In [221]:

**import** **numpy** **as** **np**  
**import** **pandas** **as** **pd**  
**import** **seaborn** **as** **sns**  
**import** **matplotlib.pyplot** **as** **plt**  
%**matplotlib** inline

In [222]:

demo = pd.read\_csv('traindemographics.csv')

In [223]:

perf = pd.read\_csv('trainperf.csv')

In [224]:

prev\_loan = pd.read\_csv('trainprevloans.csv')

**Demographics Data**[**¶**](#gjdgxs)

In [225]:

demo.head(10)

Out[225]:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **customerid** | **birthdate** | **bank\_account\_type** | **longitude\_gps** | **latitude\_gps** | **bank\_name\_clients** | **bank\_branch\_clients** | **employment\_status\_clients** | **level\_of\_education\_clients** |
| **0** | 8a858e135cb22031015cbafc76964ebd | 1973-10-10 00:00:00.000000 | Savings | 3.319219 | 6.528604 | GT Bank | NaN | NaN | NaN |
| **1** | 8a858e275c7ea5ec015c82482d7c3996 | 1986-01-21 00:00:00.000000 | Savings | 3.325598 | 7.119403 | Sterling Bank | NaN | Permanent | NaN |
| **2** | 8a858e5b5bd99460015bdc95cd485634 | 1987-04-01 00:00:00.000000 | Savings | 5.746100 | 5.563174 | Fidelity Bank | NaN | NaN | NaN |
| **3** | 8a858efd5ca70688015cabd1f1e94b55 | 1991-07-19 00:00:00.000000 | Savings | 3.362850 | 6.642485 | GT Bank | NaN | Permanent | NaN |
| **4** | 8a858e785acd3412015acd48f4920d04 | 1982-11-22 00:00:00.000000 | Savings | 8.455332 | 11.971410 | GT Bank | NaN | Permanent | NaN |
| **5** | 8a858f045bc9690c015bca251e9a4e95 | 1978-12-11 00:00:00.000000 | Savings | 3.330087 | 6.552518 | GT Bank | NaN | Permanent | NaN |
| **6** | 8a858ea05a859123015a8892914d15b7 | 1990-07-21 00:00:00.000000 | Savings | 3.365935 | 6.564823 | Access Bank | NaN | Permanent | NaN |
| **7** | 8a858f275c451af5015c546a50ce55ed | 1986-09-09 00:00:00.000000 | Savings | 4.734382 | 7.606868 | GT Bank | NaN | Permanent | NaN |
| **8** | 8a858f405d13c45f015d13dd93ec0c1c | 1992-11-17 00:00:00.000000 | Savings | 3.290590 | 6.612075 | EcoBank | NaN | Permanent | NaN |
| **9** | 8a858f735c161bda015c164d86632026 | 1965-02-23 00:00:00.000000 | Savings | 3.362783 | 6.509418 | EcoBank | NaN | Permanent | NaN |

In [226]:

demo.describe().T

Out[226]:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **longitude\_gps** | 4346.0 | 4.626189 | 7.184832 | -118.247009 | 3.354953 | 3.593302 | 6.545220 | 151.209290 |
| **latitude\_gps** | 4346.0 | 7.251356 | 3.055052 | -33.868818 | 6.470610 | 6.621888 | 7.425052 | 71.228069 |

In [227]:

demo.isna().sum()

Out[227]:

customerid 0  
birthdate 0  
bank\_account\_type 0  
longitude\_gps 0  
latitude\_gps 0  
bank\_name\_clients 0  
bank\_branch\_clients 4295  
employment\_status\_clients 648  
level\_of\_education\_clients 3759  
dtype: int64

In [228]:

demo.info()

<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 4346 entries, 0 to 4345  
Data columns (total 9 columns):  
customerid 4346 non-null object  
birthdate 4346 non-null object  
bank\_account\_type 4346 non-null object  
longitude\_gps 4346 non-null float64  
latitude\_gps 4346 non-null float64  
bank\_name\_clients 4346 non-null object  
bank\_branch\_clients 51 non-null object  
employment\_status\_clients 3698 non-null object  
level\_of\_education\_clients 587 non-null object  
dtypes: float64(2), object(7)  
memory usage: 305.7+ KB

In [229]:

type(demo['birthdate'].iloc[1])

Out[229]:

str

In [230]:

demo['birthdate'] = pd.to\_datetime(demo['birthdate'])

In [231]:

type(demo['birthdate'].iloc[1])

Out[231]:

pandas.\_libs.tslib.Timestamp

In [232]:

**def** show\_Uniques(data\_keys):  
 print('SHOWING ALL UNIQUE VALUES: ')  
 print('----------------------------')  
 **for** i **in** range(len(data\_keys)):  
 print(data\_keys[i])  
 print(demo[data\_keys[i]].unique())  
 print('total: ',len(demo[data\_keys[i]].unique()))  
 print('----------------------------')

In [233]:

show\_Uniques(['bank\_account\_type','bank\_name\_clients','bank\_branch\_clients','employment\_status\_clients','level\_of\_education\_clients'])

