Rutgers Business School

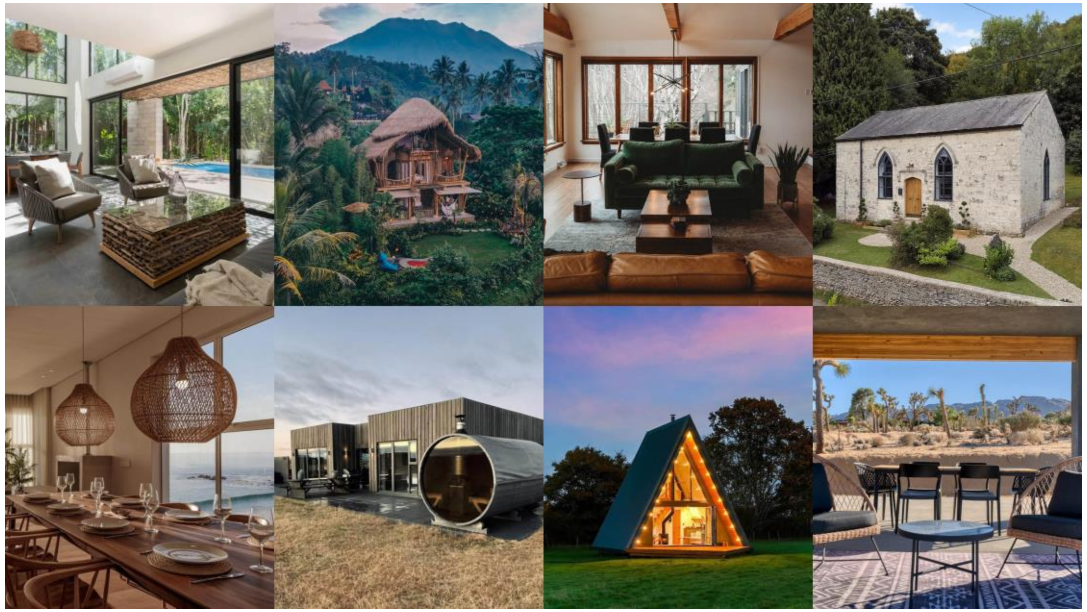
Airbnb Data Analysis: Lodge Prices

Tonghua Lin

Rutgers Business School

CATALOG

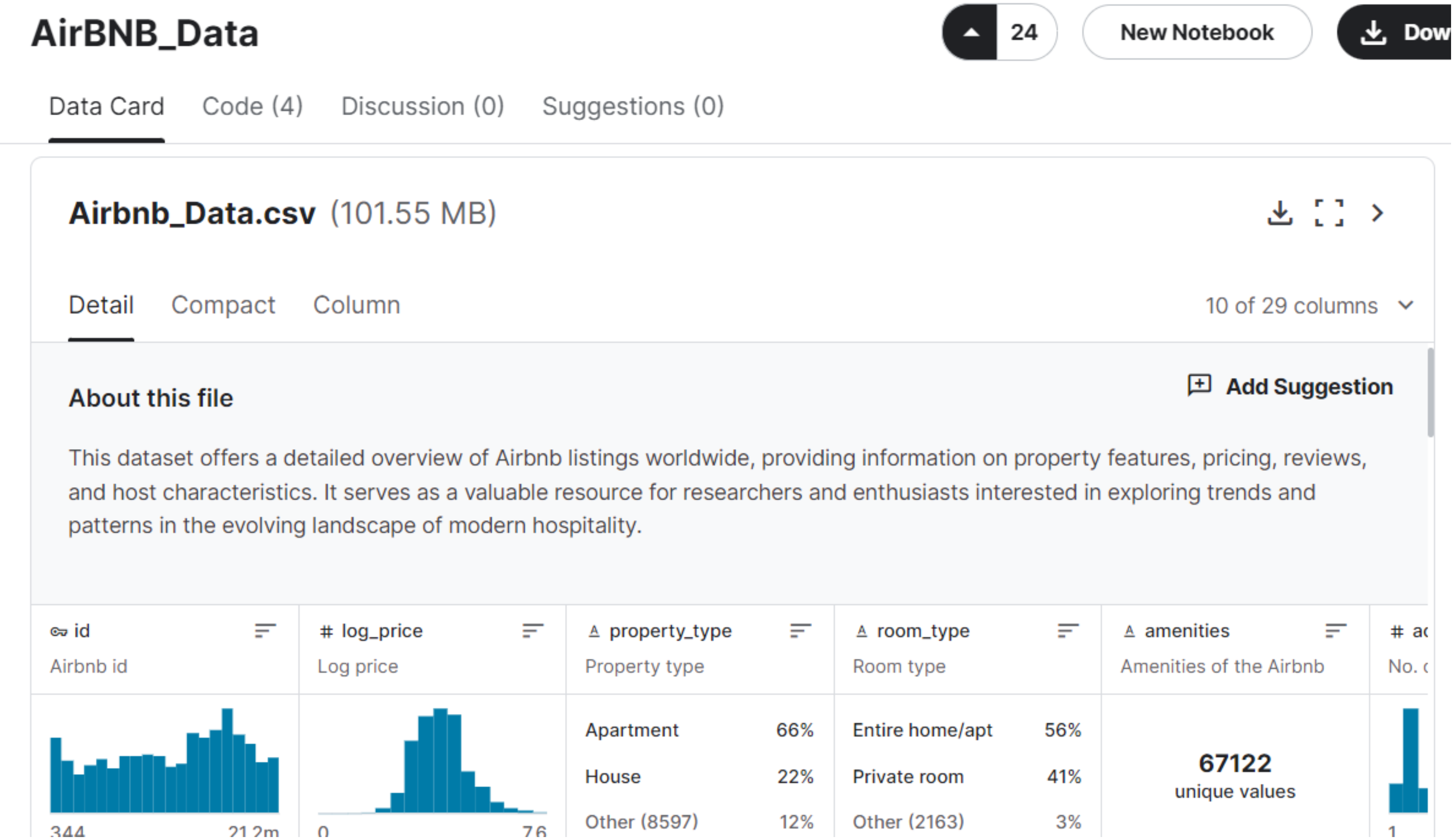
Introduction



Besides the chain hotels, Airbnb is another way to find a lodge while traveling. This project is trying to find the factors that can influence the price of a lodge.

