THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

The LA GLAM

Help you find your ideal home in LA

Lingyao Meng
Kamran Qureshi
Qinya Wang
Tianweibao Zheng

Part 1 Selection

Data Source: http://us-city.census.okfn.org/entry/losangeles/parcels

Parcels: Parcel data is data on the geographic boundaries of property. Parcels are the most specific units of geodata that governments maintain. The data is mainly used by the County Assessor's office to assess property taxes, yet is also used to keep track of addresses, other type of taxes, and zoning information. Because of their granularity, parcel data can be used for very detailed maps, visualizations, and all kinds of applications.

Official Parcel boundaries in the City of Los Angeles created and maintained by the Bureau of Engineering / GIS Mapping Division.

We want to explore the predictors of home price in LA

Filters

2394065 rows in the metadata,

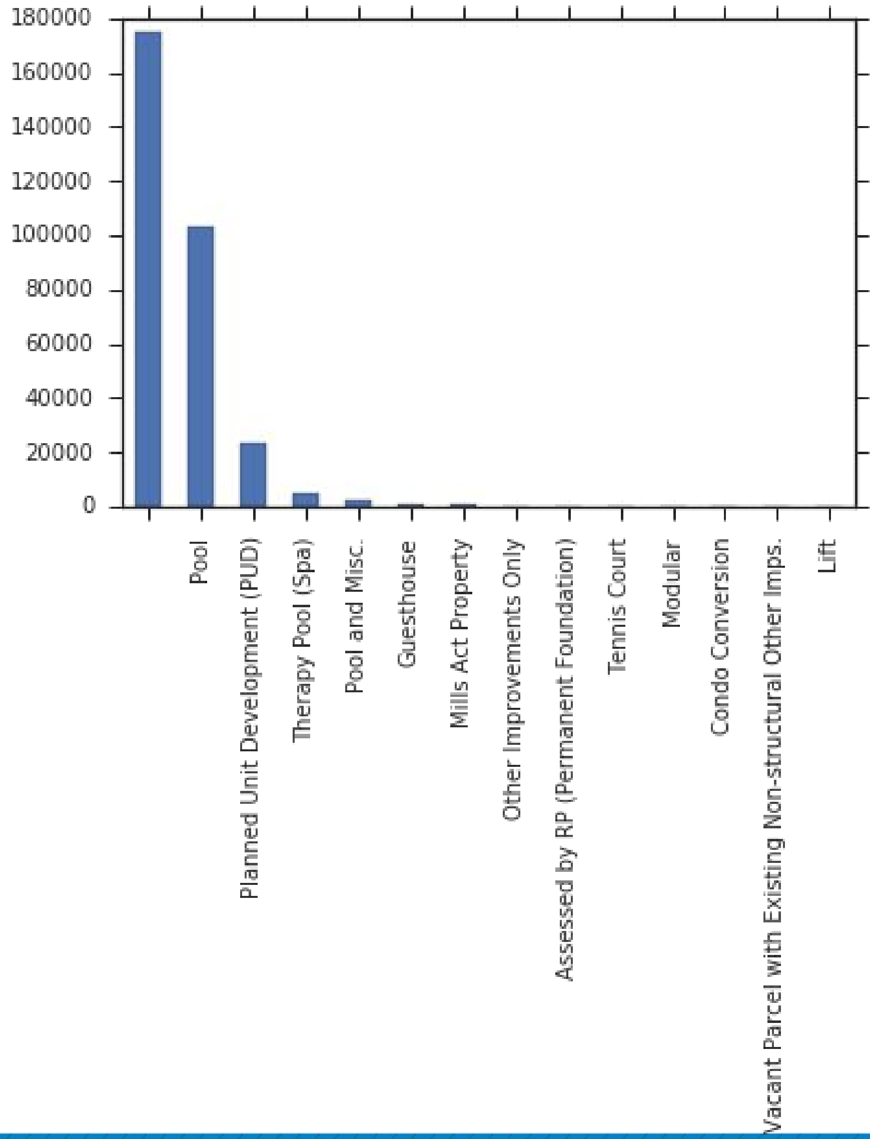
Roll year = 2016

Property type = SFR (single family residence)

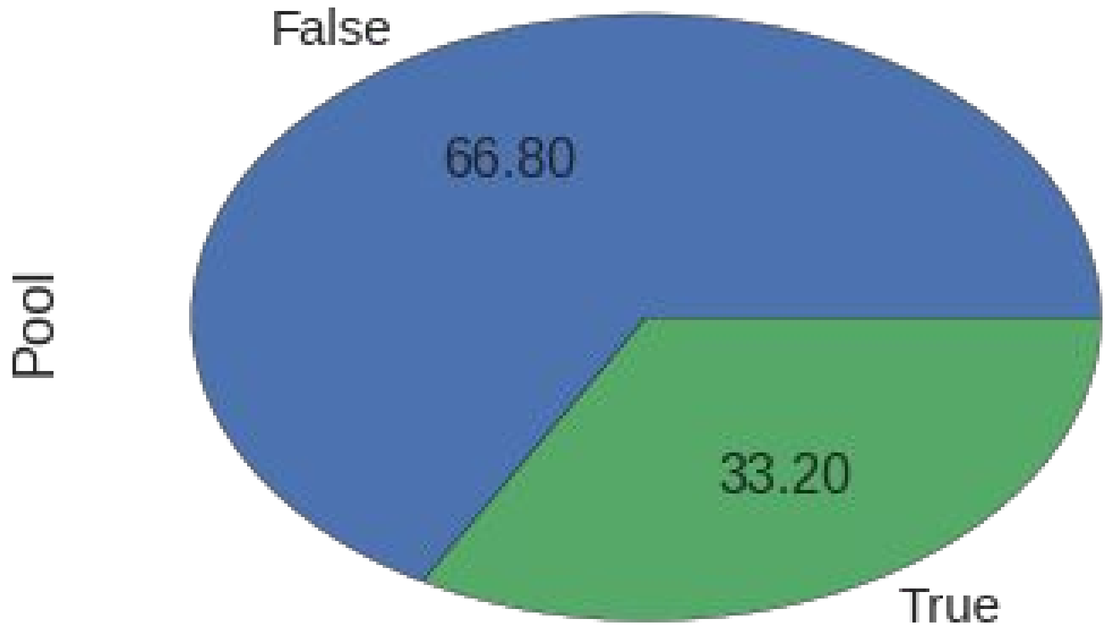
Bedroom > 3

Bathroom > 2

SpecificUseDetail2



```
df['Pool'] = (df['SpecificUseDetail2'] == 'Pool')
df['Pool']
```



df['TotalValue'].describe()

```
Count: 310757
Unique: 192671
Top: $650,000.00
Freq: 118
Name: TotalValue
dtype: object
```

```
count    3.107570e+05
mean     8.335643e+05
std      1.671932e+06
min      0.000000e+00
25%      3.194600e+05
50%      5.272000e+05
75%      8.791670e+05
max      5.830608e+08
Name: TotalValue
dtype: float64
```

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

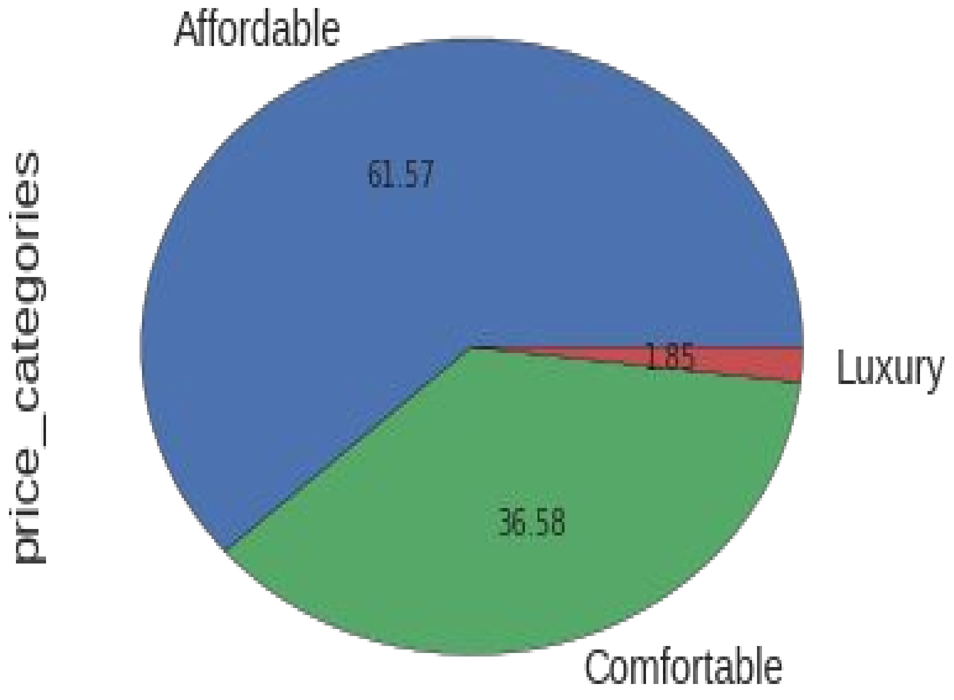
bins = [0, 650000, 4000000, 49000000]

group_names = ['Affordable', 'Comfortable', 'Luxury']

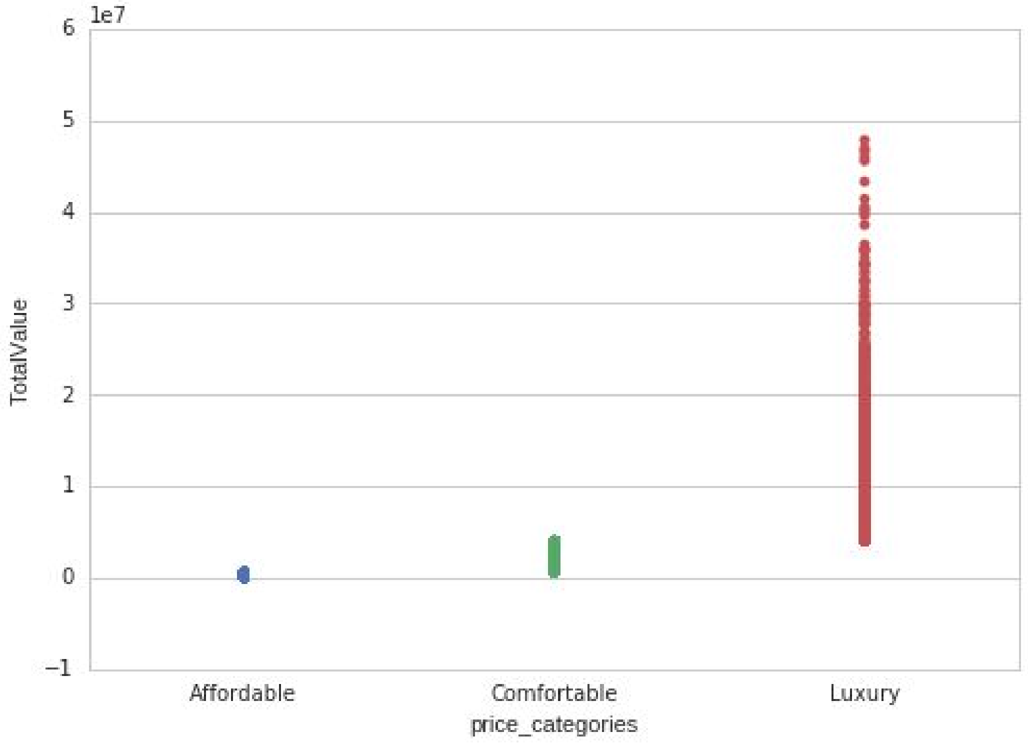
categories = pd.cut(df['TotalValue'], bins, labels=group_names)

df['price_categories'] = pd.cut(df['TotalValue'], bins, labels=group_names)

df['price_categories'].value_counts().plot(kind='pie',autopct='%.2f')



