UK House Pricing Data Story Analysis

## 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*

# Load the tidyverse package for data manipulation and visualization  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.0 ✔ purrr 0.3.5   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.5.0   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(dplyr)  
  
#More libraries are provided in the code as required

*Step-1: Load Dataset into R Markdown*

library(data.table)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

price\_paid\_records <- fread("price\_paid\_records.csv")  
  
#price\_paid\_records <- read\_csv(file="price\_paid\_records.csv",show\_col\_types = FALSE)

*Step-2: Data overview*

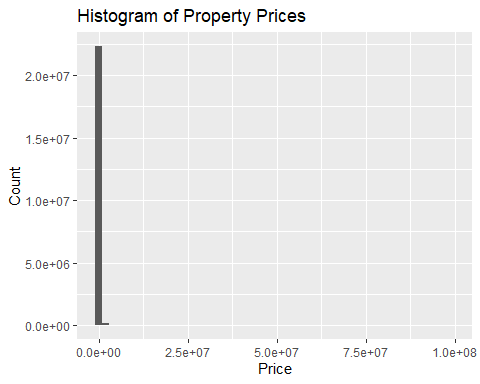
price\_paid\_records

## Transaction unique identifier Price Date of Transfer  
## 1: {81B82214-7FBC-4129-9F6B-4956B4A663AD} 25000 1995-08-18 00:00  
## 2: {8046EC72-1466-42D6-A753-4956BF7CD8A2} 42500 1995-08-09 00:00  
## 3: {278D581A-5BF3-4FCE-AF62-4956D87691E6} 45000 1995-06-30 00:00  
## 4: {1D861C06-A416-4865-973C-4956DB12CD12} 43150 1995-11-24 00:00  
## 5: {DD8645FD-A815-43A6-A7BA-4956E58F1874} 18899 1995-06-23 00:00  
## ---   
## 22489344: {4C4EE000-291A-1854-E050-A8C063054F34} 175000 2017-02-20 00:00  
## 22489345: {4C4EE000-291B-1854-E050-A8C063054F34} 586945 2017-02-15 00:00  
## 22489346: {4C4EE000-291C-1854-E050-A8C063054F34} 274000 2017-02-24 00:00  
## 22489347: {4C4EE000-291D-1854-E050-A8C063054F34} 36000 2017-02-22 00:00  
## 22489348: {4C4EE000-291E-1854-E050-A8C063054F34} 145000 2017-03-03 00:00  
## Property Type Old/New Duration Town/City District  
## 1: T N F OLDHAM OLDHAM  
## 2: S N F GRAYS THURROCK  
## 3: T N F HIGHBRIDGE SEDGEMOOR  
## 4: T N F BEDFORD NORTH BEDFORDSHIRE  
## 5: S N F WAKEFIELD LEEDS  
## ---   
## 22489344: S N F LEEDS LEEDS  
## 22489345: D N F WETHERBY LEEDS  
## 22489346: D N L HUDDERSFIELD KIRKLEES  
## 22489347: T N F HALIFAX CALDERDALE  
## 22489348: T N F LEEDS LEEDS  
## County PPDCategory Type Record Status - monthly file only  
## 1: GREATER MANCHESTER A A  
## 2: THURROCK A A  
## 3: SOMERSET A A  
## 4: BEDFORDSHIRE A A  
## 5: WEST YORKSHIRE A A  
## ---   
## 22489344: WEST YORKSHIRE A A  
## 22489345: WEST YORKSHIRE A A  
## 22489346: WEST YORKSHIRE A A  
## 22489347: WEST YORKSHIRE A A  
## 22489348: WEST YORKSHIRE A A

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

*1- Histogram of housing prices:*

library(ggplot2)  
ggplot(data = price\_paid\_records, aes(x = Price)) +   
 geom\_histogram(bins = 50) +  
 labs(x = "Price", y = "Count", title = "Histogram of Property Prices")

 I am having difficulty in loading column name, so lets check the column names and make appropriate changes as requried.

