HOW DOES YOUR NEIGHBORHOOD AFFECT YOUR HOME VALUE?

AN ANALYSIS OF THE EFFECT OF PROXIMITY TO LOCAL SCHOOLS, BUSINESSES, AND OTHER PLACES ON HOUSE PRICE

MELANIE BLUCK, TIMOTHY SHEN, MAHA SHAFAEEN, JEFF LEE

CONTENTS

INTRODUCTION

It's often said that the number one rule in real estate is "location, location, location." Certainly, location is a major predictor of real estate pricing on a regional scale, but does this principle apply within a city, neighborhood to neighborhood? It's common wisdom that a house's proximity to landmarks such as schools, parks, and businesses will affect the property's desirability, and therefore its price. While seemingly sensible, this idea is rarely presented side by side with data, and exists in the public consciousness only as an assumption. To find out if this idea is true, we analyzed houses across four metropolitan areas.

These regions are:

- Los Angeles – Anaheim – Long Beach, CA
- Chicago – Naperville – Elgin, IL
- Dallas – Fort Worth – Arlington, TX
- Washington – Arlington – Alexandria, DC-VA

These four were selected to represent varying regions of the US while being roughly similar in population.

We built a regression model to measure the effect of a house's proximity to various landmarks on its price. A landmark is defined as a location of interest and includes the following set of landmark types:

- Schools
- Grocery stores
- Other retail (Shopping)
- Parks
- Gyms
- Golf courses
- Beaches
- Hospitals
- Cemeteries

We generated samples of up to 500 single-family home sale listings for each city and compiled the addresses and GPS coordinates of each house and all landmarks. The predictor variables in the model were the distances from houses to each of the landmark types, and the response variable was the house price. If the common wisdom about real estate were true, then we'd expect to see that a larger distance to landmarks would be associated with a lower home price, except in the case of hospitals and cemeteries where it would be a higher price. However, our model showed that the effects of distance was totally inconsistent across different regions, and more often had the opposite effect on house price than what was expected. How does the surrounding neighborhood (parks, shopping centers, etc.) affect the housing prices?

DATA ACQUISITION AND GENERATION

House data was retrieved via Realty Mole's Property Data API which generated a sample of up to 500 single-family home sale listings for each city, along with the respective GPS coordinates, price, and square footage. Rows with missing square footage values were removed.

School data was sourced from the National Center for Education Statistics, which provides an online tool to generate a list of all school addresses in a given city and export it to a CSV file. The addresses of all other location types (beaches, parks, etc.) were scraped from yellowpages.com using Beautiful Soup.

To measure the distances from houses to landmarks, it was necessary to first convert addresses to GPS coordinates. This was done by querying the Position Stack geocoding API. Next, the distances between pairs of coordinates needed to be calculated. GPS coordinates are spherical coordinates, making it difficult to manually calculate distances, so we used GeoPy to obtain the distances and convert them to miles. A given house's distance to a particular landmark type, such as beaches, was defined as the shortest distance among all the landmarks of the same type present in the same metro area.

STATISTICAL METHODS AND RESULTS

EXPLORATORY DATA ANALYSIS

The house price data contained a large number of high outliers. However, due to the expensive nature of the modern housing market, many of these outliers appeared to be useful to our study, rather than radical exceptions to trends. Rather than removing any houses outside the interquartile ranges of their respective metro areas, we removed outliers subjectively, but most had a price per square foot greater than $3000.

Naturally, identical homes in different regions can have wildly different prices. To account for this regional effect, we used a MinMaxScaler for each metro area to normalize the prices relative to each region. The data was heavily right skewed, so log, square root, and cube root transformations were applied to the prices of each metro area separately. These are displayed in the pairplot below.

