In [1]:
```python
import re
import os

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import datetime as dt
```

In [2]:
```python
accept_file_path = "Data/accepted_2007_to_2018Q4.csv"
```

In [3]:
```python
#reading accept csv file
accept_df = pd.read_csv(accept_file_path)
```

```
C:\Users\glori\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3146: Dtyp
eWarning: Columns (0,19,49,59,118,129,130,131,134,135,136,139,145,146,147) have mixed
types.Specify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

In [4]:
```python
reject_file_path = "Data/rejected_2007_to_2018Q4.csv"
```

In [5]:
```python
#reading reject csv file
reject_df = pd.read_csv(reject_file_path)
```

In [6]:
```python
list(accept_df)
```

Out[6]:
```
['id',
 'member_id',
 'loan_amnt',
 'funded_amnt',
 'funded_amnt_inv',
 'term',
 'int_rate',
 'installment',
 'grade',
 'sub_grade',
 'emp_title',
 'emp_length',
 'home_ownership',
 'annual_inc',
 'verification_status',
 'issue_d',
 'loan_status',
 'pymnt_plan',
 'url',
 'desc',
 'purpose',
 'title',
 'zip_code',
 'addr_state',
 'dti',
 'delinq_2yrs',
 'earliest_cr_line',
 'fico_range_low',
 'fico_range_high',
 'inq_last_6mths',
 'mths_since_last_delinq',
 'mths_since_last_record',
 'open_acc',
 'pub_rec',
 'revol_bal',
 'revol_util',
 'total_acc',
 'initial_list_status',
 'out_prncp',
 'out_prncp_inv',
 'total_pymnt',
 'total_pymnt_inv',
 'total_rec_prncp',
 'total_rec_int',
 'total_rec_late_fee',
 'recoveries',
 'collection_recovery_fee',
 'last_pymnt_d',
 'last_pymnt_amnt',
 'next_pymnt_d',
 'last_credit_pull_d',
 'last_fico_range_high',
 'last_fico_range_low',
 'collections_12_mths_ex_med',
 'mths_since_last_major_derog',
 'policy_code',
 'application_type',
 'annual_inc_joint',
 'dti_joint',
 'verification_status_joint',
 'acc_now_delinq',
 'tot_coll_amt',
 'tot_cur_bal',
 'open_acc_6m',
 'open_act_il',
 'open_il_12m',
 'open_il_24m',
 'mths_since_rcnt_il',
 'total_bal_il',
 'il_util',
 'open_rv_12m',
 'open_rv_24m',
 'max_bal_bc',
 'all_util',
 'total_rev_hi_lim',
 'inq_fi',
 'total_cu_tl',
 'inq_last_12m',
 'acc_open_past_24mths',
 'avg_cur_bal',
 'bc_open_to_buy',
 'bc_util',
 'chargeoff_within_12_mths',
 'delinq_amnt',
 'mo_sin_old_il_acct',
 'mo_sin_old_rev_tl_op',
 'mo_sin_rcnt_rev_tl_op',
 'mo_sin_rcnt_tl',
 'mort_acc',
 'mths_since_recent_bc',
 'mths_since_recent_bc_dlq',
 'mths_since_recent_inq',
 'mths_since_recent_revol_delinq',
 'num_accts_ever_120_pd',
 'num_actv_bc_tl',
 'num_actv_rev_tl',
 'num_bc_sats',
 'num_bc_tl',
 'num_il_tl',
