ANLY-501/COSC-587 Project report Fall 2017

Analysis of Real Estate Data, Regional Economic and Demographic Data

Team Project GitHub page:
https://github.com/WuZhuoran/ANLY-501-Data-Analyse-Project

Team members:

- Armaan Khullar <ak1643@georgetown.edu>
- Kornraphop Kawintiranon <kk1155@georgetown.edu>
- Shaobo Wang <sw1001@georgetown.edu>
- Zhuoran Wu <zw118@georgetown.edu>

1 Project overview

A house is often the largest and most expensive purchase a person makes in his or her lifetime. Ensuring homeowners have a trusted way to monitor this asset is incredibly important. Meanwhile, not only the houses feature such as total area, building quality and other features will influence the house price, but also that the overall economic and demographic situation will influence the house value.

In this project, we will concentrate on conducting a data science analysis on both real estate data and regional economy and demographic data. We will use the former for predictive analysis, i.e., predict the price of a property in a certain area. Meanwhile, we will use the latter for descriptive analysis, i.e., find how the economy and security have been in a certain city for the past 5 years.

2 Data science problem

For the subsequent project assignments, we will combine the two data sets. Then, our data science problem would be:

1) Is a certain city a better place for real estate investment based on its estimated price and its regional economic and demographic data? [1]
2) How to use House Information and Regional Economy and Security to predict House Price? What factors are significant for house price?

3 Data collection

3.1 API Source

- Open Data Network API https://www.opendatanetwork.com and API docs http://docs.odn.apiary.io/data-availability/find-all-available-data-for-some-entities

- Backup API for economic data https://www.bea.gov/API/bea_web_service_api_user_guide.htm

- Zillow API for house data https://www.zillow.com/howto/api/APIOverview.htm

[1] "Real Estate Investing: A Guide | Investopedia." 13 Feb. 2017, http://www.investopedia.com/mortgage/real-estate-investing-guide/. Accessed 6 Oct. 2017.

3.2 Data Source

In our experience, we consider that the value of a house will be determined by Internal and External factors. The Internal factors are the situation of house itself. The external factors are those economic and demographic situation around the house. We have collected the above two kinds of data from 2 sources.

We will collect the real estate data from Zillow, a famous online real estate database company. As for the economic and demographic data, there are multiple public data platforms. Currently, we choose Open Data Network as the major data source. As the project requires, the two data sets will include at least 20,000 examples and 12 attributes respectively.

First, we used Zillow API to collect their data. Zillow API provides various information about one property. We will collect house information from three perspectives: 1) House Location Information. 2) House Detail Information 3) House Quality Information. In this part, we will collect more than 40000 houses in the area of Los Angeles county.

Second, we used API from *opendatanetwork* website who provide public data for each area in many perspectives. For this API, we gathered 5 data including GDP per capita for each metropolitan, crime rate per 100k people for each city, crime count for each city, graduation rate for each city and earning information for each city. Each data contains lots of information such as graduation rate data is comprised of high school graduation rate, bachelor graduation rate, etc.

The major issue in our data is each data set contains a large number of missing values. Meanwhile, we can see there are many examples with suspicious values which might be impossible. We will conduct a series of evaluations to identify the problems and create specific methods to handle the issues.

3.3 Collected Data

3.3.1 Data Size

Data	**Data size**	**Attribute number**	**Record number**
Zillow	3.29 MB	19	40000
GDP	366 KB	6	5369

Crime rate	11.9 MB	14	66067
Crime count	4.81 MB	14	66070
Graduation rate	1.6 MB	9	24121
Earning information	3.2 MB	23	24121

3.3.2 Data attributes overview

3.3.2.1 Gross Domestic Product (GDP)

Attribute	**Type**	**Description**
area_id	String	ID for each area
area_name	String	Area name
area_type	String	There can be only region.msa, which means this area is in metropolitan level
year	Integer	Year number for each record
per_capita_gdp	Integer	GDP per capita of each area
per_capita_gdp_percent_change	Float	Percentage of GDP per capita change in a year of each area

3.3.2.2 Crime Rate

Attribute	**Type**	**Description**
area_id	String	ID for each area
area_name	String	Area name
area_type	String	There can be only region.place, which means this area is in city level
year	Integer	Year number for each record

Aggravated assault	Float	Aggravated assault incident rate per 100,000 people
All Crimes	Float	All crime incident rate per 100,000 people
Burglary	Float	Burglary incident rate per 100,000 people
Larceny	Float	Larceny incident rate per 100,000 people
Motor vehicle theft	Float	Motor vehicle theft incident rate per 100,000 people
Murder and nonnegligent manslaughter	Float	Motor vehicle theft incident rate per 100,000 people
Property crime	Float	Property crime incident rate per 100,000 people
Rape (revised definition)	Float	Rape incident rate per 100,000 people
Robbery	Float	Robbery incident rate per 100,000 people
Violent crime	Float	Violent crime incident rate per 100,000 people

3.3.2.4 Crime Count

Attribute	**Type**	**Description**
area_id	String	ID for each area
area_name	String	Area name
area_type	String	There can be only region.place, which means this area is in city level
year	Integer	Year number for each record

Aggravated assault	Integer	Aggravated assault incident count
All Crimes	Integer	All crime incident count
Burglary	Integer	Burglary incident count
Larceny	Integer	Larceny incident count
Motor vehicle theft	Integer	Motor vehicle theft incident count
Murder and nonnegligent manslaughter	Integer	Motor vehicle theft incident count
Property crime	Integer	Property crime incident count
Rape (revised definition)	Integer	Rape incident count
Robbery	Integer	Robbery incident count
Violent crime	Integer	Violent crime incident count

3.3.2.5 Graduation Rate

Attribute	**Type**	**Description**
area_id	String	ID for each area
area_name	String	Area name
area_type	String	There can be only region.place, which means this area is in city level
year	Integer	Year number for each record
percent_associates_degree	Float	Percentage of people who graduated associates degree as highest education
percent_bachelors_degree_or_higher	Float	Percentage of people who graduated bachelor degree or higher as highest education
percent_graduate_or_professional_degree	Float	Percentage of people who graduated graduate or

		professional degree as highest education
percent_high_school_graduate_or_higher	Float	Percentage of people who graduated high school level or higher as highest education
percent_less_than_9th_grade	Float	Percentage of people who graduated in less than 9th grade level