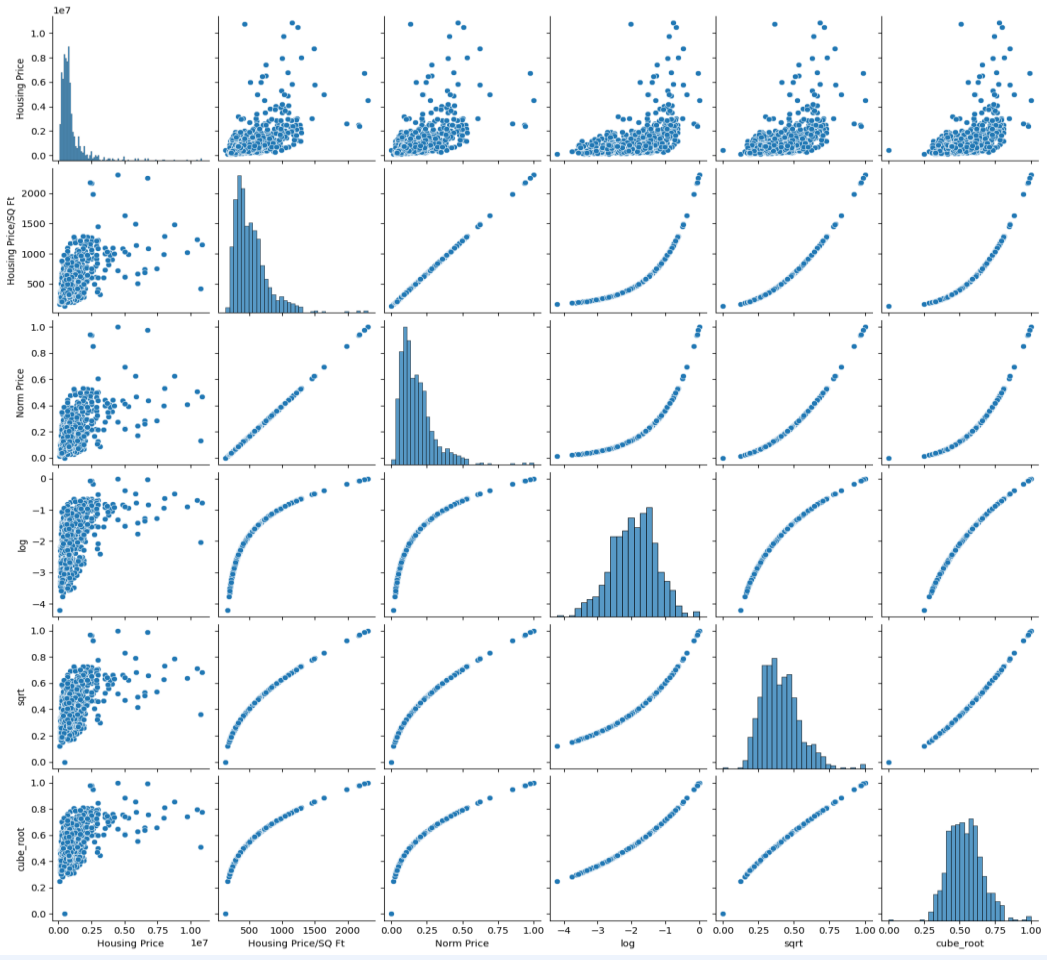

Figure 1a: pairplot for all house prices with transformations

The pair plots for individual metro areas show that different regions respond differently to the transformations, suggesting that house prices are distributed differently in different regions.

We then generated correlation heatmaps for distances to the different landmarks. The heatmap for all regions combined is below.

