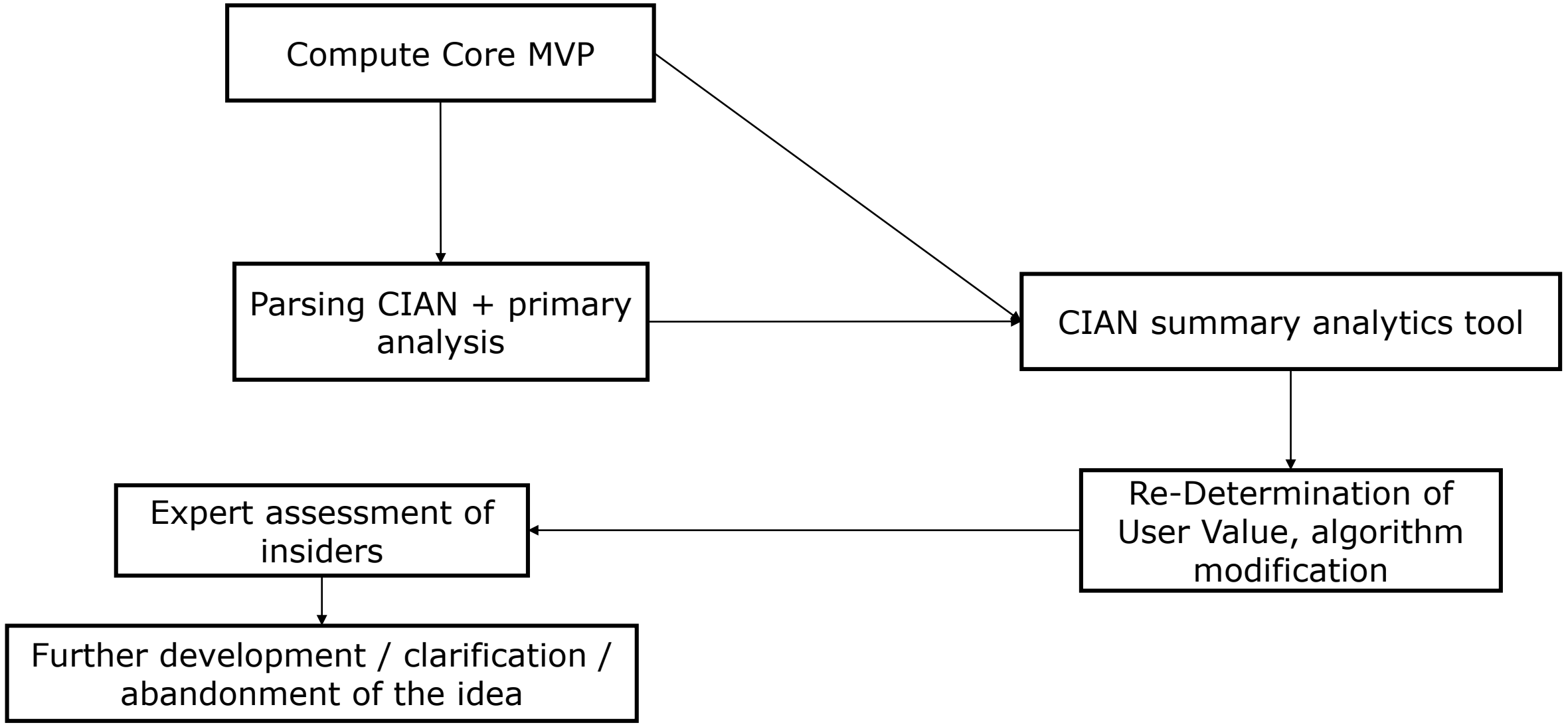
Scheme of the project Service for searching for interesting offers for the purchase of commercial real estate



Part No. 1: CYAN Parsing + Primary Analysis

Objectives: to upload advertisements for the sale of commercial real estate of the third transport ring of Moscow, the cost of which is below the market value.

Input data: data on the sale of real estate in Moscow from the CIAN website.

Problems: the presence of fakes.

Tasks:
1. Develop a dataset parser from the CIAN website (a minimum set of factors: latitude and longitude of the location of the house, total area, number of rooms, wall material, floor of the apartment and number of storeys of the building, address, district, type of sale);
2. Process omissions and search for anomalies;
3. Primary analysis of factors, feature engineering based on business logic (example - distance to the center);
4. Output ads to .xlsx document with a cost below the sample average of similar ads.

Methods and libraries: Jupyter Notebook, Python 3, pandas, scikit-learn, matplotlib, geopy.

Summary of the result of Part No. 1 (April 14 – May 22)

Sprint result: a script was developed for unloading ads every 1-7 days and an algorithm for estimating the market value (calculated using algorithm A).

Algorithm A:

1. Parsing
   1.1 Extracting data from commercial ads according to expert criteria - uploading the HTML code of the corresponding pages, the result is N HTML pages from the CYAN search query;
   1.2 Scrubbing - unloading data from HTML code; The result is a dataset of N*28 ads.
2. Data preparation
   Cleaning data from unwanted objects and other noises, converting factors into the desired format (for example, district to distance to the geographical center), the result is a dataset ready for modeling;
3. Simulation
   3.1 Data segmentation - selection of subsamples for model training;
   3.2 Predictive models for cost estimation are trained;
4. Interpretation
   Result: top interesting listings by market valuation-actual value.