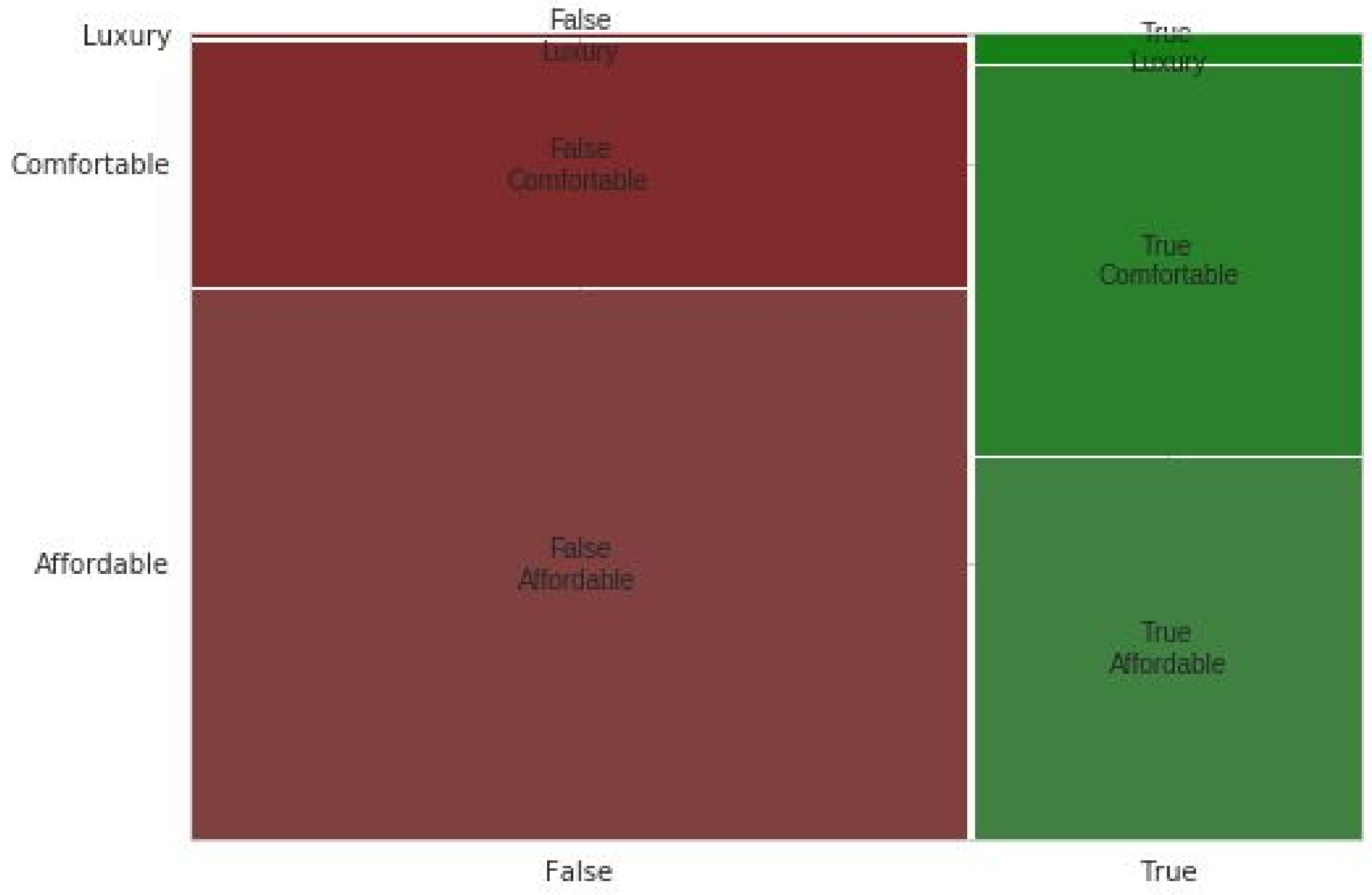
```
import numpy as np
import pandas as pd
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style="whitegrid",
color_codes=True)
sns.stripplot(df['price_categori
es'],df[ 'TotalValue'])
```



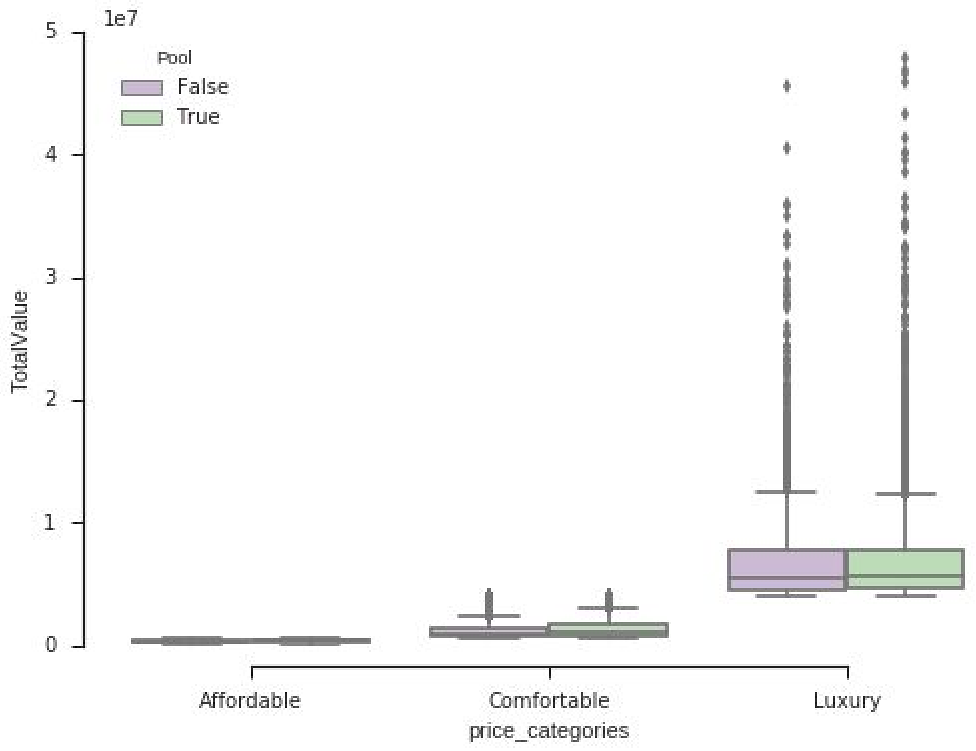
THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

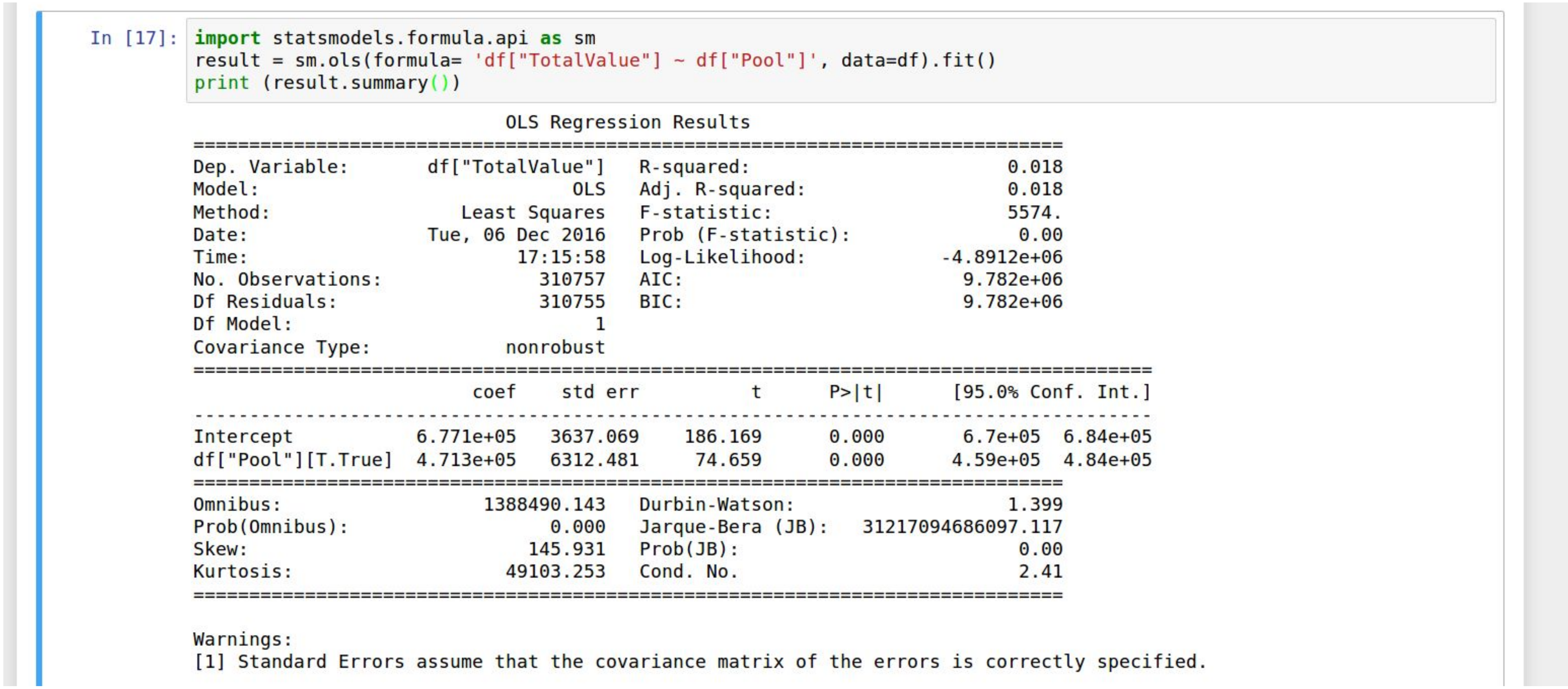
mosaic(df, ['Pool', 'price_categories'])



```
sns.boxplot(df['price_categories'],df[ 'TotalValue'], hue=df['Pool'], palette="PRGn")
sns.despine(offset=10, trim=True)
```

