Expedia_modeling

April 30, 2022

[1]:
```
import numpy as np
import math
import pandas as pd
import matplotlib.pyplot as plt


import statsmodels.api as sm
from statsmodels.sandbox.regression.gmm import IV2SLS
from sklearn.linear_model import LassoCV
import scipy.stats as stats


import warnings


warnings.filterwarnings('ignore')
pd.set_option('display.max_columns', None)
```

0.1 Import and prepare data

[2]:
```
df = pd.read_csv("data/train.csv")
df['date'] = pd.to_datetime(df['date_time']).dt.date # Create Date column for␣
 ↪fuzzier matching
df = df[df['random_bool'] == 0] # Only keep real data
df
```

[2]:

	srch_id	date_time	site_id	visitor_location_country_id	\
60	6	2013-06-05 12:27:51	14	100	
61	6	2013-06-05 12:27:51	14	100	
62	6	2013-06-05 12:27:51	14	100	
63	6	2013-06-05 12:27:51	14	100	
64	6	2013-06-05 12:27:51	14	100	
…	…	…	…	…	
9917525	665574	2013-05-21 11:06:37	24	216	
9917526	665574	2013-05-21 11:06:37	24	216	
9917527	665574	2013-05-21 11:06:37	24	216	
9917528	665574	2013-05-21 11:06:37	24	216	

```
9917529   665574  2013-05-21 11:06:37       24                          216

         visitor_hist_starrating  visitor_hist_adr_usd  prop_country_id  \
60                           NaN                   NaN              100
61                           NaN                   NaN              100
62                           NaN                   NaN              100
63                           NaN                   NaN              100
64                           NaN                   NaN              100
…                              …                     …                …
9917525                      NaN                   NaN              117
9917526                      NaN                   NaN              117
9917527                      NaN                   NaN              117
9917528                      NaN                   NaN              117
9917529                      NaN                   NaN              117

         prop_id  prop_starrating  prop_review_score  prop_brand_bool  \
60         10759                0                2.0                0
61         22135                0                5.0                0
62         52376                2                0.0                1
63        104251                3                4.0                1
64        118866                2                4.5                1
…            …                …                  …                …
9917525    32019                4                3.5                0
9917526    33959                4                3.0                1
9917527    35240                4                0.0                0
9917528    94437                4                0.0                0
9917529    99509                4                4.5                1

         prop_location_score1  prop_location_score2  \
60                       1.95                   NaN
61                       1.95                   NaN
62                       1.95                   NaN
63                       1.95                   NaN
64                       1.95                   NaN
…                           …                     …
9917525                  2.48                0.0551
9917526                  2.20                0.3344
9917527                  1.79                   NaN
9917528                  2.94                0.0928
9917529                  2.08                0.0344

         prop_log_historical_price  position  price_usd  promotion_flag  \
60                            0.00         4      97.63               0
61                            0.00         6     115.03               0
62                            0.00         2      86.03               0
63                            0.00         1     145.00               0
64                            0.00         3     183.66               0
```

```
…                              …       …       …              …
9917525                     4.53       2     66.07              0
9917526                     4.39       6     67.10              0
9917527                     4.64       8     73.91              0
9917528                     4.64       7     66.07              0
9917529                     4.64       4     82.06              0

         srch_destination_id  srch_length_of_stay  srch_booking_window  \
60                     21106                    1                    5
61                     21106                    1                    5
62                     21106                    1                    5
63                     21106                    1                    5
64                     21106                    1                    5
…                      …                      …                    …
9917525                19246                    2                    7
9917526                19246                    2                    7
9917527                19246                    2                    7
9917528                19246                    2                    7
9917529                19246                    2                    7

         srch_adults_count  srch_children_count  srch_room_count  \
60                       2                    0                1
61                       2                    0                1
62                       2                    0                1
63                       2                    0                1
64                       2                    0                1
…                      …                      …                …
9917525                  1                    0                1
9917526                  1                    0                1
9917527                  1                    0                1
9917528                  1                    0                1
9917529                  1                    0                1

         srch_saturday_night_bool  srch_query_affinity_score  \
60                              0                        NaN
61                              0                        NaN
62                              0                        NaN
63                              0                        NaN
64                              0                        NaN
…                             …                          …
9917525                         0                        NaN
9917526                         0                        NaN
9917527                         0                        NaN
9917528                         0                        NaN
9917529                         0                        NaN

         orig_destination_distance  random_bool  comp1_rate  comp1_inv  \
```

```
