FINAL PROJECT--JANE SANJEEVINI REGIS KUMAR--23205839

TASK 1

Data Set Name	WORK.MYDATA	**Observations**	324
Member Type	DATA	**Variables**	10
Engine	V9	**Indexes**	0
Created	08/14/2024 23:19:30	**Observation Length**	134
Last Modified	08/14/2024 23:19:30	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	976
Obs in First Data Page	324
Number of Data Set Repairs	0
Filename	/saswork/SAS_work047800019FE3_odaws02-euw1.oda.sas.com/SAS_work5AF100019FE3_odaws02-euw1.oda.sas.com/mydata.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	1610716770
Access Permission	rw-r--r--
Owner Name	u63920390
File Size	256KB
File Size (bytes)	262144

Alphabetic List of Variables and Attributes					
#	**Variable**	**Type**	**Len**	**Format**	**Informat**
5	C03151V03803	Char	4	$4.	$4.
7	C03440V04149	Char	4	$4.	$4.
8	Energy Rating	Char	17	$17.	$17.
1	STATISTIC	Char	13	$13.	$13.
2	Statistic Label	Char	35	$35.	$35.
3	TLIST(A1)	Char	6	$6.	$6.
6	Type of Dwelling	Char	22	$22.	$22.
9	UNIT	Char	21	$21.	$21.
10	VALUE	Char	6	$6.	$6.
4	Year	Char	6	$6.	$6.

Obs	STATISTIC	Statistic Label	TLIST(A1)	Year	C03151V03803	Type of Dwelling	C03440V04149	Energy Rating	UNIT	VALUE
1	DBEREL01C01	Household Electricity Consumption	2015	2015	01	Apartment	13	A - B	Mean kilowatt-hours	4832
2	DBEREL01C01	Household Electricity Consumption	2015	2015	01	Apartment	07	C	Mean kilowatt-hours	5420
3	DBEREL01C01	Household Electricity Consumption	2015	2015	01	Apartment	08	D	Mean kilowatt-hours	5230
4	DBEREL01C01	Household Electricity Consumption	2015	2015	01	Apartment	09	E	Mean kilowatt-hours	4827
5	DBEREL01C01	Household Electricity Consumption	2015	2015	01	Apartment	10	F- G	Mean kilowatt-hours	3908
6	DBEREL01C01	Household Electricity Consumption	2015	2015	01	Apartment	-	All BER ratings	Mean kilowatt-hours	5056
7	DBEREL01C01	Household Electricity Consumption	2015	2015	11	Mid-terrace house	13	A - B	Mean kilowatt-hours	3547
8	DBEREL01C01	Household Electricity Consumption	2015	2015	11	Mid-terrace house	07	C	Mean kilowatt-hours	5988
9	DBEREL01C01	Household Electricity Consumption	2015	2015	11	Mid-terrace house	08	D	Mean kilowatt-hours	5501
10	DBEREL01C01	Household Electricity Consumption	2015	2015	11	Mid-terrace house	09	E	Mean kilowatt-hours	4949

Analysis Variable : VALUE_num					
N	**Mean**	**Median**	**Std Dev**	**Minimum**	**Maximum**
315	5180.60	5077.00	1048.29	2977.00	8868.00

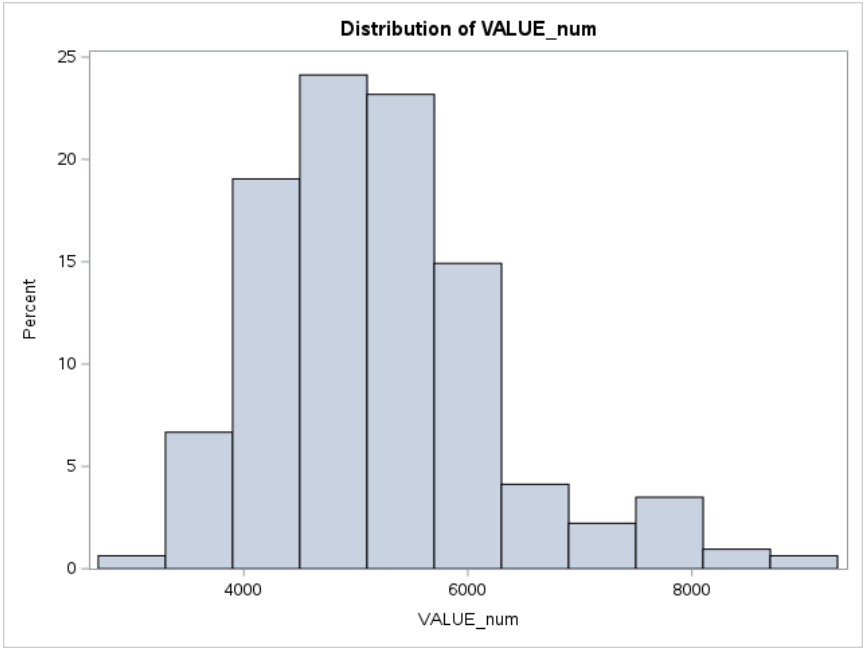
STATISTIC	Frequency	Percent	Cumulative Frequency	Cumulative Percent
DBEREL01C01	324	100.00	324	100.00

