|  |
| --- |
| **R Final Project**  Analysis & Discussion, by Tony Mazich  First Model (M1)  For the first model, I chose to examine real-estate valuation data provided in the modified UCI data ZIP. This dataset concerns historical market data collected from Sindian District in New Taipei City, Taiwan.  For the response, I chose price, as this is the variable we would primarily be concerned with as a buyer/seller.  To find viable predictors, I plotted all the data categories against each other and looked for curves that indicate a strong correlation.  We see from the scatter plots that most of the categories (except for date of purchase) show at least a weak correlation with price. The strongest potential predictors I saw were age, MRTdistance, and latitude. I also identified longitude as a possible predictor, but later analysis ruled it out.  The relationship between price and MRTdistance (The distance of the property to an MRT Station, i.e. rapid public transit) followed the most significant and predictable curve among all the relationships shown. I thought this relation roughly followed a logarithmic curve, and fit the model based on that, resulting in the following plot:    Model information:  R^2 (Adjusted): 0.5381  AIC: 3020.601  Equation: price = 95.016865 - 8.923536 \* log(MRTdistance)  Second Model (M2)  For the second model, I fit all possible predictors, and then compared it to another model with slightly pruned down terms (predictors with no apparent relationship to price or very high p-value were removed)  Model 2 (all) had an AIC of 2989.924 and an R^2 of 0.5762. A summary is seen below:    The refined Model 2 had an AIC of 2989.924 and an R^2 of 0.6481, with the following summary:    As seen above, the predictors used are the log of distance from public transit, the age of the property, its latitude, the number of convenience stores nearby, and the date sold.  Based on the AIC, the refined Model 2 is a more representative model, and based on the R^2 values, the refined Model 2 also accounts for more of the variability than Model 2 (All.)  With our winner chosen, we also want to see how Model 2 fares against Model 1. The stats are as follows:  **M1 / M2**  AIC: 3020.601 / 2989.924  R^2: 0.5381 / 0.5762  Based on these stats, the relationship between MRT distance and price is decent, but including more predictors produces a more accurate model that accounts for more of the variability in the data. |