Data Programming with SAS Final Project

Dev Walia
Student No. 23205184

Data Analysis I: Import and Preview Online UK Retail Data

Obs	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12JAN10:08:26:00	2.55	17850	United Kingdom
2	536365	71053	WHITE METAL LANTERN	6	12JAN10:08:26:00	3.39	17850	United Kingdom
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12JAN10:08:26:00	2.75	17850	United Kingdom
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12JAN10:08:26:00	3.39	17850	United Kingdom
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12JAN10:08:26:00	3.39	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12JAN10:08:26:00	7.65	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12JAN10:08:26:00	4.25	17850	United Kingdom
8	536366	22633	HAND WARMER UNION JACK	6	12JAN10:08:28:00	1.85	17850	United Kingdom
9	536366	22632	HAND WARMER RED POLKA DOT	6	12JAN10:08:28:00	1.85	17850	United Kingdom
10	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12JAN10:08:34:00	1.69	13047	United Kingdom

The CSV file has been read in successfully using a PROC import step for Data Analysis 1. Printing the first few rows to confirm this.

Detailed structure and Types of Data

The CONTENTS Procedure

Data Set Name	WORK.ONLINERETAIL	Observations	541909
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	08/14/2024 21:34:14	Observation Length	96
Last Modified	08/14/2024 21:34:14	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	398
First Data Page	1
Max Obs per Page	1363
Obs in First Data Page	1333
Number of Data Set Repairs	0
Filename	/saswork/SAS_work390200014C63_odaws02-euw1.oda.sas.com/SAS_work797900014C63_odaws02-

Engine/Host Dependent Information	
	euw1.oda.sas.com/onlineretail.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	1610688428
Access Permission	rw-r--r--
Owner Name	u63898820
File Size	50MB
File Size (bytes)	52297728

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
8	Country	Char	14	$14.	$14.
7	CustomerID	Num	8	BEST12.	BEST32.
3	Description	Char	35	$35.	$35.
5	InvoiceDate	Num	8	DATETIME.	ANYDTDTM40.
1	InvoiceNo	Num	8	BEST12.	BEST32.
4	Quantity	Num	8	BEST12.	BEST32.
2	StockCode	Char	6	$6.	$6.
6	UnitPrice	Num	8	BEST12.	BEST32.

The dataset includes categorical variables like InvoiceNo, StockCode, Description, CustomerID, and Country, along with numerical variables such as Quantity, InvoiceDate, and UnitPrice, which capture transaction details and product information in an online retail context.

Summary statistics for all numerical variables

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Quantity	541909	9.5522495	218.0811579	-80995.00	1.0000000	3.0000000	10.0000000	80995.00
UnitPrice	541909	4.6111136	96.7598531	-11062.06	1.2500000	2.0800000	4.1300000	38970.00

The statistical summary reveals substantial variability in `Quantity` and `UnitPrice` with extremes suggesting data quality issues, including negative values. Most transactions involve small quantities and lower-priced items, with the data showing a skewed distribution towards a few high-value transactions. This suggests a need for data cleaning and potential adjustments in inventory and pricing strategies.

Preview of Cleaned and Processed Retail Data

Obs	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalRevenue	NewCustomerID
1	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12JAN10:08:26:00	4.25	17850	United Kingdom	$25.50	17850
2	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12JAN10:08:26:00	7.65	17850	United Kingdom	$15.30	17850
3	536365	71053	WHITE METAL LANTERN	6	12JAN10:08:26:00	3.39	17850	United Kingdom	$20.34	17850

Obs	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalRevenue	NewCustomerID
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12JAN10:08:26:00	3.39	17850	United Kingdom	$20.34	17850
5	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12JAN10:08:26:00	3.39	17850	United Kingdom	$20.34	17850

The processed dataset now reflects transactions from active sales, with data cleansing eliminating cancellations and data errors, ensuring each record is from a unique transaction with valid quantities and prices, resulting in a consistent dataset ready for further analysis.

Summary Statistics for Quantity, UnitPrice, and TotalRevenue

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Quantity	520756	10.7909309	158.3823054	1.0000000	1.0000000	4.0000000	12.0000000	80995.00
UnitPrice	520756	3.8921442	32.4268063	0	1.2500000	2.0800000	4.1300000	13541.33
TotalRevenue	520756	20.3100334	272.2639530	0	3.9000000	9.9500000	17.7000000	168469.60

Summary statistics reveal a mean quantity of approximately 11 per transaction with significant variability, and a moderate average unit price of about 3.89 sterling, reflecting typical retail conditions. Total revenue per transaction averages around 20.31, with values ranging up to nearly 168,650, indicating occasional high-value purchases.

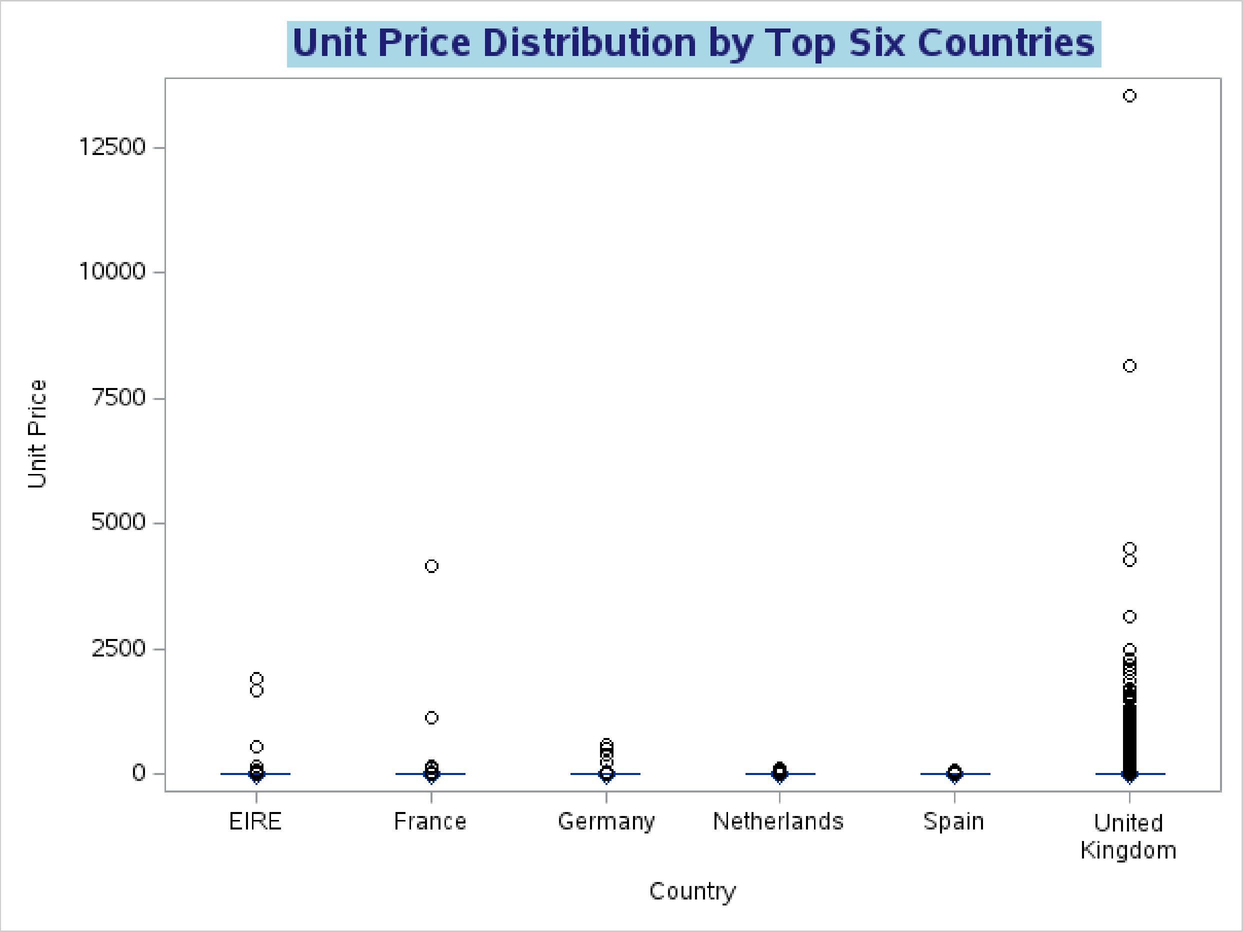
Frequency Distribution of Countries in Data

