Data Cleaning

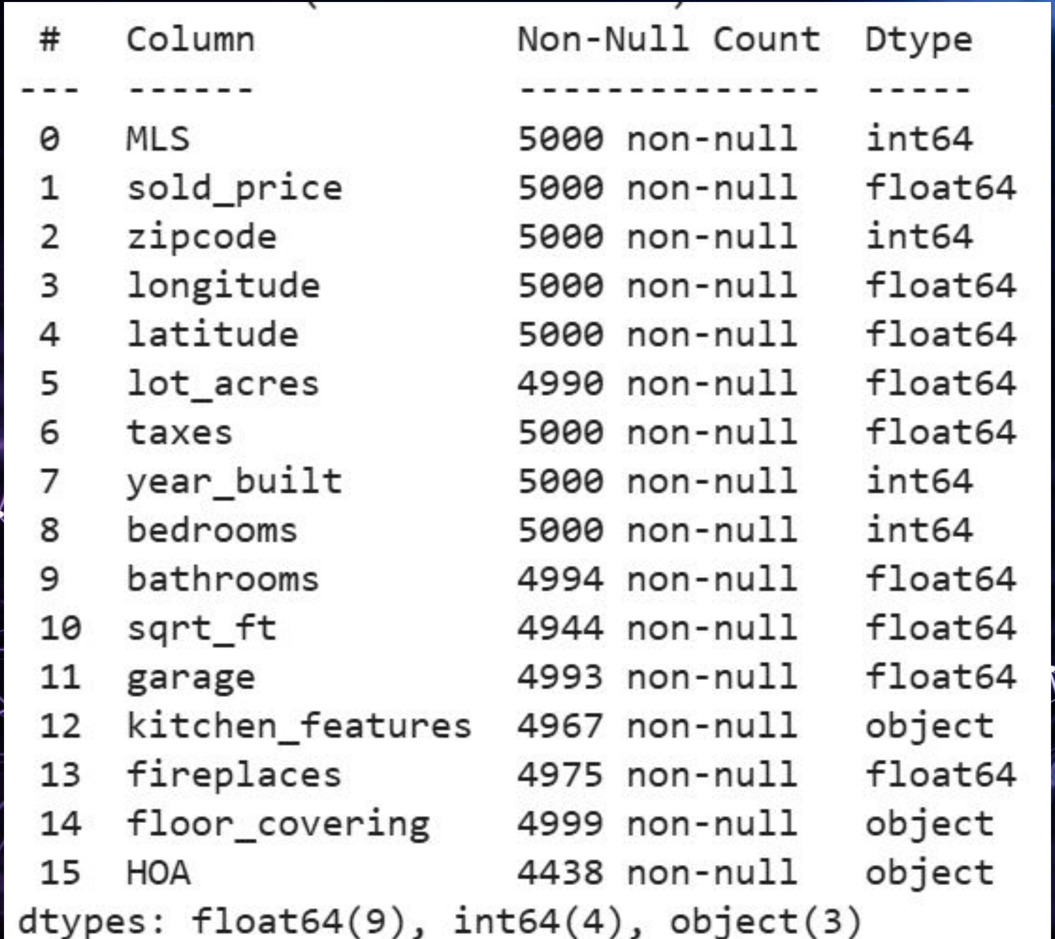
Housing Data

By :
Chandu Eddala

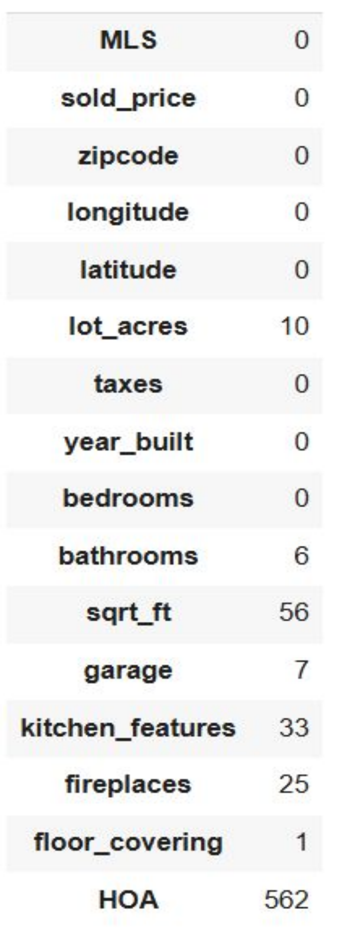
Index

- About the Data and Columns
- Filling Null Values
- Normalizing the Data
- One Hot Encoding (Kitchen Features and Floor Covering)
- Insights and Graphs
- Final Data.

About the Data and Columns



```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   MLS               5000 non-null   int64
 1   sold_price        5000 non-null   float64
 2   zipcode           5000 non-null   int64
 3   longitude         5000 non-null   float64
 4   latitude          5000 non-null   float64
 5   lot_acres         4990 non-null   float64
 6   taxes             5000 non-null   float64
 7   year_built        5000 non-null   int64
 8   bedrooms          5000 non-null   int64
 9   bathrooms         4994 non-null   float64
