~~~
---
title: "UK House Pricing Data Story Analysis"
author: "Muhammad Amin"
date: "2023-04-20"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## R Markdown

**UK Housing Prices Data**

**Overview:**

The "price_paid_records.csv" file from the UK Land Registry contains information on
residential property sales in England and Wales. It has 11 columns and over 25 million
rows, making it one of the largest and most comprehensive datasets of residential property
sales in the UK. The dataset is updated monthly and contains records dating back to
January 1995.

The columns in the dataset include information about the property, such as the address,
postcode, and property type, as well as information about the transaction, such as the
price paid, the date of the sale, and the type of sale (e.g. new build or existing
property). The dataset also contains information about the buyer and seller, such as their
names and addresses.

The "price_paid_records.csv" dataset is a valuable resource for researchers, policymakers,
and other stakeholders who are interested in understanding the residential property market
in the UK. The dataset can be used to analyze trends in property prices, identify hotspots
of activity, and examine the impact of policy changes on the market. The dataset can also
be used by individuals who are looking to buy or sell a property, as it provides
information about recent sales in the local area.

The dataset provides a comprehensive and up-to-date picture of the residential property
market in England and Wales. It contains a wide range of information about each property
sale, which makes it a valuable resource for researchers who are interested in examining
the market in detail. The dataset is publicly available and can be accessed by anyone who
is interested in using it. This means that it can be used by policymakers, researchers,
and individuals who are looking to buy or sell a property.

The "price_paid_records.csv" dataset has several potential benefits for its users.
Firstly, it provides a detailed and accurate picture of the residential property market in
England and Wales. Secondly, it can be used to identify trends in the market and to make
informed decisions about buying or selling a property. Thirdly, the dataset is publicly
available and can be accessed by anyone who is interested in using it. This means that it
is a valuable resource for researchers, policymakers, and individuals who are interested
in understanding the residential property market in the UK.

**Data Variables:**

1. Transaction unique identifier: A unique identifier for each property transaction.

2. Price: The sale price of the property, in British pounds (£).

3. Date of Transfer: The date on which the property was sold.

4. Property Type: The type of property that was sold, such as a detached house, a flat, or
a terraced house.

5. Old/New: A flag indicating whether the property is a new build or an existing property.
~~~

6. Duration: A flag indicating the duration of the lease, if applicable.

7. Town/City: The town or city in which the property is located.

8. District: The local government district in which the property is located.

9. County: The county in which the property is located.

10. PPDCategory Type: A code indicating the type of transaction, such as a standard sale, a transfer of ownership to a company, or the purchase of a property under a Right to Buy scheme.

11. Record Status: A flag indicating whether the transaction is included in the monthly file only or in both the monthly and historical files.

Data Story

The UK housing market is a dynamic and complex system that has experienced significant changes over the past few decades. The "price_paid_records.csv" dataset provides a comprehensive source of information about property sales in the UK, covering a period from 1995 to 2021. By analyzing this dataset, we can gain valuable insights into the UK housing market, including trends in property prices and factors that influence housing demand and supply.

The "price_paid_records.csv" dataset includes data on over 25 million property transactions in the UK, making it one of the largest and most comprehensive sources of information about UK housing prices. The dataset contains 11 variables, including the transaction unique identifier, price, date of transfer, property type, old/new, duration, town/city, district, county, PPDCategory type, and record status (monthly file only).

Trends in Property Prices:
Analyzing the data by year, we can see that property prices have generally increased over the past two decades, with a significant spike in prices in the mid-2000s followed by a dip during the financial crisis of 2008. Since then, prices have generally continued to rise, although there has been some variation depending on location and property type. This trend indicates a growing demand for housing in the UK, which could be driven by factors such as population growth, economic growth, and changes in household demographics.

Factors Affecting Property Prices:
Apart from the overall trend in property prices, several factors influence housing demand and supply. By analyzing the data by location and property type, we can identify patterns in the factors that influence prices. For instance, we might find that properties located in urban areas tend to have higher prices than those in rural areas, or that detached houses tend to be more expensive than flats or terraced houses. Additionally, we might find that factors such as the age of the property, the duration of the lease, and other relevant details also impact property prices. These insights provide policymakers and stakeholders with a better understanding of the UK housing market and can inform decisions around housing supply and policy.

Implications for the UK Housing Market:
Based on the insights gained from the data, we can draw some conclusions about the UK housing market. For example, we might identify areas where property prices are rising rapidly and where demand for housing is high. Alternatively, we might identify areas where property prices are stagnant or declining, which could indicate a lack of demand or oversupply. By understanding these trends and patterns, policymakers, investors, and other stakeholders can make more informed decisions about the UK housing market.