Statistic Label	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Household Electricity Consumption	324	100.00	324	100.00

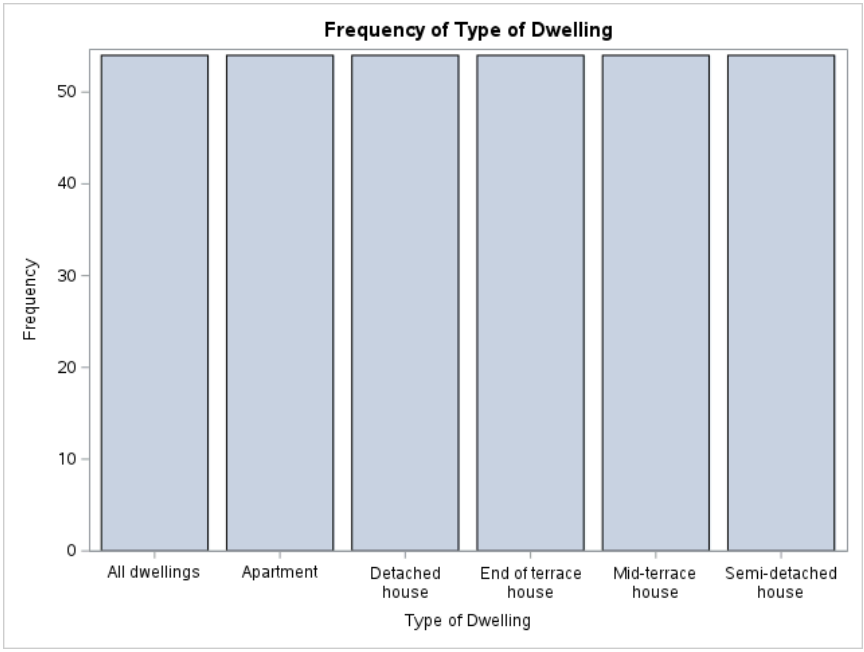
Type of Dwelling	Frequency	Percent	Cumulative Frequency	Cumulative Percent
All dwellings	54	16.67	54	16.67
Apartment	54	16.67	108	33.33

Type of Dwelling	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Detached house	54	16.67	162	50.00
End of terrace house	54	16.67	216	66.67
Mid-terrace house	54	16.67	270	83.33
Semi-detached house	54	16.67	324	100.00

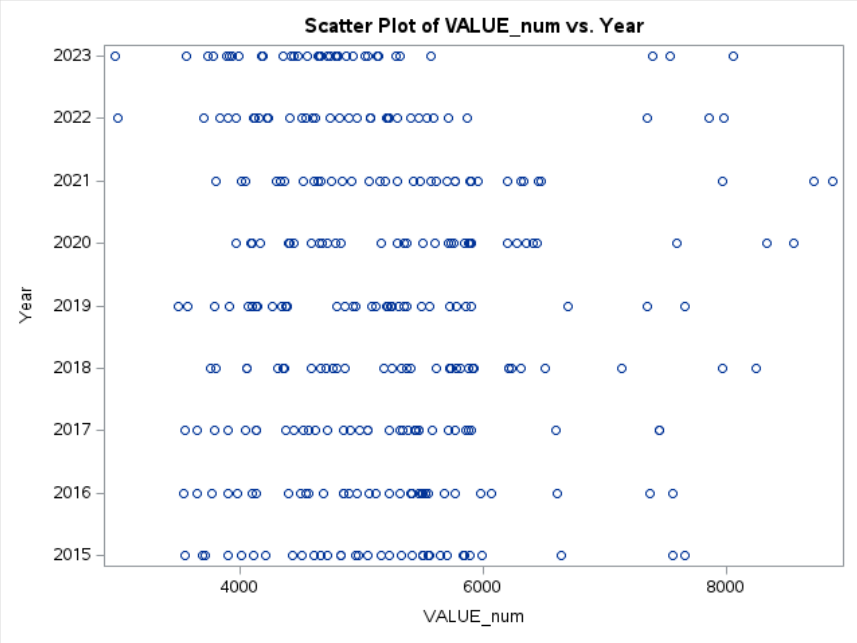
Energy Rating	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A - B	54	16.67	54	16.67
All BER ratings	54	16.67	108	33.33
C	54	16.67	162	50.00
D	54	16.67	216	66.67
E	54	16.67	270	83.33
F- G	54	16.67	324	100.00



