LD7188 APPLIED DATA SCIENCE ASSESSMENT

by Lim Shin Huey

Applied Data Science

Professor Dilek Celik

Northumbria University

London

22nd January 2024

Contents

Tables

Figures

1 Statistical Data Analysis

1.1 Domain Understanding and Research Questions

The domain of research for the dataset provided includes information on house pricing in London for the year 2023, which belongs to the field of real estate economics and urban studies. The dataset includes information such as unique identifiers, price_paid, deed_date, postcode, property_type, new_build, estate_type, street, locality, town, district, and country. The research questions suggest an interest in understanding the variation in property prices across different London districts and identifying the key features influencing property prices.

According to the academic article by Santos and Jiang (2020), the research aims to investigate the relationship between various explanatory variables and house prices in Greater London, adopting both macroeconomic and microeconomic perspectives within the housing market domain.

The study initiates by outlining two broad frameworks for housing market research: macroeconomics and microeconomics. In the macroeconomic framework, the analysis emphasizes identifying key drivers, such as income, taxes, and interest rates, and their potential impact on pricing mechanisms. Additionally, environmental statistics, including labour, capital markets, and the construction business, are considered to predict market outcomes, providing a comprehensive understanding of the broader context influencing market behaviours.

On the microeconomic side, the study underscores the significance of modeling consumer preferences and decision mechanisms. This involves a detailed examination of personal preferences and constraints to enhance the understanding of consumer behavior and its contribution to overall housing market dynamics. The article proceeds to discuss the analysis and conclusions, highlighting a crucial step involving the examination of spatial data variations and the exploration of general characteristics. The mention of quantitative spatial data analysis indicates the study's utilization of exploratory data analysis strategies to analyze datasets, explore their characteristics, and generate hypotheses.

In evaluating the research area, the article suggests a comprehensive approach to understanding the relationship between explanatory variables and house prices in London. The incorporation of both macroeconomic and microeconomic perspectives and the emphasis on exploratory data analysis and quantitative spatial data analysis reflect a rigorous methodological approach. Further exploration of specific results and conclusions from this study could offer valuable insights into property price variation across different London districts and the key features influencing these prices, aligning with the research questions posed for the provided dataset.

Research questions and appropriate hypothesis (NULL and alternative)

1. Which London District offers properties with the highest, middle, and lowest prices?

H0: There is no significant correlation between house prices in different districts of London.

H1: There is a significant correlation between house prices in different districts of London.

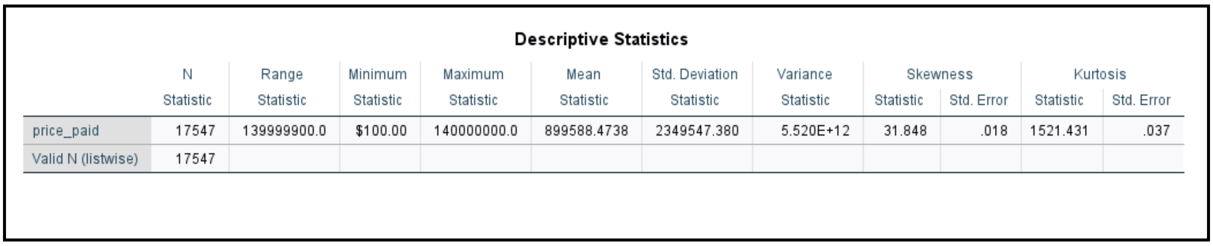
2. What are the most important features that determine the price of the property in this data?

H0: There is no significant linear relationship between the price of the property (price_paid) and any of the features included in the dataset, indicating that none of the features contribute significantly to predicting the property price.

H1: There is a significant linear relationship between the price of the property (price_paid) and at least one of the features included in the dataset. This suggests that one or more features are important in determining the property price.

1.2 Dataset and Data Preparation

Table 1: Price Paid Descriptive



Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
price_paid	17547	139999900.0	$100.00	140000000.0	899588.4738	2349547.380	5.520E+12	31.848	.018	1521.431	.037
Valid N (listwise)	17547										

This dataset contains information on London housing prices for 2023. The 'price_paid' column shows various prices, with a range of 139,999,900.0 indicating a big difference between the highest and lowest values. This suggests a lot of variability, possibly due to some unusual values.

The prices range from a minimum of £100 to a maximum of £140,000,000, showing transactions with both low and high values. The average is around £899,970.81, with a high standard deviation indicating significant changes in housing prices. The variance value highlights a large spread among the data points.

The distribution leans to the right, meaning most housing prices are lower, and there are fewer properties at higher prices. The long right tail suggests some very high-value real estate transactions. The high kurtosis value indicates more extreme values than a normal distribution, as numerically supported by Table 1 and visually depicted in Figure 1 (Kerr et al.).

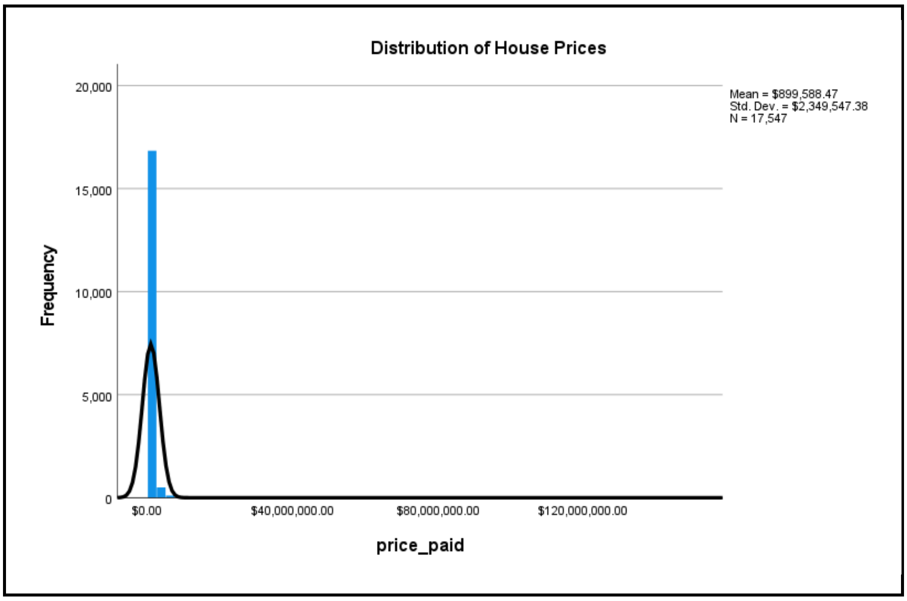
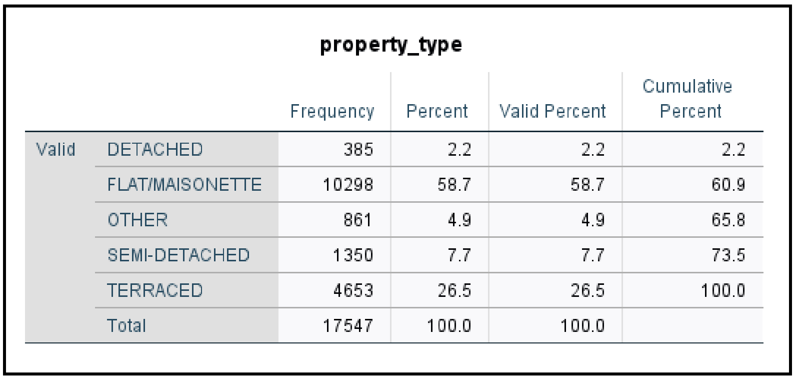


Figure 1: Distribution of House Prices

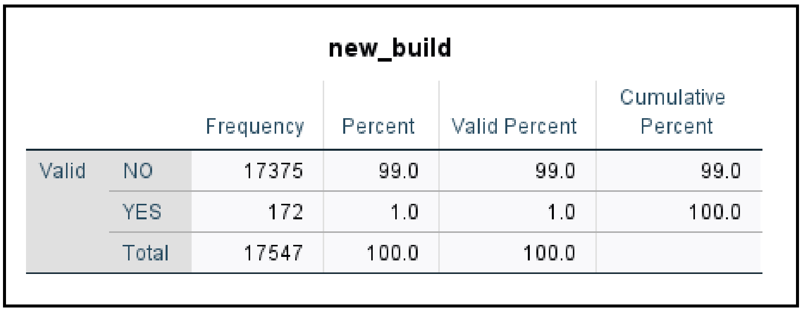
Table 2: Distribution of Property Types



property_type

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	DETACHED	385	2.2	2.2	2.2
	FLAT/MAISONETTE	10298	58.7	58.7	60.9
	OTHER	861	4.9	4.9	65.8
	SEMI-DETACHED	1350	7.7	7.7	73.5
	TERRACED	4653	26.5	26.5	100.0
	Total	17547	100.0	100.0	

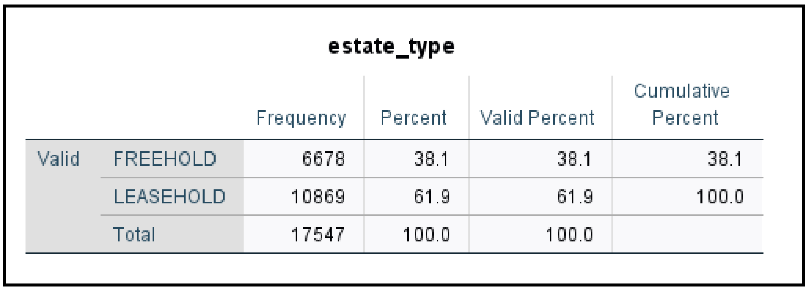
Table 3: Distribution of New Build



new_build

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NO	17375	99.0	99.0	99.0
	YES	172	1.0	1.0	100.0
	Total	17547	100.0	100.0	

Table 4: Distribution of Estate Type

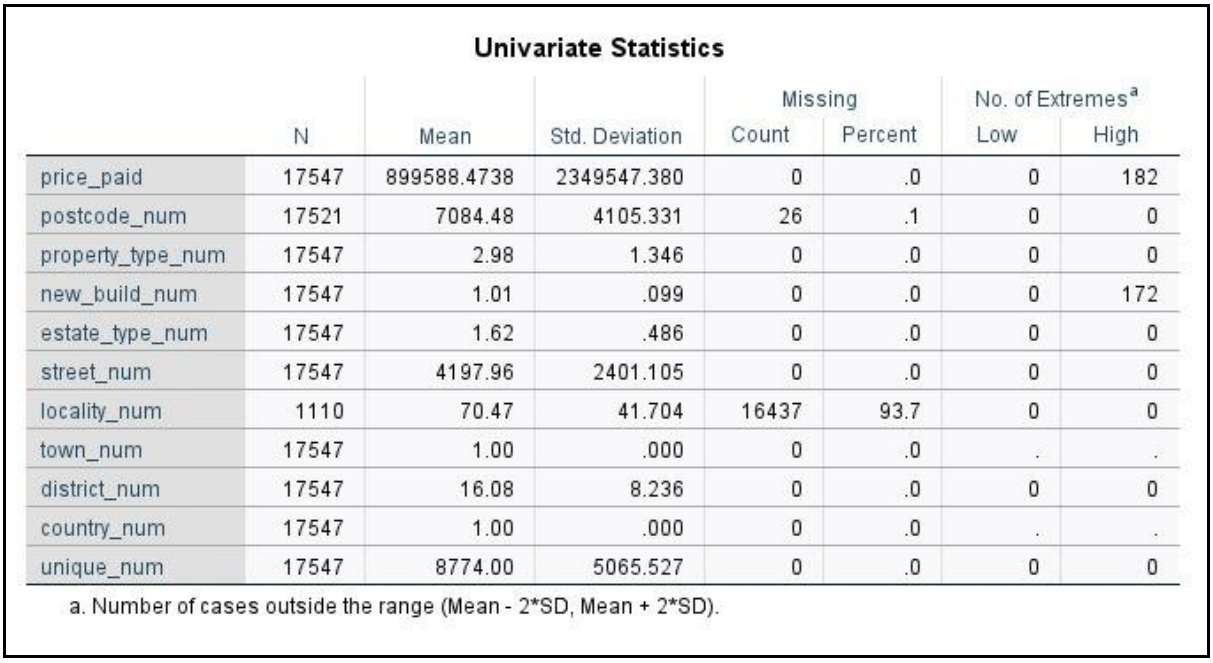


estate_type

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	FREEHOLD	6678	38.1	38.1	38.1
	LEASEHOLD	10869	61.9	61.9	100.0
	Total	17547	100.0	100.0	

The data indicates that the most prevalent property type is 'FLAT/MAISONETTE', followed by 'TERRACED' and 'SEMI-DETACHED' (Table 2). In terms of the 'new_build' status, the majority of properties are not new builds. However, there is a highly imbalanced dataset for the 'new_build' variable, where 99% of the values are 'no' and only 1% are 'yes.' It's important to be aware of potential challenges and consider appropriate strategies (Table 3). Regarding the 'estate_type,' 'LEASEHOLD' is more common than 'FREEHOLD' (Table 4).