60                          652.84            0          NaN         NaN
61                          652.84            0          NaN         NaN
62                          652.85            0          NaN         NaN
63                          652.84            0          NaN         NaN
64                          652.78            0          NaN         NaN
…                              …              …          …           …
9917525                       NaN             0          NaN         NaN
9917526                       NaN             0          NaN         NaN
9917527                       NaN             0          NaN         NaN
9917528                       NaN             0          NaN         NaN
9917529                       NaN             0          NaN         NaN

         comp1_rate_percent_diff  comp2_rate  comp2_inv  \
60                           NaN         NaN        NaN
61                           NaN         NaN        NaN
62                           NaN         NaN        NaN
63                           NaN         NaN        NaN
64                           NaN         NaN        NaN
…                              …           …          …
9917525                      NaN         1.0        0.0
9917526                      NaN         0.0        0.0
9917527                      NaN         1.0        0.0
9917528                      NaN         1.0        0.0
9917529                      NaN         0.0        0.0

         comp2_rate_percent_diff  comp3_rate  comp3_inv  \
60                           NaN         NaN        NaN
61                           NaN         NaN        NaN
62                           NaN         NaN        NaN
63                           NaN         NaN        NaN
64                           NaN         NaN        NaN
…                              …           …          …
9917525                     22.0         1.0        0.0
9917526                      NaN         0.0        0.0
9917527                     55.0         0.0        0.0
9917528                     43.0         1.0        0.0
9917529                      NaN         0.0        0.0

         comp3_rate_percent_diff  comp4_rate  comp4_inv  \
60                           NaN         NaN        NaN
61                           NaN         NaN        NaN
62                           NaN         NaN        NaN
63                           NaN         NaN        NaN
64                           NaN         NaN        NaN
…                              …           …          …
9917525                    127.0        -1.0        0.0
9917526                      NaN         0.0        0.0
```

```
9917527                      NaN         0.0        0.0
9917528                     43.0        -1.0        0.0
9917529                      NaN         0.0        0.0

         comp4_rate_percent_diff  comp5_rate  comp5_inv  \
60                           NaN         NaN        0.0
61                           NaN         NaN        0.0
62                           NaN         NaN        NaN
63                           NaN         NaN        NaN
64                           NaN         NaN        NaN
…                              …           …          …
9917525                     27.0         1.0        0.0
9917526                     16.0         1.0        0.0
9917527                     16.0         0.0        0.0
9917528                     12.0        -1.0        0.0
9917529                     16.0         0.0        0.0

         comp5_rate_percent_diff  comp6_rate  comp6_inv  \
60                           NaN         NaN        NaN
61                           NaN         NaN        NaN
62                           NaN         NaN        NaN
63                           NaN         NaN        NaN
64                           NaN         NaN        NaN
…                              …           …          …
9917525                     22.0         NaN        NaN
9917526                     22.0         NaN        NaN
9917527                      3.0         NaN        NaN
9917528                     12.0         NaN        NaN
9917529                      NaN         NaN        NaN

         comp6_rate_percent_diff  comp7_rate  comp7_inv  \
60                           NaN         NaN        NaN
61                           NaN         NaN        NaN
62                           NaN         NaN        NaN
63                           NaN         NaN        NaN
64                           NaN         NaN        NaN
…                              …           …          …
9917525                      NaN         NaN        NaN
9917526                      NaN         NaN        NaN
9917527                      NaN         NaN        NaN
9917528                      NaN         NaN        NaN
9917529                      NaN         NaN        NaN

         comp7_rate_percent_diff  comp8_rate  comp8_inv  \
60                           NaN         NaN        NaN
61                           NaN         NaN        NaN
62                           NaN         NaN        NaN
```

```
63                          NaN          NaN         NaN
64                          NaN          NaN         NaN
…                           …            …           …
9917525                     NaN          NaN         NaN
9917526                     NaN          NaN         NaN
9917527                     NaN          NaN         NaN
9917528                     NaN          NaN         NaN
9917529                     NaN          NaN         NaN

         comp8_rate_percent_diff  click_bool  gross_bookings_usd  \
60                           NaN           0                 NaN
61                           NaN           0                 NaN
62                           NaN           0                 NaN
63                           NaN           1              162.38
64                           NaN           0                 NaN
…                            …             …                 …
9917525                      NaN           0                 NaN
9917526                      NaN           1              154.34
9917527                      NaN           0                 NaN
9917528                      NaN           0                 NaN
9917529                      NaN           0                 NaN

         booking_bool        date
60                  0  2013-06-05
61                  0  2013-06-05
62                  0  2013-06-05
63                  1  2013-06-05
64                  0  2013-06-05
…                   …          …
9917525             0  2013-05-21
9917526             1  2013-05-21
9917527             0  2013-05-21
9917528             0  2013-05-21
9917529             0  2013-05-21

[6977878 rows x 55 columns]
```