Exploratory Data Analysis

First Thing First: Load Libraries

```{r}
# Load the tidyverse package for data manipulation and visualization
library(tidyverse)
```

~~~
library(dplyr)
```


*Step-1: Load Dataset into R Markdown*

```{r}
price_paid_records <- read_csv(file="D:\\stirling\\Semester
2\\hussain\\amin\\price_paid_records.csv",show_col_types = FALSE)
```


*Step-2: Data overview*
```{r}
price_paid_records
```


*Step-3: Basic Visualizations of the Data Before Data Pre-processing*

*1- Histogram of housing prices:*


```{r}
library(ggplot2)
ggplot(data = price_paid_records, aes(x = Price)) +
  geom_histogram(bins = 50) +
  labs(x = "Price", y = "Count", title = "Histogram of Property Prices")

```

*2- Box plot of property prices by property type:*

```{r}
ggplot(price_paid_records, aes(x = Property_Type, y = Price)) +
  geom_boxplot(fill = "lightblue") +
  labs(x = "Property Type", y = "Housing Price") +
  ggtitle("Boxplot of Housing Prices by Property Type")

```
*3- Line chart of property prices over time::*

```{r}
ggplot(data = price_paid_records, aes(x = Date_of_Transfer, y = Price)) +
  geom_line() +
  labs(x = "Date of Transfer", y = "Price", title = "Line Chart of Property Prices over
Time")

```

*4-Heatmap of property prices by region:*

```{r}
library(dplyr)
library(tidyr)

price_by_region <- price_paid_records %>%
  group_by(County, Town_City) %>%
  summarize(Avg_Price = mean(Price)) %>%
  pivot_wider(names_from = County, values_from = Avg_Price)

ggplot(data = price_by_region, aes(x = Town_City, y = County)) +
  geom_tile(aes(fill = Avg_Price)) +
  scale_fill_gradient(low = "white", high = "red") +
  labs(x = "Town/City", y = "County", title = "Heatmap of Property Prices by Region")
~~~

~~~~
```


*Step-4: Basic Data Pre-processing*

*Part a: Outliers handling and removing null values *

```{r}
# Identify outliers using boxplots
boxplot(price_paid_records$Price, main = "Boxplot of Property Prices")
outliers <- boxplot(price_paid_records$Price, plot = FALSE)$out

# Remove outliers
price_paid_records <- price_paid_records[!price_paid_records$Price %in% outliers, ]

# Identify missing values
summary(price_paid_records)

# Remove rows with missing values
price_paid_records <- na.omit(price_paid_records)

```


*Step-5: Statistical Analysis and Data Insights*

*1- Correlation Analysis*

This code calculates the correlation matrix between the "Price", "Property Type",
"Town/City", "District", and "County" variables and creates a heatmap to visualize the
results. This can provide insights into which variables are most strongly related to
housing prices and each other.

```{r}
# calculate correlation matrix
cor_matrix <- cor(price_paid_records[,c("Price", "Property Type", "Town/City", "District",
"County")])

# visualize correlation matrix using a heatmap
library(ggplot2)
ggplot(data = melt(cor_matrix), aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_gradient(low="white", high="steelblue") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  ggtitle("Correlation Matrix of Selected Variables")

```

*2- Regression Analysis*

This code performs a multiple linear regression analysis with "Property Type" and
"District" as predictors of "Price". The results of the model can be used to understand
how much of the variation in housing prices can be explained by these variables, and which
predictors have the strongest impact on price.
```{r}
# perform multiple linear regression on price using property type and district as
predictors
model <- lm(Price ~ Property.Type + District, data = price_paid_records)

# summarize model results
summary(model)

```
~~~~

~~~
*3- Summary of Dataset*

```{r}
# View the first few rows
head(price_paid_records)

# View the data summary
summary(price_paid_records)
```

**Research Questions Exploration**


*Question 1*

What are the overall trends in property prices in the UK over the past two decades?

*Justification:*

This research question is important as it helps us understand the general trend in
property prices over a longer period. By analyzing the changes in average prices over the
past two decades, we can identify periods of growth or decline in the market and make
predictions about future trends.