The histogram of VALUE_num suggests a roughly normal distribution, centered around 6000 with values ranging from 4000 to 8000. The distribution appears symmetrical, indicating that the mean, median, and mode are likely close. Most observations cluster near the center, with fewer data points in the tails. There are no obvious outliers, and the data is concentrated in the middle bins.



The bar chart titled Frequency of Type of Dwelling shows that all dwelling types (Apartment, Detached house, End of terrace house, Mid-terrace house, Semi-detached house) have an identical frequency of 50. The equal height of the bars indicates an even distribution of these dwelling types within the area or population studied, with no single dwelling type being more common than others. This suggests a balanced representation of all dwelling types in the dataset.



The scatter plot of VALUE_num versus Year (2015-2023) shows a range of values from 4000 to 8000 without a clear trend over time. The data points are scattered, indicating no strong linear relationship between VALUE_num and Year. While some years show clustering within certain VALUE_num ranges, this is not consistent. A few outliers are present, suggesting unusual observations. Overall, there appears to be no strong correlation between the variables, and further analysis may be needed to explore any underlying patterns.

Cross-tabulation of Type of Dwelling by Energy Rating

Frequency
Percent
Row Pct
Col Pct

Table of Type of Dwelling by Energy Rating

Type of Dwelling	Energy Rating						
	A - B	All BER ratings	C	D	E	F- G	Total
All dwellings	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	54 16.67
Apartment	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	54 16.67
Detached house	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	54 16.67
End of terrace house	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	54 16.67
Mid-terrace house	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	54 16.67
Semi-detached house	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	9 2.78 16.67 16.67	54 16.67
Total	54 16.67	54 16.67	54 16.67	54 16.67	54 16.67	54 16.67	324 100.00

Statistics for Table of Type of Dwelling by Energy Rating

Statistic	DF	Value	Prob
Chi-Square	25	0.0000	1.0000
Likelihood Ratio Chi-Square	25	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0000	1.0000
Phi Coefficient		0.0000	
Contingency Coefficient		0.0000	
Cramer's V		0.0000	

Sample Size = 324

Summary Statistics of VALUE_num by Type of Dwelling

Analysis Variable : VALUE_num						
Type of Dwelling	N Obs	N	Mean	Std Dev	Minimum	Maximum
All dwellings	54	45	5106.53	578.2548039	3783.00	5900.00
Apartment	54	54	4788.54	641.6689534	2977.00	5740.00
Detached house	54	54	6452.00	1388.09	3912.00	8868.00
End of terrace house	54	54	5037.13	819.0988156	3577.00	6506.00
Mid-terrace house	54	54	4744.70	837.5478275	3495.00	6326.00
Semi-detached house	54	54	4942.35	664.3843556	3697.00	6311.00

TASK 2

Summary Statistics of VALUE_num by Type of Dwelling

Obs	university_name	year	world_rank	country	national_rank	quality_of_education	citations	patents	score	award	pub	teaching	international	research	num_students	student_staff_ratio
1	Harvard University	2012	1	USA	1	7	1	5	100	100	100	95.8	67.5	97.4	20152	8.9
2	Harvard University	2013	1	USA	1	1	1	7	100	100	100	94.9	63.7	98.6	20152	8.9
3	Harvard University	2014	1	USA	1	1	1	2	100	100	100	95.3	66.2	98.5	20152	8.9
4	Harvard University	2015	1	USA	1	1	1	3	100	100	100	92.9	67.6	98.6	20152	8.9
5	Stanford University	2013	2	USA	2	11	2	11	93.94	80.7	69.4	95	56.6	98.8	15596	7.8

