Northumbria University NEWCASTLE

Newcastle • London • Amsterdam

COURSEWORK COVER PAGE	
Module Number:	LD7188
Module Title:	Applied Data Science
Module Tutor Name:	Dilek Celik [DC]
Coursework Title:	**Assignment**
Word count:	3286
Student Name:	Rasikh Thakur
Student ID:	W23042021
Programme of Study:	MSc Big Data & Data Science Technology [BD]

Submission Instructions

1. Name your submission in according to the name convention, <module number>_<tutor initial>_< prog code>_<assignment group>_<your id><first name>.docx, eg LD7087_NT_BD_A_w22012345John.docx is the LD7087 assignment submitted from a group A student enrolled in MSc Big Data & Data Sciences, attending Ning Tse's session.
2. Submit to Final Submission Link on Bb **before 16:00, 23 Jan 2024**

Declaration

I confirm that this assessment is my own work and that I have duly acknowledged and correctly referenced the work of others. I am aware of and understand that any breaches to the Code of Academic Conduct will be investigated and sanctioned in accordance with the Academic Conduct Regulation.

Your signature:	Rasikh Sadiq Thakur	**Date:**	21/01/2024

Table Of Content:

Task one: Domain Understanding and Research Questions

The data set is relevant to the area of real estate and housing market research, focusing on residential property transactions in London. This subject encompasses the research fields housing market dynamics, urban studies and real estate economics. Scholars in this field discuss several aspects, including property prices, modalities of transactions and product types as well as factors influencing housing market. The dataset contributes to the knowledge on dynamics occurring in real estate and provides valuable data for analysis of patterns, as wells trend forms within London' housing markets.

Paper studied:

First Paper:

Paper Name: Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism

Authors: PEI-YING WANG1 , CHIAO-TING CHEN 2 , JAIN-WUN SU1 , TING-YUN WANG1 , AND SZU-HAO HUANG 3 , (Member, IEEE) 1 Institute of Information Management, National Chiao Tung University, Hsi

Analysis:

Research Problem:

This work tackles the issue of inaccurate house price forecasts in previous research because of incomplete data. It emphasises how crucial it is to take into account both the environmental aspects of the immediate area and the home itself.

Proposed Solution:

An end-to-end joint self-attention model for predicting property prices is put out by the authors. The model uses satellite maps to analyse the surrounding area and integrates data on public amenities (parks, schools, etc.). Attention methods are used to pinpoint important aspects that potential customers are thinking about.

Important Elements of the Model:

An attention mechanism and a spatial transformer network (STN) are employed by the model to manage heterogeneous data, such as data from public facilities and satellite maps. From satellite maps, the STN extracts rotation-invariant picture characteristics.

Information Used:

The experimental data consists of satellite maps acquired via the Google Maps API, actual selling prices from real estate transactions in 2017 and 2018, and data from Taipei and New Taipei governments on public facilities.

Comparing This Model with Others:

The performance of the suggested model is contrasted with that of several machine learning-based models, including deep learning, numerous attention models, Extreme Gradient Boosting, and Light Gradient Boosted Machine.

Results of the experiment:

The experimental findings show that the suggested model works better than other models and achieves a low prediction error. According to the authors, this model is the first to predict home prices using both a STN network and an attention mechanism.

Contributions:

The primary innovations include the inclusion of a joint self-attention mechanism taking into account two-hop correlations between various variables, the incorporation of heterogeneous data, and the use of STN for picture feature extraction.

Novelty:

The research highlights its innovation by asserting that it is the first to combine STN network and attention processes for predicting housing prices.

Second Paper:

Paper Name: Boston House Price Prediction Using Regression Models

Authors: Saptarsi Sanyal, Saroj Kumar Biswas, Dolly Das, Manomita Chakraborty, Biswajit Purkayastha

Analysis:
Use machine learning algorithms to give reliable predictions on the housing values. The researchers used Ridge Regression, Lasso Regression, Polynomial regressions and Simple Linear Regression with the Boston Housing dataset. The models were evaluated using metrics such as R-Squared, Root Mean Square Error (RMSE), and Cross Validation.

Model Execution:
Among the models, Lasso Regression performed better than all others achieving maximum accuracy in R – Squared and Cross-Validation.
Ridge Regression did well also; and it had the second highest accuracy.
A two regression of polynomial degree outperformed simple linear regressions; however, it did not perform as well as a Lasso and Ridge Regression.

Pre-processing of Data:
The application of good prediction techniques such as outlier removal and log transformation requires data pre-processing procedures.
Strong correlations between features like "TAX" and "RAD" point to the need to eliminate unnecessary characteristics in order to increase model efficiency.

Future work:
By gathering more current data, the research may be expanded to take into consideration shifts in the real estate market over time.
Further insights may be obtained by evaluating other machine learning models, such as ensemble approaches, decision trees, and support vector machines.
To improve model performance, feature selection strategies and optimisation techniques, such as deep learning approaches, might be investigated.

In summary:

The paper's models can be used as a starting point to forecast home prices.

To increase forecast accuracy, further study and advancements—including the use of cutting-edge models and methods—are necessary.

Research Questions:

1. Do the average home prices in London's boroughs differ much from one another? Examining regional differences in housing costs is one such area of investigation.

   Null Hypothesis (H0):
   The average cost of a home in each given London borough does not differ much from one another.
   Theoretically: H0: μ_borough1 = μ_borough2 =... = μ_boroughN

   Alternate hypothesis (Ha):
   There exists a notable disparity in the mean property values between a minimum of two London boroughs.
   Symbolically: Ha: Among the μ_boroughs, at least one is not like the rest.

2. Does the kind of property (home, flat, etc.) affect the London real estate market price in a statistically meaningful way?
   Examining if property type affects pricing might be one area of study.

   Null Hypothesis (H0):
   The type of property in London has no discernible impact on the selling price.
   H0: μ_apartment = μ_house, symbolically

   Alternate hypothesis (Ha):
   The type of property in London has a substantial impact on the selling price.
   Ha: μ_apartment ≠ μ_house, symbolically

3. Does the kind of estate—freehold or leasehold—have a big influence on how much a London property sells for?

   Null Hypothesis (H0):

The selling price of homes does not significantly change depending on whether they are leasehold or freehold.

H0: μ_freehold = μ_leasehold, symbolically

Alternate hypothesis (Ha):

The selling price of houses varies significantly depending on whether they are leasehold or freehold.

Ha: μ_freehold ≠ μ_leasehold, symbolically

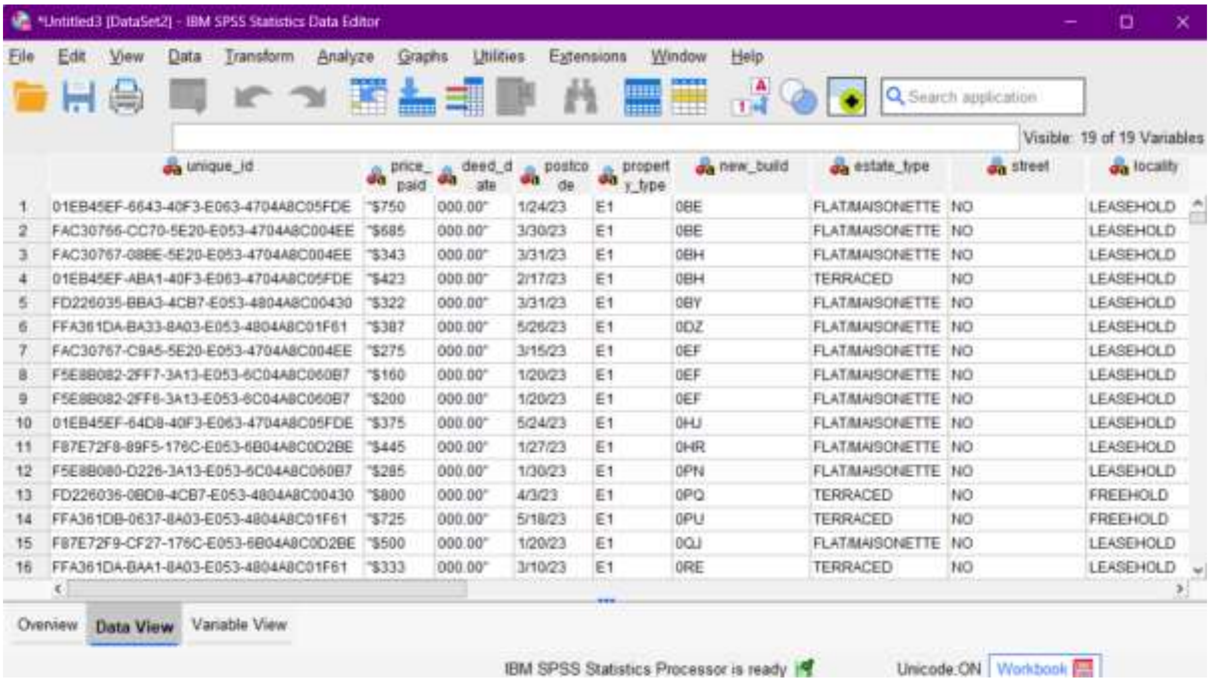
References:

IEEE Xplore Full-Text PDF: (no date). https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9395585.

https://www.researchgate.net/publication/362812590_Boston_House_Price_Prediction_Using_Regression_Models

Task two: Dataset and Data Preparation

First loaded the dataset to the SPSS

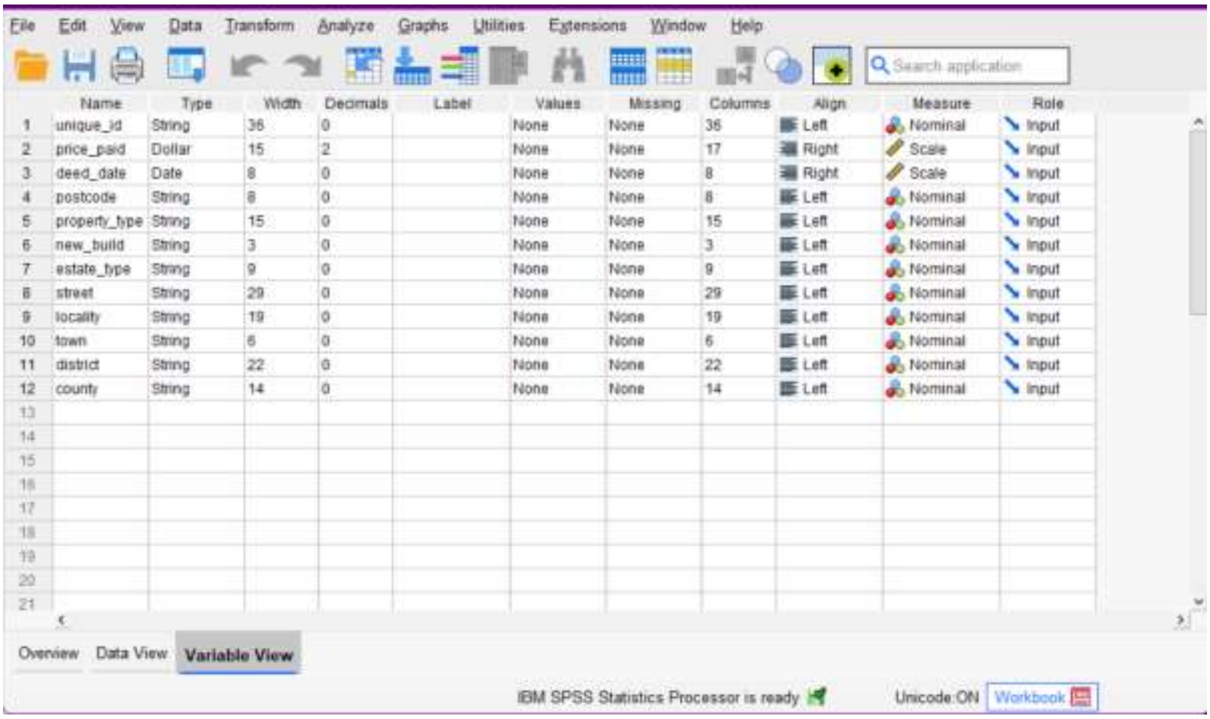


	unique_id	price_paid	deed_date	postcode	property_type	new_build	estate_type	street	locality
1	01EB45EF-6643-40F3-E063-4704A8C05FDE	"$750	000.00"	1/24/23	E1	0BE	FLAT/MAISONETTE	NO	LEASEHOLD
2	FAC30766-CC70-5E20-E053-4704A8C004EE	"$685	000.00"	3/30/23	E1	0BE	FLAT/MAISONETTE	NO	LEASEHOLD
3	FAC30767-08BE-5E20-E053-4704A8C004EE	"$343	000.00"	3/31/23	E1	0BH	FLAT/MAISONETTE	NO	LEASEHOLD
4	01EB45EF-ABA1-40F3-E063-4704A8C05FDE	"$423	000.00"	2/17/23	E1	0BH	TERRACED	NO	LEASEHOLD
5	FD226035-BBA3-4CB7-E053-4804A8C00430	"$322	000.00"	3/31/23	E1	0BY	FLAT/MAISONETTE	NO	LEASEHOLD
6	FFA361DA-BA33-8A03-E053-4804A8C01F61	"$387	000.00"	5/26/23	E1	0DZ	FLAT/MAISONETTE	NO	LEASEHOLD
7	FAC30767-C9A5-5E20-E053-4704A8C004EE	"$275	000.00"	3/15/23	E1	0EF	FLAT/MAISONETTE	NO	LEASEHOLD
8	F5E8B082-2FF7-3A13-E053-6C04A8C060B7	"$160	000.00"	1/20/23	E1	0EF	FLAT/MAISONETTE	NO	LEASEHOLD
9	F5E8B082-2FF6-3A13-E053-6C04A8C060B7	"$200	000.00"	1/20/23	E1	0EF	FLAT/MAISONETTE	NO	LEASEHOLD
10	01EB45EF-64D8-40F3-E063-4704A8C05FDE	"$375	000.00"	5/24/23	E1	0HJ	FLAT/MAISONETTE	NO	LEASEHOLD
11	F87E72F8-89F5-176C-E053-6B04A8C0D2BE	"$445	000.00"	1/27/23	E1	0HR	FLAT/MAISONETTE	NO	LEASEHOLD
12	F5E8B080-D226-3A13-E053-6C04A8C060B7	"$285	000.00"	1/30/23	E1	0PN	FLAT/MAISONETTE	NO	LEASEHOLD
13	FD226036-0BD8-4CB7-E053-4804A8C00430	"$800	000.00"	4/3/23	E1	0PQ	TERRACED	NO	FREEHOLD
14	FFA361DB-0637-8A03-E053-4804A8C01F61	"$725	000.00"	5/18/23	E1	0PU	TERRACED	NO	FREEHOLD
15	F87E72F9-CF27-176C-E053-6B04A8C0D2BE	"$500	000.00"	1/20/23	E1	0QJ	FLAT/MAISONETTE	NO	LEASEHOLD
16	FFA361DA-BAA1-8A03-E053-4804A8C01F61	"$333	000.00"	3/10/23	E1	0RE	TERRACED	NO	LEASEHOLD

Now after loading the dataset in the SPSS, exploring the dataset.

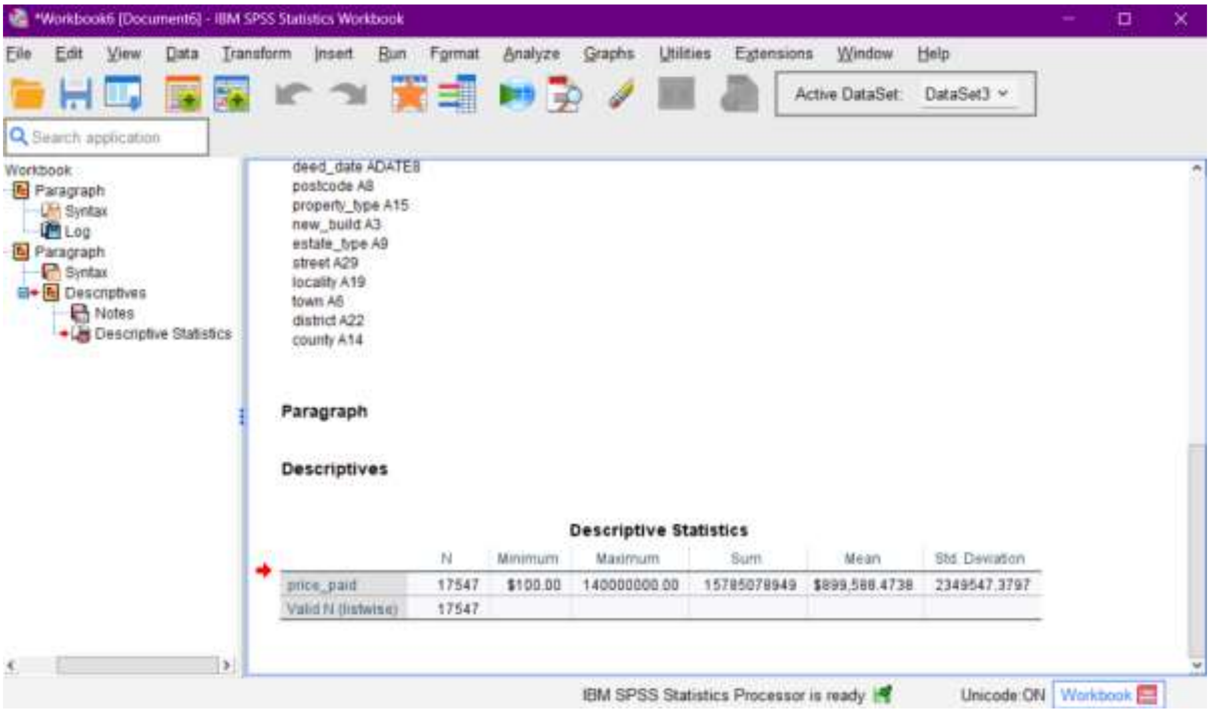
Now in the SPSS, clicked on the variable view where we can see the information about each column in the dataset.

As we can see in the below image, the dataset every variable name, variable type, width, Decimals, label, Values, Missing, Columns, Align, measure and Role.