 'num_op_rev_tl',
 'num_rev_accts',
 'num_rev_tl_bal_gt_0',
 'num_sats',
 'num_tl_120dpd_2m',
 'num_tl_30dpd',
 'num_tl_90g_dpd_24m',
 'num_tl_op_past_12m',
 'pct_tl_nvr_dlq',
 'percent_bc_gt_75',
 'pub_rec_bankruptcies',
 'tax_liens',
 'tot_hi_cred_lim',
 'total_bal_ex_mort',
 'total_bc_limit',
 'total_il_high_credit_limit',
 'revol_bal_joint',
 'sec_app_fico_range_low',
 'sec_app_fico_range_high',
 'sec_app_earliest_cr_line',
 'sec_app_inq_last_6mths',
 'sec_app_mort_acc',
 'sec_app_open_acc',
 'sec_app_revol_util',
 'sec_app_open_act_il',
 'sec_app_num_rev_accts',
 'sec_app_chargeoff_within_12_mths',
 'sec_app_collections_12_mths_ex_med',
 'sec_app_mths_since_last_major_derog',
 'hardship_flag',
 'hardship_type',
 'hardship_reason',
 'hardship_status',
 'deferral_term',
 'hardship_amount',
 'hardship_start_date',
 'hardship_end_date',
 'payment_plan_start_date',
 'hardship_length',
 'hardship_dpd',
 'hardship_loan_status',
 'orig_projected_additional_accrued_interest',
 'hardship_payoff_balance_amount',
 'hardship_last_payment_amount',
 'disbursement_method',
 'debt_settlement_flag',
 'debt_settlement_flag_date',
 'settlement_status',
 'settlement_date',
 'settlement_amount',
 'settlement_percentage',
 'settlement_term']
```

In [7]:
```python
list(reject_df)
```

Out[7]:
```
['Amount Requested',
 'Application Date',
 'Loan Title',
 'Risk_Score',
 'Debt-To-Income Ratio',
 'Zip Code',
 'State',
 'Employment Length',
 'Policy Code']
```

In [8]:
```python
print(reject_df['Loan Title'].unique())
print('-------------------')
print(accept_df['purpose'].unique())
print('-------------------')
print(accept_df['title'].unique())
```

```
['Wedding Covered but No Honeymoon' 'Consolidating Debt'
 'Want to consolidate my debt' ... 'dougie03' 'freeup'
 'Business Advertising Loan']
-------------------
['debt_consolidation' 'small_business' 'home_improvement' 'major_purchase'
 'credit_card' 'other' 'house' 'vacation' 'car' 'medical' 'moving'
 'renewable_energy' 'wedding' 'educational' nan]
-------------------
['Debt consolidation' 'Business' nan ... 'takeitaway' 'Creditt Card Loan'
 'debt reduction/hone updates']
```

In [9]:
```python
#needed columns from the accepted data
clean_accept_df = accept_df[[
    'loan_amnt',
    'int_rate',
    'issue_d',
    'fico_range_high',
    'dti',
    'zip_code',
    'addr_state',
    'emp_length',
    'purpose'
]]
```

In [10]:
```python
#columns from the rejected data
clean_reject_df = reject_df[['Amount Requested',
 'Application Date',
 'Risk_Score',
 'Debt-To-Income Ratio',