Summary Statistics of VALUE_num by Type of Dwelling

Data Set Name	WORK.UNIVERSITY	**Observations**	551
Member Type	DATA	**Variables**	16
Engine	V9	**Indexes**	0
Created	08/14/2024 23:19:31	**Observation Length**	160
Last Modified	08/14/2024 23:19:31	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	818
Obs in First Data Page	551
Number of Data Set Repairs	0
Filename	/saswork/SAS_work047800019FE3_odaws02-euw1.oda.sas.com/SAS_work5AF100019FE3_odaws02-euw1.oda.sas.com/university.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	1610716806
Access Permission	rw-r--r--
Owner Name	u63920390
File Size	256KB
File Size (bytes)	262144

Variables in Creation Order					
#	**Variable**	**Type**	**Len**	**Format**	**Informat**
1	university_name	Char	34	$34.	$34.
2	year	Num	8	BEST12.	BEST32.
3	world_rank	Num	8	BEST12.	BEST32.
4	country	Char	14	$14.	$14.
5	national_rank	Num	8	BEST12.	BEST32.
6	quality_of_education	Num	8	BEST12.	BEST32.
7	citations	Num	8	BEST12.	BEST32.
8	patents	Num	8	BEST12.	BEST32.
9	score	Num	8	BEST12.	BEST32.
10	award	Num	8	BEST12.	BEST32.
11	pub	Num	8	BEST12.	BEST32.
12	teaching	Num	8	BEST12.	BEST32.
13	international	Num	8	BEST12.	BEST32.
14	research	Num	8	BEST12.	BEST32.
15	num_students	Num	8	BEST12.	BEST32.
16	student_staff_ratio	Num	8	BEST12.	BEST32.

Summary Statistics of VALUE_num by Type of Dwelling

Analysis Variable : student_staff_ratio				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
543	15.9902394	10.2271127	2.9000000	70.4000000

Summary Statistics of VALUE_num by Type of Dwelling

Variable: num_students

Moments			
N	543	**Sum Weights**	543
Mean	24504.5175	**Sum Observations**	13305953
Std Deviation	14091.3492	**Variance**	198566122
Skewness	1.73004778	**Kurtosis**	5.91701474

Moments			
Uncorrected SS	4.33679E11	Corrected SS	1.07623E11
Coeff Variation	57.5051078	Std Error Mean	604.717675

Basic Statistical Measures			
Location		Variability	
Mean	24504.52	Std Deviation	14091
Median	22578.00	Variance	198566122
Mode	2243.00	Range	118743
		Interquartile Range	15554

Note: The mode displayed is the smallest of 45 modes with a count of 4.

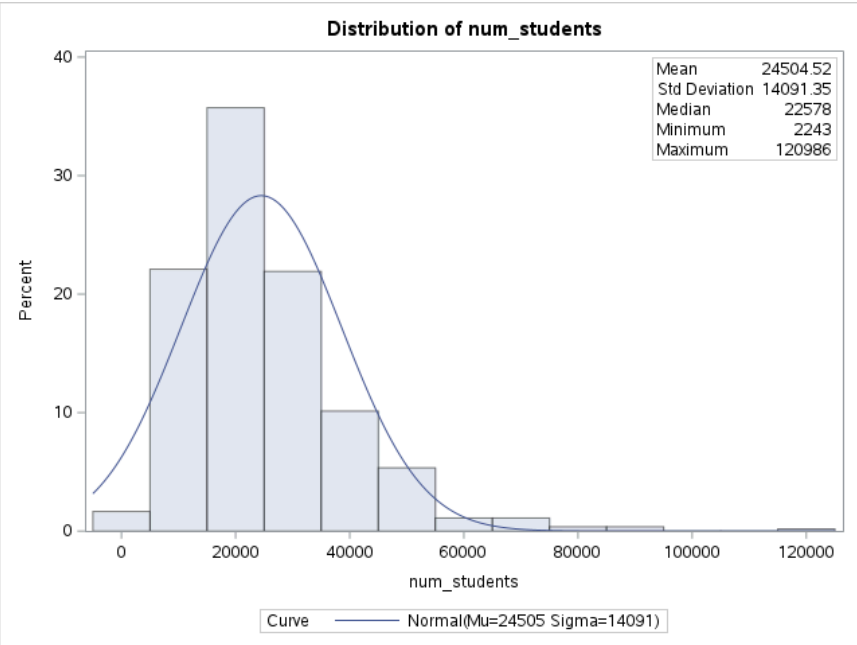
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	40.52224	Pr > |t|	<.0001
Sign	M	271.5	Pr >= |M|	<.0001
Signed Rank	S	73848	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	120986
99%	67552
95%	50152
90%	41868
75% Q3	30726
50% Median	22578
25% Q1	15172
10%	9586
5%	7426
1%	3055
0% Min	2243

