CHAPTER 2
DESCRIPTIVE STATISTICS

Learning Objectives :

1. *Understand statistics as a methodology that is concerned with formulating question, collecting data, analyzing data and interpreting results.*
2. *Use basic terminology in statistics such as random variable, population and sample.*
3. *Distinguish between the different types of data, such as qualitative and quantitative.*
4. *Construct various graphical displays of the data, and provide basic interpretations.*
5. *Compute numerical summaries of the data, and provide basic interpretations.*
6. *Analyse strength of relationship using scatter plot and correlation coefficient.*

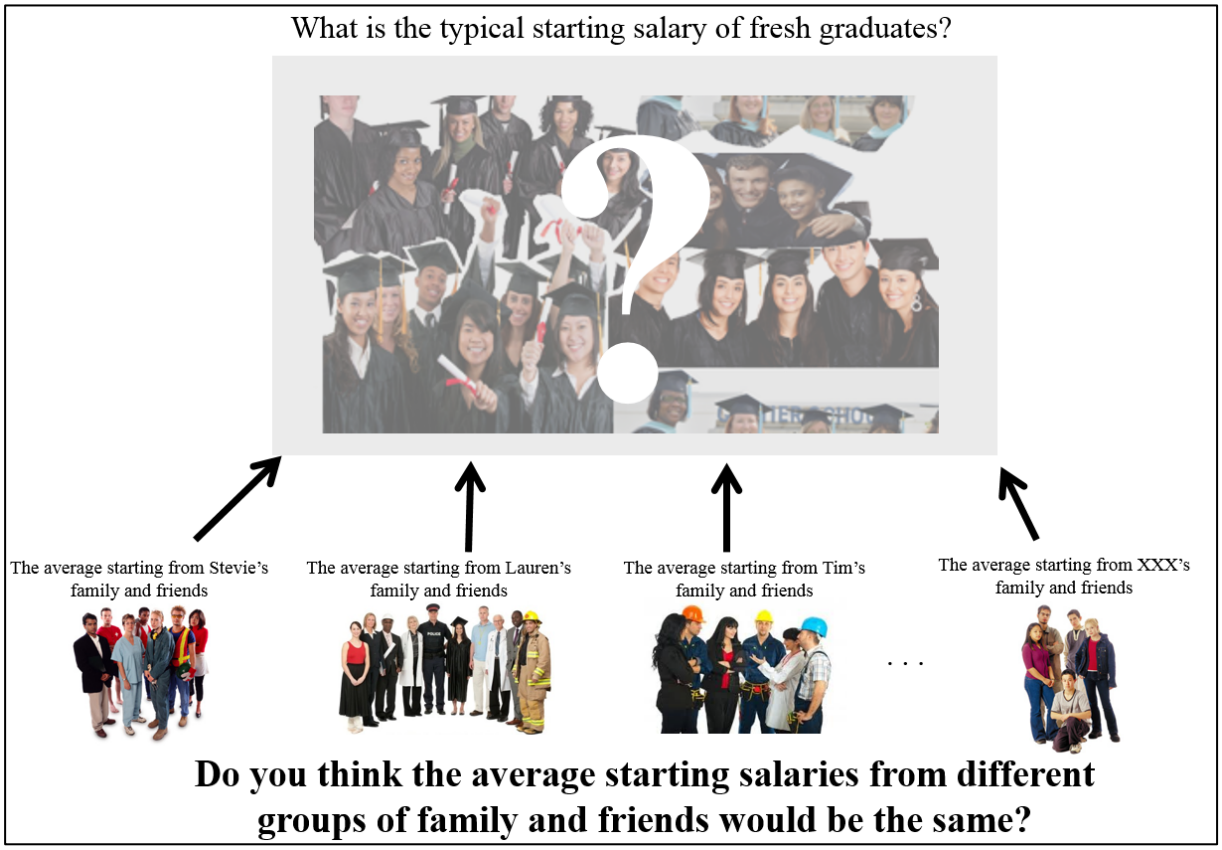
Content

1. Statistical Problem-Solving Process

Suppose we are interested to find out the *typical* starting salary of a fresh graduate, how would we go about finding an appropriate answer?

We could ask relatives and friends to share with us their starting salaries and then, perhaps we would take an average of their salaries. This average number, we presume, is possibly a typical starting salary based on the data we have collected.

Suppose now there are ten other people who are interested in the same question and they took samples of their own relatives and friends (for convenience, assume that none of these ten people are related and that they do not share the same relatives and friends), would you expect all of these ten people to arrive at the same typical starting salary as ours earlier?

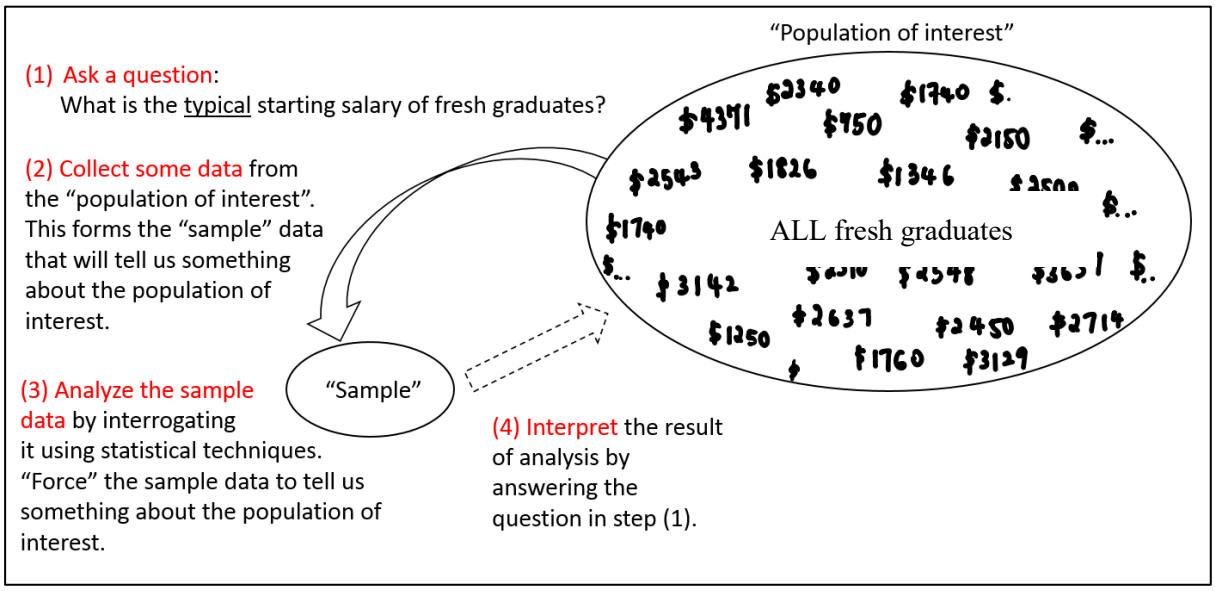


Well, it is highly likely the average salaries from different groups of family and friends would be different.

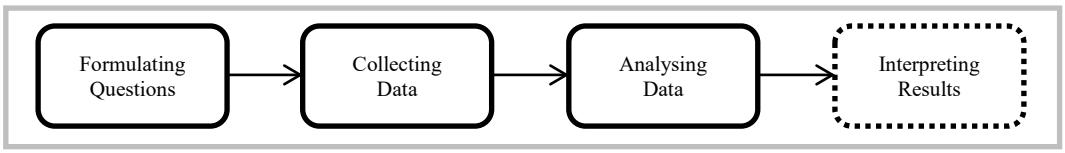
In studies where there are **variability**, we need a tool to help us capture this variability.
How do we capture the *variability* in different sample data sets and use it to make more sense of the data sets?

One tool we can use to study data – the typical value, the variability of data, and more – is **statistics**.

Essentially, this is what we are doing:



Steps (1) to (4) outlined the **statistical problem-solving process**, summarized here:

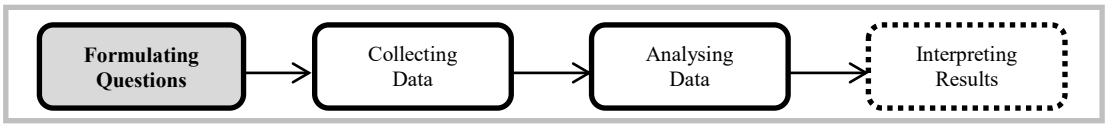


You may have learnt a lot of step (3) in school, so it is hoped that this chapter will value-add to your learning by teaching you the holistic statistical problem-solving process which always begins with a **question** of interest, then **collect** some data to help answer the question, **analyze** the data by using statistical techniques, then **interpret** the results to answer the question! ☺ Thus, the focus of this chapter is to take you through the statistical problem-solving process of steps (1) to (4) through a case study – Prestige Mall.

Remember, the main aim of this chapter is to give you an opportunity to experience statistics in a more holistic way, hence every step of the statistical problem-solving process is as important as the other steps. And to remind you of the step which you are at, the process can be found at the top of every page with the current step highlighted.

Exercise: Can you identify the population and sample in the scenarios below?

Scenario	Population	Sample
A new filtration system has been installed in the water systems of a small city. The amount of impurities (in parts per million) remaining in the filtered water is recorded over a 30-day period.	Filtered water of the city.	
To serve customers better by cutting queueing time at counters during peak period, ABC Bank recorded the queueing times (in minutes) of 20 customers.		The 20 customers whose queueing times were recorded.



2. Formulating Questions

Case Study: Prestige Mall

What businesses have you here?

One of the shops in this high-end shopping centre, *Prestige Mall*, has become vacant. You have always wanted to start your own business but do not have the capital. Hence, you have decided to write a business proposal to bring in potential investors.

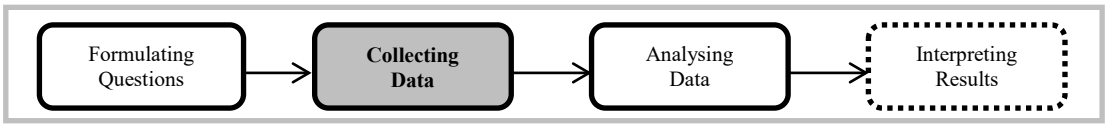
Some decisions that we make may be based on personal judgments but some may not. In this case, a proposal with support from data collected is obviously more convincing as compared to a proposal without such a support or a proposal supported with personal judgments alone. What are some of the information that you would include in your proposal, as support to the kind of business that you would like to have in Prestige Mall?

Well, for instance, you may like to give your investors an initial "feel" of the profile of customers who patronize this mall:

- What kind of job sectors are they from?
- How many times do they frequent Prestige Mall in a month?
- How much do they typically spend in Prestige Mall?
- What is their age profile?
- What is the proportion of male customers of Prestige Mall?
- What is their average monthly household income?

Think-out-loud…
So, what are some other questions that you might ask?

Investors, naturally, are interested in how much money can be made and why the customers are willing to pay for the product or service that you have to offer. So the questions listed above would generate valuable information in helping you to decide on the product or service that your business would offer.



3. Collecting Data

3.1 Sample Data

The previous section introduces a case study, specifically what kind of business opportunities there are in high-end Prestige Mall. The rest of this chapter would base its contents and discussions on the given case study, guided by the statistical problem-solving process, as indicated by the top of each page.

We have asked a few questions in the previous section and to answer those questions, we would need to collect data. There are many possible ways to collect data, such as from Prestige Mall's database, provided there is one. However, if interrogating databases is not a viable option, then another possible method is to conduct a survey. It is not uncommon to see people filling up survey forms in malls, of course after proper permission is sought. A possible survey form could look like this:

Prestige Mall Customer Survey

Dear Valued Customer,

Thank you for taking part in this survey. Your feedback will be valuable in helping us to enhance your shopping experience here in Prestige Mall. This survey will take approximately 5 minutes. All information shared with us will remain private and confidential.

---

(1) Which job sector are you in? □ IT/Engineering □ Business/Finance □ Others

(2) What is your gender? □ Male □ Female

(3) What is your age? _____

(4) How many times did you visit Prestige Mall in the past month? _____

(5) How much is your monthly household income? _____

(6) Approximately how much do you spend at Prestige Mall a month? _____

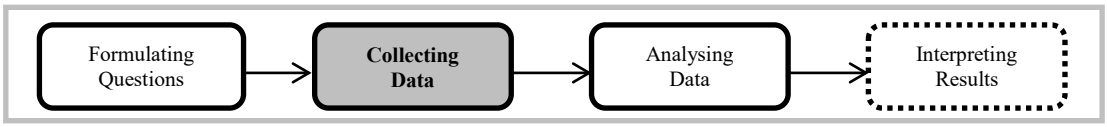
Kindly drop this survey form at the information counter and receive a token of appreciation.

Thank You!

Think-out-loud…
Why not **all** the customers?

In order to collect data, a selected group of customers of Prestige Mall is to be chosen to respond to the survey. In statistics, we are concerned with randomness and representativeness – how do we know that we have not been biased in selecting the respondents for the survey?

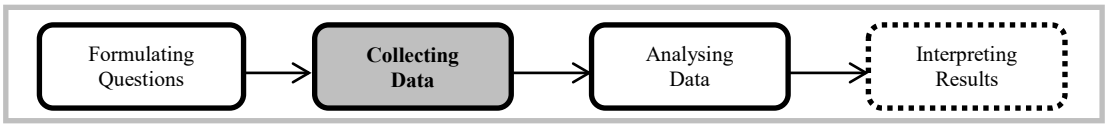
There are methods of sampling in statistics such as simple random sampling, systematic sampling, stratified sampling, and many more; but it is beyond the scope of this chapter to discuss sampling methods further. Hence we would make an assumption that all the customers who eventually are involved in the survey above are randomly selected (hence, not biased).



By the end of the survey period, suppose you have collected feedback from 200 customers. You then entered the data into an Excel spreadsheet as shown below; sorted first by gender, then by age.

