ANALYSING STUDENT SURVEY DATA: A STATISTICAL SOLUTION GUIDE

Dr. Akinlolu Akande
Dr. Syam Kumar
Dr. David Obada

Ollscoil Teicneolaíochta an Atlantaigh
Atlantic Technological University

TORR
Transforming Learning

This document was created to help undergraduate students in understanding data and statistics concepts. Additionally, postgraduate students who are embarking on their research journey and require a quick review of fundamental concepts in data and statistics will also find value in this resource.

Furthermore, we would like to acknowledge funding and support received from the National Technological University TransfOrmation for Resilience & Recovery (NTUTORR). The title of the project funded is "*Implementing and evaluating project-based learning (PBL) in undergraduate introductory statistics modules at ATU Sligo – A Pilot Study*".

akinlolu.akande@atu.ie

syam.kumar@atu.ie

david.obada@atu.ie

VERSION: 01

APRIL 2024

SCHOOL OF SCIENCE, ATLANTIC TECHNOLOGICAL UNIVERSITY, SLIGO

AVAILABLE RESOURCES FOR STUDENTS AND INSTRUCTORS

This material is part of our collection of resources that have been developed from over a decade of teaching introductory statistics and data analysis to undergraduate students at ATU Sligo and various Irish tertiary institutions, as well as providing Microsoft Excel classes and trainings as part of our Information Technology Modules. It assumes the use of computer software, particularly recommending Excel for beginners before progressing to more advanced statistical tools like SPSS or R. As such, we have used Excel to implement the various analysis and statistical tests.

Primarily aimed at undergraduate modules in science, health, business, and related disciplines that leverage Excel for data analysis, our materials are intended to be comprehensive resources. Accompanying these resources is a dedicated repository at https://thewee.link/NTUTORR-PBL-Resources, offering a lot of useful resources for both students and instructors.

- Access to pertinent datasets required for completing exercises within this resource.
- Detailed instructions on conducting data and statistical analysis using Excel.
- Provision of Excel outputs to aid students in verifying their work during exercises.
- Sample answers to questions featured within this resource.
- A compilation of notes covering introductory statistics and data analysis.
- Additional datasets suitable for project-based learning (PBL) endeavours.

NOTE

Within the afore mentioned repository, we have curated resources designed to support users as they navigate through the exercises outlined in this material.

TABLE OF CONTENTS

INTRODUCTION TO DATASET

For several years, a first-day survey has been administered to students in an introductory statistics class at one university. The data for 362 students have been collected and stored in an Excel Spreadsheet.

The datasets contain plenty of information that can be used to learn more about students. The repository contains,

- Excel Spreadsheet named **StudentSurvey.xlsx** containing the raw data collected from the cohort of 362 students.
- Excel Spreadsheet named **CleanedStudentSurvey.xlsx** containing the cleaned-up data after deleting records with missing values.

All values are self-reported.

In any dataset, it is important to understand exactly what each variable is measuring and how the values are coded.

Data Dictionary:

Variable Name	Description and Detail
Year	Year in school
Gender	Student's gender: *Gender M for male and F for female*
Smoke	Does the student smoke?
Award	Award the student prefers to win
Exercise	Hours of exercise per week
TV	Hours of TV viewing per week
Height	Height (in cm)
Weight	Weight (in kg)
Siblings	Number of siblings
BirthOrder	Birth order: *1 for first/oldest, 2 for second born, etc.*
GPA	College grade point average *on a 4-point scale*
Pulse	Pulse rate (beats per minute) *at the time of the survey*
Piercings	Number of body piercings

In what follows, you will perform statistical analysis on selected parameters relating to the students taking Introductory Statistics Module in that University.

PART A: FIRST THINGS FIRST

Dataset Insights

Prior to commencing any data analysis, it is advisable to gain an understanding of the dataset's columns and their respective data types. Additionally, it is essential to ***determine the presence of null*** or ***missing values*** within the dataset. Finding the ***total number of rows*** and ***columns*** in the dataset is also a crucial preliminary step. Furthermore, in some cases it is important to ***identify the dependent*** and ***independent variables*** within the dataset.

Insert Your Answers Below in Red Fonts

1. **Data Cleaning and Shape of the dataset:** Always check whether there are missing values in your dataset and decide on what you would like to do with them. ***For this particular project, we have deleted the rows containing the missing values and saved the new data file as*** "**CleanedStudentSurvey.xlsx**". Write down the number of rows and columns in the cleaned-up data set.

   Shape of the Dataset after cleaning:

   Number of rows = 331

   Number of columns = 13

2. **Classify Data Type:**
   a) Use the data dictionary to identify each variable in the dataset as categorical or quantitative. If the variable is categorical, further identify it as ordinal, nominal, or an identifier variable. If the variable is quantitative, identify it as discrete or continuous.

Categorical Variables are:	*Gender, Smoke, Award, Year, BirthOrder (Though numeric but could not do numerical computation with it)*
Ordinal:	*Year, BirthOrder*
Nominal:	*Gender, Smoke, Award*
Identifier:	*Nil*
Quantitative Variables are:	*Exercise, Siblings, Piercings, Height, Weight, GPA, Pulse, TV*
Discrete:	*Siblings, Piercings, Pulse*
Continuous:	*Exercise, Height, Weight, GPA, TV*

b) ***Data Preprocessing:*** If the unique categories of the categorical variables are not listed in the Data Dictionary, it is necessary to FIRST OF ALL identify them. Identify the unique categories of all the categorical variables. The first variable (first row) has been done for you, complete the rest of the following table.

Categorical Variables	*Categories*
Year	*4 Categories: FirstYear, SecondYear, ThirdYear, FourthYear*
BirthOrder	*8 Categories: 1, 2, 3, 4, 5, 6, 7, 8*
Gender	*2 Categories: M, F*
Smoke	*2 Categories: No, Yes*
Award	*3 Categories: Olympic, Academy, Nobel*

3. **Sample and Population:** Is the dataset from a sample or a population? If it is from a sample, describe a relevant population to which we might make inferences.

This is a sample dataset.
A relevant population is the whole student body in the University.

Research Questions

The formulation of meaningful statistical research questions is of importance in data analysis. These questions serve as the foundation upon which research endeavours are built, guiding the entire data analysis procedures. They drive the selection of appropriate data collection methods, statistical techniques and hypotheses to be tested. By posing precise research questions, you can clarify your objectives, focus your investigations and define the scope of your studies.

1. **Research/Statistical Questions – One Variable:** List at least two questions we might ask about any one of these individual variables.

Questions
1. *What percentage of people would like to win an Olympic reward?* 2. *What percentage of students do not smoke?* 3. *What is the average number of hours of exercise a typical student will do a week?* 4. *Etc.*

2. **Research/Statistical Questions – Two Variables (Association or Relationship):** List at least two questions we might ask about relationships between any two (or more) of these variables.

Questions
1. *Is there a difference in GPAs between different academic years?* 2. *Do students who smoke have different GPAs compared to non-smokers?* 3. *Is there a correlation between the number of hours of exercise per week and body weight?* 4. *Is there a relationship between students' GPAs and their pulse rates?* 5. *What proportion of male student have piercings in their bodies?* 6. *Is there any difference in the heights of female students compared to their male counterparts?* 7. *Etc.*

DESCRIPTIVE STATISTICS

PART B: INVESTIGATING AND DESCRIBING **CATEGORICAL** DATA

Univariate Analysis

GOAL: Our goal is to summarise the overall ***patterns/trends*** in a single categorical variable of interest. Here, we do not want to just present raw data but simple table and chart that best represent the overall ***story*** we learn from the data.

Refer to the cleaned-up Student Survey data ("**CleanedStudentSurvey.xlsx"**). Consider the variable "***Award***". Organise it into frequency distribution and use appropriate chart to display the distribution. Write a brief report summarising your findings. Be sure to answer the following questions in your report.

Insert Your Answers Below in Blue Fonts

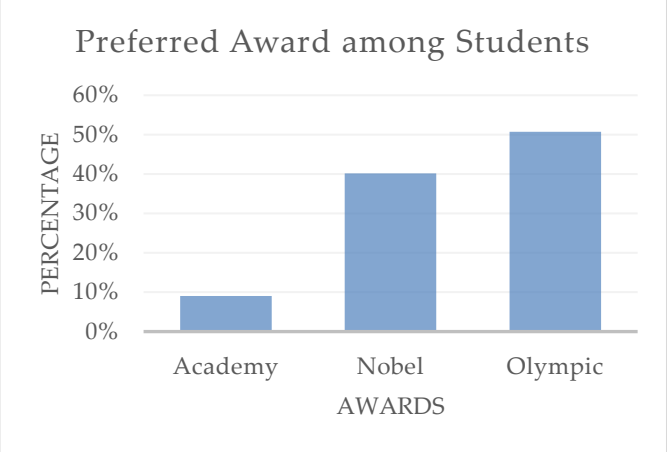
1. **Frequency Distribution**: Create a frequency table showing the frequency, relative frequency and percentage of preferred *Award* among the students surveyed.

Paste Your Table Here. Make sure the table is well labeled.

Awards	Frequency	Relative Frequency	Percentage
Academy	30	0.09	9%
Nobel	133	0.40	40%
Olympic	168	0.51	51%
Total	331	1.00	100%

2. **Bar or Pie Chart:** Produce a bar or a pie chart to display the *Preferred Award* among the students surveyed.

Paste Your Chart Here. Make sure the axes are well labeled and the chart is titled.



3. **Summarise Your Findings**: Write a sentence on the chart quoting any relevant statistics to aid your explanation.

The chart shows the distribution of awards among the surveyed individuals revealing that the majority preferred Olympic awards (51% of the students), followed by Nobel awards (40% of the students), while Academy awards were least frequent (9% of the students).

Bivariate Analysis

GOAL: Our goal is to summarise the overall ***association/relationship*** between two categorical variables of interest. Then, we present simple table and chart that best represent the overall ***story*** we learn from the data.

Refer to the cleaned-up Student Survey data ("**CleanedStudentSurvey.xlsx"**). Consider the variables "***Award***" and "***Gender***" and develop a contingency table to show the relationship between them. Use appropriate chart to display the distribution and write a brief report summarising your findings. Be sure to answer the following questions in your report.

Insert Your Answers Below in Blue Fonts

4. **Two-way Table or Crosstab or Contingency Table**: Create a contingency table to show the joint distribution of *Award* and *Gender.* Report both the count and conditional distribution cases.

Research Question: Does preferred award depend on gender?