Part No. 1 discussion

Possible ways of development:

- The main thing is to learn how to catch the sale of a business. There is an idea to start by searching for the word "rent" from the ad text, the problem is that the text in the HTML is mixed up in places.
- Use k nearest neighbors (increases interpretability) and wall material factor.
- Shell development, first GUI drafts - http://researchmachine.pythonanywhere.com/
- Expansion of the list of observations.

Part No. 2: **Analysis using Algorithm A (Sprint 1) to extract User Value - experiment based on available CIAN data**

Tasks:

1. Eliminate the deficiencies in the data found in Sprint 1 (add some factors, eliminate the sale of the business) and update the model. Prepare simulation results to evaluate User Value in the results of algorithm A.
2. Evaluate the User Value in the results of Algorithm A.
3. Assess the success of the idea: Assess how successfully the idea with Algorithm A can extract User Value from the available data and find out the reasons why this may be impossible or difficult.
4. Conduct an analysis to determine the possibility of eliminating shortcomings when extracting User Value using the available data.
5. Offer recommendations: Based on the results of the experiment, offer recommendations for further actions based on testing the idea with algorithm A in extracting User Value. If the idea is not successful enough, propose alternative approaches or improvements to the algorithm to achieve the expected User Value.

Summary of results of Part No. 2

Results of sprints 2, 3: strengthening filters, feature engineering and segment expansion led to a significant improvement in the Top; the boundaries of previously downloaded data from CIAN for modeling were determined (only for the task of ranking within the sample).

More:

I. Current Value of User is a list of N (from 20 to 70) advertisements of commercial real estate in Moscow on CIAN, which currently stand out from the crowd, taking into account geolocation and other factors specified in the advertisement. To achieve this particular goal, there are no fundamental problems regarding the set of factors. This can be useful, for example, for monitoring several cities for advertisements that want to be dropped quickly.

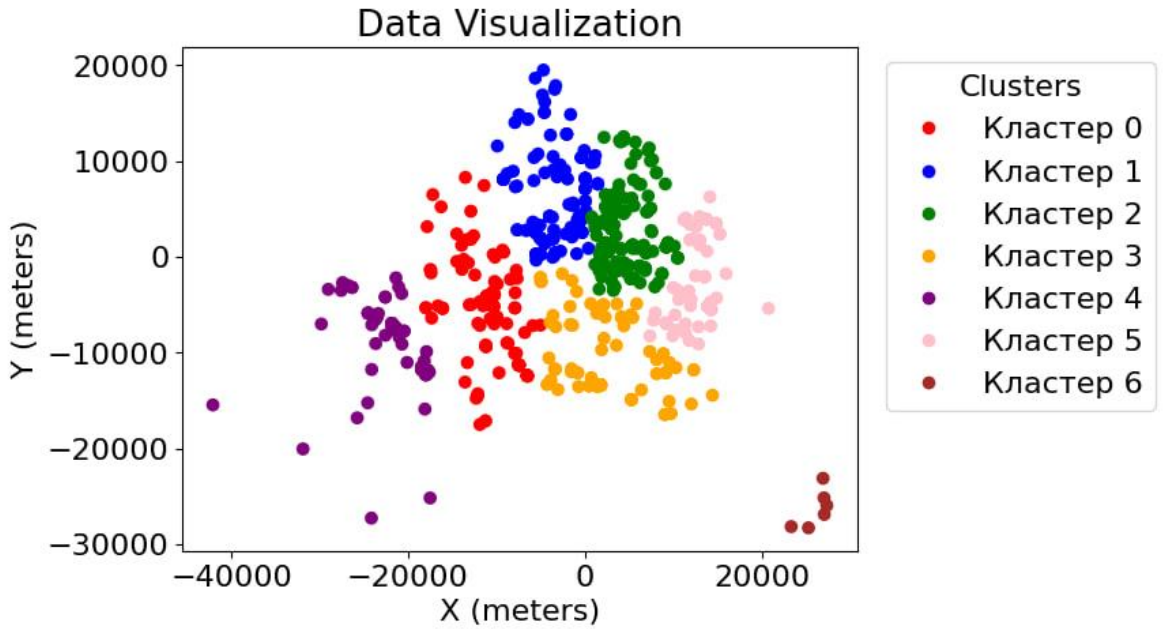
II. I found a manual on professional methods for assessing the market value of real estate by L.N. Tepman. (corresponding member of the Russian Academy of Sciences) "Real Estate Valuation", 303 pp., 2005 (see page 231). A good book for feature engineering. The main conclusion is that imitating a professional appraiser is unrealistic. Therefore, our algorithm performs only the ranking task; using this algorithm for price recommendations is incorrect from the point of view of professional evaluation methods.

III. Strengthening the filters for the sample for May 31 allowed us to filter out more than 40% of unwanted ads (111 ads out of 253).

Summary of results of Part No. 2

IV. Feature engineering made it possible to take into account geolocation when calculating the average cost. This is very important to consider when we single out the "middle" segment.
V. The algorithm showed good results in the segment from 20 million to 100 million. One target offer with an obvious low cost got into the top
https://www.cian.ru/sale/commercial/230805591/

Other: in addition to the filters from point III, I tried to remove warehouses and car washes (according to the description of the ad), they heavily clog the Top and it seems that they have their own cost formation features.



Rice. Geographic clusters of commercial advertisements (structures) CIAN in Moscow