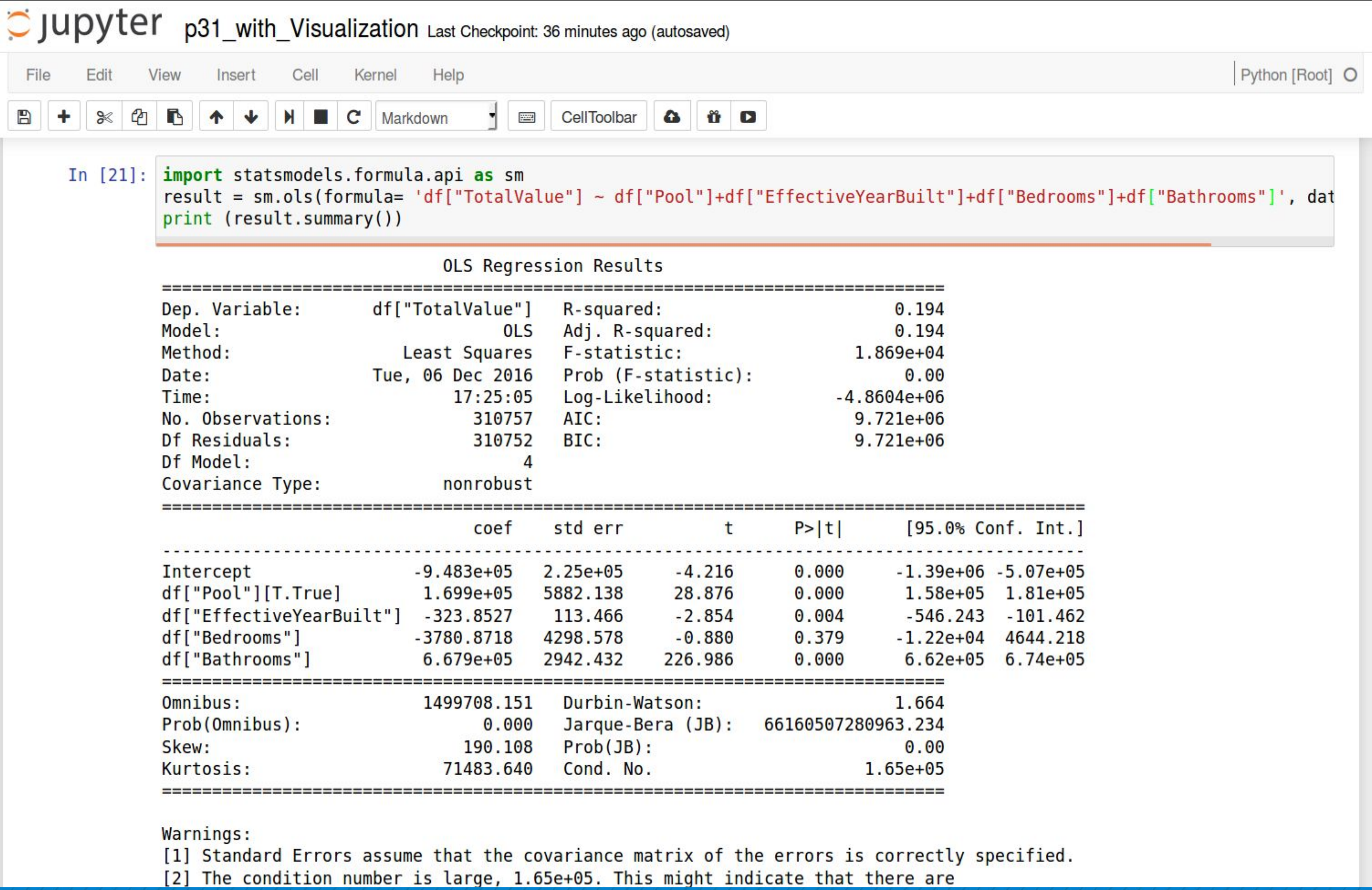


Have a pool or not?



```
In [17]: import statsmodels.formula.api as sm
         result = sm.ols(formula= 'df["TotalValue"] ~ df["Pool"]', data=df).fit()
         print (result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      df["TotalValue"]   R-squared:                       0.018
Model:                           OLS   Adj. R-squared:                  0.018
Method:                Least Squares   F-statistic:                     5574.
Date:               Tue, 06 Dec 2016   Prob (F-statistic):               0.00
Time:                       17:15:58   Log-Likelihood:            -4.8912e+06
No. Observations:             310757   AIC:                         9.782e+06
Df Residuals:                 310755   BIC:                         9.782e+06
Df Model:                          1
Covariance Type:           nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------------
Intercept           6.771e+05   3637.069    186.169      0.000       6.7e+05  6.84e+05
df["Pool"][T.True]  4.713e+05   6312.481     74.659      0.000      4.59e+05  4.84e+05
==============================================================================
Omnibus:                 1388490.143   Durbin-Watson:                   1.399
Prob(Omnibus):                 0.000   Jarque-Bera (JB):   31217094686097.117
Skew:                        145.931   Prob(JB):                         0.00
Kurtosis:                  49103.253   Cond. No.                         2.41
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```


jupyter p31_with_Visualization Last Checkpoint: 36 minutes ago (autosaved)
File Edit View Insert Cell Kernel Help
Python [Root]
Markdown
CellToolbar

```
In [21]: import statsmodels.formula.api as sm
         result = sm.ols(formula= 'df["TotalValue"] ~ df["Pool"]+df["EffectiveYearBuilt"]+df["Bedrooms"]+df["Bathrooms"]', dat
         print (result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       df["TotalValue"]   R-squared:                       0.194
Model:                            OLS   Adj. R-squared:                  0.194
Method:                 Least Squares   F-statistic:                 1.869e+04
Date:                Tue, 06 Dec 2016   Prob (F-statistic):               0.00
Time:                        17:25:05   Log-Likelihood:            -4.8604e+06
No. Observations:              310757   AIC:                         9.721e+06
Df Residuals:                  310752   BIC:                         9.721e+06
Df Model:                           4
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------------
Intercept             -9.483e+05   2.25e+05     -4.216      0.000     -1.39e+06 -5.07e+05
df["Pool"][T.True]     1.699e+05   5882.138     28.876      0.000      1.58e+05  1.81e+05
df["EffectiveYearBuilt"] -323.8527  113.466     -2.854      0.004      -546.243  -101.462
df["Bedrooms"]        -3780.8718   4298.578     -0.880      0.379     -1.22e+04  4644.218
df["Bathrooms"]        6.679e+05   2942.432    226.986      0.000      6.62e+05  6.74e+05
==============================================================================
Omnibus:                  1499708.151   Durbin-Watson:                   1.664
Prob(Omnibus):                  0.000   Jarque-Bera (JB):   66160507280963.234
Skew:                         190.108   Prob(JB):                         0.00
Kurtosis:                   71483.640   Cond. No.                     1.65e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.65e+05. This might indicate that there are
```

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Part 2 - Wrangling

Based on the data we filtered above, we wrangle the data into a format suitable for analysis.

1. we drop the redundant columns:Save the columns we want to keep in a new file using”

   !csvcut -c1,11-13,15-20,22,24,29,41,43-50 p2.csv > p31.csv"

Repeated information:2. TaxRateArea_CITY, 3. AIN: included in rowID, 4. RollYear, 5. TaxRateArea: same number as AIN. 7. PropertyLocation, 8. PropertyType, 9. PropertyUseCode, 10. GeneralUseType, 26. TotalLandImpValue=Total LandValue + ImprovementValue on this assessment roll, 34. TotalValue= LandValue + ImprovementValue + FixtureValue + PersonalPropertyValue, 35. TotalExemption= HomeownersExemption + RealEstateExemption + FixtureExemption + PersonalPropertyExemption, 36. netTaxableValue= column34 - column35 51. Location 1

Not related: 6. AssessorID 14. totBuildingDataLines   21. RecordingDate   23. LandBaseYear 25. ImpBaseYear   27. HomeownersExemptionon   28. RealEstateExemption   30. FixtureExemption   31. PersonalPropertyValue   32. PersonalPropertyExemptionur   33. isTaxableParcel?   37. SpecialParcelClassification   38. AdministrativeRegion 39. Cluster   40. ParcelBoundaryDescription 42. HouseFraction

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

```
!csvcut -n p31.csv
```

```
  1: ZIPcode
  2: SpecificUseType
  3: SpecificUseDetail1
  4: SpecificUseDetail2
  5: YearBuilt
  6: EffectiveYearBuilt
  7: SQFTmain
  8: Bedrooms
  9: Bathrooms
 10: Units
 11: LandValue
 12: ImprovementValue
 13: FixtureValue
 14: HouseNo
 15: StreetDirection
 16: StreetName
 17: UnitNo
 18: City
 19: ZIPcode5
 20: rowID
 21: CENTER_LAT
 22: CENTER_LON
```