Independent Variable: Gender

Dependent Variable: Award

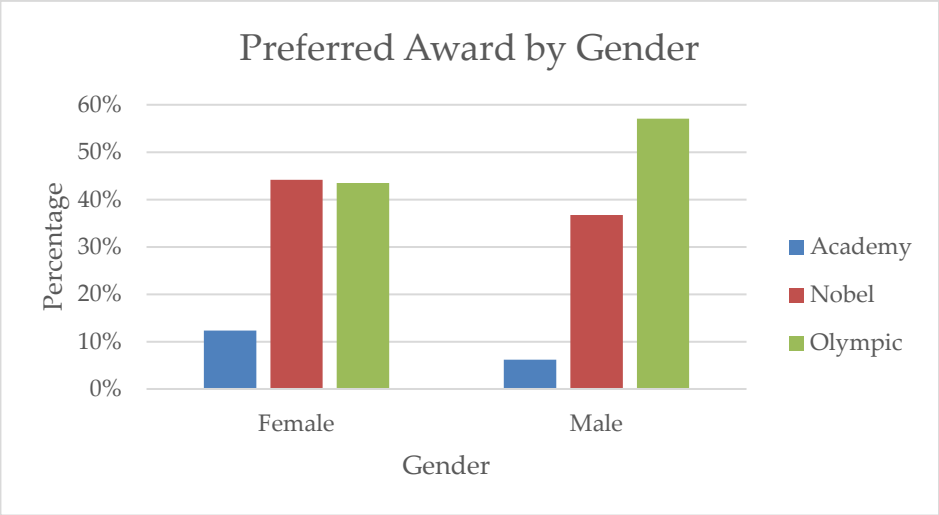
Paste Your Tables Here. Make sure the table is well labeled.

Awards	Academy	Nobel	Olympic	Total
Female	19	68	67	154
Male	11	65	101	177
Total	30	133	168	331

Awards	Academy	Nobel	Olympic	Total
Female	12%	44%	44%	100%
Male	6%	37%	57%	100%
Total	**9%**	**40%**	**51%**	**100%**

5. **Side-by-side bar charts**: Produce a side-by-side bar chart to show the *Preferred Award by Gender.*

Paste Your Chart Here. Make sure the axes are well labeled and the chart is titled.



6. **Summarise Your Findings:** Write a paragraph explaining what the chart shows quoting any relevant statistics to aid your explanation.

There appears to be an association between preferred award and gender. Olympic awards are most preferred by male students (57%). For female students, Nobel and Olympic awards had a close preference (44%). Academy awards had the least preferences in both genders, with double the number of females (12%) preferring the academy award compared to their male counterparts.

PART C: INVESTIGATING AND DESCRIBING **UNIVARIATE** QUANTITATIVE DATA

Tabular and Graphical Summaries

GOAL: Our goal is to *organise* and *graphically* summarise the overall ***patterns/trends*** in the quantitative variables of interest. Here, we do not want to just present raw data but simple tables and appropriate charts that best represent the overall ***story*** we learn from the data.

Refer to the cleaned-up Student Survey data ("**CleanedStudentSurvey.xlsx"**). Consider the variable "***Exercise***". Organise it into frequency distribution and use appropriate chart to display the distribution. Write a brief report summarising your findings. Be sure to answer the following questions in your report.

Insert Your Answers Below in Blue Fonts

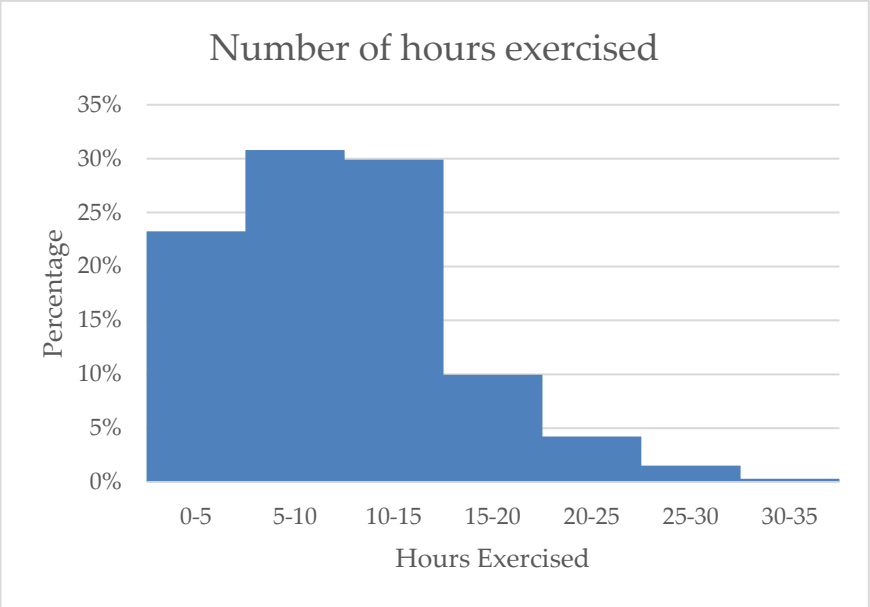
1. **Frequency Distribution**: Create a frequency table showing the frequency, relative frequency and percentage of the number of hours students exercised per week. (*Exercise*)

Paste Your Table Here. Make sure the table is well labeled.

Intervals	Frequency	Relative Frequency	Percentage
0-5	77	0.233	23.3
5-10	102	0.308	30.8
10-15	99	0.299	29.9
15-20	33	0.100	10.0
20-25	14	0.042	4.2
25-30	5	0.015	1.5
30-35	1	0.003	0.3
Total	331	1.000	100.0

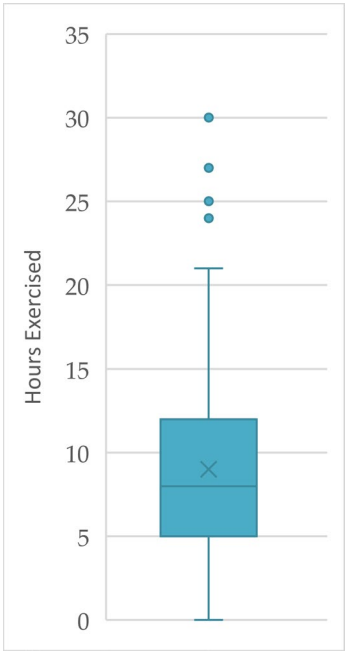
2. **Histogram:** Produce a histogram to display the number of hours students exercised per week. (*Exercise*)

Paste Your Chart Here. Make sure the axes are well labeled and the chart is titled.



3. **Boxplot:** Produce a boxplot to display the number of hours students exercised per week. (*Exercise*)

Paste Your Chart Here. Make sure the axes are well labeled and the chart is titled.



4. **Summarise Your Findings**: Describe the histogram/boxplot showing the distribution of *Exercise*. *(You can describe a* histogram/boxplot *by its* ***shape, centre, spread*** *and* ***possible outliers).***

Insert Your Answer in the Following Table

Shape
The distribution is right skewed.
Centre
The midpoint or the middle class of the distribution is about 5 to 10 hours of exercise per week.
Spread
The spread is from 0 to 30 hours of exercise per week.
Possible Outliers
Intervals between 20 to 30 hours of exercise per week are the possible outliers. The boxplot confirms that the distribution is right skewed, with 4 outliers on the higher side of the distribution. ***Note**: Be curious about outliers! They might be unusual observations but correct observations or there is an error in the data.*

Numerical Summaries

GOAL: Our goal is to *numerically* summarise the overall ***patterns/trends*** in the quantitative variables of interest. Here, we do not want to just present raw data but *numbers* that best represent the overall ***story*** we learn from the data.

Refer to the cleaned-up Student Survey data ("**CleanedStudentSurvey.xlsx"**). Consider the variable "***Exercise***". Prepare a report on its numerical summaries detailing your findings. Be sure to answer the following questions in your report.

Insert Your Answers Below in Blue Fonts

1. **Numerical Measures**: Produce numerical summaries (measures of center, spread, location and shape) of the number of hours students surveyed exercised per week. (*Exercise*). Discuss some of these measures and check their agreements with graphical summaries.

Complete the following table to 2 decimal places.

Sample Size	
Count (n)	*331.00*
Measures of Center	
Mean	*9.00*
Median	*8.00*

Measures of Location	
Minimum	*0.00*
Q1	*5.00*
Q3	*12.00*
Maximum	*30.00*
Measures of Shape	
Skewness	*0.86*
Kurtosis	*0.60*
Measures of Spread	
Range	*30.00*
IQR	*7.00*
Variance	*31.20*
Standard Deviation	*5.59*

Insert Your Discussions checking agreement with the graphical summaries in the following Box

Comparing Mean and Median: The mean (9.00) is greater than the median (8.00) and therefore the distribution is skewed to the right (positively skewed distribution).
Skewness: The skewness (0.86) is positive and suggests moderately positively skewed distribution.
The numerical summaries agree with the graphical representation showing that the distribution is moderately skewed to the right.

2. **Appropriate Numerical Measures**: Based on the shape of the distribution of *Exercise*, report the appropriate measures of center and spread for the data.

Insert Your Answer in the Following Box

Due to the moderate positive skewness and outliers present in the distribution, the median and IQR are appropriate measures of center and spread, respectively, for the distribution.
Median: *8 hours*
IQR: *7 hours*

3. **Summarise Your Findings**: Draw conclusions about the number of hours of exercise of a typical student.

Insert Your Answer in the Following Box

The typical weekly average hour of exercise is 8 hours with a spread of 7 hours. Typical Exercise Hours per week = 8 ± 7 hours

PART D: **BIVARIATE** ANALYSIS INVOLVING AT LEAST ONE QUANTITATIVE VARIABLE

One Categorical Variable and One Quantitative Variable (Comparative Analysis)

GOAL: Our goal is to *graphically* and *numerically* compare the overall ***patterns/trends*** in a quantitative variable of interest across the categories or groups of a categorical variable of interest. Here, we present numbers and appropriate charts that best represent the overall ***story*** we learn from the data.

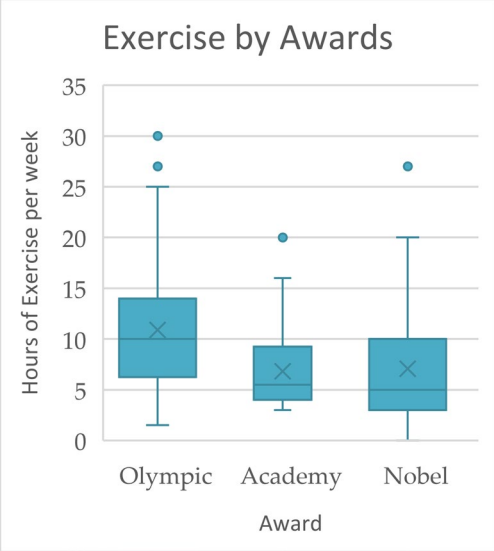
Refer to the cleaned-up Student Survey data ("**CleanedStudentSurvey.xlsx"**). Consider the variables "***Exercise***" and "***Award***" and display ***Exercise*** across all the groups of ***Award*** using an appropriate chart. What do you observe? Write a brief report summarising your findings and be sure to answer the following questions in your report.