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
2243	41	83236	216
2243	40	83236	228
2243	36	85532	346
2243	13	85532	358
3055	319	120986	239

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	8	1.45	100.00

Summary Statistics of VALUE_num by Type of Dwelling



Summary Statistics of VALUE_num by Type of Dwelling

Fitted Normal Distribution for num_students

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	24504.52

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Std Dev	Sigma	14091.35

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.1254493	Pr > D	<0.010
Cramer-von Mises	W-Sq	1.8430084	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	11.2712362	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	3055.00	-8276.86
5.0	7426.00	1326.31
10.0	9586.00	6445.73
25.0	15172.00	15000.05
50.0	22578.00	24504.52
75.0	30726.00	34008.99
90.0	41868.00	42563.31
95.0	50152.00	47682.72
99.0	67552.00	57285.90

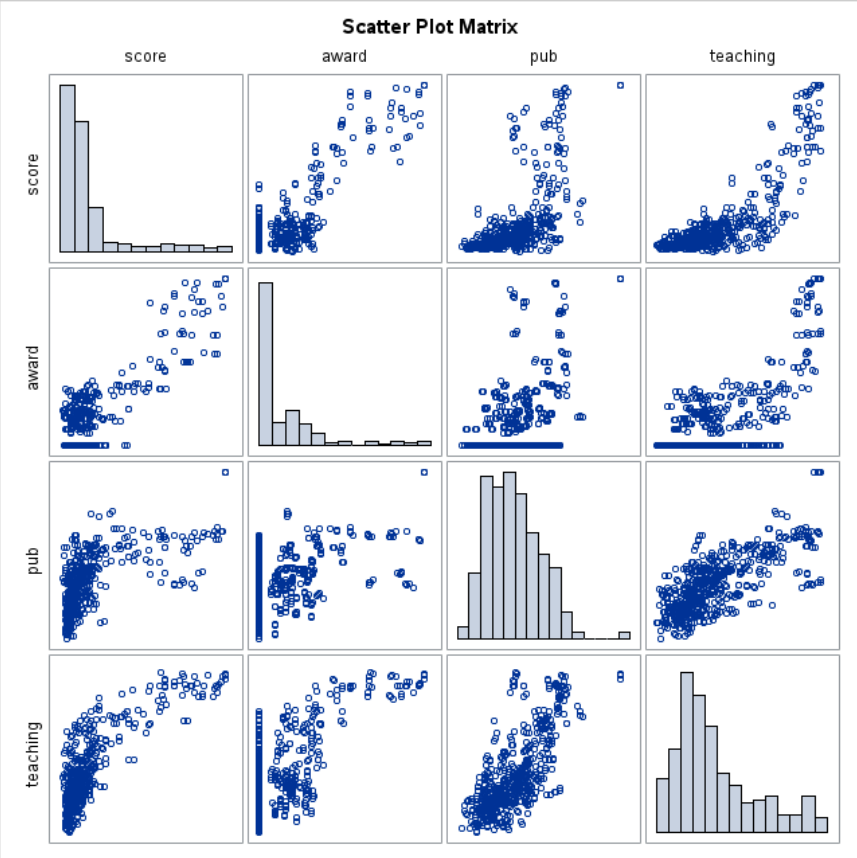
The histogram shows a bell-shaped distribution of num_students, centered around 25,000. The data spans from 0 to 120,000, with most observations clustered near the center. The overlaid normal curve suggests the data closely follows a normal distribution, with a mean (Mu) of 24,505 and a standard deviation (Sigma) of 14,091. The distribution is symmetrical, indicating that the data is evenly spread around the mean, with fewer observations as you move away from the center.

Summary Statistics of VALUE_num by Type of Dwelling

4 Variables:	score award pub teaching

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
score	551	52.63819	12.00925	29004	43.47000	100.00000
award	551	13.34846	21.80780	7355	0	100.00000
pub	551	44.42250	13.93838	24477	17.10000	100.00000
teaching	551	44.16661	19.89288	24336	15.00000	96.30000

Pearson Correlation Coefficients, N = 551 Prob > |r| under H0: Rho=0				
	score	award	pub	teaching
score	1.00000	0.86233 <.0001	0.64115 <.0001	0.82408 <.0001
award	0.86233 <.0001	1.00000	0.52702 <.0001	0.73071 <.0001
pub	0.64115 <.0001	0.52702 <.0001	1.00000	0.73511 <.0001
teaching	0.82408 <.0001	0.73071 <.0001	0.73511 <.0001	1.00000



