```
In [1]: import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        import datetime as dt
        import numpy as np
        import scipy
        %matplotlib inline
```

Describe data

```
In [7]: def change_datatype(df):
            int_cols = list(df.select_dtypes(include=['int']).columns)
            for col in int_cols:
                if ((np.max(df[col]) <= 127) and(np.min(df[col] >= -128))):
                    df[col] = df[col].astype(np.int8)
                elif ((np.max(df[col]) <= 32767) and(np.min(df[col] >= -32768))):
                    df[col] = df[col].astype(np.int16)
                elif ((np.max(df[col]) <= 2147483647) and(np.min(df[col] >= -2147483648))):
                    df[col] = df[col].astype(np.int32)
                else:
                    df[col] = df[col].astype(np.int64)

        def change_datatype_float(df):
            float_cols = list(df.select_dtypes(include=['float']).columns)
            for col in float_cols:
                df[col] = df[col].astype(np.float32) #code from https://www.kaggle.com/c/kkbox-chur
        n-prediction-challenge
```

Train table

```
In [44]: df_train = pd.read_csv('train_v2.csv')
```

The train table doesn't have any missing values or outliers.

```
In [18]: df_train.isnull().values.any()

Out[18]: False
```

```
In [20]: df_train.describe()
```

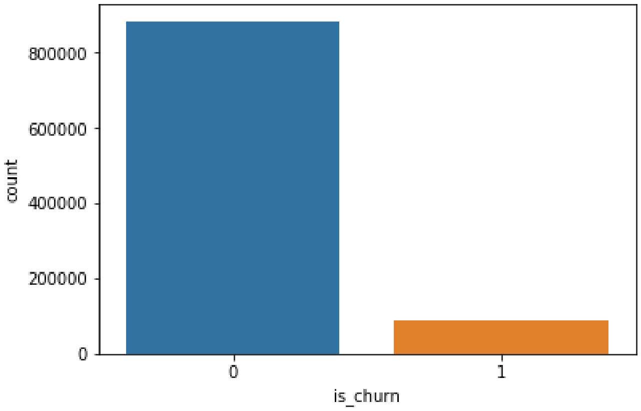
Out[20]:

	is_churn
count	970960.000000
mean	0.089942
std	0.286099
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

```
In [7]: sns.countplot(x='is_churn',data=df_train)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

```
Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x7f10288a51d0>
```



Members Table

```
In [15]: df_members = pd.read_csv('members_v3.csv')
```

```
In [5]: df_members.describe()
```

Out[5]:

	city	bd	registered_via	registration_init_time
count	6.769473e+06	6.769473e+06	6.769473e+06	6.769473e+06
mean	3.847358e+00	9.795794e+00	5.253069e+00	2.014518e+07
std	5.478359e+00	1.792590e+01	2.361398e+00	2.318601e+04
min	1.000000e+00	-7.168000e+03	-1.000000e+00	2.004033e+07
25%	1.000000e+00	0.000000e+00	4.000000e+00	2.014042e+07
50%	1.000000e+00	0.000000e+00	4.000000e+00	2.015101e+07
75%	4.000000e+00	2.100000e+01	7.000000e+00	2.016060e+07
max	2.200000e+01	2.016000e+03	1.900000e+01	2.017043e+07

The members table only has missing gender values.

```
In [39]: df_members.isnull().sum()
```

```
Out[39]: msno                            0
         city                            0
         bd                              0
         gender                    4429505
         registered_via                  0
         registration_init_time          0
         dtype: int64
```

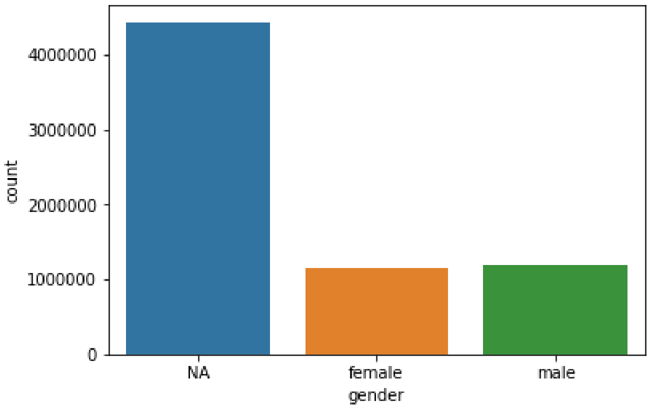
The majority of the users did not list their genders. Of the users who did list their gender, the number of male users is almost equal to the number of female users.

```
In [10]: df_members_na = df_members.fillna('NA')
In [5]: df_members['registration_init_time'] = pd.to_datetime(df_members['registration_init_time'],
         format='%Y%m%d')
```

```
In [11]: sns.countplot(x='gender',data=df_members_na)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

```
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x7f0ffe7185f8>
```



There is a large percentage of users whose gender is unknown, 65%. Since the percentage is so large, I will not use the gender in any predictive model.

```
In [23]: df_members['gender'].isnull().sum()/len(df_members)
```

```
Out[23]: 0.65433527838873129
```

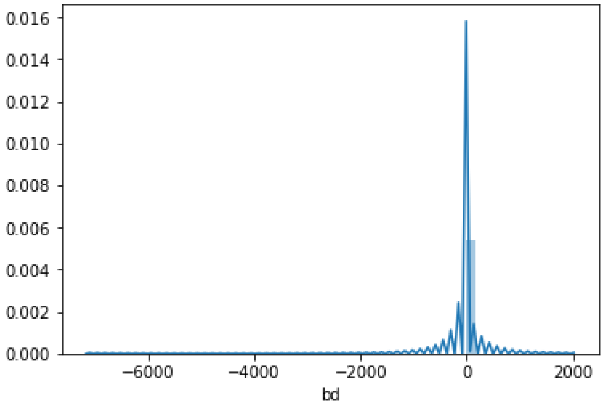
The feature bd gives the users age, and there are some values outside of a range that makes sense. The min age is -7168 and the max age is 2016. The proportion of bad ages is 0.08%. Since the bad ages comprise such a small proportion of the dataset, I will remove them.

```
In [26]: len(df_members[(df_members['bd']<0) | (df_members['bd']>100)])/len(df_members)
```

```
Out[26]: 0.00083477694644767777
```

```
In [23]: sns.distplot(df_members['bd'])
```

```
Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x7f0ffe0c7550>
```



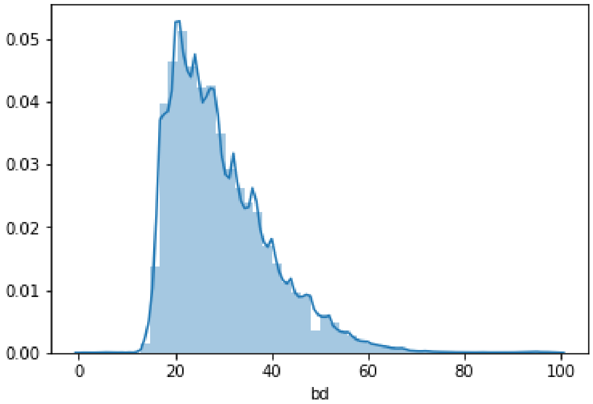
Plotting only sensible values for the age, shows that most of the users are teenageers and young adults. The distribution of ages peaks around 25, and then decreases.