 'Zip Code',
 'State',
 'Employment Length',
 'Loan Title']]
```

In [11]:
```python
clean_accept_df.shape
```

Out[11]: (2260701, 9)

In [12]:
```python
clean_reject_df.shape
```

Out[12]: (27648741, 8)

In [13]:
```python
#added target value to the accepted data
clean_accept_df['Loan Status'] = "Accepted"
clean_accept_df.head()
```

```
<ipython-input-13-841988178c2f>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  clean_accept_df['Loan Status'] = "Accepted"
```

Out[13]:

	loan_amnt	int_rate	issue_d	fico_range_high	dti	zip_code	addr_state	emp_length	purpose
0	3600.0	13.99	Dec-2015	679.0	5.91	190xx	PA	10+ years	debt_consolidation
1	24700.0	11.99	Dec-2015	719.0	16.06	577xx	SD	10+ years	small_business
2	20000.0	10.78	Dec-2015	699.0	10.78	605xx	IL	10+ years	home_improvement
3	35000.0	14.85	Dec-2015	789.0	17.06	076xx	NJ	10+ years	debt_consolidation
4	10400.0	22.45	Dec-2015	699.0	25.37	174xx	PA	3 years	major_purchase

In [14]:
```python
# adding a target value to the rejected data
clean_reject_df['Loan Status'] = "Rejected"
clean_reject_df.head()
```

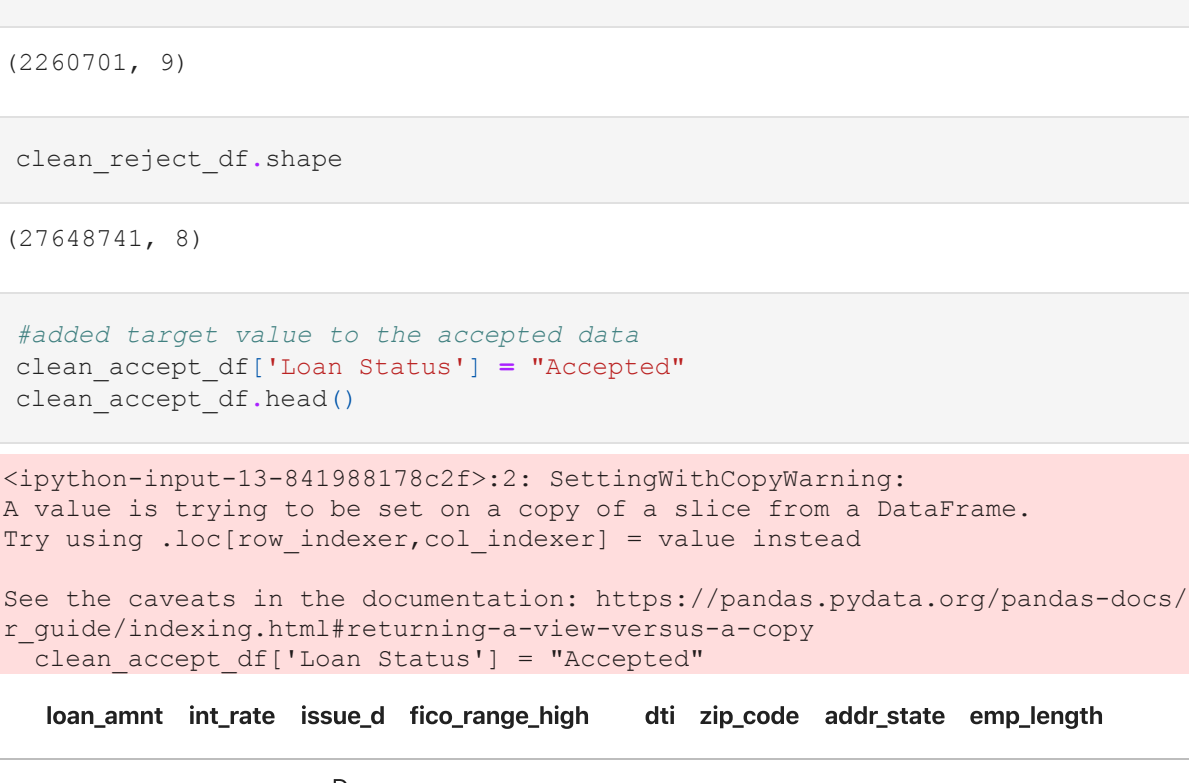
Out[14]:

	Amount Requested	Application Date	Risk_Score	Debt-To-Income Ratio	Zip Code	State	Employment Length	Loan Title	Loan Status
0	1000.0	2007-05-26	693.0	10%	481xx	NM	4 years	Wedding Covered but No Honeymoon	Rejected
1	1000.0	2007-05-26	703.0	10%	010xx	MA	< 1 year	Consolidating Debt	Rejected
2	11000.0	2007-05-27	715.0	10%	212xx	MD	1 year	Want to consolidate my debt	Rejected
3	6000.0	2007-05-27	698.0	38.64%	017xx	MA	< 1 year	waksman	Rejected
4	1500.0	2007-05-27	509.0	9.43%	209xx	MD	< 1 year	mdrigo	Rejected

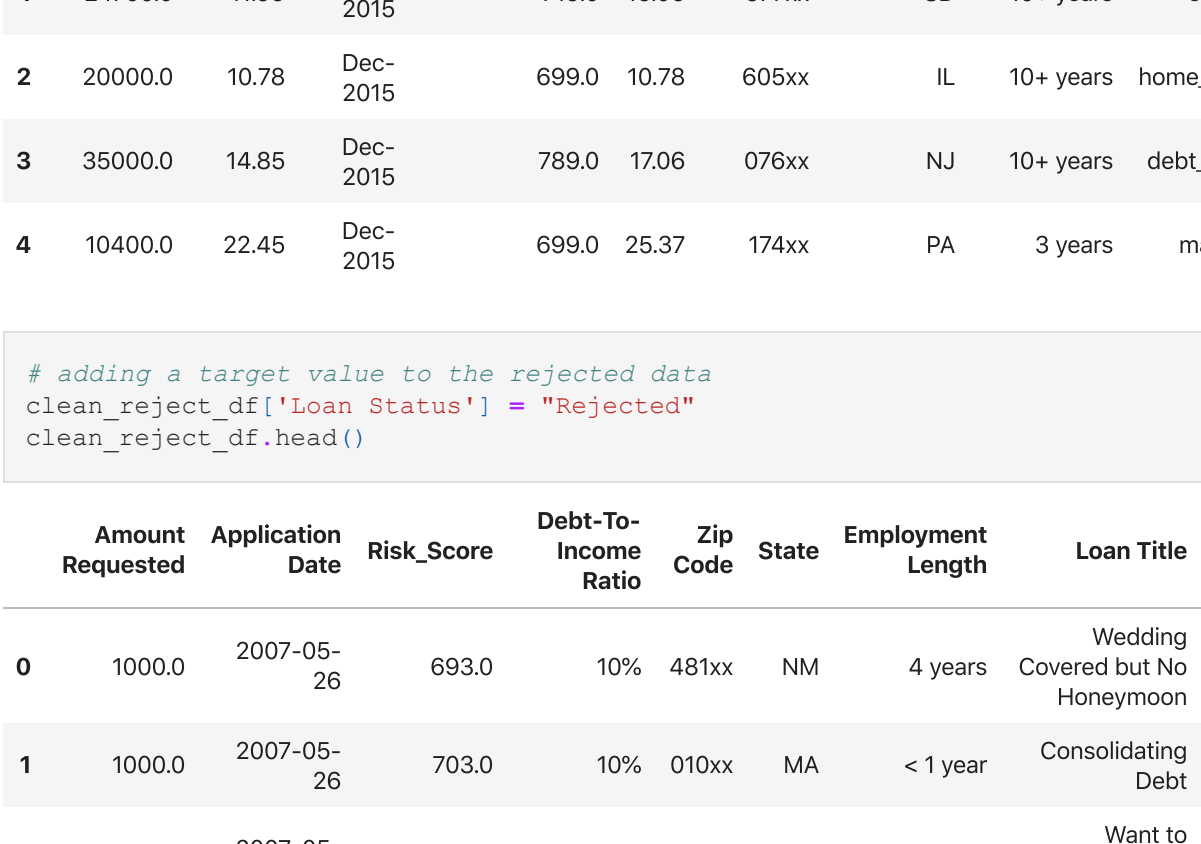
In [15]:
```python
clean_accept_df['int_rate'].describe() #interest rates in accept data (min 5.321%, ma
```

Out[15]:
```
count    2.260668e+06
mean     1.309283e+01
std      4.832138e+00
min      5.310000e+00
25%      9.490000e+00
50%      1.262000e+01
75%      1.599000e+01
max      3.099000e+01
Name: int_rate, dtype: float64
```

In [16]:
```python
#histogram correlation b/t fico score and Risk Score
plt.figure(figsize=(10,6))
clean_accept_df["fico_range_high"].hist(alpha=.5, color='green', bins=25), clean_reje
plt.show()
```



In [17]:
```python
#correlation b/t Risk Score and fico
plt.figure(figsize=(10,6))
clean_reject_df["Risk_Score"].hist(alpha=.5, color='red', bins=25), clean_accept_df["
plt.show()
```

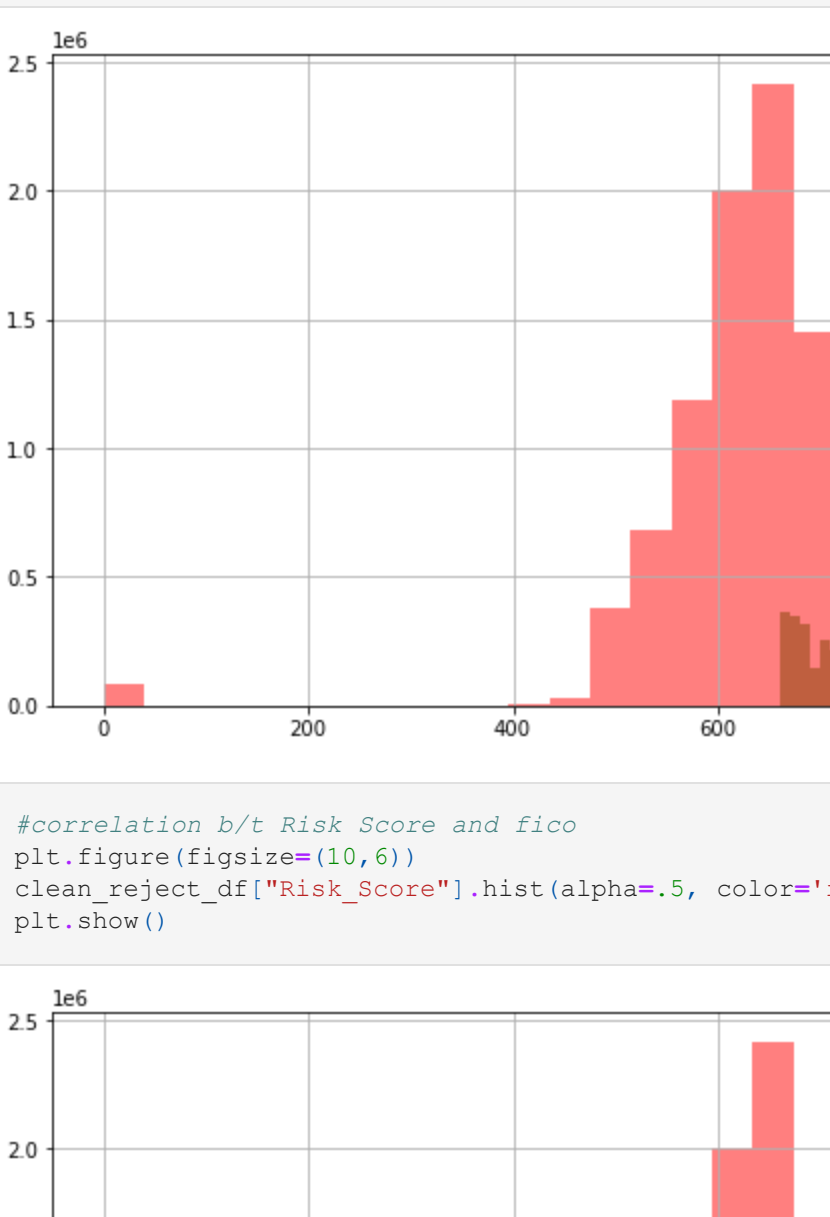


In [18]:
```python
sns.jointplot(x="fico_range_high", y="loan_amnt", data=clean_accept_df, color='purple