Summary Statistics of VALUE_num by Type of Dwelling

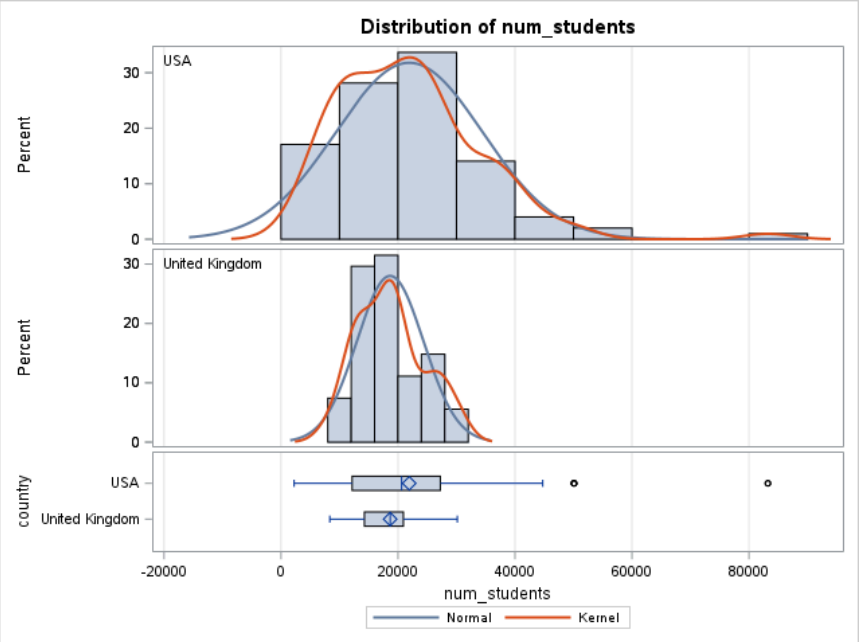
Variable: num_students

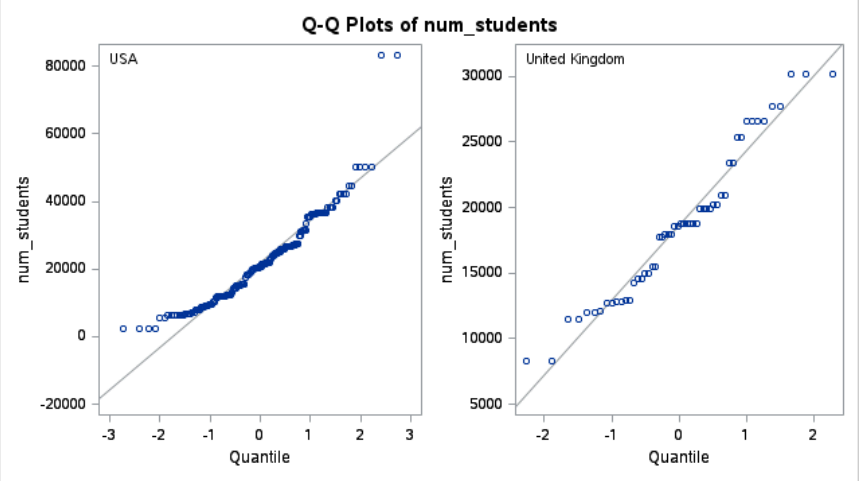
country	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
USA		199	21920.1	12548.1	889.5	2243.0	83236.0
United Kingdom		54	18658.9	5698.3	775.4	8338.0	30144.0
Diff (1-2)	Pooled		3261.2	11448.3	1756.6		
Diff (1-2)	Satterthwaite		3261.2		1180.1		

country	Method	Mean	99% CL Mean		Std Dev	99% CL Std Dev	
USA		21920.1	19606.6	24233.7	12548.1	11100.5	14392.7
United Kingdom		18658.9	16587.1	20730.8	5698.3	4546.5	7545.0
Diff (1-2)	Pooled	3261.2	-1298.2	7820.6	11448.3	10260.7	12920.9
Diff (1-2)	Satterthwaite	3261.2	191.4	6331.0			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	251	1.86	0.0646
Satterthwaite	Unequal	194.23	2.76	0.0063

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	198	53	4.85	<.0001





The Q-Q plots indicate that the distribution of num_students is not normal for both the USA and the UK. The USA shows a stronger right skew with more pronounced deviations from normality, while the UK's deviations are less severe. Histograms, kernel density plots, and box plots reveal that both countries have right-skewed distributions, with the USA having a higher median and greater variation in student numbers. The USA also shows more outliers with exceptionally high student counts compared to the UK. Overall, both countries have a larger proportion of smaller institutions, but the USA has a wider spread and higher median student numbers.