In [26]:

```
sns.distplot(df_members['bd'][(df_members.bd>0) & (df_members.bd<100)])
```

Out[26]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0ffdcd3cc0>
```



The max value of registered_via should be 16, and the min value should be 3. However, there are values that lie outside of this range. The proportion of bad values is 0.06%. Since the bad values comprise such a small proportion of the dataset, I will remove them.

In [33]:

```
len(df_members[(df_members['registered_via']<3) | (df_members['registered_via']>16)])/len(d
f_members)
```

Out[33]:

```
0.000586308564935557
```

The most popular registration method is method 4. Methods 3 and 9 are the next most popular, followed by method 7, about half as many signups as 3 & 9). The rest of the methods are not very popular.

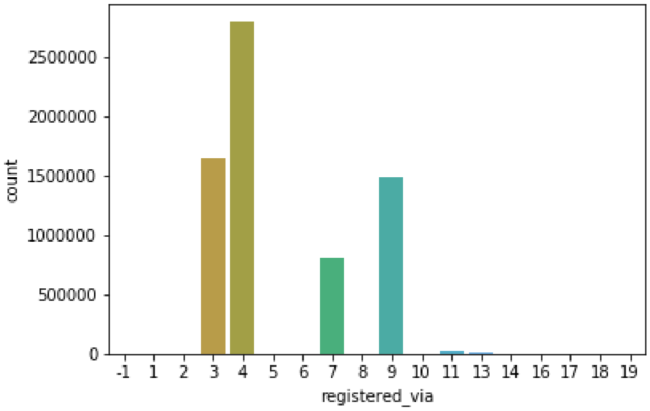
In [13]:

```
sns.countplot(x='registered_via',data=df_members_na)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

Out[13]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0ffe4f5be0>
```

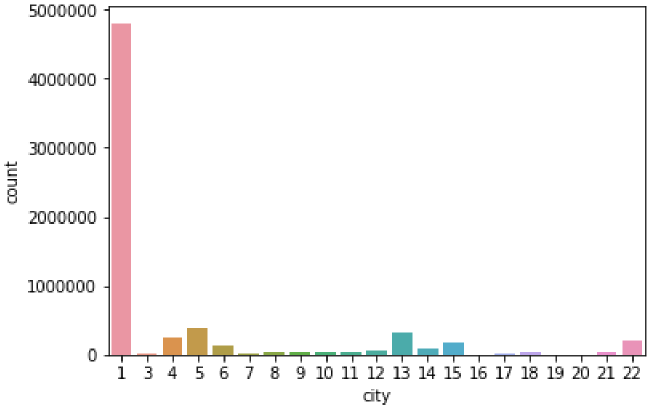


Most of the users live in the city labeled 1. The other 21 cities do not have many registrations. According to the description of the dataset, there aren't any incorrect values for registered_via.

```
In [18]: sns.countplot(x='city',data=df_members_na)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

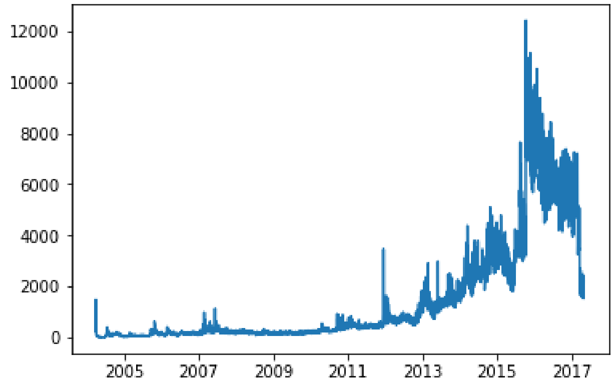
```
Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x7f0ffe0650b8>
```



The service wasn't very popular until 2010, at which point signups began to slowly increase. After 2015, signups increased strongly. However, recently new registrations have been declining.

```
In [14]: plt.plot(pd.to_datetime(df_members['registration_init_time'], format='%Y%m%d').value_counts
().sort_index())
```

```
Out[14]: [<matplotlib.lines.Line2D at 0x7f036ac904e0>]
```



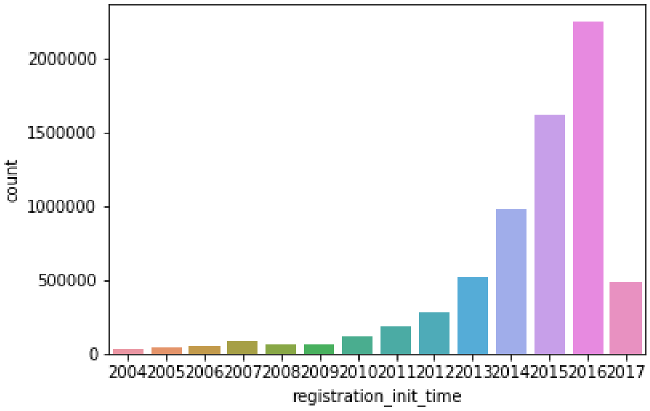
The range of the registration_init_time is correct according to the description of the dataset.

```
In [44]: year = pd.to_datetime(df_members['registration_init_time'], format='%Y%m%d').dt.year
year = year.to_frame()
```

```
In [46]: sns.countplot(x='registration_init_time',data=year)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/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[46]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1e6763f6d8>
```



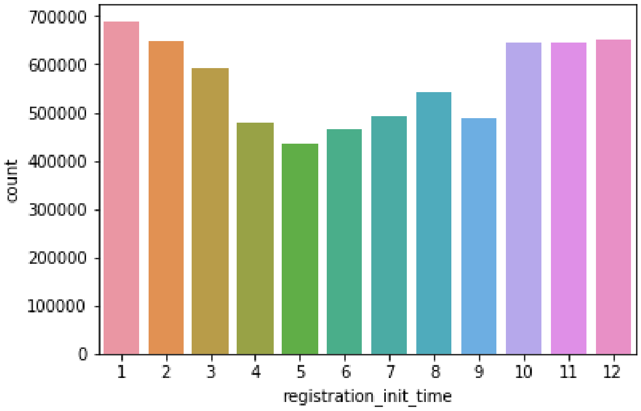
```
In [40]: months = pd.to_datetime(df_members['registration_init_time'], format='%Y%m%d').dt.month
         months=months.to_frame()
```

