Data Programming with SAS- FINAL PROJECT

Chelsea Rodrigues - 23200333

I have read and understood the Honesty Code and have neither received nor given assistance in any way with the work contained in this submission.

DATA ANALYSIS 1

The CONTENTS Procedure

Data Set Name	S40840FP.DEATH_N_BIRTH	**Observations**	630
Member Type	DATA	**Variables**	10
Engine	V9	**Indexes**	0
Created	08/13/2024 17:16:49	**Observation Length**	94
Last Modified	08/13/2024 17:16:49	**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	1392
Obs in First Data Page	630
Number of Data Set Repairs	0
Filename	/home/u63919273/death_n_birth.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	6877619971
Access Permission	rw-r--r--
Owner Name	u63919273
File Size	256KB
File Size (bytes)	262144

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
5	C02199V02655	Char	3	$3.	$3.
7	C02466V02984	Char	4	$4.	$4.
8	Criteria for Projection	Char	13	$13.	$13.
1	STATISTIC	Char	10	$10.	$10.
6	Sex	Char	12	$12.	$12.
2	Statistic Label	Char	25	$25.	$25.
3	TLIST(A1)	Char	6	$6.	$6.
9	UNIT	Char	8	$8.	$8.
10	VALUE	Char	7	$7.	$7.
4	Year	Char	6	$6.	$6.

Few rows of death and birth dataset

Obs	STATISTIC	Statistic Label	TLIST(A1)	Year	C02199V02655	Sex	C02466V02984	Criteria for Projection	UNIT	VALUE
1	PEC25C01	Projected Annual Births	2023	2023	-	Both sexes	21	Method - M1	Number	57537
2	PEC25C01	Projected Annual Births	2023	2023	-	Both sexes	22	Method - M2	Number	57537
3	PEC25C01	Projected Annual Births	2023	2023	-	Both sexes	23	Method - M3	Number	57537
4	PEC25C01	Projected Annual Births	2023	2023	1	Male	21	Method - M1	Number	29522
5	PEC25C01	Projected Annual Births	2023	2023	1	Male	22	Method - M2	Number	29522

Numerical Summaries

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Year_Num	630	2040.00	10.1075300	2023.00	2057.00
VALUE_Num	630	33505.36	12866.97	16728.00	66772.00

Conclusion:Year :As we can clearly notice from the above table that, the average year in the dataset is 2040,and the years covered range from 2023 to 2057 showing a 35-year span this indiactes that there is small variation around the average.Value: The average projected value is about 33,505. Values vary widely, from 16,728 to 66,772 indiactinglarge variation in projected values.

Frequency distribution of categorical Variables

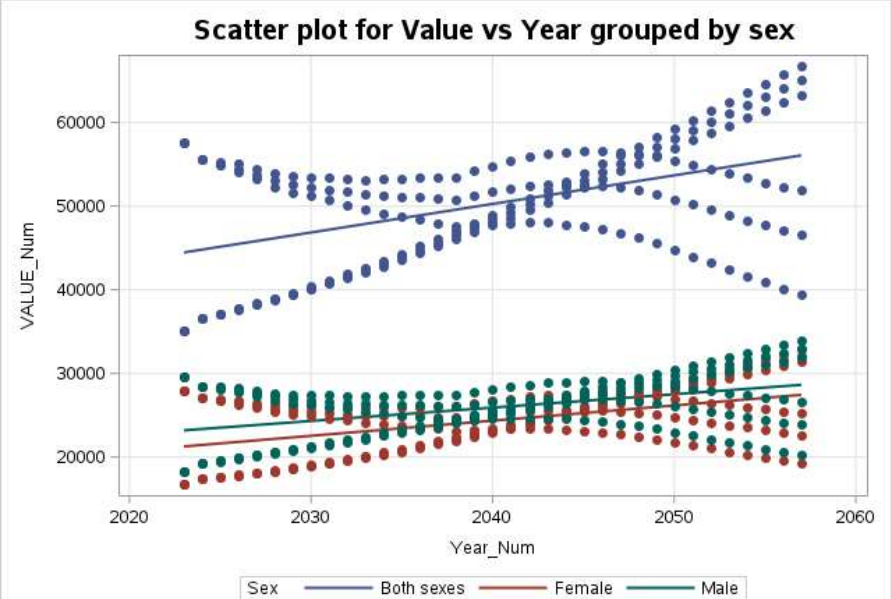
The FREQ Procedure

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Both sexes	210	33.33	210	33.33
Female	210	33.33	420	66.67
Male	210	33.33	630	100.00

Statistic Label	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Projected Annual Births	315	50.00	315	50.00
Projected Annual Deaths	315	50.00	630	100.00

UNIT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Number	630	100.00	630	100.00

Conclusion:The data is equally divided among Both sexes,Female and Male, with each category having about 33.33% of the total data.The dataset is equally divided with 50% of Projected annual birth data and remaining 50% Projected annual birth data.



