Module 1

Prediction
- Numeric (How much? How many?)
- Classes (type A or type B?)

Detection
- Structure (How is this organized?)
- Anomaly (Is it weird behaviour?) ⇒ need to know regular structure first; check for irregularity

Decision
- Action (What should be done next)

Regression ⇒ Predictor of numeric data
Classification ⇒ Predictor of classes
Clustering ⇒ For structure detection; find "groups"
Anomaly detection ⇒ find irregularity; large deviations
Adaptive learning ⇒ for decision : Action

Structured Data
→ highly organized, easy to analyze
→ numeric / factor, time series, network

Numeric / factor
→ excel sheet, SQL databases, from sensors & devices

Categorical Data
→ if there is levels in the data it is not continuous
→ excel sheet, SQL databases, from sensors & devices

numeric

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75.0	7.2
57.5	32.8	23.5	11.8
8.6	2.1	1.0	4.8
199.8	2.6	21.2	10.6
66.1	5.8	24.2	8.6
214.7	24.0	4.0	17.4
23.8	35.1	65.9	9.2

categorical

Safety	Doors	Seats	Condition
high	4	2	unacc
med	5more	more	good
high	5more	more	vgood
high	2	2	unacc
high	2	2	unacc
low	4	more	acc
med	5more	2	unacc
high	4	4	acc
med	2	more	acc
high	4	2	unacc
low	3	4	unacc
high	3	4	unacc

Can be mixed data as well

Structured Data

⇒ Time Series Data

⇒ e.g. stocks & equity markets, weather data over time, prices & promotions

⇒ Clearly defined time axis

⇒ Numeric with Timestamp

⇒ Network data

⇒ clearly defined links (e.g. Pioneer, Yew Tee, etc.)

⇒ nodes & connections

⇒ e.g. Social Networks & Web, Transport Networks (MRT), Financial Transactions

Unstructured data
→ Highly Unorganized and contextual → context sensitive
⇒ Text, Image, Voice, Videos.

Data Science
Unstructured Data
Text Data

Highly Unorganized Data
Non-Obvious Variables
Highly Context-Sensitive
Words, Phrases, Emoticons

because ⇒ difficult to mine
E.g. hashtags, hyperlinks

Example Source
- Social Networks and Web
- Text Messages / WhatsApp
- Books, Wikis, Documents

Image Data

Highly Unorganized Data
Non-Obvious Variables
Highly Context-Sensitive
Pixels and Objects

leads to

Example Source
- Social Networks and Web
- Mobile Phone Cameras
- Blogs, Wikis, Documents

Video Data

Highly Unorganized Data
Non-Obvious Variables
Highly Context-Sensitive
Images, Frames, Objects

Example Source
- YouTube and Social Media
- Video Messages and Calls
- Mobile Phone Cameras

Voice Data

Highly Unorganized Data
Non-Obvious Variables
Highly Context-Sensitive
Voice Signals and Waves → wave forms

Example Source
- Songs and Social Media
- Microphones and Cameras
- Recordings, Announcements

- Names may include special characters → need to phrase ; separate out
- Numeric values may not always be integer
  → can be levels ; categorical
- Variables like Names may not be unique
- NaN : not a number / level
- Generation : numeric but its categorical (1-6)

Module 2

- Central Tendency: descriptive summary of a dataset through a single value that reflects the center of data distribution

- Max & Min: Range
- spread of Data: Dispertion
- Central Tendency: Mean

$$\bar{x} = \frac{\text{sum of data}}{\text{Count of data}} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- Central Tendency: Median

Mid value of dataset → sort values 50/50 ; ascending order
if odd no. of data → take middle value straight away
if even no. of data → take mean of the two middle values

spread

- Dispertion: Standard deviation

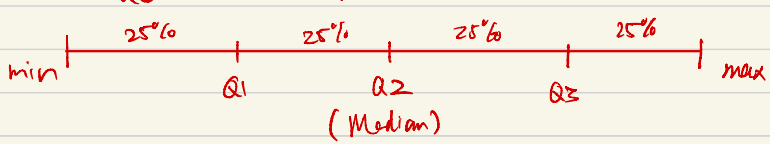
$$\sqrt{\frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots + (x_n-\bar{x})^2}{n}}$$

← Σ(deviation − mean)^2

← count of data

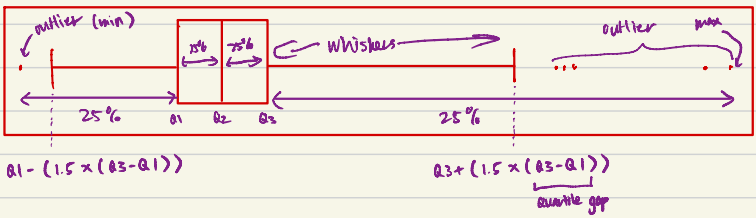
- Dispertion: Quantiles

Divide data into 25 : 50 : 25



- Statistical Summary of Data
→ table of mean, std, etc.
→ box plots

Box plots



Calculate of whiskers:

→ Q1 − (1.5 × Quartile gap) = Q1 − (1.5 × (Q3 − Q1))

→ Q3 + (1.5 × Quartile gap) = Q3 + (1.5 × (Q3 − Q1))

Outliers do not follow the norm.