 10  sqrt_ft           4944 non-null   float64
 11  garage            4993 non-null   float64
 12  kitchen_features  4967 non-null   object
 13  fireplaces        4975 non-null   float64
 14  floor_covering    4999 non-null   object
 15  HOA               4438 non-null   object
dtypes: float64(9), int64(4), object(3)
```



MLS	0
sold_price	0
zipcode	0
longitude	0
latitude	0
lot_acres	10
taxes	0
year_built	0
bedrooms	0
bathrooms	6
sqrt_ft	56
garage	7
kitchen_features	33
fireplaces	25
floor_covering	1
HOA	562

Filling Null Values:

- **Bathroom:(6)**

  Filled by grouping as per Bedrooms and mean values of bathroom

- **Sqrt_ft:(56)**

  Filled by grouping as per bedrooms,bathrooms and mean values of Sqrt_ft

- **Garage:(7)**

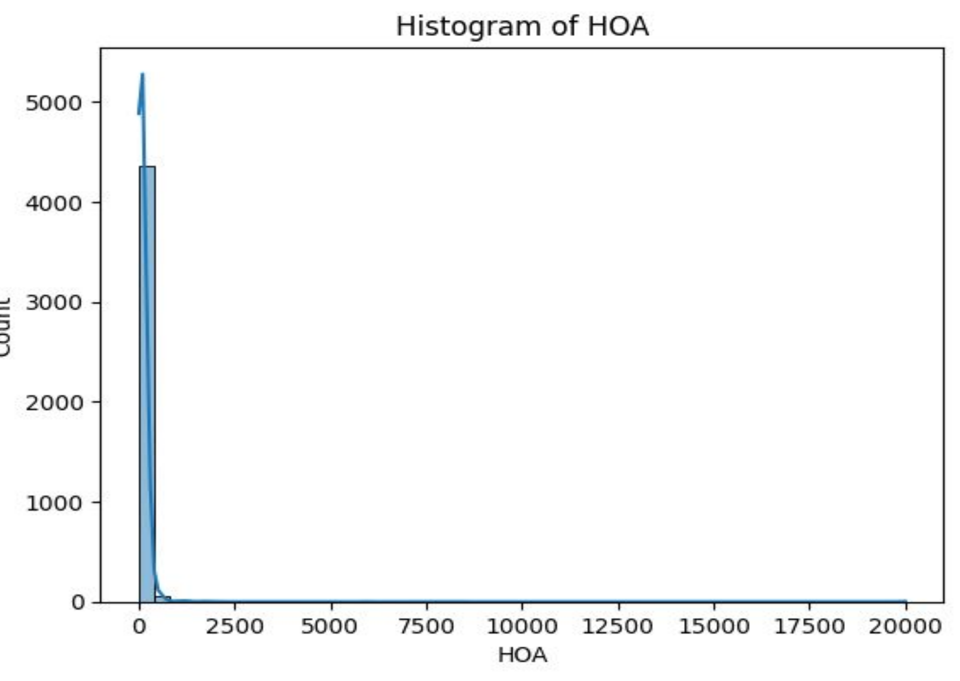
  Filled by grouping as per bedrooms,bathrooms and mean values of garage

