# UK Housing Prices Paid Analysis

Authors: Jeremy Contreras, Kanchon Bishnu, Yu Wang
Department of Information Systems, California State University Los Angeles
CIS 5200 System Analysis and Design
jcontr185@calstatela.edu, ywang289@calstatela.edu, kbishnu@calstatela.edu

Data source: https://www.kaggle.com/datasets/hm-land-registry/uk-housing-prices-paid

**Abstract:** The dataset being used is the UK Housing Prices Paid dataset, available at https://www.kaggle.com/datasets/hm-land-registry/uk-housing-prices-paid. The size of the dataset is 2.41 GB, and it includes both time and location columns. This project will conduct a Tempo-Spatial Analysis of housing prices in the United Kingdom, examining how prices have evolved over time and how they vary across different regions. Specifically, the analysis will focus on identifying trends in housing prices across different years and geographic locations, such as postal codes, counties, and cities. This analysis is important because it will provide valuable insights into the housing market in the UK, identifying areas of growth, potential investment opportunities, and regional disparities. Moreover, understanding these trends can help inform policy decisions related to housing affordability, urban development, and economic growth.

## 1. Introduction

Our group selected to analyze the UK Housing Prices Paid dataset because it can provide insight into housing prices in the United Kingdom. The UK Housing Prices Paid dataset includes transactions of property sales in the UK. We will use Hive for analysis because it can efficiently handle analyzing large datasets like the UK Housing Prices Paid dataset. For visualization of the data we will use Excel. As a result of our project we will perform big data analysis using Hive gaining hands-on practice analyzing a large dataset.

## 2. Related Work

A similar paper I found that is like our group project is Visualizing and Recognizing Important Trends In Housing Sales Data Using Hadoop. The paper is about analyzing housing data prices in King County Washington. In the study, Hive was used to query the data. The paper also includes many charts to give insights into the housing market. The difference is in the paper they are using Hortonworks Data Platform. Our group project uses Hive on Cloud Computing. Another difference is in the paper the tables are DataFrames in Apache Spark. Python programming language was used to analyze DataFrames (Kim, 2020). The next paper I found that is similar is Harnessing Big Data to Revolutionize Real Estate Financing for Low Income Earners in the U.S. The analytical framework they used is Hadoop. For visualization they used the python visualization library matplotlib. Our group project queries real estate transaction prices using Hive in cloud computing. The paper applies big data analysis to address the challenges of real estate financing (Ademowo et al., 2024). Another paper that is like our project is High-Performance Geospatial Big Data Processing System Based on MapReduce. This relates to our group project because we create a geospatial visualization of a 3D map. In the paper, they developed a framework called Marmot based on Hadoop to help process geospatial data (Jo & Lee, 2018).

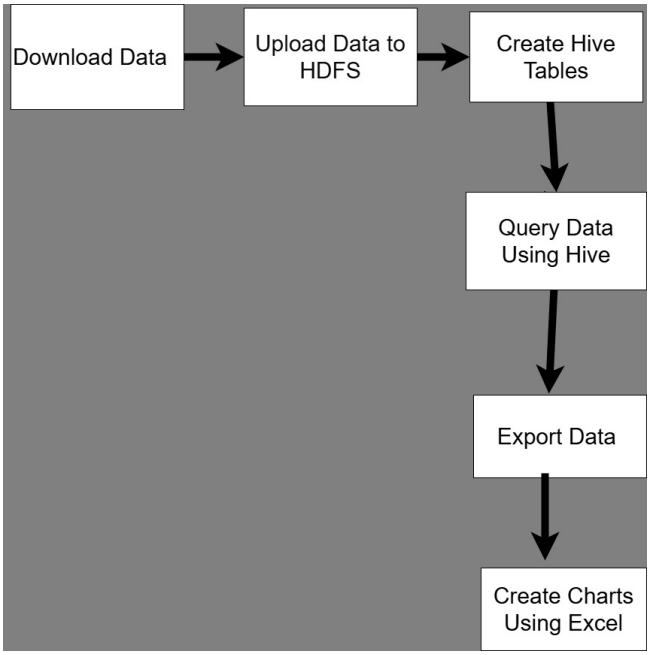
## 3. Specifications

| | |
|---|---|
| Dataset: UK Housing Prices Paid Dataset | Size (Total 2.41 GB ) |