Conclusion:The above graph indicates that Males have higher values of Projected death and birth than Females and both sexes category have very high values than individual ma1e and female category.

DATA ANALYSIS 2

The CONTENTS Procedure

Data Set Name	S40840FP.UNIVERISITIES	Observations	551
Member Type	DATA	Variables	16
Engine	V9	Indexes	0

Created	08/09/2024 14:30:26	**Observation Length**	160
Last Modified	08/09/2024 14:30:26	**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	/home/u63919273/univerisities.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	6853530928
Access Permission	rw-r--r--
Owner Name	u63919273
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.

The output shows contents of Univeristy dataset

Q1. Few obs of Universities dataset

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

The above table displays few rows of Unversities dataset

Q2. Numerical Summary of student_staff_ratio

The MEANS Procedure

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

The above table displays Numerical Summaries of student_staff_ratio variable

Numerical summary rounded to 2 decimals

mean	stddev	min	max
15.99	10.23	2.9	70.4

The above table displays Numerical Summaries rounded to 2 decimals of student_staff_ratio variable

Q3. Univariate analysis of 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
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

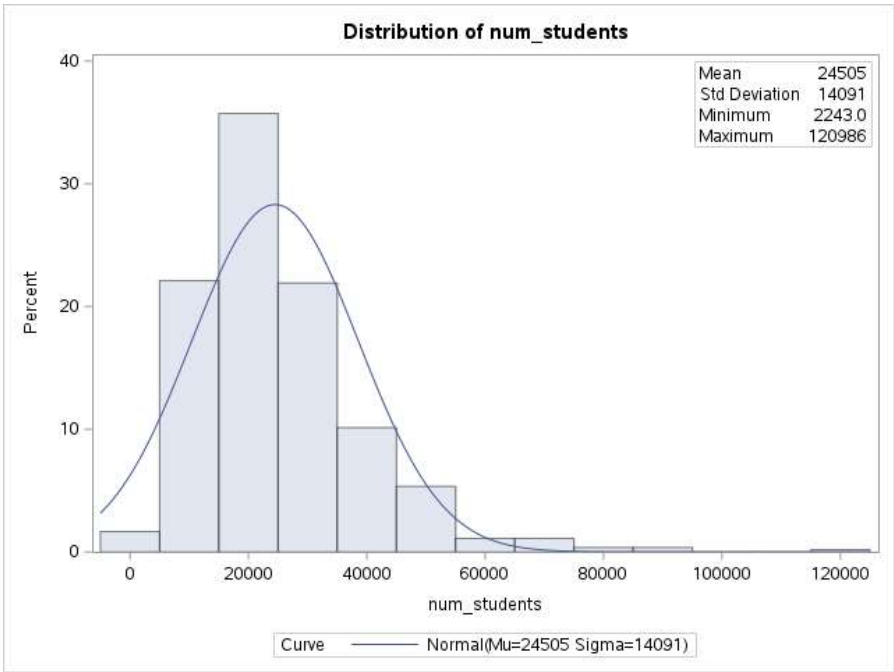
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
2243	41	83236	216
2243	40	83236	228
2243	36	85532	346
2243	13	85532	358
3055	319	120986	239

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

The mean number of students enrolled in universities is approximately 24505.Each university has a very different student population; some universities have very few students, while others have a large number.The average difference is 14091 students. Histogram - As we can clearly observe from the histogram that the graph is right skewed which indicates that most universities have a moderate number of students, but a few universities have a very large number of students.Normal Probability Plot (Q-Q Plot)-Even this plot indicates the same that the number of students isn't normally distributed. The plots points deviate from the straight line, particularly at universities with large numbers of students, suggesting that the data is skewed.