SHOWING ALL UNIQUE VALUES:   
----------------------------  
bank\_account\_type  
['Savings' 'Other' 'Current']  
total: 3  
----------------------------  
bank\_name\_clients  
['GT Bank' 'Sterling Bank' 'Fidelity Bank' 'Access Bank' 'EcoBank' 'FCMB'  
 'Skye Bank' 'UBA' 'Zenith Bank' 'Diamond Bank' 'First Bank' 'Union Bank'  
 'Stanbic IBTC' 'Standard Chartered' 'Heritage Bank' 'Keystone Bank'  
 'Unity Bank' 'Wema Bank']  
total: 18  
----------------------------  
bank\_branch\_clients  
[nan 'LAGOS' 'LADIPO' 'AWOLOWO' 'OGBA' 'AJOSE ADEOGUN' 'ABULE EGBA'  
 'OREGUN' 'ABEOKUTA' 'TRANS AMADI' 'OBA AKRAN BERGER PAINT' 'APAPA'  
 'BOSSO ROAD, MINNA' 'DUGBE, IBADAN' 'ALAUSA' 'OJUELEGBA' 'MUSHIN BRANCH'  
 'AKUTE' 'OBA ADEBIMPE' 'STERLING BANK PLC 102, IJU ROAD, IFAKO BRANCH'  
 'MEDICAL ROAD IKEJA' 'ADENIRAN OGUNSANYA' 'OBA AKRAN'  
 '40,SAPELE ROAD ,OPPOSITE DUMAZ JUNCTION BENIN CITY EDO STATE.'  
 'ACCESS BANK PLC, CHALLENGE ROUNDABOUT IBADAN, OYO STATE.'  
 'ABULE EGBA U-TURN,BRANCH' 'OGUDU, OJOTA' 'OBA AKRAN ROAD, IKEJA,'  
 '47 LAGOS ROAD, IKORODU' 'OAU ILE IFE' 'RING ROAD'  
 'PLOT 999C DANMOLE STREET, ADEOLA ODEKU, VICTORIA ISLAND, LAGOS'  
 'HERITAGE BANK, DUGBE, IBADAN' 'ADEOLA HOPEWELL' 'WHARF ROAD, APAPA'  
 'ILUPEJU' 'AKOWONJO' 'DUGBE,IBADAN' 'HEAD OFFICE' 'GBAGADA' 'LEKKI EPE'  
 'MAFOLUKU' 'AKURE BRANCH'  
 '17, SANUSI FAFUNWA STREET, VICTORIA ISLAND, LAGOS' ' IDI - ORO MUSHIN'  
 'TINCAN']  
total: 46  
----------------------------  
employment\_status\_clients  
[nan 'Permanent' 'Student' 'Self-Employed' 'Unemployed' 'Retired'  
 'Contract']  
total: 7  
----------------------------  
level\_of\_education\_clients  
[nan 'Secondary' 'Graduate' 'Post-Graduate' 'Primary']  
total: 5  
----------------------------

**Performance Data**[**¶**](#30j0zll)

In [234]:

perf.head(10)

Out[234]:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **customerid** | **systemloanid** | **loannumber** | **approveddate** | **creationdate** | **loanamount** | **totaldue** | **termdays** | **referredby** | **good\_bad\_flag** |
| **0** | 8a2a81a74ce8c05d014cfb32a0da1049 | 301994762 | 12 | 2017-07-25 08:22:56.000000 | 2017-07-25 07:22:47.000000 | 30000.0 | 34500.0 | 30 | NaN | Good |
| **1** | 8a85886e54beabf90154c0a29ae757c0 | 301965204 | 2 | 2017-07-05 17:04:41.000000 | 2017-07-05 16:04:18.000000 | 15000.0 | 17250.0 | 30 | NaN | Good |
| **2** | 8a8588f35438fe12015444567666018e | 301966580 | 7 | 2017-07-06 14:52:57.000000 | 2017-07-06 13:52:51.000000 | 20000.0 | 22250.0 | 15 | NaN | Good |
| **3** | 8a85890754145ace015429211b513e16 | 301999343 | 3 | 2017-07-27 19:00:41.000000 | 2017-07-27 18:00:35.000000 | 10000.0 | 11500.0 | 15 | NaN | Good |
| **4** | 8a858970548359cc0154883481981866 | 301962360 | 9 | 2017-07-03 23:42:45.000000 | 2017-07-03 22:42:39.000000 | 40000.0 | 44000.0 | 30 | NaN | Good |
| **5** | 8a8589f35451855401546b0738c42524 | 301986516 | 8 | 2017-07-19 21:46:24.000000 | 2017-07-19 20:46:18.000000 | 30000.0 | 39000.0 | 60 | NaN | Good |
| **6** | 8a858e095c59b91b015c5e5cea3719bc | 301972027 | 3 | 2017-07-10 19:25:33.000000 | 2017-07-10 18:25:28.000000 | 10000.0 | 13000.0 | 30 | NaN | Good |
| **7** | 8a858e1158dc4d830158f7bde4f47ea7 | 301994428 | 10 | 2017-07-24 23:42:27.000000 | 2017-07-24 22:42:21.000000 | 30000.0 | 34500.0 | 30 | NaN | Good |
| **8** | 8a858e185b4923b4015b4ae48d28646a | 301996687 | 4 | 2017-07-26 10:53:05.000000 | 2017-07-26 09:52:57.000000 | 10000.0 | 11500.0 | 15 | NaN | Good |
| **9** | 8a858e1d5cd58f9e015ceda4bdb63673 | 301981931 | 2 | 2017-07-17 13:48:42.000000 | 2017-07-17 12:48:35.000000 | 10000.0 | 11500.0 | 15 | NaN | Good |

In [235]:

perf.describe().T

Out[235]:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **systemloanid** | 4368.0 | 3.019810e+08 | 13431.151516 | 301958485.0 | 301969128.0 | 301980100.0 | 301993503.5 | 302004050.0 |
| **loannumber** | 4368.0 | 5.172390e+00 | 3.653569 | 2.0 | 2.0 | 4.0 | 7.0 | 27.0 |
| **loanamount** | 4368.0 | 1.780907e+04 | 10749.694571 | 10000.0 | 10000.0 | 10000.0 | 20000.0 | 60000.0 |
| **totaldue** | 4368.0 | 2.125738e+04 | 11943.510416 | 10000.0 | 13000.0 | 13000.0 | 24500.0 | 68100.0 |
| **termdays** | 4368.0 | 2.926168e+01 | 11.512519 | 15.0 | 30.0 | 30.0 | 30.0 | 90.0 |

In [236]:

perf.info()

<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 4368 entries, 0 to 4367  
Data columns (total 10 columns):  
customerid 4368 non-null object  
systemloanid 4368 non-null int64  
loannumber 4368 non-null int64  
approveddate 4368 non-null object  
creationdate 4368 non-null object  
loanamount 4368 non-null float64  
totaldue 4368 non-null float64  
termdays 4368 non-null int64  
referredby 587 non-null object  
good\_bad\_flag 4368 non-null object  
dtypes: float64(2), int64(3), object(5)  
memory usage: 341.3+ KB

In [237]:

perf.isna().sum()

Out[237]:

customerid 0  
systemloanid 0  
loannumber 0  
approveddate 0  
creationdate 0  
loanamount 0  
totaldue 0  
termdays 0  
referredby 3781  
good\_bad\_flag 0  
dtype: int64