```
!csvcut -c20,21,22 p31.csv | csvstat
```

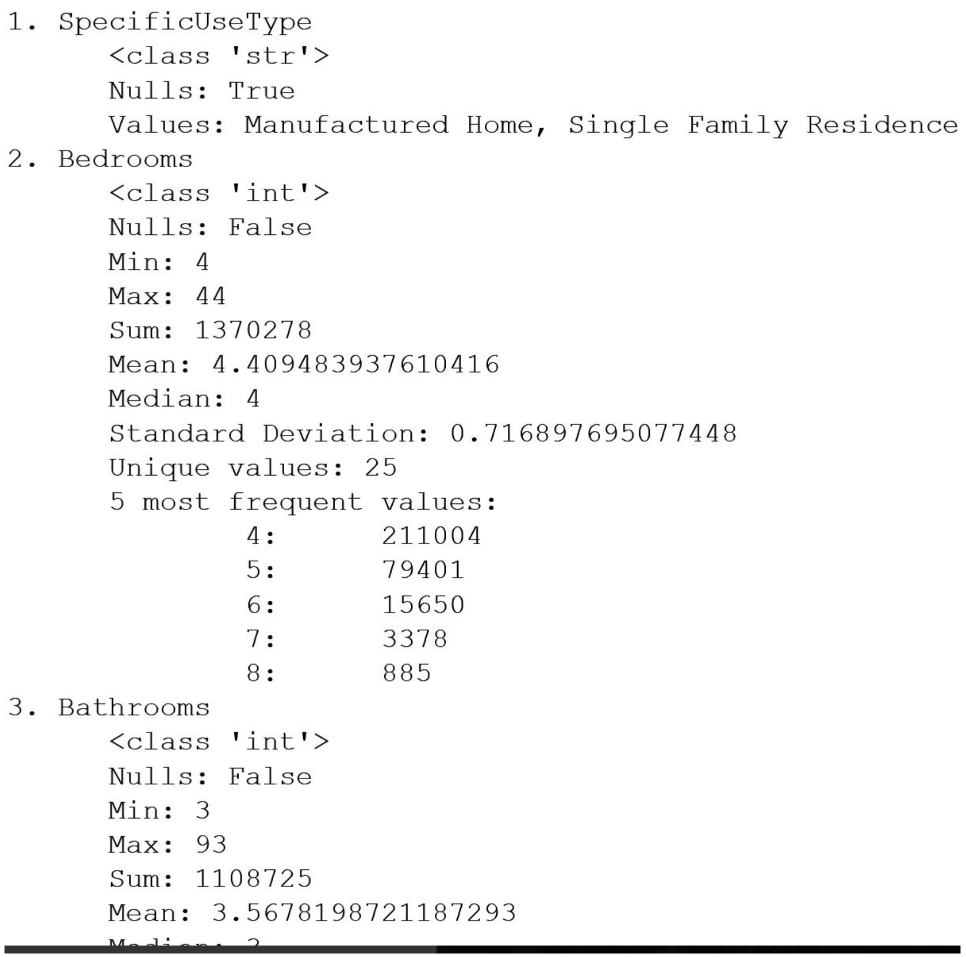
```
  1. rowID
        <class 'int'>
        Nulls: False
        Min: 20162004001003
        Max: 20168765022045
        Sum: 6266415875014577263
        Mean: 20165003121456.887
        Median: 20164411014009
        Standard Deviation: 2296140842.4706125
        Unique values: 310757
  2. CENTER_LAT
        <class 'float'>
        Nulls: True
        Min: 33.33971975
        Max: 34.81962976
        Sum: 10607209.542702254
        Mean: 34.13466156507969
        Median: 34.117165985
        Standard Deviation: 0.22305192815144012
        Unique values: 306772
        5 most frequent values:
                34.12514642:    71
                34.12231882:    61
                33.82320686:    43
                34.41000082:    37
                34.04246356:    36
```

2. Examine and filter again:

Take a look these columns, we may need further filtering.

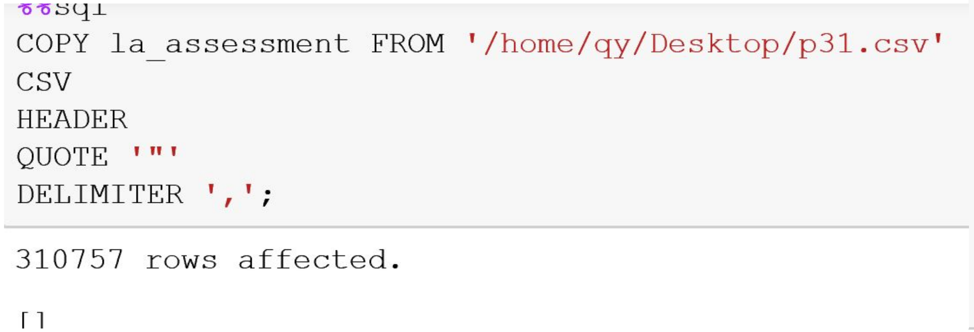
Because first, we are explorating Single Family Residence; Second, the number of rooms need to be controlled in reasonable ranges.

we assume the range of Bedrooms is[5,10], and that of Bathrooms is[3,6];finally, to precise our analysis, we assume Units to be 1. The filters are added below in the SQL part.



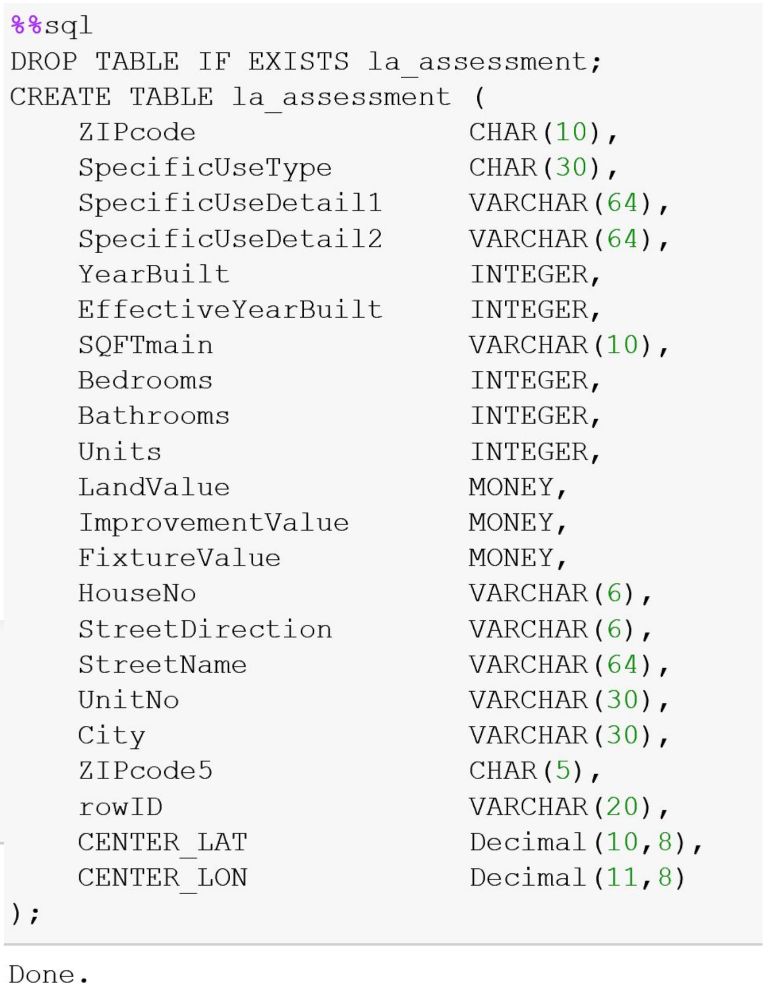
```
1. SpecificUseType
        <class 'str'>
        Nulls: True
        Values: Manufactured Home, Single Family Residence
2. Bedrooms
        <class 'int'>
        Nulls: False
        Min: 4
        Max: 44
        Sum: 1370278
        Mean: 4.409483937610416
        Median: 4
        Standard Deviation: 0.716897695077448
        Unique values: 25
        5 most frequent values:
                4:      211004
                5:      79401
                6:      15650
                7:      3378
                8:      885
3. Bathrooms
        <class 'int'>
        Nulls: False
        Min: 3
        Max: 93
        Sum: 1108725
        Mean: 3.5678198721187293
        Median: 3
```

Load our data into database, introduce postgresql



```
%%sql
COPY la_assessment FROM '/home/qy/Desktop/p31.csv'
CSV
HEADER
QUOTE '"'
DELIMITER ',';
```

```
310757 rows affected.

[]
```



```
%%sql
DROP TABLE IF EXISTS la_assessment;
CREATE TABLE la_assessment (
    ZIPcode                 CHAR(10),
    SpecificUseType         CHAR(30),
    SpecificUseDetail1      VARCHAR(64),
    SpecificUseDetail2      VARCHAR(64),
    YearBuilt               INTEGER,
    EffectiveYearBuilt      INTEGER,
    SQFTmain                VARCHAR(10),
    Bedrooms                INTEGER,
    Bathrooms               INTEGER,
    Units                   INTEGER,
    LandValue              MONEY,
    ImprovementValue       MONEY,
    FixtureValue           MONEY,
    HouseNo                 VARCHAR(6),
    StreetDirection         VARCHAR(6),
    StreetName              VARCHAR(64),
    UnitNo                  VARCHAR(30),
    City                    VARCHAR(30),
    ZIPcode5                CHAR(5),
    rowID                   VARCHAR(20),
    CENTER_LAT              Decimal(10,8),
    CENTER_LON              Decimal(11,8)
);
```

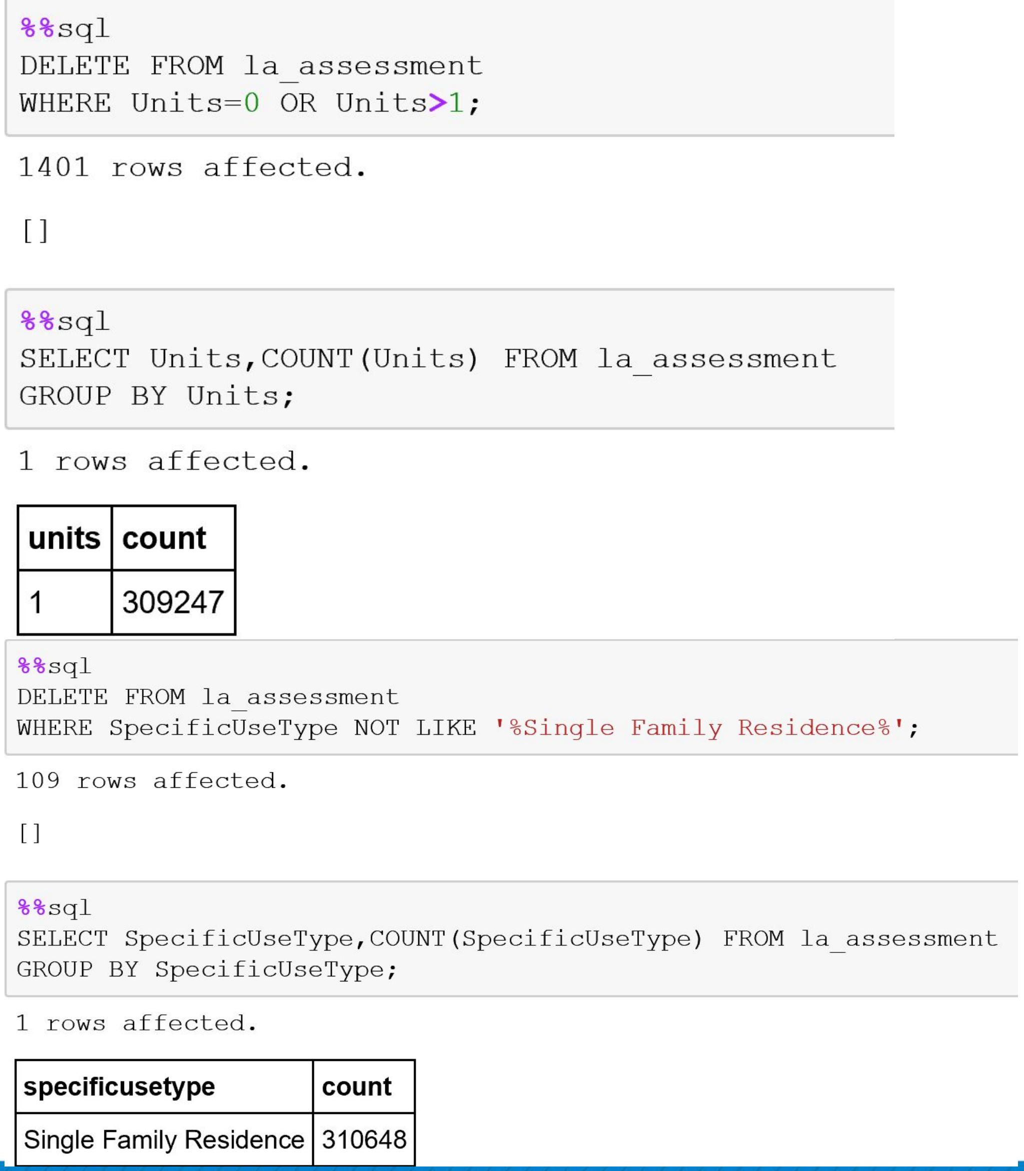
```
Done.
```

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

We set our data as:

Special Use Type=Single Family Residence

Units = 1 (Total number of living units)



```
%%sql
DELETE FROM la_assessment
WHERE Units=0 OR Units>1;
```

```
1401 rows affected.

