Property prices in Tartu compared to environmental conditions

Team members: Laur Edvard Lindmaa, Artur Eksi, Argo Aljand

Business understanding

Business goals

Background

Property prices are influenced by many factors, including location, infrastructure, economic trends and environmental conditions. Among these, environmental conditions such as air quality, noise levels and proximity to green spaces play a critical role in determining the desirability of the area. We have chosen to do the data analysis on Tartu as we currently study there and we would like to gather more information on how the property prices are shaped and whether it would be possible to predict them.

Business goals

The primary goal of this project is to analyze the relationship between property prices in Tartu and the environmental conditions to find related patterns. Specifically with this project we aim to identify key environmental factors, quantify their impacts to property valuations and inform the local population.

Business success criteria

The success of this project will be measured by its ability to provide actionable and reliable insights into the relationship between property prices and environmental conditions in Tartu. The success criteria consists of accuracy and reliability of insights, ease of interpretation and relevance to the local community.

Assessing the situation

Inventory of resources

For this project we have the following data resources:

- National noise data (https://avaandmed.eesti.ee/datasets/eesti-strateegilised-murakaardid-(wfs))
- Data published by the city government (https://geohub.tartulv.ee/pages/avaandmed)
- Current property prices of Tartu (https://www.kv.ee/)

Regarding human resources we have three people, who are all students in the University of Tartu. All of whom are in computer engineering and robotics master's.
From the computational power perspective each of the project members has access to a laptop which can be used to perform necessary data analysis.

Requirements, assumptions and constraints

This project has multiple requirements to ensure the quality of the results:

- Data requirements
    - Comprehensive property price data giving a good overview of the prices in each city district
    - Accurate and up-to-date environmental data
    - Relevant contextual data, such as infrastructure details
- Analytical and computational requirements
    - Statistical tools and techniques for identifying correlations and trends
    - Sufficient computational power to process the entire data before the deadline
- Reporting requirements
    - Clear and accessible visualizations and documentation for communicating findings to non-technical people

The following are the projects assumptions:

- The datasets are sufficient to represent Tartus real estate market and environmental conditions
- External factors have a consistent or minimal impact to the prices
- Environmental factors like air quality and noise pollution influence property prices significantly

Finally there are some constraints that have to be taken into account:

- The project has to be completed by the 9th of december, meaning the project has to be done within a week
- Limited availability of real estate market data as it is dependent on the people making the posts, not real prices that they actually get traded with in the end
- Reliance on publicly available data, which might be limited
- The methods used might not capture the factors as intended and may oversimplify the findings
- Some of the data might not be available

Risks and contingencies

Data related risks:

- Missing or incomplete real estate prices or environmental data
    - Contingency: Identifying alternative data sources such as other real estate trading websites or other publicly available datasets.
- Inconsistent data formats and resolutions (for example different data uses different coordinate systems)
    - Contingency: Normalizing and aggregating data to a common scale

Interpolation methods can also be used when there are minor gaps in data.

Analysis related risks:

- Weak or non-significant correlations between environmental conditions and property prices
    - Contingency: Explore interactions with additional variables, for example with proximity to infrastructure or other socioeconomic factors
- Overfitting the statistical models, leading to wrongful results
    - Contingency: Use techniques like cross-validation to improve model robustness

Resource constraints:

- Insufficient time to complete a thorough analysis due to project deadlines
    - Contingency: Focus on the smaller areas of interest first and scaling down the non-essential tasks
- Limited computing power for processing the dataset if it should become too large
    - Contingency: Use cloud based solutions or reduce dataset size through sampling techniques