Table 5: Missing Value Detection



Univariate Statistics

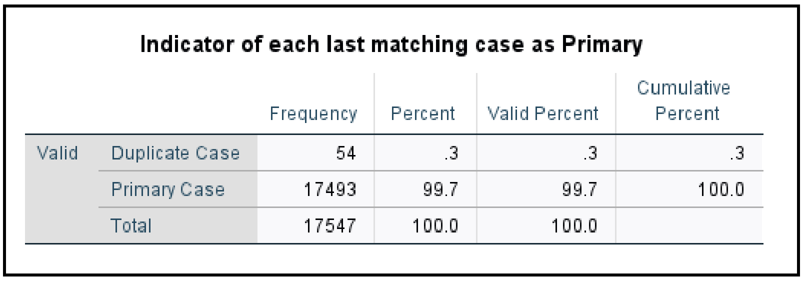
	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
price_paid	17547	899588.4738	2349547.380	0	.0	0	182
postcode_num	17521	7084.48	4105.331	26	.1	0	0
property_type_num	17547	2.98	1.346	0	.0	0	0
new_build_num	17547	1.01	.099	0	.0	0	172
estate_type_num	17547	1.62	.486	0	.0	0	0
street_num	17547	4197.96	2401.105	0	.0	0	0
locality_num	1110	70.47	41.704	16437	93.7	0	0
town_num	17547	1.00	.000	0	.0	.	.
district_num	17547	16.08	8.236	0	.0	0	0
country_num	17547	1.00	.000	0	.0	.	.
unique_num	17547	8774.00	5065.527	0	.0	0	0

a. Number of cases outside the range (Mean - 2*SD, Mean + 2*SD).

During the examination of the dataset, a crucial step involved checking the data types to ensure consistency and accuracy. In the SPSS software, a specific focus was placed on identifying numeric values and addressing potential issues with missing data. While checking the dataset, the variable 'postcode_num' was identified to have 26 instances of missing data out of a total of 17,547 observations (Table 5). Considering the inherent regional specificity of postal code data, where each area corresponds to a unique postal code, a decision was made to replace these 26 missing values with 'unknown' instead of using mean or median imputations. To execute this replacement, simply navigate to the 'Data View' tab, identify the variable with missing values, and replace the blanks with 'unknown.'

Simultaneously, the variable 'locality_num' exhibits a considerable number of missing values, accounting for 16,437 out of the total 17,547 entries (Table 5). Recognizing the substantial proportion (93.7%) of missing data in 'locality_num', a strategic decision was made to exclude this variable from the dataset. The rationale behind this choice is that the amount of missing information is significant, and attempting to estimate such a substantial portion may compromise the overall integrity of the dataset (Kwak and Kim).

Table 6: Duplicate Data Detection

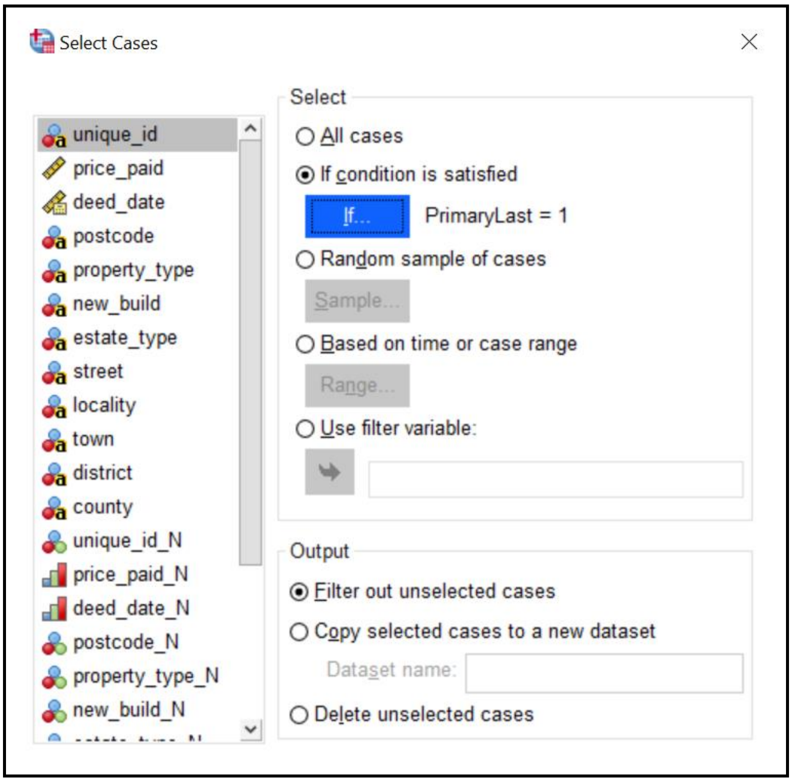


Indicator of each last matching case as Primary

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Duplicate Case	54	.3	.3	.3
	Primary Case	17493	99.7	99.7	100.0
	Total	17547	100.0	100.0	

Upon identifying 54 instances of duplicated data in Table 6, a removal step was taken to enhance data cleanliness and accuracy. ~~Figure 2~~Figure 2 and ~~Figure 3~~Figure 3 below illustrate the steps involved in this data-cleansing procedure, showcasing the implementation of the removal process.

This action aims to ensure the reliability and integrity of the dataset by eliminating redundancies and promoting consistency in subsequent analyses.



Select Cases

unique_id
price_paid
deed_date
postcode
property_type
new_build
estate_type
street
locality
town
district
county
unique_id_N
price_paid_N
deed_date_N
postcode_N
property_type_N
new_build_N

Select

○ All cases

● If condition is satisfied

If... PrimaryLast = 1

○ Random sample of cases

○ Based on time or case range

○ Use filter variable:

Output

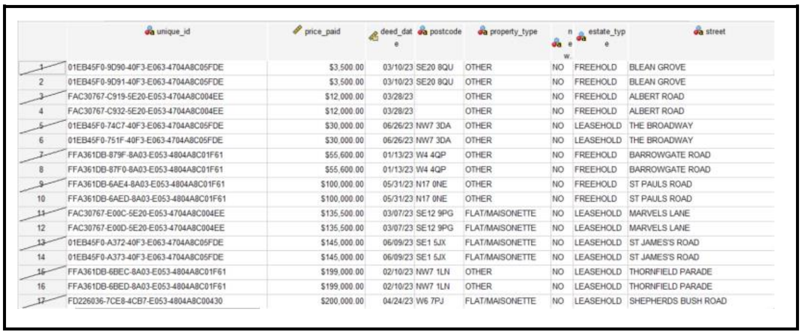
● Filter out unselected cases

○ Copy selected cases to a new dataset

Dataset name:

○ Delete unselected cases

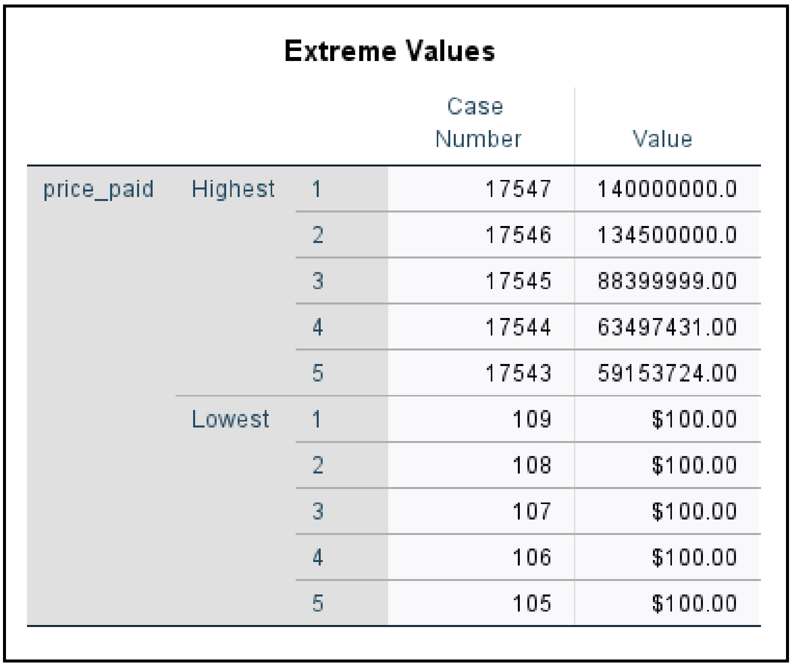
Figure 2: Removing Duplicate Data using Select Cases in SPSS



