House Price Modelling Exploratory Data Analysis

Atalay İlhanlı
Aydın Bayraktar
Emre Boran
Gül Evin Yılmaz
Orhan Yaman

1) The source of data

We took our data from Kaggle. Data set was constructed from the publicly available data in Domain.com.au.

Data cleaning: Dropping missing values

Dividing the features into 2 groups:(Numerical and categorical)

Eliminating the outliers.

Checking for duplicate variables

Num of missing values

```
Suburb              0
Address             0
Rooms               0
Type                0
Price            4344
Method              0
SellerG             0
Date                0
Distance            1
Postcode            1
Bathroom         4055
Car              4055
Landsize         4082
BuildingArea    11359
YearBuilt       10092
CouncilArea      4085
Lattitude        3937
Longtitude       3937
Regionname          1
Propertycount       1
Age             10092
Historic            0
dtype: int64
```

2) Feature list and their descriptions

Suburb: Suburb

Address: Address

Rooms: Number of rooms

Price: Price in dollars

Method: S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.

Type: br - bedroom(s); h - house,cottage,villa, semi,terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.

SellerG: Real Estate Agent

Date: Date sold

Distance: Distance from CBD

2) Feature list and their descriptions (cont.)

Regionname: General Region (West, North West, North, North east ...etc)

Propertycount: Number of properties that exist in the suburb.

Bedroom2 : Scraped # of Bedrooms (from different source)