In [238]:

type(perf['approveddate'].iloc[1])

Out[238]:

str

In [239]:

type(perf['creationdate'].iloc[1])

Out[239]:

str

In [240]:

perf['approveddate'] = pd.to\_datetime(perf['approveddate'])

In [241]:

perf['creationdate'] = pd.to\_datetime(perf['creationdate'])

In [242]:

type(perf['approveddate'].iloc[1])

Out[242]:

pandas.\_libs.tslib.Timestamp

In [243]:

type(perf['creationdate'].iloc[1])

Out[243]:

pandas.\_libs.tslib.Timestamp

In [244]:

perf['referredby'].fillna(0,inplace=**True**)

**Previous Loans Data**[**¶**](#1fob9te)

In [245]:

prev\_loan.head(10)

Out[245]:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **customerid** | **systemloanid** | **loannumber** | **approveddate** | **creationdate** | **loanamount** | **totaldue** | **termdays** | **closeddate** | **referredby** | **firstduedate** | **firstrepaiddate** |
| **0** | 8a2a81a74ce8c05d014cfb32a0da1049 | 301682320 | 2 | 2016-08-15 18:22:40.000000 | 2016-08-15 17:22:32.000000 | 10000.0 | 13000.0 | 30 | 2016-09-01 16:06:48.000000 | NaN | 2016-09-14 00:00:00.000000 | 2016-09-01 15:51:43.000000 |
| **1** | 8a2a81a74ce8c05d014cfb32a0da1049 | 301883808 | 9 | 2017-04-28 18:39:07.000000 | 2017-04-28 17:38:53.000000 | 10000.0 | 13000.0 | 30 | 2017-05-28 14:44:49.000000 | NaN | 2017-05-30 00:00:00.000000 | 2017-05-26 00:00:00.000000 |
| **2** | 8a2a81a74ce8c05d014cfb32a0da1049 | 301831714 | 8 | 2017-03-05 10:56:25.000000 | 2017-03-05 09:56:19.000000 | 20000.0 | 23800.0 | 30 | 2017-04-26 22:18:56.000000 | NaN | 2017-04-04 00:00:00.000000 | 2017-04-26 22:03:47.000000 |
| **3** | 8a8588f35438fe12015444567666018e | 301861541 | 5 | 2017-04-09 18:25:55.000000 | 2017-04-09 17:25:42.000000 | 10000.0 | 11500.0 | 15 | 2017-04-24 01:35:52.000000 | NaN | 2017-04-24 00:00:00.000000 | 2017-04-24 00:48:43.000000 |
| **4** | 8a85890754145ace015429211b513e16 | 301941754 | 2 | 2017-06-17 09:29:57.000000 | 2017-06-17 08:29:50.000000 | 10000.0 | 11500.0 | 15 | 2017-07-14 21:18:43.000000 | NaN | 2017-07-03 00:00:00.000000 | 2017-07-14 21:08:35.000000 |
| **5** | 8a858970548359cc0154883481981866 | 301832407 | 5 | 2017-03-06 13:25:17.000000 | 2017-03-06 12:25:10.000000 | 20000.0 | 23800.0 | 30 | 2017-04-04 15:46:56.000000 | NaN | 2017-04-05 00:00:00.000000 | 2017-04-04 15:31:47.000000 |
| **6** | 8a858970548359cc0154883481981866 | 301926513 | 8 | 2017-06-04 01:00:23.000000 | 2017-06-04 00:00:16.000000 | 30000.0 | 39000.0 | 60 | 2017-07-03 23:35:38.000000 | NaN | 2017-07-04 00:00:00.000000 | 2017-07-03 23:25:29.000000 |
| **7** | 8a8589f35451855401546b0738c42524 | 301912443 | 6 | 2017-05-18 09:42:18.000000 | 2017-05-18 08:41:12.000000 | 20000.0 | 24500.0 | 30 | 2017-06-19 10:10:35.000000 | NaN | 2017-06-19 00:00:00.000000 | 2017-06-19 10:00:21.000000 |
| **8** | 8a858e095c59b91b015c5e5cea3719bc | 301936857 | 2 | 2017-06-13 08:26:31.000000 | 2017-06-13 07:26:24.000000 | 10000.0 | 13000.0 | 30 | 2017-07-10 13:32:03.000000 | NaN | 2017-07-13 00:00:00.000000 | 2017-07-10 13:21:53.000000 |
| **9** | 8a858e1158dc4d830158f7bde4f47ea7 | 301811291 | 3 | 2017-02-06 18:55:30.000000 | 2017-02-06 17:55:20.000000 | 10000.0 | 11500.0 | 15 | 2017-02-21 05:34:20.000000 | NaN | 2017-02-21 00:00:00.000000 | 2017-02-21 05:19:09.000000 |

In [246]:

prev\_loan.describe().T

Out[246]:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **systemloanid** | 18183.0 | 3.018395e+08 | 93677.672704 | 301600134.0 | 301776577.0 | 301854965.0 | 301919682.5 | 302000275.0 |
| **loannumber** | 18183.0 | 4.189353e+00 | 3.249490 | 1.0 | 2.0 | 3.0 | 6.0 | 26.0 |
| **loanamount** | 18183.0 | 1.650124e+04 | 9320.547516 | 3000.0 | 10000.0 | 10000.0 | 20000.0 | 60000.0 |
| **totaldue** | 18183.0 | 1.957320e+04 | 10454.245277 | 3450.0 | 11500.0 | 13000.0 | 24500.0 | 68100.0 |
| **termdays** | 18183.0 | 2.669279e+01 | 10.946556 | 15.0 | 15.0 | 30.0 | 30.0 | 90.0 |

In [247]:

prev\_loan.info()

<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 18183 entries, 0 to 18182  
Data columns (total 12 columns):  
customerid 18183 non-null object  
systemloanid 18183 non-null int64  
loannumber 18183 non-null int64  
approveddate 18183 non-null object  
creationdate 18183 non-null object  
loanamount 18183 non-null float64  
totaldue 18183 non-null float64  
termdays 18183 non-null int64  
closeddate 18183 non-null object  
referredby 1026 non-null object  
firstduedate 18183 non-null object  
firstrepaiddate 18183 non-null object  
dtypes: float64(2), int64(3), object(7)  
memory usage: 1.7+ MB

