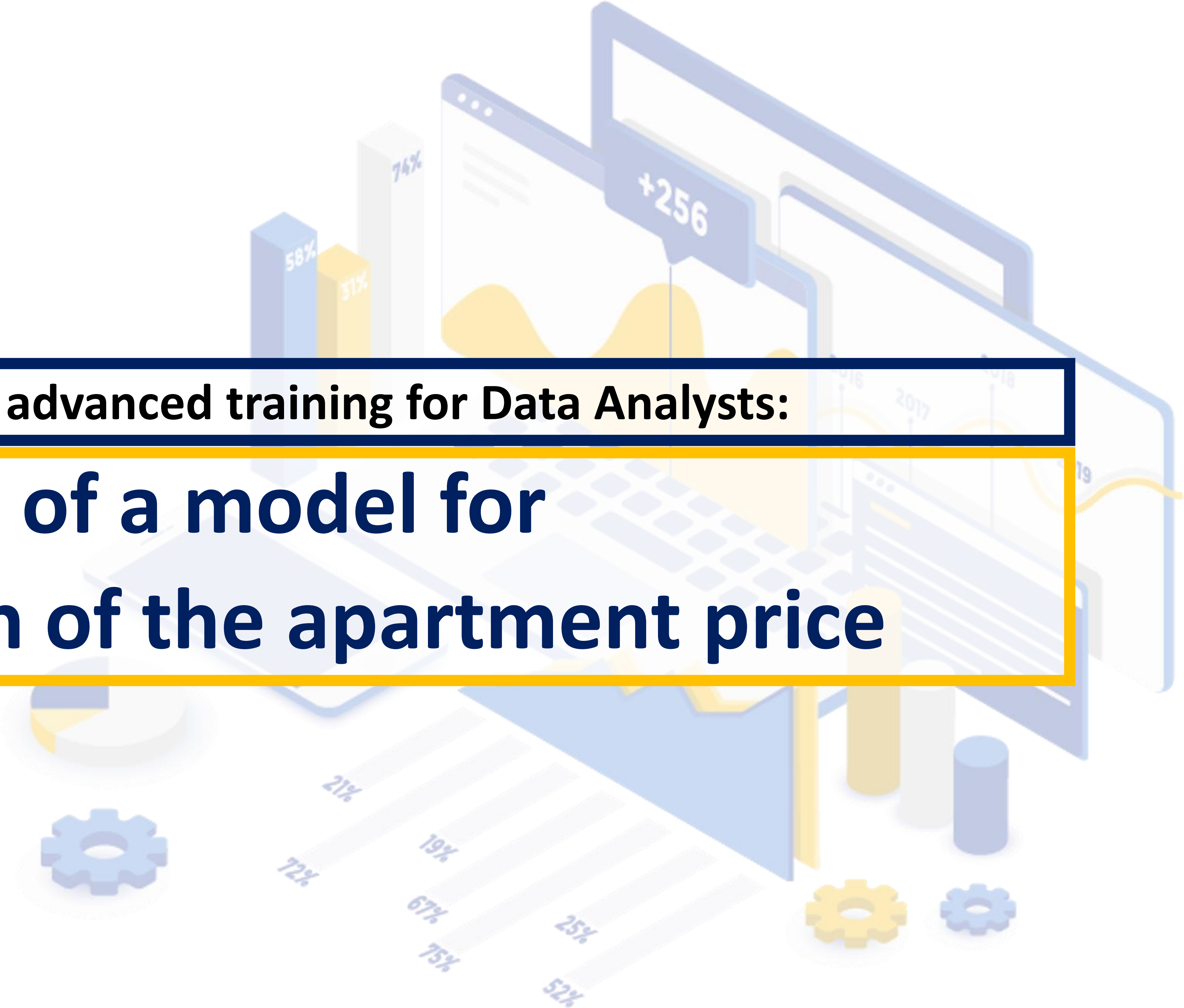
The final project of the advanced training for Data Analysts:

Development of a model for the prediction of the apartment price



The main idea

Aim:

build a model to predict the cost of the apartment.

