Question: Can we produce an analytical tool that helps real estate agents and home buyers better understand the apartment market in Riyadh?

Can we develop an ML model that would help say a real estate agent filter only those apartments that are offered for a great price saving him a lot of time?

Can we validate our intuition on price appropriateness using such predictive model?

Finally, there is only one dataset on Kaggle on Aqar data, so I wanted to contribute to the community by scraping and cleaning the website and providing the dataset for other to use.

The source and structure of my data: The source is Aqar’s website, Saudi Arabia’s unparalleled online real estate market. The structure of the data:

[Data dictionary] [problems with]

The data is initially in raw format (unclean), so I developed a function that cleans the dataset.

Assumptions:

* People agree on the definition of “living room”, “bed room” etc,.
  + Some cases I saw in the dataset were relatively large apartments with 3 bedrooms and 1 living room, and another apartment with the same layout would be described as having 2 bedrooms and 2 living rooms.
* Advertisers provide fair estimates of the price
* We can estimate the price using only general information on the apartment.
  + The data doesn’t include the exsistance of maintenance costs, water bill cost, elevator, etc,.

Steps to find solution:

* Scrape the data 5 times (once for each area in Riyadh) and combine them in a single dataset
  + Find post()
* Clean the dataset
  + [Clean\_post()] [final data dictionary]
* EDA
  + I utilized libraries like pandas reports to automize a lot of my EDA
* Develop ML models
  + Train-test split
  + Fillna()
  + Normalize features
  + Standarize features
  + Dummy encoding
  + Impute missing values using knn
  + Noramilize and standardize price using box-cox
* Develop Good Deal Indicator
  + Good\_deal\_indicator()
* Develop Analytical dashboard

Challenges:

* Poor data quality (many people opt to provide erroneous information instead of leaving them blank)
* Difficulty interpreting the results of the ML model when we preprocess the price

Take home message:

* We have been successful in developing an analytical tool that achieves the goals we set out for it: better understand the apartment real estate market in Riyadh and utilize publicaly available data to help home buyers make better purchases.

Outlook for future development:

* Imporove data quality
* Utilize text description of properties in making price predictions
* Utilize the provided images of properties in making price predictions
* Utilize a google maps-like API to expand the features in the dataset to include things like: distance\_to\_major\_highway, distance\_to\_grocery\_store, etc.