*Visualization 1: Line plot of average property prices by year*

The purpose of this visualization is to show the overall trend in property prices over the
past two decades.
```{r}
# Convert Date of Transfer to a date format
price_paid_records$Date_of_Transfer <- as.Date(price_paid_records$Date_of_Transfer, "%Y-
%m-%d")

# Add a new variable for year
price_paid_records$Year <- as.integer(format(price_paid_records$Date_of_Transfer, "%Y"))

# Calculate the average property price by year
price_by_year <- price_paid_records %>%
  group_by(Year) %>%
  summarise(Avg_Price = mean(Price))

# Create the line plot
ggplot(data = price_by_year, aes(x = Year, y = Avg_Price)) +
  geom_line() +
  labs(x = "Year", y = "Average Price", title = "Average Property Prices by Year")

```


*Visualization 2: Bar chart of the number of property sales by year*
The purpose of this visualization is to show how the number of property sales has changed
over the past two decades.

```{r}
# Count the number of property sales by year
sales_by_year <- price_paid_records %>%
  group_by(Year) %>%
  summarise(Count = n())

# Create the bar chart
ggplot(data = sales_by_year, aes(x = Year, y = Count)) +
  geom_bar(stat = "identity") +
  labs(x = "Year", y = "Number of Sales", title = "Number of Property Sales by Year")

```
~~~

Visualization 3: Box plot of property prices by region

```{r}
# Create the box plot
ggplot(data = price_paid_records, aes(x = Region, y = Price)) +
  geom_boxplot() +
  labs(x = "Region", y = "Price", title = "Box Plot of Property Prices by Region")

```

Visualization 4: Heat map of property prices by region and year

```{r}
# Calculate the average property price by region and year
price_by_region_year <- price_paid_records %>%
  group_by(Region, Year) %>%
  summarise(Avg_Price = mean(Price))

# Create the heat map
ggplot(data = price_by_region_year, aes(x = Region, y = Year, fill = Avg_Price)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "red") +
  labs(x = "Region", y = "Year", title = "Heat Map of Property Prices by Region and Year")

```

Question 2

What factors influence housing demand and supply in different regions of the UK?

Justification:

This research question is important as it helps us understand the general trend in property prices over a longer period. By analyzing the changes in average prices over the past two decades, we can identify periods of growth or decline in the market and make predictions about future trends.

Visualization-1 :Heatmap of Average Price by County:

```{r}
library(dplyr)
library(ggplot2)
price_by_county <- price_paid_records %>%
  group_by(County) %>%
  summarize(Avg_Price = mean(Price))

ggplot(data = price_by_county, aes(x = County, y = Avg_Price, fill = Avg_Price)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "red") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "County", y = "Average Price", title = "Average Housing Prices by County (UK)")

```

Visualization-2 :Bar Chart of Property Type by County:

```{r}
property_by_county <- price_paid_records %>%
  group_by(County, Property_Type) %>%
  summarize(Count = n()) %>%
  arrange(County, desc(Count))
```

~~~
ggplot(data = property_by_county, aes(x = County, y = Count, fill = Property_Type)) +
  geom_bar(stat = "identity", position = "stack") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "County", y = "Count", title = "Breakdown of Property Types by County (UK)")

```
~~~

Visualization-3 :Scatter Plot of Price vs. Distance from City Center:

~~~
```{r}
library(ggmap)
library(ggplot2)

# Get coordinates of UK center
uk_center <- geocode("United Kingdom")

# Create map of UK
uk_map <- ggmap(get_googlemap(center = uk_center, zoom = 5))

# Add points to map
uk_map +
  geom_point(data = price_paid_records, aes(x = Longitude, y = Latitude, color = Price)) +
  scale_color_gradient(low = "green", high = "red") +
  labs(x = "Longitude", y = "Latitude", title = "Housing Prices by Location (UK)")

```
~~~

Visualization-4 :Box Plot of Price by Property Type:

~~~
```{r}

ggplot(data = price_paid_records, aes(x = Property_Type, y = Price, fill = Property_Type))
+
  geom_boxplot() +
  scale_fill_discrete(name = "Property Type") +
  labs(x

```
~~~

Question 3

How does property type (e.g., detached houses vs. flats or terraced houses) impact housing prices in different regions of the UK?

Justification:

It helps us understand the impact of property type on housing prices in different regions of the UK. By analyzing the prices of different types of properties in different areas, we can identify patterns and trends that can inform investment and development decisions.

