Business understanding

Background

We want to analyze the trends in the Estonian real estate and housing market. The housing market is a dynamic and multifaceted ecosystem influenced by a myriad of factors. In order to gain a comprehensive understanding of the housing landscape, a detailed analysis of housing prices by district is imperative. We want to unravel the intricacies of housing price variations, identify long and short term trends in the market and possibly even predict future housing prices.

Business goals

Market Understanding: Our primary business goal revolves around fostering a comprehensive understanding of the Estonian housing market, with a specific focus on dissecting the dynamics across various districts. By delving into the intricacies of market trends and identifying areas of high and low demand, we aim to provide readers with a nuanced perspective on the current landscape.

Predictive Analysis: Building upon a foundation of historical time series data, our secondary objective is to develop predictive models that illuminate future movements in housing prices. By analyzing trends and patterns, we seek to equip readers with insights that go beyond mere retrospective observations. This predictive analysis serves as a vital tool for strategic planning and decision-making, offering a forward-looking perspective to guide investments and navigate the ever-evolving Estonian real estate market.

Business success criteria

Data collection, coverage and refactoring: One key point of our project's success is finding and gathering relevant and useful data about the housing market. The public sources that give historical data about the real estate market in Estonia are limited and don't contain a lot of information. The reshaping and manipulation of that data also plays a significant role in the success of our project.

Accurate Predictive Models: The secondary criteria of our project's success lies in the development of data heatmaps and predictive models that exhibit a high degree of accuracy in forecasting housing price and trend movement. These models are designed to leverage historical time series data, discern patterns, and extrapolate future trends, for anticipating market dynamics.

End goal of the analysis: The analysis can be reflected as an success if we can find and visualize specific trends, such as:

- Price and advertisement count differences (percentage) in Tartu and Tallinn districts. Displayed on geospatial plots as a heatmap for the following timeframes: long term (15 years), medium term (7 years), short term (2 years).
- Price and advertisement count changes for Tartu and Tallinn districts on line plots.
- The current average price and advertisement count for Tartu and Tallinn districts as a geospatial heatmap.
- The average housing price in relation to average income (without tax) in Tartu and Tallinn.
- The average housing price and advertisement count in relation to household loan interest rates in Tartu and Tallinn as a line plot.

Inventory of resources

Our resources for gathering data about the housing market are real estate websites that we can scrape with automatic bots.

1. https://www.kv.ee/hinnastatistika (main data)
2. https://andmed.stat.ee/et/stat/Lepetatud_tabelid__Majandus.%20Arhiiv__Palk%20ja%20toojeukulu.%20Arhiiv/PA21 (average quarterly income 2000-2017 by county)
3. https://andmed.stat.ee/et/stat/Lepetatud_tabelid__Majandus.%20Arhiiv__Palk%20ja%20toojeukulu.%20Arhiiv/PA004/table/tableViewLayout2 (average quarterly income 2018-2022 by county)
4. https://andmed.stat.ee/et/stat/majandus__palk-ja-toojeukulu__palk__luhiajastatistika/PA119 (average quarterly income 2021-2023 by county)
5. https://statistika.eestipank.ee/#/et/p/979/r/3997/3746 (household loan interest rates 1997-2023 by month)

Requirements, assumptions, and constraints

Requirements:

1. A significant amount of historical time series market data
2. Tools for creating geospatial heat maps
3. Machine learning models for predicting future values of time series data

Assumptions:

1. The real estate market follows reliable trends
2. Price movement and sales activity are major indicators of market activity
3. The geospatial shape files used for plotting for our desired districts is publicly available

Constrains:

1. Limited knowledge about real estate
2. Limited amounts of publicly available housing market data

Risks and contingencies

1. The geospatial shape files needed for plotting heat maps are not publicly available
2. Unpredictable price changes**:** One of the biggest risks is that the housing market is so inefficient that it's borderline impossible to predict. In such a case, developing an accurate prediction model is not a primary objective.

Terminology

Active advertisements - Active ads average count during a month.

Advertised listings - Total active ads during the month.

Price - Average m2 price during the month.

District - Part of a city.

