EDA REPORT

Foundations of Data Science

OCTOBER 14, 2024

DUBLIN BUSINESS SCHOOL
Wicklow House, 84/88 South Great George's Street, Dublin

Assignment Cover Sheet

Student Name: Daniel Fernandes Moreira Neto

Student Number: 10564203

Programme: BSc (Hons) Computing Yr. 3

Lecturer Name: Oleksandr Bezrukavyi

Module/Subject Title: Foundations in Data Science

Assignment Title: Data Understanding - EDA

No of Words: 780

Contents

Report

This report looks at the analysis of a dataset from daft.ie. The dataset starts with 22 columns and 3952 rows, except for the propertySize column, which has missing data (flagged in bold).

Information

#	Column	Non-	Null Count	Dtype
---	------	----	----------	-----
0	id	3952	non-null	int64
1	title	3952	non-null	object
2	featuredLevel	3952	non-null	object
3	publishDate	3952	non-null	object
4	price	3952	non-null	int64
5	numBedrooms	3952	non-null	int64
6	numBathrooms	3952	non-null	int64
7	propertyType	3952	non-null	object
8	propertySize	**3604**	non-null	float64
9	category	3952	non-null	object
10	AMV_price	3952	non-null	int64
11	sellerId	3952	non-null	float64
12	seller_name	3952	non-null	object
13	seller_branch	3952	non-null	object
14	sellerType	3952	non-null	object
15	m_totalImages	3952	non-null	float64
16	m_hasVideo	3952	non-null	bool
17	m_hasVirtualTour	3952	non-null	bool
18	m_hasBrochure	3952	non-null	bool
19	ber_rating	3952	non-null	object
20	longitude	3952	non-null	float64
21	latitude	3952	non-null	float64

EDA

The main challenge was handling the missing values in propertySize, which I believed would have a strong correlation to the price. The end goal is to predict house prices, so to fill the missing values, I used the IterativeImputer function from Scikit-learn. Instead of just using the most important columns based on the heatmap, I decided to apply one-hot encoding to the categorical columns and include all the features to impute propertySize.

Before this, I checked for outliers. The initial R^2 score was very low at -0.002. I found some extreme outliers—properties with a large size but only one bedroom and bathroom. These seemed to be free land, so I removed them from the dataset since the focus was on houses. I also limited the propertySize to 600 sqm, which helped improve the R^2 score to 0.47, a big improvement.

After that, I created two datasets: one with the categorical variables added back on again and another for numerical variables to predict house prices. For the visualization dataset (the ones with categorical variables) I reduced the dataset to 11 columns (focused on the main columns for visualization).

Variables for the visualization (Categorical & Continuous)

```
#      Column           Non-Null Count    Dtype
---    ------           --------------    -----
0      price            3578 non-null     int64
1      numBedrooms      3578 non-null     int64
2      numBathrooms     3578 non-null     int64
3      propertySize     3578 non-null     float64
4      total_images     3578 non-null     float64
5      longitude        3578 non-null     float64
6      latitude         3578 non-null     float64
7      ber_rating       3578 non-null     category
8      featuredLevel    3578 non-null     category
9      propertyType     3578 non-null     category
10     sellerType       3578 non-null     category
```

Visualization - (Interpret each feature & relationship between the features)

Folium

In the visualization part, we explored various options. I started by importing Folium, a geospatial library that helps map the location of properties based on latitude and longitude. This allowed me to see an outlier where a house was incorrectly placed in the USA. By coming back to the EDA (previous dataset) part and checking the title column (which contained the address), I corrected the latitude and longitude to its actual location in County Cork.

Histograms

Next, we analyzed the price distribution. The graphs showed a Gamma distribution, with house prices ranging from €20,000 to €4,500,000. Most prices were between €20,000 and €1 million, so I capped the visualization at €1 million and rechecked the data. The Gamma distribution shape remained the same. This pattern also repeated for features like property size, number of bedrooms, and total images.

Scatter Plots

For the categorical data, we noticed the graphs didn't follow a specific trend. In the correlation heatmap, the strongest correlation was 0.61 between number of bedrooms and bathrooms, while property size and price had a 0.45 correlation. We also found a negative correlation, with price and total images having no correlation at all ($R^2 = -0.019$).

Box Plots

The box plot for BER rating and price showed that most homes had a median price between €0 and €1 million, but there were outliers. One notable outlier was a house priced at €4,500,000 with a BER rating of F, which suggests high energy consumption. Looking further into the data, we found this house is located in Foxrock, Dublin 18—a very upscale neighbourhood. The house was built in 1906, which explains its F1 rating. However, the house has since been sold, refurbished, and now holds a B1 BER rating.

Pie Charts

In the pie charts, we see that most houses have a standard ad on Daft.ie, but a few have featured or premium ads, which provide better visibility on the site. Regarding the seller types, the dataset includes three categories: unbranded agent, branded agent, and private user, which are explained within the dataset.

This concludes our visualizations. All visualizations were defined in functions to maintain clean and organized code, making it easier to understand and follow best coding practices.

Pairplot

For the pairplot, I used Seaborn to quickly visualize the relationships between different features in the dataset. It shows scatter plots for each pair of variables and histograms on the diagonal. From the pairplot, I could see a positive relationship between the number of bedrooms and bathrooms and some correlation between property size and price. This visualization helped me spot patterns and some outliers in the data.