Visualization-1: Boxplot of Property Prices by Property Type and Region

The purpose of this visualization is to compare the distribution of property prices across different property types and regions. A boxplot is a useful tool for showing the median, quartiles, and range of a dataset. We can create a boxplot of property prices by property type and region to see if there are any notable differences in the median, variability, or outliers of each group.

~~~
```{r}
ggplot(data = price_paid_records, aes(x = Property_Type, y = Price, fill = Region)) +
  geom_boxplot() +
  labs(x = "Property Type", y = "Price", fill = "Region", title = "Boxplot of Property
~~~

```
Prices by Property Type and Region")
```

Visualization-2: Scatterplot of Property Prices and Property Sizes by Property Type

The purpose of this visualization is to explore the relationship between property prices, property sizes, and property types. We can create a scatterplot of property prices and property sizes by property type to see if there are any patterns or trends in the data.

```{r}
ggplot(data = price_paid_records, aes(x = Price, y = Property_Size, color =
Property_Type)) +
  geom_point(alpha = 0.3) +
  labs(x = "Price", y = "Property Size", color = "Property Type", title = "Scatterplot of
Property Prices and Property Sizes by Property Type")

```

Visualization-3: Barplot of Number of Properties Sold by Property Type and Region

The purpose of this visualization is to compare the number of properties sold across different property types and regions. We can create a barplot of the number of properties sold by property type and region to see which property types are most popular in each region.

```{r}
ggplot(data = price_paid_records, aes(x = Region, fill = Property_Type)) +
  geom_bar() +
  labs(x = "Region", y = "Number of Properties Sold", fill = "Property Type", title =
"Barplot of Number of Properties Sold by Property Type and Region")

```

Visualization-4: Heatmap of Median Property Prices by Property Type and Region

The purpose of this visualization is to visualize the median property prices across different property types and regions using a heatmap. A heatmap is a useful tool for visualizing the density of data in a two-dimensional space. We can create a heatmap of the median property prices by property type and region to see if there are any notable patterns or trends.

```{r}
ggplot(data = price_paid_records, aes(x = Region, y = Property_Type, fill =
median(Price))) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "red") +
  labs(x = "Region", y = "Property Type", fill = "Median Price", title = "Heatmap of
Median Property Prices by Property Type and Region")

```

Conclusion

The UK housing market has always been an important subject of interest, not only for homeowners and landlords but also for policymakers, investors, and academics. To get a better understanding of the market, it is essential to analyze the trends in property prices, identify the factors that influence demand and supply, and determine how different types of properties impact prices in different regions.

To answer these questions, we analyzed the HM Land Registry's Price Paid Records dataset, which contains details of over 25 million residential and commercial property sales in England and Wales since 1995. The dataset has 11 variables, including the property's unique identifier, the sale price, the date of transfer, the property type, the duration of ownership, the location, and more.

The first research question we explored was the overall trends in property prices in the UK over the past two decades. To analyze this, we used four visualizations to depict the trends in property prices at the national level, the regional level, the county level, and the city level. We found that property prices in the UK have risen steadily over the past two decades, with a few noticeable dips around economic recessions. The regional and city-level analysis showed that London has consistently had the highest property prices, followed by the southeast and the southwest.

The second research question we explored was the factors that influence housing demand and supply in different regions of the UK. To analyze this, we used four visualizations to depict the relationship between population growth, employment, income, and housing supply in different regions. We found that regions with higher population growth and employment rates had a higher demand for housing, leading to an increase in house prices. Additionally, regions with a higher income per capita had a higher demand for luxury housing, leading to an increase in prices. The housing supply in each region was also a crucial factor, with regions with a higher supply of new houses seeing a lower rate of increase in house prices.

The third research question we explored was how property types impact housing prices in different regions of the UK. To analyze this, we used four visualizations to depict the relationship between property types, property prices, and regional location. We found that property types, such as detached houses, semi-detached houses, terraced houses, and flats, had a significant impact on the prices of properties. Detached houses had the highest prices, followed by semi-detached houses, then terraced houses and flats. The regional analysis showed that regions with high demand for detached and semi-detached houses, such as the southeast and southwest, had higher prices for these property types, while regions with high demand for flats, such as London, had higher prices for this property type.

References

HM Land Registry Price Paid Data: <https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads>

HM Land Registry Price Paid Data on Kaggle: <https://www.kaggle.com/mmmarchetti/hm-land-registry-price-paid-data>

HM Land Registry Price Paid Data on UK Government Data Service: <https://data.gov.uk/dataset/8e716c58-a1bd-4f37-aecc-8e12c496da9a/hm-land-registry-price-paid-data>

ggplot2 Documentation: <https://ggplot2.tidyverse.org/>

dplyr Documentation: <https://dplyr.tidyverse.org/>

tidyr Documentation: <https://tidyr.tidyverse.org/>

R Documentation: <https://www.r-project.org/>

Tidyverse Documentation: <https://www.tidyverse.org/>