Q3. Univariate analysis of number_of_students

The UNIVARIATE Procedure



The mean number of students enrolled in universities is approximately 24505.Each university has a very different student population; some universities have very few students, while others have a large number.The average difference is 14091 students. Histogram - As we can clearly observe from the histogram that the graph is right skewed which indicates that most universities have a moderate number of students, but a few universities have a very large number of students.Normal Probability Plot (Q-Q Plot)-Even this plot indicates the same that the number of students isn't normally distributed. The plots points deviate from the straight line, particularly at universities with large numbers of students, suggesting that the data is skewed.

Q3. Univariate analysis of 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

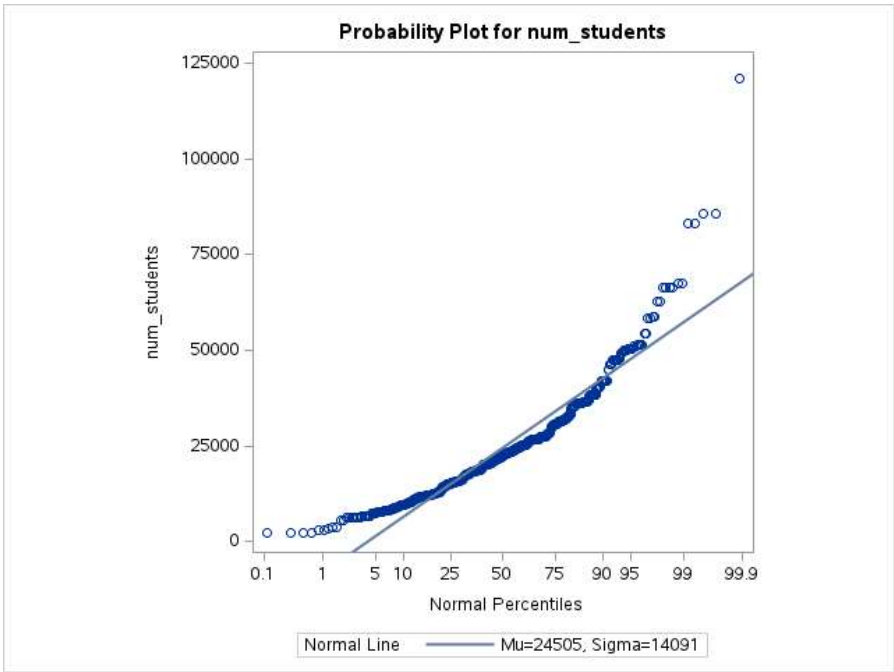
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.1254493	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	1.8430084	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	11.2712362	**Pr > A-Sq**	<0.005

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

The mean number of students enrolled in universities is approximately 24505.Each university has a very different student population; some universities have very few students, while others have a large number.The average difference is 14091 students. Histogram - As we can clearly observe from the histogram that the graph is right skewed which indicates that most universities have a moderate number of students, but a few universities have a very large number of students.Normal Probability Plot (Q-Q Plot)-Even this plot indicates the same that the number of students isn't normally distributed. The plots points deviate from the straight line, particularly at universities with large numbers of students, suggesting that the data is skewed.

Q3. Univariate analysis of number_of_students

The UNIVARIATE Procedure



The mean number of students enrolled in universities is approximately 24505.Each university has a very different student population; some universities have very few students, while others have a large number.The average difference is 14091 students. Histogram - As we can clearly observe from the histogram that the graph is right skewed which indicates that most universities have a moderate number of students, but a few universities have a very large number of students.Normal Probability Plot (Q-Q Plot)-Even this plot indicates the same that the number of students isn't normally distributed. The plots points deviate from the straight line, particularly at universities with large numbers of students, suggesting that the data is skewed.