The beginning of the year and the end of the year have higher initial registrations than the other months. The least popular months are April and May. Note that this trend might not hold over all years.

```
In [43]: sns.countplot(x='registration_init_time',data=months)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/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[43]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1e676bb320>
```



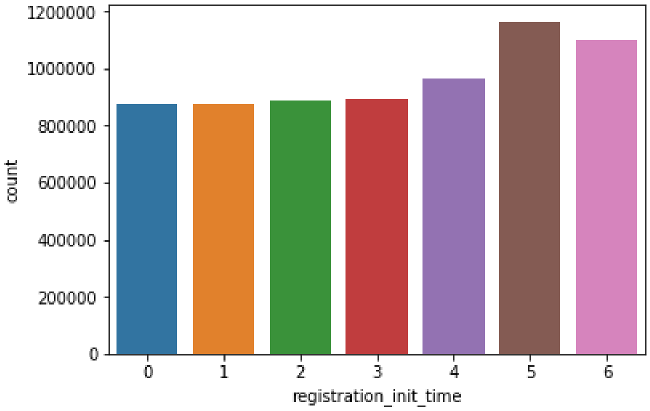
```
In [45]: dayofweek = pd.to_datetime(df_members['registration_init_time'], format='%Y%m%d').dt.dayofweek
         dayofweek = dayofweek.to_frame()
```

Weekend days have higher initial registrations than the rest of the week. The weekday with the most registrations is Friday.

```
In [47]: sns.countplot(x='registration_init_time',data=dayofweek)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

```
Out[47]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1e675d7e80>
```



Transactions Table

```
In [57]: df_trans = pd.read_csv('transactions_v2.csv')
```

```
In [21]: mem = df_trans.memory_usage(index=True).sum()
         print(mem/ 1024**2," MB")
```

```
98.2596664429  MB
```

```
In [22]: change_datatype(df_trans)
```

```
In [23]: mem = df_trans.memory_usage(index=True).sum()
         print(mem/ 1024**2," MB")
```

```
34.1179895401  MB
```

```
In [22]: df_trans['transaction_date'] = pd.to_datetime(df_trans['transaction_date'], format='%Y%m%d'
         )
```

```
In [23]: df_trans['membership_expire_date'] = pd.to_datetime(df_trans['membership_expire_date'], for
         mat='%Y%m%d')
```

There aren't any missing values in the transactions table.

```
In [6]: df_trans.isnull().values.any()
```

```
Out[6]: False
```

```
In [4]: df_trans.describe()
```

Out[4]:

	payment_method_id	payment_plan_days	plan_list_price	actual_amount_paid	is_auto_renew	transaction
count	1.431009e+06	1.431009e+06	1.431009e+06	1.431009e+06	1.431009e+06	1.431009e+0
mean	3.791835e+01	6.601770e+01	2.817870e+02	2.813172e+02	7.853025e-01	2.016848e+0
std	4.964805e+00	1.024864e+02	4.351861e+02	4.354200e+02	4.106124e-01	4.858797e+0
min	2.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.015010e+0
25%	3.600000e+01	3.000000e+01	9.900000e+01	9.900000e+01	1.000000e+00	2.017023e+0
50%	4.000000e+01	3.000000e+01	1.490000e+02	1.490000e+02	1.000000e+00	2.017031e+0
75%	4.100000e+01	3.000000e+01	1.490000e+02	1.490000e+02	1.000000e+00	2.017032e+0
max	4.100000e+01	4.500000e+02	2.000000e+03	2.000000e+03	1.000000e+00	2.017033e+0

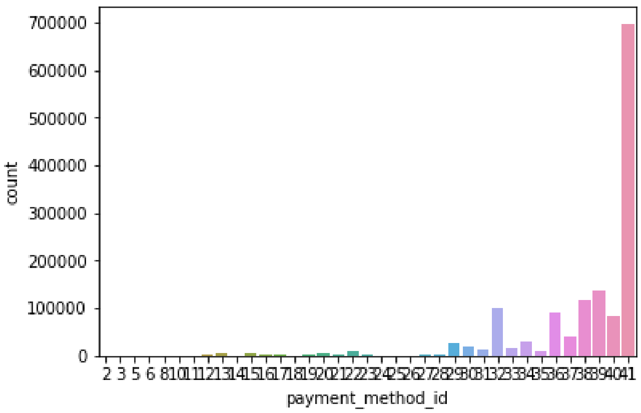
Transactions data table doesn't have any missing values and the ranges of the features look good.

The most popular payment method is 41. Otherwise, payment methods greater than 30 are the most popular.

```
In [53]: sns.countplot(x='payment_method_id',data=df_trans)

         /home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
         l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
           stat_data = remove_na(group_data)

Out[53]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1e6754fe48>
```

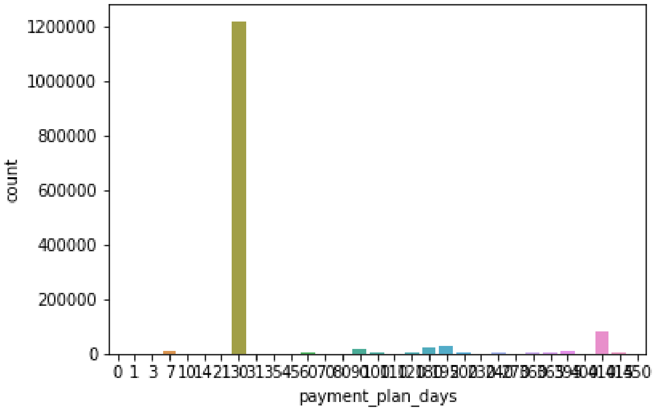


The most popular payment plan duration is 30 days, followed by multiples of months (60 days, 90 days) and 1 week.

In [57]:
```
sns.countplot(x='payment_plan_days',data=df_trans)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

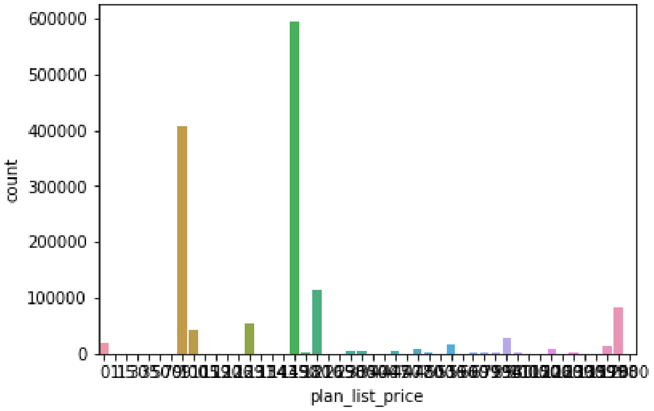
Out[57]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1df239acc0>
```



In [59]:
```
sns.countplot(x='plan_list_price',data=df_trans)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