3.3.2.6 Earning Information

Attribute	**Type**	**Description**
area_id	String	ID for each area
area_name	String	Area name
area_type	String	There can be only region.place, which means this area is in city level
year	Integer	Year number for each record
female_full_time_median_earnings	Integer	The median earnings of only female people who work full time
female_median_earnings	Integer	The median earnings of female people
male_full_time_median_earnings	Integer	The median earnings of only male people who work full time
male_median_earnings	Integer	The median earnings of male people
median_earnings	Integer	The median earnings of each area
median_earnings_bachelor_degree	Integer	The median earnings of people who graduated bachelor degree as highest

		education
median_earnings_graduate_or_professional_degree	Integer	The median earnings of people who graduated graduate or professional degree as highest education
median_earnings_high_school	Integer	The median earnings of people who graduated high school level as highest education
median_earnings_less_than_high_school	Integer	The median earnings of people who did not graduate high school level
median_earnings_some_college_or_associates	Integer	The median earnings of people who graduated from colleges or associates as highest education
percent_with_earnings_10000_to_14999	Float	Percentage of people who earn $10,000 to $14,999 per year
percent_with_earnings_15000_to_24999	Float	Percentage of people who earn $15,000 to $24,999 per year
percent_with_earnings_1_to_9999	Float	Percentage of people who earn $1 to $9,999 per year
percent_with_earnings_25000_to_34999	Float	Percentage of people who earn $25,000 to $34,999 per year
percent_with_earnings_35000_to_49999	Float	Percentage of people who earn $35,000 to $49,999 per year
percent_with_earnings_50000_to_64999	Float	Percentage of people who earn $50,000 to $64,999 per year
percent_with_earnings_65000_to_74999	Float	Percentage of people who earn $65,000 to $74,999 per year
percent_with_earnings_7500	Float	Percentage of people who

0_to_99999		earn $75,000 to $99,999 per year
percent_with_earnings_over_100000	Float	Percentage of people who earn more than $100,000 per year

3.3.2.8 Zillow House Info

Attribute	**Type**	**Description**
zpid	Integer	Unique ID for each house
latitude	Integer (by 10e6)	Latitude of the location of this house.
longitude	Integer (by 10e6)	Longitude of the location of this house.
cityid	Integer	City ID in which the house is located
countyid	Integer	County ID in which the house is located
zipcode	Integer	Zip Code in which the house is located
amount	Integer	House price

3.3.2.9 Zillow House Detail

Attribute	**Type**	**Description**
zpid	Integer	Unique ID for each house
bathrooms	Float	The number of bathrooms
bedrooms	Integer	The number of bedrooms
fullbathrooms	Integer	The number of bathrooms

		with shower.
finishedSqFt	Integer	The area of house in Square Feet
lotsizeSqFt	Integer	The area of land in Square Feet
yearbuilt	Integer	The build year of this house
latitude	Integer (by 10e6)	Latitude of the location of this house.
longitude	Integer (by 10e6)	Longitude of the location of this house.
amount	Integer	The house price of house.

3.3.2.10 Zillow House Updated

Attribute	**Type**	**Description**
zpid	Integer	Unique ID for each house
rooms	Integer	The number of rooms in the house
units	Integer	Number of units the structure is built in
airconditiontype	Integer	The type id of air condition system in the house.
heatsystemtype	Integer	The type id of heating system in the house
buildingquality	Integer	The quality of the house
pools	Integer	The number of pool in the house

3.3.2.11 Heating System Type Dictionary

Heating System Type ID	Heating System Type Description
1	Baseboard
2	Central
3	Coal
4	Convection
5	Electric
6	Forced air
7	Floor/Wall
8	Gas
9	Geo Thermal
10	Gravity
11	Heat Pump
12	Hot Water
13	None
14	Other
15	Oil
16	Partial
17	Propane
18	Radiant
19	Steam
20	Solar
21	Space/Suspended
22	Vent

23	Wood Burning
24	Yes
25	Zone

3.3.2.12 Air Condition Type Dictionary

Air Condition Type ID	Air Condition Type Description
1	Central
2	Chilled Water
3	Evaporative Cooler
4	Geo Thermal
5	None
6	Other
7	Packaged AC Unit
8	Partial
9	Refrigeration
10	Ventilation
11	Wall Unit
12	Window Unit
13	Yes

4 Data cleanliness evaluation

Define:
Clean data fraction = 1 - Missing value fraction - Noise value fraction
Data quality score = sum(Clean data fraction) / number of columns

4.1 Zillow data:

Use combined zillow data (join tables on zpid)
data size: rows: 40000, columns: 19
Data quality score: 0.831468421053
Three attributes with the worst data quality: pools, airconditiontype, units

Attributes	**Missing values fraction**	**Noise values fraction**	**Clean values fraction**
zpid	0	0	1
bathrooms	0	0.05725	0.94275
bedrooms	0	0	1
fullbathrooms	0.033425	0.023925	0.94265
finishedSqFt	0.012225	0.04795	0.939825
lotsizeSqFt	0.068575	0.1353	0.796125
yearbuilt	0.013775	0.003575	0.98265
latitude	0	0	1
longitude	0	0	1
cityid	0.017	0.033625	0.949375
countyid	0	0	1
zipcpde	0.000675	0.0001	0.999225
rooms	0	0	1
units	0.314425	0.071425	0.61415
airconditiontype	0.73005	0.025525	0.244425
heatsystemtype	0.377375	0.0067	0.615925
buildingquality	0.327925	0.001275	0.6708
pools	0.82965	0	0.17035
amount	0.00785	0.0625	0.92965

4.2 Economic and demographic data:

4.2.1 gdp_info:

Data size: 5369 rows, 6 columns
Data quality score: 0.989693922

Attributes	Missing frac	Noise frac	Clean frac
area_id	0	0	1
area_name	0	0	1
area_type	0	0	1
year	0	0	1
per_capita_gdp	0	0	1
per_capita_gdp_percent_change	0	0.061836	0.938164