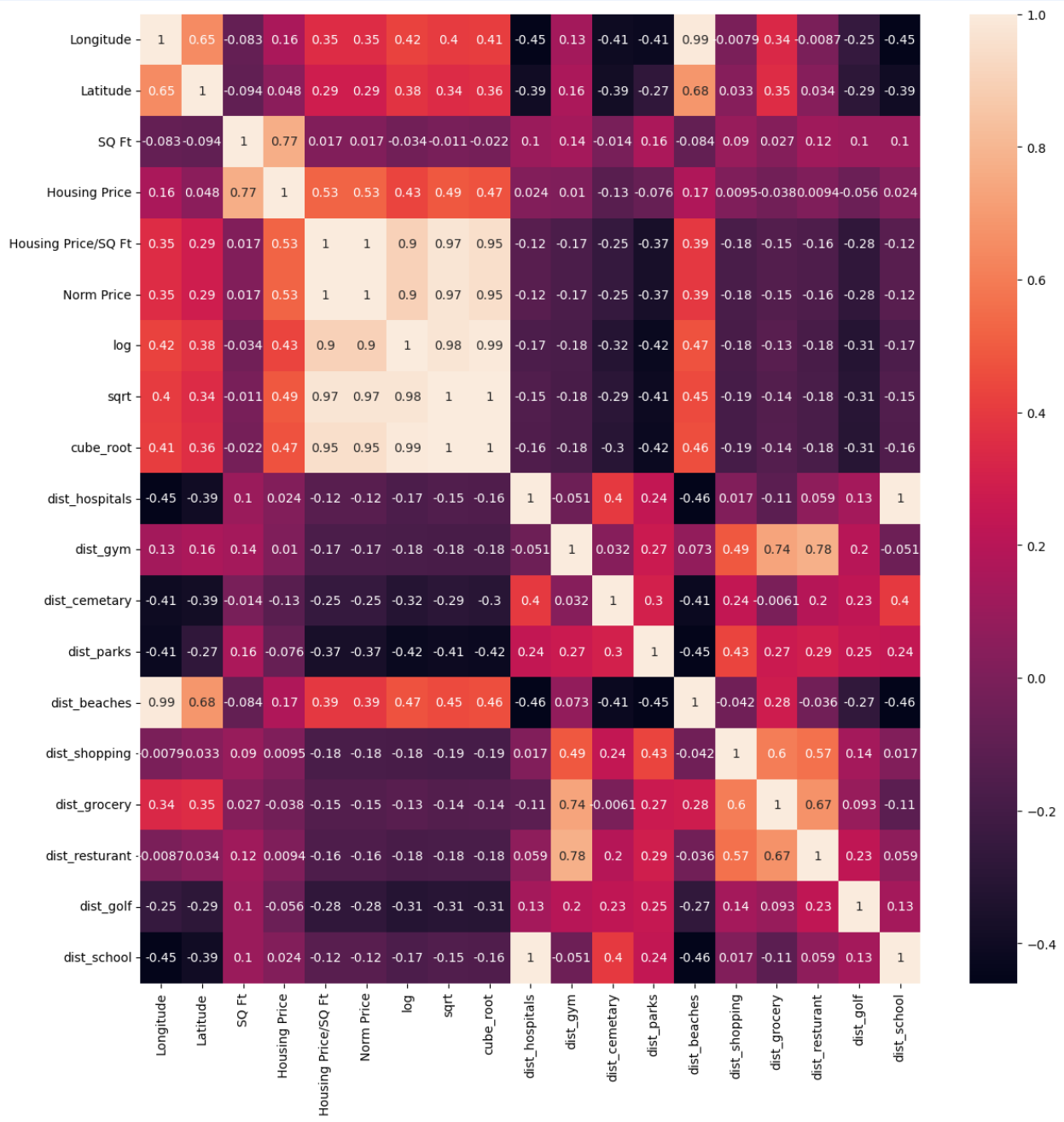

Figure 1b: heatmap for all regions

Although the housing prices typically decreased when further away from landmarks, the coefficients varied by region. Most variables have a negative coefficient, which is to be expected. The only positive coefficients were parks in the LA metro and beaches in the DC metro. As seen in the combined heat map above, the positive correlation for beaches in DC was large enough to offset the negative coefficients in other regions. This represents a significant departure from expectation.

LINEAR REGRESSION ANALYSIS

For Model I, we set the *Normalized Price* as the response variable and the 10 predictor variables as the distance to the closest *Hospital*, *Gym*, *Cemetery*, *Park, Beach, Shopping, Grocery, Restaurant, Golf,* and *School*.

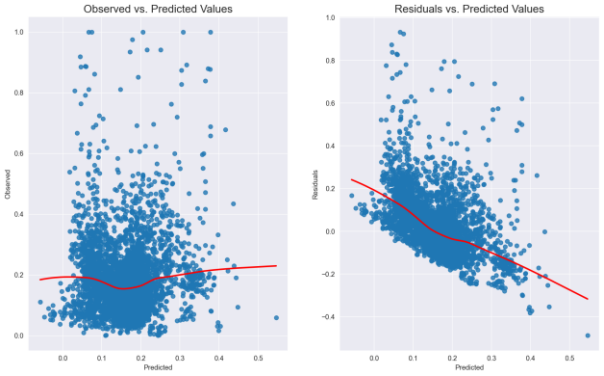

Figure 2a: Observed vs Predicted Values Plot **and** Residuals vs Predicted Values Plot for Model I

The model appears to have a trend in the Residuals vs Predicted Values plot (Figure 2a), suggesting that the data is non-linear and errors are correlated. Furthermore, running a Normal Probability Plot of the model shows significant departures from the straight line pattern, suggesting the data is non-normal(Figure 2b). The residual plot also shows a lot of heteroskedasticity (Figure 2a).

