MA26620 Assignment 01 – Regression and Housing Data

Ben Jaeger | [bej29@aber.ac.uk](mailto:bej29@aber.ac.uk) | 170033626

Contents

[1. Regression 2](#_Toc531157814)

[2. Housing 3](#_Toc531157815)

[2.1. Housing Data Introduction 3](#_Toc531157816)

[2.2. Exploring the data 3](#_Toc531157817)

[2.2.1. The effect of location on price 3](#_Toc531157818)

[2.2.2. The effect of house material on price 3](#_Toc531157819)

[2.2.3. The effect of location on material 4](#_Toc531157820)

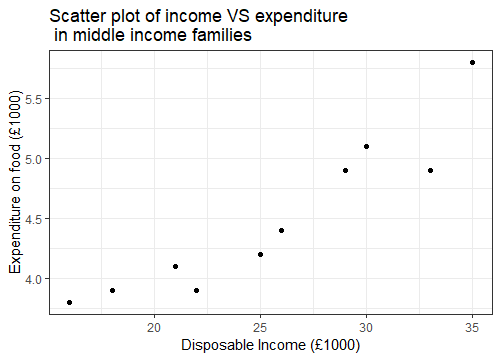
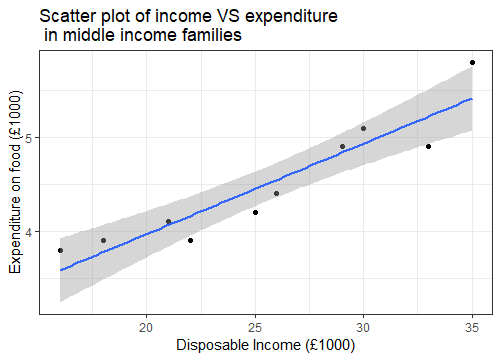
[2.2.4. The effect of house type on price and location on type 5](#_Toc531157821)

[2.2.5. The effect of price on offers received and relation to location 6](#_Toc531157822)

[2.3. Conclusion 7](#_Toc531157823)

[2.4. Future Investigation 7](#_Toc531157824)

# 1. Regression

We are given the disposable income and expenditure on food of 10 middle income families of the same size, these plotted on a scatterplot look like the diagram you see to the right. By running the command 100\*cor(income, expenditure)^2 in RStudio we can see that the percentage of variation in expenditure accounted for by its linear relationship with income is 86.96%. from this and the plot we can assume a linear regression model would be very appropriate.

We can see the linear regression line added to the plot, it has the equation y = 2.03891 + 0.09651x. Using the fact that S­­xy = 1182.1 and Sxx = 6861, we can work out β1 (whose equation is Sxy/Sxx) as being 0.09651 and β0 (whose equation is y-β1x) as being 2.03891. Using the equation, we can predict a family earning £29,000 would spend £4,837.70 on food (approx..) and a family earning £19,800 would spend £3,949.81 on food (approx.)

If a family were to reduce their income by £3,500, under this model, they’d reduce their expenditure by £337.76. this would be the same for however much the family earned as it’s a linear model being used

# 2. Housing

## 2.1. Housing Data Introduction

In this report I will be looking into a subset of some data collected by housing agents in Douglasville, Ohio, USA, this data is mostly details about the house that was sold (number of bedrooms, bathrooms, neighbourhood, etc) but also contains information on the sale such as the price it was sold at and the number of offers received.

In my report I will be generally exploring the data, but I also want to see what factors affect the final price and are there any trends related to it.

## 2.2. Exploring the data

### 2.2.1. The effect of location on price

The first thing I want to explore is the effect of the house’s location on price, do some neighbourhoods have noticeably different average prices than others or are they uniform across the board. I plotted the prices in a boxplot and got the plot you see in Figure 1. Clearly Spectralia’s house prices are much higher on average, digging into the data revealed that the mean price of houses in Spectralia is $159,294.90 against $125,231.10 and $110,154.50 in Realtown and Normsville respectively, a difference of 1.272 and 1.446 times respectively!