[3]:
```
df.columns
```

[3]:
```
Index(['srch_id', 'date_time', 'site_id', 'visitor_location_country_id',
       'visitor_hist_starrating', 'visitor_hist_adr_usd', 'prop_country_id',
       'prop_id', 'prop_starrating', 'prop_review_score', 'prop_brand_bool',
       'prop_location_score1', 'prop_location_score2',
       'prop_log_historical_price', 'position', 'price_usd', 'promotion_flag',
       'srch_destination_id', 'srch_length_of_stay', 'srch_booking_window',
       'srch_adults_count', 'srch_children_count', 'srch_room_count',
       'srch_saturday_night_bool', 'srch_query_affinity_score',
```

```
       'orig_destination_distance', 'random_bool', 'comp1_rate', 'comp1_inv',
       'comp1_rate_percent_diff', 'comp2_rate', 'comp2_inv',
       'comp2_rate_percent_diff', 'comp3_rate', 'comp3_inv',
       'comp3_rate_percent_diff', 'comp4_rate', 'comp4_inv',
       'comp4_rate_percent_diff', 'comp5_rate', 'comp5_inv',
       'comp5_rate_percent_diff', 'comp6_rate', 'comp6_inv',
       'comp6_rate_percent_diff', 'comp7_rate', 'comp7_inv',
       'comp7_rate_percent_diff', 'comp8_rate', 'comp8_inv',
       'comp8_rate_percent_diff', 'click_bool', 'gross_bookings_usd',
       'booking_bool', 'date'],
      dtype='object')
```

0.2 Filter df to get "same searches"

[4]:
```
duplicatedMask = df.duplicated(subset=['date','srch_destination_id', 'prop_id',
 ↪'srch_length_of_stay','srch_booking_window'],
                               keep=False)
dup = df[duplicatedMask] # All rows with duplicates where the 5 parameters are␣
 ↪exactly the same
dup # Unique rows are dropped
```

[4]:

	srch_id	date_time	site_id	visitor_location_country_id	\
498	45	2013-05-18 09:25:08	14	100	
499	45	2013-05-18 09:25:08	14	100	
500	45	2013-05-18 09:25:08	14	100	
501	45	2013-05-18 09:25:08	14	100	
502	45	2013-05-18 09:25:08	14	100	
…	…	…	…	…	
9916418	665500	2012-12-22 06:55:24	5	219	
9916419	665500	2012-12-22 06:55:24	5	219	
9916420	665500	2012-12-22 06:55:24	5	219	
9916424	665500	2012-12-22 06:55:24	5	219	
9916426	665500	2012-12-22 06:55:24	5	219	