```

Out[18]: <seaborn.axisgrid.JointGrid at 0x291dc103e50>

In [19]:
```python
sns.jointplot(x="fico_range_high", y="int_rate", data=clean_accept_df, color='purple')
```

Out[19]: <seaborn.axisgrid.JointGrid at 0x29197979d60>



In [20]:
```python
print(clean_accept_df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2260701 entries, 0 to 2260700
Data columns (total 10 columns):
 #   Column           Dtype  
---  ------           -----  
 0   loan_amnt        float64
 1   int_rate         float64
 2   issue_d          object 
 3   fico_range_high  float64
 4   dti              float64
 5   zip_code         object 
 6   addr_state       object 
 7   emp_length       object 
 8   purpose          object 
 9   Loan Status      object 
dtypes: float64(4), object(6)
memory usage: 172.5+ MB
None
```

In [21]:
```python
print(clean_reject_df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27648741 entries, 0 to 27648740
Data columns (total 9 columns):
 #   Column                Dtype  
---  ------                -----  
 0   Amount Requested      float64
 1   Application Date      object 
 2   Risk_Score            float64
 3   Debt-To-Income Ratio  object 
 4   Zip Code              object 
 5   State                 object 
 6   Employment Length     object 
 7   Loan Title            object 
 8   Loan Status           object 
dtypes: float64(2), object(7)
memory usage: 1.9+ GB
None
```

In [22]:
```python
#accepted issue date as date
clean_accept_df['issue_d'] = pd.to_datetime(clean_accept_df['issue_d'])
clean_accept_df.head()
```

```
<ipython-input-22-51609ae8d551>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  clean_accept_df['issue_d'] = pd.to_datetime(clean_accept_df['issue_d'])
```

Out[22]:

	loan_amnt	int_rate	issue_d	fico_range_high	dti	zip_code	addr_state	emp_length	purpose
0	3600.0	13.99	2015-12-01	679.0	5.91	190xx	PA	10+ years	debt_consolidation
1	24700.0	11.99	2015-12-01	719.0	16.06	577xx	SD	10+ years	small_business
2	20000.0	10.78	2015-12-01	699.0	10.78	605xx	IL	10+ years	home_improvement
3	35000.0	14.85	2015-12-01	789.0	17.06	076xx	NJ	10+ years	debt_consolidation
4	10400.0	22.45	2015-12-01	699.0	25.37	174xx	PA	3 years	major_purchase

In [50]:
```python
#2018 accept data
accept_filtered_df = clean_accept_df[clean_accept_df['issue_d'].dt.year==2018]
accept_filtered_df.head()
```

Out[50]:

	loan_amnt	int_rate	issue_d	fico_range_high	dti	zip_code	addr_state	emp_length	pur
421097	5000.0	20.39	2018-03-01	669.0	21.80	740xx	OK	8 years	