Research Question: Is there a difference in the mean of the hours of exercise per week across the groups of preferred awards among the students surveyed?

Insert Your Answers Below in Blue Fonts

1. **Side-by-side Boxplots:** Produce side-by-side box plots illustrating the distribution of *Exercise* by *Award*.

Paste Your Chart Here. Make sure the axes are well labeled and the chart is titled.



2. **Compare Numerical Measures:** Compare the measures of center and measures of spread of *Exercise* by *Award* by completing the following table.

Complete the following table to 2 decimal places.

Exercise	*Award*		
	Academy	*Olympic*	*Nobel*
Count	*30.00*	*168.00*	*133.00*
Mean	*6.83*	*10.91*	*7.08*
Standard Deviation	*4.15*	*5.68*	*4.91*
Median	*5.50*	*10.00*	*5.00*
IQR	*4.75*	*7.25*	*7.00*
Standard Error	*0.76*	*0.44*	*0.43*

3. **Summarise Your Findings**: Summarise your results about the relationship of *Exercise* by *Award* in the students surveyed.

Insert Your Answer in the Following Box

All the distributions are right skewed, with outliers in the upper distributions. Students who preferred Olympic awards exhibit the highest measure of center (median = 10 hours), indicating a propensity for increased physical activity, likely associated with their involvement in sports. There is much variation in Olympic and Nobel. Academy has the least spread suggesting a more consistent exercise pattern within this group. The data thus suggests a discernible relationship between academic/athletic achievements and exercise habits, with students who preferred Olympic awards exhibiting higher and less consistent exercise levels than their peers who preferred Academy and Nobel categories.

Two Quantitative Variables (Correlation and Regression)

GOAL: Our goal is to summarise the overall ***association/relationship*** between two quantitative variables of interest. Then, we present numbers and charts that best represent the overall ***story*** we learn from the data.

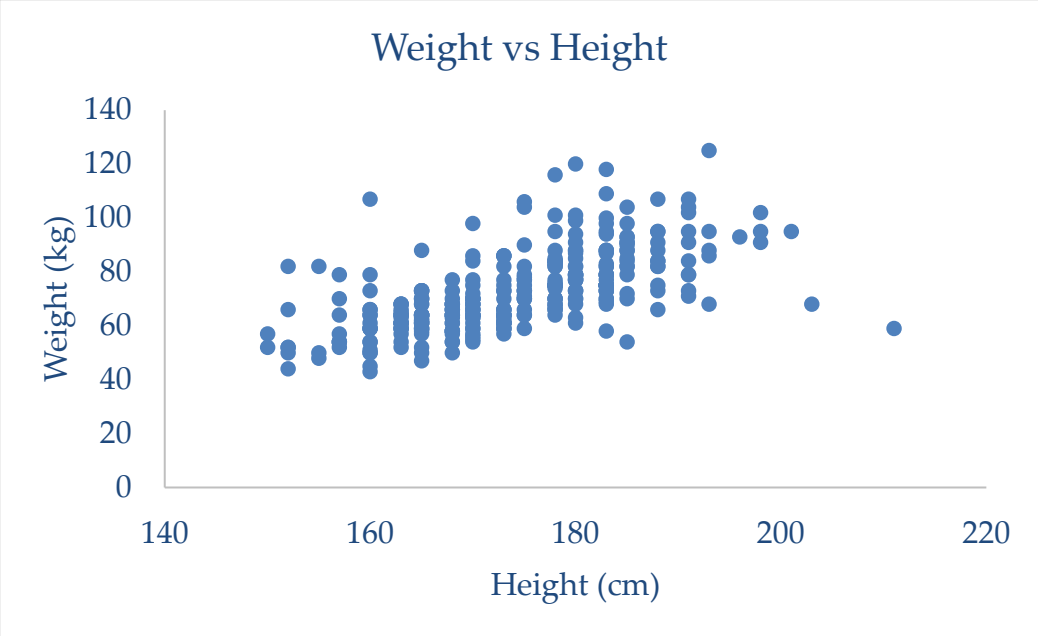
Research Question: Is there a relationship between "**Weight**" and "**Height**" of the students surveyed?

Refer to the cleaned-up Student Survey data ("**CleanedStudentSurvey.xlsx"**). Develop a regression equation that expresses the relationship between "**Weight**" and "**Height**". Let "**Weight**" be the dependent variable and "**Height**" be the independent variable. Write a brief report summarising your findings and be sure to answer the following questions in your report.

Insert Your Answers Below in Blue Fonts

1. **Scatterplot:** Make a scatterplot to display the relationship between *Height* (in cm) and *Weight* (in kg).

Paste Your Chart Here. Make sure the axes are well labeled and the chart is titled.



2. **Comment** on direction, form and strength: Describe any patterns apparent from the scatterplot. (Does there appear to be a positive or negative relationship between height and weight? How strong does the trend appear to be? Does it appear to be approximately a linear trend?). Is there any outlier? Describe the outlier(s), if any.

Insert Your Answer in the Following Table

Direction
The scatterplot has a positive direction.
Form
The form of the plot is linear.
Strength
The plot is moderately correlated.
Possible Outliers
There are potential outliers in the data set which include (160,107), (211,59), (203,68), etc.

3. **Correlation**: Obtain the correlation between *Height* and *Weight* and interpret. (The correlation coefficient is a measure of strength and direction of the association between the variables) Does the value support your observation in Deliverable 2?

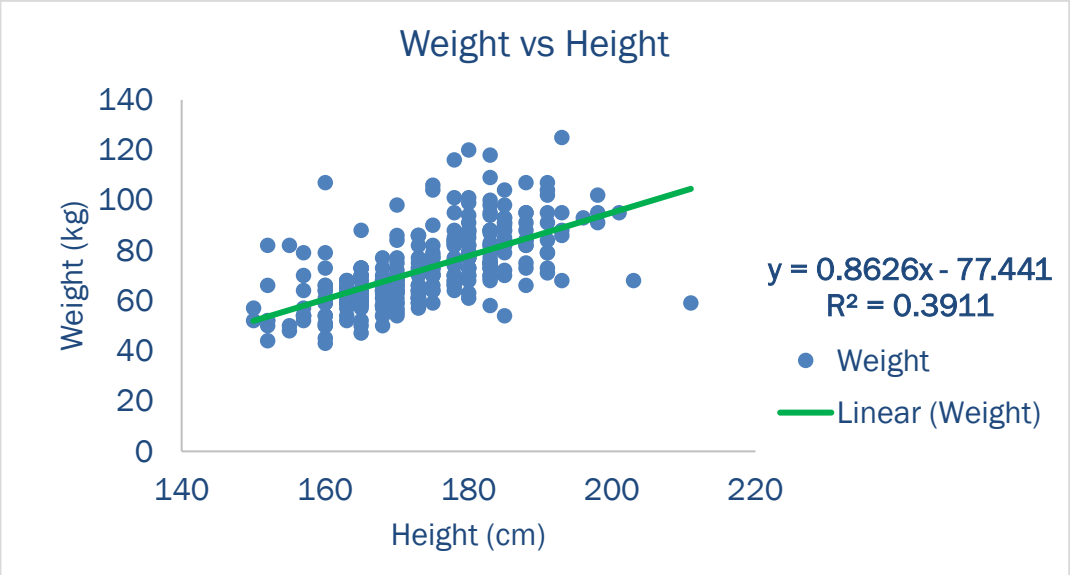
Insert Your Answer in the Following Box

Correlation coefficient, $r = 0.63$
The *positive* correlation coefficient between Height and Weight, at 0.63, supports the observation of a positive relationship between the two variables. As Height increases, Weight tends to increase as well.

Also, the closer the absolute value of the correlation coefficient is to unity, the stronger the correlation. Here the absolute value of correlation coefficient is 0.63, suggesting that the variables are moderately correlated as observed in Deliverable 2.

4. **Regression Line and Equation:** Modify the scatterplot from Deliverable 1 to include the regression line and the regression equation that relates *Height* and *Weight*.

Paste Your Chart Here. Make sure the axes are well labeled and the chart is titled.



Insert the Appropriate Equation of the Line in the Following Box

$$\boldsymbol{Weight = 0.8626 \times Height - 77.441}$$

5. **Interpret Slope and Intercept:** What is the slope of the line? Interpret it in context. What is the intercept of the line? If it is reasonable to do so, interpret it in context. If it is not reasonable, explain why not.

Insert Your Answer in the Following Table

Slope
Slope = 0.8626 **Interpretation:** The positive sign in the slope represents a positive relationship between the Weight and Height. The slope of 0.8626 indicates that for every one-unit increase in height, weight increases by 0.8626.
Intercept on y-axis
Intercept = - 77.441 **Interpretation:** - 77.441 is the value at which the line crosses the Weight-axis when Height is zero. However, zero is not a reasonable value for Height. Hence, this value does not have a physical meaning in this case.

6. **Prediction when appropriate:** Use the regression line to determine the expected weight for a who is 5 feet tall (152 cm). What weight is predicted for someone 6 feet tall (183 cm).

Insert Your Answer in the Following Table

Height	*Expected Weight*
5 feet (152 cm)	*53.67 kg*
6 feet (183 cm)	*80.41 kg*

7. **Prediction when it is not appropriate:** What weight does the regression line predict for a baby who is 51 cm long? Why is it not appropriate to use the regression line in this case?

Insert Your Answer in the Following Table

Height	*Expected Weight*
51 cm	*-33.45 kg*

Explanation: *The predicted weight of -33.45 kg for a baby with a height of 51 cm, obtained from the regression line, raises concerns about the appropriateness of using the model for extrapolation. This prediction falls outside the range of the original dataset and is physically implausible since weight cannot be negative. Extrapolating beyond the range of the dataset can result in unreliable predictions, as we do not know if the linear relationship continues beyond the range used in this model.*

8. **Coefficient of determination, R^2, and its meaning:** Obtain R^2 for this relationship. Interpret it in context of *Height* and *Weight*. (Does *Height* seem to be a good predictor of *Weight*?)

Insert Your Answer in the Following Box

The coefficient of determination, R^2 for this relationship is 0.3911. This means that 39.11% of the variability in the Weight can be explained by variations in Height according to the linear regression model. This suggests that Height alone accounts for a moderate amount of the observed variation in Weight, implying that there is still a substantial amount of variability in Weight that is not explained by Height alone.

STATISTICAL INFERENCE

PART E: SAMPLING DISTRIBUTION AND CENTRAL LIMIT THEOREM

1. **Compute Numerical Measures of the Entire Set:** Report the measures of center and measures of spread of the *GPA* of the students surveyed by *Gender* by completing the following table.

