ANLY-501/COSC-587 Project report Fall 2017

Analysis of Real Estate, 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>

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.

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]

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

2) How to use House Information and Regional Economy and Security to predict House Price.

Data collection

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

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.

Collected Data

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

Data attributes overview

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

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

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

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

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_75000_to_99999	Float	Percentage of people who 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

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

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.

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

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

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

Data cleanliness evaluation

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

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

Economic and demographic data:

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

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

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 definition)	0.845073	0.01839	0.136537
Robbery	6.05E-05	0.149584	0.850356
Violent crime	0.023702	0.131921	0.844377

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

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_les	0.00E+00	0	1

s_than_high_school			
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_earnings_25000_to_34999	0	0.02815	0.97185
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

Data cleaning

Data editing strategies

Deletion:

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.

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.

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.

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”.

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.

Imputation:

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.

GDP Data:

There is no imputation necessary for the GDP data.

Graduation Data:

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

Crime Data:

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

Earning Info Data:

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

Leave unchanged

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

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

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