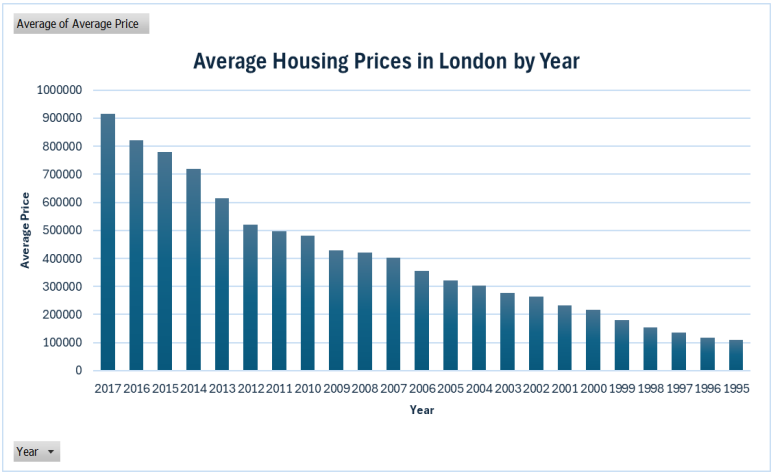
| | |
|---|---|
| Cluster Version | Hadoop 3.1.2 |
| Cluster Number of Nodes | 5 nodes: 2 master, 3 work |
| Memory Size | 155 GB/ 31 GB Per node |
| CPU Speed | 2.45 GHz |

**Implementation Flowchart:**

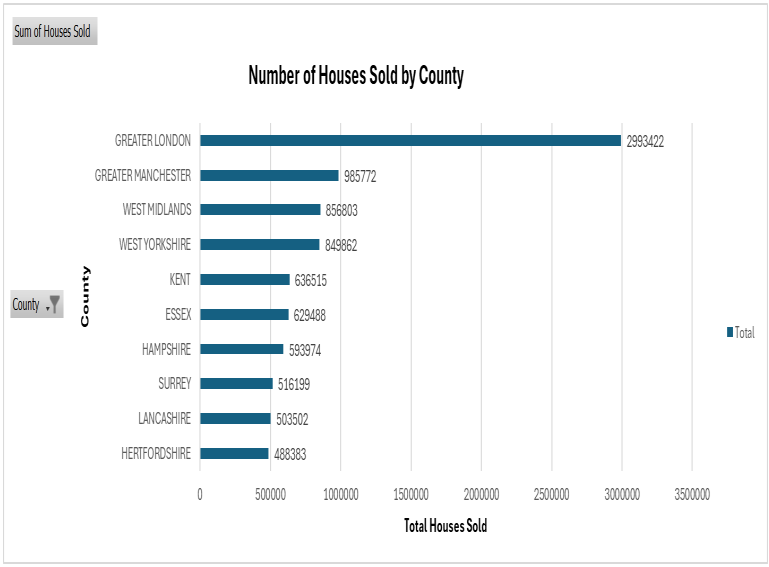
Our group downloaded the data using the Kaggle API because the data was located on Kaggle. After downloading the data, we uploaded the data to Hadoop Distributed File System (HDFS). We then created external Hive tables. We performed HiveQL queries on the data to analyze price transactions. We Exported the data to our local file system. Then we created charts and a 3D map for data visualization using Excel.