Introduction

kaggle



29 columns and 73000 rows

- ID
- Log_price
- Property_type
- Room_type
- Bathroom
- City
- ….

Introduction

Project Target:

- Predict the price by different factors
- Find a good city with low lodging prices

Data Cleaning

Exclude text data and location name related data:

id amenities description name neighbourhood thumbnail_url

id
6901257
6304928
7919400
13418779
3808709
12422935
11825529
13971273
180792

amenities
{"Wireless Internet","Air conditioning",Kitchen,
{"Wireless Internet","Air conditioning",Kitchen,
{TV,"Cable TV","Wireless Internet","Air conditic
{TV,"Cable TV",Internet,"Wireless Internet",Kitc
{TV,Internet,"Wireless Internet","Air conditionir
{TV,"Wireless Internet",Heating,"Smoke detect
{TV,Internet,"Wireless Internet","Air conditionir
{TV,"Cable TV","Wireless Internet","Wheelchair
{TV,"Cable TV","Wireless Internet","Pets live on

description
Beautiful, sunlit brownstone 1-bedroom in t
Enjoy travelling during your stay in Manhatt
The Oasis comes complete with a full backy
This light-filled home-away-from-home is s
Cool, cozy, and comfortable studio located i
Beautiful private room overlooking scenic vi
Warm and cozy studio with full kitchen and
Arguably the best location (and safest) in dc
Garden Studio with private entrance from tl

name	neighbourhood
Beautiful brownstone 1-	Brooklyn Heights
Superb 3BR Apt Located	Hell's Kitchen
The Garden Oasis	Harlem
Beautiful Flat in the Hea	Lower Haight
Great studio in midtowr	Columbia Heights
Comfort Suite San Franc	Noe Valley
Beach Town Studio and	Parking!!!11h
Near LA Live, Staple's. St	Downtown
Cozy Garden Studio - P	Richmond District

thumbnail_url	zipcode
https://a0.muscache.com/im/pic	11201
https://a0.muscache.com/im/pic	10019
https://a0.muscache.com/im/pic	10027
https://a0.muscache.com/im/pic	94117
	20009
https://a0.muscache.com/im/pic	94131
https://a0.muscache.com/im/pic	90292
https://a0.muscache.com/im/pic	90015
https://a0.muscache.com/im/pic	94121

Data Cleaning

Deal with the NA missing data

```
           log_price        property_type              room_type
                   0                    0                      0
        accommodates            bathrooms               bed_type
                   0                  200                      0
 cancellation_policy         cleaning_fee                   city
                   0                    0                      0
        first_review host_has_profile_pic host_identity_verified
               15864                  188                    188
  host_response_rate           host_since      instant_bookable
               18299                  188                      0
         last_review             latitude              longitude
               15827                    0                      0
   number_of_reviews review_scores_rating               bedrooms
                   0                16722                     91
                beds
                 131
```

- For host_identity_verified related: Delete
- For bathrooms: 0
- For review related: 2017/11/01
- For beds and bedrooms and scores: average

Data Cleaning

Deal with the NA missing data

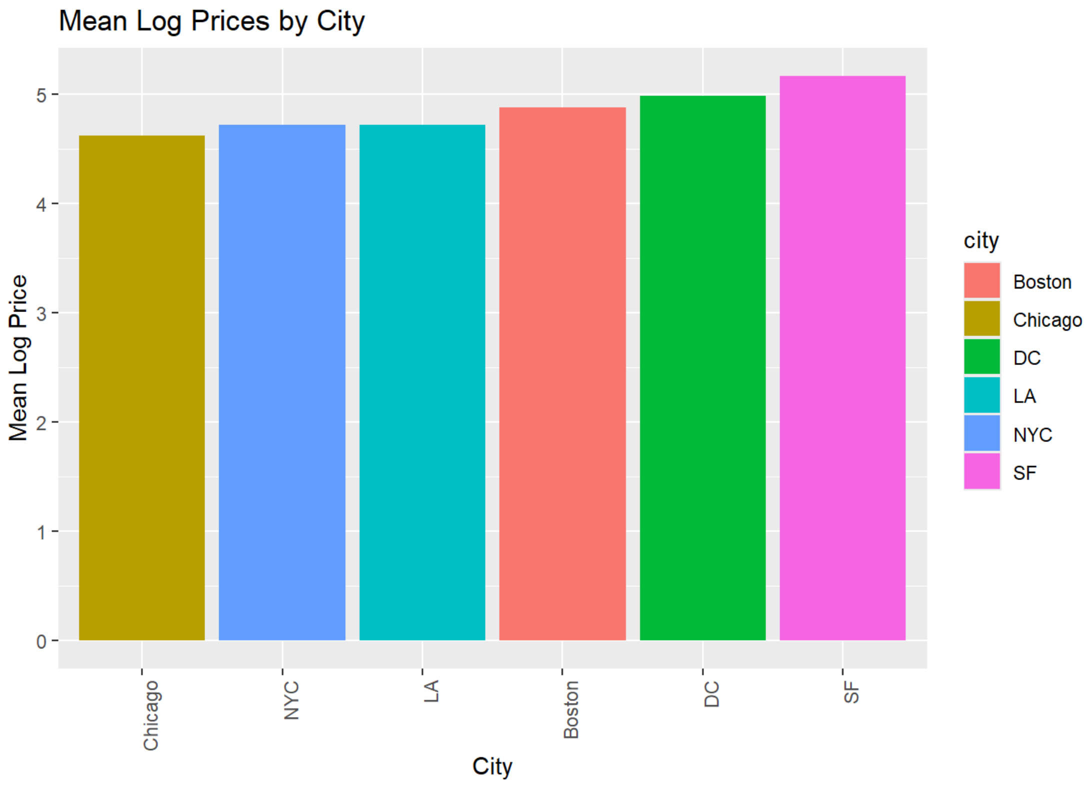
```
##        log_price       property_type               room_type
##                0                   0                       0
##     accommodates           bathrooms                bed_type
##                0                   0                       0
## cancellation_policy      cleaning_fee                    city
##                0                   0                       0
##     first_review host_has_profile_pic host_identity_verified
##                0                   0                       0
## host_response_rate          host_since        instant_bookable
##                0                   0                       0
##      last_review            latitude               longitude
##                0                   0                       0
## number_of_reviews review_scores_rating               bedrooms
##                0                   0                       0
##             beds
##                0
```

