Report

October 7, 2017

1 Predicting AirBnB bookings

Given various data about an AirBnB property, including items such as number of bedrooms, cost, and the economic situation of the surrounding area, we would like to see whether it's possible to predict the average booking rate of an AirBnB property.

To do this, we look at the following datasets: - listings: information about each listing - econ_state: economic information for each state - calendar: day-to-day listing data for five cities

---

2 Examining daily rental information for five cities

We would first like to see detailed rental information for a few cities, to get an idea of how things like price and availability change with location and time. To begin, we load and format the data.

```
In [1]: %matplotlib inline
        import matplotlib
        import matplotlib.pyplot as plt
        import pandas as pd
        import numpy as np
```

```
/Users/mattzhang/py2_kernel/lib/python2.7/site-packages/matplotlib/font_manager.py:
  warnings.warn('Matplotlib is building the font cache using fc-list. This may take
```

```
In [ ]: calendar = pd.read_csv("Data/calendar.csv", parse_dates=["date"], index_col
```

```
In [19]: calendar.keys()
```

```
Out[19]: Index([u'listing_id', u'available', u'price', u'metro_area'], dtype='objec
```

```
In [21]: calendar[:5]
```

```
Out[21]:             listing_id available   price metro_area
         date
         2018-03-05        2515         t  $69.00        NYC
         2018-03-04        2515         t  $69.00        NYC
         2018-03-03        2515         t  $69.00        NYC
         2018-03-02        2515         t  $69.00        NYC
         2018-03-01        2515         t  $69.00        NYC
```

```
In [22]: # convert 't' and 'f' in dataset to 1 and 0
         calendar['available'] = (calendar['available']=='t').astype(int)
```

```
In [24]: # convert price to float
         calendar['price'] = calendar['price'].replace('[\$,)]', '', regex=True).as
```

```
In [25]: calendar[:5]
```

```
Out[25]:             listing_id  available  price metro_area
         date
         2018-03-05        2515          1   69.0        NYC
         2018-03-04        2515          1   69.0        NYC
         2018-03-03        2515          1   69.0        NYC
         2018-03-02        2515          1   69.0        NYC
         2018-03-01        2515          1   69.0        NYC
```

Each unique property has one listing for each day that it's on AirBnB. We would like to rebin so that we get the average availability and price for each property on a monthly basis.

```
In [28]: unique_calendar = calendar.groupby([pd.TimeGrouper('M'), 'metro_area', 'li
```

Let's look at the average rental price for each city.

```
In [40]: calendar['metro_area'].unique()
```

```
Out[40]: array(['NYC', 'denver', 'chicago', 'boston', 'dc'], dtype=object)
```

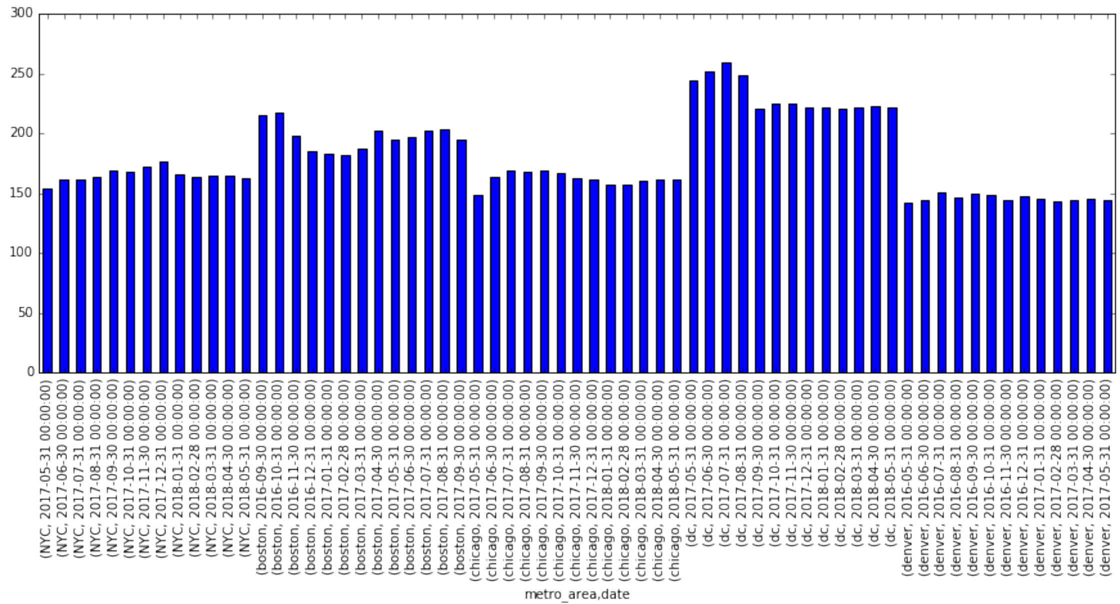
```
In [41]: # the average rental price for each city
         calendar.groupby('metro_area')['price'].aggregate(np.mean)
```

```
Out[41]: metro_area
         NYC        165.904913
         boston     198.438909
         chicago    167.022702
         dc         243.322393
         denver     145.892388
         Name: price, dtype: float64
```