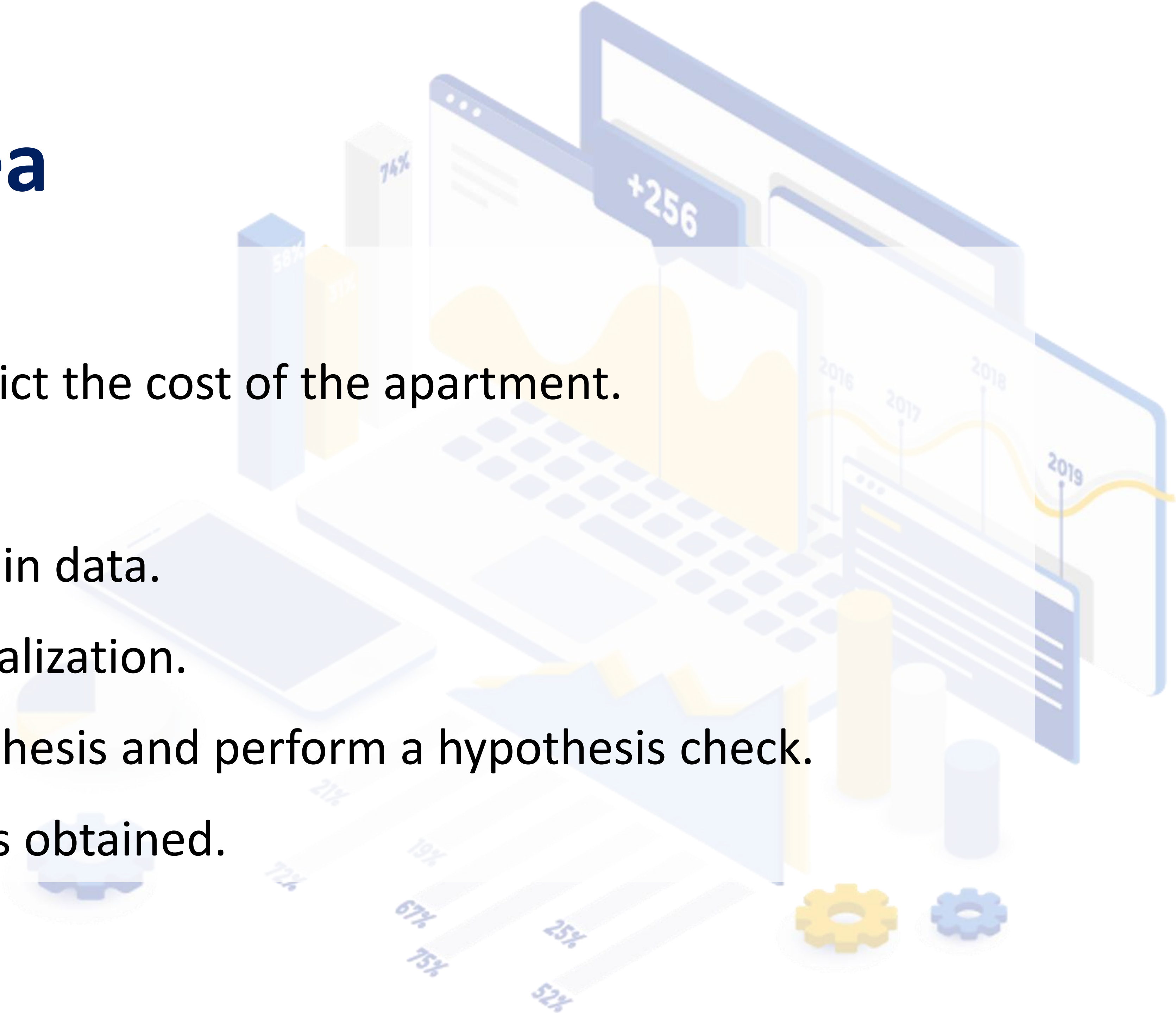
Tasks:

→Search for patterns in data.

→Perform a data visualization.

→Create a data hypothesis and perform a hypothesis check.

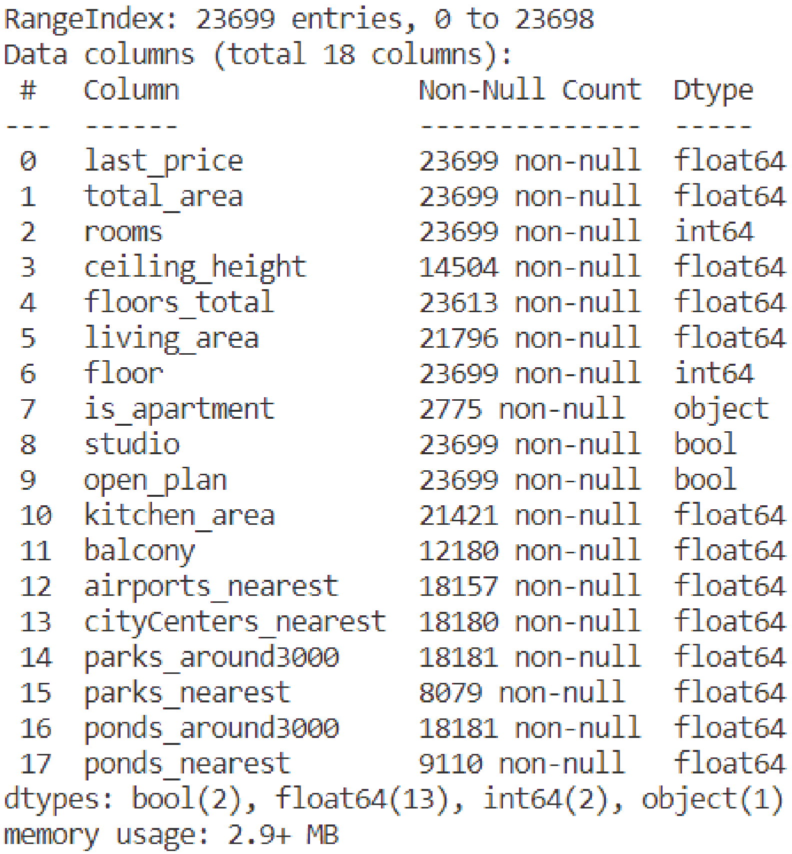
→Interpret the results obtained.