4.2.2 crime rates:

Data size: 66067 rows, 14 columns
Data quality score: 0.893945324

Attributes	Missing frac	Noise frac	Clean frac
area_id	0	0	1
area_name	0	0	1
area_type	0	0	1
year	0	0	1
Aggravated assault	0.001771	0.063375	0.934854
All Crimes	1.51E-05	0.035237	0.964748
Burglary	0.000893	0.048859	0.950247
Larceny	0.000938	0.036433	0.962629

Motor vehicle theft	0.000348	0.069324	0.930328
Murder and nonnegligent manslaughter	0	0.214631	0.785369
Property crime	0.001983	0.034904	0.963113
Rape (revised definition)	0.796843	0.012291	0.190867
Robbery	6.05E-05	0.083294	0.916645
Violent crime	0.023658	0.059909	0.916433

4.2.3 crime counts:

Data size: 66070 rows, 14 columns
Data quality score: 0.861089969513

Attributes	**Missing frac**	**Noise frac**	**Clean frac**
area_id	0	0	1
area_name	0	0	1
area_type	0	0	1
year	0	0	1
Aggravated assault	0.001771	0.131951	0.866278
All Crimes	1.51E-05	0.123702	0.876283
Burglary	0.000893	0.129393	0.869714
Larceny	0.000938	0.119192	0.87987
Motor vehicle theft	0.000348	0.143469	0.856183
Murder and nonnegligent manslaughter	0	0	1
Property crime	0.001998	0.12234	0.875662
Rape (revised	0.845073	0.01839	0.136537

definition)			
Robbery	6.05E-05	0.149584	0.850356
Violent crime	0.023702	0.131921	0.844377

4.2.4 graduation rates:

Data size: 24121 rows, 9 columns
Data quality score: 0.97002611832

Attributes	**Missing frac**	**Noise frac**	**Clean frac**
area_id	0	0	1
area_name	0	0	1
area_type	0	0	1
year	0	0	1
percent_associates_degree	0	0.031839	0.968161
percent_bachelors_degree_or_higher	0	0.059409	0.940591
percent_graduate_or_professional_degree	0	0.075577	0.924423
percent_high_school_graduate_or_higher	0	0.037146	0.962854
percent_less_than_9th_grade	0	0.065793	0.934207

4.2.5 earning info:

Data size: 24121 rows, 23 columns
Data quality score: 0.983815294

Attributes	**Missing frac**	**Noise frac**	**Clean frac**

area_id	0	0	1
area_name	0	0	1
area_type	0	0	1
year	0	0	1
female_full_time_median_earnings	0	0	1
female_median_earnings	0.00E+00	0	1
male_full_time_median_earnings	0	0	1
male_median_earnings	0	0	1
median_earnings	0	0	1
median_earnings_bachelor_degree	0	0	1
median_earnings_graduate_or_professional_degree	0	0	1
median_earnings_high_school	0	0	1
median_earnings_less_than_high_school	0.00E+00	0	1
median_earnings_some_college_or_associates	0	0	1
percent_with_earnings_10000_to_14999	0	0.056217	0.943783
percent_with_earnings_15000_to_24999	0	0.032213	0.967787
percent_with_earnings_1_to_9999	0	0.064674	0.935326
percent_with_earning	0	0.02815	0.97185

s_25000_to_34999			
percent_with_earnings_35000_to_49999	0	0.031881	0.968119
percent_with_earnings_50000_to_64999	0	0.029974	0.970026
percent_with_earnings_65000_to_74999	0	0.025828	0.974172
percent_with_earnings_75000_to_99999	0	0.021268	0.978732
percent_with_earnings_over_100000	0	0.082045	0.917955

5 Data cleaning

5.1 Data editing strategies

5.1.1 Deletion:

5.1.1.1 Zillow data:

For the Zillow data, we delete rows where the year is under 2000 and is above 2017. In addition, we also remove rows where the values of bathrooms, bedrooms and full bathrooms are negative. In addition, we delete any rows where the finished square foot area is greater than the size of the lot. This is because the finished area is the area of the building, while the lot size also includes the land. Furthermore, we also drop all rows where the amount of the house is less than $10,000 as it seems suspiciously cheap and highly likely to be incorrect. Likewise, we also remove rows where the amount of the house is greater than $50 million. Although $50 million itself seems pretty high, it is possible that it is a mansion or belongs to a celebrity. Anything higher than $50 million would be considered as noisy data. After making these changes, we remove any rows that are missing any values.

5.1.1.2 GDP Data:

For GDP data, we delete rows that have the year below 2000 or above 2017. In addition, if the per capita gdp for a row is below $2000, which would imply that the average inhabitant earns less than $2000, then this would be considered noisy data and would need to be deleted.

5.1.1.3 Graduation Data:

For graduation data, we delete all rows that are missing any observations. After that, we delete rows where the year is below 2000 and over 2017. From there, we drop any rows that have negative values for each percentage category (pertaining to academic credentials). Also, we remove any rows where the sum of all the percentage values within the row do not sum up to 100.

5.1.1.4 Crime Data:

For crime data, we delete any rows where the year is below 2000 or above 2017. For crime_counts, we drop any rows where the sum of all the crimes does not add up to “All Crimes”. If the value of “All Crimes” is negative, we delete the row. We delete any row where a crime value is negative. For the crime_rate data, we allow the sum of the rates of all crimes to differ from “All Crimes” by a value of 2000. This is because not all crimes may have been categorized in the data; another reason is that the data is in terms of floating-point values, and the sum may not add up perfectly with “All Crimes”.

5.1.1.5 Earning Info Data:

For the data on earnings, we delete all rows that are missing observations. We also drop all rows with the year below 2000 or above 2017. In addition, we also drop all rows where the percentage of each category (for each row) does not add up to 100. In addition, we delete any row that contains negative values as it would not make any sense.