No.	Job sector	Age	No. of visits per month	Gender	Household income (in $)	Amount spent per month (in $)
1	Bus/Fin	18	6	Female	8852.32	441.73
2	Bus/Fin	19	3	Female	7889.24	663.63
3	Bus/Fin	22	1	Female	6901.79	489.26
4	Bus/Fin	22	6	Female	8566.96	412.19
5	Bus/Fin	23	6	Female	7144.45	188.59
6	Bus/Fin	24	6	Female	8032.08	253.59
7	IT/Eng	25	5	Female	9405.84	421.35
8	Bus/Fin	25	3	Female	7694.62	538.59
9	Bus/Fin	25	3	Female	8727.36	489.49
10	Bus/Fin	25	5	Female	7723.2	514.52
11	Bus/Fin	26	3	Female	9598.72	528.6
12	Others	27	3	Female	6943.93	617.12
13	Others	27	4	Female	7628.78	387.54
14	Bus/Fin	28	6	Female	6584.36	607.96
15	Bus/Fin	28	4	Female	8018.32	395.89
16	Bus/Fin	28	2	Female	8532.96	445.36
17	Bus/Fin	29	4	Female	8000	430.04
18	Bus/Fin	30	6	Female	7940.7	529.17
19	IT/Eng	30	4	Female	7041.06	485.61
20	IT/Eng	30	4	Female	8008.96	549.98
21	IT/Eng	30	5	Female	9364.32	634.15
22	IT/Eng	31	4	Female	7801.48	249.33
23	Bus/Fin	32	6	Female	7686.12	393.87
24	Others	32	2	Female	6260.83	546.15
25	IT/Eng	33	2	Female	8072.96	462.99
26	Bus/Fin	34	3	Female	8921.52	695.09
27	Bus/Fin	34	5	Female	8976.72	537.01
28	IT/Eng	34	2	Female	7955.15	279.04
29	Bus/Fin	34	5	Female	10313.68	686.41
30	IT/Eng	35	2	Female	7380.6	500.96
31	IT/Eng	35	1	Female	9686.88	271.49
32	IT/Eng	35	3	Female	6945.16	419.45
33	IT/Eng	35	3	Female	10130	710.38
34	IT/Eng	35	1	Female	9454.8	519.65
35	IT/Eng	35	5	Female	6875.78	454.07
36	IT/Eng	36	1	Female	7940.91	749.45
37	IT/Eng	37	5	Female	9186.56	723.74
38	IT/Eng	37	5	Female	10041.2	568.12
39	IT/Eng	37	4	Female	9498.4	501.65
40	Others	39	1	Female	9218.24	518.06
41	Bus/Fin	39	3	Female	8553.44	577.6
42	IT/Eng	39	6	Female	7666.1	460
43	Others	39	1	Female	11288.16	471.31
44	Others	39	4	Female	7056.47	555.35
45	Bus/Fin	39	5	Female	10167.28	398.03
46	Bus/Fin	40	3	Female	8024.72	547.03
47	IT/Eng	40	6	Female	7081.9	536.28
48	IT/Eng	40	4	Female	9330.32	500.5
49	Bus/Fin	41	1	Female	8336.96	444.38
50	IT/Eng	41	4	Female	7603.2	419.53
51	Others	42	4	Female	5028.62	621.32
52	IT/Eng	42	3	Female	7378.14	466.73
53	IT/Eng	43	6	Female	8563.44	508.67
54	IT/Eng	43	2	Female	10868.96	296.47
55	IT/Eng	44	5	Female	8351.12	557.35
56	Others	44	1	Female	9864.56	310.47
57	Others	44	3	Female	7546.02	460.37
58	Others	44	6	Female	8315.12	130.03
59	Others	44	3	Female	7977.22	536.22
60	IT/Eng	44	4	Female	7640.8	606.88
61	IT/Eng	44	6	Female	9520	437.9
62	IT/Eng	45	6	Female	10125.68	633.5
63	Others	46	1	Female	7417.85	565.32
64	IT/Eng	46	4	Female	7405.18	345.18
65	IT/Eng	48	3	Female	6864.86	335.57
66	IT/Eng	48	4	Female	9822.88	523.66
67	Others	48	4	Female	7775.92	504.83
68	IT/Eng	49	6	Female	9200	305.35
69	IT/Eng	49	5	Female	9683.44	546.46
70	Others	50	1	Female	8175.84	545.24
71	Bus/Fin	50	5	Female	7151.78	623.88
72	Others	52	4	Female	12000	423.39
73	IT/Eng	52	6	Female	7570.6	397.94
74	IT/Eng	52	5	Female	8558.88	479.93
75	Bus/Fin	52	2	Female	7681.66	799.55
76	Bus/Fin	52	3	Female	7229.6	273.78
77	Others	52	5	Female	11091.04	438.76
78	Others	53	2	Female	9180.48	434.59
79	IT/Eng	53	3	Female	7735.74	501.85
80	IT/Eng	53	6	Female	8987.84	743.84
81	IT/Eng	53	4	Female	7340.88	460.21
82	Others	55	4	Female	7628.06	577.98
83	IT/Eng	55	3	Female	7301.79	480.58
84	IT/Eng	55	5	Female	9389.68	469.83
85	IT/Eng	55	1	Female	6739.38	303.77
86	IT/Eng	55	5	Female	7432.54	423.54
87	IT/Eng	57	6	Female	7804.9	572.23
88	IT/Eng	57	3	Female	6604.22	617.24
89	IT/Eng	57	1	Female	8014.4	386.85
90	IT/Eng	58	6	Female	7980.16	429.63
91	IT/Eng	58	1	Female	7912.84	421.49
92	Others	59	1	Female	9509.76	704.26
93	IT/Eng	59	6	Female	7341.42	530.9
94	IT/Eng	59	1	Female	7141.75	485.78
95	IT/Eng	60	4	Female	6775.33	572.48
96	IT/Eng	60	2	Female	8119.36	706.38
97	IT/Eng	60	2	Female	6147.52	384.43
98	Others	60	3	Female	9280.56	440.37
99	IT/Eng	60	1	Female	9950.08	572.09
100	Bus/Fin	60	5	Female	10594.08	439.01

No.	Job sector	Age	No. of visits per month	Gender	Household income (in $)	Amount spent per month (in $)
101	IT/Eng	61	1	Female	8937.2	604.36
102	IT/Eng	61	2	Female	6324.46	701.94
103	IT/Eng	61	1	Female	9111.84	406.34
104	IT/Eng	61	6	Female	6754.44	494.46
105	IT/Eng	61	3	Female	10053.84	493.23
106	Bus/Fin	61	1	Female	8960	583.01
107	IT/Eng	62	5	Female	8245.12	398.3
108	IT/Eng	62	5	Female	8577.28	595.97
109	IT/Eng	62	6	Female	8079.12	718.32
110	IT/Eng	62	6	Female	8575.68	448.86
111	Bus/Fin	20	3	Male	8078.08	615.07
112	Bus/Fin	22	4	Male	9205.28	718.42
113	Bus/Fin	23	6	Male	7703.13	557.55
114	Others	23	1	Male	6948.54	328.81
115	Bus/Fin	24	6	Male	10140.16	400.01
116	Bus/Fin	24	2	Male	8066.56	380.55
117	Bus/Fin	24	5	Male	7931.98	633.94
118	Bus/Fin	25	3	Male	7876.96	623.06
119	Bus/Fin	25	6	Male	8068.16	422.38
120	IT/Eng	27	5	Male	7538.61	458
121	Bus/Fin	27	5	Male	9517.76	714.08
122	IT/Eng	27	2	Male	8850.88	578.7
123	Bus/Fin	27	2	Male	6702.81	759.77
124	Bus/Fin	28	2	Male	9738.88	547.47
125	Bus/Fin	28	6	Male	8095.68	435.69
126	Bus/Fin	28	3	Male	7918.79	150.35
127	Bus/Fin	29	3	Male	6955.31	516.49
128	Bus/Fin	29	6	Male	8203.28	357.14
129	Bus/Fin	29	1	Male	8062.32	697.92
130	IT/Eng	29	6	Male	8370	498.49
131	IT/Eng	30	1	Male	5924.37	377.35
132	Bus/Fin	30	2	Male	7834.53	626.42
133	Others	31	5	Male	6065.54	708.58
134	IT/Eng	32	2	Male	8617.44	629.97
135	Bus/Fin	32	6	Male	11016.16	617.78
136	Others	33	4	Male	5882.7	565.15
137	IT/Eng	33	2	Male	7315.18	454.91
138	Bus/Fin	33	4	Male	9264.24	470.76
139	IT/Eng	33	4	Male	9554.32	757.38
140	Bus/Fin	34	2	Male	6645.55	106.67
141	Bus/Fin	34	6	Male	7828.06	714.31
142	IT/Eng	35	1	Male	7455	367.63
143	Bus/Fin	35	3	Male	8718.88	461.5
144	IT/Eng	35	4	Male	8168.56	358.09
145	IT/Eng	35	1	Male	8322.48	771.98
146	Bus/Fin	36	3	Male	7439.84	365.68
147	IT/Eng	36	4	Male	7603.98	505.87
148	IT/Eng	36	5	Male	6453.19	439.28
149	Bus/Fin	36	6	Male	10400	513.48
150	Bus/Fin	37	5	Male	7334.46	659.28
151	IT/Eng	37	1	Male	8349.68	303.19
152	IT/Eng	38	6	Male	7071.22	457.92
153	IT/Eng	38	1	Male	8602.64	591.45
154	Others	39	5	Male	7326.22	258.44
155	Others	40	5	Male	8346.56	606
156	IT/Eng	40	2	Male	8021.2	504.37
157	IT/Eng	40	5	Male	8092.08	689.16
158	IT/Eng	40	4	Male	9600	514.32
159	Others	41	2	Male	7272.61	270.5
160	Others	41	3	Male	7477.06	412.38
161	Bus/Fin	42	5	Male	8290.88	650.01
162	IT/Eng	43	5	Male	8746.64	544.08
163	Bus/Fin	43	2	Male	8697.12	508.22
164	Others	43	4	Male	10144	475.03
165	Others	44	5	Male	7223	766.51
166	IT/Eng	44	4	Male	10484.48	398.39
167	Others	44	1	Male	7433.19	419.39
168	Others	44	1	Male	8571.12	683.45
169	IT/Eng	44	6	Male	9293.76	573.89
170	IT/Eng	45	5	Male	9772	526
171	IT/Eng	45	1	Male	7750.62	564.51
172	Others	45	3	Male	8265.44	385.24
173	Bus/Fin	45	4	Male	6181.74	763.24
174	IT/Eng	45	4	Male	6980.31	562.92
175	IT/Eng	46	2	Male	12000	773.94
176	IT/Eng	48	6	Male	8324.96	431.06
177	Others	48	6	Male	8492.32	214.46
178	IT/Eng	48	3	Male	10422.72	655.75
179	IT/Eng	49	3	Male	8199.76	326.25
180	Others	50	6	Male	9836.88	539.39
181	IT/Eng	50	6	Male	7995.7	525.93
182	Others	51	4	Male	6206.42	665.62
183	Others	51	4	Male	6791.86	609.67
184	Bus/Fin	51	2	Male	8369.52	561.95
185	IT/Eng	51	6	Male	6914.22	400.74
186	IT/Eng	51	2	Male	6426.92	331.43
187	IT/Eng	52	2	Male	8118.24	513.57
188	Others	52	1	Male	8008.48	660.83
189	Bus/Fin	54	3	Male	8881.44	355.79
190	IT/Eng	54	4	Male	9967.12	314.15
191	IT/Eng	55	4	Male	10329.04	397.82
192	IT/Eng	57	2	Male	8080.88	483.82
193	IT/Eng	57	2	Male	7289.73	484.15
194	Others	59	5	Male	8378.32	52.98
195	IT/Eng	59	1	Male	8499.84	616.73
196	Bus/Fin	60	1	Male	6616.04	580.12
197	IT/Eng	60	2	Male	7268.83	441.53
198	IT/Eng	60	6	Male	7450.53	521.33
199	IT/Eng	61	5	Male	6089.39	526.55
200	IT/Eng	61	5	Male	8091.92	460.2

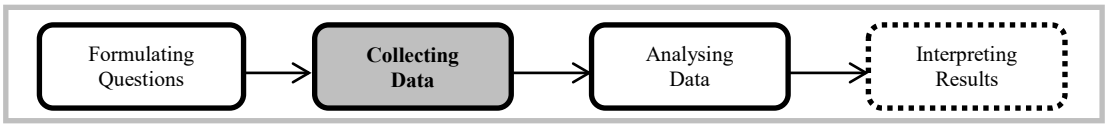


3.2 Common Statistical Terms

It is not uncommon to handle large amount of data in statistics. But with such large amount of data, what statistical techniques are there to churn these data into information?

Let us first define some terminologies in statistics, shown as follows, before we look at types of data in statistics:

Terminology	Definition	Example from the Case Study
Variable	A quantity that can be measured and may take on different values within a problem.	
Data	Observations or responses collected for the selected variable. (A single observation is called *datum*.)	
Population	The complete set of items which we are studying. This is usually too large for the collection of data.	
Population Size	The number of items in the population.	
Sample	A subset of items selected from the population. When the population is too large, a representative sample is usually selected instead.	
Sample Size	The number of items in the sample.	



3.3 Types of Data

There are mainly two types of data:

- **Qualitative** data are non-numerical values (or text) that are descriptive in nature. It is often used interchangeably with the term **categorical** data.
- **Quantitative** data take on values measured on a numerical scale.

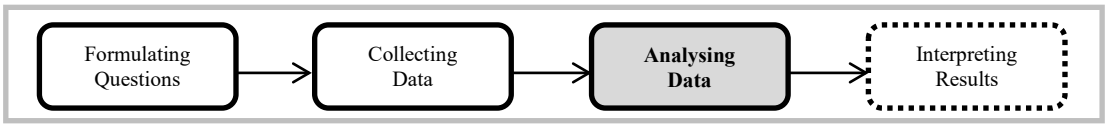
Qualitative data can be further classified by **nominal** or **ordinal** scale.
Nominal scale data are identified by names or labels only, whereas ordinal scale data can be ordered or ranked.

Quantitative data can be further classified into **discrete** or **continuous** data.
Discrete data can only take on certain values, whereas continuous data can take on any value within a range. We say that discrete data is counted, whereas continuous data is measured.

Example: Classify the type of data given below.

Data	*Qualitative/Quantitative?*	*Nominal/Ordinal/Discrete/Continuous?*
Exam grade		
Height		
Weight		
Number of children		
Blood type		
Colour of eyes		
Shoe size		
Number of heads in coin tosses		

CHECKPOINT #1. Think of one more example and classify the type of data.
#2. Sketch a map of the types of data.



4. Analysing Data

4.1 Summarizing Data

How do we describe a set of data? We can group them and present their pattern or distribution in a tabular or graphical form. We can also describe data by using a few well-chosen numbers that summarise meaningfully the entire data set. Hence, we can summarize the data, in two ways – by **graphical summary** and by **numerical summary**.

Nowadays, your calculators are equipped with statistical functions which enable almost effortless computations of the numerical summaries. Furthermore, software packages are able to produce sophisticated graphs easily. In this course, you will learn how to produce the numerical summaries and generate graphs using the statistical software **Minitab Express**. As such, the focus will be to learn how to interpret the summaries, rather than the "formulae" behind the summaries.

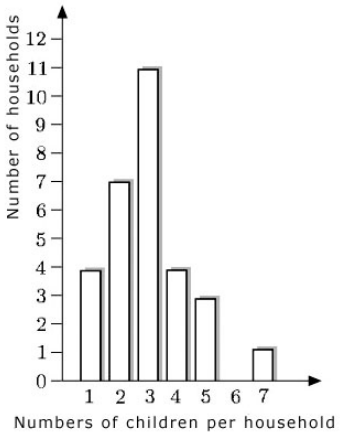
4.2 Graphical Summaries

An effective way to present a set of data to a team of decision makers is to use diagrams or graphs. Pattern exhibited by a variable and comparisons between variables become visual.

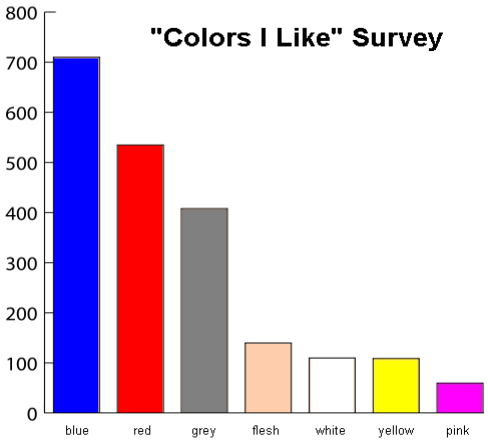
In this course, we will cover the more commonly used graphs – bar graph, pie chart, histogram and box plot.

- **Bar graph**

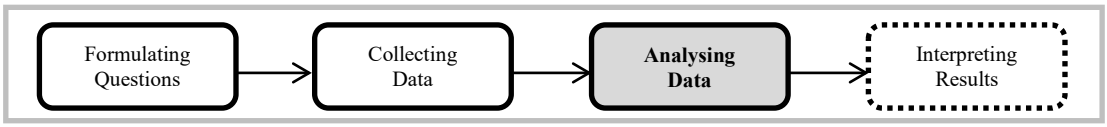
  Typically used to represent quantitative or qualitative data. It gives a visual overview of differences in frequency (or percentage) between categories.



An example bar graph of a quantitative variable (i.e. number of children per household)

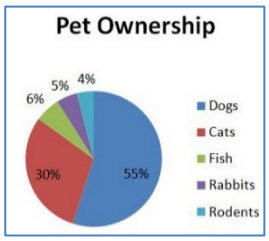


An example bar graph of a qualitative variable (i.e. favourite colour)



- **Pie chart**

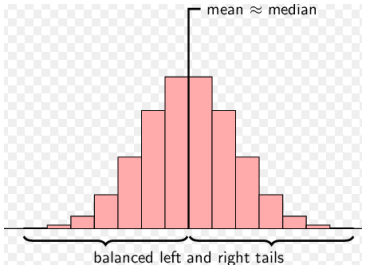
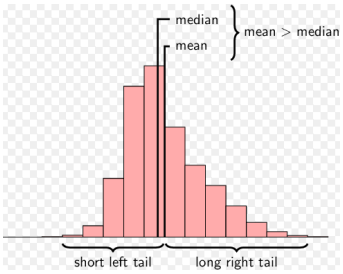
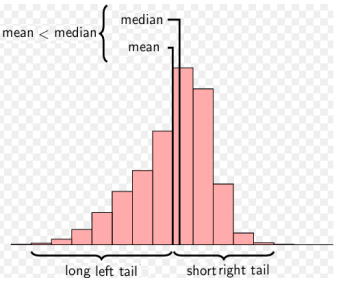
  Typically used to represent qualitative data. It gives a visual overview of proportions belonging to each category.