Insert Your Answers Below in Blue Fonts

Complete the following table to 2 decimal places

Exercise of the Entire Set	*Gender*	
	Female	*Male*
Count	*154*	*177*
Mean	*3.24*	*3.09*
Standard Deviation	*0.38*	*0.40*
Standard Error	*0.03*	*0.03*

2. **Random Selection of Cases:** Select a simple random sample of 50 cases from Male and 50 cases from Female (Making a total of 100 cases). Report the measures of center and measures of spread of *GPA* by *Gender* by completing the "Random Selection 1" row in the following table. Repeat the random selection process 25 times and record the means and standard deviations for Male and Female.

Complete the following table to 2 decimal places

Random Selections	*Female*		*Male*	
	Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
1	*3.24*	*0.36*	*3.01*	*0.46*
2	*3.21*	*0.29*	*3.13*	*0.41*
3	*3.20*	*0.39*	*3.08*	*0.40*
4	*3.20*	*0.34*	*3.11*	*0.42*
5	*3.28*	*0.35*	*3.07*	*0.32*

6	*3.27*	*0.40*	*2.99*	*0.43*
7	*3.19*	*0.45*	*3.12*	*0.42*
8	*3.23*	*0.35*	*3.16*	*0.37*
9	*3.24*	*0.38*	*3.13*	*0.26*
10	*3.34*	*0.34*	*3.13*	*0.36*
11	*3.26*	*0.41*	*3.06*	*0.43*
12	*3.27*	*0.40*	*3.06*	*0.36*
13	*3.21*	*0.40*	*3.00*	*0.44*
14	*3.31*	*0.37*	*3.10*	*0.43*
15	*3.32*	*0.37*	*3.16*	*0.39*
16	*3.23*	*0.35*	*3.09*	*0.38*
17	*3.25*	*0.36*	*3.08*	*0.42*
18	*3.27*	*0.27*	*3.17*	*0.37*
19	*3.14*	*0.36*	*3.06*	*0.40*
20	*3.17*	*0.42*	*3.20*	*0.36*
21	*3.24*	*0.40*	*3.15*	*0.33*
22	*3.15*	*0.40*	*3.10*	*0.34*
23	*3.13*	*0.39*	*3.09*	*0.36*
24	*3.21*	*0.41*	*3.15*	*0.37*
25	*3.27*	*0.37*	*3.14*	*0.33*
AVERAGE OF MEANS	*3.23*		*3.10*	
STANDARD DEVIATION OF MEANS		*0.05*		*0.05*

3. **Compare the Averages of Means and Standard Deviation of the Randomly Selected Cases with the Mean and the Standard Error of Entire Set:** Compare the average of the means

and the standard deviation of the randomly selected cases of *GPA* by *Gender* with the mean and the standard error of the entire set of *GPA* by *Gender*. Comment on your observation.

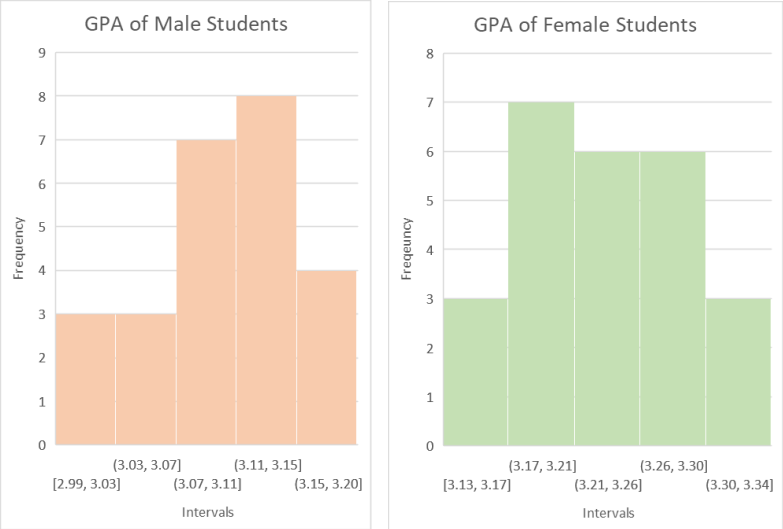
The mean and standard deviation of the sample means of the randomly selected cases (sample) for female are 3.23 and 0.05 respectively, whereas the mean for the entire set (population) were obtained as 3.24 and standard error of 0.03.

This is confirming one of the key results of CLT that states that, given a sufficiently large sample, the mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the ***standard error of the mean***, is nearly equal to the population standard deviation divided by the square root of the sample size (n) – standard error.

Similarly, the mean and standard deviation of the sample means of the randomly selected cases (sample) for male are 3.10 and 0.05 respectively, whereas the mean for the entire set (population) were obtained as 3.09 and standard error of 0.03.
This confirms the CLT.

4. **Sampling Distribution of the Sample Means**: Construct histograms showing the distribution of the means of the *GPA* by *Gender* of the randomly selected cases.

   Paste Your Chart Here. Make sure the axes are well labeled and the chart is titled.



5. **Comment on the histograms**: Comment in relation to Central Limit Theorem.

The Central Limit Theorem (CLT) states that, for a sufficiently large sample size, the distribution of the sample means will be approximately normally distributed, regardless of the shape of the original population distribution. In line with the CLT, the histogram's shape for both genders tends to approach a normal distribution.

PART F: INFERENCE ABOUT A POPULATION MEAN AND CONFIDENCE INTERVALS

Research Question (a): Suppose the national average of GPA is 2.9. Is there evidence that the students in this college have a GPA more than the national average?

1. List the variable(s) involved and the type(s) (Qualitative or Quantitative?).

Insert Your Answer in the Following Box

How many variables are involved? One continuous variable **Variable(s) involved:** GPA **Types of variables:** Quantitative (Continuous) variable

2. Identify the null and alternative hypotheses.

Insert Your Answer in the Following Box

Null Hypothesis (H_0)
The mean GPA of students in the college is less than or equal to 2.9. $\mu \leq 2.9$
Alternative Hypothesis (H_1)
The mean GPA of students in the college is greater than 2.9. $\mu > 2.9$

3. Identify the appropriate statistical test for these hypotheses and when the test can be used.

Insert Your Answer in the Following Box

Statistical test for these hypotheses
Appropriate statistical test: One-sample *t* test **When to use:** Only use the one-sample t-test to analyse a continuous (interval or ratio scale) dependent variable from a single sample (group) by comparing the sample mean to a hypothesised test value. The population standard deviation is not known.

4. Verify that the assumptions for using that test are met.

Insert Your Answer in the Following Box

Assumptions and Verification for the test
i. The data should be collected through a random sampling method to ensure that the sample is representative of the population: *The data was collected through a random sampling method.*

ii. The t-test assumes that the population from which the sample is drawn is normally distributed (Normality can be tested with the histogram, boxplot, comparing mean and median, etc.) or the sample size is sufficiently large (typically n > 30) for the Central Limit Theorem to apply: *The sample size is sufficiently larger than 30.*

iii. The data should be at the interval or ratio scale. The t-test is not appropriate for nominal or ordinal data: *The variable is (continuous) quantitative*.

iv. The observations in the sample should be independent of each other: *The observations in the sample were independent of each other.*

5. Calculate the sampling distribution of the sample mean under the null hypothesis.

Insert Your Answer in the Following Box

Under the null hypothesis, the sample mean has a t-distribution with n−1 degrees of freedom. The mean is equal to the mean under the null hypothesis.

Sample Mean = 3.16
Sample Size = 331
Sample Standard Deviation = 0.40
Degree of Freedom = 330
Standard Error of the Mean = 0.02

Under the null hypothesis and subject to the assumptions we checked earlier, the sample mean has a t-distribution with 330 degrees of freedom. The mean is 3.16 and the standard error is 0.02.

6. Conduct the hypothesis test and report your conclusion at the alpha = 0.05 significance level.

Insert Your Answer in the Following Box

We have a one-sided (more than) alternative hypothesis. As this is a right tailed hypothesis test, the critical t-value was found to be 1.65.

The calculated t-value is 11.81 whereas the critical t-value is 1.65.

Conclusion: Absolute value of the calculated test statistic is greater than absolute value of the critical value. Hence, it falls within the area of rejecting the null hypothesis.

Plain English: We reject the null hypothesis. There is evidence to support that the mean GPA of students in the college is greater than 2.9.

ALTERNATIVELY
The p-value was calculated to be 0.00. As this is less than the significance level (α=0.05), the result is statistically significant. This also implies that we reject the null hypothesis.

7. Report and interpret a 95% confidence interval for the true mean of *GPA*.

Insert Your Answer in the Following Box

For a left-tailed test, the one-sided CI is $(-\infty, \mathbf{95\%\ Upper\ Bound})$
where the 95% Upper Bound **=** $\bar{x} + t_{\propto} \frac{s}{\sqrt{n}}$ and $t_{\propto}$ is the one tailed critical t-value.

For a right tailed test, the one-sided CI is **(95% Lower Bound,** ∞**)**,
where the 95% Lower Bound **=** $\bar{x} - t_{\propto} \frac{s}{\sqrt{n}}$ and $t_{\propto}$ is the one tailed critical t-value.
As this is a right tailed test,
95% Lower Bound = 3.12
95% Upper Bound = ∞

Conclusion: We can be 95% confident that the interval between 3.12 and ∞ contains the population mean GPA of the college students. As the interval does not contain the hypothesised mean $(\mu = 2.9)$, it can be concluded that the result is statistically significant. This is the same conclusion reached previously that we reject the null hypothesis.

8. **Summarise the test results: As** a minimum, the following information should be reported in the results section of any research report when using one-sample t-test:
   - null hypothesis that is being evaluated to include test value,
   - descriptive statistics (e.g., mean, SD, n),
   - statistical test used (i.e., one-sample t-test),
   - results of evaluation of test assumptions, and
   - test results.

Insert Your Answer in the Following Box

A one-sample t-test was conducted to evaluate the null hypothesis that mean GPA of students in the college is less than or equal to the norm or hypothesised value of 2.9 (n = 331). The test showed that the sample mean (mean = 3.16, SD = 0.40) was significantly different than the test value of 2.9, t(df = 330) = 1.65 critical t-value, p = 0.00 (1-tailed). Consequently, there was sufficient evidence to reject the null hypothesis. We have sufficient evidence to support the alternative hypothesis that the mean GPA is greater than 2.9.

Research Question (b): Using the entire dataset, will the mean pulse rate of students surveyed be different from 72 beats per minute?