5.1.2 Imputation:

5.1.2.1 Zillow data:

1) Delete the "rooms" column (Because they are all 0).
2) Find the rows in which"fullbathrooms" is larger than "bathrooms", and set the “fullbathrooms” value as the “bathrooms” value. This is because the number of full bathrooms of a house is a subset of the number of bathrooms a house has. As a result, we are correcting the wrong number.
3) The missing value of “pools” value will be set to 0. Many houses don’t have a pool, so it is likely that this feature was forgotten about. Conversely, if a house has a pool then it is considered to be a very fine addition to the house, and it is very unlikely that this feature would go ignored.
5) For rows that have a missing “buildingquality”, we will find the mode of all the “buildingquality” values and then set it as the value for the missing “buildingquality”.

6) Heating System and Air condition could also be the mode number among all the houses, because there is a greater likelihood that the missing value could be that of the majority.

5.1.2.3 GDP Data:

There is no imputation necessary for the GDP data.

5.1.2.4 Graduation Data:

For the graduation data, we split “area_name” into two columns: “City” and “State”.

5.1.2.5 Crime Data:

For the crime data, we split “area_name” into two columns: “City” and “State”.

5.1.2.6 Earning Info Data:

For the crime data, we split “area_name” into two columns: “City” and “State”.

5.1.3 Leave unchanged

Leave a data value unchanged if it doesn’t fit the rules for deletion and imputation.

5.2 Results comparison

We only compare the results for Zillow data because Eco and Security data is relatively clean and good to use.

Cleaned zillow data size: 25725 rows, 18 rows (attribute ‘rooms’ is removed since it is a sparse attribute)

New data quality score: 0.95738955407 much better than old score (0.83)

Attributes	**Missing frac**	**Noise frac**	**Clean frac**
zpid	0	0	1
bathrooms	0	0.026668	0.973332
bedrooms	0	0.009991	0.990009

fullbathrooms	0	0.026668	0.973332
finishedSqFt	0	0.050809	0.949191
lotsizeSqFt	0	0.158374	0.841626
latitude	0	0.055901	0.944099
longitude	0	0.00241	0.99759
yearbuilt	0	0.001322	0.998678
amount	0	0.068574	0.931426
cityid	0	0.050692	0.949308
countryid	0	0	1
zipcpde	0	0.053802	0.946198
units	0	0.088206	0.911794
airconditiontype	0	0	1
heatsystemtype	0	0.000972	0.999028
buildingquality	0	0	1
pools	0	0.172601	0.827399

6 Basic Statistical Analysis and data cleaning insight & Additional LOF Part

6.1 Statistical description

We will study the statistics of crime_counts data. The data set is the crime counts in each city for 9 years (2006 - 2014). We can tell that, among all crimes, property crime is the most frequently happened crime in the entire nation.

	Aggravated assault	All Crimes	Burglary	Larceny	Motor vehicle theft	Murder and nonnegligent	Property crime	Rape (revised definition)	Robbery	Violent crime

						manslaughter				
count	65953	66069	66011	66008	66047	66070	65938	10236	66066	64504
mean	69.425667	1818.647505	171.0295	544.238971	77.750768	1.392357	791.515075	8.590563	40.0051	116.381775
std	518.372369	9694.955163	858.9081	2683.71527	573.714116	12.868114	3996.952944	48.086239	405.269	907.024846
min	0	0	0	0	0	0	0	0	0	0
25%	2	92	8	29	1	0	42	0	0	3
50%	8	307	27	103	5	0	140	1	1	12
75%	30	1080	94	363	23	0	490	5	9	46
max	31767	411044	29279	117931	25389	596	153436	2190	23511	52993

6.2 Clean outliers in Zillow data

We have delivered an detailed explanation on how to clean the data in project assignment 1 along with python code. In this section, we only focus on dealing with potential outliers in zillow data.

Attributes with outliers in zillow data:
bathrooms/fullbathrooms: max == 14
finishedSqFt: max == 12000
lotsizeSqft: max == 3500000
amount: max == 4 * 10^7

We apply the code for evaluation in assignment 1, using box plots, to detect the outliers. We consider the largest few points out of interquartile range (IQR) as the outliers. In our case, there are only a few outliers in each attribute. So we simply use Excel to sort the data according to each attribute above. Then we remove the top 5 to 10 largest records.

6.3 Binning data

The binning strategy we chose is equal-width, because we care more about the frequency distribution in each attribute. In this section, we choose to bin the median_earnings from earning_info data to see which earning interval contains the largest number of cities. As it shows on the histograms, we partition the data into 10 bins with equal width. We can tell most cities have median earning in between 20k to 40k. For the purpose of consistency, we output the column that includes which bin each example belongs to in a separated .csv file (median_earnings_bin.csv), instead of adding a new column in our original data.

6.4 Local Outlier Factor

We apply LOF algorithm to find outliers in our data for each dataset. However, this algorithm's result depends on parameter *k*, which is number of neighbors that will be considered (We will not explain how the algorithm works because it is not in the scope of this project). After we try different *k* including 5, 20 and 50, the results of outlier detection using different *k*, have same outlier numbers for same dataset. After investigating more, we have multidimensional outliers the outliers. Moreover, outlier numbers are same but the data that are considered as outliers are changed. For example, when using k=5, assume that outliers are [Yes, No, No], which means percentage of outliers is 33.33%. After we change k=20, the percentage is still same at 33.33% but the outliers are change to be the second data point instead [No, Yes, No]. The result summary is shown as following table.