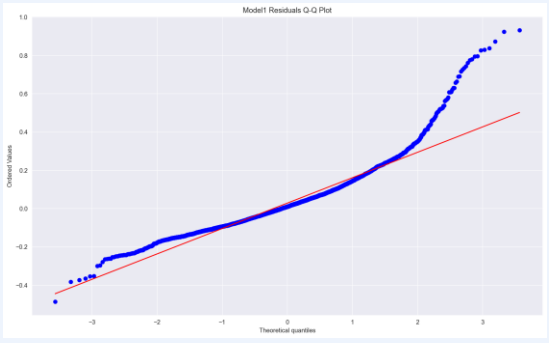

Figure 2b: Normal Probability Plot of Model I

To remedy non-normality, we applied a *reciprocal transformation* (1/y) to the response variable, *Normalized Price*. To remedy constant variance and linearizing the data, we applied a *reciprocal transformation* (1/x) to the 10 predictor variables. The improved model we will denote as Model II.

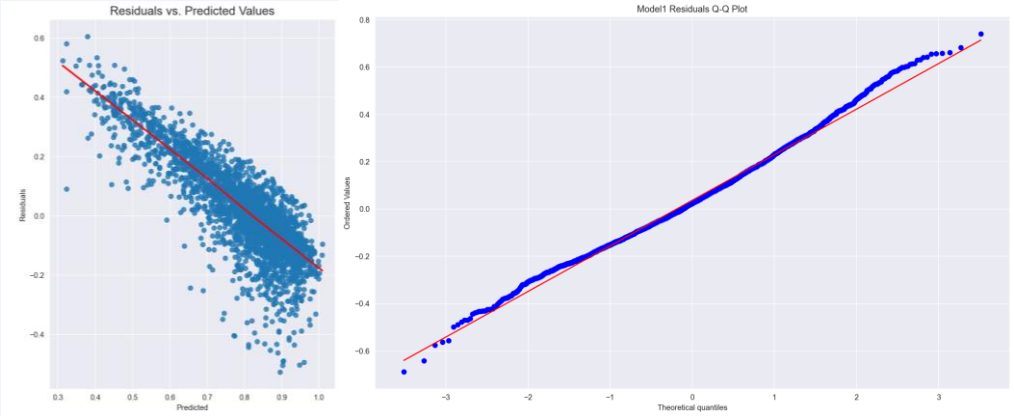

Figure 2c: Residuals vs Predicted Values Plot **and** Normal Probability Plot for Model II

The normal probability plot (Figure 2c) appears approximately normal, but the Jarque-Bera test outputs a nearly zero p-value, suggesting that house prices are still not normally distributed. The appearance of residual plot improves, but the right side of the plot still contains many departures from the trend line, and the Breuch Pagan test still outputs a p-value near zero, suggesting that heteroskedasticity remains (Figure 2c).

There appeared to be no high multicollinearity in Model II, with no variables having a VIF greater than 3. The resulting multiple linear regression produced the following results: (Note all variables are 1/predictor).

	Estimate	**P-Value**		**Estimate**	**P-Value**
Distance to Hospitals	-0.0516	0.003	*Distance to Shopping*	0.0366	0.090
Distance to Gym	0.1968	< 2e-16	*Distance to Grocery*	0.0314	0.121
Distance to Cemetary	-0.0539	0.004	*Distance to Restaurant*	-0.0065	0.753
Distance to Parks	-0.0629	< 2e-16	*Distance to Golf*	0.13333	< 2e-16
Distance to Beaches	-0.6043	0.015	*Distance to School*	0.9825	< 2e-16

Figure 2d: Summary of Linear Regression Model II

The model produced an R^2 value of 0.943, suggesting that 94.3% of the variance of the *Normalized Price* can be explained by the model. With p-values greater than 0.05, the effects of the distances to a shopping center, grocery store, and restaurant are not significant (Figure 2d). While statistically significant, the effects of the distances to hospitals and cemeteries are negligible, and the effect of the distances to parks is arguably not meaningful either.

Despite the high R^2 value, the model still violates the assumptions of normally distributed data and homoscedasticity. Therefore, we investigated the effects of landmark proximity for each metro area separately. For each metro, we ran an initial linear regression model where *Normalized Price* was its response variable and the following 10 predictors as the distance closest to nearest *Hospital*, *Gym*, *Cemetery*, *Parks, Beaches, Shopping, Grocery, Restaurants, Golf,* and *School*.

Los Angeles Metro Area

The initial model violated the assumptions of normally distributed data and homoscedasticity just as Model I did. We applied a *reciprocal transformation* to both the response variable and its predictor variables, denoted as Model LA. The Jarque-Bera test now produces a p-value of 0.35, which is high enough to conclude that the prices are normally distributed. However, heteroskedasticity is still present as the Breusch Pagan test still returns a p-value near zero.