In the following plot we see the average rental price as a function of time for each city. For example, for NYC we have monthly data from May 2017 to May 2018, and we see that the average price hovered around $160. One thing we notice is that while prices may differ by a large amount between cities, the average price within a city does not change much as a function of time. This means that when extracting information such as city or state GDP, we should pay more attention to location than to the year. This is fortunate, since our yearly state economics dataset does not overlap in time with most of our listings data.

```
In [30]: # all 5 cities shown below - monthly average price by all listings
         price = unique_calendar.groupby(level=['metro_area', 'date']).mean().dropr
         price.plot(kind='bar', figsize=(15,5))
```

```
Out[30]: <matplotlib.axes._subplots.AxesSubplot at 0x10f5a8c10>
```

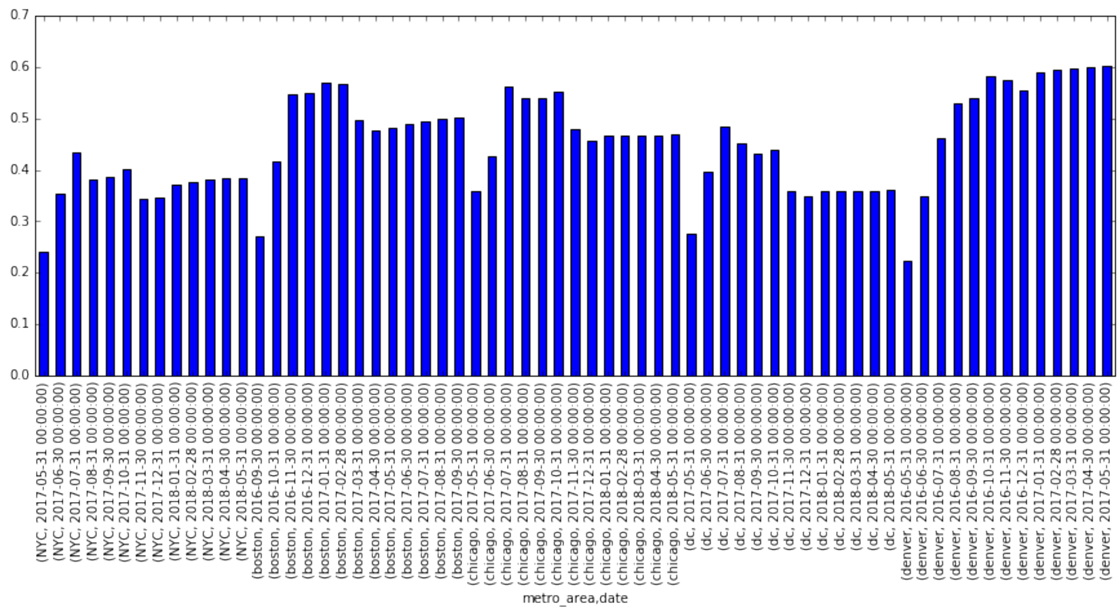


From the next plot we see that each city sees roughly the same trend in availability over the course of a year. Availability is lowest (meaning that bookings are highest) around May. There are large differences in availability for different cities. Combined with the previous plot, this demonstrates that rental rates fluctuate by a large amount with the month, though we would not be able to determine these changes from indicators like rental price.

This is significant because our listings data does not include month! If we had this data, our predictions could be much more accurate!

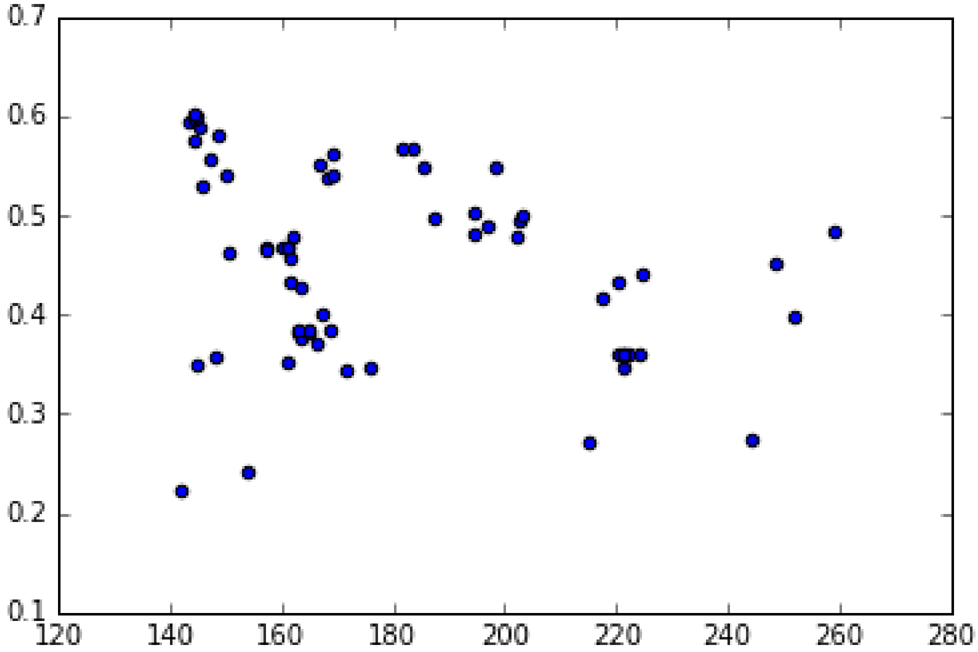
```
In [33]: # all 5 cities shown below - monthly average availability by all listings
         availability = unique_calendar.groupby(level=['metro_area', 'date']).mean(
         availability.plot(kind='bar', figsize=(15,5))
```

```
Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x1203e2590>
```