The FREQ Procedure

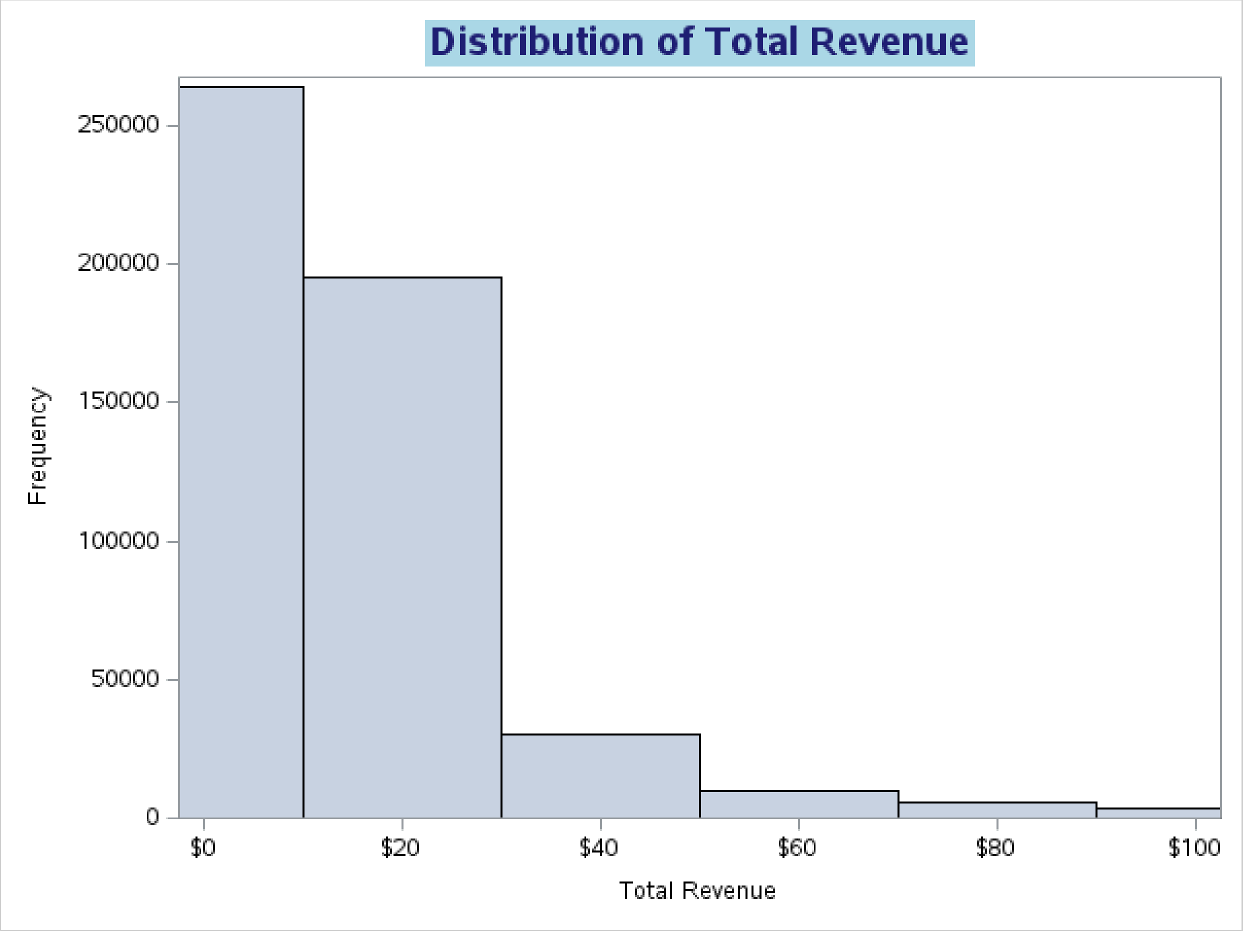
Country	Frequency
Australia	1183
Austria	398
Bahrain	18
Belgium	2031
Brazil	32
Canada	151
Channel Island	747
Cyprus	599
Czech Republic	25
Denmark	379
EIRE	7879
European Commu	60
Finland	681
France	8373
Germany	9015
Greece	145
Hong Kong	274
Iceland	182
Israel	291

Country	Frequency
Italy	758
Japan	321
Lebanon	45
Lithuania	35
Malta	112
Netherlands	2363
Norway	1069
Poland	330
Portugal	1477
RSA	58
Saudi Arabia	9
Singapore	218
Spain	2464
Sweden	450
Switzerland	1945
USA	179
United Arab Em	68
United Kingdom	475959
Unspecified	433

The frequency distribution analysis of the dataset confirms that the majority of transactions are concentrated in six primary countries: United Kingdom, Germany, France, EIRE, Spain, and the Netherlands.

