A black background with blue text

Description automatically generated

**Individual Assignment**

Sandesh Sapkota

Softwarica College of IT and E-Commerce, Coventry University

ST5014CEM Data Science for Developers

Siddhartha Neupane

August 2024

# Table of Contents

[Table of Contents 2](#_Toc174572438)

[Table of Figures 3](#_Toc174572439)

[Introduction 4](#_Toc174572440)

[Cleaning Data 5](#_Toc174572441)

[Housing Data 5](#_Toc174572442)

# Table of Figures

# No table of figures entries found.

# Introduction

This assignment shows the process and outcomes of using data analysis to recommend the appropriate city in the United Kingdom to purchase a property in the United Kingdom and particularly in Cornwall or Bristol counties. Each city has their own advantages but buying a house requires detailed analysis of several factors influencing property valuation and quality of life. The primary factors that are usually considered while buying a property include the housing price, crime rate in the neighborhood, internet connectivity, education and so on. This project utilizes the data made available by the government of United Kingdom and public sources to facilitate buyers to make informed decisions.

The objective of this project is to develop a recommendation system that uses data to analyze and compare the cities of Bristol and Cornwall based on several factors. The system provides a score to each city of the two counties based on the given criteria and rank the towns accordingly. The final goal is to suggest the top three cities to purchase a house to the friend in the context. This report outlines the process of obtaining the data, cleaning and preprocessing, performing exploratory data analysis and linear modeling to develop the recommendation system.

# Cleaning Data

Data cleaning is an important step in the data science lifecycle which comes right after obtaining the datasets. This step ensures that the datasets are accurate and consistent and prepares them appropriately to be used for analysis. The datasets obtained from the UK government and other public institutions also had several inconsistencies and inaccuracies. Each dataset were carefully cleaned and prepared to be used for the further analysis and development of the recommendation system.

## Cleaning Housing Data

The housing price data from 2020 and 2023 were combined into one using the bind\_rows() function in R. Then the data was filtered to include the housing data of Cornwall and Bristol only. A new column was added named “Year” by extracting the year from the existing column “Transaction\_Date”. After that only the required columns were selected. Finally, the null values and redundant entries were removed using na.omit() and distinct() respectively. Now, the dataset was saved using the write\_csv() function.

## Cleaning Broadband Speed

The broadband speed dataset was loaded using read\_csv() function selecting the relevant columns. The the columns were renamed to simpler names for easier operation. For example, “Median download speed (Mbit/s)” was renamed to “MedianDownSpeed”. The final columns were ‘Postcode’, ‘MedianDownSpeed’, ‘MedianUpSpeed’, ‘AvgUpSpeed’, ‘MaxUpSpeed’, ‘AvgDownSpeed’ and ‘MaxDownSpeed’. Then the null values were removed using the na.omit() function. After that, the broadband data was merged with housing dataset using inner\_join() function in R. The two datasets were joined based on the common field “Postcode”. Finally, the redundant rows were removed using the distinct() function. In this way, the broadband speed dataset was cleaned and processed for further analysis. The cleaned dataset was then saved as a CSV using the write\_csv() function.

## Cleaning Crime Data

The crime datasets from Bristol and Devon & Cornwall in the years 2020 to 2024 were loaded using the read\_csv(). Then all the datasets were combined together using the rbind() functions and converted into a tibble. Since the crime dataset did not include postcodes but includes the LSOA codes, another dataset postcodes to LSOA was also loaded and cleaned to join with the crime dataset. Only the required columns were selected from the LSOA and crime data. From the combined crime dataset, month, LSOA code, crime type, and falls within columns were selected. The columns were renamed for simplicity in further processing. Similarly, from the LSOA to postcode dataset, only ‘lsoa11cd’, ‘lsoa11nm’, ‘ladnm’, ‘pcds’ were selected. The data frame was then filtered to include the counties Bristol and Cornwall only. The duplicate values for LSOA code were checked and removed from both crime and lsoa datasets. Finally, the selected crime dataset was merged with the lsoa dataset using left join by the column LSOA code common in both datasets. Then two new columns ‘Year’ and ‘Month’ were created using the mutate function from the original Month column by trimming from 1 to 4 and 6 to 7 respectively. Now, the population data was also merged using left join. Finally, distinct() and na.omit() functions were used to remove the redundant rows and null values to clean the final data. Now, the dataset was saved using the write\_csv() function.

## Cleaning School Data

The performance school datasets from the year 2021 to 2022 were loaded using the read\_csv() function from the readR library. Then the datasets were filtered to only include the relevant columns