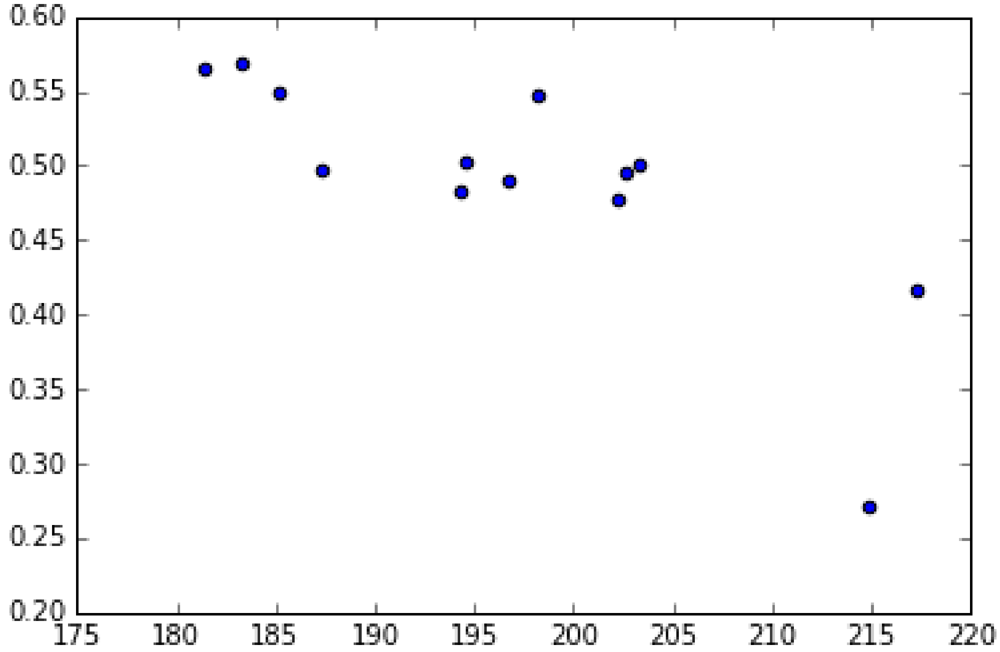


Finally, we look at the relationship between rental price and booking rate. There appears to be little correlation between price and availability when looking across all data, but the difference between cities is surprisingly large. In Boston the most expensive properties are the ones which are booked most often, while in Chicago the opposite is true. This indicates that location data should play a very large role in predicting rental rates.

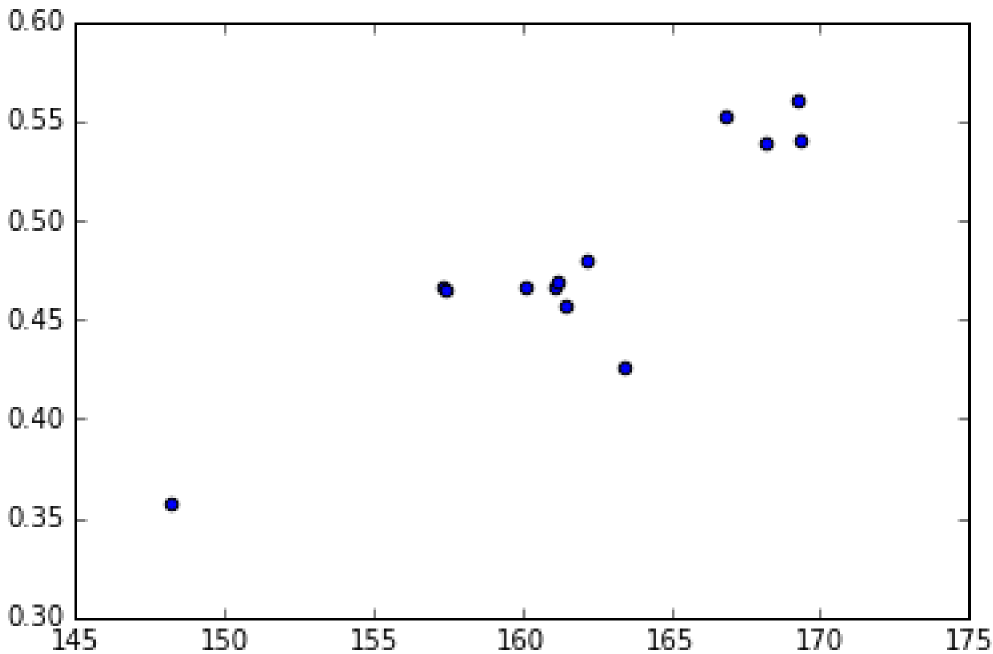
```
In [34]: # monthly price vs. availability by all unique listings
         plt.scatter(price, availability)
         del price, availability
```




```
In [35]: # monthly price vs. availability by all unique listings for Boston
         metro = 'boston'
         metro_unique_calendar = unique_calendar.xs(metro, level='metro_area', drop
         price = metro_unique_calendar.groupby(level=['date']).mean().dropna()['pri
         availability = metro_unique_calendar.groupby(level=['date']).mean().dropna
         plt.scatter(price, availability)
         del price, availability
```