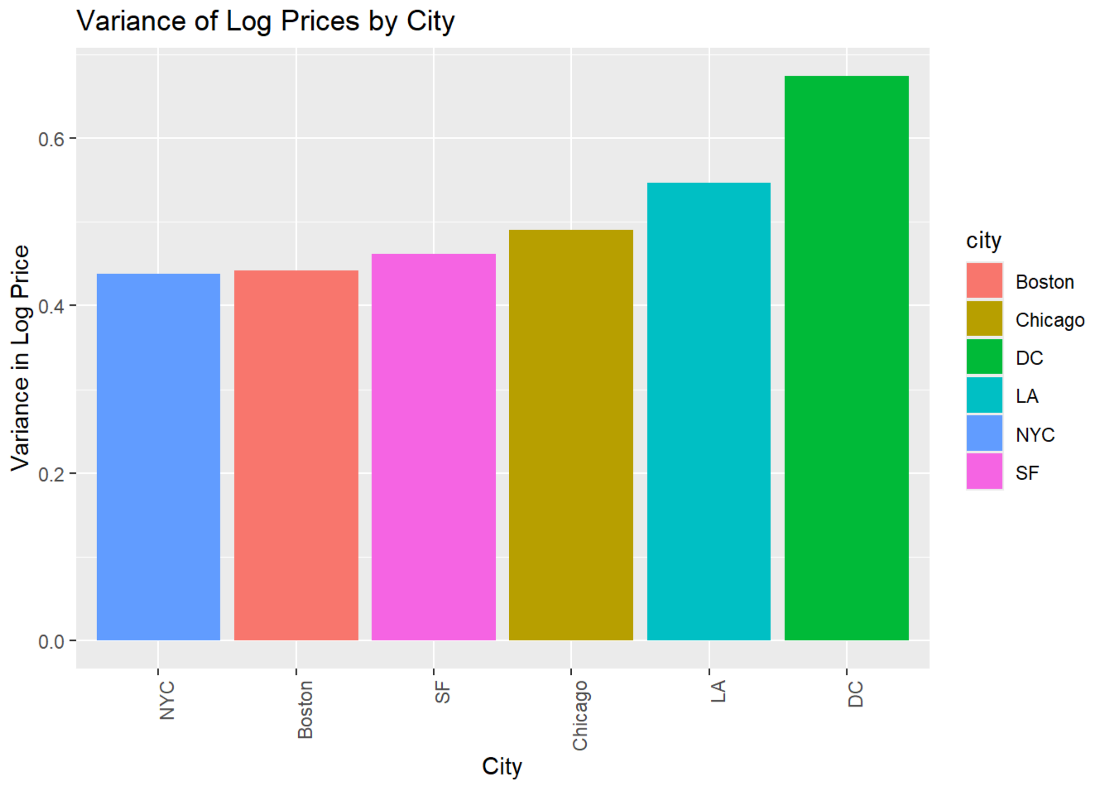
Data Cleaning

Transform the date data to number: using 2017/11/01 as baseline

```
## # A tibble: 6 × 22
##   log_price property_type room_type       accommodates bathrooms bed_type
##       <dbl> <chr>         <chr>                  <dbl>     <dbl> <chr>
## 1      5.01 Apartment     Entire home/apt            3         1 Real Bed
## 2      5.13 Apartment     Entire home/apt            7         1 Real Bed
## 3      4.98 Apartment     Entire home/apt            5         1 Real Bed
## 4      6.62 House         Entire home/apt            4         1 Real Bed
## 5      4.74 Apartment     Entire home/apt            2         1 Real Bed
## 6      4.44 Apartment     Private room               2         1 Real Bed
## # ℹ 16 more variables: cancellation_policy <chr>, cleaning_fee <lgl>,
## #   city <chr>, first_review <dbl>, host_has_profile_pic <lgl>,
## #   host_identity_verified <lgl>, host_response_rate <dbl>, host_since <dbl>,
## #   instant_bookable <lgl>, last_review <dbl>, latitude <dbl>, longitude <dbl>,
## #   number_of_reviews <dbl>, review_scores_rating <dbl>, bedrooms <dbl>,
## #   beds <dbl>
```