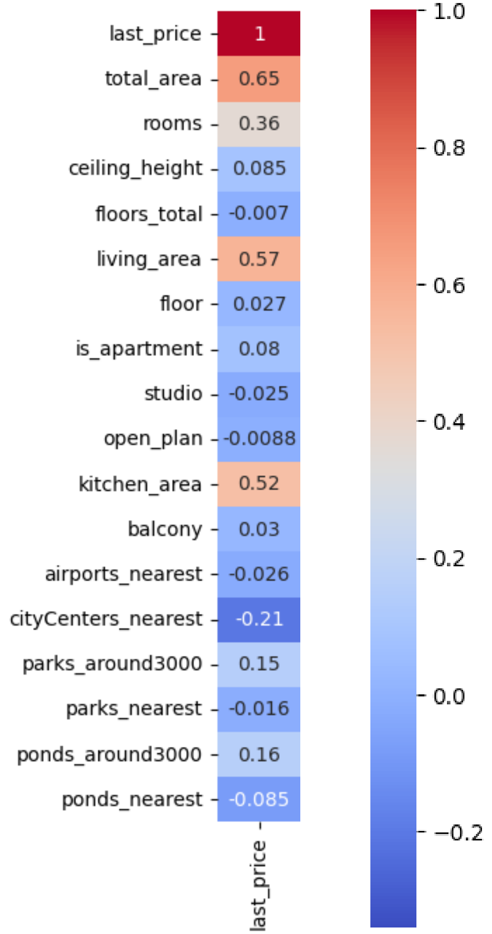
Raw data

Different format values and many data are not available in the source data.

Price at the time of withdrawal
Apartment area (m^2)
Number of rooms
Ceiling height (m)
Total floors of the house
Living area (m^2)
Floor
Is it an apartment? (Boulevard type)
Is it a studio apartment? (Boulevard style)
Is it a free layout? (Boulevard type)
Kitchen area (m^2)
Number of balconies
Distance to the nearest airport (m)
Distance to the city center (m)
Number of parks within 3 km
Distance to the nearest park (m)
Number of ponds within 3 km
Distance to the nearest pond (m)



```
RangeIndex: 23699 entries, 0 to 23698
Data columns (total 18 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   last_price            23699 non-null  float64
 1   total_area            23699 non-null  float64
 2   rooms                 23699 non-null  int64
 3   ceiling_height        14504 non-null  float64
 4   floors_total          23613 non-null  float64
 5   living_area           21796 non-null  float64
 6   floor                 23699 non-null  int64
 7   is_apartment          2775 non-null   object
 8   studio                23699 non-null  bool
 9   open_plan             23699 non-null  bool
 10  kitchen_area          21421 non-null  float64
 11  balcony               12180 non-null  float64
 12  airports_nearest      18157 non-null  float64
 13  cityCenters_nearest   18180 non-null  float64
 14  parks_around3000      18181 non-null  float64
 15  parks_nearest         8079 non-null   float64
 16  ponds_around3000      18181 non-null  float64
 17  ponds_nearest         9110 non-null   float64
dtypes: bool(2), float64(13), int64(2), object(1)
memory usage: 2.9+ MB
```



Data cleaning

The missing and abnormal values were replaced with modal values during data pre-processing.
In the graph «number of parks in a radius of 3 km» unique statistically significant values are not found.

Price at the time of withdrawal

Apartment area (m²)

Number of rooms

Ceiling height (m)

Total floors of the house

Living area (m²)

Floor

Is it an apartment? (Boulevard type)

Is it a studio apartment? (Boulevard style)

Is it a free layout? (Boulevard type)

Kitchen area (m²)

Number of balconies

Distance to the nearest airport (m)

