***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**.

**Models and Interpretation:**

As it is already said, we used three models for our estimation. At **Appendix 2**, we can see the Linear Probability Model. The lnPrice and the stars of the hotel determines mainly the chance, whether a hotel will be highly rated or not, while the distance is just a minor part of the increasing chance. For every star, the chance that a hotel will be highly rated, increases by around 16.4%, while for the lnPrice, if it increases by one unit, the chance increases by 7.7%.

At **Appendix 3**, we can see the used logit and probit model, where the logit assumes a logistic distribution error, while the probit assumes a normal distributed errors. The logit and probit models are better indicators, than the LPM, for example by the fact, that the estimations will be between 0 and 1, while at the LPM, it can go trough than 1, than 100%. It can be seen, that the logit and probit models have roughly the same estimation and are similar than the LPM (45 degree line), except at the tails.

**Appendix 1.:**



**Appendix 2.:**



**Appendix 3.:**

