Group **E**

Cancer Patients Data Set

Class F

Brainnest | Group E

THE **TEAM**

Koketso Bess Mangwale
Analyst Programmer, HR Systems, Payroll
30 years old

Leonne Tuairau Maraiauria
Data Analyst student, Coursera Google
37 years old

Guilherme Joaquim
Brazilian Technology student
19 years old

Brainnest | Group E

THE **TEAM**

Muhsin Fıratoğlu
Data Analyst,Madrid
28 years old

Francesca Maltese
Economics and Statistics student
20 years old

Brainnest | Group E

Week **ONE**

Agenda

- What is our data talking about?
- Cleaning Data
- Descriptive Analysis
- Crosstab

The **Data Base**

The database shows a questionnaire made to patients who have cancer. The database computes age, gender, symptoms and external factors the patients were submitted on a scale of 1 to 9. For example, factors such as smoking, alcoholism and air pollution. And some symptoms like coughing up blood and chest pain.

Cleaning Data

In the data base, most of the variables are categoricals and ordinals (scale from 1 to 9). And we could find some missing data and some wrong values (values that exceeds 9 or values with decimals, such as 3.5).

Therefore, for the wrong values, we deleted the values that exceeds 9, so they could be missing values. For the values with decimals, we rounded the values. If it was 3.7, we rounded to 4.

	Name	Type
1	PatientId	String
2	Age	Numeric
3	Gender	String
4	AirPollution	Numeric
5	Alcoholuse	Numeric
6	DustAllergy	Numeric
7	OccuPationalHazards	Numeric
8	GeneticRisk	Numeric
9	chronicLungDisease	Numeric
10	BalancedDiet	Numeric
11	Obesity	Numeric
12	Smoking	Numeric
13	PassiveSmoker	Numeric
14	ChestPain	Numeric
15	CoughingofBlood	Numeric
16	Fatigue	Numeric
17	WeightLoss	Numeric
18	ShortnessofBreath	Numeric
19	Wheezing	Numeric
20	SwallowingDifficulty	Numeric
21	ClubbingofFingerNails	Numeric
22	FrequentCold	Numeric
23	DryCough	Numeric
24	Snoring	Numeric
25	Level	String

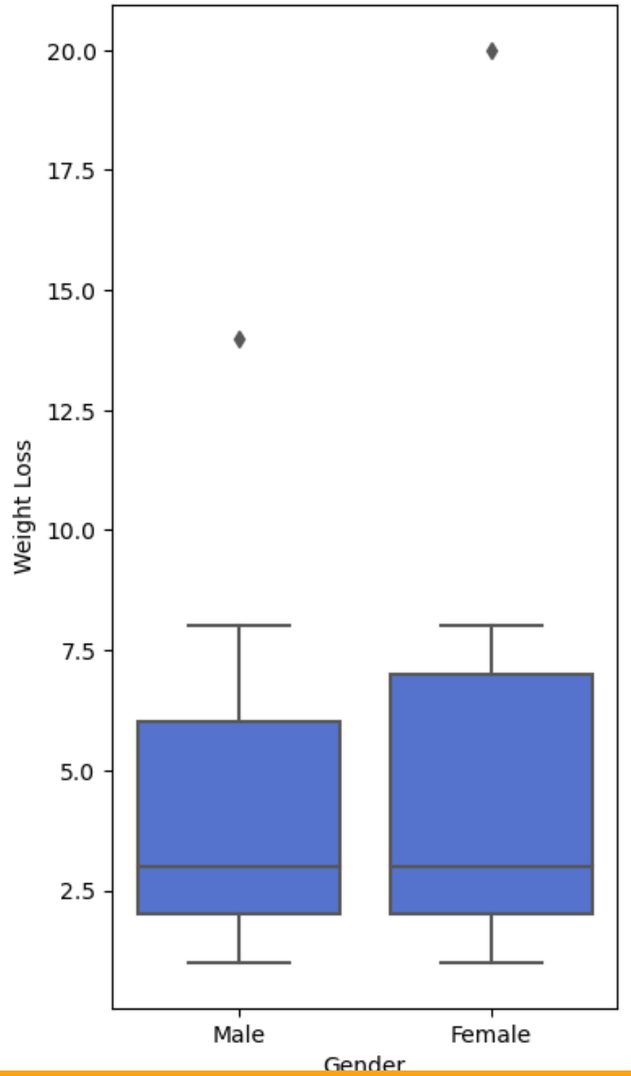
Cleaning Data

Here we can see what variables were having missing data and how much missing values in each column. We also can see five columns had no missing data.

```
: Coughing of Blood            13
  Alcohol use                  12
  Shortness of Breath          10
  Air Pollution                 9
  Obesity                       8
  Chest Pain                    8
  Smoking                       7
  Dust Allergy                  7
  Swallowing Difficulty         7
  Gender                        6
  OccuPational Hazards          6
  Genetic Risk                  6
  Wheezing                      6
  Fatigue                       5
  Weight Loss                   5
  Balanced Diet                 5
  Clubbing of Finger Nails      4
  Passive Smoker                4
  Age                           3
  Frequent Cold                 3
  Dry Cough                     0
  Snoring                       0
  Patient Id                    0
  chronic Lung Disease          0
  Level                         0
```

Identifying Outliers

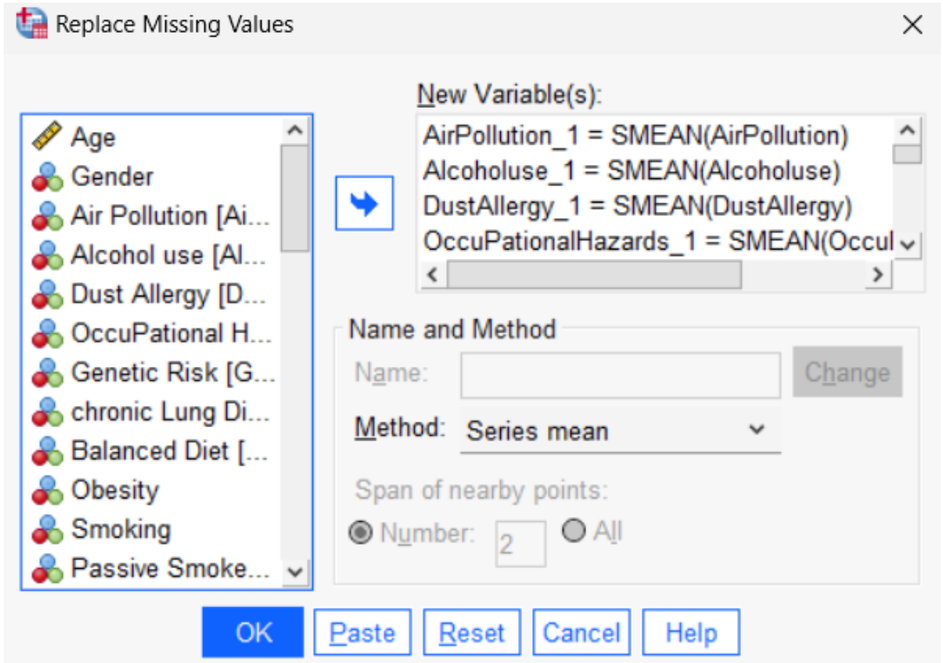
Here is an example of the values exceeding 9 (outliers) using the boxplot.



Cleaning Data

After that, we replaced the missing values using the mean substitution method.

Transform → Replace Missing Values → Select all variables except the ID → Method: Series Mean



Cleaning Data

After replacing the missing data with mean substitution, we compared the new values with the old ones using the Paired-sample-t-test to guarantee the new and old values were alike.

Pair 14	Weight Loss	3,86[a]	1000	2,200	,070
	SMEAN(WeightLoss)	3,862[a]	1000	2,2004	,0696
Pair 15	Shortness of Breath	4,24[a]	1000	2,277	,072
	SMEAN (ShortnessofBreath)	4,238[a]	1000	2,2774	,0720
Pair 16	Wheezing	3,79[a]	1000	2,033	,064
	SMEAN(Wheezing)	3,789[a]	1000	2,0329	,0643
Pair 17	Swallowing Difficulty	3,75[a]	1000	2,264	,072
	SMEAN (SwallowingDifficulty)	3,751[a]	1000	2,2637	,0716
Pair 18	Clubbing of Finger Nails	3,93[a]	1000	2,386	,075
	SMEAN (ClubbingofFingerNails)	3,929[a]	1000	2,3857	,0754
Pair 19	Frequent Cold	3,54[a]	1000	1,828	,058
	SMEAN(FrequentCold)	3,537[a]	1000	1,8281	,0578
Pair 20	Dry Cough	3,85[a]	1000	2,039	,064
	SMEAN(DryCough)	3,853[a]	1000	2,0390	,0645
Pair 21	Snoring	2,93[a]	1000	1,475	,047
	SMEAN(Snoring)	2,926[a]	1000	1,4747	,0466
Pair 22	Age	37,19[a]	1000	11,986	,379
	SMEAN(Age)	37,186[a]	1000	11,9859	,3790

a. The correlation and t cannot be computed because the standard error of the difference is 0.

Brainnest | Group E

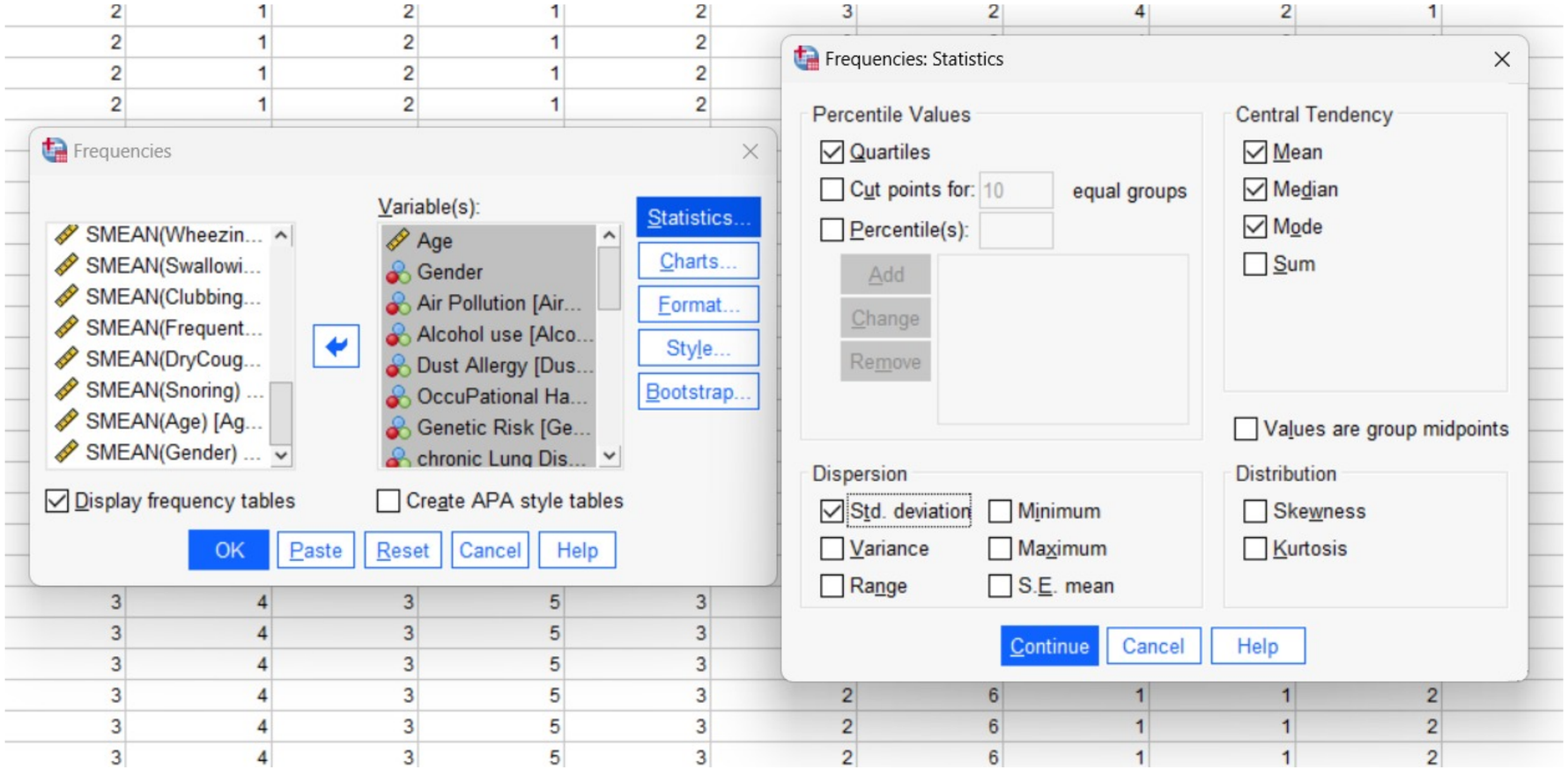
Descriptive **Analysis**

(Frequency, central tendecy, dispersion, position)

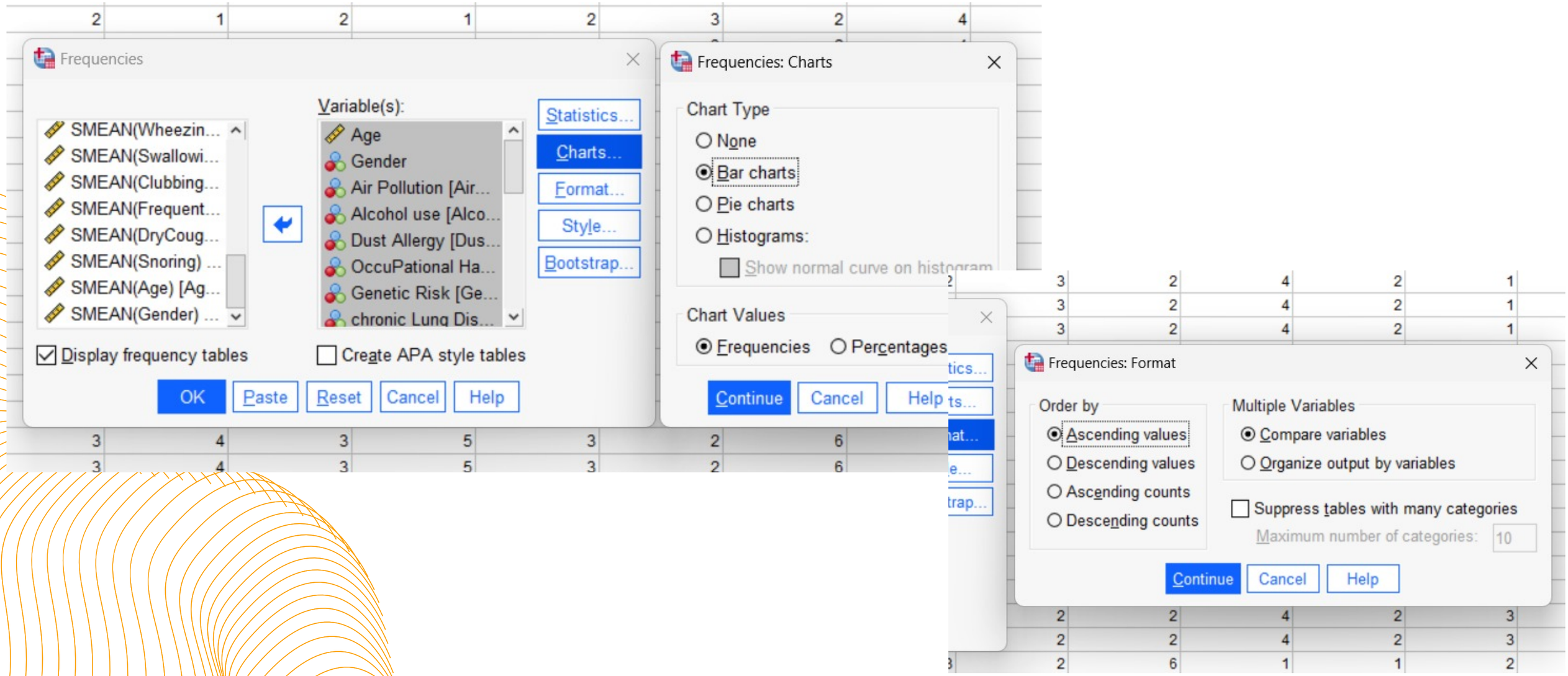
Brainnest | Group E

Descriptive **Analysis**

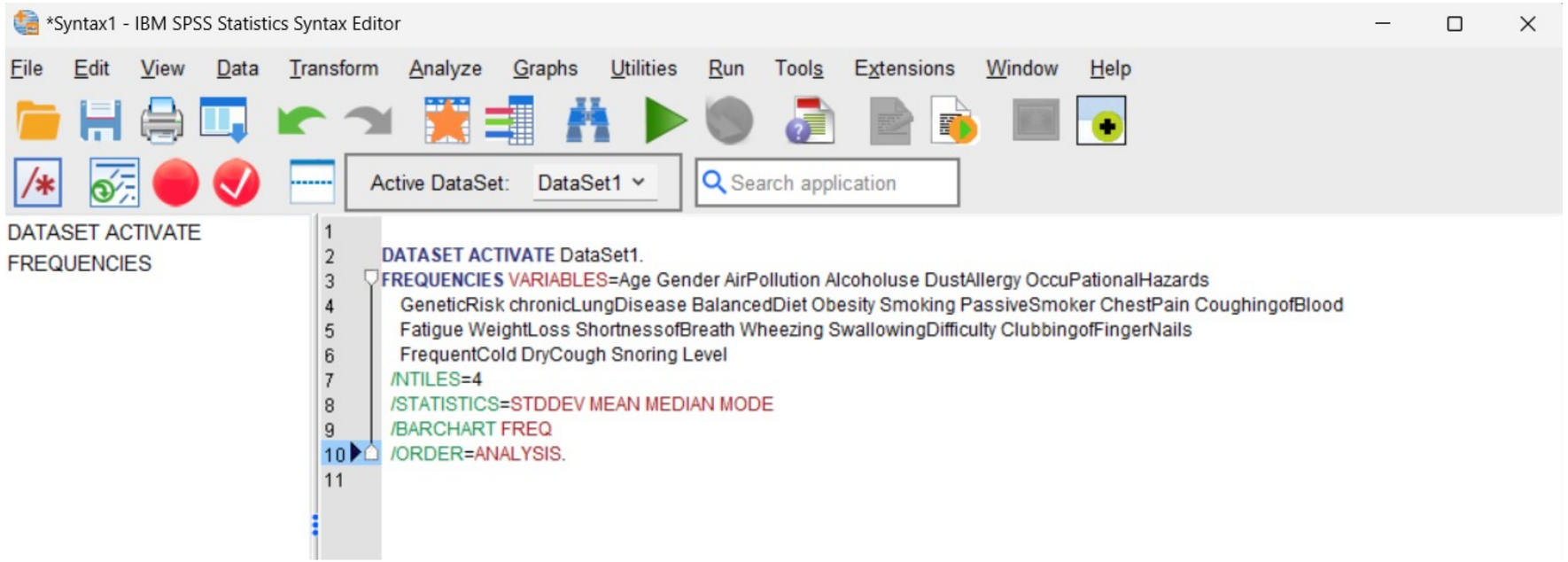
Analyze → Descriptive Statistics → Frequencies