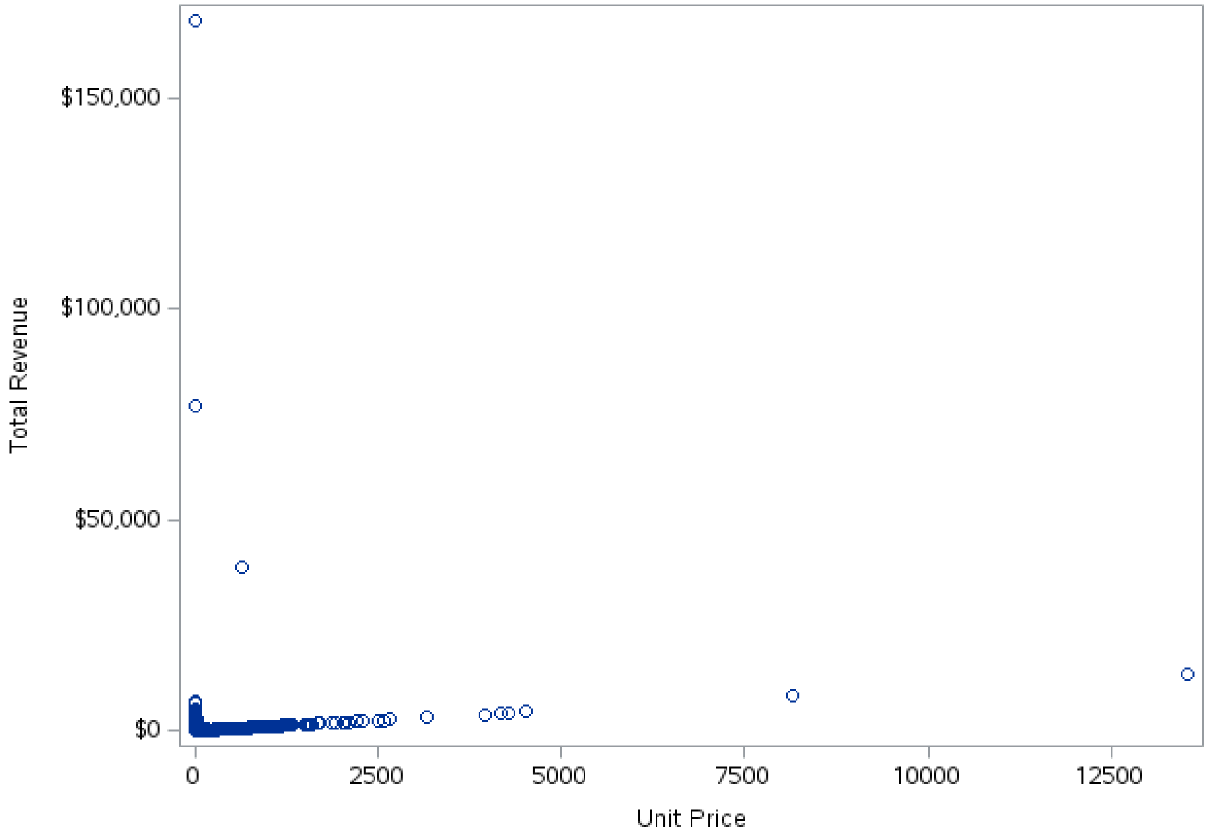


The box plot illustrates that while most countries have similar unit price distributions with few outliers, the United Kingdom exhibits significantly more variation with several high outliers, indicating occasional very high-priced items.



The histogram of Total Revenue shows a heavily right-skewed distribution with most transactions generating less than $20, indicating that lower-priced purchases dominate this retail data set. A significant drop in frequency beyond the $20 mark suggests fewer high-value transactions.

Scatter Plot of Total Revenue vs. Unit Price



The scatter plot reveals that most transactions involve low unit prices and generate modest total revenues, with a few outlier transactions showing exceptionally high total revenues at moderate unit prices, suggesting bulk purchases or high-value sales.

The End For Data Analysis I, I selected the Online UK Retail dataset which includes both categorical and numerical variables such as InvoiceNo, StockCode, Description (categorical), and Quantity, InvoiceDate, UnitPrice (numerical). Using SAS, I successfully imported the data, displayed its structure, and provided descriptive statistics highlighting substantial variability in 'Quantity' and 'UnitPrice'. Further data cleaning processed transactions by removing cancellations and data errors to ensure integrity. I utilized graphical summaries to showcase distributions and relationships, like the histogram of total revenue and box plots for unit prices by country, reflecting the dominant low-value transactions and variation in pricing. This comprehensive analysis utilized SAS functionalities effectively to illustrate the retail dataset's characteristics and dynamics.

Data Analysis II: Q1.a. Preview of University Data: First 5 Observations

Obs	university_name	year	world_rank	country	national_rank
1	Harvard University	2012	1	USA	1
2	Harvard University	2013	1	USA	1
3	Harvard University	2014	1	USA	1
4	Harvard University	2015	1	USA	1
5	Stanford University	2013	2	USA	2

A1.a. Successfully imported university data includes key variables like university name, year, world ranking, country, and national ranking, providing a snapshot of the data's structure for analysis.

Q1. b.Variable Information and Order

The CONTENTS Procedure

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.

A1.b. Displayed here is a sorted list of the variables from the 'university' dataset, detailing attributes such as name, type, and format, which are essential for subsequent data handling and analysis.

Q2. Descriptive Statistics for Student/Staff Ratio

The MEANS Procedure

Analysis Variable : student_staff_ratio

N	Mean	Median	Minimum	Maximum	Std Dev
543	15.99	14.10	2.90	70.40	10.23

A2. The mean of the student/staff ratio is 15.99.

Q3. Univariate Analysis: Number of Students

The UNIVARIATE Procedure
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
Uncorrected SS	4.33679E11	Corrected SS	1.07623E11

Moments			
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

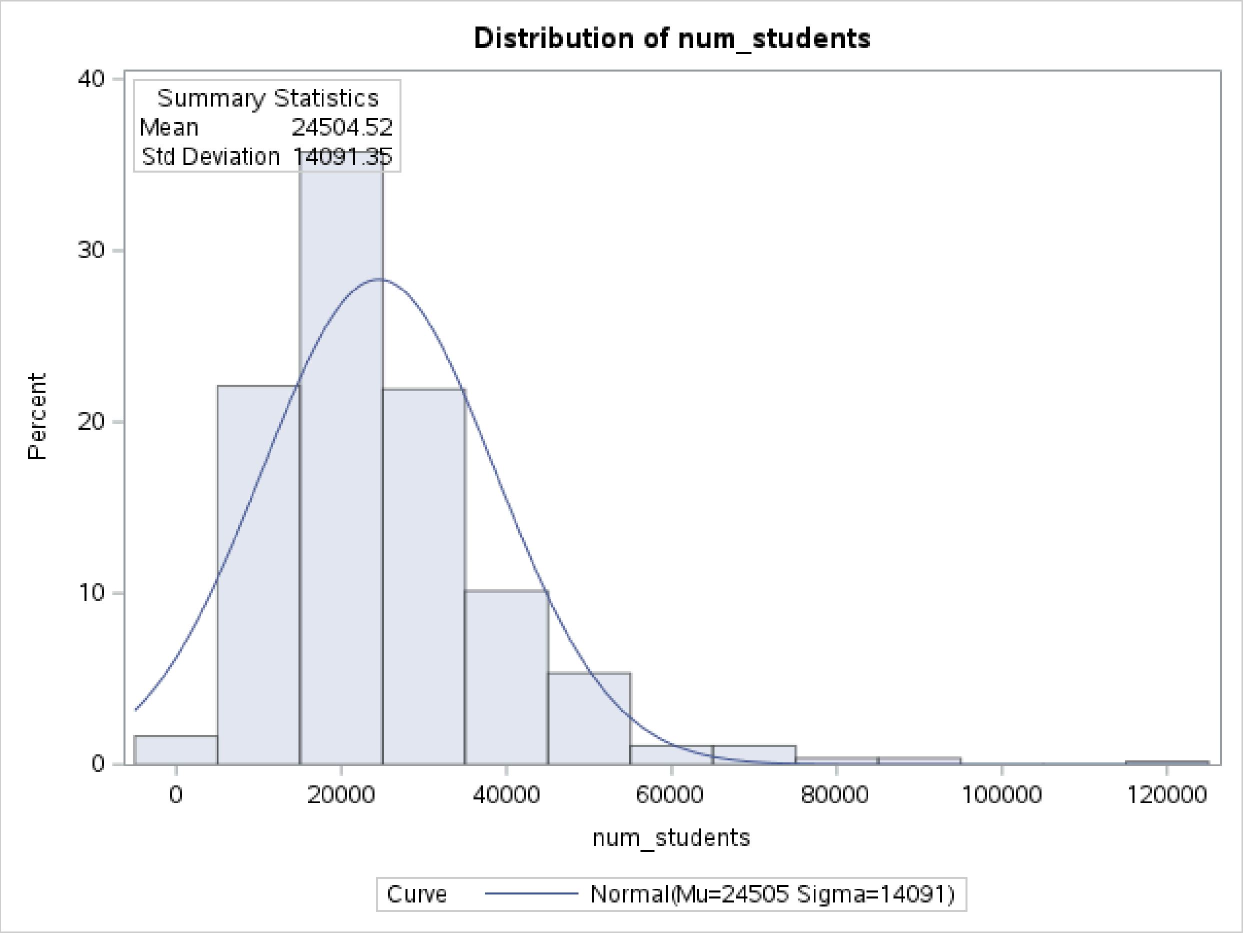
Quantiles (Definition 5)	
Level	Quantile
100% Max	120986
99%	67552
95%	50152
90%	41868
75% Q3	30726
50% Median	22578
25% Q1	15172
10%	9586
5%	7426
1%	3055
0% Min	2243