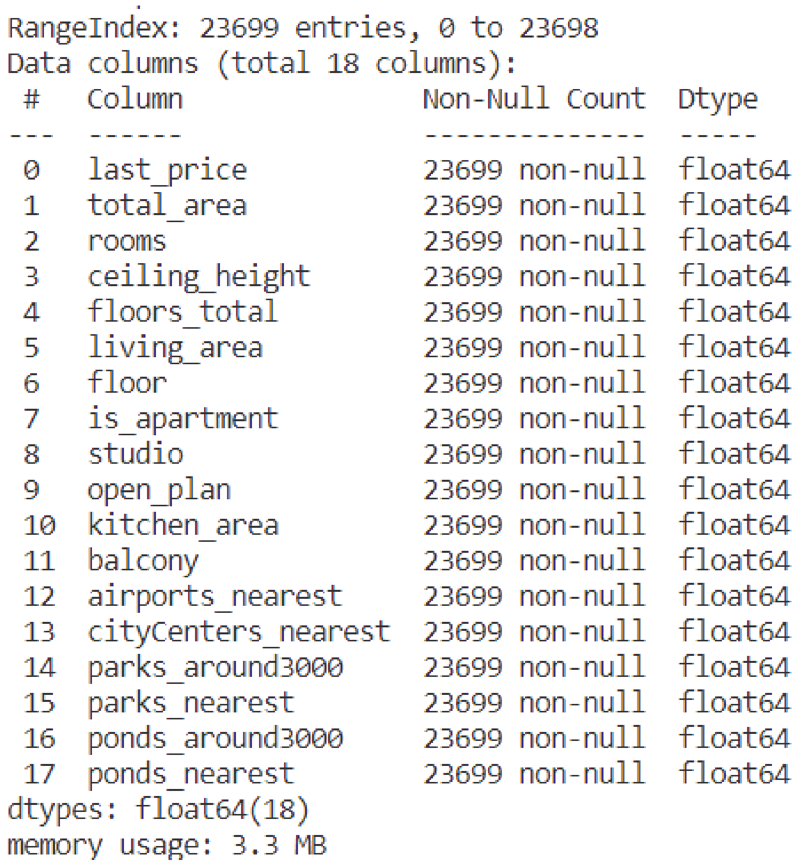
Distance to the city center (m)

Number of parks within 3 km

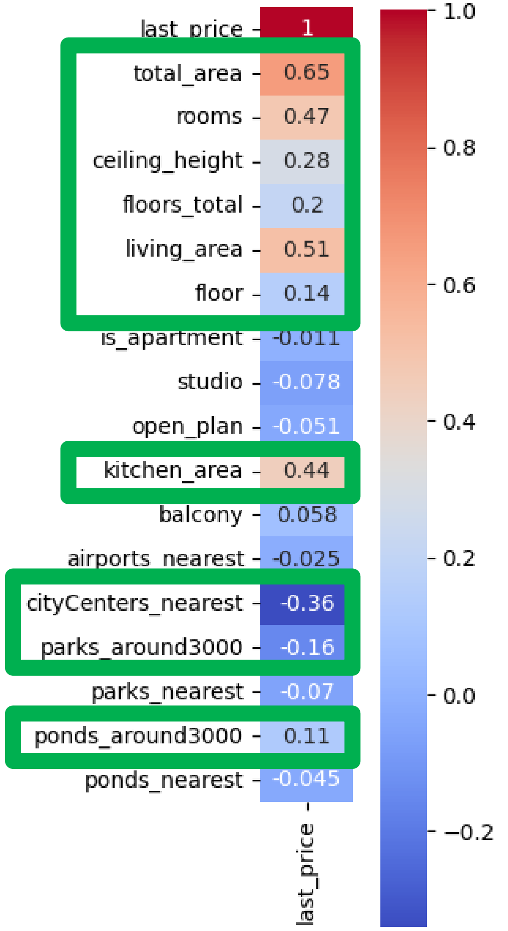
Distance to the nearest park (m)

Number of ponds within 3 km

Distance to the nearest pond (m)



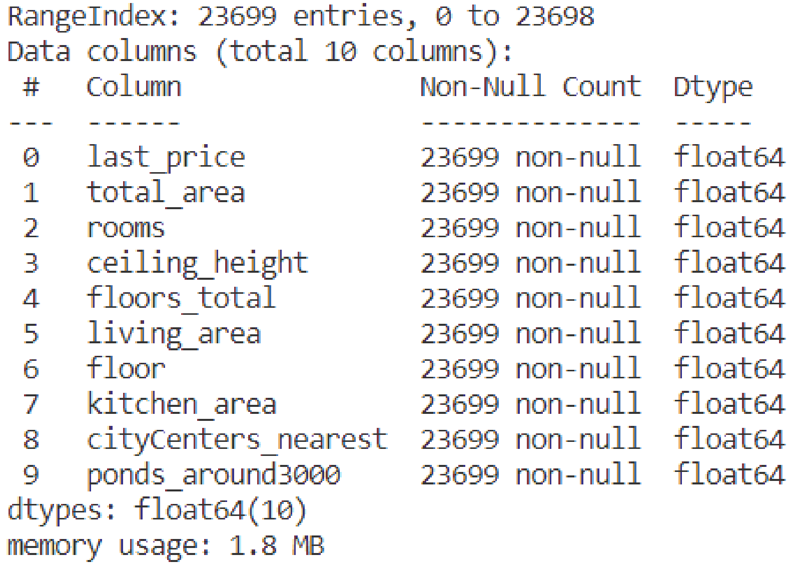
```
RangeIndex: 23699 entries, 0 to 23698
Data columns (total 18 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   last_price          23699 non-null  float64
 1   total_area          23699 non-null  float64
 2   rooms               23699 non-null  float64
 3   ceiling_height      23699 non-null  float64
 4   floors_total        23699 non-null  float64
 5   living_area         23699 non-null  float64
 6   floor               23699 non-null  float64
 7   is_apartment        23699 non-null  float64
 8   studio              23699 non-null  float64
 9   open_plan           23699 non-null  float64
 10  kitchen_area        23699 non-null  float64
 11  balcony             23699 non-null  float64
 12  airports_nearest    23699 non-null  float64
 13  cityCenters_nearest 23699 non-null  float64
 14  parks_around3000    23699 non-null  float64
 15  parks_nearest       23699 non-null  float64
 16  ponds_around3000    23699 non-null  float64
 17  ponds_nearest       23699 non-null  float64
dtypes: float64(18)
memory usage: 3.3 MB
```