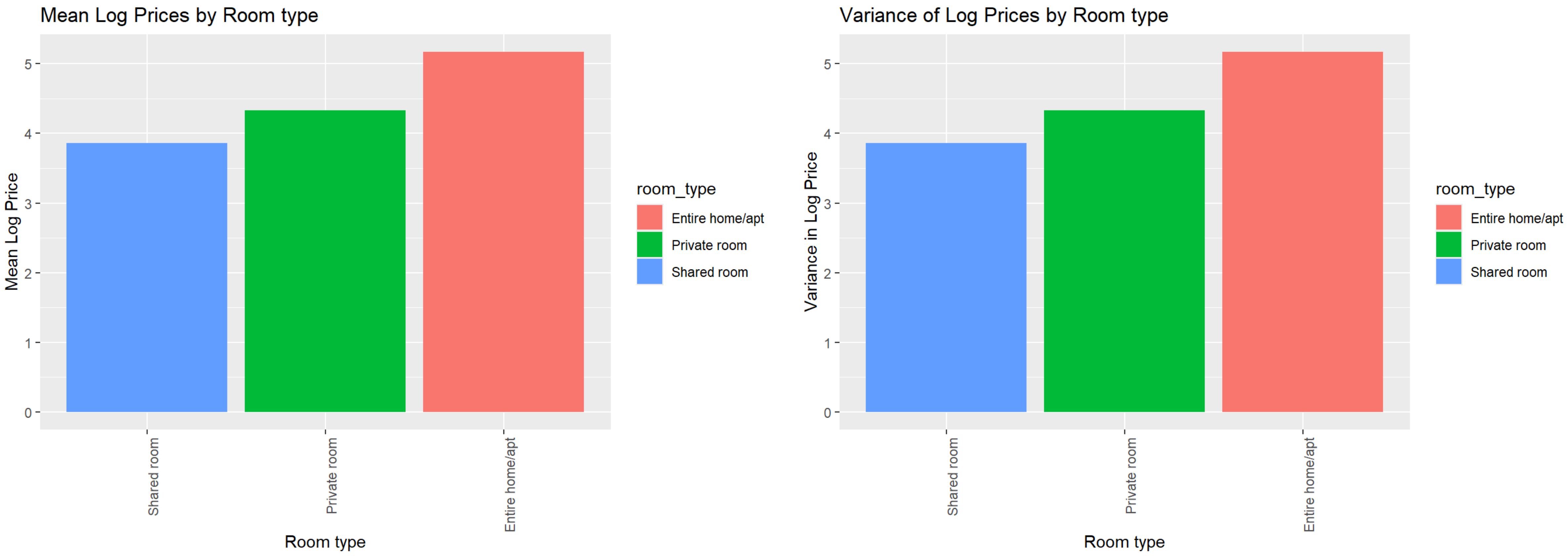
Explore





Means and variances across cities

Explore



Means and variances across room types

Explore

```
##  Two Sample t-test
##
## data:  DataSet$log_price[DataSet$city == "NYC"] and DataSet$log_price[DataSet$city == "LA"]
## t = -0.21408, df = 54615, p-value = 0.4152
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf 0.008644347
## sample estimates:
## mean of x mean of y
##  4.719238  4.720531
```

```
##
##  Two Sample t-test
##
## data:  DataSet$log_price[DataSet$city == "NYC"] and DataSet$log_price[DataSet$city == "Boston"]
## t = -13.927, df = 35639, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1453337
## sample estimates:
## mean of x mean of y
##  4.719238  4.884035
```

T-test results

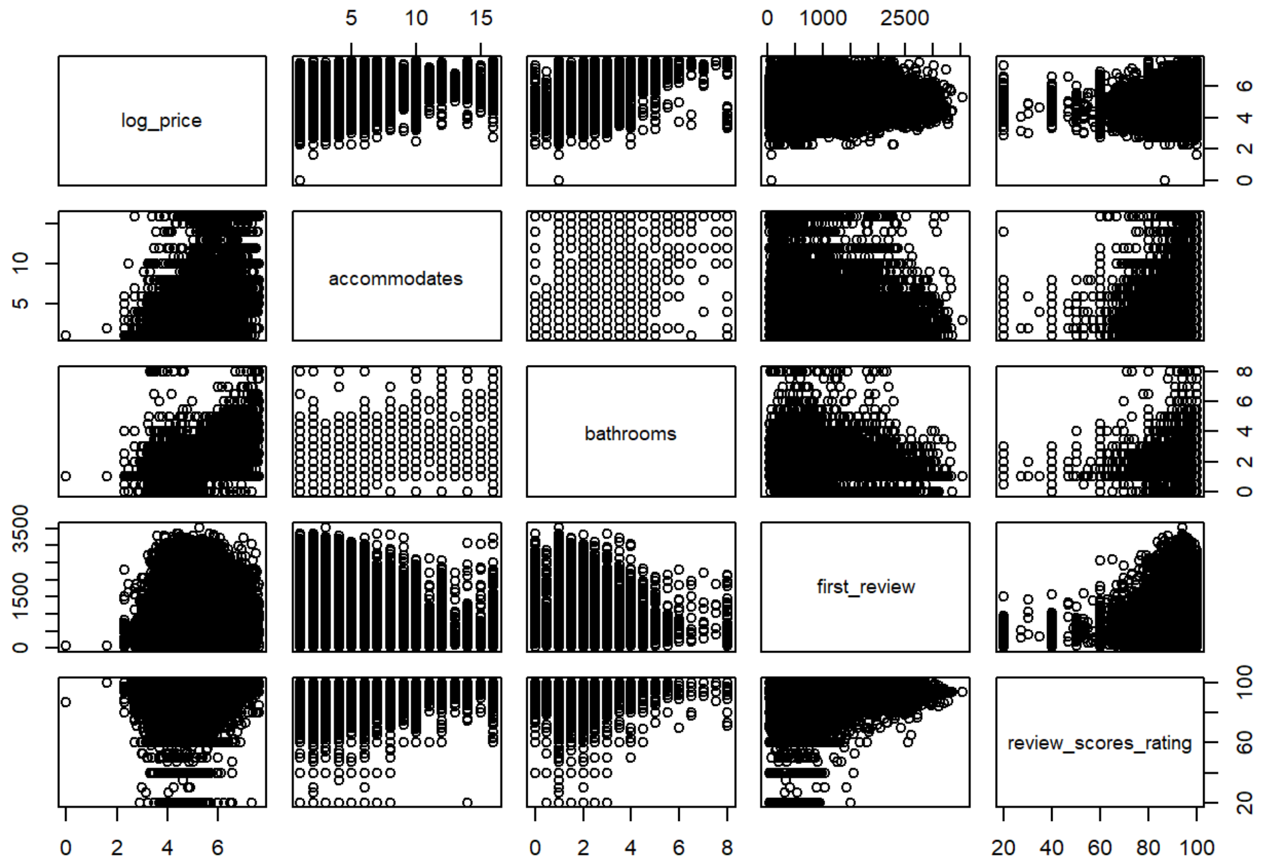
Explore

```
##  Two Sample t-test
##
## data:  DataSet$log_price[DataSet$room_type == "Private room"] and DataSet$log_price[DataSet$room_type == "Shared room"]
## t = 41.153, df = 32698, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.4854585
## sample estimates:
## mean of x mean of y
##  4.327647  3.860847
```

```
##  Two Sample t-test
##
## data:  DataSet$log_price[DataSet$room_type == "Private room"] and DataSet$log_price[DataSet$room_type == "Entire home/apt"]
## t = -196.9, df = 71760, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -0.8325653
## sample estimates:
## mean of x mean of y
##  4.327647  5.167226
```