[]
```

```
%%sql
SELECT Units,COUNT(Units) FROM la_assessment
GROUP BY Units;
```

```
1 rows affected.
```

units	count
1	309247

```
%%sql
DELETE FROM la_assessment
WHERE SpecificUseType NOT LIKE '%Single Family Residence%';
```

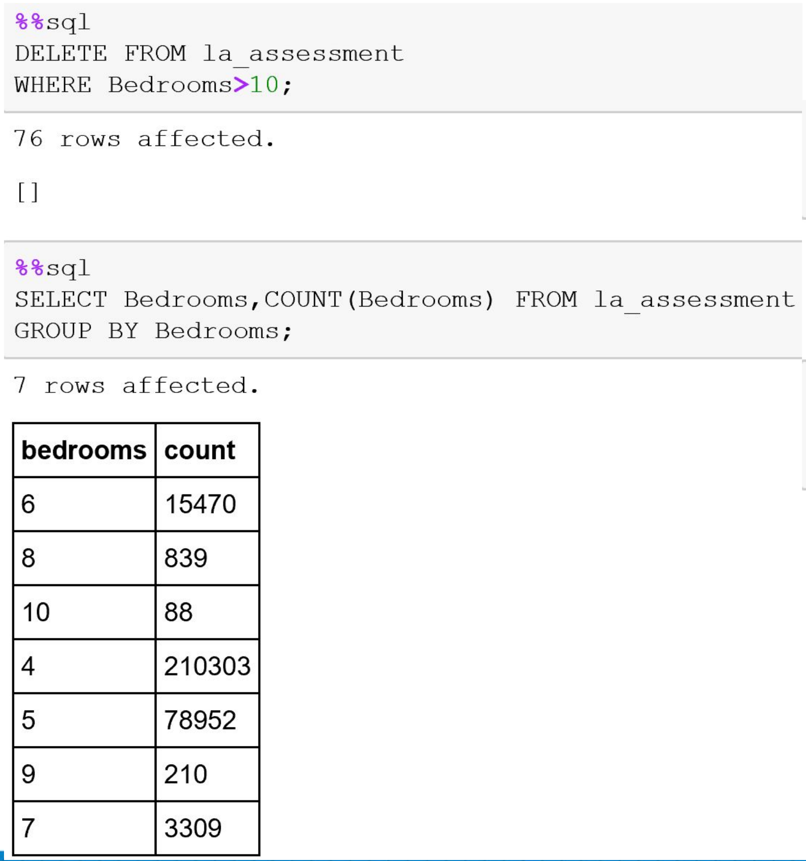
```
109 rows affected.

[]
```

```
%%sql
SELECT SpecificUseType,COUNT(SpecificUseType) FROM la_assessment
GROUP BY SpecificUseType;
```

```
1 rows affected.
```

specificusetype	count
Single Family Residence	310648

4=<Bedrooms<=10
3=<Bathrooms <=6



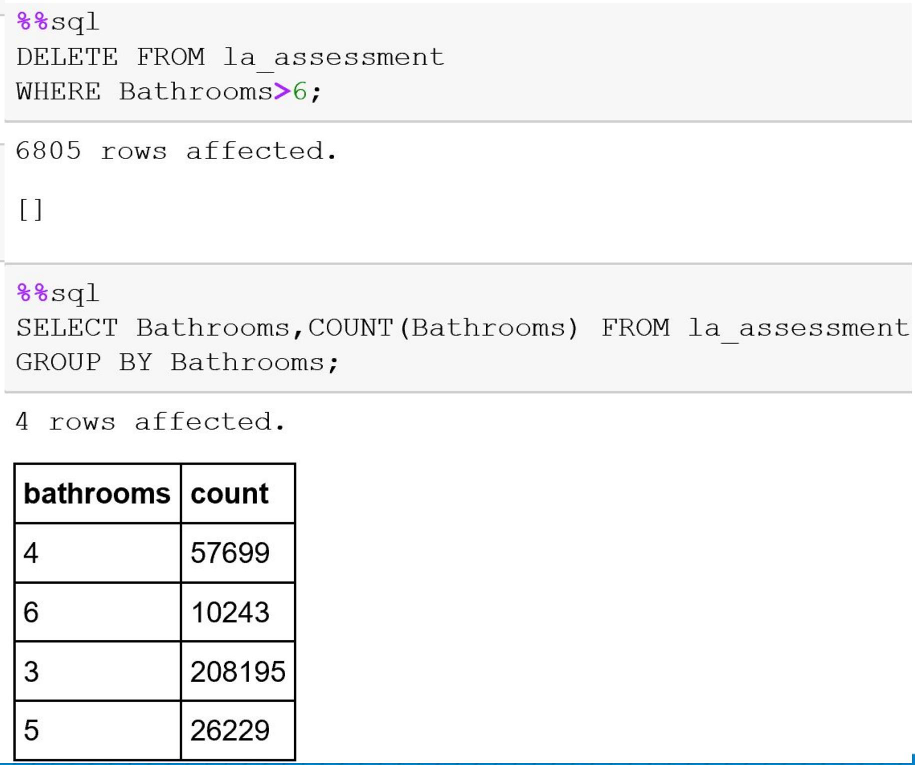
```
%%sql
DELETE FROM la_assessment
WHERE Bedrooms>10;
```

```
76 rows affected.

[]
```

```
%%sql
SELECT Bedrooms,COUNT(Bedrooms) FROM la_assessment
GROUP BY Bedrooms;
```

```
7 rows affected.
```

bedrooms	count
6	15470
8	839
10	88
4	210303
5	78952
9	210
7	3309



```
%%sql
DELETE FROM la_assessment
WHERE Bathrooms>6;
```

```
6805 rows affected.

[]
```

```
%%sql
SELECT Bathrooms,COUNT(Bathrooms) FROM la_assessment
GROUP BY Bathrooms;
```

```
4 rows affected.
```

bathrooms	count
4	57699
6	10243
3	208195
5	26229

Drop the SpecificUseType and Units, identical values

```
%%sql
ALTER TABLE la_assessment
DROP COLUMN SpecificUseType;
```

```
Done.

[]
```

```
%%sql
ALTER TABLE la_assessment
DROP COLUMN Units;
```

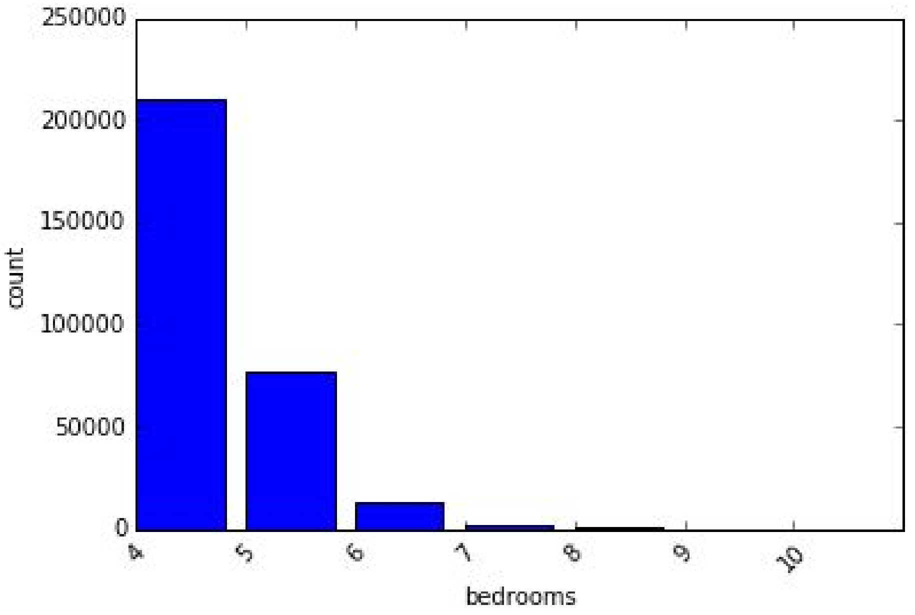
```
Done.
```

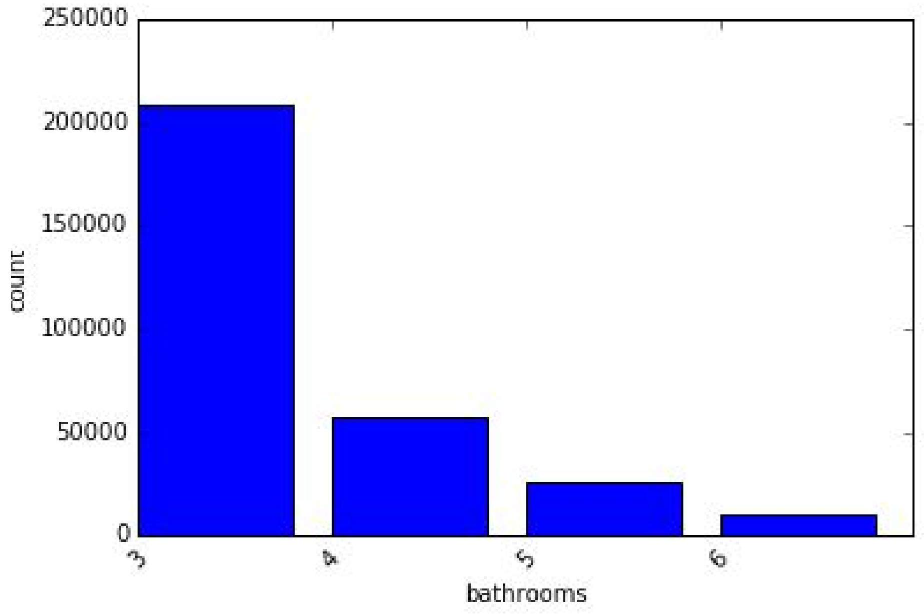
Part 3 Analysis

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Number of properties by Number of Bedrooms & Bathrooms





THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Total Value

V.S.

Number of Bedrooms

```
SELECT Bedrooms, AVG(CAST(Totalvalue AS decimal)) FROM la_assessment
GROUP BY Bedrooms
ORDER BY Bedrooms;
```

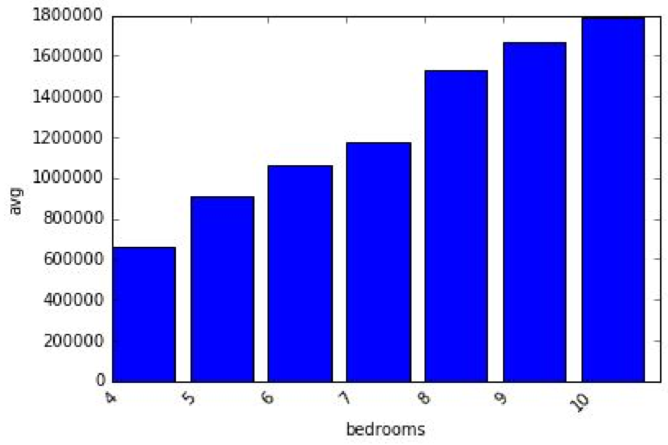
```
7 rows affected.
```

Out[62]:

bedrooms	avg
4	657018.350425027771
5	908849.820506141162
6	1062606.003751052591
7	1174554.085096803242
8	1532844.014018691589
9	1666046.385416666667
10	1788516.300000000000

```
In [47]: bedrooms = _
         bedrooms.bar()
```

```
Out[47]: <Container object of 7 artists>
```



Total Value

V.S.

Number of Bathrooms