	visitor_hist_starrating	visitor_hist_adr_usd	prop_country_id	\
498	NaN	NaN	219	
499	NaN	NaN	219	
500	NaN	NaN	219	
501	NaN	NaN	219	
502	NaN	NaN	219	
…	…	…	…	
9916418	NaN	NaN	219	
9916419	NaN	NaN	219	
9916420	NaN	NaN	219	
9916424	NaN	NaN	219	
9916426	NaN	NaN	219	

```
         prop_id  prop_starrating  prop_review_score  prop_brand_bool  \
498         2924                4                4.5                1
499        25444                4                4.0                1
500        31792                2                4.0                1
501        34700                2                3.0                1
502        38213                0                4.0                0
…            …                …                …                …
9916418    80475                3                4.5                0
9916419    98736                4                4.5                1
9916420   110940                3                0.0                0
9916424   126657                4                4.0                1
9916426   135537                5                4.5                1

         prop_location_score1  prop_location_score2  \
498                      2.30                0.0499
499                      2.30                0.0860
500                      2.08                0.0085
501                      2.77                0.0027
502                      0.69                0.0019
…                           …                   …
9916418                  5.65                0.1359
9916419                  5.63                0.0879
9916420                  5.74                0.0695
9916424                  5.96                0.1755
9916426                  3.97                0.0069

         prop_log_historical_price  position  price_usd  promotion_flag  \
498                           5.27         3     174.11               0
499                           5.02         9     174.11               0
500                           4.68        13     102.13               0
501                           4.58        16      87.54               0
502                           4.52        21      68.09               0
…                                …         …         …               …
9916418                       5.82        22     161.00               0
9916419                       5.92        25     143.00               1
9916420                       5.39        18     105.00               0
9916424                       6.02        14     199.00               1
9916426                       6.11        30     179.00               0

         srch_destination_id  srch_length_of_stay  srch_booking_window  \
498                    10948                    1                    0
499                    10948                    1                    0
500                    10948                    1                    0
501                    10948                    1                    0
502                    10948                    1                    0
…                          …                    …                    …
```

```
9916418                4562                     1                    14
9916419                4562                     1                    14
9916420                4562                     1                    14
9916424                4562                     1                    14
9916426                4562                     1                    14

         srch_adults_count  srch_children_count  srch_room_count  \
498                      2                    0                1
499                      2                    0                1
500                      2                    0                1
501                      2                    0                1
502                      2                    0                1
…                      …                    …                …
9916418                  2                    0                1
9916419                  2                    0                1
9916420                  2                    0                1
9916424                  2                    0                1
9916426                  2                    0                1

         srch_saturday_night_bool  srch_query_affinity_score  \
498                             1                        NaN
499                             1                        NaN
500                             1                        NaN
501                             1                        NaN
502                             1                        NaN
…                             …                          …
9916418                         1                        NaN
9916419                         1                        NaN
9916420                         1                        NaN
9916424                         1                        NaN
9916426                         1                        NaN

         orig_destination_distance  random_bool  comp1_rate  comp1_inv  \
498                            NaN            0         NaN        NaN
499                            NaN            0         NaN        NaN
500                            NaN            0         NaN        NaN
501                            NaN            0         NaN        NaN
502                            NaN            0         NaN        NaN
…                              …              …         …          …
9916418                     183.97            0         NaN        NaN
9916419                     185.19            0         NaN        NaN
9916420                     184.51            0         NaN        NaN
9916424                     184.56            0         NaN        NaN
9916426                     187.56            0         NaN        NaN

         comp1_rate_percent_diff  comp2_rate  comp2_inv  \
498                          NaN         0.0        0.0
