Data Analysis

- Repository: `real-estate-price-prediction`
- Type of Challenge: `Consolidation`
- Duration: `5 days`
- Deadline: `14/10/2022 12:30 PM`
- Presentation: `14/10/2022 01:30 PM`
- Team challenge : Solo

Mission objectives

- Be able to use `pandas`.
- Be able to use Data visualization libraries.(`matplotlib`, `seaborn` or `plotly`).
- Be able to clean a dataset for analysis.
- Be able to use colors in visualizations correctly.
- Be able to establish conclusions about a dataset.
- Be able to find and answer creative questions about data.
- Be able to think outside the box.

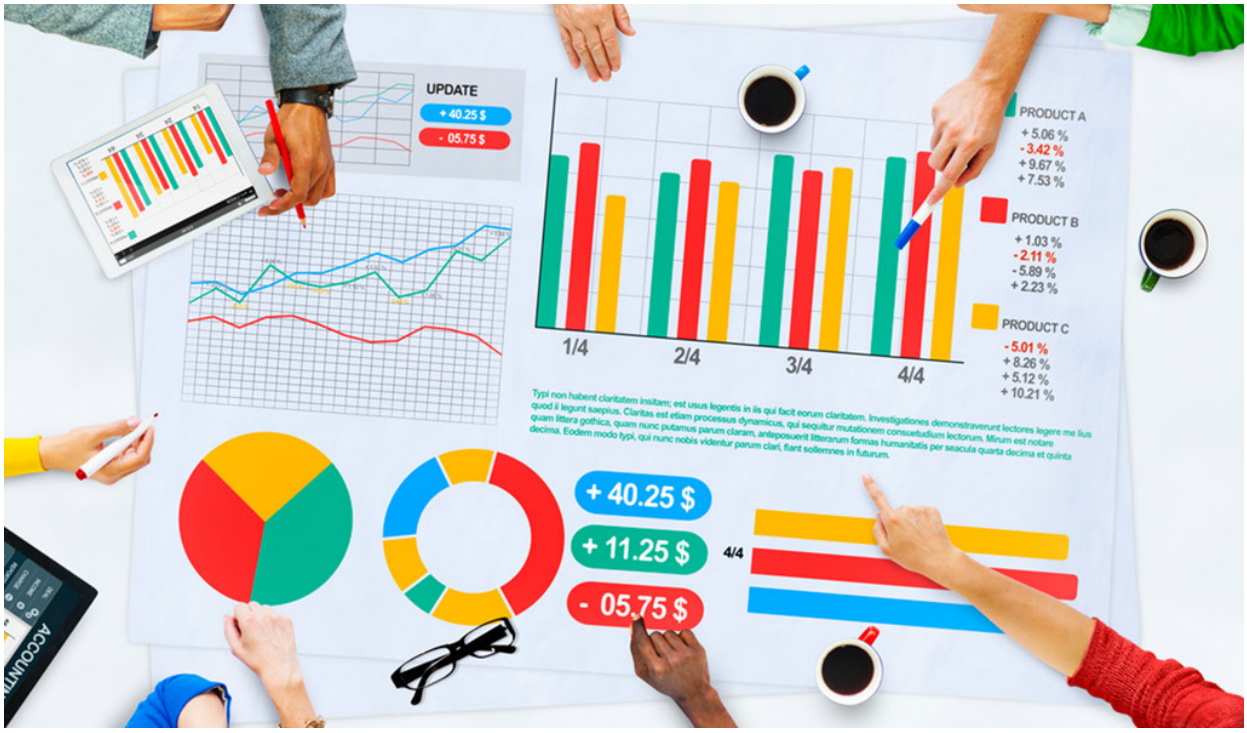
The Mission

The real estate company "ImmoEliza" wants to establish itself as the biggest one in all of Belgium. To pursue this goal, it needs to create a machine learning model to predict prices on Belgium's sales. That way, they can pick out the properties that are the most valuable to them.

But for this, it needs to do a preliminary analysis to gather some information. Having no in-house data scientist, they are looking for talented people to do it for them.

Since your last encounter with them went great, they reached out to you to do this job. Everything is in your hands now!

Take the dataset previously scraped to do the analysis.



Must-have features

Step 1 : Data Cleaning

For this step we strongly encourage you to use pandas. Take the time to go in the material to get familiar with that library which is unavoidable in data science.

You have collected your data! So it's time to do a cleaning on it. A cleaned dataset is a dataset that doesn't contain any duplicates, is blank spaces or error-free. The rest of the analysis can be discarded if you neglect this step!

- No duplicates
- No blank spaces (ex: `" I love python "` => `"I love python"`)
- No errors
- No empty values

Step 2 : Data Analysis

Now that the data has been collected and cleaned, it is time for the analysis. How many variables and inputs do you have? And so on...

Use the tools such as matplotlib / seaborn / plotly. Take the time to play with those libraries before starting.

Answer the following questions with a vizualization if appropriate:

- How many rows and columns?
- What is the correlation between the variables and the price? (Why might that be?)
- How are variables correlated to each other? (Why?)
- Which variables have the greatest influence on the price?
- Which variables have the least influence on the price?
- How many qualitative and quantitative variables are there? How would you transform these values into numerical values?
- What is the percentage of missing values per column?

Step 3 : Data Interpretation

After analyzing your data, it's finally time to interpret your results. You have to communicate your analysis using simple words and a table or graph, then use the results to decide on your best course of action.

For starting, you **must** be able to answer :

- Plot the outliers.
- Which variables would you delete and why ?
- Represent the number of properties according to their surface using a histogram.

130 lines (93 sloc) | 6.52 KB

price, price per square meter)
- What are the **most** expensive municipalities in Wallonia? (Average price, median price, price per square meter)
- What are the **most** expensive municipalities in Flanders? (Average price, median price, price per square meter)
- What are the **less** expensive municipalities in Belgium? (Average price, median price, price per square meter)
- What are the **less** expensive municipalities in Wallonia? (Average price, median price, price per square meter)
- What are the **less** expensive municipalities in Flanders? (Average price, median price, price per square meter)

Then we ask you to prepare some visualizations that bring powerful and original insights about the dataset. Be creative !

Deliverables

As a deliverable, we expect you to give a short presentation of your findings to the team (max. 2 minutes). For this presentation select your two most powerful visualizations and explain them clearly.

Your code should be clean, structured and documented as always. It should be pushed on the repository of the project by using branches and pull requests.

Don't forget to increase and pimp your README file:

- Description
- Installation
- Usage
- (Visuals)
- (Contributors)
- (Timeline)
- (Personal situation)

Steps

1. Create a new branch in your repository
2. Add a folder `data_analysis`
3. Study the request (What & Why ?)
4. Identify technical challenges (How ?)

Plots must-have

- Title
- Legend
- Axis labels (do not forget units)
- Correct usage of colors
- Comparable scales
- No overlapping text
- No screenshots (look how to export the plots in `png`)

Evaluation criteria

Criteria	Indicator	Yes/No
1. Is complete	Know how to answer all the above questions.	
	Use `pandas` and `matplotlib` / `seaborn` / `plotly`	
	Nice presentation on the subject	
	Code is clean and structured	
	README is complete and nicely formatted	
2. Is great	Additional questions were answered.	
	Bonus was answered.	
	The colors are chosen carefully.	

Quotes

“The lottery is a tax on people who don't understand the statistics.” - *Anonymous*