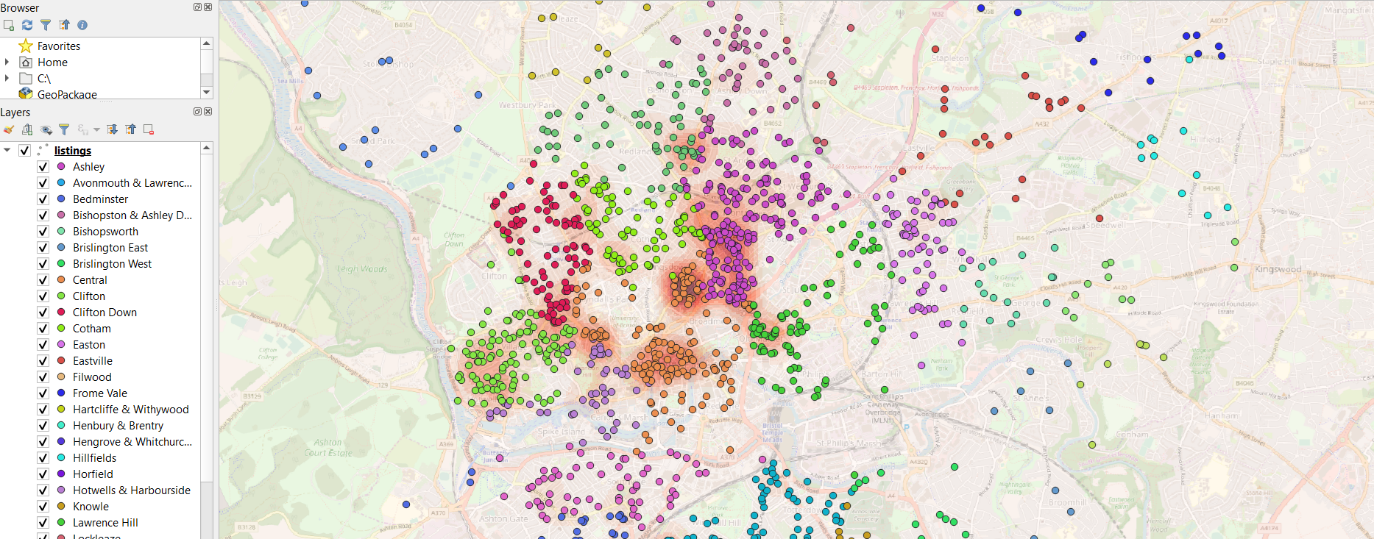
The Bristol Conundrum! Solving with CRISP-DM



**Image** – Bristol listing GIS grouped by area with an underlying Price heat-map

I was asked to do an analysis of a chosen data set using the CRISP-DM methodology . Basically this requires 6 steps :-

1) Business Understanding – what are we trying to answer/solve?