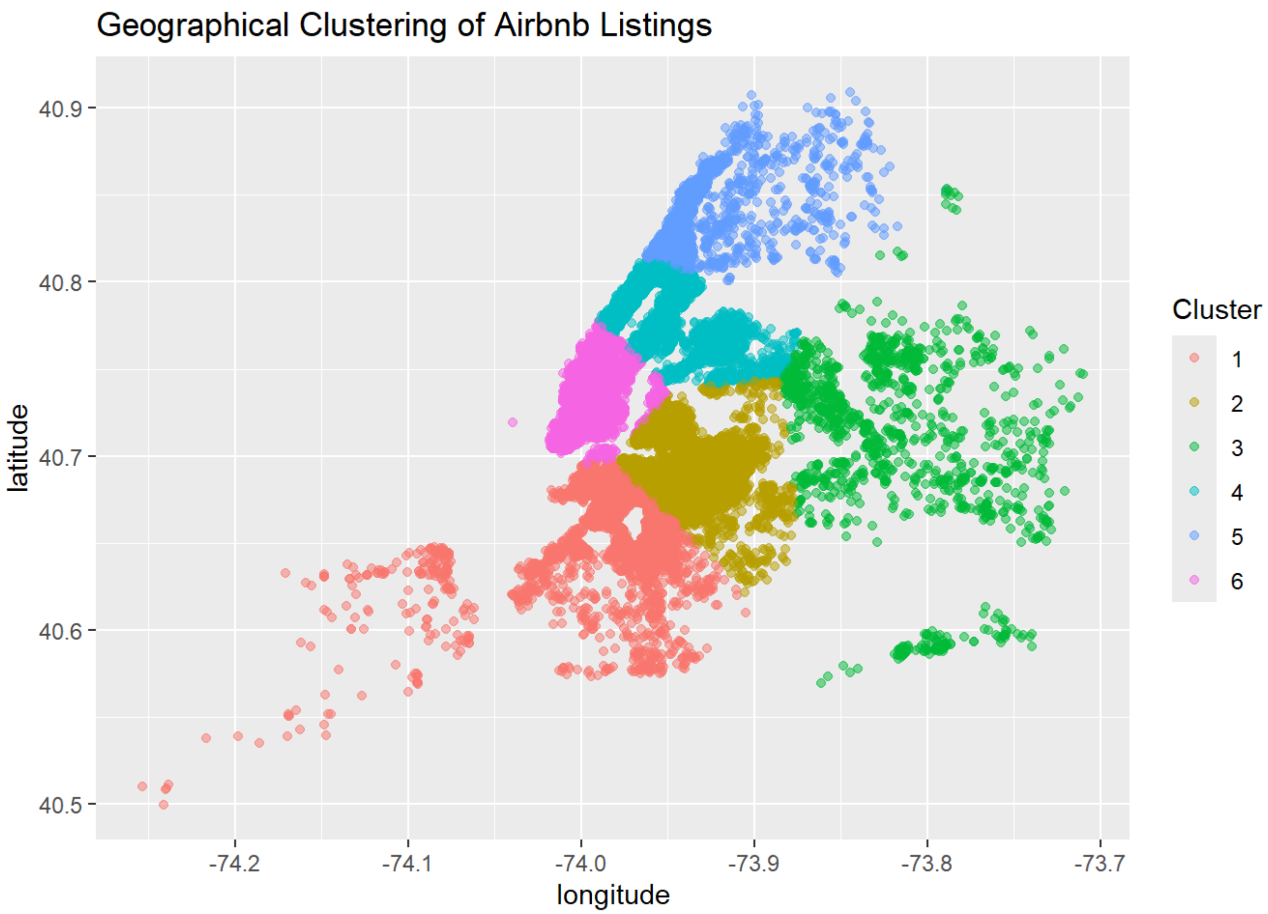
T-test results

Explore



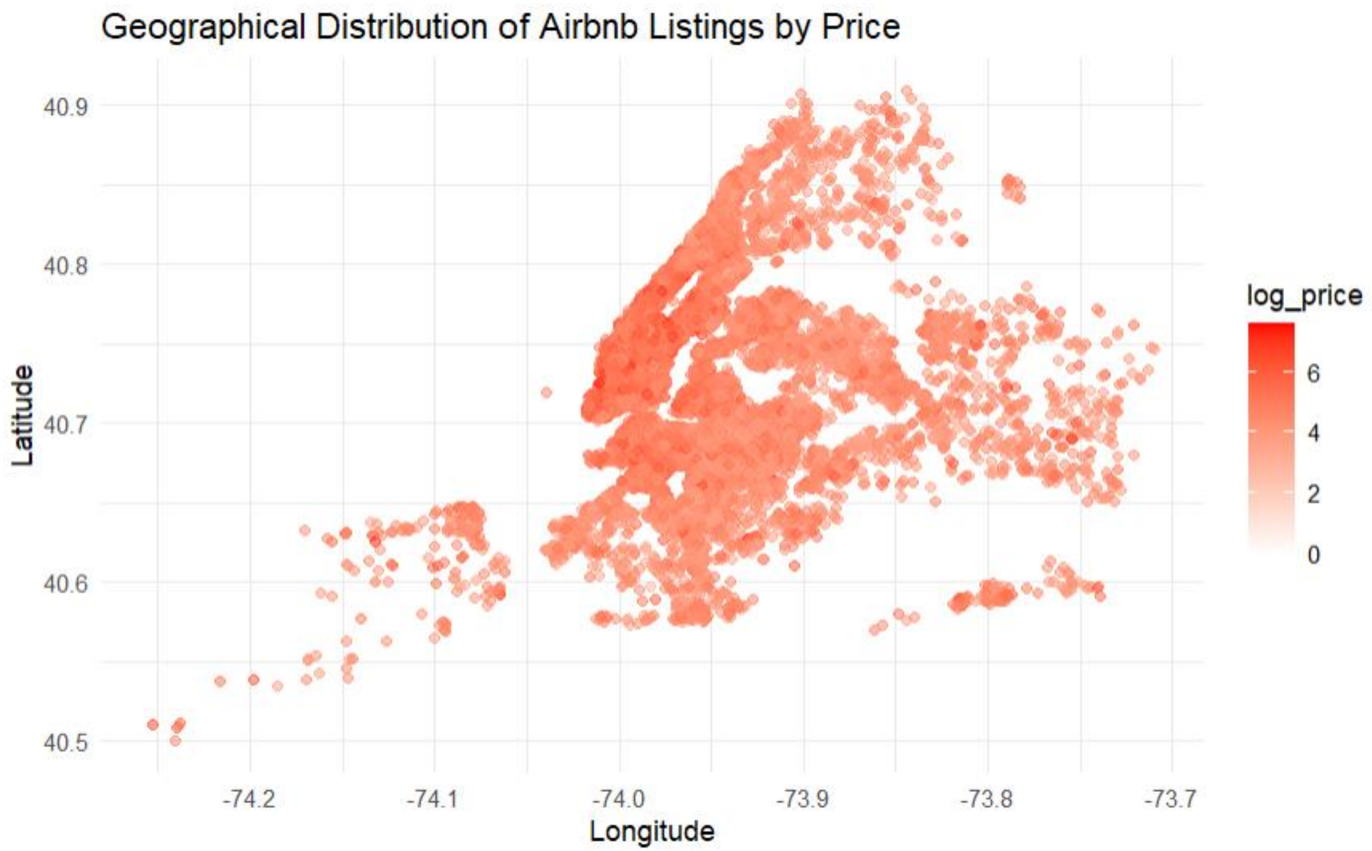
Multi-variables visualization: significant relationship pattern

Explore



K-means cluster in NYC

Explore



Prices across latitude and longitude

Explore

```
## $ log_price                : num [1:73923] 5.01 5.13 4.98 6.62 4.74 ...
## $ property_type            : Factor w/ 35 levels "Apartment","Bed & Breakfast",..: 1 1 1 18 1 1 1 12 18 18 ...
## $ room_type                : Factor w/ 3 levels "Entire home/apt",..: 1 1 1 1 1 2 1 1 2 2 ...
## $ accommodates             : num [1:73923] 3 7 5 4 2 2 3 2 2 2 ...
## $ bathrooms                : num [1:73923] 1 1 1 1 1 1 1 1 1 1 ...
## $ bed_type                 : Factor w/ 5 levels "Airbed","Couch",..: 5 5 5 5 5 5 5 5 5 5 ...
## $ cancellation_policy      : Factor w/ 5 levels "flexible","moderate",..: 3 3 2 1 2 3 2 2 2 2 ...
## $ cleaning_fee             : Factor w/ 2 levels "FALSE","TRUE": 2 2 2 2 2 2 2 2 2 2 ...
## $ city                     : Factor w/ 6 levels "Boston","Chicago",..: 5 5 5 6 3 6 4 4 6 4 ...
## $ first_review             : num [1:73923] 501 88 185 927 904 66 236 320 627 212 ...
## $ host_has_profile_pic     : Factor w/ 2 levels "FALSE","TRUE": 2 2 2 2 2 2 2 2 2 2 ...
## $ host_identity_verified: Factor w/ 2 levels "FALSE","TRUE": 2 1 2 2 2 2 1 2 1 1 ...
## $ host_response_rate       : num [1:73923] 94.4 100 100 94.4 100 ...
## $ host_since               : num [1:73923] 2046 135 372 927 976 ...
## $ instant_bookable         : Factor w/ 2 levels "FALSE","TRUE": 1 2 2 1 2 2 2 1 1 2 ...
```