	unique_id	price_paid	deed_date	postcode	property_type	new	estate_type	street
1	01EB45F0-9D90-40F3-E063-4704A8C05FDE	$3,500.00	03/10/23	SE20 8QU	OTHER	NO	FREEHOLD	BLEAN GROVE
2	01EB45F0-9D91-40F3-E063-4704A8C05FDE	$3,500.00	03/10/23	SE20 8QU	OTHER	NO	FREEHOLD	BLEAN GROVE
3	FAC30767-C919-5E20-E053-4704A8C004EE	$12,000.00	03/28/23		OTHER	NO	FREEHOLD	ALBERT ROAD
4	FAC30767-C932-5E20-E053-4704A8C004EE	$12,000.00	03/28/23		OTHER	NO	FREEHOLD	ALBERT ROAD
5	01EB45F0-74C7-40F3-E063-4704A8C05FDE	$30,000.00	06/26/23	NW7 3DA	OTHER	NO	LEASEHOLD	THE BROADWAY
6	01EB45F0-751F-40F3-E063-4704A8C05FDE	$30,000.00	06/26/23	NW7 3DA	OTHER	NO	LEASEHOLD	THE BROADWAY
7	FFA361DB-879F-8A03-E053-4804A8C01F61	$55,600.00	01/13/23	W4 4QP	OTHER	NO	FREEHOLD	BARROWGATE ROAD
8	FFA361DB-87F0-8A03-E053-4804A8C01F61	$55,600.00	01/13/23	W4 4QP	OTHER	NO	FREEHOLD	BARROWGATE ROAD
9	FFA361DB-6AE4-8A03-E053-4804A8C01F61	$100,000.00	05/31/23	N17 0NE	OTHER	NO	FREEHOLD	ST PAULS ROAD
10	FFA361DB-6AED-8A03-E053-4804A8C01F61	$100,000.00	05/31/23	N17 0NE	OTHER	NO	FREEHOLD	ST PAULS ROAD
11	FAC30767-E00C-5E20-E053-4704A8C004EE	$135,500.00	03/07/23	SE12 9PG	FLAT/MAISONETTE	NO	LEASEHOLD	MARVELS LANE
12	FAC30767-E00D-5E20-E053-4704A8C004EE	$135,500.00	03/07/23	SE12 9PG	FLAT/MAISONETTE	NO	LEASEHOLD	MARVELS LANE
13	01EB45F0-A372-40F3-E063-4704A8C05FDE	$145,000.00	06/09/23	SE1 5JX	FLAT/MAISONETTE	NO	LEASEHOLD	ST JAMES'S ROAD
14	01EB45F0-A373-40F3-E063-4704A8C05FDE	$145,000.00	06/09/23	SE1 5JX	FLAT/MAISONETTE	NO	LEASEHOLD	ST JAMES'S ROAD
15	FFA361DB-6BEC-8A03-E053-4804A8C01F61	$199,000.00	02/10/23	NW7 1LN	OTHER	NO	LEASEHOLD	THORNFIELD PARADE
16	FFA361DB-6BED-8A03-E053-4804A8C01F61	$199,000.00	02/10/23	NW7 1LN	OTHER	NO	LEASEHOLD	THORNFIELD PARADE
17	FD226036-7CE8-4CB7-E053-4804A8C00430	$200,000.00	04/24/23	W6 7PJ	FLAT/MAISONETTE	NO	LEASEHOLD	SHEPHERDS BUSH ROAD

Figure 3: Duplicate Data

Table 7: Outliers Finding



Extreme Values

			Case Number	Value
price_paid	Highest	1	17547	140000000.0
		2	17546	134500000.0
		3	17545	88399999.00
		4	17544	63497431.00
		5	17543	59153724.00
	Lowest	1	109	$100.00
		2	108	$100.00
		3	107	$100.00
		4	106	$100.00
		5	105	$100.00

In the process of examining outliers in the dataset, it was observed that the top 5 highest outlier values for the 'price_paid' variable ranged from £59,153,724.00 to £140,000,000.0. Conversely, the last 5 lowest outlier values were noted to be £100 (Table 7). Recognizing the need to enhance the robustness of the dataset and ensure the reliability of subsequent analyses, a decision was made to apply both the Interquartile Range (IQR) and Z-score methods to effectively identify these outliers. This approach aims to mitigate the influence of extreme values, promoting a more accurate representation of the central tendency within the dataset.

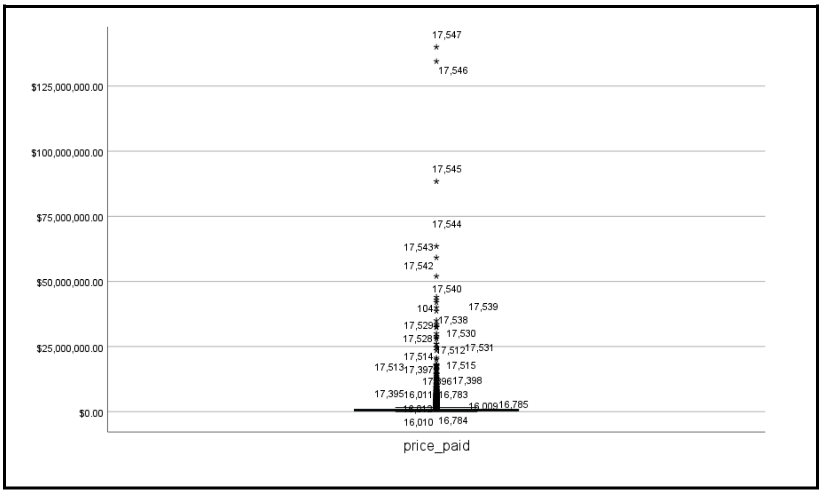
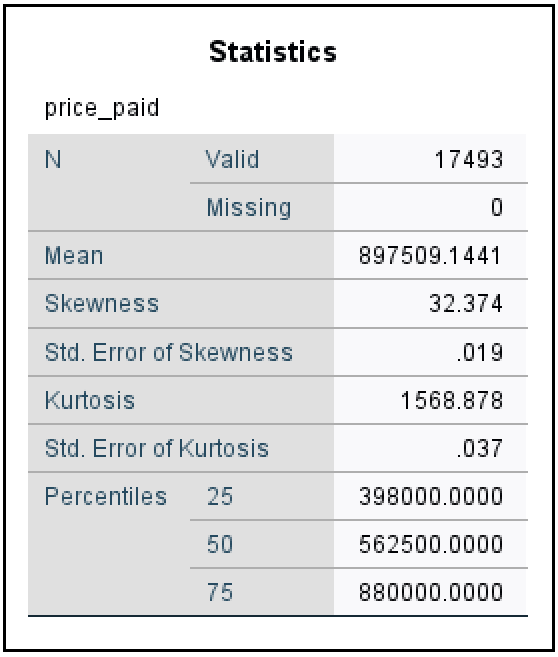


Figure 4: Boxplot of Outliers

In this case, Figure 4 displays a multitude of extreme outliers, denoted by the star sign. To address these outliers, both the Interquartile Range (Table 8)and Z-score methods are applied.



Statistics

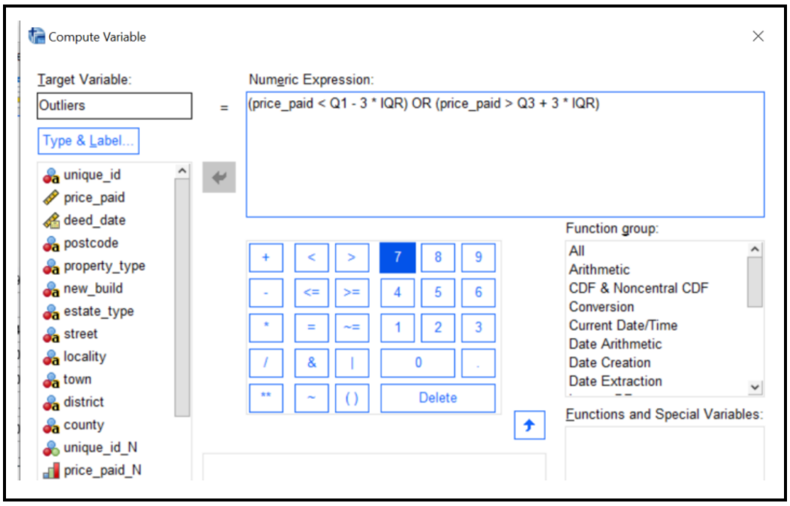
price_paid

N	Valid	17493
	Missing	0
Mean		897509.1441
Skewness		32.374
Std. Error of Skewness		.019
Kurtosis		1568.878
Std. Error of Kurtosis		.037
Percentiles	25	398000.0000
	50	562500.0000
	75	880000.0000

Table 8: Interquartile Range (IQR) of 'price_paid' by the 25th, 50th and 75th

In the Interquartile Range outlier detection method, the function (price_paid < Q1 - 3 * IQR) OR (price_paid > Q3 + 3 * IQR) is employed to identify extreme outliers in the dataset Figure 5). This method utilizes a more stringent 3 times IQR threshold to capture data points with significantly skewed values. This adjustment was made to enhance the reliable identification of extreme values. However, it is important to note that using this stringent criterion results in the exclusion of a large number of data points, potentially leading to a loss of valuable information (Figure 6 and Figure 7) (Barbato et al.).

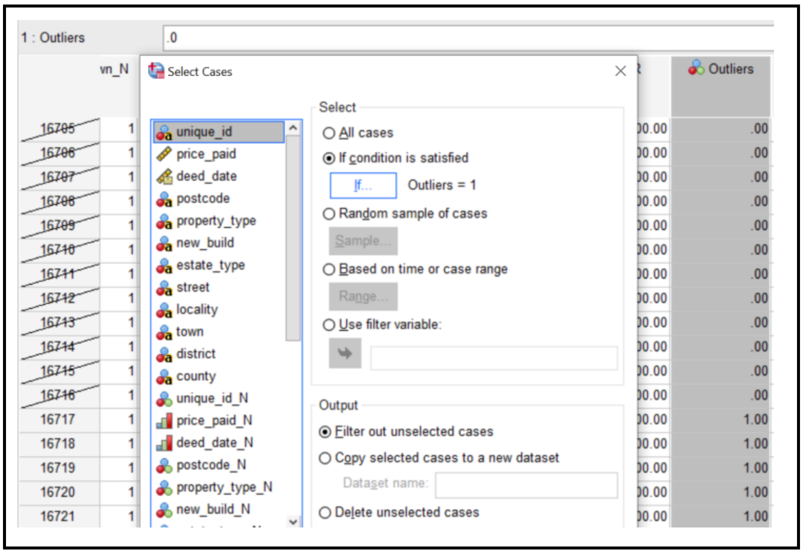
Furthermore, a comparison between the Interquartile Range and Z-score methods for outlier detection revealed differences, with the Z-score failing to detect certain outliers. London maintains its status as the region with the highest average house prices in the UK, reaching £516,000 in October 2023 ("UK House Price Index - Office for National Statistics"). However, the dataset contains notably low house prices (e.g., £100, £600, and £1000), potentially considered outliers. Despite this, Z-score analysis did not identify outliers below £1,000 (Figure 8, Figure 9, and Figure 10). Consequently, a decision was made to retain all outliers to preserve crucial data and prevent the potential loss of valuable information.



Target Variable: Outliers

Numeric Expression: (price_paid < Q1 - 3 * IQR) OR (price_paid > Q3 + 3 * IQR)

Figure 5: Extreme Outliers Condition in IQR



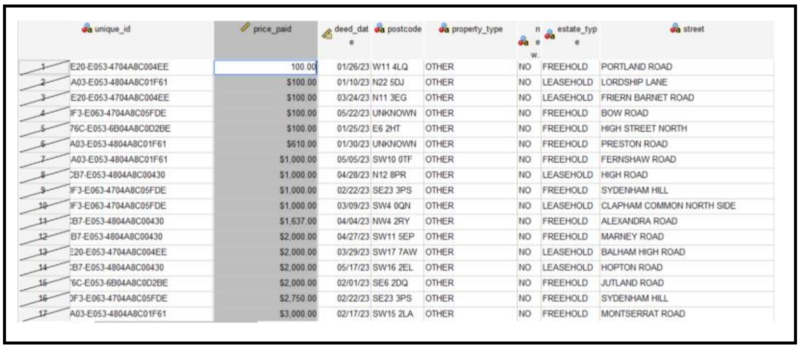
1 : Outliers .0

Select: ● If condition is satisfied — If... Outliers = 1

Output: ● Filter out unselected cases

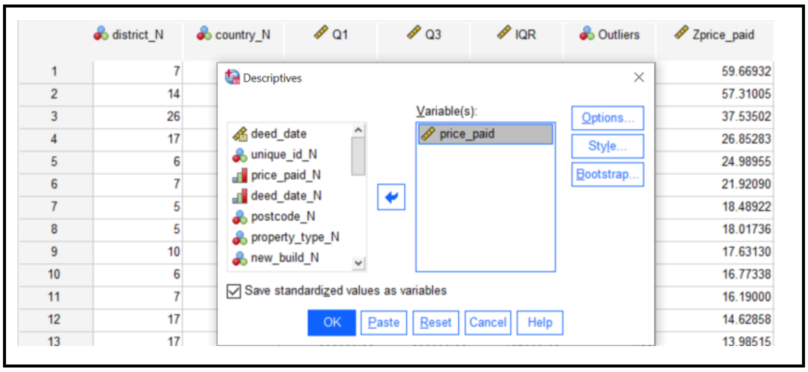
	vn_N		Outliers
~~16705~~	1	)0.00	.00
~~16706~~	1	)0.00	.00
~~16707~~	1	)0.00	.00
~~16708~~	1	)0.00	.00
~~16709~~	1	)0.00	.00
~~16710~~	1	)0.00	.00
~~16711~~	1	)0.00	.00
~~16712~~	1	)0.00	.00
~~16713~~	1	)0.00	.00
~~16714~~	1	)0.00	.00
~~16715~~	1	)0.00	.00
~~16716~~	1	)0.00	.00
16717	1	)0.00	1.00
16718	1	)0.00	1.00
16719	1	)0.00	1.00
16720	1	)0.00	1.00
16721	1	)0.00	1.00