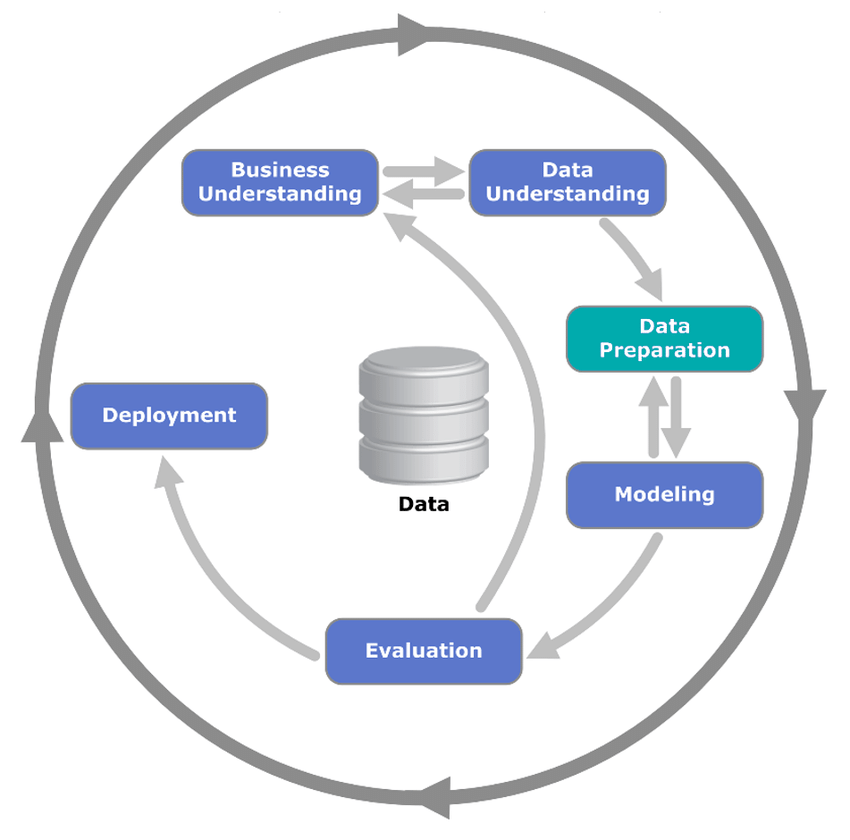
2) Data understanding – do we have the data necessary to answer the question? Do we need to collect more data

3) Data preparation – cleanup the data by removing or imputing data .

4) Modelling -

5)Evaluation/results- use the results of the modelling to answer the questions

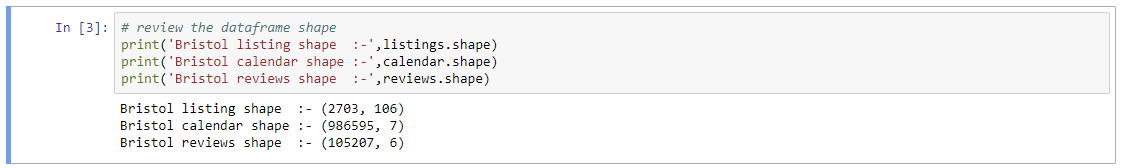
6)Deployment – communicate the results/interpretation of the analysis .

  
Figure 1: CRISP-DM Process

This blog is not a line by line of the above process but rather the final step of the CRISP-DM process . I will however touch on some of the steps to give you a better understanding of how I came to conclusions and tried to answer the proposed questions .

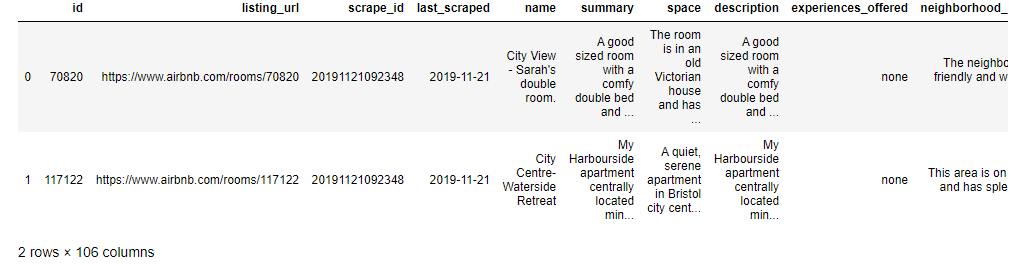
Why Bristol and Airbnb? My son is potentially looking to do his Masters in Music in Bristol. It’s only 1.5hrs down the road but I know he won’t be coming home so we would be visiting him. It would be good to spend a couple of days in Bristol rather than do day trips . The obvious Choice would be staying in an Airbnb .