421098	15000.0	9.92	2018-03-01	704.0	18.29	337xx	FL	2 years	debt_consolic
421099	11200.0	30.79	2018-03-01	669.0	43.97	030xx	NH	< 1 year	me
421100	25000.0	21.85	2018-03-01	669.0	12.89	361xx	AL	10+ years	debt_consolic
421101	3000.0	7.34	2018-03-01	764.0	0.58	988xx	WA	9 years	major_purc

In [51]:
```python
#see the array
accept_filtered_df ['issue_d'].dt.year.unique()
```

Out[51]: array([2018], dtype=int64)

In [52]:
```python
#rejected application date as date
clean_reject_df['Application Date'] = pd.to_datetime(clean_reject_df['Application Date
clean_reject_df.head()
```

Out[52]:

	Amount Requested	Application Date	Risk_Score	Debt-To-Income Ratio	Zip Code	State	Employment Length	Loan Title	Loan Status
0	1000.0	2007-05-26	693.0	10%	481xx	NM	4 years	Wedding Covered but No Honeymoon	Rejected
1	1000.0	2007-05-26	703.0	10%	010xx	MA	< 1 year	Consolidating Debt	Rejected
2	11000.0	2007-05-27	715.0	10%	212xx	MD	1 year	Want to consolidate my debt	Rejected
3	6000.0	2007-05-27	698.0	38.64%	017xx	MA	< 1 year	waksman	Rejected
4	1500.0	2007-05-27	509.0	9.43%	209xx	MD	< 1 year	mdrigo	Rejected

In [53]:
```python
#2018 reject data
reject_filtered_df = clean_reject_df[clean_reject_df['Application Date'].dt.year==201
reject_filtered_df.head()
```

Out[53]:

	Amount Requested	Application Date	Risk_Score	Debt-To-Income Ratio	Zip Code	State	Employment Length	Loan Title	Loan Status
4404427	3000.0	2018-07-01	NaN	100%	925xx	CA	< 1 year	Debt consolidation	Rejected
4404428	40000.0	2018-07-01	NaN	7.45%	335xx	FL	< 1 year	Major purchase	Rejected
4404429	16000.0	2018-07-01	NaN	34.93%	156xx	PA	< 1 year	Debt consolidation	Rejected
4404430	40000.0	2018-07-01	NaN	27.87%	957xx	CA	< 1 year	Debt consolidation	Rejected
4404431	300000.0	2018-07-01	NaN	-1%	258xx	TN	< 1 year	Business Loan	Rejected

In [54]:
```python
#see the array
reject_filtered_df ['Application Date'].dt.year.unique()
```

Out[54]: array([2018], dtype=int64)

In [55]:
```python
reject_filtered_df.rename(
columns=({"Amount Requested": "loan_amt", "Application Date": "issue_d", "Risk_Score"
)
reject_filtered_df.head()
```

```
C:\Users\glori\anaconda3\lib\site-packages\pandas\core\frame.py:4296: SettingWithCopyW
arning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  return super().rename(
```

Out[55]:

	loan_amt	issue_d	fico_range_high	dti	zip_code	addr_state	emp_length	purpose	Loan Sta
4404427	3000.0	2018-07-01	NaN	100%	925xx	CA	< 1 year	Debt consolidation	Rejec
4404428	40000.0	2018-07-01	NaN	7.45%	335xx	FL	< 1 year	Major purchase	Rejec
