Data Preprocessing and EDA

Review 1

Digital Enigma

Data Preprocessing

Data preparation

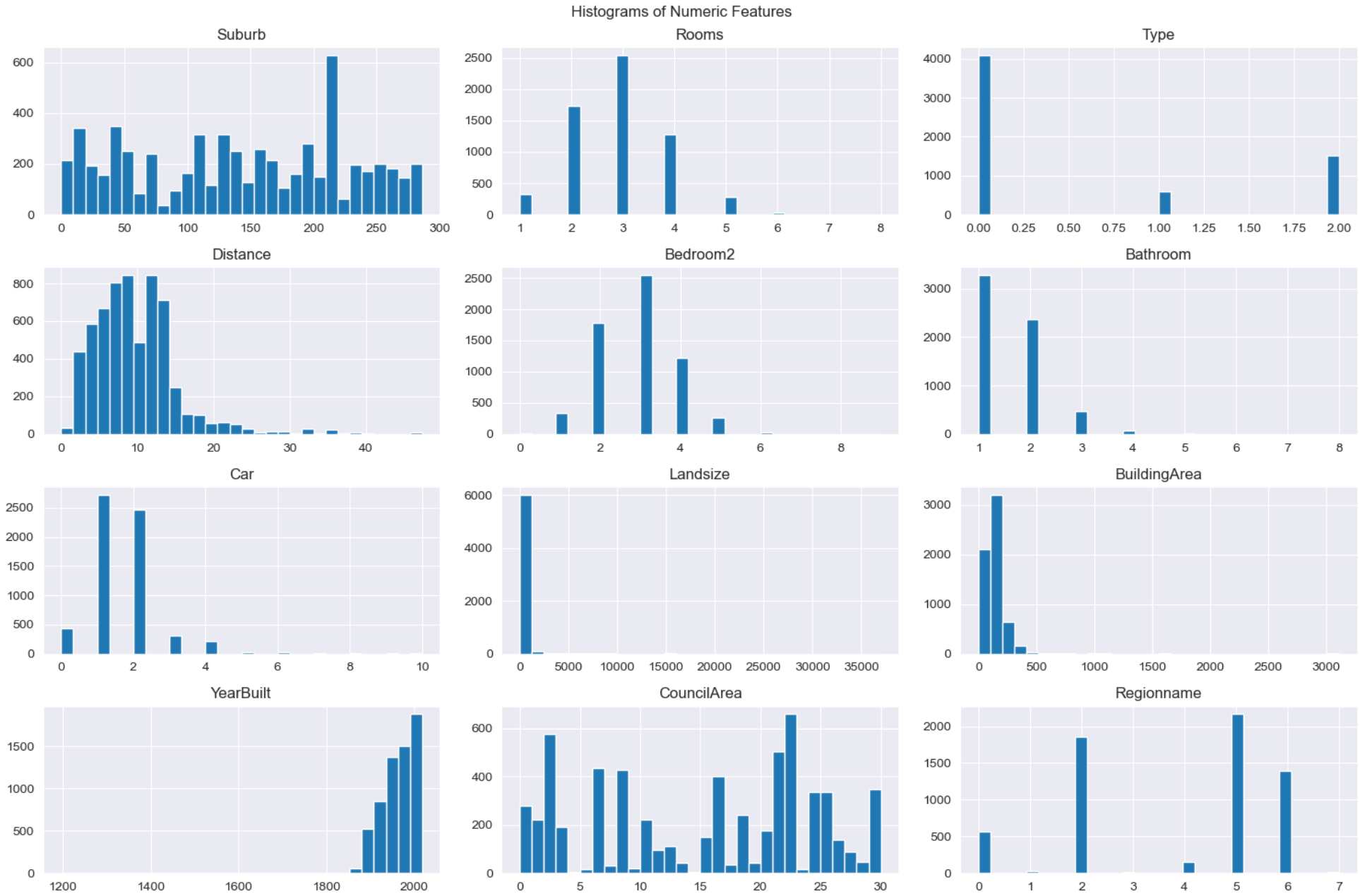
1. Dropping columns that are obvious like address, seller, date which do not contribute to price
2. Dropping rows which have NA values
3. Convert non-numerical values (like suburb, region and type) into integers for numerical analysis