Violin Plot : Box plot + Density plot

⇒ most comprehensive & compact manner

Histogram y-axis is usually the frequency

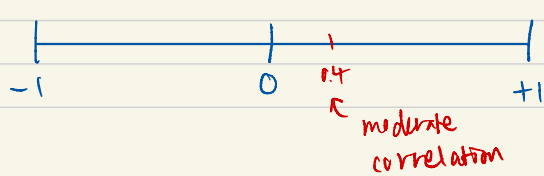
Correlation Coefficient

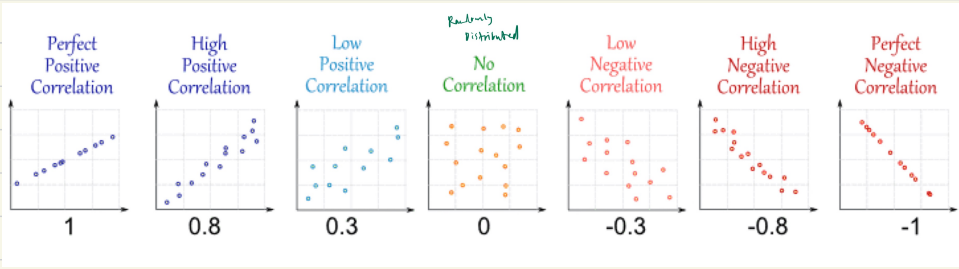
$$\rho_{xy} = \text{co-variance} \,/\, \text{St. Dev Product} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\ \sqrt{\sum (y_i - \bar{y})^2}}$$

→ No Dependence : Corr = 0 (no relationship)

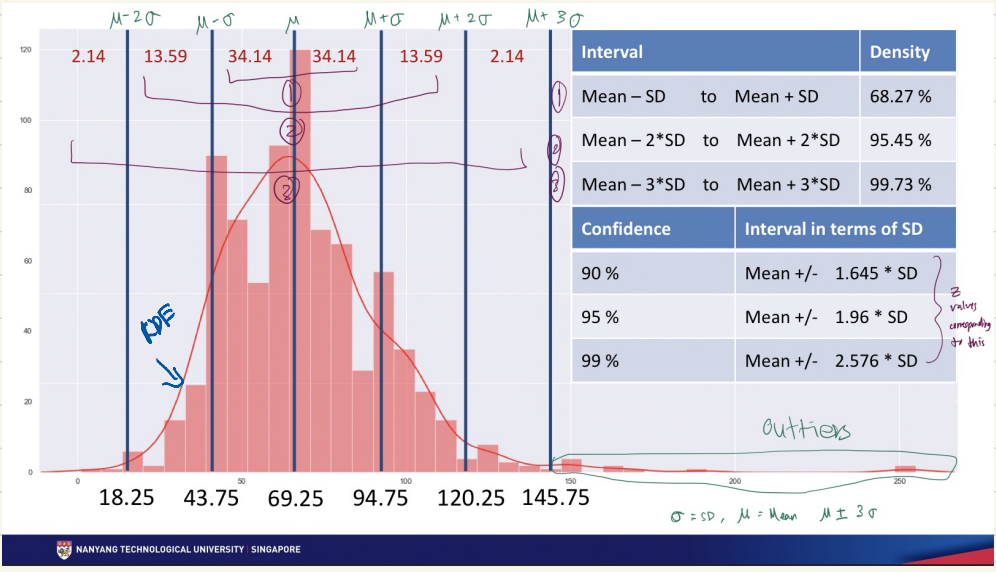
→ Perfect positive : Corr = +1 (if y ↑, x ↑)

→ Perfect Negative : Corr = −1 (in y ↑, x ↓)





KDE : smoother version of histogram



Interval			Density
Mean – SD	to	Mean + SD	68.27 %
Mean – 2*SD	to	Mean + 2*SD	95.45 %
Mean – 3*SD	to	Mean + 3*SD	99.73 %

Confidence	Interval in terms of SD
90 %	Mean +/- 1.645 * SD
95 %	Mean +/- 1.96 * SD
99 %	Mean +/- 2.576 * SD

Z values corresponding to this

Watch lecture for this

Module 3

- Machine Learning : Learning & Testing

- Prediction : Numeric → how much? how many?
  → regression
    ⇒ Model : Total = f (Variables)
    ↓ response ↓ predictions

- Supervised Learning : given something to train; train & understand the model f, from data given. → train set

- Prediction : Classes → Is it type A or type B?
  → Classification
    ⇒ Model : P (Type) = f (Variables)
    ⇒ Predict P
      ↳ Low probability : not legendary
      ↳ high probability : legendary
      ↳ only you know f, can predict

- Detection : Structure → How is this organized?
  → Can you group them?
    ↳ smaller subsections (Clustering)
  → Clustering
    ⇒ Similarities in data points
    ⇒ Unsupervised learning : not indication of name of group but asked to group
    ⇒ Grouping depends on "distance" → find optimal groups
      ↳ justify interpretation

- Detection : Anomaly → is it weird behaviour
  - ⇒ unsupervised learning
  - → really different from other data e.g. really strong pokemon
  - ⇒ identify the differences.

HP	Attack	Defense	Sp. Atk	Sp. Def	Speed
100	125	52	105	52	71
70	95	85	55	65	70
46	57	40	40	40	50
50	92	108	92	108	35
75	75	130	75	130	95
75	65	55	65	55	69
95	23	48	23	48	23
70	80	70	80	70	110
80	120	130	55	65	45
40	50	45	70	45	70

Labels

Total	Legendary
505	False
440	False
273	False
485	False
580	True
384	False
260	False
480	False
495	False
320	False