4404429	16000.0	2018-07-01	NaN	34.93%	156xx	PA	< 1 year	Debt consolidation	Rejec
4404430	40000.0	2018-07-01	NaN	27.87%	957xx	CA	< 1 year	Debt consolidation	Rejec
4404431	300000.0	2018-07-01	NaN	-1%	258xx	TN	< 1 year	Business Loan	Rejec

In [56]:
```python
accept_filtered_df.head()
```

Out[56]:

	loan_amnt	int_rate	issue_d	fico_range_high	dti	zip_code	addr_state	emp_length	pur
421097	5000.0	20.39	2018-03-01	669.0	21.80	740xx	OK	8 years	
421098	15000.0	9.92	2018-03-01	704.0	18.29	337xx	FL	2 years	debt_consolic
421099	11200.0	30.79	2018-03-01	669.0	43.97	030xx	NH	< 1 year	me
421100	25000.0	21.85	2018-03-01	669.0	12.89	361xx	AL	10+ years	debt_consolic
421101	3000.0	7.34	2018-03-01	764.0	0.58	988xx	WA	9 years	major_purc

In [57]:
```python
#categorical data to numerical data (accept)
accept_filtered_df.loc[(accept_filtered_df.emp_length == '< 1 year'), "emp_length"] =
accept_filtered_df.loc[(accept_filtered_df.emp_length == '1 year'), "emp_length"] = 1
accept_filtered_df.loc[(accept_filtered_df.emp_length == '2 years'), "emp_length"] = 
accept_filtered_df.loc[(accept_filtered_df.emp_length == '3 years'), "emp_length"] = 
accept_filtered_df.loc[(accept_filtered_df.emp_length == '4 years'), "emp_length"] = 
accept_filtered_df.loc[(accept_filtered_df.emp_length == '5 years'), "emp_length"] = 
accept_filtered_df.loc[(accept_filtered_df.emp_length == '6 years'), "emp_length"] = 
accept_filtered_df.loc[(accept_filtered_df.emp_length == '7 years'), "emp_length"] = 
accept_filtered_df.loc[(accept_filtered_df.emp_length == '8 years'), "emp_length"] = 
accept_filtered_df.loc[(accept_filtered_df.emp_length == '9 years'), "emp_length"] = 
accept_filtered_df.loc[(accept_filtered_df.emp_length == '10+ years'), "emp_length"] 
```

```
C:\Users\glori\anaconda3\lib\site-packages\pandas\core\indexing.py:1765: SettingWithCo
pyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  isetter(loc, value)
```

In [58]:
```python
#categorical data to numerical data (reject)
reject_filtered_df.loc[(reject_filtered_df.emp_length == '< 1 year'), "emp_length"] =
reject_filtered_df.loc[(reject_filtered_df.emp_length == '1 year'), "emp_length"] = 1
reject_filtered_df.loc[(reject_filtered_df.emp_length == '2 years'), "emp_length"] = 
reject_filtered_df.loc[(reject_filtered_df.emp_length == '3 years'), "emp_length"] = 
reject_filtered_df.loc[(reject_filtered_df.emp_length == '4 years'), "emp_length"] = 
reject_filtered_df.loc[(reject_filtered_df.emp_length == '5 years'), "emp_length"] = 
reject_filtered_df.loc[(reject_filtered_df.emp_length == '6 years'), "emp_length"] = 