Bathroom: Number of Bathrooms

Car: Number of carspots

Landsize: Land Size

BuildingArea: Building Size

YearBuilt: Year the house was built

CouncilArea: Governing council for the area

Lattitude: Self explanitory

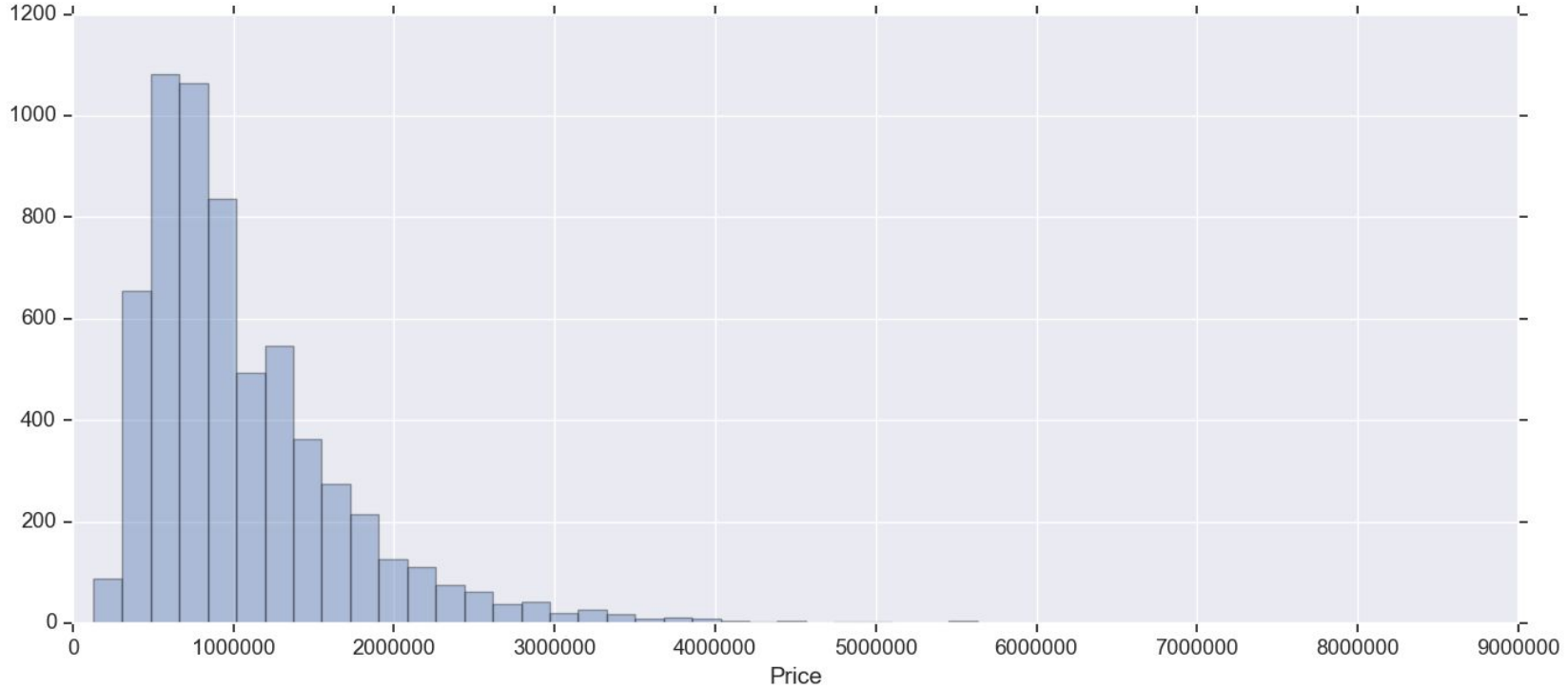
Longtitude: Self explanitory

3) Statistical exploration of features

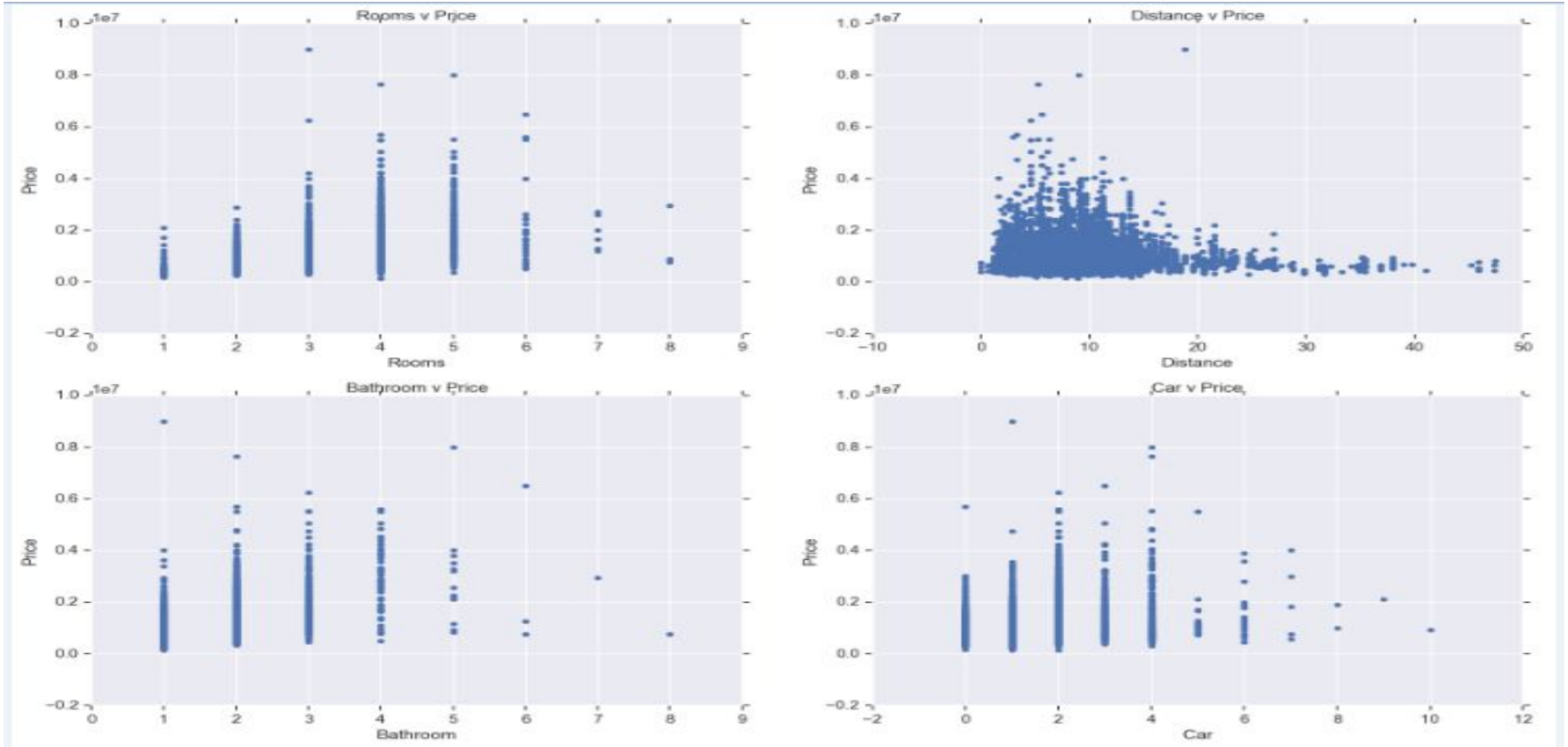
	count	mean	std	min	25%	50%	75%	max
Rooms	6195.0	2.931558e+00	0.971085	1.00000	2.000000	3.0000	4.000000e+00	8.000000e+00
Price	6195.0	1.068865e+06	675204.719649	131000.00000	620000.000000	880000.0000	1.325000e+06	9.000000e+06
Distance	6195.0	9.752300e+00	5.611720	0.00000	5.900000	9.0000	1.240000e+01	4.740000e+01
Bathroom	6195.0	1.576433e+00	0.711382	1.00000	1.000000	1.0000	2.000000e+00	8.000000e+00
Car	6195.0	1.573688e+00	0.929993	0.00000	1.000000	1.0000	2.000000e+00	1.000000e+01
Landsize	6195.0	4.710483e+02	897.516427	0.00000	152.000000	373.0000	6.280000e+02	3.700000e+04
BuildingArea	6195.0	1.415915e+02	90.824342	1.00000	91.000000	124.0000	1.700000e+02	3.112000e+03
YearBuilt	6195.0	1.964076e+03	38.106016	1196.00000	1940.000000	1970.0000	2.000000e+03	2.018000e+03
Lattitude	6195.0	-3.780791e+01	0.075856	-38.16492	-37.855455	-37.8023	-3.775820e+01	-3.745709e+01
Longtitude	6195.0	1.449902e+02	0.099171	144.54237	144.926195	144.9958	1.450527e+02	1.455264e+02
Propertycount	6195.0	7.435589e+03	4338.042029	389.00000	4382.500000	6567.0000	1.017500e+04	2.165000e+04
Age	6195.0	5.292381e+01	38.106016	-1.00000	17.000000	47.0000	7.700000e+01	8.210000e+02