Simple analysis about category variables

RUTGERS Business School

Explore

```
##                    property_type
## room_type          Apartment Bed & Breakfast  Boat Boutique hotel Bungalow Cabin
##   Entire home/apt      28820              57    56              22      313    58
##   Private room         18648             361     9              47       51    12
##   Shared room           1372              43     0               0        2     1
##                    property_type
## room_type          Camper/RV Casa particular Castle  Cave Chalet Condominium
##   Entire home/apt         73               0      7     1      3        1631
##   Private room            17               1      6     1      3         984
##   Shared room              4               0      0     0      0          39
##                    property_type
## room_type            Dorm Earth House Guest suite Guesthouse Hostel House   Hut
##   Entire home/apt       3           3          71        412      2  7513     5
##   Private room         66           1          52         68     23  8540     2
##   Shared room          73           0           0         18     45   450     1
##                    property_type
## room_type          In-law Island Lighthouse  Loft Other Parking Space
##   Entire home/apt      62      0          1   784   353             0
##   Private room          8      1          0   408   223             0
##   Shared room           1      0          0    49    31             1
##                    property_type
## room_type          Serviced apartment  Tent Timeshare  Tipi Townhouse Train
##   Entire home/apt                  16     6        59     2       787     1
##   Private room                      5    11        17     1       872     1
##   Shared room                       0     1         0     0        26     0
##                    property_type
## room_type          Treehouse Vacation home Villa  Yurt
##   Entire home/apt          4             9    83     6
##   Private room             3             2    92     3
##   Shared room              0             0     4     0
```

```
##                    broad_category
## room_type           Apartment House Other Unique Stays
##   Entire home/apt       31251  8761   609          602
##   Private room          20045  9566   632          296
##   Shared room            1460   483    80          138
```

- Apartment
- House
- Unique Stays
- Other

Explore

```
##                    bed_type
## room_type          Airbed Couch Futon Pull-out Sofa Real Bed
##   Entire home/apt     100    53   168           211    40691
##   Private room        275    61   485           256    29462
##   Shared room         102   153    97           115     1694
```

```
##                    host_has_profile_pic
## room_type           FALSE  TRUE
##   Entire home/apt     127 41096
##   Private room         90 30449
##   Shared room           9  2152
```

```
##                    cancellation_policy
## room_type          flexible moderate strict super_strict_30 super_strict_60
##   Entire home/apt      9762    10544  20788             112              17
##   Private room        11767     8079  10693               0               0
##   Shared room           955      391    815               0               0
```

```
##                    city
## room_type          Boston Chicago    DC    LA   NYC    SF
##   Entire home/apt    2146    2217  3882 12998 16163  3817
##   Private room       1276    1401  1668  8472 15205  2517
##   Shared room          46     101   137   974   805    98
```

Regression Model

```
Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                29.248008   0.600812   48.681  < 2e-16 ***
accommodates                0.273277   0.004733   57.739  < 2e-16 ***
bathrooms                   0.109863   0.003081   35.664  < 2e-16 ***
first_review                0.050307   0.004387   11.467  < 2e-16 ***
host_response_rate         -0.022973   0.002441   -9.412  < 2e-16 ***
host_since                 -0.012076   0.003161   -3.820 0.000134 ***
last_review                 0.055619   0.003949   14.085  < 2e-16 ***
latitude                    0.439013   0.107776    4.073 4.64e-05 ***
longitude                 -29.682742   0.583046  -50.910  < 2e-16 ***
number_of_reviews          -0.037284   0.003188  -11.697  < 2e-16 ***
review_scores_rating        0.062223   0.002419   25.726  < 2e-16 ***
bedrooms                    0.164799   0.003909   42.154  < 2e-16 ***
beds                       -0.083314   0.004403  -18.921  < 2e-16 ***
property_typeHouse         -0.026082   0.006361   -4.101 4.12e-05 ***
property_typeOther          0.164041   0.018252    8.988  < 2e-16 ***
property_typeUnique Stays  -0.147914   0.020725   -7.137 9.62e-13 ***
room_typePrivate room      -0.832712   0.005918 -140.701  < 2e-16 ***
room_typeShared room       -1.444050   0.015385  -93.862  < 2e-16 ***
bed_typeCouch               0.202804   0.049919    4.063 4.86e-05 ***
bed_typeFuton              -0.018888   0.038075   -0.496 0.619852    
bed_typePull-out Sofa       0.083466   0.040135    2.080 0.037561 *  
bed_typeReal Bed            0.085573   0.030014    2.851 0.004359 ** 
cancellation_policymoderate -0.032916  0.006855   -4.802 1.57e-06 ***
cancellation_policystrict   0.017643   0.006361    2.774 0.005545 ** 
cleaning_feeTRUE           -0.054722   0.005999   -9.122  < 2e-16 ***
cityChicago               -23.045924   0.446837  -51.576  < 2e-16 ***
cityDC                     -7.617708   0.206776  -36.840  < 2e-16 ***
cityLA                    -63.654096   1.321949  -48.152  < 2e-16 ***
cityNYC                    -3.756324   0.099393  -37.793  < 2e-16 ***
citySF                    -69.149846   1.399629  -49.406  < 2e-16 ***
host_has_profile_picTRUE   -0.141855   0.043393   -3.269 0.001079 ** 
host_identity_verifiedTRUE -0.042862   0.005556   -7.714 1.23e-14 ***
instant_bookableTRUE       -0.022839   0.005690   -4.014 5.98e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Exclude some insignificant variables