```

```
499                          NaN          NaN         NaN
500                          NaN          NaN         NaN
501                          NaN          NaN         NaN
502                          NaN          NaN         NaN
…                            …            …           …
9916418                      NaN          NaN         1.0
9916419                      NaN          1.0         0.0
9916420                      NaN          NaN         NaN
9916424                      NaN          0.0         0.0
9916426                      NaN          NaN         NaN

         comp2_rate_percent_diff  comp3_rate  comp3_inv  \
498                          NaN          0.0         0.0
499                          NaN          0.0         0.0
500                          NaN          0.0         0.0
501                          NaN          NaN         NaN
502                          NaN          NaN         NaN
…                            …            …           …
9916418                      NaN          NaN         NaN
9916419                     11.0          1.0         0.0
9916420                      NaN          NaN         NaN
9916424                      NaN          0.0         0.0
9916426                      NaN          0.0         0.0

         comp3_rate_percent_diff  comp4_rate  comp4_inv  \
498                          NaN          NaN         NaN
499                          NaN          NaN         NaN
500                          NaN          NaN         NaN
501                          NaN          NaN         NaN
502                          NaN          NaN         NaN
…                            …            …           …
9916418                      NaN          NaN         1.0
9916419                     13.0          1.0         0.0
9916420                      NaN          NaN         NaN
9916424                      NaN          0.0         0.0
9916426                      NaN          NaN         NaN

         comp4_rate_percent_diff  comp5_rate  comp5_inv  \
498                          NaN          NaN         NaN
499                          NaN          NaN         NaN
500                          NaN          NaN         NaN
501                          NaN          NaN         NaN
502                          NaN          NaN         NaN
…                            …            …           …
9916418                      NaN          NaN         NaN
9916419                     25.0          0.0         0.0
9916420                      NaN          NaN         NaN
```

```
9916424                       NaN         0.0        0.0
9916426                       NaN         0.0        0.0

         comp5_rate_percent_diff  comp6_rate  comp6_inv  \
498                          NaN         NaN        NaN
499                          NaN         NaN        NaN
500                          NaN         NaN        NaN
501                          NaN         NaN        NaN
502                          NaN         NaN        NaN
…                            …           …          …
9916418                      NaN         NaN        NaN
9916419                     10.0         NaN        NaN
9916420                      NaN         NaN        NaN
9916424                     15.0         NaN        NaN
9916426                      NaN         NaN        NaN

         comp6_rate_percent_diff  comp7_rate  comp7_inv  \
498                          NaN         NaN        NaN
499                          NaN         NaN        NaN
500                          NaN         NaN        NaN
501                          NaN         NaN        NaN
502                          NaN         NaN        NaN
…                            …           …          …
9916418                      NaN         NaN        NaN
9916419                      NaN         NaN        NaN
9916420                      NaN         NaN        NaN
9916424                      NaN         NaN        NaN
9916426                      NaN         NaN        NaN

         comp7_rate_percent_diff  comp8_rate  comp8_inv  \
498                          NaN         0.0        0.0
499                          NaN         0.0        0.0
500                          NaN        -1.0        0.0
501                          NaN         0.0        0.0
502                          NaN         NaN        NaN
…                            …           …          …
9916418                      NaN         NaN        NaN
9916419                      NaN         0.0        0.0
9916420                      NaN         NaN        NaN
9916424                      NaN         0.0        0.0
9916426                      NaN         0.0        0.0

         comp8_rate_percent_diff  click_bool  gross_bookings_usd  \
498                          NaN           0                 NaN
499                          NaN           0                 NaN
500                          9.0           0                 NaN
501                          NaN           0                 NaN
```

```
502                          NaN           0                  NaN
…                            …             …                  …
9916418                      NaN           0                  NaN
9916419                     10.0           0                  NaN
9916420                      NaN           0                  NaN
9916424                      NaN           0                  NaN
9916426                      NaN           0                  NaN

         booking_bool        date
498                 0  2013-05-18
499                 0  2013-05-18
500                 0  2013-05-18
501                 0  2013-05-18
502                 0  2013-05-18
…                   …          …
9916418             0  2012-12-22
9916419             0  2012-12-22
9916420             0  2012-12-22
9916424             0  2012-12-22
9916426             0  2012-12-22

[85144 rows x 55 columns]
```