Figure 6: 16,716 Outliers Filtered using IQR Method



	unique_id	price_paid	deed_date	postcode	property_type	n e w..	estate_type	street
~~1~~	E20-E053-4704A8C004EE	100.00	01/26/23	W11 4LQ	OTHER	NO	FREEHOLD	PORTLAND ROAD
~~2~~	A03-E053-4804A8C01F61	$100.00	01/10/23	N22 5DJ	OTHER	NO	LEASEHOLD	LORDSHIP LANE
~~3~~	E20-E053-4704A8C004EE	$100.00	03/24/23	N11 3EG	OTHER	NO	LEASEHOLD	FRIERN BARNET ROAD
~~4~~	)F3-E063-4704A8C05FDE	$100.00	05/22/23	UNKNOWN	OTHER	NO	FREEHOLD	BOW ROAD
~~5~~	76C-E053-6B04A8C0D2BE	$100.00	01/25/23	E6 2HT	OTHER	NO	FREEHOLD	HIGH STREET NORTH
~~6~~	A03-E053-4804A8C01F61	$610.00	01/30/23	UNKNOWN	OTHER	NO	FREEHOLD	PRESTON ROAD
~~7~~	A03-E053-4804A8C01F61	$1,000.00	05/05/23	SW10 0TF	OTHER	NO	FREEHOLD	FERNSHAW ROAD
~~8~~	CB7-E053-4804A8C00430	$1,000.00	04/28/23	N12 8PR	OTHER	NO	LEASEHOLD	HIGH ROAD
~~9~~	)F3-E063-4704A8C05FDE	$1,000.00	02/22/23	SE23 3PS	OTHER	NO	FREEHOLD	SYDENHAM HILL
~~10~~	)F3-E063-4704A8C05FDE	$1,000.00	03/09/23	SW4 0QN	OTHER	NO	LEASEHOLD	CLAPHAM COMMON NORTH SIDE
~~11~~	CB7-E053-4804A8C00430	$1,637.00	04/04/23	NW4 2RY	OTHER	NO	FREEHOLD	ALEXANDRA ROAD
~~12~~	B7-E053-4804A8C00430	$2,000.00	04/27/23	SW11 5EP	OTHER	NO	FREEHOLD	MARNEY ROAD
~~13~~	E20-E053-4704A8C004EE	$2,000.00	03/29/23	SW17 7AW	OTHER	NO	LEASEHOLD	BALHAM HIGH ROAD
~~14~~	CB7-E053-4804A8C00430	$2,000.00	05/17/23	SW16 2EL	OTHER	NO	LEASEHOLD	HOPTON ROAD
~~15~~	'6C-E053-6B04A8C0D2BE	$2,000.00	02/01/23	SE6 2DQ	OTHER	NO	FREEHOLD	JUTLAND ROAD
~~16~~	)F3-E063-4704A8C05FDE	$2,750.00	02/22/23	SE23 3PS	OTHER	NO	FREEHOLD	SYDENHAM HILL
~~17~~	A03-E053-4804A8C01F61	$3,000.00	02/17/23	SW15 2LA	OTHER	NO	FREEHOLD	MONTSERRAT ROAD

Figure 7: Some Outliers Data using IQR Method



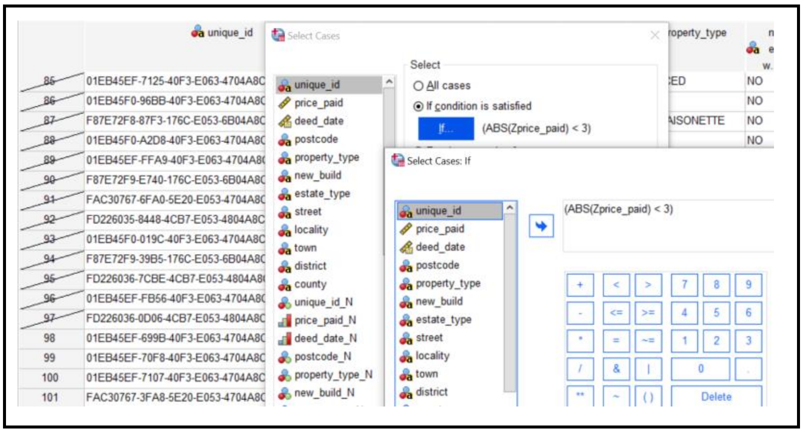
	district_N	country_N	Q1	Q3	IQR	Outliers	Zprice_paid
1	7						59.66932
2	14						57.31005
3	26						37.53502
4	17						26.85283
5	6						24.98955
6	7						21.92090
7	5						18.48922
8	5						18.01736
9	10						17.63130
10	6						16.77338
11	7						16.19000
12	17						14.62858
13	17						13.98515

Descriptives

Variable(s): price_paid

☑ Save standardized values as variables

Figure 8: Applied Z-Score Method of 'price_paid'



	unique_id	roperty_type	new
~~85~~	01EB45EF-7125-40F3-E063-4704A8C	ED	NO
~~86~~	01EB45F0-96BB-40F3-E063-4704A8C		NO
~~87~~	F87E72F8-87F3-176C-E053-6B04A8C	AISONETTE	NO
~~88~~	01EB45F0-A2D8-40F3-E063-4704A8C		NO
~~89~~	01EB45EF-FFA9-40F3-E063-4704A8C		
~~90~~	F87E72F9-E740-176C-E053-6B04A8C		
~~91~~	FAC30767-6FA0-5E20-E053-4704A8C		
~~92~~	FD226035-8448-4CB7-E053-4804A8C		
~~93~~	01EB45F0-019C-40F3-E063-4704A8C		
~~94~~	F87E72F9-39B5-176C-E053-6B04A8C		
~~95~~	FD226036-7CBE-4CB7-E053-4804A8		
~~96~~	01EB45EF-FB56-40F3-E063-4704A8C		
~~97~~	FD226036-0D06-4CB7-E053-4804A8C		
98	01EB45EF-699B-40F3-E063-4704A8C		
99	01EB45EF-70F8-40F3-E063-4704A8C		
100	01EB45EF-7107-40F3-E063-4704A8C		
101	FAC30767-3FA8-5E20-E053-4704A8C		

Select Cases

Select

○ All cases

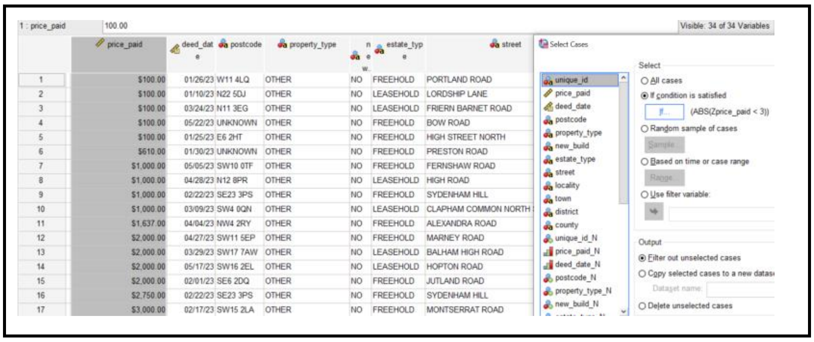
● If condition is satisfied

If... (ABS(Zprice_paid) < 3)

Select Cases: If

(ABS(Zprice_paid) < 3)

Figure 9: Outliers Condition in Z-Score using Threshold 3



1 : price_paid 100.00 — Visible: 34 of 34 Variables

	price_paid	deed_date	postcode	property_type	new	estate_type	street
1	$100.00	01/26/23	W11 4LQ	OTHER	NO	FREEHOLD	PORTLAND ROAD
2	$100.00	01/10/23	N22 5DJ	OTHER	NO	LEASEHOLD	LORDSHIP LANE
3	$100.00	03/24/23	N11 3EG	OTHER	NO	LEASEHOLD	FRIERN BARNET ROAD
4	$100.00	05/22/23	UNKNOWN	OTHER	NO	FREEHOLD	BOW ROAD
5	$100.00	01/25/23	E6 2HT	OTHER	NO	FREEHOLD	HIGH STREET NORTH
6	$610.00	01/30/23	UNKNOWN	OTHER	NO	FREEHOLD	PRESTON ROAD
7	$1,000.00	05/05/23	SW10 0TF	OTHER	NO	FREEHOLD	FERNSHAW ROAD
8	$1,000.00	04/28/23	N12 8PR	OTHER	NO	LEASEHOLD	HIGH ROAD
9	$1,000.00	02/22/23	SE23 3PS	OTHER	NO	FREEHOLD	SYDENHAM HILL
10	$1,000.00	03/09/23	SW4 0QN	OTHER	NO	LEASEHOLD	CLAPHAM COMMON NORTH
11	$1,637.00	04/04/23	NW4 2RY	OTHER	NO	FREEHOLD	ALEXANDRA ROAD
12	$2,000.00	04/27/23	SW11 5EP	OTHER	NO	FREEHOLD	MARNEY ROAD
13	$2,000.00	03/29/23	SW17 7AW	OTHER	NO	LEASEHOLD	BALHAM HIGH ROAD
14	$2,000.00	05/17/23	SW16 2EL	OTHER	NO	LEASEHOLD	HOPTON ROAD
15	$2,000.00	02/01/23	SE6 2DQ	OTHER	NO	FREEHOLD	JUTLAND ROAD
16	$2,750.00	02/22/23	SE23 3PS	OTHER	NO	FREEHOLD	SYDENHAM HILL
17	$3,000.00	02/17/23	SW15 2LA	OTHER	NO	FREEHOLD	MONTSERRAT ROAD

Select Cases

Select

○ All cases

● If condition is satisfied

If... (ABS(Zprice_paid < 3))

○ Random sample of cases

○ Based on time or case range

○ Use filter variable:

Output

● Filter out unselected cases

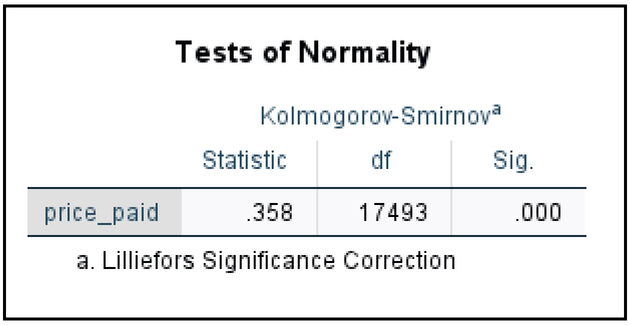
○ Copy selected cases to a new datase

○ Delete unselected cases

Figure 10: Undetected Outliers: Data Points Not Identified by Z-Score Method

When assessing the normality of a dataset, the choice often lies between the Shapiro-Wilk test and the Kolmogorov-Smirnov test. While the Shapiro-Wilk test is renowned for its accuracy in evaluating normality, it was dismissed from consideration due to its recommendation for small to moderately-sized samples. Instead, the Kolmogorov-Smirnov test emerged as the preferred option for this analysis, particularly given the larger sample size and its simplicity in application (Steinskog et al.). The primary focus of the normality test was directed toward continuous variables, with particular emphasis on the 'price_paid'. To enhance the statistical evaluation, visual methods such as histograms (Figure 11) and Q-Q plots (Figure 12 and Figure 13) were incorporated. These visual aids complemented the analysis by providing a more intuitive understanding of the distribution, ensuring a comprehensive approach to the assessment of the dataset's normality.