- **Histogram**

  A histogram displays frequencies of quantitative data that have been sorted into intervals. These give visual overview of the shape of distribution of the data values. Specifically, **skewness** is a measure of symmetry of the data distribution, or rather, asymmetry.

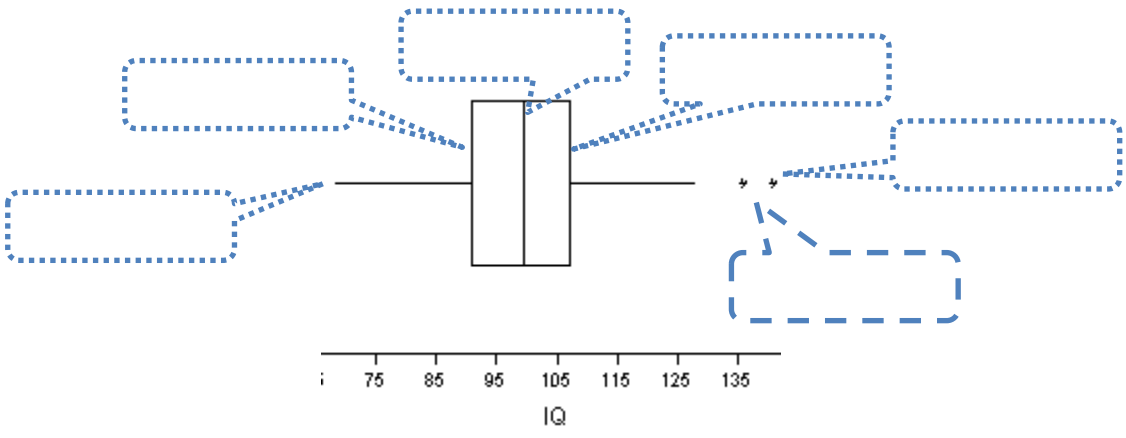
  Histogram is similar to a bar chart in that they both use bars, either horizontal or vertical, to represent the number of data points in each category or interval. However, a histogram has no spaces between bars.

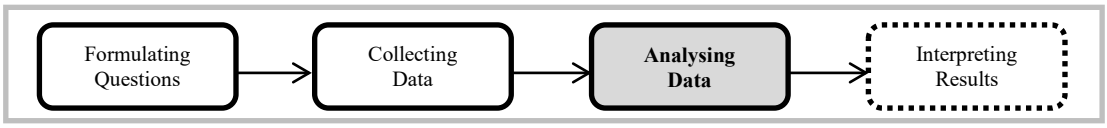
Symmetrical distribution	**Positively-skewed** distribution	**Negatively-skewed** distribution
		
In such cases, we can see (approximately) mirror images about the centre. Note that mean ≈ median.	The tail extends to the right Note that mean > median.	The tail extends to the left. Note that mean < median.

- **Box plot**

  Also known as **box and whiskers plot**, is another way to display quantitative data. It is especially effective for comparing multiple groups of data sets.

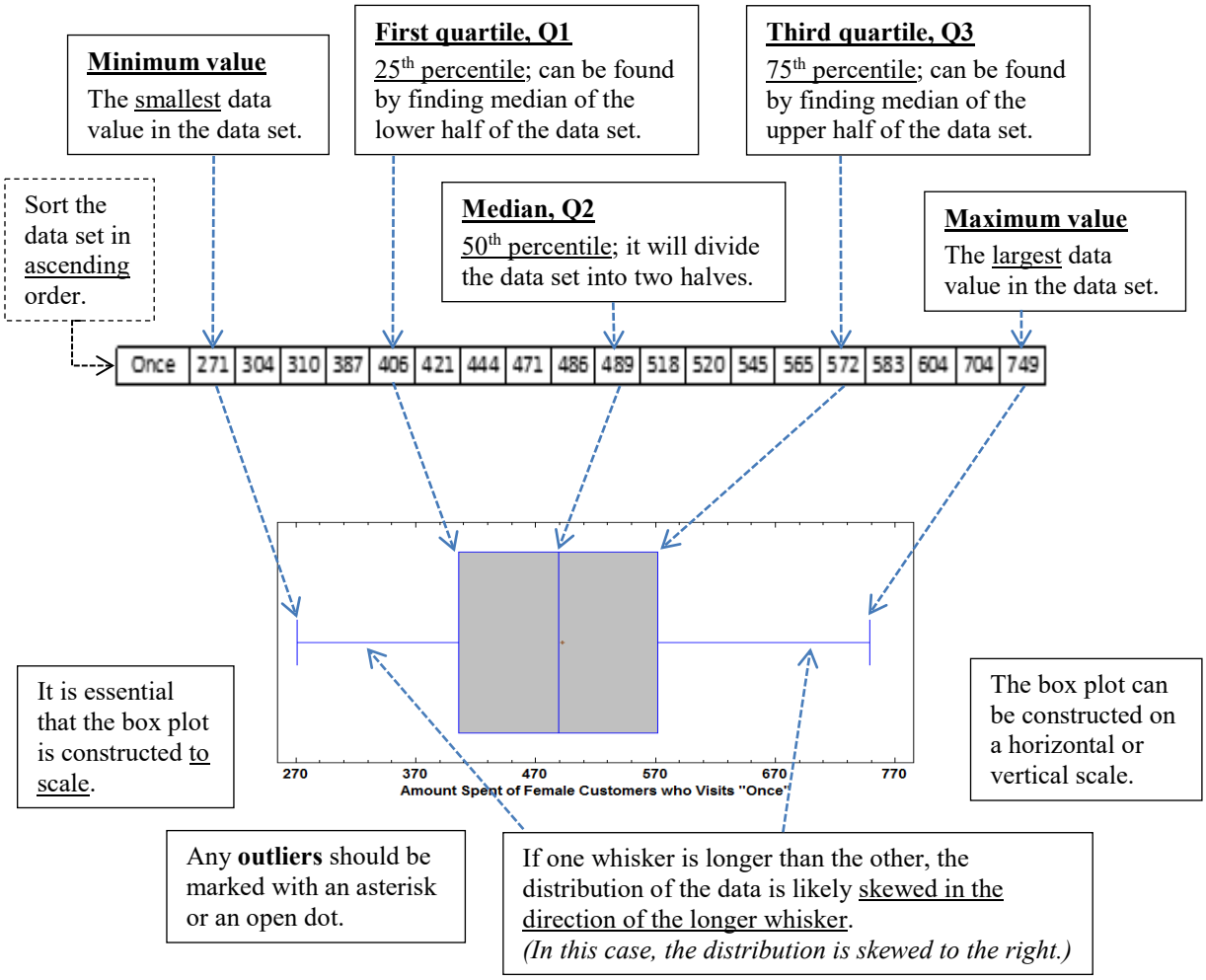
  We will need to generate a **five-number summary** in order to construct boxplot.





The box plot shows much of the structure of the data at a quick glance:

- the centre
- two measures of spread (interquartile range and range)
- skewness
- existence of outliers (extreme data)



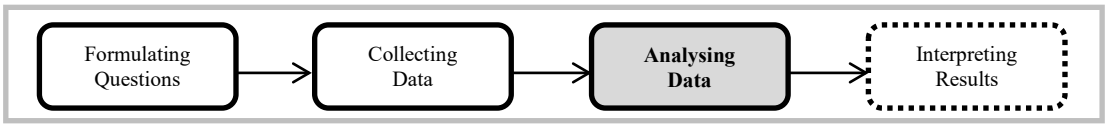
To identify outliers, we compute the values of the **fences**:

- **Lower fence** can be calculated by the formula: Q1 – 1.5×IQR

  In the "Once" data set: lower fence =
- **Upper fence** can be calculated by the formula: Q3 + 1.5×IQR

  In the "Once" data set: upper fence =

Any extreme data values that fall outside the fences are considered to be **outliers**.
Note that fences are not indicated in the box plot.
In the "Once" data set: since all data values fall within the fences, there is no outlier.



CHECKPOINT

#1. Which graph is appropriate for the following visualisation?

- ✓ To visualise proportions of categories _____
- ✓ To visualise shape of distribution _____
- ✓ To visualise differences in frequencies between categories _____
- ✓ To visualise comparison of numerical data between groups _____

#2. List the items in a five-number summary and match the location of each item to the parts of a box plot.

#3. How do you identify outlier(s)?

4.3 Numerical Summaries

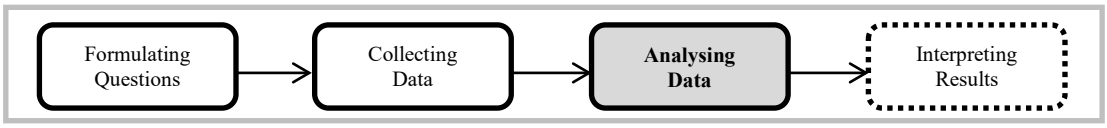
We can describe data by using a few well-chosen numbers that summarise meaningfully the entire data set.

Typically, it is useful to know where the centre or middle of the data set is, referred to as **measures of centre**. It is also known as measures of central tendencies or measures of central location. This is a single value that best represents the concentration of data, and suggests the "average" value of a distribution.

However, measures of centre alone provide only a partial description of a data set. We need a measure to indicate the spread or variation of quantitative data values. These measures are called **measures of dispersion**. In fact, these measures are of essential importance in statistics which is, mainly, the study of variability.

The various measures of centre and dispersion are listed here:

Measures of Centre	Measures of Dispersion
Mode **Median** **Mean**	**Range & Interquartile Range** **Standard Deviation & Variance**



The selection of the appropriate measures of centre is described in the table below:

Measures	**Mode**	**Median**	**Mean**
Method	The most-likely occurring data value.	The centre or middle data value.	The numerical average of the data values.
Application	Most useful in, but not limited to, qualitative data.	Good for quantitative data that has outliers and/or is skewed.	Good for quantitative data that are quite symmetrical and has no outlier.

Furthermore, comparing the values of these measures of centre (usually mean and median suffice) gives a quick sense of the distribution of the data in terms of **skewness**.

Distribution	Negatively-skewed	Symmetrical	Positively-skewed
Comparison	mean < median < mode	mean = median = mode	mean > median > mode

Further elaboration on each measure of centre follows…

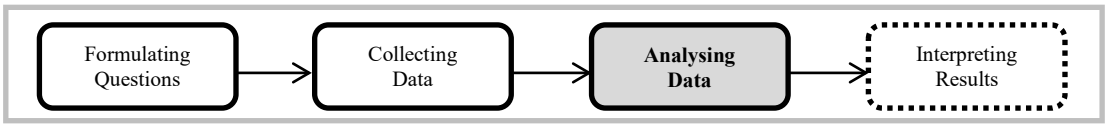
- **Mode**
  - The mode of a data set is the data value that occurs with the greatest frequency.
  - If all data values have same frequencies, then the data set has no mode.
  - If two data values occur with the same greatest frequency, then both the data values are considered modes. Such data with two modes are known as bimodal.

  Examples:

  5, 8, 13, 15, 17 3, 5, 7, 13, 3, 7, 9, 3 1, 1, 2, 2, 2, 2, 3, 4, 5, 5, 5, 5, 6, 7, 9

- **Median**
  - The median of a data set is the value that lies in the middle when the data set is ordered. It is also known as the **second quartile** (**Q2**) or the **50th percentile**.
  - If the data set has an even number of observations, then the median is the midpoint of the two middle data observations. If the data set has an odd number of observations, then the median is the middle data observation.
  - The median is not influenced by extreme data values.
  - In addition, since half of the data values fall below Q2 and the other half of the data values fall above Q2, the median of the lower half of the data values is known as **lower quartile** or **first quartile** (**Q1**) or **25th percentile**.
  - Similarly, the median of the upper half of the data values is known as **upper quartile** or **third quartile** (**Q3**) or **75th percentile**.

  Examples: 4, 7, 9, 11, 12, 20 5, 8, 10, 10, 15, 18, 99



- **Mean**
  - This is the most popular and arguably, most accurate measure of centre.
  - Its value is obtained by "levelling out" the entire data set, hence every data value is used.
  - As a result, mean can be heavily influenced by extreme data values.
  - Mean is meaningless as a measure for qualitative data.
  - The notation for **population mean** is $\boldsymbol{\mu}$ and for **sample mean** is $\bar{\boldsymbol{x}}$.

    Example: 16, 17, 10, 13, 20, 18, 13, 14, 18

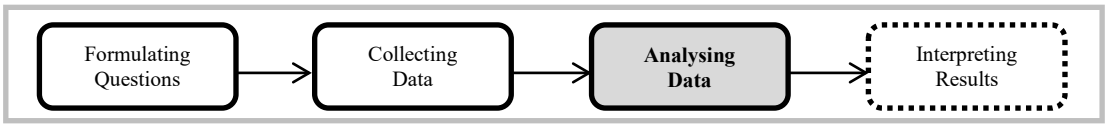
The selection of the appropriate measures of dispersion, paired with the corresponding measure of centre, is described in the table below:

Measures of Dispersion	**Inter-Quartile Range & Range**	**Standard Deviation & Variance**
Application	Range is a quick and easy measure but sensitive to outliers; whereas IQR is not sensitive to outliers. Both are good for skewed data.	Good for data that are quite symmetrical. SD is more commonly used than its squared counterpart, variance.
Corresponding Measure of Centre	Median	Mean

Further elaboration on each measure of dispersion follows…

- **Range & Inter-Quartile Range**
  - The **range** of a data set is simply the difference between the largest and the smallest data values. Although it serves as a quick and easy measure of variability, it might not reflect the typical variability if either the largest or smallest (or both) data value is an extreme data value.
  - **Inter-quartile range** is the difference between the lower and upper quartiles, that is, IQR = Q3 – Q1. Since it measures variation of data values in the middle 50% of the data set, hence it is not affected by extreme data values.
  - Nevertheless, both range and inter-quartile range are based on only two data values in the whole data set. It does not reveal any information about the dispersion of the rest of the data values.

    Examples: 3, 4, 6, 7, 9 15, 15, 20, 25, 25, 30, 30, 30, 35, 75, 85



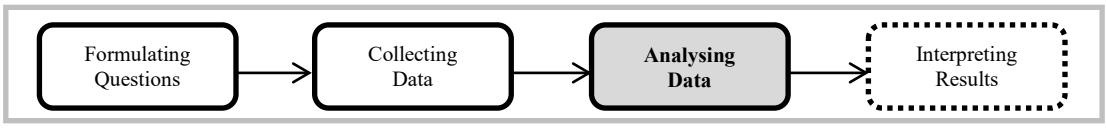
- **Standard Deviation & Variance**
  - The **standard deviation** is considered a more powerful measure of dispersion because it takes into account every data value in the data set, by summarising the amount by which each data value deviates from the mean.
  - Effectively, it indicates how tightly the data values in the data set are "bunched" around the mean value.
  - A small standard deviation implies that the data values are tightly bunched together, whereas a large standard deviation implies that the data values are spread apart.
  - The notation for **population standard deviation** is $\boldsymbol{\sigma}$ and for **sample standard deviation** is $\boldsymbol{s}$.
  - **Variance** is mathematically the square of standard deviation. It represents the average squared deviation from the mean of the data.
  - The notation for **population variance** is $\boldsymbol{\sigma^2}$ and for **sample variance** is $\boldsymbol{s^2}$.

CHECKPOINT

#1. List down all 3 measures of centre and all 4 measures of dispersion.

#2. Describe how you can tell the shape of distribution from the following summaries:
- ✓ Histogram
- ✓ Box plot
- ✓ Measures of centre

#3. Which measures of centre and dispersion will you select for each of the following data?
- ✓ Qualitative data
- ✓ Skewed quantitative data
- ✓ Symmetrical quantitative data



4.4 Analysing Relationships

Let's use the scenario of investigating the **relationship** between motorboat propellers in Florida waterways and manatee fatalities from 1977 to 2011.

4.4.1 HOW CAN WE **VISUALIZE RELATIONSHIPS**?

The number of deaths and the number of powerboat registrations are both quantitative variables. That means they can be measured numerically, and we can plot their values.