So I had the area of interest and the [data](http://insideairbnb.com/get-the-data.html) I now needed to come up with some questions . For me the areas of interest would be focused on price and availability . With that in mind I took a closer look at the data . There were 7 files shared for Bristol . A quick look at them pointed me to look deeper at 3 specific csv files i.e. listings.csv , calendar.csv and reviews.csv

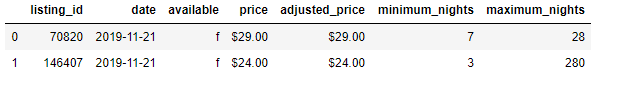


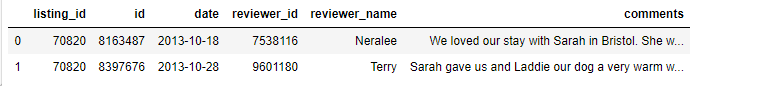
Initial investigation shows there is quite a bit of data to build my questions on . I needed to look at what was actually contained within the files

**listings.csv**



**calendar.csv**

**reviews.csv**

Although the reviews file had some interesting data , producing incites from it is a bit out of my league at my current level of expertise . I would definitely like to look at it more at a later stage of my studies .

Focusing on the calendar and listings data I came up with the following questions to try answer .

1) Are there any clear trends in the availability of the listings ?

2) Are there any trends in the pricing of listings across the year ?

3) How are the listings distributed across the neighbourhoods and is there a clear price grouping ?

4) Do any of the columns have a direct affect on the pricing ?

5) Could I train a model to predict the pricing of the listings ?

Before actually starting to answer these questions I had to check and clean-up the data to enable the analysis .This was done using a few techniques

i) remove columns or rows with high levels of missing data

ii) fill in missing data with either ‘0’ value or mean values

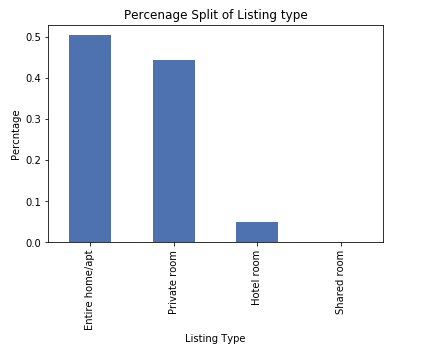
iii) impute values for categorical

iv)remove/filter out categories that could potentially create misleading trends

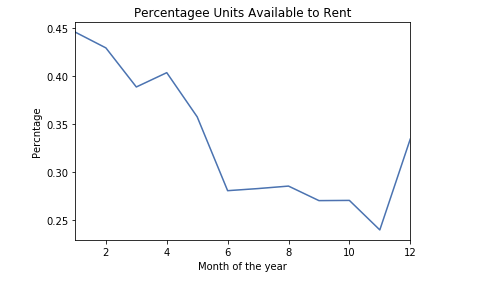
v) clean-up and convert data e.g. outliers and date formats

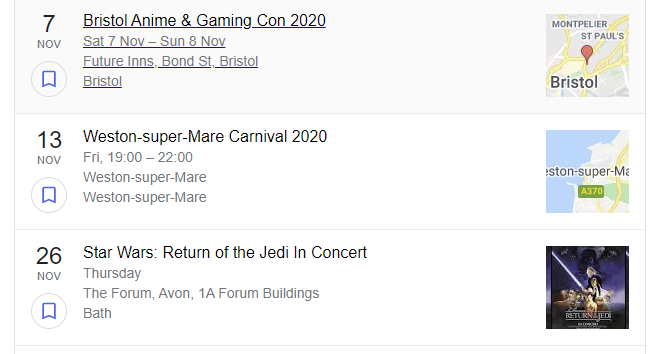
**Question 1 :- Are there any clear trends in the availability of listings ?**

To answer this question I used the calendar data set . The column that potentially gives this insight was **available** . However I wasn't sure exactly what this particular column meant . I looked for a description on the insideairbnb website but they don't provide (please advise if otherwise ) My thought was it could have one of two meanings 1) the home owner has made the listing available for renting or not ? 2) the home is booked or not booked ? I did a bit of reading on the internet to see if other users had the same uncertainty about the exact meaning . Seems to be leaning to the first possible meaning . I decided for this question I didn't need to know the exact meanings because I wanted to see if there was a trend on the availability not why it wasn't available . I did however only want to consider entire homes/apartments . This didn't dilute the data because this listing accounted for 50% of the listings close on 500,000 rows to work with .