Out[59]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1df2201208>
```



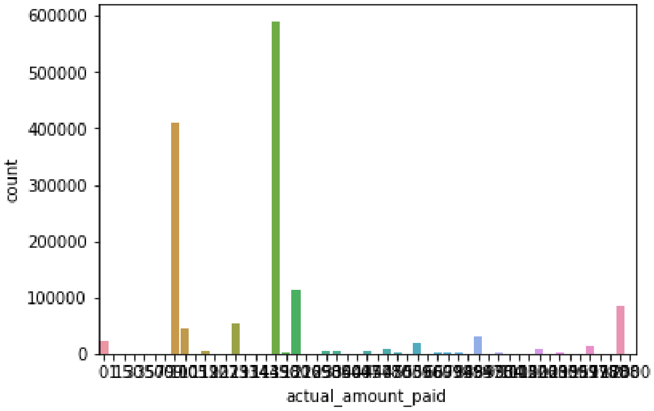
In [60]:

```
sns.countplot(x='actual_amount_paid',data=df_trans)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

Out[60]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1df2085b00>
```



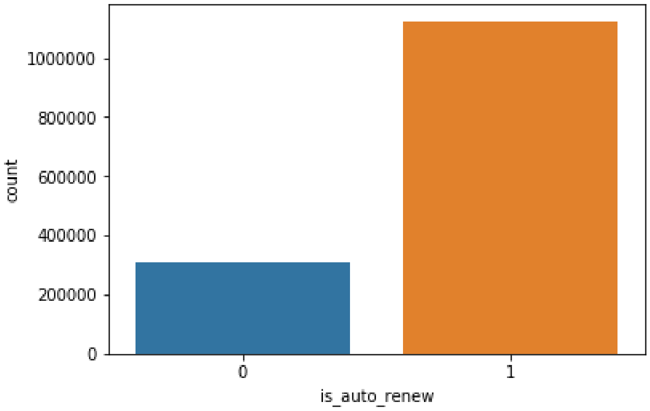
In [9]:

```
sns.countplot(x='is_auto_renew',data=df_trans)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

Out[9]:

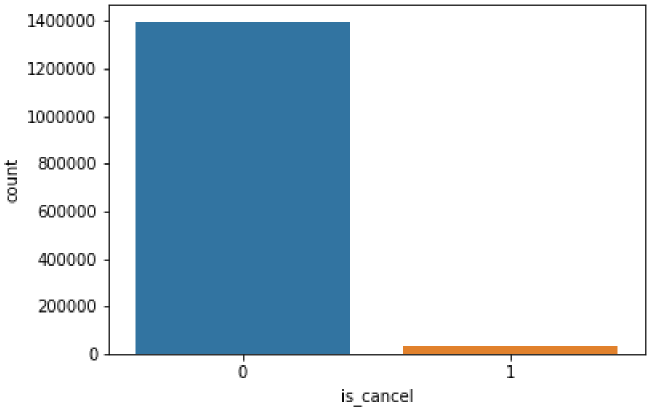
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f70e78ed6d8>
```



```
In [11]: sns.countplot(x='is_cancel',data=df_trans)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

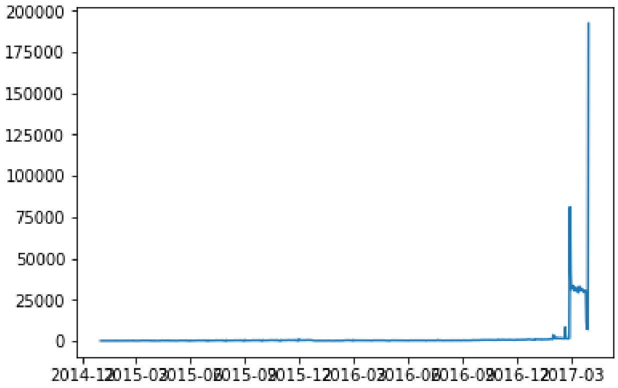
```
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x7f70e7873128>
```



Most of the trasactions occur in early 2017.

```
In [15]: plt.plot(pd.to_datetime(df_trans['transaction_date'], format='%Y%m%d').value_counts().sort_
index())
```

```
Out[15]: [<matplotlib.lines.Line2D at 0x7f0384872eb8>]
```



```
In [25]: mask = (df_trans['membership_expire_date'].dt.year<=2017) & (df_trans['membership_expire_da
te'].dt.year>2016)
```

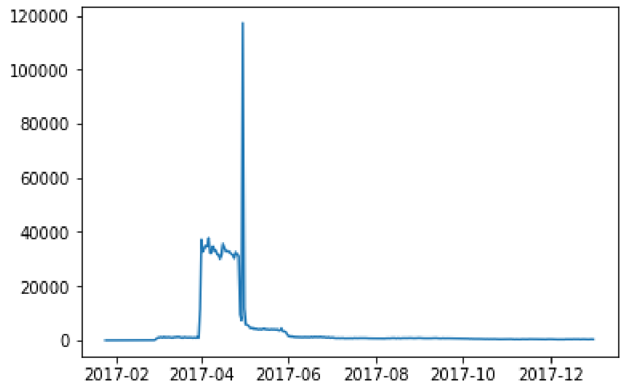
Most of the memberships expire in April and May 2017.

In [27]:

```
plt.plot(df_trans['membership_expire_date'][mask].value_counts().sort_index())
```

Out[27]:

```
[<matplotlib.lines.Line2D at 0x7f03813dada0>]
```



Logs table

In [5]:

```
df_logs = pd.read_csv('user_logs_v2.csv')
```

In [6]:

```
mem = df_logs.memory_usage(index=True).sum()
print(mem/ 1024**2," MB")
change_datatype_float(df_logs)
change_datatype(df_logs)

mem = df_logs.memory_usage(index=True).sum()
print(mem/ 1024**2," MB")
df_logs['date'] = pd.to_datetime(df_logs['date'], format='%Y%m%d')
```

```
1263.17800903  MB
526.324214935  MB
```

The logs table doesn't have any null values, and the ranges for the features seems sensible.

