounts is between $14,000 and $15,000 dollars which confirms the average loan borrowed is at $14,755.26. This amount is also the median amount for loans that have been "Charged Off" therefore the loan amount is not a reliable indicator of a loan being charged off.

```
ax= sns.boxplot(x="loan_amnt", y="loan_status", data=loan);
plt.figure(figsize=(16,6))
```



TOP 10 STATES

LOCATION OF BORROWERS

```
loan.addr_state.value_counts()
```

CA	129517
NY	74086
TX	71138
FL	60935
IL	35476
NJ	33256
PA	31393
OH	29631
GA	29085
VA	26255

- California is #1 and accounts for 15% of the total loans granted. New York and Texas are next at 8% and Virginia at the #10 accounts for 3% of all loans granted. It must be noted that Lending Club is headquartered in San Francisco, California so that could explain why as a state it has the most loans.

- This chart indicates that the state of California is #1 in loans that have been charged off at 7,332 loans. New York is second at 4,124 loans and the state of Texas is third at 3,035.
- Interestingly, Maine and North Dakota have zero charged off loans.

```
address_loan = ['addr_state', 'loan_status']
cm = sns.light_palette("green", as_cmap=True)
pd.crosstab(loan[address_loan[0]],
            loan[address_loan[1]]).style.background_gradient(cmap = cm)
```

addr_state	Charged Off	Current	Default	Does not meet the credit policy. Status:Charged Off	Does not meet the credit policy. Status:Fully Paid	Fully Paid	In Grace Period	Issued	Late (16-30 days)	Late (31-120 days)
AK	96	1469	2	1	4	567	15	14	6	31
AL	662	7576	9	8	24	2485	111	122	43	160
AR	337	4637	8	6	9	1417	57	70	13	86
AZ	1049	13577	39	18	33	5028	143	193	50	282
CA	7332	81851	211	101	223	35778	906	1147	327	1641
CO	784	12573	25	13	52	4829	106	166	57	202
CT	614	9353	8	12	50	3067	126	139	29	133
DC	87	1543	2	2	8	750	10	13	0	17
DE	121	1730	5	4	18	546	21	28	7	31
FL	3524	40999	95	72	160	14021	408	607	162	887
GA	1360	19993	36	35	69	6654	211	308	75	344
HI	276	2894	8	2	5	1202	45	42	13	83
IA	1	1	0	2	5	5	0	0	0	0
ID	1	3	0	0	3	5	0	0	0	0
IL	1542	25098	28	36	111	7711	219	298	56	377
IN	626	10529	19	7	3	2174	95	122	34	180
KS	356	5605	5	5	21	1727	35	82	21	69
KY	436	6016	15	10	22	1832	40	76	16	87
LA	566	7216	18	5	20	2386	81	110	30	155
MA	1017	13667	29	24	70	5122	140	211	54	259
MD	1100	14179	23	23	47	4907	175	215	55	307
ME	0	478	0	0	0	13	1	33	0	0
MI	1150	16147	32	19	55	4846	152	222	76	286
MN	803	10944	20	11	25	3657	112	148	37	200
MO	781	9773	16	26	53	3173	78	124	33	150
MS	86	3228	4	4	3	335	31	52	17	59
MT	95	1724	6	6	5	641	14	21	6	40
NC	1306	16835	33	12	29	5613	198	272	76	346
ND	0	444	0	0	0	8	1	23	2	1
NE	4	1086	0	3	3	34	8	29	2	7
NH	157	3005	3	2	14	991	36	39	12	35
NJ	1841	22411	49	26	107	7760	251	296	96	419
NM	269	3372	8	3	12	1108	39	45	15	68
NV	803	8136	27	16	13	3006	79	115	22	226
NY	4124	49670	106	57	191	17214	619	722	233	1150
OH	1472	20873	40	18	85	6266	162	272	67	376
OK	421	5620	8	3	14	1710	71	76	25	137
OR	544	7201	8	6	11	2809	80	96	22	116
PA	1557	21812	39	43	89	6842	243	252	93	423
RI	191	2661	5	2	7	896	27	37	8	59
SC	449	7469	15	4	13	2369	64	110	27	119
SD	90	1200	1	1	2	454	8	18	3	38
TN	563	9944	22	4	11	1863	114	146	28	192
TX	3035	49254	111	55	126	16308	455	691	183	920
UT	349	3935	8	5	14	1763	43	45	22	80
VA	1438	17263	29	16	63	6504	205	258	92	387
VT	70	1324	1	1	2	358	7	15	5	14
WA	986	12879	22	16	30	4929	118	188	64	202
WI	536	8107	14	14	42	2542	56	104	23	136
WV	159	3104	6	2	8	978	28	32	14	55
WY	82	1371	1	0	4	520	9	16	6	19

CONCLUSION

- In conclusion, of the features used(employment title, loan amount, interest rate, and location), interest rate appears to be a major indicator of a loan being charged off.
- The location of the borrower is biased towards the state of California which is the headquarters of the lending company. It therefore has the most loan applications, charge offs, defaults, and borrowers who are current.
- The loan amount is consistent across all of the loan status' and therefore not a reliable indicator of a loan being charged off.
- The data in the employment title column is imprecise and can not be used as an accurate indicator of a loan being charged off.