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

Q3. Univariate Analysis: Number of Students

The UNIVARIATE Procedure



Q3. Univariate Analysis: Number of Students

The UNIVARIATE Procedure
Fitted Normal Distribution for num_students

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	24504.52
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

A3.The univariate analysis for the variable number of students across 543 observations reveals a mean of approximately 24,504, a median of 22,578, and a mode of 2,243 students (the most frequently occurring value, present in only four instances, suggesting multiple modes). The data exhibits significant variability with a standard deviation of about 14,091, a variance of 198,566,122, and a range of 118,743 (between 2,243 and 120,986 students). The interquartile range is 15,554, indicating that the middle 50% of data points are spread across a relatively wide range. The distribution's skewness of 1.73 and a kurtosis of 5.917 suggest a right-skewed and peakier distribution compared to a normal distribution, which is further confirmed by goodness-of-fit tests indicating poor fit to a normal model. These statistical indicators highlight a diverse dataset with a significant spread and multiple peaks in the distribution of the number of students.

Q4. Correlation Analysis Among Measures

The CORR Procedure

4 Variables:	score award pub teaching

Pearson Correlation Coefficients, N = 551

	score	award	pub	teaching
score	1.00000	0.86233	0.64115	0.82408
award	0.86233	1.00000	0.52702	0.73071
pub	0.64115	0.52702	1.00000	0.73511
teaching	0.82408	0.73071	0.73511	1.00000

A4. Correlation analysis explores relationships between university scores, awards, publications, and teaching quality, Yes these correlations statistically significant different from Zero (less than 1 & equal to 1).

Q5. Hypothesis Testing: USA vs UK Universities

The TTEST Procedure

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		

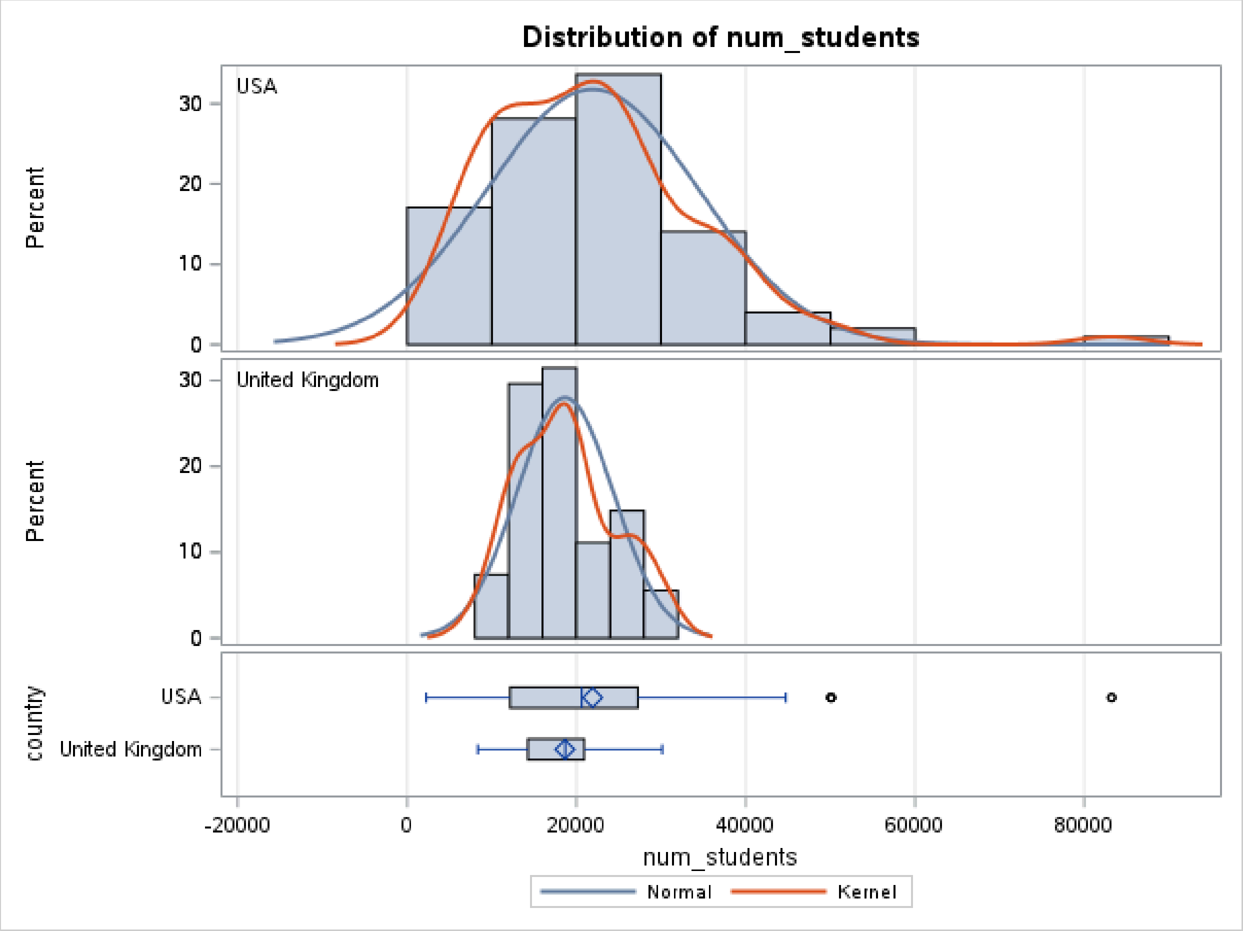
country	Method	Mean	99% CL Mean		Std Dev	99% CL Std Dev	
USA		21920.1	19606.6	24233.7	12548.1	11100.5	14392.7
United Kingdom		18658.9	16587.1	20730.8	5698.3	4546.5	7545.0
Diff (1-2)	**Pooled**	3261.2	-1298.2	7820.6	11448.3	10260.7	12920.9
Diff (1-2)	**Satterthwaite**	3261.2	191.4	6331.0			