Instead of looking at a single variable, we can create a **scatter plot** to consider the relationship between these two variables.

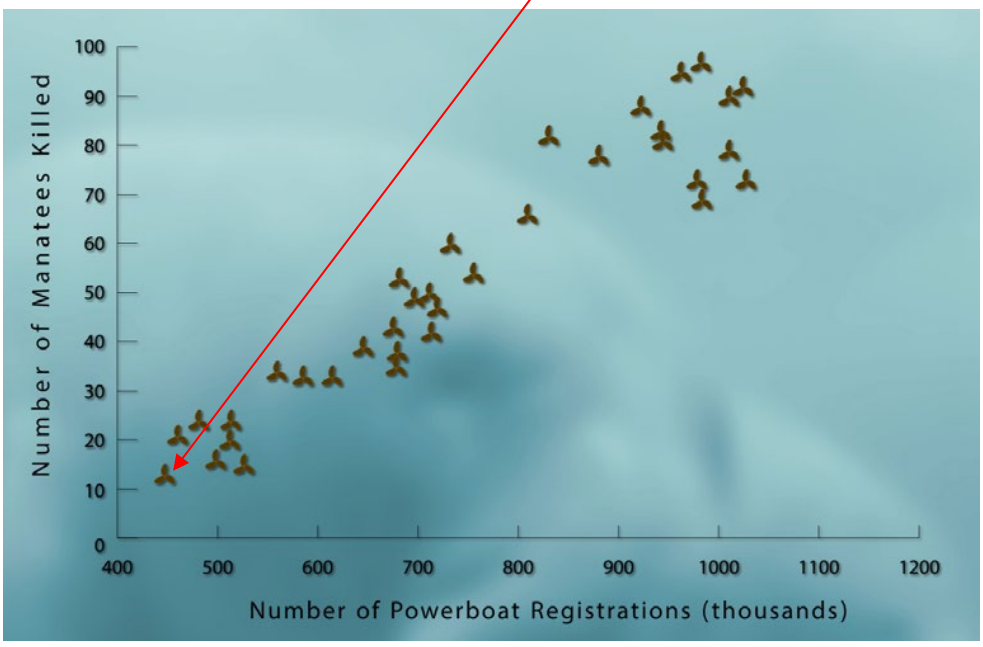
4.4.2 HOW DO WE **PRODUCE A SCATTER PLOT**?

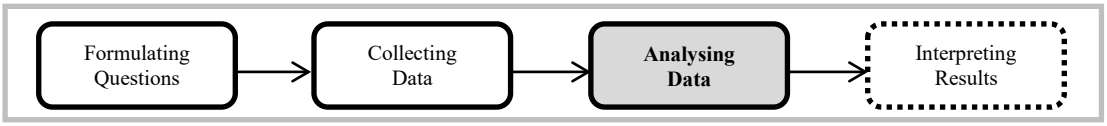
To make a scatter plot, we first draw horizontal and vertical axes.

Since the number of powerboats in the water helps explain the number of manatees killed, thus the number of powerboat registrations is called the **explanatory variable**.
The explanatory variable always goes on the horizontal axis.

We expect that the more boats that are in the water, the more manatees will be killed. That is, we assume that the number of manatees killed is a response to the number of boats in the water, thus we call the number of manatees killed the **response variable**.
The response variable always goes on the vertical axis.

Each point represents a datum. For example, the first point represents that (in 1977) the number of the registrations was 447,000 and the number of manatees killed by boats was 13.





4.4.3 WHAT DOES A **SCATTER PLOT SHOW**?

As the number of powerboat registrations increased, the number of manatees killed increased. This is called a **positive** association.
A **negative** association would be when one variable increases while the other decreases.

The points roughly fall in a line. We call this pattern **linear**.
In fact, since the points do not deviate much from a line, we can say that the linear relationship is **strong** between boats in the water and dead manatees.
If our data were all over the place with much deviations from the line, we would call the relationship **weak**.
However, not all relationships are linear; some show a curved pattern while some have no pattern at all.

When looking at scatter plot, we should look out for:

- Overall pattern – how strong it is and its direction
- Deviations from pattern
- Outliers

A scatter plot show the nature of a relationship between two variables, but it does not prove why the relationship exists. The changes in one variable do not necessarily *cause* the changes in the other; there could be other factors. **(Note: correlation does not imply causality.)**

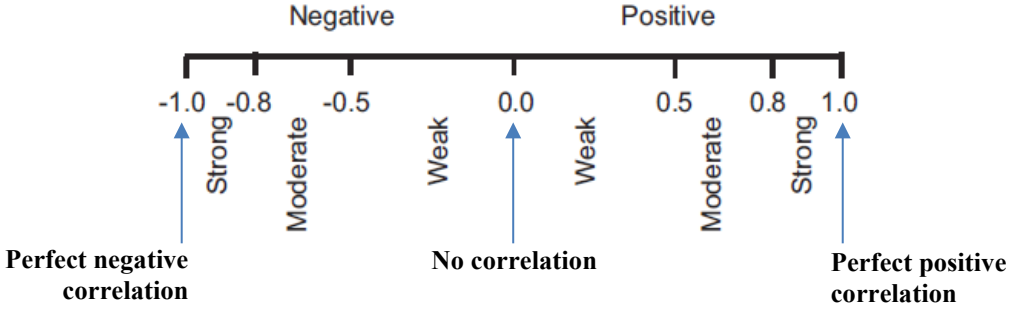
4.4.4 WHAT NUMERICAL SUMMARY CAN **MEASURE RELATIONSHIP**?

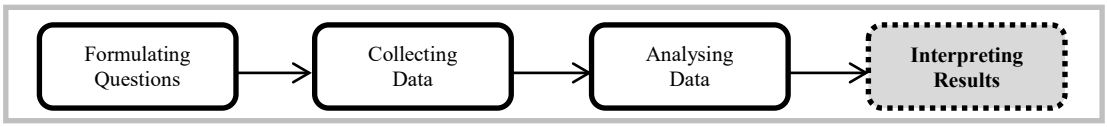
The sample **correlation coefficient**, denoted by ***r***, measures strength and direction of a linear relationship between two quantitative variables.

Basic properties of r :

- The sign of r shows positive or negative association.
- The value of r always satisfies $-1 \le r \le 1$.
- The value of r remains the same when the two variables are interchanged or when the units of the variables are changed.

Guidelines on interpreting r :





CHECKPOINT

#1. In a scatter plot, what are the variables plotted on the horizontal and vertical axes?

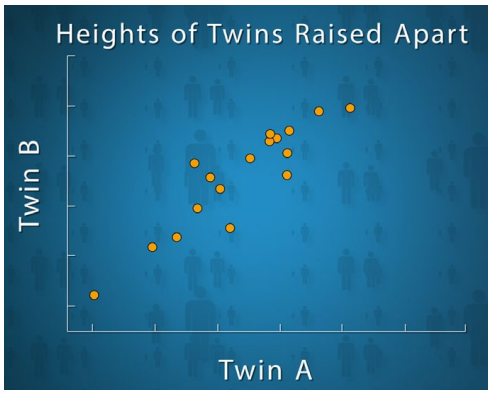
#2. What can a scatter plot tells us about the relationship between two variables?

#3. What can correlation coefficient tells us about the relationship between two variables?

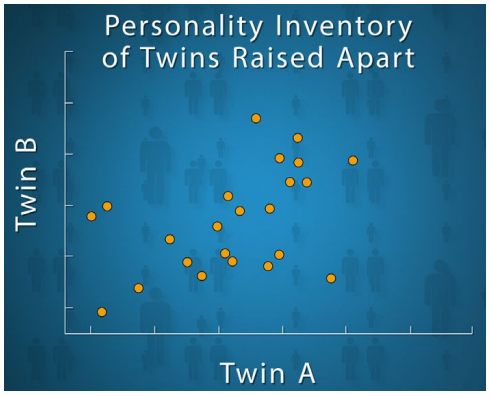
5. Interpreting Results

5.1 Interpreting Relationships

The following graphs and correlation values are produced from studying the physical and personality traits of identical twins who have been raised apart.



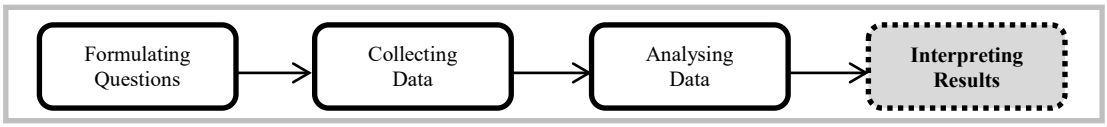
$r = 0.92$



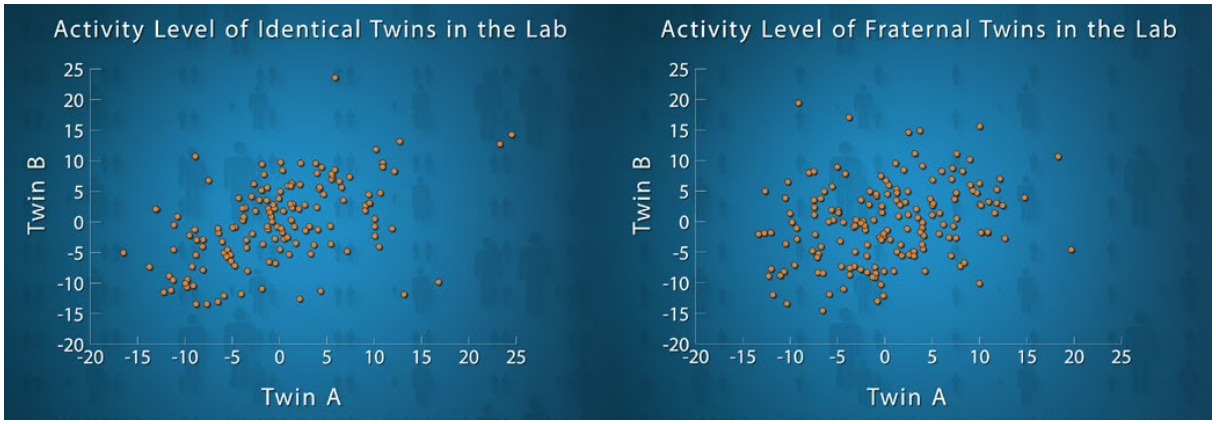
$r = 0.49$

We can observe the following:

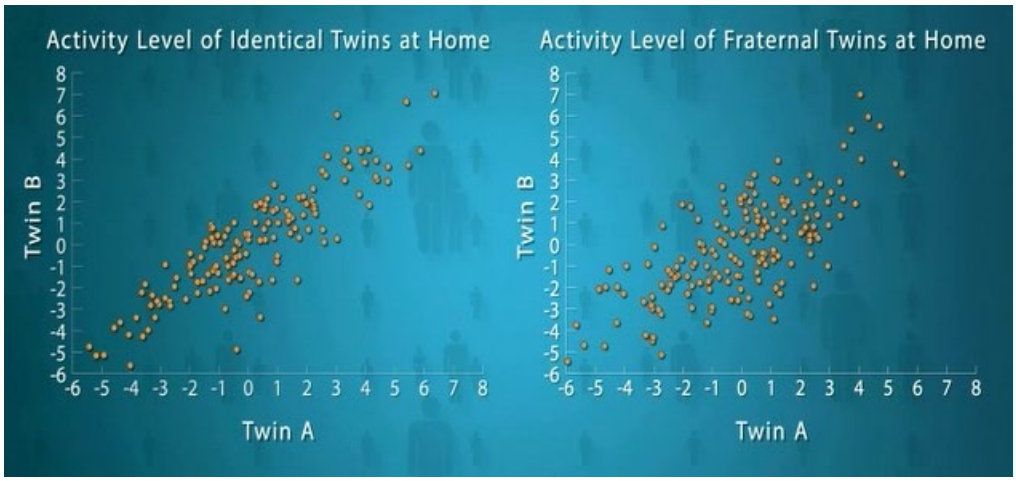
- From the plot on heights, the taller one twin is, the taller is the other. There is a positive association with strong pattern. Since $r = 0.92$, which is very close to 1, it indicates a strong, positive, linear association between heights of twins.
- From the plot on personality, though the relationship is not as clear as it was for height, the points do tend to increase together. Since $r = 0.49$, the relationship is not as strong as for height, but only moderate.



The following graphs and correlation values are produced from studying the activity level of twins in lab setting and at home; identical twins are on the left and fraternal twins on the right.



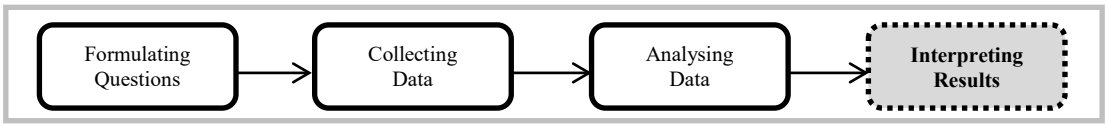
$r = 0.48$ $r = 0.26$



$r = 0.87$ $r = 0.7$

We can observe the following:

- In lab setting, there is moderate positive association between activity levels of identical twins, but weak positive association between activity levels of fraternal twins.
- Hence, in lab setting, the correlation between the activity levels of fraternal twins is much less than that between identical twins.
- In home setting, both plots show strong patterns.
- The correlation of activity levels in both identical and fraternal twins are much higher in the home setting (moderate to strong) than in the lab setting (weak to moderate).

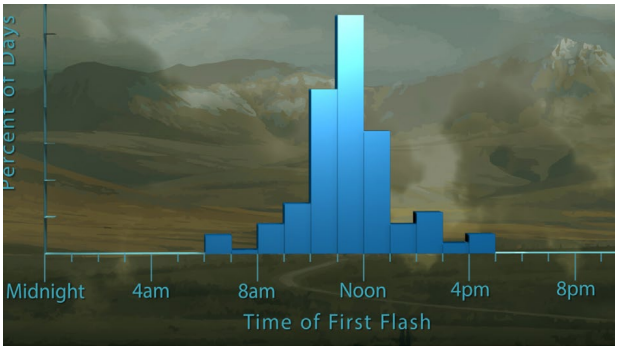


5.2 Interpreting Graphical Summaries

This histogram shows the time of first lightning strike collected over a particular year in a small area of Colorado, US.

We can observe the following from the graph:

- horizontal axis represents time of day
- vertical axis represents percentage of days
- each bar represents one hour
- roughly symmetrical about the tallest bar between 11am and 12 noon
- data is tightly clustered around the central bar, between 10am to 1pm
- no first strikes at night

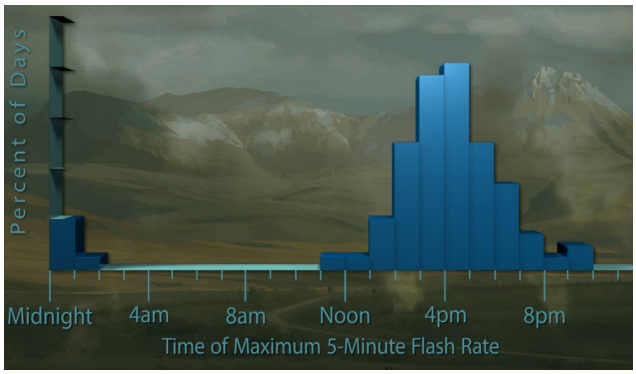


This histogram shows the time of day when the maximum number of lightning flashes (in 5 mins) were recorded in the same year and area as above.

We can observe the following from the graph:

- a peak shows that most flashes occur between 4pm and 5pm
- there are outliers where maximum flashes occur between 12am and 2am

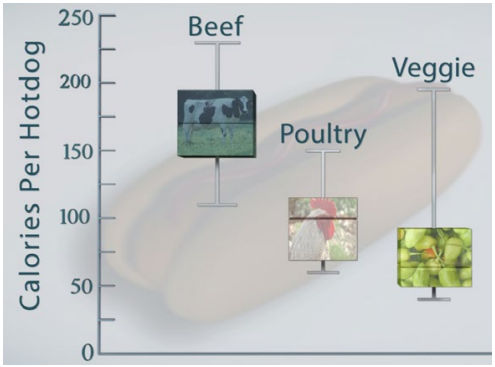
It is important when plotting a histogram to choose the best class size, that is, the width of intervals along the horizontal axis.