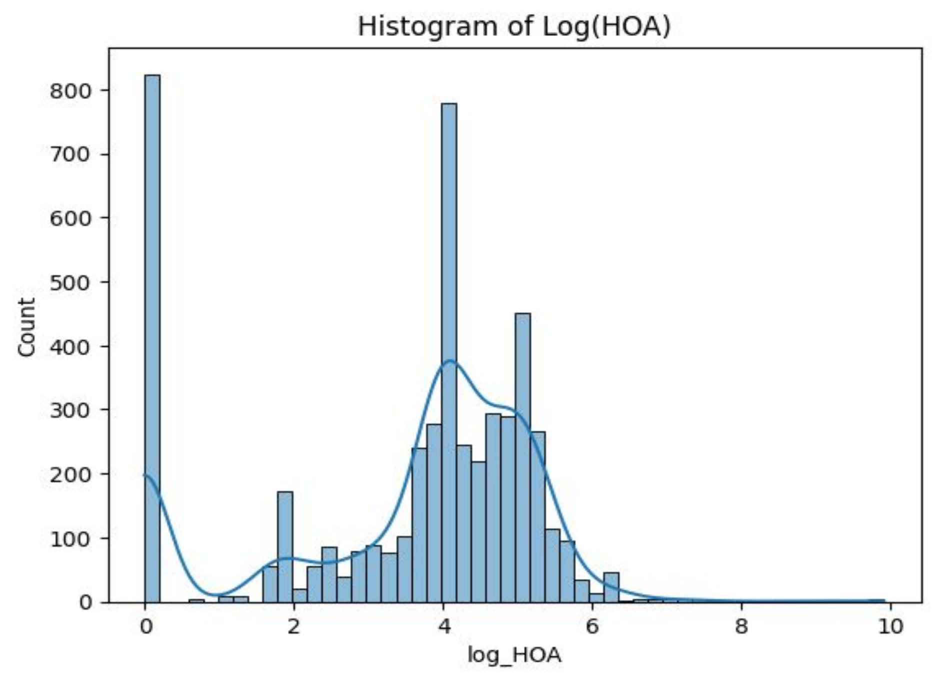
- **Fireplaces_avg:(25)**

  Filled by grouping as per bedrooms,bathrooms and garage,mean values of Fireplaces

HOA (Missing and Normalization)

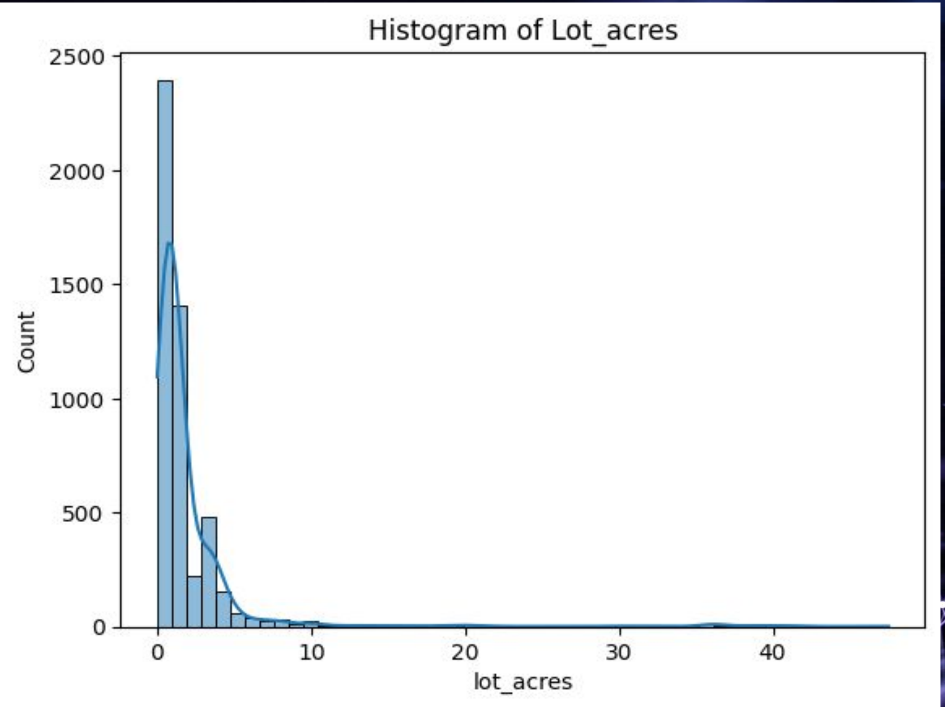
- Initially the spread is right skewed .
- We apply log to the HOA and get the Normalized form
- After the Log,fill the Null values with medium .