reject_filtered_df.loc[(reject_filtered_df.emp_length == '7 years'), "emp_length"] = 
reject_filtered_df.loc[(reject_filtered_df.emp_length == '8 years'), "emp_length"] = 
reject_filtered_df.loc[(reject_filtered_df.emp_length == '9 years'), "emp_length"] = 
reject_filtered_df.loc[(reject_filtered_df.emp_length == '10+ years'), "emp_length"] 
```

In [60]:
```python
final_accept_df = accept_filtered_df.drop(['zip_code', 'issue_d'], axis = 1)
final_reject_df = reject_filtered_df.drop(['zip_code', 'issue_d'], axis = 1)
```

In [61]:
```python
print(final_accept_df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 495242 entries, 421097 to 1611876
Data columns (total 8 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   loan_amnt        495242 non-null  float64
 1   int_rate         495242 non-null  float64
 2   fico_range_high  495242 non-null  float64
 3   dti              494110 non-null  float64
 4   addr_state       495242 non-null  object 
 5   emp_length       453255 non-null  object 
 6   purpose          495242 non-null  object 
 7   Loan Status      495242 non-null  object 
dtypes: float64(4), object(4)
memory usage: 34.0+ MB
None
```

In [62]:
```python
print(final_reject_df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9496782 entries, 4404427 to 19699075
Data columns (total 7 columns):
 #   Column           Dtype  
---  ------           -----  
 0   loan_amt         float64
 1   fico_range_high  float64
 2   dti              object 
 3   addr_state       object 
 4   emp_length       object 
 5   purpose          object 
 6   Loan Status      object 
dtypes: float64(2), object(5)
memory usage: 579.6+ MB
None
```

In [64]:
```python
#drop na values in accept df
final_accept_df = final_accept_df.dropna()
final_accept_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 453176 entries, 421097 to 1611876
Data columns (total 8 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   loan_amnt        453176 non-null  float64
 1   int_rate         453176 non-null  float64
 2   fico_range_high  453176 non-null  float64
 3   dti              453176 non-null  float64
 4   addr_state       453176 non-null  object 
 5   emp_length       453176 non-null  object 
 6   purpose          453176 non-null  object 
 7   Loan Status      453176 non-null  object 
dtypes: float64(4), object(4)
memory usage: 31.1+ MB
```

In [65]:
```python
#drop na values in reject df
final_reject_df = final_reject_df.dropna()
final_reject_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 628344 entries, 4404458 to 19699073