[5]:
```
dupGroups = dup.groupby(['date','srch_destination_id', 'prop_id',
                         ␣
 ↪'srch_length_of_stay','srch_booking_window'])[['price_usd', 'position']].agg(
                                  ['median', 'std', 'mean', 'count', 'min',␣
 ↪'max'])
dupGroups['range'] = dupGroups['price_usd']['max'] -␣
 ↪dupGroups['price_usd']['min']
dupGroups # Group by the aforementioned criteria
```

[5]:
```
price_usd  \
  median
date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
309.00
                               25075   1                   1
259.00
                               34036   1                   1
200.00
                               41278   1                   0
499.63
                               53204   1                   1
135.00
…
…
```

```
2013-06-30 11230               118327  1                   1
107.00
                               122691  1                   1
174.00
                               125320  1                   1
89.00
                               135372  1                   1
75.00
                               138986  1                   1
154.00

           \
      std
date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
14.142136
                               25075   1                   1
0.000000
                               34036   1                   1
0.000000
                               41278   1                   0
0.890955
                               53204   1                   1
0.000000
…
…
2013-06-30 11230               118327  1                   1
0.000000
                               122691  1                   1
7.071068
                               125320  1                   1
0.000000
                               135372  1                   1
0.000000
                               138986  1                   1
0.000000

        \
  mean
date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
309.00
                               25075   1                   1
259.00
                               34036   1                   1
200.00
                               41278   1                   0
```

```
499.63
                              53204   1                   1
135.00
…
…
2013-06-30 11230              118327  1                   1
107.00
                              122691  1                   1
174.00
                              125320  1                   1
89.00
                              135372  1                   1
75.00
                              138986  1                   1
154.00

      \
count
date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
2
                              25075   1                   1
2
                              34036   1                   1
2
                              41278   1                   0
2
                              53204   1                   1
2
…
…
2013-06-30 11230              118327  1                   1
2
                              122691  1                   1
2
                              125320  1                   1
2
                              135372  1                   1
2
                              138986  1                   1
2

       \
  min
date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
299.0
```

```
                              25075   1                   1
259.0
                              34036   1                   1
200.0
                              41278   1                   0
499.0
                              53204   1                   1
135.0
…
…
2013-06-30 11230              118327  1                   1
107.0
                              122691  1                   1
169.0
                              125320  1                   1
89.0
                              135372  1                   1
75.0
                              138986  1                   1
154.0

        \
   max
date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
319.00
                              25075   1                   1
259.00
                              34036   1                   1
200.00
                              41278   1                   0
500.26
                              53204   1                   1
135.00
…
…
2013-06-30 11230              118327  1                   1
107.00
                              122691  1                   1
179.00
                              125320  1                   1
89.00
                              135372  1                   1
75.00
                              138986  1                   1
154.00
```

```
position  \
 median
date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
3.0
                               25075   1                   1
5.0
                               34036   1                   1
23.5
                               41278   1                   0
1.5
                               53204   1                   1
26.5
…
…
2013-06-30 11230               118327  1                   1
31.0
                               122691  1                   1
18.5
                               125320  1                   1
7.5
                               135372  1                   1
28.5
                               138986  1                   1
14.0

          \
     std
date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
0.000000
                               25075   1                   1
1.414214
                               34036   1                   1
3.535534
                               41278   1                   0
0.707107
                               53204   1                   1
3.535534
…
…
2013-06-30 11230               118327  1                   1
4.242641
                               122691  1                   1
4.949747
                               125320  1                   1
7.778175
```

```
                              135372  1                   1
0.707107
                              138986  1                   1
8.485281

      \
mean
date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
3.0
                              25075   1                   1
5.0
                              34036   1                   1
23.5
                              41278   1                   0
1.5
                              53204   1                   1
26.5
…
…
2013-06-30 11230               118327  1                   1
31.0
                              122691  1                   1
18.5
                              125320  1                   1
7.5
                              135372  1                   1
28.5
                              138986  1                   1
14.0

      \
count
date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
2
                              25075   1                   1
2
                              34036   1                   1
2
                              41278   1                   0
2
                              53204   1                   1
2
…
…
2013-06-30 11230               118327  1                   1
```

```
2
                              122691  1                   1
2
                              125320  1                   1
2
                              135372  1                   1
2
                              138986  1                   1
2

    \
min
date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
3
                               25075   1                   1
4
                               34036   1                   1
21
                               41278   1                   0
1
                               53204   1                   1
24
…
..
2013-06-30 11230               118327  1                   1
28
                              122691  1                   1
15
                              125320  1                   1
2
                              135372  1                   1
28
                              138986  1                   1
8

    \
max
date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
3
                               25075   1                   1
6
                               34036   1                   1
26
                               41278   1                   0
2
```

```
                                53204   1                   1
29
…
..
2013-06-30 11230                118327  1                   1
34
                                122691  1                   1
22
                                125320  1                   1
13
                                135372  1                   1
29
                                138986  1                   1
20

range

date       srch_destination_id prop_id srch_length_of_stay srch_booking_window
2012-11-01 4562                21155   1                   1
20.00
                                25075   1                   1
0.00
                                34036   1                   1
0.00
                                41278   1                   0
1.26
                                53204   1                   1
0.00
…
…
2013-06-30 11230                118327  1                   1
0.00
                                122691  1                   1
10.00
                                125320  1                   1
0.00
                                135372  1                   1
0.00
                                138986  1                   1
0.00

[41685 rows x 13 columns]
```