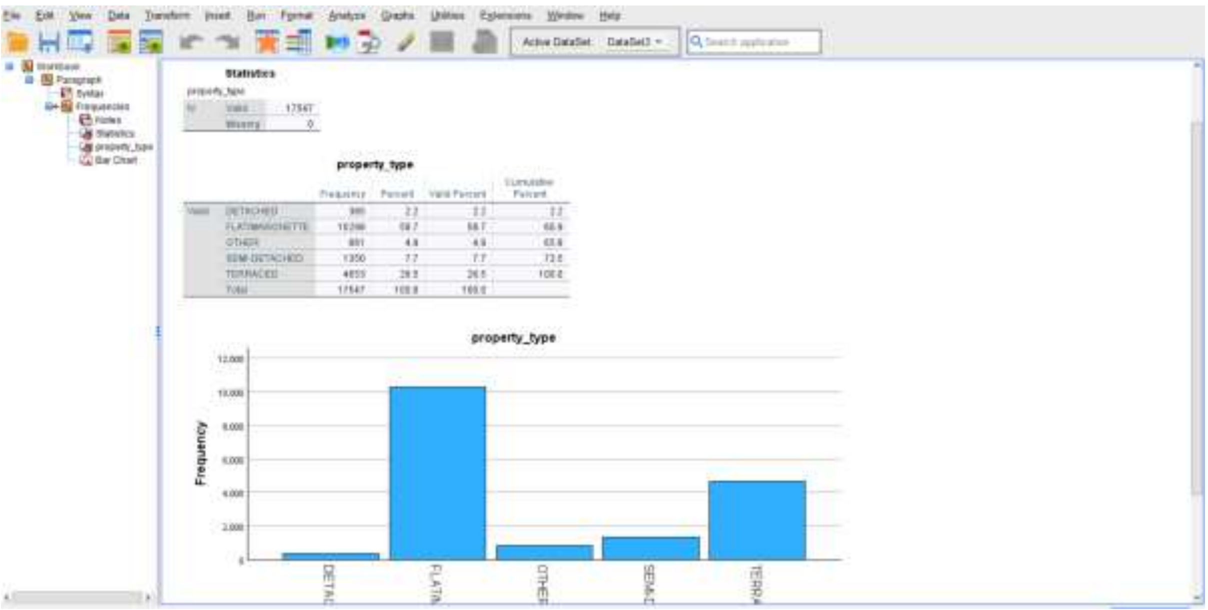


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	unique_id	String	36	0		None	None	36	Left	Nominal	Input
2	price_paid	Dollar	15	2		None	None	17	Right	Scale	Input
3	deed_date	Date	8	0		None	None	8	Right	Scale	Input
4	postcode	String	8	0		None	None	8	Left	Nominal	Input
5	property_type	String	15	0		None	None	15	Left	Nominal	Input
6	new_build	String	3	0		None	None	3	Left	Nominal	Input
7	estate_type	String	9	0		None	None	9	Left	Nominal	Input
8	street	String	29	0		None	None	29	Left	Nominal	Input
9	locality	String	19	0		None	None	19	Left	Nominal	Input
10	town	String	6	0		None	None	6	Left	Nominal	Input
11	district	String	22	0		None	None	22	Left	Nominal	Input
12	county	String	14	0		None	None	14	Left	Nominal	Input

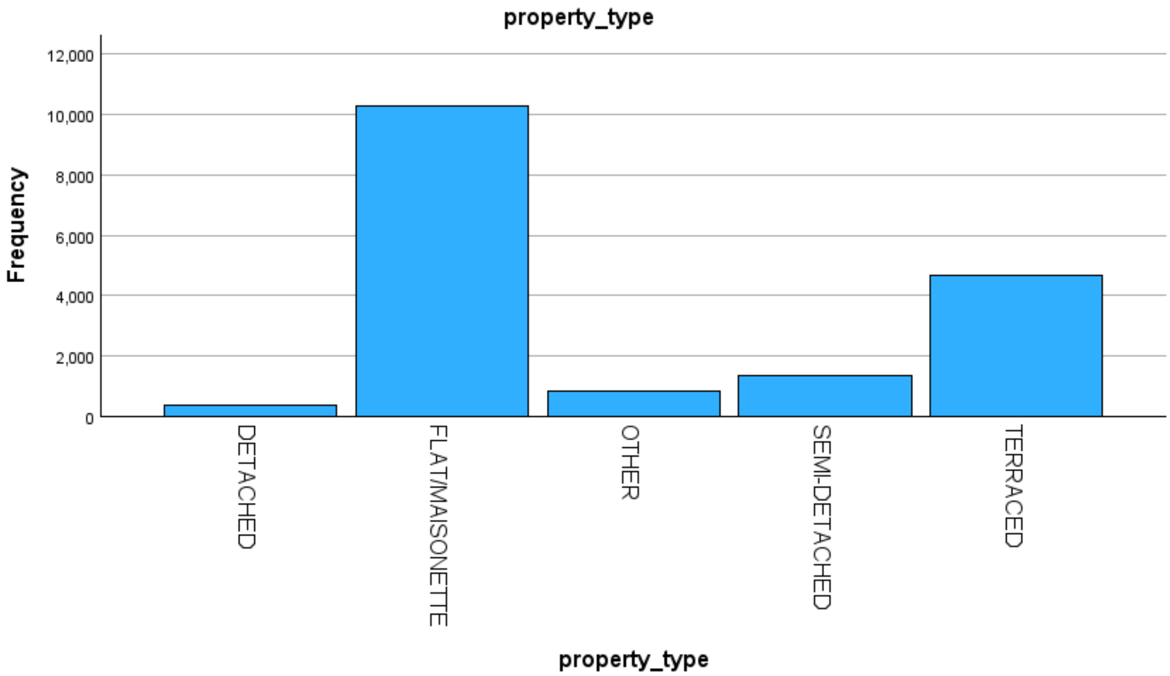
Then calculated the descriptive statistics. So from the above when explored the dataset we can see that of the variable price_paid as its type is dollar so we can calculate the descriptives i.e Mean, median, Sum ,Standard Deviation, Maximum, Total Count (N).



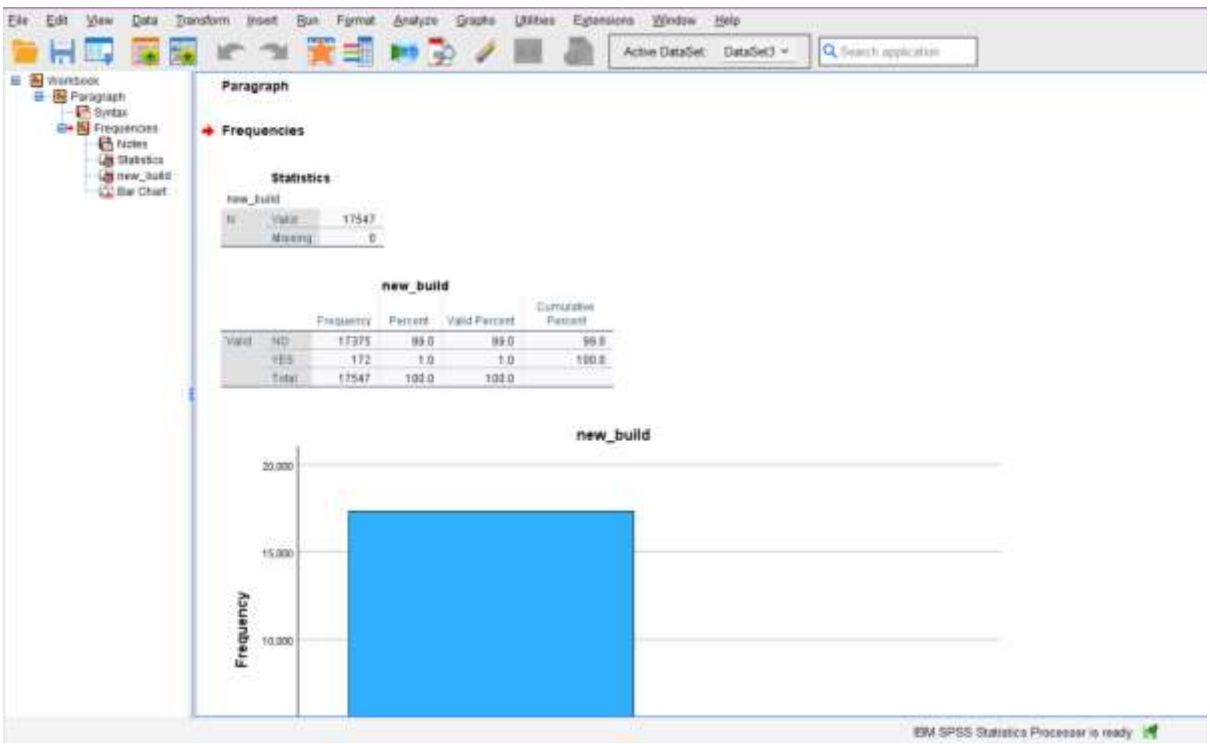
And then for the variable property_type we are calculating the statistics, then for the property_type variable where we have DETACHED, FLAT/MAISONETTE, OTHER,SEMI-DETACHED, TERRACED. So for each of this shows the Frequency, Percent, valid Percent, and Cumulative Percent.

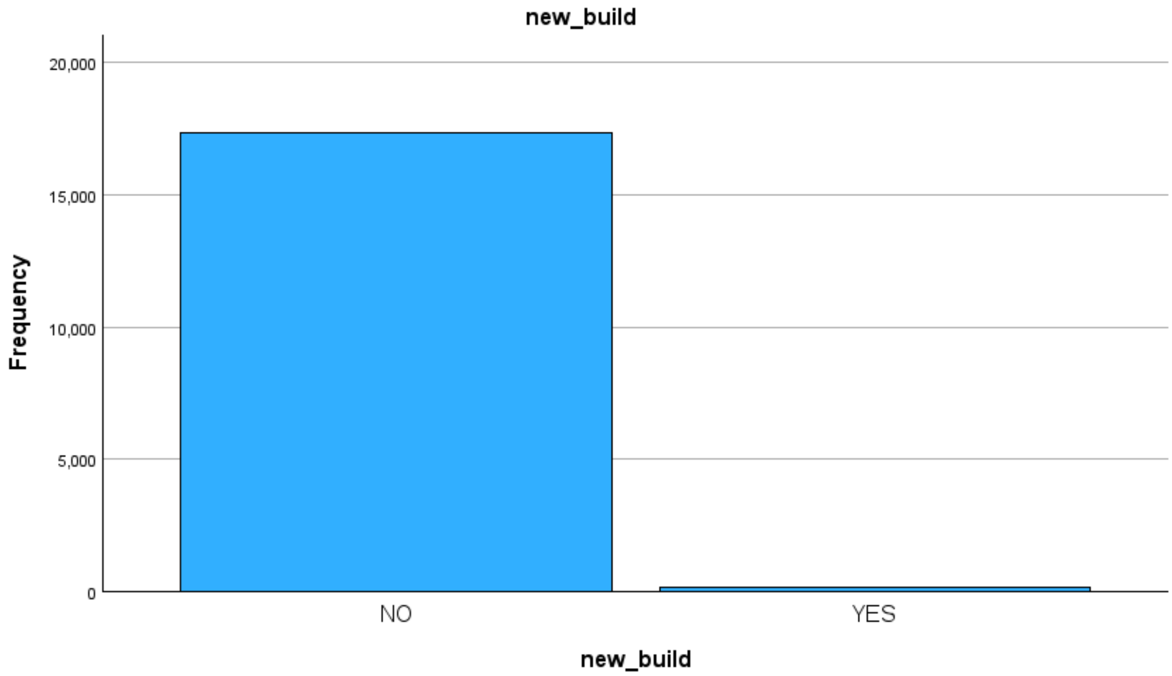


Also plotted the bar chart for the property_type variable.



Similarly for the new_build variable also we are plotting the bar chart and calculated the statistics, Frequency, percent, Valid Percent, Cumulative Percent.





Similarly, for esate_type variable we have calculated the statistics and Frequency, Percent, valid percent, and Cumulative Percent for both the type i.e for FREEHOLD and LEASEHOLD of the estate_type.

Statistics

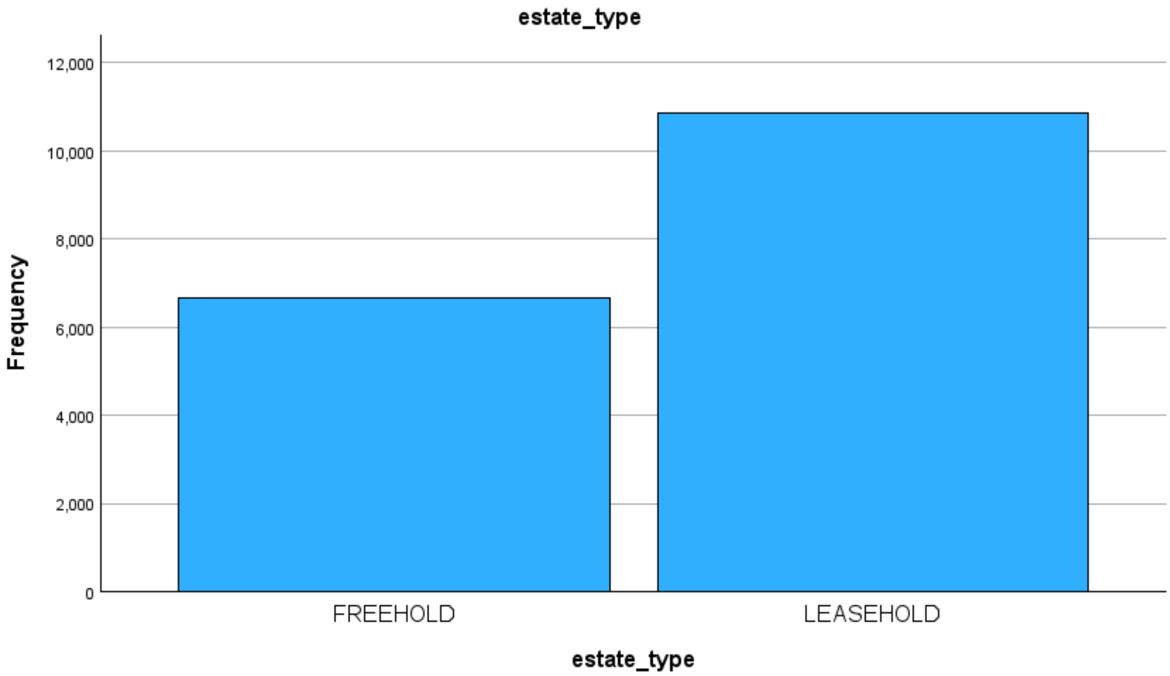
estate_type

N	Valid	17547
	Missing	0

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	

And also plotted the bar chart for estate_type.



Task three: Data Analytical Methods

So now for first research question:

Do the average home prices in London's boroughs differ much from one another?