Data Science

Machine Learning

Supervised Learning

Regression

Classification

Given training data & labels

Un-Supervised Learning

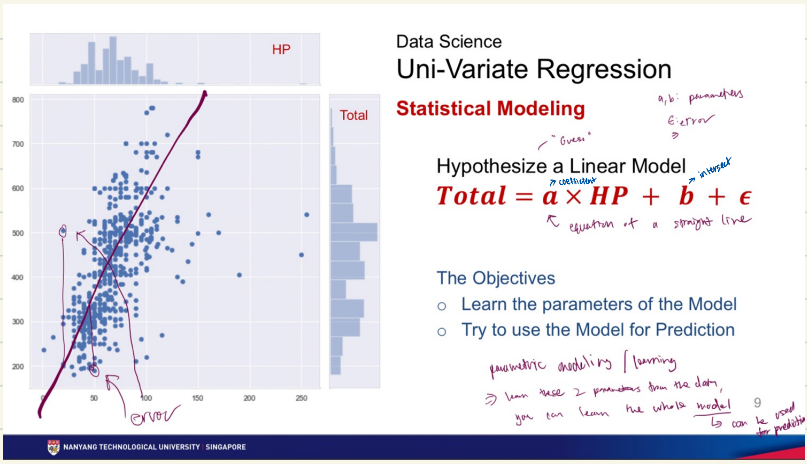
Clustering

Anomaly Detection

nothing that tells you exactly what to do, but want to find patterns

Photos : Pokedex Website | https://www.pokemon.com/us/pokedex/

→ Given whole dataset, no label, required to find structure



Cost Function to Minimize

$$J(a,b) = \sum (\underbrace{Total}_{\text{actual}} - \underbrace{a \times HP - b}_{\text{predicted}})^2$$

RSS = Residual Sum of Squares

Steps in Linear Regression

1. Guess parameters a & b in the model
2. predict the values of Total in Train Data
3. Calculate the Errors compared to actual (choose specific cost function)
4. Tune parameter to minimize errors

RSS / n

Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum \underbrace{(Total - a \times HP - b)^2}_{RSS}$$

* Lower MSE → better model

Explained Variance (R^2)

$$R^2 = 1 - \frac{\sum (Total - a \times HP - b)^2}{\sum (Total - \overline{Total})^2}$$

← RSS

← Total sum of squares (TSS)

↑ max 1.

$$= 1 - \frac{MSE}{VAR}$$

$$0 \leq RSS \leq TSS$$

* Higher R^2 → better Model

* $\frac{1}{n} TSS = Var(Total)$

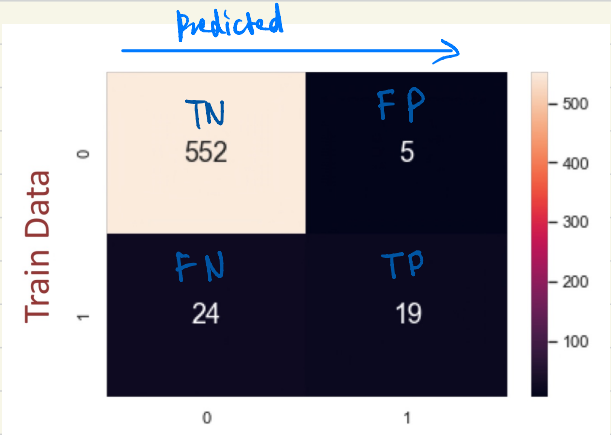
Chapter 4

- Lower gini → more confident
  → confidence depends on the number of data in the zone.

Gini Index

$$gini = \frac{x}{n}\left(1 - \frac{x}{n}\right) + \frac{y}{n}\left(1 - \frac{y}{n}\right)$$

- x: value
- $\frac{x}{n}$: one class
- $1 - \frac{x}{n}$: error on classification
- $\frac{y}{n}$: another class
- n: total



Accuracy : $\frac{TP + TN}{Total}$

False Positive Rate : $\frac{FP}{FP + TN}$

True Positive Rate : $\frac{TP}{FN + TP}$

False Negative Rate : $\frac{FN}{TP + FN}$

True Negative Rate : $\frac{TN}{FP + TN}$

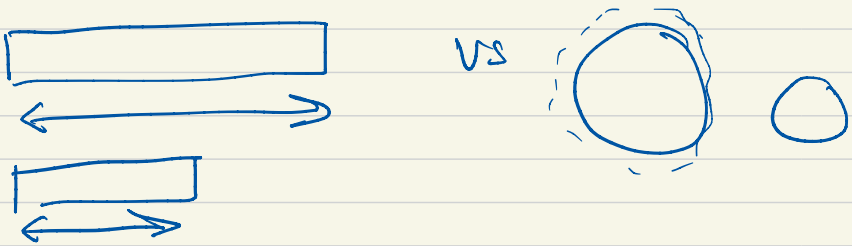
Module 5

- convey your story

→ effectiveness

⇒ more readily perceived

→ length conveys more information than area

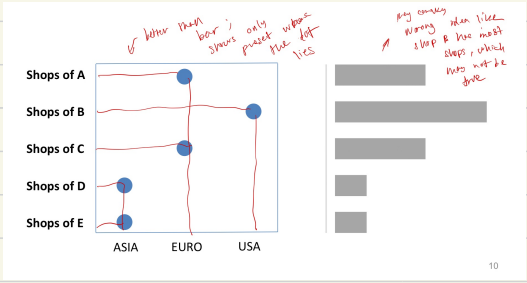


- establish credibility

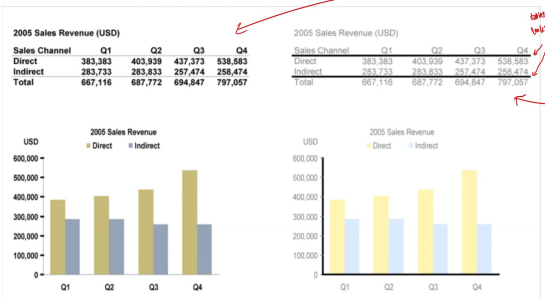
→ expressiveness

⇒ express all facts & only the facts

eg. don't truncate data just cause it seems neater



convey your story	establish credibility
Effectiveness	**Expressiveness**
Use encodings that people decode better.	Tell the truth and nothing but the truth.
Better means more accurate and faster.	Do not lie, and do not lie by omission.