Table 9: Normality Test Result



Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
price_paid	.358	17493	.000

a. Lilliefors Significance Correction

The Kolmogorov-Smirnov test statistic, reported as 0.358 in (Table 9), represents the maximum vertical distance between the observed cumulative distribution and the expected cumulative distribution under the assumption of normality. The associated p-value is reported as 0.000, indicating the probability of observing a test statistic as extreme as the calculated one, assuming the data follows a normal distribution.

In the context of the K-S test, the null hypothesis posits that the sample is drawn from a population adhering to a specified distribution, in this case, a normal distribution. Since the p-value (Sig) is less than the conventional significance level of 0.05, it is typically grounds to reject the null hypothesis. The very small p-value of 0.000 suggests substantial evidence to reject the null hypothesis, signifying that the data's distribution significantly deviates from a normal distribution (Leech et al.). Consequently, it is advisable to employ non-parametric tests in subsequent analysis processes.

After the preprocessing step, the following figures provide insights into the normality of the 'price_paid' variable in London. The histogram (Figure 11) illustrates a right-skewed distribution of London property prices, indicating that the majority of prices are concentrated on the lower end, with a tail extending toward higher values. The normal Q-Q plot (Figure 12) and detrended normal Q-Q plot (Figure 13) further reveal deviations from a theoretical normal distribution. These visualizations, in conjunction with the right-skewed histogram, collectively suggest that the 'price_paid' variable in London deviates from a normal population distribution.

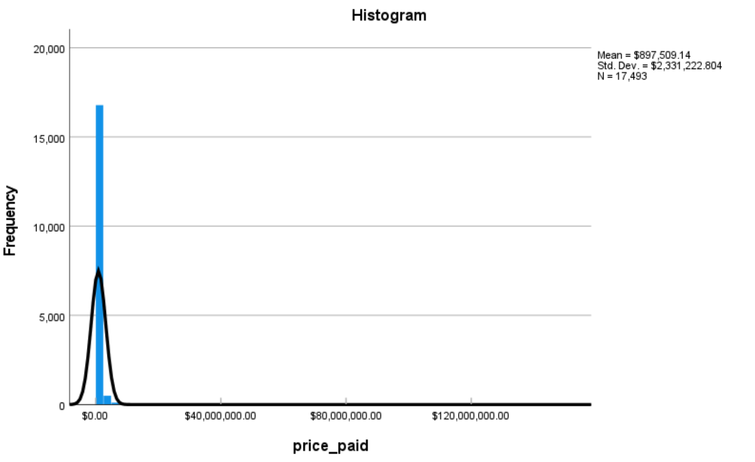


Figure 11 'price_paid' Normality Histogram

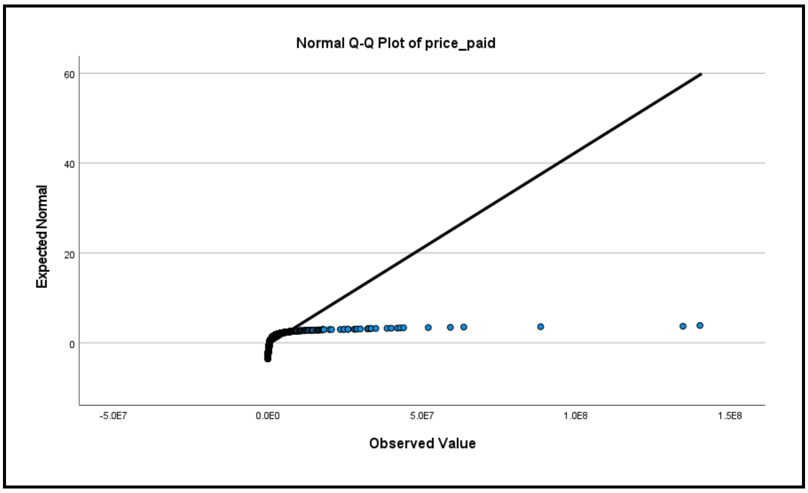


Figure 12: Normal Q-Q plot ‘price_paid’

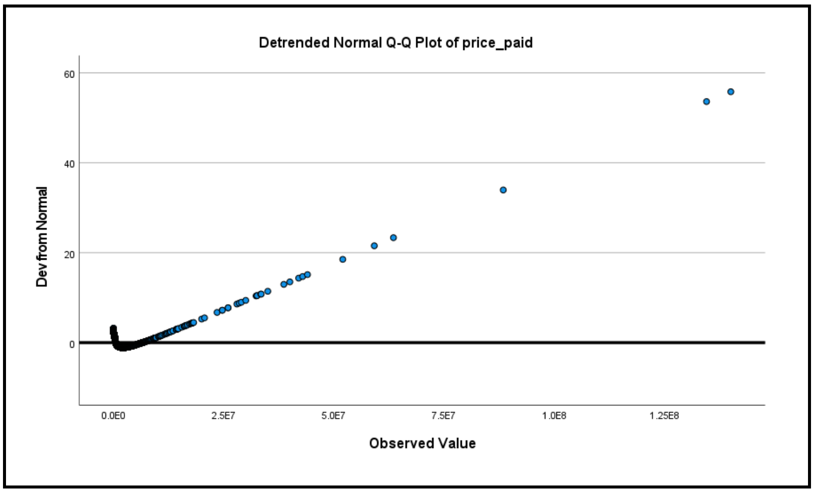
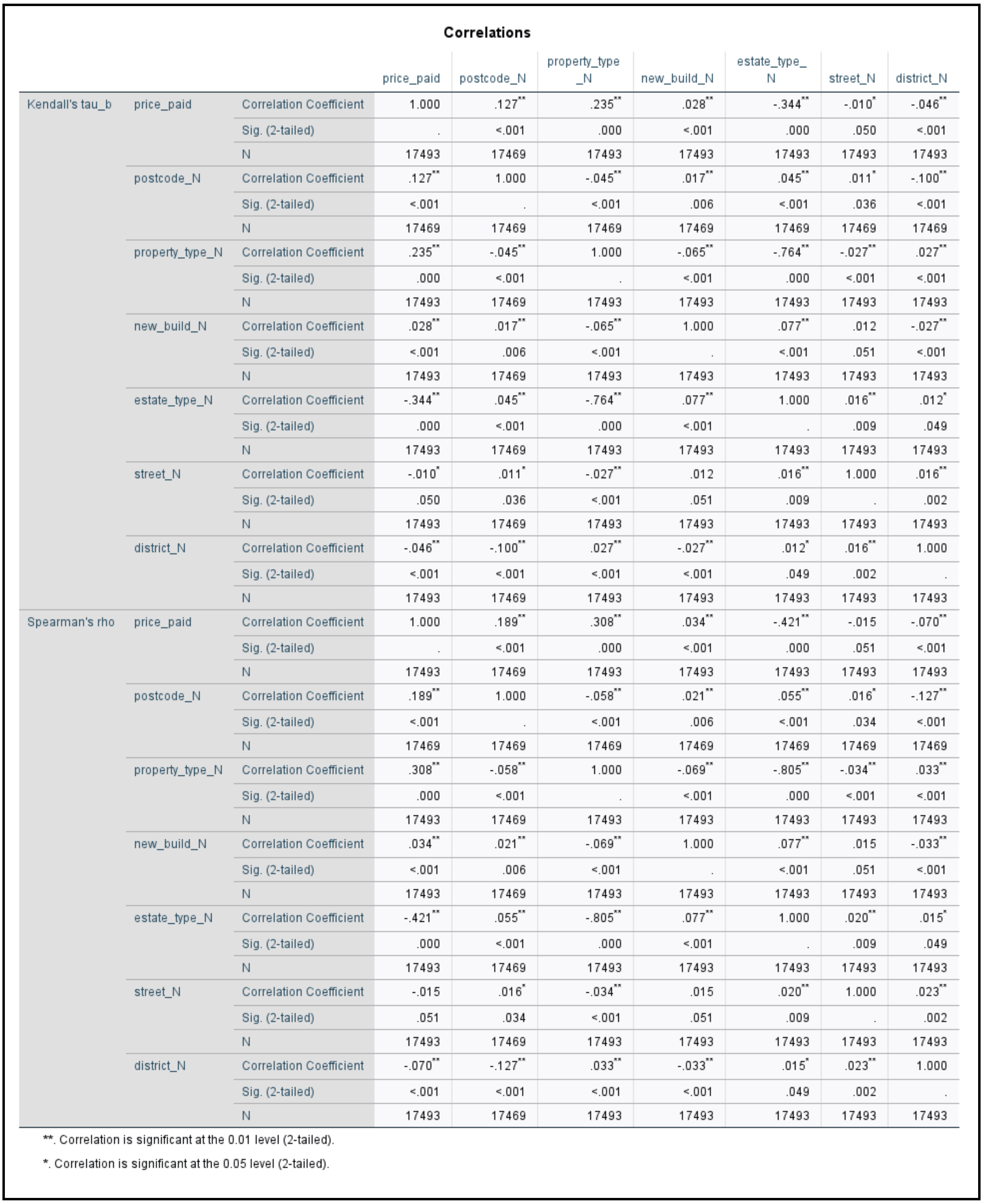


Figure 13: Detrended Normal Q-Q plot ‘price_paid’