Descriptive **Analysis**



Descriptive **Analysis**

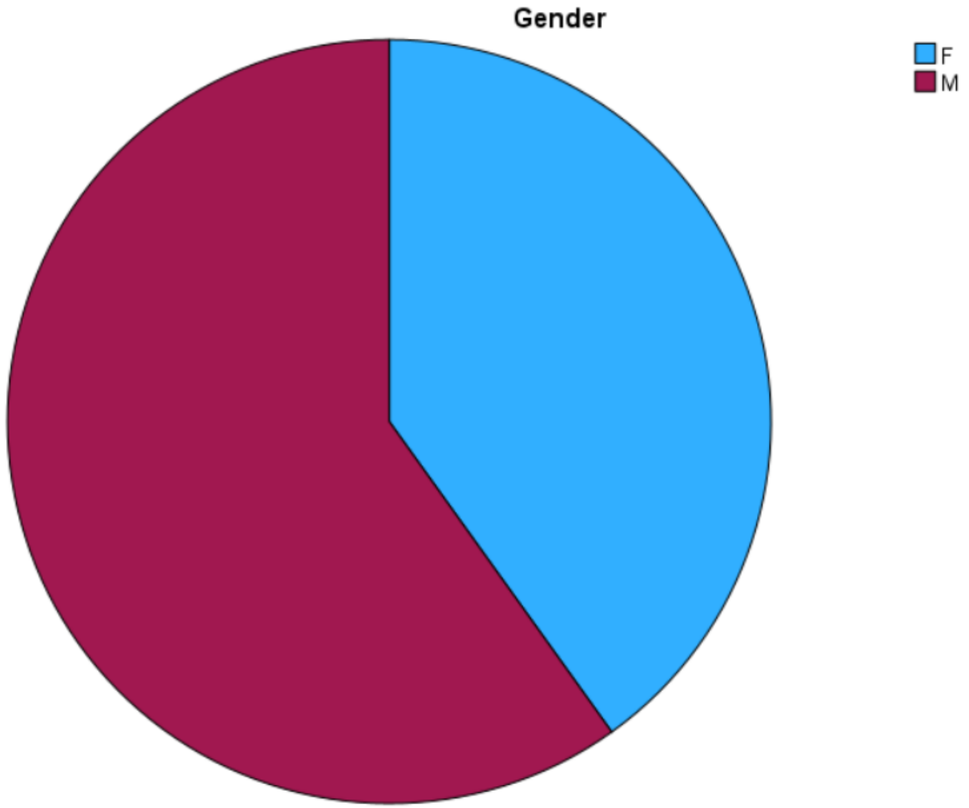


Syntax

Descriptive **Analysis**

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	F	401	40,1	40,1	40,1
	M	599	59,9	59,9	100,0
	Total	1000	100,0	100,0	



Descriptive **Analysis**

Statistics

		Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker
N	Valid	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
	Missing	0	0	0	0	0	0	0	0	0	0
Mean		3,84	4,58	5,16	4,84	4,59	4,38	4,49	4,48	3,95	4,19
Median		3,00	5,00	6,00	5,00	5,00	4,00	4,00	4,00	3,00	4,00
Mode		6	2	7	7	7	6	7	7	2	2
Std. Deviation		2,023	2,604	1,975	2,101	2,120	1,849	2,130	2,120	2,486	2,306
Percentiles	25	2,00	2,00	4,00	3,00	3,00	3,00	2,00	3,00	2,00	2,00
	50	3,00	5,00	6,00	5,00	5,00	4,00	4,00	4,00	3,00	4,00
	75	6,00	7,00	7,00	7,00	7,00	6,00	7,00	7,00	7,00	7,00

a. Multiple modes exishown

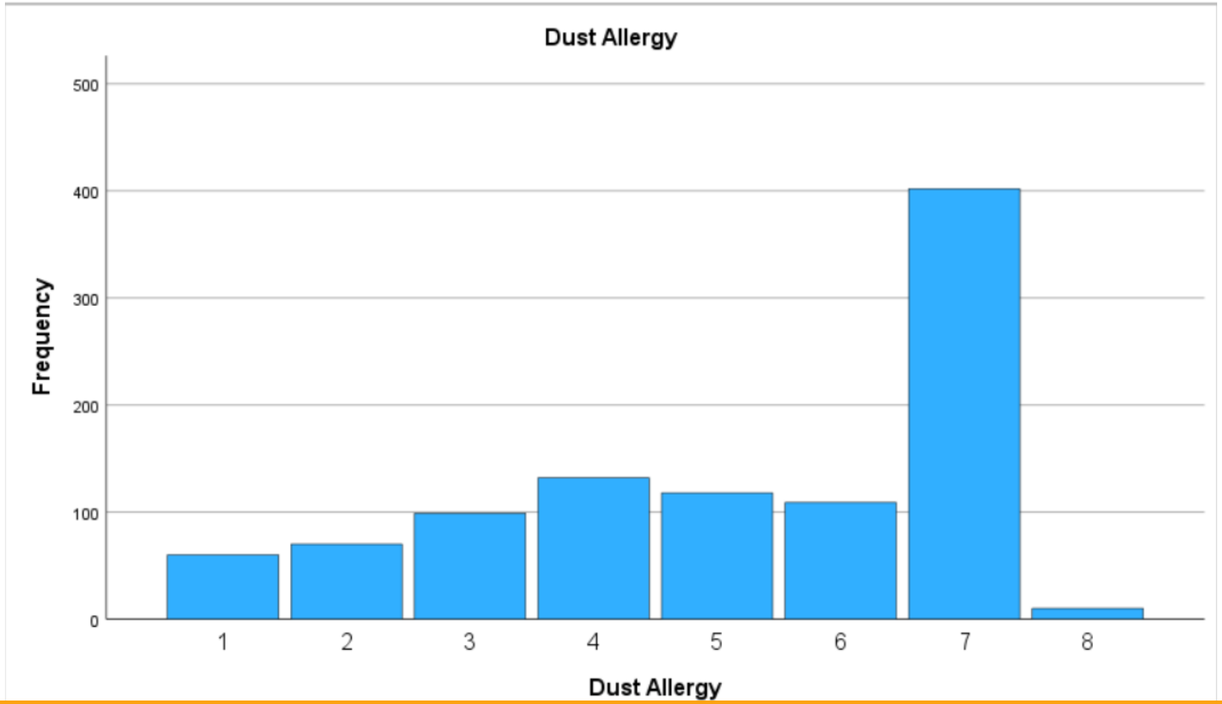
Statistics

		Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring
N	Valid	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
	Missing	0	0	0	0	0	0	0	0	0	0
Mean		4,87	3,86	3,86	4,24	3,79	3,75	3,93	3,54	3,85	2,93
Median		4,00	3,00	3,00	4,00	4,00	4,00	4,00	3,00	4,00	3,00
Mode		7	2[a]	2	2	2	1	2	3	2	2
Std. Deviation		2,411	2,243	2,200	2,277	2,033	2,264	2,386	1,828	2,039	1,475
Percentiles	25	3,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00
	50	4,00	3,00	3,00	4,00	4,00	4,00	4,00	3,00	4,00	3,00
	75	7,00	5,00	6,00	6,00	5,00	5,00	5,00	5,00	6,00	4,00

Descriptive **Analysis**

Dust Allergy

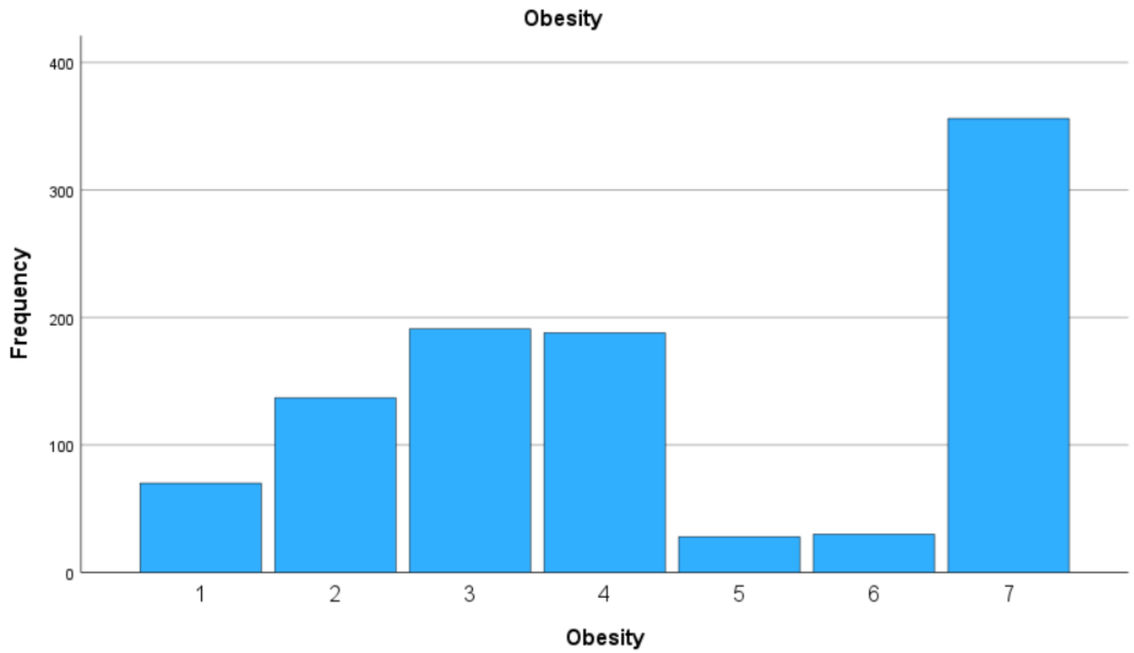
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	60	6,0	6,0	6,0
	2	70	7,0	7,0	13,0
	3	99	9,9	9,9	22,9
	4	132	13,2	13,2	36,1
	5	118	11,8	11,8	47,9
	6	109	10,9	10,9	58,8
	7	402	40,2	40,2	99,0
	8	10	1,0	1,0	100,0
	Total	1000	100,0	100,0	



Descriptive **Analysis**

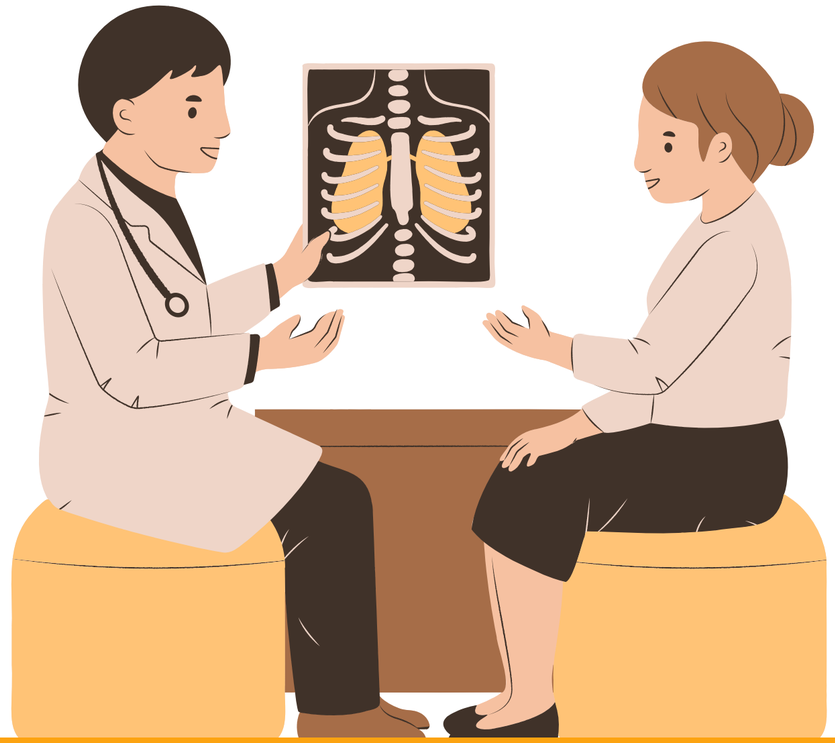
Obesity

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	70	7,0	7,0	7,0
	2	137	13,7	13,7	20,7
	3	191	19,1	19,1	39,8
	4	188	18,8	18,8	58,6
	5	28	2,8	2,8	61,4
	6	30	3,0	3,0	64,4
	7	356	35,6	35,6	100,0
	Total	1000	100,0	100,0	



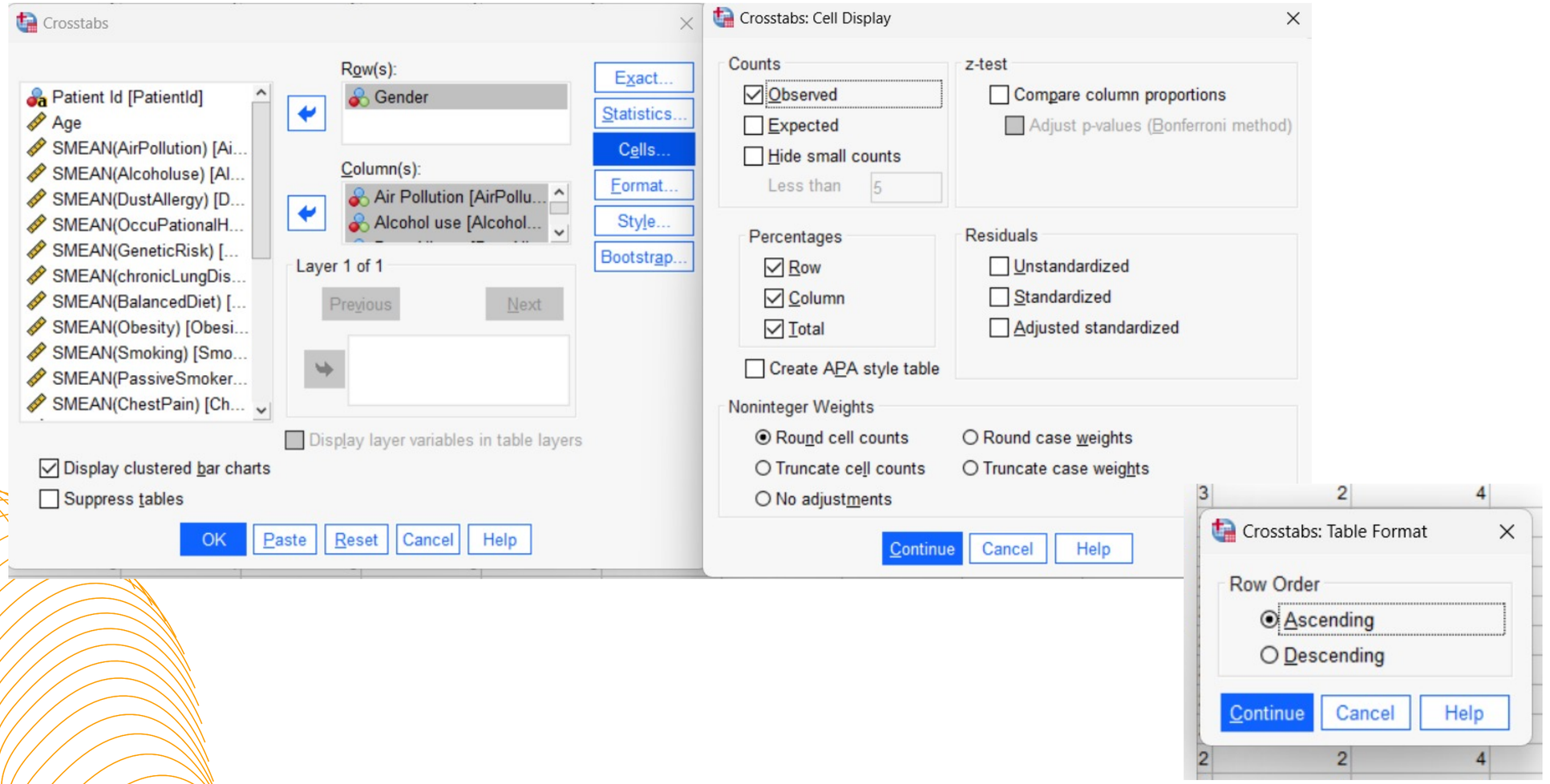
Brainnest | Group E

Crosstab



Crosstab

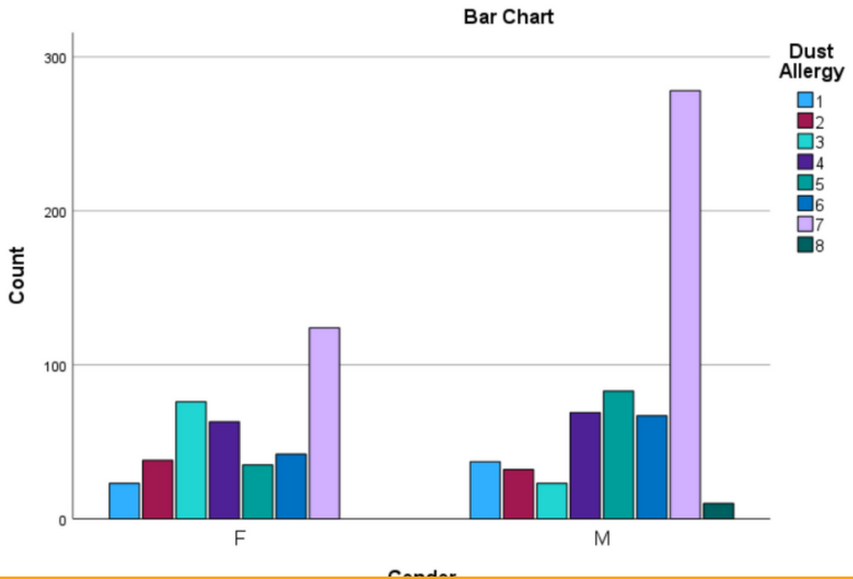
Analyze → Descriptive Statistics →Crosstabs