In [28]:

```
df_logs.describe()
```

Out[28]:

	date	num_25	num_50	num_75	num_985	num_100	num_unq	to
count	1.839636e+07	1.839636e+07	1.839636e+07	1.839636e+07	1.839636e+07	1.839636e+07	1.839636e+07	1.839
mean	2.017032e+07	6.191401e+00	1.508789e+00	9.413759e-01	1.079905e+00	3.028246e+01	2.903615e+01	7.904
std	8.916720e+00	1.342827e+01	3.908539e+00	1.924840e+00	3.518409e+00	4.203641e+01	3.219866e+01	1.013
min	2.017030e+07	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000
25%	2.017031e+07	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	7.000000e+00	8.000000e+00	1.959
50%	2.017032e+07	2.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	1.700000e+01	1.800000e+01	4.582
75%	2.017032e+07	7.000000e+00	2.000000e+00	1.000000e+00	1.000000e+00	3.700000e+01	3.800000e+01	9.848
max	2.017033e+07	5.639000e+03	9.120000e+02	5.080000e+02	1.561000e+03	4.110700e+04	4.925000e+03	9.194

In [30]:

```
df_logs.isnull().values.any()
```

Out[30]:

```
False
```

total_secs gives the total amount of listening time for the day in seconds. Tha maximum value is 11534622 seconds. This is much more than 24 hours, unless multiple people were using this users' account. Additionally, this amount of listening is also large based on the users previous listening. So I will remove all values of total_secs greater than 24 hours.

```
In [22]: df_logs['total_secs'].values.argmax()
```

```
Out[22]: 11534622
```

```
In [24]: df_logs.msno[11534622]
```

```
Out[24]: 'sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4='
```

```
In [26]: df_logs[df_logs.msno=='sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4='].sort_values(by='date'
         )
```

Out[26]:

	msno	date	num_25	num_50	num_75	num_985	nu
2471755	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170301	5	0	0	2	11
6217341	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170302	1	1	0	1	3
15065790	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170303	1	1	1	0	10
5101805	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170305	0	0	0	0	8
15974974	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170306	7	0	0	0	4
1511565	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170307	0	2	0	0	10
9762376	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170308	23	3	3	2	10
4827279	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170309	12	3	0	2	17
3849933	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170310	1	0	0	1	22
17440397	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170311	32	16	11	4	38
13201372	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170312	6	0	1	0	20
16141473	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170313	0	0	0	1	9
3876130	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170314	3	3	0	0	15
7445604	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170315	7	2	1	1	3
5911716	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170316	1	1	0	0	13
8130440	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170317	0	0	0	0	3
6642669	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170318	20	10	5	2	26
2756732	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170319	1	0	0	0	23
300123	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170321	1	0	0	1	31
7605466	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170322	2	1	0	3	12
4578398	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170323	1	0	0	0	6
11534622	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170324	227	104	95	108	411
15192982	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170325	0	0	0	0	27
11214738	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170326	39	5	6	2	10
17290938	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170327	14	5	1	2	5
10003199	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170328	4	0	1	1	4
14335551	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170329	1	0	1	0	13
7125010	sfWgePzzK7p+HF5X/IzTitF34mnDy6LvFqHEOIvRPc4=	20170330	5	1	0	2	19

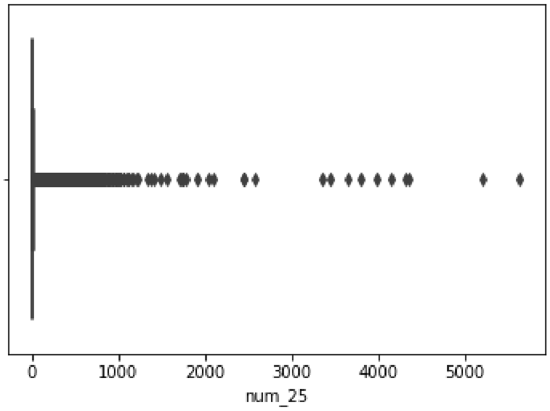
```
In [38]: mask = df_logs['total_secs']/60/60 <=24
         df_logs = df_logs[mask]
```

The boxplot for num_25, shows the number of songs listened up to 25% of their length. The boxplot shows that there are several values that should be considered outliers based on the IQR.

```
In [28]: sns.boxplot(x='num_25',data=df_logs)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

```
Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1fb7d05128>
```

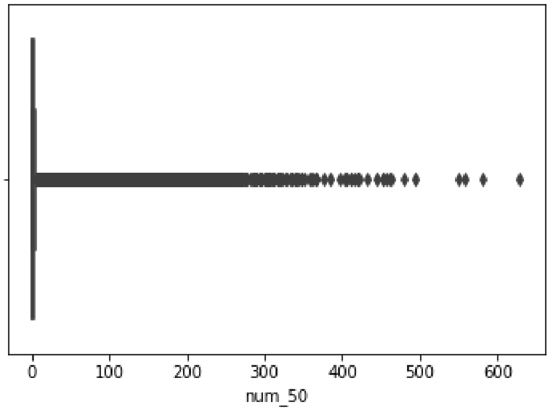


The boxplot for num_50, shows the number of songs listened up to 50% of their length. The boxplot shows that there are several values that should be considered outliers based on the IQR.

```
In [29]: sns.boxplot(x='num_50',data=df_logs)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

```
Out[29]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1fb76cca58>
```

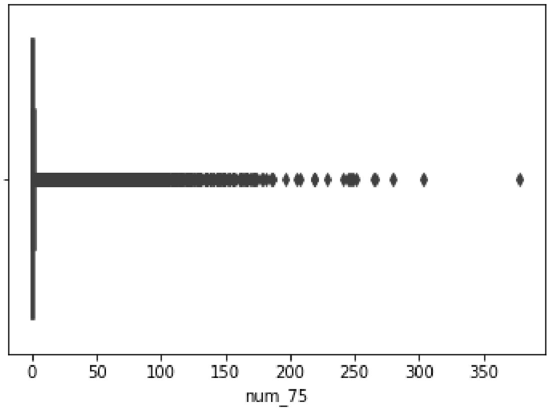


The boxplot for num_75, shows the number of songs listened up to 75% of their length. The boxplot shows that there are several values that should be considered outliers based on the IQR.

In [30]:
```
sns.boxplot(x='num_75',data=df_logs)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

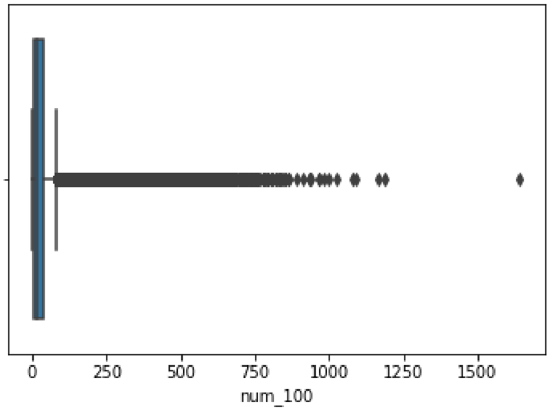
Out[30]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1fb7f199b0>
```



In [31]:
```
sns.boxplot(x='num_100',data=df_logs)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

Out[31]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1fb8022978>
```