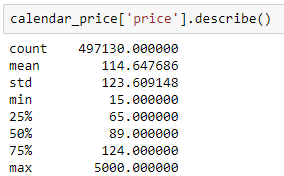
There does seem to be a clear trend of low availability between June and August . This would make sense as these are summer months and would be very popular so aligns closely with expectations . What did surprise was a continued downwards trend into November. November appears to have the lowest availability of the year . I looked for local events that could possibly have caused that dip but couldn't find anything of major importance .



****

**2) Are there any trends in the pricing of listings ?**

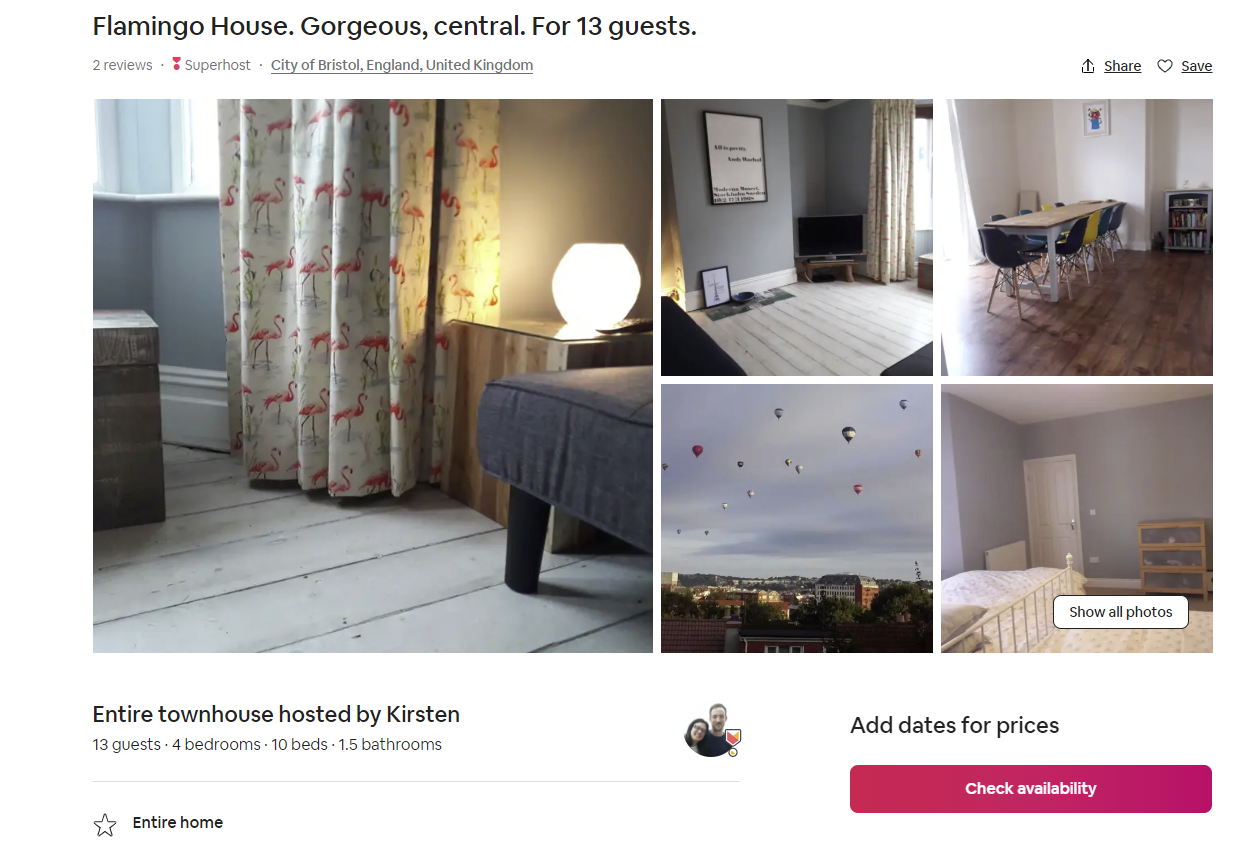
With a better understanding of when listings would be more readily available I started to look at trends in the per night pricing . Was there a cheaper time to go and did it match the availability trend above ? Were the months that had higher availability cheaper i.e. supply and demand? Again only looking at entire listings . I started to see the range of prices across the year .