Costs and benefits

Costs: Time

Benefits:

a. Potential investment strategies
b. Better understanding of the housing market and the trends present
c. Experience working with time series and geospatial data

Data-mining goals

For this project, our data-mining goals are centered on extracting valuable insights from the available data to comprehensively understand and predict trends in the Estonian real estate and housing market. We aim to uncover patterns and trends that contribute to housing price variations. The data-mining process will focus on identifying key drivers of market dynamics, such as demographic trends and economic indicators.

Our overarching objective is to develop a refined understanding of how these factors interplay within different districts, allowing for a nuanced analysis of the real estate landscape in Estonia.

Additionally, the data-mining goals encompass the creation of predictive models that leverage historical time series data to forecast future housing price movements. These models will serve as tools for making informed decisions and anticipating market trends in the dynamic real estate environment.

Data-mining success criteria

Data-mining success criteria include finding a significant amount of public market data, uncovering meaningful patterns, and housing price trends then possibly developing an effective predictive model.

Data understanding

Gathering data

- Outline data requirements
    - Monthly time series (at least 10 years)
    - Average monthly price and number of advertisements organized by county and district
    - No big caps for long periods in district data. This would make finding trends, analyzing price action and predicting future values using time series forecasting a lot more difficult.
- Verify data availability
    - All the needed data is available and provided by www.kv.ee and accessible here: https://www.kv.ee/hinnastatistika. The data is stored as a table and contains relevant monthly market info for the past 16 years. Since the data is not given as a dataset nor is it available through an API, we have to use web scraping in Python to gather it and format it into a csv file.
- Define selection criteria
    - We will start our analysis with the data we have scraped to the csv file. We chose to start with only gathering data about Tallinn and Tartu, because that way we can be sure that enough data per district is present. Since we tend to use almost all the data fields and the whole time range provided by www.kv.ee, all the relevant info for this project is already stored in our csv file after scraping.

Describing data:

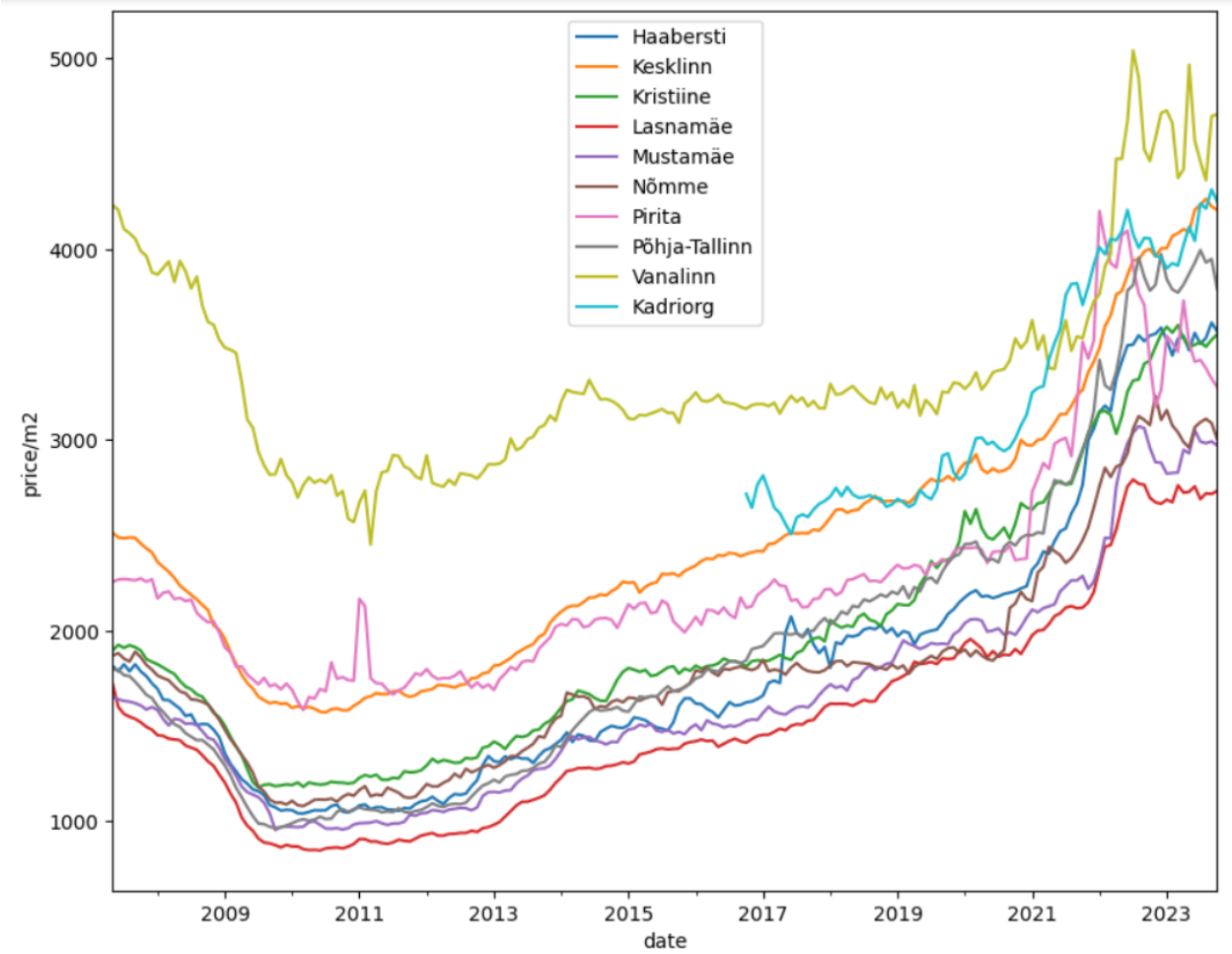
The data we have gathered is grouped into 6 columns:

1. Date - We have collected data for every month since 05.2007, in exception to some districts that don't have readily available data going back that far. The data is formatted as month and year put together, example: 12-2018.
2. City - We have data on two cities: Tallinn and Tartu.

3. District - For each city we have every district of that city and some close-by towns.
4. Price - Average price of living spaces sold as represented by a number calculated as totalPrice/m2. Usually ranging from
5. Advertisements - The total amount of advertisements posted to the site in a specific month.
6. Active advertisements - The total amount of advertisements active on the site in a specific month.

Exploring data

First thing that we saw when plotting the data was that ‘Kadriorg’ did not have any data before 2016. This is something we have to keep in mind when moving forward and calculating price change percentages and plotting data past a certain point in time. If we analyze the trend by looking at the simple plot of the housing price changes, we can see that most values are in the range 1000 - 5000 eur/m^2. The first time series values of our data are from 2007. From that point we can see a down trend in the housing prices that ends at around 2010. From 2010 till present day we have a bullish trend. The same trends apply for both Tallinn and Tartu.



Verifying data quality

Overall, the data quality is quite reliable. Some districts that have less sales activity have less accurate data. That's why we only collected data about Tallinn and Tartu. Smaller towns have so small sales activity that we won't get any noteworthy data out of them.
As You can see, for some reason, Kadriorg has data starting from 2016. That means that if we want to use longer timeframes in our research, we have to deal with the missing data.

Project plan

1. Project planning - 5h
2. Python scraper for web scraping kv.ee data - 5h
3. Tallinn and Tartu geospatial map creation - 10h
    a. Finding publicly available geospatial maps for both cities
    b. Learning to work with and plot our data onto the geospatial plots (for example as heatmaps)
4. Creating plots for analyzing housing market trends ~25h
    a. Heatmaps for displaying price/m^2 by district in Tallinn and Tartu - 5h
    b. Heatmaps for displaying relative price change by district over different time periods in Tallinn and Tartu - 10h
    c. Heatmaps for displaying popular districts in Tallinn and Tartu - 10h
    d. …
5. Investigating new interesting possible data combinations to display on new maps - 10h
    a. Identifying possible factors that are relevant to and in correlation with the housing market prices
    b. Finding new time series datasets that we could incorporate into our geospatial plots.
6. District popularity and average housing price time series forecasting model - 20h
    a. Learning about different possible time series forecasting models (not covered in our course material)
    b. Training different models on training data and evaluating their accuracy on test data.
    c. Possibly finding and adding into our models as variables other relevant time series datasets that cause changes in the housing market.
7. Creating poster for the project: 15h (Discussion about the subjects we choose to show on the poster, designing the poster, creating the poster and presenting the poster)

Total time predicted: ~90h