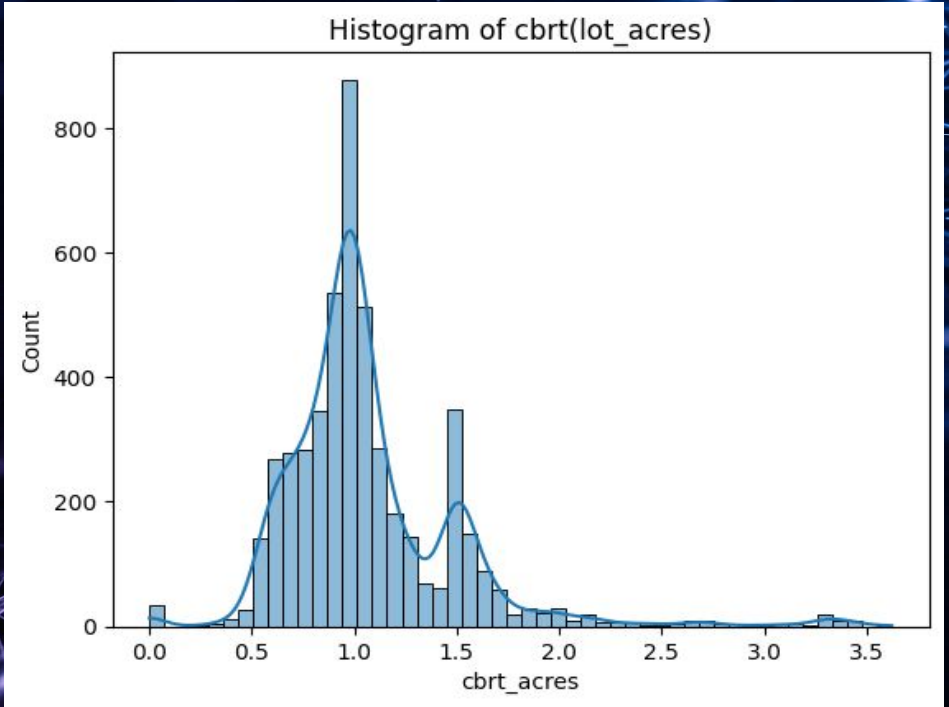




Lot Acres(Missing and Normalization)

- Initially the spread is right skewed . and also only 64 values are more than 50.
- We apply cube_root to the Acres and get the Normalized form
- After the Log,fill the Null values with medium .





Kitchen Feature and Floor Covering

1. Split the string and counted the frequency
2. Considered ,A feature found in at least 5% of houses .(Columns)
3. Filled the null values with ,features that had found at least 60% of Houses .