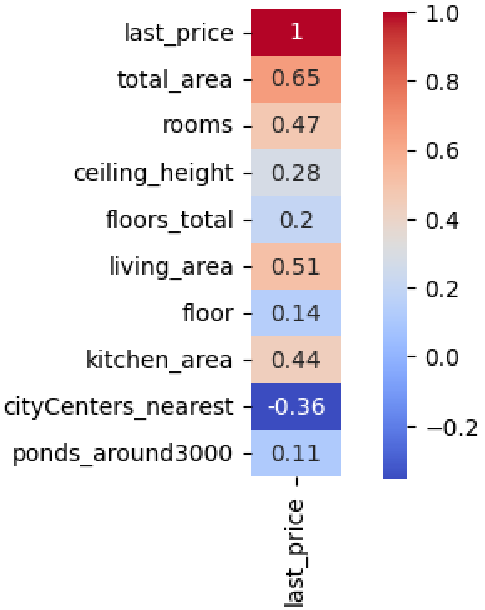
Data cleaning

As a result, significant parameters have been selected, based on which hypotheses will be built and tested.

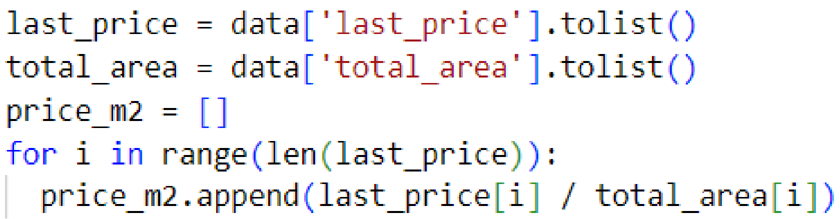
Price at the time of withdrawal
Apartment area (m^2)
Number of rooms
Ceiling height (m)
Total floors of the house
Living area (m^2)
Floor
Kitchen area (m^2)
Distance to the city center (m)
Number of ponds within 3 km



```
RangeIndex: 23699 entries, 0 to 23698
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   last_price          23699 non-null  float64
 1   total_area          23699 non-null  float64
 2   rooms               23699 non-null  float64
 3   ceiling_height      23699 non-null  float64
 4   floors_total        23699 non-null  float64
 5   living_area         23699 non-null  float64
 6   floor               23699 non-null  float64
 7   kitchen_area        23699 non-null  float64
 8   cityCenters_nearest 23699 non-null  float64
 9   ponds_around3000    23699 non-null  float64
dtypes: float64(10)
memory usage: 1.8 MB
```



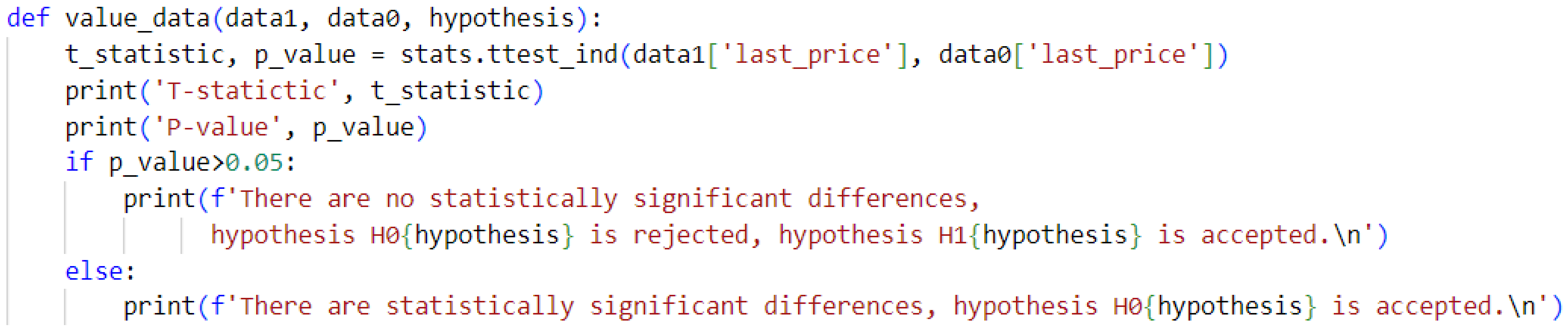
The parameters of «price at the time of withdrawal» and «apartment area(m^2) have a high correlation, which allowed the addition of a new parameter «cost of m^2», necessary for the forecast of the price of the apartment.



```
last_price = data['last_price'].tolist()
total_area = data['total_area'].tolist()
price_m2 = []
for i in range(len(last_price)):
    price_m2.append(last_price[i] / total_area[i])
```

