## Introduction

This report provides an analysis of the housing market and a technical overview on the development of a location recommendation system. The initial challenge was set out by the need of understanding what the markers of an appropriate property are. A recommendation system requires input data for analysis to determine the suitability of a potential location for properties. Consequently, broadband speed, house price and crime rate were chosen as the main deterministic factors. Besides finding possible location candidates, the relation and interaction between each piece of data was analysed. Upon a quick inspection of the data, the hypothesis that crime rate and house prices have a strong inverse relationship was formulated. A critical review of the literature about crime and housing prices by Ihlanfeldt and Mayock (2009) suggest that the assertion stands to be correct. Thus, another goal of the report was to see if the findings corroborated with theirs.

## Data collection

The data was collected from different governmental and public bodies. For properties, HM Land Registry department provides data on properties sold in the United Kingdom and details of the transaction such as price, property type, build type, postcode, etc. A sample of two thousand properties which represent the total sold in April 2019 were collected from the counties of Warwickshire and Leicestershire. Criminality of a location was determined with the use of a police's public database who provides a range of methods for procuring the data(*Home | data.police.uk*, 2020). These two proved to be the most important since broadband speed remained similar across regions and had a lower variability. A regional, constituency-based dataset from a governmental analysis of Ofcom's data was utilised.(UK Parliament, 2020)

As described above, finding the data was not an issue however linking different pieces together to create a unified model proved to be a challenge. The ONS coding system uses geocodes that represent a wide range of geographical areas of the United Kingdom. (A Beginners Guide to UK Geography, 2020) Geographical units are subdivided and range from representing large areas (SOA's) to small locations that have a population of a hundred households (OA's, LSOA's).

## Data rationale

Early-stage findings suggested that the correlation between crime rate and house prices was lower than expected (-0.05) however this was as a result of bad methodology. Initially, the crime rate was determined by counting all incidents that occurred in the same LSOA as the house, but this provided an inaccurate representation. The solution to this issue was to use the public Postcodes.io database(Ideal Postcodes, 2020.). Postcodes.io falls under the MIT license and collects its data from governmental and public bodies. A simple API call for a postcode returns a JSON response that contains the longitude, latitude and all the ONS codes which the location falls under. This allowed a more restrictive, narrow search of crime that occurred on a specific postcode since the police's crime database also allows users to request a criminality report at a fixed geographical position. Furthermore, the broadband data was linked to the house with the administration ward code that is found in the postcode API call result.

### 2.3 Data collection and processing

To avoid any service interruptions that may stop the collection process abruptly and reduce the amount of load the public external service is experiencing, a Docker image of the service (Docker Hub, 2020) has been initialised on the localhost domain. This allows for internal calls to

be made and retrieved. The fetched results are parsed, and the longitude and latitude are used to create another request towards the police crime database (Police API Documentation | data.police.uk, 2020). As previously described, this allows specific data at a certain date to be collected. Although the properties used for this report are from the month of April 2019, all the crime collected were from the year 2018. The reason behind is the assumption that a potential house buyer would most likely check for a longer period of criminality history rather than a specific  month. Also, Carbone-Lopez and Lauritsen (2012) suggests a pattern between the season of the year and property crimes. By only considering a month of the year, this could make the sample not representative of the entire population. Furthermore, not all crimes were collected as a dictionary with crimes and their score was created and utilised to filter and create an overall criminality score in that location. Crimes such as shoplifting and bicycle theft were given a lower score whereas robberies, violent crimes and vehicle crime a higher one.

## Database rationale

An SQLite database consisting of four tables in a primary-foreign key relationship were created.(figure 1). Data redundancy is avoided by mapping each administration ward to a broadband speed.