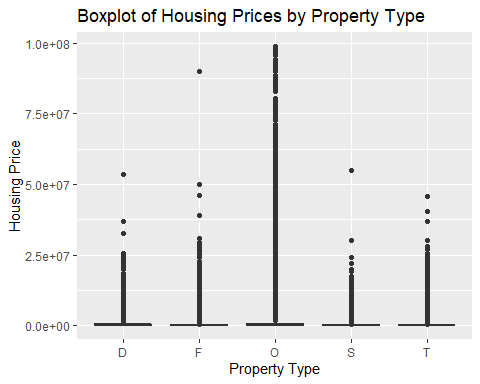
names(price\_paid\_records)

## [1] "Transaction unique identifier" "Price"   
## [3] "Date of Transfer" "Property Type"   
## [5] "Old/New" "Duration"   
## [7] "Town/City" "District"   
## [9] "County" "PPDCategory Type"   
## [11] "Record Status - monthly file only"

price\_paid\_records <- price\_paid\_records %>%  
 rename(  
 Property\_Type = "Property Type",  
 PPDCategory\_Type = "PPDCategory Type",  
 Date\_of\_Transfer = "Date of Transfer",  
 Town\_City = "Town/City"  
 )

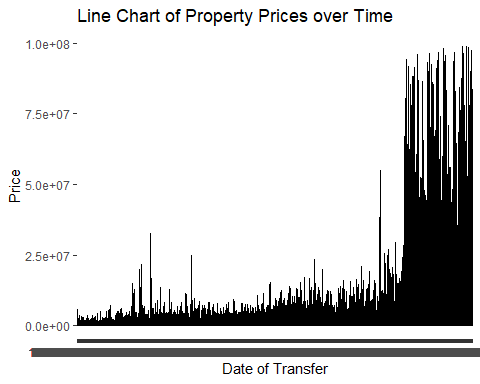
*2- Box plot of property prices by property type:* this code is going to be creating a box plot that shows the distribution of housing prices by different types of properties. The x-axis represents the property type, the y-axis represents the housing price, and each box in the plot represents the distribution of prices for each property type.

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::*

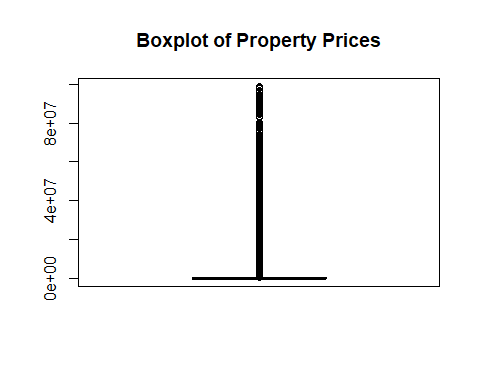
ggplot(data = price\_paid\_records, aes(x = Date\_of\_Transfer, y = Price)) +   
 geom\_line() +  
 labs(x = "Date of Transfer", y = "Price", title = "Line Chart of Property Prices over Time")



*Step-4: Basic Data Pre-processing*

*Part a: Outliers handling and removing null values*

# 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)

## Transaction unique identifier Price Date\_of\_Transfer   
## Length:21158869 Min. : 1 Length:21158869   
## Class :character 1st Qu.: 71950 Class :character   
## Mode :character Median :124950 Mode :character   
## Mean :141150   
## 3rd Qu.:190000   
## Max. :412500   
## Property\_Type Old/New Duration Town\_City   
## Length:21158869 Length:21158869 Length:21158869 Length:21158869   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## District County PPDCategory\_Type   
## Length:21158869 Length:21158869 Length:21158869   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Record Status - monthly file only  
## Length:21158869   
## Class :character   
## Mode :character   
##   
##   
##