From Kitchen Feature (19): dishwasher,freezer,refrigerator,oven,garbage disposal,double sink, microwave,compactor,'electric range,island,appliance color,gas range,prep sink,countertops granite,desk,lazy susan,pantry walk-in,pantry closet,pantry cabinet.

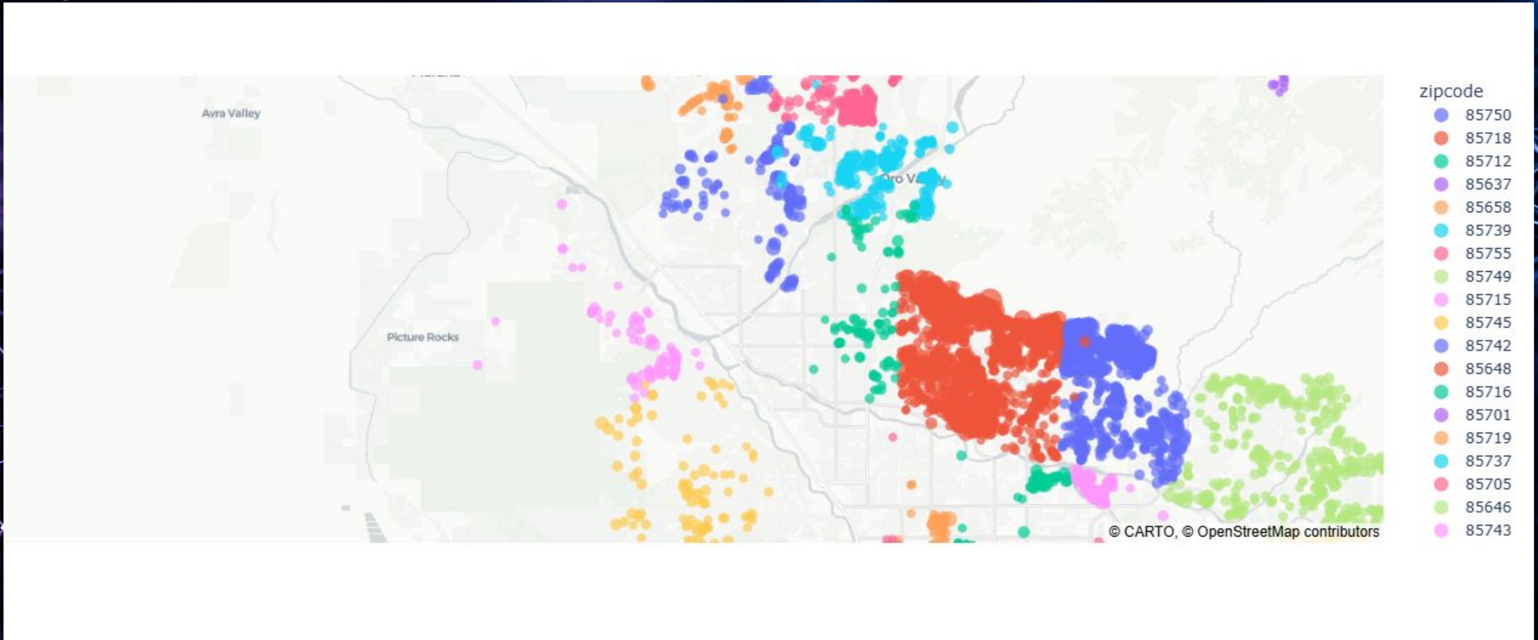
Null :dishwasher,refrigerator,garbage disposal

From Floor Covering(7):mexican tile,wood,natural stone,other,ceramic tile,carpet,concrete.