In [248]:

prev\_loan.isna().sum()

Out[248]:

customerid 0  
systemloanid 0  
loannumber 0  
approveddate 0  
creationdate 0  
loanamount 0  
totaldue 0  
termdays 0  
closeddate 0  
referredby 17157  
firstduedate 0  
firstrepaiddate 0  
dtype: int64

In [249]:

type(prev\_loan['approveddate'][1])

Out[249]:

str

In [250]:

type(prev\_loan['creationdate'][1])

Out[250]:

str

In [251]:

type(prev\_loan['firstduedate'][1])

Out[251]:

str

In [252]:

type(prev\_loan['firstrepaiddate'][1])

Out[252]:

str

In [253]:

prev\_loan['approveddate'] = pd.to\_datetime(prev\_loan['approveddate'])

In [254]:

prev\_loan['creationdate'] = pd.to\_datetime(prev\_loan['creationdate'])

In [255]:

prev\_loan['firstduedate'] = pd.to\_datetime(prev\_loan['firstduedate'])

In [256]:

prev\_loan['firstrepaiddate'] = pd.to\_datetime(prev\_loan['firstrepaiddate'])

In [257]:

prev\_loan['closeddate'] = pd.to\_datetime(prev\_loan['closeddate'])

In [258]:

prev\_loan.info()

<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 18183 entries, 0 to 18182  
Data columns (total 12 columns):  
customerid 18183 non-null object  
systemloanid 18183 non-null int64  
loannumber 18183 non-null int64  
approveddate 18183 non-null datetime64[ns]  
creationdate 18183 non-null datetime64[ns]  
loanamount 18183 non-null float64  
totaldue 18183 non-null float64  
termdays 18183 non-null int64  
closeddate 18183 non-null datetime64[ns]  
referredby 1026 non-null object  
firstduedate 18183 non-null datetime64[ns]  
firstrepaiddate 18183 non-null datetime64[ns]  
dtypes: datetime64[ns](5), float64(2), int64(3), object(2)  
memory usage: 1.7+ MB

**Analytical Base Table**[**¶**](#3znysh7)

In [259]:

master = pd.merge(demo,perf,on='customerid')

In [260]:

master = pd.merge(master,prev\_loan,on='customerid')

In [261]:

master.head().T

Out[261]:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **0** | **1** | **2** | **3** | **4** |
| **customerid** | 8a858e135cb22031015cbafc76964ebd | 8a858e275c7ea5ec015c82482d7c3996 | 8a858e5b5bd99460015bdc95cd485634 | 8a858e5b5bd99460015bdc95cd485634 | 8a858e5b5bd99460015bdc95cd485634 |
| **birthdate** | 1973-10-10 00:00:00 | 1986-01-21 00:00:00 | 1987-04-01 00:00:00 | 1987-04-01 00:00:00 | 1987-04-01 00:00:00 |
| **bank\_account\_type** | Savings | Savings | Savings | Savings | Savings |
| **longitude\_gps** | 3.31922 | 3.3256 | 5.7461 | 5.7461 | 5.7461 |
| **latitude\_gps** | 6.5286 | 7.1194 | 5.56317 | 5.56317 | 5.56317 |
| **bank\_name\_clients** | GT Bank | Sterling Bank | Fidelity Bank | Fidelity Bank | Fidelity Bank |
| **bank\_branch\_clients** | NaN | NaN | NaN | NaN | NaN |
| **employment\_status\_clients** | NaN | Permanent | NaN | NaN | NaN |
| **level\_of\_education\_clients** | NaN | NaN | NaN | NaN | NaN |
| **systemloanid\_x** | 301964962 | 301972172 | 301976271 | 301976271 | 301976271 |
| **loannumber\_x** | 2 | 2 | 4 | 4 | 4 |
| **approveddate\_x** | 2017-07-05 14:29:48 | 2017-07-10 21:21:46 | 2017-07-13 15:40:27 | 2017-07-13 15:40:27 | 2017-07-13 15:40:27 |
| **creationdate\_x** | 2017-07-05 13:29:42 | 2017-07-10 20:21:40 | 2017-07-13 14:40:19 | 2017-07-13 14:40:19 | 2017-07-13 14:40:19 |
| **loanamount\_x** | 10000 | 10000 | 10000 | 10000 | 10000 |
| **totaldue\_x** | 13000 | 13000 | 13000 | 13000 | 13000 |
| **termdays\_x** | 30 | 30 | 30 | 30 | 30 |
| **referredby\_x** | 8a858899538ddb8e0153a780c56e34bb | 0 | 0 | 0 | 0 |
| **good\_bad\_flag** | Good | Bad | Good | Good | Good |
| **systemloanid\_y** | 301943951 | 301929966 | 301931228 | 301895253 | 301919784 |
| **loannumber\_y** | 1 | 1 | 3 | 1 | 2 |
| **approveddate\_y** | 2017-06-19 17:55:26 | 2017-06-07 12:47:30 | 2017-06-08 11:49:34 | 2017-05-08 11:07:01 | 2017-05-27 17:10:41 |
| **creationdate\_y** | 2017-06-19 16:54:19 | 2017-06-07 11:46:22 | 2017-06-08 10:49:27 | 2017-05-08 10:06:40 | 2017-05-27 16:10:34 |
| **loanamount\_y** | 10000 | 10000 | 10000 | 10000 | 10000 |
| **totaldue\_y** | 11500 | 13000 | 13000 | 11500 | 11500 |
| **termdays\_y** | 15 | 30 | 30 | 15 | 15 |
| **closeddate** | 2017-07-04 18:09:47 | 2017-07-10 08:52:54 | 2017-07-11 10:12:20 | 2017-05-27 13:02:53 | 2017-06-08 11:13:50 |
| **referredby\_y** | 8a858899538ddb8e0153a780c56e34bb | NaN | NaN | NaN | NaN |
| **firstduedate** | 2017-07-05 00:00:00 | 2017-07-07 00:00:00 | 2017-07-10 00:00:00 | 2017-05-23 00:00:00 | 2017-06-12 00:00:00 |
| **firstrepaiddate** | 2017-07-04 17:59:36 | 2017-07-10 08:42:44 | 2017-07-11 10:02:11 | 2017-05-27 12:52:45 | 2017-06-08 11:03:40 |