Figure 1

### 2.2.2. The effect of house material on price

The second thing I decided to investigate was the effect of housing material on price, the data provided only distinguishes between brick and non-brick materials, so we can’t do a more in-depth analysis but comparing the two variables, as seen in Figure 2:

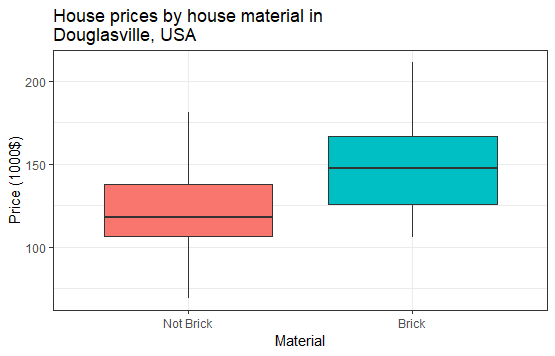
Here we can see that brick houses have a marked price increase over non-brick houses, with an average price of $147,769 for brick houses and $121,958.1, which is a 17.46% increase in price. I wanted to see if this was due to a number of outliers or outside factors such as house type (such as apartment, bungalow, etc), so I did a QQ plot of prices against the normal distribution, separated by house material, resulting in Figure 3:

Figure 2

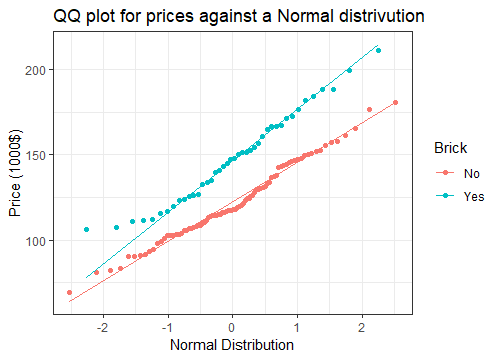
As we can see both are distributed pretty similarly to the normal distribution, so brick houses are just worth more than other material houses overall.

Figure 3

### 2.2.3. The effect of location on material

I said previously that I would be mainly investigating the effect of different factors on price, however the above two investigations made me curious as to whether there was a link between neighbourhood and material type.

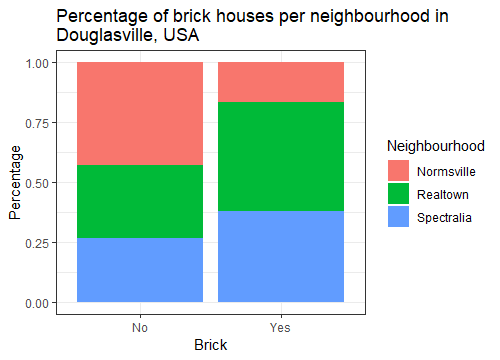
As we can see in Figure 4, there is disproportionally more non-brick houses in Normsville, thus explaining the price difference between Normsville and Spectralia, since non-brick houses are worth significantly less. However, the proportion of brick to non-brick in Realtown is remarkably similar, meaning the lower price in this neighbourhood is still a mystery.

Figure 4

### 2.2.4. The effect of house type on price and location on type

Given that there is no known reason for Realtown’s prices for being markedly lower so far, I decided to look into how the effect of house type on prices, and to see if Realtown had significantly more houses of a lower price or less of a higher price.

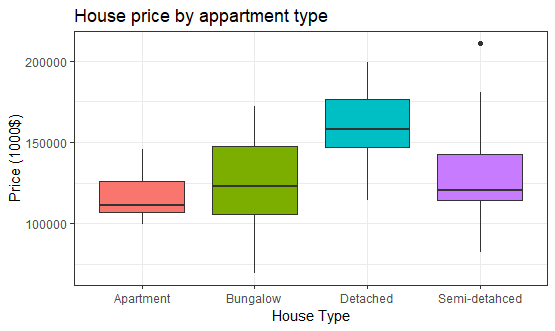
I started by seeing how the prices of the different house types (apartment, bungalow, semi-detached and detached) compared with each other using this graph:

Figure 5

As you can see, detached houses have a significantly higher median price compared to apartments, bungalows and semi-detached at $158,000 compared to $111,350, $123,100 and $120,500 respectively.