Null:carpet

Longitude and Latitude

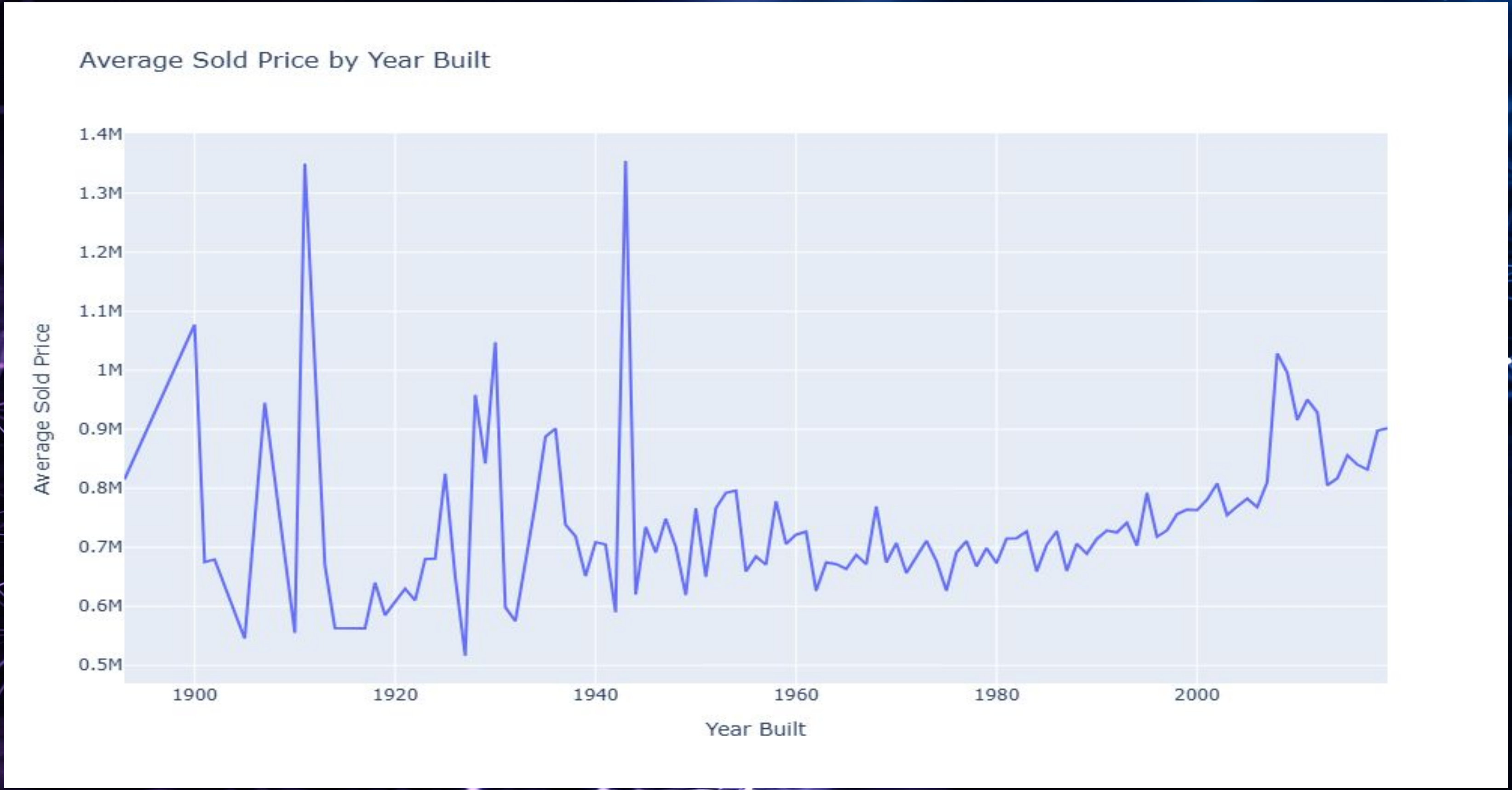
- The distance calculated from the center .
- 2 more columns created for the directions the house located .