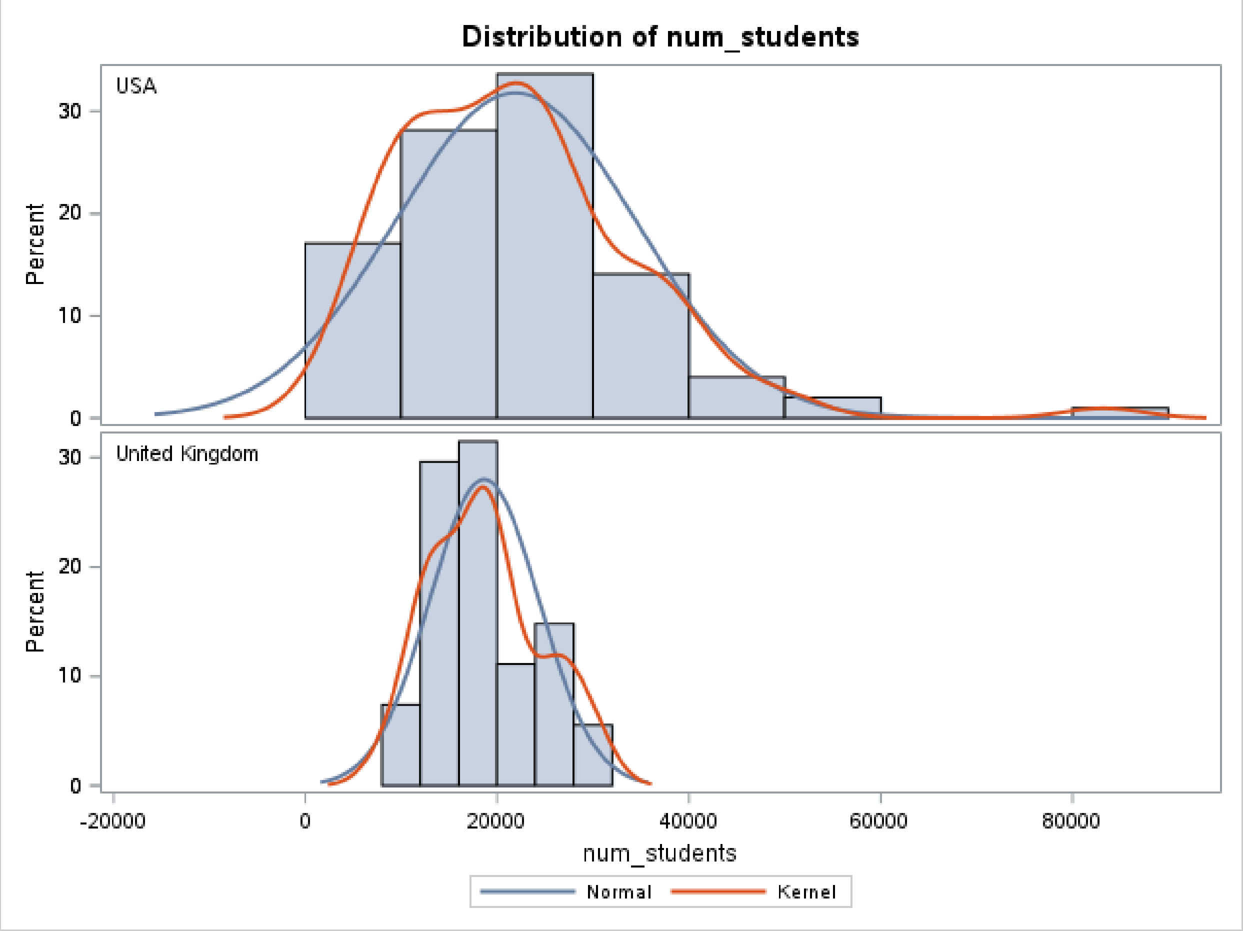
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	251	1.86	0.0646
Satterthwaite	Unequal	194.23	2.76	0.0063

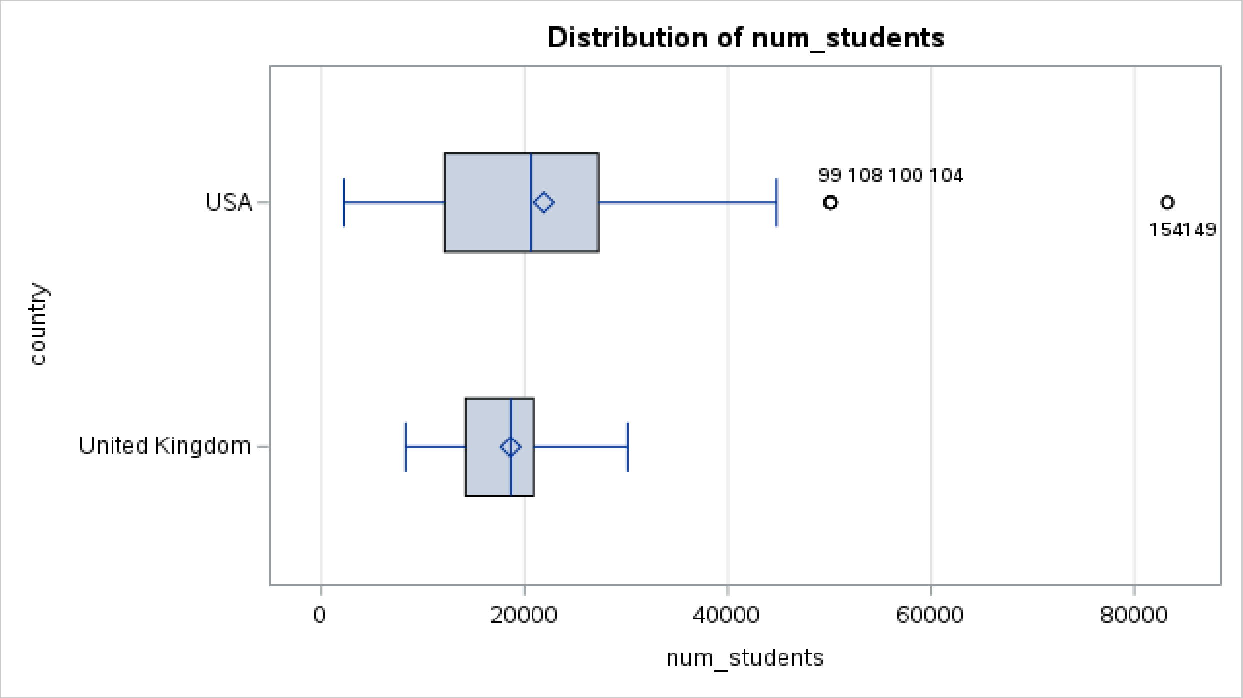
Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F