Q4.Correlation Analysis

The CORR Procedure

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

All these correlations statistically significant different from 0

Q5. Hypothesis test

The TTEST Procedure

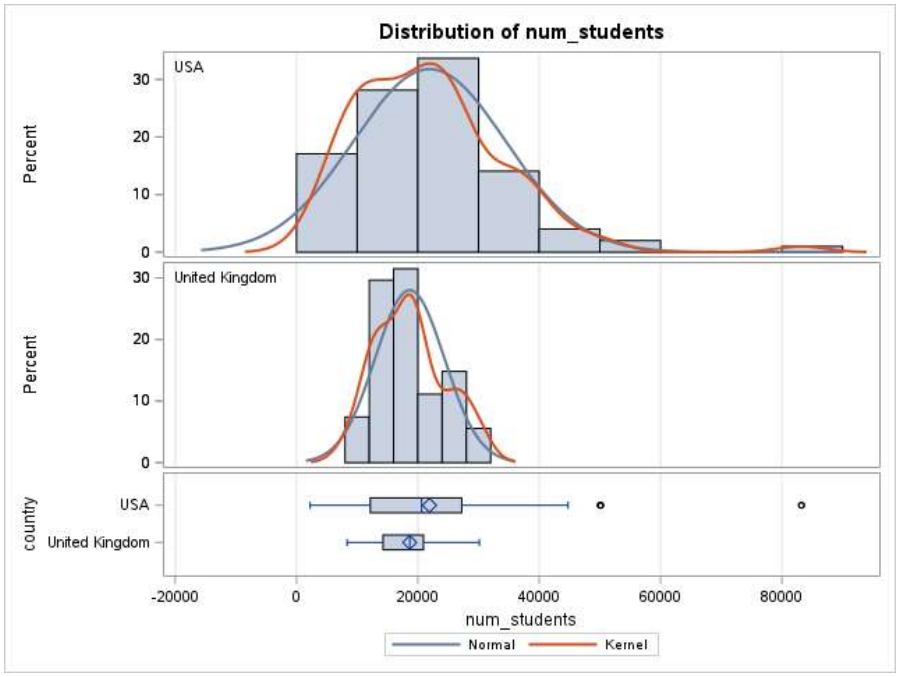
Variable: num_students

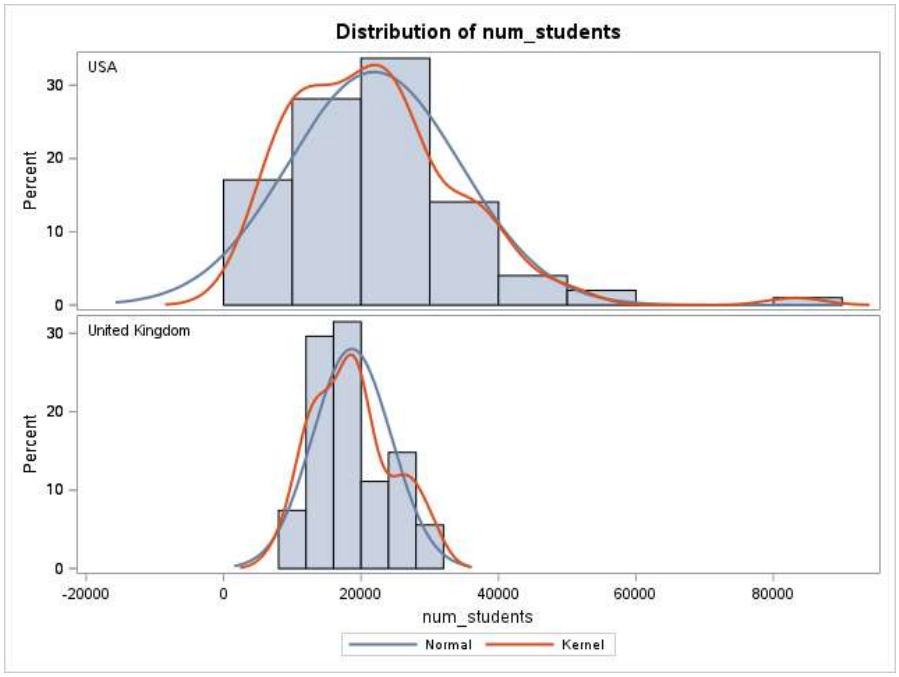
country	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
USA		199	21920.1	12548.1	889.5	2243.0	83236.0
United Kingdom		54	18658.9	5698.3	775.4	8338.0	30144.0
Diff (1-2)	Pooled		3261.2	11448.3	1756.6		
Diff (1-2)	Satterthwaite		3261.2		1180.1		