In [38]:
```
# monthly price vs. availability by all unique listings for Chicago
metro = 'chicago'
metro_unique_calendar = unique_calendar.xs(metro, level='metro_area', drop
price = metro_unique_calendar.groupby(level=['date']).mean().dropna()['pri
availability = metro_unique_calendar.groupby(level=['date']).mean().dropna
plt.scatter(price, availability)
del price, availability
```



---

3 Loading and cleaning listings dataset

This dataset shows around 60k listings along the U.S. midwest and northeast. Information such as number of beds, the location, and the rental price are shown.

```
In [10]: pd.set_option('display.max_columns', None)
         listings = pd.read_csv("Data/listings.csv")
```

```
In [3]: listings
```

```
Out[3]:        accommodates                                          amenities  \
        0               2.0  {"Cable TV","Wireless Internet","Air condition...
        1               4.0  {TV,Internet,"Wireless Internet","Air conditio...
        2               4.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
        3               3.0  {TV,Internet,"Wireless Internet","Air conditio...
        4               4.0  {Internet,"Wireless Internet","Air conditionin...
        5               2.0  {TV,"Wireless Internet","Air conditioning",Kit...
        6               4.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
        7               3.0  {"Cable TV",Internet,"Wireless Internet","Air ...
        8               5.0  {TV,Internet,"Wireless Internet","Air conditio...
        9               8.0  {TV,Internet,"Wireless Internet","Air conditio...
        10              2.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
        11              1.0  {TV,Internet,"Wireless Internet","Air conditio...
```

```
12                4.0  {TV,Internet,"Wireless Internet","Air conditio...
13                2.0  {TV,"Cable TV",Internet,"Wireless Internet",Ki...
14                1.0  {TV,"Wireless Internet","Air conditioning",Kit...
15                1.0  {"Cable TV",Internet,"Wireless Internet",Kitch...
16                2.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
17                2.0  {TV,Kitchen,"Free parking on premises",Heating...
18                2.0  {TV,"Wireless Internet","Air conditioning",Kit...
19                3.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
20                4.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
21                2.0  {TV,Internet,"Wireless Internet",Kitchen,Heati...
22                2.0  {TV,Internet,"Wireless Internet",Kitchen,"Free...
23               16.0  {TV,Internet,"Wireless Internet","Air conditio...
24                1.0  {"Cable TV","Wireless Internet",Kitchen,"Pets ...
25                3.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
26                2.0  {TV,"Wireless Internet","Air conditioning",Kit...
27                2.0  {"Cable TV",Internet,"Wireless Internet",Kitch...
28                2.0  {Internet,"Wireless Internet",Kitchen,"Free pa...
29                3.0  {Internet,"Wireless Internet",Kitchen,"Smoking...
...               ...                                                ...
59794             6.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
59795             5.0  {TV,Internet,"Wireless Internet","Air conditio...
59796             2.0  {Internet,"Wireless Internet","Air conditionin...
59797             1.0  {"Wireless Internet","Air conditioning",Kitche...
59798             4.0  {TV,"Wireless Internet","Air conditioning",Kit...
59799             4.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
59800            12.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
59801             6.0  {TV,"Wireless Internet","Air conditioning",Kit...
59802             3.0  {TV,"Air conditioning",Kitchen,"Smoking allowe...
59803             4.0  {TV,Internet,"Wireless Internet","Air conditio...
59804             2.0  {Internet,"Wireless Internet","Air conditionin...
59805             2.0  {TV,"Wireless Internet","Air conditioning",Kit...
59806             6.0  {TV,Internet,"Wireless Internet","Air conditio...
59807             2.0  {TV,Internet,"Wireless Internet","Pets live on...
59808             4.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
59809             2.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
59810             4.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
59811             4.0  {TV,"Wireless Internet","Air conditioning",Kit...
59812             5.0  {TV,"Wireless Internet","Air conditioning",Kit...
59813             1.0  {"Wireless Internet","Suitable for events",Ess...
59814             4.0  {TV,"Cable TV",Internet,"Wireless Internet",Ki...
59815             1.0  {TV,"Wireless Internet","Air conditioning",Kit...
59816             8.0  {TV,"Wireless Internet",Kitchen,"Free parking ...
59817             6.0  {TV,"Wireless Internet","Air conditioning",Kit...
59818             2.0  {TV,"Wireless Internet","Air conditioning",Kit...
59819             2.0  {Internet,"Wireless Internet","Air conditionin...
59820             5.0  {TV,"Cable TV","Wireless Internet","Air condit...
59821             6.0  {TV,Internet,"Wireless Internet","Air conditio...
59822             2.0  {TV,"Cable TV",Internet,"Wireless Internet","A...
```

```
59823           2.0  {TV,Internet,"Wireless Internet","Air conditio...

       availability_30  bathrooms  bed_type  bedrooms  beds  \
0                   24        1.0  Real Bed       1.0   1.0
1                   30        1.0  Real Bed       1.0   1.0
2                   30        3.0  Real Bed       3.0   3.0
3                    8        1.0  Real Bed       1.0   1.0
4                   17        1.0  Real Bed       1.0   1.0
5                   23        1.0  Real Bed       0.0   1.0
6                   15        1.0  Real Bed       1.0   2.0
7                    5        1.0  Real Bed       1.0   2.0
8                   17        1.0  Real Bed       1.0   1.0
9                   12        1.0  Real Bed       1.0   3.0
10                   7        1.5  Real Bed       1.0   1.0
11                   2        1.5  Real Bed       1.0   1.0
12                   0        1.0  Real Bed       2.0   2.0
13                   7        1.5  Real Bed       1.0   1.0
14                   1        1.0  Real Bed       1.0   1.0
15                   0        1.0  Real Bed       1.0   1.0
16                   7        1.0  Real Bed       1.0   1.0
17                   0        1.0  Real Bed       1.0   1.0
18                   6        1.0  Real Bed       1.0   2.0
19                   4        1.0  Real Bed       1.0   2.0
20                   8        1.0  Real Bed       1.0   1.0
21                  26        1.0  Real Bed       1.0   1.0
22                  28        1.0  Real Bed       1.0   2.0
23                  22        1.0  Real Bed       2.0   4.0
24                   0        1.0  Real Bed       1.0   1.0
25                   6        1.0  Real Bed       1.0   2.0
26                   0        1.0  Real Bed       1.0   1.0
27                  11        1.5  Real Bed       1.0   1.0
28                  28        1.5  Real Bed       1.0   2.0
29                  28        1.0  Real Bed       1.0   2.0
...                ...        ...       ...       ...   ...
59794                9        1.0  Real Bed       2.0   3.0
59795                7        1.0  Real Bed       2.0   2.0
59796               10        1.0  Real Bed       0.0   1.0
59797                0        1.0  Real Bed       1.0   1.0
59798                0        1.0  Real Bed       1.0   2.0
59799               15        1.0  Real Bed       1.0   2.0
59800                0        3.5  Real Bed       3.0   7.0
59801                0        1.0  Real Bed       2.0   2.0
59802                0        1.0  Real Bed       1.0   1.0
59803                0        1.0  Real Bed       1.0   1.0
59804                0        1.5  Real Bed       1.0   1.0
59805                0        1.0  Real Bed       1.0   1.0
59806                9        2.0  Real Bed       3.0   3.0
59807                7        1.0  Real Bed       1.0   1.0
```

```
59808                4          1.0  Real Bed        1.0   2.0
59809                0          1.0  Real Bed        0.0   1.0
59810               14          1.0  Real Bed        1.0   2.0
59811               17          1.5  Real Bed        1.0   2.0
59812                0          2.0  Real Bed        3.0   2.0
59813               29          1.0  Real Bed        1.0   1.0