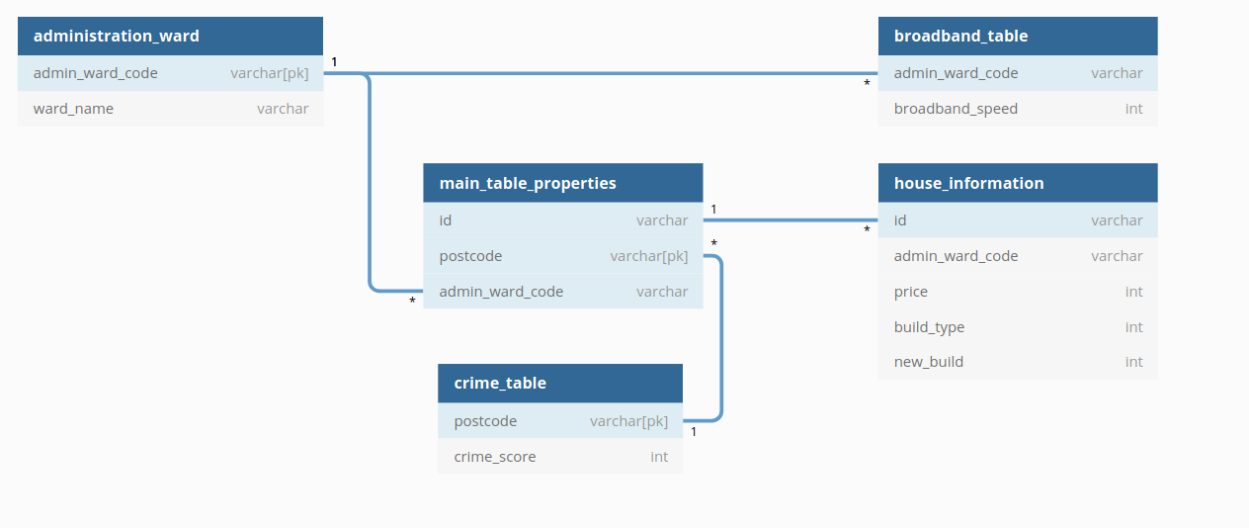


*Figure 1*

Also only one entry for each postcode and its corresponding crime rate value is allowed since the possibility of having multiple records of properties that occupy the same postcode. By enforcing these rules, the search time through each comma separated value files is reduced since the database is searched for the value needed in the case where it has been previously inserted.

## Data harvesting

The exact procedure of procuring the data is described as following. Each row of the file that contains the properties has the data described in the “house_information” table extracted, this includes a categorical variable that expressed if the property is a new built. To facilitate its use in creating a linear regressive model, it has been transformed from a string to an integer-based representation (one and zero). The postcode obtained is utilised to query the local postcodes’ database to determine its ONS administration ward code, longitude and the latitude. The administration ward code from the query response is required to read through the broadband speed CSV file and do a search for the speed at the “admin_ward_code” index. With the help of

geographical coordinates, a string request is formatted for each of the twelve months of the year and sent for processing to the online police database. The data received in JSON format is unpacked and run through a local dictionary that creates a score for each of the crime categories found. Insertion inside the database occurs gradually and data is checked for anomalies to avoid any unexpected behaviour and data corruption.

## Data analysis

A set of plots and graphs were created to explore the relationships between data and suggest hypotheses about the cause of observed effects. The Pearson correlation coefficient is a statistic that measures linear correlation between two variables. The values of the relationship coefficient are between -1 and +1. A relationship coefficient of +1 demonstrates that two factors are related in a positive straight sense, a relationship coefficient of -1 shows that two factors are related negatively.