Crosstab

Gender * Dust Allergy Crosstabulation

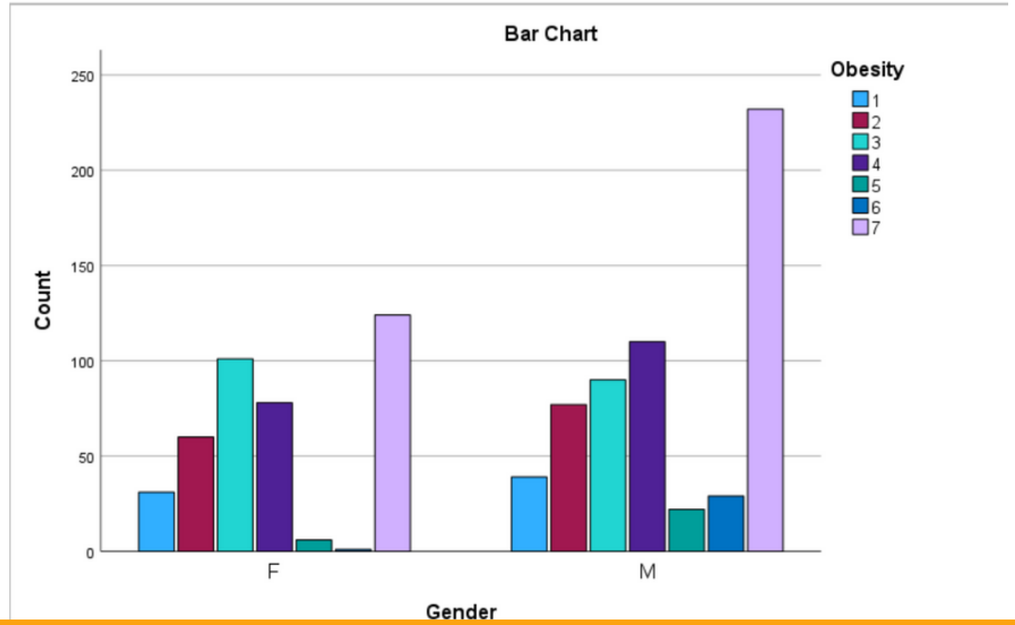
			Dust Allergy								
			1	2	3	4	5	6	7	8	Total
Gender	F	Count	23	38	76	63	35	42	124	0	401
		% within Gender	5,7%	9,5%	19,0%	15,7%	8,7%	10,5%	30,9%	0,0%	100,0%
		% within Dust Allergy	38,3%	54,3%	76,8%	47,7%	29,7%	38,5%	30,8%	0,0%	40,1%
		% of Total	2,3%	3,8%	7,6%	6,3%	3,5%	4,2%	12,4%	0,0%	40,1%
	M	Count	37	32	23	69	83	67	278	10	599
		% within Gender	6,2%	5,3%	3,8%	11,5%	13,9%	11,2%	46,4%	1,7%	100,0%
		% within Dust Allergy	61,7%	45,7%	23,2%	52,3%	70,3%	61,5%	69,2%	100,0%	59,9%
		% of Total	3,7%	3,2%	2,3%	6,9%	8,3%	6,7%	27,8%	1,0%	59,9%
Total		Count	60	70	99	132	118	109	402	10	1000
		% within Gender	6,0%	7,0%	9,9%	13,2%	11,8%	10,9%	40,2%	1,0%	100,0%
		% within Dust Allergy	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% of Total	6,0%	7,0%	9,9%	13,2%	11,8%	10,9%	40,2%	1,0%	100,0%



Crosstab

Gender * Obesity Crosstabulation

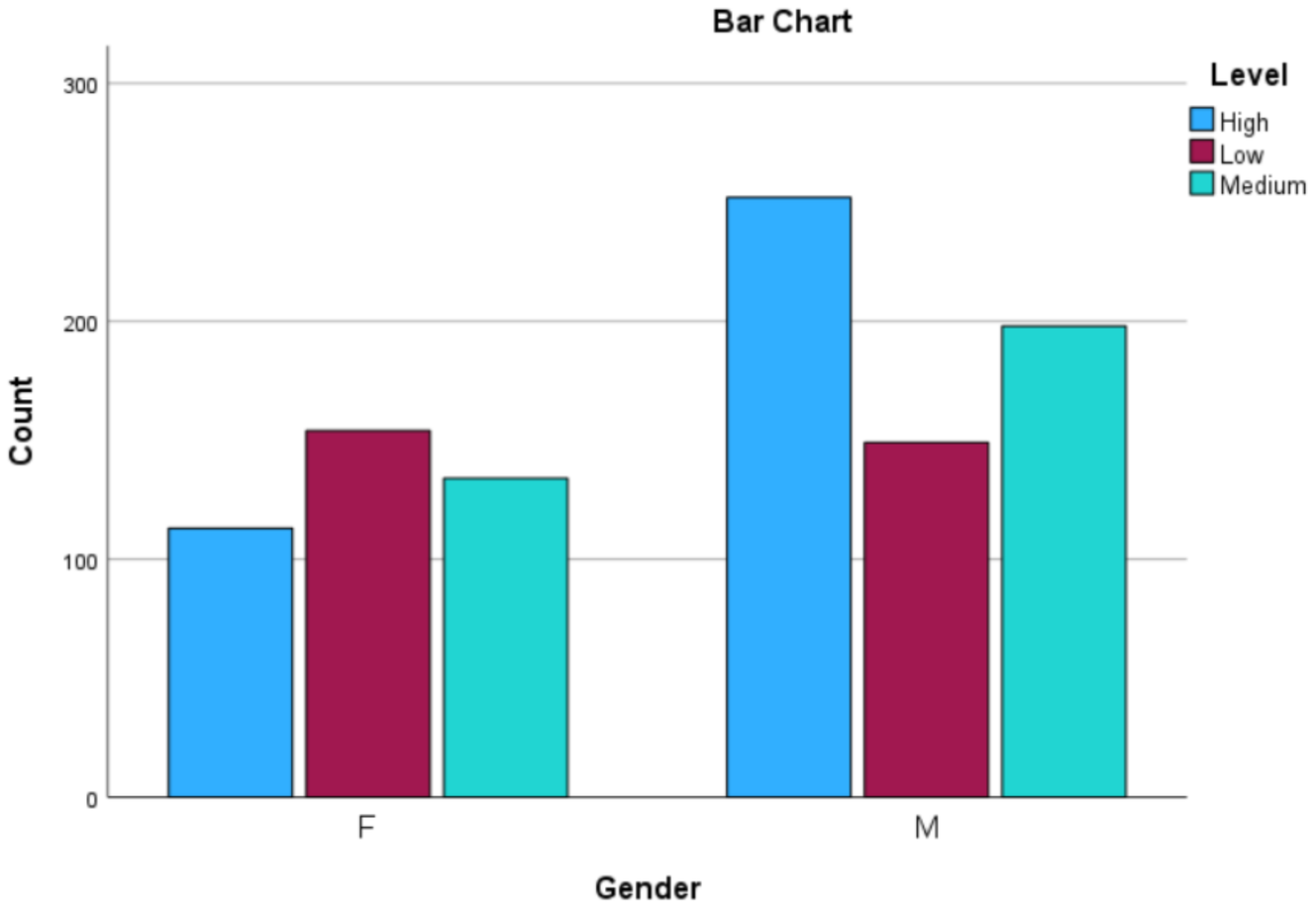
			Obesity							
			1	2	3	4	5	6	7	Total
Gender	F	Count	31	60	101	78	6	1	124	401
		% within Gender	7,7%	15,0%	25,2%	19,5%	1,5%	0,2%	30,9%	100,0%
		% within Obesity	44,3%	43,8%	52,9%	41,5%	21,4%	3,3%	34,8%	40,1%
		% of Total	3,1%	6,0%	10,1%	7,8%	0,6%	0,1%	12,4%	40,1%
	M	Count	39	77	90	110	22	29	232	599
		% within Gender	6,5%	12,9%	15,0%	18,4%	3,7%	4,8%	38,7%	100,0%
		% within Obesity	55,7%	56,2%	47,1%	58,5%	78,6%	96,7%	65,2%	59,9%
		% of Total	3,9%	7,7%	9,0%	11,0%	2,2%	2,9%	23,2%	59,9%
Total		Count	70	137	191	188	28	30	356	1000
		% within Gender	7,0%	13,7%	19,1%	18,8%	2,8%	3,0%	35,6%	100,0%
		% within Obesity	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% of Total	7,0%	13,7%	19,1%	18,8%	2,8%	3,0%	35,6%	100,0%



Crosstab

Gender * Level Crosstabulation

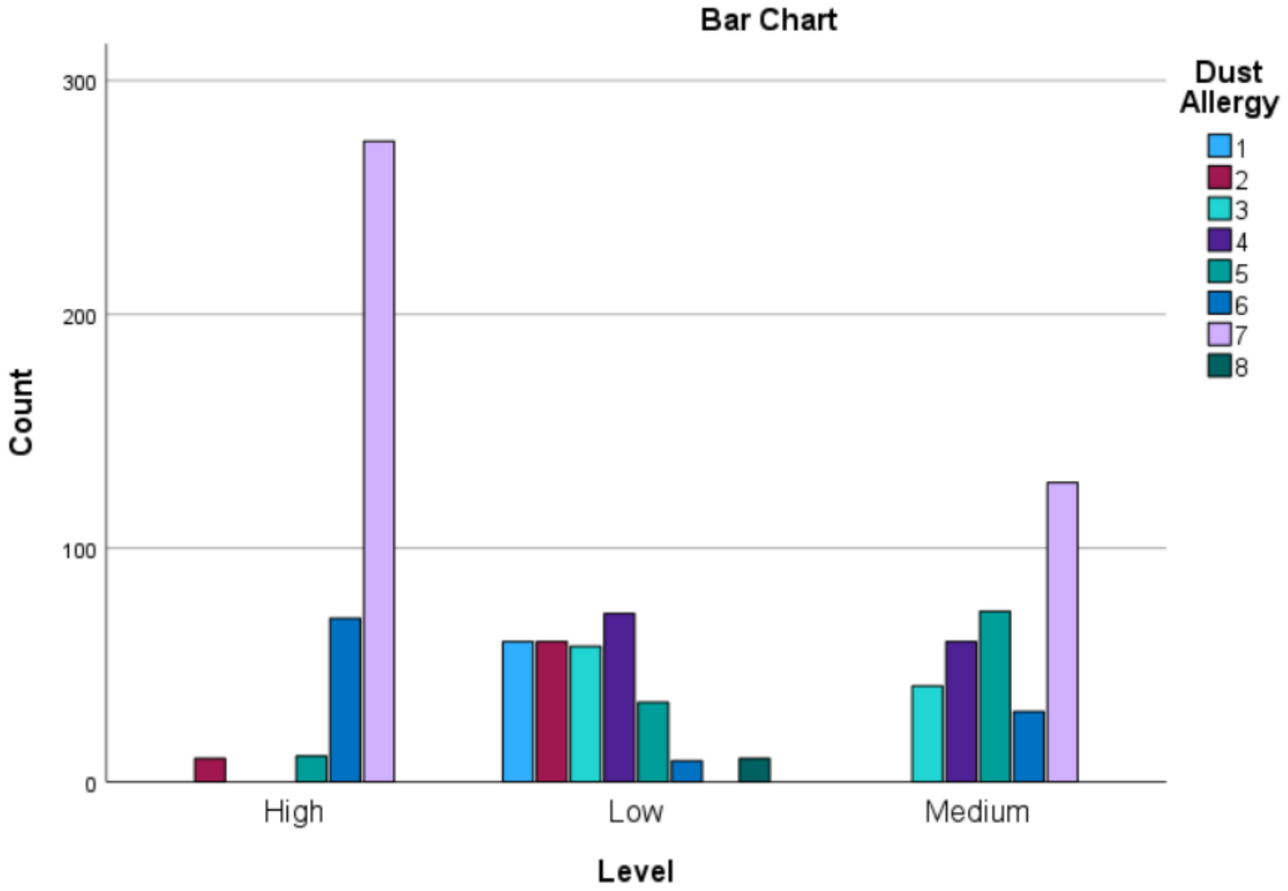
			Level			
			High	Low	Medium	Total
Gender	F	Count	113	154	134	401
		% within Gender	28,2%	38,4%	33,4%	100,0%
		% within Level	31,0%	50,8%	40,4%	40,1%
		% of Total	11,3%	15,4%	13,4%	40,1%
	M	Count	252	149	198	599
		% within Gender	42,1%	24,9%	33,1%	100,0%
		% within Level	69,0%	49,2%	59,6%	59,9%
		% of Total	25,2%	14,9%	19,8%	59,9%
Total		Count	365	303	332	1000
		% within Gender	36,5%	30,3%	33,2%	100,0%
		% within Level	100,0%	100,0%	100,0%	100,0%
		% of Total	36,5%	30,3%	33,2%	100,0%



Crosstab

Level * Dust Allergy Crosstabulation

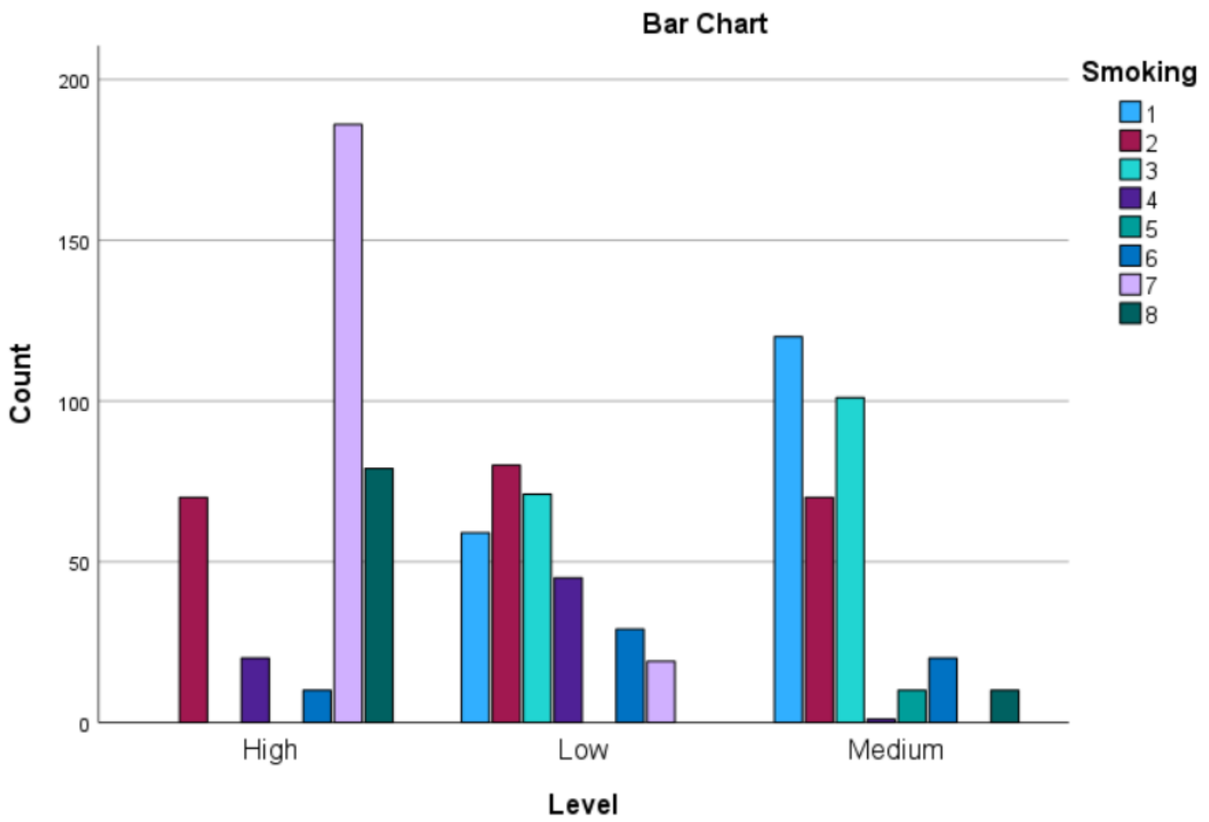
			Dust Allergy								
			1	2	3	4	5	6	7	8	Total
Level	High	Count	0	10	0	0	11	70	274	0	365
		% within Level	0,0%	2,7%	0,0%	0,0%	3,0%	19,2%	75,1%	0,0%	100,0%
		% within Dust Allergy	0,0%	14,3%	0,0%	0,0%	9,3%	64,2%	68,2%	0,0%	36,5%
		% of Total	0,0%	1,0%	0,0%	0,0%	1,1%	7,0%	27,4%	0,0%	36,5%
	Low	Count	60	60	58	72	34	9	0	10	303
		% within Level	19,8%	19,8%	19,1%	23,8%	11,2%	3,0%	0,0%	3,3%	100,0%
		% within Dust Allergy	100,0%	85,7%	58,6%	54,5%	28,8%	8,3%	0,0%	100,0%	30,3%
		% of Total	6,0%	6,0%	5,8%	7,2%	3,4%	0,9%	0,0%	1,0%	30,3%
	Medium	Count	0	0	41	60	73	30	128	0	332
		% within Level	0,0%	0,0%	12,3%	18,1%	22,0%	9,0%	38,6%	0,0%	100,0%
		% within Dust Allergy	0,0%	0,0%	41,4%	45,5%	61,9%	27,5%	31,8%	0,0%	33,2%
		% of Total	0,0%	0,0%	4,1%	6,0%	7,3%	3,0%	12,8%	0,0%	33,2%
Total		Count	60	70	99	132	118	109	402	10	1000
		% within Level	6,0%	7,0%	9,9%	13,2%	11,8%	10,9%	40,2%	1,0%	100,0%
		% within Dust Allergy	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% of Total	6,0%	7,0%	9,9%	13,2%	11,8%	10,9%	40,2%	1,0%	100,0%