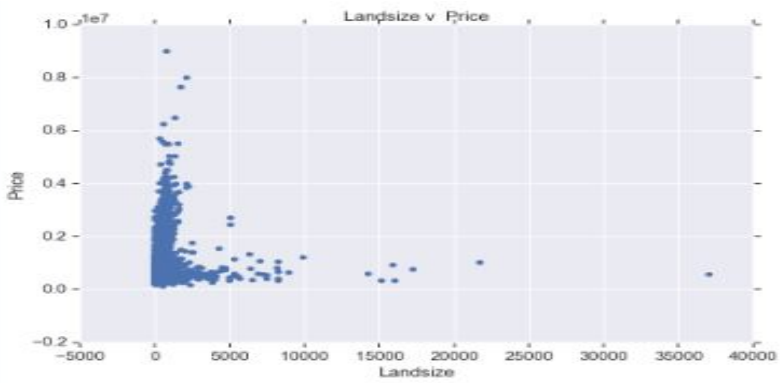
Our dependent variable in this analysis is Price. This variable appears to be normally distributed and skewed to the right.

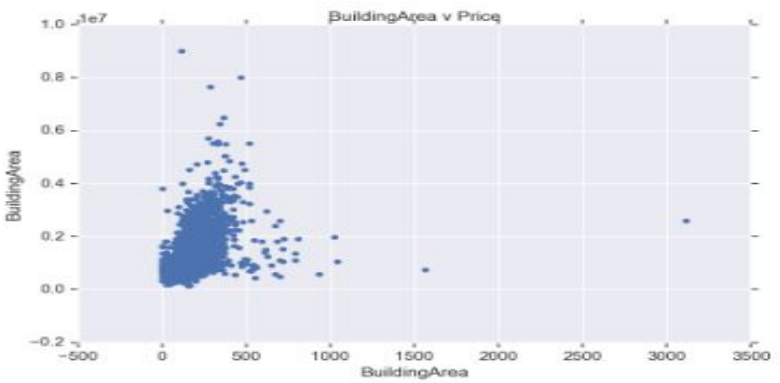
That is, the majority of homes around $800k with some outliers.

