Derek Chang

Assessment_Data.csv contains housing data for 265 Massachusetts towns and cities over 7 years. There are 34585 rows and 26 columns, with many categorical/logical variables and numeric columns. City is where the house is located, ls_year is the list year, price is the list price, assess is the assessment value, ls_month is the list month, res_area is the residential area, house_age is the house's age, style is the house's style, num_rooms is the number of rooms, stories is the number of stories, inc_medianhhd_blkgrp is the median income of a household on the block, age_med_blk is the median age of a household on the block, pop_blk is the population of the block, black_share is the number of African American households on the block, vacant_share is the number of vacant houses on the block, latitude and longitude pinpoint the house's location, distance_firestation is the distance to the nearest fire station, distance_hospital is the distance to the nearest hospital, distance_police is the distance to the nearest police station, distance_prischool is the distance to the nearest private school, distance_pubschool is the distance to the nearest public school, distance_townhall is the distance to the nearest town hall, and distance_train is the distance to the nearest train station.

Since all real estate is local, some of my analysis focuses on several in my area. I start off with a barplot of monthly listings, also adding a trendline; there is a clear peak in the summer, fairly steady number of listings from October to December, a low point in January and February, and an upward trend in the spring. I then display a distribution of local list prices overlaid with a normal distribution, which is skewed right. Next, I build a contingency table comparing the vacancy of houses on the block with the house's distance from the public school. I also include a graph not encountered yet in this class to show the list price based on house style.

In my analysis, I surprisingly find that houses' list price is not dependent on the distance from public school through a permutation test. Using a p-value based on a normal distribution, I find that the data is not consistent with the normal distribution as there is sufficient evidence against the null hypothesis. Analyzing a contingency table, I conclude that vacancy and proximity to public school are likely dependent. With the chi-square method, I find that a house's price and vacancy are likely dependent. Further, I build a statistic whose distribution is standard normal for the list price of local houses. I then take a Monte Carlo simulation approach to finding the average size of a local house, which turns out to be a great approximation. I use linear regression to show that residential area is a good predictor of a houses' assessment value—there is a positive relationship. I calculate the correlation between the assessed value and residential area, which is quite strong. Additionally, I calculate a 95% confidence interval for houses' list price. I then introduce novel statistics for further analysis. Finally, I conclude with using quantiles to compare distributions.