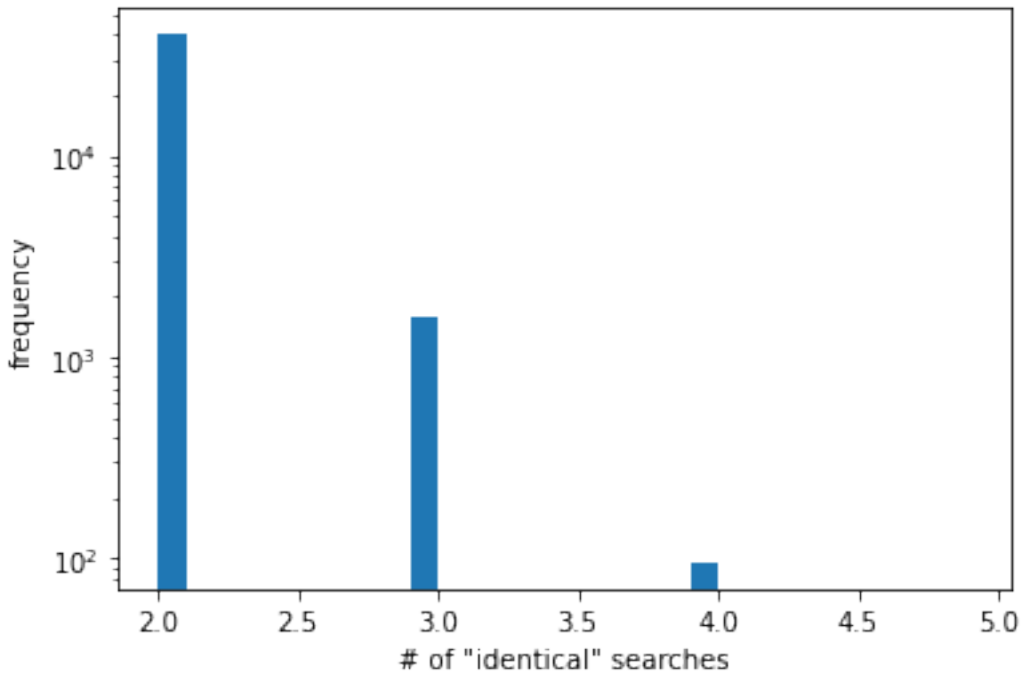
[6]:
```
dupGroups.to_csv('grouped.csv')
```

[7]:
```
dupCount = dupGroups['price_usd']['count']
print("Max number of duplicates: ", max(dupCount))
```

```
bins = np.arange(dupCount.min(), dupCount.max() + 1, 0.1)
plt.hist(dupCount, bins=bins)
plt.yscale('log')
plt.ylabel('frequency')
plt.xlabel('# of "identical" searches')
```

```
Max number of duplicates:  4
```

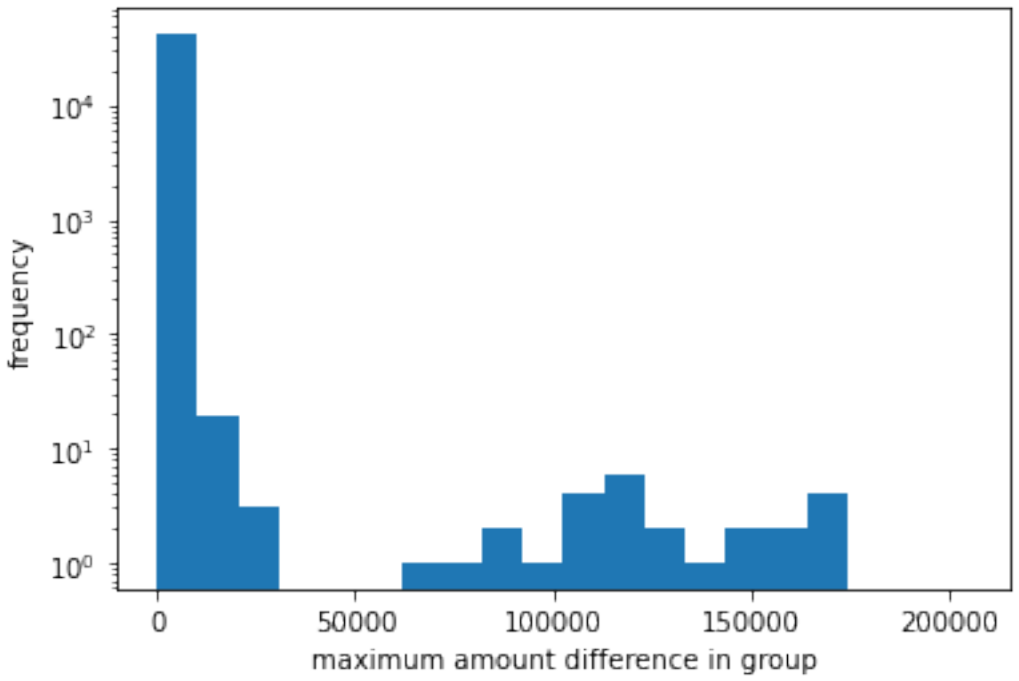
[7]: Text(0.5, 0, '# of "identical" searches')