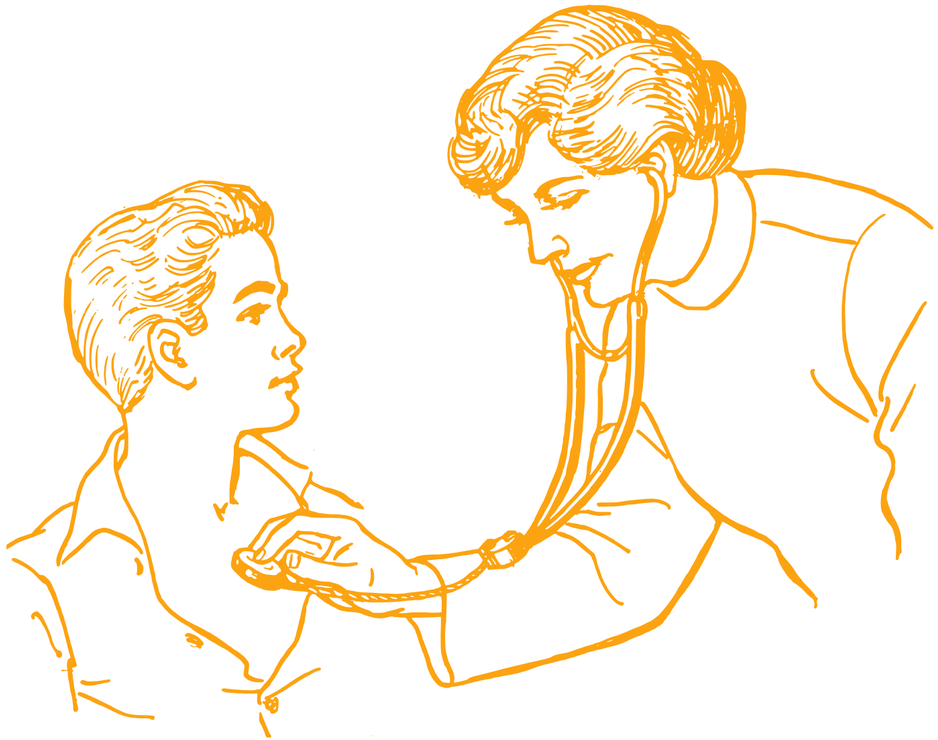
Crosstab

Level * Smoking Crosstabulation

			Smoking								
			1	2	3	4	5	6	7	8	Total
Level	High	Count	0	70	0	20	0	10	186	79	365
		% within Level	0,0%	19,2%	0,0%	5,5%	0,0%	2,7%	51,0%	21,6%	100,0%
		% within Smoking	0,0%	31,8%	0,0%	30,3%	0,0%	16,9%	90,7%	88,8%	36,5%
		% of Total	0,0%	7,0%	0,0%	2,0%	0,0%	1,0%	18,6%	7,9%	36,5%
	Low	Count	59	80	71	45	0	29	19	0	303
		% within Level	19,5%	26,4%	23,4%	14,9%	0,0%	9,6%	6,3%	0,0%	100,0%
		% within Smoking	33,0%	36,4%	41,3%	68,2%	0,0%	49,2%	9,3%	0,0%	30,3%
		% of Total	5,9%	8,0%	7,1%	4,5%	0,0%	2,9%	1,9%	0,0%	30,3%
	Medium	Count	120	70	101	1	10	20	0	10	332
		% within Level	36,1%	21,1%	30,4%	0,3%	3,0%	6,0%	0,0%	3,0%	100,0%
		% within Smoking	67,0%	31,8%	58,7%	1,5%	100,0%	33,9%	0,0%	11,2%	33,2%
		% of Total	12,0%	7,0%	10,1%	0,1%	1,0%	2,0%	0,0%	1,0%	33,2%
Total		Count	179	220	172	66	10	59	205	89	1000
		% within Level	17,9%	22,0%	17,2%	6,6%	1,0%	5,9%	20,5%	8,9%	100,0%
		% within Smoking	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% of Total	17,9%	22,0%	17,2%	6,6%	1,0%	5,9%	20,5%	8,9%	100,0%



Thank you**!**



Brainnest | Group E

Week TWO

Agenda

- Check the Normality
- Normality
  - *Cancer patients' Age*
  - *Cancer level and Gender*
- Research Question
- Insights and Visualizations
- Conclusion

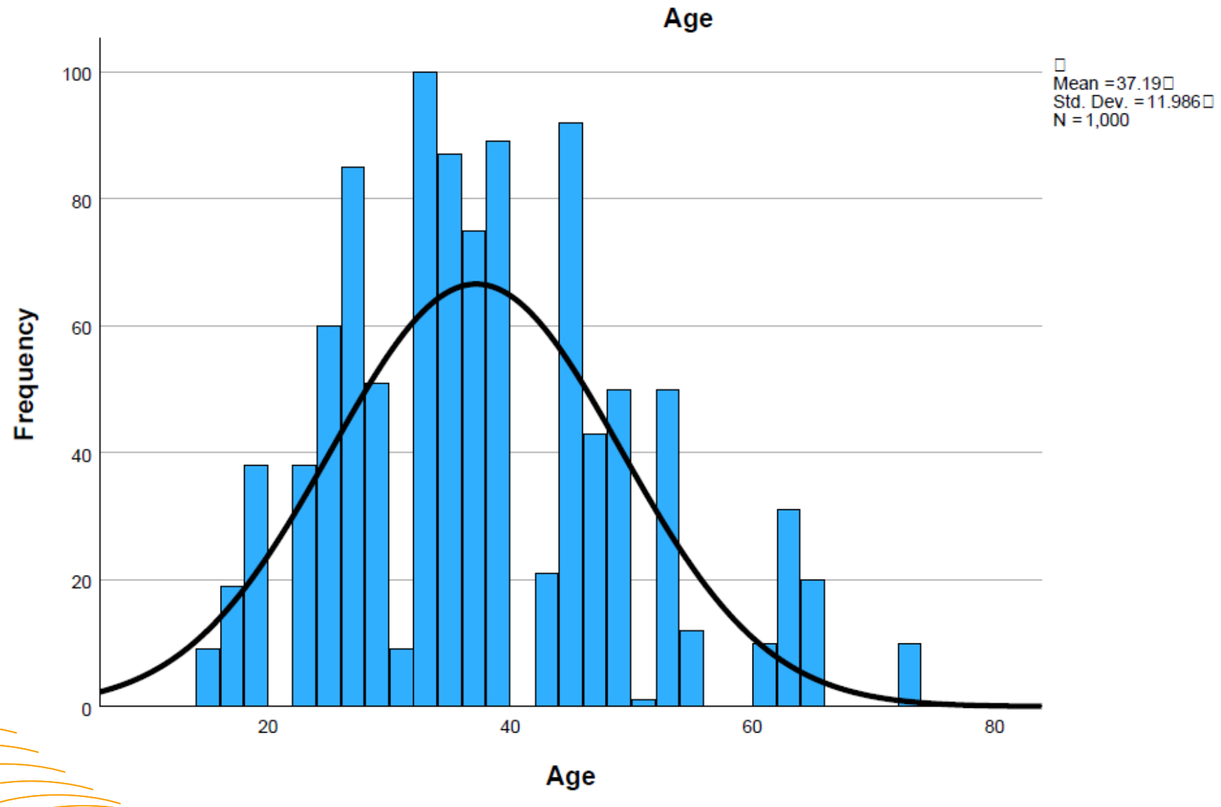
Check for **Normality**

To check for normality, we use summary statistics to assess the cancer patients clinical data.

The histogram and box plot visualizations will help with descriptive statistics. Q-Q Plots and PP-Plots will show the correlations and any deviations from normality.

Brainnest | Group E

Distribution of **Age**



Distribution of Age for Cancer Patients:

- The data appears to be skewed to the right.
- The average age (=37.19) of cancer patients is greater than the median age (=36).
- There is a high frequency of patients of ages between 34 - 36 years.

Kurtosis and **Skewness of Age**

Descriptives

			Statistic	Std. Error
Age	Mean		37.19	.379
	95% Confidence Interval for Mean	Lower Bound	36.44	
		Upper Bound	37.93	
	5% Trimmed Mean		36.75	
	Median		36.00	
	Variance		143.662	
	Std. Deviation		11.986	
	Minimum		14	
	Maximum		73	
	Range		59	
	Interquartile Range		17	
	Skewness		.555	.077
	Kurtosis		.069	.155

Kurtosis **and Skewness of Age**

Descriptives

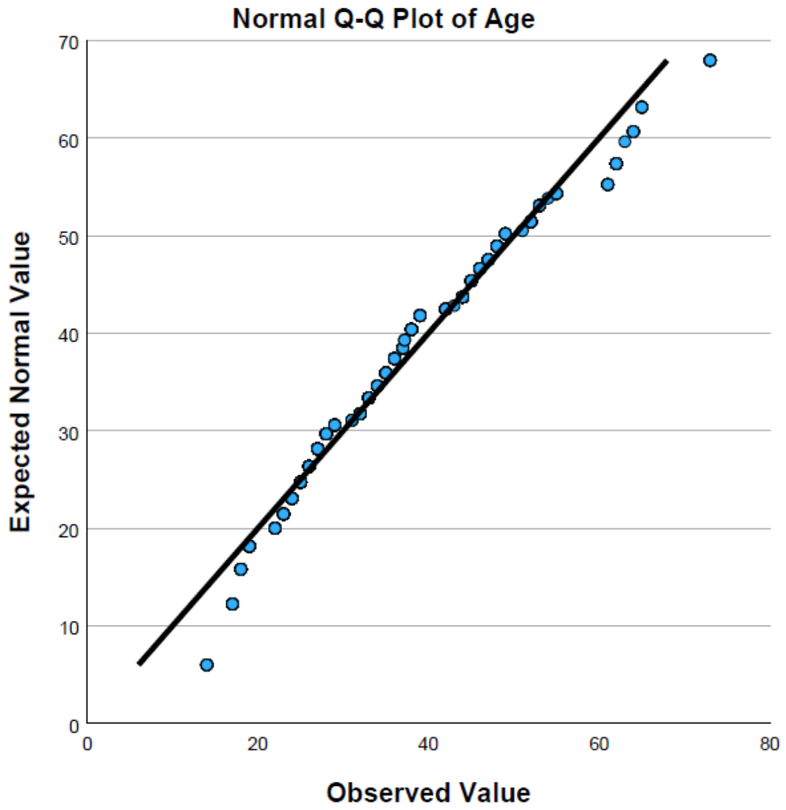
			Statistic	Std. Error
Age	Mean		37.19	.379
	95% Confidence Interval for Mean	Lower Bound	36.44	
		Upper Bound	37.93	
	5% Trimmed Mean		36.75	
	Median		36.00	
	Variance		143.662	
	Std. Deviation		11.986	
	Minimum		14	
	Maximum		73	
	Range		59	
	Interquartile Range		17	
	Skewness		.555	.077
	Kurtosis		.069	.155

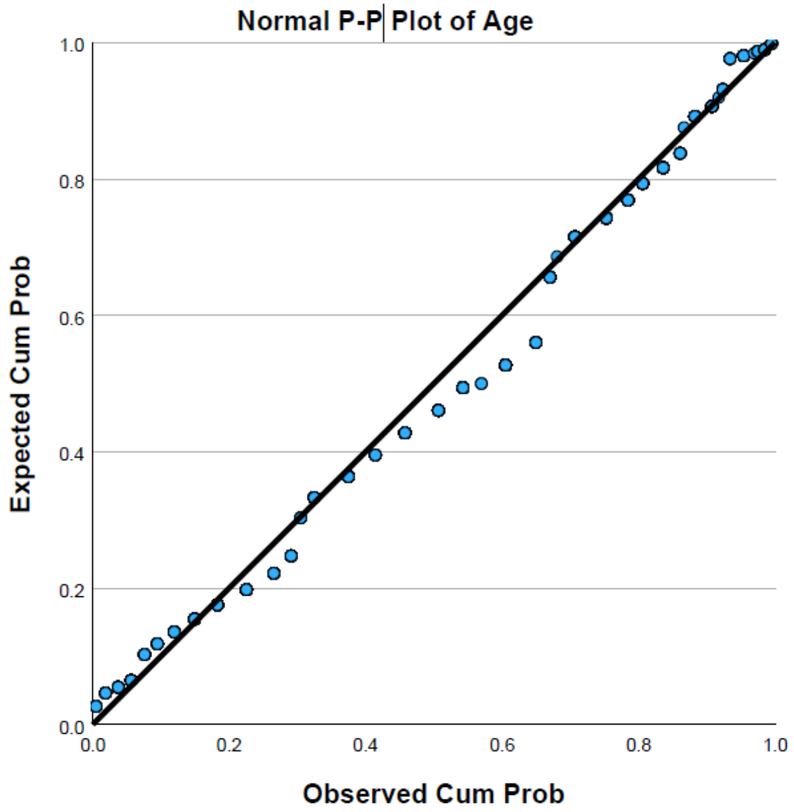
Calculating the z-score

Skewness = 0.555/0.077; **z-score = 7.20**
kurtosis = 0.069/0.155; **z-score = 0.44**

kurtosis is close to 0 but the **skewness of data does deviate away from zero.**
The results of the z-score for kurtosis is close to 1.96, but it is not within +-1.96 for skewness, so it is very skewed.

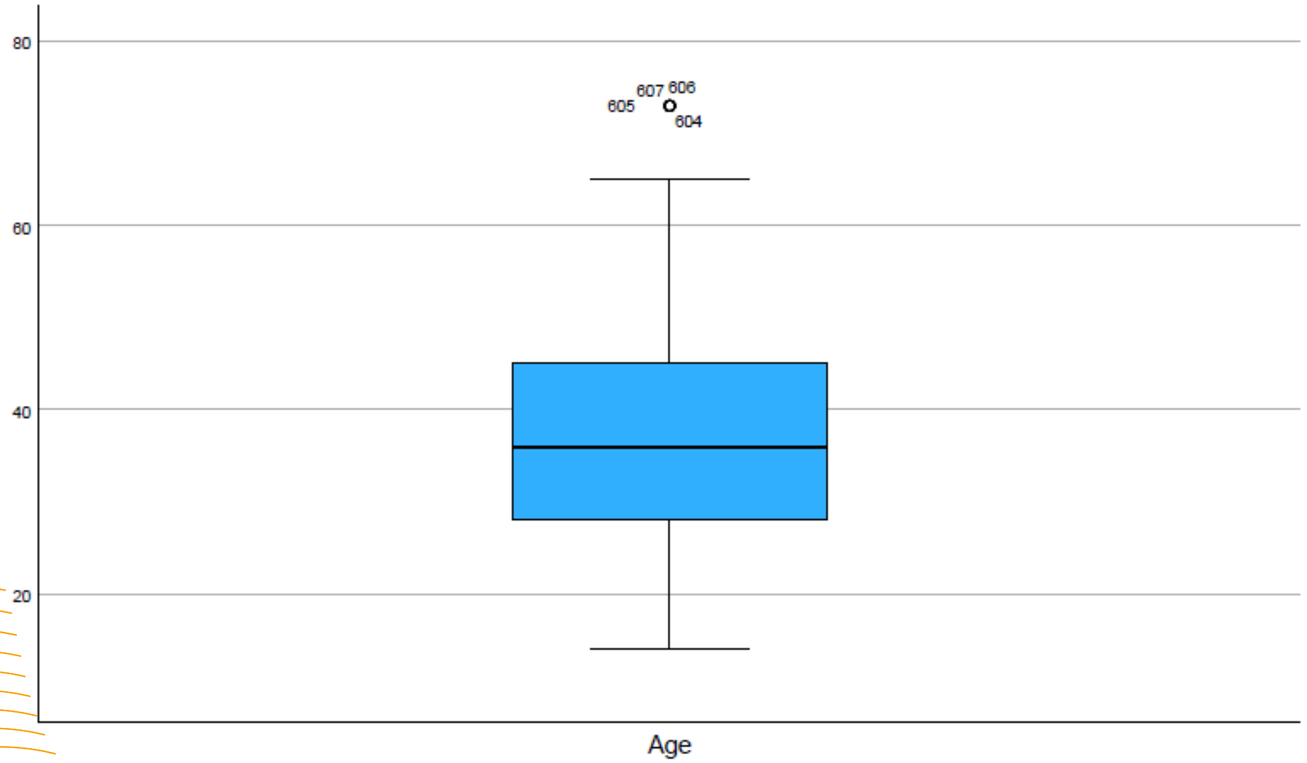
Q-Q and **P-P Plots**





Graph shows there is deviation from the straight line, which indicates deviation from normality.

Bloxplot - **Age**



<u>Indication of the range of Age</u>

- mean = 37.19; mode = 35; median = 36.
- It is observed that the average age of cancer patients is slightly greater than the median age. The mean and median are roughly the same at 36 years.
- The number of patients in the upper quartile are more than those in the lower quartile.