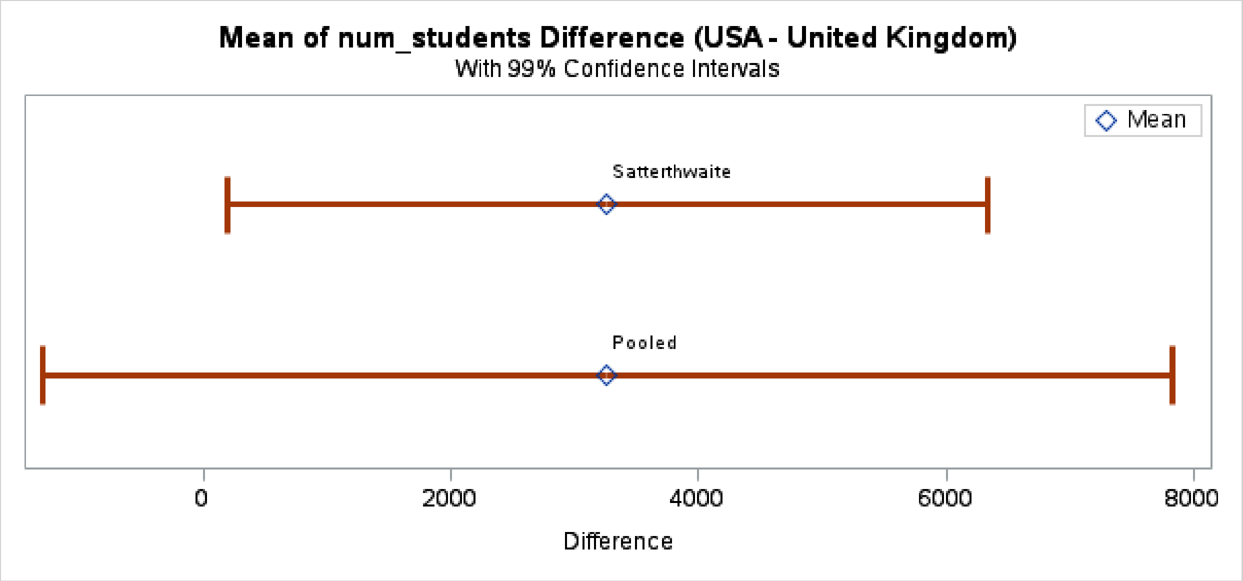
Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	198	53	4.85	<.0001





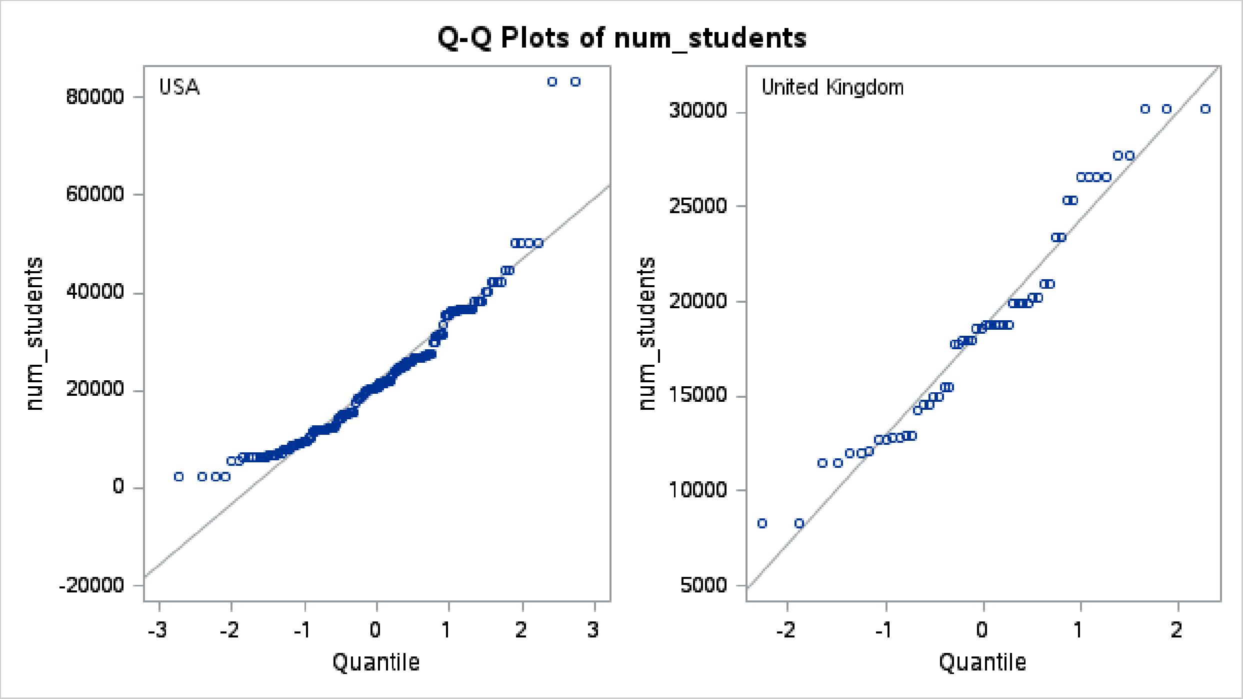


Distribution of num_students



Mean of num_students Difference (USA - United Kingdom)

With 99% Confidence Intervals



Q-Q Plots of num_students

A5. Null Hypothesis (H0): No significant difference in mean student numbers between USA and UK universities. Assumptions Checked: Normality via Q-Q plots; variance equality rejected, leading to Satterthwaite's approximation for t-tests. T-Test Result: Significant at α = 0.01 with a p-value of 0.0063, indicating a statistical difference in student numbers, favoring higher numbers in USA universities. Distribution Analysis: USA shows a broader range with more outliers compared to the UK, supported by histograms and box plots. Conclusion: Statistical analysis supports a significant difference in student populations, with implications for educational policies and resource allocation between the two countries.

A6. Analysis of Top Universities in Selected Countries

Obs	university_name	year	world_rank	country	national_rank
10	University College London	2013	30	United Kingdom	4
11	University College London	2014	30	United Kingdom	3
12	University College London	2012	31	United Kingdom	4
13	University of Nottingham	2012	97	United Kingdom	6
14	University of Bonn	2014	98	Germany	3
15	University of Bristol	2012	98	United Kingdom	7
16	Sapienza University of Rome	2015	112	Italy	1
17	University of Bristol	2014	123	United Kingdom	8

Sapienza University of Rome Italian university is the highest ranked

A7.Mean quality of education overall and for scores > 100

The MEANS Procedure

Analysis Variable : quality_of_education
Mean
213.5543478

A7.Mean quality of education overall and for scores > 100

The MEANS Procedure

Analysis Variable : quality_of_education
Mean
266.3661972

The average quality of education across the entire uni1 dataset is 213.55, while for the subset where the quality score exceeds 100, it stands at 266.366.