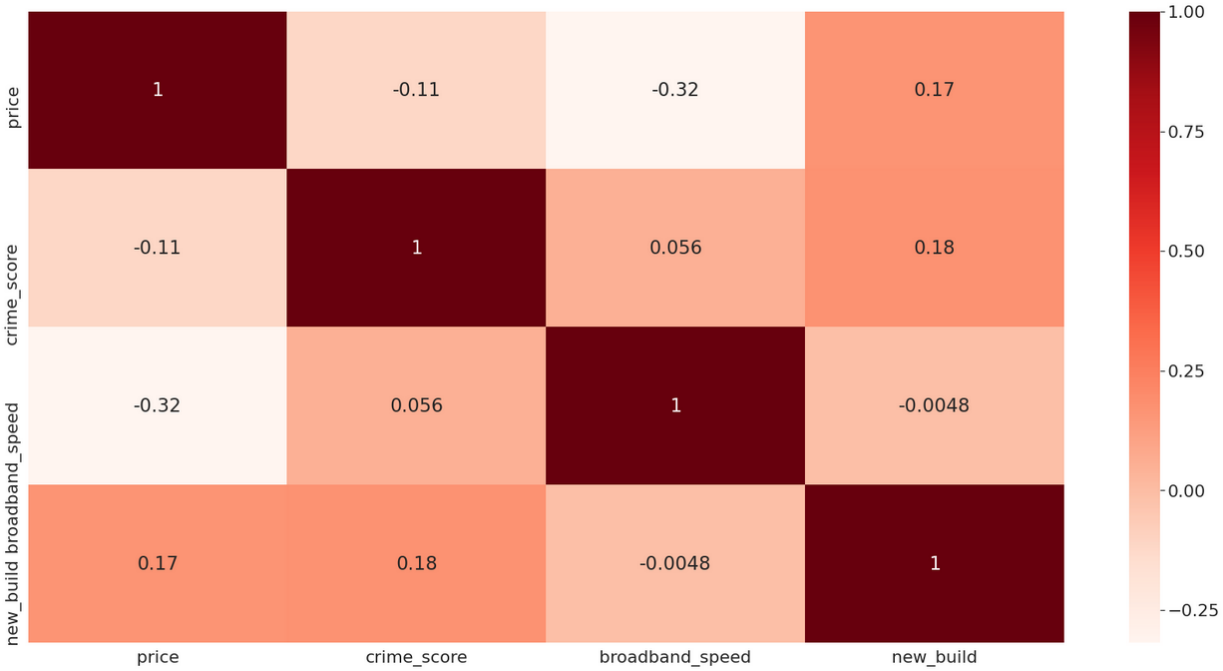


*Figure 2*

The matrix in figure 2 shows a correlation of -0.32 between broadband speeds and house prices which is more significant than the relationship with area criminality score (-0.11) and 0.17 for the type of property. Statistically these are considered insignificant.

Properties that have outliers situated 3 standard deviations away from the mean are calculated in the “detect_outlier” function and displayed in blue colour (figure 3).

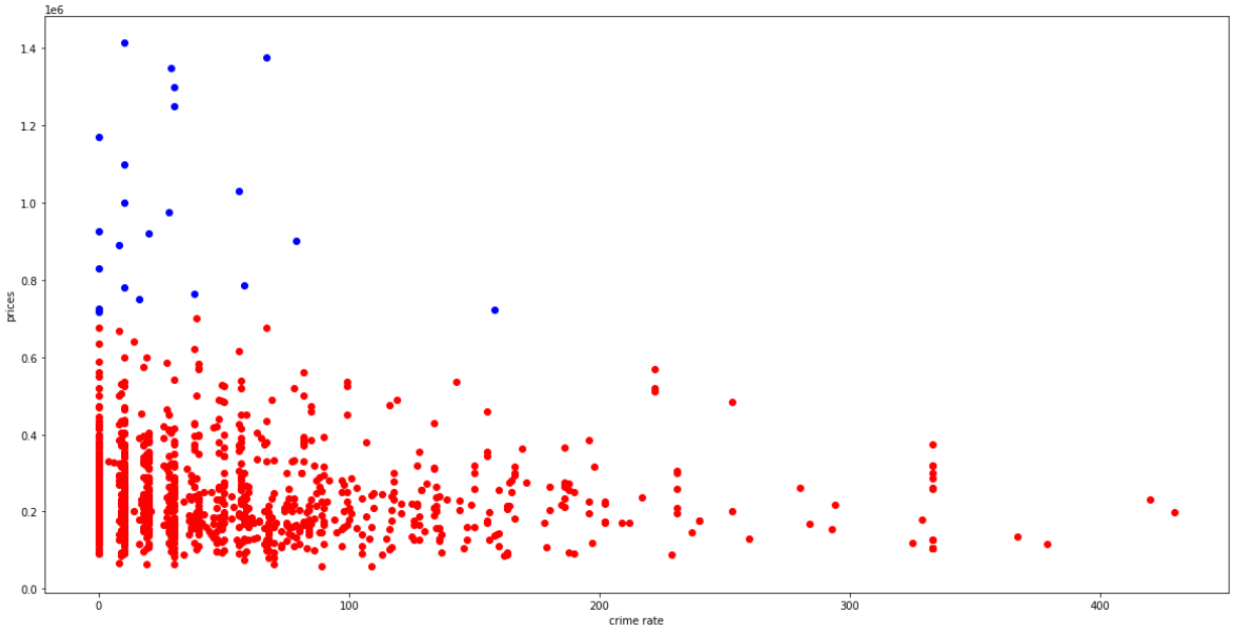


*Figure 3*

As a third factor the type of property is added along crime score and price to get an estimate of how much the type of house influences the price. (figure 4)