Statistical Tests **on Normality**

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Age	.112	1000	<.001	.969	1000	<.001
SMEAN(AirPollution)	.228	1000	<.001	.881	1000	<.001
SMEAN(Alcoholuse)	.186	1000	<.001	.867	1000	<.001
SMEAN(DustAllergy)	.236	1000	<.001	.852	1000	<.001
SMEAN (OccuPationalHazards)	.240	1000	<.001	.877	1000	<.001
SMEAN(GeneticRisk)	.198	1000	<.001	.843	1000	<.001
SMEAN (chronicLungDisease)	.225	1000	<.001	.897	1000	<.001
SMEAN(BalancedDiet)	.212	1000	<.001	.840	1000	<.001
SMEAN(Obesity)	.239	1000	<.001	.847	1000	<.001
SMEAN(Smoking)	.220	1000	<.001	.851	1000	<.001
SMEAN(PassiveSmoker)	.181	1000	<.001	.866	1000	<.001
SMEAN(ChestPain)	.213	1000	<.001	.889	1000	<.001
SMEAN(CoughingofBlood)	.175	1000	<.001	.924	1000	<.001
SMEAN(Fatigue)	.189	1000	<.001	.885	1000	<.001
SMEAN(WeightLoss)	.198	1000	<.001	.864	1000	<.001
SMEAN(ShortnessofBreath)	.165	1000	<.001	.917	1000	<.001
SMEAN(Wheezing)	.196	1000	<.001	.911	1000	<.001
SMEAN (SwallowingDifficulty)	.158	1000	<.001	.903	1000	<.001
SMEAN (ClubbingofFingerNails)	.179	1000	<.001	.884	1000	<.001
SMEAN(FrequentCold)	.175	1000	<.001	.908	1000	<.001
SMEAN(DryCough)	.188	1000	<.001	.898	1000	<.001
SMEAN(Snoring)	.205	1000	<.001	.910	1000	<.001

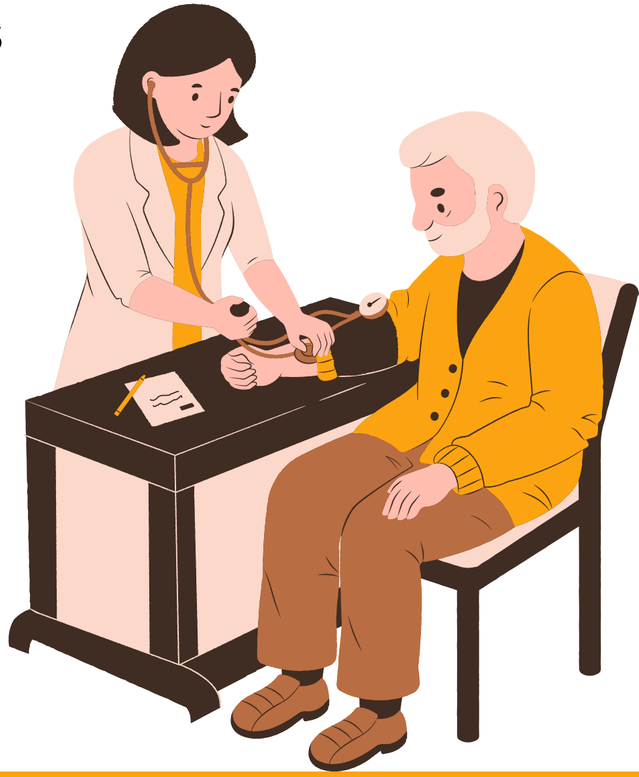
a. Lilliefors Significance Correction

if significance value of Sig. > 0.05 then the data is normal, otherwise the data is not. Here, we can see there's no Sig > 0.05, only Sig < 0.01

Research **Questions**

Since we don't have any variables following the normal distribution, we cannot ask questions about probability. However, we still can ask questions about proportions.

Research Question : Is the proportion of cancer levels the same across females and males?

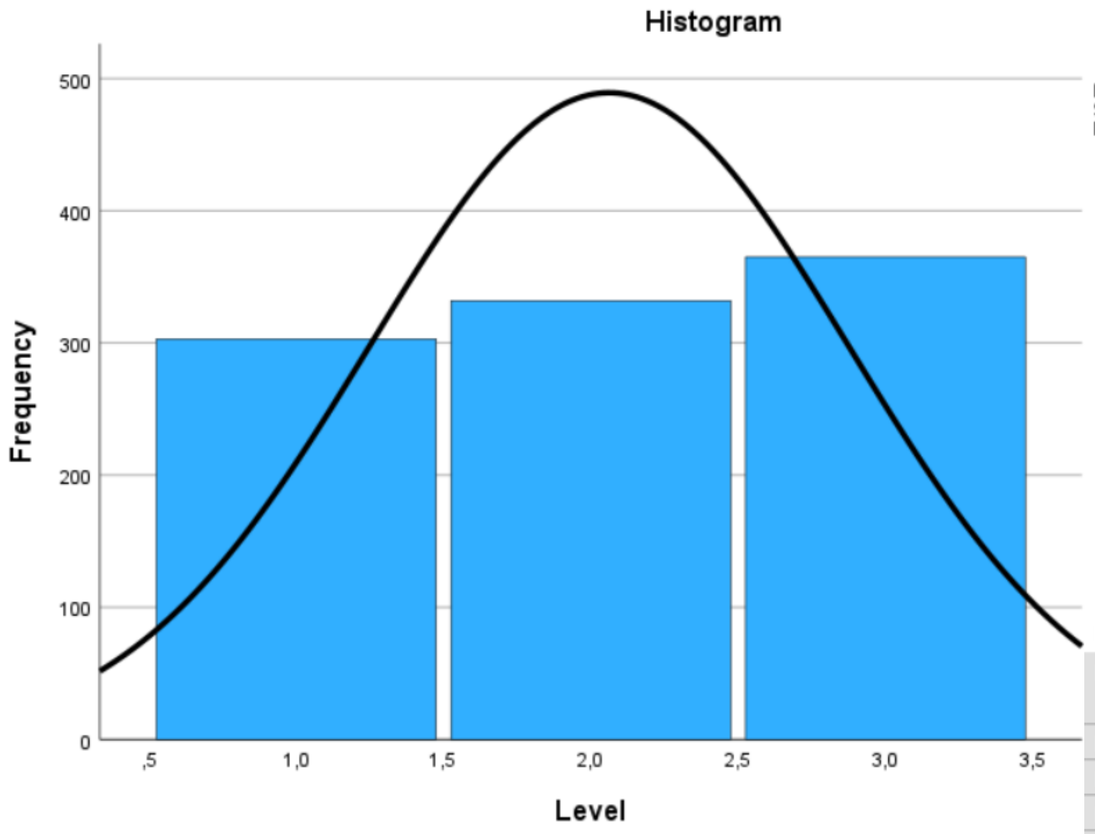


Brainnest | Group E

Checking **Normality**

Brainnest | Group E

Distribution of **Cancer Level**



Statistics

Level

N	Valid	1000
	Missing	0
Mean		2,06
Median		2,00
Mode		3
Std. Deviation		,815
Skewness		-,114
Std. Error of Skewness		,077
Kurtosis		-1,487
Std. Error of Kurtosis		,155

The histogram show us cancer level data does not follow a normal distribution. As we can see, the data is skewed to the right.

.

Kurtosis and **Skewness - Gender by Level**

Descriptives

	Gender			Statistic	Std. Error
Level	Male	Mean		2,17	,033
		95% Confidence Interval for Mean	Lower Bound	2,11	
			Upper Bound	2,24	
		5% Trimmed Mean		2,19	
		Median		2,00	
		Variance		,641	
		Std. Deviation		,801	
		Minimum		1	
		Maximum		3	
		Range		2	
		Interquartile Range		1	
		Skewness		-,320	,100
		Kurtosis		-1,372	,199
	Female	Mean		1,90	,040
		95% Confidence Interval for Mean	Lower Bound	1,82	
			Upper Bound	1,98	
		5% Trimmed Mean		1,89	
		Median		2,00	
		Variance		,657	
		Std. Deviation		,811	
		Minimum		1	
		Maximum		3	
		Range		2	
		Interquartile Range		2	
		Skewness		,189	,122
		Kurtosis		-1,454	,243

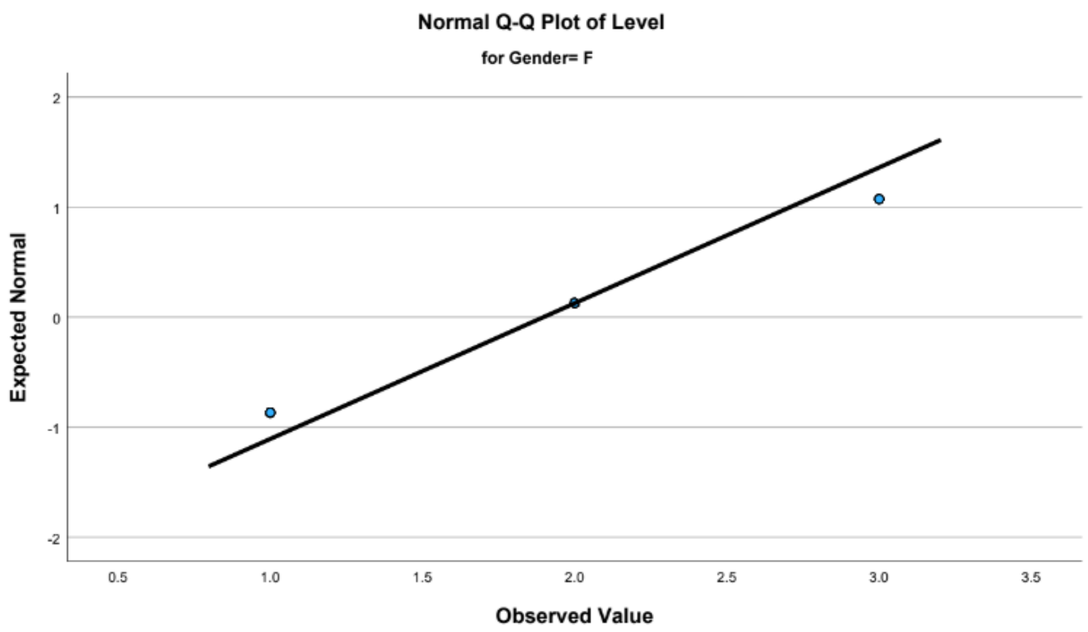
Male
Skewness: -0.320/0.1; **z-score = -3.2**
Kurtosis: -1.372/0.199; **z-score = -6.89**

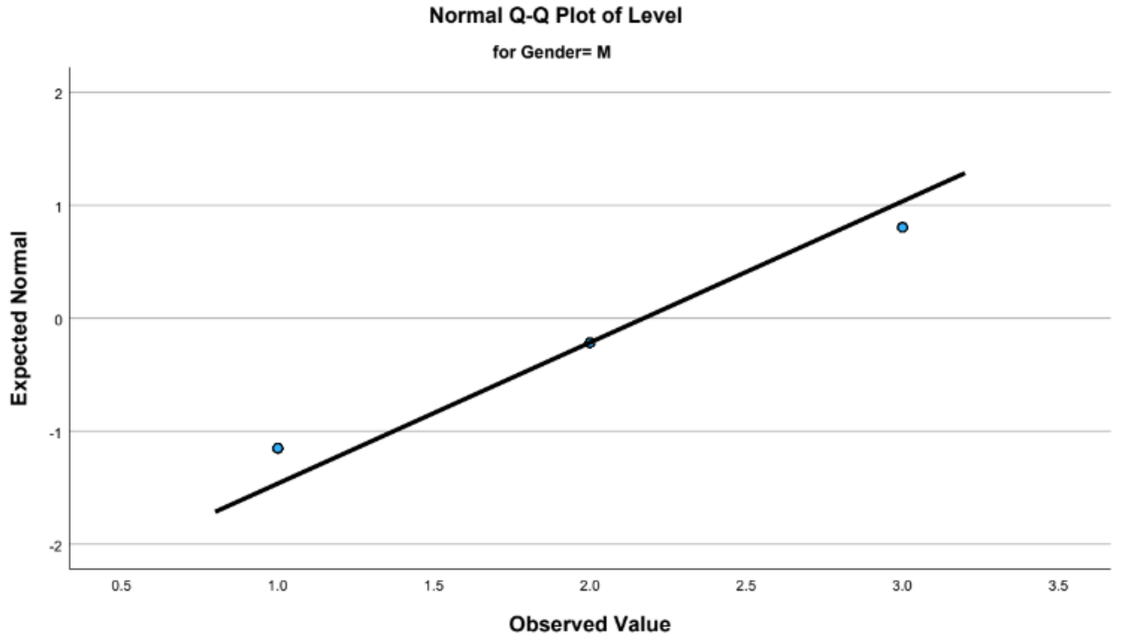
Female
Skewness: 0.189/0.122; **z-score = 1.54**
Kurtosis: -1.454/0.243; **z-score = -5.98**

As we can see, the Skewness and Kurtosis are not between +/-1.96. Then, they are not normally distributed.

Q-Q and **P-P Plots - Gender and Level**

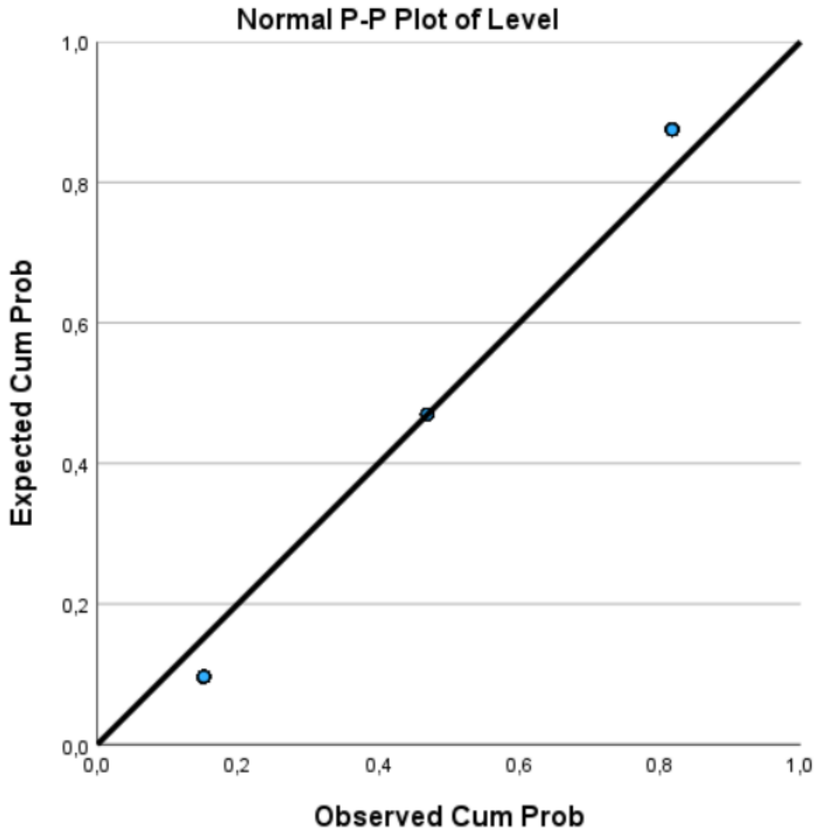
Normal Q-Q Plots

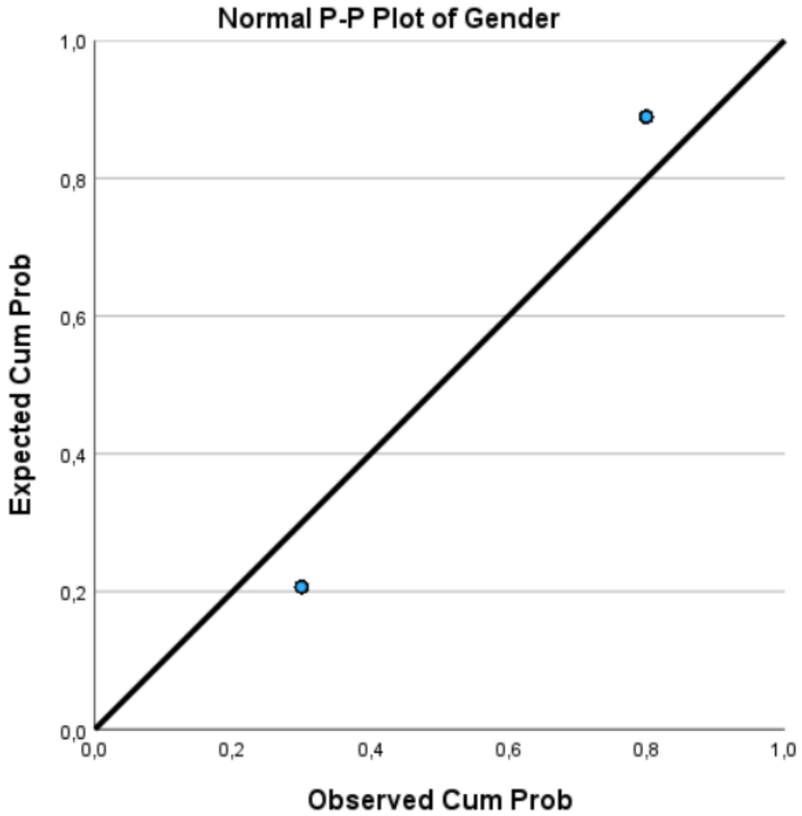




Brainnest | Group E

Q-Q and **P-P Plots - Gender and Level**





As we can see from the Q-Q and P-P Plots, level and gender are deviating from the straight line. Therefore, they don't follow the normal distribution