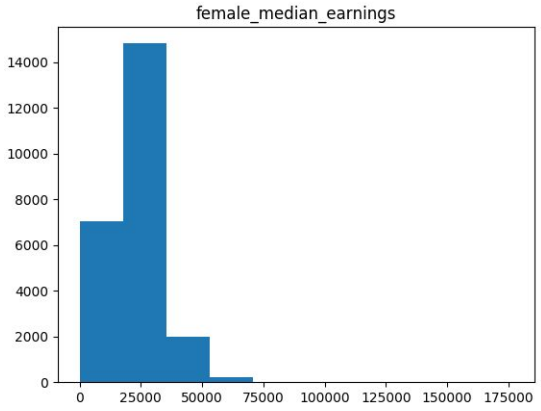
Dataset	k-neighbors	Total	Outliers	Non-outliers	% outlier
Zillow	5	25719	2572	23147	11.11%

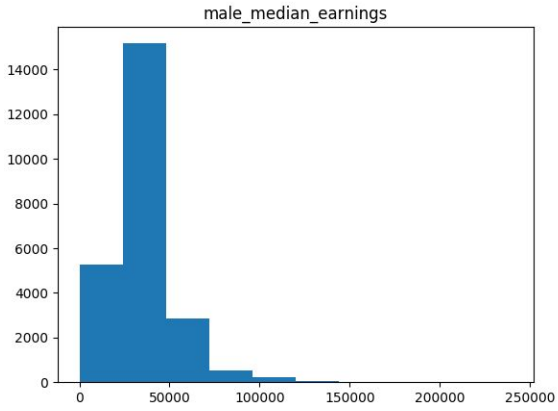
	20	25719	2572	23147	11.11%
	50	25719	2572	23147	11.11%
graduation_rates	5	322	33	289	11.42%
	20	322	33	289	11.42%
	50	322	33	289	11.42%
gdp_info	5	5369	537	4832	11.11%
	20	5369	537	4832	11.11%
	50	5369	537	4832	11.11%
earning_info	5	2823	283	2540	11.14%
	20	2823	283	2540	11.14%
	50	2823	283	2540	11.14%
crime_rates	5	13378	1338	12040	11.11%
	20	13378	1338	12040	11.11%
	50	13378	1338	12040	11.11%
crime_counts	5	1394	140	1254	11.16%
	20	1394	140	1254	11.16%
	50	1394	140	1254	11.16%

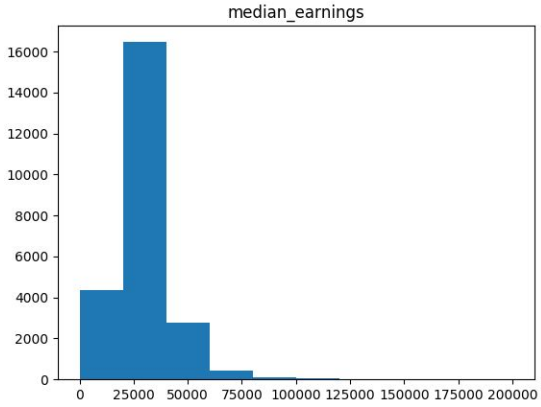
7 Histograms and Correlations

7.1 Histograms

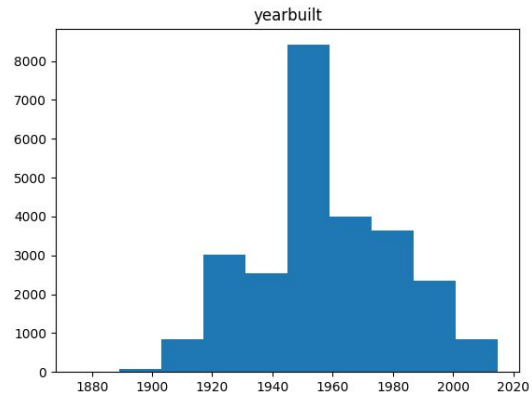
We plot the histograms of female_median_earnings, male_median_earnings and median_earnings from earning_info data (equal-width, 10 bins). We can see female median earning is less than male median earning. Meanwhile, earnings from most people are less than 50k.

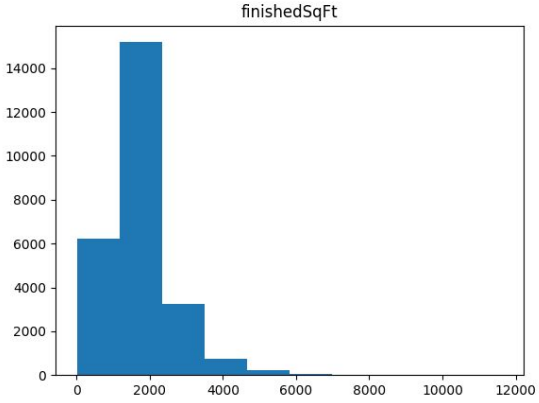


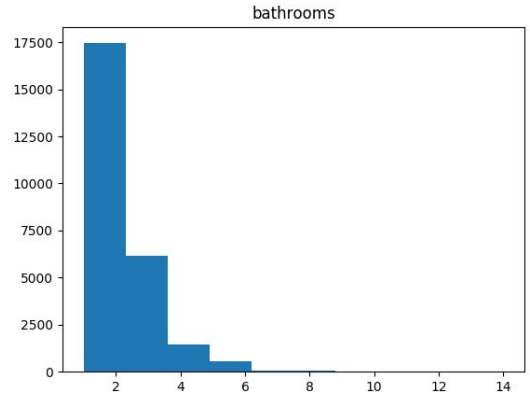




We plot the histograms of yearbuilt, finishedSqFt and bathrooms from zillow data (equal-width, 10 bins). It is easy to tell what kind of houses take the largest portion of all sales (year built: 1950 - 1960, area: 1000 - 2000 sqft, bathrooms: 1 - 3).



