Group Project Proposal

Group Name: Group 8
Group Members: Amal Byju, Xin Wan, Yi Hua Huang

Proposed Data Set: House Price Data, England & Wales, 2002 to 2021. The data set consists of transaction prices of properties in England and Wales collected from 2002 to 2021. The same property may be sold multiple times within 19 years, but the transaction identifier of each time is unique. The dataset contains 19,595,552 records and it takes 2.73 GB. There are 16 columns that indicate the following:

- Transaction Unique Identifier (string)
- Price (decimal)
- Date of Transfer (date)
- Postcode (string)
- Property Type (string)
- Old/New (string)
- Duration (string)
- PAON (Primary Addressable Object Name) (string)
- SAON (Secondary Addressable Object Name) (string)
- Street (string)
- Locality (string)
- Town/City (string)
- District (string)
- County (string)
- PPD Category Type (string)
- Record Status - monthly file only (string)

We have also chosen a second dataset that will help with our analysis. The file size is 470 kB and contains the median earnings in England & Wales over the years. There are 668 records and the columns are:
Country (string), Authority Name (Authorities in England & Wales) (string), one column for each year from 2002 to 2021 (contains median income values) (decimal) and another set of columns for years from 2002 to 2021 (where each column contains ratios of median house price to median earnings) (decimal).

Business Questions:

1. What are the different factors that affect house price?
2. How house prices have changed over the years in different counties of England and Wales. Moreover, we'll be comparing median house price to median earnings over the years.
3. Predict the price of a house given some input features.

Analysis:
We plan to load both datasets to S3 first. Then we'll spin up an EMR cluster for some initial data wrangling and transformation using Pig, Hive and Impala. Column names need to be revised and bad data/ missing values need to be removed/ imputed with calculated values. SparkR, Jupyterlab, Tableau and R will be used for exploratory analysis, predictive analysis and data visualization.

Sources:

1. https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/ratioofhousepricetoresidencebasedearningslowerquartileandmedian
2. https://www.kaggle.com/datasets/dmaso01dsta/house-price-data-england-wales-1995-to-2019
3. https://www.kaggle.com/datasets/bkoohy/uk-house-price-paid-19952022