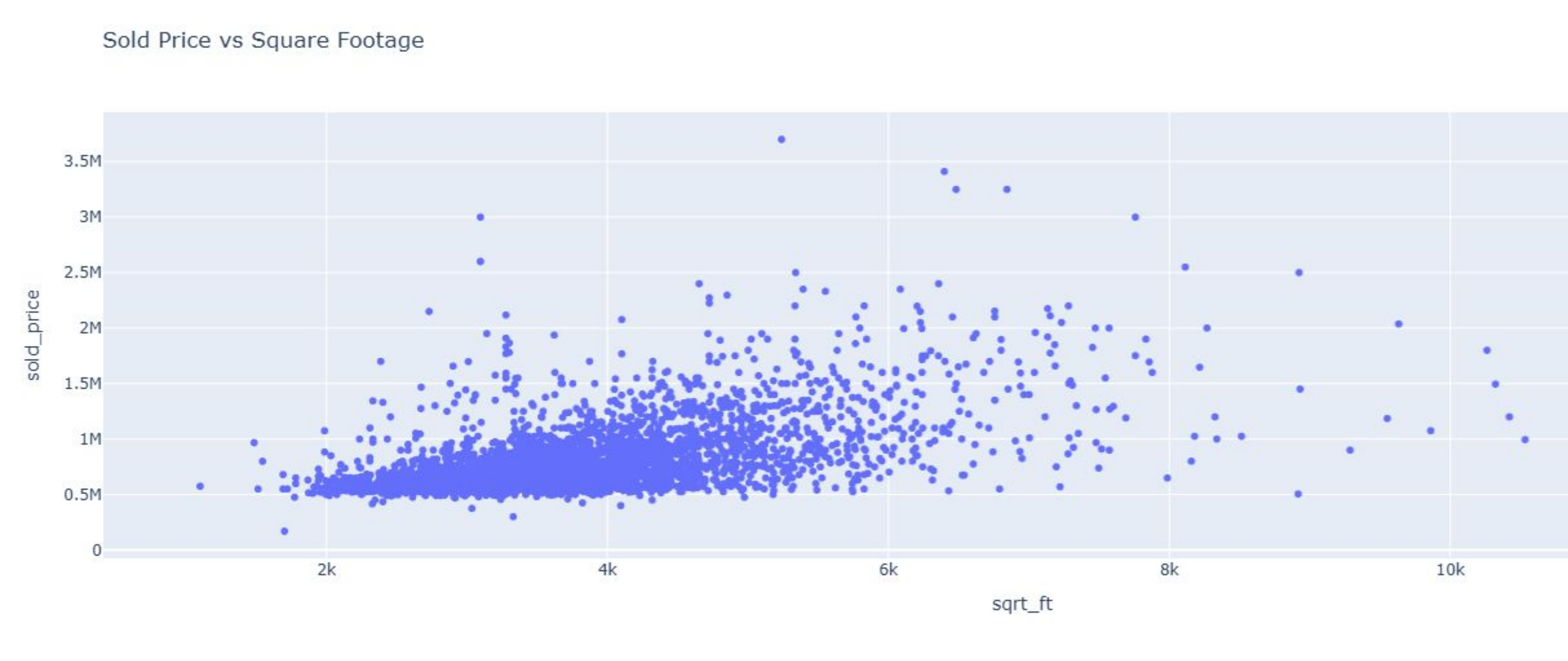
© CARTO, © OpenStreetMap contributors

Average Sold Price by year Built.

- 3 Columns are Removed as they have 0 as year



Sold Price Vs Square Footage



Final Conclusion :

- Finally Cleaned and Process Data had 40 Columns and 4929 rows .
- Considered most of the feature in Kitchen and in Floor that are appear at least 5% of houses if we wanted to reduced the number of columns ,we can consider fewer columns depending on the requirement
- Added the 3 more columns based on the Longitude and Latitude like distance from center and the direction of longitude and latitude .
- The HOA into Log(HOA) and Lot_acers into Cube_Root(Lot_acers) for normalization .
- While doing the data profiling ,removed few rows :

1.64 records removed ,as per the lot acres which are greater than 50.

2.Like 3(Zeros) in year removed those rows.

3.one Outlier in tax removed.

THANK YOU