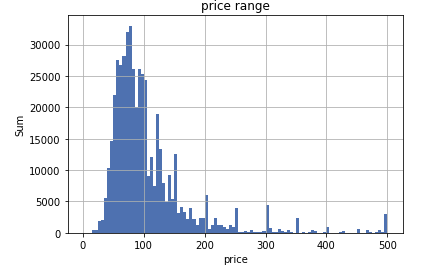
 There is quite a range starting at £15 per night at the lower end to £5,000 per night at the higher end.

Looking to get rid of the outliers I cut-off all listings above £500 using a rough equation of :-

Mean + (3\*Std Deviation)

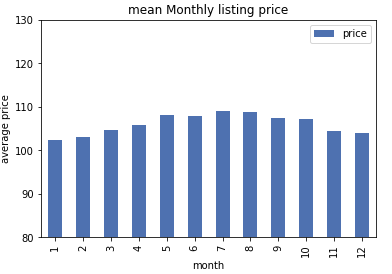
I was super curious what would warrant £5000 per night !!

 Its an impressive place that hosts 13 people but not worth £5000 a night . Closer investigation shows that the property has an £800 price tag for 5 nights of the week but Mondays and Tuesday are at the extreme rate . A mistake or cheeky risk ?

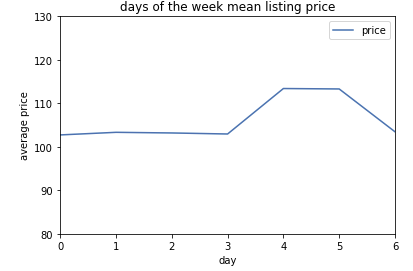
****

**Histogram above shows the range with outliers above £500 removed**

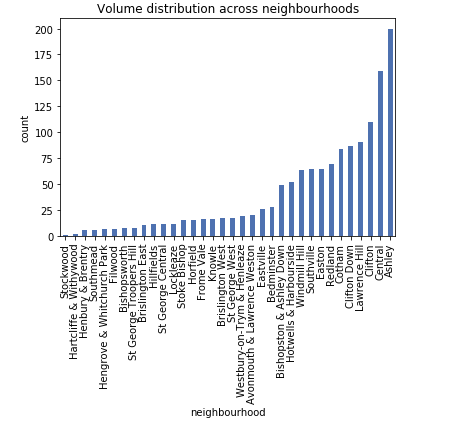
The two plots below show a trend in the average listing price across the year . The one on the left is the more granular month average and the one on the right is the weekly average . There is clearly a trend for an increase in prices across the warmer months of the year . What the month bar chart doesn't capture is the dip before Christmas and the spike again for New Year celebrations presumably .

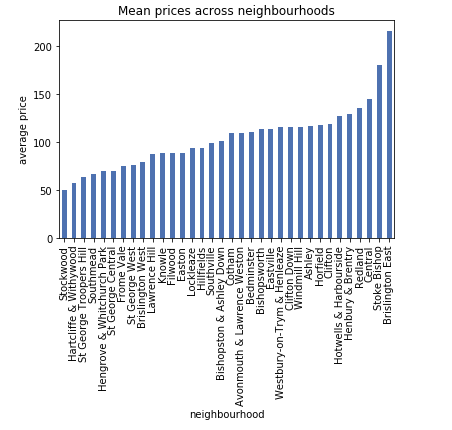


With a clear seasonal price variation it was interesting to see if there was a change in pricing across the week . There is in fact a jump on Friday and Saturday for the average pricing across the week . This is presumably to take advantage of the weekend breaks .