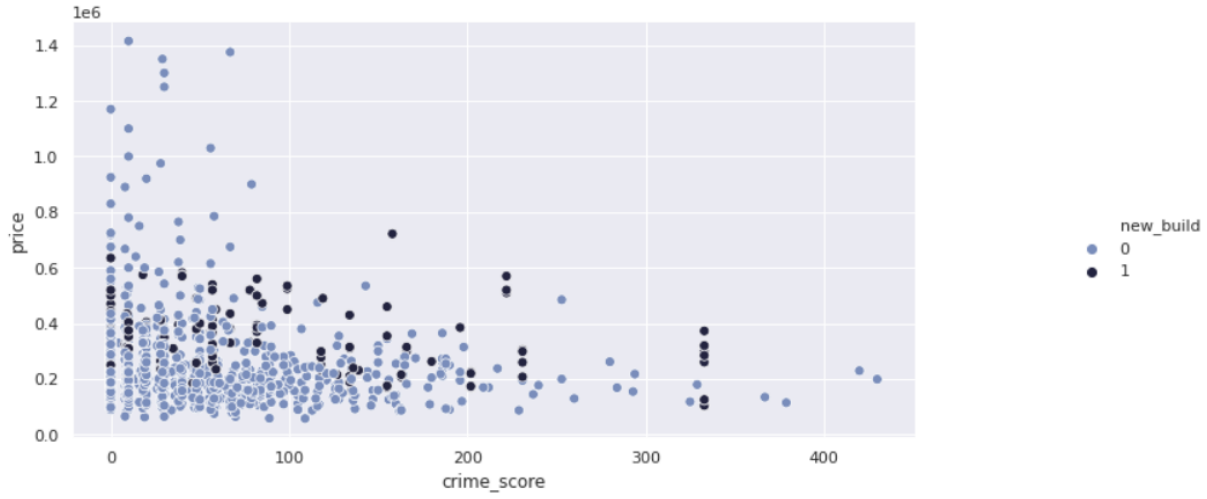


*Figure 4*

New built houses are slightly more expensive and as a feature of the property market in the area there are no new built houses in the outlier region.

By box plotting the building's categorical type attributes, most expensive properties appear to be detached houses. (figure 5)