Data Ink vs. Non-Data Ink

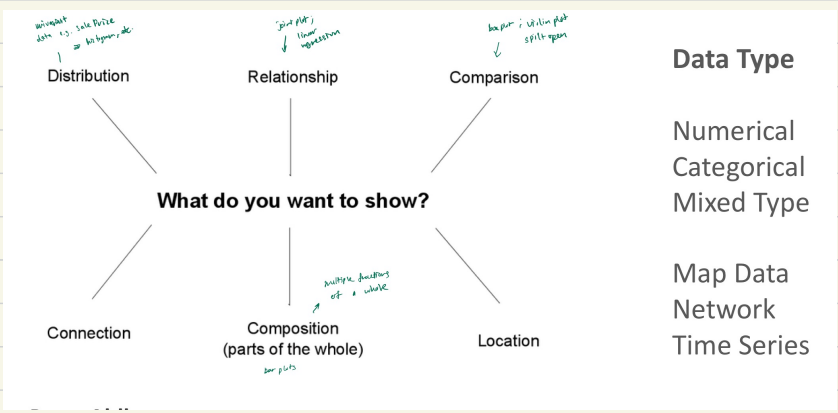
The data-ink ratio must be as high as possible.

- Tufte

Interpretation

- realizing the context and the most effective way to present your data.

Above all else, show the Data



Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit; climate change

Diverging bar

A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar

Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine

Splits a single value into two contrasting components (eg male/female).

Surplus/deficit filled line

The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy

Scatterplot

The standard way to show the relationship between two continuous variables, each of which has its own axis.

Column + line timeline

A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot

Usually used to show how the relationship between 2 variables has changed over time.

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap

A good way of showing the patterns between 2 categories of data, less effective at showing fine differences in amounts.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

Ordered bar

Standard bar charts display the ranks of values much more easily when sorted into order.

Ordered column

See above.

Ordered proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot

Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope

Perfect for showing how ranks have changed over time or vary between categories.

Lollipop

Lollipops draw more attention to the data value than standard bar/column and can also show rank and value effectively.

Bump

Effective for showing changing rankings across multiple dates. For large datasets, consider grouping lines using colour.

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (age/sex) distribution, revealing inequality

Histogram

The standard way to show a statistical distribution – keep the gaps between columns small to highlight the 'shape' of the data.

Dot plot

A simple way of showing the change or range (min/max) of data across multiple categories.

Dot strip plot

Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Barcode plot

Like dot strip plots, good for displaying all the data in a table, they work best when highlighting individual values.

Boxplot

Summarise multiple distributions by showing the median (centre) and range of the data

Violin plot

Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple average).

Population pyramid

A standard way for showing the age and sex breakdown of a population distribution; effectively, back to back histograms.

Cumulative curve

A good way of showing how unequal a distribution is; y axis is always cumulative frequency, x axis is always a measure.

Frequency polygons

For displaying multiple distributions of data. Like a regular line chart, best limited to a maximum of 3 or 4 datasets.

Beeswarm

Use to emphasise individual points in a distribution. Points can be sized to an additional variable. Best with medium-sized datasets

Change over Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries: Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses
Share price movements, economic time series, sectoral changes in a market

Line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column

Columns work well for showing change over time – but usually best with only one series of data at a time.

Column + line timeline

A good way of showing the relationship over time between an amount (columns) and a rate (line).

Slope

Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

Area chart

Use with care – these are good at showing changes to total, but seeing change in components can be very difficult.

Candlestick

Usually focused on day-to-day activity, these charts show opening/closing and high/low points of each day.

Fan chart (projections)

Use to show the uncertainty in future projections – usually this grows the further forward to projection.

Connected scatterplot

A good way of showing changing data for two variables whenever there is a relatively clear pattern of progression.

Calendar heatmap

A great way of showing temporal patterns (daily, weekly, monthly) – at the expense of showing precision in quantity.

Priestley timeline

Great when date and duration are key elements of the story in the data.

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Example FT uses
Commodity production, market capitalisation, volumes in general

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

See above. Good when series and labels have long category names.

Paired column

As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar

See above.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictogram)

Excellent solution in some instances – use only with whole numbers (do not slice off an arm to represent a decimal).

Lollipop

Lollipop charts draw more attention to the data value than standard bar/column – does not have to start at zero (but preferable).

Radar

A space-efficient way of showing value of multiple variables- but make sure they are organised in a way that makes sense to reader.

Parallel coordinates

An alternative to radar charts – again, the arrangement of the variables is important. Usually benefits from highlighting values.

Part-to-whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

Stacked column/bar

A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie

A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut

Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg total).

Treemap

Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments.

Voronoi

A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Arc

A hemicycle, often used for visualising parliamentary composition by number of seats.

Gridplot

Good for showing % information, they work best when used on whole numbers and work well in small multiple layout form.

Venn

Generally only used for schematic representation.

Waterfall

Can be useful for showing part-to-whole relationships where some of the components are negative.

Spatial

Aside from locator maps only used when precise locations or geographical patterns in data are more important to the reader than anything else.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

Basic choropleth (rate/ratio)

The standard approach for putting data on a map – should always be rates rather than totals and use a sensible base geography.

Proportional symbol (count/magnitude)

Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map

For showing unambiguous movement across a map.

Contour map

For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values

Equalised cartogram

Converting each unit on a map to a regular and equally-sized shape – good for representing voting regions with equal value.

Scaled cartogram (value)

Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density

Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map

Grid-based data values mapped with an intensity colour scale. As choropleth map – but not snapped to an admin/political unit.

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrants, lawsuits, information; relationship graphs.

Sankey

Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

Waterfall

Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord

A complex but powerful diagram which can illustrate 2-way flows (and net winner) in a matrix.

Network

Used for showing the strength and inter-connectedness of relationships of varying types.

http://ft.com/vocabulary

Visual vocabulary

Cx1115 : Theory Quiz : Sample

- There are 5 questions in this quiz. Total points are 10. Marks for the questions may be different. **Attempt ALL.**
- In there is any confusion regarding the correct answer(s), choose the one(s) that seem(s) the most appropriate.
- You are allowed 7 minutes to complete the quiz. All questions will appear together on a single page (scrolling).
- Please remember to "Save and Submit" the quiz, once you are done. You will see a timer and relevant warnings.