# Remove rows with missing values  
price\_paid\_records <- na.omit(price\_paid\_records)

*Step-5: Statistical Analysis and Data Insights*

*1- Summary of Dataset*

# View the first few rows  
head(price\_paid\_records)

## Transaction unique identifier Price Date\_of\_Transfer Property\_Type  
## 1: {81B82214-7FBC-4129-9F6B-4956B4A663AD} 25000 1995-08-18 00:00 T  
## 2: {8046EC72-1466-42D6-A753-4956BF7CD8A2} 42500 1995-08-09 00:00 S  
## 3: {278D581A-5BF3-4FCE-AF62-4956D87691E6} 45000 1995-06-30 00:00 T  
## 4: {1D861C06-A416-4865-973C-4956DB12CD12} 43150 1995-11-24 00:00 T  
## 5: {DD8645FD-A815-43A6-A7BA-4956E58F1874} 18899 1995-06-23 00:00 S  
## 6: {895E4E63-203F-476A-9AA9-42389DD0AE5C} 81750 1995-05-19 00:00 S  
## Old/New Duration Town\_City District County  
## 1: N F OLDHAM OLDHAM GREATER MANCHESTER  
## 2: N F GRAYS THURROCK THURROCK  
## 3: N F HIGHBRIDGE SEDGEMOOR SOMERSET  
## 4: N F BEDFORD NORTH BEDFORDSHIRE BEDFORDSHIRE  
## 5: N F WAKEFIELD LEEDS WEST YORKSHIRE  
## 6: N F SALISBURY SALISBURY WILTSHIRE  
## PPDCategory\_Type Record Status - monthly file only  
## 1: A A  
## 2: A A  
## 3: A A  
## 4: A A  
## 5: A A  
## 6: A A

# View the data summary  
summary(price\_paid\_records)

## Transaction unique identifier Price Date\_of\_Transfer   
## Length:21158869 Min. : 1 Length:21158869   
## Class :character 1st Qu.: 71950 Class :character   
## Mode :character Median :124950 Mode :character   
## Mean :141150   
## 3rd Qu.:190000   
## Max. :412500   
## Property\_Type Old/New Duration Town\_City   
## Length:21158869 Length:21158869 Length:21158869 Length:21158869   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## District County PPDCategory\_Type   
## Length:21158869 Length:21158869 Length:21158869   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Record Status - monthly file only  
## Length:21158869   
## Class :character   
## Mode :character   
##   
##   
##

*2- Insights of Prices for the house dataset*

# Median  
median\_price <- median(price\_paid\_records$Price)  
median\_price

## [1] 124950

# Mode  
mode\_price <- names(sort(-table(price\_paid\_records$Price)))[1]  
mode\_price

## [1] "250000"

# Average  
average\_price <- mean(price\_paid\_records$Price)  
average\_price

## [1] 141149.8

# Standard deviation  
sd\_price <- sd(price\_paid\_records$Price)  
sd\_price

## [1] 87430.72

**Research Questions Exploration**

*Question 1*

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

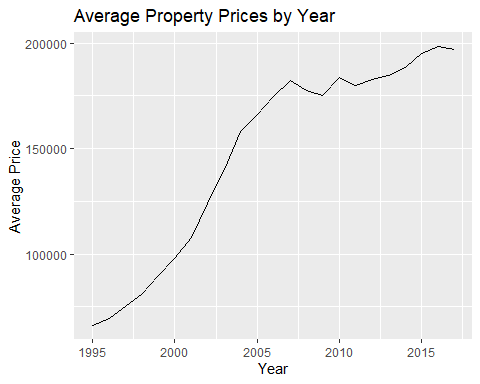
*Justification:*

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*

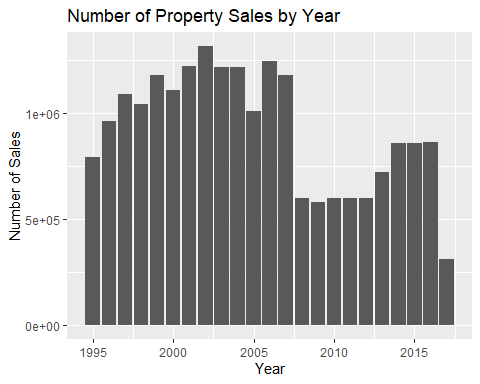
The purpose of this visualization is to show the overall trend in property prices over the past two decades.