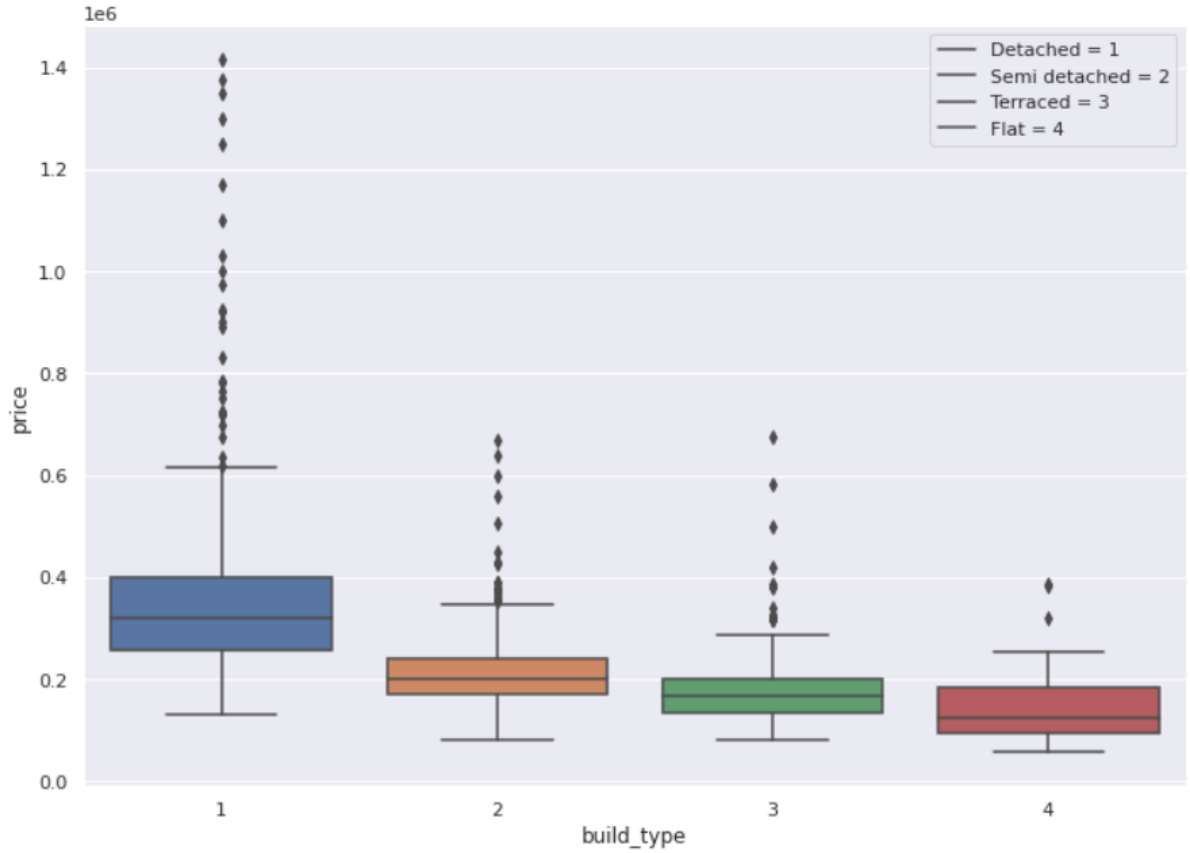


*Figure 5*

## Multiple regression

Pearson's correlation coefficient and linear regression both quantify the relationship and strength between variables and tend to indicate similar statistical outcomes however regression produces an outcome for the entire population. The 'skilearn' Python module allows users to easily create regressive models to test and predict values (Pedregosa et al., 2011). Data is split

into random partitions and a linear model is created with crime score, broadband speed and the type of house as an explanatory variable. (figure 6)

```
                 Coefficient
crime_score             -314
new_build              82672
broadband_speed         -3485
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.191
Model:                            OLS   Adj. R-squared:                  0.188
Method:                 Least Squares   F-statistic:                     51.00
Date:                Thu, 03 Dec 2020   Prob (F-statistic):           1.36e-29
Time:                        03:04:28   Log-Likelihood:                -8589.7
No. Observations:                 650   AIC:                         1.719e+04
Df Residuals:                     646   BIC:                         1.721e+04
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          4.206e+05   1.74e+04     24.132      0.000    3.86e+05    4.55e+05
crime_score     -314.0993     75.648     -4.152      0.000    -462.646    -165.553
new_build       8.267e+04   1.37e+04      6.031      0.000    5.58e+04     1.1e+05
broadband_speed -3485.4522    344.585    -10.115      0.000   -4162.094   -2808.811
==============================================================================
Omnibus:                      418.846   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5669.391
Skew:                           2.676   Prob(JB):                         0.00
Kurtosis:                      16.442   Cond. No.                         320.
==============================================================================
```

*Figure 6*

The intercept value of 420612.56 is the expected house price mean value when the predictors are zero. In this model the value has no meaning since broadband speed never touches the X axis. The model predicts values based on the previously split test data and plots them against real values. At this moment the results obtained so far describe a poor statistical model. The p-value which indicates statistical significance is 0 thus the initial hypothesis is null. R-squared value determines to what extent the variance of one variable explains the variance of the other values in the model. The regression returned the score of 0.191 which is low thus most the

variance cannot be explained. Residual values determine how well the predictions were. In this context they are described as the difference between the predicted house prices and the observed house prices. The visualiser in figure 7 shows non-constant variance between the residuals and variance in homogeneity is low as the distribution is poor since the residuals are grouping together to form a pattern. Also there exists an unequal number of predictions that exceed the expected value. On top of that there are a disproportionate number of negative outliers. In conclusion the residual values describe a poor model.