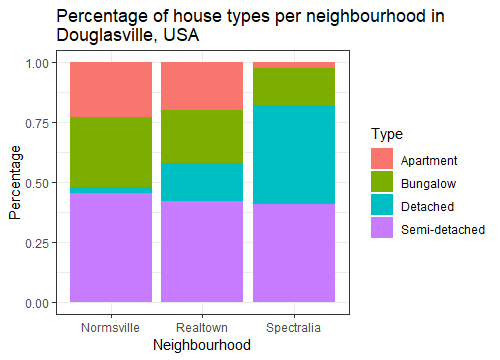
Now that we know the detached houses have a higher price on average, we can predict that the Realtown neighbourhood will have less of this type than Spectralia, graphing house type against neighbourhood we get the plot in Figure 6.

Figure 6

Our above prediction seems to be proven quite correct, there appears to be many more detached houses in Spectralia with 16, compared to Realtown at 7 and especially Normsville with only 1 detached house. As well as this we can see that Spectralia also has less of all the other housing types (apartments=, bungalows and semi-detached) than Realtown and Normsville which we saw above were all worth considerably less than detached houses.

### 2.2.5. The effect of price on offers received and relation to location

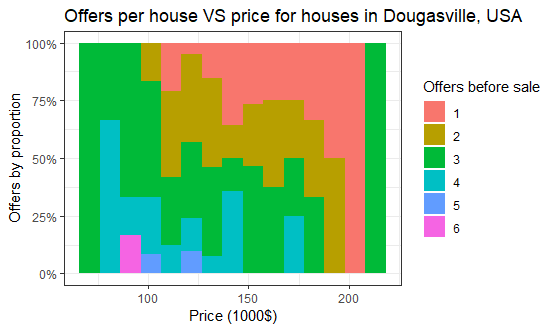
One thing I was curious about was how would the price affect something like the number of offers so I made the graph in Figure 7.

Figure 7

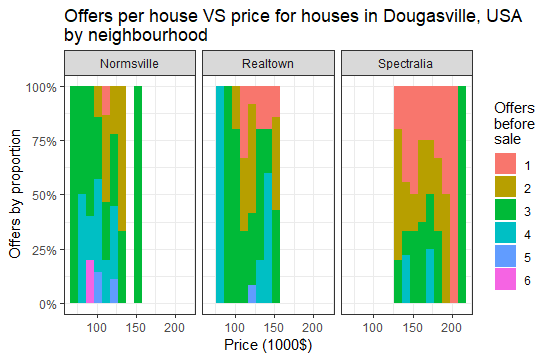
This is a modified histogram; the height is always full and is split into the number of offers a price band received. As you can see, as the price increases, the number of offers decreases, for example, no house in a price band below $100,000 received 6 offers and none above around $140,000 received more than 5. Initially, I was confused by this since I assumed a more expensive house would receive more offers as it was more desirable, however I realised it may be more desirable, but it was also less accessible and so, due to unaffordability, it would receive less, which is what we see in Figure 7.

Figure 8

I was curious to see how adding the neighbourhood datapoint would also affect this visualisation, so I divided up the data by neighbourhood and created the plot seen in Figure 8.

As we discovered earlier, Spectralia’s prices tend to be much higher, here signified by the fact that bins of a greater value tend to be filled in Spectralia and not in Realtown and Normsville. We can also see that Spectralia receives less offers per house on the whole, here shown by the fact that the Spectralia plot has more of the colours that signify less offers, this is in line with our discoveries about house price in Spectralia.

## 2.3. Conclusion

In summary, I concluded that brick houses are more expensive than non-brick houses by a significant amount and also that detached houses are the most expensive housing types to live in compared to other types of housing such as bungalows, semi-detached houses and apartments.

This fed into the discovery that Spectralia is the most expensive neighbourhood because it had a higher proportion of brick houses than other neighbourhoods and also both a higher proportion of apartments and lower proportion of other house types.

Lastly, I discovered that houses of a higher price, such as those in Spectralia, tend to get less offers before they sell since, presumably, they are more unaffordable to most people.

## 2.4. Future Investigation

In future, I would like to have investigated the bedrooms, bathrooms and sqft datapoints as I didn’t get the chance to here. I believe that the sqft datapoint specifically could have revealed that price is positively correlated to size and therefore go on to reinforce the fact that Spectralia is the most expensive neighbourhood.