In [262]:

len(master)

Out[262]:

13693

In [263]:

master.isna().sum()

Out[263]:

customerid 0  
birthdate 0  
bank\_account\_type 0  
longitude\_gps 0  
latitude\_gps 0  
bank\_name\_clients 0  
bank\_branch\_clients 13589  
employment\_status\_clients 1363  
level\_of\_education\_clients 10229  
systemloanid\_x 0  
loannumber\_x 0  
approveddate\_x 0  
creationdate\_x 0  
loanamount\_x 0  
totaldue\_x 0  
termdays\_x 0  
referredby\_x 0  
good\_bad\_flag 0  
systemloanid\_y 0  
loannumber\_y 0  
approveddate\_y 0  
creationdate\_y 0  
loanamount\_y 0  
totaldue\_y 0  
termdays\_y 0  
closeddate 0  
referredby\_y 12891  
firstduedate 0  
firstrepaiddate 0  
dtype: int64

In [264]:

master.describe().T

Out[264]:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **longitude\_gps** | 13693.0 | 4.482207e+00 | 8.639590 | -1.182470e+02 | 3.354793e+00 | 3.576022e+00 | 6.364658e+00 | 1.512093e+02 |
| **latitude\_gps** | 13693.0 | 7.291442e+00 | 3.436232 | -3.386882e+01 | 6.471885e+00 | 6.617633e+00 | 7.421333e+00 | 7.122807e+01 |
| **systemloanid\_x** | 13693.0 | 3.019801e+08 | 13435.313603 | 3.019585e+08 | 3.019678e+08 | 3.019787e+08 | 3.019929e+08 | 3.020040e+08 |
| **loannumber\_x** | 13693.0 | 8.401081e+00 | 4.318780 | 2.000000e+00 | 5.000000e+00 | 8.000000e+00 | 1.100000e+01 | 2.700000e+01 |
| **loanamount\_x** | 13693.0 | 2.568867e+04 | 12625.374212 | 1.000000e+04 | 1.000000e+04 | 3.000000e+04 | 4.000000e+04 | 6.000000e+04 |
| **totaldue\_x** | 13693.0 | 2.981267e+04 | 13772.770273 | 1.000000e+04 | 1.300000e+04 | 3.300000e+04 | 4.400000e+04 | 6.810000e+04 |
| **termdays\_x** | 13693.0 | 3.323377e+01 | 14.440482 | 1.500000e+01 | 3.000000e+01 | 3.000000e+01 | 3.000000e+01 | 9.000000e+01 |
| **systemloanid\_y** | 13693.0 | 3.018395e+08 | 93046.899548 | 3.016001e+08 | 3.017767e+08 | 3.018545e+08 | 3.019193e+08 | 3.020003e+08 |
| **loannumber\_y** | 13693.0 | 4.200760e+00 | 3.269552 | 1.000000e+00 | 2.000000e+00 | 3.000000e+00 | 6.000000e+00 | 2.600000e+01 |
| **loanamount\_y** | 13693.0 | 1.656635e+04 | 9375.687619 | 3.000000e+03 | 1.000000e+04 | 1.000000e+04 | 2.000000e+04 | 6.000000e+04 |
| **totaldue\_y** | 13693.0 | 1.964661e+04 | 10512.238823 | 3.900000e+03 | 1.150000e+04 | 1.300000e+04 | 2.450000e+04 | 6.810000e+04 |
| **termdays\_y** | 13693.0 | 2.674213e+01 | 10.997624 | 1.500000e+01 | 1.500000e+01 | 3.000000e+01 | 3.000000e+01 | 9.000000e+01 |

In [265]:

master.dtypes

Out[265]:

customerid object  
birthdate datetime64[ns]  
bank\_account\_type object  
longitude\_gps float64  
latitude\_gps float64  
bank\_name\_clients object  
bank\_branch\_clients object  
employment\_status\_clients object  
level\_of\_education\_clients object  
systemloanid\_x int64  
loannumber\_x int64  
approveddate\_x datetime64[ns]  
creationdate\_x datetime64[ns]  
loanamount\_x float64  
totaldue\_x float64  
termdays\_x int64  
referredby\_x object  
good\_bad\_flag object  
systemloanid\_y int64  
loannumber\_y int64  
approveddate\_y datetime64[ns]  
creationdate\_y datetime64[ns]  
loanamount\_y float64  
totaldue\_y float64  
termdays\_y int64  
closeddate datetime64[ns]  
referredby\_y object  
firstduedate datetime64[ns]  
firstrepaiddate datetime64[ns]  
dtype: object

**Dropping Unnecessary Values**[**¶**](#2et92p0)

In [266]:

master.columns

Out[266]:

Index(['customerid', 'birthdate', 'bank\_account\_type', 'longitude\_gps',  
 'latitude\_gps', 'bank\_name\_clients', 'bank\_branch\_clients',  
 'employment\_status\_clients', 'level\_of\_education\_clients',  
 'systemloanid\_x', 'loannumber\_x', 'approveddate\_x', 'creationdate\_x',  
 'loanamount\_x', 'totaldue\_x', 'termdays\_x', 'referredby\_x',  
 'good\_bad\_flag', 'systemloanid\_y', 'loannumber\_y', 'approveddate\_y',  
 'creationdate\_y', 'loanamount\_y', 'totaldue\_y', 'termdays\_y',  
 'closeddate', 'referredby\_y', 'firstduedate', 'firstrepaiddate'],  
 dtype='object')

In [267]:

master.drop(['customerid','longitude\_gps','latitude\_gps','bank\_branch\_clients','systemloanid\_x','systemloanid\_y'],axis=1,inplace=**True**)

In [268]:

master.columns

Out[268]:

Index(['birthdate', 'bank\_account\_type', 'bank\_name\_clients',  
 'employment\_status\_clients', 'level\_of\_education\_clients',  
 'loannumber\_x', 'approveddate\_x', 'creationdate\_x', 'loanamount\_x',  
 'totaldue\_x', 'termdays\_x', 'referredby\_x', 'good\_bad\_flag',  
 'loannumber\_y', 'approveddate\_y', 'creationdate\_y', 'loanamount\_y',  
 'totaldue\_y', 'termdays\_y', 'closeddate', 'referredby\_y',  
 'firstduedate', 'firstrepaiddate'],  
 dtype='object')