Model LA has the following results summary and an R^2 of 0.967 (initial model: 0.742).

	Estimate	**P-Value**		**Estimate**	**P-Value**
Distance to Hospitals	0.0694	0.007	*Distance to Shopping*	0.1320	< 2e-16
Distance to Gym	0.2410	< 2e-16	*Distance to Grocery*	0.1869	< 2e-16
Distance to Cemetary	-0.0725	0.002	*Distance to Restaurant*	0.0323	0.220
Distance to Parks	-0.2149	< 2e-16	*Distance to Golf*	0.0201	0.420
Distance to Beaches	-0.3456	< 2e-16	*Distance to School*	0.6432	< 2e-16

Figure 2e: Summary of Linear Regression Model LA

The effects of the distances to a restaurant and a golf course are not statistically significant (Figure 2e). All the remaining predictors are significant, though the effects of the distances to a hospital or cemetery are negligible. This is in line with Model II, though the remaining estimates vary in magnitude when compared to Model II. Schools and beaches play much less of a role in Model LA, while parks, shopping, and grocery stores are much more important than in Model II.

DC Metro Area

The initial model violated the assumptions of normally distributed data and homoscedasticity just as Model I did. Furthermore, both *Distance to Hospital* and *Distance to School* exhibited high multicollinearity and produced VIFs greater than 5, so we removed them.

We applied a *reciprocal transformation* to both the response variable and its predictor variables, denoted as Model DC. This led to the Jarque-Bera test producing a p-value of 0.18, which is high enough to assume normally distributed data. However, the p-value from the Breusch Pagan test is still close to zero, suggesting that heteroskedasticity remains.

Model DC has the following results summary and an R^2 of 0.935 (initial model: 0.782).

	Estimate	**P-Value**		**Estimate**	**P-Value**
Distance to Gym	0.4773	< 2e-16	*Distance to Shopping*	0.0972	0.046
Distance to Cemetary	0.1860	< 2e-16	*Distance to Grocery*	-0.2784	< 2e-16
Distance to Parks	0.5352	< 2e-16	*Distance to Restaurant*	0.1873	0.001
Distance to Beaches	1.1e+07	< 2e-16	*Distance to Golf*	0.1164	0.009

Figure 2f: Summary of Linear Regression Model DC

All the predictors are statistically significant, but the effect of the distance to a beach is close to zero and meaningless (Figure 2f).

Chicago Metro Area

The initial model violated the assumptions of normally distributed data and homoscedasticity just as Model I did. Furthermore, *Distance to Cemetery*, *Distance to Beach*, *Distance to Shopping*, and *Distance to Grocery* exhibited high multicollinearity with VIFs greater than 5, so we removed them.

We applied a *reciprocal transformation* to both the response and predictor variables, denoted as Model CHI. This led to the Jarque-Bera test producing a p-value of 0.62, which is high enough to assume normally distributed data. However, the p-value from the Breusch Pagan test is still close to zero, suggesting that heteroskedasticity remains.

Model CHI has the following results summary and an R^2 of 0.920 (initial model: 0.502)

	Estimate	**P-Value**		**Estimate**	**P-Value**
Distance to Hospitals	0.0198	-0.689	*Distance to Restaurant*	-0.1455	0.012
Distance to Gym	0.3764	< 2e-16	*Distance to Golf*	0.2733	<2e-16
Distance to Parks	0.1337	< 2e-16	*Distance to School*	0.8470	<2e016

Figure 2g: Summary of Linear Regression Model CHI

All predictors are statistically significant except for *Distance to Hospitals*.

Dallas Metro Area

The initial model violated the assumptions of normally distributed data and homoscedasticity just as Model I did. Thus we applied a *reciprocal transformation* to the response variable, denoted as Model DAL, which led to a much improved Jarque-Bera statistic, but the p-value was still close to zero, along with the p-value from the Breusch Pagan test. Applying a *reciprocal* or *log* transformation to the predictor variables reduced the value of R^2, so they were left untransformed.