Rules

You are allowed to use your Pen/Pencil and Calculator. You will be provided with a blank sheet of paper for rough work. Possession of any personal gadget or electronic device (other than calculator) is prohibited during the quiz. Please keep all your belongings in a bag, and store away for the entire duration of the test. Please talk to the Lab In-Charge and your TA immediately if you face any problem whatsoever with the Lockdown Browser or with accessing the quiz questions.

Questions

1. Which kind of Data Science problem does Spam Filters solve in your email service? They generally decide if an email is a genuine email or spam.

   ☐ Regression – numeric
   ☐ Clustering – find "groups"
   ☐ Classification – type A or type B?
   ☐ Forecasting – predict future event or trend

   (1 marks)

2. What is the most appropriate characterization of the MRT Map of Singapore?

   ☐ Structured Categorical Data
   ☐ Numeric Time Series Data
   ☐ Unstructured Text Data
   ☐ Network or Graph Data

   (1 marks)

3. What proportion of a data is greater than or equal to its Third Quartile (Q3)?

   ☐ 25% ☐ 50% ☐ 75% ☐ 95%

   (2 marks)

4. If after Linear Regression, the Mean Squared Error (MSE) is 25 and the Variance of the Response Variable is 100, what is the value of the Explained Variance (R^2)?

   ☐ 0.25 ☐ 0.50 ☐ 0.75 ☐ 1.00

   (4 marks)

5. Which mode of visualization is the best for comparing the prices of two products?

   ☐ Bubble-Chart, with Prices as Area
   ☐ Pie-Chart, with Prices as Sectors
   ☐ Bar-Plot, with Prices as Length
   ☐ Heatmap, with Prices as Colors

   (2 marks)

Quizzes : Module 1 Part 1

Data Analytic Thinking

In the LAMS Sequence, you have learned the theory behind this module. It is also expected that you have attempted the quizzes embedded within the LAMS Sequence, and have used the “unlimited attempts” opportunity to score 100%. Here are the quiz questions, consolidated with their answers and corresponding feedback. This is for your after-LAMS revision.

Question 1

What is the first step in the Data Science pipeline?

Answer Choice	Verdict	Explanation
Problem Formulation	Correct	Yup! You can't even start a Data Science process without having a well-formed problem.
Machine Learning	Wrong	Really? How do you know what to solve using Machine Learning? Think again.
Data Visualization	Wrong	Sometimes it good to visualize the data, but only when you don't even have a problem at hand. Think again.
Statistical Inference	Wrong	Nopes! You can't infer things from the data unless you know what is the problem at hand. Think again.

Reference Module 1 Topic 1 : What is Data Science? Slide 4 and Slide 10

Question 2

Which step in Data Science relies heavily on Algorithmic Optimization?

Answer Choice	Verdict	Explanation
Machine Learning	Correct	Yes, this is correct. In fact, most Machine Learning algorithms are clever optimizations with respect to a "cost function".
Data Collection	Wrong	Data collection is a huge part of Data Science, but we are still not in the zone of algorithms. Think again.
Data Visualization	Wrong	Sometimes, visualization requires some clever algorithms to handle big data. But this is not the main algorithmic sector.
Digital Storytelling	Wrong	Nopes! There's a lot of artistic elements (and programming) in storytelling, but algorithmic optimization is not a part of it.

Reference Module 1 Topic 1 : What is Data Science? Slide 7

Question 3

Match the Data Science problems with their respective Problem Types.

Problem Definition	Problem Type	Explanation
How many students will visit K-Cuts (the hair-cutting place) at North Spine this Friday?	Numeric Prediction	That's pretty intuitive. You are literally predicting the "number" of students.
Will K-Cuts (the hair-cutting place) at North Spine remain open at 11 am this Sunday?	Prediction of Class	The two classes are "YES" and "NO", and you are predicting which one.

Reference Module 1 Topic 2 : Data Science Problems Slide 3 and Slide 4

Question 4

Suppose you want to buy a Condominium in Singapore, and you want to find the best locality. What type of problem would this be in a Data Science context? Choose the most appropriate answer.

Answer Choice	Verdict	Explanation
Both Numeric and Class Prediction	Correct	That's correct! There are several factors associated with the problem, and you may encounter both numeric predictions (like estimating the price of the place) and class predictions (like choosing the best neighborhood in the city). So, both of them.
Prediction of a Numeric Variable	Wrong	Somewhat correct, as you will probably need to estimate/predict some numeric values, like the price of the place. However, there's more to a Condominium than just its price. Think again.
Prediction of a specific Class	Wrong	Somewhat correct, as you will probably need to estimate/predict some class values, like the best neighborhood in the city. However, there's more to a Condominium than just its neighborhood. Think again.
Neither Numeric nor Class Prediction	Wrong	Well, we do not know what else it could be. There are two major zones for prediction. Think again.

Reference Module 1 Topic 2 : Data Science Problems Slide 3 and Slide 4

Question 5

Match the Data Science problems with their respective Problem Types.

Problem Definition	Problem Type	Explanation
How many "types" of product does the Prime Supermarket at North Spine carry?	Detection of Structure in Data	As you do not know “what” the products are, you can’t perform classification. You can only look at the pattern of products, and detect a structure within, using clustering.
Is there a product in the Prime Supermarket of North Spine which does not belong there?	Detection of Anomaly in Data	As you do not always know the exact characterization of products, you can’t perform classification. More appropriate will be to find the “weird” product within all products, as an anomaly.

Reference Module 1 Topic 2 : Data Science Problems Slide 5 and Slide 6

Question 6

Suppose you want to design a Computer Game where the computer plays chess against a human player. What category of Data Science will the problem belong? Choose the most appropriate answer.