59814                0          1.5  Real Bed        2.0   2.0
59815               27          1.0  Real Bed        0.0   1.0
59816                0          3.5  Real Bed        3.0   4.0
59817               24          2.0  Real Bed        3.0   3.0
59818               28          1.0  Real Bed        1.0   1.0
59819                0          1.0  Real Bed        1.0   1.0
59820               20          1.0  Real Bed        1.0   2.0
59821               13          1.0  Real Bed        1.0   3.0
59822               12          1.0  Real Bed        1.0   1.0
59823               24          1.0  Real Bed        1.0   1.0

      cancellation_policy              city  has_availability   ...      \
0                moderate    sunnysidebronx               NaN   ...
1                flexible    sunnysidebronx               NaN   ...
2                  strict    sunnysidebronx               NaN   ...
3                  strict  long island city               NaN   ...
4                moderate    sunnysidebronx               NaN   ...
5                moderate    sunnysidebronx               NaN   ...
6                flexible  long island city               NaN   ...
7                  strict    sunnysidebronx               NaN   ...
8                moderate    sunnysidebronx               NaN   ...
9                  strict    sunnysidebronx               NaN   ...
10                 strict    sunnysidebronx               NaN   ...
11                 strict    sunnysidebronx               NaN   ...
12                 strict    sunnysidebronx               NaN   ...
13               moderate    sunnysidebronx               NaN   ...
14               moderate    sunnysidebronx               NaN   ...
15                 strict    sunnysidebronx               NaN   ...
16               moderate    sunnysidebronx               NaN   ...
17               flexible    sunnysidebronx               NaN   ...
18               flexible    sunnysidebronx               NaN   ...
19                 strict    sunnysidebronx               NaN   ...
20               moderate    sunnysidebronx               NaN   ...
21               flexible    sunnysidebronx               NaN   ...
22                 strict    sunnysidebronx               NaN   ...
23                 strict    sunnysidebronx               NaN   ...
24                 strict    sunnysidebronx               NaN   ...
25                 strict    sunnysidebronx               NaN   ...
26                 strict    sunnysidebronx               NaN   ...
27               moderate    sunnysidebronx               NaN   ...
28                 strict    sunnysidebronx               NaN   ...
29                 strict    sunnysidebronx               NaN   ...
```

```
...                 ...               ...                ...   ...
59794            strict        washington                NaN   ...
59795            strict        washington                NaN   ...
59796          flexible        washington                NaN   ...
59797            strict        washington                NaN   ...
59798            strict        washington                NaN   ...
59799            strict        washington                NaN   ...
59800            strict        washington                NaN   ...
59801          flexible        washington                NaN   ...
59802          flexible        washington                NaN   ...
59803            strict        washington                NaN   ...
59804          flexible        washington                NaN   ...
59805          flexible        washington                NaN   ...
59806            strict        washington                NaN   ...
59807          flexible        washington                NaN   ...
59808          moderate        washington                NaN   ...
59809            strict        washington                NaN   ...
59810          moderate       takoma park                NaN   ...
59811            strict      temple hills                NaN   ...
59812            strict       takoma park                NaN   ...
59813          flexible        washington                NaN   ...
59814          flexible          bethesda                NaN   ...
59815          flexible     mount rainier                NaN   ...
59816            strict          bethesda                NaN   ...
59817          flexible       hyattsville                NaN   ...
59818          flexible        washington                NaN   ...
59819          flexible     silver spring                NaN   ...
59820          flexible          bethesda                NaN   ...
59821            strict      temple hills                NaN   ...
59822          moderate     silver spring                NaN   ...
59823          flexible   capitol heights                NaN   ...

       review_scores_checkin  review_scores_cleanliness  \
0                       10.0                       10.0
1                        NaN                        NaN
2                        NaN                        NaN
3                       10.0                       10.0
4                       10.0                       10.0
5                       10.0                       10.0
6                       10.0                       10.0
7                        9.0                        9.0
8                       10.0                       10.0
9                       10.0                        9.0
10                      10.0                        9.0
11                       9.0                       10.0
12                      10.0                       10.0
13                      10.0                        9.0
14                      10.0                       10.0
```

```
15                        10.0                        9.0
16                        10.0                        9.0
17                         NaN                        NaN
18                         NaN                        NaN
19                         9.0                        9.0
20                         9.0                        9.0
21                        10.0                        8.0
22                        10.0                        9.0
23                        10.0                        9.0
24                        10.0                       10.0
25                         9.0                        9.0
26                        10.0                       10.0
27                         9.0                        9.0
28                        10.0                        9.0
29                        10.0                        8.0
...                        ...                        ...
59794                     10.0                        9.0
59795                     10.0                       10.0
59796                     10.0                       10.0
59797                      NaN                        NaN
59798                      8.0                        8.0
59799                     10.0                       10.0
59800                      NaN                        NaN
59801                      NaN                        NaN
59802                      NaN                        NaN
59803                     10.0                       10.0
59804                      9.0                        9.0
59805                      9.0                       10.0
59806                     10.0                       10.0
59807                     10.0                       10.0
59808                     10.0                       10.0
59809                      9.0                       10.0
59810                      NaN                        NaN
59811                      NaN                        NaN
59812                      NaN                        NaN
59813                      NaN                        NaN
59814                      NaN                        NaN
59815                     10.0                       10.0
59816                      NaN                        NaN
59817                      NaN                        NaN
59818                      NaN                        NaN
59819                      9.0                        8.0
59820                      NaN                        NaN
59821                      NaN                        NaN
59822                     10.0                        9.0
59823                      NaN                        NaN

      review_scores_communication  review_scores_location  \
```