9. List the variable(s) involved and the type(s) (Qualitative or Quantitative?).

Insert Your Answer in the Following Box

How many variables are involved? One discrete variable **Variable(s) involved:** Pulse **Types of variables:** Quantitative (Discrete) variable

10. Identify the null and alternative hypotheses.

Insert Your Answer in the Following Box

Null Hypothesis (H_0)
Mean pulse rate of students surveyed is equal to 72 beats per minute. $\mu = 72$
Alternative Hypothesis (H_1)
Mean pulse rate of students surveyed is different from 72 beats per minute. $\mu \neq 72$

11. Identify the appropriate statistical test for these hypotheses and when the test can be used.

Insert Your Answer in the Following Box

Statistical test for these hypotheses
Appropriate statistical test: One-sample t-test **When to use:** Only use the one-sample t-test to analyse a continuous (interval or ratio scale) dependent variable from a single sample (group) by comparing the sample mean to a hypothesised test value. The population standard deviation is not known.

12. Verify that the assumptions for using that test are met.

Insert Your Answer in the Following Box

Assumptions and Verification for the test
i. The data should be collected through a random sampling method to ensure that the sample is representative of the population: *The data was collected through a random sampling method.* ii. The t-test assumes that the population from which the sample is drawn is normally distributed (Normality can be tested with the histogram, boxplot, comparing mean and median, etc.) or the sample size is sufficiently large

(typically n > 30) for the Central Limit Theorem to apply: *The sample size is sufficiently larger than 30.*

iii. The data should be at the interval or ratio scale. The t-test is not appropriate for nominal or ordinal data: *The variable is (discrete) quantitative*.

iv. The observations in the sample should be independent of each other: *The observations in the sample were independent of each other.*

13. Calculate the sampling distribution of the sample mean under the null hypothesis.

Insert Your Answer in the Following Box

Under the null hypothesis, the sample mean has a t-distribution with n−1 degrees of freedom. The mean is equal to the mean under the null hypothesis.

Sample Mean = 69.87
Sample Size = 331
Sample Standard Deviation = 12.07
Degree of Freedom = 330
Standard Error of the Mean = 0.66

Under the null hypothesis and subject to the assumptions we checked earlier, the sample mean has a t-distribution with 330 degrees of freedom. The mean is 69.87 and the standard error is 0.66.

14. Conduct the hypothesis test and report your conclusion at the alpha = 0.05 significance level.

Insert Your Answer in the Following Box

We have a two-sided (not equal to) alternative hypothesis.

The critical t-value = 1.97

The calculated t-value = -3.21

Conclusion: Absolute value of the calculated test statistic is greater than absolute value of the critical value, hence, it falls within the area of rejecting the null hypothesis.

Plain English: We reject the null hypothesis. There is evidence to support that the mean pulse rate of students surveyed is different from 72bpm.

ALTERNATIVELY
The p-value was calculated to be 0.0014. As this is less than the significance level (α=0.05), the result is statistically significant. This also implies that we reject the null hypothesis.

15. Report and interpret a 95% confidence interval for the true mean of *Pulse*.

Insert Your Answer in the Following Box

95% CI Lower Bound = 68.56

95% CI Upper Bound = 71.17

Conclusion: We can be 95% confident that the interval between 68.56 and 71.17 contains the population mean pulse rate of the college students. As the interval does not contain the hypothesised mean (μ=72), it can be concluded that the result is statistically significant. This is the same conclusion reached previously that we reject the null hypothesis.

16. **Summarise the test results: As** a minimum, the following information should be reported in the results section of any research report when using one-sample t-test:

- null hypothesis that is being evaluated to include test value,
- descriptive statistics (e.g., mean, SD, n),
- statistical test used (i.e., one-sample t-test),
- results of evaluation of test assumptions, and
- test results.

Insert Your Answer in the Following Box

A one-sample t-test was conducted to evaluate the null hypothesis that there is no difference in the mean pulse rate of students surveyed and the norm or hypothesised value of 72 beats per minutes (n = 331). The test showed that the sample mean (mean = 69.87, SD = 12.07) was significantly different than the test value of 72, t(df = 330) = 1.97 critical t-value, p = 0.0014 (2-tailed). Consequently, there was sufficient evidence to reject the null hypothesis. We have sufficient evidence to support the alternative hypothesis that the mean pulse rate of students surveyed is different from 72 beats per minutes.

PART G: COMPARING TWO POPULATION MEANS AND CONFIDENCE INTERVALS

Research Question (a): Using the entire dataset, is there any difference between the average GPA by Gender?

1. List the variable(s) involved and the type(s) (Qualitative or Quantitative?).

Insert Your Answer in the Following Box

How many variables are involved? Two variables are involved.

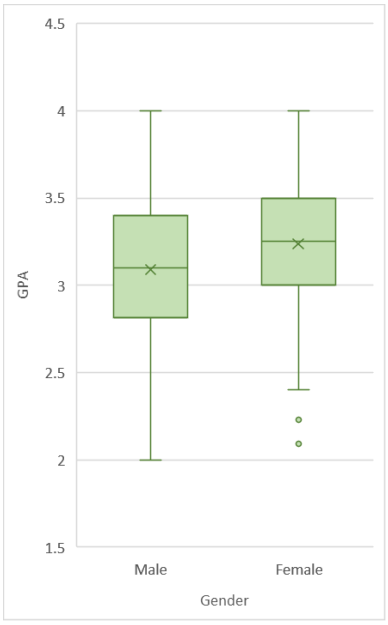
Variable(s) involved: GPA and Gender

Types of variables:

GPA: Quantitative (continuous) variable
Gender: Categorical variable with two categories (Male, Female)

2. Report Comparative Analysis from the Sample Data using either the side-by-side boxplots or bar charts with error bars. Comment on the chart.

Paste Your Chart Here. Make sure the axes are well labeled.



Insert Your Comment in the Following Box

Both distributions are approximately symmetric.
Mean GPA (3.24) of female students surveyed is slightly higher than those of their male counterparts (3.09), The variations in the two groups are nearly the same. There are two outliers in the lower distribution of female students.

3. Identify the null and alternative hypotheses.

Insert Your Answer in the Following Box

Null Hypothesis (H_0)
There is no difference in the mean GPA among the male and female students. $\mu_{male} = \mu_{female}$ OR $\mu_{male} - \mu_{female} = 0$
Alternative Hypothesis (H_1)
The mean GPA among the male and female students are different. $\mu_{male} \neq \mu_{female}$ OR $\mu_{male} - \mu_{female} \neq 0$

4. Identify the appropriate statistical test for these hypotheses and when to use the test.

Insert Your Answer in the Following Box

Statistical test for these hypotheses
Appropriate statistical test: Two-Sample Independent *t* test **When to use:** This test is used to assess whether the population means of two independent groups are statistically different from each other. In other words, it allows researchers to evaluate the mean difference between two populations using the data from two different samples. The population standard deviation is unknown.

5. Verify that the assumptions for using that test are met.

Insert Your Answer in the Following Box

Assumptions and Verification for the test
i. The data in each group should be independent of each other: *The data in each group were independent of each other.* ii. The t-test assumes that the populations from which the samples are drawn are normally distributed (Normality can be tested with the histogram, boxplot, comparing mean and median, etc.) or the sample size is sufficiently large (typically n > 30) for the Central Limit Theorem to apply: *The populations are normally distributed, the samples in each group are normally distributed and the sample size is sufficiently larger than 30.* iii. The data should be at the interval or ratio scale. The t-test is not appropriate for nominal or ordinal data. *The DV is (continuous) quantitative and IV is categorical*. iv. The samples should be randomly selected to ensure the results generalise to the larger population: *The samples are randomly selected.* v. The variances of the two groups should be homogeneous (must be equal). *This is tested with an additional test – F test or Levene's test.*

Is any additional test required? If yes, name and conduct the test. Yes. The F-test.

If the assumptions for the equal variances are not sure, the F-test can be used to determine if there is a significant difference between the variances of the groups.

This test is appropriate when the following observations are met:

- *Independence of observations*.
- *Variables: Continuous.*
- *Normality:* Both populations are normally distributed.
- *Sample size*. Sample size should be sufficiently large.

Null Hypothesis: Variances are the same.

$$\sigma^2_{male} = \sigma^2_{female}$$

Alternative Hypothesis: Variances are different.

$$\sigma^2_{male} \neq \sigma^2_{female}$$

F test two samples for variances		
Descriptive statistics		
	Male	Female
Sample size	177	154
Mean	3.09	3.24
Variance	0.16	0.15
P value	0.3323	

Summary and Decision: The F-test of equality of variance provided evidence that the variance in GPA for male and female were statistically non-significant, p = 0.3323. Since the P-value is greater than the significance level of 0.05, we fail to reject the null hypothesis. Therefore, we conduct the test assuming equal variances.

6. Conduct the hypothesis test and report your conclusion at the alpha = 0.05 significance level.

Insert Your Answer in the Following Box

Report the well-tidied Excel outputs (Using Analysis ToolPak) here:

Equal variances assumed or Equal variances not assumed? Equal Variances assumed.

t-test:

t-Test: Two-Sample Assuming Equal Variances		
	Male	Female

Mean	3.09	3.24
Variance	0.16	0.15
Observations	177	154
Pooled Variance	0.15	
Hypothesised Mean Difference	0	
df	329	
t Stat	-3.44	
P(T<=t) two-tail	6.62E-04	
t Critical two-tail	1.97	

Your conclusion based on the outputs: This summary shows that there is sufficient evidence to reject the null hypothesis that there is no difference in mean GPA between male and female because the p = 0.0007 (two-tailed) is less than the significance level of 0.05. Consequently, the arithmetic difference between the two means is statistically significant and cannot be attributed to chance.

7. **Practical significance (Cohen's d):** What is the effect size of the finding?

Insert Your Answer in the Following Box

Effect size = 0.38
An effect size (Cohen's d) of 0.38 falls within the range typically considered as having a small effect size. In practical terms, this suggests a small difference in mean GPA between male and female students surveyed.

8. Report and interpret a 95% confidence interval for the difference in the mean *GPA* by *Gender*.

Insert Your Answer in the Following Box

95% CI Lower Bound $= -0.23$
95% CI Upper Bound $= -0.06$

Conclusion: The 95% confidence interval of the difference with equal variances assumed is $[-0.23, -0.06]$. These intervals represent the estimated range of values that is 95% likely to include the population difference in means. As the interval does not contain zero (the hypothesised mean, $\boldsymbol{\mu_{male} - \mu_{female} = 0}$), it can be concluded that the result is statistically significant. Therefore, the average GPA of male is significantly less than the average GPA of female. This is the same conclusion reached previously that we reject null hypothesis.

9. **Summarise the test results:** As a minimum, the following information should be reported in the results section of any research report when using the independent two sample t-test:

- null hypothesis that is being evaluated,
- descriptive statistics (e.g., mean, SD, n, n_1, n_2),
- statistical test used (i.e., independent t-test),

- results of evaluation of test assumptions, and
- test results.