Summary Statistics of VALUE_num by Type of Dwelling

Obs	university_name	year	world_rank	country	national_rank	quality_of_education	citations	patents	score	award	pub	teaching	international	research	num_students	student_staff_ratio
10	University College London	2013	30	United Kingdom	4	24	21	73	56	29.7	67.5	83.5	89	88.8	26607	10.7
11	University College London	2014	30	United Kingdom	3	20	18	121	61.05	29.5	71.6	70.5	90.2	77.5	26607	10.7
12	University College London	2012	31	United Kingdom	4	35	33	86	55.21	30.4	67.1	77.8	91.8	84.3	26607	10.7
13	University of Nottingham	2012	97	United Kingdom	6	101	101	92	43.79	20.6	45.8	40.2	72.6	40	30144	15
14	University of Bonn	2014	98	Germany	3	23	187	227	51.37	19.8	40.5	35.8	54.2	20.4	32474	70.4
15	University of Bristol	2012	98	United Kingdom	7	101	86	101	43.77	16.8	45.3	44.2	73.7	47.7	17906	14
16	Sapienza University of Rome	2015	112	Italy	1	67	212	312	49.97	13.3	52	32.3	37.5	28.1	120986	32.3
17	University of Bristol	2014	123	United Kingdom	8	177	99	338	49.97	16.3	46.3	39.6	75	41.2	17906	14

Summary Statistics of VALUE_num by Type of Dwelling

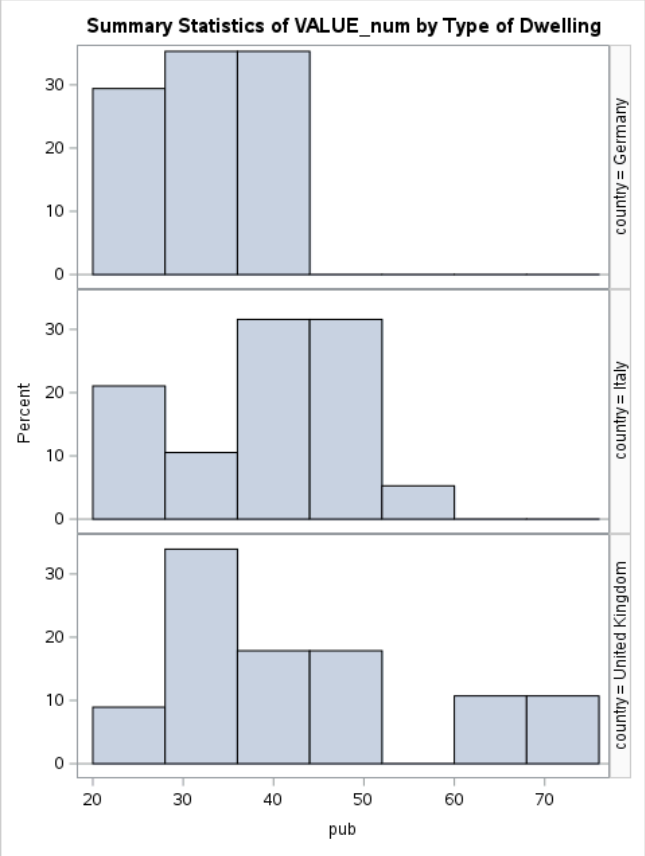
Analysis Variable : quality_of_education
Mean
213.5543478

Summary Statistics of VALUE_num by Type of Dwelling

Analysis Variable : quality_of_education
Mean
266.3661972

Summary Statistics of VALUE_num by Type of Dwelling

Analysis Variable : patents						
country	N Obs	N	Mean	Std Dev	Minimum	Maximum
Germany	17	17	386.4705882	187.5646947	138.0000000	774.0000000
Italy	19	19	532.2105263	121.0980223	312.0000000	737.0000000
United Kingdom	56	56	305.8392857	204.6968292	15.0000000	871.0000000