```
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)    
## (Intercept)                 29.510709   0.598271   49.327  < 2e-16 ***
## accommodates                 0.274852   0.004729   58.126  < 2e-16 ***
## bathrooms                    0.108939   0.003063   35.564  < 2e-16 ***
## first_review                 0.043438   0.003877   11.205  < 2e-16 ***
## host_response_rate          -0.023557   0.002438   -9.661  < 2e-16 ***
## last_review                  0.057329   0.003856   14.866  < 2e-16 ***
## latitude                     0.469393   0.107767    4.356 1.33e-05 ***
## longitude                  -29.945648   0.580588  -51.578  < 2e-16 ***
## number_of_reviews           -0.037664   0.003182  -11.837  < 2e-16 ***
## review_scores_rating         0.060655   0.002415   25.120  < 2e-16 ***
## bedrooms                     0.162782   0.003885   41.902  < 2e-16 ***
## beds                        -0.082977   0.004402  -18.850  < 2e-16 ***
## property_typeOther           0.172636   0.018159    9.507  < 2e-16 ***
## property_typeUnique Stays   -0.142209   0.020631   -6.893 5.50e-12 ***
## room_typePrivate room       -0.838806   0.005773 -145.301  < 2e-16 ***
## room_typeShared room        -1.432679   0.015241  -94.001  < 2e-16 ***
## bed_typeReal Bed             0.047190   0.014833    3.181 0.001466 ** 
## cleaning_feeTRUE            -0.055430   0.005773   -9.601  < 2e-16 ***
## cityChicago                -23.245231   0.444999  -52.237  < 2e-16 ***
## cityDC                      -7.664944   0.206408  -37.135  < 2e-16 ***
## cityLA                     -64.152643   1.317369  -48.698  < 2e-16 ***
## cityNYC                     -3.774646   0.099272  -38.023  < 2e-16 ***
## citySF                     -69.734870   1.394238  -50.016  < 2e-16 ***
## host_has_profile_picTRUE    -0.143900   0.043421   -3.314 0.000920 ***
## host_identity_verifiedTRUE  -0.048705   0.005340   -9.121  < 2e-16 ***
## instant_bookableTRUE        -0.019855   0.005678   -3.497 0.000472 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression Model

```
##                  (Intercept)                 accommodates
##                  29.51070890                   0.27485198
##                    bathrooms                 first_review
##                   0.10893871                   0.04343812
##           host_response_rate                  last_review
##                  -0.02355706                   0.05732861
##                     latitude                    longitude
##                   0.46939271                 -29.94564776
##            number_of_reviews         review_scores_rating
##                  -0.03766385                   0.06065545
```

```
##                  cityChicago                       cityDC
##                 -23.24523069                  -7.66494418
##                       cityLA                      cityNYC
##                 -64.15264317                  -3.77464607
##                       citySF  host_has_profile_picTRUE
##                 -69.73487033                  -0.14390002
## host_identity_verifiedTRUE         instant_bookableTRUE
##                  -0.04870497                  -0.01985495
```

```
##                     bedrooms                         beds
##                   0.16278239                  -0.08297697
##          property_typeOther  property_typeUnique Stays
##                   0.17263609                  -0.14220902
##       room_typePrivate room        room_typeShared room
##                  -0.83880579                  -1.43267918
##            bed_typeReal Bed              cleaning_feeTRUE
##                   0.04719036                  -0.05543024
```

R-squared: 0.58

Find the variables that have coefficients higher than 0.5

1 Introduction

2 Data Cleaning

3 Explore

4 Regression Model

5 Insights

6 Conclusion and Extension

Insights

From the results, there are several intuitive insights.

- More bathrooms, bedrooms, accommodates can increase the price.
- However, more beds can lead to lower price. It may because more beds implied sharing room.
- Shared room and Private room can decrease the price compared with Entire Home/Apt

bathrooms	bedrooms	accommodates	beds
0.10893871	0.16278239	0.27485198	-0.08297697

room_typePrivate room	room_typeShared room
-0.83880579	-1.43267918

Insights

Also, there are some interesting results.

- In property_type, others will increase the price while unique stays decrease the price. We can see some fantastic types in unique stays, such as castle, island. But their price is lower than normal house.
- Review scores don't increase the price significantly. That's kind of reasonable since we have good lodges in every point of price range.
- host_has_profile_pic seems quite significant. It may be caused by amount of data. There are only 1% of hosts don't have a picture for their lodges.
- The latitude and longitude is significantly influencing the price. Which is wield. From the early analysis, we can see that dealing with location related number is difficult.

```
property_typeOther
        0.17263609

property_typeUnique Stays
             -0.14220902

review_scores_rating
          0.06065545

latitude
0.46939271

longitude
-29.94564776
```

Insights

Finally, we can get the city related influence. Boston is the most expensive city for lodging, followed by NYC and DC. While SF, La and Chicago are much cheaper.

```
 cityChicago                  cityDC
-23.24523069             -7.66494418
      cityLA                 cityNYC
-64.15264317             -3.77464607
      citySF  host_has_profile_picTRUE
-69.73487033             -0.14390002
```

Boston → NYC → DC → Chicago → LA → SF

Conclusion and Extension

- Cleaned up the data set about Airbnb lodges
- Analyzed data by preliminary methods like MVA and T-test
- Used multiple regression to get a predicting model
- Draw some insights from the model

There are some following direction can be explored afterwards.

- The influence of property_type can be analyzed further, since the board classification is still crude.
- The location of data, namely the latitude and longitude can be explored more. Classic cluster and regression methods may not be able to deal with this part.
- About the text type of data, we can use text mining methods to put them to use.

Conclusion and Extension

- There is a lots of works before we actually run the regression. We need to clean up the data, explore with many methods for a clear direction.
- Analysis is a dynamic process; we need to run the model and do the analysis iteratively.
- Finally, I should go to SF, LA or Chicago for a holiday travel. Otherwise, I should prepare for an expensive lodging.

Thank you for your attention!

Tonghua Lin
Email: tonghua.lin@rutgers.edu
Tel: 862-423-9940
Cubicle: 1WP 951A