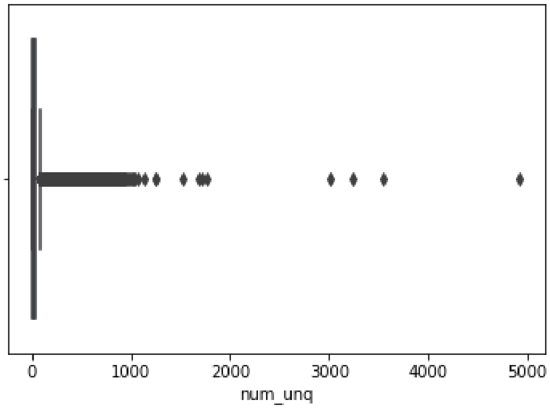


The boxplot for num_100, shows the number of songs listened up to 100% of their length. The boxplot shows that there are several values that should be considered outliers based on the IQR.

```
In [32]: sns.boxplot(x='num_unq',data=df_logs)
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

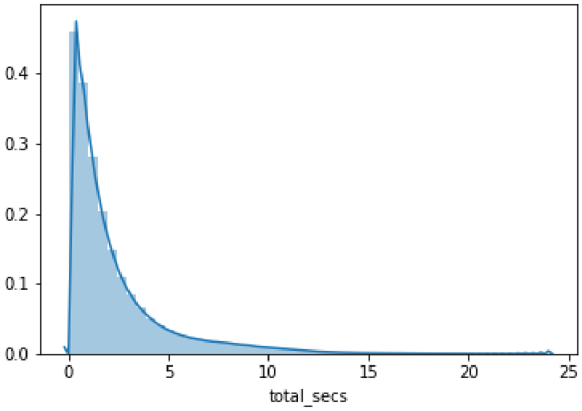
```
Out[32]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1fb8154860>
```



The distribution of total_secs (in hours) is exponentially distributed.

```
In [33]: sns.distplot(df_logs['total_secs']/60/60)
```

```
Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1fb7628b00>
```



After removing the implausible values for num_secs the num_unq distribution is highly left skewed.

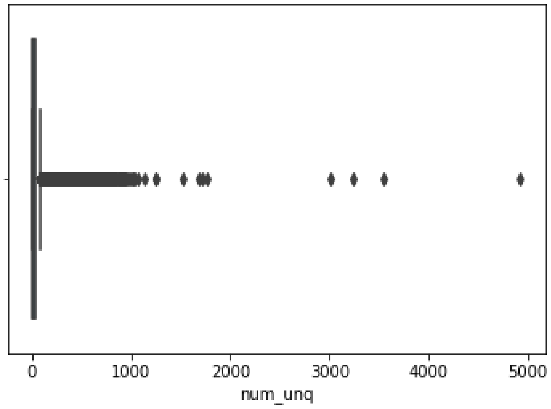
In [40]:

```
sns.boxplot(df_logs['num_unq'])
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

Out[40]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5ea3546be0>
```



Feature Engineering

Construct a new feature 'discount' that is the difference between the price of the plan and the price that the user actually paid.

In [10]:

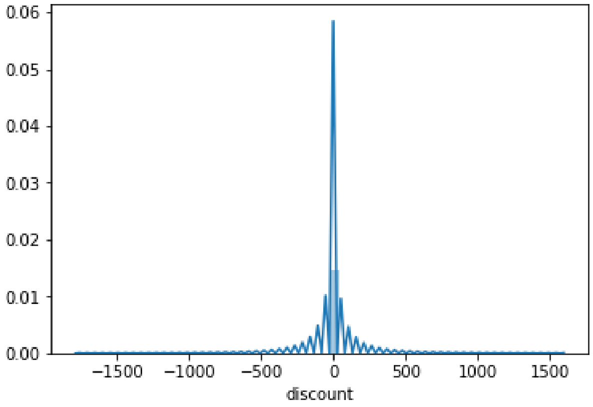
```
df_trans['discount'] = df_trans['plan_list_price'] - df_trans['actual_amount_paid']
```

In [38]:

```
sns.distplot(df_trans['discount'])
```

Out[38]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f03b78fd198>
```



In addition to the discount amount, I will make a new feature 'is_discount' that indicates whether or not this transaction was a discount transaction.

In [11]:

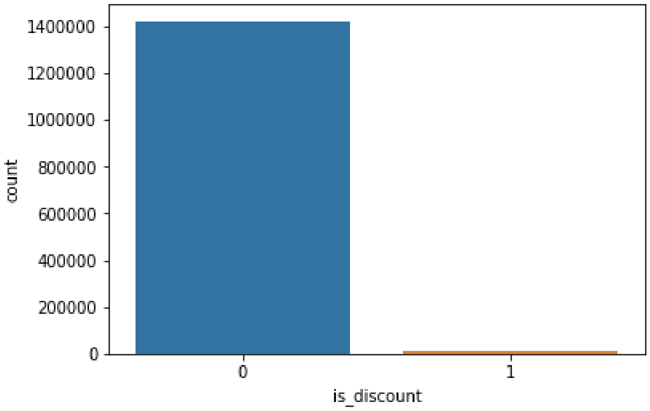
```
df_trans['is_discount'] = df_trans.discount.apply(lambda x: 1 if x > 0 else 0)
```

Most of the transactions were not discount transactions.

In [14]:
```
sns.countplot(df_trans['is_discount'])
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

Out[14]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff2d97a52e8>
```



Here I engineer new features to take the total count of the songs listened to for each user.

In [7]:
```
df_logs_new = pd.DataFrame()
```

In [9]:
```
df_logs_new[['total_num_25','total_num_50','total_num_75','total_num_985','total_num_100',
'total_num_unq','total_secs']] = df_logs.groupby('msno')[['num_25','num_50','num_75','num_9
85','num_100','num_unq','total_secs']].sum()
```

Instead of using data for all the transactions, use the mode for payment_method_id, payment_plan_days, plan_list_price, actual_amount_paid, is_auto_renew.

In [29]:
```
df_trans_new = pd.DataFrame()
```

In [30]:
```
df_trans_new = df_trans.groupby('msno')[['payment_method_id','payment_plan_days','plan_list
_price','actual_amount_paid','is_auto_renew']].agg(lambda x: scipy.stats.mode(x)[0])
```

Instead of keeping every transaction date, keep the count of how many transactions a user had.