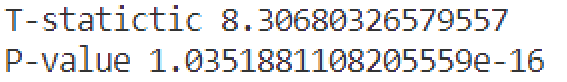
Research data analysis

Based on significant parameters, hypotheses were proposed:



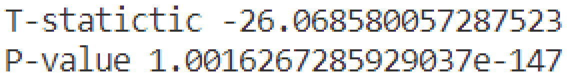
```
def value_data(data1, data0, hypothesis):
    t_statistic, p_value = stats.ttest_ind(data1['last_price'], data0['last_price'])
    print('T-staticti c', t_statistic)
    print('P-value', p_value)
    if p_value>0.05:
        print(f'There are no statistically significant differences,
              hypothesis H0{hypothesis} is rejected, hypothesis H1{hypothesis} is accepted.\n')
    else:
        print(f'There are statistically significant differences, hypothesis H0{hypothesis} is accepted.\n')
```

01. Prices for apartments in the center and far from the center are significantly different



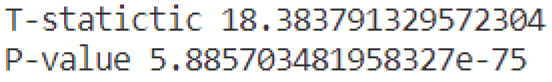
```
T-staticti c 8.30680326579557
P-value 1.0351881108205559e-16
```

02. Prices for apartments near the water body and far from the water body differ significantly



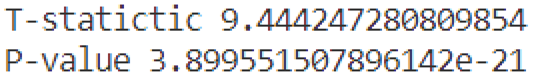
```
T-staticti c -26.068580057287523
P-value 1.0016267285929037e-147
```

03. Prices for apartments with a large and small kitchen differ significantly



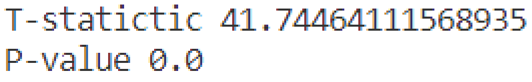
```
T-staticti c 18.383791329572304
P-value 5.885703481958327e-75
```

04: Prices of apartments with high and low ceilings differ significantly



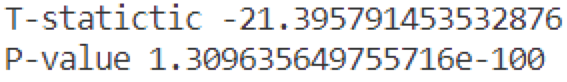
```
T-staticti c 9.44247280809854
P-value 3.899551507896142e-21
```

05: Prices for apartments in houses with floors more and less than 5 floors differ significantly



```
T-staticti c 41.7446411568935
P-value 0.0
```

06: prices for apartments on the ground floor from apartments on other floors differ significantly

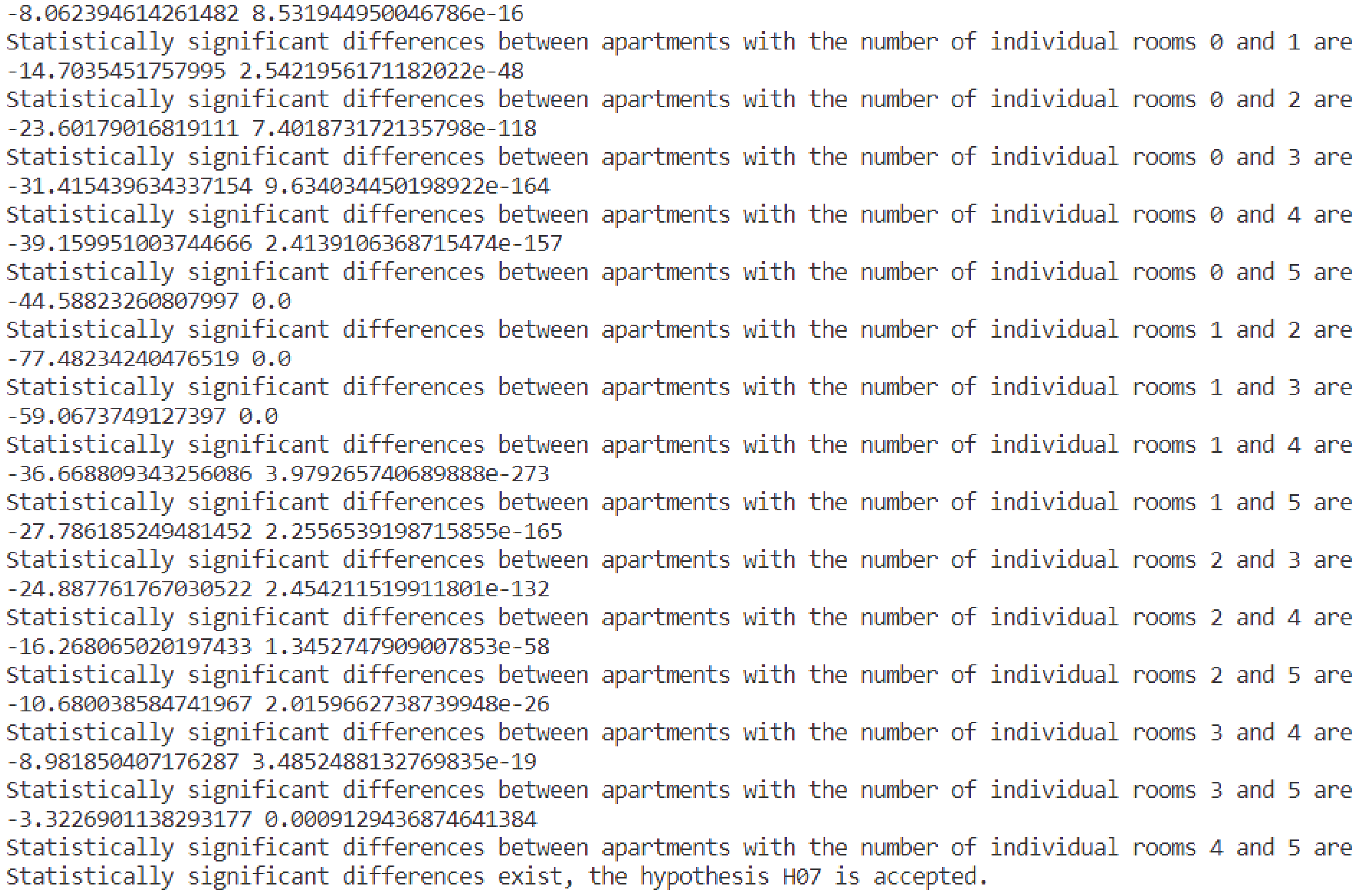


```
T-staticti c -21.395791453532876
P-value 1.309635649755716e-100
```