Terminology

- Real estate prices - The monetary value at which properties are bought or sold in Tartu. This can include average price per square meter or the total price. Relevance: Understanding the fluctuation of real estate prices is one of the goals of this project
- Environmental conditions - Refers to various environmental factors that could influence property prices, including air quality, noise pollution and green space availability. Relevance: This project aims to find correlations between property prices and environmental conditions.
- Noise pollution - The presence of harmful or disturbing levels of sound, often measured in decibels (dB). Sources can be traffic, industrial activity and constructions. Relevance: Noise pollution can be a significant factor in property price fluctuations.
- Green space - Public areas in urban environments that contain a lot of vegetation, for example parks and gardens. Relevance: Access to green spaces can increase property prices due to the aesthetics and health benefits they offer.
- Correlation - A statistical relationship or connection between two variables, indicating how they change in relation to each other. Relevance: The analysis will look for correlations between environmental factors and property prices to understand potential impacts.
- Geospatial data - Data that is associated with a specific location, often used for mapping and analysis. Relevance: Geospatial data will be used to understand the spatial distribution of both property prices and environmental factors.

Costs and benefits

Regarding this project there are no costs planned as all of the data used will be open source and cloud based computational power shall be looked for for free.

This project comes with benefits of increasing public awareness on how different sources of pollution affect real estate prices and it will give the people more informed real estate market insights.

Data-mining goals

The data-mining goals of this project are to:

- Identify correlations between environmental factors and property prices
- Predict property prices based on environmental and other relevant variables
- Uncover hidden patterns in real estate price fluctuations
- Identify environmental factors that can increase the property value
- Segment areas based on environmental and market characteristics

Data-mining success criteria

The success criteria consists of the following items:

- The created model has a high classification accuracy
- The results would show meaningful patterns in pollution and pricings
- Model is robust and must be able to generalize well to new, unseen data and not be overfit
- The data that is gathered is complete and contains only minor gaps at most
- The results of the model should be interpretable by non-technical people

Data understanding

Gathering data

Data requirements

As the plan is focused on the city of Tartu, all data needs to describe different parameters affecting the living conditions or real estate pricing in Tartu. Additionally the data cannot be generalised data on the entire city or district as factors such as noise pollution and prices can vary greatly within city blocks. Furthermore the data needs to have sufficient variability across the town to yield meaningful analysis.

Data availability

The city government and different ministries gather data on the environment, as the second largest city in the nation a lot of the data is gathered from Tartu which gives a good selection of publicly available data. Additionally third party sources such as KV.ee for real estate pricing are available.

Selection criteria

Data sources that we are using are publicly available and are accessible from the following links:

- Noise pollution map: https://avaandmed.eesti.ee/datasets/eesti-strateegilised-murakaardid-(wfs)
- Accessibility map detailing the level of access to various everyday facilities: https://geohub.tartulv.ee/maps/f2e1d970c4284f2d899a810508b253f0/about
- Property pricing with size and location in Estonia: https://www.kv.ee/

The noise pollution map is a heatmap containing numerical data about noise levels recorded at various locations across Estonia. This information allows us to analyze specific areas in Tartu and assign a noise pollution index to each property for further evaluation.

The accessibility map provides numerical data to assess the ease of access to everyday facilities for properties in different locations. Using this data, we can assign a numerical rating to each property based on its accessibility.

Data from the property pricing website, kv.ee, allows us to extract detailed information about various properties, including their size, location, and price. By mining this data, we gain essential insights into property characteristics.

In conclusion, we utilize two geographical data sources, extracting one feature from each to analyze and evaluate properties in Tartu. By combining these features, we examine potential correlations between property pricing, noise pollution levels, and accessibility to facilities. Additionally, we incorporate three supplementary features—price, location, and size—to enhance our analysis and provide a more comprehensive understanding.