Descriptives

price_paid

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
01/01/23	1	3534503.0000	.	.	.	.	$3,534,503.00	$3,534,503.00
01/03/23	38	1797471.7105	4807332.8013	779852.35585	$217,340.7387	3377602.6823	$209,750.00	28000000.00
01/04/23	111	$789,434.2703	979113.55259	$93,441.76575	$623,972.2759	$953,796.2647	$169,000.00	$7,900,000.00
01/05/23	115	1092165.6087	1921269.5910	179158.28469	$617,554.4269	1357379.7909	$13,000.00	14500000.00
01/06/23	302	$787,519.8113	1093742.6437	$62,937.81099	$663,662.9701	$911,370.6524	$101,693.00	16200000.00
01/07/23	1	$800,000.0000	.	.	.	.	$800,000.00	$800,000.00
01/09/23	151	$848,211.0397	830791.57482	$67,631.53544	$714,636.8055	$981,785.2740	$132,000.00	$5,325,000.00
01/10/23	112	$713,462.3750	504444.57133	$47,665.53164	$618,039.6430	$807,934.8080	$100.00	$3,062,500.00
01/11/23	100	$839,678.0000	1152109.0113	115210.90113	$611,073.1722	1068279.8278	$4,000.00	$9,169,300.00
01/12/23	163	1245900.0773	4972574.7143	389490.94344	$476,869.9375	2014932.0171	$3,000.00	63497431.00
01/13/23	239	$843,948.2847	2027990.0903	131430.03971	$585,017.8243	1102880.0051	$4,000.00	26000000.00
01/14/23	1	2850000.0000	.	.	.	.	$2,850,000.00	$2,850,000.00
01/15/23	3	$321,666.6667	235283.51683	135841.00193	-262809.9907	$906,143.3240	$110,000.00	$575,000.00
01/16/23	169	$930,228.6639	1122788.6957	$85,399.20787	$749,720.0273	1090733.7005	$5,000.00	11160000.00
01/17/23	86	$710,943.0698	577076.86761	$62,227.71375	$587,214.7055	$834,665.4340	$15,000.00	$3,750,000.00
01/18/23	109	$817,124.6165	913373.99300	$87,486.35230	$643,713.6710	$990,535.9620	$20,000.00	$6,000,000.00
01/19/23	107	$758,729.0561	724372.89263	$70,027.77064	$630,893.1998	$898,565.9224	$4,000.00	$4,850,000.00
01/20/23	356	1021883.7135	1999200.5725	105957.41843	$813,500.5547	1230266.8722	$28,000.00	20865187.00
01/23/23	137	2330862.3822	13955798.901	1149585.9163	$56,669.0376	4605898.6968	$125,000.00	140000000.00
01/24/23	76	$752,359.1447	761635.83118	$85,687.55754	$573,891.8458	$930,845.3397	$108,746.00	$5,825,000.00
01/25/23	105	$790,724.6566	964286.24342	$93,857.91099	$605,017.9089	$976,430.2066	$130.00	$5,900,000.00
01/26/23	93	$966,185.3871	1893307.4059	196296.83459	$576,321.6577	1356054.8165	$130.00	18000000.00
01/27/23	399	$826,603.1959	1141205.6915	$57,293.49203	$734,143.5745	$939,062.7972	$32,000.00	13000000.00
01/28/23	2	$479,278.0000	242184.37299	171228.00000	-1696271.591	2654827.5910	$307,050.00	$650,500.00
01/30/23	143	$907,069.9021	2178478.3578	182256.96902	$546,802.3221	1267377.4821	$610.00	24700000.00
01/31/23	261	$939,913.2464	1214215.5913	$75,159.09304	$818,915.9100	1114669.8827	$25,000.00	10700000.00
02/01/23	114	$826,597.3333	823152.15306	$77,091.66415	$674,043.2875	$979,151.3891	$2,000.00	$5,400,000.00
02/02/23	115	$995,165.6207	2612961.2819	242623.00154	$514,792.6457	1475579.5957	$100,000.00	26000000.00
02/03/23	278	$677,283.6504	559237.27087	$33,091.03849	$612,288.6595	$742,219.2413	$5,000.00	$4,065,000.00
02/06/23	90	$811,515.5111	936536.07733	$98,719.46510	$615,363.4367	1007670.3855	$122,000.00	$6,200,000.00
02/07/23	83	$901,825.1429	1673269.6765	213606.77005	$480,422.4170	1323227.8687	$27,000.00	13500000.00
02/08/23	86	$884,565.2500	1293848.1261	137824.78027	$590,424.8145	1138785.6855	$138,000.00	$8,250,000.00
02/09/23	94	$806,542.5532	1149992.5750	118611.64607	$571,003.3312	1042081.7752	$8,000.00	10410000.00
02/10/23	320	$794,386.6438	1090603.6283	$60,966.54033	$674,409.3417	$914,363.9458	$13,000.00	16000000.00
02/11/23	1	$755,000.0000	.	.	.	.	$755,000.00	$755,000.00
02/13/23	120	$889,817.6587	1587267.1487	139617.90892	$713,499.4881	1266138.8294	$113,750.00	14700000.00
02/14/23	95	$826,103.9474	861552.57810	$82,237.55932	$662,819.3455	$989,388.5492	$145,250.00	$6,560,000.00
02/15/23	100	$777,050.1500	769673.58197	$76,967.35819	$625,129.6349	$930,171.6651	$140,000.00	$5,609,000.00
02/16/23	102	$788,336.3039	713610.41565	$70,331.50715	$648,975.0236	$928,793.5853	$120,000.00	$3,780,000.00
02/17/23	261	$669,497.5594	464311.01203	$27,692.50729	$614,965.6304	$724,009.5004	$3,000.00	$3,425,000.00
02/18/23	1	1925000.0000	.	.	.	.	$1,925,000.00	$1,925,000.00
02/20/23	135	$756,967.3037	689327.43067	$60,198.52205	$638,924.8945	$878,009.7130	$13,000.00	$4,300,000.00
02/21/23	69	$939,924.4928	2118949.5289	255055.43092	$429,978.2750	1447899.6530	$41,750.00	17500000.00
02/22/23	83	$656,234.1368	475681.69942	$49,316.52492	$558,289.2505	$754,179.0230	$1,000.00	$2,660,000.00
02/23/23	101	$791,669.6020	745612.33592	$74,191.26017	$644,493.4549	$938,980.1491	$29,000.00	$4,851,750.00
02/24/23	337	$793,919.0000	811186.38475	$44,196.72230	$696,982.8381	$870,827.4649	$5,100.00	$6,650,000.00
02/25/23	1	$210,500.0000	.	.	.	.	$210,500.00	$210,500.00
02/27/23	133	$895,203.4586	1251984.6073	108526.09764	$690,528.1118	1109878.8055	$90,000.00	10750000.00
02/28/23	249	$953,309.4819	2241369.6799	142040.39200	$673,640.1840	1233158.7798	$30,000.00	33396000.00
03/01/23	101	$807,359.2444	844306.79329	$69,798.71240	$671,037.4333	$943,181.2555	$50,000.00	$7,750,000.00
03/02/23	71	$705,455.9718	502364.36164	$59,619.88161	$586,589.2874	$824,404.6762	$21,000.00	$3,300,000.00
03/03/23	338	$754,255.5462	732070.50025	$47,453.08264	$658,911.8335	$847,779.3580	$33,250.00	$8,300,000.00
03/04/23	1	$237,000.0000	.	.	.	.	$237,000.00	$237,000.00
03/05/23	1	$417,500.0000	.	.	.	.	$417,500.00	$417,500.00
03/06/23	120	$770,135.6163	754070.57979	$67,717.98061	$636,082.9391	$904,190.1416	$110,750.00	$6,750,000.00
03/07/23	91	$638,889.9231	562637.22309	$59,001.37265	$521,819.3982	$756,052.4480	$10,000.00	$1,900,000.00
03/08/23	80	1376013.9535	2186838.8579	235791.33204	$907,197.8151	1844830.2919	$122,000.00	11977000.00
03/09/23	95	$825,268.0211	2604875.8290	268974.31786	$316,037.4860	1335494.5561	$1,000.00	24668123.00
03/10/23	241	$717,738.7718	675663.56369	$43,608.34675	$631,834.7928	$803,642.7510	$3,500.00	$5,400,000.00
03/12/23	1	$104,500.0000	.	.	.	.	$104,500.00	$104,500.00
03/13/23	86	$754,521.0698	664996.78427	$71,705.21344	$612,091.8971	$897,190.2426	$50,000.00	$3,500,000.00
03/14/23	79	1027808.6203	1151774.5062	129584.75547	$769,829.1887	1285783.0589	$51,000.00	$6,500,000.00
03/15/23	131	$888,268.1221	1650471.0721	144202.33711	$702,972.0938	1173546.1905	$100,000.00	16500000.00
03/16/23	98	$738,901.6429	665645.70182	$66,361.53723	$607,192.3613	$870,610.9245	$10,000.00	$3,950,000.00
03/17/23	359	$723,302.5073	583326.16802	$30,246.08232	$651,937.1112	$794,678.7035	$20,000.00	$4,100,000.00
03/18/23	1	$990,000.0000	.	.	.	.	$990,000.00	$990,000.00
03/19/23	1	1500000.0000	.	.	.	.	$1,500,000.00	$1,500,000.00
03/20/23	125	1131320.6000	4618268.1134	413098.65261	$313,440.8488	1948881.1582	$80,000.00	93300000.00
03/21/23	89	1022184.6477	1956897.2291	208499.34518	$607,769.6225	1436599.6729	$100,000.00	18180000.00
03/22/23	98	$826,598.2083	990032.27825	100326.59482	$615,190.4531	1038001.9655	$4,000.00	$6,031,574.00
03/23/23	111	$923,244.6036	1193524.6787	104647.03961	$716,888.7330	1136630.4722	$65,000.00	$7,325,000.00
03/24/23	271	$883,909.1771	2143751.4544	130223.62682	$627,425.3302	1140191.0240	$100.00	32350000.00
03/25/23	2	$592,500.0000	342846.78888	242500.00000	-2488754.649	3673754.6488	$350,000.00	$835,000.00
03/26/23	1	$770,000.0000	.	.	.	.	$770,000.00	$770,000.00
03/27/23	139	$685,981.5077	803271.75205	$82,810.43280	$596,987.3200	$770,366.4946	$10,000.00	$6,100,000.00
03/28/23	91	2086962.6714	9631378.4647	969158.74210	$130,967.3487	4078967.7932	$12,000.00	88369696.00
03/29/23	119	1397878.2941	4501657.7074	412669.95225	$580,771.4398	2215189.1487	$2,000.00	46000000.00
03/30/23	147	$904,742.6365	1663623.3246	133614.86288	$646,603.8412	1162882.8377	$10,000.00	17360000.00
03/31/23	502	$939,070.4601	1549396.1536	$69,152.93463	$803,204.9956	1074935.9646	$15,000.00	18000000.00
04/01/23	1	$660,500.0000	.	.	.	.	$660,500.00	$660,500.00
04/02/23	1	1150000.0000	.	.	.	.	$1,150,000.00	$1,150,000.00
04/03/23	150	$886,118.1267	1234327.6029	100782.37776	$686,970.8201	1085265.4332	$100,000.00	12500000.00
04/04/23	119	1112268.6303	3126741.2273	286636.22862	$544,836.8885	1679678.3720	$1,637.00	33500000.00
04/05/23	150	$703,057.6733	434972.79356	$35,515.37987	$632,879.0147	$773,236.7320	$74,000.00	$2,550,000.00
04/06/23	179	$723,117.4160	843281.36806	$62,805.57733	$599,178.0930	$847,056.7490	$55,000.00	$6,520,000.00
04/09/23	1	$369,354.0000	.	.	.	.	$369,354.00	$369,354.00
04/11/23	68	$809,807.7059	1015571.5598	123158.31980	$563,762.7156	1055452.6962	$160,000.00	$8,100,000.00
04/12/23	99	$693,872.0404	428650.37128	$43,070.93289	$608,369.1759	$779,344.9050	$120,000.00	$2,100,000.00
04/13/23	62	1387840.7317	3687119.3480	404985.58590	$582,086.2440	2193595.2194	$142,500.00	32580000.00
04/14/23	240	$889,809.6867	918646.91800	$59,291.94860	$773,008.1207	1006611.2127	$30,000.00	$7,500,000.00
04/15/23	1	$640,500.0000	.	.	.	.	$640,500.00	$640,500.00
04/17/23	96	$848,776.6625	1147133.2988	117078.90214	$616,345.7531	1081207.3719	$130,000.00	10250000.00
04/18/23	66	$750,969.3333	622602.96870	$76,583.26543	$598,060.8336	$903,865.7331	$10,000.00	$4,100,000.00
04/19/23	79	$861,462.3924	1169459.9997	131574.19211	$599,519.2977	1123406.4871	$260,000.00	10260000.00
04/20/23	87	$763,608.1379	687836.21218	$73,754.39641	$586,389.2504	$888,827.0254	$69,234.00	$6,300,000.00
04/21/23	239	$822,223.6333	1397963.4930	$92,687.36386	$638,783.0486	1004664.6180	$29,250.00	26000000.00
04/24/23	89	$882,317.4157	970288.28234	102850.46929	$677,933.7404	1086711.0911	$150,000.00	$6,000,000.00
04/25/23	67	$803,926.6866	609606.41748	$74,476.64381	$655,231.4601	$952,622.3011	$300,000.00	$3,409,001.00
04/26/23	64	$794,963.2188	896393.59634	112049.19954	$541,050.7605	$998,875.6770	$82,000.00	$5,550,000.00
04/27/23	89	1252226.2360	4426162.3413	469171.20963	$319,847.0459	2184605.4261	$2,000.00	42360000.00
04/28/23	360	$845,015.6361	834699.21514	$43,987.24069	$758,519.5946	$931,520.0774	$1,000.00	$6,500,000.00
05/02/23	88	$888,480.6510	838881.71084	$84,537.91968	$720,698.3604	1056264.7216	$152,000.00	$6,200,000.00
05/03/23	67	1028410.4478	1613069.4717	197067.80167	$634,961.8880	1421868.0076	$101,500.00	10500000.00
05/04/23	80	$752,326.1125	561839.92157	$62,815.60167	$627,294.7849	$877,357.4401	$70,000.00	$3,200,000.00
05/08/23	208	1035821.1871	2476033.4660	171681.62317	$697,362.1323	1374280.2619	$1,000.00	32500000.00
05/09/23	65	1224268.9385	3501969.5107	434316.72560	$356,637.5409	2091940.3360	$75,000.00	26500000.00
05/10/23	92	$962,842.6217	1287208.1985	134200.73164	$696,269.2643	1229415.7882	$150,000.00	$9,000,000.00
05/11/23	82	$962,977.6366	1431128.6168	158041.73925	$648,524.0495	1177431.0237	$109,000.00	12000000.00
05/12/23	152	$782,472.6510	785326.02782	$68,676.02420	$670,881.3476	$894,263.9545	$140,000.00	$8,500,000.00
05/13/23	2	$578,750.0000	111369.31804	$78,750.00000	-421863.6230	1579363.6230	$500,000.00	$657,500.00
05/15/23	102	$827,945.3627	944137.95959	$93,483.56009	$642,499.0700	1013391.6547	$95,000.00	$6,700,000.00
05/16/23	74	$788,032.6243	1339165.0349	155674.77318	$477,773.4469	1098282.1988	$150,000.00	11750000.00
05/17/23	92	$751,508.5978	805190.26092	$84,051.14494	$584,551.3237	$918,465.8719	$2,000.00	$6,690,000.00
05/18/23	73	$920,780.7945	1276251.1567	149373.89949	$623,009.3081	1218552.2210	$104,000.00	$9,325,000.00
05/19/23	225	$690,431.6933	581177.64476	$38,745.17632	$614,021.2461	$767,842.1406	$122,500.00	$4,850,000.00
05/22/23	90	$729,110.3322	520353.15751	$54,850.03985	$614,124.4634	$833,095.1811	$100.00	$3,325,000.00
05/23/23	80	$700,373.6000	889888.32338	125379.86469	$448,410.6703	$952,339.8287	$150,000.00	$8,400,000.00
05/24/23	90	$910,474.5714	902984.59460	$91,105.61196	$749,502.3929	1071446.7600	$195,000.00	$6,425,000.00
05/25/23	131	$892,478.4159	892819.03909	$78,035.96030	$738,190.9561	1016801.8836	$135,000.00	$6,853,000.00
05/26/23	266	$767,836.6267	744661.67105	$44,026.86457	$681,177.7757	$854,495.8817	$91,875.00	$7,175,000.00
05/28/23	1	$525,000.0000	.	.	.	.	$525,000.00	$525,000.00
05/29/23	2	$420,000.0000	$7,071.06781	$5,000.00000	$356,469.8763	$483,531.0237	$415,000.00	$425,000.00
05/30/23	119	$854,003.5714	1096232.2940	100491.44972	$655,003.1442	1053003.9986	$155,000.00	10800000.00
05/31/23	219	$958,798.4977	1458518.6884	$98,557.38244	$764,561.3006	1153045.7968	$80,000.00	14200000.00
06/01/23	133	$799,317.6917	643622.63715	$55,809.16053	$688,921.6559	$909,713.7275	$78,000.00	$3,400,000.00
06/02/23	189	$693,229.9788	571078.75241	$41,539.93959	$611,285.6999	$775,174.2579	$130,000.00	$6,150,000.00
06/05/23	78	$851,118.9872	739674.90637	$83,751.59921	$684,346.3471	1017887.7272	$80,000.00	$5,485,000.00
06/06/23	72	1311178.4722	1734393.3593	204393.14553	$903,630.1133	1718726.8312	$290,000.00	$9,300,000.00
06/07/23	80	$694,891.6625	462644.29939	$50,696.98081	$594,182.7619	$795,600.3631	$109,000.00	$2,700,000.00
06/08/23	67	$746,411.6322	655490.79626	$79,274.91760	$606,709.7152	$886,113.5451	$140,000.00	$5,250,000.00
06/09/23	245	$818,539.7429	809896.72863	$51,694.53162	$716,731.9839	$920,341.5020	$65,000.00	$9,300,000.00
06/10/23	1	1295000.0000	.	.	.	.	$1,295,000.00	$1,295,000.00
06/12/23	89	$794,154.6169	771445.90782	$81,772.09209	$631,647.8797	$956,661.3540	$141,250.00	$6,350,000.00
06/13/23	68	$813,522.3214	1061492.6361	141847.91996	$529,252.7374	1097781.9054	$5,000.00	$6,960,000.00
06/14/23	74	$943,558.2973	780434.65119	$90,723.68539	$762,738.2531	1124362.3415	$195,000.00	$3,550,000.00
06/15/23	103	$903,480.8735	1020862.5434	101083.63130	$702,997.8337	1103963.9114	$75,000.00	$7,865,000.00
06/16/23	241	1241457.7065	4453669.7814	286887.31465	$676,319.6681	1806556.3510	$70,000.00	43960000.00
06/19/23	91	$783,405.4935	623474.73596	$65,357.91110	$653,560.5053	$913,250.3817	$150,000.00	$4,115,000.00
06/20/23	79	2774569.7867	16992642.338	1911474.4126	-1030727.330	6587846.9013	$178,000.00	134500000.00
06/21/23	71	$721,669.1549	515830.92918	$61,217.66955	$599,593.9660	$843,744.3439	$160,000.00	$2,750,000.00
06/22/23	88	$875,352.6512	834326.52368	$88,987.74270	$698,472.6758	1054232.6267	$168,000.00	$5,359,000.00
06/23/23	316	$976,308.6102	1136758.7569	$77,493.71943	$823,585.7682	1129031.8311	$68,546.00	11560000.00
06/24/23	3	$263,219.0000	140873.76493	$81,160.30103	-$85,985.5907	$612,423.5907	$137,500.00	$415,000.00
06/26/23	61	1447063.3077	6192473.6824	846032.66697	$163,966.2560	2730460.3594	$30,000.00	59163724.00
06/27/23	73	$903,785.6712	1042320.2167	121994.35414	$660,594.3709	1146976.9715	$40,000.00	$6,900,000.00
06/28/23	85	$888,564.4235	802331.37085	$87,871.57870	$730,038.6560	1094282.7911	$225,000.00	$6,909,000.00
06/29/23	84	1056951.3723	2070863.7239	313690.24369	$632,783.6485	1481109.1962	$135,000.00	17750000.00
06/30/23	305	1197048.5839	3031299.6263	173571.77966	$856,054.4386	1536202.6893	$100,500.00	38840000.00
07/03/23	88	$681,919.3182	457738.55741	$48,791.60616	$584,940.1413	$778,858.4951	$80,000.00	$3,350,000.00
07/04/23	49	$870,235.1837	869920.79131	124131.54162	$620,651.9916	1119818.3756	$65,000.00	$4,425,000.00
07/05/23	62	$753,468.1260	513236.33922	$64,830.03178	$623,880.4272	$883,031.8208	$215,000.00	$2,750,000.00
07/06/23	55	$845,303.6364	834524.27821	112527.23045	$619,700.1227	1070907.1500	$167,500.00	$4,550,000.00
07/07/23	113	$862,251.3894	771947.81584	$72,618.74187	$718,366.6609	1006139.1179	$168,000.00	$6,800,000.00
Total	17547	$885,588.4738	2346547.3787	$17,737.10831	$864,821.0859	$934,354.9618	$100.00	140000000.00

ANOVA

price_paid

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.118E+15	152	7.358E+12	1.337	.004
Within Groups	9.574E+16	17394	5.504E+12		
Total	9.686E+16	17546			

ANOVA Effect Sizes[a,b]

		Point Estimate	95% Confidence Interval	
			Lower	Upper
price_paid	Eta-squared	.012	.001	.006
	Epsilon-squared	.003	-.008	-.003
	Omega-squared Fixed-effect	.003	-.008	-.003
	Omega-squared Random-effect	.000	.000	.000

a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

b. Negative but less biased estimates are retained, not rounded to zero.

As from the ANNOVA we can see that we got the Sig. value as 0.004 which is less than 0.005 so we can reject the Null hypothesis.

Null Hypothesis (H0):

The average cost of a home in each given London borough does not differ much from one another.

Theoretically: H0: μ_borough1 = μ_borough2 =... = μ_boroughN

Now for the Second Research Question:

Does the kind of property (home, flat, etc.) affect the London real estate market price in a statistically meaningful way?

Examining if property type affects pricing might be one area of study.

In order to examine if the kind of property (house, flat, etc.) influences the price of real estate in London, we may run an Kruskal-Wallis test. The Kruskal Wallis test is suitable for ordinal and nominal data.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of price_paid is the same across categories of property_type.	Independent-Samples Kruskal-Wallis Test	<.001	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

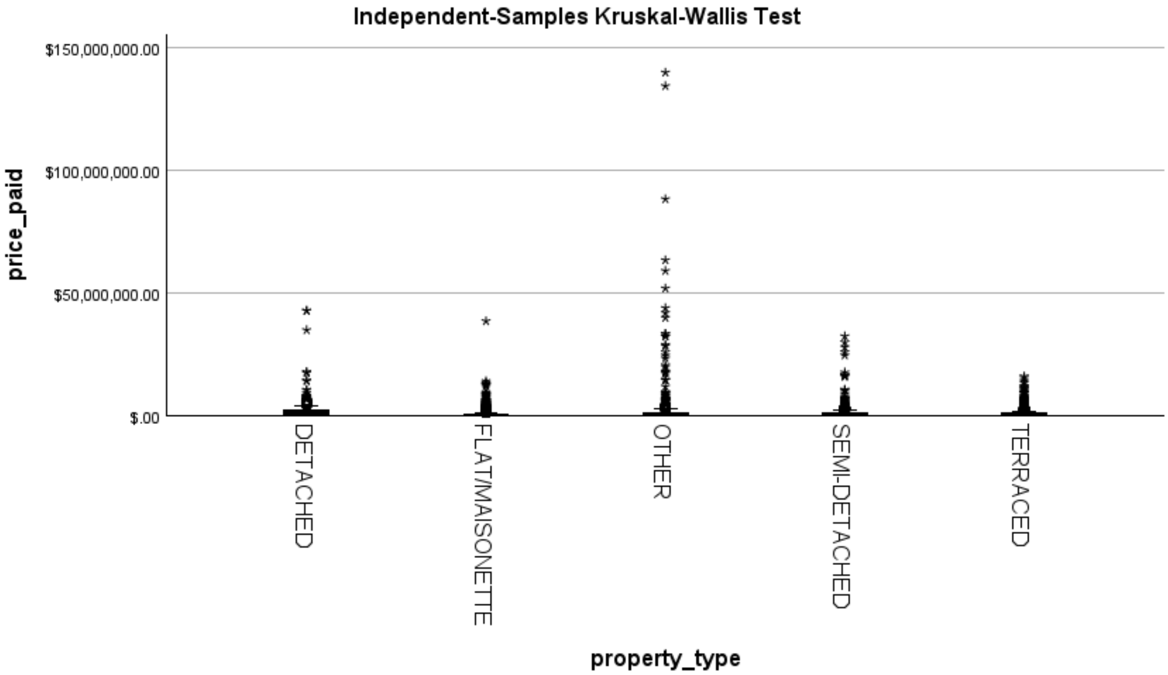
Independent-Samples Kruskal-Wallis Test

price_paid across property_type

Independent-Samples Kruskal-Wallis Test Summary

Total N	17547
Test Statistic	3234.539[a]
Degree Of Freedom	4
Asymptotic Sig.(2-sided test)	<.001

a. The test statistic is adjusted for ties.