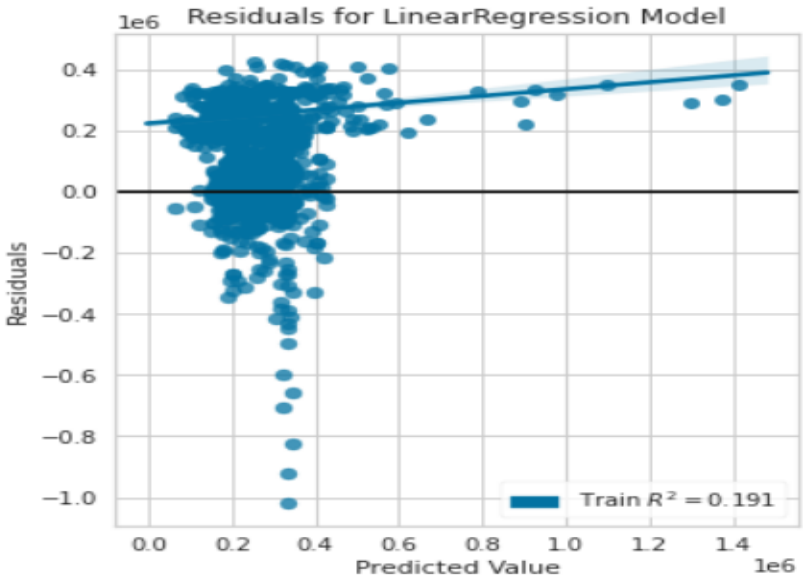


*Figure 7*

## Recommendation System

The recommendation system is created to display the top three towns or cities with the best overall score. The score is based on the price, broadband speed and criminality of the area.

A rank percentile-based approach has been designed to assess each property and location. In statistics percentiles describe a quantile which divides the given probability distribution into equal intervals. (Giri & Banerjee, 2014, p. 157). For example, a property situated in the 90th rank percentile is the value where 90% of the observations are found.

The data is retrieved locally and adjusted to fit the system. A house price in the 90th percentile describes in simple terms that the property is more expensive than 90% of the others in the population which would bring its score down. Not the same can be said about broadband since speeds in the 90% represent a positive factor. The value of 1 has been used to subtract the percentile of house price and crime score to adjust the opposing meaning in the representation of these two predictors. Therefore, a percentile of 0.9 receives a score of 0.1 which is then multiplied by 10 to give a final score of 1. Each set of values is compared against the population, gets its percentile and the resulting overall score is determined by the mean of the three attributes. The existing data frame has its score row populated and sorted from highest to lowest. Properties with the highest score have their location retrieved and once a count of three is found, the search is stopped, and results are presented to the user (figure 8).

```
[48]: precentile_based_non_weighted()
```

```
Fetching best locations...
Coalville North
Whitwick
Mallory
                                     id admin_ward_code  postcode   price  \
255  8F1B26BD-DE2E-53DB-E053-6C04A8C03649       E05010112  LE67 2HG  125000
288  87E1551E-540B-6405-E053-6C04A8C0B2EE       E05010099  LE67 5PH  131250
435  87E1551E-E51D-6405-E053-6C04A8C0B2EE       E05005481   LE9 8HA   85000
429  8A78B2AF-BD8D-5CB0-E053-6B04A8C0F504       E05005481   LE9 8DE   93000
571  8A78B2B0-2937-5CB0-E053-6B04A8C0F504       E05007489  CV10 7BL  128500
..                                    ...             ...       ...     ...
663  8A78B2B0-2838-5CB0-E053-6B04A8C0F504       E05007490  CV11 6SB  246000
80   8A78B2AF-BF8F-5CB0-E053-6B04A8C0F504       E05005534   LE2 4FW  232500
247  919FEC05-71A4-9A90-E053-6C04A8C0A300       E05010112  LE67 2EG  276950
666  87E1551E-B098-6405-E053-6C04A8C0B2EE       E05007490  CV11 6WL  239995
248  8F1B26BD-E2AB-53DB-E053-6C04A8C03649       E05010112  LE67 2EG  276950

     broadband_speed  crime_score  final_score
255         73.40000            0      9.20000
288         68.60000            0      9.00000
435         61.70000            9      8.70000
429         61.70000            9      8.60000
571         61.90000            0      8.60000
..               ...          ...          ...
663         63.30000            0      7.10000
80          63.10000            9      7.10000
247         73.40000            0      7.10000
666         63.30000            0      7.10000
248         73.40000            0      7.10000

[100 rows x 7 columns]
```