In [31]:
```
#check this
df_trans_new[['transaction_date_count']] = df_trans.groupby('msno')[['transaction_date']].c
ount()
```

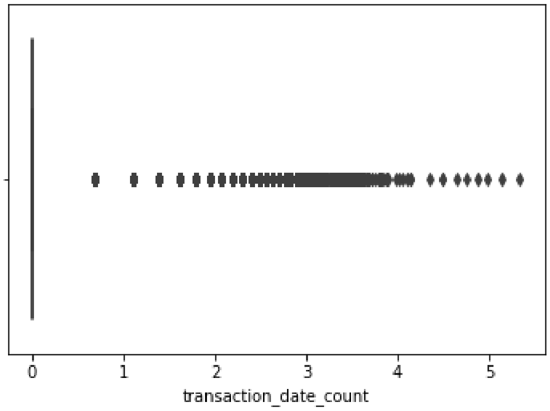
In [ ]:
```

```

```
In [48]: sns.boxplot(np.log(all_data['transaction_date_count']))
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

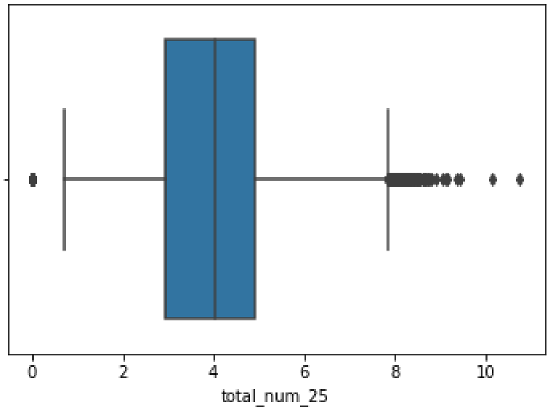
```
Out[48]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847ab8df98>
```



```
In [51]: sns.boxplot(np.log(all_data['total_num_25']))
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/ipykernel/__main_
_.py:1: RuntimeWarning: divide by zero encountered in log
  if __name__ == '__main__':
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

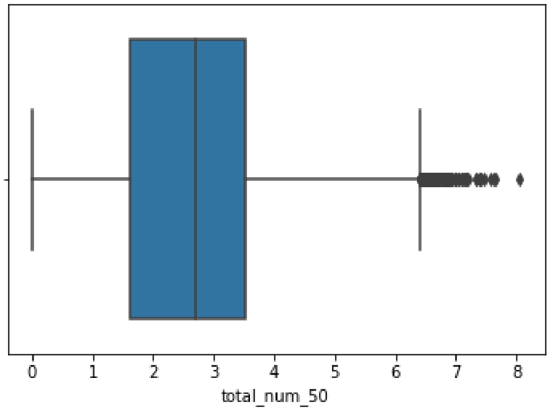
```
Out[51]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847aaaac50>
```



```
In [52]: sns.boxplot(np.log(all_data['total_num_50']))
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/ipykernel/__main_
_.py:1: RuntimeWarning: divide by zero encountered in log
  if __name__ == '__main__':
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

```
Out[52]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847aa70278>
```

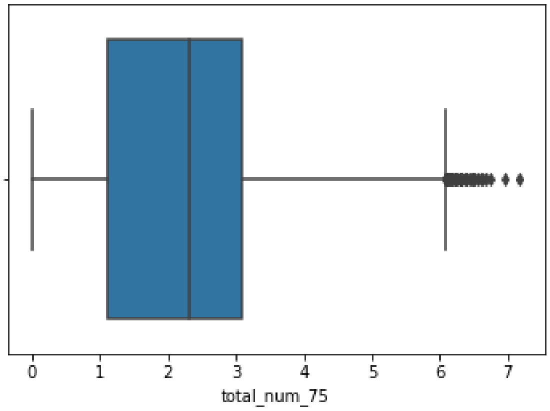


Create a feature that is true if a user ever canceled.

```
In [53]: sns.boxplot(np.log(all_data['total_num_75']))
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/ipykernel/__main_
_.py:1: RuntimeWarning: divide by zero encountered in log
  if __name__ == '__main__':
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

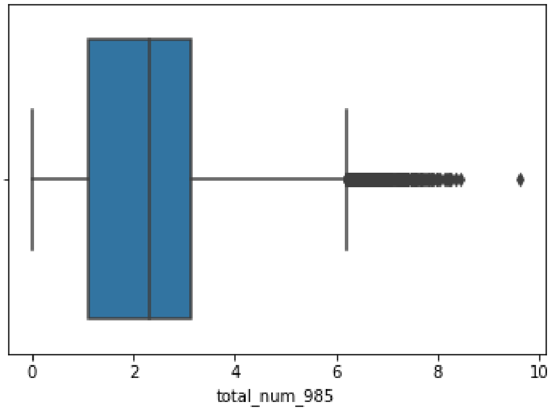
```
Out[53]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847aa10d30>
```



```
In [77]: test = df_trans.groupby('msno')[['is_cancel']].max()
In [54]: sns.boxplot(np.log(all_data['total_num_985']))
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/ipykernel/__main_
_.py:1: RuntimeWarning: divide by zero encountered in log
  if __name__ == '__main__':
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

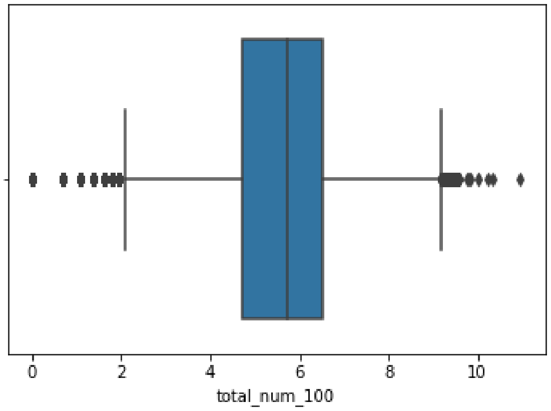
```
Out[54]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847a93b208>
```



```
In [55]: sns.boxplot(np.log(all_data['total_num_100']))
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/ipykernel/__main_
_.py:1: RuntimeWarning: divide by zero encountered in log
  if __name__ == '__main__':
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

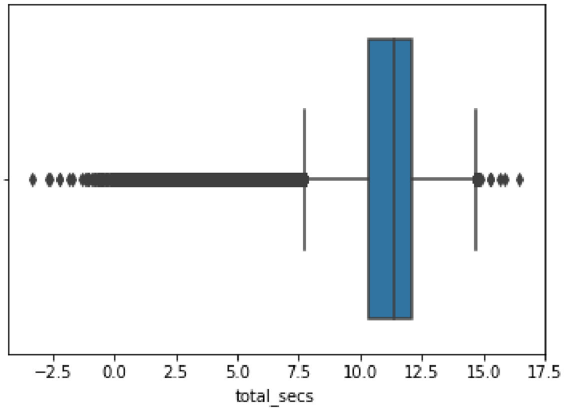
```
Out[55]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847aa10dd8>
```



```
In [57]: sns.boxplot(np.log(all_data['total_secs']))
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