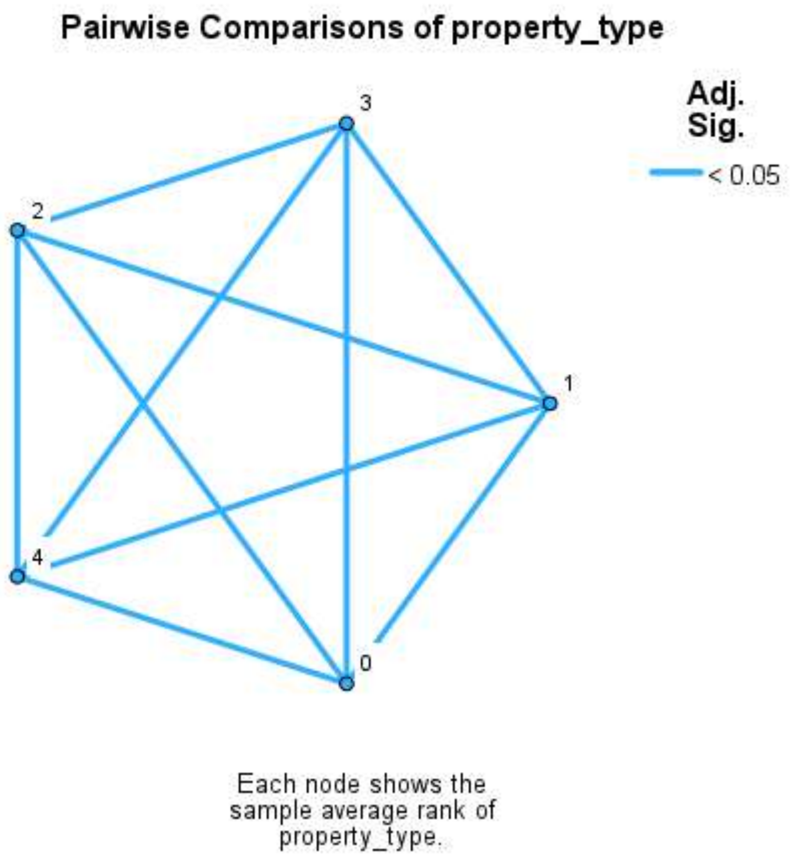


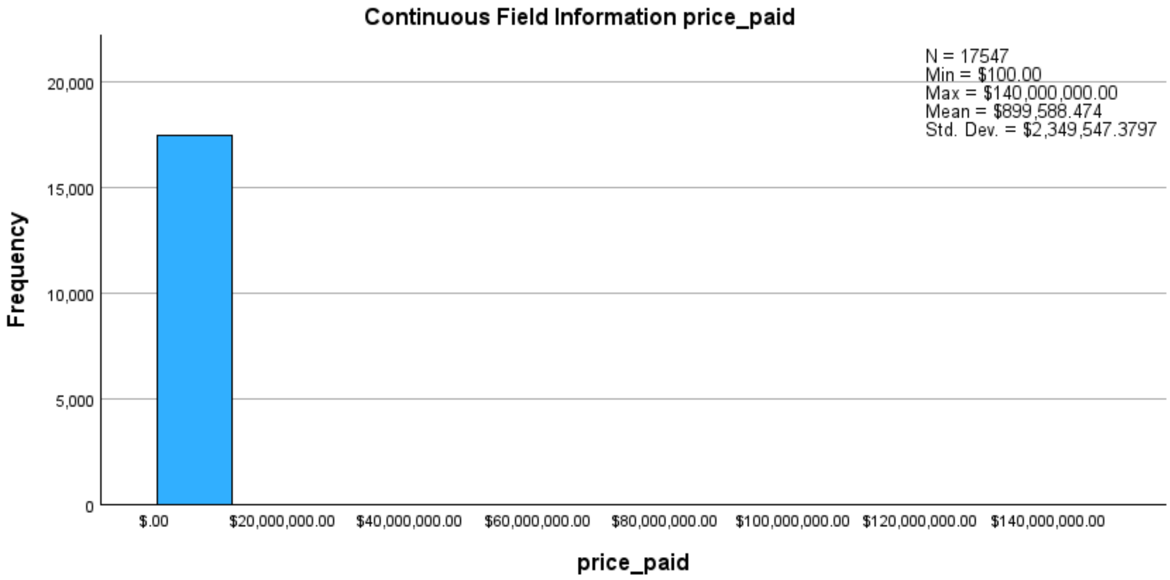
Pairwise Comparisons of property_type

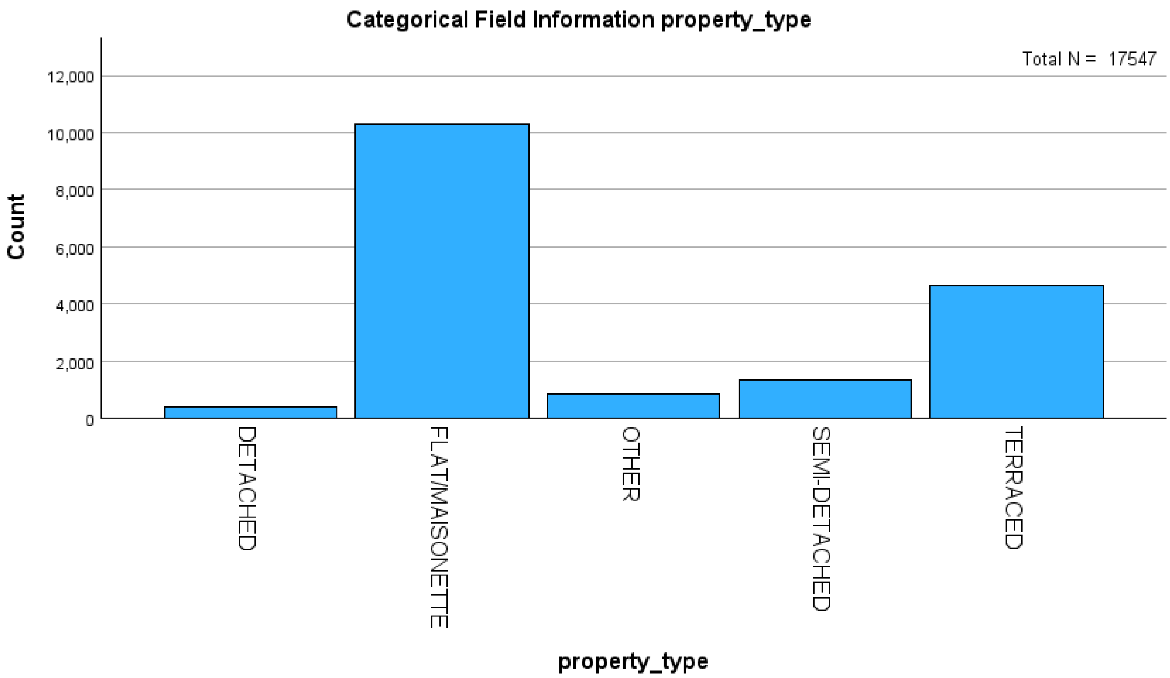
Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
FLAT/MAISONETTE-OTHER	-1237.784	179.703	-6.888	<.001	.000
FLAT/MAISONETTE-TERRACED	-3979.150	89.477	-44.471	<.001	.000
FLAT/MAISONETTE-SEMI-DETACHED	-5210.031	146.623	-35.533	<.001	.000
FLAT/MAISONETTE-DETACHED	7342.151	262.942	27.923	<.001	.000
OTHER-TERRACED	-2741.366	187.925	-14.588	<.001	.000
OTHER-SEMI-DETACHED	-3972.247	220.926	-17.980	<.001	.000
OTHER-DETACHED	6104.368	310.561	19.656	<.001	.000
TERRACED-SEMI-DETACHED	1230.881	156.592	7.860	<.001	.000
TERRACED-DETACHED	3363.002	268.629	12.519	<.001	.000
SEMI-DETACHED-DETACHED	2132.120	292.666	7.285	<.001	.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .050.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.







So here also we'll reject the null hypothesis.

Null Hypothesis (H0):

The type of property in London has no discernible impact on the selling price.

H0: μ_apartment = μ_house, symbolically

Alternate hypothesis (Ha):

The type of property in London has a substantial impact on the selling price.

Ha: μ_apartment ≠ μ_house, symbolically

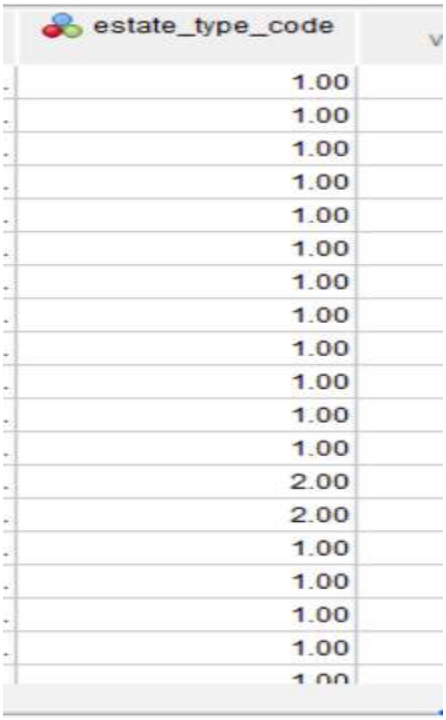
Now for 3rd Research Question:

Does the kind of estate—freehold or leasehold—have a big influence on how much a London property sells for?

For this first converted the variable estate_type which was string having the FREEHOLD and LEASEHOLD, so converting them to binary i.e
LEASEHOLD changed to 1 and FREEHOLD changed to 2.

This is done by by clicking on Transform-> Recode into Different Variables.
And then selecting the estate_type form list of variables on the left and moved it to the Input variable box on the right.
And then in the Output variable entered the new variable name i.e estate_type_code.
And then clicked on the change button.
Enter the value you wish to replace for each entry in the "Old Value" field.
Set the previous leasehold value to the appropriate amount (for example, 1).
Change the previous freehold value to the other matching value (for example, 2).

And finally got the variable, as can be seen below



estate_type_code
1.00
1.00
1.00
1.00
1.00
1.00
1.00
1.00
1.00
1.00
1.00
1.00
2.00
2.00
1.00
1.00
1.00
1.00
1.00

So now,
Depending on how your data is distributed, you may use a t-test or a Mann-Whitney U test to determine if the kind of estate (freehold or leasehold) has a substantial influence on the selling price of houses in SPSS.

Go to "Compare Means" > "Analyse" > "Independent Samples T-Test":

Navigate to Analyse from the menu bar, then choose Compare Means, and lastly, Independent Samples T-Test.
Choose Your Variables:

In the dialogue box for the Independent Samples T-Test:
Place the variable (such as price_paid) that represents the selling price in the "Test Variable(s)" box.
Place the estate type variable (such as estate_type) in the "Grouping Variable" box.
Describe Groups:

To designate the test groups, click the "Define Groups" option.
In "Group 1," type the freehold-corresponding value.

Enter the value that corresponds to leasehold for "Group 2".

Choices (if required):

Depending on your data's properties, you can investigate other choices, such correcting for uneven variances.

To run the analysis, click "OK":

T-Test

Group Statistics

	estate_type_code	N	Mean	Std. Deviation	Std. Error Mean
price_paid	2.00	6678	1300835.3327	3403862.6052	$41,653.24246
	1.00	10869	$653,059.2140	1278453.6182	$12,262.81516

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
price_paid	Equal variances assumed	247.459	<.001	17.893	17545	<.001	<.001	647776.11873	$36,203.74250	576813.19187	718739.04560
	Equal variances not assumed			14.919	7848.365	<.001	<.001	647776.11873	$43,420.83882	562659.71196	732892.52550

Independent Samples Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval	
				Lower	Upper
price_paid	Cohen's d	2328466.6979	.278	.248	.309
	Hedges' correction	2328566.2393	.278	.248	.309
	Glass's delta	1278453.6182	.507	.475	.538

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control (i.e., the second) group.

So here also rejecting the Null Hypothesis.

Frequency Statistics:

Statistics

		estate_type	postcode	new_build	property_type
N	Valid	17547	17547	17547	17547
	Missing	0	0	0	0

estate_type

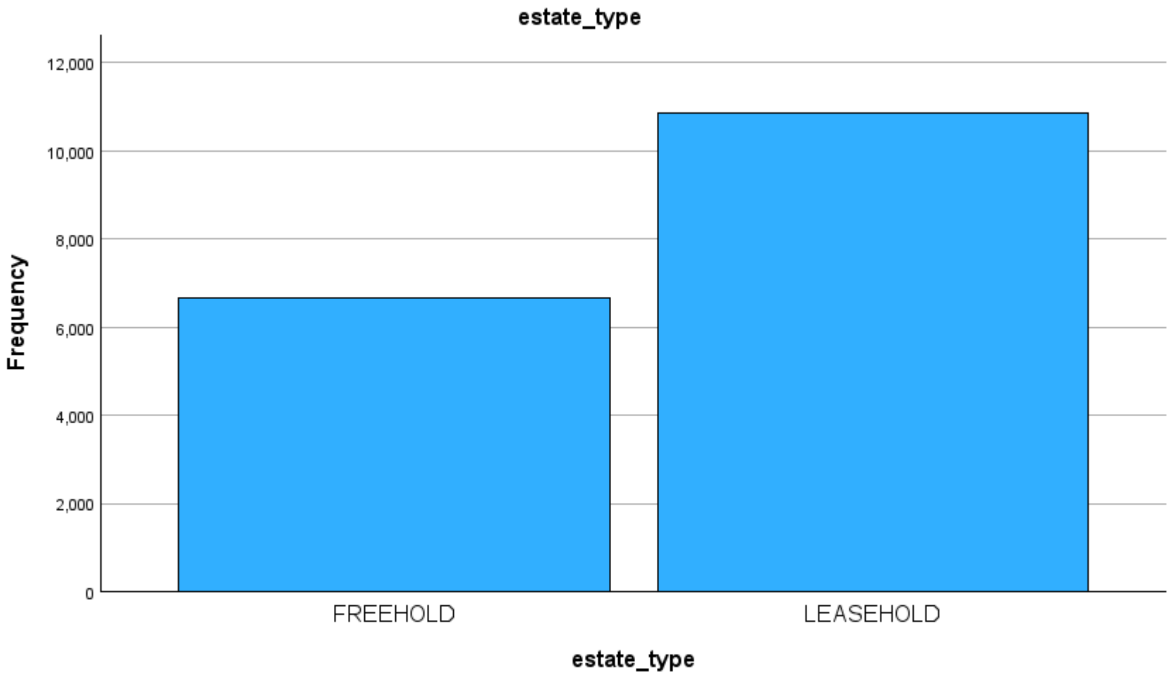
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	FREEHOLD	6678	38.1	38.1	38.1
	LEASEHOLD	10869	61.9	61.9	100.0
	Total	17547	100.0	100.0	