Insert Your Answer in the Following Box

An Independent t-test (equal variances assumed) was conducted to evaluate the null hypothesis that there is no difference in average GPA between male and female students surveyed (n = 331). Test results provided evidence that the difference between male (mean = 3.09, SD = 0.40, n1 = 177) and female (mean = 3.24, SD = 0.39, n2 = 154) was statistically significant, $t(df = 329) = -3.44$ critical t-value, p = 0.0007 (2-tailed). Therefore, there is sufficient evidence to reject the null hypothesis. There is sufficient evidence that mean GPA among the students surveyed is different between the two genders.

Research Question (b): Using the entire dataset, is there any difference between the average Pulse of a smoker and non-smoker among the students surveyed?

10. List the variable(s) involved and the type(s) (Qualitative or Quantitative?).

Insert Your Answer in the Following Box

How many variables are involved?

Variable(s) involved: Pulse and Smoke

Types of variables: Pulse – Quantitative discrete variable.

Smoke – Categorical variable with two categories (Yes and No).

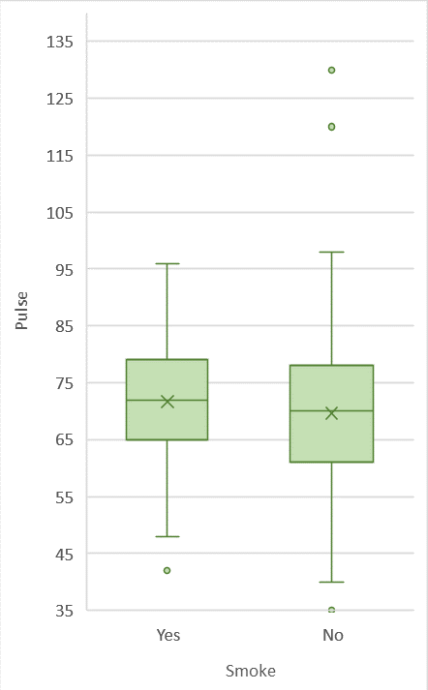
11. Determine the experimental design in this scenario. Is this a matched pairs design or are we comparing two independent means?

Insert Your Answer in the Following Box

Two Independent means

12. Report Comparative Analysis from the Sample Data using either the side-by-side boxplots or bar charts with error bars. Comment on the chart.

Paste Your Chart Here. Make sure the axes are well labeled.



Insert Your Comment in the Following Box

Both distributions are approximately symmetric.
Mean pulse rate (71.66) of smokers is slightly higher than those who did not smoke (69.61), The variation amongst the non-smokers is slightly higher than smokers. There is one outlier each in the lower distributions of both groups while there are two outliers in the upper distribution of the non-smokers.

13. Identify the null and alternative hypotheses.

Insert Your Answer in the Following Box

Null Hypothesis (H_0)
There is no difference between the average Pulse of a smoker and a non-smoker among the students surveyed. $\mu_{Smoker} = \mu_{Non-smoker}$ OR $\mu_{Smoker} - \mu_{Non-smoker} = 0$
Alternative Hypothesis (H_1)
The average Pulse of a smoker and a non-smoker among the students surveyed is different. $\mu_{Smoker} \neq \mu_{Non-smoker}$ OR $\mu_{Smoker} - \mu_{Non-smoker} \neq 0$

14. Identify the appropriate statistical test for these hypotheses and when to use the test.

Insert Your Answer in the Following Box

Statistical test for these hypotheses
Appropriate statistical test: Two independent-samples *t*-test **When to use:** This test is used to determine whether the sample mean difference obtained in a research study indicates a real mean difference between the two populations (or treatments) or whether the obtained difference is simply the result of sampling error. The population standard deviation is unknown.

15. Verify that the assumptions for using that test are met.

Insert Your Answer in the Following Box

Assumptions and Verification for the test
i. The data in each group should be independent of each other: *The data in each group were independent of each other.* ii. The t-test assumes that the populations from which the samples are drawn are normally distributed (Normality can be tested with the histogram, boxplot, comparing mean and median, etc.) or the sample size is sufficiently large (typically n > 30) for the Central Limit Theorem to apply: *The populations are normally distributed, the samples in each group are normally distributed and the sample size is sufficiently larger than 30.* iii. The data should be at the interval or ratio scale. The t-test is not appropriate for nominal or ordinal data. *The DV is (continuous) quantitative and IV is categorical*. iv. The samples should be randomly selected to ensure the results generalize to the larger population: *The samples are randomly selected.* v. The variances of the two groups should be homogeneous (must be equal). *This is tested with an additional test – F test or Levene's test.*
Is any additional test required? If yes, name and conduct the test: Yes. The F-test. If the assumptions for the equal variances are not sure, the F-test can be used to determine if there is a significant difference between the variances of the groups. This test is appropriate when the following observations are met: • *Independence of observations*. • *Variables: Continuous.* • *Normality:* Both populations are normally distributed. • *Sample size*. Sample size should be sufficiently large.

Null Hypothesis: Variances are the same.

$$\sigma^2_{Smoker} = \sigma^2_{Non-smoker}$$

Alternative Hypothesis:

$$\sigma^2_{Smoker} \neq \sigma^2_{Non-smoker}$$

F test two samples for variances		
Descriptive statistics		
	Smoker	**Non-smoker**
Sample size	41	290
Mean	71.66	69.61
Variance	131.38	147.77
P value	0.34	

Summary and Decision: The F-test of equality of variance provided evidence that the variance in the Pulse of Smoking and Non-smoking students were statistically equal, p = 0.34. Since the P-value is greater than the significance level of 0.05, we fail to reject the null hypothesis that the variances in the two smoking categories are the same. Therefore, we conduct the test assuming equal variances.

16. Conduct the hypothesis test and report your conclusion at the alpha = 0.05 significance level.

Insert Your Answer in the Following Box

Report the well-tidied Excel outputs (Using Analysis ToolPak) here:

Equal variances assumed or Equal variances not assumed? Equal variances assumed.

t-test: Two-sample assuming equal variances:

	Smoker	Non-smoker
Mean	71.66	69.61
Variance	131.38	147.77
Observations	41	290
Pooled Variance	145.77	
Hypothesised Mean Difference	0	
df	329	
t Stat	1.02	
P(T<=t) two-tail	0.31	
t Critical two-tail	1.97	

Your conclusion based on the outputs:

This summary shows that there is sufficient evidence to fail to reject the null hypothesis that there is no difference in average Pulse of smoking and non-smoking students because the p = 0.31 (two-tailed) is greater than the significance level of 0.05. Consequently, the arithmetic difference between the two means is statistically non-significant and can be attributed to chance.

17. **Practical significance (Cohen's d):** What is the effect size of the finding?

Insert Your Answer in the Following Box

Effect size = 0.17
An effect size (Cohen's d) of 0.17 falls within the range typically considered as having a trial effect size. In practical terms, this suggests a trivial difference in mean pulse rate between smoking and non-smoking students surveyed.

18. Report and interpret a 95% confidence interval for the difference in the mean *Pulse* by *Smoke*.

Insert Your Answer in the Following Box

95% CI Lower Bound = -1.92
95% CI Upper Bound = 6.01
Conclusion: The 95% confidence interval of the difference with equal variances assumed is [-1.92, 6.01]. These intervals represent the estimated range of values that is 95% likely to include the population difference in means. As the interval contains zero (the hypothesised mean, ($\boldsymbol{\mu_{Smoker} - \mu_{Non-smoker} \neq 0}$), it can be concluded that the result is statistically non-significant. Therefore, the average Pulse of smoking students is not significantly different from the average Pulse of non-smoking students. This is the same conclusion reached previously that we fail to reject H_0 that their means are the same.

19. **Summarise the test results:** As a minimum, the following information should be reported in the results section of any research report when using the independent two sample t-test:

- null hypothesis that is being evaluated,
- descriptive statistics (e.g., mean, SD, n, n_1, n_2),
- statistical test used (i.e., independent t-test),
- results of evaluation of test assumptions, and
- test results.

Insert Your Answer in the Following Box

An Independent t-test (equal variances assumed) was conducted to evaluate the null hypothesis that there is no difference in the average pulse rate between smoking students and non-smoking students that were surveyed (n = 331). Test results provided evidence that the difference between smokers (mean = 71.66, SD = 11.46, n1 = 41) and female (mean = 3.24, SD = 12.16, n2 = 290) was statistically non-significant, t(df = 320) = 1.02 critical t-value, p = 0.34 (2-tailed). Therefore, there is sufficient evidence not to reject the null hypothesis. There is sufficient evidence that mean pulse rate among the students surveyed is not different between students who smoked and those that did not smoke.

PART H: COMPARING THREE OR MORE POPULATION MEANS (ANOVA)

Research Question (a): Is the average height of students different by their Year of study?

1. List the variable(s) involved and the type(s) (Qualitative or Quantitative?).

Insert Your Answer in the Following Box

How many variables are involved? Two variables are involved.

Variable(s) involved: Height and Year

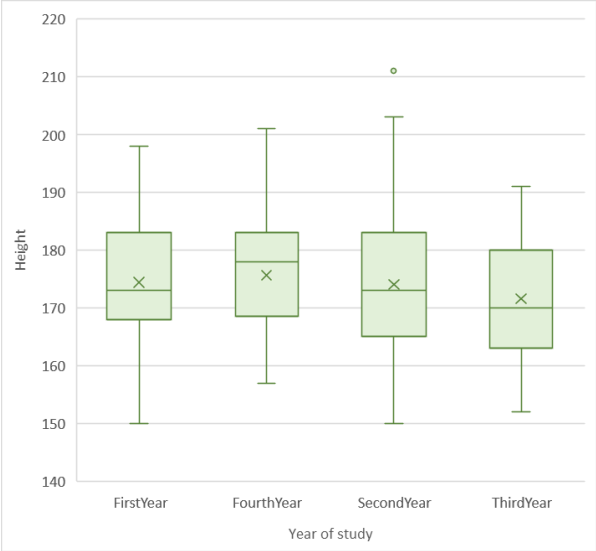
Types of variables:

Height: Quantitative (continuous) variable – *DV*

Year: Categorical variable with four categories: FirstYear, SecondYear, ThirdYear and FourthYear – *IV*

2. Report Comparative Analysis from the Sample Data using either the side-by-side boxplots or bar charts with error bars. Comment on the chart.

Paste Your Chart Here. Make sure the axes are well labeled.



Insert Your Comment in the Following Box

All the distributions except that of the FourthYear are slightly positively skewed.
FourYear has the highest mean height while ThirdYear has the least.
The variations among all the groups are nearly the same.
Only one outlier is present the distribution of the SecondYear group.

3. Identify the null and alternative hypotheses.

Insert Your Answer in the Following Box

Null Hypothesis (H_0)
There is no difference in mean Height across the 4 Years of study or there is no difference between the population height means of the four groups, $\mu_1 = \mu_2 = \mu_3 = \mu_4$.
Alternative Hypothesis (H_1)
At least one of the Year's means is not the same as the others.