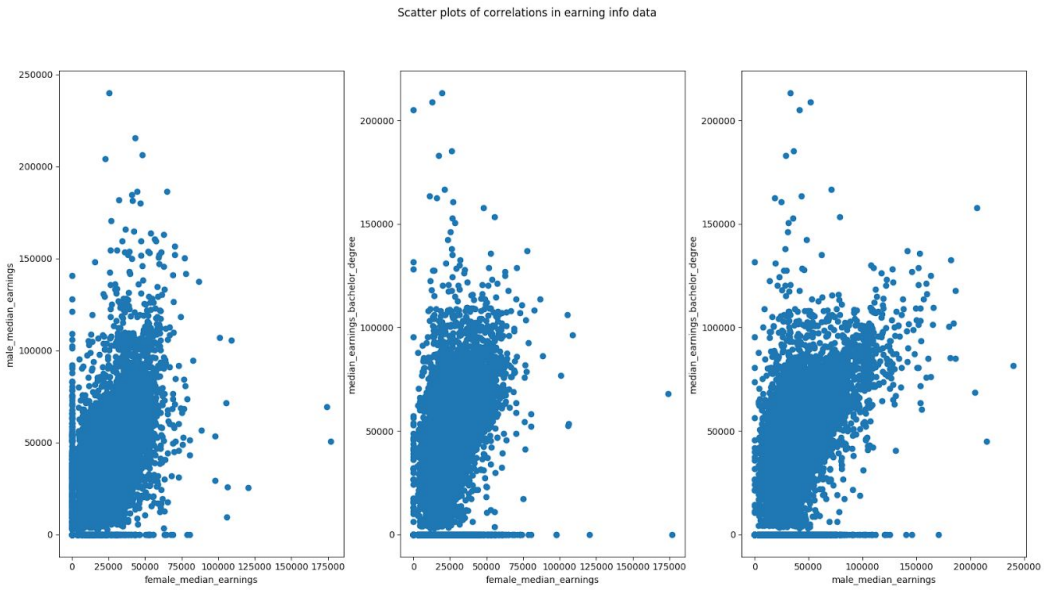




7.2 Correlations

For eco data, we choose female_median_earnings, male_median_earnings and median_earnings_bachelor_degree from earning_info data.

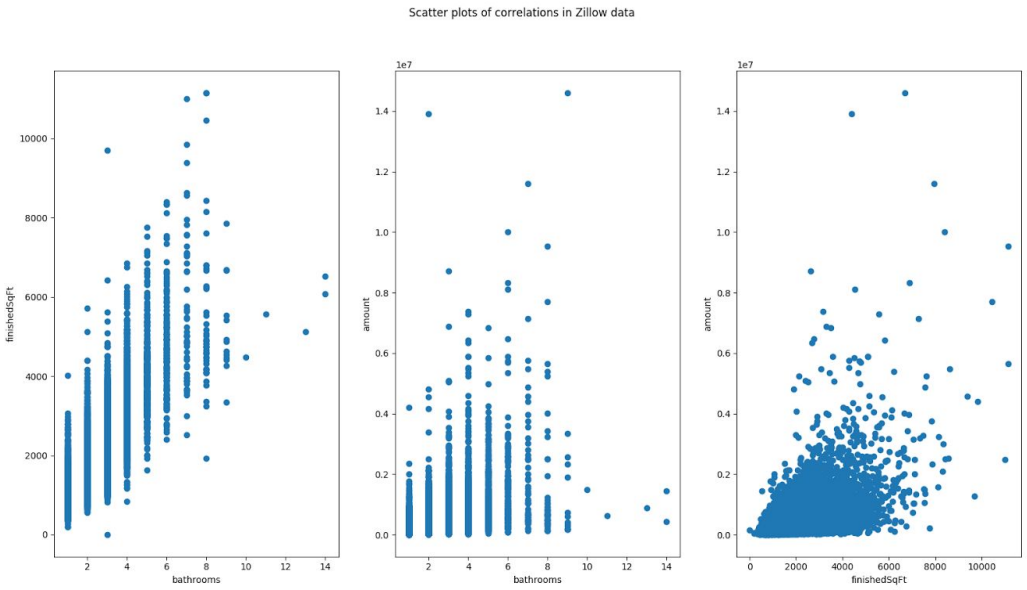
	female_median_earnings	male_median_earnings	median_earnings_bachelor_degree
female_median_earnings	1	0.607516	0.528004
male_median_earnings	0.607516	1	0.546437
median_earnings_bachelor_degree	0.528004	0.546437	1



All pairs are approximately positively correlated. Generally, higher median earnings means the city is well developed and the economy is better than others. Some points out the central group possibly means that the major labor force in those cities is either male or female, not both (i.e. male_median_earnings = 200000, female_median_earnings = 25000 possibly means the major labor force is male in this city).

For zillow data, we choose bathrooms, amount and finishedSqFt. The correlations are as below:

	bathrooms	finishedSqFt	amount
bathrooms	1	0.789838	0.372687
finishedSqFt	0.789838	1	0.496968
amount	0.372687	0.496968	1



We can see the first pair (bathrooms / finishedSqFt) is approximately positively correlated, while the other two are not obvious. It's easy to infer that large houses probably have more bathrooms. However, from the other two histograms, the amount of a house is not only determined by its size, although size might be an important reference. We think that location, security and economical environment are also important to evaluate the house price.

8 Cluster Analysis

With regards to the cluster analysis, we conducted K-means, hierarchical, and DBSCAN clustering to three different sources of data.

In particular, we applied K-means clustering to "Zillow_Cleaned.csv", which contains detailed information about each home, and performed our clustered based on attributes ranging from "bathrooms" and "bedrooms", to "State". Since the dimensionality of our data was high, we decided to plot a PCA projection of the clusters for different values of K. Initially, we plotted the clusters for a few values of K and noticed that K = 4 provided clusters that are well-separated and non-overlapping. To confirm our hypothesis, we decided to do K-means with K taking on values 3, 4, 5, 6, 7, 8, 9, and 10, and searched for the value of K that provided the highest silhouette score. We found that K = 4 provided a silhouette score of 0.545, which was the highest. This meant that items within a cluster are more similar to other items within its cluster and more dissimilar to items in other clusters.

Furthermore, we applied Hierarchical Clustering to "graduation_rates_CLEANED.csv", which contains graduation rates for a given city and state. We applied the "Ward" algorithm and plotted a dendrogram showing the nested clusters. According to the dendrogram, the number of clusters is 4, and were able to proceed to fft the hierarchical clustering to the graduation

dataset. Given the high dimensionality of the data, we plotted a PCA projection of the clusters to 2-D space. After plotting the clusters, we noticed that the clusters are not as well-separated and dense as we would have liked. This might be due to the high-dimensionality of the data. The silhouette score is 0.4667, which is low and indicates that the clusters are not as well-separated and dense.