Q8. Grouped Summary Statistics for Patents

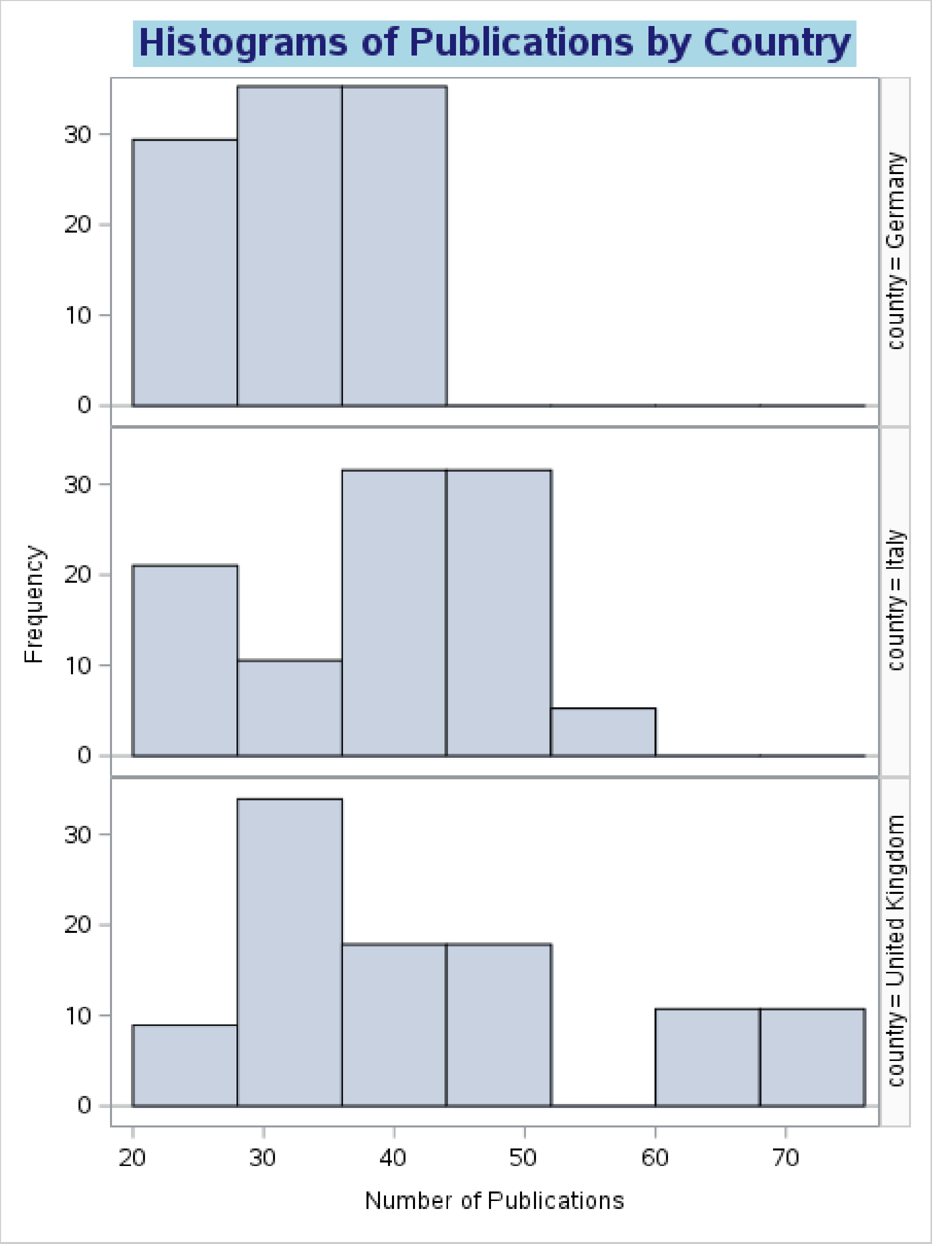
The MEANS Procedure

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

Analysis Variable : patents						
country	N Obs	N	Mean	Std Dev	Minimum	Maximum
United Kingdom	56	56	305.8392857	204.6968292	15.0000000	871.0000000

A8. The table presents grouped summary statistics for patents across universities in Germany, Italy, and the United Kingdom. Italy shows the highest average number of patents per university (532.21), suggesting a strong focus on innovation and research, while the UK, despite having the most observations, exhibits the lowest average (305.84). The standard deviation indicates significant variability in the number of patents across universities in each country, reflecting differences in research output and capabilities.

Q9. A plot of the publications variable by country



A9. The histograms depicting the distribution of publications across universities in Germany, Italy, and the United Kingdom reveal distinct patterns in each country. German universities show a relatively uniform distribution, mostly clustering between

30 and 50 publications, indicating consistent output. In contrast, Italian universities exhibit a bimodal distribution with notable peaks around 30 and 50 publications, reflecting a more varied publication count. The UK shows the greatest variability, with a primary peak at 30 publications but extending up to 70, suggesting a broader range of publication activity among its universities. These observations indicate that German universities maintain a steady publication rate, Italian institutions vary more broadly, and UK universities display the most diverse range of publication outputs.

Task Demonstration: Principal Component Analysis ->PCA is a statistical technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to reduce the dimensions of the data by transforming it into a new set of variables, the principal components, which are uncorrelated and which maximize the variance. This analysis helps in understanding the data structure, detecting outliers, and performing feature reduction for other machine learning tasks.

Obs	Species	SepalLength	SepalWidth	PetalLength	PetalWidth
1	Setosa	50	33	14	2
2	Setosa	46	34	14	3
3	Setosa	46	36	10	2
4	Setosa	51	33	17	5
5	Setosa	55	35	13	2

We using inbuilt iris datset to perform the task PCA

Performing PCA: The method involves loading the dataset, performing PCA, and then visualizing the results to interpret the principal components.

The PRINCOMP Procedure

Observations	150
Variables	4

Simple Statistics

	SepalLength	SepalWidth	PetalLength	PetalWidth
Mean	58.43333333	30.57333333	37.58000000	11.99333333
StD	8.28066128	4.35866285	17.65298233	7.62237669

Correlation Matrix

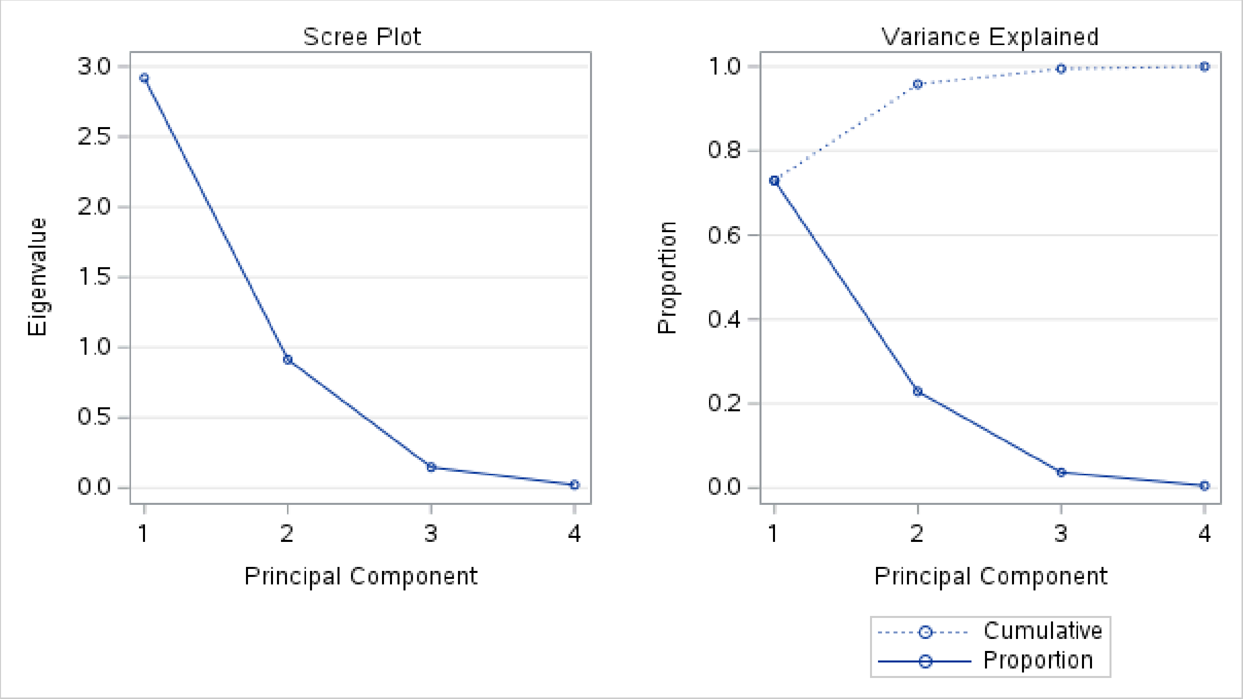
		SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	Sepal Length (mm)	1.0000	-.1176	0.8718	0.8179
SepalWidth	Sepal Width (mm)	-.1176	1.0000	-.4284	-.3661
PetalLength	Petal Length (mm)	0.8718	-.4284	1.0000	0.9629
PetalWidth	Petal Width (mm)	0.8179	-.3661	0.9629	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.91849782	2.00446735	0.7296	0.7296
2	0.91403047	0.76727360	0.2285	0.9581
3	0.14675688	0.12604204	0.0367	0.9948