Table 10: Correlation using Kendall's Tau and Spearman



Correlations

			price_paid	postcode_N	property_type_N	new_build_N	estate_type_N	street_N	district_N
Kendall's tau_b	price_paid	Correlation Coefficient	1.000	.127**	.235**	.028**	-.344**	-.010*	-.046**
		Sig. (2-tailed)	.	<.001	.000	<.001	.000	.050	<.001
		N	17493	17469	17493	17493	17493	17493	17493
	postcode_N	Correlation Coefficient	.127**	1.000	-.045**	.017**	.045**	.011*	-.100**
		Sig. (2-tailed)	<.001	.	<.001	.006	<.001	.036	<.001
		N	17469	17469	17469	17469	17469	17469	17469
	property_type_N	Correlation Coefficient	.235**	-.045**	1.000	-.065**	-.764**	-.027**	.027**
		Sig. (2-tailed)	.000	<.001	.	<.001	.000	<.001	<.001
		N	17493	17469	17493	17493	17493	17493	17493
	new_build_N	Correlation Coefficient	.028**	.017**	-.065**	1.000	.077**	.012	-.027**
		Sig. (2-tailed)	<.001	.006	<.001	.	<.001	.051	<.001
		N	17493	17469	17493	17493	17493	17493	17493
	estate_type_N	Correlation Coefficient	-.344**	.045**	-.764**	.077**	1.000	.016**	.012*
		Sig. (2-tailed)	.000	<.001	.000	<.001	.	.009	.049
		N	17493	17469	17493	17493	17493	17493	17493
	street_N	Correlation Coefficient	-.010*	.011*	-.027**	.012	.016**	1.000	.016**
		Sig. (2-tailed)	.050	.036	<.001	.051	.009	.	.002
		N	17493	17469	17493	17493	17493	17493	17493
	district_N	Correlation Coefficient	-.046**	-.100**	.027**	-.027**	.012*	.016**	1.000
		Sig. (2-tailed)	<.001	<.001	<.001	<.001	.049	.002	.
		N	17493	17469	17493	17493	17493	17493	17493
Spearman's rho	price_paid	Correlation Coefficient	1.000	.189**	.308**	.034**	-.421**	-.015	-.070**
		Sig. (2-tailed)	.	<.001	.000	<.001	.000	.051	<.001
		N	17493	17469	17493	17493	17493	17493	17493
	postcode_N	Correlation Coefficient	.189**	1.000	-.058**	.021**	.055**	.016*	-.127**
		Sig. (2-tailed)	<.001	.	<.001	.006	<.001	.034	<.001
		N	17469	17469	17469	17469	17469	17469	17469
	property_type_N	Correlation Coefficient	.308**	-.058**	1.000	-.069**	-.805**	-.034**	.033**
		Sig. (2-tailed)	.000	<.001	.	<.001	.000	<.001	<.001
		N	17493	17469	17493	17493	17493	17493	17493
	new_build_N	Correlation Coefficient	.034**	.021**	-.069**	1.000	.077**	.015	-.033**
		Sig. (2-tailed)	<.001	.006	<.001	.	<.001	.051	<.001
		N	17493	17469	17493	17493	17493	17493	17493
	estate_type_N	Correlation Coefficient	-.421**	.055**	-.805**	.077**	1.000	.020**	.015*
		Sig. (2-tailed)	.000	<.001	.000	<.001	.	.009	.049
		N	17493	17469	17493	17493	17493	17493	17493
	street_N	Correlation Coefficient	-.015	.016*	-.034**	.015	.020**	1.000	.023**
		Sig. (2-tailed)	.051	.034	<.001	.051	.009	.	.002
		N	17493	17469	17493	17493	17493	17493	17493
	district_N	Correlation Coefficient	-.070**	-.127**	.033**	-.033**	.015*	.023**	1.000
		Sig. (2-tailed)	<.001	<.001	<.001	<.001	.049	.002	.
		N	17493	17469	17493	17493	17493	17493	17493

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Kendall's Tau and Spearman are employed as nonparametric correlation coefficients, particularly in scenarios where nonlinearity is plausible. In comparison to Pearson's correlation, which might struggle to effectively capture nonlinear relationships, Spearman and Kendall's Tau exhibit greater resilience to such complexities. Their reliance on rankings rather than precise values is a significant advantage, imparting robustness in the presence of outliers. Both correlation coefficients span from -1 to 1: a value of 1 denotes perfect agreement, -1 indicates complete disagreement, and 0 signifies no correlation. The significance value (Sig) of the correlation aids in determining its statistical significance, with a small p-value (usually less than 0.05) indicating a significant correlation and providing a reliable indication of the strength and directionality of the relationship between variables (Chok).

However, the results in Table 10: Correlation using Kendall's Tau and Spearman reveal lower correlations and smaller p-values. In this case, we deduce that despite the weak correlation, it holds statistical significance and should not be dismissed, particularly when considering multivariate confounders. Employing both Kendall's Tau and Spearman's correlation tests allows for a comprehensive examination of the relationship, leveraging their unique strengths in different situations. Identifying relevant variables requires extensive discussion between analysts and investigators to ensure accurate inferences are drawn. Therefore, caution is warranted, and hasty conclusions should be avoided in this scenario. As a result, these correlation measures will not be utilized for the following analytics.

1.3 Data Analytical Method

According to the test results above, conclude that the dataset follows a non-normal distribution, indicating the need for non-parametric tests.

Research Question 1:

Which London district offers properties with the highest, middle, and lowest prices?

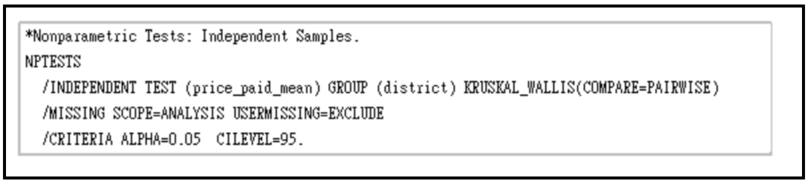
Hypotheses:

H0: There is no significant correlation between house prices in different districts of London.
H1: There is a significant correlation between house prices in different districts of London.

To address the first research question and hypothesis, the Kruskal-Wallis method was applied to test the hypothesis. This non-parametric test serves as an alternative to one-way analysis of variance (ANOVA), examining the null hypothesis that populations from different groups share the same distribution. The analysis utilized the variables 'price_paid_mean' and 'district' (Figure 14).

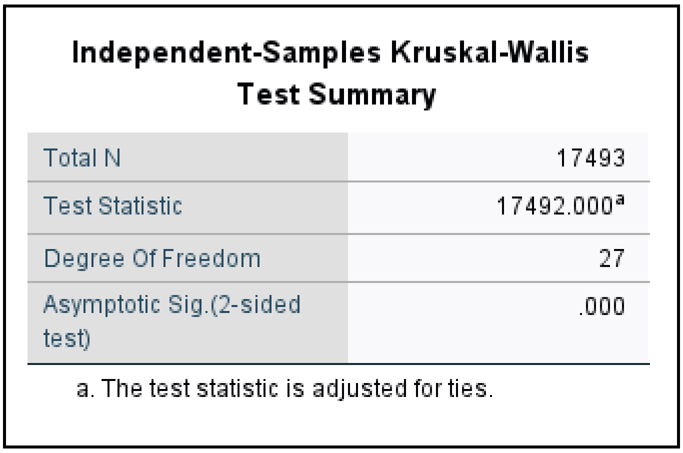
The Kruskal-Wallis test yielded a substantial statistic of 17492.000, indicating strong evidence against the null hypothesis (Figure 15). Furthermore, Figure 16 provides a summary test result, affirming the rejection of the null hypothesis.

Upon examining the output, the p-value is found to be less than the significance level (0.05). This leads to the rejection of the null hypothesis (H0) and the acceptance of the alternative hypothesis (H1), suggesting significant differences in median house prices among different districts. Visual representations of the Kruskal-Wallis Test results can be found in Figure 17 and Figure 18.



```
*Nonparametric Tests: Independent Samples.
NPTESTS
  /INDEPENDENT TEST (price_paid_mean) GROUP (district) KRUSKAL_WALLIS(COMPARE=PAIRWISE)
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.
```

Figure 14: Kruskal-Wallis Test in SPSS

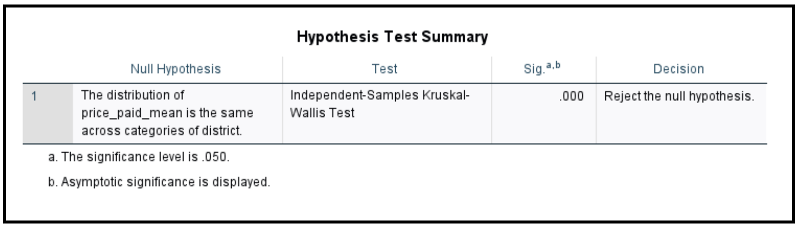


Independent-Samples Kruskal-Wallis Test Summary

Total N	17493
Test Statistic	17492.000[a]
Degree Of Freedom	27
Asymptotic Sig.(2-sided test)	.000

a. The test statistic is adjusted for ties.

Figure 15: Kruskal-Wallis Test Statistic Result



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.^{a,b}	Decision
1	The distribution of price_paid_mean is the same across categories of district.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Figure 16: Kruskal-Wallis Test Summary