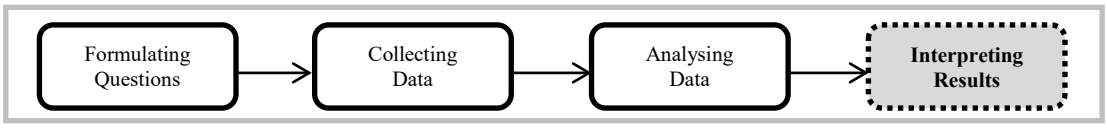


This box plot compares calories of beef, poultry and veggie hotdogs.

We can observe the following from the graph:

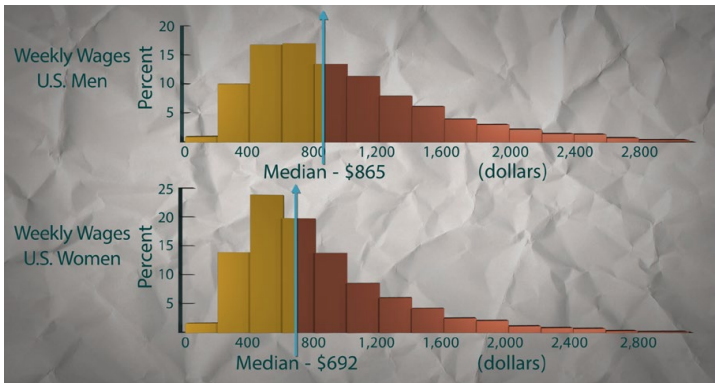
- The median of the poultry hotdogs lies below the minimum value for beef hotdogs, meaning the *typical* poultry hotdog has fewer calories than any beef brand.
- Overall, the veggie hotdogs have the lowest calories. But, the whiskers show that at least one veggie brand has more calories than ¾ of the beef hotdogs.





5.3 Interpreting Numerical Summaries

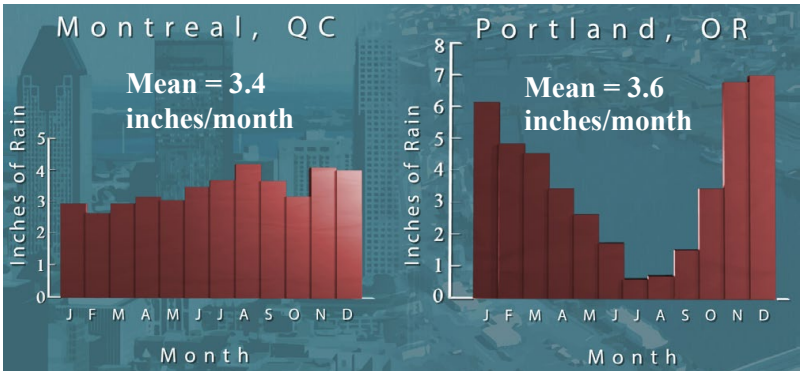
These histograms, marked with the respective medians, show the weekly wages of Americans in 2011, separated by gender.



We can observe the following:

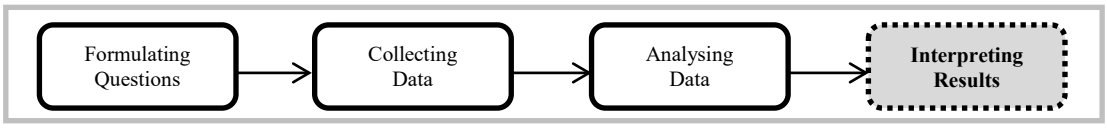
- Both histograms are skewed to the right, with most people making moderate salaries, while a few make much more.
- The median weekly salary for men in 2011 was $865. This means that half of all men made more than $865, and half earned less.
- The median wage for women was only $692, just 80% of what men make.

These graphs and statistics show the rain distribution of Montreal, Quebec and Portland, Oregon in a year.

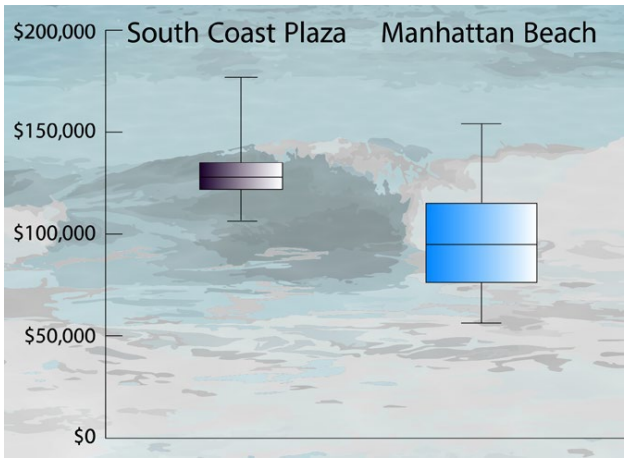


We can observe the following:

- The mean values show that average monthly rainfall for both cities are about the same, but they have very different climate.
- From the graph, Montreal's rainfall is relatively consistent, measuring between 2 to 4 inches monthly.
- However, Portland's rainfall is much more varied, concentrated in the winter months, which can get almost 7 inches of rain, while summer months get less than 1 inch.



These box plots and statistics show the sales from two Wahoo's Fish Taco stores over four-week periods, one located at South Coast Plaza and the other located at Manhattan Beach.



Store	South Coast Plaza	Manhattan Beach
Mean	$130,675 per month	$97,429 per month
SD	$17,164	$31,075

We can observe the following:

- From the boxplots: The median sales of South Coast Plaza location is higher than that of Manhattan Beach location. But the interquartile range (represented by the widths of the boxes) for Manhattan Beach location is wider than South Coast Plaza location.
- South Coast Plaza location has higher mean sales than Manhattan Beach location.
- The SD values also show that the sales for Manhattan Beach location has greater variability than South Coast Plaza location.

Statistical Problem-Solving Process
Case Study: Prestige Mall

Student Name	Student Number	Class
(#1)		
(#2)		

Please refer to the case study of "Prestige Mall" on the chapter of Descriptive Statistics. Minitab Express will be used to analyzed the data.

Q1: What is the aim of this case study?

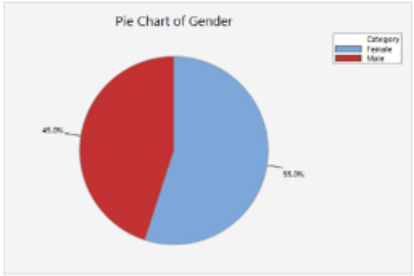
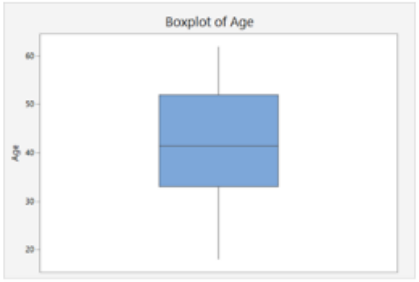
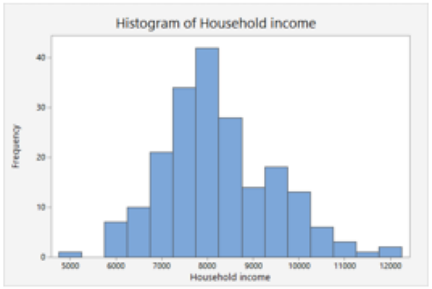
Q2: What is the sample of this case study? And what is the targeted population?

Q3: How were the data collected, as recorded in the data file named "Prestige Mall"?

Q4: What information does the data file named "Prestige Mall" hold?

Statistical Problem-Solving Process
Case Study: Prestige Mall

			Numerical and graphical summaries	Generalization to the target population
Q5	What is the proportion of customers in the IT/Eng and Bus/Fin sectors?	*Which variable data would you use to answer this question?* *What kind of data is this?*	Bus/Fin: IT/Eng: Total proportion:	*Describing the sample:*
Q6	How often do the customers visit Prestige Mall in the last month?	*Which variable data would you use to answer this question?* *What kind of data is this?*	Mean: Median: SD:	
Q7	How much did the customers spent last month at Prestige Mall?	*Which variable data would you use to answer this question?* *What kind of data is this?*		*Generalizing to the population:*

			Numerical and graphical summaries	Generalization to the target population
Q8	What is the proportion of male and female customers of Prestige Mall?	*Which variable data would you use to answer this question?* **Gender** *What kind of data is this?* **Qualitative (nominal)**	**Male: 45%** **Female: 55%** 	*Describing the sample:* There is a slightly higher proportion of female customers visiting Prestige Mall compared to male customers. The boxplot shows that the distribution of age of customers is quite symmetrical ranging from 18 to 62 years. The mean age is 42 years, with SD of about 12 years. Hence, customers who frequent Prestige Mall are more likely to be mature adults. The histogram for household income shows slight positive skewness. Many customers cluster around moderately low household income. The median household income is about $8k, with IQR of about $1.6k. *Generalizing to the population:* Generally, customers of Prestige Mall are mature adults, slightly more likely to be female, and could have moderate household income.
Q9	What is the age profile of the customers?	*Which variable data would you use to answer this question?* **Age** *What kind of data is this?* **Quantitative (discrete)**	**Mean: 42.0 years** **Median: 41.5 years** **SD: 11.9 years** 	
Q10	What is the distribution of income of the customers?	*Which variable data would you use to answer this question?* **Household income** *What kind of data is this?* **Quantitative (continuous)**	**Mean: $8231.29** **Median: $8067.36** **IQR: $1598.32** 	

			Numerical and graphical summaries	**Generalization to the target population**
Q11	Is there any preliminary evidence to claim that female customers who went to the mall 6-times spent more than female customers who went to Prestige Mall only once last month?	*Which variable data would you use to answer this question?* *What kind of data is this?* *Which variable data is used for grouping?*	See table below	*Describing the sample:* *Generalizing to the population:*

Amount spent/ month	Visit the mall once / month	Visit the mall 6 times / month
n	**19**	
Mean	**$ 492.24**	
SD	**$ 127.24**	
Minimum	**$ 271.49**	
Q1	**$ 406.34**	
Median (Q2)	**$ 489.26**	
Q3	**$ 572.09**	
Maximum	**$ 749.45**	

			Numerical and graphical summaries	Generalization to the target population
Q12	*(Ask a question about the relationship between 2 variables, then proceed to investigate.)* **Is there a relationship between…**	*Which variable data would you use to answer this question?* *What kind of data are these?*		*Describing the sample:* *Generalizing to the population:*

TUTORIAL 2

1. To investigate the driving habits of Singaporeans, you would like to design a survey to collect data from a sample of 100 drivers.

   (a) Define the population and sample in this context.

   (b) Decide which of the following variables is relevant to your investigation and classify the type of data to be collected.

	Variable	**Relevant or not?**	**Type of data**
i	Age of driver		
ii	Height of driver		
iii	Weight of driver		
iv	Gender of driver		
v	Capacity of car (eg. 1600 cc)		
vi	Number of trips made per day		
vii	Distance covered per day		
viii	Amount of money spent on petrol per month		
ix	Colour of car		
x	Make (model) of car		
xi	Purchase price of car		

   (c) Select one of the relevant variables as indicated in part (b) and justify why this variable is relevant in your investigation.

   (d) Which type of graphs is suitable to present the data of the following variables?

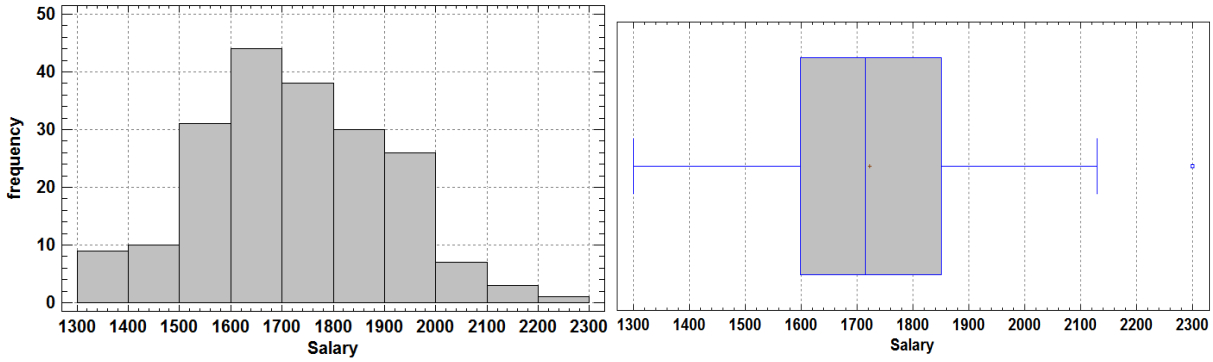
   I. Age of driver

   II. Gender of driver

   III. Number of trips made per day

   For each of the graphs selected to present the data in parts I to III, what information can be obtained from the graph?

2. Two hundred staff were randomly selected from a company and their salaries were presented using two charts, as shown in the following.



(a) What is the median salary of these 200 staff?

(b) Find the range and interquartile range of the salaries.

(c) What are the cut-off salaries for the bottom 25% and top 25% earners?

(d) Is there any outlier salary? What are the values of the fences?

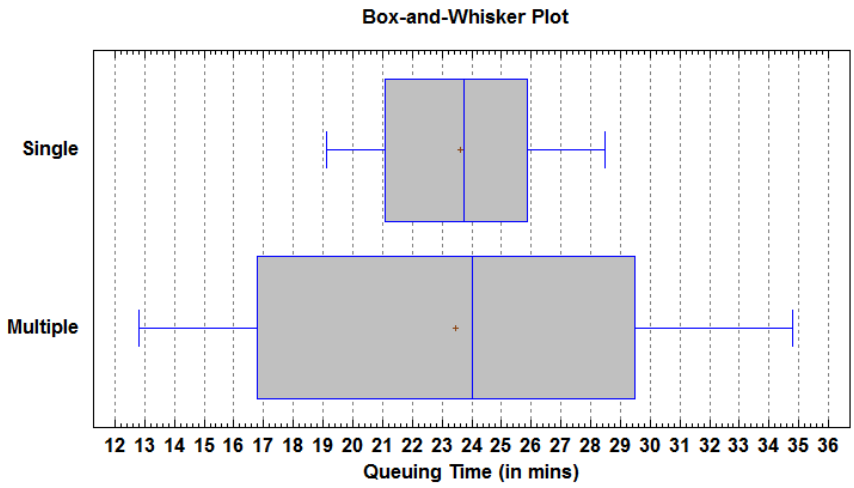
(e) How many staff earn between \$1800 and \$2000?

(f) Andrew earns \$1600. At which percentile is his salary?

(g) What is the shape of the distribution of salaries?

3. To serve customers better by cutting the queuing time at the counters, ABC Bank experimented with two types of queue system:
   - a single queue that feeds to all counters, or
   - multiple queues, one for each counter.

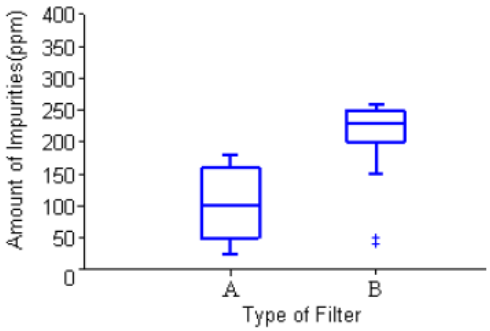
   The queuing times (in minutes) for 20 customers during the peak period before being served were recorded for each queue system. The results are displayed in the following box plots, where "+" inside the box represents the mean queuing time.



   Compare the two types of queue systems.

   Hints:
   - *Compare and comment on the measures of centre of both systems.*
   - *Compare and comment on the measures of dispersion of both systems, and discuss their pros and cons.*

4. Two new filtration systems A and B have been proposed for use in the water systems of a small city. The amount of impurities (in parts per million) remaining in the water after the water passes through each filter is recorded over a 30-day period. The average daily values for the two systems are plotted using a side-by-side box plot as follows:

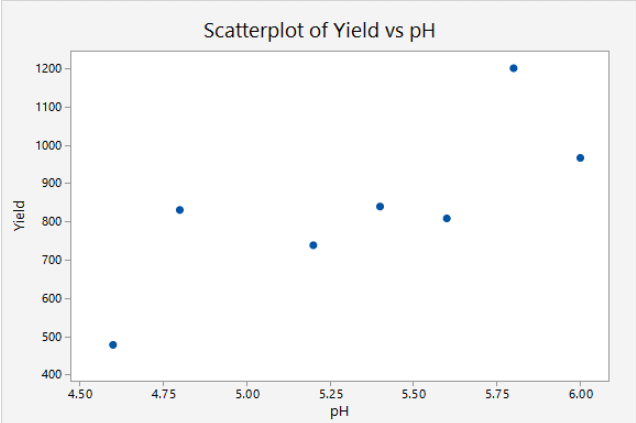


(a) For each filter, describe the shape of the distribution of the amount of impurities.
(b) Estimate the median, lower quartile and upper quartile for each filter.
(c) Which filter, A or B, produces less variability? Briefly explain.
(d) Which filter, A or B, appears to generally filter water more thoroughly?