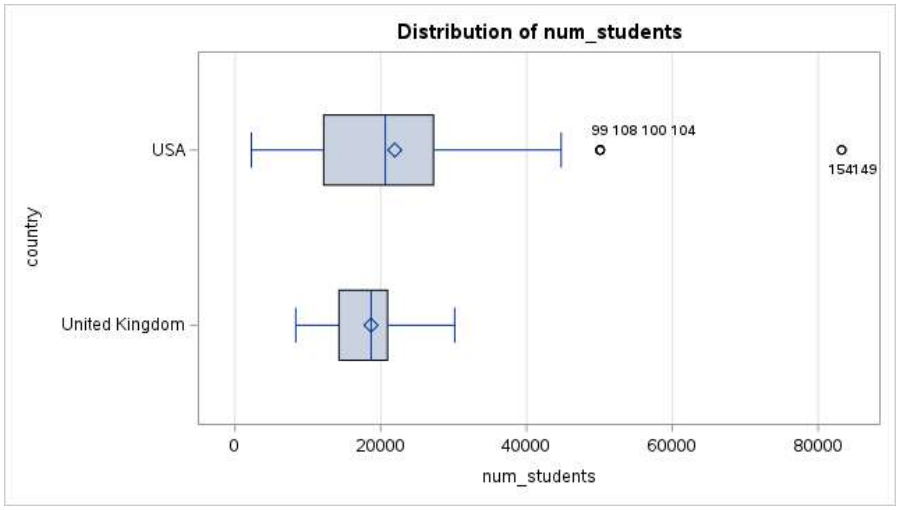
country	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
USA		21920.1	20166.0	23674.3	12548.1	11424.5	13918.7
United Kingdom		18658.9	17103.6	20214.3	5698.3	4790.1	7034.6
Diff (1-2)	Pooled	3261.2	-198.4	6720.8	11448.3	10528.4	12545.6
Diff (1-2)	Satterthwaite	3261.2	933.8	5588.6			

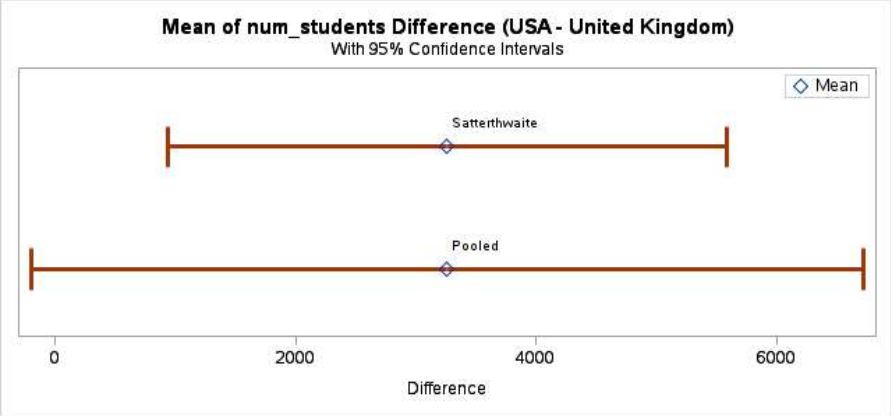
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	251	1.86	0.0646
Satterthwaite	Unequal	194.23	2.76	0.0063

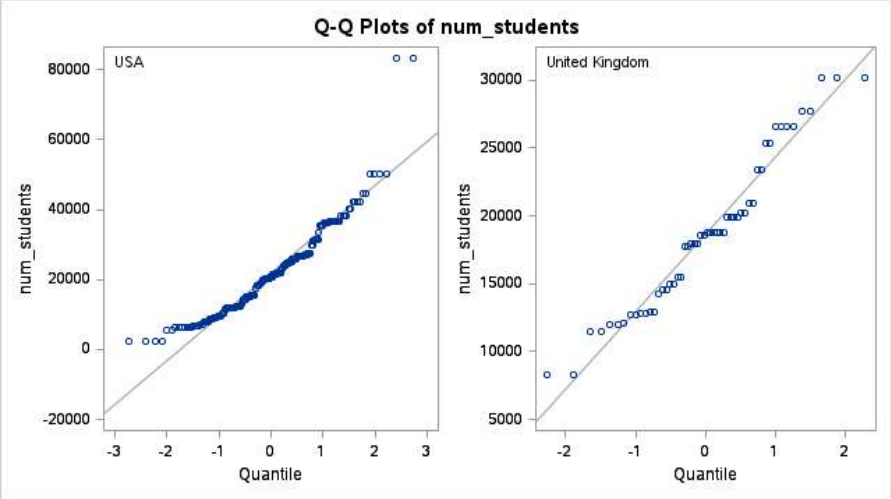
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	198	53	4.85	<.0001