Brainnest | Group E

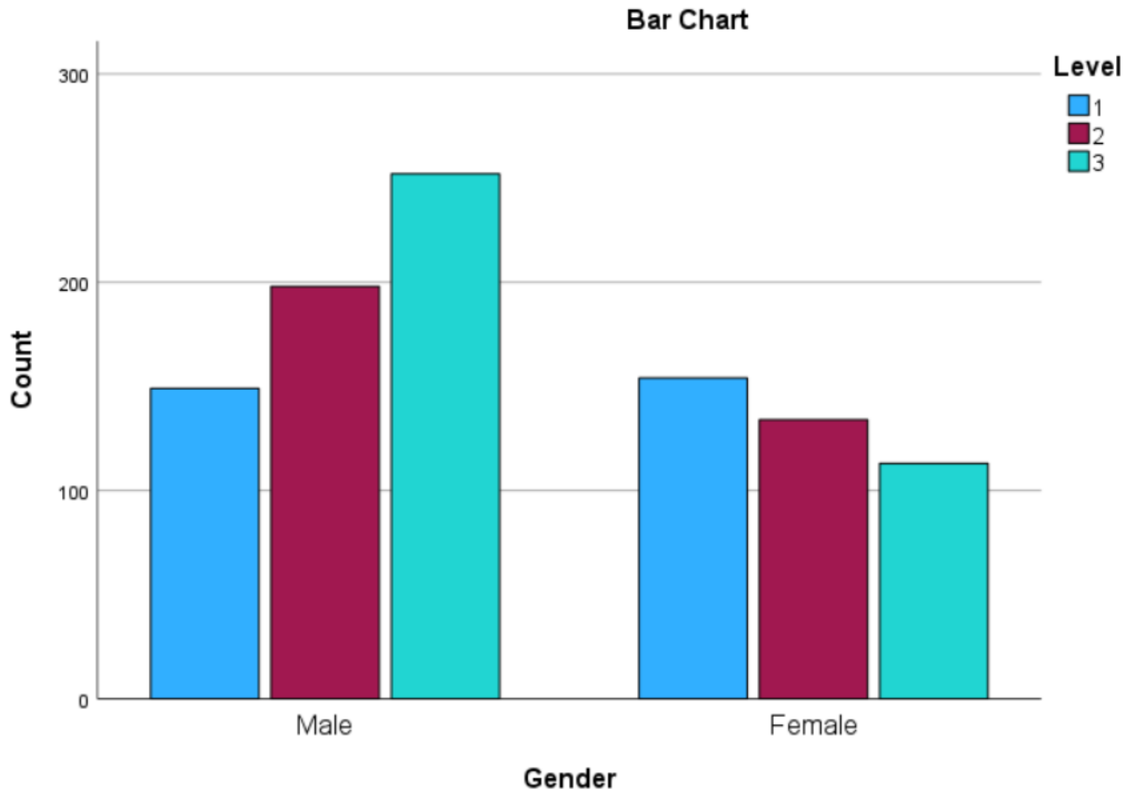
Statistical **Tests**

Brainnest | Group E

Proportions of Gender **by Level Cross- Tab**

Gender * Level Crosstabulation

			Level			
			1	2	3	Total
Gender	Male	Count	149	198	252	599
		% within Gender	24,9%	33,1%	42,1%	100,0%
		% within Level	49,2%	59,6%	69,0%	59,9%
		% of Total	14,9%	19,8%	25,2%	59,9%
	Female	Count	154	134	113	401
		% within Gender	38,4%	33,4%	28,2%	100,0%
		% within Level	50,8%	40,4%	31,0%	40,1%
		% of Total	15,4%	13,4%	11,3%	40,1%
Total		Count	303	332	365	1000
		% within Gender	30,3%	33,2%	36,5%	100,0%
		% within Level	100,0%	100,0%	100,0%	100,0%
		% of Total	30,3%	33,2%	36,5%	100,0%



Level 1 = Low
Level 2 = Medium
Level 3 = High

Brainnest | Group E

Proportions of Gender **by Level Group Statistics**

Independent-Samples Proportions Group Statistics

	Gender	Successes	Trials	Proportion	Asymptotic Standard Error
Level = 1	= M	149	599	.249	.018
	= F	154	401	.384	.024

The proportion is greater for female in level 1 (Low). (0.135 of difference).

	Gender	Successes	Trials	Proportion	Asymptotic Standard Error
Level = 2	= M	198	599	.331	.019
	= F	134	401	.334	.024

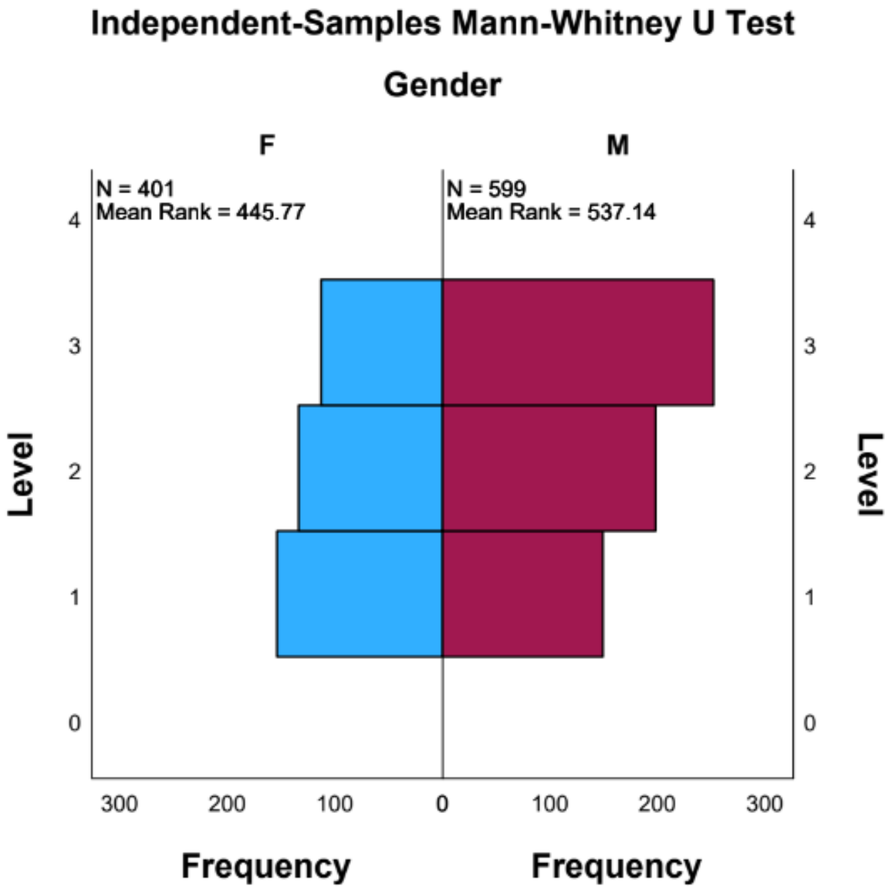
The proportion is greater for female in level 2 (Medium). (0.003 of difference).

	Gender	Successes	Trials	Proportion	Asymptotic Standard Error
Level = 3	= M	252	599	.421	.020
	= F	113	401	.282	.022

The proportion is greater for male in level 3 (High). (0.139 of difference)

Brainnest | Group E

Proportions of Gender **by Level Group Statistics**



Independent-Samples Mann-Whitney U Test Summary

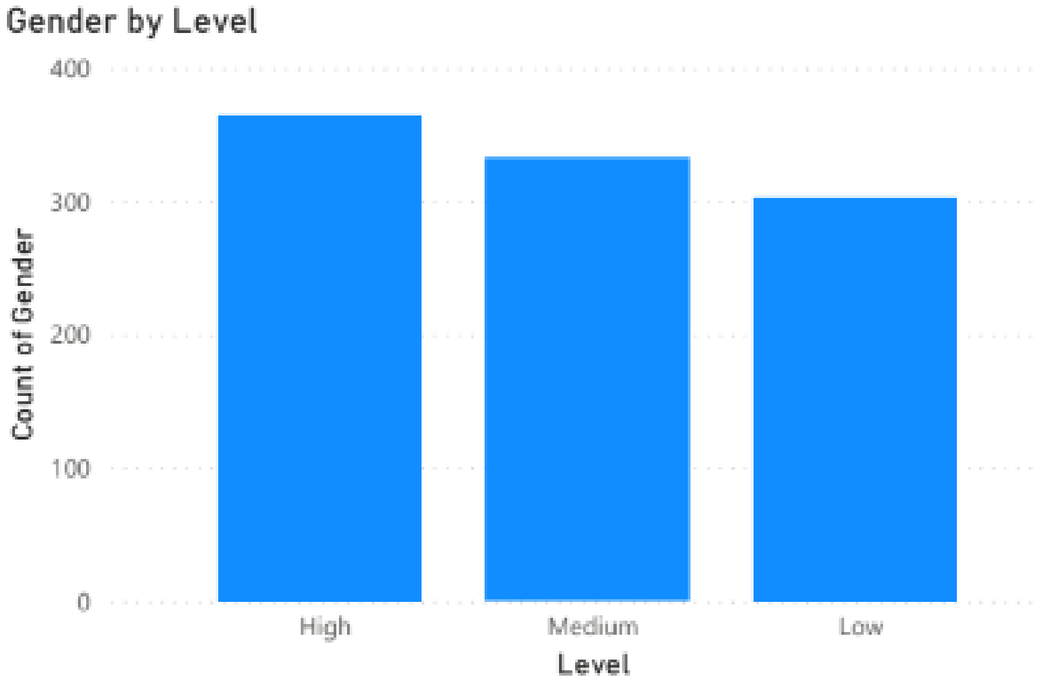
Total N	1000
Mann-Whitney U	142045.000
Wilcoxon W	321745.000
Test Statistic	142045.000
Standard Error	4215.648
Standardized Test Statistic	5.206
Asymptotic Sig.(2-sided test)	<.001

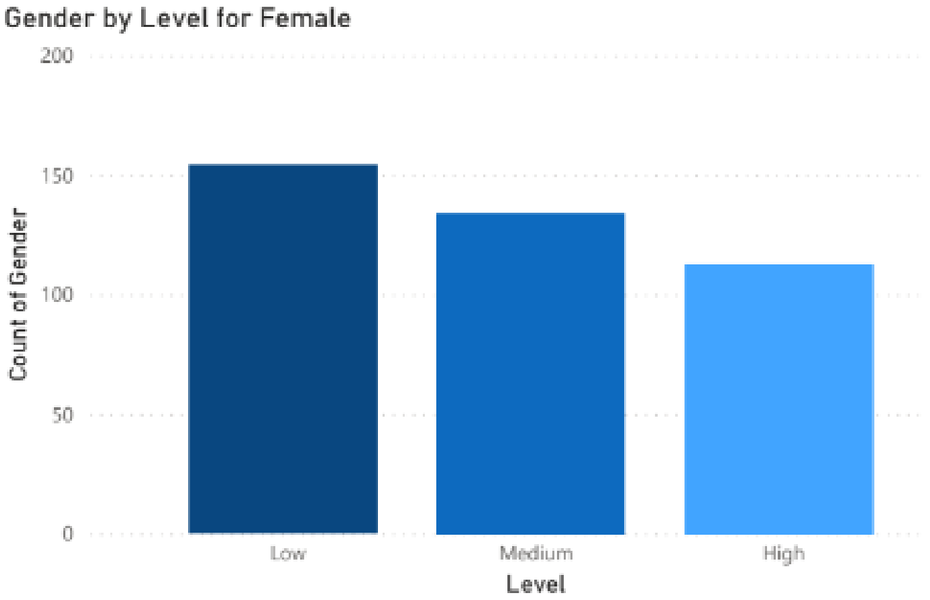
This chart shows us exactly the proportions of gender by Level Group Statistics side to side. As it was told, we can see clearly how there's a greater proportion of "High" level in male than female, while the proportion of "Low" level is greater in female than male.

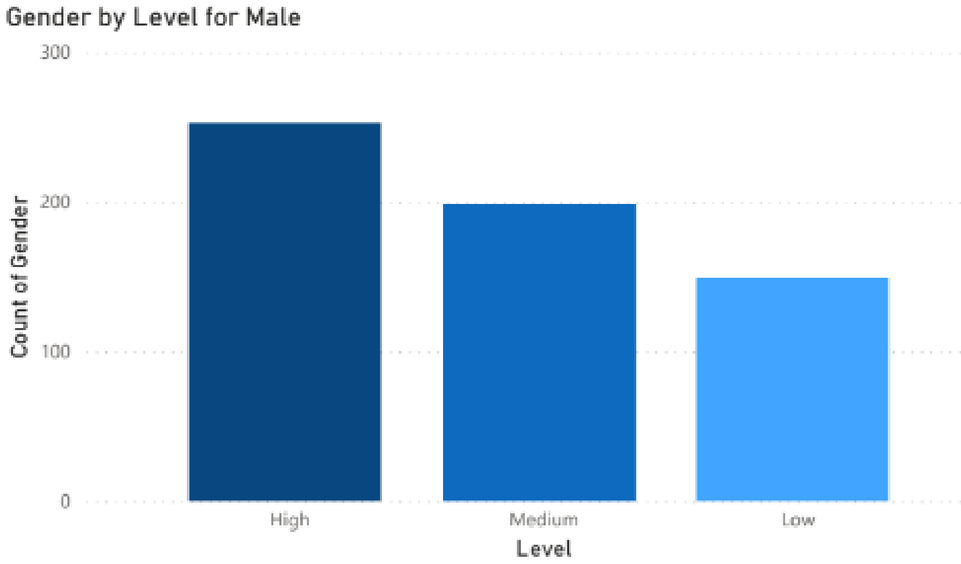
Brainnest | Group E

Insights and **Visualizations**

Brainnest | Group E

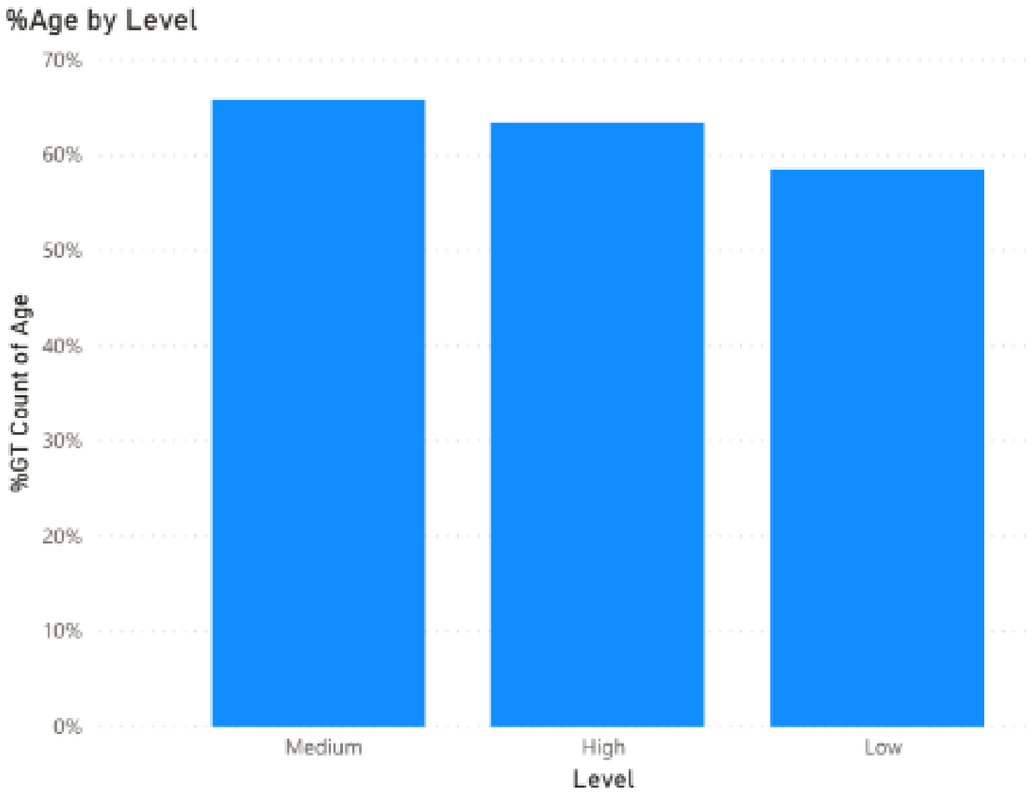


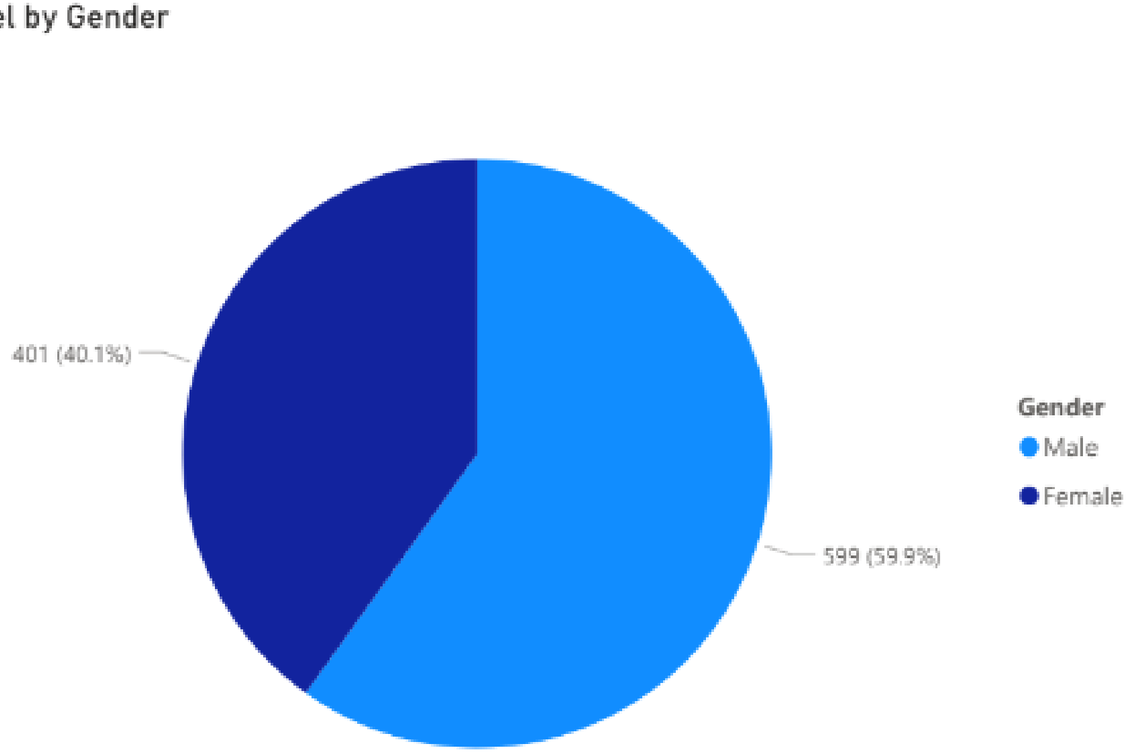




Even though we didn't use parametric statistical test, we can still see with the charts below the proportion of the cancer lever for males and females. We also can see for females the highest cancer level is "Low", on the other hands for males the highest cancer level is "High".