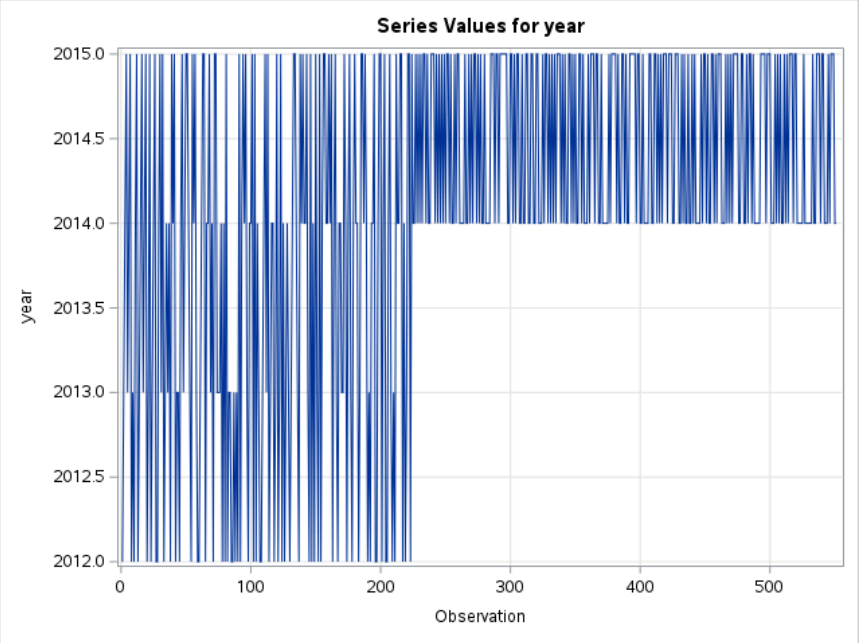
The histograms depict varying distributions of pub values by country. Germany shows a relatively uniform distribution, indicating consistent pub values. Italy's distribution is right-skewed, with a higher concentration of lower values. The United Kingdom has a bimodal distribution, suggesting two distinct clusters of pub values. The UK also exhibits the widest spread, while Germany has the narrowest. These differences highlight distinct patterns in pub values across the three countries.

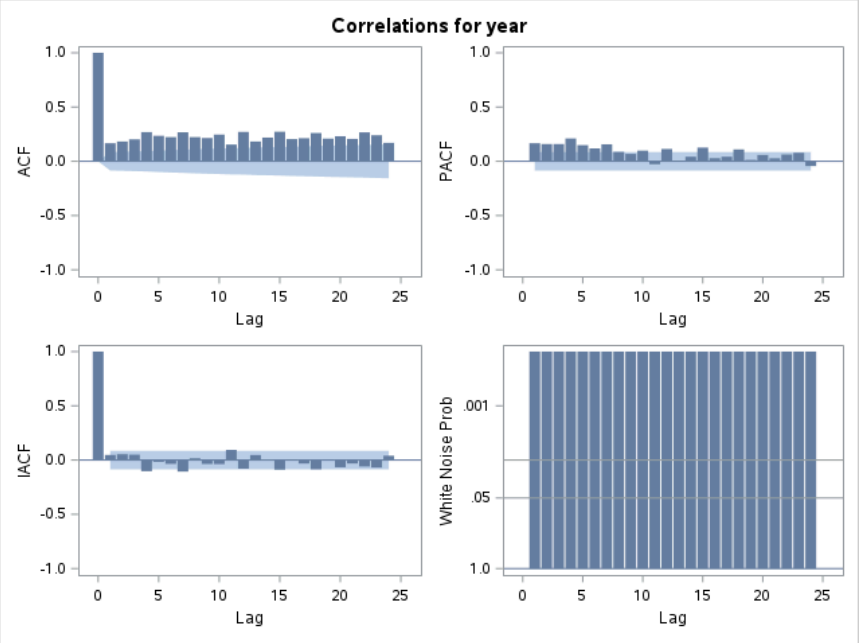
TASK 3

Summary Statistics of VALUE_num by Type of Dwelling

Input Data Set	
Name	WORK.UNIVERSITY
Label	
Length of Seasonal Cycle	1

Variable Information	
Name	year
Label	
Number of Observations Read	551





The scatter plot shows Series Values distributed across the years 2012 to 2015, with no clear trend or strong linear relationship between the variables. Data points are scattered, with some clustering in specific year ranges but without consistency. A few outliers are present, indicating unusual observations. Overall, the plot suggests no strong correlation between Series Values and year, and further analysis might be needed to explore any underlying patterns.The ACF and PACF plots reveal significant spikes at certain lags, indicating potential autocorrelation in the time series data, which suggests that past values may influence future values. The IACF plot, although less commonly used, also shows some structure, supporting the presence of autocorrelation. The White Noise Probability plot indicates that the series is unlikely to be white noise, as p-values are below the significance level, suggesting the data has an underlying structure rather than being purely random. This analysis hints at the need for a more sophisticated time series model, such as AR, MA, or ARIMA, to capture the patterns in the data.