



# ISE 291 Term Project

## Data Science on **askaan** business data <sup>4</sup>

Due date: Dec 25, 2021.


A national level investment company, called **askaan** business co. bearing logo , is looking for data scientists to help them understand the possible patterns that will affect the real estate prices. Currently the company purchases and sells real estates across the country. The company is interested in **estimating the price of real estate after 5 years from the date of purchase**. Such prediction system will help the company to invest in potential estates that will generate substantial profit margins. The company has provided the relevant data that they have collected over the years. Following table presents an overview of the given data:

Table 1: Data Description (Any non-applicable value is set to NA)

Fields	Description
Sale-Price	Sale Price of the property after 5 years from the date of purchase in millions of SAR.
Purchase-Date	Month and year, when the property was purchased.
Purchase-Price	Property's price at the time of purchase in millions of SAR.
Type	Type of the property. The property could be <b>open-land, villa, duplex, flat</b> .
Class	Legal classification of the property, could be one of the following options: <b>residential, industrial, or commercial</b> .
Location	Where the property is located w.r.t nearby city. 'Center' implies center of the city, 'Border' implies at the entry/exit of city, 'Outskirts' implies on the outskirts of the city.
Shape	Shape of the property. It could be <b>rectangle, trapezoid, irregular</b> .
U-Index	Index based on number of utilities available on a <b>scale of 1 to 5</b> . A value of <b>5</b> indicates <b>all utilities are available</b> .
Proximity	Proximity to the nearest metro station in meters.
N-Rank	Rank based on neighborhood facilities that will make the property attractive on a scale of <b>1 to 10</b> . A value of <b>1</b> indicates <b>the best neighborhood</b> .
P-Chance	Probability of finding parking space on adjacent roads at a given time. It is a value between <b>0 and 1</b> , where <b>1</b> indicates <b>sure availability of parking space</b> .
Built	Original <b>year of construction</b> . Applicable for villa, duplex, flat.
Renovate	Latest renovation year. Applicable for villa, duplex, flat. A value of <b>0</b> implies <b>no renovation</b> done so far or renovation not applicable.
Access	Type of direct access to the property, which could be <b>street, alley or highway</b> .
Crime-Rate	Average number of crimes reported per year in the neighborhood.
C-Rating	Pleasantness of the climate throughout the year on a <b>scale of 1 to 5</b> . A value of <b>5</b> indicates <b>pleasant climate</b> .
Gov-Index	Expected level of government infrastructure project and/or developments in the neighborhood on a <b>scale of 1 to 10</b> . A value of <b>10</b> indicates that there are <b>huge developments</b> planned by the government.
Contour	<b>Flatness of the property</b> . Applicable <b>only for the open land type</b> property. A value of <b>C</b> indicates the <b>slope</b> of the property is <b>irregular</b> . A value of <b>F</b> indicates the <b>property</b> has a <b>smooth slope</b> .
Garage	Is there a private parking garage? <b>Yes or No</b> . Applicable to the flat or duplex type. All villas have private garage.
Swimming	Is there a swimming pool? <b>Yes or No</b> . Applicable to the villa type.

**Aim.** The aim of this project is to explore the data, and find possible patterns/relationships in the data. The key variable of interest to **askaan** business co. is Sale-Price. **Any patterns that shows connections of input variables to the output variable (Sale-Price) will be considered fruitful by askaan**. Assume that the properties that **appreciate by 100% or less over the five years are low potential estates**, and those that **appreciate by 400% or more are high potential estates**. The percentage increase or decrease is defined as  $\frac{[Sale-Price] - [Purchase-Price]}{[Purchase-Price]} * 100$ .

**Data.** The data related to the project is provided in three different files, named in the following format: Group\_XX\_A, Group\_XX\_B and Group\_XX\_C files, where XX is your group number. In addition to that, Table 1 presents the meta data related to the given data.

**Expectations.** At the end of this project, you are expected to provide **askaan** with **answers to the following questions**. Support your answers with corresponding/appropriate data science methods and visualizations (wherever applicable).

<sup>4</sup> Case study for ISE 291 Term 211 students. Developed by: Dr. Mujahid Syed.



For the following task use Group\_XX\_A file:

Task-1: Prepare the data given in Group\_XX\_A file, i.e., handle the missing values, remove outliers, and fix inconsistencies. You can pick any set of methods, but clearly justify your approach.

For the following task use Group\_XX\_B file:

Task-2: Draw the pair-wise plots between all the input variables and the output variable (Sale-Price).

Task-3: Identify top and bottom three numerical variables that are strongly related to the output variable (Sale-Price)? Use the relevant analysis approach.

Task-4: Show if the input variables have the information to separate low and high performing estates? Use plots to justify.

Task-5: What are the common patterns for the low performance of the estates? Use plots to justify.

Task-6: What are the common patterns for the high performance of the estates? Use plots to justify.

For the following task use Group\_XX\_B and Group\_XX\_C files:

Task-7: From the input and output columns given in Group\_XX\_B file; identify how the input variables together are related to the output. Assume that all the input variables are relevant to output variable (Sale-Price).

Task-8: It was observed that some of the input columns are correlated, and this may make the above analysis unreliable. Redo Task-(7), with the consideration of correlation issue between input variables.

Task-9: It was observed that some of the input columns may not be relevant to the output variable, and this may make the above analysis unreliable. Redo Task-(7), with the consideration of possible unrelated input variables.

Task-10: It was observed that some of the input columns are correlated, and this may make the above analysis unreliable. Redo the above task, with incorporating the correlation between input variables.

Task-11: Predict the estimated Sale-Price values given in Group\_XX\_C file. Consider all the numerical and categorical variables for the analysis. If you skip any column, then provide strong justification. Also, justify your transformation and modification of the columns for the analysis.