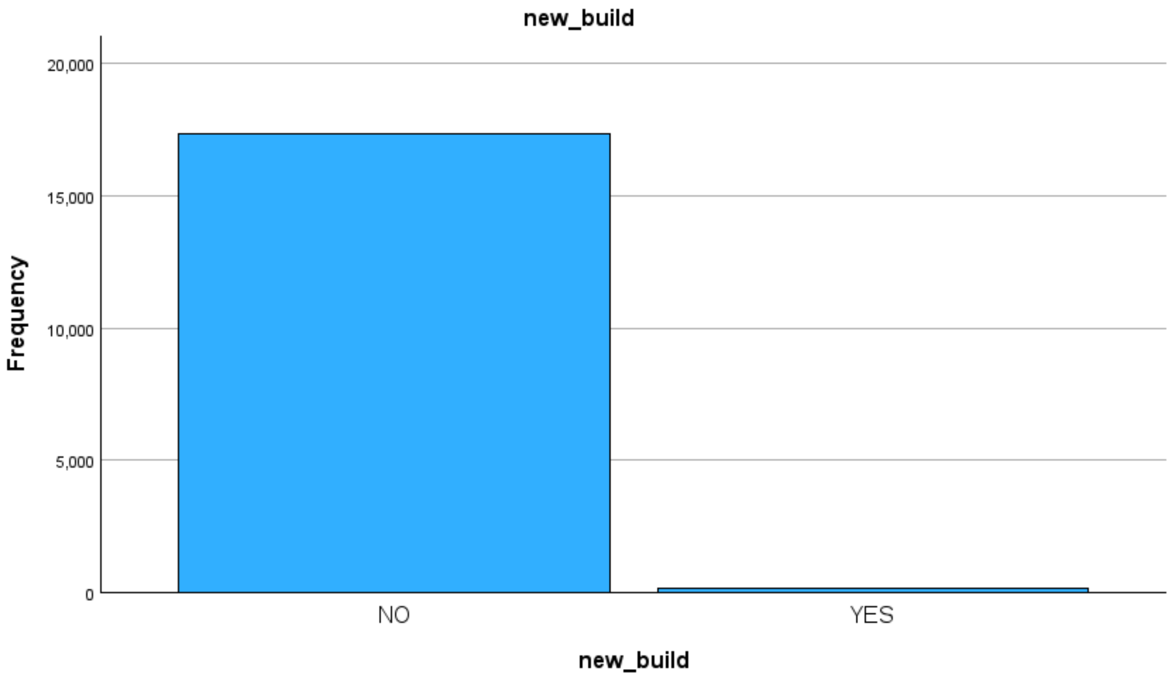
new_build

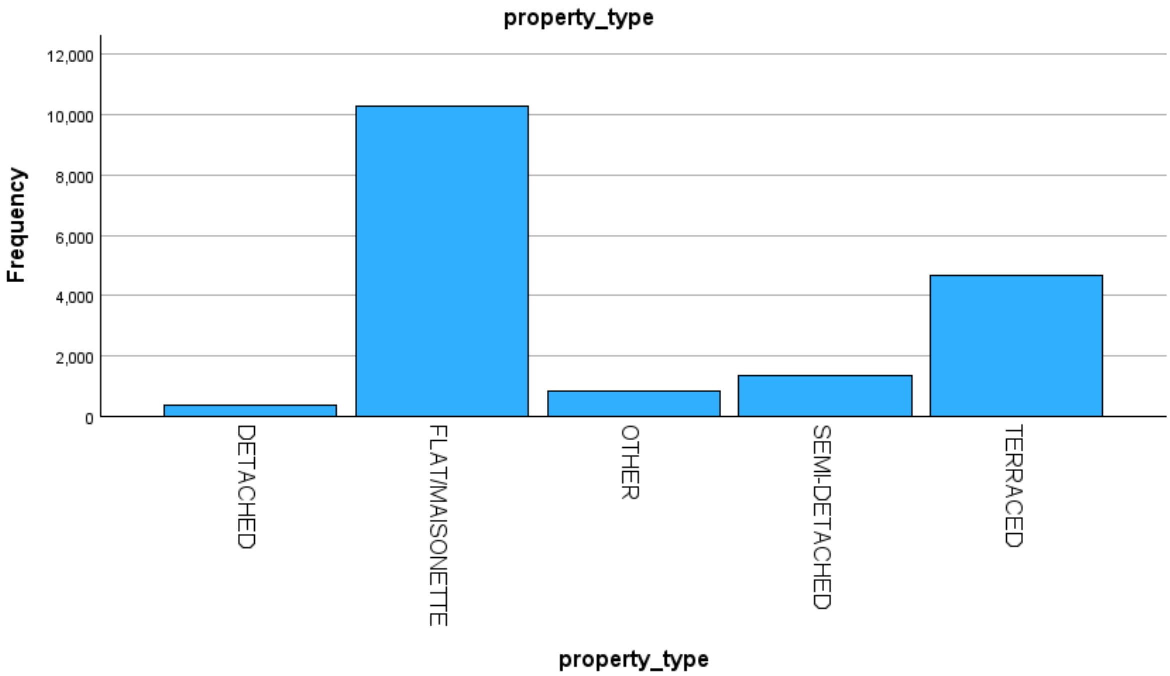
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NO	17375	99.0	99.0	99.0
	YES	172	1.0	1.0	100.0
	Total	17547	100.0	100.0	

property_type

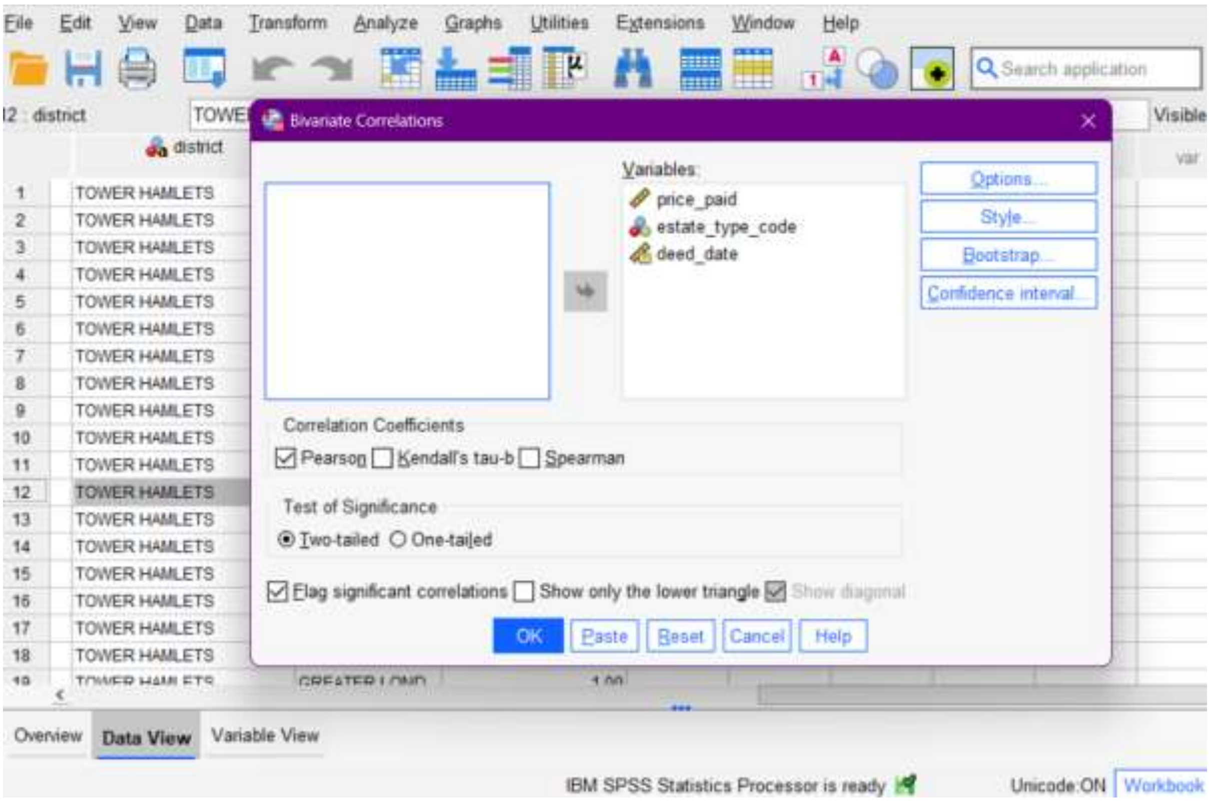
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	DETACHED	385	2.2	2.2	2.2
	FLAT/MAISONETTE	10298	58.7	58.7	60.9
	OTHER	861	4.9	4.9	65.8
	SEMI-DETACHED	1350	7.7	7.7	73.5
	TERRACED	4653	26.5	26.5	100.0
	Total	17547	100.0	100.0	


estate_type
12,000
10,000
8,000
6,000
4,000
2,000
0
Frequency
FREEHOLD
LEASEHOLD
estate_type


new_build
20,000
15,000
10,000
5,000
0
Frequency
NO
YES
new_build



Corelations:



Descriptive Statistics

	Mean	Std. Deviation	N
price_paid	$899,588.4738	2349547.3797	17547
estate_type_code	1.3806	.48554	17547
deed_date	03/29/23	53 16:56:34.7...	17547

Correlations

		price_paid	estate_type_code	deed_date
price_paid	Pearson Correlation	1	.134**	.007
	Sig. (2-tailed)		<.001	.372
	N	17547	17547	17547
estate_type_code	Pearson Correlation	.134**	1	.032**
	Sig. (2-tailed)	<.001		<.001
	N	17547	17547	17547
deed_date	Pearson Correlation	.007	.032**	1
	Sig. (2-tailed)	.372	<.001	
	N	17547	17547	17547

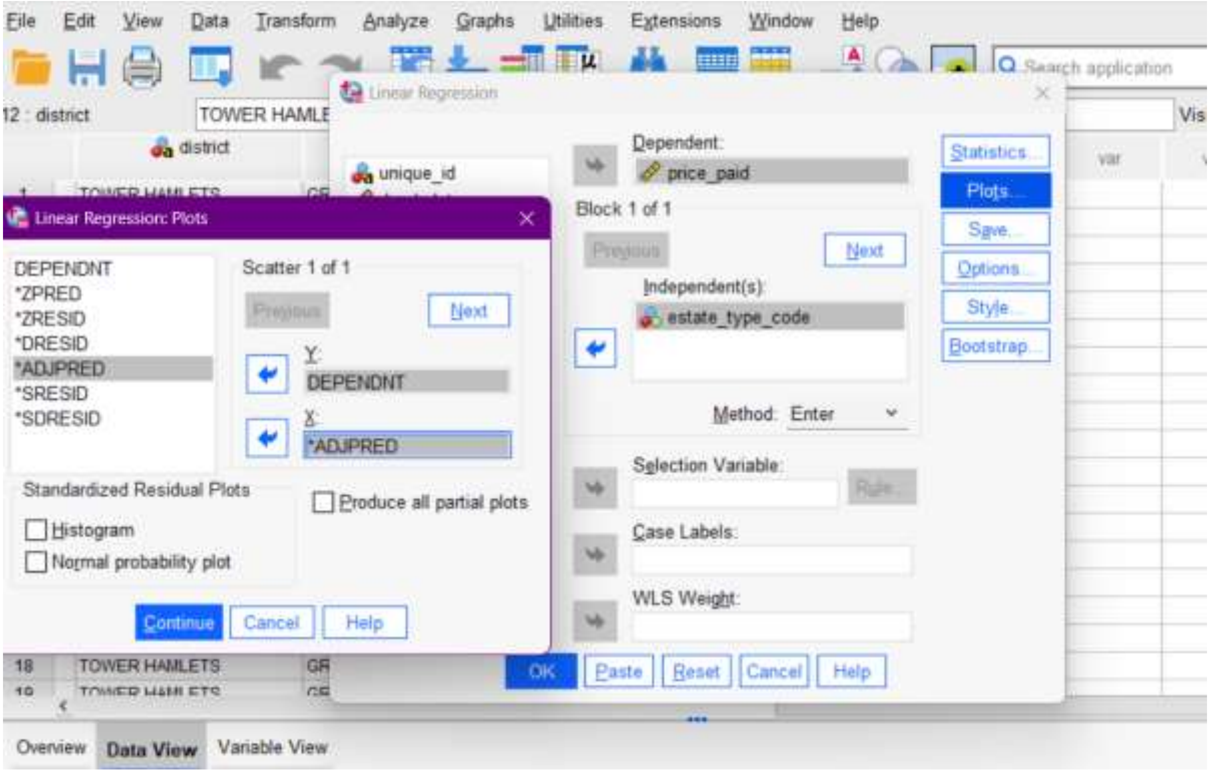
**. Correlation is significant at the 0.01 level (2-tailed).

Regression Analysis:

Does the type of estate (leasehold/freehold) impact housing prices?

Independent Variable: estate_type

Dependent variable: price_paid



Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	estate_type_code[b]	.	Enter

a. Dependent Variable: price_paid

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.134[a]	.018	.018	2328466.6979	.018	320.142	1	17545	<.001

a. Predictors: (Constant), estate_type_code

b. Dependent Variable: price_paid

Coefficients[a]

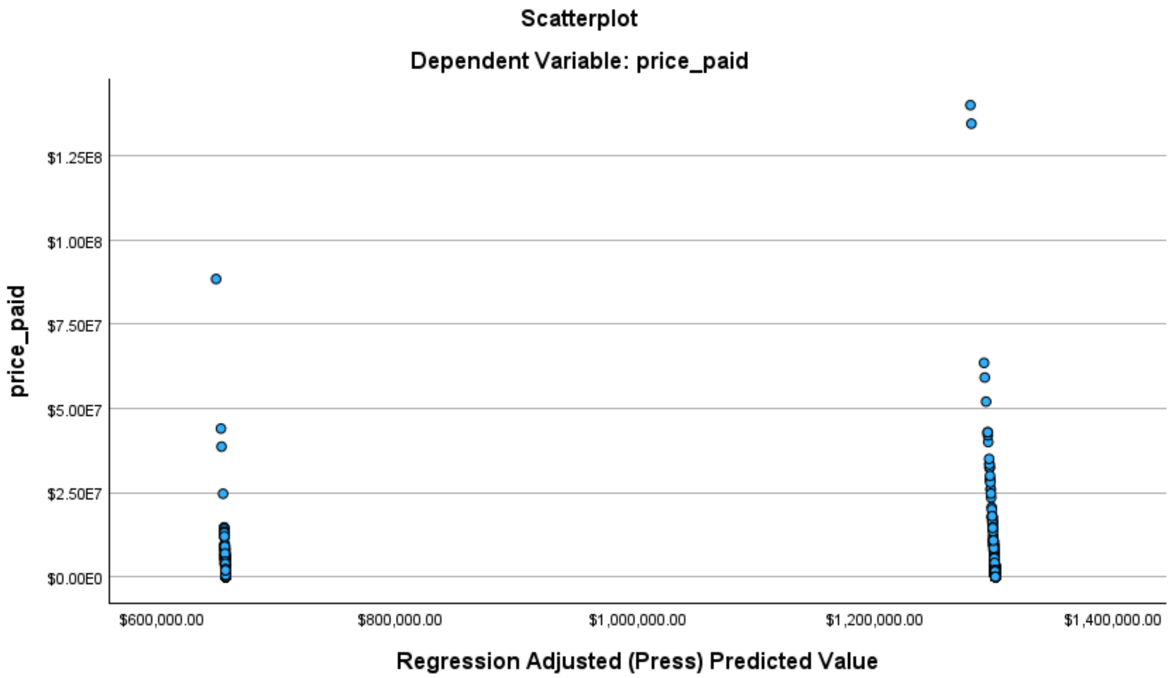
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	5283.095	52982.957		.100	.921
	estate_type_code	647776.119	36203.742	.134	17.893	<.001

a. Dependent Variable: price_paid

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	$653,059.1875	1300835.3750	$899,588.4738	314523.01795	17547
Std. Predicted Value	-.784	1.276	.000	1.000	17547
Standard Error of Predicted Value	22334.449	28493.566	24678.473	2990.516	17547
Adjusted Predicted Value	$644,985.3125	1301030.1250	$899,588.4738	314523.18880	17547
Residual	-1300735.375	138699168.00	$0.00000	2328400.3438	17547
Std. Residual	-.559	59.567	.000	1.000	17547
Stud. Residual	-.559	59.571	.000	1.000	17547
Deleted Residual	-1300930.125	138719936.00	$0.00000	2328723.9525	17547
Stud. Deleted Residual	-.559	66.695	.001	1.053	17547
Mahal. Distance	.614	1.627	1.000	.492	17547
Cook's Distance	.000	.266	.000	.003	17547
Centered Leverage Value	.000	.000	.000	.000	17547

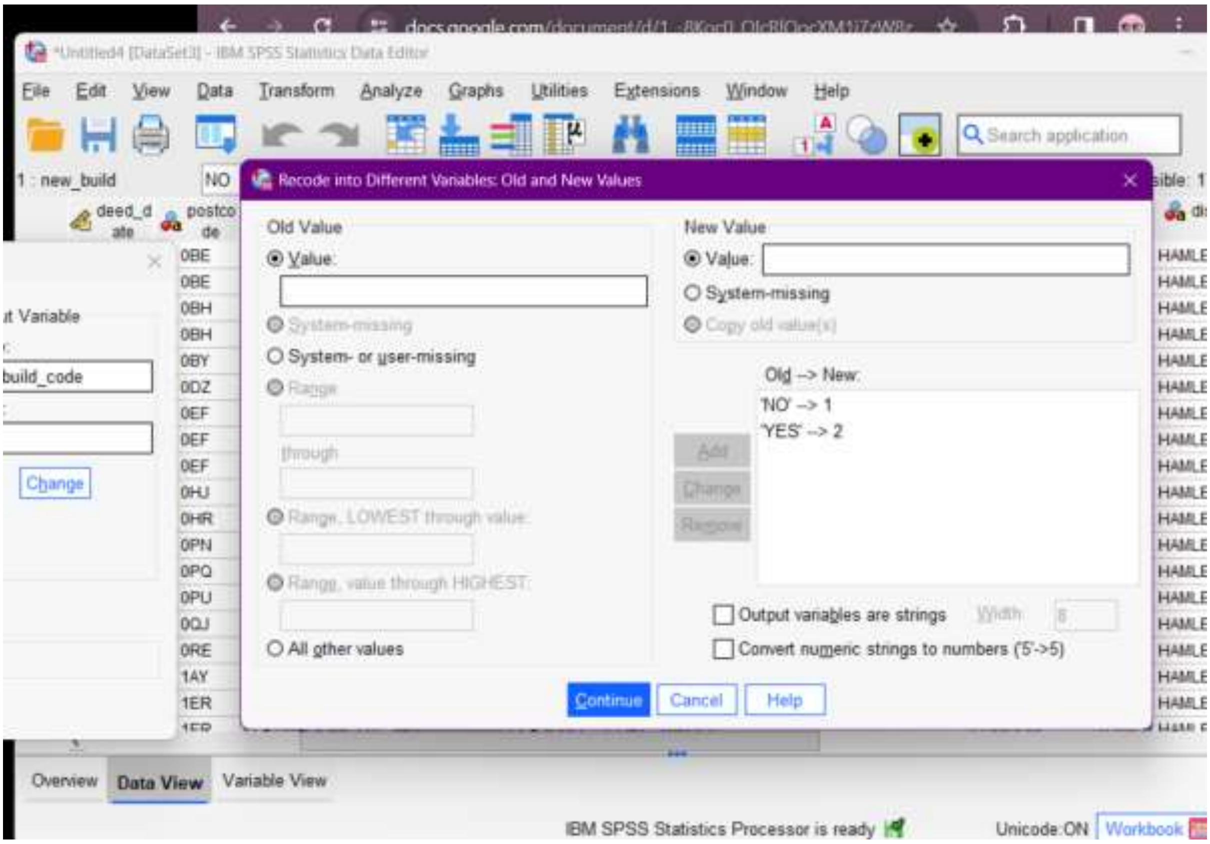
a. Dependent Variable: price_paid



How does the newness of a property (new build or not) relate to housing prices?

Independent Variable: new_build

Dependent variable: price_paid



Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	new_build_code[b]	.	Enter

a. Dependent Variable: price_paid

b. All requested variables entered.