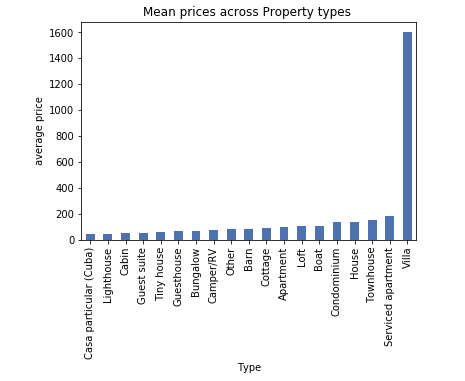
****

**3)How are the listings distributed across the neighbourhoods and is there a clear price grouping ?**

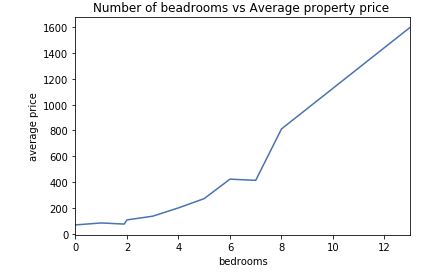
**Ashley has the largest volume of listings per neighbourhood and Stockwood the lowest .**

**Brislington East has the highest average price per list while Stockwood has the lowest .**

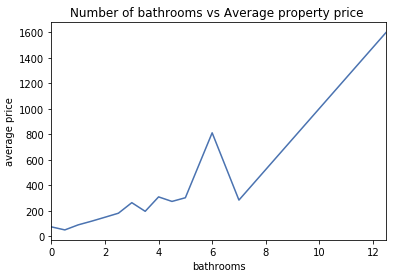
**4) Do any of the columns have a direct affect on the pricing ?**

****

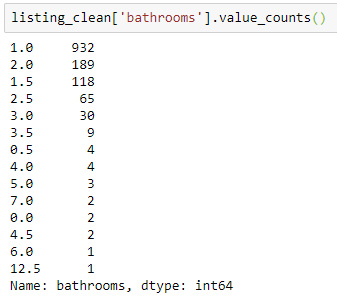
Villa has the highest mean price of all the property types with an average price of £1600 while all the other property types are below the £200 mark . This might have something to do with other factors like features offered  **T**here are some interesting e.g. Tiny House , Cabin & Lighthouse

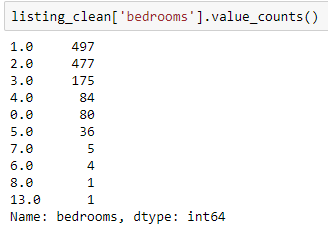
****

Definitely a relationship between the number of bedrooms and the average price of the listing . There is a slight deviation with property with 7 bedrooms but generally the relationship is linear

****

Again there is a clear relationship between the number of bathrooms and the average listing price . The only distortion is the pricing for the listings with six and seven bathrooms .

