TASK

Task 1: House prices in Birmingham (40%)

For this task you will need to download the csv file ‘Prices.csv’ from Blackboard. The file ‘Prices.csv’ contains the prices paid for all houses sold in Birmingham over a 25 year period from 1st January 1995 to 31st December 2019 and was obtained from the UK land registry website (<https://landregistry.data.gov.uk/app/ppd>).

The file ‘Prices.csv’ has 14 columns:

unique id, price paid (in £), deed date (in the form day/month/year), postcode, property type (‘D’ for Detached, ‘S’ for Semi-detached, ‘T’ for Terraced, ‘F’ for Flat, ‘O’ for Other), new build (‘Y’ for Yes, ‘N’ for No), estate type (‘F’ for Freehold, ‘L’ for Leasehold), saon, paon, street, locality, district, county, transaction category (‘A’ for Standard, ‘B’ for Additional).

For this task you need to write a code to do the following:

1. Read in the csv file to create a dataframe. Remove the columns named unique id, saon, paon, street, locality, district and county. Further, remove all the rows which have ‘O’ in the property type column. In the new build column, convert all of the ‘Y’s to 1 and all of the ‘N’s to 0. In the estate type column, convert all of the ‘F’s to 1 and all of the ‘L’s to 0. In the transaction category column, convert all of the ‘A’s to 1 and all of the ‘B’s to 0. Remove all of the rows which don’t have a postcode. Now make a copy of the postcode column and name it ‘PC’. ‘PC’ represents a simplified numerical postcode. Now for each element in ‘PC’ remove all the letters and spaces to yield a 2 or 3 digit number.
2. Create a function to turn a date from the form day/month/year into a single number representing the number of days starting from 1st January 1995. Ensure your function starts with 1st January 1995 as Day 1 and ends with 31st December 2019 as Day 9131. In the data frame create a new column called ‘days’ and using your function convert the date given in the deed date column into a number in the column called ‘days’.
3. Create 4 bar charts to illustrate the number of sales for: property type, new build, estate type and transaction category. [Do not include ‘Other’ in property type.]
4. Create a data frame only containing data on detached houses. Let n be the number of days starting at n=1 for 1st January 1995 and ending at n=9131 for 31st December 2019. Consider the number of sales of detached houses over each 365 day period. (Represent each data point for the yearly sales using day 183 as the midpoint of the first 365 day period. Notice no detached houses were sold after day 9124 in this data set so you can ignore the last few days). Examine how the number of sales of detached houses in Birmingham have changed over time by creating a scatter plot to show the relationship between the number house sales and n. Include the straight line of best fit. How well does your line fit the data?
5. Determine the mean price for a detached house in each 365 day period. Use the mean house price over each 365 day period to represent the average price of the house on day 183 of each 365 day period. To examine how the prices of detached houses in Birmingham have changed over time, create a scatter plot to show the relationship between the mean house price and n. State and include the straight line of best fit. How well does your line fit the data? Using the line of best fit, normalise all of the price paid data in your detached houses data frame and store these values in a column named ‘normalised price’.
6. Create a new Data frame only including ‘PC’ and the mean normalised price paid data for each ‘PC’. Determine the ‘PC’s with the 10 largest normalised price paid for detached houses.
7. For detached houses, split the data into (80%) training data and (20%) test data. Use an appropriate regression model from sklearn to predict the price paid using the variables: ‘days’, ‘PC’, ‘new build’, ‘estate type’, and ‘transaction category’. Test your model using the test data set. Discuss your results.