Model Summary[b]

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.008[a]	.000	.000	2349536.5690	.000	1.161	1	17545	.281

a. Predictors: (Constant), new_build_code

b. Dependent Variable: price_paid

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.412E+12	1	6.412E+12	1.161	.281[b]
	Residual	9.685E+16	17545	5.520E+12		
	Total	9.686E+16	17546			

a. Dependent Variable: price_paid

b. Predictors: (Constant), new_build_code

Coefficients[a]

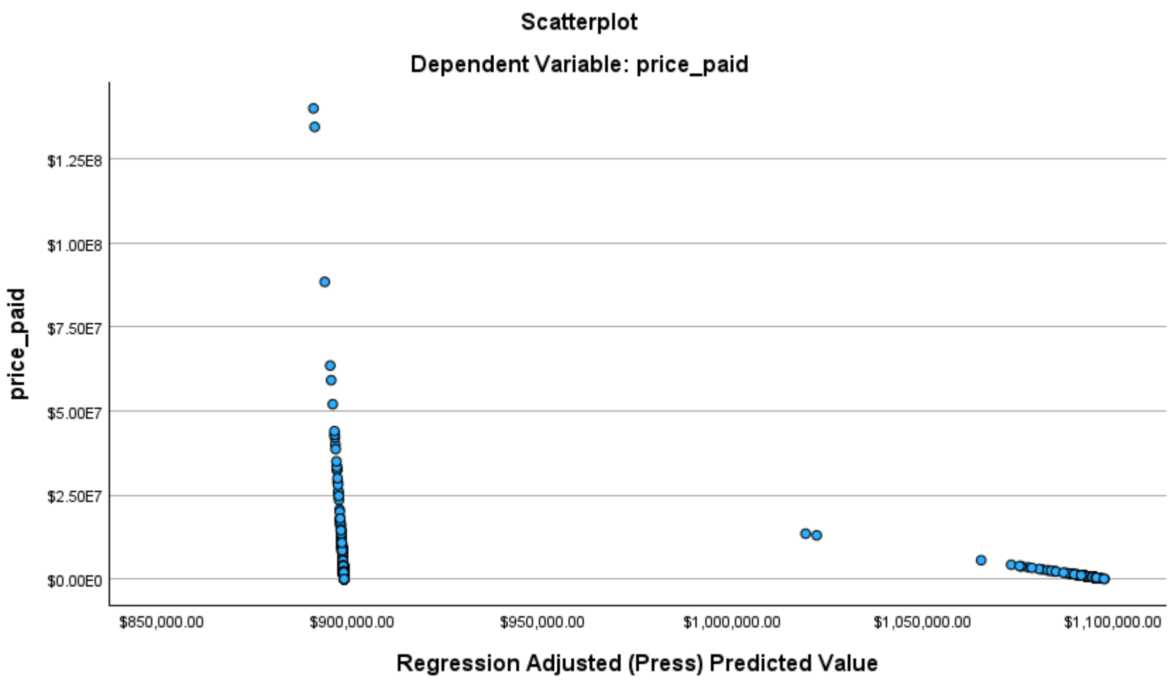
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	703660.336	182662.972		3.852	<.001
	new_build_code	194026.245	180035.031	.008	1.078	.281

a. Dependent Variable: price_paid

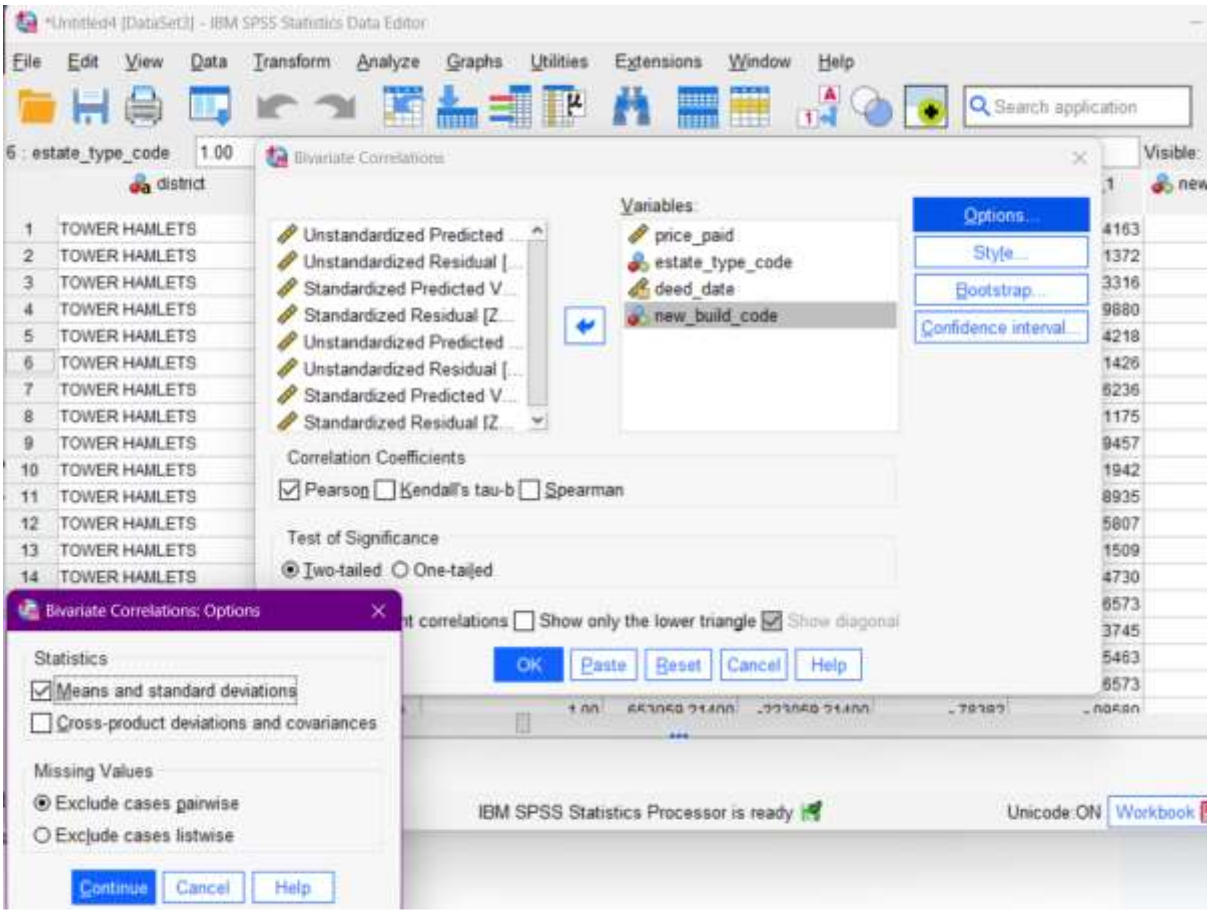
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	$897,686.5625	1091712.8750	$899,588.4738	$19,115.98176	17547
Std. Predicted Value	-.099	10.050	.000	1.000	17547
Standard Error of Predicted Value	17824.600	179150.484	19405.956	15894.255	17547
Adjusted Predicted Value	$889,680.2500	1097804.7500	$899,588.4738	$19,137.70602	17547
Residual	-1041712.812	139102320.00	$0.00000	2349469.6144	17547
Std. Residual	-.443	59.204	.000	1.000	17547
Stud. Residual	-.445	59.206	.000	1.000	17547
Deleted Residual	-1047804.688	139110320.00	$0.00000	2349663.5893	17547
Stud. Deleted Residual	-.445	66.184	.001	1.051	17547
Mahal. Distance	.010	101.012	1.000	9.951	17547
Cook's Distance	.000	.101	.000	.001	17547
Centered Leverage Value	.000	.006	.000	.001	17547

a. Dependent Variable: price_paid



Correlation matrix:



Descriptive Statistics

	Mean	Std. Deviation	N
price_paid	$899,588.4738	2349547.3797	17547
estate_type_code	1.3806	.48554	17547
deed_date	03/29/23	53 16:56:34.7...	17547
new_build_code	1.0098	.09852	17547

Correlations

		price_paid	estate_type_code	deed_date	new_build_code
price_paid	Pearson Correlation	1	.134**	.007	.008
	Sig. (2-tailed)		<.001	.372	.281
	N	17547	17547	17547	17547
estate_type_code	Pearson Correlation	.134**	1	.032**	-.078**
	Sig. (2-tailed)	<.001		<.001	<.001
	N	17547	17547	17547	17547
deed_date	Pearson Correlation	.007	.032**	1	-.075**
	Sig. (2-tailed)	.372	<.001		<.001
	N	17547	17547	17547	17547
new_build_code	Pearson Correlation	.008	-.078**	-.075**	1
	Sig. (2-tailed)	.281	<.001	<.001	
	N	17547	17547	17547	17547

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

Task four: Evaluation and Conclusion

Regional Variations in the Cost of Homes:

Is there a significant difference in average housing prices between the boroughs of London?

Conclusion: By using an ANOVA analysis, we were able to identify notable variations in the average cost of homes in each London borough. This implies that borough-to-borough housing values differ considerably.

Property Type's Effect on Price:

Research Question: Does the kind of property (house, flat, etc.) have a statistically significant impact on the price of real estate in London?

In summary, we found that property type does affect price in a statistically significant way using ANOVA. The average selling price of various property kinds varies.

Estate Type's Effect on Property Prices

Research Question: Is there a significant difference in the sale price of a London home depending on whether the property is freehold or leasehold?
In summary: Using a t-test for independent samples, we discovered evidence that refuted the null hypothesis. The selling prices of freehold and leasehold homes varies significantly.

Comprehensive Analysis of the Dataset:
Descriptive Statistics: For the key variables, descriptive statistics such as means median and standard deviations were computed. It uncovered the facts of the variation and central tendency of the dataset.

Data Cleaning and Preparation:
Data Transformation: Adequate preparation of the data was also ensured in this step by using transformation and cleansing processes to ensure reliability and accuracy for later analysis.

Additional Understanding:

Correlation Matrix: The correlations among the variables were calculated by building a correlation matrix. The diagram is used to determine the relationship between variables.

In the end, detailed interdependent patterns between variables that determine real estate price emerged through housing data analysis in London. The results can help practitioners in the real estate field as well as contribute to future research or decision making.

Section 2: Data Modelling

Entities and Attributes:

Task1: Design a relational Database

Customer

- CustomerID (PK)
- FirstName
- LastName
- Email
- Phone
- Address
- RegistrationDate

LoyaltyPoints

Product

ProductID (PK)

ProductName

Description

Price

StockQuantity

Weight

Dimensions

Order

OrderID (PK)

OrderDate

ShippingAddress

PaymentMethod

OrderStatus

TotalAmount

TaxAmount

ShippingFee

OrderItem

OrderItemID (PK)

Quantity

Subtotal

Vendor

VendorID (PK)

VendorName

ContactName

ContactEmail

Phone

Address

Review

ReviewID (PK)

Rating

ReviewText

ReviewDate

Relationships:

Customer-Order Relationship:

A client can place a single order or several orders.

A client can have only one single order.

Order-OrderItem Relationship:

Each order may contain one or more items from the list of orders.

Each order contains one item order.

Product-OrderItem Relationship:

A product may be in any quantity or zero.

An order item is one product.

Vendor-Product Relationship:

One vendor can offer one or several products.

A seller provides one product.

Customer-Review Relationship:

This single consumer can write one review or several reviews.

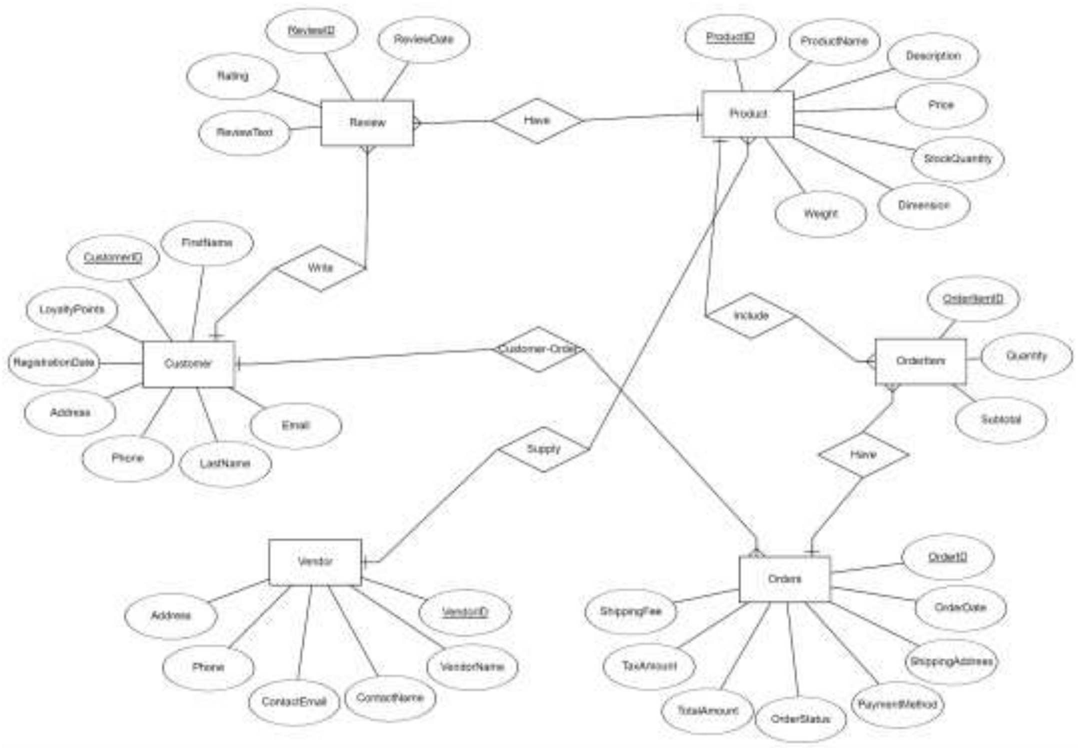
One customer writes one review.

Product-Review Relationship:

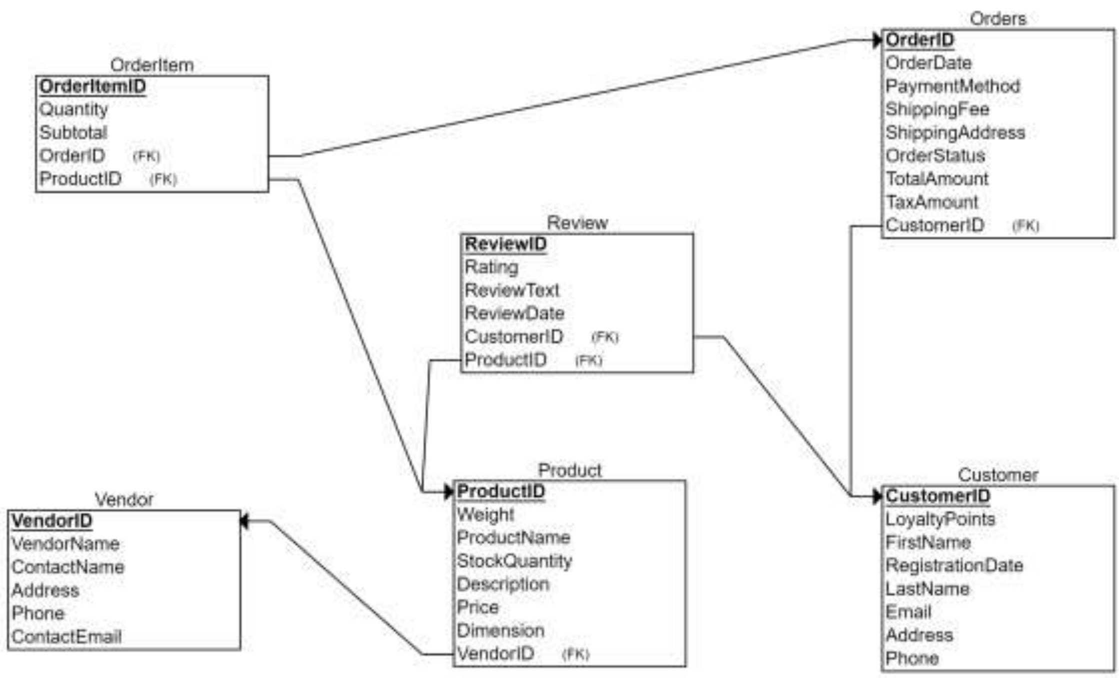
A product may have one review or many.

Every product gets one review.

ERD



Relation Table



Task 2: Write Code

INSERT INTO Customer (CustomerID, FirstName, LastName, Email, Phone, Address, RegistrationDate, LoyaltyPoints)

VALUES

(1, 'John', 'Doe', 'john.doe@email.com', '123-456-7890', '123 Main St', '2022-01-01', 100),

(2, 'Jane', 'Smith', 'jane.smith@email.com', '987-654-3210', '456 Oak St', '2022-01-02', 75),

(3, 'Bob', 'Johnson', 'bob.johnson@email.com', '555-123-4567', '789 Pine St', '2022-01-03', 50),

(4, 'Alice', 'Williams', 'alice.williams@email.com', '111-222-3333', '987 Cedar St', '2022-01-04', 120),

(5, 'Charlie', 'Brown', 'charlie.brown@email.com', '333-444-5555', '654 Walnut St', '2022-01-05', 80),

(6, 'Eva', 'Miller', 'eva.miller@email.com', '666-777-8888', '321 Birch St', '2022-01-06', 60),

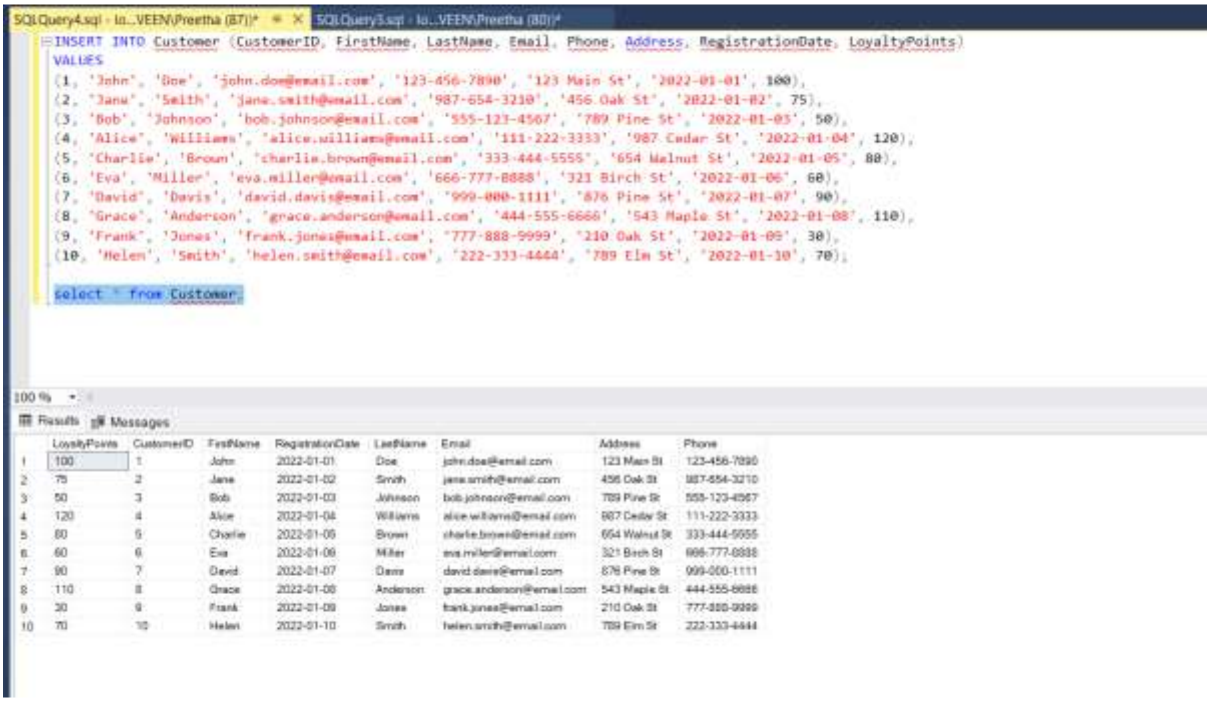
(7, 'David', 'Davis', 'david.davis@email.com', '999-000-1111', '876 Pine St', '2022-01-07', 90),

(8, 'Grace', 'Anderson', 'grace.anderson@email.com', '444-555-6666', '543 Maple St', '2022-01-08', 110),