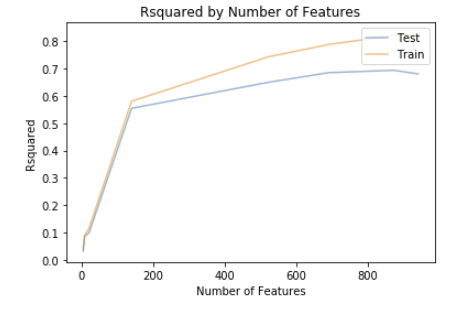
****

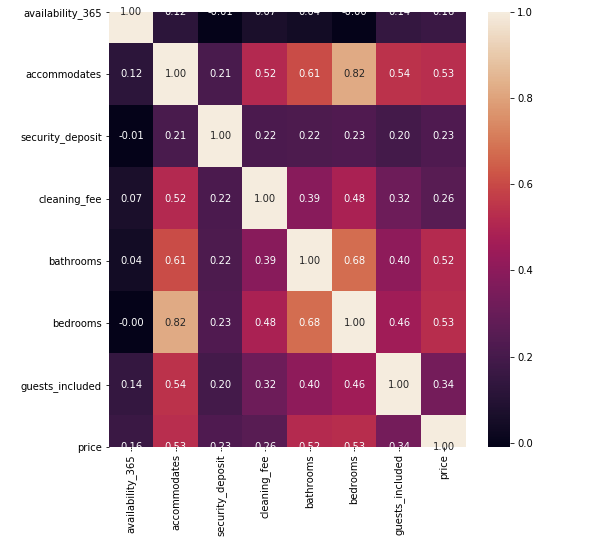
****

Looking at the volumes of listings in each of these categories you can see 0-5 range for bedrooms and 1-3.5 range for bathrooms . There is a concern that 2 of the listings have 0 bathrooms !

**5) Could I train a model to predict the pricing of the listings ?**

I started with a LinearRegression model to train and test my data but was getting really bad results and was starting to doubt my method . I had done all the house keeping , data cleansing , followed all the steps and had expected my data to produce good results . Then I remembered a graph in one of the lessons that illustrated what happens when we introduce too many features into the model.

Using the above I started to re-look at my data and started with only a few key columns that I felt had a direct bearing on the price . I would run a correlation matrix to validate my assumptions . After a number of trials I settled on the following columns



The results were better but not as good as I’d expected . The results are as follows

**Mean squared error of 6252.29**

**R^2 score: 0.36 .**

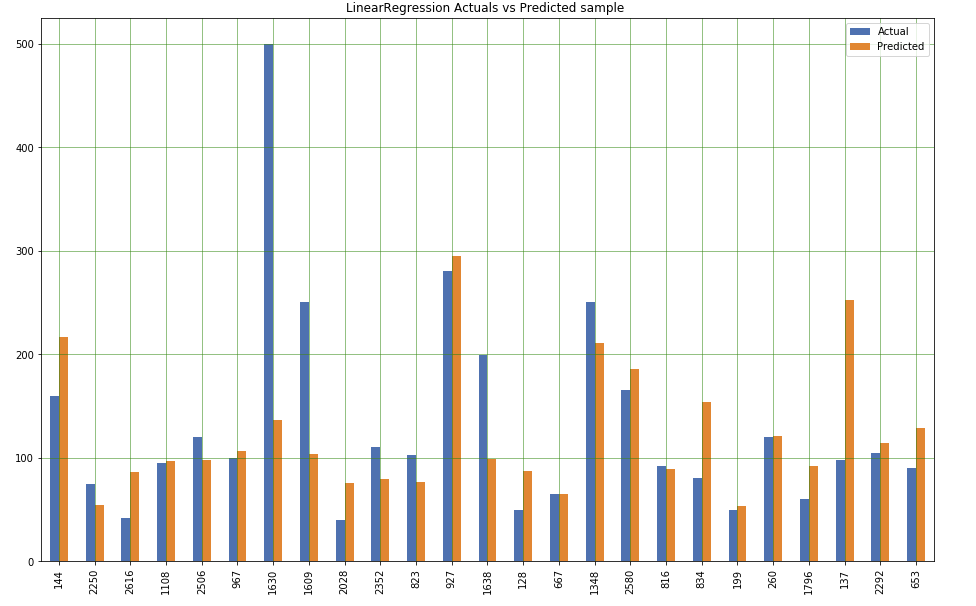
After a bit of reading I decided to try a RandomForestRegressor model . The results were better but still not great .