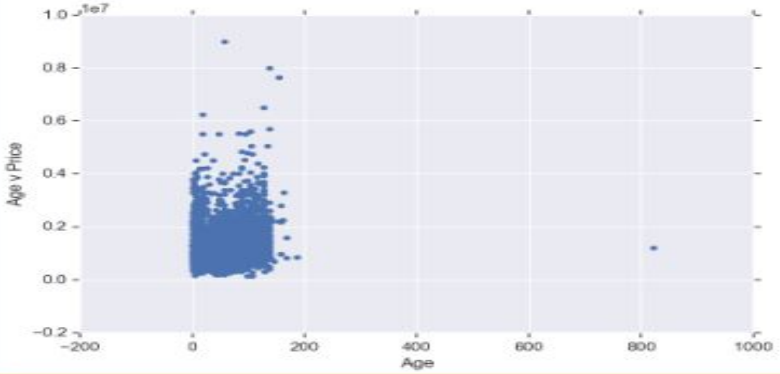


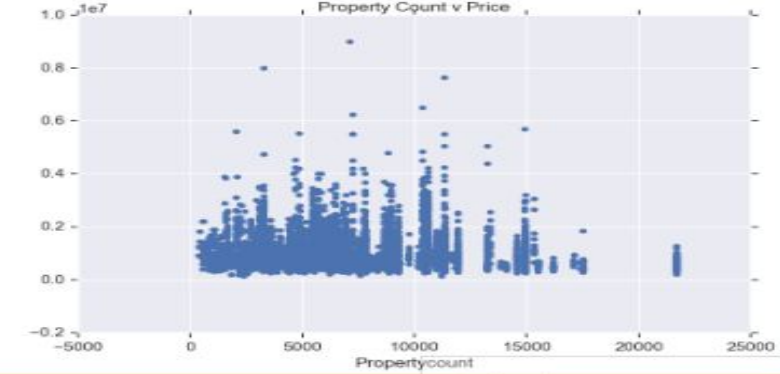
Numerical Features



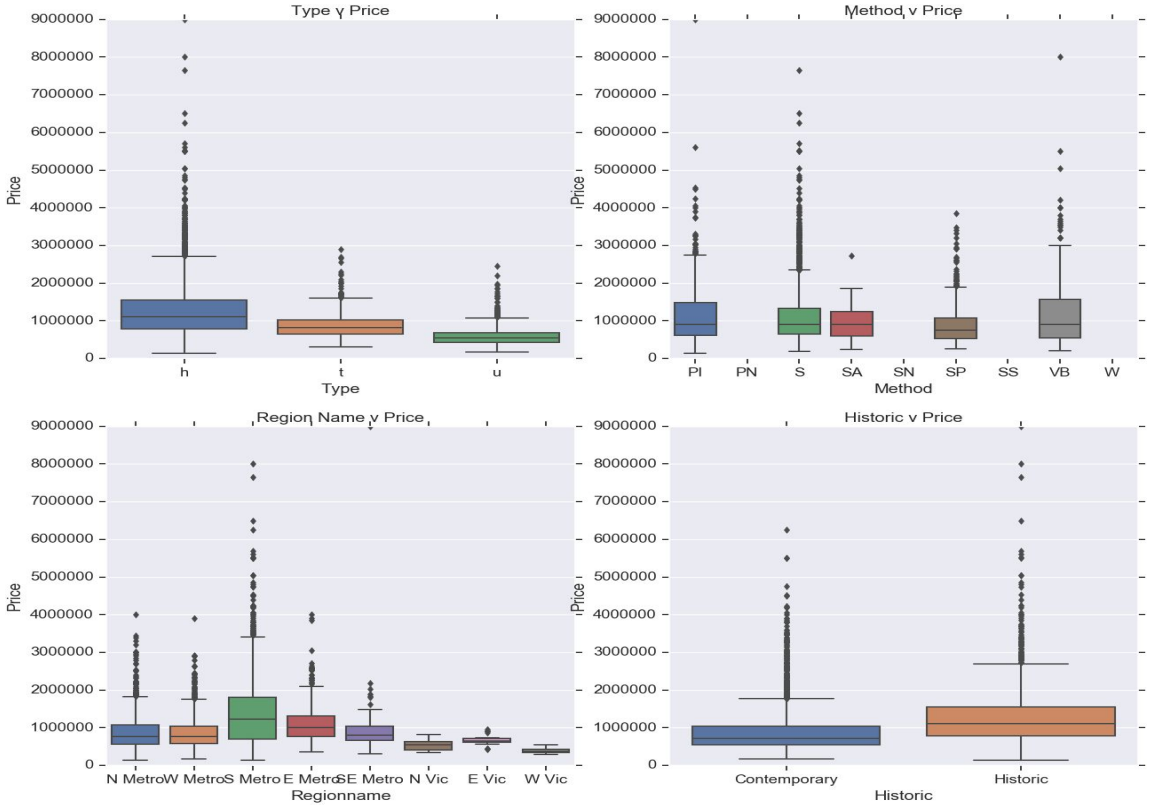

Landsize v Price
1e7
Price
Landsize


BuildingArea v Price
1e7
BuildingArea
BuildingArea


1e7
Age v Price
Age

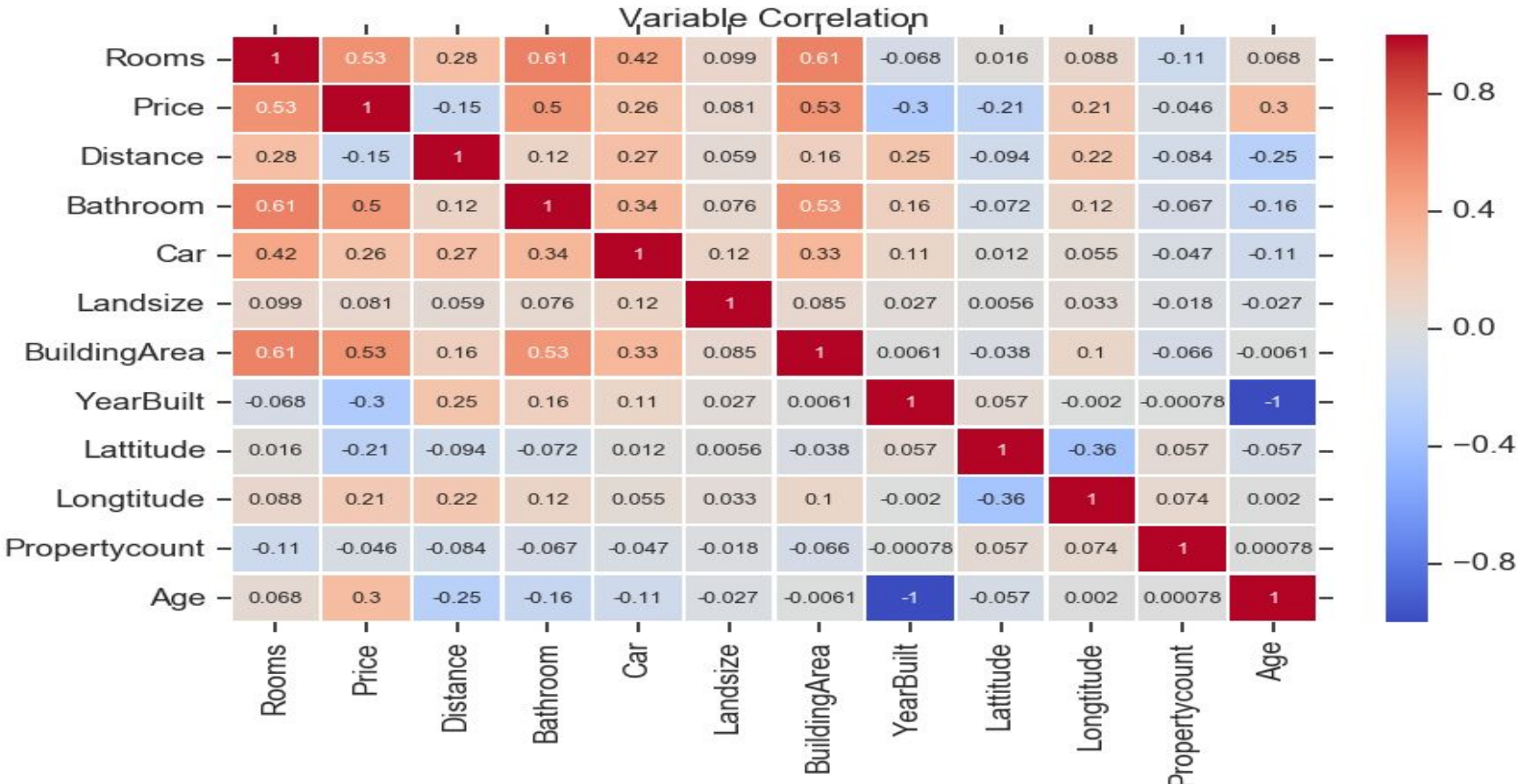

Property Count v Price
1e7
Propertycount

Categorical Features



- Median prices for houses are over $1M, townhomes around $800k and units are approx $500k.
- There isn't much difference between selling methods.
- Houses in the Metropolitan region are more expensive than the Victoria Region.
- Historic homes are much more expensive than the never homes.

Correlation



4) ML methods

We will use linear regression model and try to predict housing prices using our independent variables.

We will divide our data into two groups (test group and train group) randomly.

We will analyze the coefficients which will be outputs of our regression.