```
In [48]: %%sql
SELECT Bathrooms, AVG(CAST(Totalvalue AS decimal)) FROM la_assessment
GROUP BY Bathrooms
ORDER BY Bathrooms;
```

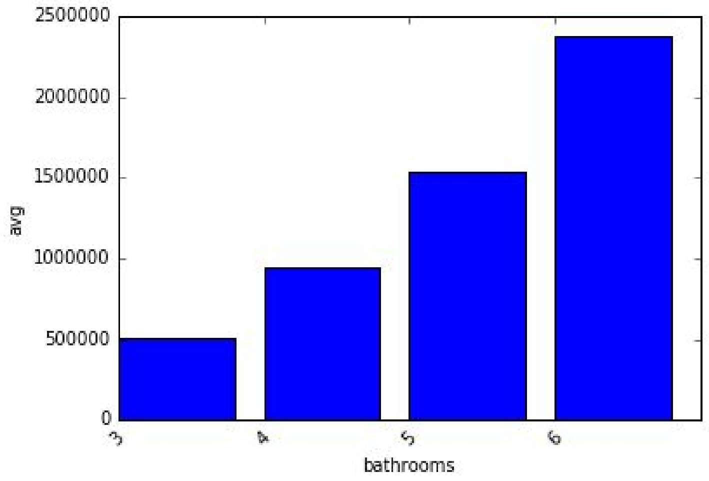
```
4 rows affected.
```

Out[48]:

bathrooms	avg
3	507537.415816902423
4	945144.767084351549
5	1540008.173853368409
6	2377686.627452894660

```
In [49]: bathrooms = _
bathrooms.bar()
```

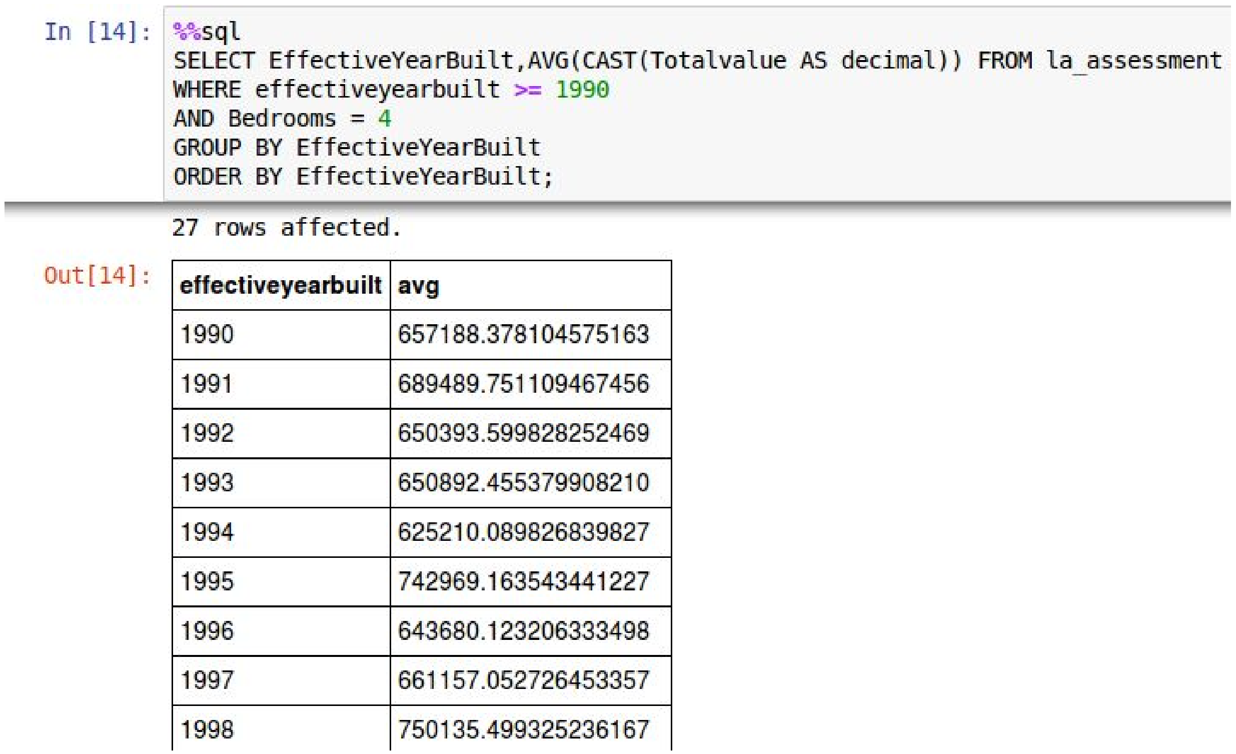
```
Out[49]: <Container object of 4 artists>
```



Total Value

V.S.

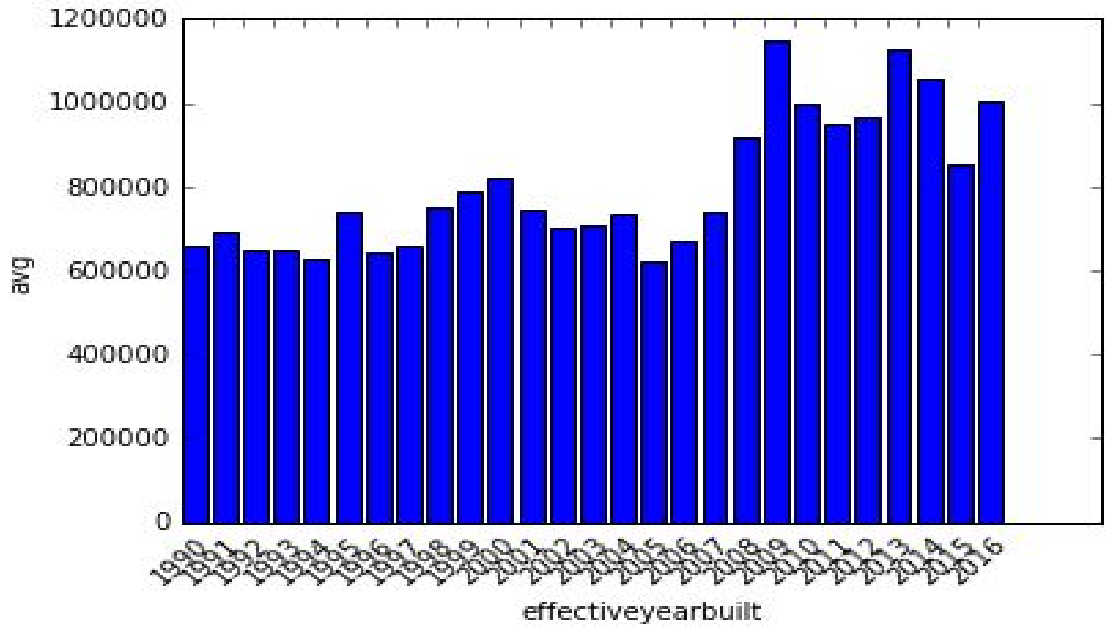
Year Built



```
In [14]: %%sql
         SELECT EffectiveYearBuilt,AVG(CAST(Totalvalue AS decimal)) FROM la_assessment
         WHERE effectiveyearbuilt >= 1990
         AND Bedrooms = 4
         GROUP BY EffectiveYearBuilt
         ORDER BY EffectiveYearBuilt;

         27 rows affected.
```

Out[14]:

effectiveyearbuilt	avg
1990	657188.378104575163
1991	689489.751109467456
1992	650393.599828252469
1993	650892.455379908210
1994	625210.089826839827
1995	742969.163543441227
1996	643680.123206333498
1997	661157.052726453357
1998	750135.499325236167

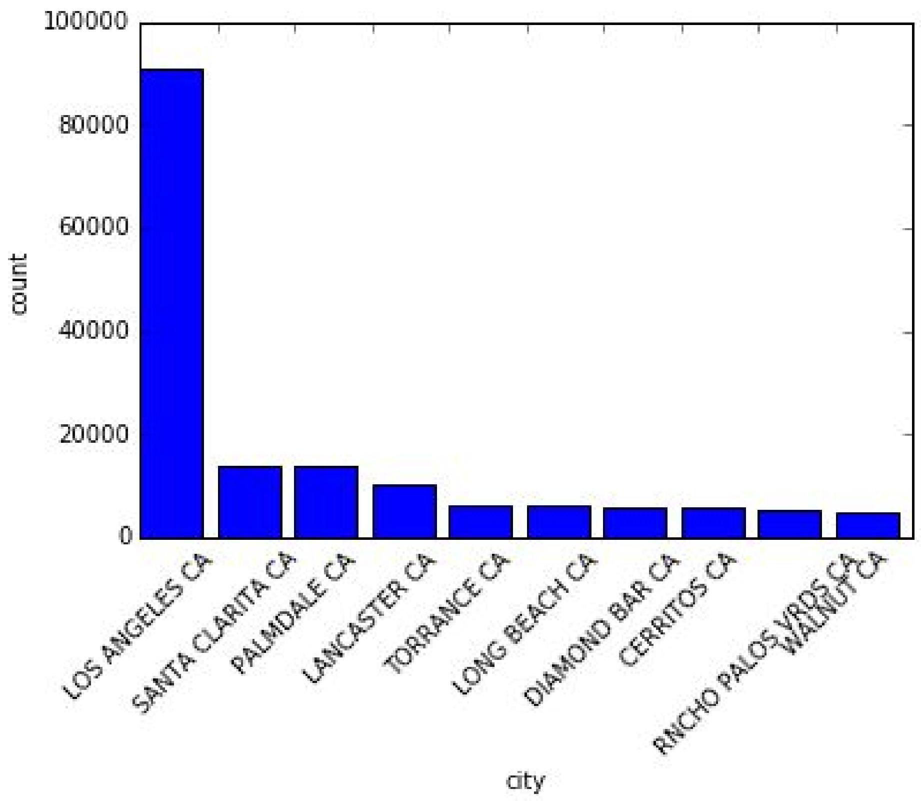


City with most properties

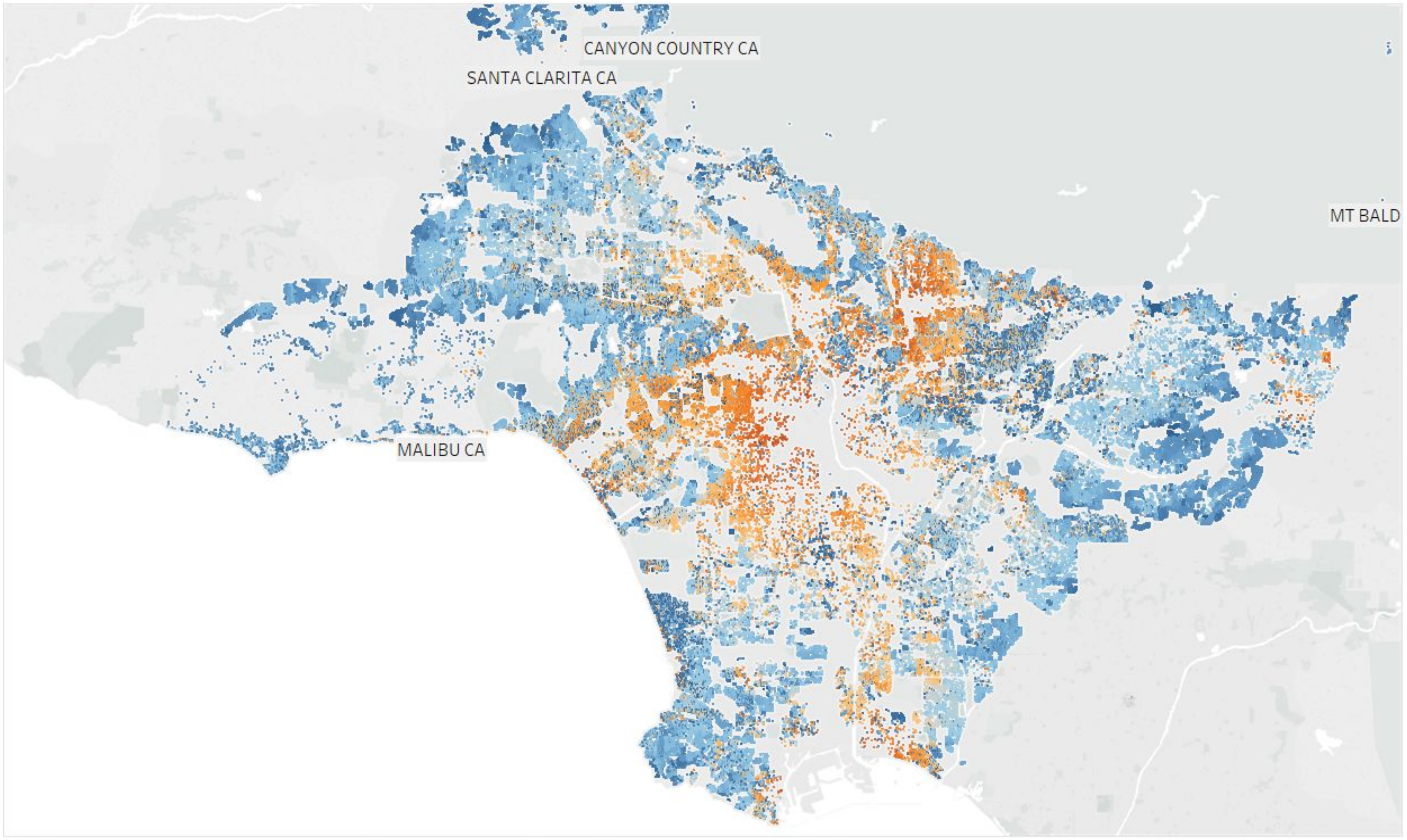
```
%%sql
SELECT City,COUNT(*) AS COUNT FROM la_assessment
GROUP BY City
ORDER BY COUNT DESC
LIMIT 10;
```

