Partners: Corinne & Myitzu

Major decisions: Focus on hotels in the city of Amsterdam with 1,837 observations. We added a column to the dataframe called 'highly_rated' which was simply the 'rating' column when hotels were rated 4 or higher. We then made the 'highly_rated' variable a dummy variable (1 if highly_rated (rating of 4 or larger), 0 if not (rating less than 4)).

LPM: We performed a linear probability model regression (LPM) in order to see how unit changes in *distance* and *stars* would affect the probability of the dependent variable being 1 (a hotel being highly rated). The constant coefficient doesn't seem to have an interpretable meaning here. It is also negative (despite the dependent variable ranging from 0-1 (limitations of LPM)). For the distance coefficient, for one unit increase in distance, hotels are more likely to be highly rated by 2.67 percentage points. p value is 0 (or near 0) meaning distance is statistically significant (most likely a relationship between distance and probability of a hotel being highly rated (having more than 4 stars)). For the coefficient on stars, for one unit increase in stars, hotels are more likely to be highly rated by 30.83 percentage points. p value is 0 (or near 0) meaning stars are statistically significant (most likely a relationship between stars and probability of a hotel being highly rated. We also produced graphs to visualize the predicted probabilities of hotels being highly rated by distance and stars respectively in order to see the overall trend of the relationship.

Logit: In order to assure y values are bound within [0,1] and linearity isn't necessarily assumed, we produced a table with logit regression results, however for logit regressions, the coefficients are hard to interpret, so we must use marginal effects in order to get coefficients similarly interpretable to a LPM. We then create a logit marginal effects table. For the distance coefficient, for one unit increase in distance, hotels are more likely to be highly rated by 1.99 percentage points. p value is 0 (or near 0) meaning distance is statistically significant (most likely a relationship between distance and probability of a hotel being highly rated. For the stars coefficient, for one unit increase in stars, hotels are more likely to be highly rated by 30.23 percentage points. p value is 0 (or near 0) meaning stars are statistically significant (most likely a relationship between stars and probability of a hotel being highly rated.

Probit: We produced a table with probit regression results, however similarly to logit, the coefficients are hard to interpret. Therefore, we created a probit marginal effects table. For the distance coefficient, for one unit increase in distance, hotels are more likely to be highly rated by 1.76 percentage points. p value is 0 (or near 0) meaning distance is statistically significant (most likely a relationship between distance and probability of a hotel being highly rated. For the stars coefficient, for one unit increase in stars, hotels are more likely to be highly rated by 30.45 percentage points. p value is 0 (or near 0) meaning stars are statistically significant (most likely a relationship between stars and probability of a hotel being highly rated.

Predicted Probabilities: The predicted probabilities for all three models (LPM, Logit, and Probit) were all similarly in the 0.5 region indicating that ambiguity in predicting the probability of hotels being highly rated. However, looking at the fit of the predicted probabilities (using R-squared and Brier-score), the Logit and Probit models proved to provide the best fit for the predicted probability of hotels being highly rated. (Logit and Probit were very similar).

Overall Summary: Distance from city center and Stars seem to both have a positive relationship with hotels being highly rated (would reject the null hypothesis that they didn't have a relationship). We expected this from stars, but not from distance. However, both variables are statistically significant and therefore we can assume the relationship is meaningful and we can generalize to the population of all hotels in Amsterdam.