Describing data

The data for this project is sourced from three datasets: noise pollution data, accessibility data, and property data. The noise pollution data provides numerical geospatial values measuring noise levels in Estonia, focusing on Tartu, with its key feature being the noise pollution index for each location. The accessibility data, derived from a map, is categorical geospatial data with three categories assessing ease of access to facilities, providing an accessibility rating for each location. The property data, obtained from kv.ee, is tabular and includes property size, number of rooms, location, and price. These datasets comprehensively cover properties in Tartu. Fields include a noise pollution index, accessibility rating, property size, location, and price. Geospatial and tabular data require integration using location-based matching. Data validation ensures accuracy and consistency across geographic regions. Normalization of numerical values and addressing any missing or incomplete data are essential steps to maintain data integrity. This preparation ensures the datasets are suitable for analysis, enabling a thorough evaluation of property-related factors.

Exploring data

The noise pollution map provides a numerical value in decibels representing the noise levels in specific areas. This data will be analyzed to determine the range and distribution of noise levels across the dataset.
The accessibility map includes a categorical feature with three distinct categories: good accessibility, base accessibility, and unknown accessibility. These categories can be converted into numerical values, with good accessibility assigned a value of 1, base accessibility assigned 2, and unknown accessibility assigned 0. The unknown category is predominantly associated with areas near the city edges, where accessibility is generally assumed to be worse than in the city center.
The property data from kv.ee includes three key features: size, price, and location. The size and price features are numeric, while location is categorical, representing the geographic area of the property. Since size, price, and location are mandatory fields for posting listings on the website, their completeness can be reasonably assured.

Verifying data quality

The collected data is sufficient for analyzing correlations between property prices, noise pollution, and accessibility. The noise pollution map provides reliable numerical data without significant quality issues. The accessibility map's categorical data can be converted into numerical values, with "unknown accessibility" assigned 0 due to its likely lower accessibility at city edges. Property data from kv.ee, including size, price, and location, is complete and consistent, as these are mandatory for listings. No severe data quality issues, such as missing fields or inaccuracies, have been found. Minor challenges like data normalization and merging geospatial and tabular data are manageable. The data is suitable for analysis, and no alternative sources are needed.

Project plan

Ma actually ei tea.. Mõelge kratile (mulle) midagi välja ja andke teada mida vaja. Mul on kamputer power ja viitsimist vahest brute force midagi teha :3

1. Data Collection and Validation (6 hours per member)

- Collect noise pollution, accessibility, and property data.
- Validate data accuracy, consistency, and completeness.
- Address missing or incorrect values.
  Methods and Tools:
  - Tools: Python (pandas, NumPy), QGIS.
  - Comments: Data integration may require geospatial matching and preprocessing to ensure compatibility.

2. Data Preprocessing (8 hours per member)

- Normalize numerical data (noise index, price, size).
- Convert categorical accessibility data into numerical values.
- Integrate geospatial and tabular data using location-based matching.
  Methods and Tools:
    - Tools: Python (scikit-learn), SQL for integration, GeoPandas for geospatial processing.
    - Comments: Careful handling of unknown accessibility categories is crucial for accuracy.

3. Exploratory Data Analysis (6 hours per member)

- Analyze distributions, detect outliers, and visualize data.
- Summarize trends and identify potential correlations.
  Methods and Tools:
    - Tools: Matplotlib, Seaborn, Jupyter Notebook.
    - Comments: Visualizations will help identify initial hypotheses and inform model preparation.

4. Model Development and Analysis (10 hours per member)

- Create regression models to assess correlations between price, noise pollution, and accessibility.
- Validate model performance using cross-validation techniques.
  Methods and Tools:
    - Tools: Python (statsmodels, scikit-learn).
    - Comments: Ensure sufficient data splits for training and testing to avoid overfitting.

5. Report and Presentation (5 hours per member)

- Compile findings into a final report with visualizations and interpretations.
- Prepare a presentation summarizing the project.
  Methods and Tools:
    - Tools: Microsoft Word, PowerPoint, Google Slides.
    - Comments: Include actionable insights and any limitations encountered.

Important Notes

- Total time contribution per member: 35 hours.
- Clear documentation is essential at every step for transparency and reproducibility.
- Collaboration will be facilitated through a shared workspace, such as Google Drive and GitHub, to ensure seamless team coordination.