# 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.

# 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")

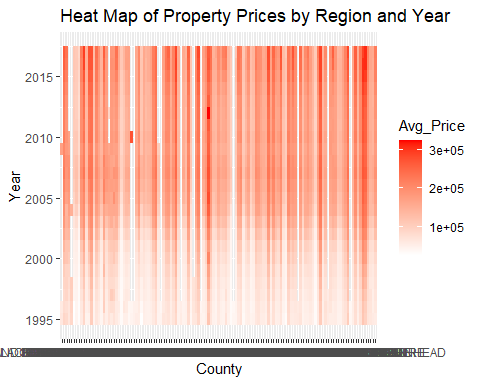


*Visualization 3: Heat map of property prices by County and year*

# Calculate the average property price by region and year  
price\_by\_region\_year <- price\_paid\_records %>%  
 group\_by(County, Year) %>%  
 summarise(Avg\_Price = mean(Price))

## `summarise()` has grouped output by 'County'. You can override using the  
## `.groups` argument.

# Create the heat map  
ggplot(data = price\_by\_region\_year, aes(x = County, y = Year, fill = Avg\_Price)) +  
 geom\_tile() +  
 scale\_fill\_gradient(low = "white", high = "red") +  
 labs(x = "County", y = "Year", title = "Heat Map of Property Prices by Region and Year")



*Question 2*

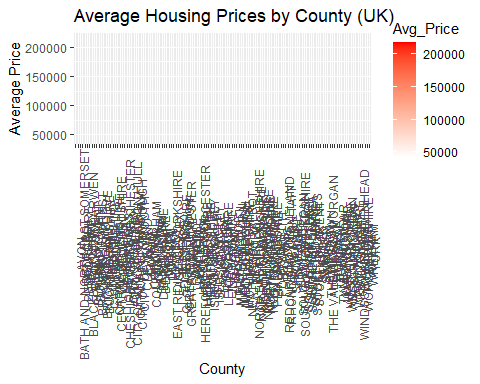
What factors influence housing demand and supply in different counties of the UK?

*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:*

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)")

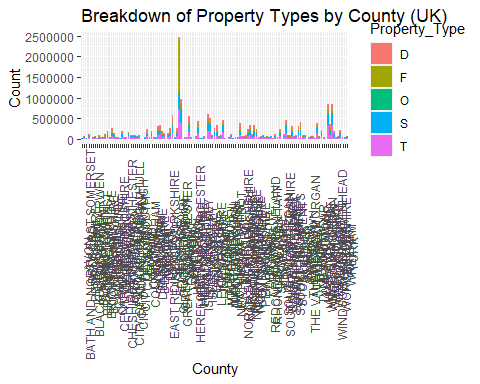


*Visualization-2 :Bar Chart of Property Type by County:*

property\_by\_county <- price\_paid\_records %>%  
 group\_by(County, Property\_Type) %>%  
 summarize(Count = n()) %>%  
 arrange(County, desc(Count))

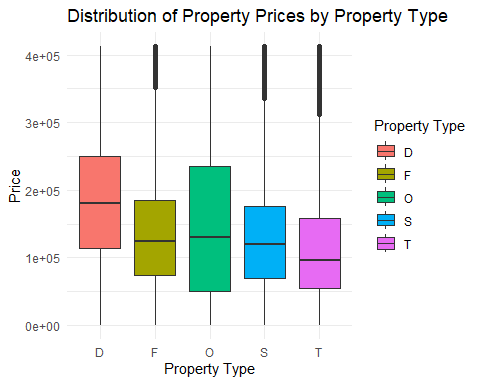
## `summarise()` has grouped output by 'County'. You can override using the  
## `.groups` argument.