Distribution of num_students
USA
United Kingdom
Percent
country
USA
United Kingdom
-20000
0
20000
40000
60000
80000
num_students
Normal
Kernel


Distribution of num_students
USA
United Kingdom
Percent
-20000
0
20000
40000
60000
80000
num_students
Normal
Kernel


Distribution of num_students
USA
United Kingdom
country
99 108 100 104
154149
0
20000
40000
60000
80000
num_students





Hypotheses:Null Hypothesis (H_0): There is no significant difference between the mean number of students in USA universities and UK universities.$H0{:}\mu USA{=}\mu UK$Alternative Hypothesis (H_1): There is a significant difference between the mean number of students in USA universities and UK universities.$H1{:}\mu USA{\neq}\mu UK$ Plot Interpretation:Boxplot - The box plots indicates that, on average, USA universities enrolls more students than UK universities as USA box plots are positioned on higher scale.The statistical test and the differences in these plots verify that the average student population in US and UK universities differs significantly.Q-Qplot - The distribution of number of students for USA and UK universities is normal as data points do not deviate much for both USA and UK universities. Conclusion: We reject the null hypothesis (H_0) because the p-value is less than the significance level of 0.01.Interpretation: There is enough evidence to reach the conclusion that the mean number of students attending US and UK universities differs significantly.

Question 6.

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 is the highest ranked Italian university

Q7. Mean Quality of education for uni1 dataset

The MEANS Procedure

Analysis Variable : quality_of_education
Mean
213.5543478

Mean quality of education for whole uni1 dataset is 213.5543478

Mean Quality of Education for the Subset Where quality_of_education > 100

The MEANS Procedure

Analysis Variable : quality_of_education
Mean
266.3661972

Mean quality of education for subset where quality_of_education >100 is 266.3661972

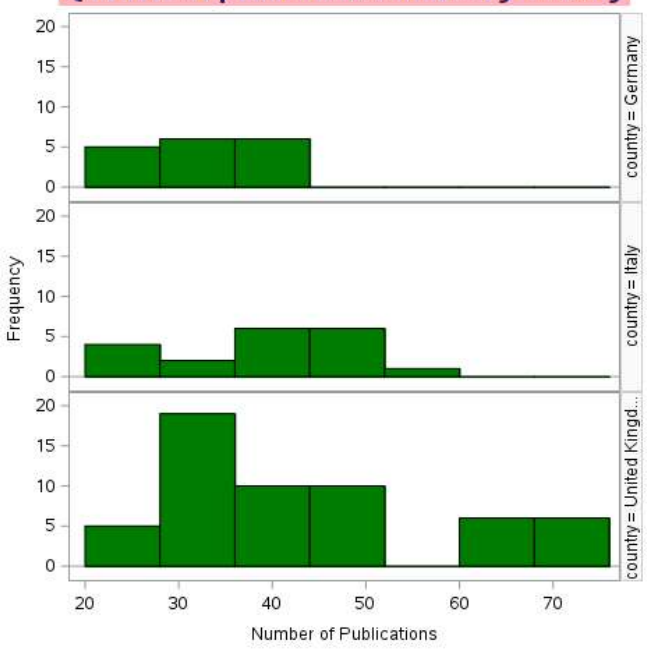
Q8.Summary statistics

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
United Kingdom	56	56	305.8392857	204.6968292	15.0000000	871.0000000

From the above output we can clearly see that,Mean Patents:Italy has the highest average number of patents (532.21),Germany comes in second with 386.47 patents on average.The United Kingdom has the lowest average, with 305.84 patents.Standard Deviation Patents:The United Kingdom shows the most variability, with a standard deviation of 204.70, indicating that the number of patents varies widely between universities.Italy has the least standard deviation (121.10), indicating that the majority of Italian universities have a similar number of patents.Germany exhibits a moderate variability (187.56).Range of patents:The largest range is seen in the United Kingdom, where patent counts range from 15 to 871, exhibiting both extremely lowand extremely high numbers.Germany and Italy have smaller ranges of between 138 and 774 patents, respectively, and between 312 and 737 patents, respectively.

Q9.Plots of publication variable by country



The above plot is the visual comparison of publications output across UK, German and Italian universities.1.United Kingdom:We can notice that UK has more diversity in publication numbers, most of the universities have moderate number of publications,while few universities have very high number of publications.2. Germany:We can notice that number of publications for German universities is between specific range and has few outliers compared to UK universities. 3: Italy:Even, Italian universities have publication counts that are fairly close to each other, with some variation.