```
10 rows affected.
```

city	count
LOS ANGELES CA	90865
SANTA CLARITA CA	14116
PALMDALE CA	13812
LANCASTER CA	10469
TORRANCE CA	6422
LONG BEACH CA	6279
DIAMOND BAR CA	5999
CERRITOS CA	5913
RNCHO PALOS VRDS CA	5517
WALNUT CA	5014



Effective year built V.S. Location



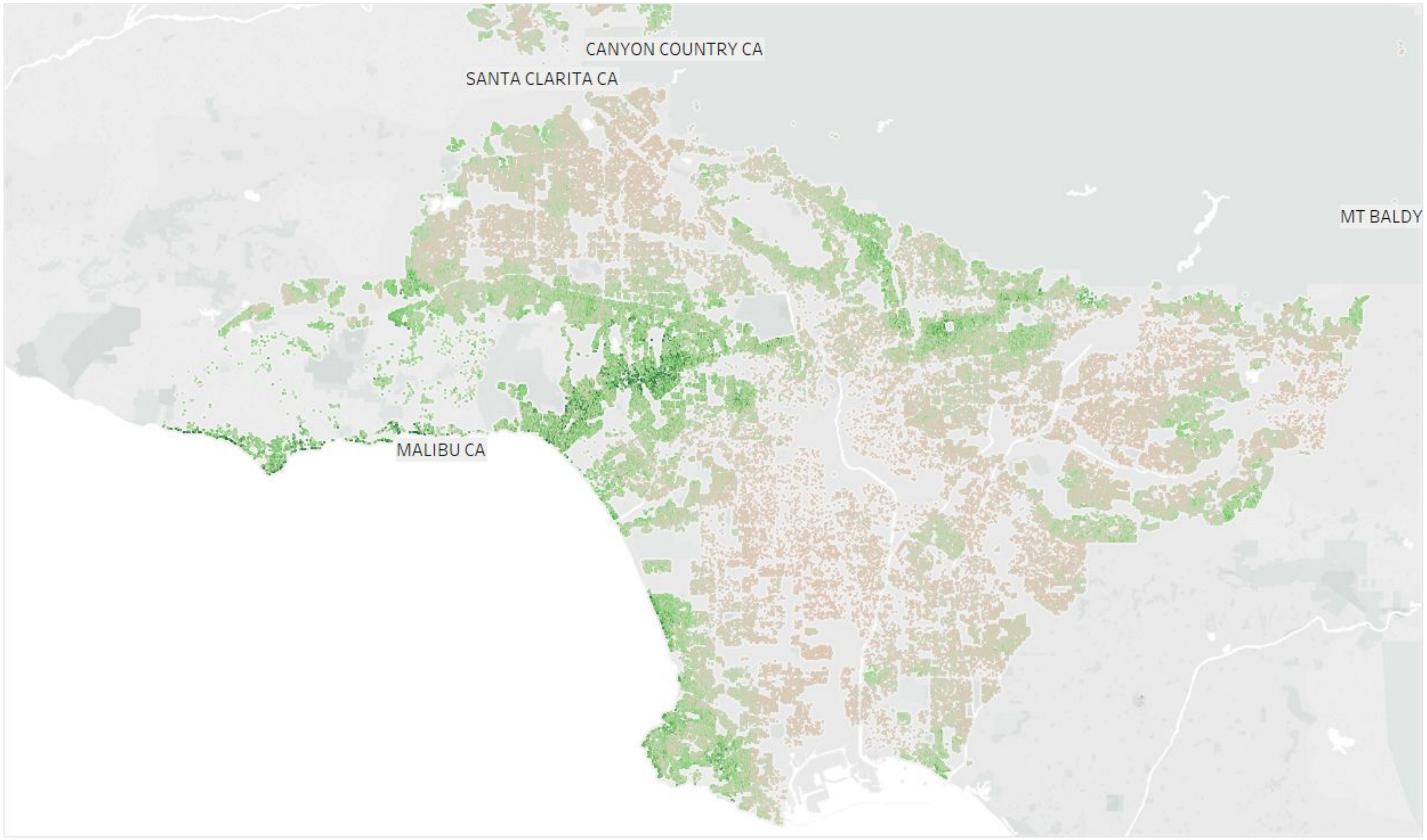
Map based on Center Lon1 and Center Lat1. Color shows average of Yearbuilt. The marks are labeled by City.

1,880.0
2,016.0

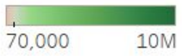
THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Total value V.S. Location



Map based on Center Lon1 and Center Lat1. Color shows average of Totalvalue. The marks are labeled by City as an attribute.

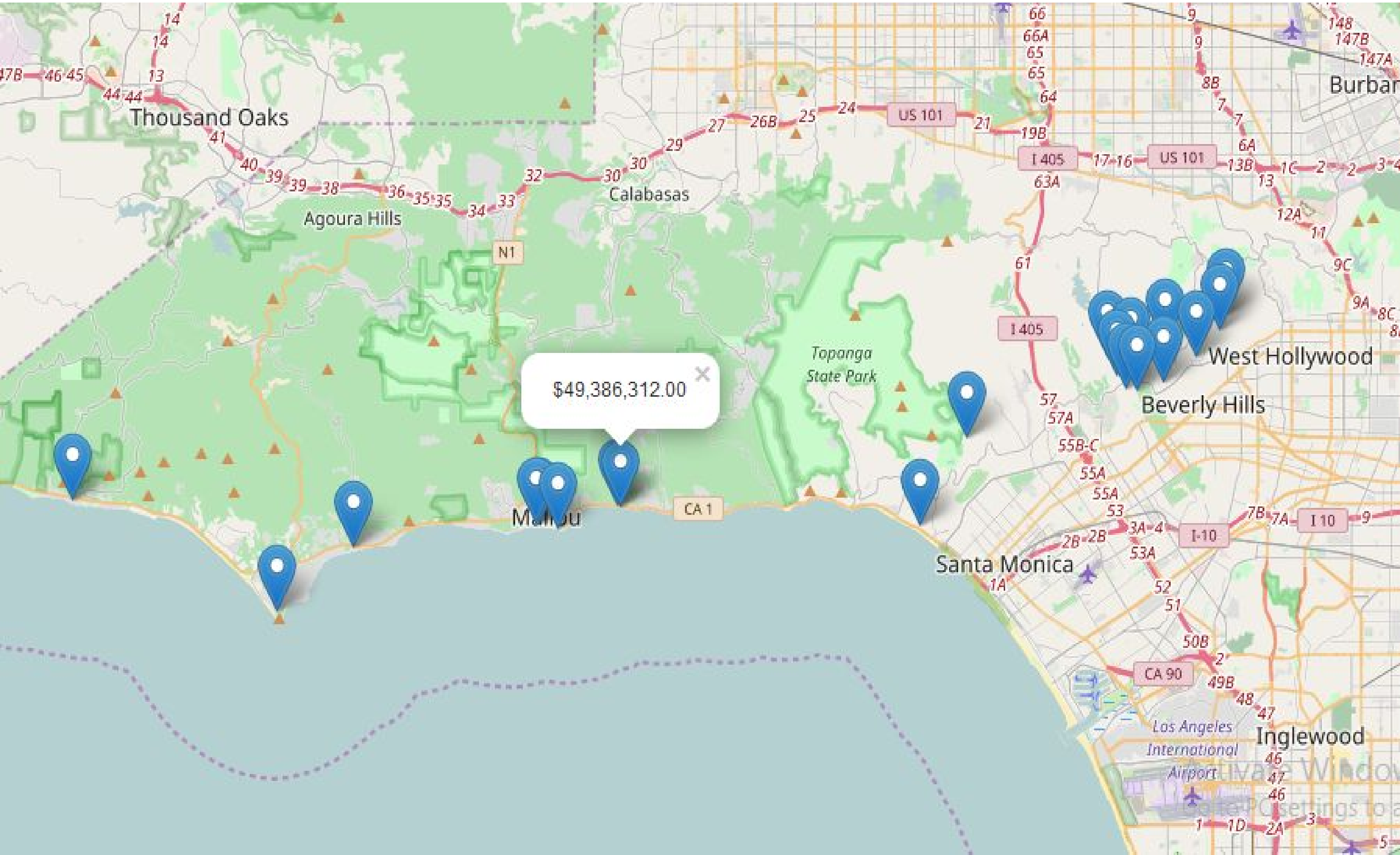


THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Dr. NotSoStrange's dilemma

- Dr. NotSoStrange just moved to LA with his wife and 4 teenage kids