Exploratory Data Analysis

Suburb	Rooms	Type	Price	Distance	Bedroom2	Bathroom	Car	Landsize	Building…	Year…	Cou…	Reg…
6196.0	6196.0	6196.0	6196.0	6196.0	6196.0	6196.0	6196.0	6196.0	6196.0	6196.0	6196.0	6196.0
141.733376…	2.93140735…	0.5832795…	1068828.202065849	1097482246…	2.90203357004…	1.57633957391…	1.573595…	471.006939961…	141.568644544…	1964.0819…	14.67850…	3.824402…
83.2021911…	0.97107880…	0.8539919…	675156.427…	5.612065327996…	0.97005482126…	0.71136188787…	0.929946…	897.449880543…	90.8348237271…	38.105673…	9.217255…	1.986514…
0.0	1.0	0.0	131000.0	0.0	0.0	1.0	0.0	0.0	0.0	1196.0	0.0	0.0
64.0	2.0	0.0	620000.0	5.9	2.0	1.0	1.0	152.0	91.0	1940.0	6.0	2.0
141.0	3.0	0.0	880000.0	9.0	3.0	1.0	1.0	373.0	124.0	1970.0	16.0	5.0
214.25	4.0	1.0	1325000.0	12.4	3.0	2.0	2.0	628.0	170.0	2000.0	22.0	5.0
286.0	8.0	2.0	9000000.0	47.4	9.0	8.0	10.0	37000.0	3112.0	2018.0	30.0	7.0