5. A scientist planted alfalfa on several plots of land, identical except for the soil pH. The data collected and shown below give the yields (in kilograms per acre) for each plot. The scatterplot and correlation coefficient are also produced below.

pH	Yield
4.6	479
4.8	831
5.2	739
5.4	840
5.6	809
5.8	1201
6.0	967

$r = 0.78$



(a) Which is the explanatory variable and which is the response variable?
(b) Comment on the relationship between variables pH and yield.

6. You wish to compare the weight reducing program offered by two programmes, Program A and Program B. You have 60 participants and you randomly assigned thirty of them to each program. The data on the weight loss (in kg) of the participants two months after attending the programs were collected. Minitab Express gave the following summary:

Descriptive Statistics: Programme A, Programme B

Statistics

Variable	N	Mean	StDev
Programme A	30	4.0833	0.6086
Programme B	30	4.9633	0.5798

Which program is more effective in weight reducing? Explain.

ANSWERS

1. (a) Population: all Singaporean drivers
   Sample: the 100 Singaporean drivers surveyed

   (b) *<As long as you can justify, there is no correct or wrong answers to "relevance".>*

(i) Quantitative	(ii) Quantitative	(iii) Quantitative
(iv) Qualitative	(v) Quantitative	(vi) Quantitative
(vii) Quantitative	(viii) Quantitative	(ix) Qualitative
(x) Qualitative	(xi) Quantitative	

   (c) *<Sample answer>* For example, capacity of car: more powerful cars in the hands of amateur drivers may cause more reckless driving.

   (d) I. Histogram; to see the distribution of the age data
   II. Pie chart; to see the proportion of male and female drivers
   III. Bar chart; to see the differences between the number of trips recorded

2. (a) \$1720 (b) \$1000, \$250 (c) \$1600, \$1850
   (d) Yes, \$2300, LF = \$1225, UF = \$2225 (e) About 56 staff
   (f) About 25th percentile (g) Slightly positively-skewed

3. The mean and median for both the system is approximately the same but the variation (as measured by the "box") of the multiple queue system is greater than that of single queue system. The minimum time for single system is higher than that of multiple queue system, but the maximum queue time for single system is lower than that of the multiple queue system. Although there is a possibility that a customer may be have a shorter queue time in a multiple queue system, but queue time for multiple queue system is not as consistent as single queue system.

4. (a) A is roughly symmetric; B is negatively-skewed with 2 outliers.
   (b) Filter A: Q1 ≈ 50 ppm , Q2 ≈ 100 ppm , Q3 ≈ 160 ppm
   Filter B: Q1 ≈ 200 ppm , Q2 ≈ 230 ppm , Q3 ≈ 250 ppm
   (c) Ignoring the outliers, B seems to produce less variability, as evident from the shorter width of the box, which represents an IQR of approximately 50 ppm.
   (d) Filter A

5. (a) Explanatory: pH; Response: Yield
   (b) Scatterplot shows that the data points are somewhat close to a straight line. As pH increases, yield also increases. Since $r = 0.78$, it indicates quite a strong positive linear relationship.

6. B, higher mean weight loss with about the same variation in weight loss.

LAB 2 : Descriptive Statistics

Learning Objectives:

1. *Enter and import data into Minitab Express.*
2. *Generate numerical summaries using Minitab Express.*
3. *Generate graphical summaries using Minitab Express.*
4. *Generate correlation for bivariate data using Minitab Express.*

Task 1A

Input data into Minitab Express.

There are two ways to input data: type manually or copy-paste from another source.

You can type data manually into cells in the Minitab worksheet. Press *Enter* to go to the next row, and press *Tab* to go to the next column.

Notice that the un-numbered row contains the label of each column.

	C1	C2	C3	C4
	School	GPA	Starting Salary	
1	EEE	2.90	1520.10	
2	EEE	3.81	1819.50	
3	SB	3.50	1584.90	
4	CLS	3.92	2022.00	
5	MAE	3.61	1724.50	
6	SB	3.72	1685.78	
7	MAE	3.13	1582.70	
8	MAE	3.47	1633.41	
9	MAE	3.91	1826.32	
10	CLS	3.75	1758.40	
11	SB	3.22	1515.50	
12	EEE	3.34	1560.33	
13	SB	3.68	1696.00	

Step 1: Open Minitab Express software.

Step 2: Go to **File > Save Project As**, name your project "*Lab2*" and save in your preferred location.

Step 3: Enter data into the cells in the worksheet.

You can copy and paste data from a different Minitab worksheet or from a different application, such as Microsoft Excel or Numbers by Apple.

If you paste data into a single cell of the worksheet, Minitab overwrites the contents of neighbouring cells in order to paste all of the data from the clipboard.

If you select multiple cells and paste data, Minitab pastes data into the selected cells and omits or repeats values to fill the paste area.

Step 1: Open Excel file "*SAE_Data_AY1920.xlsx*", go to worksheet tab "*Lab2*".

Step 2: Copy all data with headers "*School*", "*GPA*" and "*Starting Salary*", including the header.

Step 3: Go to Minitab Express, in the same Minitab worksheet, paste the data in column C1, starting from just under label "*C1*". Note the un-numbered row which is supposed to contain the label/header of each column.

Task 1B

Use the data set in "*Lab2*" Excel worksheet to construct various graphical summaries and provide basic interpretations. The graphs include:

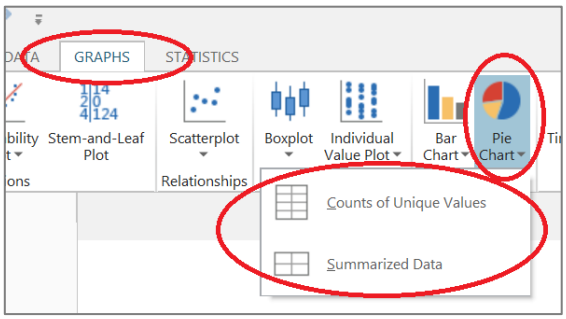
- **pie charts**
- **bar graphs**
- **histograms**
- **boxplots**

(I) CONSTRUCTING PIE CHARTS

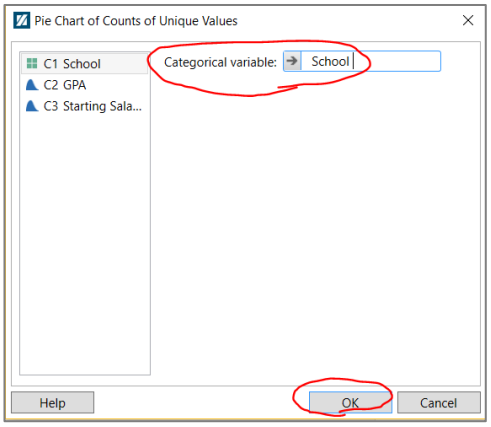
Use a pie chart to compare the proportion of data in each category or group.
A pie chart is a circle ("pie") that is divided into segments ("slices") to represent the proportion of observations that are in each category.

There are two types of pie charts in Minitab Express:

- **Counts of Unique Values**
  Create a pie chart that displays counts of unique values for one variable.
- **Summarized Data**
  Create a pie chart that displays the summarized data for values of a categorical variable.

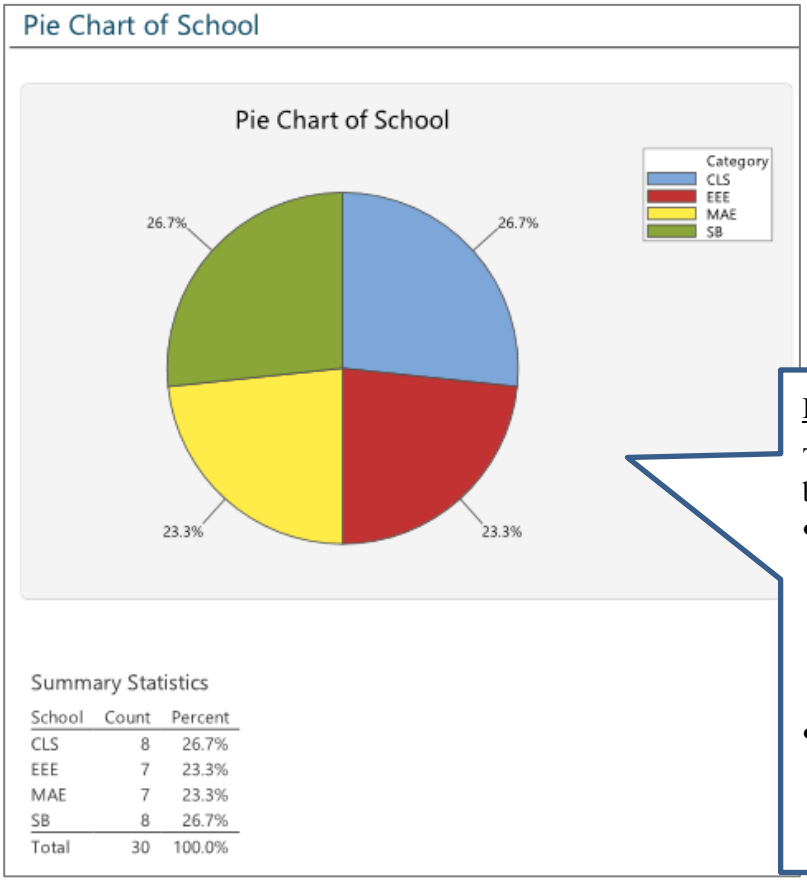


Step 1: Select **GRAPHS** > **Pie Chart** > **Counts of Unique Values**



Step 2: For **Categorical variable**, select *School*.
Click **OK**.

Step 3: The graph will be displayed in the output window.



Basic Interpretations of Pie Charts

To interpret a pie chart, compare between groups.

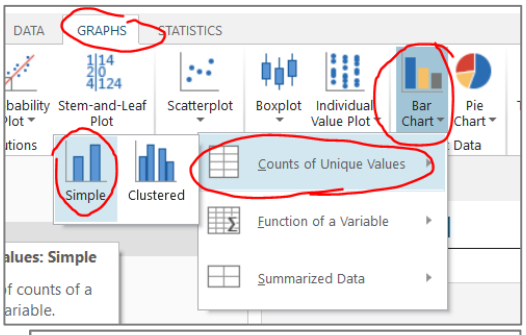
- When you interpret single pie chart, look for differences in the size of the slices. The size of a slice shows the proportion of observations that are in that group.
- When you compare multiple pie charts, look for differences in the size of slices for the same categories in all the pie charts.

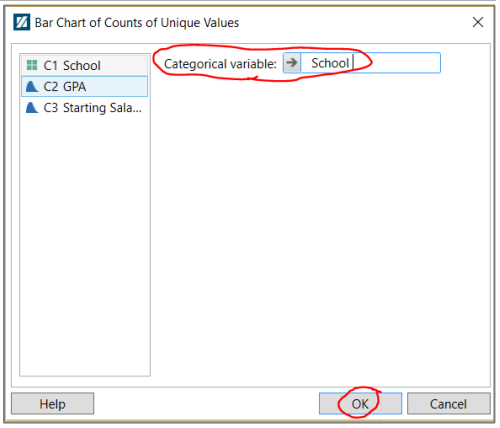
(II) CONSTRUCTING BAR CHARTS

Use a bar chart to compare the counts or frequencies (and sometimes summarised statistics like mean) using bars to represent groups or categories.

We shall use the following to create bar charts in Minitab Express:

- **Counts of Unique Values**
  Create a bar chart of counts of a single categorical variable.
- **Summarized Data**
  Create a bar chart of a single column of summarized data for one categorical variable. Each observation summarizes a category. Summarized data can be a count or a calculated value, such as a mean.

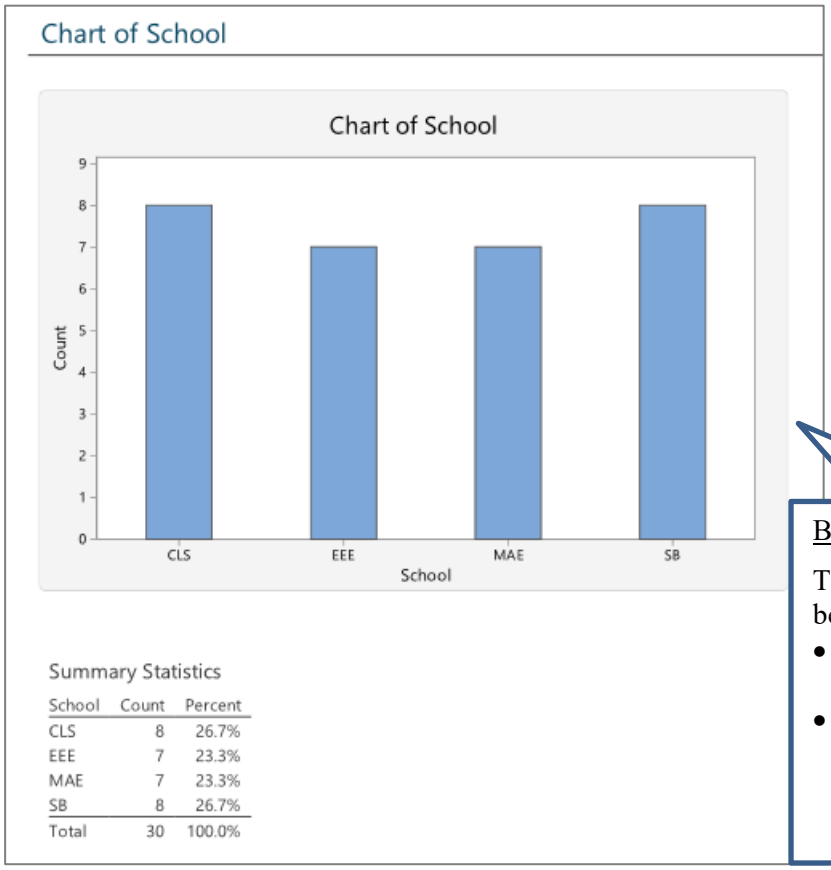




Step 1: Select **GRAPHS** > **Bar Chart** > **Counts of Unique Values** > **Simple**.

Step 2: For **Categorical variable**, select *School*. Click **OK**.

Step 3: The graph will be displayed in the output window.



Basic Interpretations of Bar Charts

To interpret a bar chart, compare between groups.

- Look for differences in the heights of the bars.
- The bars show the value for the groups. Refer to the scale range of the *y*-axis to determine the actual differences.

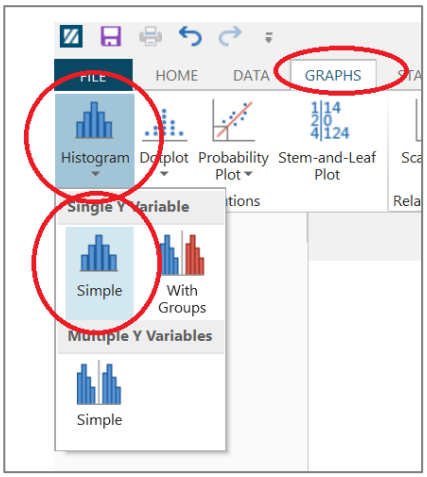
(III) CONSTRUCTING HISTOGRAMS

Use a histogram to examine the **shape** and **spread** of your data.