Answer Choice	Verdict	Explanation
Taking Adaptive Decisions	Correct	Yup! That's correct. We are in the zone of Artificial Intelligence in this case, where learning means adaptive learning.
Identifying Structure in Data	Wrong	There is of course some structure in any game. But this is not the main task for the computer to master in case of Chess. Think again.
Identifying Anomalies in Data	Wrong	Anomaly detection is probably not the most appropriate task in playing Chess. Think again.

Reference Module 1 Topic 2 : Data Science Problems Slide 7

Question 7

Prediction of Numeric Values (Regression) can be accomplished using …

Answer Choice	Verdict	Explanation
Any of the models mentioned below	Correct	That's right. Almost any model can do anything in practice.
Only the Linear Regression Models	Wrong	Models are not limited -- almost any model can do anything in practice. Think again.
Only the Tree-based Models	Wrong	Somewhat correct, as you will probably need to estimate/predict some class values, like the best neighborhood in the city. However, there's more to a Condominium than just its neighborhood. Think again.
Only Neural Network Models	Wrong	Well, we do not know what else it could be. There are two major zones for prediction. Think again.

Reference Module 1 Topic 3 : Data Science Solutions Slide 5

Question 8

Suppose you want to know if "that specific boy/girl" will go out with you. You model it as a Classification Problem, and use the data from all your previous attempts at dating to learn from. What will you learn at the end?

Answer Choice	Verdict	Explanation
The probability of "that specific boy/girl" going out with you.	Correct	That's correct. Any classification problem will produce a probability, and not just a class.
Whether "that specific boy/girl" will go out with you.	Wrong	Yes, that's the goal, right? But remember, classification generally returns probability. Think again.
Whether "that specific boy/girl" will go out with your arch-enemy.	Wrong	Well, you would want to know that too. But that's not the problem we are solving. Think again.
The probability of "that specific boy/girl" already having a girl/boyfriend.	Wrong	Hmm, tricky! Very hard to predict that from the data we have. Try again.

Reference Module 1 Topic 3 : Data Science Solutions Slide 7

Question 9

What is the most crucial concept in identifying Clusters in data?

Answer Choice	Verdict	Explanation
The notion of "distance" to identify close and far points.	Correct	Yes, indeed. The notion of distance is the most crucial aspect of identifying groups or clusters. We will see more of it later.
Techniques of visualizing the data to identify groups.	Wrong	Visualization is extremely important, no doubt, but in high dimensional datasets, you will not be able to do that effectively.
The notion of "anomalies" to weed out extreme points.	Wrong	Weeding out anomalies is a different ball-game altogether. Think again.
The expertise of a "domain expert" to identify clusters.	Wrong	You do need domain expertise, no doubt. But often, machine learning reveals more than an expert knows. Think again.

Reference Module 1 Topic 3 : Data Science Solutions Slide 10

Question 10

How would you spot the next "Joseph Schooling", that is, the next biggest star in Swimming from Singapore?

Answer Choice	Verdict	Explanation
Anomaly Detection across all Swimmers at the school-level.	Correct	This is probably the best approach, as of course, Joseph Schooling is an exception within swimmers from Singapore.
Clustering groups across all Athletes at the school-level.	Wrong	Clustering will surely help you with spotting the general structure in swimmers of Singapore. But Schooling is an exception, isn't it? Think again.
Predicting the best lap-times of all Swimmers in Singapore.	Wrong	You can surely do that. However, you will have to look for exceptions at the end of the day. Think again.
Predicting the School/JC/Poly to produce the best Swimmer.	Wrong	This classification can work, but it's not just the School/JC/Poly that matters. We will need to find these exceptional talent, as is Joseph Schooling. Think again.

Reference Module 1 Topic 3 : Data Science Solutions Slide 13

Quizzes : Module 1 Part 2

The Data Pipeline

In the LAMS Sequence, you have learned the theory behind this module. It is also expected that you have attempted the quizzes embedded within the LAMS Sequence, and have used the “unlimited attempts” opportunity to score 100%. Here are the quiz questions, consolidated with their answers and corresponding feedback. This is for your after-LAMS revision.

Question 1

Which of the variables (out of the four options) in this dataset are "Categorical" (with Factors/Levels) in nature? Choose all answers that seem correct to you.

Year	Sex	Level of school	Age	No of vice principals (numbers)
2019	MF	PRIMARY	30 - 34	1
2019	F	PRIMARY	30 - 34	1
2019	MF	SECONDARY	30 - 34	2
2019	F	SECONDARY	30 - 34	0
2019	MF	PRE-UNIVERSITY	30 - 34	1
2019	F	PRE-UNIVERSITY	30 - 34	1
2019	MF	PRIMARY	35 - 39	23
2019	F	PRIMARY	35 - 39	19
2019	MF	SECONDARY	35 - 39	37
2019	F	SECONDARY	35 - 39	20
2019	MF	PRE-UNIVERSITY	35 - 39	7
2019	F	PRE-UNIVERSITY	35 - 39	3

Answer Choice	Verdict	Explanation
Sex	Correct	Yes, of course. This is a categorical variable.
Level of School	Correct	Yes. It has levels. Naturally a categorical variable.
Age	Correct	Age is generally numeric, but in this case, it is broken up in classes or bands. Hence, categorical.
No of Vice Principals	Wrong	This is numeric for sure.

