STAT40840 - Data Programming with SAS
Final Project

Divya Pariti - 23200831

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 Task 1

Task-1 : Loading the data into SAS

Obs	country_code	country_name	indicator_id	indicator_name	index_id	index_name	value	year
1	#country+code	#country+name	#indicator+id	#indicator+name	#index+id	#index+name	#indicator+value+num	#date+year
2	CHE	Switzerland	abr	Adolescent Birth Rate (births per 1,000 women ages 15-19)	GII	Gender Inequality Index	7.56	1990
3	CHE	Switzerland	abr	Adolescent Birth Rate (births per 1,000 women ages 15-19)	GII	Gender Inequality Index	8.28	1991
4	CHE	Switzerland	abr	Adolescent Birth Rate (births per 1,000 women ages 15-19)	GII	Gender Inequality Index	7.83	1992
5	CHE	Switzerland	abr	Adolescent Birth Rate (births per 1,000 women ages 15-19)	GII	Gender Inequality Index	7.02	1993

The Human Development Indicators dataset has been loaded successfully using a PROC import step.

source: hdro_indicators_swd.csv

Task-2 : Analysing the data

Printing Contents table

Data Set Name	WORK.HDRO_DATA	**Observations**	895
Member Type	DATA	**Variables**	8
Engine	V9	**Indexes**	0
Created	08/14/2024 21:42:58	**Observation Length**	152
Last Modified	08/14/2024 21:42:58	**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	2
First Data Page	1
Max Obs per Page	861
Obs in First Data Page	842
Number of Data Set Repairs	0
Filename	/saswork/SAS_workBE9C000183C6_odaws02-euw1.oda.sas.com/SAS_workC9C3000183C6_odaws02-euw1.oda.sas.com/hdro_data.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	1610683360
Access Permission	rw-r--r--
Owner Name	u63920100
File Size	384KB
File Size (bytes)	393216

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
1	country_code	Char	13	$13.	$13.
2	country_name	Char	13	$13.	$13.
5	index_id	Char	9	$9.	$9.
6	index_name	Char	23	$23.	$23.
3	indicator_id	Char	13	$13.	$13.
4	indicator_name	Char	59	$59.	$59.

Task-2 : Analysing the data

Printing Contents table

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
8	value	Num	8		
7	year	Num	8		

We have successfully converted the value and year variables into their proper data types(numerical).

Table 3 from the default output we can see that there are 8 variables.

6 variables are categorical variables and 2 are numerical variables.

Task-3 : Creating tabluar summaries for the numerical variables

Variable	Mean	Median	Std Dev	Minimum	Maximum
value	7032.23	16.5800000	20343.33	-6.0000000	84820.40
year	2006.80	2007.00	9.6212576	1990.00	2022.00

Summary of tabluar summaries for the numerical values from the dataset have been printed successfully.

From the table the mean of value variable is 7032.23. Followed by the largest standard deviation of 20343.33.

After careful observation we can conclude that the distribution is skewed due to its lower median value with a range of -8.00 to 84820.40

year variable has a mean of 2008.80 with a standard deviation of 9.62.

From this we can conclude that the distribution is concetrated around a particular value, here, the mean value of 2008.80

Task-4 : Creating frequency table summaries for the categorical variables

country_name	Frequency	Percent	Cumulative Frequency	Cumulative Percent
#country+name	1	0.11	1	0.11
Switzerland	894	99.89	895	100.00

country_code	Frequency	Percent	Cumulative Frequency	Cumulative Percent
#country+code	1	0.11	1	0.11
CHE	894	99.89	895	100.00

index_id	Frequency	Percent	Cumulative Frequency	Cumulative Percent
#index+id	1	0.11	1	0.11
GDI	331	36.98	332	37.09
GII	265	29.61	597	66.70
HDI	133	14.86	730	81.56
IHDI	65	7.26	795	88.83
PHDI	100	11.17	895	100.00

index_name	Frequency	Percent	Cumulative Frequency	Cumulative Percent
#index+name	1	0.11	1	0.11
Gender Development Inde	331	36.98	332	37.09
Gender Inequality Index	265	29.61	597	66.70
Human Development Index	133	14.86	730	81.56
Inequality-adjusted Hum	65	7.26	795	88.83
Planetary pressures–a	100	11.17	895	100.00

Task-4 : Creating frequency table summaries for the categorical variables

indicator_id	Frequency	Percent	Cumulative Frequency	Cumulative Percent
#indicator+id	1	0.11	1	0.11
abr	33	3.69	34	3.80
co2_prod	33	3.69	67	7.49
coef_ineq	13	1.45	80	8.94
diff_hdi_phdi	33	3.69	113	12.63
eys	33	3.69	146	16.31
eys_f	33	3.69	179	20.00
eys_m	33	3.69	212	23.69
gdi_group	1	0.11	213	23.80
gii_rank	1	0.11	214	23.91
gni_pc_f	33	3.69	247	27.60
gni_pc_m	33	3.69	280	31.28
gnipc	33	3.69	313	34.97
hdi_f	33	3.69	346	38.66
hdi_m	33	3.69	379	42.35
hdi_rank	1	0.11	380	42.46
ineq_edu	13	1.45	393	43.91
ineq_inc	13	1.45	406	45.36
ineq_le	13	1.45	419	46.82
le	33	3.69	452	50.50
le_f	33	3.69	485	54.19
le_m	33	3.69	518	57.88
lfpr_f	33	3.69	551	61.56
lfpr_m	33	3.69	584	65.25
loss	13	1.45	597	66.70
mf	33	3.69	630	70.39
mmr	33	3.69	663	74.08
mys	33	3.69	696	77.77
mys_f	33	3.69	729	81.45
mys_m	33	3.69	762	85.14
pr_f	33	3.69	795	88.83
pr_m	33	3.69	828	92.51
rankdiff_hdi_	1	0.11	829	92.63
se_f	33	3.69	862	96.31
se_m	33	3.69	895	100.00

Task-4 : Creating frequency table summaries for the categorical variables

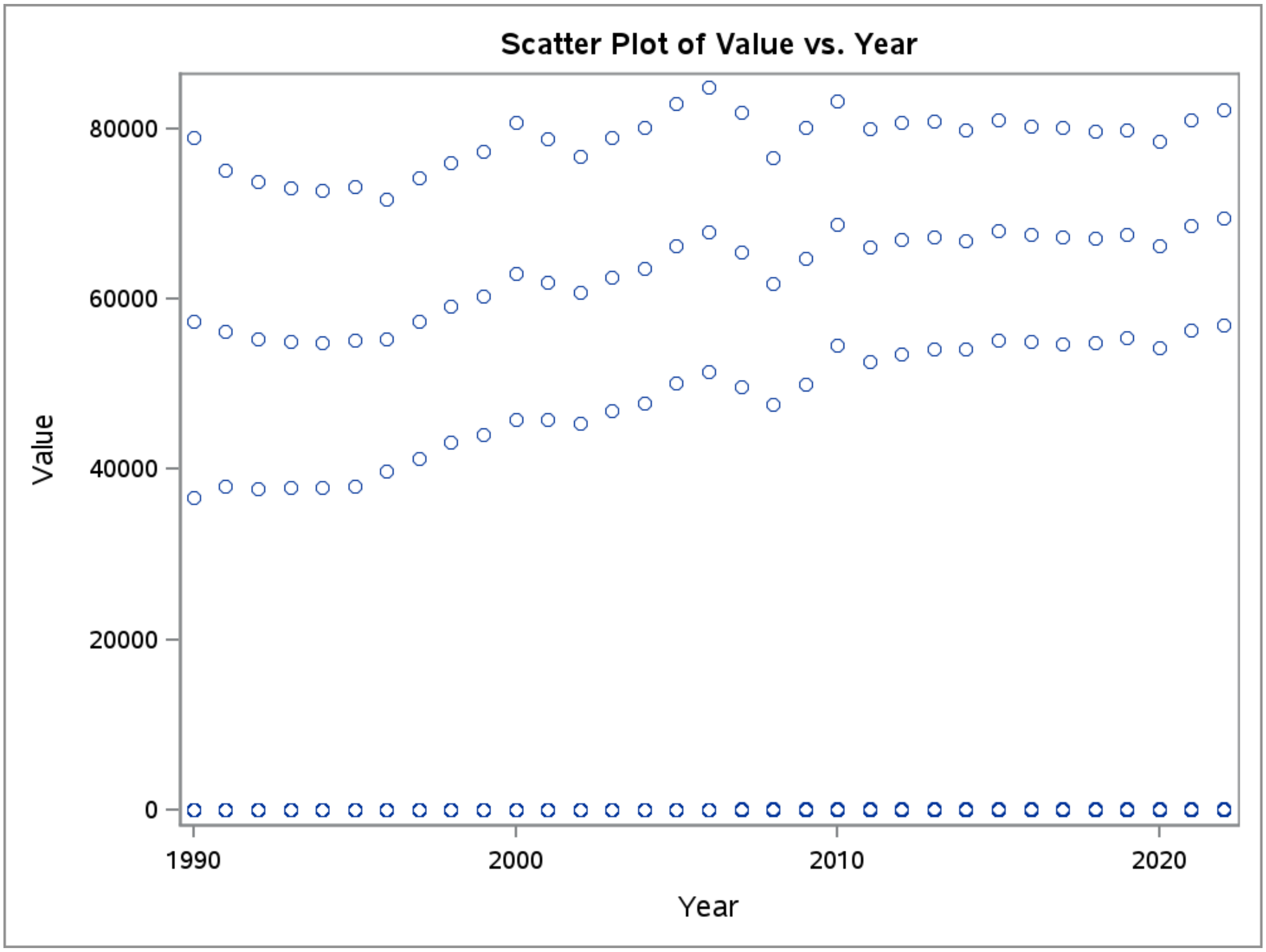
indicator_name	Frequency	Percent	Cumulative Frequency	Cumulative Percent
#indicator+name	1	0.11	1	0.11
Adolescent Birth Rate (births per 1,000 women ages 15-19)	33	3.69	34	3.80
Carbon dioxide emissions per capita (production) (tonnes)	33	3.69	67	7.49
Coefficient of human inequality	13	1.45	80	8.94
Difference from HDI rank	1	0.11	81	9.05
Difference from HDI value (%)	33	3.69	114	12.74
Expected Years of Schooling (years)	33	3.69	147	16.42
Expected Years of Schooling, female (years)	33	3.69	180	20.11
Expected Years of Schooling, male (years)	33	3.69	213	23.80
GDI Group	1	0.11	214	23.91
GII Rank	1	0.11	215	24.02
Gross National Income Per Capita (2017 PPP$)	33	3.69	248	27.71
Gross National Income Per Capita, female (2017 PPP$)	33	3.69	281	31.40
Gross National Income Per Capita, male (2017 PPP$)	33	3.69	314	35.08
HDI Rank	1	0.11	315	35.20
HDI female	33	3.69	348	38.88
HDI male	33	3.69	381	42.57
Inequality in eduation	13	1.45	394	44.02
Inequality in income	13	1.45	407	45.47
Inequality in life expectancy	13	1.45	420	46.93
Labour force participation rate, female (% ages 15 and olde	33	3.69	453	50.61
Labour force participation rate, male (% ages 15 and older)	33	3.69	486	54.30
Life Expectancy at Birth (years)	33	3.69	519	57.99
Life Expectancy at Birth, female (years)	33	3.69	552	61.68
Life Expectancy at Birth, male (years)	33	3.69	585	65.36
Material footprint per capita (tonnes)	33	3.69	618	69.05
Maternal Mortality Ratio (deaths per 100,000 live births)	33	3.69	651	72.74
Mean Years of Schooling (years)	33	3.69	684	76.42
Mean Years of Schooling, female (years)	33	3.69	717	80.11
Mean Years of Schooling, male (years)	33	3.69	750	83.80
Overall loss (%)	13	1.45	763	85.25
Population with at least some secondary education, female (	33	3.69	796	88.94
Population with at least some secondary education, male (%	33	3.69	829	92.63
Share of seats in parliament, female (% held by women)	33	3.69	862	96.31
Share of seats in parliament, male (% held by men)	33	3.69	895	100.00

Frequency tables for all the categorical variables have been printed successfully.

From the above table, we can say that the HDI indicators dataset of Switzerland contains 894 records.

There are 5 different indicators that used to measure different aspects like Gender Development(GDI), Gender Inequality(GII), Human Developer Index(HDI), Inequality adjusted Human Development Index(IHDI) and Planetary pressures-adjusted Human Development Index(PHDI).

From the analysis we can say that the most frequent indicator is GDI with 37.09% followed by GII at 29.61%. The least being PHDI with 11.17%



A Scatter Plot showing the relatioship between value and year is visualized successfully.

The graph shows how the value variable performs across the past three decades.

From the above graph we can say that most of the data points follow a similar trends(Due to the bands) in the years 1990-2022.

Value variable increases over the years.

Alot of values are clustered at the bottom of the graph indicating an index where the values remain low.

Data Analysis Task 2

Task-1

Step-1 : Loading the data into SAS. Printing first 5 observations and first 5 variables from the dataset

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

Universities dataset has been loaded successfully using PROC IMPORT step.

souce: universities.csv

Step-2 : sort the variables in creation order.

Data Set Name	WORK.UNIVERSITY	**Observations**	551
Member Type	DATA	**Variables**	16
Engine	V9	**Indexes**	0
Created	08/14/2024 21:42:59	**Observation Length**	176
Last Modified	08/14/2024 21:42:59	**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	743
Obs in First Data Page	551
Number of Data Set Repairs	0
Filename	/saswork/SAS_workBE9C000183C6_odaws02-euw1.oda.sas.com/SAS_workC9C3000183C6_odaws02-euw1.oda.sas.com/university.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	1610683362
Access Permission	rw-r--r--
Owner Name	u63920100
File Size	256KB
File Size (bytes)	262144

Step-2 : sort the variables in creation order.

Variables in Creation Order

#	Variable	Type	Len	Format	Informat
1	university_name	Char	51	$51.	$51.
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	Char	6	$6.	$6.
16	student_staff_ratio	Char	4	$4.	$4.

The variables are sorted in creation order.

Task-2

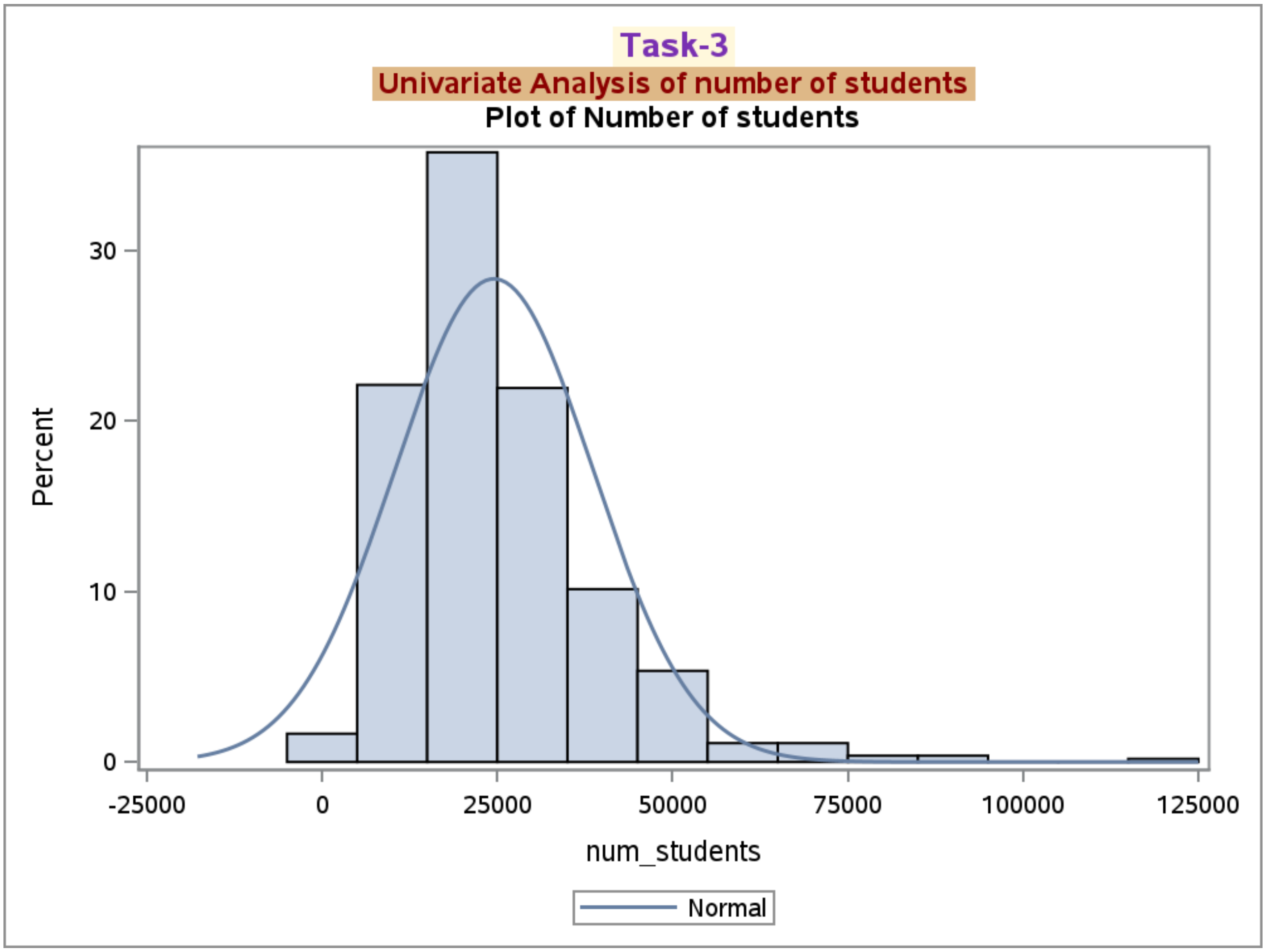
Student/Staff Ratio

Analysis Variable : student_staff_ratio			
Mean	**Std Dev**	**Minimum**	**Maximum**
15.99	10.23	2.90	70.40

The table provides a summary of the student-to-staff ratio, highlighting key statistics such as the mean, standard deviation, as well as the minimum and maximum values.

The average student to staff ratio is 15.99. The variability in the above ratio is 10.23

The dataset shows a minimum staff-student ratio of 2.9 and a maximum of 70.4.



Histogram illustrates the distribution of variable number of students

From the above graph we can say that the univariate analysis depicts a right-skewed distribution.

Most of the student population is concentrated between 10,000 and 40,000. There is a peak around 25,000 suggesting that it is the most common value in the dataset.

This skewness indicates that most institutions fall within a moderate range of student populations, a few have significantly larger or smaller enrollments.

Task-4

Correlation Analysis

Correlation Table for Score, Awards, Publications, and Teaching

4 Variables:	score award pub teaching

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

Correlation table between score, awards, publications, and teaching is printed successfully.

We can see strong positive correlations between all pairs of variables: score, awards, publications, and teaching

The correlations are statistically significantly different from 0.

All correlations are statistically significant, with p-values less than 0.0001.

higher values in score are associated with higher values in the others (awards, publications, teaching), reflecting that there is a possibility that they are interconnected.

Task-5

Hypothesis Test

Variable: num_students

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

Normality Check: Histogram and Probability Plot of Number of Students

Variable: num_students
country = USA

Moments			
N	199	**Sum Weights**	199
Mean	21920.1457	**Sum Observations**	4362109
Std Deviation	12548.0639	**Variance**	157453908
Skewness	1.40342129	**Kurtosis**	4.40904209
Uncorrected SS	1.26794E11	**Corrected SS**	3.11759E10
Coeff Variation	57.2444366	**Std Error Mean**	889.508667

Basic Statistical Measures			
Location		**Variability**	
Mean	21920.15	**Std Deviation**	12548
Median	20626.00	**Variance**	157453908
Mode	2243.00	**Range**	80993
		Interquartile Range	15072

Note: The mode displayed is the smallest of 29 modes with a count of 4.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	24.64298	**Pr > |t|**	<.0001
Sign	**M**	99.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	9950	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	83236
99%	83236
95%	42056
90%	36534
75% Q3	27233
50% Median	20626
25% Q1	12161
10%	7929
5%	6333
1%	2243
0% Min	2243

Normality Check: Histogram and Probability Plot of Number of Students

Variable: num_students
country = USA

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
2243	41	50095	100
2243	40	50095	104
2243	36	50095	108
2243	13	83236	149
5495	223	83236	154

Missing Values			
		Percent Of	
Missing Value	**Count**	**All Obs**	**Missing Obs**
.	5	2.45	100.00

Normality Check: Histogram and Probability Plot of Number of Students

Variable: num_students
country = United Kingdom

Moments			
N	54	**Sum Weights**	54
Mean	18658.9444	**Sum Observations**	1007583
Std Deviation	5698.25826	**Variance**	32470147.1
Skewness	0.36753622	**Kurtosis**	-0.5983921
Uncorrected SS	2.05214E10	**Corrected SS**	1720917799
Coeff Variation	30.5390172	**Std Error Mean**	775.43473

Basic Statistical Measures			
Location		**Variability**	
Mean	18658.94	**Std Deviation**	5698
Median	18670.50	**Variance**	32470147
Mode	18812.00	**Range**	21806
		Interquartile Range	6665

Note: The mode displayed is the smallest of 3 modes with a count of 4.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	24.06256	**Pr > |t|**	<.0001
Sign	**M**	27	**Pr >= |M|**	<.0001
Signed Rank	**S**	742.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	30144.0
99%	30144.0
95%	30144.0
90%	26607.0
75% Q3	20925.0
50% Median	18670.5
25% Q1	14260.0
10%	12001.0
5%	11512.0
1%	8338.0
0% Min	8338.0

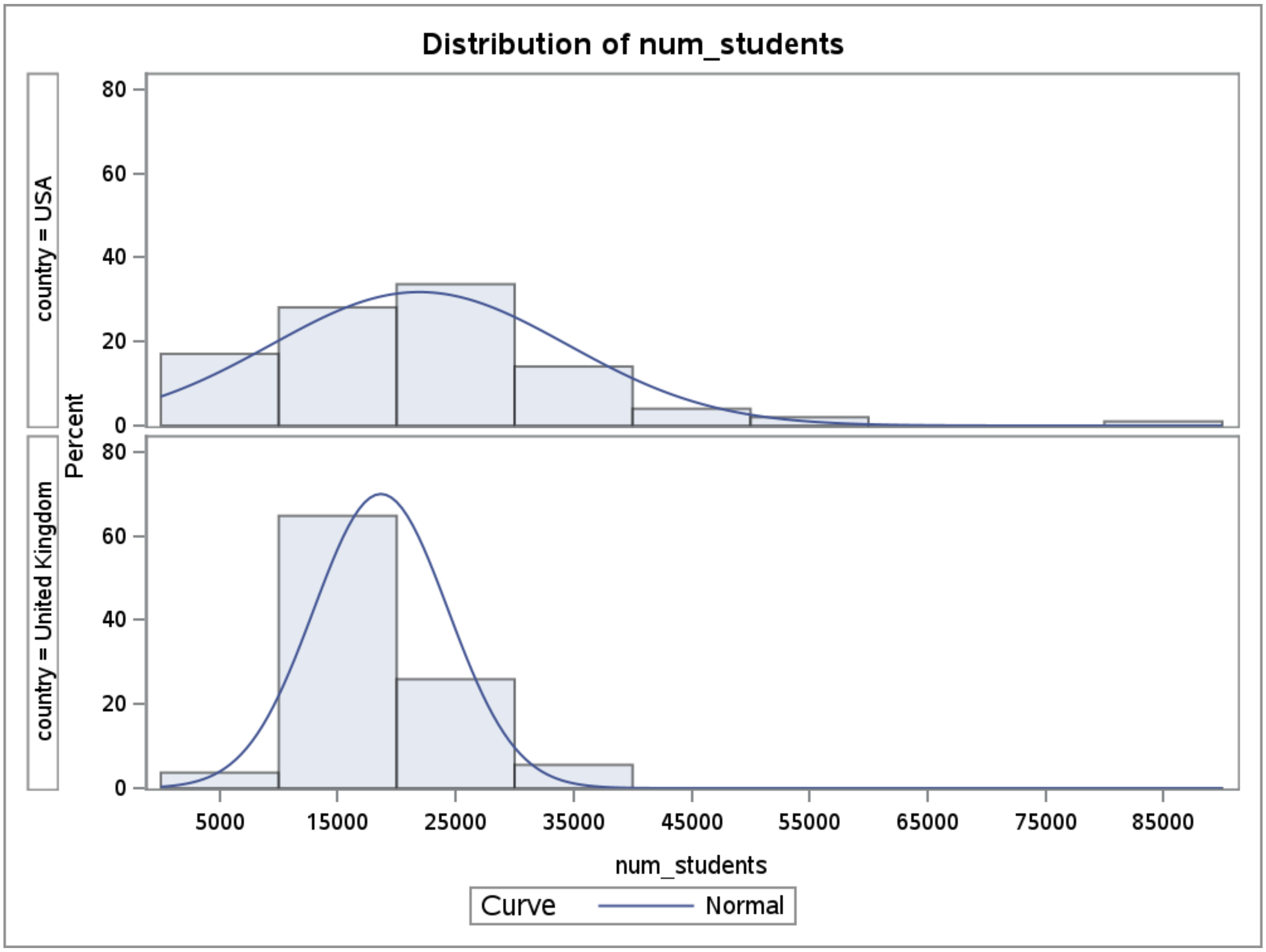
Normality Check: Histogram and Probability Plot of Number of Students

Variable: num_students
country = United Kingdom

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
8338	234	27703	175
8338	230	27703	180
11512	254	30144	150
11512	252	30144	168
12001	212	30144	173

Missing Values			
		Percent Of	
Missing Value	Count	All Obs	Missing Obs
.	2	3.57	100.00

Normality Check: Histogram and Probability Plot of Number of Students



Normality Check: Histogram and Probability Plot of Number of Students

country = USA
Fitted Normal Distribution for num_students

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	21920.15
Std Dev	Sigma	12548.06

Goodness-of-Fit Tests for Normal Distribution

Test	Statistic		p Value	
Kolmogorov-Smirnov	**D**	0.10642207	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	0.35505341	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	2.47946388	**Pr > A-Sq**	<0.005

Quantiles for Normal Distribution

	Quantile	
Percent	Observed	Estimated
1.0	2243.00	-7271.02
5.0	6333.00	1280.42
10.0	7929.00	5839.15
25.0	12161.00	13456.61
50.0	20626.00	21920.15
75.0	27233.00	30383.69
90.0	36534.00	38001.14
95.0	42056.00	42559.87
99.0	83236.00	51111.31

Normality Check: Histogram and Probability Plot of Number of Students

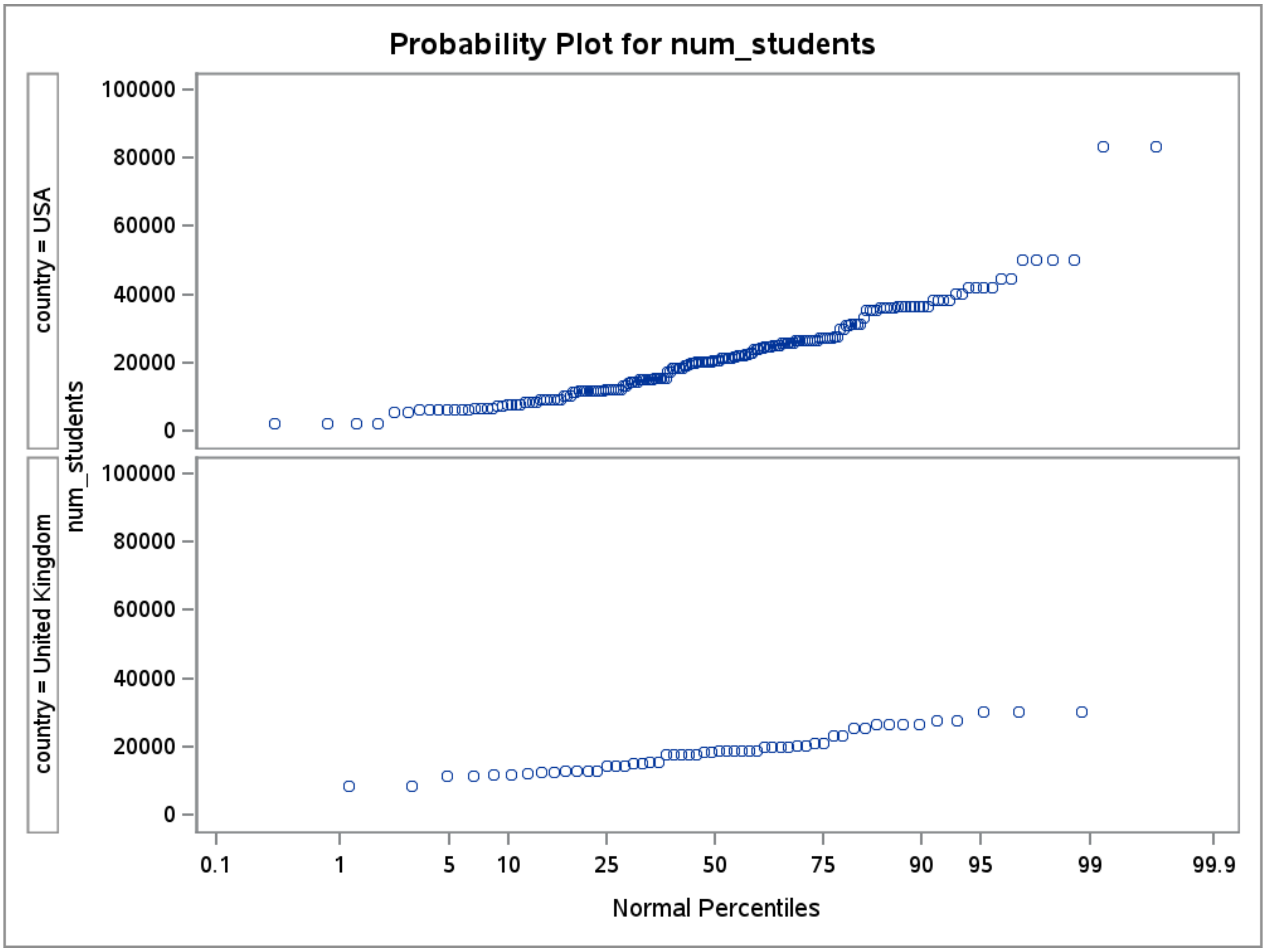
country = United Kingdom
Fitted Normal Distribution for num_students

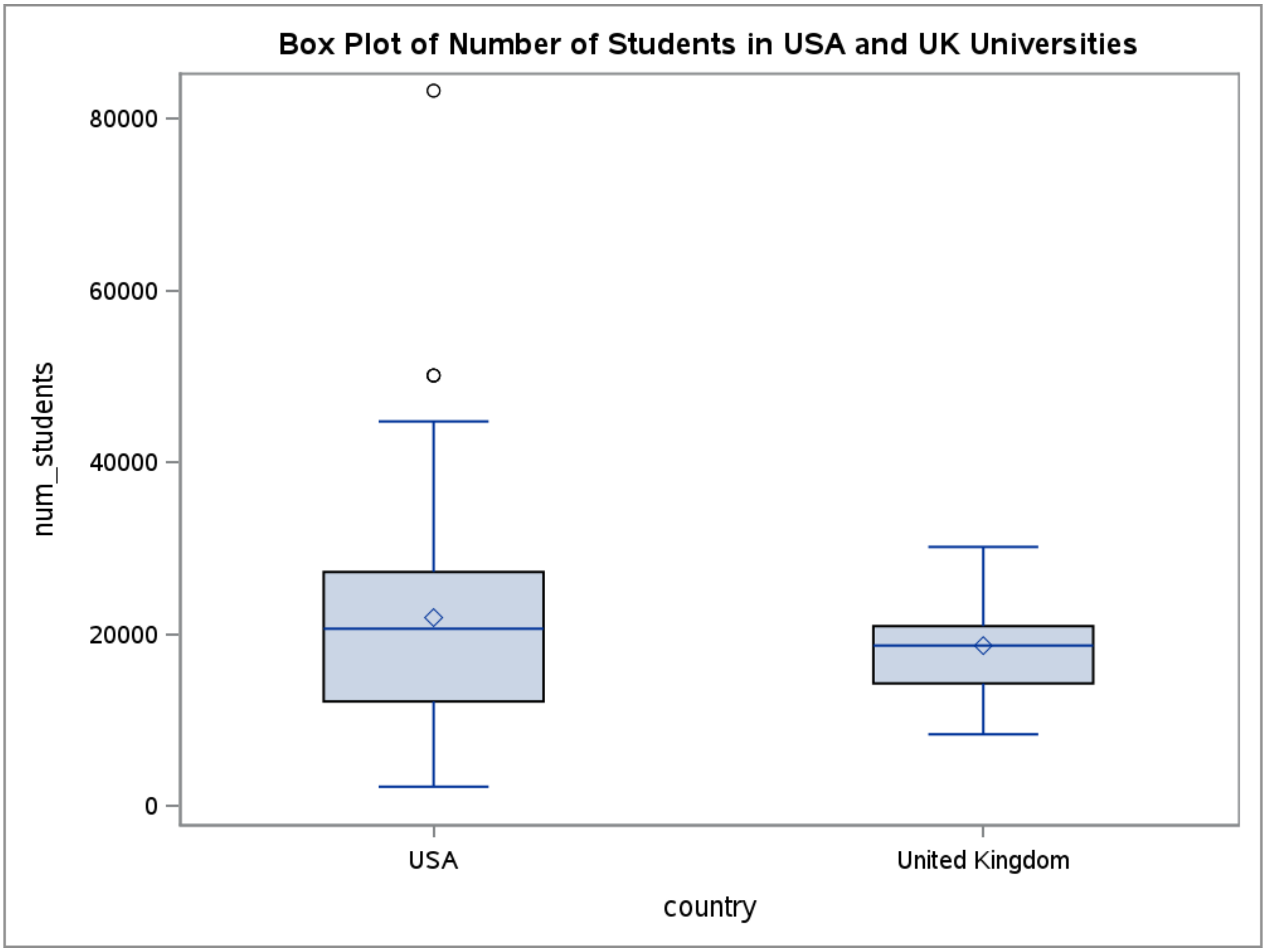
Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	18658.94
Std Dev	Sigma	5698.258

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.11738786	**Pr > D**	0.063
Cramer-von Mises	**W-Sq**	0.13591278	**Pr > W-Sq**	0.037
Anderson-Darling	**A-Sq**	0.86293850	**Pr > A-Sq**	0.025

Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	8338.00	5402.81
5.0	11512.00	9286.14
10.0	12001.00	11356.33
25.0	14260.00	14815.53
50.0	18670.50	18658.94
75.0	20925.00	22502.36
90.0	26607.00	25961.56
95.0	30144.00	28031.75
99.0	30144.00	31915.08

Normality Check: Histogram and Probability Plot of Number of Students





Summaries of the hypothesis tests are printed successfully.

From the analysis, the reports indicate a statistically significant difference between the mean number of students in USA and UK universities.

The box plot also illustrates the differences in student populations between the two countries with outliers in USA data.

The p-value associated with the t-test is less than the significance level of 0.01, leading to the rejection of the null hypothesis.

The normality assumption results show that the data for both countries USA and UK did not perfectly follow a normal distribution.

F test indicates a significant difference in variances between the groups with F-value = 4.85, and p < 0.0001, suggesting that the assumption of equal variances is not met.

The normality assumption results show that the data for both countries USA and UK did not perfectly follow a normal distribution.

Task-6

Subsetted Dataset of United Kingdom, Germany, and Italy

Obs	university_name	world_rank	country	year	quality_of_education
1	University of Oxford	3	United Kingdom	2013	7
2	University of Cambridge	4	United Kingdom	2012	10
3	University of Cambridge	4	United Kingdom	2014	2
4	University of Cambridge	4	United Kingdom	2015	2
5	University of Cambridge	5	United Kingdom	2013	3

First few observations from universities dataset is printed successfully using PROC PRINT step.

Step-1 : Printing observations from 10 to 17 and the first 5 variables.

Obs	university_name	world_rank	country	year	quality_of_education
10	University College London	30	United Kingdom	2013	24
11	University College London	30	United Kingdom	2014	20
12	University College London	31	United Kingdom	2012	35
13	University of Nottingham	97	United Kingdom	2012	101
14	University of Bonn	98	Germany	2014	23
15	University of Bristol	98	United Kingdom	2012	101
16	Sapienza University of Rome	112	Italy	2015	67
17	University of Bristol	123	United Kingdom	2014	177

Observations from 10 to 17 from universities dataset is printed successfully.

Step-2 : Printing highest ranked Italian university.

Highest Ranked Italian University

Obs	university_name	world_rank	country
16	Sapienza University of Rome	112	Italy

Sapienza University of Rome is the highest ranked Italian university. It is ranked 112th in the world.

Task-7

Mean Quality of Education

Mean Quality of Education for the Whole uni1 Dataset

Analysis Variable : quality_of_education
Mean
213.55

Task-7

Mean Quality of Education

Mean Quality of Education for the Whole uni1 Dataset

Mean Quality of Education for Universities with Quality of Education > 100

Analysis Variable : quality_of_education
Mean
266.37

Mean tables for mean quality for education for universities with quality education is printed successfully.

The mean quality of education for this new dataset(uni1) is 213.55

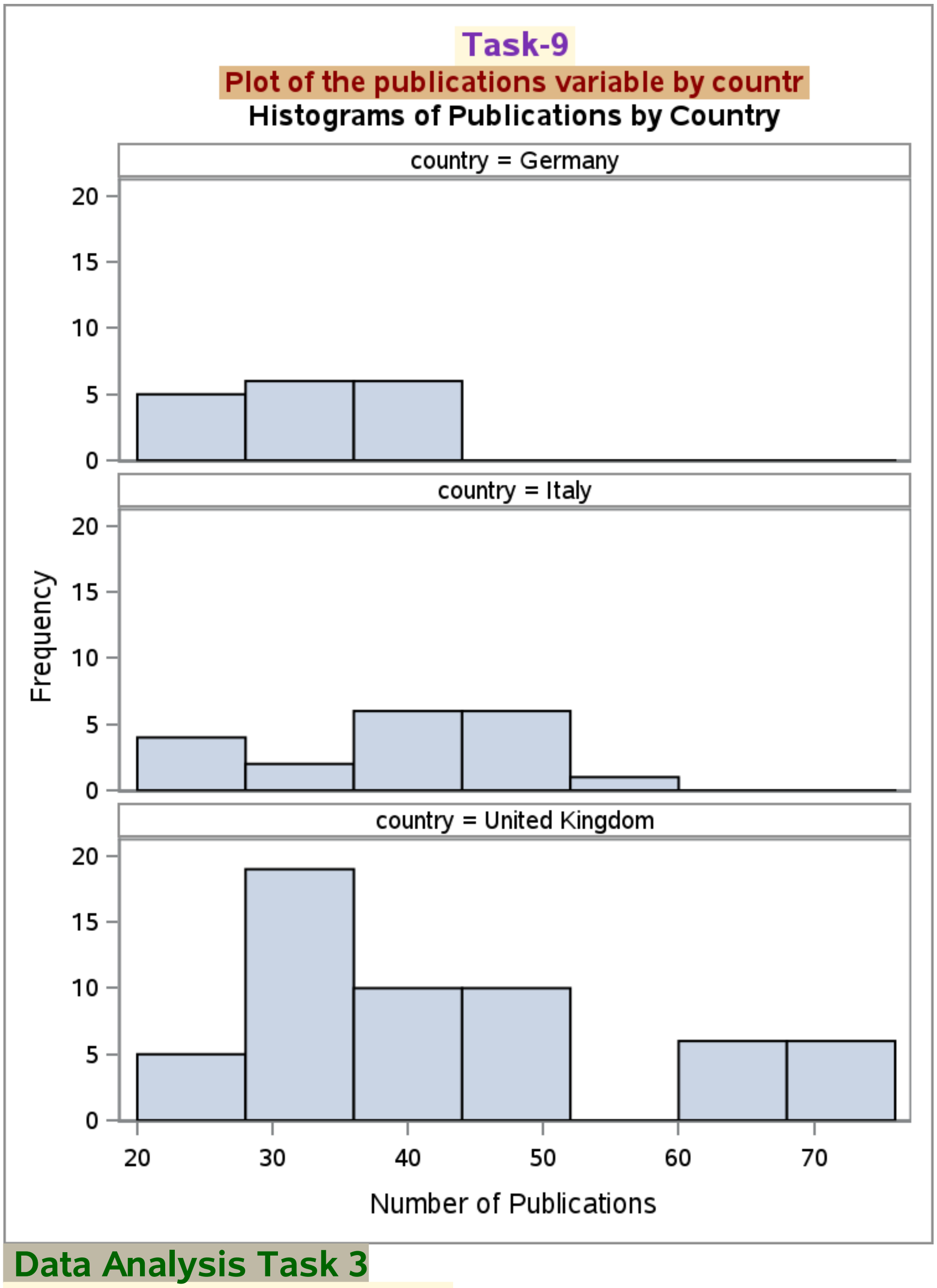
The mean quality of education greater than 100 is 286.37

Task-8

Summary Statistics for the Patents variable

Analysis Variable : patents						
	N					
country	Obs	N	Mean	Std Dev	Minimum	Maximum
Germany	17	17	386.47	187.56	138.00	774.00
Italy	19	19	532.21	121.10	312.00	737.00
United Kingdom	56	56	305.84	204.70	15.00	871.00

The Summary statistics summary table for the patents variable has been printed successfully.
Italy has the highest mean number of patents, while UK has the widest and highest number of patents
Germany shows moderate valyes with a narrower range.



Task-9

Plot of the publications variable by countr

Histograms of Publications by Country

Data Analysis Task 3

Tasks & Utilities : Data Mining

Task-1 : Rapid Predictive Modeler

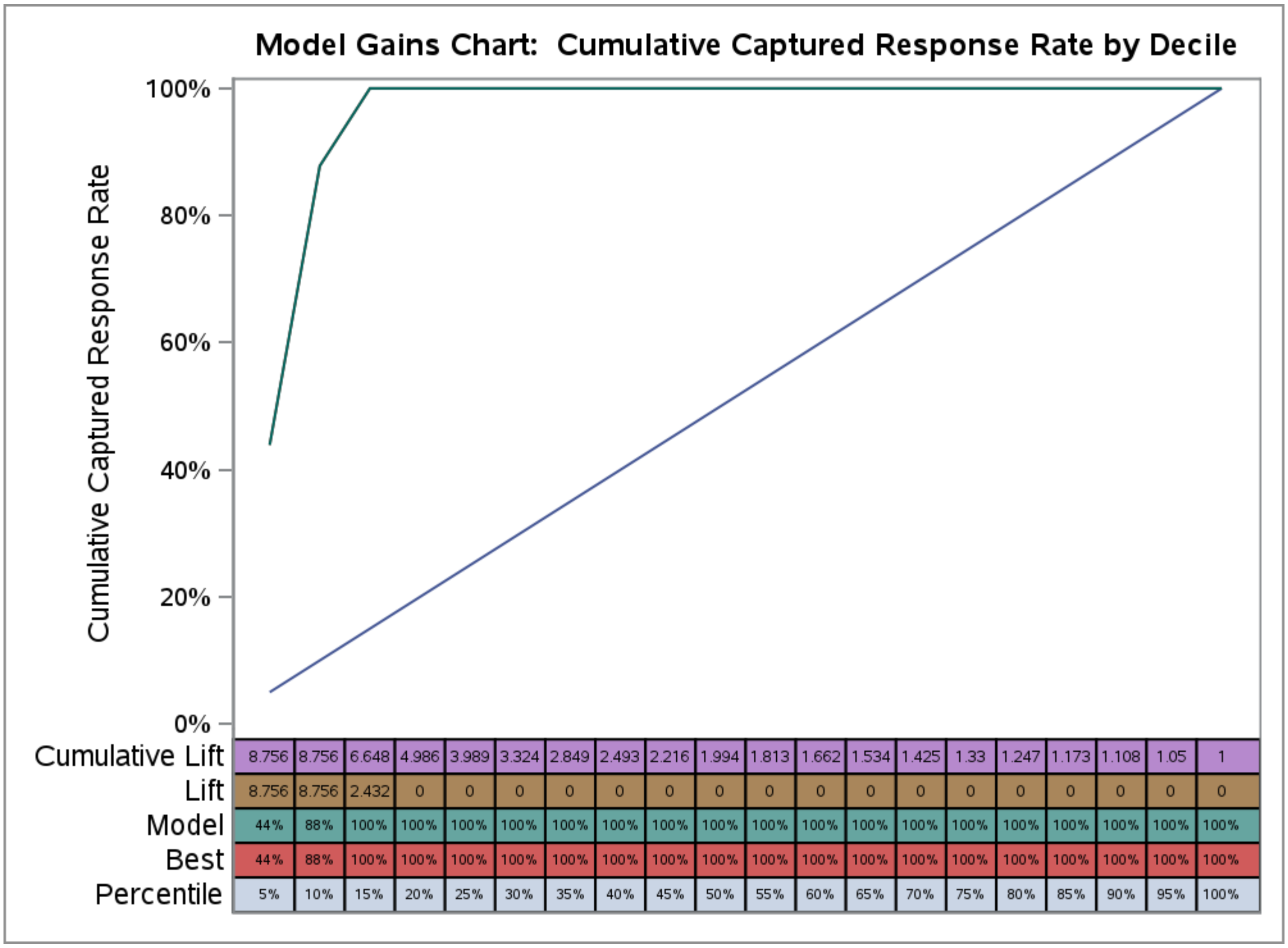
Histogram displaying the distribution of number of pulications by country has been printed successfully.

Germany's distribution is uniform but with a narrow range. Most of the publicaitons are concentrated between 30 and 60.

Italy follows the similar trend but contains variability and wider spread.