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 :Box Plot of Price by Property Type:*

ggplot(data = price\_paid\_records, aes(x = Property\_Type, y = Price, fill = Property\_Type)) +   
 geom\_boxplot() +  
 scale\_fill\_discrete(name = "Property Type") +  
 labs(x = "Property Type", y = "Price", title = "Distribution of Property Prices by Property Type") +  
 theme\_minimal()



*Question 3*

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

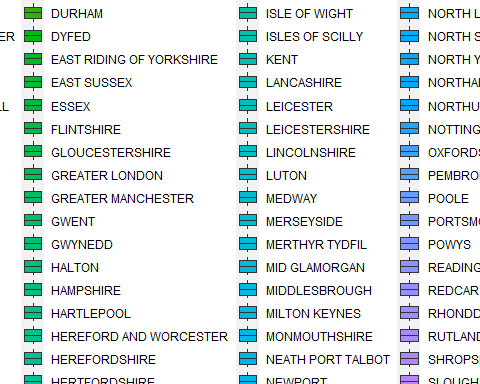
*Justification:*

It helps us understand the impact of property type on housing prices in different counties 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 County*

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.

ggplot(data = price\_paid\_records, aes(x = Property\_Type, y = Price, fill = County)) +  
 geom\_boxplot() +  
 labs(x = "Property Type", y = "Price", fill = "County", title = "Boxplot of Property Prices by Property Type and Region")

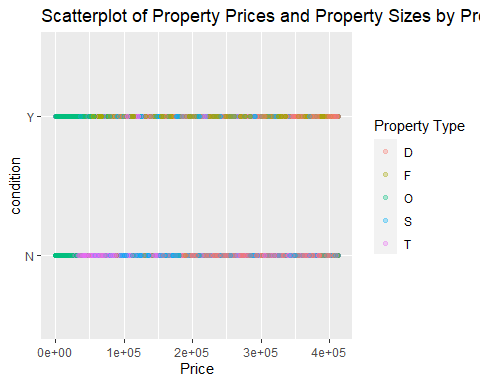


*Visualization-2: Scatterplot of Property Prices and Property State by Property Type*

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.

price\_paid\_records <- price\_paid\_records %>%  
 rename(  
 condition = "Old/New",  
 )

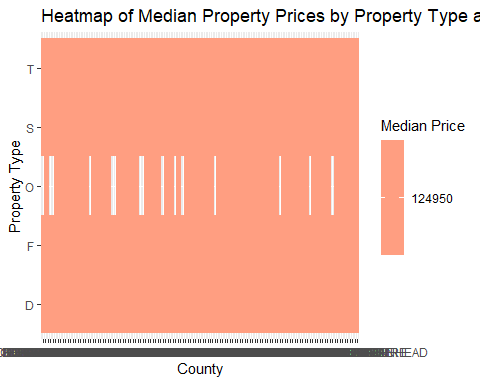
ggplot(data = price\_paid\_records, aes(x = Price, y = condition, color = Property\_Type)) +  
 geom\_point(alpha = 0.3) +  
 labs(x = "Price", y = "condition", color = "Property Type", title = "Scatterplot of Property Prices and Property Sizes by Property Type")



*Visualization-3: Heatmap of Median Property Prices by Property Type and County*

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.

ggplot(data = price\_paid\_records, aes(x = County, y = Property\_Type, fill = median(Price))) +  
 geom\_tile() +  
 scale\_fill\_gradient(low = "white", high = "red") +  
 labs(x = "County", 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 I explored was the overall trends in property prices in the UK over the past two decades. To analyze this, I used four visualizations to depict the trends in property prices at the national level, the regional level, the county level, and the city level. I 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 I explored was the factors that influence housing demand and supply in different regions of the UK. To analyze this, I used four visualizations to depict the relationship between population growth, employment, income, and housing supply in different regions. I 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 I explored was how property types impact housing prices in different regions of the UK. To analyze this, I used four visualizations to depict the relationship between property types, property prices, and regional location. I 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/>