Model DAL has the following results summary and an R^2 of 0.903 (initial model: 0.483)

	Estimate	**P-Value**		**Estimate**	**P-Value**
Distance to Hospitals	0.0239	0.001	*Distance to Shopping*	0.1091	<2e-16
Distance to Gym	-0.0619	2e-16	*Distance to Grocery*	0.0106	0.272
Distance to Cemetary	0.0062	0.527	*Distance to Restaurant*	.0505	<2e016
Distance to Parks	0.0121	0.197	*Distance to Golf*	0.0356	0.006
Distance to Beaches	0.0417	<2e-16	*Distance to School*	0.335	<2e-16

Figure 2h: Summary of Linear Regression Model DAL

DISCUSSION

The results of the analysis run contrary to expectations. It's generally believed that a house's price should rise with proximity to all of the landmarks, except for hospitals and cemeteries, which should lower the price. We found that the real life effect of proximity to these locations was very inconsistent among different cities, and was generally associated with a *decrease* in home value, rather than an increase.

The following tables compare the estimated linear coefficients of the various landmarks across the four metro areas.

	Hospital	Gym	Cemetery	Park	Beach
LA	0.0694	0.2410	-0.0725	-0.2149	-0.3456
DC	High VIF	0.4773	0.1860	0.5352	>0
Chicago	0.0198	0.3764	High VIF	0.1337	High VIF
Dallas	0.0239	-0.0619	0.0062	0.0121	0.0417

Figure 3a: Summary of linear coefficients

	Shopping	Grocery	Restaurant	Golf	School
LA	0.1320	0.1869	0.0323	0.0201	0.6432
DC	0.0972	-0.2784	0.1873	0.1164	High VIF
Chicago	High VIF	High VIF	-0.1455	0.2733	0.8470
Dallas	0.1091	0.0106	0.0505	0.0356	0.335

Figure 3b: Summary of linear coefficients continued

Summarized in words, the results for each landmark type is as follows. Keep in mind that a positive coefficient indicates that proximity to the landmark has a negative effect on the house price.

- Hospitals: No meaningful effect anywhere.
- Gyms: Modest negative effect in most places, but no effect in Dallas.
- Cemeteries: Small negative effect in DC, but no effect anywhere else.
- Parks: Modest negative effect in DC, small negative effect in Chicago, small positive effect in LA, and no effect in Dallas.
- Beaches: Modest positive effect in LA, no effect anywhere else.
- Shopping: Small negative effect in LA and Dallas, no effect elsewhere.
- Groceries: Small negative effect in LA and DC, small positive effect in Chicago, no effect in Dallas.

- Restaurants: Small negative effect in DC, small positive effect in Chicago, no effect elsewhere.
- Golf: Small negative effect in DC, modest negative effect in Chicago, no effect elsewhere.
- Schools: Large negative effect in LA and Chicago and small negative effect in Dallas.

Not a single landmark type has a consistent relationship with home value. The one with the most impact, schools, even has the opposite relationship from what was expected. Unraveling why the effects are so inconsistent across regions would require an in-depth analysis of the urban design of all nine cities involved.

It's important to note that our regression models were flawed. Some of them did not meet the normality assumption, and none met the assumption of equal variance. The extent to which the models' accuracy was affected is unclear.

As to why the effects were so different across regions, it could very well be due to their differing spatial makeups, though the full answer is certainly not this simple. Los Angeles and Dallas are known as highly car-dependent cities. This means that residential areas are kept almost entirely separate from commercial land use, as opposed to the mixed-use zoning used in most of the world, where small commercial properties are distributed throughout residential areas. While Chicago and Washington DC are known for comparatively better public transit systems and urban design that predates the era of car-centrism, most of the actual land area of their metropolitan areas are also car dependent. This sort of design means that measuring the distances to places like gyms and restaurants is often just measuring the distance to one strip mall that houses both. Additionally, an entire neighborhood will often share the same one or two strip malls that all the houses are closest to. It might be that more expensive houses tend to be those which are newly built around the outskirts of a suburb rather than the older houses which are next to already developed commercial areas. This explanation doesn't address the influence of houses downtown, which tend to be the most expensive as well as the closest to a variety of land uses. Of course, such a statement is only true about cities broadly, and not necessarily true of our chosen sample of cities. In particular, Dallas lacks the high density associated with a major city's downtown. All in all, it would require a much more in-depth study of urban design to find any answers as to why the effects of the distances to landmarks are so unpredictable.

CONCLUSION

Our analysis did not find any evidence that the distance to locations like schools, parks, or businesses has a consistent effect on home value. While these effects are likely to exist in some places on a local scale, our data does not support this as a universal rule of thumb, which is how it's so often presented. This acts as an example of how sensible, common wisdom is not always the full truth.