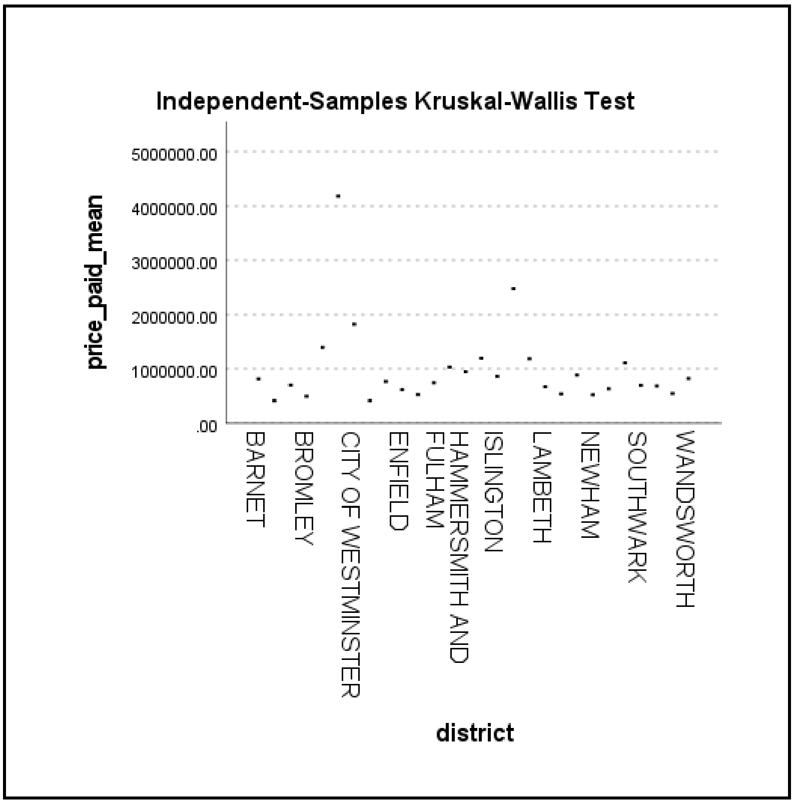


Figure 17: Scatter Plot of Mean Price Paid by District

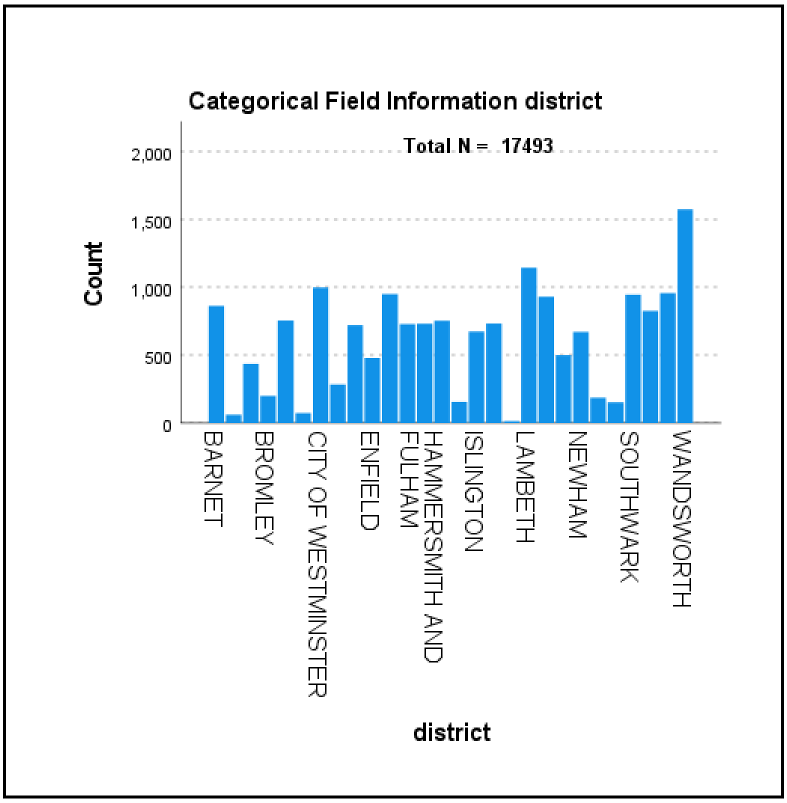
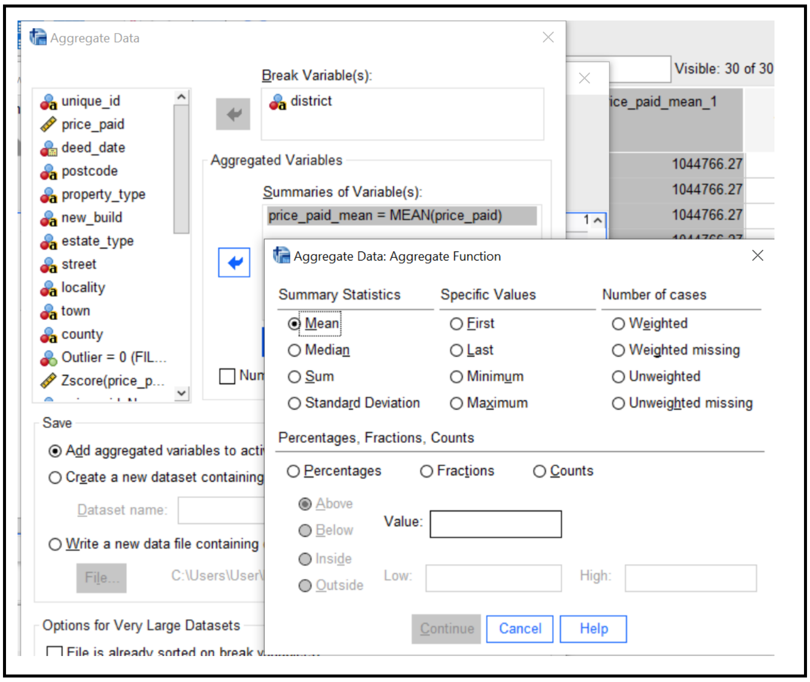


Figure 18: District Count Histogram

After confirming a relationship between house prices in different districts of London, used the Aggregate Data function to aggregate the 'price_paid' variable by mean, revealing the mean value for each London district (Figure 19). Subsequently, created a new variable 'price_paid_max' to denote the highest mean value of the price paid (Figure 20). The same methodology was then applied to identify the middle and lowest mean prices (Figure 21 and Figure 22).



Aggregate Data

Break Variable(s): district

Summaries of Variable(s): price_paid_mean = MEAN(price_paid)

Aggregate Data: Aggregate Function

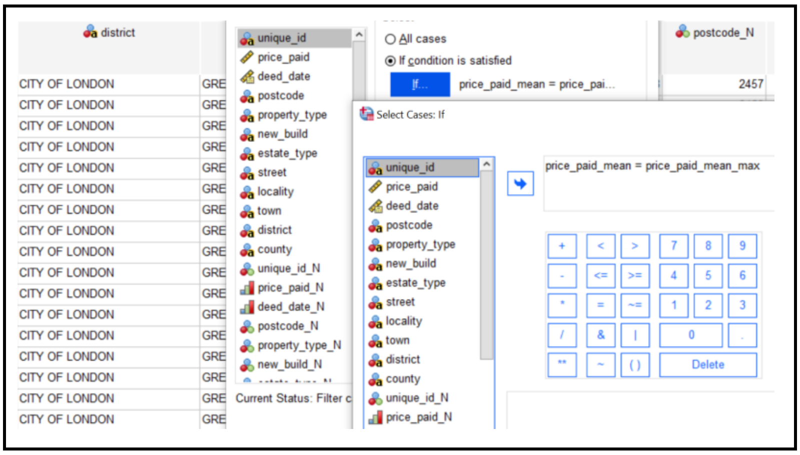
Summary Statistics: ● Mean

Save: ● Add aggregated variables to acti

Visible: 30 of 30

ice_paid_mean_1
1044766.27
1044766.27
1044766.27

Figure 19: Average Property Prices by London District (Aggregate Method)



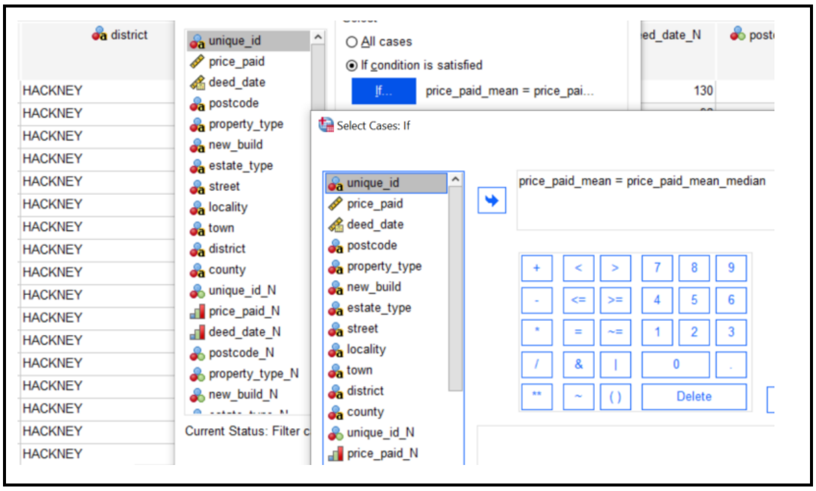
Select Cases: If

● If condition is satisfied

price_paid_mean = price_paid_mean_max

district	postcode_N
CITY OF LONDON	2457
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	
CITY OF LONDON	

Figure 20: Max Mean Price by District



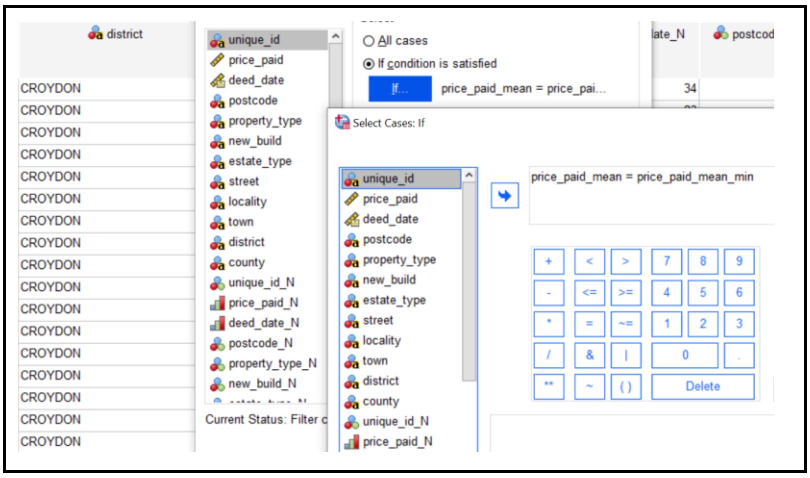
Select Cases: If

● If condition is satisfied

price_paid_mean = price_paid_mean_median

district	ed_date_N
HACKNEY	130
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	
HACKNEY	

Figure 21: Middle Mean Price by District



price_paid_mean = price_paid_mean_min

Figure 22: Min Mean Price by District

Table 11: London District Prices Overview (Highest, Middle, and Lowest)

	District		Price Paid
Top District	CITY OF LONDON		4184777.73
Middle District	HACKNEY		737358.46
Bottom District	CROYDON		412087.03

In Table 11, obtain the maximum, median, and minimum mean values of the price paid by utilizing the Select Cases function in SPSS. Subsequently, review the district names in the Variable View window.

Research Question 2:

What are the most important features that determine the price of the property in this data?

Hypotheses:

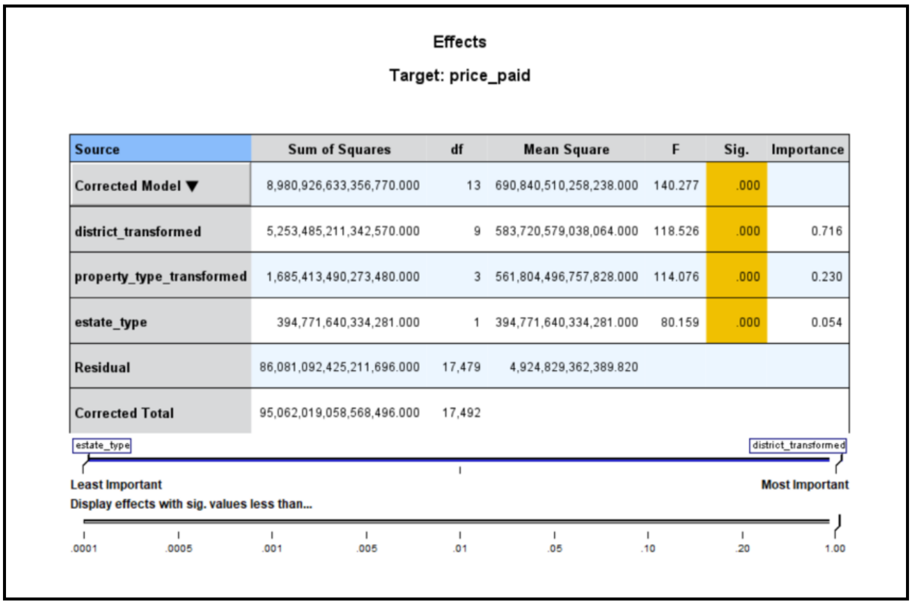
H0: There is no significant linear relationship between the price of the property (price_paid) and any of the features included in the dataset, indicating that none of the features contribute significantly to predicting the property price.

H1: There is a significant linear relationship between the price of the property (price_paid) and at least one of the features included in the dataset. This suggests that one or more features are important in determining the property price.

To address the second research question and the hypothesis, a regression method was employed to analyze 'price_paid' as the dependent variable and features ('property_type', 'estate_type', 'district', 'postcode') as independent variables. 'new_build' was considered as a feature, but due to data imbalance, where 99% of the values are 'no' and only 1% are 'yes', it was excluded from the analysis. Additionally, 'street' was not considered because a street can be grouped into different boroughs or postcodes in London, leading to potential fluctuations in prices. 'town' and 'country' were excluded from the independent features as they are both represented in a single category, Greater London.

According to (Yang), Automatic Linear Modeling (ALM) was employed to automatically build linear regression models, facilitating an exploration of the relationship between property prices and various features, including 'property_type,' 'estate_type,' 'district,' and 'postcode' (Figure 23).