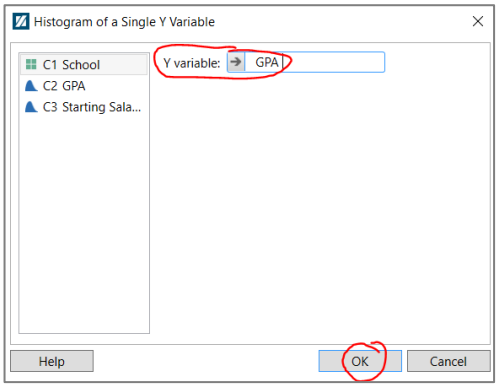
A histogram works best when the sample size is at least 20. If the sample size is too small, each bar on the histogram may not contain enough data points to accurately show the distribution of the data.

A histogram divides sample values into many intervals and represents the frequency of data values in each interval with a bar.

After you create a histogram, you can add a normal distribution fit line, change the scale type, etc.

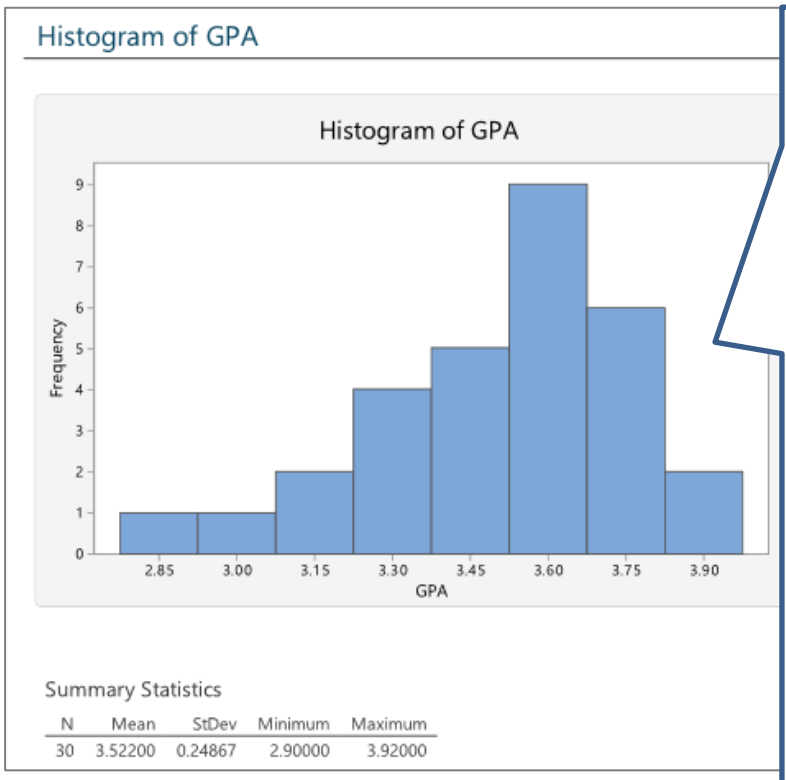


Step 1: Select **GRAPHS > Histogram > Single Y Variable: Simple.**



Step 2: For **Y variable**, select *GPA*.
Click **OK**.
(Try constructing histogram for *Starting Salary* too.)

Step 3: The graph will be displayed in the output window.



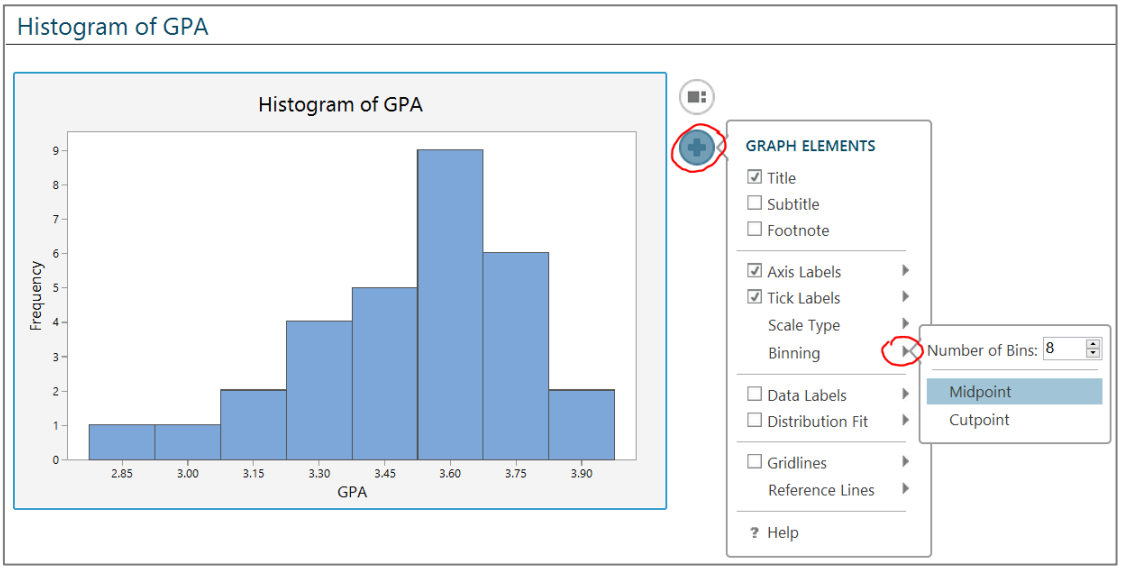
Basic Interpretations of Histogram

To interpret a Histogram, assess the key characteristics:

- Identify the peaks, which are the tallest clusters of bars. The peaks represent the most common values.
- Assess the spread of your sample to understand how much your data varies.
- Observe the skewness. When data are skewed, the majority of the data are located on the high or low side of the graph. Skewness indicates that the data may not be normally distributed.
- Outliers, which are data values that are far away from other data values, can strongly affect your results. Often, outliers are easiest to identify on a boxplot. On a histogram, isolated bars at the ends identify outliers.

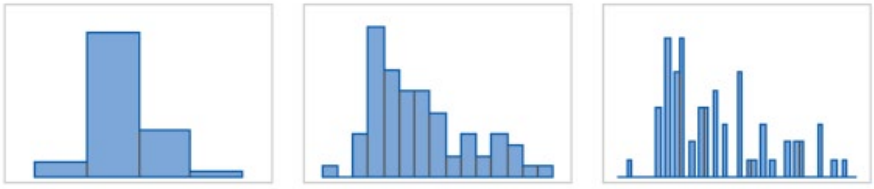
CUSTOMIZING THE HISTOGRAM

Click the graph to select it, then click the plus sign beside the graph and select items to display on the graph. The following information describes some of the items on the **Graph Elements** menu:



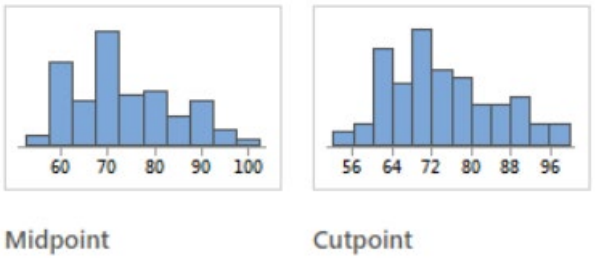
Choose the **number of bins**:

- The number of bins affects the appearance of a graph. If there too few bins, the graph will be unrefined and will not represent the data well.
- If there are too many bins, many of the bins will be unoccupied and the graph may have too much detail. For example, these histograms represent the same data with different numbers of bins.



Choose where to display the **tick labels**:

- Bins can be defined by either their midpoints (centre values) or their cut points (boundaries).
- The appearance of the graph changes if you change the bin definition method.

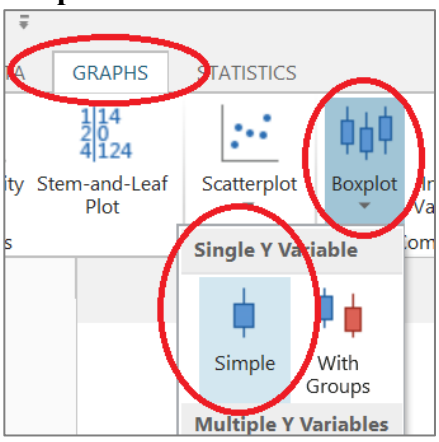


(IV) CONSTRUCTING BOXPLOTS

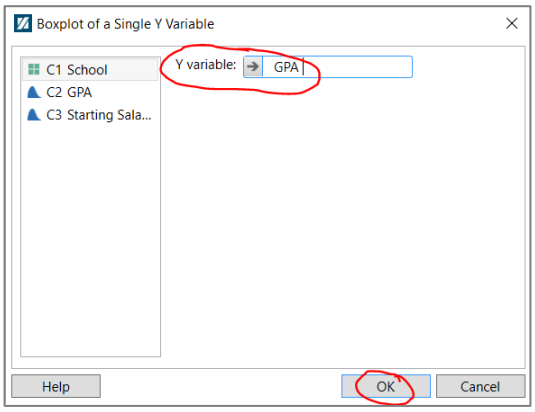
Use a boxplot to assess and compare the shape, central tendency, and variability of sample distributions and to look for outliers.

A boxplot works best when the sample size is at least 20. A boxplot shows the median, interquartile range, range and outliers.

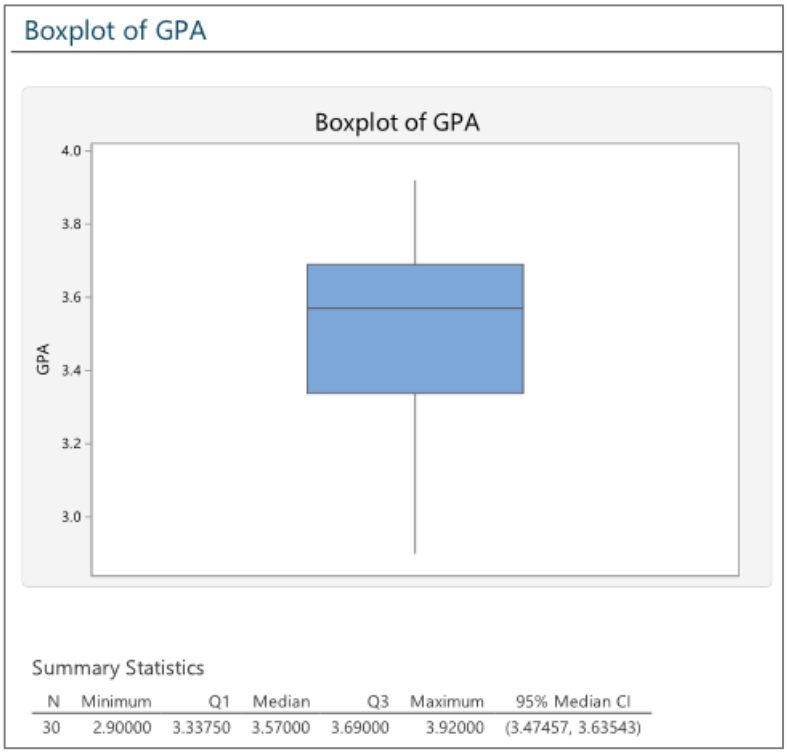
Step 1: Select **GRAPHS > Boxplot > Single Y Variable: Simple**



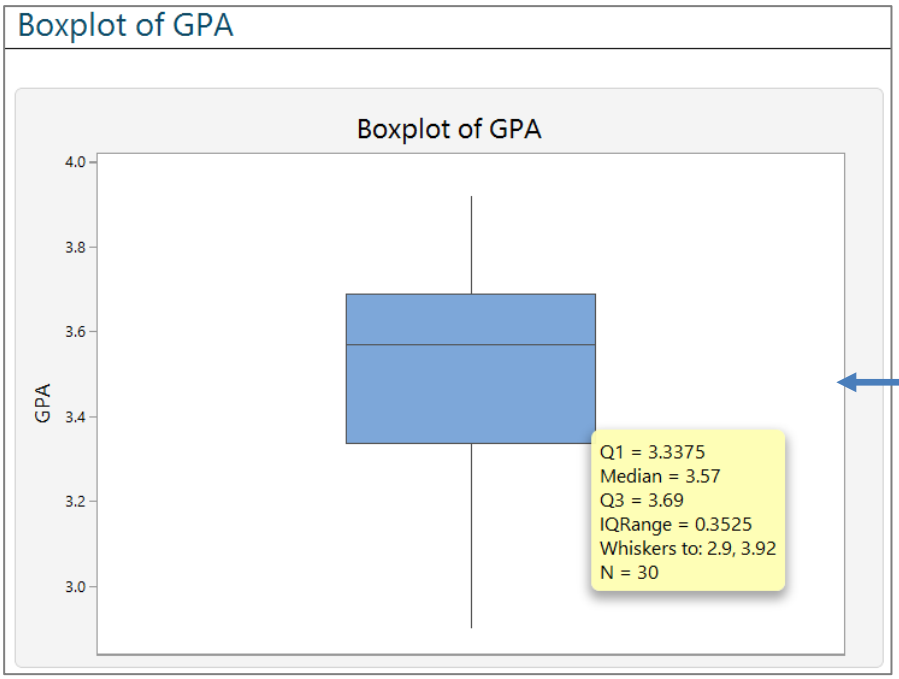
Step 2: For **Y variable**, select *GPA*.
Click **OK**.
(Try constructing histogram for *Starting Salary* too.)



Step 3: The graph will be displayed in the output window.



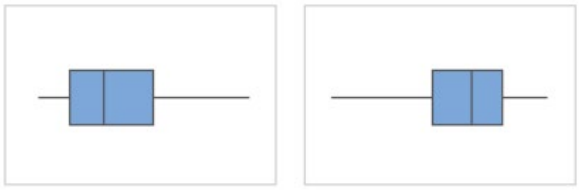
Hover the pointer over the boxplot to display a tooltip that shows numerical statistics.



Basic Interpretations of Boxplot

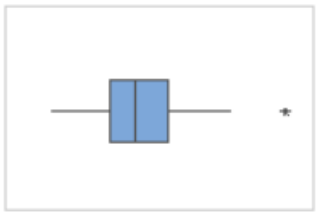
To interpret a boxplot:

- Examine the following elements to learn more about the centre and spread of your sample data.
  - The median is represented by the line in the box. The median is a common measure of the centre of your data.
  - The interquartile range box represents the middle 50% of the data.
  - The whiskers extend from either side of the box. The whiskers represent the ranges for the bottom 25% and the top 25% of the data values, excluding outliers.
- Skewed data
  - When data are skewed, the majority of the data are located on the high or low side of the graph. Skewness indicates that the data may not be normally distributed (you will learn Normal distribution in Chapter 3).



Right-skewed Left-skewed

- Outliers, which are data values that are far away from other data values, can strongly affect results. Often, outliers are easiest to identify on a boxplot. On a boxplot, outliers are identified by asterisks (*).



Task 1C

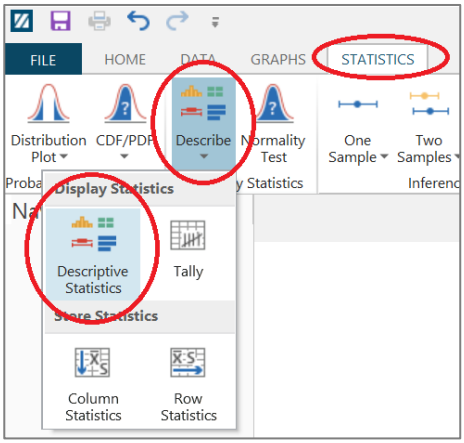
Use the dataset in "*Lab2*" Excel worksheet to compute numerical summaries of data and provide basic interpretations.

Use this analysis to summarize numeric data with a variety of statistics such as the sample size, mean, median and standard deviation. This analysis also provides graphs of your data.

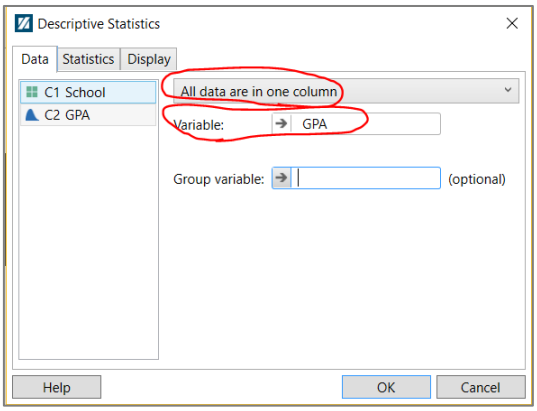
The data must be numeric. You must have continuous data, such as the weights of packages, or discrete data, such as the number of complaints.