(9, 'Frank', 'Jones', 'frank.jones@email.com', '777-888-9999', '210 Oak St', '2022-01-09', 30),

(10, 'Helen', 'Smith', 'helen.smith@email.com', '222-333-4444', '789 Elm St', '2022-01-10', 70);

select * from Customer;



INSERT INTO OrderS (OrderID, OrderDate, ShippingAddress, PaymentMethod, OrderStatus, TotalAmount, TaxAmount, ShippingFee, CustomerID)

VALUES

(1, '2022-01-15', '123 Main St', 'Credit Card', 'Shipped', 150.00, 15.00, 10.00, 3),

(2, '2022-01-16', '456 Oak St', 'PayPal', 'Processing', 200.00, 20.00, 15.00, 7),

(3, '2022-01-17', '789 Pine St', 'Credit Card', 'Delivered', 100.00, 10.00, 8.00, 5),

(4, '2022-01-18', '987 Cedar St', 'Cash on Delivery', 'Shipped', 120.00, 12.00, 5.00, 1),

(5, '2022-01-19', '654 Walnut St', 'Credit Card', 'Processing', 180.00, 18.00, 12.00, 10),

(6, '2022-01-20', '321 Birch St', 'PayPal', 'Delivered', 90.00, 9.00, 7.00, 2),

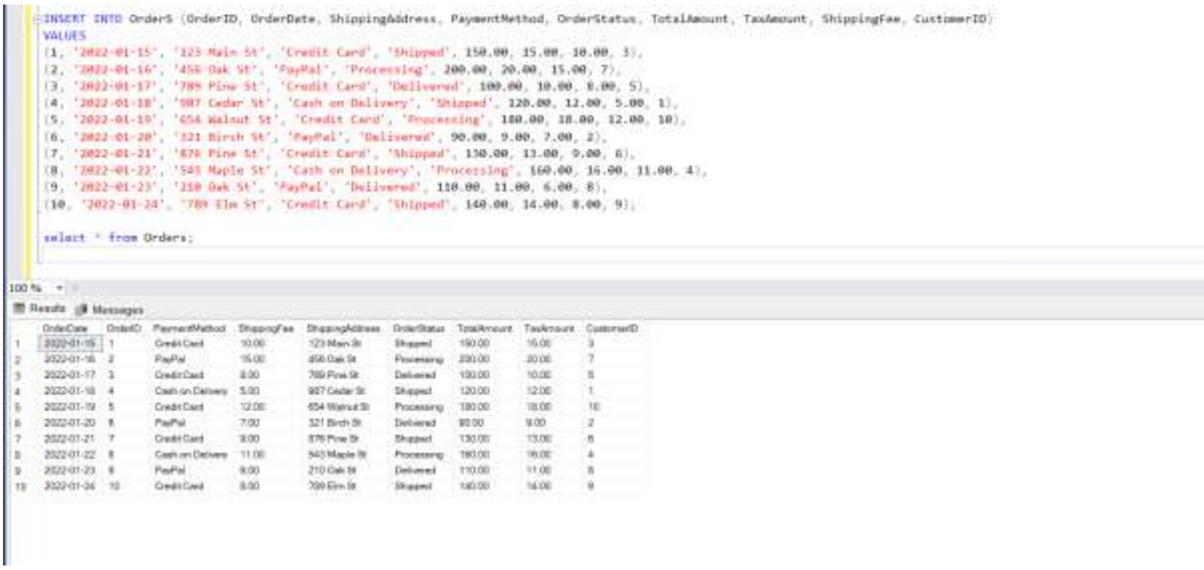
(7, '2022-01-21', '876 Pine St', 'Credit Card', 'Shipped', 130.00, 13.00, 9.00, 6),

(8, '2022-01-22', '543 Maple St', 'Cash on Delivery', 'Processing', 160.00, 16.00, 11.00, 4),

(9, '2022-01-23', '210 Oak St', 'PayPal', 'Delivered', 110.00, 11.00, 6.00, 8),

(10, '2022-01-24', '789 Elm St', 'Credit Card', 'Shipped', 140.00, 14.00, 8.00, 9);

select * from Orders;



INSERT INTO Vendor (VendorID, VendorName, ContactName, ContactEmail, Phone, Address)

VALUES

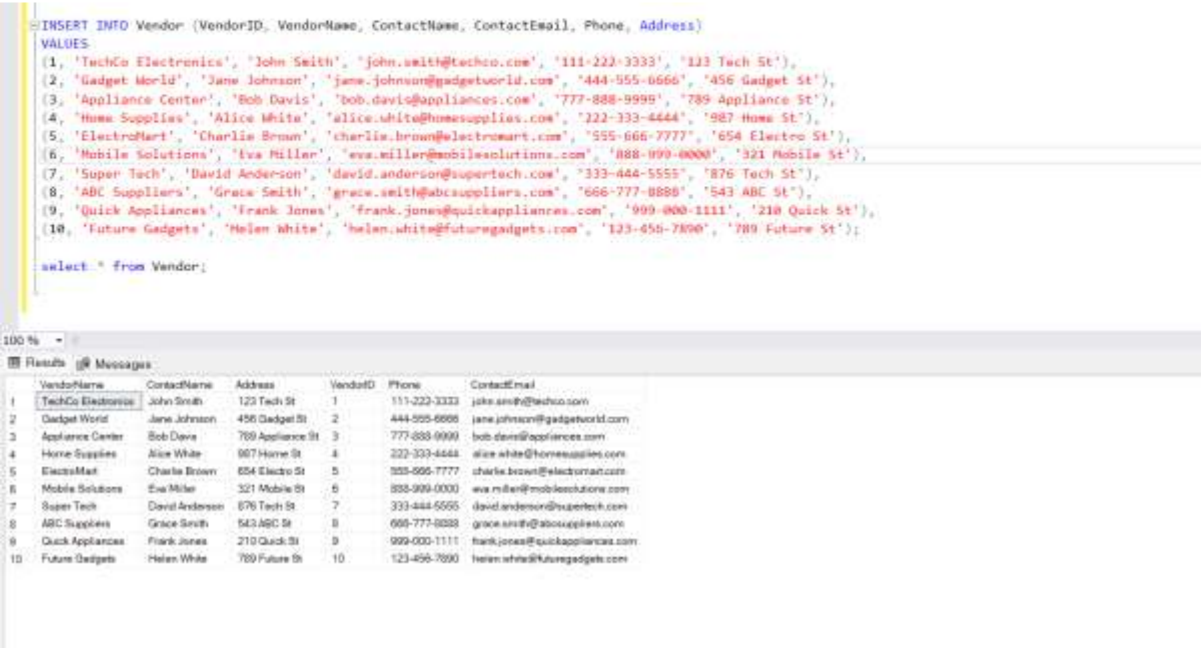
(1, 'TechCo Electronics', 'John Smith', 'john.smith@techco.com', '111-222-3333', '123 Tech St'),

(2, 'Gadget World', 'Jane Johnson', 'jane.johnson@gadgetworld.com', '444-555-6666', '456 Gadget St'),

(3, 'Appliance Center', 'Bob Davis', 'bob.davis@appliances.com', '777-888-9999', '789 Appliance St'),

(4, 'Home Supplies', 'Alice White', 'alice.white@homesupplies.com', '222-333-4444', '987 Home St'),
(5, 'ElectroMart', 'Charlie Brown', 'charlie.brown@electromart.com', '555-666-7777', '654 Electro St'),
(6, 'Mobile Solutions', 'Eva Miller', 'eva.miller@mobilesolutions.com', '888-999-0000', '321 Mobile St'),
(7, 'Super Tech', 'David Anderson', 'david.anderson@supertech.com', '333-444-5555', '876 Tech St'),
(8, 'ABC Suppliers', 'Grace Smith', 'grace.smith@abcsuppliers.com', '666-777-8888', '543 ABC St'),
(9, 'Quick Appliances', 'Frank Jones', 'frank.jones@quickappliances.com', '999-000-1111', '210 Quick St'),
(10, 'Future Gadgets', 'Helen White', 'helen.white@futuregadgets.com', '123-456-7890', '789 Future St');

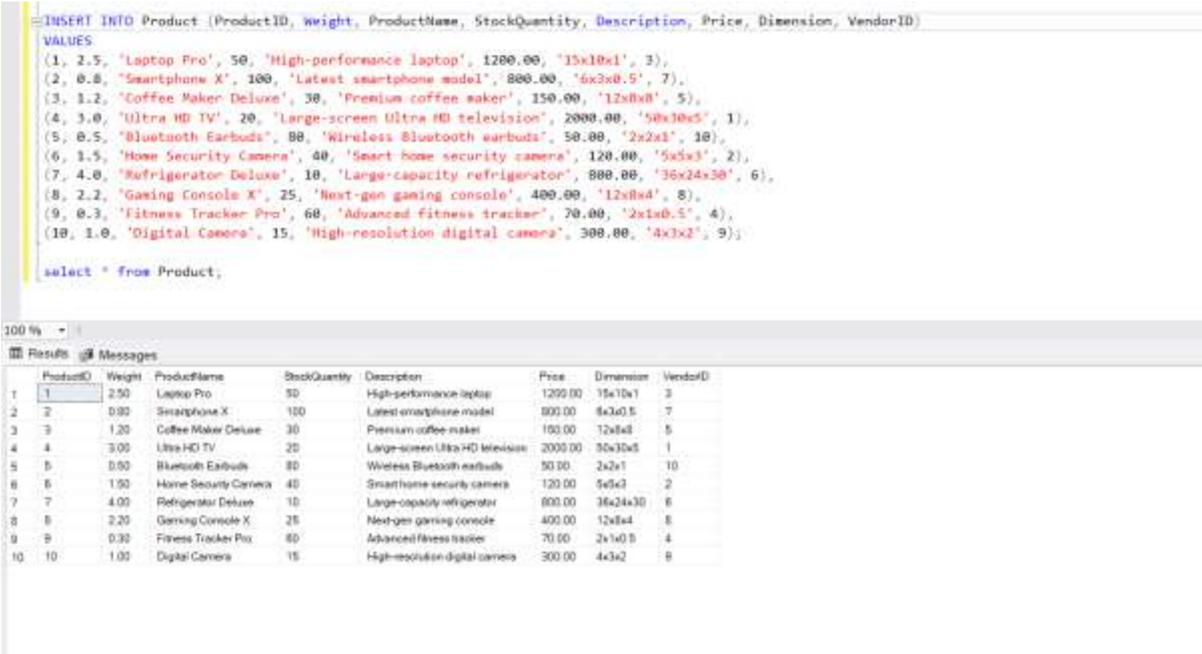
select * from Vendor;



INSERT INTO Product (ProductID, Weight, ProductName, StockQuantity, Description, Price, Dimension, VendorID)
VALUES
(1, 2.5, 'Laptop Pro', 50, 'High-performance laptop', 1200.00, '15x10x1', 3),
(2, 0.8, 'Smartphone X', 100, 'Latest smartphone model', 800.00, '6x3x0.5', 7),

(3, 1.2, 'Coffee Maker Deluxe', 30, 'Premium coffee maker', 150.00, '12x8x8', 5),
(4, 3.0, 'Ultra HD TV', 20, 'Large-screen Ultra HD television', 2000.00, '50x30x5', 1),
(5, 0.5, 'Bluetooth Earbuds', 80, 'Wireless Bluetooth earbuds', 50.00, '2x2x1', 10),
(6, 1.5, 'Home Security Camera', 40, 'Smart home security camera', 120.00, '5x5x3', 2),
(7, 4.0, 'Refrigerator Deluxe', 10, 'Large-capacity refrigerator', 800.00, '36x24x30', 6),
(8, 2.2, 'Gaming Console X', 25, 'Next-gen gaming console', 400.00, '12x8x4', 8),
(9, 0.3, 'Fitness Tracker Pro', 60, 'Advanced fitness tracker', 70.00, '2x1x0.5', 4),
(10, 1.0, 'Digital Camera', 15, 'High-resolution digital camera', 300.00, '4x3x2', 9);

select * from Product;



```sql
INSERT INTO Product (ProductID, Weight, ProductName, StockQuantity, Description, Price, Dimension, VendorID)
VALUES
(1, 2.5, 'Laptop Pro', 50, 'High-performance laptop', 1200.00, '15x10x1', 3),
(2, 0.8, 'Smartphone X', 100, 'Latest smartphone model', 800.00, '6x3x0.5', 7),
(3, 1.2, 'Coffee Maker Deluxe', 30, 'Premium coffee maker', 150.00, '12x8x8', 5),
(4, 3.0, 'Ultra HD TV', 20, 'Large-screen Ultra HD television', 2000.00, '50x30x5', 1),
(5, 0.5, 'Bluetooth Earbuds', 80, 'Wireless Bluetooth earbuds', 50.00, '2x2x1', 10),
(6, 1.5, 'Home Security Camera', 40, 'Smart home security camera', 120.00, '5x5x3', 2),
(7, 4.0, 'Refrigerator Deluxe', 10, 'Large-capacity refrigerator', 800.00, '36x24x30', 6),
(8, 2.2, 'Gaming Console X', 25, 'Next-gen gaming console', 400.00, '12x8x4', 8),
(9, 0.3, 'Fitness Tracker Pro', 60, 'Advanced fitness tracker', 70.00, '2x1x0.5', 4),
(10, 1.0, 'Digital Camera', 15, 'High-resolution digital camera', 300.00, '4x3x2', 9);

select * from Product;
```

100 %

Results | Messages

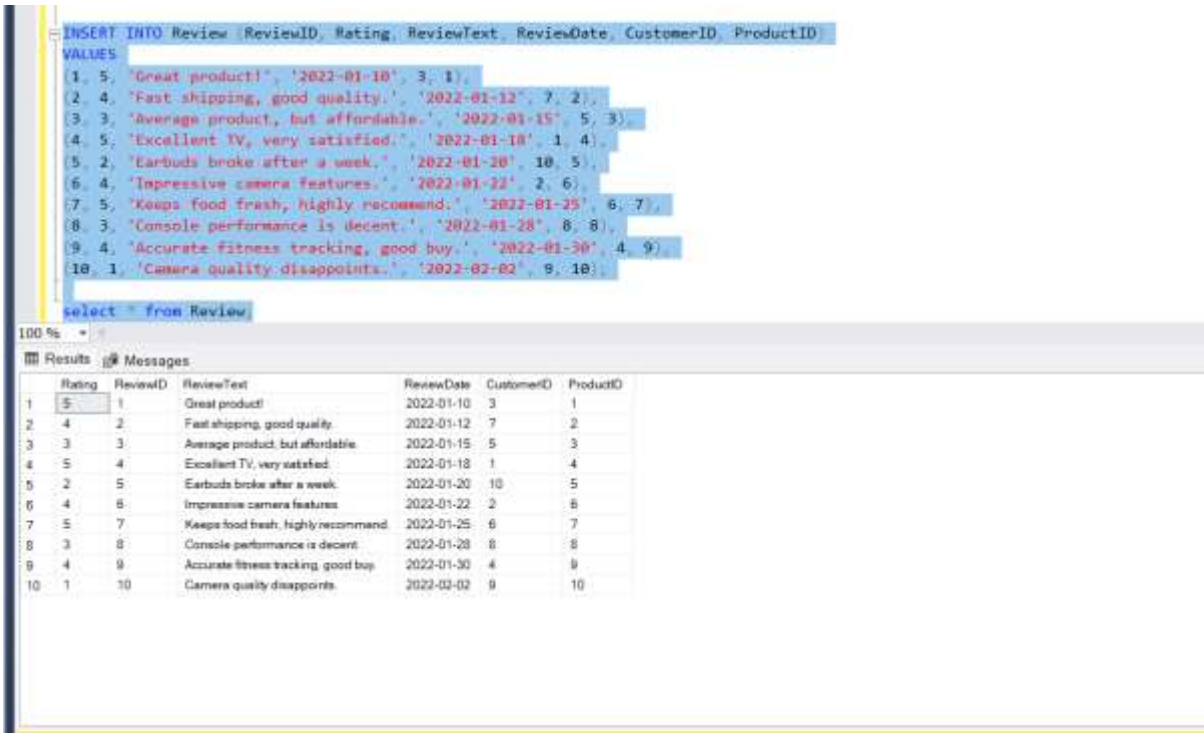
	ProductID	Weight	ProductName	StockQuantity	Description	Price	Dimension	VendorID
1	1	2.50	Laptop Pro	50	High-performance laptop	1200.00	15x10x1	3
2	2	0.80	Smartphone X	100	Latest smartphone model	800.00	6x3x0.5	7
3	3	1.20	Coffee Maker Deluxe	30	Premium coffee maker	150.00	12x8x8	5
4	4	3.00	Ultra HD TV	20	Large-screen Ultra HD television	2000.00	50x30x5	1
5	5	0.50	Bluetooth Earbuds	80	Wireless Bluetooth earbuds	50.00	2x2x1	10
6	6	1.50	Home Security Camera	40	Smart home security camera	120.00	5x5x3	2
7	7	4.00	Refrigerator Deluxe	10	Large-capacity refrigerator	800.00	36x24x30	6
8	8	2.20	Gaming Console X	25	Next-gen gaming console	400.00	12x8x4	8
9	9	0.30	Fitness Tracker Pro	60	Advanced fitness tracker	70.00	2x1x0.5	4
10	10	1.00	Digital Camera	15	High-resolution digital camera	300.00	4x3x2	9

INSERT INTO Review (ReviewID, Rating, ReviewText, ReviewDate, CustomerID, ProductID)
VALUES
(1, 5, 'Great product!', '2022-01-10', 3, 1),
(2, 4, 'Fast shipping, good quality.', '2022-01-12', 7, 2),
(3, 3, 'Average product, but affordable.', '2022-01-15', 5, 3),
(4, 5, 'Excellent TV, very satisfied.', '2022-01-18', 1, 4),
(5, 2, 'Earbuds broke after a week.', '2022-01-20', 10, 5),
(6, 4, 'Impressive camera features.', '2022-01-22', 2, 6),
(7, 5, 'Keeps food fresh, highly recommend.', '2022-01-25', 6, 7),
(8, 3, 'Console performance is decent.', '2022-01-28', 8, 8),

(9, 4, 'Accurate fitness tracking, good buy.', '2022-01-30', 4, 9),

(10, 1, 'Camera quality disappoints.', '2022-02-02', 9, 10);

select * from Review;



```
INSERT INTO Review (ReviewID, Rating, ReviewText, ReviewDate, CustomerID, ProductID)
VALUES
(1, 5, 'Great product!', '2022-01-10', 3, 1),
(2, 4, 'Fast shipping, good quality.', '2022-01-12', 7, 2),
(3, 3, 'Average product, but affordable.', '2022-01-15', 5, 3),
(4, 5, 'Excellent TV, very satisfied.', '2022-01-18', 1, 4),
(5, 2, 'Earbuds broke after a week.', '2022-01-20', 10, 5),
(6, 4, 'Impressive camera features.', '2022-01-22', 2, 6),
(7, 5, 'Keeps food fresh, highly recommend.', '2022-01-25', 6, 7),
(8, 3, 'Console performance is decent.', '2022-01-28', 8, 8),
(9, 4, 'Accurate fitness tracking, good buy.', '2022-01-30', 4, 9),
(10, 1, 'Camera quality disappoints.', '2022-02-02', 9, 10);

select * from Review;
```