TASKS DEMONSTRATION
Regression Analysis of University Scores on Publications, Teaching, and International

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	551
Number of Observations Used	551

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	54251	18084	394.55	<.0001
Error	547	25071	45.83361		
Corrected Total	550	79322			

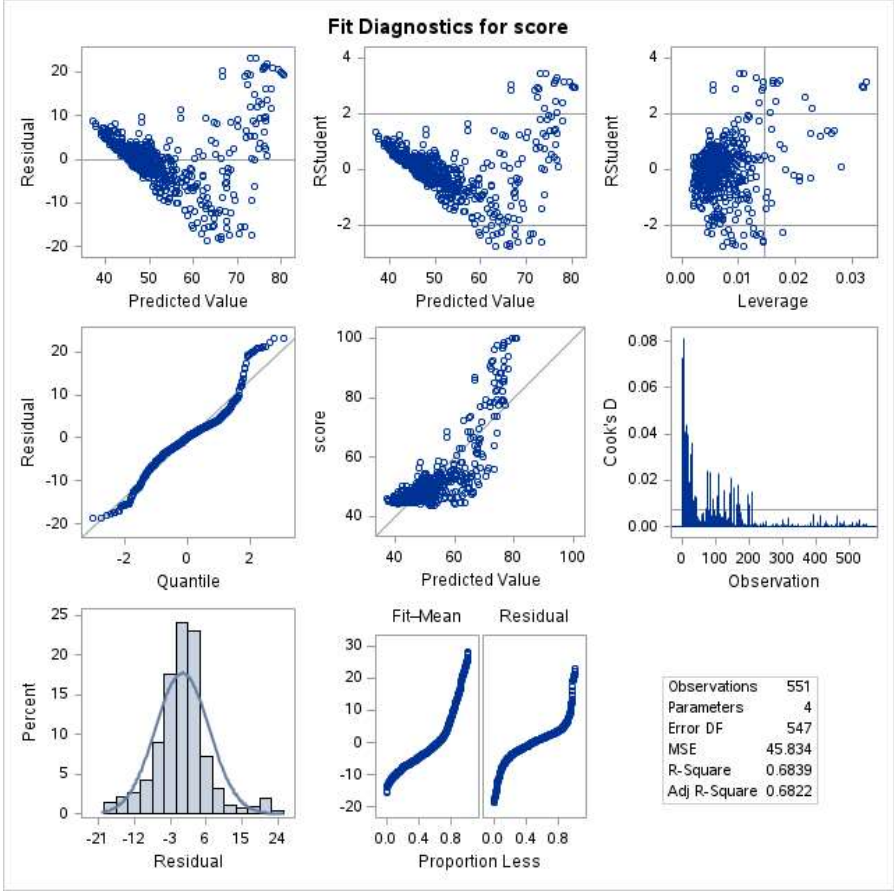
Root MSE	6.77005	R-Square	0.6839
Dependent Mean	52.63819	Adj R-Sq	0.6822
Coeff Var	12.86149		

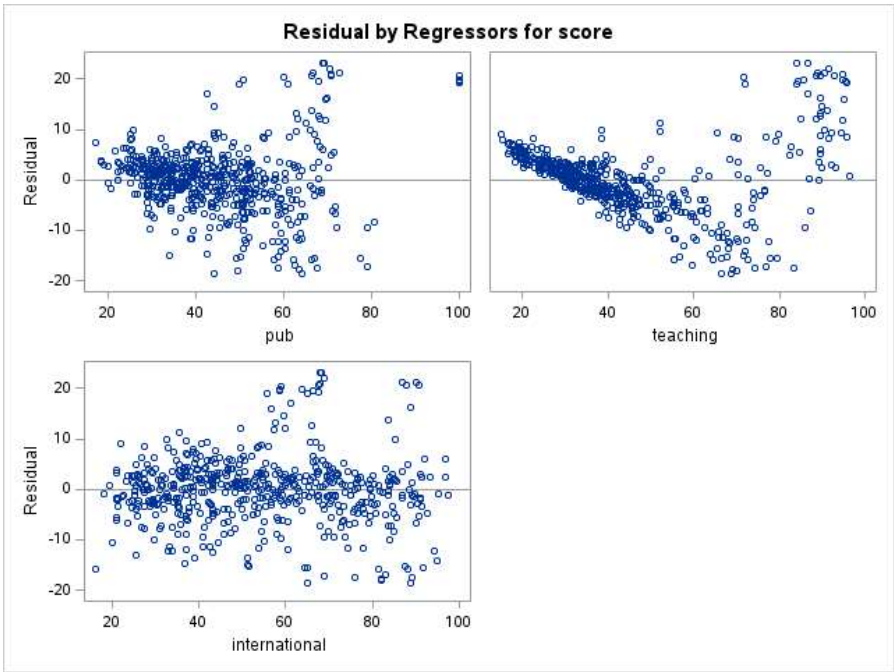
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	27.70655	1.25225	22.13	<.0001
pub	1	0.06843	0.03057	2.24	0.0256
teaching	1	0.46147	0.02143	21.54	<.0001
international	1	0.02775	0.01453	1.91	0.0567