*Figure 8*

## User based recommendation system

While the results are representative of what a good property, location is they do not consider the users' preference. In addition, a user-based recommendation system has been created. This facilitates the user by giving him the choice of expressing how important each attribute is when deciding a potential property location. While the previous method used percentiles, in this case the z-score has been chosen. The standard score represents the number of standard deviations the value is from the population mean. The user centrism in this approach comes from computing a percentage of importance for each of the characteristics by requesting a score from 1 to 10 for the three attributes. In this case, adjustments are made before the multiplication and the sign value of the broadband score is reverse to solve the representation

error. The figure below outlines a user that prefers an area with lower house prices in the detriment of criminality score and broadband. Each standard score has been multiplied with its weighted score and a mean of each value has populated the “final_score” column.(figure 9)

```
[54]: # precentile_based_non_weighted()
      weighted_z_based_recommandation_system()

      Select a level of importance from 1-10 for the following
      Crime Rate: 1
      House Price: 10
      Broadband Speed: 1
      There are the top 10 houses sold which fit your criteria best
                                           id admin_ward_code  postcode  price  \
      585  8A78B2B0-78AB-5CB0-E053-6B04A8C0F504       E05007477  CV10 8EN  63000
      714  8A78B2B0-28C6-5CB0-E053-6B04A8C0F504       E05008982  CV21 1HH  66000
      572  8A78B2B0-2980-5CB0-E053-6B04A8C0F504       E05007475  CV10 7BY  65000
      43   8CAC1318-6AC9-0253-E053-6B04A8C08E51       E05005536  LE18 4AB  65000
      435  87E1551E-E51D-6405-E053-6C04A8C0B2EE       E05005481   LE9 8HA  85000
      735  98C75472-7CCD-72E9-E053-6B04A8C042F0       E05008981  CV21 2JY  58000
      296  87E1551E-E537-6405-E053-6C04A8C0B2EE       E05010113  LE67 6LL  90000
      429  8A78B2AF-BD8D-5CB0-E053-6B04A8C0F504       E05005481   LE9 8DE  93000
      428  8CAC1318-6BDA-0253-E053-6B04A8C08E51       E05005481   LE9 8DD  88000
      185  8A78B2AF-BD7C-5CB0-E053-6B04A8C0F504       E05005452   LE4 8JF  82000

           broadband_speed  crime_score  final_score
      585         59.20000           19     -1.17146
      714         55.50000            8     -1.14800
      572         48.20000           30     -1.08489
      43          56.30000           70     -1.08065
      435         61.70000            9     -1.07822
      735         52.10000          109     -1.04649
      296         59.40000           10     -1.03679
      429         61.70000            9     -1.03466
      428         61.70000           34     -1.03072
      185         53.90000           30     -1.02444
      Best locations based on your criteria are:
      Nuneaton and Bedworth
      North West Leicestershire
      Rugby
      North Warwickshire
      Charnwood
      Oadby and Wigston
      Hinckley and Bosworth
```

*Figure 9*

As expected the displayed results present the cheapest properties sold in Warwickshire and Leicestershire in the month of April 2019.

## Ethical discussion

All data used in this report has been collected or originated from governmental institutions which are compliant with the Data Protection Act 2018. Properties sold in the United Kingdom have their transaction information made available through the HM Land Registry. Data used for this report does not contain any personal details such as the registered owners or lenders of the

property. Such details have not been requested and would not be feasible to request given that a monetary charge is required.

While the property and broadband data require no ethical considerations, not the same can be said about crime. To discern which locations are better than others, a crime score has been calculated based on the postcode of the property's location. These matters are dismissed since the data provider has created a feature which anonymises the exact location of a crime (About | data.police.uk, 2016). Therefore, no postcode has its own unique crime statistic but shares it with other postcodes and is linked to a main representative randomized point situated into its vicinity.