100 %

Results | Messages

	Rating	ReviewID	ReviewText	ReviewDate	CustomerID	ProductID
1	5	1	Great product!	2022-01-10	3	1
2	4	2	Fast shipping, good quality.	2022-01-12	7	2
3	3	3	Average product, but affordable.	2022-01-15	5	3
4	5	4	Excellent TV, very satisfied.	2022-01-18	1	4
5	2	5	Earbuds broke after a week.	2022-01-20	10	5
6	4	6	Impressive camera features.	2022-01-22	2	6
7	5	7	Keeps food fresh, highly recommend.	2022-01-25	6	7
8	3	8	Console performance is decent.	2022-01-28	8	8
9	4	9	Accurate fitness tracking, good buy.	2022-01-30	4	9
10	1	10	Camera quality disappoints.	2022-02-02	9	10

INSERT INTO OrderItem (OrderItemID, Quantity, Subtotal, OrderID, ProductID)

VALUES

(1, 2, 2400.00, 1, 5),

(2, 3, 2400.00, 2, 8),

(3, 1, 150.00, 3, 3),

(4, 2, 4000.00, 4, 7),

(5, 5, 250.00, 5, 2),

(6, 1, 120.00, 6, 10),

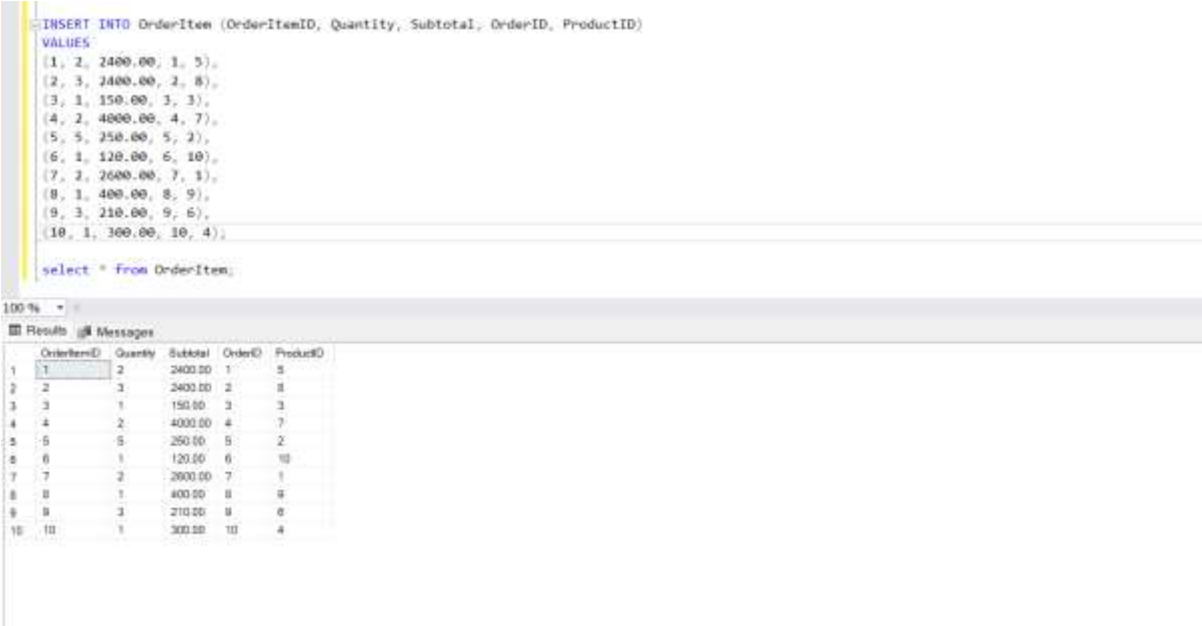
(7, 2, 2600.00, 7, 1),

(8, 1, 400.00, 8, 9),

(9, 3, 210.00, 9, 6),

(10, 1, 300.00, 10, 4);

select * from OrderItem;



```sql
INSERT INTO OrderItem (OrderItemID, Quantity, Subtotal, OrderID, ProductID)
VALUES
(1, 2, 2400.00, 1, 5),
(2, 3, 2400.00, 2, 8),
(3, 1, 150.00, 3, 3),
(4, 2, 4000.00, 4, 7),
(5, 5, 250.00, 5, 2),
(6, 1, 120.00, 6, 10),
(7, 2, 2600.00, 7, 1),
(8, 1, 400.00, 8, 9),
(9, 3, 210.00, 9, 6),
(10, 1, 300.00, 10, 4);

select * from OrderItem;
```

100 %

Results Messages

	OrderItemID	Quantity	Subtotal	OrderID	ProductID
1	1	2	2400.00	1	5
2	2	3	2400.00	2	8
3	3	1	150.00	3	3
4	4	2	4000.00	4	7
5	5	5	250.00	5	2
6	6	1	120.00	6	10
7	7	2	2600.00	7	1
8	8	1	400.00	8	9
9	9	3	210.00	9	6
10	10	1	300.00	10	4

Task 3: Explanations with Queries

1. Retrieve Products and Their Average Rating

SELECT

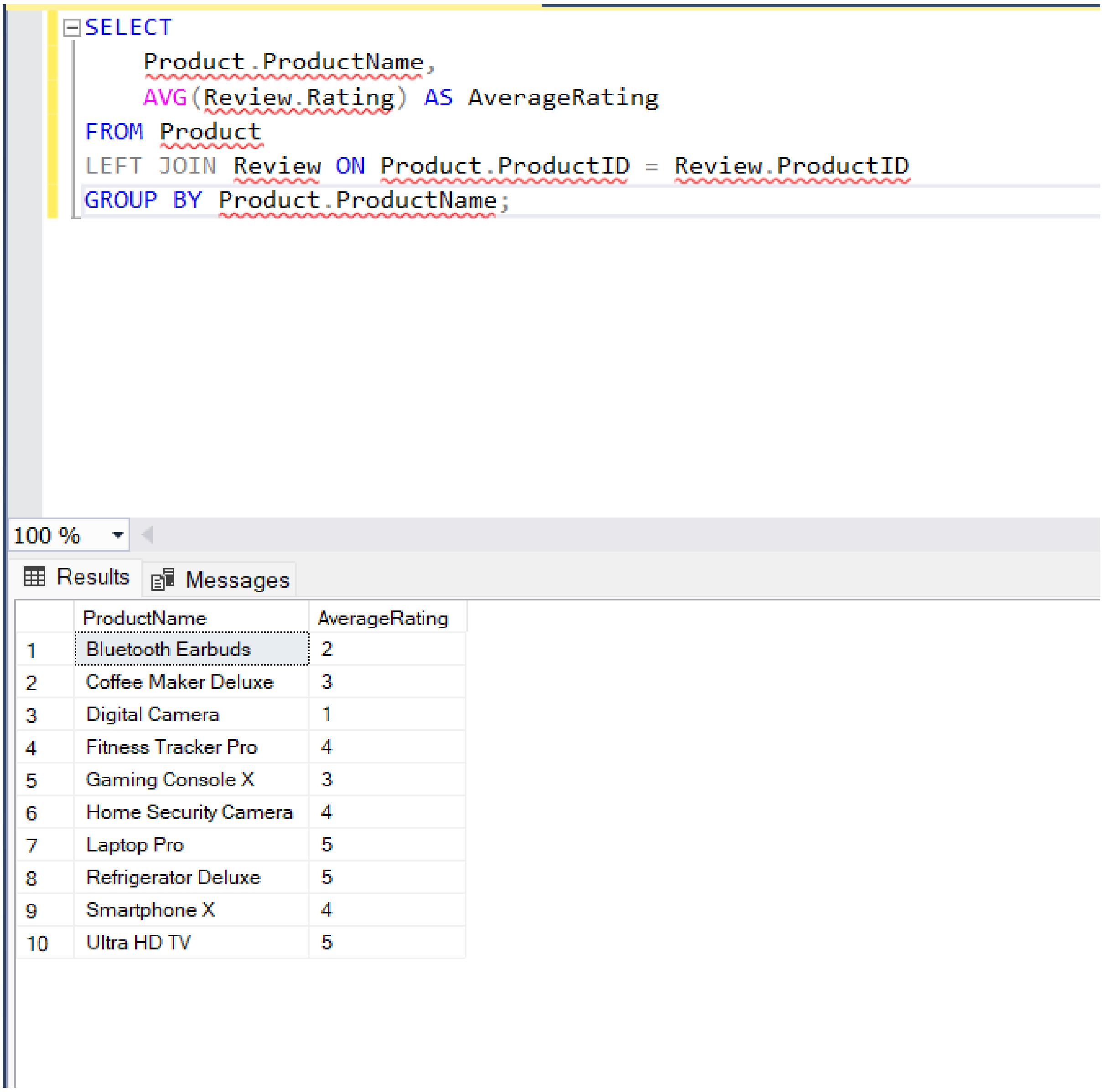
Product.ProductName,

AVG(Review.Rating) AS AverageRating

FROM Product

LEFT JOIN Review ON Product.ProductID = Review.ProductID

GROUP BY Product.ProductName;



```
SELECT
    Product.ProductName,
    AVG(Review.Rating) AS AverageRating
FROM Product
LEFT JOIN Review ON Product.ProductID = Review.ProductID
GROUP BY Product.ProductName;
```

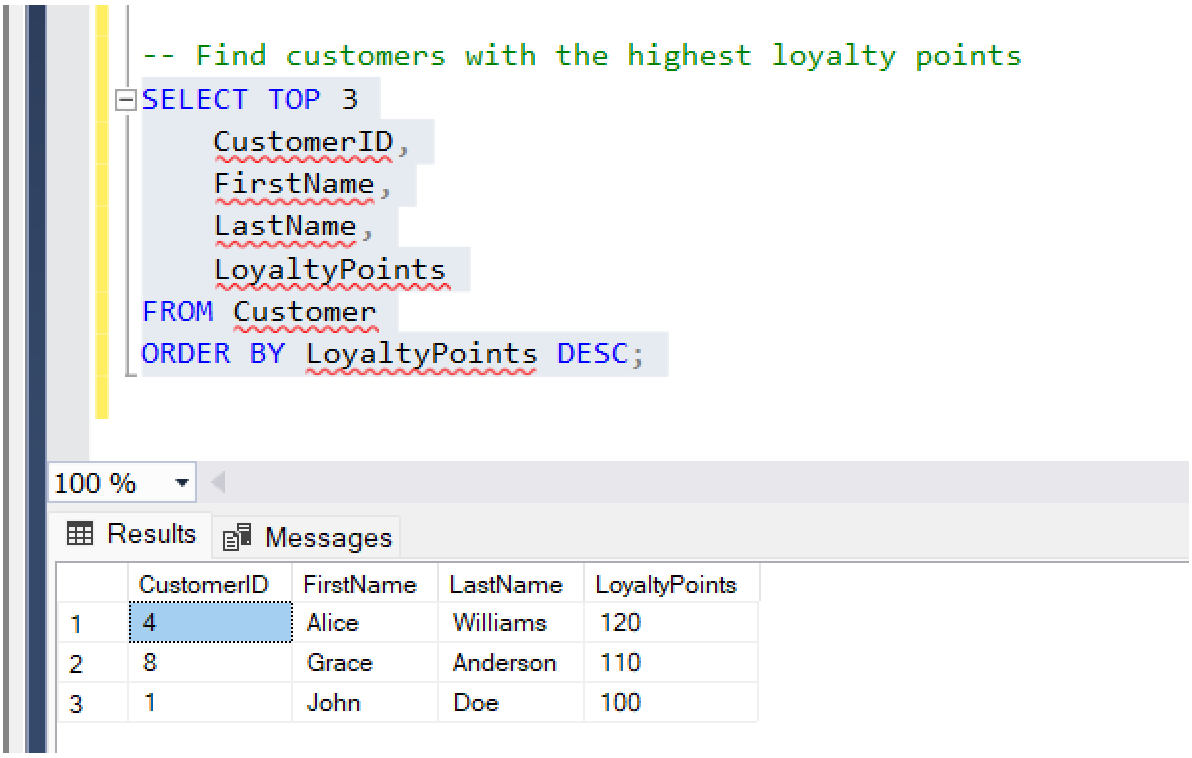
100 %

Results | Messages

	ProductName	AverageRating
1	Bluetooth Earbuds	2
2	Coffee Maker Deluxe	3
3	Digital Camera	1
4	Fitness Tracker Pro	4
5	Gaming Console X	3
6	Home Security Camera	4
7	Laptop Pro	5
8	Refrigerator Deluxe	5
9	Smartphone X	4
10	Ultra HD TV	5

2. Find Customers with the Highest LoyaltyPoints

```
SELECT TOP 3
    CustomerID,
    FirstName,
    LastName,
    LoyaltyPoints
FROM Customer
ORDER BY LoyaltyPoints DESC;
```



3. Identify top 5 Best-Selling Products

```
SELECT TOP 5
    Product.ProductID,
    Product.ProductName,
    SUM(OrderItem.Quantity) AS TotalQuantitySold
FROM Product
JOIN OrderItem ON Product.ProductID = OrderItem.ProductID
GROUP BY Product.ProductID, Product.ProductName
ORDER BY TotalQuantitySold DESC;
```

```sql
-- Identify best-selling products
SELECT TOP 5
    Product.ProductID,
    Product.ProductName,
    SUM(OrderItem.Quantity) AS TotalQuantitySold
FROM Product
JOIN OrderItem ON Product.ProductID = OrderItem.ProductID
GROUP BY Product.ProductID, Product.ProductName
ORDER BY TotalQuantitySold DESC;


-- Calculate revenue by vendor
```

.00 %

Results Messages

	ProductID	ProductName	TotalQuantitySold
1	2	Smartphone X	5
2	6	Home Security Camera	3
3	8	Gaming Console X	3
4	1	Laptop Pro	2
5	5	Bluetooth Earbuds	2

4. Calculate Revenue by Vendor

```
SELECT
    Vendor.VendorID,
    Vendor.VendorName,
    SUM(OrderItem.Subtotal) AS TotalRevenue
FROM Vendor
JOIN Product ON Vendor.VendorID = Product.VendorID
JOIN OrderItem ON Product.ProductID = OrderItem.ProductID
GROUP BY Vendor.VendorID, Vendor.VendorName
ORDER BY TotalRevenue DESC;
```

```sql
-- Calculate revenue by vendor
SELECT
    Vendor.VendorID,
    Vendor.VendorName,
    SUM(OrderItem.Subtotal) AS TotalRevenue
FROM Vendor
JOIN Product ON Vendor.VendorID = Product.VendorID
JOIN OrderItem ON Product.ProductID = OrderItem.ProductID
GROUP BY Vendor.VendorID, Vendor.VendorName
ORDER BY TotalRevenue DESC;
```

100 %

Results | Messages

	VendorID	VendorName	TotalRevenue
1	6	Mobile Solutions	4000.00
2	3	Appliance Center	2600.00
3	8	ABC Suppliers	2400.00
4	10	Future Gadgets	2400.00
5	4	Home Supplies	400.00
6	1	TechCo Electronics	300.00
7	7	Super Tech	250.00
8	2	Gadget World	210.00
9	5	ElectroMart	150.00
10	9	Quick Appliances	120.00

5. Retrieve Customer Reviews

   SELECT

       Review.ReviewID,

       Review.Rating,

       Review.ReviewText,

       Review.ReviewDate,

       Customer.FirstName,

       Customer.LastName,

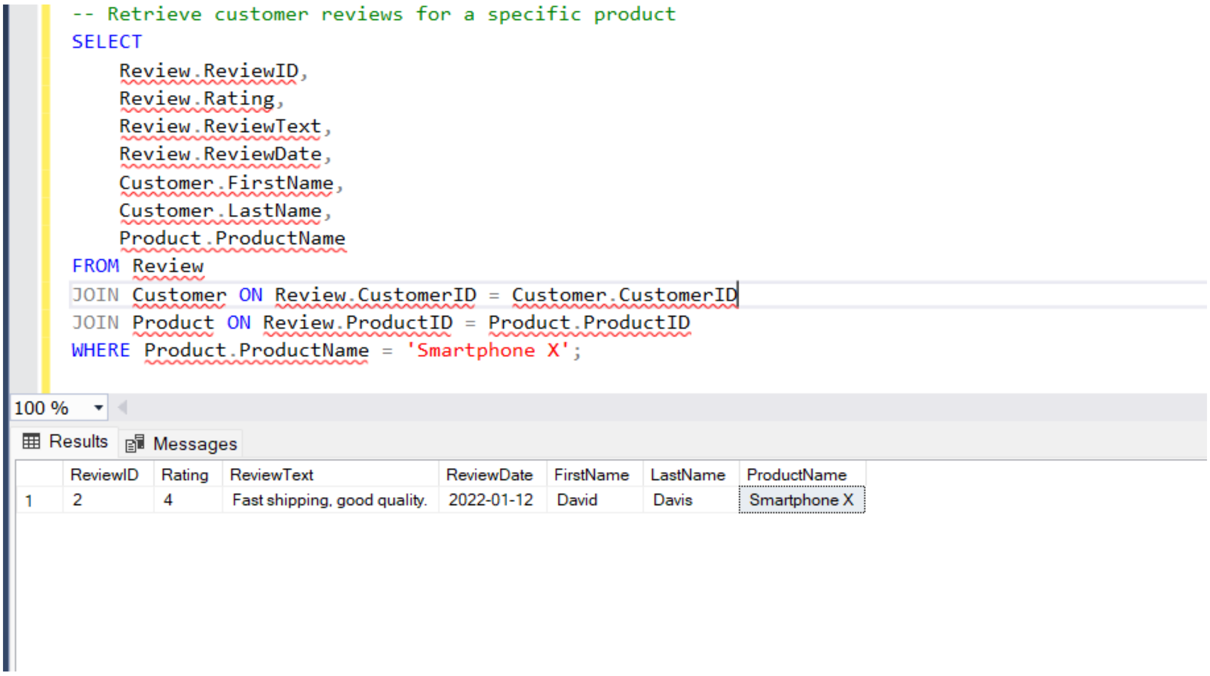
       Product.ProductName

FROM Review

JOIN Customer ON Review.CustomerID = Customer.CustomerID

JOIN Product ON Review.ProductID = Product.ProductID

WHERE Product.ProductName = 'Smartphone X';



```sql
-- Retrieve customer reviews for a specific product
SELECT
    Review.ReviewID,
    Review.Rating,
    Review.ReviewText,
    Review.ReviewDate,
    Customer.FirstName,
    Customer.LastName,
    Product.ProductName
FROM Review
JOIN Customer ON Review.CustomerID = Customer.CustomerID
JOIN Product ON Review.ProductID = Product.ProductID
WHERE Product.ProductName = 'Smartphone X';
```

100 %

Results Messages

	ReviewID	Rating	ReviewText	ReviewDate	FirstName	LastName	ProductName
1	2	4	Fast shipping, good quality.	2022-01-12	David	Davis	Smartphone X