Technical Documentation: Scraping and Cleaning Data from Idealista

Introduction:

The purpose of this project was to scrape data from the website https://www.idealista.com/en/ and clean the data to prepare it for analysis. The data obtained from this website was in a raw format and required cleaning in order to be used effectively for analysis.

Methodology:

1. Data Scraping:

The first step in the process was to scrape data from the website https://www.idealista.com/en/. The scraping was performed using the Beautiful Soup library in Python. The library was used to extract specific elements from the HTML of the website and store them in a list.

1. Data Cleaning:

The second step was to clean the data that was obtained from the website. The data was exported into an Excel spreadsheet and the cleaning process was performed in the spreadsheet. The cleaning process included the following tasks:

* Removing useless rows: Rows that contained no relevant information were removed to reduce the size of the dataset and to ensure that the data was relevant to the analysis.
* Converting the row with parking lots into boolean: The row that indicated whether a property had a parking lot or not was converted into a boolean format. This allowed for easier analysis and interpretation of the data.
* Removing Duplicates: All duplicates were removed from the dataset to ensure that the data was accurate and reliable.
* Improving Readability: The data was made more readable by cleaning up the formatting and ensuring that all values were consistent.
* Dividing the location row into two: The location row was divided into two columns, Province and City. This allowed for a more granular analysis of the data and made it easier to compare properties in different locations.

Conclusion:

In conclusion, the process of scraping and cleaning data from the website https://www.idealista.com/en/ was successfully completed. The data was cleaned to prepare it for analysis and was stored in an Excel spreadsheet for future use. The data will now be used for further analysis and to answer questions about the real estate market in the region.