**Exploratory Data Analysis**[**¶**](#tyjcwt)

In [269]:

master.isna().sum()

Out[269]:

birthdate 0  
bank\_account\_type 0  
bank\_name\_clients 0  
employment\_status\_clients 1363  
level\_of\_education\_clients 10229  
loannumber\_x 0  
approveddate\_x 0  
creationdate\_x 0  
loanamount\_x 0  
totaldue\_x 0  
termdays\_x 0  
referredby\_x 0  
good\_bad\_flag 0  
loannumber\_y 0  
approveddate\_y 0  
creationdate\_y 0  
loanamount\_y 0  
totaldue\_y 0  
termdays\_y 0  
closeddate 0  
referredby\_y 12891  
firstduedate 0  
firstrepaiddate 0  
dtype: int64

In [270]:

master.describe().T

Out[270]:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **loannumber\_x** | 13693.0 | 8.401081 | 4.318780 | 2.0 | 5.0 | 8.0 | 11.0 | 27.0 |
| **loanamount\_x** | 13693.0 | 25688.673045 | 12625.374212 | 10000.0 | 10000.0 | 30000.0 | 40000.0 | 60000.0 |
| **totaldue\_x** | 13693.0 | 29812.666494 | 13772.770273 | 10000.0 | 13000.0 | 33000.0 | 44000.0 | 68100.0 |
| **termdays\_x** | 13693.0 | 33.233769 | 14.440482 | 15.0 | 30.0 | 30.0 | 30.0 | 90.0 |
| **loannumber\_y** | 13693.0 | 4.200760 | 3.269552 | 1.0 | 2.0 | 3.0 | 6.0 | 26.0 |
| **loanamount\_y** | 13693.0 | 16566.347769 | 9375.687619 | 3000.0 | 10000.0 | 10000.0 | 20000.0 | 60000.0 |
| **totaldue\_y** | 13693.0 | 19646.612269 | 10512.238823 | 3900.0 | 11500.0 | 13000.0 | 24500.0 | 68100.0 |
| **termdays\_y** | 13693.0 | 26.742131 | 10.997624 | 15.0 | 15.0 | 30.0 | 30.0 | 90.0 |

In [271]:

master.info()

<class 'pandas.core.frame.DataFrame'>  
Int64Index: 13693 entries, 0 to 13692  
Data columns (total 23 columns):  
birthdate 13693 non-null datetime64[ns]  
bank\_account\_type 13693 non-null object  
bank\_name\_clients 13693 non-null object  
employment\_status\_clients 12330 non-null object  
level\_of\_education\_clients 3464 non-null object  
loannumber\_x 13693 non-null int64  
approveddate\_x 13693 non-null datetime64[ns]  
creationdate\_x 13693 non-null datetime64[ns]  
loanamount\_x 13693 non-null float64  
totaldue\_x 13693 non-null float64  
termdays\_x 13693 non-null int64  
referredby\_x 13693 non-null object  
good\_bad\_flag 13693 non-null object  
loannumber\_y 13693 non-null int64  
approveddate\_y 13693 non-null datetime64[ns]  
creationdate\_y 13693 non-null datetime64[ns]  
loanamount\_y 13693 non-null float64  
totaldue\_y 13693 non-null float64  
termdays\_y 13693 non-null int64  
closeddate 13693 non-null datetime64[ns]  
referredby\_y 802 non-null object  
firstduedate 13693 non-null datetime64[ns]  
firstrepaiddate 13693 non-null datetime64[ns]  
dtypes: datetime64[ns](8), float64(4), int64(4), object(7)  
memory usage: 2.5+ MB

In [272]:

plt.figure(figsize=(12,7))  
sns.heatmap(master.corr(),annot=**True**)

Out[272]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x35c5b2c7b8>

In [273]:

plt.figure(figsize=(12,7))  
sns.countplot(master['employment\_status\_clients'])

E:\Python\Anaconda\lib\site-packages\seaborn\categorical.py:1460: FutureWarning: remove\_na is deprecated and is a private function. Do not use.  
 stat\_data = remove\_na(group\_data)

Out[273]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x35c5721a58>

In [274]:

plt.figure(figsize=(12,7))  
sns.countplot(master['level\_of\_education\_clients'])

E:\Python\Anaconda\lib\site-packages\seaborn\categorical.py:1460: FutureWarning: remove\_na is deprecated and is a private function. Do not use.  
 stat\_data = remove\_na(group\_data)

Out[274]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x35c0e15ac8>

In [275]:

plt.figure(figsize=(12,7))  
sns.countplot(master['good\_bad\_flag'])

E:\Python\Anaconda\lib\site-packages\seaborn\categorical.py:1460: FutureWarning: remove\_na is deprecated and is a private function. Do not use.  
 stat\_data = remove\_na(group\_data)

Out[275]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x35c5bf32e8>

In [276]:

pd.crosstab(master['employment\_status\_clients'],master['level\_of\_education\_clients'])

Out[276]:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **level\_of\_education\_clients** | **Graduate** | **Post-Graduate** | **Primary** | **Secondary** |
| **employment\_status\_clients** |  |  |  |  |
| **Permanent** | 1520 | 289 | 27 | 239 |
| **Retired** | 12 | 0 | 0 | 0 |
| **Self-Employed** | 752 | 106 | 9 | 208 |
| **Student** | 131 | 40 | 0 | 15 |
| **Unemployed** | 79 | 0 | 0 | 23 |

In [277]:

pd.crosstab(master['employment\_status\_clients'],master['bank\_account\_type'])

Out[277]:

|  |  |  |  |
| --- | --- | --- | --- |
| **bank\_account\_type** | **Current** | **Other** | **Savings** |
| **employment\_status\_clients** |  |  |  |
| **Contract** | 10 | 0 | 0 |
| **Permanent** | 136 | 3074 | 5928 |
| **Retired** | 0 | 19 | 10 |
| **Self-Employed** | 7 | 1644 | 668 |
| **Student** | 0 | 341 | 274 |
| **Unemployed** | 0 | 123 | 96 |