Table 12: Hypothesis Test Result



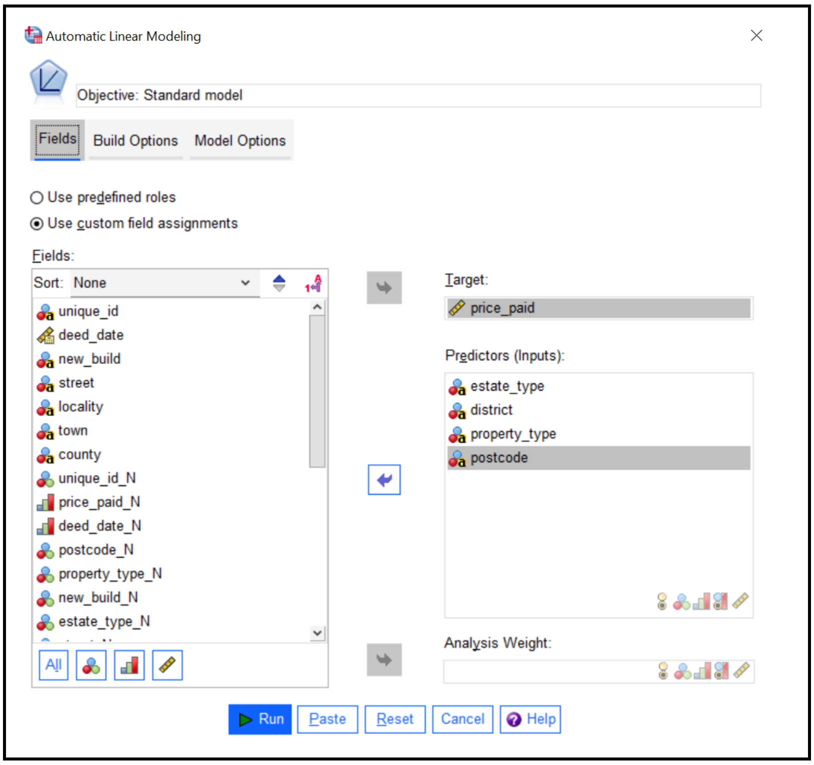
Effects

Target: price_paid

Source	Sum of Squares	df	Mean Square	F	Sig.	Importance
Corrected Model ▼	8,980,926,633,356,770.000	13	690,840,510,258,238.000	140.277	.000	
district_transformed	5,253,485,211,342,570.000	9	583,720,579,038,064.000	118.526	.000	0.716
property_type_transformed	1,685,413,490,273,480.000	3	561,804,496,757,828.000	114.076	.000	0.230
estate_type	394,771,640,334,281.000	1	394,771,640,334,281.000	80.159	.000	0.054
Residual	86,081,092,425,211,696.000	17,479	4,924,829,362,389.820			
Corrected Total	95,062,019,058,568,496.000	17,492				

To test hypotheses regarding the relationship between property prices (‘price_paid’) and the features, ALM proved to be a viable option. In ALM, statistical testing focused on assessing the significance of coefficients in the linear model, where each coefficient represents the contribution of a specific predictor variable to the prediction of the dependent variable Table 12).

The hypothesis test results (Table 12) indicate that in the Source of the Corrected Model, the sig (p-value) is 0.000. Therefore, the null hypothesis is rejected, suggesting a significant linear relationship between the price of the property and at least one of the features. Notably, the results reveal that 'district' plays the most crucial role in house prices, accounting for approximately 72%, followed by 'property_type' with just over 23%, and 'estate_type' with around 0.05%. However, it's important to note that the significance of these features does not necessarily correlate with the model's accuracy. The result was visualized in Figure 24.



Objective: Standard model

○ Use predefined roles

● Use custom field assignments

Fields:

Sort: None

unique_id
deed_date
new_build
street
locality
town
county
unique_id_N
price_paid_N
deed_date_N
postcode_N
property_type_N
new_build_N
estate_type_N

Target: price_paid

Predictors (Inputs):

estate_type
district
property_type
postcode

Analysis Weight:

Figure 23: Automatic Linear Modeling in SPSS – Exploring the Relationship Between Property Prices and Various Features

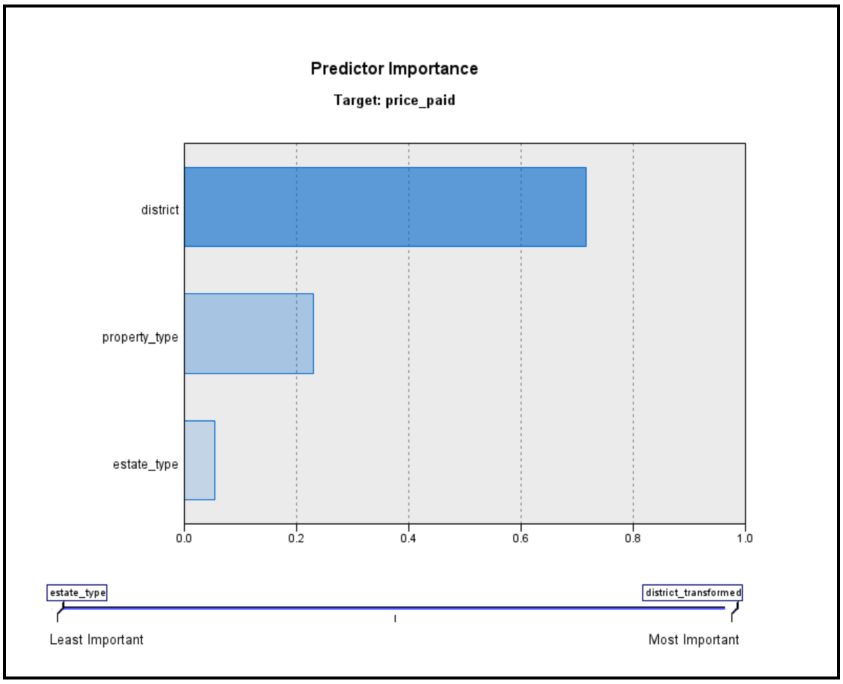
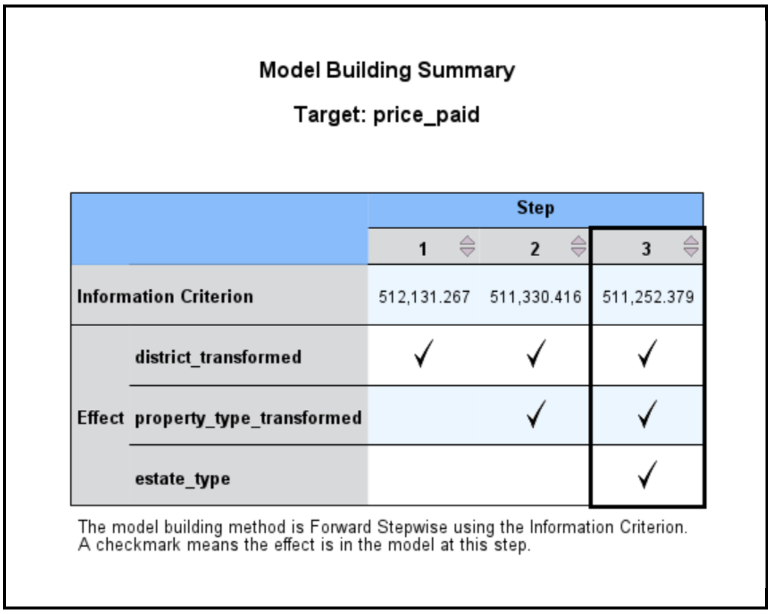


Figure 24: Histogram of Important Feature (‘district’, ‘property_type’ and ‘estate_type’)

The models were built using stepwise selection due to the limited number of predictors. Stepwise selection initiates with an empty model and adds predictors sequentially, selecting the variable that contributes the most at each step. At each stage, the model fit is assessed based on statistical criteria such as the F-statistic or p-value. The advantage lies in systematically considering predictors to maximize the improvement in model fit and potentially leading to a more parsimonious model compared to including all predictors (Yang). Notably, it may not consider all possible predictor combinations, possibly missing the best subset. Additionally, the final model depends on the order in which predictors are added. Figure 25 illustrates the forward stepwise selection process.



Model Building Summary

Target: price_paid

		Step		
		1	2	3
Information Criterion		512,131.267	511,330.416	511,252.379
Effect	district_transformed	✓	✓	✓
	property_type_transformed		✓	✓
	estate_type			✓

The model building method is Forward Stepwise using the Information Criterion.
A checkmark means the effect is in the model at this step.

Figure 25: Summary of Model Building

1.4 Evaluation and Conclusion

Research Questions:
1. Which London district offers properties with the highest, middle, and lowest prices?
2. What are the most important features that determine the price of the property in this dataset?

To address the research questions, an exploration of the dataset was conducted to gain a better understanding of the analysis and to explore variables descriptively. Following this, data cleaning procedures were implemented as part of the data preprocessing, encompassing the handling of missing values, removal of duplicated data, and addressing outliers in the price-paid variable. Subsequently, a test of normality revealed that the dataset is not normally distributed, leading to the adoption of nonparametric methods for hypothesis testing.

In response to the hypothesis test result (Kruskal-Wallis) for the first question, the null hypothesis was rejected, signifying a correlation between the price paid and different districts. Applying the aggregated method revealed, in Table 11, that the top district by mean price is the City of London, the middle district is Hackney, and the bottom district is Croydon. Transitioning to the second question, the result of the automatic linear model led to the rejection of the null hypothesis, indicating a significant linear relationship between the property price and at least one of the features. Figure 24 illustrates that the most influential feature in London house pricing in 2023 is the district.

This comprehensive report holds insights for diverse stakeholders in the real estate sector. Investors, developers, and policymakers can leverage the findings to make decisions on property investments, urban development, and policy formulation. The report's identification of significant features influencing property prices, particularly the pivotal role of districts, provides a foundation for predictive modeling and strategic planning. Additionally, the report serves as a benchmark for assessing property performance and offers educational value for students and professionals in real estate analytics. As the real estate market evolves, continuous monitoring and updates can ensure the report's ongoing relevance, allowing stakeholders to adapt to changing dynamics and make data-driven decisions in the future

2 References

Barbato, G., Barini, E.M., Genta, G. and Levi, R. (2011). Features and performance of some outlier detection methods. *Journal of Applied Statistics*, 38(10), pp.2133–2149. doi:https://doi.org/10.1080/02664763.2010.545119.

Chok, N.S. (2010). *Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data*. [online] d-scholarship.pitt.edu. Available at: https://d-scholarship.pitt.edu/8056/.

Kerr, A.W., Hall, H.K. and Kozub, S.A. (2003). *Doing statistics with SPSS*. London ; Thousand Oaks ; New Delhi: Sage.

Kwak, S.K. and Kim, J.H. (2017). Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*, [online] 70(4), p.407. doi:https://doi.org/10.4097/kjae.2017.70.4.407.

Leech, N.L., Karen Caplovitz Barrett and Morgan, G.A. (2008). *SPSS for intermediate statistics : use and interpretation*. New York: L. Erlbaum Associates.

Santos, L.L. and Jiang, R. (2020). *Spatial Analysis for House Price Determinants*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/340398763_Spatial_Analysis_for_House_Price_Determinants [Accessed 5 Dec. 2023].

Steinskog, D.J., Tjøstheim, D.B. and Kvamstø, N.G. (2007). A Cautionary Note on the Use of the Kolmogorov–Smirnov Test for Normality. *Monthly Weather Review*, 135(3), pp.1151–1157. doi:https://doi.org/10.1175/mwr3326.1.

www.ons.gov.uk. (2023). *UK House Price Index - Office for National Statistics*. [online] Available at: https://www.ons.gov.uk/economy/inflationandpriceindices/bulletins/housepriceindex/october2023#house-prices-by-region-in-england [Accessed 20 Dec. 2023].

Yang, H. (2013). Automatic Linear Modeling Multiple Linear Regression Viewpoints. [online] 39(2), p.27. Available at: https://www.statwks.com/wp-content/uploads/2018/11/Yang-39_2_proof_27.pdf.