4. Identify the appropriate statistical test for these hypotheses and when to use the test.

Insert Your Answer in the Following Box

Statistical test for these hypotheses
Appropriate statistical test: One-way Analysis of Variance (ANOVA) **When to use:** Only use the one-way between subjects ANOVA to analyse a continuous (interval or ratio scale) dependent variable when the purpose of the test is to determine if there is a difference in three or more independent groups in only one independent variable.

5. Verify that the assumptions for using that test are met.

Insert Your Answer in the Following Box

Assumptions and Verification for the test
i. Each group should be approximately normally distributed. This assumption is more critical with smaller sample sizes: Some deviation from normality is acceptable unless the populations are highly skewed. *Each group is not highly skewed.* ii. The variances within each group should be roughly equal (Homogeneity of variance). This can require additional tests such as Levene's or Bartlett's test of equality of variances. It has been suggested that violations of the ANOVA homogeneity of variance assumption have negligible consequences on the accuracy of the probability statements when the n's are equal: *The variance within each group is not too far apart and also for sufficiently large samples, the equality of variances assumption is not required.* *When sample sizes are relatively large and approximately equal in size, ANOVA test is fairly robust to violations of the assumptions of normality and homogeneity of variance.* iii. Observations within each group should be independent of each other: *Observations within each group were independent of each other.* iv. Data points should be randomly and independently selected from the population: *Data points were randomly and independently selected from the population.* *v.* The DV should be measured on an interval or ratio scale: *The DV was measured as a quantitative variable.*

6. Conduct the hypothesis test and report your conclusion at the alpha = 0.05 significance level.

Insert Your Answer in the Following Box

Report the well-tidied Excel outputs (Using Analysis ToolPak) here:

Groups	*Count*	*Sum*	*Average*	*Variance*
FirstYear	79	13778	174.41	101.40
SecondYear	183	31847	174.03	111.37
ThirdYear	33	5663	171.61	131.62
FourthYear	36	6323	175.64	86.52
		F	*P-value*	*F crit*
		0.92	0.43	2.63

Your conclusion based on the outputs: The above summary shows that the ANOVA is not significant since the p-level is greater than 0.05 (the assumed à priori significance level). Since the ANOVA is non-significant, post hoc multiple comparison tests are not required.

7. Is there a need to conduct a post hoc test? If no, why? If yes, conduct the test and report your conclusion.

Insert Your Answer in the Following Box

The is no need to conduct a post hoc test and we failed to reject the null hypothesis.

8. **Summarise the test results:** As a minimum, the following information should be reported in the results section of any research report when using One way ANOVA test:
    - null hypothesis that is being evaluated,
    - descriptive statistics (e.g., mean, SD, n, n_1, n_2, n_3, n_4...),
    - statistical test used (i.e., one-way between subjects ANOVA),
    - results of evaluation of ANOVA assumptions, and
    - ANOVA results.

Insert Your Answer in the Following Box

One-way analysis of variance (ANOVA) was conducted to compare means of heights between different years of study among the students surveyed: FirstYear, SecondYear, ThirdYear and FourthYear. The mean (± standard deviation) Heights for the FirstYear, SecondYear, ThirdYear and FourthYear were 174.41 ± 10.07, 174.03 ± 10.55, 171.61 ± 11.47 and 175.64 ± 9.30 respectively. The ANOVA was non-significant, F (3, 327) = 0.92; p = 0.43. Consequently, there was sufficient evidence not to reject the null hypothesis of no difference in mean heights across the 4 years of study. Post hoc multiple comparison *t*-tests were not conducted to compare each of the groups. An alpha level of 0.05 was used for all tests of statistical significance. The sample size of the FirstYear, SecondYear, ThirdYear and FourthYear groups were 79, 183, 33 and 36 respectively.

Research Question (b): Is the average students' GPA different for the Preferred Award students would like to win?

9. List the variable(s) involved and the type(s) (Qualitative or Quantitative?).

Insert Your Answer in the Following Box

How many variables are involved? Two variables are involved.

Variable(s) involved: GPA and Preferred Awards

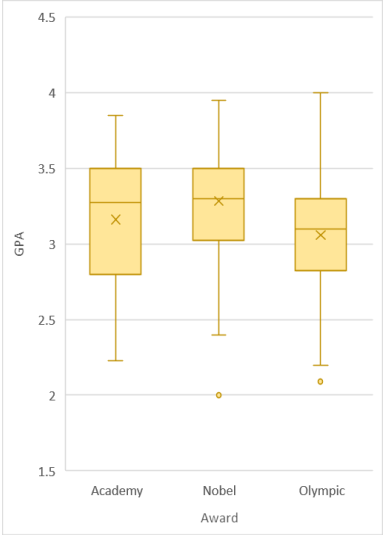
Types of variables:

GPA: Quantitative (continuous) variable – *DV*

Awards: Categorical variable with three categories: Academy, Nobel and Olympic – *IV*

10. Report Comparative Analysis from the Sample Data using either the side-by-side boxplots or bar charts with error bars. Comment on the chart.

Paste Your Chart Here. Make sure the axes are well labeled.



Insert Your Comment in the Following Box

All the distributions are slightly negatively skewed.
Those who preferred the Nobel award has the highest mean GPA while Olympic has the least.
There is much variation in Academy award compared to those of Noble and Olympic.
Variations among all the Nobel and Olympic are nearly the same.
One outlier is present for both Nobel and Olympic distributions.

11. Identify the null and alternative hypotheses.

Insert Your Answer in the Following Box

Null Hypothesis (H_0)
There is no difference in the average student's GPA based on the Preferred Award or there is no difference between the GPA of students based on the Preferred Award, $\mu_1 = \mu_2 = \mu_3$.
Alternative Hypothesis (H_1)
The average GPA based on at least one of the Preferred Award is not same as the others.

12. Identify the appropriate statistical test for these hypotheses and when to use the test.

Insert Your Answer in the Following Box

Statistical test for these hypotheses
Appropriate statistical test: One-way Analysis of Variance (ANOVA) **When to use:** Only use the one-way between subjects ANOVA to analyse a continuous (interval or ratio scale) dependent variable when the purpose of the test is to determine if there is a difference in three or more independent groups in only one independent variable.

13. Verify that the assumptions for using that test are met.

Insert Your Answer in the Following Box

Assumptions and Verification for the test

i. Each group should be approximately normally distributed. This assumption is more critical with smaller sample sizes: Some deviation from normality is acceptable unless the populations are highly skewed. *Each group is not highly skewed.*

ii. The variances within each group should be roughly equal (Homogeneity of variance). This can require additional tests such as Levene's or Bartlett's test of equality of variances. It has been suggested that violations of the ANOVA homogeneity of variance assumption have negligible consequences on the accuracy of the probability statements when the n's are equal: *The variance within each group is not too far apart and also for sufficiently large samples, the equality of variances assumption is not required.*

When sample sizes are relatively large and approximately equal in size, ANOVA test is fairly robust to violations of the assumptions of normality and homogeneity of variance.

iii. Observations within each group should be independent of each other: *Observations within each group were independent of each other.*

iv. Data points should be randomly and independently selected from the population: *Data points were randomly and independently selected from the population.*

v. The DV should be measured on an interval or ratio scale: *The DV was measured as a quantitative variable.*

14. Conduct the hypothesis test and report your conclusion at the alpha = 0.05 significance level.

Insert Your Answer in the Following Box

Report the well-tidied Excel outputs (Using Analysis ToolPak) here:

Groups	*Count*	*Sum*	*Average*	*Variance*
Academy	30	94.86	3.16	0.18
Nobel	133	436.68	3.28	0.14
Olympic	168	513.96	3.06	0.15

F	*P-value*	*F crit*
12.55	0.000006	3.02

Your conclusion based on the outputs: The above summary shows that the ANOVA is significant since the p-level is less than 0.05 (the assumed à priori significance level). Since the ANOVA is significant, post hoc multiple comparison tests are required to identify pairwise differences.

15. Is there a need to conduct a post hoc test? If no, why? If yes, conduct the test and report your conclusion.

Insert Your Answer in the Following Box

Yes, there is a need to conduct a post hoc test.

Comparison among groups					
t-test among pairs of means					
Group vs Group	Differences	95% CI Lower Bound	95% CI Upper Bound	Cohen's d value	P-value
Academy vs Nobel	-0.12	-0.27	0.03	-0.32	0.12
Academy vs Olympic	0.10	-0.05	0.26	0.26	0.19
Nobel vs Olympic	0.22	0.14	0.31	0.59	0.0000007

Conclusion: Post hoc multiple comparison t-tests revealed only one pairwise significant difference: Nobel vs Olympic.

16. **Summarise the test results:** As a minimum, the following information should be reported in the results section of any research report when using One way ANOVA test:

- null hypothesis that is being evaluated,

- descriptive statistics (e.g., mean, SD, n, n_1, n_2, n_3, $n_{4...}$),
- statistical test used (i.e., one-way between subjects ANOVA),
- results of evaluation of ANOVA assumptions, and
- ANOVA results.

Insert Your Answer in the Following Box

One-way analysis of variance (ANOVA) was conducted to compare means of GPA between different awards students surveyed preferred: Academy, Nobel and Olympic. The mean (± standard deviation) GPA for the Academy, Nobel and Olympic preferred awards were 3.16 ± 0.42, 3.28 ± 0.37 and 3.06 ± 0.39 respectively. The ANOVA was significant, F (2, 328) = 12.55; p = 0.000006. Consequently, there was sufficient evidence to reject the null hypothesis of no difference in mean GPA across the 3 preferred awards. Post hoc multiple comparison *t*-tests were conducted to compare each of the groups. The tests revealed one pairwise significant difference: Nobel vs Olympic. An alpha level of 0.05 was used for all tests of statistical significance. The sample size of the Academy, Nobel and Olympic groups were 30, 133 and 168 respectively.

PART I: CHI SQUARE TEST

Research question: Does preferred award depend on gender?

1. List the variable(s) involved and the type(s) (Qualitative or Quantitative?).

Insert Your Answer in the Following Box

How many variables are involved? Two variables.

Variable(s) involved:

Preferred Award: Categorical variable with three categories: Academy, Nobel and Olympic – *DV*

Gender: Categorical variable with two categories: Male and Female – *IV*

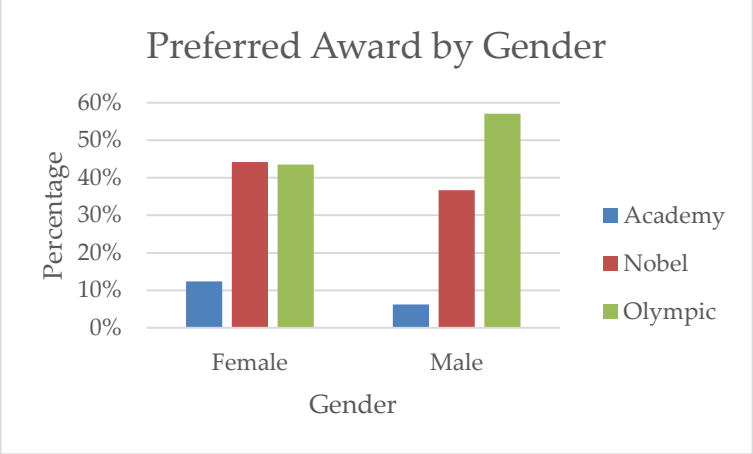
2. Report the two-way table and its corresponding clustered column chart. (Remember that this has been done previously). Comment on the chart.