Lastly, we applied DBSCAN to “crime_counts_CLEANED.csv”, which contains detailed information about the crime rates by city and state. From plotting the clusters, we can see that the clusters are well-separated, but not as dense. We see that there are two clusters and that one cluster is very dense and the other is not very dense. This may be because there might be some states that have cities with both high and low crime. This could explain why the clusters in the top of the graph are not very dense. We also find that the silhouette average is 0.7373, which indicates that the clusters are well-separated and dense, which indicates the clusters are well-separated and dense. For our DBSCAN clustering, we had the value of eps, which is the radius for points to be part of a neighborhood, to be 0.8; we had the value of 30 be the minimum number of elements for a neighborhood.

9 Association Rules / Frequent Itemset Mining Analysis

From the instruction, we follow it by using apriori algorithm in sample of datasets. Since our data are very large and have many columns so the algorithm run very slow. Therefore, we will apply the algorithm to find association rules of only sample of each dataset (3,000 itemsets as maximum). Additionally, we do binning in some columns to make it workable in association rules algorithm.

We try 3 different support levels [0.2, 0.5, 0.8] and 3 different confidence levels [0.2, 0.5, 0.8]. Then the combination of these experiments will have 9 different styles. In addition, we also consider only rules comprised of more than 2 items. We run our experiment on each dataset separately to find the patterns (rules) of each dataset. The result summary of this experiment is shown as following table. (We will show only using min_conf=0.2 since it is the minimum confidence level we used in this experiment)

Dataset	min_sup	Number of rule sets	Most frequent itemset	support
Zillow	0.2	659	{'lotsizeSqFt_Mean', 'units_1.0', 'pools_0.0'}	0.699
	0.5	11		
	0.8	0		

graduation_rates	0.2	16	{'percent_associates_degree_Mean', 'percent_graduate_or_professional_degree_Mean', 'percent_bachelors_degree_or_higher_Mean'}	0.786
	0.5	16		
	0.8	0		
gdp_info	0.2	0		
	0.5	0		
	0.8	0		
earning_info	0.2	36	{'male_full_time_median_earnings_Mean', 'male_median_earnings_Mean', 'median_earnings_Mean'}	0.270
	0.5	0		
	0.8	0		
crime_rates	0.2	3128	{'All Crimes_Mean', 'Property crime_Mean', 'Larceny_Mean'}	1.0
	0.5	526		
	0.8	198		
crime_counts	0.2	1470	{'All Crimes_Mean', 'Property crime_Mean', 'Larceny_Mean'}	0.885
	0.5	466		
	0.8	54		

From this result table, we have shown the most frequent itemset for each dataset. Also, we found that gdp_info has no any association rule. This is not surprising because in the dataset, there is a year column varying from 2001 to 2013. This makes it is really hard to get higher support score. If we set min_sup down, we could get some rules for this.

After we investigate more detail in the result, we found many facts from the rules. For example, in Zillow data, the most frequent itemset is {'lotsizeSqFt_Mean', 'units_1.0', 'pools_0.0'} and the most confident rule of this itemset is {'pools_0.0', 'units_1.0'} ---> {'lotsizeSqFt_Mean'} with confident score 0.95. This shows the unsurprising fact that houses with 1 unit and no pool will have only moderate size. Moreover, we found the additional rule from the previous example, {'pools_0.0', 'units_1.0', 'yearbuilt_Mean'} ---> {'lotsizeSqFt_Mean'}. This also means the same thing as previous but its confidence score is higher to 0.99 because of more information about build year of houses.

10 Hypothesis Testing

10.1 Hypotheses 0:

Hypotheses 0: The true mean of house value in CA is larger than 253700.5 dollars
Alternative Hypotheses 0: The true mean of house value is CA is smaller than 253700.5

We will apply **T-Test** to analyse this hypothesis.

Ttest_indResult(statistic=-9.5375443666596578, pvalue=1.5875360149285238e-21)

According to the result of T-Test, we could not accept the Hypotheses 0.

10.2 Hypotheses 1:

Hypotheses 1: The houses built before 1970 and after 1970 have no difference in house money.
Alternative Hypotheses 1: The houses built before 1970 and after 1970 do have difference in house value.

We will apply **T-Test** to analyse this hypothesis. Here is the result:

T-Test on analyses the mean about houses built before and after 1960
Ttest_indResult(statistic=-0.41120619462865038, pvalue=0.68096503148403242)

According to the result of T-Test, we could accept the hypotheses 1. A house will not be more expensive than the house which are built after than 1960.

10.3 Hypotheses 2:

Hypotheses 2: The attribute “finishedSqFt” (which means the total living area for a house) has the most coefficient relationship with house value;
Alternative Hypotheses 2: There are other attributes which has more coefficient relationship.

We will apply **OLS Regression** to analyse this hypothesis. Here is the result:

OLS Regression Result						
Dep. Variable	amount			R-squared	0.317	
Model	OLS			Adj. R-squared	0.317	
Method	Least Squares			F-statistic	852.1	
Date	Sun, 05 Nov 2017			Prob (F-statistic)	0.00	
Time	12:19:38			Log-Likelihoo	-3.7100e+05	
No. Observations	25719			AIC	7.420e+05	
Df Residuals	25704			BIC	7.422e+05	
Df Model	14					
Covariance Type	nonrobust					
	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.738e+07	1.71e+06	-10.173	0.000	-2.07e+07	-1.4e+07
bathrooms	2.502e+04	2596.869	9.635	0.000	1.99e+04	3.01e+04
bedrooms	-9.912e+04	3637.397	-27.250	0.000	-1.06e+05	-9.2e+04
fullbathrooms	2.502e+04	2596.869	9.635	0.000	1.99e+04	3.01e+04
finishedSqFt	392.4646	5.889	66.640	0.000	380.921	404.008
lotsizeSqFt	-0.0109	0.031	-0.350	0.726	-0.072	0.050
latitude	-2.192e+05	1.56e+04	-14.089	0.000	-2.5e+05	-1.89e+05

longitude	-2.225e+05	1.39e+04	-16.015	0.000	-2.5e+05	-1.95e+05
yearbuilt	-731.9847	162.457	-4.506	0.000	-1050.410	-413.560
cityid	-0.0491	0.048	-1.024	0.306	-0.143	0.045
zipcpde	-0.9639	0.847	-1.137	0.255	-2.625	0.697
units	-1.499e+05	7017.608	-21.362	0.000	-1.64e+05	-1.36e+05
airconditiontype	-9436.0336	8060.058	-1.171	0.242	-2.52e+04	6362.135
heatsystemtype	-2590.7922	1521.789	-1.702	0.089	-5573.585	392.000
buildingquality	2.513e+04	1897.663	13.241	0.000	2.14e+04	2.88e+04
pools	6.26e+04	7865.077	7.959	0.000	4.72e+04	7.8e+04
Omnibus	71325.460			Durbin-Watson	1.790	
Prob(Omnibus)	0.000			Jarque-Bera (JB)	9281156011.070	
Skew	34.961			Prob(JB)	0.00	
Kurtosis	2945.097			Cond. No.	8.10e+18	

According to the result OLS, we found that, the ‘FinishedSqFt’ has significant coefficient with the house value, which is obviously true. The most important of a house is its area.