Overview of data

Analyzing price distribution



Analysis of features

01

Detecting and visualising outliers using box plots

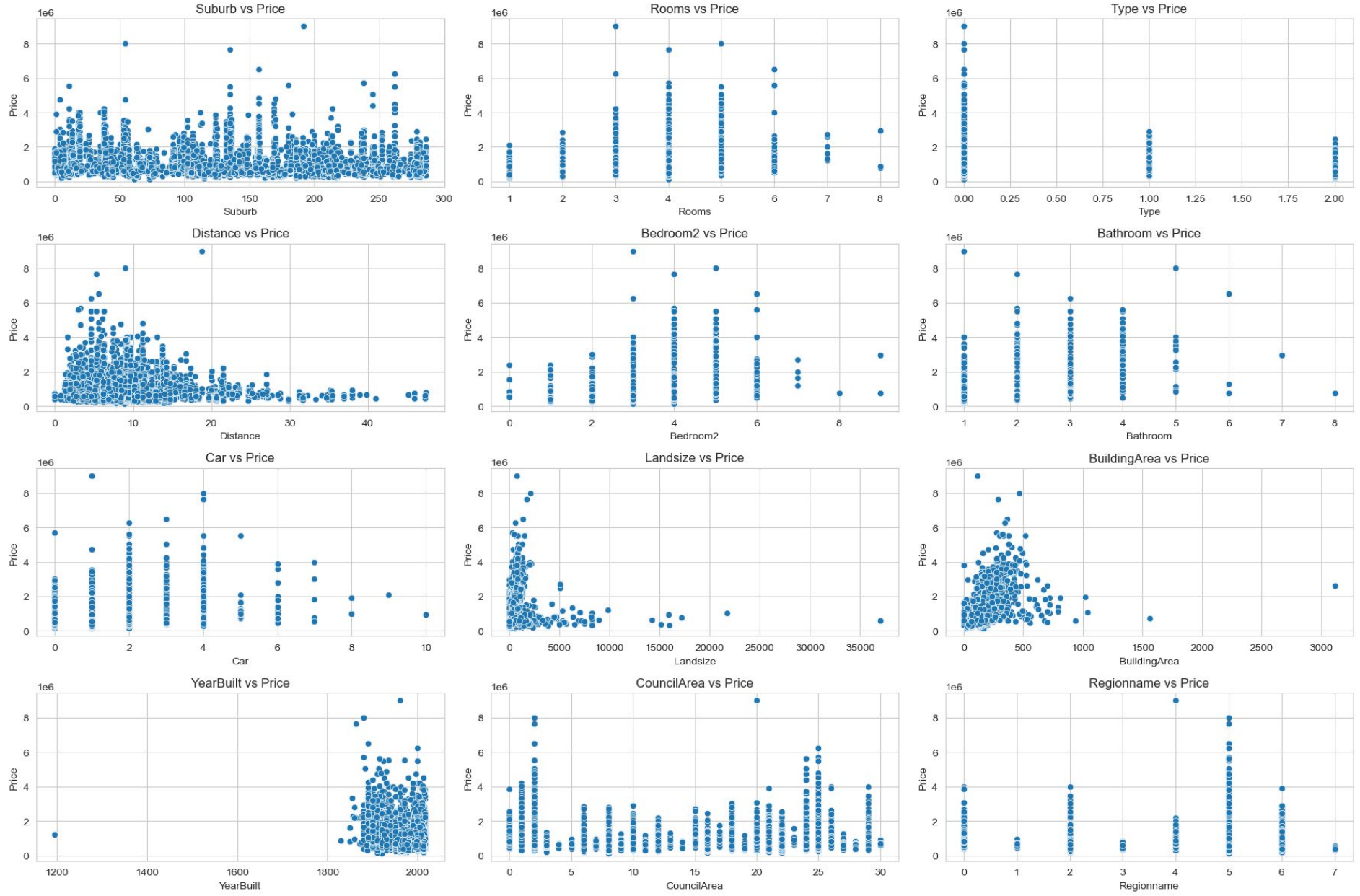
02

Removing outliers by doing calculations on IQR

Handling outliers

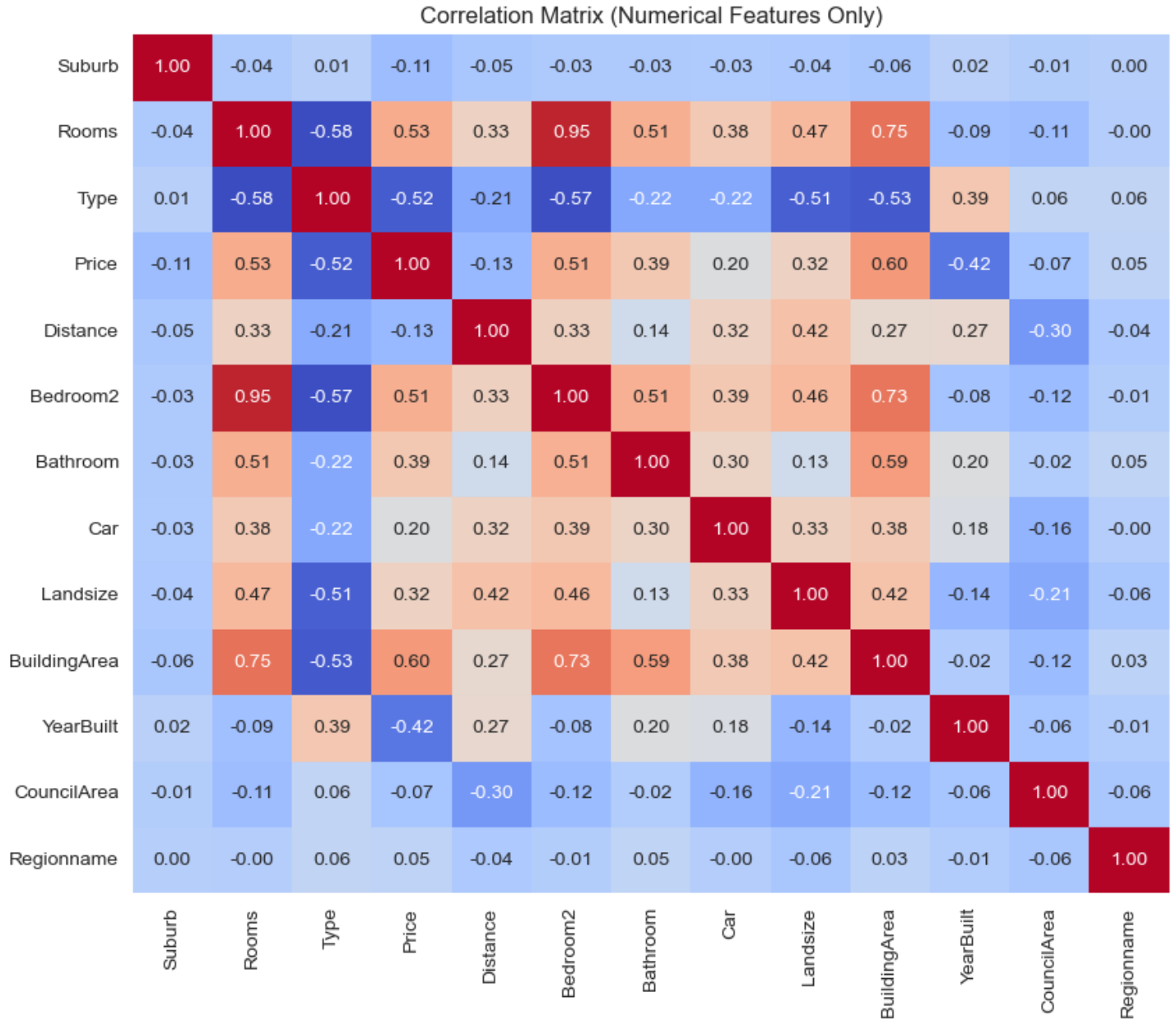
Handling outliers

- Detecting and visualising outliers using box plots
- Removing outliers by doing calculations on IQR



Relationship between price and features

Final correlation matrix



Relevant features

After completing EDA, the features to focus on are:

- No. of rooms
- Type of house
- Scraped no. of bedrooms
- Building size

Thank you!