Paste Your Tables Here. Make sure the table is well labeled.

Awards	Academy	Nobel	Olympic	Total
Female	19	68	67	154
Male	11	65	101	177
Total	30	133	168	331

Awards	Academy	Nobel	Olympic	Total
Female	12%	44%	44%	100%
Male	6%	37%	57%	100%
Total	9%	40%	51%	100%

Paste Your Chart Here. Make sure the axes are well labeled and the chart is titled.



Insert Your Comment in the Following Box

There appears to be an association between preferred award and gender. Olympic awards are most preferred by male students (57%). For female students, Nobel and Olympic awards had a close preference (44%). Academy awards had the least preferences in both genders, with double the number of females (12%) preferring the academy award compared to their male counterparts.

3. Identify the null and alternative hypotheses.

Insert Your Answer in the Following Box

Null Hypothesis (H_0)
There is no dependence between the preferred award and the gender. **OR** There is no association or relationship between the preferred award and gender.
Alternative Hypothesis (H_1)
The preferred award depends on gender. **OR** There is association or relationship between the preferred award and gender.

4. Identify the appropriate statistical test for these hypotheses and when to use the test.

Insert Your Answer in the Following Box

Statistical test for these hypotheses
Appropriate statistical test: Chi-square test for independence (or no association). **When to use:** Only use chi-square contingency table analysis to determine if a relationship exists between two categorical (nominal scale) variables. If a relationship exists, this test will not indicate strength of relationship.

5. Verify that the assumptions for using that test are met.

Insert Your Answer in the Following Box

Assumptions and Verification for the test
i. The observations in the sample must be independent: *The observations in the sample are independent.* ii. The data should be collected through random sampling to ensure that the sample is representative of the population: *The data were collected through random sampling to ensure that the sample was representative of the population.* iii. The data must be in the form of frequencies or counts (not percentages) for different categories: *The data are in the form of frequencies or counts for different categories.* iv. The expected frequency for each cell in the contingency table should be at least 5: *The contingency table's expected frequency for each cell is more than 5 as shown in the following table.*

Expected Data				
Female	13.96	61.88	78.16	154.00
Male	16.04	71.12	89.84	177.00
Total	30.00	133.00	168.00	331.00

6. Conduct the hypothesis test and report your conclusion at the alpha = 0.05 significance level.

Insert Your Answer in the Following Box

Report the well-tidied Excel outputs here:

Observed Values				
Gender/Awards	*Academy*	*Nobel*	*Olympic*	*Total*
Female	19.00	68.00	67.00	154.00
Male	11.00	65.00	101.00	177.00
Total	30.00	133.00	168.00	331.00
Expected Data				
Female	13.96	61.88	78.16	154.00
Male	16.04	71.12	89.84	177.00
Total	30.00	133.00	168.00	331.00

f_o	f_e	f_o-f_e	$(f_o-f_e)^2$	$(f_o-f_e)^2/f_e$
19.000	13.960	5.040	25.402	1.820
68.000	61.880	6.120	37.454	0.605
67.000	78.160	-11.160	124.546	1.593
11.000	16.040	-5.040	25.402	1.584
65.000	71.120	-6.120	37.454	0.527
101.000	89.840	-11.160	250.258	6.129
331.000	331.000	-22.320	500.515	12.257

Results		Remarks
df	2	*df = (rows-1)*(columns-1)*
Critical Value	5.991	*Critical Value < Test Statistics: Reject the hypothesis*
Chi-Square test statistic	12.257	
p-value	0.023	*p-value < alpha (0.05): Reject the hypothesis*

Your conclusion based on the outputs: The calculated test statistic is 12.257 is greater than the critical value of 5.991 so we reject H_0. There is sufficient evidence, at the $\alpha = 0.05$ level of significance, to conclude that Preferred Award and Gender are related.

7. **Summarise the test results:** As a minimum, the following information should be reported in the results section of any research report when using Chi square test:

    - null hypothesis that is being evaluated,

- descriptive statistics (e.g., observed frequency counts, proportion, n),
- statistical test used (i.e., Pearson χ^2 contingency table analysis or χ^2 goodness-of-fit Test),
- results of evaluation of test assumptions if violated, and
- χ^2 goodness-of-fit or χ^2 contingency table analysis results.

Insert Your Answer in the Following Box

The chi-square test for independence (no association) was used to evaluate the null hypothesis that there is no difference between preferred award and gender. The test showed a statistically significant difference between preferred award and gender, $\chi^2(2, n = 331) = 12.26$, $p = 0.023$. Since the p-value is less than the chosen significance level $\alpha = 0.05$, there is sufficient evidence to reject the null hypothesis and conclude that there is association between preferred award and gender.

PART J: LINEAR REGRESSION TEST

Research question: Do Weights of the students really depend on their Heights?

1. List the variable(s) involved and the type(s) (Qualitative or Quantitative?).

Insert Your Answer in the Following Box

How many variables are involved?

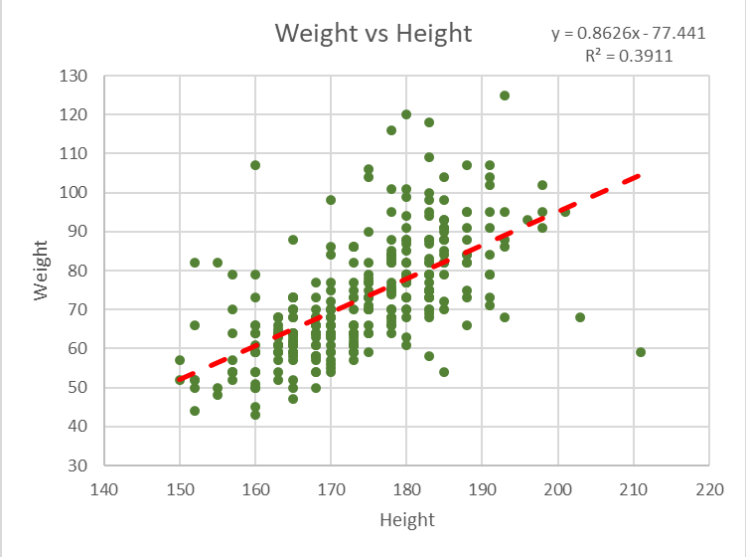
Two quantitative variables

Variable(s) involved: Weight (*DV*) and Height (*IV*)

Types of variables: Quantitative Continuous

2. Report the scatterplot containing regression line and equation. (Remember that this has been done previously). Comment on the chart.

Paste Your Chart Here. Make sure the axes are well labeled.



Insert Your Comment in the Following Box

Weight = 0.8626*(Height) -77.441

There is a weak, positive relationship between Weight and Height, indicating that as Height increases, Weight also increases. Height explained 39% of the variance in Weight (R^2 = 0.39). Some noticeable outlying data points.

3. Identify the null and alternative hypotheses.

Insert Your Answer in the Following Box

Null Hypothesis (H_0)
$y = a + bx$ $\boldsymbol{b = 0}$, i.e. the slope of the population regression model equals zero and there is no linear relationship between Weight and Height.
Alternative Hypothesis (H_1)
$\boldsymbol{b \neq 0}$, i.e. the slope of the population regression model is not equal zero and there is real linear relationship between Weight and Height.

4. Conduct the hypothesis test and report your conclusion at the alpha = 0.05 significance level.

Insert Your Answer in the Following Box

Report the well-tidied Excel outputs here:

	Coefficients	*t Stat*	*P-value*	*Individually Significant*
Intercept	-77.44079067	-7.48405391	6.64662E-13	Yes
Slope	0.862611337	14.53556081	2.55537E-37	Yes

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	26576.30994	26576.30994	211.282528	2.55537E-37
Residual	329	41383.47858	125.7856492		
Total	330	67959.78852			

Your conclusion based on the outputs: The p-value is less than the chosen significance level. So, we reject the null hypothesis and conclude that there is a real linear trend.

5. **Summarise the test results:** As a minimum, the following information should be reported in the results section of any research report when using Chi square test:

   - null hypothesis that is being evaluated,
   - correlation
   - statistical test used (i.e., bivariate regression Test),
   - results of evaluation of test assumptions if violated,
   - the amount of variance explained by the model (i.e., R^2),
   - regression equation and the standard error of the estimate and
   - the significance of the model.

Insert Your Answer in the Following Box

Simple linear regression is used to examine whether Weight depends on Height among the students surveyed. The statistical significance of the slope of the regression coefficient was examined using a t-test with an alpha level of 0.05. The coefficient of determination (R^2) was calculated to represent the percentage of variation in Weight, explained by Height.

Height is a significant predictor of Weight ($b = 0.86$; $p < 0.00001$) and explained 39% of the variance in Weight ($R^2 = 0.39$). The regression equation developed based on the analysis was:

Weight = 0.8626*(Height) -77.441

This material is part of our collection of resources that have been developed from over a decade of teaching introductory statistics and data analysis to undergraduate students at ATU Sligo and various Irish tertiary institutions, as well as providing Microsoft Excel classes and trainings as part of our Information Technology Modules.

Primarily aimed at undergraduate modules in science, health, business, and related disciplines that leverage Microsoft Excel for data analysis, our materials are intended to be comprehensive resources. Accompanying these resources is a dedicated repository at https://thewee.link/NTUTORR-PBL-Resources, offering a lot of useful resources for both students and instructors.

Dr. Akinlolu Akande is a Lecturer in Mathematics and Information Technology in the Faculty of Science at Atlantic Technological University (ATU) Sligo, Ireland. He is a Senior Fellow of the Higher Education Academy (SFHEA), a recognition of his expertise in teaching and learning in higher education.

Dr. Syam Kumar R. is a Lecturer in Mathematics and Information Technology in the Faculty of Science at ATU Sligo, Ireland. He is committed to training and teaching students in the faculty, providing valuable support to students and colleagues.

Dr. David Obada is a Postdoctoral Fellow at the ATU Sligo, Ireland. He was also a research and teaching fellow at the Massachusetts Institute of Technology (MIT) in the USA and holds a Kaufmann Teaching Certificate from MIT.

Oliscoil Teicneolaíochta an Atlantaigh
Atlantic Technological University
NTUTORR Transforming Learning