In [278]:

pd.crosstab(master['employment\_status\_clients'],master['good\_bad\_flag']).T

Out[278]:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **employment\_status\_clients** | **Contract** | **Permanent** | **Retired** | **Self-Employed** | **Student** | **Unemployed** |
| **good\_bad\_flag** |  |  |  |  |  |  |
| **Bad** | 0 | 1647 | 0 | 347 | 140 | 40 |
| **Good** | 10 | 7491 | 29 | 1972 | 475 | 179 |

In [279]:

pd.crosstab(master['level\_of\_education\_clients'],master['good\_bad\_flag']).T

Out[279]:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **level\_of\_education\_clients** | **Graduate** | **Post-Graduate** | **Primary** | **Secondary** |
| **good\_bad\_flag** |  |  |  |  |
| **Bad** | 415 | 89 | 1 | 120 |
| **Good** | 2093 | 346 | 35 | 365 |

In [280]:

**def** Graph\_Summary(input\_value):  
 **if**((type(master[input\_value].iloc[0]) == np.int64) | (type(master[input\_value].iloc[0]) == np.float64) ) :  
   
 f = plt.figure(figsize=(20,20))  
 f.set\_figheight(8)  
 f.set\_figwidth(25)  
  
 plt.subplot(131)  
 sns.set()  
 plt.hist(master[input\_value],bins = 50,normed=**True**)  
 plt.xlabel(input\_value)  
 plt.ylabel('F(X)')  
   
  
  
 plt.subplot(132)  
 sns.set()  
 X = np.sort(master[input\_value])  
 Y = np.arange(1,len(X)+1)/len(X)  
 plt.plot(X,Y,marker='.',linestyle='none')  
 plt.xlabel(input\_value)  
 plt.ylabel('ECDF')  
  
  
  
 plt.subplot(133)  
 sns.set()  
 plt.boxplot(master[input\_value])  
 plt.xlabel(input\_value)  
   
 plt.show()  
 **else**: **return** **None**

In [281]:

print('Percentages of Good and Bad!')  
print('Percentage of good: ',len(master[master['good\_bad\_flag']=='Good'])/len(master)\*100,'%')  
print('Percentage of bad: ',len(master[master['good\_bad\_flag']=='Bad'])/len(master)\*100,'%')

Percentages of Good and Bad!  
Percentage of good: 81.53801212298255 %  
Percentage of bad: 18.461987877017457 %

**Top 10 Banks having the customers flagged as good:**[**¶**](#3dy6vkm)

In [282]:

master[master['good\_bad\_flag']=='Good']['bank\_name\_clients'].value\_counts().head(10)

Out[282]:

GT Bank 4131  
First Bank 1569  
Access Bank 1031  
UBA 892  
Diamond Bank 880  
Zenith Bank 745  
Stanbic IBTC 441  
FCMB 345  
EcoBank 317  
Fidelity Bank 209  
Name: bank\_name\_clients, dtype: int64

**Top 10 Banks having the customers flagged as bad:**[**¶**](#1t3h5sf)

In [283]:

master[master['good\_bad\_flag']=='Bad']['bank\_name\_clients'].value\_counts().head(10)

Out[283]:

GT Bank 903  
First Bank 363  
Access Bank 239  
UBA 234  
Zenith Bank 203  
EcoBank 133  
Diamond Bank 103  
Skye Bank 84  
FCMB 75  
Fidelity Bank 58  
Name: bank\_name\_clients, dtype: int64

In [284]:

master.groupby('good\_bad\_flag').mean()

Out[284]:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **loannumber\_x** | **loanamount\_x** | **totaldue\_x** | **termdays\_x** | **loannumber\_y** | **loanamount\_y** | **totaldue\_y** | **termdays\_y** |
| **good\_bad\_flag** |  |  |  |  |  |  |  |  |
| **Bad** | 7.671282 | 22689.873418 | 26796.914557 | 32.865902 | 3.835839 | 15189.873418 | 18113.770728 | 25.864320 |
| **Good** | 8.566323 | 26367.666816 | 30495.498639 | 33.317062 | 4.283386 | 16878.011644 | 19993.681093 | 26.940887 |

**Termdays and Loan Amount (X)**[**¶**](#4d34og8)

In [285]:

plt.figure(figsize=(12,7))  
sns.boxplot(x=master['termdays\_x'],y=master['loanamount\_x'])

E:\Python\Anaconda\lib\site-packages\seaborn\categorical.py:462: FutureWarning: remove\_na is deprecated and is a private function. Do not use.  
 box\_data = remove\_na(group\_data)

Out[285]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x35c6e43ef0>

In [286]:

master.groupby('termdays\_x')['loanamount\_x'].median()

Out[286]:

termdays\_x  
15 10000.0  
30 25000.0  
60 40000.0  
90 50000.0  
Name: loanamount\_x, dtype: float64

**Termdays and Loan Amount (Y)**[**¶**](#2s8eyo1)

In [287]:

plt.figure(figsize=(12,7))  
sns.boxplot(x=master['termdays\_y'],y=master['loanamount\_y'])

E:\Python\Anaconda\lib\site-packages\seaborn\categorical.py:462: FutureWarning: remove\_na is deprecated and is a private function. Do not use.  
 box\_data = remove\_na(group\_data)

Out[287]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x35c78ec198>

In [288]:

master.groupby('termdays\_y')['loanamount\_y'].median()

Out[288]:

termdays\_y  
15 10000.0  
30 10000.0  
60 30000.0  
90 50000.0  
Name: loanamount\_y, dtype: float64

**Termdays and Total Due (X)**[**¶**](#17dp8vu)

In [289]:

plt.figure(figsize=(12,7))  
sns.boxplot(x=master['termdays\_x'],y=master['totaldue\_x'])

E:\Python\Anaconda\lib\site-packages\seaborn\categorical.py:462: FutureWarning: remove\_na is deprecated and is a private function. Do not use.  
 box\_data = remove\_na(group\_data)

Out[289]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x35c7f95940>

In [290]:

master.groupby('termdays\_x')['totaldue\_x'].median()

Out[290]:

termdays\_x  
15 11500.0  
30 28750.0  
60 48000.0  
90 57500.0  
Name: totaldue\_x, dtype: float64