Samples that have at least 20 observations are often adequate to represent the distribution of your data. However, to better represent the distribution with a histogram, some practitioners recommend that you have at least 50 observations. Larger samples also provide more precise estimates of the process parameters, such as the mean and standard deviation.

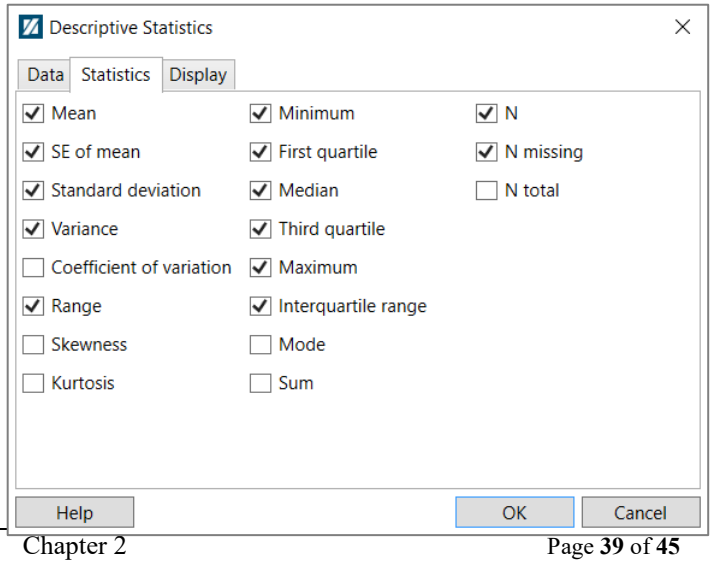
Step 1: Select **STATISTICS** > **Describe** > **Descriptive Statistics**



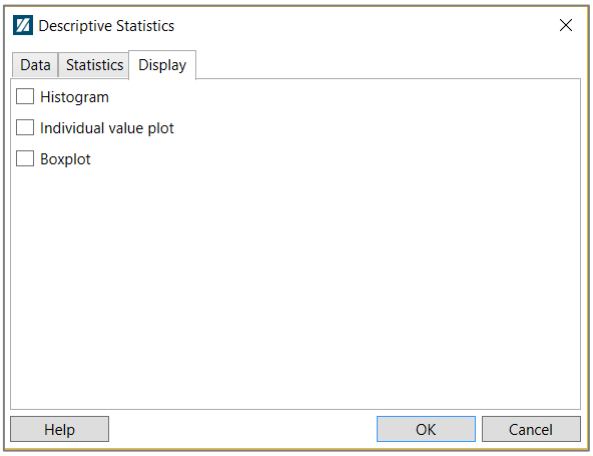
Step 2: In **Data** tab, select **All data are in one column**. For **Variable**, select *GPA*.
(Try displaying summary statistics for *Starting Salary* too.)



Step 3: In **Statistics** tab, select the numerical statistics you wish to display.



Step 4: (Optional) In **Display** tab, select the graph you wish to display.
Click **OK**.



Step 5: The results will be displayed in the output window.

Descriptive Statistics: GPA

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum	Range	IQR
GPA	30	0	3.52200	0.04540	0.24867	0.06184	2.90000	3.33750	3.57000	3.69000	3.92000	1.02000	0.35250

Basic Interpretations of Numerical Summaries of Data

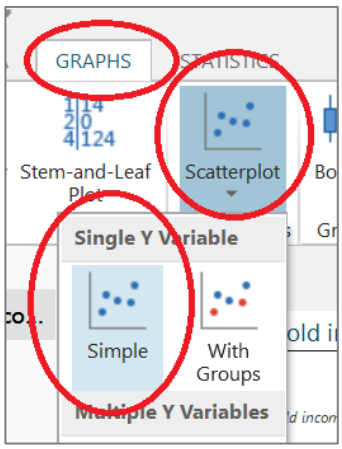
- Describe the size of your sample
  - Use "N" to know how many observations are in your sample. Minitab does not include missing values in this count.
- Describe the centre of your data
  - Use the mean to describe the sample with a single value that represents the centre of the data. Many statistical analyses use the mean as a standard measure of the centre of the distribution of the data.
  - The median and the mean both measure central tendency. But unusual values, called outliers, affect the median less than they affect the mean. When you have unusual values, you can compare the mean and the median to decide which the better measure to use. If your data are symmetric, the mean and median are similar.
- Describe the spread of your data
  - Use the standard deviation (or IQR) to determine how spread out the data are from the mean. A higher standard deviation value indicates greater spread in the data.

Task 1D

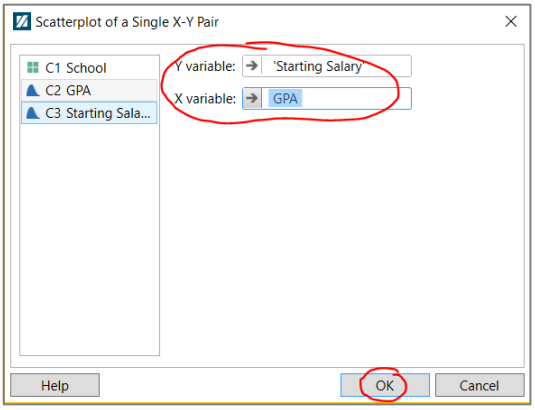
Use the dataset in "*Lab2*" Excel worksheet to construct scatterplot and compute correlation coefficient.

(I) CONSTRUCT SCATTERPLOT

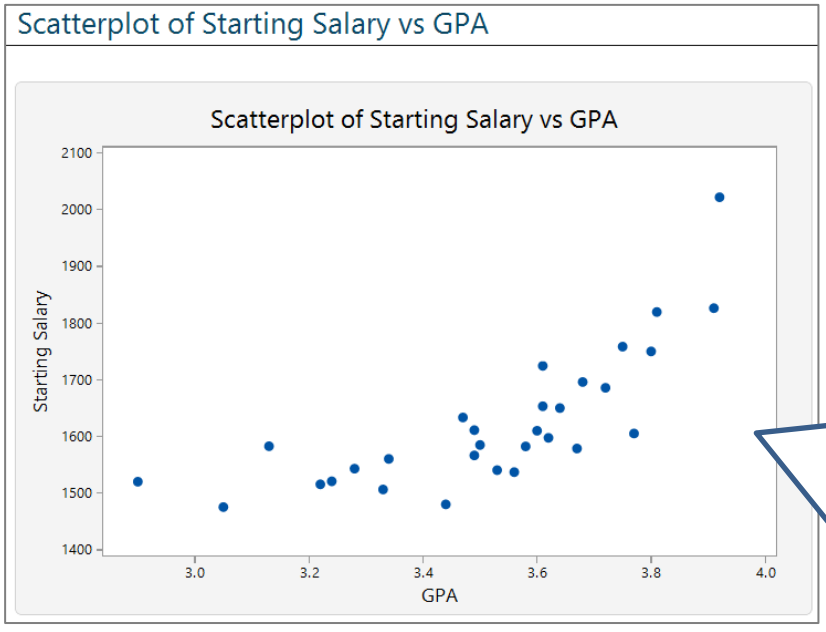
Step 1: Select **GRAPHS** > **Scatterplot** > **Single Y Variable: Simple**



Step 2: For **Y variable**, select *Starting Salary*.
For **X variable**, select *GPA*.
Click **OK**.



Step 3: The graph will be displayed in the output window.

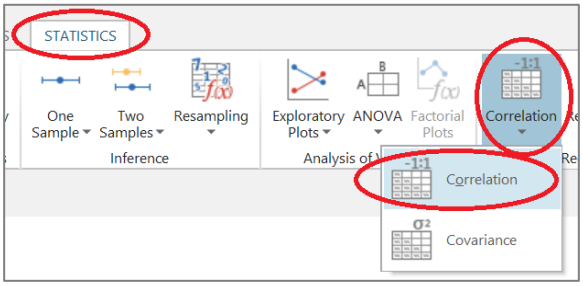


Basic Interpretations of Scatterplots

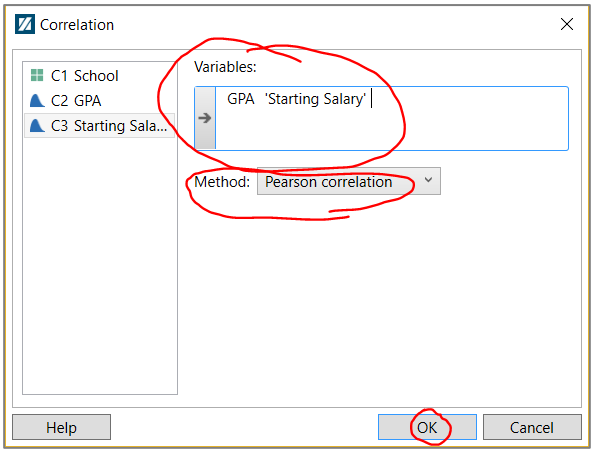
- Are the points close to an "imaginary" linear line?
- When the explanatory variable increase, does the response variable increase too or decrease?
- Hence, is this indicative of a positive or negative relationship?

(II) COMPUTE CORRELATION COEFFICIENT

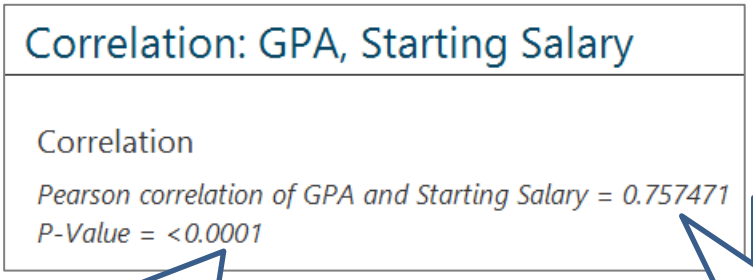
Step 1: Select **STATISTICS** > **Correlation** > **Correlation**



Step 2: For **Variables**, select *GPA* and *Starting Salary*. (Note that the order does not matter)
For **Method**, keep the default **Pearson correlation**.
Click **OK**.



Step 3: The results will be displayed in the output window.



"*P-Value*" of correlation coefficient is not covered in this module.

Basic Interpretations of Correlation Coefficients

- Does the value indicate a linear relationship?
- Does the value indicate a positive or negative relationship?
- Does the value indicate a strong, moderate or weak linear relationship?

Task 1E (OPTIONAL)

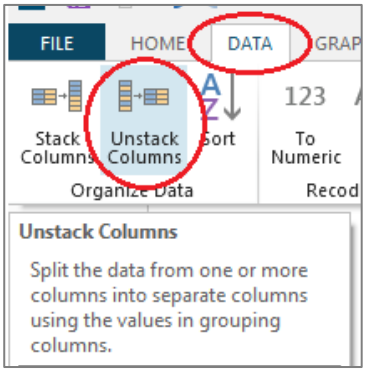
Use the dataset in "*Lab2*" Excel worksheet to stack and unstack data in Minitab Express.
(Alternatively, filter data in Excel before copying data over to Minitab Express.)

(I) UNSTACK DATA

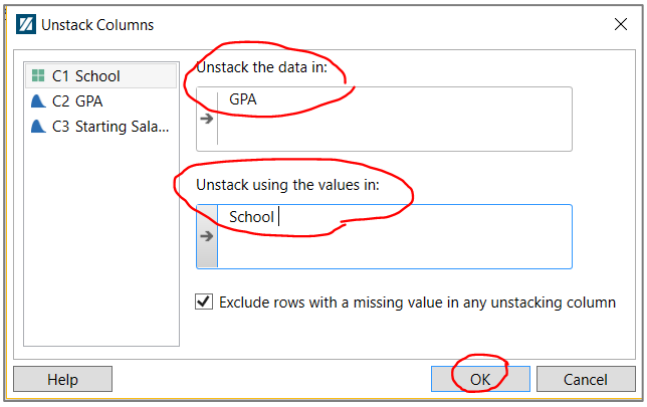
How to separate data from a single column (stacked) to different columns (unstacked)?

We will use the *Lab2* data set. Let's separate the "GPA" according to the "School".

Step 1: Select **DATA** > **Unstack Columns**.



Step 2: For **Unstack the data in**, select *GPA*.
For **Unstack using the values in**, select *School*.
Click **OK**.



Step 3: You should be able to see new columns, with "GPA" separated into different columns according to "School":

	C1	C2	C3	C4	C5	C6	C7	C8
	School	GPA	Starting Salary	GPA_CLS	GPA_EEE	GPA_MAE	GPA_SB	
1	EEE	2.90	1520.10	3.92	2.90	3.61	3.50	
2	EEE	3.81	1819.50	3.75	3.81	3.13	3.72	
3	SB	3.50	1584.90	3.64	3.34	3.47	3.22	

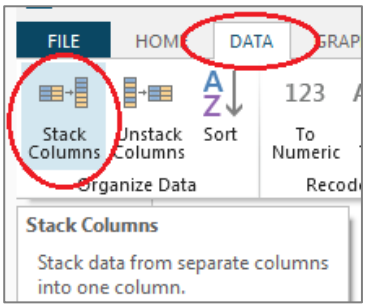
(Note: You may choose to rename the column names.)

(II) STACK DATA

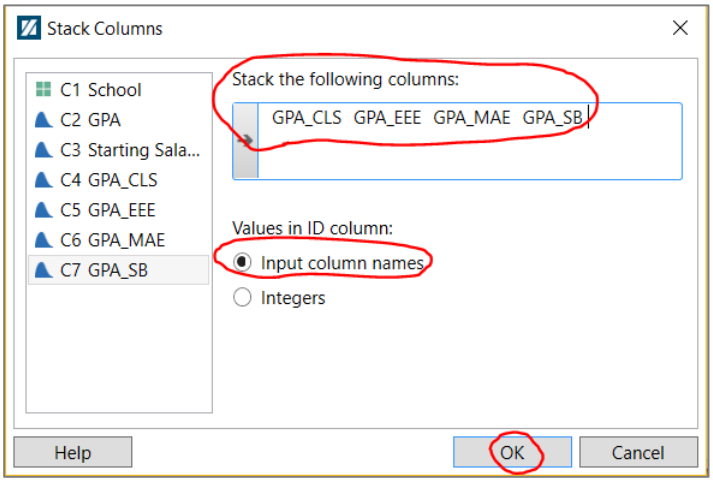
Now, how do we combine different columns into a single column?

The "GPA" according to "School" were separated into three different columns just now. Let's combine the "GPA" of different "School" back into a single column.

Step 1: Select **DATA** > **Stack Columns**.



Step 2: For **Stack the following columns**, select *C4 GPA_CLS* to *C7 GPA_SB*.
For **Values in ID column**, select *Input column names*.
Click **OK**.



Step 3: You should be able to see two new columns – an "ID" denoting the names from which the columns were combined from, and a "Stack" column that combined all the *GPA_CLS*, *GPA_EEE*, *GPA_MAE* and *GPA_SB*.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
	School	GPA	Starting Salary	GPA_CLS	GPA_EEE	GPA_MAE	GPA_SB	ID	Stack	
1	EEE	2.90	1520.10	3.92	2.90	3.61	3.50	GPA_CLS	3.92	
2	EEE	3.81	1819.50	3.75	3.81	3.13	3.72	GPA_CLS	3.75	

Task 2

Moto Automobile's would like to know if its newly developed petrol additive is useful in increasing car mileage significantly. Fifty car owners were randomly asked to include additives into their cars, of which 25 car owners were given the petrol additives and 25 others were given placebos. All the car owners were asked to diligently and carefully record their car mileage (in km) per litre of petrol used. The results are shown as follows:

Without additive (Placebos)					With additive				
7.2	7.7	6.1	11.9	9.5	7.4	7.3	7.6	12.2	9.3
8.6	10.9	7.2	6.9	15.2	9.1	10.4	6.6	6.9	15.2
5.3	8.6	10.2	8.4	9.2	5.3	9.5	9.5	8.2	9.7
9.0	8.2	13.0	15.3	8.4	8.4	7.9	12.9	15.6	8.3
9.0	5.3	11.9	8.5	11.7	9.7	4.7	11.9	7.6	12.2

(This data set can be found in "Lab2" Excel worksheet..)

Here is the comment from one of the owners:

Is there enough evidence to support this user's comment? Justify using the data given.

Formulating Questions	
Collecting Data	Sample size, n = "Additive type" is "Mileage" is
Analysing Data	
Interpreting Results	