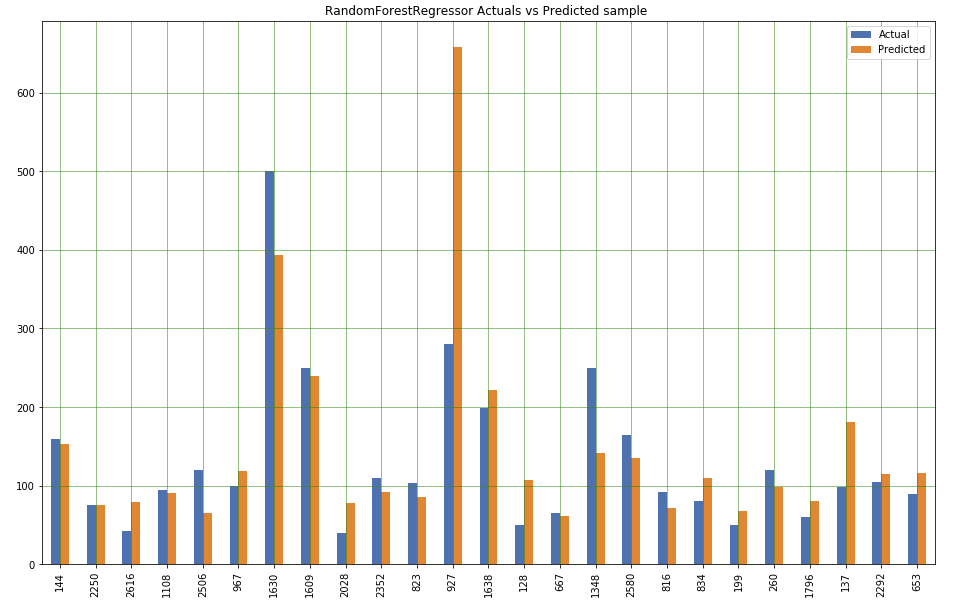
**RMSE test: 67.153**

**R^2 score: 0.540**

Feeling a bit out of my depth I decided to stop the testing as I had managed to improve the scores three times but wasn't sure of the next steps . Was this as good as it could get based on the limited data and features ? Or was there room for more improvement . Maybe something to come back and look at once I have a better understanding .

Below are sample graphs for each of the predicted vs actual models . There are a few big differences in the samples but on a whole they are quite close . Could those big differences be throwing off the scores ? I will come back at a later stage to see if I can answer and improve the scores .





**Conclusion**

I was able to show trends in the availability , coupled with the pricing trend and distribution across the neighbourhoods . Using this I would be able to choose the best time and location to visit Bristol . The worst time would be July to August because of low availability and higher prices . Best option would be to go in winter on a midweek break . This would fit in nicely with my sons timetable as the University would be closed for summer break during those peak times anyway . The other factor that affects price are number of bedrooms , bathrooms , type and area . It would be interesting in the future to see if I could pull some crime stats for the areas and give it a price to crime rating to find the optimum neighbourhood . Trying to train a model to predict was more challenging and needs a bit more knowledge and finesse . It did however give some insights into what is possible

It would have also been nice to use GIS in the future to represent some of the findings . I did create a simple layer, grouping the listing in neighbourhoods overlayed on a heat map of price and used for the title image .

The full Jupyter notebook is available on Github

**References :-**

Udacity Data Science course notes and worksheets [https://classroom.udacity.com/](https://classroom.udacity.com/me)

Data - <http://insideairbnb.com/get-the-data.html> then search for Bristol .

Figure 1 - <https://www.researchgate.net/figure/Cross-Industry-Standard-Process-for-Data-Mining-CRISP-DM-12_fig1_320100474>

<https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f>

<https://towardsdatascience.com/random-forest-in-python-24d0893d51c0>

<https://medium.com/@josh_2774/a-comparison-of-airbnb-homes-seattle-vs-boston-cdc0df2cfcd7>