# Answers

1. **Before running any model, pick any two variables and attempt to show an (interesting) relation via visualization. This requires one to formulate a question, and to communicate clearly a conclusion based on data visualization. Ideally this would suggest ways to act on the issue.**

   *We have picked **price** and **rooms** because they have not bad interrelation. Graph indicates that there are too many apartments with 1 - 4 rooms in Almaty, additionally prices of real estates with 2 rooms and with 3 rooms have a big difference. (Figure 1)*
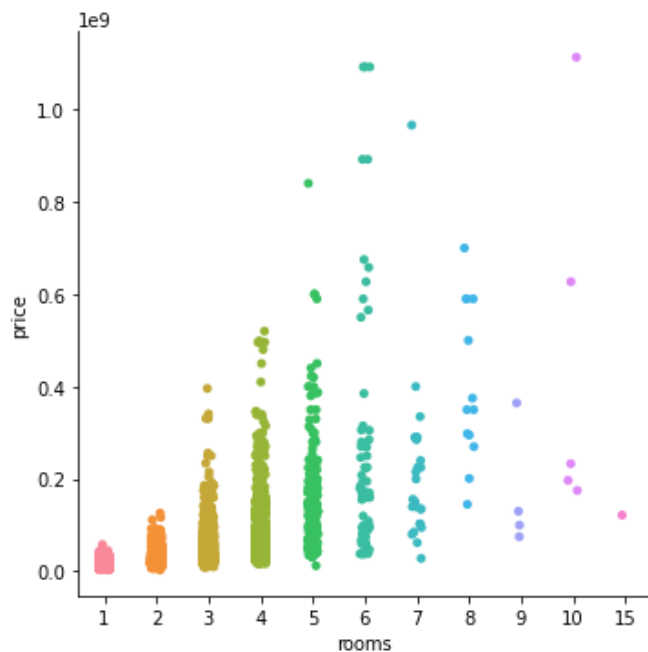
2. **Try to explain** price **with** area. **What can you conclude?**

Pearson Correlation of Features
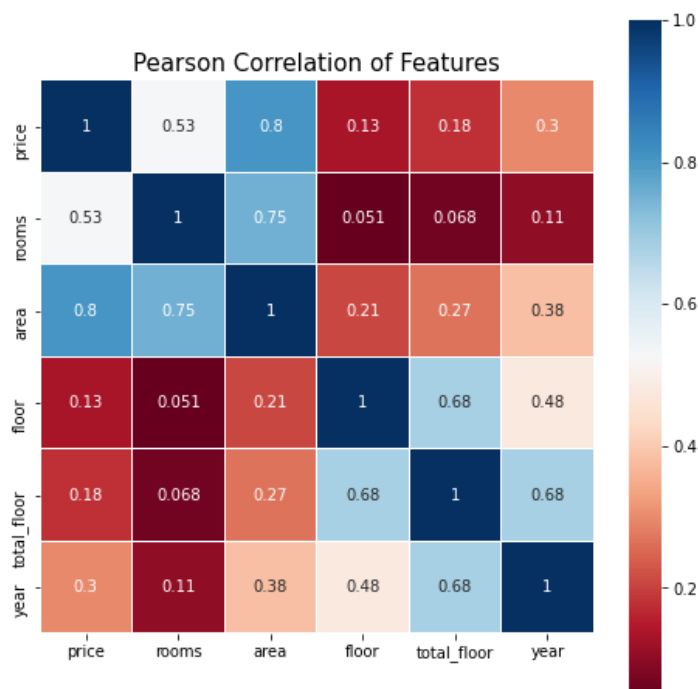
*Figure 2 shows Pearson Correlation of features in the dataset of apartments. As you see correlation coefficient between **price** and **area** is the highest one in this pattern. It is exact that all apartment prices are estimated according to their area in the major cities. So the real estate's price prediction strongly requires the area section.*

3. **In your opinion, can any of the variables provided in Exhibit 1 help to predict apartment prices? Since opinions do not provide strong arguments, provide a simple evidence based on data.**

Yes, of course. As I said, one of evidence could be the correspondence coefficients (Figure 2) between the variables described in Exhibit 1. Correlation strongly affects to the prediction. Additionally, we have tested it.

```
In [63]: lm = smf.ols(formula='price ~ rooms', data=data).fit()
         print(lm.rsquared)

         lmm = smf.ols(formula='price ~ area', data=data).fit()
         print(lmm.rsquared)

         0.2812326555835555
         0.6475658133031146
```

*Above you can be aware of coefficient of determination with only **rooms** and **area** separately. 64% of predicting prices have exactly determined by giving only **area** to the model which is hardly more precise than the giving only **rooms.***

4. **Run a multiple regression at least 4 of the variables described in Exhibit 1. Which variables are statistically significant?**

*I thought that multiple regression helps to improve the prediction accuracy of my model. But as I said only **area** has a priority to change it better unfortunately. In Figure 4, you can see firstly there very low coefficient of determination without **area** and **rooms**. But then adding them consecutively the coefficient of determination raised rapidly.*

Coefficient of Determination

```
In [64]: # only include rooms and area in the model
         lm = smf.ols(formula='price ~ year + type + district + floor + total_floor', data=data).fit()
         lm.rsquared

Out[64]: 0.09716984724864697
```

Accuracy is too low with these columns.

```
In [65]: lm = smf.ols(formula='price ~ rooms + year + type + district + floor + total_floor', data=data).fit()
         lm.rsquared

Out[65]: 0.34384194699124293
```

```
In [66]: lm = smf.ols(formula='price ~ rooms + area + year + type + district + floor + total_floor', data=data).fit()
         lm.rsquared

Out[66]: 0.667501386032185
```

5. **Do you think you can improve the model with additional predictors? Where/how can we get them and introduce to our model?**

*I guess we can improve our model by adjusting categorical predictors with more than two categories. We can create dummy variables instead of giving numeric values which causes an **ordered relationship** between categories. In addition, we should use different models of regression such as logistic regression or lasso. We can introduce them from NumPy, scikit-learn.*

6. **Hypothetical case illustrates one possible use case of such model use. Can you think of other examples where such model will be useful? Give at least 3 examples.**

- Mr. Duman wants to buy an apartment from the luxury residential complex, by connecting with Abraca-data Group which is a real estate agent company, he could be aware of which residential complexes are for wealthy people. Besides, he knows how much the average prices of them.
- Mrs. Sarah is a realtor in the RC Asyl Arman. After the quarantine measures, housing market fell rapidly and she needs a help to estimate apartments of her clients. By going to our office and using our model, she can predict the prices of houses entering its' characters. Also, she knows that in which residential complexes have more houses to sell. Then she helps to owners of those real estates.
- Almas is owner of the big house that's area contains 900 square meter. House was built in 2010. However, it looks new and beautiful. Almas confused with predicting its price as this house is extraordinary one in Almaty city. He searched the real estate agent who helps him about this problem. Then he found Mrs. Sarah who has our model. Almas estimated his strange house by our model.

Written by
Team 7 - Abraca-data:

- Muratbekov Maksatbek
- Sarkulov Yerden
- Tolbassy Bakdaulet
- Yegemberdi Nurgeldi
- Zhakypov Aslan