Research data analysis

Based on significant parameters, hypotheses were proposed:

07: prices for apartments with different number of rooms differ significantly



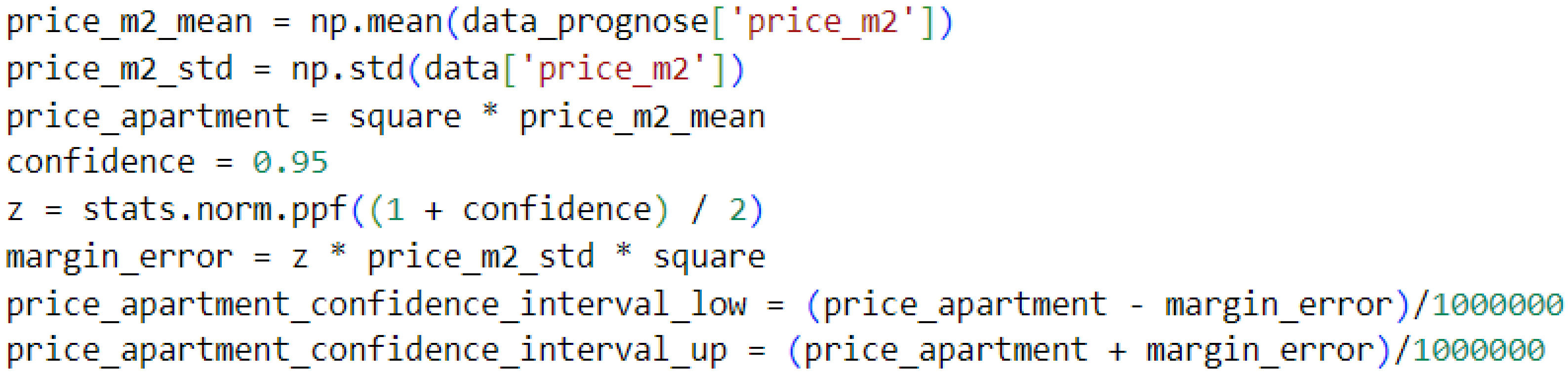
```
-8.062394614261482 8.531944950046786e-16
Statistically significant differences between apartments with the number of individual rooms 0 and 1 are
-14.7035451757995 2.5421956171182022e-48
Statistically significant differences between apartments with the number of individual rooms 0 and 2 are
-23.60179016819111 7.401873172135798e-118
Statistically significant differences between apartments with the number of individual rooms 0 and 3 are
-31.415439634337154 9.634034450198922e-164
Statistically significant differences between apartments with the number of individual rooms 0 and 4 are
-39.159951003744666 2.4139106368715474e-157
Statistically significant differences between apartments with the number of individual rooms 0 and 5 are
-44.58823260807997 0.0
Statistically significant differences between apartments with the number of individual rooms 1 and 2 are
-77.48234240476519 0.0
Statistically significant differences between apartments with the number of individual rooms 1 and 3 are
-59.0673749127397 0.0
Statistically significant differences between apartments with the number of individual rooms 1 and 4 are
-36.668809343256086 3.979265740689888e-273
Statistically significant differences between apartments with the number of individual rooms 1 and 5 are
-27.786185249481452 2.2556539198715855e-165
Statistically significant differences between apartments with the number of individual rooms 2 and 3 are
-24.887761767030522 2.454211519911801e-132
Statistically significant differences between apartments with the number of individual rooms 2 and 4 are
-16.268065020197433 1.3452747909007853e-58
Statistically significant differences between apartments with the number of individual rooms 2 and 5 are
-10.680038584741967 2.0159662738739948e-26
Statistically significant differences between apartments with the number of individual rooms 3 and 4 are
-8.981850407176287 3.4852488132769835e-19
Statistically significant differences between apartments with the number of individual rooms 3 and 5 are
-3.3226901138293177 0.0009129436874641384
Statistically significant differences between apartments with the number of individual rooms 4 and 5 are
Statistically significant differences exist, the hypothesis H07 is accepted.
```

All the hypotheses put forward were accepted as a result of statistical analysis.