- Being a super specialist, life has been kind to him and he has earned enough to buy a house without having to pay for mortgage later
- He wants to explore areas in LA where he could buy a house for his family to settle in

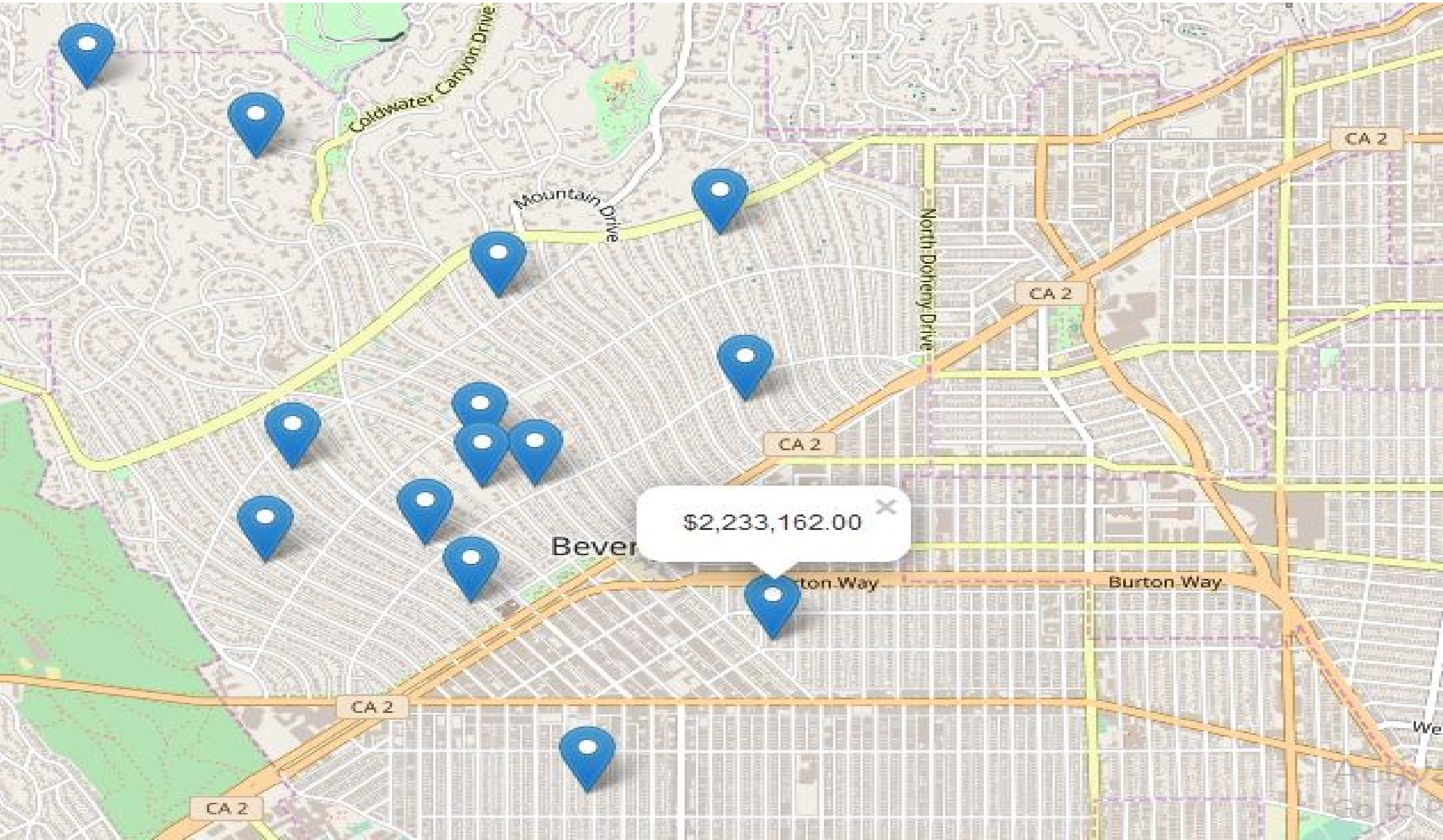
Let's show Dr NotSoStrange highest valued properties in LA and drop a bomb on his bank account



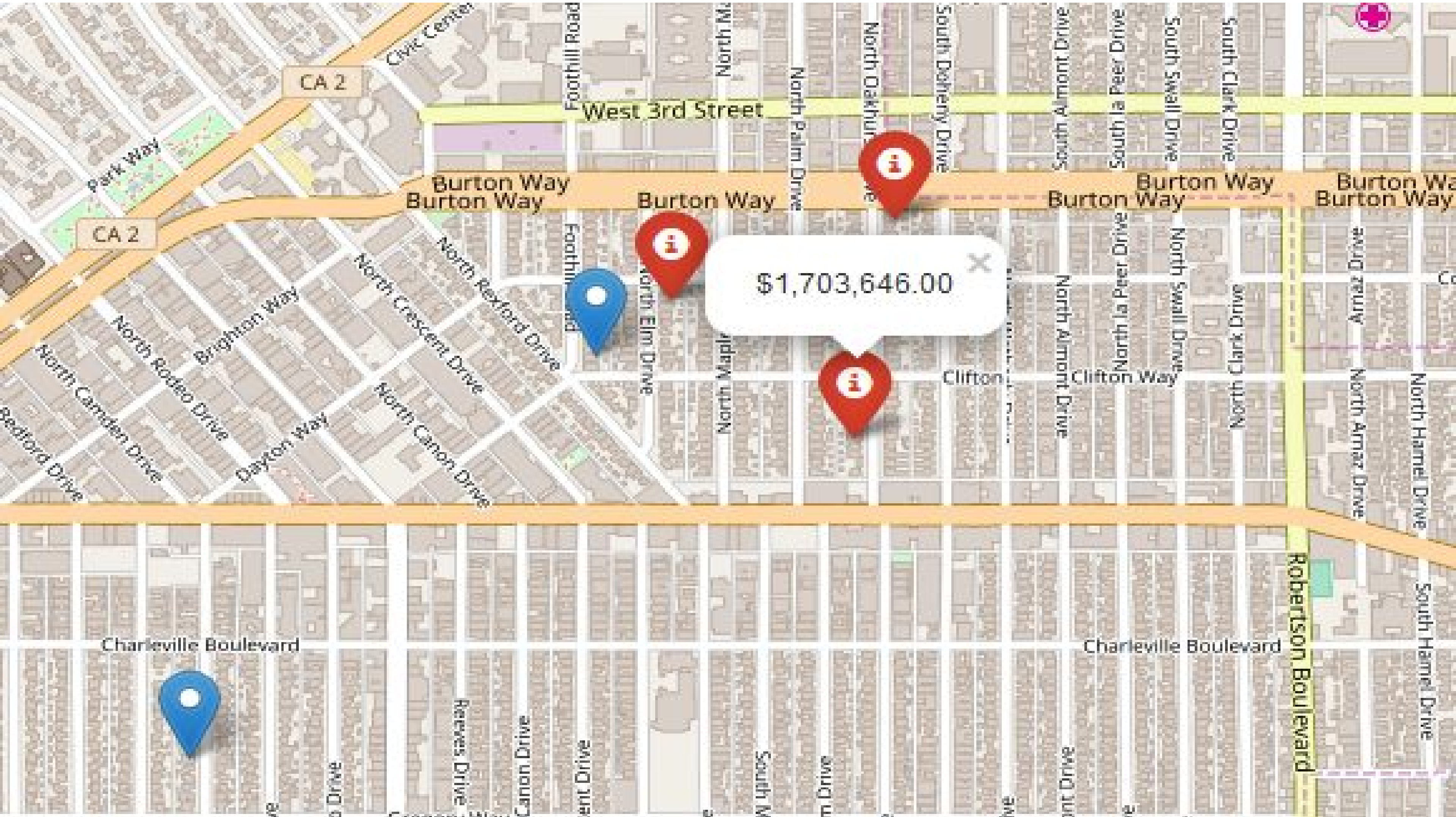
House for a big family!

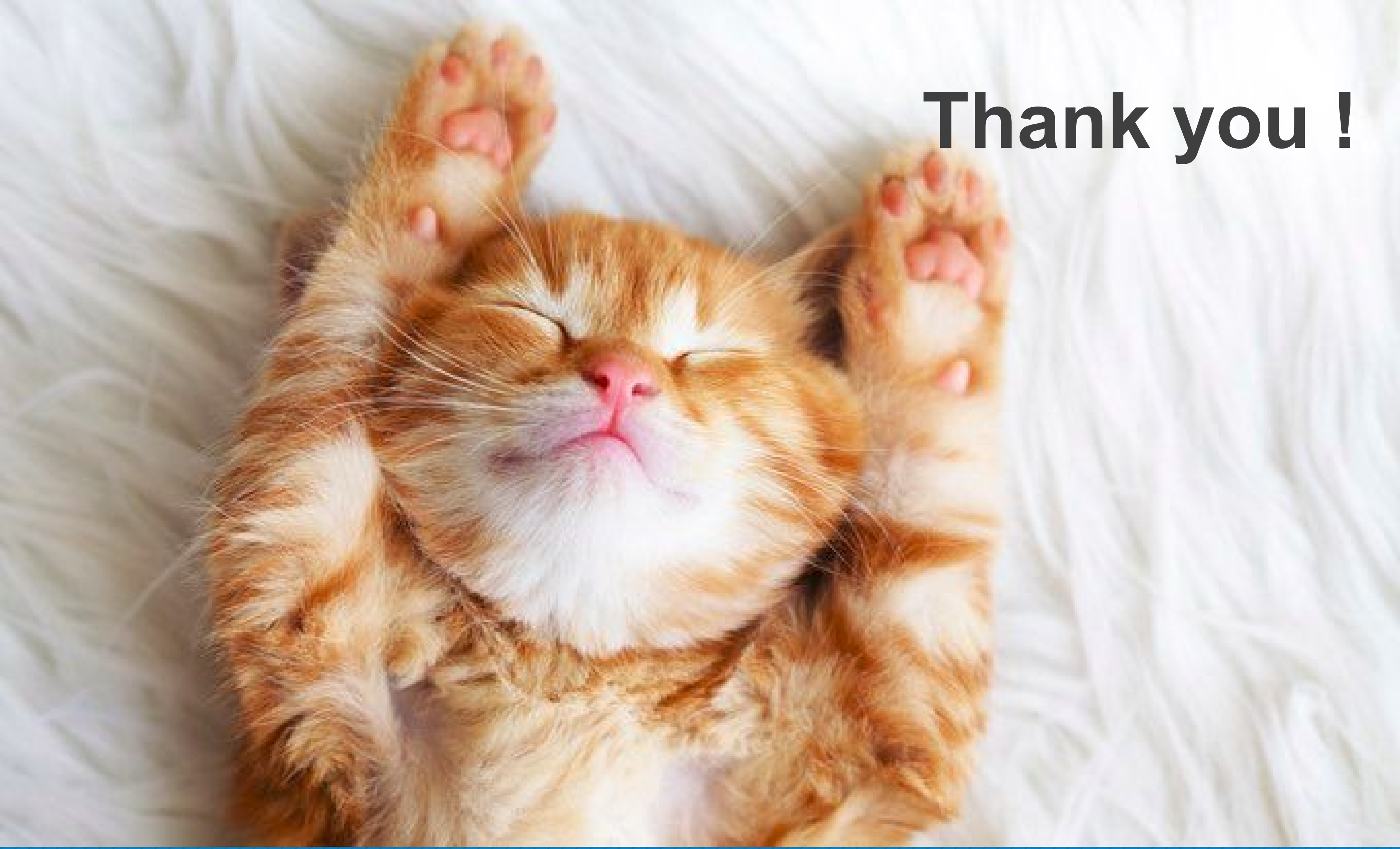
- He wants to find a house in Beverly Hills.
- House should have more than 5 bedrooms
- It shouldn’t be too old
- It should have a pool

14 houses matched the criteria



A cheaper house few blocks away!





Thank you !

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

References

http://cattime.com/cat-facts/kittens

https://docs.google.com/document/d/1Z-LBxAGf25yRocLuJrmZ-Gi3XIFFgN2b1W9tr3zbxJQ/edit#bookmark=id.liqixfol2pm