***Data Analysis 2: Assignment 2***

***(Moscow)***

*Ian Brandenburg*

*Dávid Szabados; ID 2302806*

**Introduction:**

As part of the assignment the goal is to see if there’s a correlation between the *highly\_rated* Boolean variable *(True / False),* and the assigned explanatory variables: *stars* and *distance*, and also the *price* variable, as our extra choice. *highly\_rated*’s value is decided by the *rating* variable, where if a hotel has a rating more or equal than 4, we give it a 1 (True) value, otherwise, we give it a 0 (False) value. We are going to use the *log of price*, as we want to see the relative price differences between the hotels, rather than the price itself. During the assignment, we will use the linear probability, logit, and probit models, as it was seen during the classes.

**Filtering, Cleaning and basic attributes of the Data:**

The data is filtered and cleaned by the following way:

We only included those **Hotels** in the data, which resides in the main city of **Moscow** (not other urban agglomeration). We chose **2017 November** as the date of our data, including only the **weekdays**. After these filtering, we checked the distribution of the prices and we concluded that because of the very long right tail, we should include only those observations, where the price is **less than $500**.

Next, those rows where the *rating* and/or *rating\_reviewcount* have a missing value, were discarded, as we don’t think these rows could get a 1 for *highly\_rated*, and giving a 0 would distort our data. The basic attributes of this data can be seen at **Appendix 1**.

**Used models:**

**Interpretations:**

**Appendix 1.:**