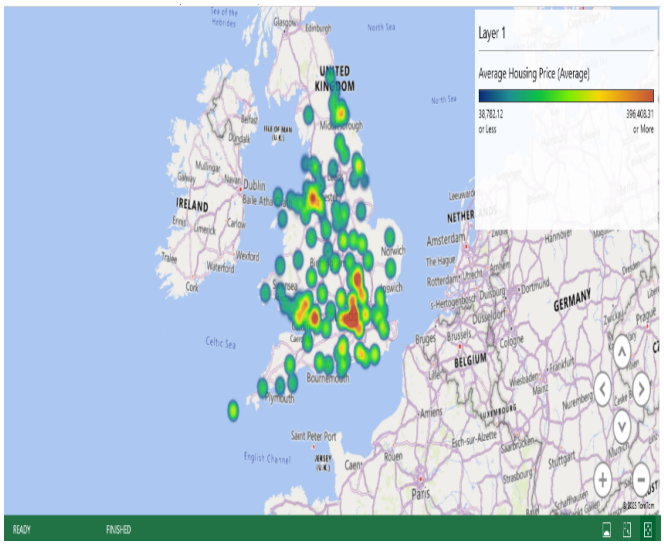

**Analysis and Visualization:**




The column chart displays the average housing prices in London by year. There is an increasing trend in average housing prices in London from the years: 1995 to 2016. In the year 1995 prices were under £100,000 and by year 2016 prices exceeded £900,000. Over two decades prices have increased significantly. Another insight we gained is prices remained stable between the years: 2007 to 2009.

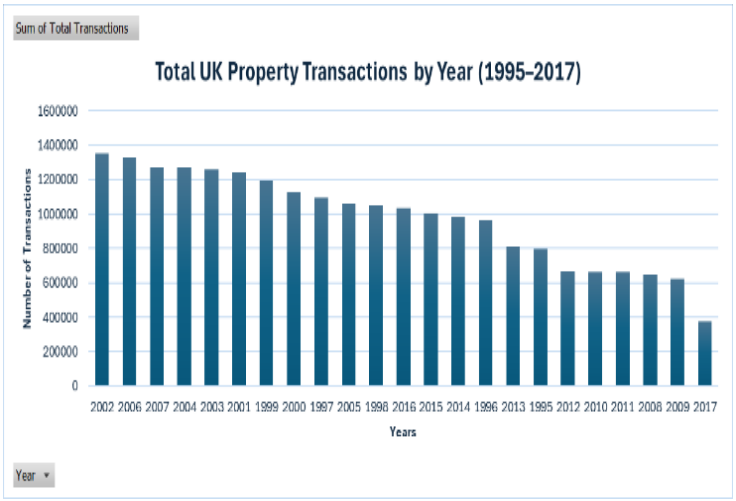



The horizontal bar chart displays the number of Houses Sold By County in the United Kingdom. The county with the most property transactions is Greater London. There were nearly three million property transactions in Greater London. The next county with the second highest property transactions is Greater Manchester with just under one million property transactions. The other counties within the top five most property transactions included West Midlands, West Yorkshire, and Kent.

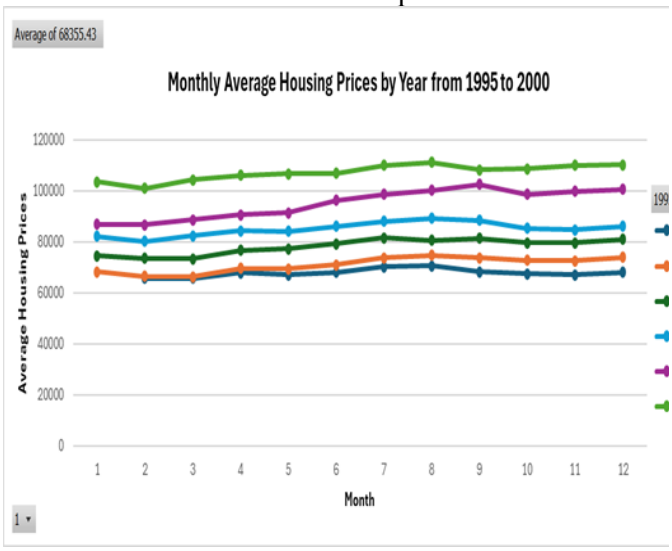



The 3D map heatmap displays the average housing prices across the United Kingdom. Brighter colors

represent such as yellow, orange or red represent higher average prices. Green and blue colors represent lower average prices. Based on the 3D map Greater London and its surrounding Southeast region have higher average prices. Another insight is the Southwest and Midlands regions of the UK show moderate housing prices based on the color zones of green and yellow.