[8]:
```
dupRange = dupGroups['price_usd']['max'] - dupGroups['price_usd']['min']
print("Maximum Range: ", max(dupRange))
plt.hist(dupRange, bins=np.arange(dupRange.min(), dupRange.max() + 1, (dupRange.
  ↪max() - dupRange.min()) // 20))
plt.yscale('log')
plt.ylabel('frequency')
plt.xlabel('maximum amount difference in group')
```

```
Maximum Range:  205305.0
```

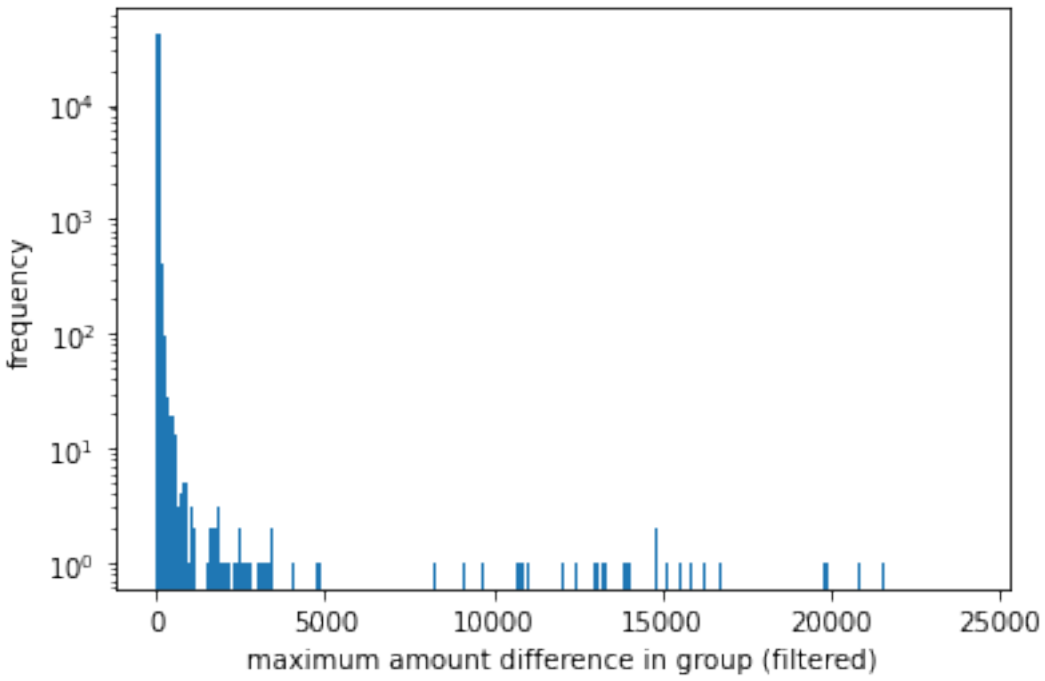
[8]: Text(0.5, 0, 'maximum amount difference in group')



[9]:
```
dupNoOutliers = dupGroups[dupGroups['range'] < 50000.0]['range']
print("Maximum Range: ", max(dupNoOutliers))
plt.hist(dupNoOutliers, bins=np.arange(dupNoOutliers.min(),
                                       dupNoOutliers.max() + 1,
                                       100))
plt.yscale('log')
plt.ylabel('frequency')
plt.xlabel('maximum amount difference in group (filtered)')
```

```
Maximum Range:  24131.0
```

[9]:
```
Text(0.5, 0, 'maximum amount difference in group (filtered)')
```


frequency
10^4
10^3
10^2
10^1
10^0
0
5000
10000
15000
20000
25000
maximum amount difference in group (filtered)

[ ]: