STAT40840 - FINAL PROJECT

Sarvesh Sairam Naik - 22204841

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 I

Importing the dataset.

Obs	ID	Gender	Age	Customer Type	Type of Travel	Class	Flight Distance	Departure Delay	Arrival Delay	Departure and Arrival Time Conve	Ease of Online Booking	Check-in Service	Online Boarding	Gate Location	On-board Service	Seat Comfort	Leg Room Service	Cleanliness	Food and Drink	In-flight Service	In-flight Wifi Service	In-flight Entertainment	Baggage Handling	Satisfaction
1	1	Male	48	First-time	Business	Business	821	2	5	3	3	4	3	3	3	5	2	5	5	5	3	5	5	Neutral or Dissatisfied
2	2	Female	35	Returning	Business	Business	821	26	39	2	2	3	5	2	5	4	5	5	3	5	2	5	5	Satisfied
3	3	Male	41	Returning	Business	Business	853	0	0	4	4	4	5	4	3	5	3	5	5	3	4	3	3	Satisfied
4	4	Male	50	Returning	Business	Business	1905	0	0	2	2	3	4	2	5	5	5	4	4	5	2	5	5	Satisfied
5	5	Female	49	Returning	Business	Business	3470	0	1	3	3	3	5	3	3	4	4	5	4	3	3	3	3	Satisfied

About the dataset :Customer satisfaction scores from 120,000+ airline passengers, including additional information about each passenger, their flight, and type of travel, as well as ther evaluation of different factors like cleanliness, comfort, service, and overall experience.

Which percentage of airline passengers are satisfied? Does it vary by customer type? What about type of travel?

'

Satisfaction Percentage by Overall

The FREQ Procedure

Satisfaction	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Neutral or Dissatisfied	73452	56.55	73452	56.55
Satisfied	56428	43.45	129880	100.00

'

Obs	COUNT	PERCENT
1	73452	56.5537
2	56428	43.4463

From the tables shown above we can infer that about 73452 that is 56.5537% of airline customers are Neutral or Dissatisfied with the overall service. And remaining 56428 or 43.4463% of the customers are satisfied by the overall service.

Satisfaction Percentage by Customer Type

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of Satisfaction by Customer Type			
	Customer Type		
Satisfaction	First-time	Returning	Total
Neutral or Dissatisfied	18080 13.92 24.61 76.03	55372 42.63 75.39 52.19	73452 56.55
Satisfied	5700 4.39 10.10 23.97	50728 39.06 89.90 47.81	56428 43.45
Total	23780 18.31	106100 81.69	129880 100.00

'

Obs	Customer Type	COUNT	PERCENT

Obs	Customer Type	COUNT	PERCENT
1	First-time	18080	13.9205
2	Returning	55372	42.6332
3	First-time	5700	4.3887
4	Returning	50728	39.0576

We can infer from the dataset that about First-time customers who account to about 13.9205% and Returning customers who account to 42.6332% in the dataset are Neutral/Dissatisfied with the service. Very low in number 5700 of First-time customers were satisfied with the service. And, Returning about 50728 (39.0576%) of customers were satisfied with the service. Notably, First-time customers show a higher proportion of Neutral/Dissatisfied responses, while Returning customers exhibit a more favorable satisfaction rate.

Satisfaction Percentage by Class of Travel

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of Satisfaction by Class			
	Class		
Satisfaction	Business	Economy	Total
Neutral or Dissatisfied	18994 14.62 25.86 30.56	54458 41.93 74.14 80.42	73452 56.55
Satisfied	43166 33.24 76.50 69.44	13262 10.21 23.50 19.58	56428 43.45
Total	62160 47.86	67720 52.14	129880 100.00

Satisfaction Percentage by Class of Travel

Obs	Class	COUNT	PERCENT
1	Business	18994	14.6243
2	Economy	54458	41.9295
3	Business	43166	33.2353
4	Economy	13262	10.2110

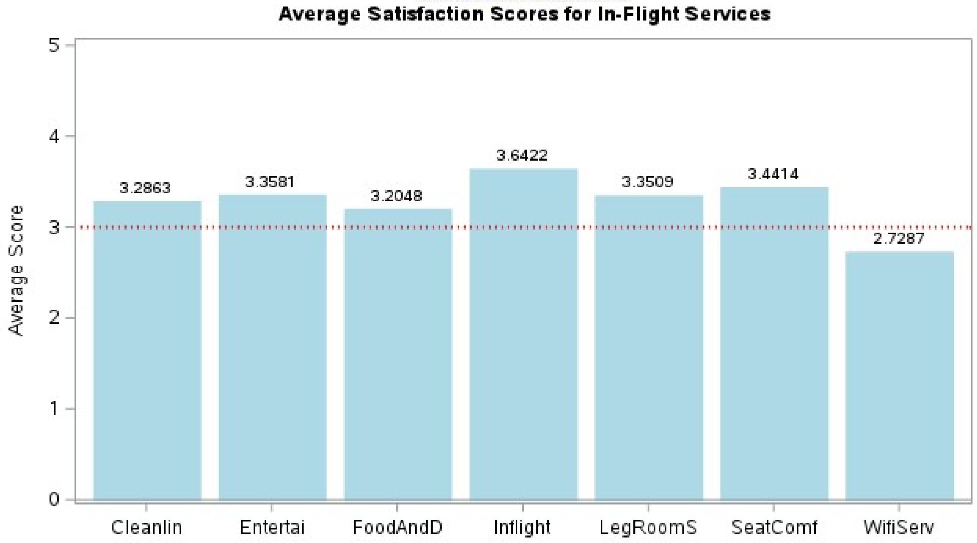
It is observed that 18994 cutomers flying Business class and 54458 customers flying Economy are Neutral/Dissatisfied with the service. And, 43166 Business and 13262 Economy flyers were satisfied with the service of the airline. Business class flyers display a higher satisfaction rate (43.17%) compared to Economy class flyers (13.26%). These findings emphasize the need for improvements to enhance customer satisfaction, particularly among First-time customers and Economy class flyers.

'

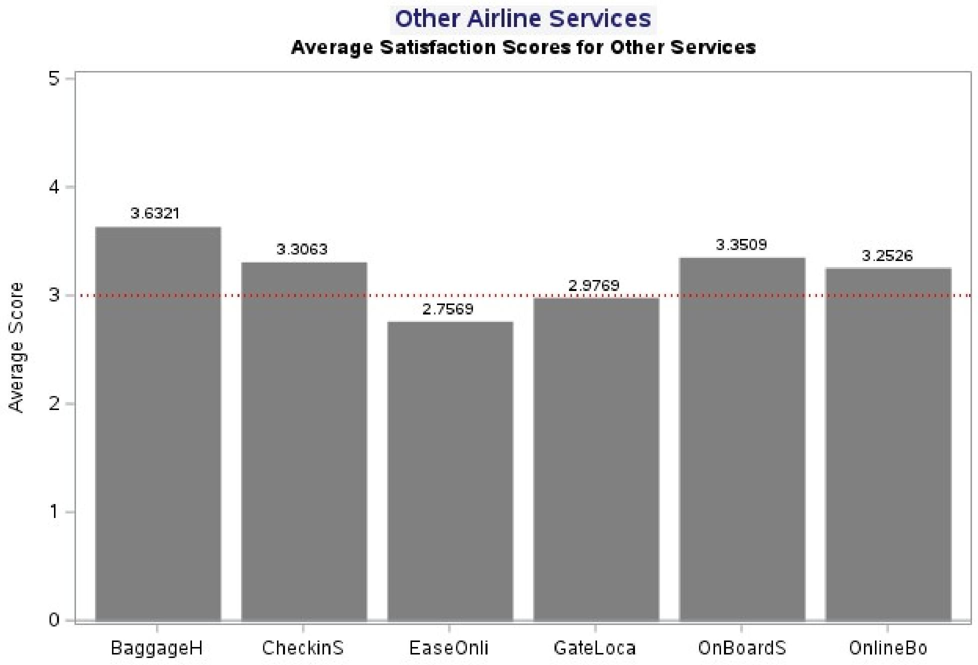
To enhance the overall customer experience, the airline should prioritize addressing the concerns of First-time customers and Economy class flyers. Additionally, it is vital to maintain the high satisfaction levels among Returning customers and Business class flyers, as they currently represent the more satisfied customer segments.

Which factors contribute to customer satisfaction the most?

In-Flight Services


Average Satisfaction Scores for In-Flight Services

We can see from the chart that passengers are not happy with the Inflight Wifi Service. The Food and Drink service can also be improved. The Inflight steward service has the highest average score and thus is positively affecting the satisfaction of the customers.



We can infer that lack of Ease of Online booking and outlying Gate Locations are critical factors affecting the passenger satisfaction. Meanwhile customers are happy with the Baggage Handling of the airline.

Is there any underlying relation between Arrival Delay, Departure Delay and Flight Distance ?

Obs	_TYPE_	_FREQ_	Mean_Std	Std_Std
1	0	129880	1190.3163921	997.45247733

In summary, there is no strong linear relationship between Flight Distance and either Arrival Delay or Departure Delay. However, there is a very strong positive linear relationship between Arrival Delay and Departure Delay, indicating that they are closely related.

Is there any underlying relation between Arrival Delay, Departure Delay and Flight Distance ?

The CORR Procedure

3 Variables:	Flight Distance Arrival Delay Departure Delay

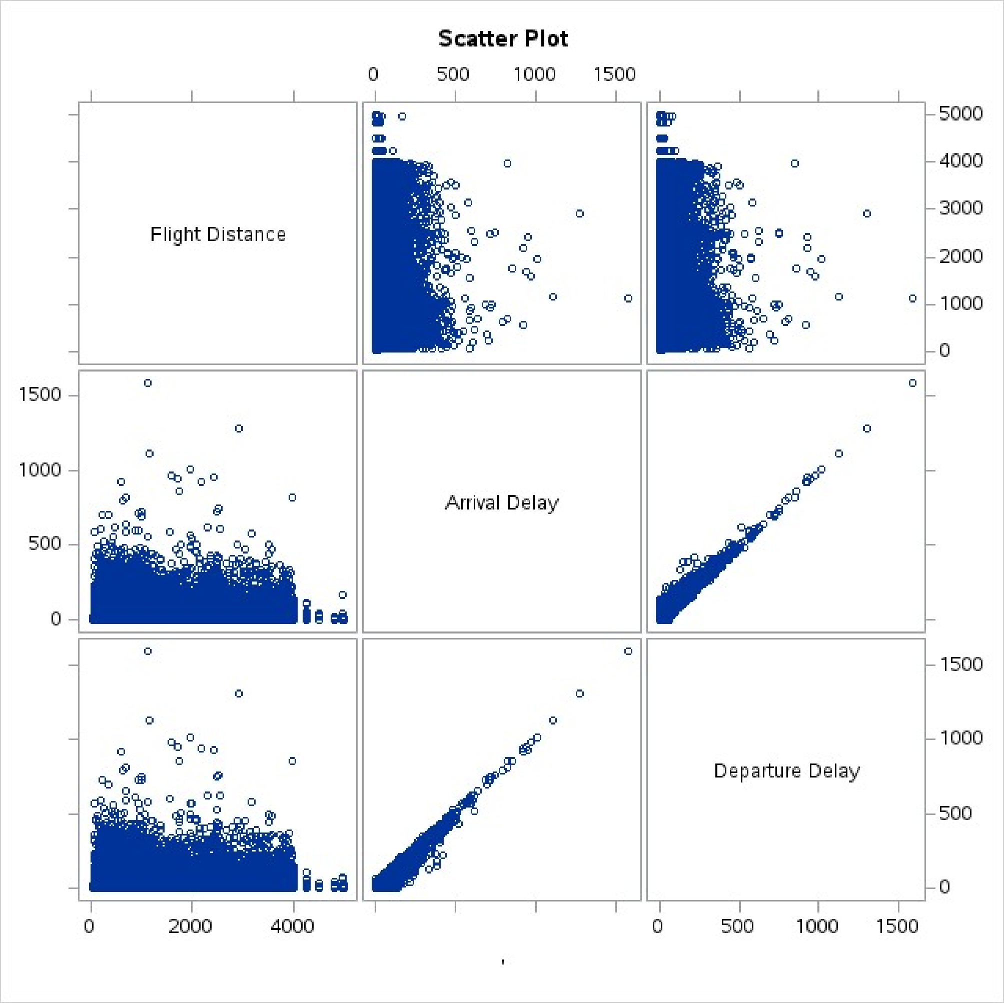
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Flight Distance	129880	1190	997.45248	154598293	31.00000	4983
Arrival Delay	129487	15.09113	38.46565	1954105	0	1584
Departure Delay	129880	14.71371	38.07113	1911017	0	1592

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Flight Distance	Arrival Delay	Departure Delay
Flight Distance	1.00000 129880	-0.00193 0.4863 129487	0.00240 0.3867 129880
Arrival Delay	-0.00193 0.4863 129487	1.00000 129487	0.96529 <.0001 129487
Departure Delay	0.00240 0.3867 129880	0.96529 <.0001 129487	1.00000 129880

In summary, there is no strong linear relationship between Flight Distance and either Arrival Delay or Departure Delay. However, there is a very strong positive linear relationship between Arrival Delay and Departure Delay, indicating that they are closely related.



Data Analysis II

1. Read the dataset erasmus.csv into SAS and call the resulting table erasmus, saving it in the s40840 library. The file contains column names on the first row, with the first observation starting on the second row. You should ensure your code will overwrite any previous object of the same name.

a)Print the 1st 4 rows of the resulting erasmus table

Obs	academic_year	duration	nationality	gender	age	sending_country	sending_city	receiving_country	receiving_city
1	2014-2015	1	AT	Female	13	AT	Dornbirn	AT	Dornbirn

Obs	academic_year	duration	nationality	gender	age	sending_country	sending_city	receiving_country	receiving_city
2	2014-2015	1	AT	Female	14	AT	Dornbirn	AT	Dornbirn
3	2014-2015	1	AT	Female	15	AT	Dornbirn	AT	Dornbirn
4	2014-2015	1	AT	Male	14	AT	Dornbirn	AT	Dornbirn

b)The duration variable is stored in months. Find the mean duration spent in the programme by students of Irish nationality (`IE'). How many students of Irish nationality are in the dataset?

Obs	nationality	mean_duration	_FREQ_
1	IE	1.4148282098	2765

As seen, the mean duration spent in programme by 2765 Irish students is around 43 days.

c) One student is older than all other participants. What is the age of this student? In what city did this student study? In what academic year did they start?

Oldest Student Details

Obs	age_oldest	sending_city	academic_year
1	80	Valencia	2018-2019

The age of the student is 80. The student studied in Valencia. The student's academic year started in 2018.

d) Create a table of the nationality variable for students who are not from Ireland (that is, their nationality is not `IE') and whose receiving city is Dublin. The table should be ordered from highest to lowest frequency. What is the most frequent nationality of non-Irish students who studied in Dublin?

Frequency Table of Non-Irish Students Studying in Dublin

Obs	nationality	COUNT
1	UK	37
2	CZ	14
3	IT	14
4	BE	13
5	ES	11
6	NO	11
7	PL	11
8	AM	10
9	AT	9
10	DE	9
11	NL	8
12	EL	7
13	FR	6
14	AL	4
15	CA	4
16	HU	4
17	IS	4

Obs	nationality	COUNT
18	LU	4
19	RO	4
20	SO	4
21	TR	4
22	FI	3
23	IN	3
24	NG	3
25	JM	2
26	LT	2
27	ZM	2
28	ZW	2
29	CL	1
30	GM	1
31	KR	1
32	RS	1
33	SK	1
34	US	1

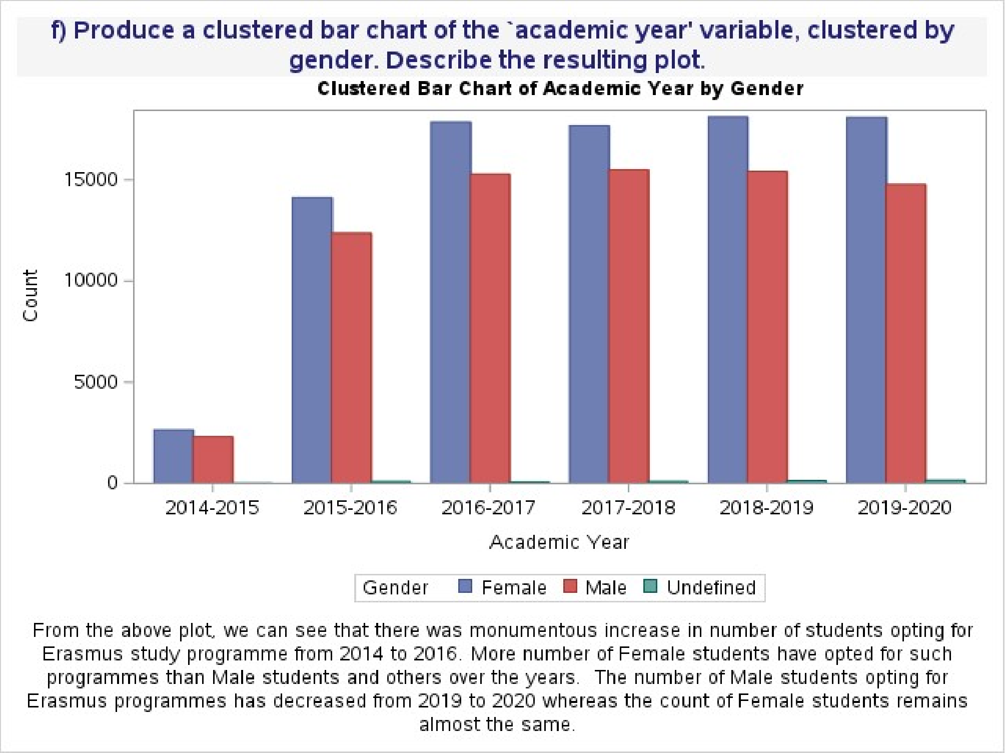
The most frequent nationality of non-irish students who studied in Dublin are from UK count:37

e) In a single table, print the summary statistics for the age variable, divided into groups by both gender and academic year. Which cohort had the greatest mean age?

Summary Statistics for Age by Gender and Academic Year

Obs	gender	academic_year	mean_age
1	Female	2018-2019	25.021495275
2	Undefined	2018-2019	24.971830986
3	Male	2016-2017	24.866618649
4	Female	2017-2018	24.864323315
5	Female	2016-2017	24.752730328
6	Male	2018-2019	24.740490718
7	Female	2019-2020	24.635437995
8	Male	2017-2018	24.525653437
9	Male	2015-2016	24.450853491
10	Male	2019-2020	24.433254318
11	Male	2014-2015	24.099001303
12	Female	2015-2016	24.065140346
13	Female	2014-2015	23.941220799
14	Undefined	2019-2020	21.932098765
15	Undefined	2017-2018	21.242424242
16	Undefined	2016-2017	20.712328767
17	Undefined	2015-2016	20.625
18	Undefined	2014-2015	18

As observed from the table, Female in academic year 2018-2019 have the greatest mean age -25.021495275



f) Produce a clustered bar chart of the `academic year' variable, clustered by gender. Describe the resulting plot.

From the above plot, we can see that there was monumentous increase in number of students opting for Erasmus study programme from 2014 to 2016. More number of Female students have opted for such programmes than Male students and others over the years. The number of Male students opting for Erasmus programmes has decreased from 2019 to 2020 whereas the count of Female students remains almost the same.

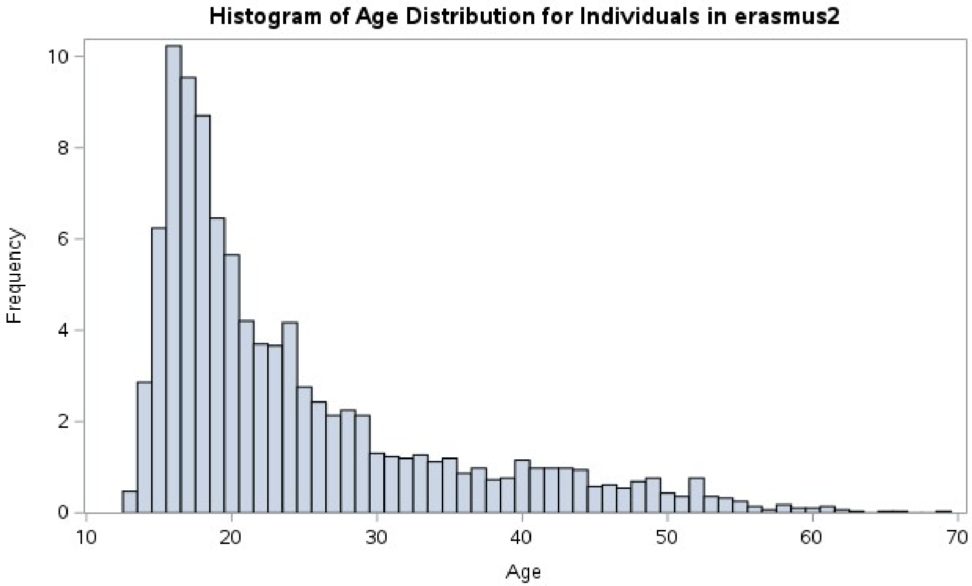
2. For this question, create a subset of the erasmus dataset which contains only those individuals whose receiving country is Ireland (`IE'). Call this subset erasmus2 and use this subset for all of the follwing parts

'

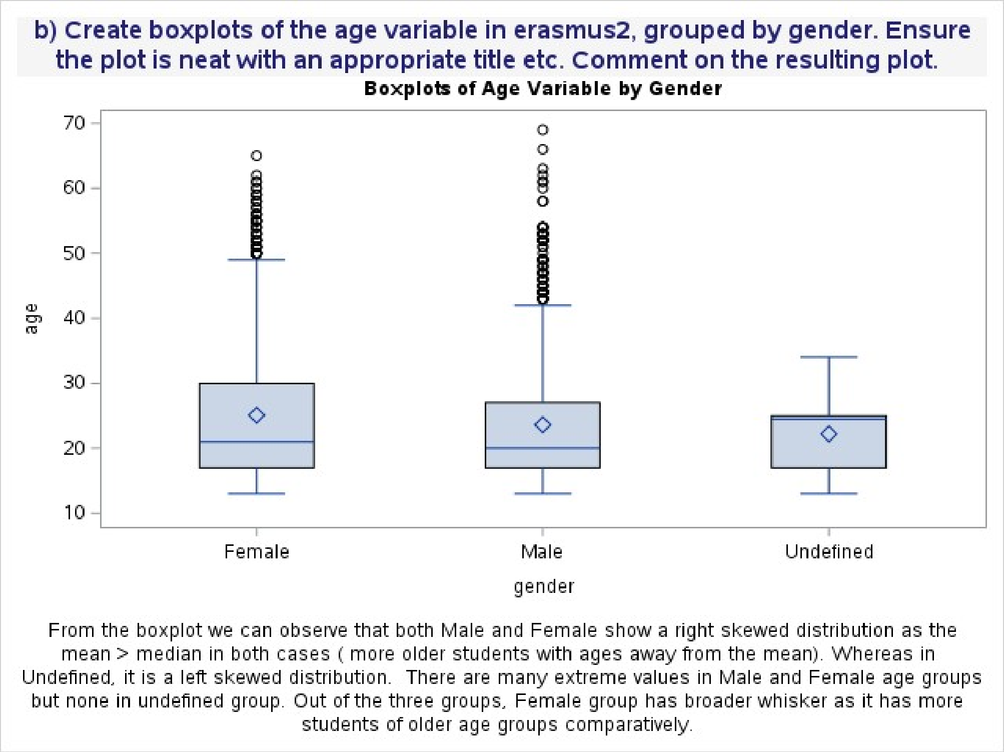
a) Conduct a univariate analysis of the age variable for those individuals in erasmus2. Write a short description of your findings, including key statistics and discussion of any plots produced.

Obs	mean_age	max_age	median_age	min_age
1	24.4084	69	20	13

The average age of individuals in the dataset is around 24.41 years, with the age range spanning from 13 to 69 years. The median age of 20 years suggests that half of the individuals are younger than 20, and the other half are older.



As seen from the histogram, since the mean (24.41 years) is greater than the median (20 years), the age distribution is right-skewed. In a right-skewed distribution, the tail of the distribution extends more to the right, indicating that there are relatively more older individuals with ages farther from the mean. In this case, the maximum age (69 years) is causing the rightward tail in the distribution, pulling the mean higher than the median.



b) Create boxplots of the age variable in erasmus2, grouped by gender. Ensure the plot is neat with an appropriate title etc. Comment on the resulting plot.

Boxplots of Age Variable by Gender

From the boxplot we can observe that both Male and Female show a right skewed distribution as the mean > median in both cases (more older students with ages away from the mean). Whereas in Undefined, it is a left skewed distribution. There are many extreme values in Male and Female age groups but none in undefined group. Out of the three groups, Female group has broader whisker as it has more students of older age groups comparatively.

c) Conduct a hypothesis test to see if there is a statictically significant difference between the mean ages of female and male students, using as your sample data those students in erasmus2. State your hypotheses carefully, check all assumptions necessary, run your chosen test, comment on the resulting plots and state your conclusion clearly. Use a significance level of α = 0.01.

The TTEST Procedure

Variable: age

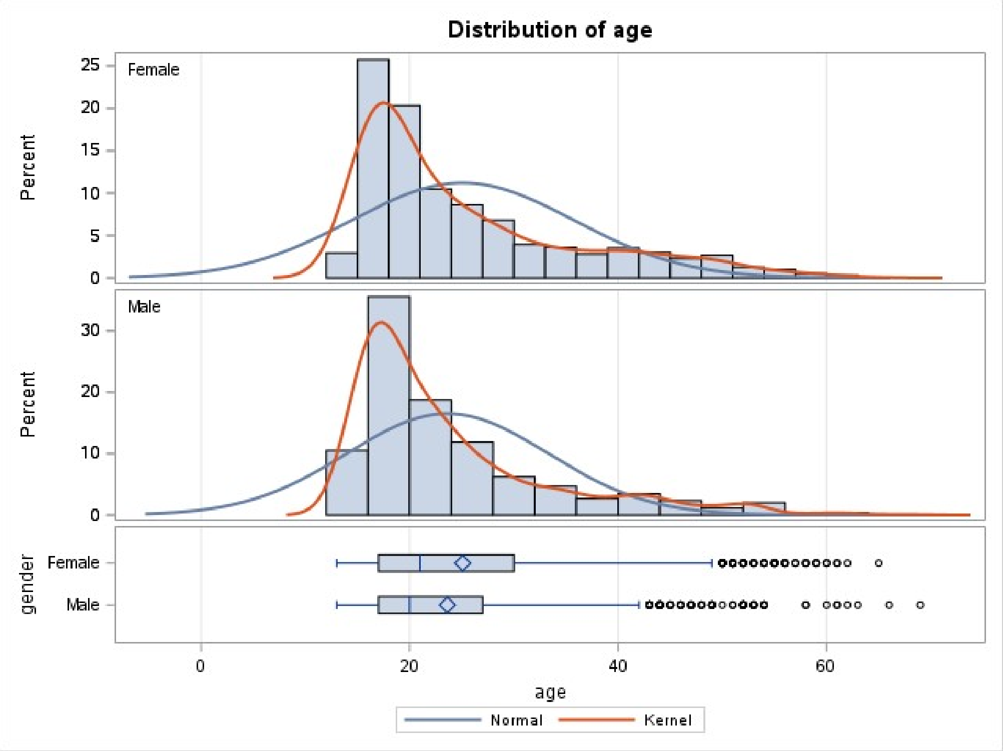
gender	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female		1496	25.0909	10.6835	0.2762	13.0000	65.0000
Male		1237	23.6257	9.6632	0.2748	13.0000	69.0000
Diff (1-2)	Pooled		1.4652	10.2343	0.3933		
Diff (1-2)	Satterthwaite		1.4652		0.3896		

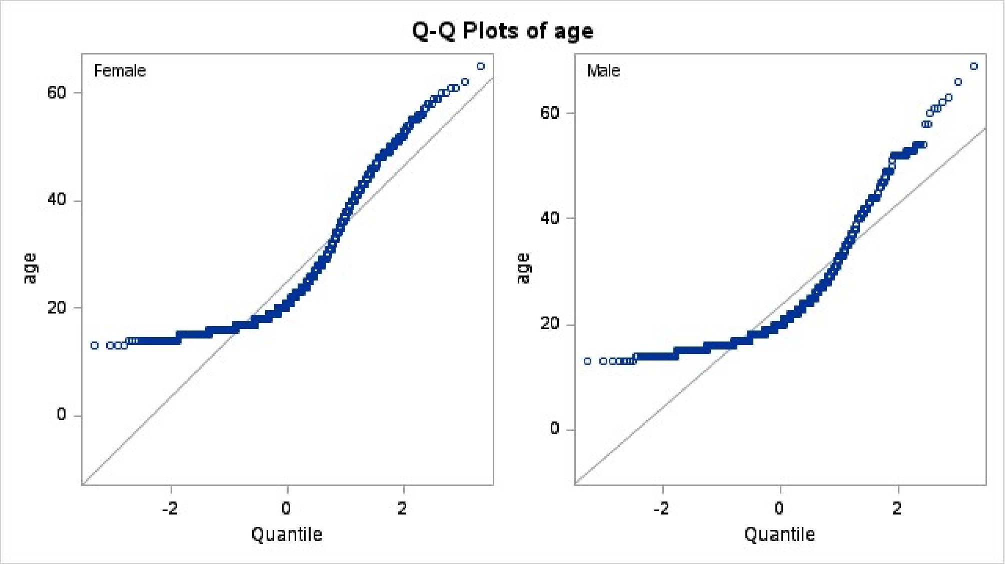
gender	Method	Mean	99% CL Mean		Std Dev	99% CL Std Dev	
Female		25.0909	24.3785	25.8033	10.6835	10.2015	11.2098
Male		23.6257	22.9169	24.3345	9.6632	9.1858	10.1892
Diff (1-2)	Pooled	1.4652	0.4514	2.4790	10.2343	9.8889	10.6030
Diff (1-2)	Satterthwaite	1.4652	0.4610	2.4694			

Method	Variances	DF	t Value	Pr > |t|

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2731	3.73	0.0002
Satterthwaite	Unequal	2709.1	3.76	0.0002

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	1495	1236	1.22	0.0002





Here is a summary of the key findings:

1. Sample Sizes and Descriptive Statistics: - For Female students: The sample size is 1496, with a mean age of 25.0909 years and a standard deviation of 10.6835 years. - For Male students: The sample size is 1237, with a mean age of 23.6257 years and a standard deviation of 9.6632 years.
2. Difference in Means (Pooled and Satterthwaite): - The difference in mean ages between Female and Male students is approximately 1.4652 years. - The standard error of the difference is 0.3933 for the Pooled method and 0.3896 for the Satterthwaite method.
3. Confidence Intervals: - For Female students, the 99% confidence interval for the mean age is between 24.3785 and 25.8033 years. - For Male students, the 99% confidence interval for the mean age is between 22.9169 and 24.3345 years. - The 99% confidence interval for the difference in means (Pooled method) is between 0.4514 and 2.4790 years.
4. T-Test Results: - The t-value for the t-test is approximately 3.73 for the Pooled method and 3.76 for the Satterthwaite method. - The p-values for both t-tests are very small (less than 0.0002), indicating statistical significance at the α = 0.01 level.
5. Equality of Variances: - The test for equality of variances (Folded F-test) indicates that the variances of the age distributions for Female and Male students are statistically significantly different (p-value < 0.0002).

Conclusion:Based on the t-test results, we reject the null hypothesis that there is no statistically significant difference between the mean ages of Female and Male students in the `erasmus2` subset. The data provide strong evidence to support the alternative hypothesis that there is a significant difference in the mean ages of Female and Male students. The confidence intervals for the mean ages of each group and the difference between the groups' means further confirm the significance of the findings. Additionally, the unequal variances suggest that the groups might have different age distributions.

Tasks demonstration

Statistics Task

The SAS Statistics Task provides a user-friendly interface to perform various statistical analyses on data. It offers a range of options to explore data, calculate summary statistics, Distribution Analysis, Correlation Analysis, Table Analysis and conduct tests.

1. Summary Statistics

Summary Statistics for Width

Obs	_TYPE_	_FREQ_	Mean_Std	Min_Std	Max_Std	P25_Std	P50_Std	P75_Std
1	0	159	4.41749	1.0476	8.142	3.3756	4.2485	5.589

Summary statistics are a fundamental aspect of data analysis, providing valuable insights into the distribution and characteristics of numerical variables. In this demonstration, we calculate summary statistics for the Width in the SASHELP.FISH dataset.

The output displays the summary statistics for the selected numerical variable in the SASHELP.FISH dataset. The computed statistics include the mean, standard deviation, minimum, maximum, and quartiles for each variable.

2. T-Test

The TTEST Procedure

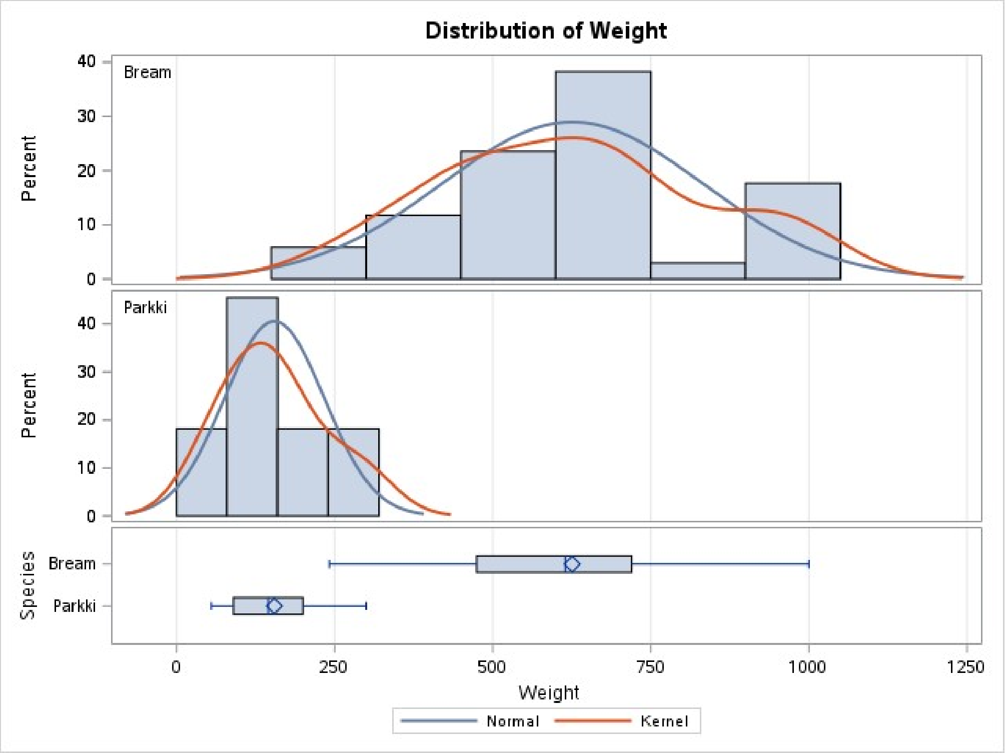
Variable: Weight

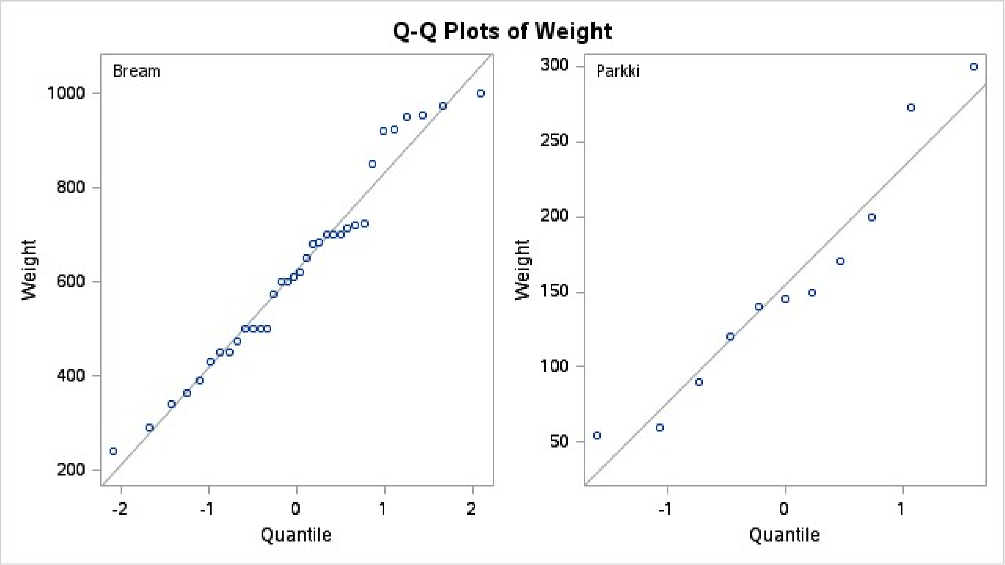
Species	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Bream		34	626.0	206.6	35.4324	242.0	1000.0
Parkki		11	154.8	78.7551	23.7456	55.0000	300.0
Diff (1-2)	**Pooled**		471.2	184.9	64.1490		
Diff (1-2)	**Satterthwaite**		471.2		42.6533		

Species	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Bream		626.0	553.9	698.1	206.6	166.6	271.9
Parkki		154.8	101.9	207.7	78.7551	55.0275	138.2
Diff (1-2)	**Pooled**	471.2	341.8	600.6	184.9	152.8	234.3
Diff (1-2)	**Satterthwaite**	471.2	385.1	557.3			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	43	7.35	<.0001
Satterthwaite	Unequal	41.605	11.05	<.0001

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	33	10	6.88	0.0027


Distribution of Weight
Bream
Parkki
Percent
0
10
20
30
40
Species
Bream
Parkki
0
250
500
750
1000
1250
Weight
Normal
Kernel


Q-Q Plots of Weight
Bream
Weight
200
400
600
800
1000
-2
-1
0
1
2
Quantile
Parkki
Weight
50
100
150
200
250
300
-1
0
1
Quantile

The t-test output provides the results of the two-sample t-test between the Weight variable for the Bream and Parkki species. The t-test results include the means and standard deviations of both groups, the t-value, degrees of freedom, and p-value.In conclusion, the t-test results indicate a statistically significant difference in the mean weights of Bream and Parkki fish species. The species Bream tends to have a significantly higher mean weight compared to Parkki.

The SAS Statistics Task is a valuable tool that simplifies statistical analysis by providing a user-friendly interface. In this report, we demonstrated how to use the Statistics Task to compute summary statistics and conduct a t-test using the SASHELP.FISH dataset.

The task's functionality allows users to explore and analyze data efficiently, gaining valuable insights into their datasets and making informed decisions based on statistical findings.

The SAS Statistics Task's capabilities extend far beyond the examples presented in this report, making it a versatile tool for various statistical analyses. Users can leverage its power to perform Distribution Analysis, Correlation Analysis, Table Analysis, create informative visualizations, and gain deeper insights into their data.

By harnessing the capabilities of the SAS Statistics Task, analysts and researchers can make data-driven decisions with confidence.