1.Purpose:The Linear Regression task is used to model and analyze how several characterstics like publications, teaching quality, and international outlook—influence a universities overall rating (score). By examining these relationships, we can identify which factors have the most significant impact on the universities score, with this we can better understand what influences higher rankings.2. Key Functionality of Linear Regression task:Model Relationships:These enable us to determine and measure the influence of several independent variables such as pub, teaching and international on dependent variable score.Predict Outcomes:By using the values of the independent variables, it assists in making predictions about the dependent variable.Assess Model Fit: It provides statistical measures (e.g., coefficients, R-squared) and visualizations to evaluate how effectively the model explains the data variability and the degree of correlation between variables.

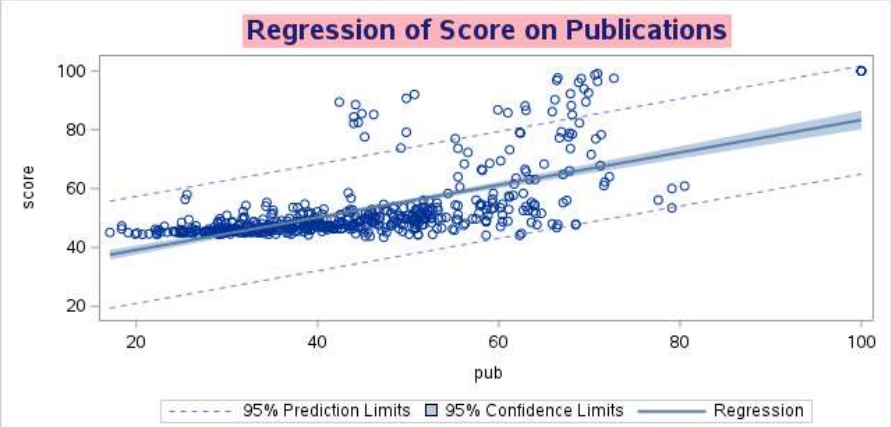
TASKS DEMONSTRATION
Regression Analysis of University Scores on Publications, Teaching, and International

The REG Procedure
Model: MODEL1
Dependent Variable: score





1.Purpose:The Linear Regression task is used to model and analyze how several characterstics like publications, teaching quality, and international outlook—influence a universities overall rating (score). By examining these relationships, we can identify which factors have the most significant impact on the universities score, with this we can better understand what influences higher rankings.2. Key Functionality of Linear Regression task:Model Relationships:These enable us to determine and measure the influence of several independent variables such as pub, teaching and international on dependent variable score.Predict Outcomes:By using the values of the independent variables, it assists in making predictions about the dependent variable.Assess Model Fit: It provides statistical measures (e.g., coefficients, R-squared) and visualizations to evaluate how effectively the model explains the data variability and the degree of correlation between variables.



1.Purpose:The Linear Regression task is used to model and analyze how several characterstics like publications, teaching quality, and international outlook—influence a universities overall rating (score). By examining these relationships, we can identify which factors have the most significant impact on the universities score, with this we can better understand what influences higher rankings.2. Key Functionality of Linear Regression task:Model Relationships:These enable us to determine and measure the influence of several independent variables such as pub, teaching and international on dependent variable score.Predict Outcomes:By using the values of the independent variables, it assists in making predictions about the dependent variable.Assess Model Fit: It provides statistical measures (e.g., coefficients, R-squared) and visualizations to evaluate how effectively the model explains the data variability and the degree of correlation between variables.