***Statistics for Business Assignment 1 FAQs***

**30/10/2023**

Please see some commonly asked questions about assignment 1. Note that this document will be updated as questions arise.

|  |  |
| --- | --- |
| **Question** | **Response** |
| How many variables should be used in the analysis? | In total you may use anywhere from 10-20 variables for the full analysis.  Hypotheses should be specified for five ‘focal’ variables. But visualisations can include more than five variables (each visualisation can include multiple variables – you are not restricted to five variables).  The regression analysis should include five focal variables, as well as a series of control variables / variables to improve accuracy. Here you may end up using up to 20 variables to build a ‘good’ model. |
| Is it necessary to address data quality issues for the entire dataset? | The recommendation is to provide a general overview of the quality of the dataset, based on the summary statistics.  Beyond this, the crucial task is to address any quality issues with the variables that you use. |
| How many outliers should be removed / corrected? | This is largely down to judgement. Outliers should be identified by reviewing summary statistics, and crucially by visualising the data.  Only a vey small number of outliers should be removed / corrected (maybe no more than around 10).  It is important to justify in the written report why any data was removed / corrected, and what the implications of this might be. |
| Gr living area is in the data dictionary but not in the dataset. | Please ignore this – Gr living area is the sum of floor1\_sf and floor2\_sf. Later when we look at regression you will see that it would cause a big problem to include both. |
| There are some discrepancies between the data dictionary and the specific observations in the dataset. | For a small number of variables there may be discrepancies between the data dictionary and the values in the dataset (e.g. extra categories, slightly different variable names). This should not impact the analysis – if you notice this then you can note it as a data quality issue in your report. |