0	10.0	10.0
1	NaN	NaN
2	NaN	NaN
3	10.0	10.0
4	10.0	10.0
5	10.0	10.0
6	10.0	10.0
7	9.0	9.0
8	10.0	10.0
9	10.0	9.0
10	10.0	9.0
11	7.0	10.0
12	10.0	9.0
13	10.0	9.0
14	10.0	10.0
15	10.0	9.0
16	10.0	9.0
17	NaN	NaN
18	NaN	NaN
19	9.0	8.0
20	10.0	9.0
21	10.0	6.0
22	10.0	9.0
23	10.0	9.0
24	10.0	9.0
25	9.0	9.0
26	10.0	9.0
27	10.0	9.0
28	10.0	8.0
29	10.0	9.0
...	...	...
59794	9.0	10.0
59795	10.0	10.0
59796	10.0	9.0
59797	NaN	NaN
59798	8.0	6.0
59799	10.0	10.0
59800	NaN	NaN
59801	NaN	NaN
59802	NaN	NaN
59803	10.0	10.0
59804	10.0	9.0
59805	10.0	10.0
59806	10.0	10.0
59807	10.0	10.0
59808	10.0	10.0
59809	10.0	10.0
59810	NaN	NaN

```
59811                  NaN                      NaN
59812                  NaN                      NaN
59813                  NaN                      NaN
59814                  NaN                      NaN
59815                 10.0                     10.0
59816                  NaN                      NaN
59817                  NaN                      NaN
59818                  NaN                      NaN
59819                  9.0                      8.0
59820                  NaN                      NaN
59821                  NaN                      NaN
59822                 10.0                     10.0
59823                  NaN                      NaN

       review_scores_rating review_scores_value        room_type state  \
0                     100.0                10.0     Private room    NY
1                       NaN                 NaN     Private room    NY
2                       NaN                 NaN  Entire home/apt    NY
3                      93.0                10.0  Entire home/apt    NY
4                      97.0                10.0     Private room    NY
5                      97.0                10.0  Entire home/apt    NY
6                      98.0                10.0  Entire home/apt    NY
7                      90.0                 9.0     Private room    NY
8                     100.0                10.0  Entire home/apt    NY
9                      92.0                 9.0  Entire home/apt    NY
10                     93.0                 9.0     Private room    NY
11                     93.0                 9.0     Private room    NY
12                     85.0                10.0  Entire home/apt    NY
13                     96.0                10.0     Private room    NY
14                    100.0                10.0     Private room    NY
15                     96.0                10.0     Private room    NY
16                     97.0                10.0     Private room    NY
17                      NaN                 NaN     Private room    NY
18                      NaN                 NaN     Private room    NY
19                     89.0                 9.0  Entire home/apt    NY
20                     93.0                 9.0  Entire home/apt    NY
21                    100.0                10.0  Entire home/apt    NY
22                     89.0                10.0     Private room    NY
23                     95.0                 9.0  Entire home/apt    NY
24                     94.0                10.0     Private room    NY
25                     87.0                 9.0     Private room    NY
26                     96.0                 9.0     Private room    NY
27                     89.0                 9.0     Private room    NY
28                     89.0                 9.0     Private room    NY
29                     87.0                10.0     Private room    NY
...                     ...                 ...              ...   ...
59794                  91.0                10.0  Entire home/apt    DC
59795                  96.0                 9.0  Entire home/apt    DC
```