Reference Module 1 Topic 4 : Structured Data in Practice Slide 3 and Slide 4 and Slide 5

Question 2

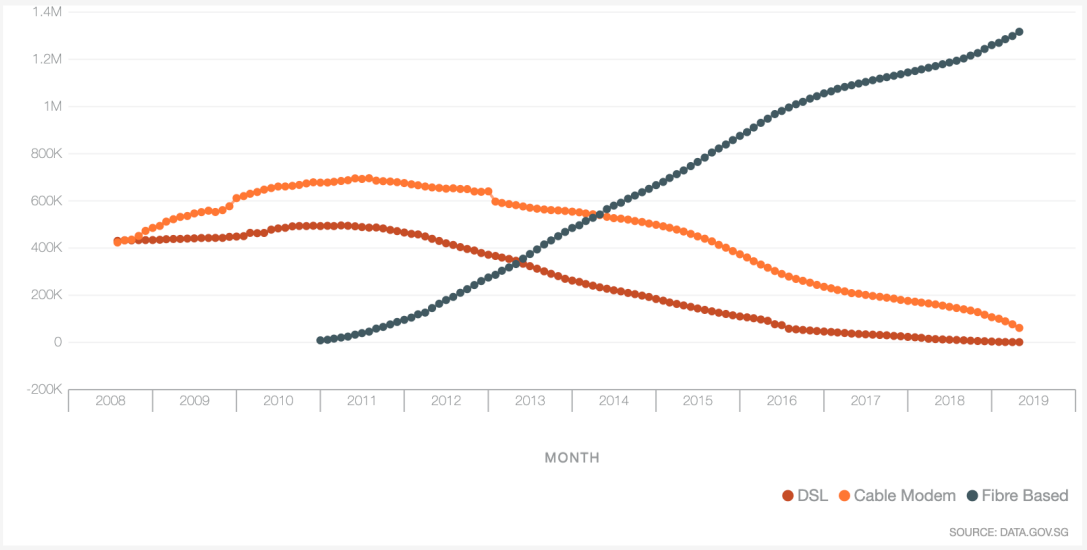
Suppose you find a variable (column) in a structured dataset with ten values : [1, 0, 1, 1, 1, 0, 0, 1, 0, 0]. Which one of the following conclusions can you draw from this information? Choose all answers that seem correct to you.

Answer Choice	Verdict	Explanation
Half of the values (the 1's) are different from the other half (the 0's).	Correct	This is correct, as the levels for categorical can also be "different", as is true for numeric values. But remember, you can't assume some variable is numeric if their values look like numbers. Be careful.
The average value (or mean) of the variable is 5/10 = 1/2 = 0.5.	Wrong	Are you assuming that the variable is "Numeric", as the values are 0 and 1. What if it's gender, with 0 = Male and 1 = Female? Would Mean make sense in that case? Think again.
Half of the values (the 1's) are greater than the other half (the 0's).	Wrong	Are you assuming that the variable is "Numeric", as the values are 0 and 1. What if it's gender, with 0 = Male and 1 = Female? Would "greater" or "lower" make sense in that case? Think again.
The maximum value of the variable is 1, and the minimum value is 0.	Wrong	Are you assuming that the variable is "Numeric", as the values are 0 and 1. What if it's gender, with 0 = Male and 1 = Female? Would "maximum" or "minimum" make sense in that case? Think again.

Reference Module 1 Topic 4 : Structured Data in Practice Slide 3 and Slide 4 and Slide 5

Question 3

The following Time Series presents the Monthly number of Broadband Subscriptions across Singapore. If we take every datapoint in this dataset, and subtract the previous datapoint from it, the resulting dataset will be:



Answer Choice	Verdict	Explanation
Time Series, with different interpretation and labels for the time axis	Correct	Yes, correct. The data format is still a time series. Just the time axis is now "difference" in time.
Time Series, with the same interpretation and labels for the time axis	Wrong	It is a time series, no doubt, but the time axis has changed. Think again.
Structured Data of type "Numeric", but NOT a Time Series any more	Wrong	Think again. Time series means a numeric value varying with time. Isn't it the same?
Structured Data of type "Mixed", but NOT a Time Series any more	Wrong	Think again. Time series means a numeric value varying with time. Isn't it the same?

Reference Module 1 Topic 4 : Structured Data in Practice Slide 6

Question 4

In which of the following ways can you interpret Facebook data as a Network (or Graph)? Note : Multiple answers may be correct. Choose all answers that seem correct to you.

Answer Choice	Verdict	Explanation
Facebook Profiles as nodes and Facebook Friendships (yes/no) as links	Correct	Yes, right. This is of course one interpretation. Two nodes (profiles) are connected if they are "friends" on Facebook.
Facebook Profiles as nodes and Number of Common Interests as links	Correct	Yes, right. This is of course one interpretation. Two nodes (profiles) are connected if they have "common interests" on Facebook, and the number of common interests become the weight of the edge in the network/graph.
Facebook Profiles as nodes and Number of Individual Friends as links	Wrong	Think again. Number of individual friends is a property of the nodes independent of one another. How does it affect any sort of connection or edge between them? For graphs, we need something to define an edge.
Facebook Profiles as nodes and Number of Photos Uploaded as links	Wrong	Think again. Number of individual photos is a property of the nodes independent of one another. How does it affect any sort of connection or edge between them? For graphs, we need something to define an edge.

Reference Module 1 Topic 4 : Structured Data in Practice Slide 7

Question 5

Suppose you are designing a Survey Questionaire to collect the feedback of your fellow students about this course. If you want to obtain a Structured Dataset of Mixed type as a result of this survey, what type of questions can you include in the Questionaire?

Answer Choice	Verdict	Explanation
Multiple Choice Questions with fixed number of options	Correct	Yes of course. The responses will be Categorical. But you need Numeric too.
Rating for the Course, within a specific range (say 0 to 10)	Correct	Yes of course. The responses will be Numeric. But you need Categorical too.
Detailed feedback/comments on how to improve the course	Wrong	Not quite. The responses will be Unstructured Text. Think again.

Reference Module 1 Topic 5 : Unstructured Data in Practice Slide 2

Quizzes : Module 2 Part 1

Statistics and Visualization

In the LAMS Sequence, you have learned the theory behind this module. It is also expected that you have attempted the quizzes embedded within the LAMS Sequence, and have used the “unlimited attempts” opportunity to score 100%. Here are the quiz questions, consolidated with their answers and corresponding feedback. This is for your after-LAMS revision.