## Critical reflection

All code presented in this report has been written in Python3 (Guido Rossum & Drake Jr, 1995). It has been an implementation decision made with the goal of creating modularizable code. At this moment most tasks are split into different functions and utilise some variables declared in the global namespace. To create, manipulate data and visualise it, Jupyer Notebooks, a web-based interactive computing environment have been utilised (Levy et al., 2016). One of the reasons for using Jupyer has been the fact that it allows users to run small sections of code and retain variables, results in memory until the kernel is closed.

A future addition and improvement to the project would be the creation of a separate module that deals specifically with data harvesting since this is the most intensive part of the project. Rather than skipping early through retrieved incomplete CSV records, testing would allow to craft rules that explicitly define what a bad record is. This would be especially useful if the project was extended to retrieve data not for a specific month but for a year or more.

The data collection can be time-consuming because it is dependent on the external calls made to web services to retrieve data however this allowed a more flexible way of linking pieces of data together. The local Docker database contains an abundance of data that can be used to extend the functionality and trajectory of the project. Execution performance was acceptable but can be improved because the API providers allow up to 15 requests per second but currently only an average of 2 are made. Making use of 15 requests per second with a concurrent implementation would make the data harvesting process faster.

In the domain of data science there exists a conundrum between “more data or better algorithms” for better models. In this project, results have been very poor in the model presented and this is due to the amount of data utilised. The coefficients obtained in Pearson‘s correlation measure revealed the mediocre choice of data. More important data points could have been selected for example: area surface of property, number of rooms, proximity to local institutions such as hospitals.

In the future a more advanced recommendation system can be implemented. A memory based collaborative item would require a data set with how different individuals rated properties. Once the user’s preference is collected then the cosine distance between each data point can be found and a score of similarity is extrapolated.

The foundations for the project have been laid, the only requirement is the addition of other data sets to improve the regressive model and the creation of a better recommendation system.

## References

A Beginners Guide to UK Geography. (2020, October 23). Geoportal Statistics https://geoportal.statistics.gov.uk/

About | data.police.uk. (2016). Police.Uk. https://data.police.uk/about/#location-anonymisation

Carbone-Lopez, K., & Lauritsen, J. (2012). Seasonal Variation in Violent Victimization: Opportunity and the Annual Rhythm of the School Calendar. Journal of Quantitative Criminology, 29(3), 399–422. https://doi.org/10.1007/s10940-012-9184-8

Docker Hub. (2020, April 1). Hub.Docker.com. https://hub.docker.com/r/idealpostcodes/postcodes.io

Giri, P. K., & Banerjee, J. (2014). Introduction To Statistics (9th ed., p. 157).

Guido Rossum, & Drake Jr, F. L. (1995). Python reference manual. Cwi -01-01.

Home | data.police.uk. (2020). Police.Uk. https://data.police.uk/

Ideal Postcodes. (n.d.). Postcodes.io. Postcodes.Io. Retrieved December 2, 2020, from https://postcodes.io/

Ihlanfeldt, K., & Mayock, T. (2009). Crime and Housing Prices. https://coss.fsu.edu/dmc/wp-content/uploads/sites/8/2020/09/02.2009-Crime-and-Housing-Prices.pdf

Levy, N., Aurélien Naldi, Céline Hernandez, Gautier Stoll, Thieffry, D., Andrei Zinovyev, Calzone, L., & Loï Paulevé. (2016). Prediction of Mutations to Control Pathways

Enabling Tumour Cell Invasion with the CoLoMoTo Interactive Notebook (Tutorial). Computational Engineering & Design Group.

Pedregosa, F., Pedregosa@inria, F., Fr, Org, G., Michel, V., Fr, B., Grisel, O., Grisel@ensta, O., Blondel, M., Prettenhofer, P., Weiss, R., Com, V., Vanderplas, J., Com, A., Cournapeau, D., Varoquaux, G., Gramfort, A., Thirion, B., Dubourg, V., … Duchesnay@cea, E. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot Edouard Duchesnay. Journal of Machine Learning Research, 12, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Police API Documentation | data.police.uk. (n.d.). Data.Police.Uk. https://data.police.uk/docs/

UK Parliament. (2020). Constituency data: broadband coverage and speeds. *Commonslibrary.Parliament.Uk*. https://commonslibrary.parliament.uk/constituency-data-broadband-coverage-and-speeds/