Data requirements and selection

To solve the main business problems a data scientist needs additional information. A good practice is to use official public databases.

1. The main database for these business problems is Foursquare. It's convenient for exploring the city neighborhoods. We may collect all the beer bars, shops according to them locations. Also we may collect data about another public places (for example: public parks etc.), but there are some problems to use this database. This database does not have enough information about Moscow city, so I decided to extend Foursquare database with additional queries to Yandex GEO API https://tech.yandex.com/maps/.
   This API has a free version. In the free version of the API, the total number of requests must not exceed 25,000 requests per day so all the data will be collected to local files in some days. Also the collected data may be corrected according to Open Street Map https://www.openstreetmap.org/ (if necessary).

   In this part I'll extract all common features of the public palaces and save them to "JSON" format (it has approximately similar format for these geo services).
   The main features are:
   1. Total count of bear shops per district (or part of district)
   2. Total count of public places (parks etc.)
   3. Total count of elementary schools (may be helpful because of prohibition law to sell alcohol near the schools)

2. The Administrative division of Moscow is available on Wikipedia web site.
   https://en.wikipedia.org/wiki/Administrative_divisions_of_Moscow

   The final table:

District Name	Latitude	longitude

3. To indicate the standard of living may be helpful to collect data about total value of real estate in the city by districts. All the data of this year represented by Moscow Marker Indicator.
   https://www.irn.ru/rating/moscow/

   The final table:

District Name	Total value per square meter

   From this site is useful to collect average rental price:

District Name	Average Rental Price

4. For additional standards of living exploring the main database is official Federal State of statistics. Moscow city represents in amount of several excel and word files.
   http://moscow.gks.ru/wps/wcm/connect/rosstat_ts/moscow/ru/statistics/standards_of_life/

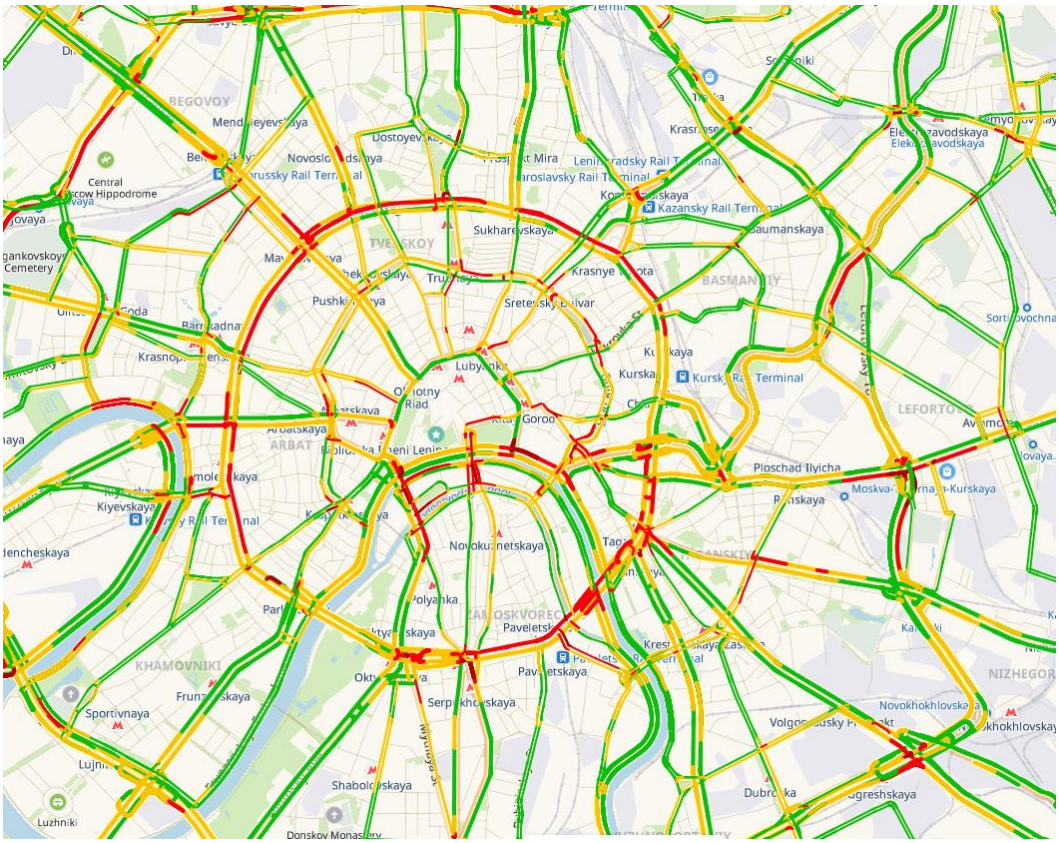
   Main features in this section:
   1. Total approximate amount of month salary (grouped by district). It may be collected only approximately by the whole site statistics.
   2. Structure of residents
      a. Sex (male/female)
      b. Average age of residents
      c. Total resident count per district.

5. For the calculation of distance between some city points in Moscow the best choice is Yandex maps.
   https://yandex.com/maps/213/moscow/?l=trf%2Ctrfe&ll=37.633950%2C55.754925&z=13. It also may be correct by additional database (if necessary).
   For example https://maps.openrouteservice.org/

   The final table:

Probably place Name	Total distance from home (km)

6. To explore the Moscow traffic jam the best database is Yandex. https://yandex.com/support/maps/concept/stoppers.html
   Yandex stores all the traffic jam statistics by all days and hours. This service is not free but I can collect this data manually (not by API) if necessary.



Average traffic jam map on Wednesdays at 18:05

The final table:

Probably place Name	Day Of Week	Total time to distance

Is this section may be useful to collect data by time. But I have to connect my customer and ask him.

7. To explore additional beer shop franchises (if necessary) may be used:

    1. Russian Business Consulting web site http://biztorg.ru/franchises

(All the data in Russian language only)

2. Top Franchises Russia aggregator
   https://topfranchise.ru/catalog/
   (All the data in Russian language only).

The final table in this section:

Trademark Name	Capital investment	Estimated Monthly Profit	Approximate payback period

Feature engendering

Most of the data is well structured but some of that should be collected by a grabber or manually.

In this section the best practice is to apply feature engendering.

For example:

1. Distance rate (speed) = Total distance from home / Total time to distance (according traffic jam).
2. “Crowded” rate = Total count of resident (in area) / total count of public places near the point.

This section might be corrected a few times when the main algorithm is adjusted.

Result of data collection

The final result of this section:

Tables:

1. All the data collected about the city districts (data fields are described above).
2. All the data collected about the possible franchises is this district (data fields are described above).

All the data has to be converted in uniform format (for example .csv) and saved to disk to avoid additional access to limit GEO services and etc.