59796	96.0	10.0	Entire home/apt	DC
59797	NaN	NaN	Private room	DC
59798	80.0	6.0	Entire home/apt	DC
59799	98.0	10.0	Entire home/apt	DC
59800	NaN	NaN	Entire home/apt	DC
59801	NaN	NaN	Entire home/apt	DC
59802	NaN	NaN	Entire home/apt	DC
59803	99.0	10.0	Entire home/apt	DC
59804	92.0	9.0	Private room	DC
59805	100.0	10.0	Entire home/apt	DC
59806	98.0	10.0	Entire home/apt	DC
59807	95.0	10.0	Entire home/apt	DC
59808	94.0	10.0	Private room	DC
59809	100.0	10.0	Entire home/apt	DC
59810	NaN	NaN	Entire home/apt	MD
59811	NaN	NaN	Private room	MD
59812	NaN	NaN	Entire home/apt	MD
59813	NaN	NaN	Shared room	DC
59814	NaN	NaN	Entire home/apt	MD
59815	100.0	10.0	Private room	MD
59816	NaN	NaN	Entire home/apt	MD
59817	NaN	NaN	Entire home/apt	MD
59818	NaN	NaN	Private room	DC
59819	80.0	9.0	Private room	MD
59820	NaN	NaN	Entire home/apt	MD
59821	NaN	NaN	Private room	MD
59822	100.0	10.0	Entire home/apt	MD
59823	NaN	NaN	Private room	MD

	weekly_price	zipcode
0	NaN	10464
1	NaN	10464
2	NaN	10464
3	$775.00	10464
4	$350.00	10464
5	$550.00	10464
6	NaN	10464
7	NaN	10467
8	NaN	10469
9	NaN	10469
10	$350.00	10469
11	$300.00	10469
12	NaN	10462
13	$305.00	10469
14	NaN	10467
15	$250.00	10469
16	NaN	10469
17	NaN	10469

```
18             NaN    10467
19             NaN    10467
20             NaN    10469
21             NaN    10467
22             NaN    10469
23             NaN    10469
24             NaN    10469
25             NaN    10467
26             NaN    10469
27         $370.00    10469
28         $280.00    10469
29             NaN    10467
...            ...      ...
59794          NaN    20003
59795          NaN    20003
59796      $800.00    20003
59797          NaN    20003
59798          NaN    20002
59799          NaN    20002
59800          NaN    20003
59801          NaN    20003
59802          NaN    20003
59803          NaN    20003
59804          NaN    20003
59805          NaN    20003
59806          NaN    20003
59807          NaN    20003
59808          NaN    20003
59809    $1,250.00    20002
59810          NaN    20912
59811          NaN    20748
59812          NaN    20912
59813          NaN    20006
59814          NaN    20816
59815          NaN    20712
59816          NaN    20816
59817          NaN    20782
59818          NaN    20020
59819          NaN    20910
59820          NaN    20816
59821          NaN    20748
59822          NaN    20910
59823          NaN    20743

[59824 rows x 29 columns]
```

There is a lot of interesting information here, but it requires extensive cleaning before it's useable. For instance, we would like to convert features such as the amenities list to a categorical format. In other words, we would like to create columns labelled has_cable, has_wifi, etc.

Next, we include columns for the average unemployment rate from 2011-2016 in each location's state, since we have established that location is important but the year is not. We would like to include GDP and personal income for each state as well, but unfortunately this information appears to be either corrupted or inaccurate in our dataset.