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
4	0.02071484		0.0052	1.0000

Eigenvectors					
		Prin1	Prin2	Prin3	Prin4
SepalLength	Sepal Length (mm)	0.521066	0.377418	-.719566	-.261286
SepalWidth	Sepal Width (mm)	-.269347	0.923296	0.244382	0.123510
PetalLength	Petal Length (mm)	0.580413	0.024492	0.142126	0.801449
PetalWidth	Petal Width (mm)	0.564857	0.066942	0.634273	-.523597



Purpose: Conducts PCA on the iris dataset considering all four primary measurements. The output dataset Work.PcaOut contains the principal components.

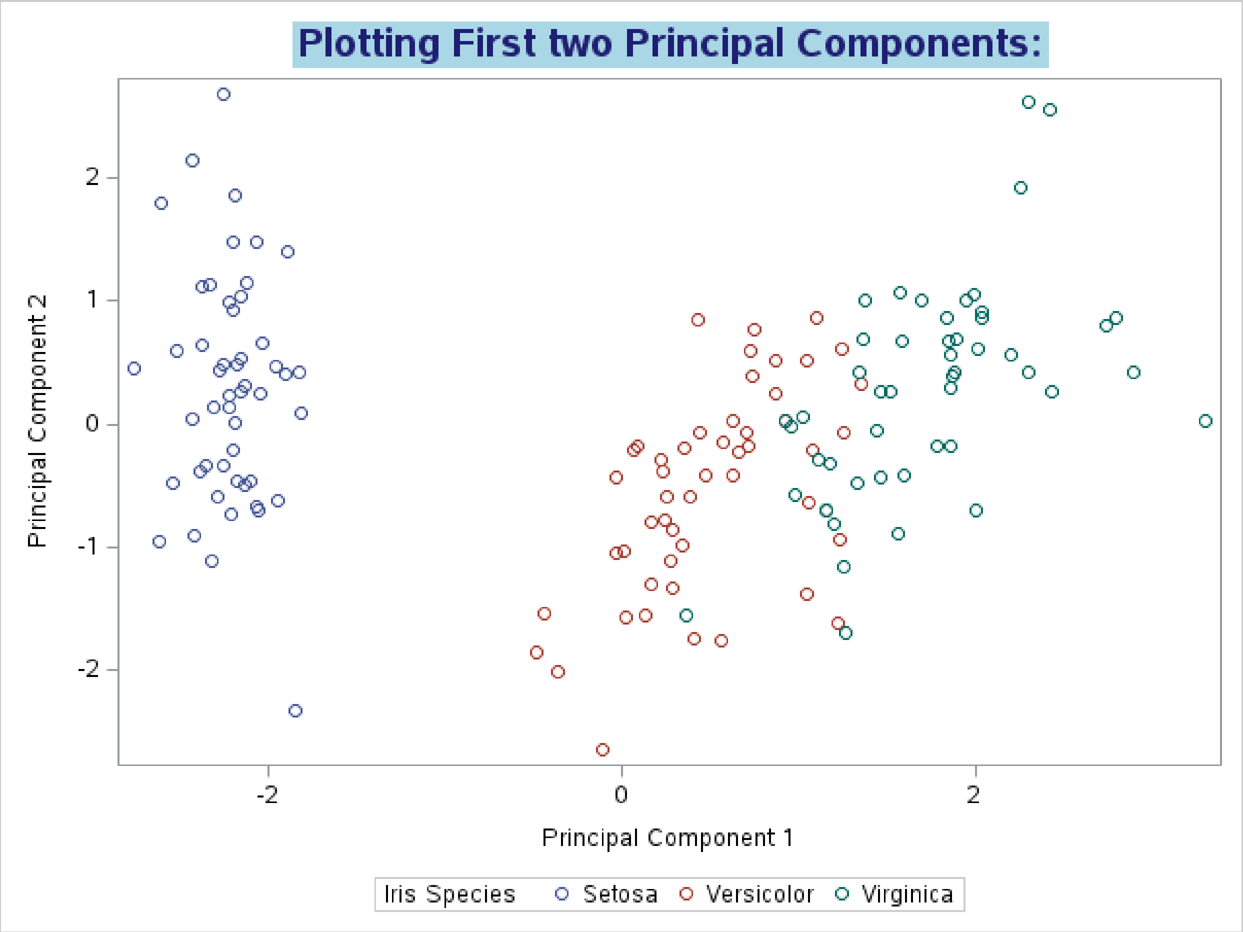
Findings after performing PCA on Iris Dataset: The Principal Component Analysis (PCA) of the Iris dataset reveals that the first principal component explains approximately 72.96% of the variance, indicating a strong pattern across the four measurements, while the first two components together account for about 95.81% of the variance. This suggests that most information about the Iris flowers can be captured in just two dimensions, which simplifies visualization and analysis. The plot of the first two principal components shows distinct clustering by species, particularly Setosa, which is well-separated from Versicolor and Virginica, demonstrating PCA's effectiveness in reducing dimensionality while preserving significant classification information.

Displaying PCA Output:

Obs	Prin1	Prin2	Prin3	Prin4
1	-2.19648	0.00919	-0.15252	-0.04921
2	-2.43587	0.04749	0.33435	0.03665
3	-2.76508	0.45681	0.33107	-0.01958
4	-1.81260	0.08527	0.03437	-0.15064

Obs	Prin1	Prin2	Prin3	Prin4
5	-2.03832	0.65935	-0.48292	-0.19570

Purpose: Displays the first five observations from the PCA output, showing the principal components for the first few data points, which helps in understanding the immediate transformation results.



Purpose: Visualizes the first two principal components, providing a scatter plot to evaluate how well PCA separated different species based on their transformed features. This is an excellent way to visually assess the effectiveness of PCA.

Conclusion: This report demonstrates how PCA can be utilized to reduce dimensionality and uncover patterns in multivariate data. The principal components provide a way to visualize complex data structures, helping in easier interpretation and analysis.