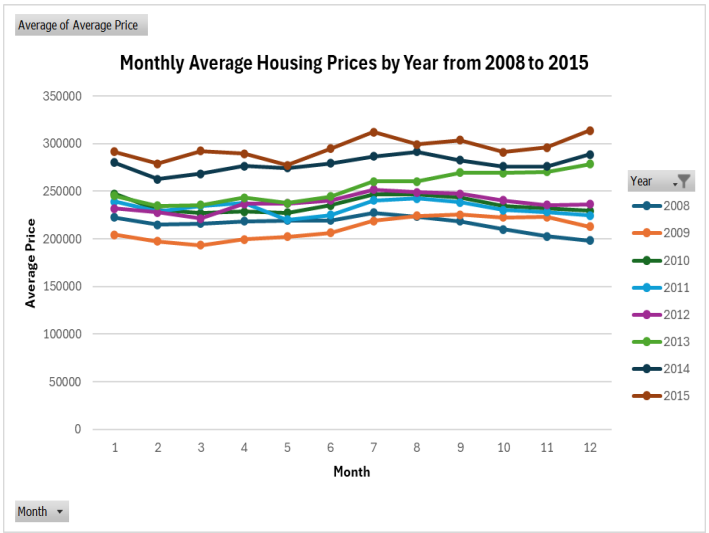



The column chart displays the total UK property transactions by year from 1995 to 2017. The year with the most property transactions in the UK is 2002 at 1.3 million. The years with the least property transactions occurred between 2009 and 2012. The year 2017 might be an outlier because of an incomplete dataset.




The line chart displays monthly average housing prices by year from 1995 to 2000. There is a year over year increase in average housing prices from 1995 to 2000. For each year prices peak around July to August. The years with most housing price growth occurred between 1999 to 2000.


The line chart displays the monthly average housing prices by year from 2001 to 2007. There is also consistent year-over-year growth. The year 2007 had the highest average housing price for this period. There is also price growth between the years 2004 to 2007.




The line chart displays the monthly average housing prices by year from 2008 to 2015. There is a drop in housing prices in the year 2008. From the years 2013 to 2015 there is a steady increase in average housing prices. The years with the most price growth occurred in 2014 to 2015 particularly around July to December for months.

## Conclusion

Our group performed a tempo-spatial analysis of UK Housing Prices Paid Dataset. We used Hive to query a massive dataset over 2GB in size. In Hive we used date functions to analyze data. We used group by queries in Hive for analysis. For visualizations we used Excel. The visualizations we created include bar charts, line charts, and a 3D heat map to analyze housing prices based on time and geographic region in the UK. The visualizations helped us discover insights in the data such as steady rise in housing prices in the past two decades. Our work is important because it provides insights into the economic issue of housing affordability. The findings and insights can be used by policymakers or investors for data driven decision making processes about UK housing prices. From this project we gained hands on experience with big data processing tools such as Hive and HDFS. We also learned to gain insights into a massive dataset.

**References**

[1] Ademowo, Adedotun & Osigwe, Ifeanyi & Nkurumah, Kwame. (2024). Harnessing Big Data to Revolutionize Real Estate Financing for Low-Income Earners in the U.S. INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING AND APPLIED SCIENCES, 2024, 373.

[2] Jo, J., & Lee, K.-W. (2018). High-Performance Geospatial Big Data Processing System Based on MapReduce. ISPRS International Journal of Geo-Information, 7(10), 399. https://doi.org/10.3390/ijgi7100399

[3] Kim, M. (2020, March 23). Visualizing and Recognizing Important Trends in Housing Sales Data using Hadoop (Graduate project). California State University. https://scholarworks.calstate.edu/concern/projects/4q77ft47c

**Github link**: https://github.com/Yuwang-maker/CIS-5200-Housing-Price-Project-

**Data Source**: https://www.kaggle.com/datasets/hm-land-registry/uk-housing-prices-paid