We also remove the following columns, which provide information unsuitable for a neural net regressor: ['amenities', 'bed_type', 'has_availability', 'host_id', 'id', 'latitude', 'longitude', 'name', 'state', 'zipcode']

Finally, we normalize each feature to lie between 0 and 1.

```
In [92]: import data_processing
         listings = data_processing.clean_listings()
```

```
In [72]: X = listings.drop('availability_30', 1)
         y = listings['availability_30']
```

```
In [84]: X.values[:5]
```

```
Out[84]: array([[ 0.06666667,  0.125     ,  0.1       , ...,  1.        ,
                  1.        ,  1.        ],
                [ 0.13333333,  0.125     ,  0.1       , ...,  1.        ,
                  1.        ,  1.        ],
                [ 0.2       ,  0.125     ,  0.1       , ...,  1.        ,
                  1.        ,  1.        ],
                [ 0.06666667,  0.125     ,  0.        , ...,  1.        ,
                  1.        ,  1.        ],
                [ 0.2       ,  0.125     ,  0.1       , ...,  1.        ,
                  1.        ,  1.        ]])
```

```
In [85]: y.values[:5]
```

```
Out[85]: array([24, 30, 30,  8, 17])
```

```
In [104]: # shuffle and split into test and train
          from sklearn.utils import shuffle
          listings = shuffle(listings)
          listings = listings.dropna()

          testRatio = 0.2
          nSamples = len(X)
          nTest = int(nSamples * testRatio)

          X_test = X[:nTest].values
          y_test = y[:nTest].values
          X_train = X[nTest:].values
          y_train = y[nTest:].values
```

4 Predicting booking rate with a neural net

```
In [64]: from keras.models import Sequential
         from keras.layers import Dense
         from keras.layers import Dropout
```

```
In [65]: # set random seed for reproducibility
         random_seed = 13
         np.random.seed(random_seed)
```

```
In [87]: model = Sequential()
         model.add(Dense(20, input_dim=X_train.shape[1], init='normal', activation=
         model.add(Dropout(0.5))
         model.add(Dense(20, init='normal'))
         model.add(Dropout(0.5))
         model.add(Dense(1, init='normal'))
         model.compile(loss='mean_squared_error', optimizer='adam')
```

```
/Users/mattzhang/py2_kernel/lib/python2.7/site-packages/ipykernel/__main__.py:2: Us
  from ipykernel import kernelapp as app
/Users/mattzhang/py2_kernel/lib/python2.7/site-packages/ipykernel/__main__.py:4: Us
/Users/mattzhang/py2_kernel/lib/python2.7/site-packages/ipykernel/__main__.py:6: Us
```

```
In [105]: model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=20,
```

```
Train on 36088 samples, validate on 9021 samples
Epoch 1/20
36088/36088 [==============================] - 2s - loss: 95.5321 - val_loss: 93.43
Epoch 2/20
36088/36088 [==============================] - 1s - loss: 84.4840 - val_loss: 88.97
Epoch 3/20
36088/36088 [==============================] - 1s - loss: 82.2417 - val_loss: 87.24
Epoch 4/20
36088/36088 [==============================] - 1s - loss: 80.8809 - val_loss: 86.17
Epoch 5/20
36088/36088 [==============================] - 1s - loss: 79.4377 - val_loss: 86.20
Epoch 6/20
36088/36088 [==============================] - 1s - loss: 78.8778 - val_loss: 86.04
Epoch 7/20
36088/36088 [==============================] - 1s - loss: 78.3192 - val_loss: 85.53
Epoch 8/20
36088/36088 [==============================] - 1s - loss: 77.5360 - val_loss: 85.68
Epoch 9/20
36088/36088 [==============================] - 1s - loss: 77.6245 - val_loss: 85.51
Epoch 10/20
36088/36088 [==============================] - 1s - loss: 77.4793 - val_loss: 85.79
Epoch 11/20
36088/36088 [==============================] - 1s - loss: 76.8658 - val_loss: 85.29
```

```
Epoch 12/20
36088/36088 [==============================] - 1s - loss: 76.9841 - val_loss: 85.21
Epoch 13/20
36088/36088 [==============================] - 1s - loss: 76.6333 - val_loss: 86.43
Epoch 14/20
36088/36088 [==============================] - 1s - loss: 76.3039 - val_loss: 84.86
Epoch 15/20
36088/36088 [==============================] - 1s - loss: 76.2772 - val_loss: 85.80
Epoch 16/20
36088/36088 [==============================] - 1s - loss: 76.2586 - val_loss: 86.16
Epoch 17/20
36088/36088 [==============================] - 1s - loss: 75.7701 - val_loss: 84.72
Epoch 18/20
36088/36088 [==============================] - 1s - loss: 75.5828 - val_loss: 84.78
Epoch 19/20
36088/36088 [==============================] - 1s - loss: 75.5200 - val_loss: 84.98
Epoch 20/20
36088/36088 [==============================] - 1s - loss: 75.2841 - val_loss: 84.87
```

```
Out[105]: <keras.callbacks.History at 0x12425c950>
```

---

5 Conclusion

Using a regression model based on a two-layer neural net, we are able to predict booking rate for a 30-day period with a mean-squared-error of around 9 days.

There are several methods of improving our prediction results. First, we have shown that the difference between cities is large, but we have not included a good method of capturing this information. We have a one-hot encoding for each city, but this is an inefficient method of using the information. It would be more sensible to capture relevant info such as a city's population and GDP instead. We have included unemployment information at the state level, but this is a poor proxy.

Second, we showed that rental month is a very strong indicator of availability. For example, for Denver we see a three-fold difference in average availability based on the month. However, the listings dataset we accessed did not include the months during which each data point was gathered. We predict that with this additional information, our regression model could be much more effective.

```
In [ ]:
```