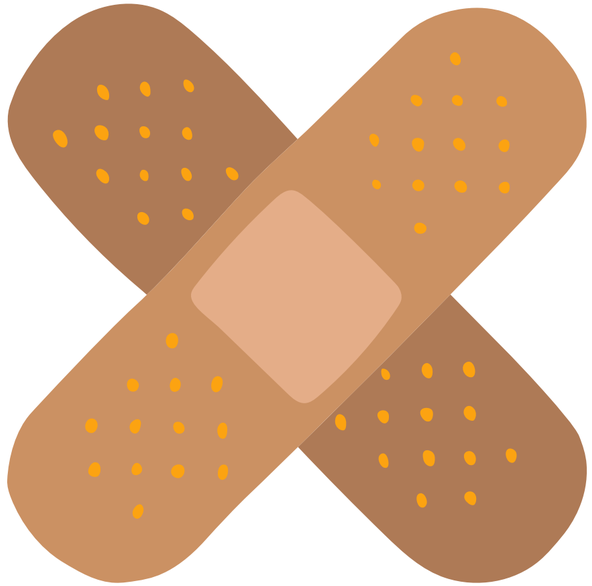
Brainnest | Group E





Conclusion

Visual and statistical tests like Shapiro Wilk, Kolmogorov Smirnov, Histogram, Q-Q and P-P plots, showed that the cancer level were not approximately normally distributed for both males and females with skewness of 0.189 (SE 0.122) and kurtosis of -1.454 (SE 0.243) for the females and skewness of 0.320 (SE 0.1) and kurtosis of -1.372 (SE 0.199) for the males.



Brainnest | Group E

Thank you!

Brainnest | Group E

Week THREE

Agenda

- Hypothesis 1
  - Hypothesis
  - One Sample T-Test
  - Conclusion with statistics
- Hypothesis 2
  - Hypothesis
  - Unpaired Sample T-Test
  - Conclusion with statistics
- Hypothesis 3
  - Hypothesis
  - ANOVA Two Factors
  - Conclusion with statistics

Hypothesis **Test 1**

Is there a difference between the sample and the population when looking at the average smokers?

Null Hypothesis (H0)	**Alternative Hypothesis (H1)**
There is no difference between the mean of the sample and the mean of the population	There is a difference between the mean of the sample and the mean of the population

In that case, we are comparing a **group** (sample) with the **population**. Therefore, our method to validate or not our hypothesis will be **One Sample T-Test**.

Smoking **Effect Size, Power**

Tests of Between-Subjects Effects

Dependent Variable: Smoking

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	,000[a]	0	.	.	.	,000	,000	.
Intercept	1600,000	1	1600,000	271,233	<,001	,733	271,233	1,000
Error	584,000	99	5,899					
Total	2184,000	100						
Corrected Total	584,000	99						

Effect size = 0.733 for a sample of 100 respondents and power is 100%.
The effect size means the differences will be meaningful.

So we have a 100% chance of rejecting the null hypothesis.

There is very little overlapping in the sample size as the effect size is greater.

One sample **T-test**

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Smoking	1000	1	8	3,95	2,486
Valid N (listwise)	1000				

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Smoking	100	1	8	4,00	2,429
Valid N (listwise)	100				

The mean of the smoking population is 3.95.

So our **t-value** = 3.95

The mean of the smoking sample is 4.00

One sample **T- test**

One-Sample Test

Test Value = 3.95

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Smoking	,206	99	,419	,837	,050	-,43	,53

One-Sample Effect Sizes

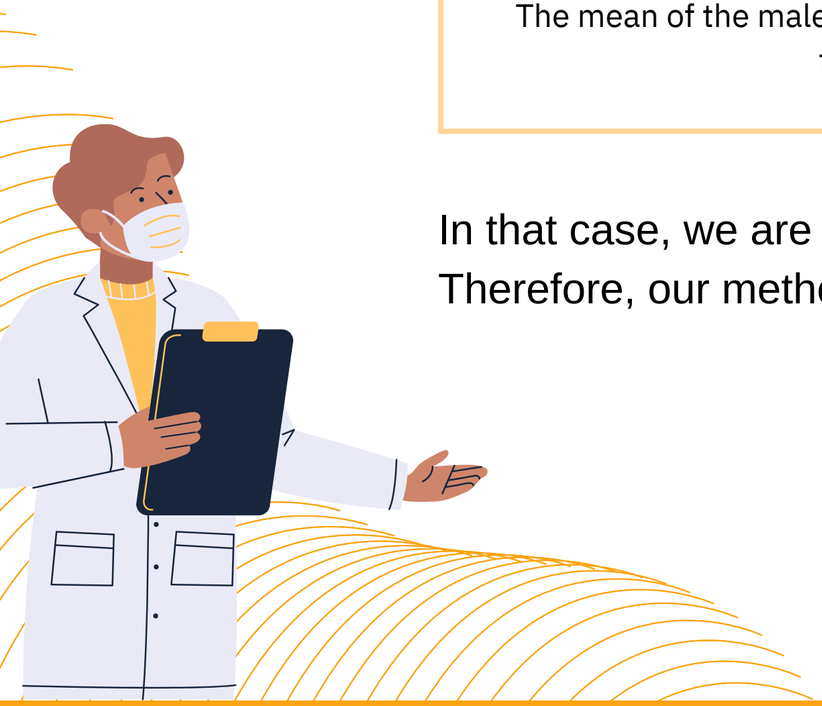
		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Smoking	Cohen's d	2,429	,021	-,175	,217
	Hedges' correction	2,447	,020	-,174	,215

- We get critical t = 1.660 from the t-table and test statistic t = 0.206, df = 99
  - If the test statistic > critical t, then we reject H0
  - In our case, the test statistic is not as extreme as the critical t.
    - 0.206 < 1.660 --> the null hypothesis is retained i.e. there is no difference in the means.
- If p < significance level, we reject the null hypothesis
  - In our case the two sided p-value corresponding to the Smoking variable is 0.837
  - When comparing this value to the significance level 5%
    - 0.837 > 0.05 --> the null hypothesis is retained.
- Conclusion:
  - The difference of the sample mean and the population mean for the Smoking variable is statistically insignificant.

Hypothesis **Test 2**

We want to check whether there are more male smokers than females.

Null Hypothesis (H0)	**Alternative Hypothesis (H1)**
The mean of the male smokers is the same as the mean of female smokers	The mean of the male smokers is **not** the same as the mean of female smokers



In that case, we are comparing two **independents groups** (male and female) with smoking. Therefore, our method to validate or not our hypothesis will be **Unpaired Sample T-Test**.

Unpaired (independent) samples **T-test**

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Smoking	M	62	4,16	2,497	,317
	F	38	3,74	2,321	,377

From the sample of 100 respondents, 62 are males and 38 are females with a mean of 4.16 and 3.74 respectively and their standard deviations are about the same.

Unpaired samples **T-test**

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Smoking	Equal variances assumed	,762	,385	,847	98	,200	,399	,424	,501	-,570	1,419
	Equal variances not assumed			,862	82,828	,196	,391	,424	,492	-,555	1,404

For comparing the two groups (male and female) of cancer patients on smoking we use the independent samples t-tests with **equal variances assumed**

There is a mean difference of 0.424

Limitation:

- Gender and Smoking are not normally distributed, so with the parametic test perfomed, normality is assumed

Unpaired samples **T-test**

We get critical t = 1.660 from the t-table and test statistic t = 0.847 with equal variances assumed, df = 98

- In our case, the test statistic is not as extreme as the critical t.
  - 0.847 < 1.660 --> the null hypothesis is retained

The two-sided p-value corresponding to the Smoking variable between males and females is 0.339

- When comparing this value to the significance level 5%
  - 0.339 > 0.05 --> the null hypothesis is retained.

Conclusion: There is no statistical significance in the difference of mean scores between the two groups. i.e. there is no difference in the means of female and male smokers.

Hypothesis **Test 3**

Besides smoking we want to check whether chest pain has an effect on cancer levels (with the assumption that the data follows a normal distribution).

Null Hypothesis (H0)	**Alternative Hypothesis (H1)**
There are no significant differences in the mean between the cancer levels and smoking	There is a significant difference in the mean between the cancer levels and smoking
There are no significant differences in the mean between the cancer levels and chest pain.	There is a significant difference in the mean between the cancer levels and chest pain.
Smoking has no effect on the effect of chest pains	Smoking has an effect on the effect of chest pains

In that case, we are comparing **two independents factors** (smoking and chest pain) to see their effect on a **dependent variable** (Level). Therefore, our method to validate or not our hypothesis will be **ANOVA Two Factors**.

ANOVA **Two Factors**

Tests of Between-Subjects Effects

Dependent Variable: Level

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	47,202[a]	22	2,146	9,123	<,001	,723	200,710	1,000
Intercept	96,537	1	96,537	410,493	<,001	,842	410,493	1,000
Smoking	3,282	7	,469	1,994	,067	,153	13,958	,742
ChestPain	13,185	7	1,884	8,009	<,001	,421	56,063	1,000
Smoking * ChestPain	5,020	8	,627	2,668	,012	,217	21,346	,905
Error	18,108	77	,235					
Total	519,000	100						
Corrected Total	65,310	99						

- Smoking's Sig (0.067) is greater than 5% (0.05), so it doesn't have any effect on cancer level. **That means we can accept the Null Hypothesis (Smoking has no effect on cancer level).**
- ChestPain's Sig (<0.001) is less than 5% (0.05), then it does have an effect on cancer level. **Because of that, we reject the Null Hypothesis (accept H1): Chest Pain has an effect on cancer level.**
- Smoking*ChestPain's Sig (0.012) is less than 5% (0.05), then it does have an effect on Level. **Therefore, we can accept the Alternative Hypothesis: Smoking has effect on the effect of chest pains**

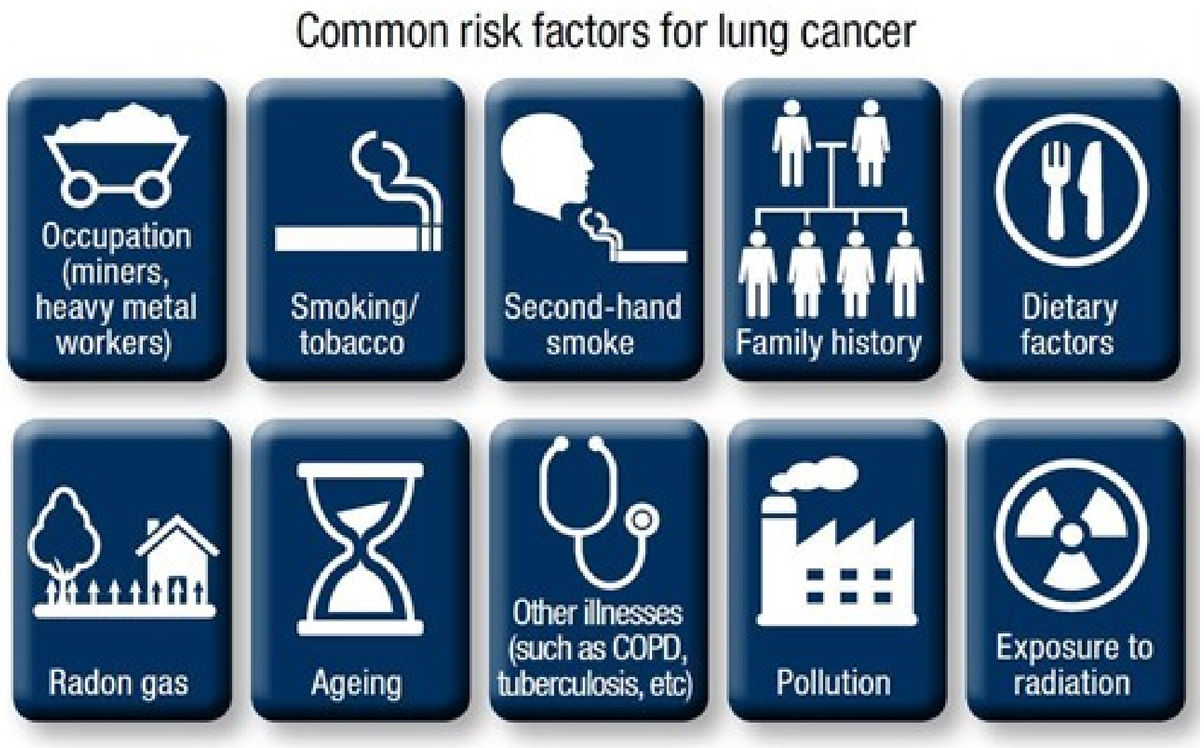
Brainnest | Group E

Thank you!

Brainnest | Group E

Week FOUR

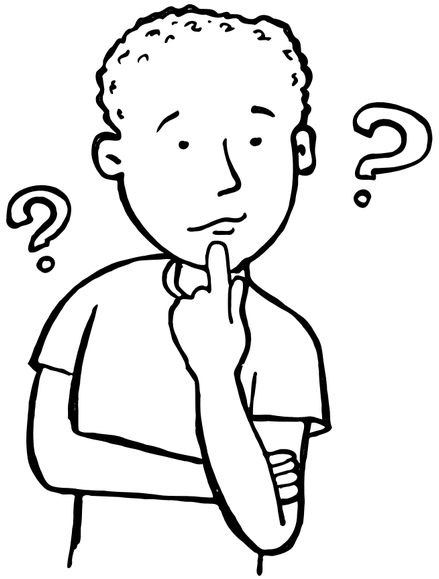
INTRO



According to the ***World Health Organization*** *(WHO)*, the most common cause of cancer deaths is **lung cancer** with **1.80 million deaths in 2020**.

In the cancer patients data, we see that **lifestyle factors** (*Smoking Passive Smoking, Alcohol use, Obesity Balanced Diet*), **Environmental factors** (*Air pollution, Occupational Hazards*), **Genetics**, and other illnesses such as **Chronic Lung** diseases are ***risk factors for cancer***.

INTRO

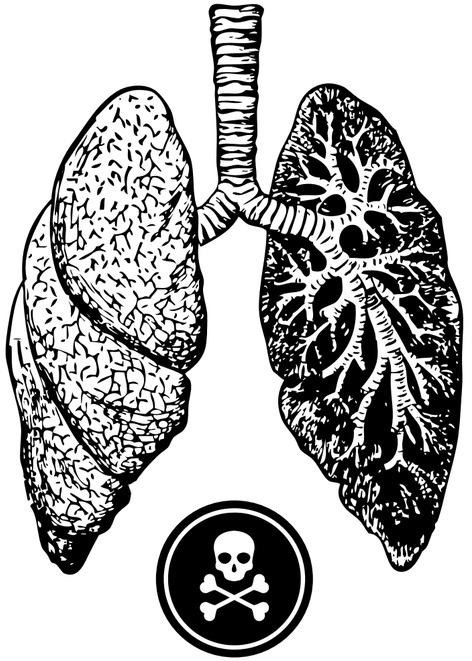


Looking at the symptoms/factors that have a strong correlation with the cancer level, we concluded that the cancer specificity is ***lung cancer***.

We want to check: ***Can reducing exposure to these risk factors also reduce the risk of lung cancer?***

First, we perform non-parametric tests on the exposure to passive smoking across gender groups as well as across categories of cancer levels

Then, we will also perform regression on variables in order to predict cancer levels.



Passive Smoker **and Gender**

According to the **American Cancer Society**, lung cancer rates have been ***higher in men,*** mainly because of their smoking patterns.

However, during the most recent period of the study, ***the rate was higher in woman from 30 to 49 in 6 countries: Canada, Denmark, Germany, New Zealand, the Netherlands and the US***.

We wanted to see the distribution of the risk factor: **--> Passive Smoker across Gender**

So we tested the following hypotheses statements:

Null Hypothesis (H0)	**Alternative Hypothesis (H1)**
The distribution of Passive Smoker is the ***same across*** all categories of Gender	The distribution of Passive Smoker ***is not the same*** across all categories of Gender

We have **two independent groups** - ***Male*** and ***Female*** - and a **dependent variable**: ***Passive Smoker*** and we will use the ***Mann Whitney*** non-parametric test for the above hypothesis.