Data columns (total 7 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   loan_amt         628344 non-null  float64
 1   fico_range_high  628344 non-null  float64
 2   dti              628344 non-null  object 
 3   addr_state       628344 non-null  object 
 4   emp_length       628344 non-null  object 
 5   purpose          628344 non-null  object 
 6   Loan Status      628344 non-null  object 
dtypes: float64(2), object(5)
memory usage: 38.4+ MB
```

In [66]:
```python
filenames = [final_accept_df, final_reject_df]
combined_data = pd.concat(filenames)
```

In [67]:
```python
combined_data.head()
```

Out[67]:

	loan_amnt	int_rate	fico_range_high	dti	addr_state	emp_length	purpose	Loan Status	lo
421097	5000.0	20.39	669.0	21.8	OK	8	other	Accepted	
421098	15000.0	9.92	704.0	18.29	FL	2	debt_consolidation	Accepted	
421099	11200.0	30.79	669.0	43.97	NH	0	medical	Accepted	
421100	25000.0	21.85	669.0	12.89	AL	10	debt_consolidation	Accepted	
421101	3000.0	7.34	764.0	0.58	WA	9	major_purchase	Accepted	

In [70]:
```python
combined_data.to_csv('Data/model_1_combine_data.csv')
```

In [79]:
```python
combined_data.shape
```

Out[79]: (1081520, 9)

In [73]:
```python
accept_filtered_df.to_csv('Data/model_1_accept_data.csv')
```

In [74]:
```python
accept_filtered_df.head(5)
```

Out[74]:

	loan_amnt	int_rate	issue_d	fico_range_high	dti	zip_code	addr_state	emp_length	pur
421097	5000.0	20.39	2018-03-01	669.0	21.80	740xx	OK	8	
421098	15000.0	9.92	2018-03-01	704.0	18.29	337xx	FL	2	debt_consolic
421099	11200.0	30.79	2018-03-01	669.0	43.97	030xx	NH	0	me
421100	25000.0	21.85	2018-03-01	669.0	12.89	361xx	AL	10	debt_consolic
421101	3000.0	7.34	2018-03-01	764.0	0.58	988xx	WA	9	major_purc

In [75]:
```python
accept_filtered_df.shape
```

Out[75]: (495242, 10)

In [76]:
```python
reject_filtered_df.to_csv('Data/model_1_reject_data.csv')
```

In [77]:
```python
reject_filtered_df.head(5)
```

Out[77]:

	loan_amt	issue_d	fico_range_high	dti	zip_code	addr_state	emp_length	purpose	Lo Sta
4404427	3000.0	2018-07-01	NaN	100%	925xx	CA	0	Debt consolidation	Rejec
4404428	40000.0	2018-07-01	NaN	7.45%	335xx	FL	0	Major purchase	Rejec
4404429	16000.0	2018-07-01	NaN	34.93%	156xx	PA	0	Debt consolidation	Rejec
4404430	40000.0	2018-07-01	NaN	27.87%	957xx	CA	0	Debt consolidation	Rejec
4404431	300000.0	2018-07-01	NaN	-1%	258xx	TN	0	Business Loan	Rejec

In [78]:
```python
reject_filtered_df.shape
```

Out[78]: (9496782, 9)

In [ ]:
```python
!pip install sklearn --upgrade
```

In [ ]: