Metadata and Data Quality Insights

Data Quality Insight

Cost and Revenue datasets are handled separately in an attempt to enrich the data quality for exploratory data analysis.

	Zillow	Airbnb
observations	8946	48895
dimensions	262	106

Data Quality Check

In Zillow data, Except price data in early years, only `Metro` contains missing value. It's an irrelevant column for analysis so let it be.

column_name	missing_count	percentage
RegionID	0	0
RegionName	0	0
City	0	0
State	0	0
Metro	250	0.027945
CountyName	0	0
SizeRank	0	0
1996-04	2662	0.297563
1996-05	2582	0.288621
1996-06	2582	0.288621
1996-07	2577	0.288062
1996-08	2576	0.287950
1996-09	2576	0.287950
1996-10	2576	0.287950
1996-11	2566	0.286832
1997-01	2542	0.284149
1997-02	2113	0.236195
1997-03	2093	0.233959
...	...	...

Airbnb data

Below are the key columns that are considered for data quality checks from the airbnb data.

column_name	missing_count	pct
neighbourhood_group_cleansed	0	0
city	62	0.001268
state	6	0.000123
zipcode	517	0.010574
latitude	0	0
longitude	0	0
property_type	0	0
room_type	0	0
bathrooms	56	0.001145
bedrooms	22	0.000450
square_feet	48487	0.991656
price	0	0
weekly_price	42891	0.877206
monthly_price	43647	0.892668
security_deposit	17317	0.354167
cleaning_fee	10645	0.217711
number_of_reviews	0	0
review_scores_rating	11022	0.225422
review_scores_value	11080	0.226608
reviews_per_month	10052	0.205583

Quality Insights from Initial Processing

After initial examination into both dataset,

Data Integrity and Data Quality

Completeness

There were plenty of columns in which missing values were present. The variables important for our analysis are:

Zillow's data

- `Monthly median price` by zipcode for the year 1996-2009: up to 2662 missing values per column of this time period. This is one of reason why I have not used data of earlier years for model making.

Airbnb data

- `bedrooms`: 22 missing values or 0.045%
- `zipcode`: 50 or 0.77%
- `weekly_price`,`monthly_price`

Accuracy

Negative or zero values

There are varibles having negative or zero numbers:

column_name	Zero_count	pct
longitude	48895	1.000000
bathrooms	126	0.002577
bedrooms	4569	0.093445
square_feet	34	0.000695
number_of_reviews	10052	0.205583

There are 4569 properties having 0 bedrooms in airbnb dataset.

Conformity

1. `zipcode`:
    - Some zipcodes have length other than 5 which questions the data credibility. 9 digit (Zip code + 4 code) values in the Zipcode column are trimmed to 5 digits. Eg, 11103-3233,11426-1175; Zipcodes less than 5 digit, eg,7310, are padded with 0 in front. Assume it happened because they were treated as integer when loaded.
    - convert zipcode into strings
2. `price`:
    - dollar sign and thousand comma are removed to beter analysis
    - convert price into integer
3. Zillow data is given for 1996-2017 by each column wise, which was later pivoted to better analysis.

Integrity

1. Zipcode `and` neighborhood ` Consistancy

   As long as column neighborhood is used, we should make sure zipcode and neighborhood is one to one mapped. There are some Zipcode `10013` is belongs to two different neighborhoods in the data, Manhattan and Brooklyn respectively.

```
zipcode  neighborhood       count
10013    Brooklyn           1
         Manhattan          70
```

2. Duplicaiton: Neithor dataset has duplicated records.

Timeliness

Zillow house price data was last scraped in June 2017 and Airbnb data was crawled in July 2019. The analysis was done in January 2020. Prediction should be done (is done but not used) for Zillow properties to keep both datasets in the same frame of time.

Metadata Created After Data Filtering and Mungling

Metadata Documentation

Airbnb_Zillow_Analysis is a file created by a candidate for Data Challenge in Capital One recruiting process. It includes datasets using for calculate investment ROI and generate business insights for a real estate company.

Datasets include `cost_2019`, `revenue_clean`, `payback`

`cost_2019`

This dataset is cost data filtered and processed from zillow dataset

Field	Description	Type
zipcode	zipcode that the properties are in, only contains NYC zipcodes appears Zillow. There are **25 unique zipcode** after filtering.	string
county	Political and administrative division of a state, referred to as a particular part of the state.	string
cost_2019	Median price of properties in the same zipcode in July 2019 calculated based on historical data	integer

`revenue_clean`

This dataset is revenue data derived from airbnb data after data processing.

Field	Description	
zipcode	zipcode that the properties are in, contains zipcodes appears in airbnb filtered by 2 bedrooms. There are **169 unique zipcode** after mungling	string
neighbourhood	Name of the area where the property is located, orignally derived from `neighbourhood_group_cleansed`, which has been cleaned as one-to-one mapping to zipcode	string
price	Price the host is charging to stay per night	integer

`payback`

After data filtering and cleaning, I have included important variables used in ROI analysis in payback table

Field	Description	Type
zipcode	zipcode that the properties are in, only contains zipcodes appears in both cost and revenue There are **24 unique zipcode** after mungling. ['10306', '10303', '11434', '10304', '10308', '10305', '11234', '10309', '10314', '10036', '10022', '11201', '11215', '10025', '11231', '10028', '11217', '10011', '10023', '10021', '10003', '10014', '10128', '10013']	string
neighbourhood	Name of the area where the property is located, orignally derived from `neighbourhood_group_cleansed`, which has been cleaned as one-to-one mapping to zipcode	string
rent_daily	Median daily rental price on airbnb grouped by zipcode	integer
rent_annual	Annual revenue coming from rental business rent_annual = rent_daily * 365 * occupancy	integer
cost	Median house price in July 2019 calculated based on historical data on zillow grouped by zipcode	float
num_property	number of property listed on airbnb for each zipcode	integer
breakeven	Break even time period measured in years	float
cost_20XX	Median price of properties in the same zipcode in 20XX calculated based on historical data	float
appreciation	housing price appreciation after purchsing	float
ROI	return of invenstment	float