10.4 Hypotheses 3:

Hypotheses 3: The houses with more bedrooms and bathrooms will have a larger lot area.
Alternative Hypotheses 3: There is no obvious relationship with these two attributes.

We will apply **Random Forest Regression** to analyse this hypothesis. Here is the result:

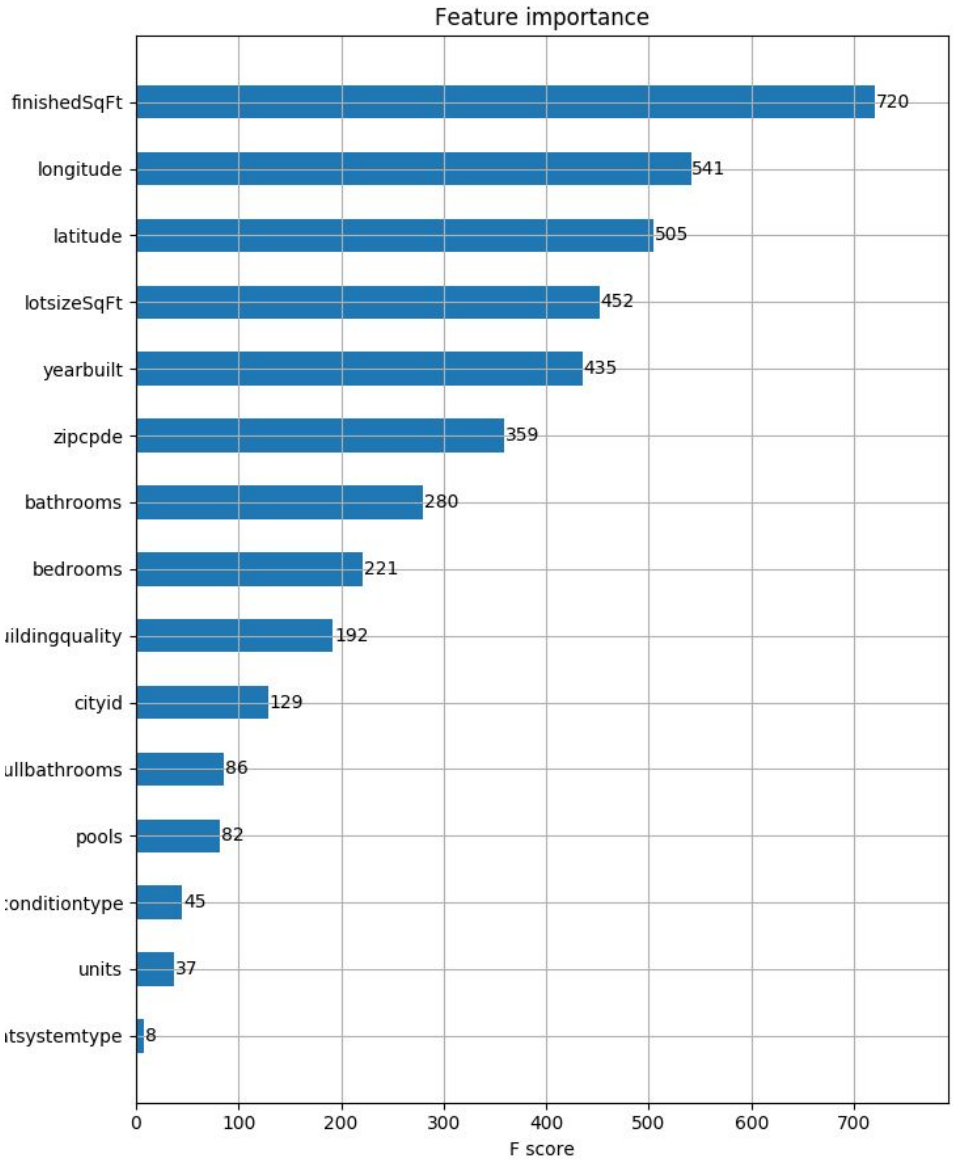
RandomForestRegressor Result Scores: 0.0149353264259

Additionally, We apply the **XGBoost** and **Support Vector Regressor** to predict the house value based on the house conditions. Consider it is not a classification problem, it is a regression problem. So we apply RMSE(Root Mean Square Error) to evaluate our result. Furthermore, we apply 3-fold and 10-fold Cross Validation to get the result:

Build Support Vector Regression and Linear Regression to predict house value
Support Vector Regressor Result Scores: -0.0602669961327
Linear Regressor Result Scores: 0.395702635318

[0] train-rmse:639648 test-rmse:634563
[50] train-rmse:302668 test-rmse:396918
[100] train-rmse:258518 test-rmse:391102
[150] train-rmse:242604 test-rmse:390274
[200] train-rmse:231823 test-rmse:389576

Also, we draw our Feature-Importance for every attributes. According to Feature Importance, we could see that, ‘FinishedSqFt’ is really important for a house, which is right with our previous conclusion.


Feature importance
finishedSqFt
longitude
latitude
lotsizeSqFt
yearbuilt
zipcpde
bathrooms
bedrooms
uildingquality
cityid
ullbathrooms
pools
:onditiontype
units
itsystemtype
720
541
505
452
435
359
280
221
192
129
86
82
45
37
8
0
100
200
300
400
500
600
700
F score