```
Out[57]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847a89e780>
```



```
In [84]: test = test.reset_index()
         df_trans = pd.merge(test,df_trans,on='msno')
```

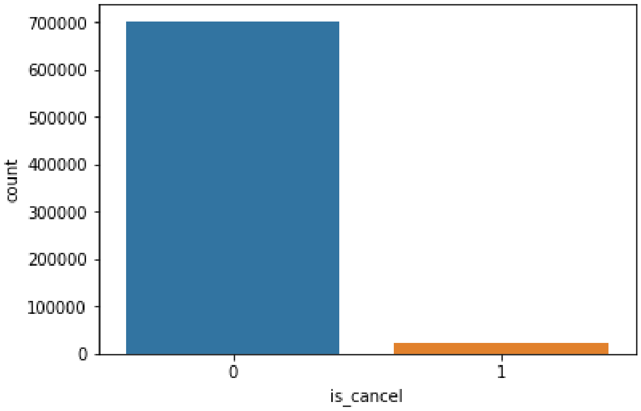
Out[84]:

	msno	is_cancel_x	payment_method_id	payment_plan_days	pla
0	+++IZseRRiQS9aaSkH6cMYU6bGDcxUieAi/tH67sC5s=	0	22	395	159
1	+++hVY1rZox/33YtvDgmKA2Frg/2qhkz12B9ylCvh8o=	0	41	30	99
2	+++l/EXNMLTijfLBa8p2TUVVVp2aFGSuUI/h7mLmthw=	0	39	30	149
3	+++snpr7pmobhLKUgSHTv/mpkqgBT0tQJ0zQj6qKrqc=	0	41	30	149
4	++/9R3sX37CjxbY/AaGvbwr3QkwElKBCtSvVzhCBDOk=	0	41	30	149

```
In [41]: sns.countplot(df_trans['is_cancel'])
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:1460: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

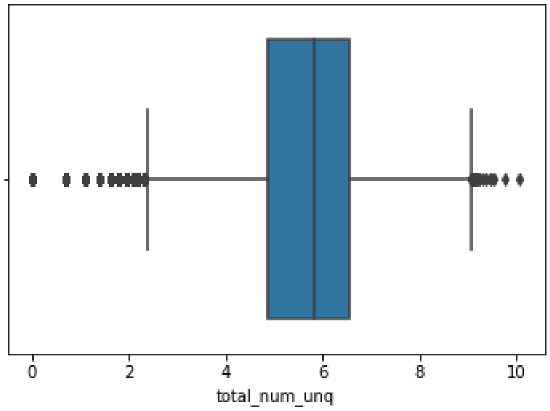
```
Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847afbafd0>
```



```
In [56]: sns.boxplot(np.log(all_data['total_num_unq']))
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

```
Out[56]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847a8b4278>
```

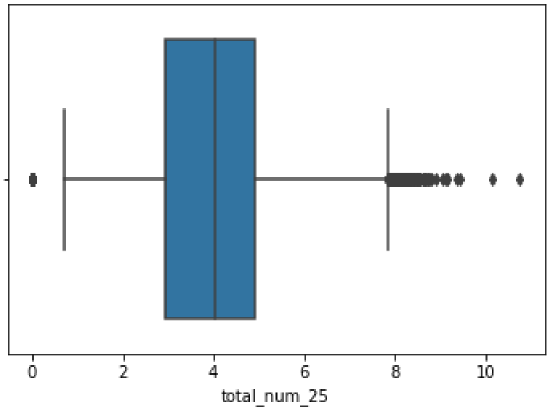


Make all tables include the same msno numbers.

```
In [51]: sns.boxplot(np.log(all_data['total_num_25']))
```

```
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/ipykernel/__main_
_.py:1: RuntimeWarning: divide by zero encountered in log
  if __name__ == '__main__':
/home/rebecca/anaconda3/envs/my_projects_env/lib/python3.6/site-packages/seaborn/categorica
l.py:462: FutureWarning: remove_na is deprecated and is a private function. Do not use.
  box_data = remove_na(group_data)
```

```
Out[51]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847aaaac50>
```



```
In [26]: df_trans = df_trans[df_trans.msno.isin(memb.msno)]
```

```
In [17]: len(set(df_members.msno) - set(df_trans.msno))

Out[17]: 5692039
```

```
In [19]: len(set(df_members.msno))

Out[19]: 6769473
```

```
In [20]: len(set(df_trans.msno))

Out[20]: 1197050
```

```
In [13]: df_logs_new.to_csv('logs_summed.csv')
```

```
In [26]: df_members_new = df_members[df_members.msno.isin(df_logs.msno)]
```

```
In [108]: df_members_new.to_csv('members_new.csv')
```

```
In [45]: df_train_new = df_train[df_train.msno.isin(df_logs.msno)]
```

```
In [47]: df_train_new.to_csv('train_new.csv')
```

```
In [88]: df_trans_new.to_csv('trans_new.csv')
```

```
In [41]: df_members = pd.read_csv('members_new.csv')
         df_train = pd.read_csv('train_new.csv')
         df_trans = pd.read_csv('trans_new.csv')
         df_logs = pd.read_csv('logs_summed.csv')
```

```
In [55]: df_members['city'] = df_members['city'].astype('category')
         df_members['gender'] = df_members['gender'].astype('category')
         df_members['registered_via'] = df_members['registered_via'].astype('category')
         df_members['registration_init_time'] = pd.to_datetime(df_members['registration_init_time'], format='%Y%m%d').dt.year.astype('category')
         df_trans['payment_method_id'] = df_trans['payment_method_id'].astype('category')
```

```
In [113]: all_data = pd.merge(df_logs,df_trans,on='msno',how='inner')

          ---------------------------------------------------------------------------
          ValueError                                Traceback (most recent call last)
          ValueError: Buffer dtype mismatch, expected 'Python object' but got 'long'

          Exception ignored in: 'pandas._libs.lib.is_bool_array'
          ValueError: Buffer dtype mismatch, expected 'Python object' but got 'long'
```

```
In [114]: all_data = pd.merge(all_data,df_members,on='msno',how='inner')

          ---------------------------------------------------------------------------
          ValueError                                Traceback (most recent call last)
          ValueError: Buffer dtype mismatch, expected 'Python object' but got 'long'

          Exception ignored in: 'pandas._libs.lib.is_bool_array'
          ValueError: Buffer dtype mismatch, expected 'Python object' but got 'long'

          ---------------------------------------------------------------------------
          ValueError                                Traceback (most recent call last)
          ValueError: Buffer dtype mismatch, expected 'Python object' but got 'long'

          Exception ignored in: 'pandas._libs.lib.is_bool_array'
          ValueError: Buffer dtype mismatch, expected 'Python object' but got 'long'

          ---------------------------------------------------------------------------
          ValueError                                Traceback (most recent call last)
          ValueError: Buffer dtype mismatch, expected 'Python object' but got 'long'

          Exception ignored in: 'pandas._libs.lib.is_bool_array'
          ValueError: Buffer dtype mismatch, expected 'Python object' but got 'long'
```