Research data analysis

Not all parameters have been taken for estimating the price of the apartment, as when they are taken into account, the sample of data becomes unaffordable, and some parameters have ambiguous attitudes on the part of the user.

The price of the apartment is calculated based on a cut-down sample created based on the parameters of the search of the apartment (total area, number of rooms, kitchen area, distance from the center). It is issued as a range from the lowest to the highest value of the apartment under these parameters.

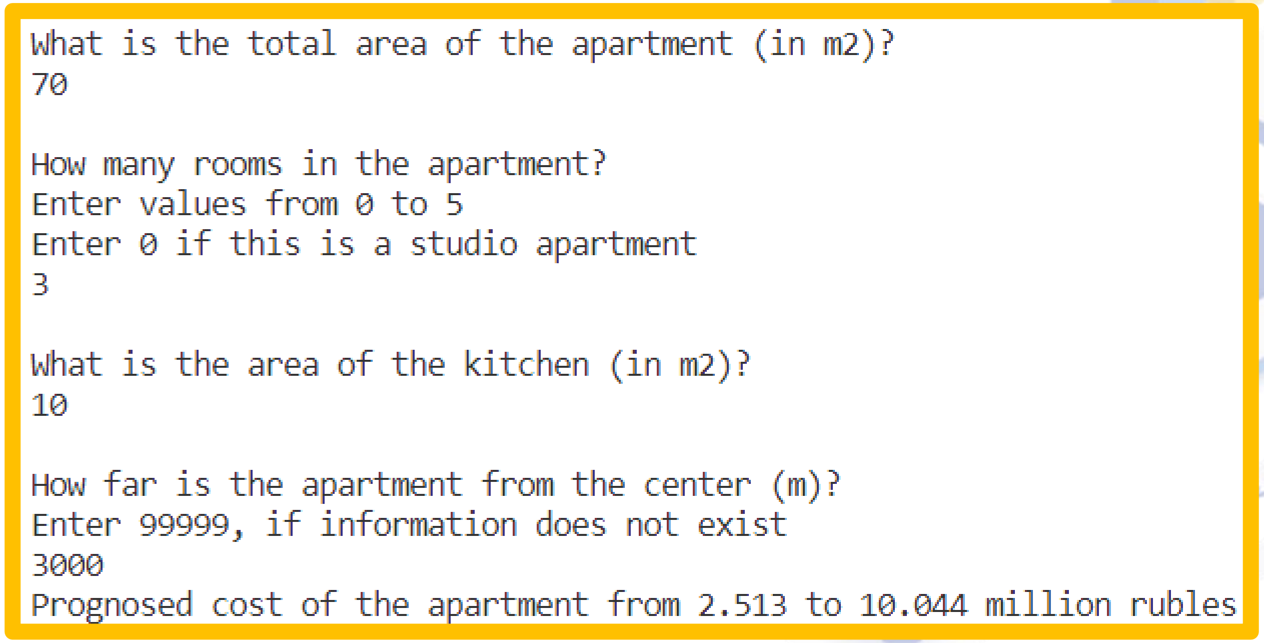


```
price_m2_mean = np.mean(data_prognose['price_m2'])
price_m2_std = np.std(data['price_m2'])
price_apartment = square * price_m2_mean
confidence = 0.95
z = stats.norm.ppf((1 + confidence) / 2)
margin_error = z * price_m2_std * square
price_apartment_confidence_interval_low = (price_apartment - margin_error)/1000000
price_apartment_confidence_interval_up = (price_apartment + margin_error)/1000000
```

Example

A family with kids needs a new apartment with 3 rooms near the city center.

The total area is about 70 m^2 and the kitchen area is about 10 m^2 for comfort.



```
What is the total area of the apartment (in m2)?
70

How many rooms in the apartment?
Enter values from 0 to 5
Enter 0 if this is a studio apartment
3

What is the area of the kitchen (in m2)?
10

How far is the apartment from the center (m)?
Enter 99999, if information does not exist
3000
Prognosed cost of the apartment from 2.513 to 10.044 million rubles
```

Conclusion

As a result of the work, it was noted that the following parameters correlate with the value of the apartment: area of the apartment in square meters (m^2), number of rooms, height of ceilings (m), total floors, living area in square meters (m^2), floor, kitchen area in square meters (m^2), distance to the city center (m), number of bodies of water in a radius of 3 km.

A heat map illustrates the correlation of the data concerned.

Based on the examination of the hypotheses put forward, it was possible to prove that the previously noted parameters affect the price of the apartment. However, not all the influencing parameters were used to forecast the price of the apartment.

As a model for forecasting the cost of the apartment, a function calculating the range of possible value of the apartment, taking into account the wishes of the user, has been proposed.