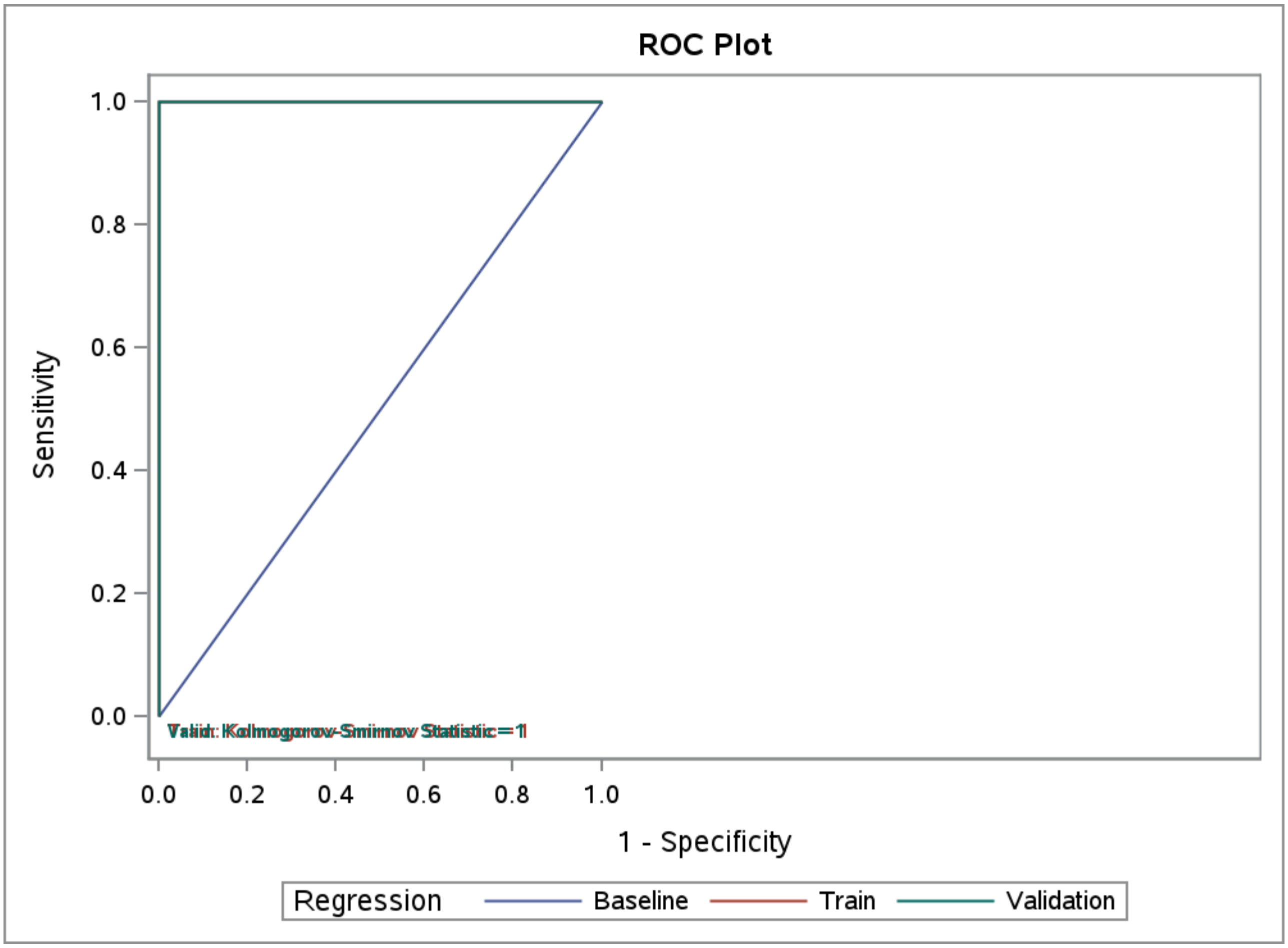
UK's distribution is more varied, most of the publications are concentrated between 30-40.

```
The computed Average Squared Error for index_id is 3.7411555E-7.

This indicates a possible target duplication issue.

Please review the list of inputs used in the model.
```



Model Gains Chart: Cumulative Captured Response Rate by Decile

Cumulative Lift	8.756	8.756	6.648	4.986	3.989	3.324	2.849	2.493	2.216	1.994	1.813	1.662	1.534	1.425	1.33	1.247	1.173	1.108	1.05	1
Lift	8.756	8.756	2.432	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Model	44%	88%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Best	44%	88%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Percentile	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%


ROC Plot
1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
Train: Kolmogorov-Smirnov Statistic = 1
Valid: Kolmogorov-Smirnov Statistic = 1
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
Regression
Baseline
Train
Validation

Variable Attribute Importance

		Scorecard Points
index_name	#INDEX+NAME	0.00
	GENDER DEVELOPMENT INDE	868.00
	GENDER INEQUALITY INDEX	903.00
	HUMAN DEVELOPMENT INDEX	934.00
	INEQUALITY-ADJUSTED HUM	963.00
	PLANETARY PRESSURES–A	1000.00

Project Information

Property	Value
Name	RPM
Diagram	RPM2 - index_id
Path	/saswork/SAS_workBE9C000183C6_odaws02-euw1.oda.sas.com/ SAS_workC9C3000183C6_odaws02-euw1.oda.sas.com./ u63920100/RPM
Date Created	14Aug2024:21:43:20

Data Mining using SAS Rapid Predicitive Modeler on Human Development Indicators dataset of Switzerland.

Rapid Predictive Modeler is designed to automate the process of predictive modelling.

Here we will demonstrate how to analyze the dataset, build the predicitive model and evaluate the performance using RPM.

The purpose of the data mining task is streamiling the creation of predictive models. RPM reduces the complexity of predictive modeling.

The selected dataset is Human Development Indicators dataset of Switzerland where the variable `index_id` is used for the predicitions.

The system then automatically identifies other variables(here value and year) such as inputs that contribute towards building the predicitive model.

RPM selects appropriate features from the dataset, runns the standard modeling procedures and then automatically determines the best model type based on the data and task requirements.

The primary metric that is used to evaluate the performance is Average Squared Error. In our case it is calculated as `3.7411555E-7`. This lower value suggests that the model performed well.

This suggests that the model predictions were highly accurate, with minimal deviation from the actual target values in the dataset.

Key variables such as the Gender Inequality Index and the Human Development Index were identified as significant contributors to the model's predictions.

Gender Inequality Index (GII) and Human Development Index (HDI) were among the top variables contributing to the model's predictions.

he scorecard provides a detailed breakdown of the points assigned to each variable based on their importance in the model. It allows for a clear comparison of how different variables contribute to the final prediction.

The ROC curve for the validation data shows a sharp increase from (0,0) to (0,1), indicating that the model achieves perfect sensitivity with no false positives.

From the ROC plot, The KS Statistic of 1 suggests that there is a perfect distinction between the positive and negative classes in the validation set.

The gains chart shows a steep rise in the cumulative captured response rate within the first two deciles, reaching close to 100% by the second decile.

This indicates that the model is extremely effective in ranking the positive responses.

The steepness of the curve in the first few deciles and the flatness thereafter imply that the model is highly effective in concentrating positive responses at the top of the ranking.

The green line being significantly above the diagonal baseline line throughout the chart confirms that the model is performing much better than random guessing.

The detailed table under the graph shows that the model captures 100% of positive responses by the 4th decile

The cumulative lift values confirm that most of the lift is achieved within the first two deciles, which is typical for a highly effective model.

These findings collectively highlight the strengths of using the SAS Rapid Predictive Modeler for data mining tasks, particularly its ability to automate complex processes while maintaining a high level of accuracy in its predictions.