**Termdays and Total Due (Y)**[**¶**](#3rdcrjn)

In [291]:

plt.figure(figsize=(12,7))  
sns.boxplot(x=master['termdays\_y'],y=master['totaldue\_y'])

E:\Python\Anaconda\lib\site-packages\seaborn\categorical.py:462: FutureWarning: remove\_na is deprecated and is a private function. Do not use.  
 box\_data = remove\_na(group\_data)

Out[291]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x35c6efd710>

In [292]:

master.groupby('termdays\_y')['totaldue\_y'].median()

Out[292]:

termdays\_y  
15 11500.0  
30 13000.0  
60 36800.0  
90 57500.0  
Name: totaldue\_y, dtype: float64

**Total Due and Loan Amount**[**¶**](#26in1rg)

In [314]:

plt.figure(figsize=(12,7))  
plt.subplot(211)  
plt.scatter('totaldue\_x','loanamount\_x',data=master)  
plt.xlabel('X\_Total\_Due')  
plt.subplot(212)  
plt.scatter('totaldue\_y','loanamount\_y',data=master)  
plt.xlabel('Y\_Total\_Due')

Out[314]:

Text(0.5,0,'Y\_Total\_Due')

In [294]:

master.groupby('good\_bad\_flag').sum()

Out[294]:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **loannumber\_x** | **loanamount\_x** | **totaldue\_x** | **termdays\_x** | **loannumber\_y** | **loanamount\_y** | **totaldue\_y** | **termdays\_y** |
| **good\_bad\_flag** |  |  |  |  |  |  |  |  |
| **Bad** | 19393 | 57360000.0 | 67742600.0 | 83085 | 9697 | 38400000.0 | 45791612.4 | 65385 |
| **Good** | 95643 | 294395000.0 | 340482242.3 | 371985 | 47824 | 188443000.0 | 223229449.4 | 300795 |

In [301]:

**for** i **in** master.columns:  
 Graph\_Summary(i)

In [302]:

sns.pairplot(master)

Out[302]:

<seaborn.axisgrid.PairGrid at 0x35cb82bcf8>

In [303]:

master.info()

<class 'pandas.core.frame.DataFrame'>  
Int64Index: 13693 entries, 0 to 13692  
Data columns (total 23 columns):  
birthdate 13693 non-null datetime64[ns]  
bank\_account\_type 13693 non-null object  
bank\_name\_clients 13693 non-null object  
employment\_status\_clients 12330 non-null object  
level\_of\_education\_clients 3464 non-null object  
loannumber\_x 13693 non-null int64  
approveddate\_x 13693 non-null datetime64[ns]  
creationdate\_x 13693 non-null datetime64[ns]  
loanamount\_x 13693 non-null float64  
totaldue\_x 13693 non-null float64  
termdays\_x 13693 non-null int64  
referredby\_x 13693 non-null object  
good\_bad\_flag 13693 non-null object  
loannumber\_y 13693 non-null int64  
approveddate\_y 13693 non-null datetime64[ns]  
creationdate\_y 13693 non-null datetime64[ns]  
loanamount\_y 13693 non-null float64  
totaldue\_y 13693 non-null float64  
termdays\_y 13693 non-null int64  
closeddate 13693 non-null datetime64[ns]  
referredby\_y 802 non-null object  
firstduedate 13693 non-null datetime64[ns]  
firstrepaiddate 13693 non-null datetime64[ns]  
dtypes: datetime64[ns](8), float64(4), int64(4), object(7)  
memory usage: 3.1+ MB

In [304]:

**def** hist\_of\_levels(categorical,continuous):  
 fig = plt.figure(figsize=(12,7))  
 fig.set\_figheight(8)  
 fig.set\_figwidth(15)  
 **for** i **in** range(len(master[categorical].unique())):  
 uni = master[categorical].unique()  
 plt.subplot(2,len(uni),i+1)  
 plt.hist(master[master[categorical]==uni[i]][continuous],bins=50,normed=**True**)  
 plt.xlabel(uni[i])

In [305]:

hist\_of\_levels('employment\_status\_clients','termdays\_x')

E:\Python\Anaconda\lib\site-packages\numpy\lib\function\_base.py:838: RuntimeWarning: invalid value encountered in true\_divide  
 return n/db/n.sum(), bin\_edges

In [306]:

hist\_of\_levels('good\_bad\_flag','totaldue\_x')

In [307]:

an = np.round(master[master['loannumber\_x'] >= np.mean(master['loannumber\_y'])])

In [308]:

plt.figure(figsize=(12,7))  
plt.hist(an['loannumber\_x'],bins=50,normed=**True**)

Out[308]:

(array([0.21123345, 0. , 0.20123549, 0. , 0.24250366,  
 0. , 0.24399272, 0. , 0. , 0.23654743,  
 0. , 0.27377387, 0. , 0.16592356, 0. ,  
 0.20102277, 0. , 0. , 0.15060754, 0. ,  
 0.10231953, 0. , 0.0982778 , 0. , 0. ,  
 0.04148089, 0. , 0.03382288, 0. , 0.01084885,  
 0. , 0.02297403, 0. , 0. , 0.00808346,  
 0. , 0. , 0. , 0.01786869, 0. ,  
 0.0046799 , 0. , 0. , 0. , 0. ,  
 0. , 0. , 0. , 0. , 0.00553079]),  
 array([ 5. , 5.44, 5.88, 6.32, 6.76, 7.2 , 7.64, 8.08, 8.52,  
 8.96, 9.4 , 9.84, 10.28, 10.72, 11.16, 11.6 , 12.04, 12.48,  
 12.92, 13.36, 13.8 , 14.24, 14.68, 15.12, 15.56, 16. , 16.44,  
 16.88, 17.32, 17.76, 18.2 , 18.64, 19.08, 19.52, 19.96, 20.4 ,  
 20.84, 21.28, 21.72, 22.16, 22.6 , 23.04, 23.48, 23.92, 24.36,  
 24.8 , 25.24, 25.68, 26.12, 26.56, 27. ]),  
 <a list of 50 Patch objects>)

In [309]:

hist\_of\_levels('employment\_status\_clients','loannumber\_x')

E:\Python\Anaconda\lib\site-packages\numpy\lib\function\_base.py:838: RuntimeWarning: invalid value encountered in true\_divide  
 return n/db/n.sum(), bin\_edges