README

This file gives an overview of this coding project and how it is organized.

Table of Contents:

Purpose

The purpose of the program is to help decide which apartment to buy.

Consideration	Risk	Reward
1.	cost to purchase	savings left over
2.	depreciation during time I live there	appreciation during time I live there
3.	inability to get paying roommate	ability to get paying roommate
4.	inability to find work if unemployed	ability to find work if unemployed
5.	distant from friends	close to friends
6.	distant from social and networking events	close to social and networking events
7.	distant from forest and water	close to forest and water
8.	distant from central train station	close to central train station
9.	will want to move again	will want to stay

Considerations 1. and 2. will likely be modelled with some kind of multiple linear regression model.

Consideration 3. will have to do with first if the apartment is at least 2.5 rooms and the location. Will have to find a way to model: occupancy rate of rentors in the market * rate I can charge a rentor for the market.

Consideration 4. will have to do with unemployment rates in the two markets. I may be able to buffer myself from this if I can negotiate 100% remote work.

Considerations 5.–7. all intuitively favor Stockholm over Uppsala, but I haven't considered how to model them yet.

Consideration 8. mitigates some of the other potential negatives if this is positive.

Consideration 9. makes me realize that I do not want to go through this process over and over... Since I know I cannot currently afford to get what I want, I do not need to stress about solving this problem right now. Instead, I can keep renting and saving until I know a bit more about what I want (perhaps I will get a partner in the meantime who will want to help pay, then it becomes a co-decision and the budget also changes?). I also have not been considering all of the other costs involved in buying a property (capital gains tax if sell, maintenance, real-estate agent fees, potential property taxes). If I'm not able to afford something I really love (i.e., it falls into the reward category for basically all of the considerations above), then it's not really worth all of the headache involved and I'd rather risk wasting some money on renting while continuing to live a more hassle-free lifestyle.

Scraping and pre processing

- Clean file names- put in folder with date scraped
- Script that checks latest date scraped, runs scraper program for new hits since last scraped date so as to not re-scrape everything
- Distribution plots for QA
- Feature engineer dates, etc
- Subset to only include 5.5 msek selling price or less

Start building models

Models to build

1. Regression to predict the final sales price. This determines how much I should be willing to bid at a maximum.
    1. Inputs: all data I can find on sold apartments in the last 10 years with a final price of 5 800 000 SEK or less.
2. Regression to predict appreciation.
    1. Inputs: all data I can find on apartments that were bought and sold more than once in the last 10 years where the first price was 5 800 000 SEK or less.

Kinds of models

1. Linear regression
2. Random forest regression
3. Neural network regression
4. Ensemble method to bring them together?

Splitting the data for tuning and grading the models

1. Split into train and test
2. Split the train set further inte train and validate for optimizing hyper-parameters of model
    1. Use k-fold cross-validation for tuning

Scoring/grading the models

1. Accuracy

2. Precision
3. Recall
4. Sensitivity
5. Specificity
6. ROC curve
7. FA score
8. Youden's J
9. F1 score

Variables to consider bringing in

1. Lat/Long of addresses

Structure

The program is written predominantly in R. Not all files are published publicly.

File Organization

- `.gitignore`
  - File with instructions of what not to make public (not public).
- `00_readme/00_readme.md`
  - The file you are currently reading.
- `01_file_organization.pptx`
  - A powerpoint to visualize the dendrogram of the file structure (in planning).
- `02_ws.code-workspace`
  - VS code workspace (not public).
- `03_batch_program.r`
  - Running this program will run all other scripts, process the data, and output it to the target files.
- `04_my_fxns_this_project.r`
  - Define functions and parameters for this project (not public).
- `05_scripts/`
  - Contains the scripts run by `01_batch_program.r`.
    - `<>.r`
      - Description of specific program...
- `06_inputs/`
  - Contains the files containing data used by the scripts (not public).
- `07_outputs/`
  - Contains the files newly generated by the scripts.
    - `<>`
      - Description of specific output file...
- `08_logs/`
  - Contains logs of when `01_batch_program.r` was executed (not public - in planning).