Brainnest | Group E

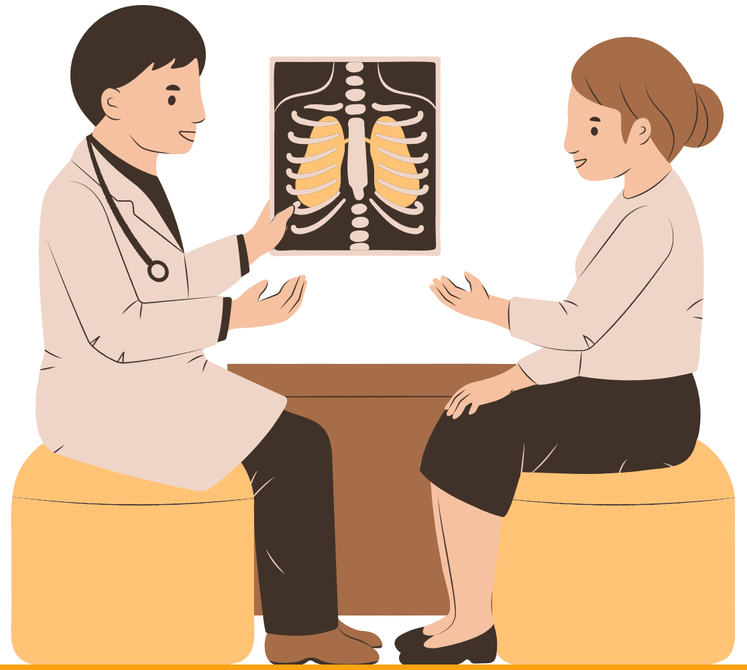
Passive Smoking and **Gender**

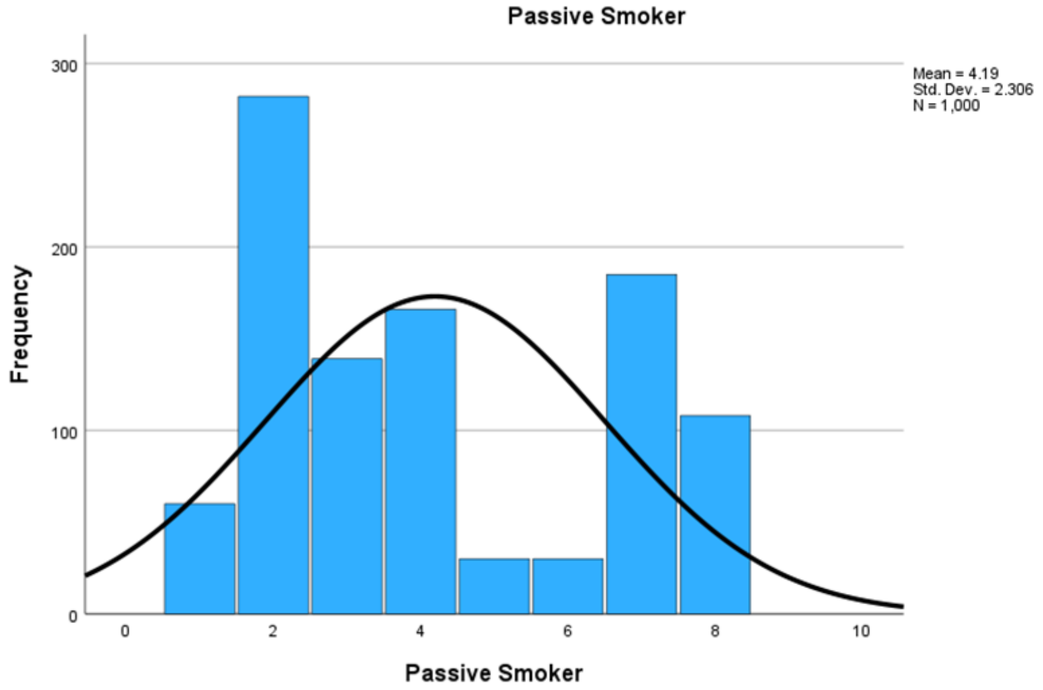
Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Passive Smoker	1000	4,19	2,306	1	8
Gender	1000	1,40	,490	1	2

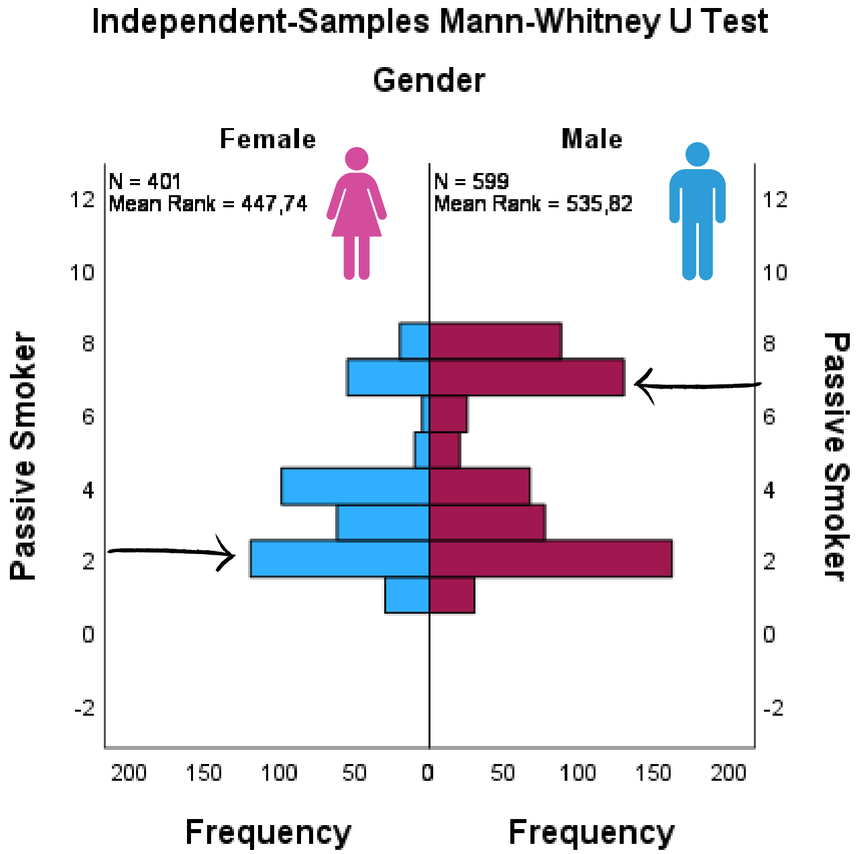
- Our histogram below shows that our outcome variable - **Passive Smoker - is *not normally distributed* for both genders.**

--> Which means, the ***data is right skewed***.





Passive Smoking and **Gender**



So we proceeded to an **Independent-Samples Man-Whitney U Test**

- The **most frequent risk score** for both genders is **2**.
- **There is a change in pattern**. The second most frequent rating is:
  - 7 for Males
  - 4 for Females

As mentioned before, the recent study made by the ***American Cancer Society*** showed an **increase of Female smokers** so we can suppose that due to that increase, **Males** may be **more exposed to passive smokers** (***second hand smoking***).

Brainnest | Group E

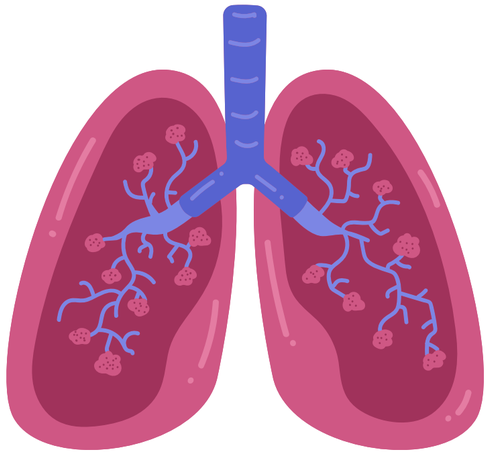
Passive Smoking and **Gender**

Ranks

	Gender	N	Mean Rank	Sum of Ranks
Passive Smoker	Male	599	535,82	320958,00
	Female	401	447,74	179542,00
	Total	1000		

Additionally:

- The **sum of the rankings** for males is **greater** than the one for females ***(=320 958)*** VS ***(=179 542)***



The **Hypothesis Test summary** tells us if the distribution of Passive Smoker is the same across categories of Gender.

Because p-value is < 0.001 < 0.05, than we **reject the null hypothesis**

Once again, we can suppose that Males are more exposed than Females to passive smoker.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Passive Smoker is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	<,001	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

Passive Smoking and **Gender**

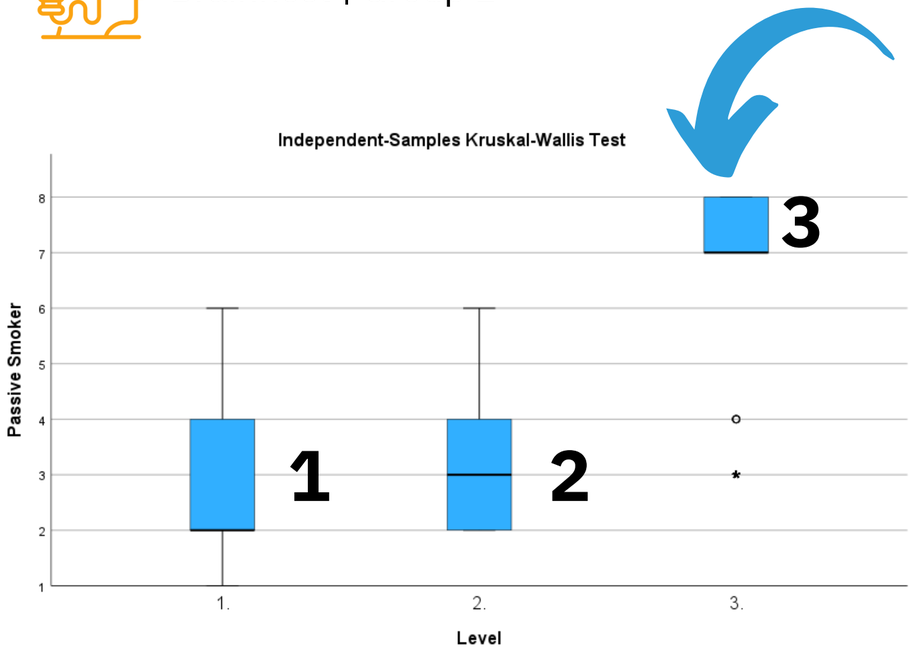
We tortured our data a bit further by checking the **distribution of the risk factor**: ***Passive Smoker across cancer levels***. So we tested the following hypothesis statements:

Null Hypothesis (H0)	**Alternative Hypothesis (H1)**
The distribution of Passive Smoker **is the same** across all categories of Level	The distribution of Passive Smoker **is not the same** across all categories of Level.

We have **three independent groups of Cancer *Level*** - ***Low, Medium and High*** - and **a dependent variable: *Passive Smoker*** and we will use the ***Kruskal-Wallis non-parametric test*** for the above hypothesis.

Brainnest | Group E

Passive Smoking and **Level**



Ranks

	Level	N	Mean Rank
Passive Smoker	1	303	309,39
	2	332	380,76
	3	365	768,06
	Total	1000	

Pairwise Comparisons of Level

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
1-2	-71,365	22,512	-3,170	,002	,005
1-3	-458,673	22,021	-20,829	<,001	,000
2-3	-387,308	21,489	-18,024	<,001	,000

- Box plots on the left **are not symmetric** for level 1 ,2 and 3. This means our **outcome variable - Passive Smoker** - is ***not normally distributed*** for all cancer level.

- The **median risk score increases** as the **cancer level increases**.
  - Level 1, median = 2
  - Level 2, median = 3
  - Level 3, median = 7,

--> so we can say that **high exposure to passive smoking can lead to high levels of cancer.**

- The **highest mean rank** is for level 3(=768.06), which is the **highest level of cancer** followed by level 2(=380.76)

- As the respondents have a **high risk score for passive smoking**, the **result is a high level of cancer**.

- **Conclusion**: ***The distribution of Passive Smoker is not the same across all categories of Gender, so we reject the Null Hypothesis.***

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Passive Smoker is the same across categories of Level.	Independent-Samples Kruskal-Wallis Test	<,001	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

Cancer Level **and** Symptoms **Correlations**

As shown in the figure below, Spearman's correlation shows that there is a strong positive relationship with symptoms such as:

- Obesity, Coughing of Blood, Dust Allergy, Passive Smoking having a coefficient closer to 1
- The correlations are statistically significant at $p < 0.05$ and $N = 1000$. There is a strong linear relationship between these symptoms and cancer levels.

There is also noticeable weak relationships with symptoms such as:

- Snoring : 0.282, Weight Loss: 0.355, and Smoking: 0.481 having a coefficient closer to 0
- We were surprised that Passive Smoking has a stronger correlation to Level than Smoking but Smoking and Passive Smoking are strongly positive correlated.

The factors with a strong linear relationship with cancer levels are actually the most common risk factors for lung cancer.

Correlations

		Obesity	Snoring	Alcohol use	Air Pollution	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Smoking	Passive Smoker	Chest Pain	Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath
Spearman's rho Level	Correlation Coefficient	.811**	.282**	.682**	.618**	.703**	.659**	.676**	.614**	.696**	.481**	.681**	.649**	.765**	.627**	.355**	.486**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	N	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Cancer Level **Prediction**

We want to know : ***Can we predict lung cancer level from risk scores of the symptoms?***
(With the assumption that our data follows a normal distribution).

We performed regression on a number of symptoms that have a linear relationship with cancer levels and assessed the model and its predictive accuracy.

Dependent Variable = Level
Independent Variables = Genetic Risk, Weight Loss and Alcohol use

Null Hypothesis (H0)	**Alternative Hypothesis (H1)**
There is no relationship between the cancer level and symptoms risk score	There is a relationship between the cancer level and symptoms risk score

Regression Model **Summary**

1

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.756[a]	.571	.570	.535

a. Predictors: (Constant), Genetic Risk, Weight Loss, Alcohol use

b. Dependent Variable: Level

2

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.830[a]	.689	.687	.456

a. Predictors: (Constant), Passive Smoker, Weight Loss, Alcohol use, Smoking, Genetic Risk

b. Dependent Variable: Level

From regression model 1 to 2, as we added more factors that are the most common risk factors for lung cancer, we see that R increases from 0.756 to 0.830

Regression Model **Summary**

We decided to remove weight loss and Smoking as it does not have a linear relationship with Level. From model 3 to 4 regression is performed between 9 factors and level.

We end up with a final model summary (4) with:

- R = 0.917, which is close to 1
  - the correlation between the predictors and the dependent variable is strong.
- R Square and Adjusted R Square are closer to each other with 0.841 and 0.840 respectively
  - R Square indicates that the 9 predictors accounts for some 84.1% of the variance in level. That is, they predict lung cancer level very well.

3

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.913[a]	.833	.831	.335

a. Predictors: (Constant), OccuPational Hazards, Weight Loss, Fatigue, Passive Smoker, Shortness of Breath, Coughing of Blood, Smoking, Balanced Diet, chronic Lung Disease, Dust Allergy, Alcohol use, Genetic Risk

b. Dependent Variable: Level

4

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.917[a]	.841	.840	.327	.841	476.313	11	988	<.001

a. Predictors: (Constant), OccuPational Hazards, Fatigue, Shortness of Breath, Air Pollution, Passive Smoker, Coughing of Blood, Balanced Diet, chronic Lung Disease, Obesity, Dust Allergy, Alcohol use

b. Dependent Variable: Level

Regression Model **Results**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.160	.037		4.375	<.001
	Alcohol use	.066	.011	.210	6.219	<.001
	Passive Smoker	.081	.009	.228	9.421	<.001
	Air Pollution	.007	.009	.017	.785	.433
	Obesity	.080	.011	.208	7.200	<.001
	Shortness of Breath	.060	.007	.169	9.019	<.001
	Fatigue	.080	.006	.220	12.818	<.001
	Coughing of Blood	.058	.008	.173	6.907	<.001
	chronic Lung Disease	.053	.013	.121	4.063	<.001
	Balanced Diet	-.010	.009	-.026	-1.057	.291
	Dust Allergy	.040	.012	.096	3.288	.001
	OccuPational Hazards	-.076	.018	-.195	-4.148	<.001

a. Dependent Variable: Level

1. Most coefficients for the independent variables are positive, so they have a positive relationship with Level. Occupational Hazards and Chest Pain have a negative relationship with Level though.
2. P-value < 0.05 for most of the predictor coefficients and the p-value of the F-statistics(<0.001) < 0.05 which indicates that there is a relationship between the features(predictors) and the dependent variable: Level
3. **On the table on the next page, we see that using low scores for factor risk results in low cancer level. The higher the score is used for prediction the higher levels of cancer**
    a. **Therefore, we conclude that we can reduce the risk of level of cancer by reducing exposure to factors that are avoidable.**

Predicting **Cancer Level**

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k$$

$$\hat{y} = 0.160 + 0.060x_1 + 0.081x_2 + 0.080x_3 + 0.060x_4 + 0.080x_5 + 0.058x_6 + 0.053x_7 + 0.040x_8 - 0.076x_9$$

	994	605	184	635	891	803	115	838	266	917
PatientId	P500	P911	P837	P469	P656	P714	P979	P171	P961	P608
Age	26.0	73.0	17.0	46.0	38.0	36.0	45.0	52.0	44.0	64.0
Gender	2	1	2	2	2	1	1	1	1	1
Alcoholuse	8	6	2	6	8	7	1	8	2	8
PassiveSmoker	7	5	2	3	2	7	4	2	2	8
Obesity	7	5	1	2	4	7	3	4	3	7
ShortnessofBreath	4	6	3	7	6	7	2	6	2	5
Fatigue	3	4	1	3	2	8	3	2	4	9
CoughingofBlood	9	5	2	2	3	7	1	3	5	7
chronicLungDisease	6	5	3	2	6	7	3	6	4	6
DustAllergy	7	6	3	7	7	7	4	7	3	7
OccuPationalHazards	7	5	4	5	7	7	3	7	4	7
Level	3	2	1	2	2	3	1	2	2	3
Predicted_Level	2.835	2.42	0.873	1.705	1.882	3.292	1.293	1.882	1.44	3.34

Logistic **Regression**

Do the **Passive Smoker** and **Genetic Risk** of the patients have an influence on the probability they have a high **Level of Cancer** ?

(With the assumption that our data follows a normal distribution).

Here, our dependent variable (Level of Cancer) is categorical, therefore we are dealing with a Logistic Regression.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,783[a]	,613	,612	,508

a. Predictors: (Constant), Genetic Risk, Passive Smoker

b. Dependent Variable: Level

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	,644	,040		16,235	<,001	,566	,722
	Passive Smoker	,155	,009	,438	17,698	<,001	,138	,172
	Genetic Risk	,168	,010	,436	17,591	<,001	,149	,186

a. Dependent Variable: Level

Looking at the Model Summary, we can see R is a high number, i.e. we have a great variability on Cancer Level that can be explained by the variables **Passive Smoker** and **Genetic Risk**.

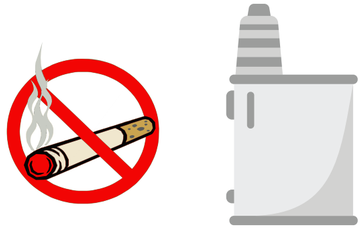
On the coefficients, we can see the Passive Smoker and Genetic Risk p-value is less than 0.001, **i.e. there's a strong influence on Cancer Level.**

Medical **Recommendations**

Some recommandations to avoid Lung Cancer

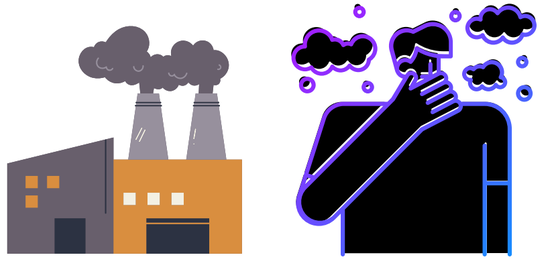
Avoid or reduce consumption of alcohol

Avoid second hand smoke



Maintain a healthy body weight

Minimizing occupational exposure



Have enough sleep to avoid fatigue

References' **Links**

- **World Health Organization**: https://www.who.int/news-room/fact-sheets/detail/cancer
- **Statistics By Jim**: https://statisticsbyjim.com/basics/graph-groups-boxplots-individual-values/
- **Statistics By Jim**: https://statisticsbyjim.com/graphs/x-and-y-axis/
- **Better Health Channel**: https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/passive-smoking

Brainnest | Group E

Thank you!