Question 1

Suppose the Male vs Female ratio in a specific class of NTU is 70% vs 30%, whereas you know that the Male vs Female ratio in Singapore is 960 against 1000. Do you think that the selection process for admission was biased towards male applicants?

Answer Choice	Verdict	Explanation
I can't be sure, unless I get the data for the applications, and know the gender ratio within applicants.	Correct	This is more accurate. Once you know the applicants' profile and ratio, you can judge the selection process better.
Possibly, as gender ratio in the specific class is unexpectedly different from that in the country.	Wrong	Think again. Was the proportion of applicants to the specific program exactly the same as the ratio in Singapore? If not, then you can't conclude this.
No. As the gender ratio in the specific class is same as the gender ratio in that specific field of study across Singapore.	Wrong	Yes, you are thinking in the right track. But is the field of study everything? Think again. Was the proportion of applicants to the specific program exactly the same as the ratio in the field of study? If not, then you can't conclude this.
No. As the gender ratio in the specific class is same as the gender ratio in this field across all JCs and Polys.	Wrong	Yes, you are thinking in the right track. But is the JC or Poly background everything? Think again. Was the proportion of applicants to the specific program exactly the same as the ratio in the JCs or Polys? If not, then you can't conclude this.

Reference Module 2 Topic 1 : Uni-Variate Statistics No specific slide. The overall concept matters.

Question 2

Suppose the average household income of Singapore is SGD 12,000 per month, while the average household income of USA is SGD 15,000 per month. What can you conclude?

Answer Choice	Verdict	Explanation
Can't compare the household income distributions of the two countries with just this information.	Correct	True. You do not have enough data yet. You will need Median, Standard Deviation and Number of Households to know more.
The top 50% households in USA earn more per month than the top 50% households in Singapore.	Wrong	Do you know the Median in either case? If not, then you DO NOT know about top 50% of either country. Think again.
Household income inequality in USA is significantly more than the income inequality in Singapore.	Wrong	Do you know the Standard Deviation in either case? If not, then you DO NOT know about income inequality or spread of either country. Think again.
Total household income across USA is higher than the total household income across Singapore.	Wrong	Do you know the total number of households in either case? If not, then you DO NOT know about total household income of either country. Think again.

Reference Module 2 Topic 1 : Uni-Variate Statistics Slide 6

Question 3

Suppose you have a dataset X, with n numeric values in total. Which of the following values will be the largest?

Answer Choice	Verdict	Explanation
Impossible to tell. Depends on the actual data.	Correct	True. You can't tell without looking at the data. In fact, you can come up with example datasets where each of the other items will be the greatest.
Mean	Wrong	Think again. Can you say for sure, without looking at the actual data? Can't mean be 0 with a high standard deviation?
Mean Absolute Deviation	Wrong	Think again. Can you say for sure, without looking at the actual data? What if all the values are the same?
Standard Deviation	Wrong	Think again. Can you say for sure, without looking at the actual data? What if all the values are the same?

Reference Module 2 Topic 1 : Uni-Variate Statistics Slide 6 and Slide 7

Question 4

Suppose the average household income of Singapore is SGD 12,000 per month, and the standard deviation of the household income is SGD 3,000 p/m. What can you conclude?

Answer Choice	Verdict	Explanation
None of the other statements are valid under all cases, as it depends on various other factors.	Correct	True. You DO NOT know enough about the distribution yet. Only Mean and Standard Deviation is not enough, unless it is truly Normal Distribution. And you can't assume that.
More than 50% of the households in Singapore earn between SGD 9,000 to SGD 15,000 per month.	Wrong	You are assuming a Normal Distribution, implicitly. Can you assume that? Think again.
Maximum household income inequality in Singapore is no more than SGD 6,000 per month.	Wrong	Not true. You just know that SGD 6,000 per month is twice the standard deviation. You don't know the minimum or maximum.
Minimum and maximum possible household income in Singapore is SGD 3,000 and SGD 21,000.	Wrong	Not true. You can't say this without knowing the actual minimum and maximum household income.

Reference Module 2 Topic 1 : Uni-Variate Statistics Slide 6 and Slide 7

Question 5

Suppose that the mean score of your class is 75, the median score of your class is 77, the standard deviation of the scores is 5, and your own score is 81. What can you conclude?

Answer Choice	Verdict	Explanation
You definitely have a score better than 50% (or more) students in your class.	Correct	True, as you are above the Median for sure.
Definitely more than 50% students in the class scored above-average marks.	Correct	True, as the Median is higher than the Mean.
You definitely scored better than 84% (or more) students in your class.	Wrong	Not sure, unless you know that it is a Normal Distribution (you can't assume that).
Your score is above-average, but may be in the lower 50% students in the class.	Wrong	Can't be, as you scored above the Median.

Reference Module 2 Topic 1 : Uni-Variate Statistics Slide 6 and Slide 7 and Slide 8

Question 6

Suppose the median household income in Singapore is SGD 9,000, and the quartiles are Q1 = SGD 4,500 and Q3 = SGD 10,500. What can you infer from this data?

Answer Choice	Verdict	Explanation
70% or more of the households in Singapore earn below SGD 10,500 per month.	Correct	True, as 75% of the data must lie below the third quartile.
25% or more households in Singapore earn more than SGD 10,000 per month.	Correct	True, as 25% of the data must lie above the third quartile.
70% or more of the households in Singapore earn between SGD 4,500 to SGD 10,500 per month.	Wrong	Not true, as only 50% of the data lies within the first and third quartiles.
None of the households in Singapore earn more than SGD 1 Million per month.	Wrong	You don't know that. There may always be outliers, and quartiles do not tell you anything about them.

Reference Module 2 Topic 1 : Uni-Variate Statistics Slide 8 and Slide 9

Question 7

Suppose the median household income in Singapore is SGD 9,000, and the quartiles are Q1 = SGD 4,500 and Q3 = SGD 10,500. What can you infer about the outliers?

Answer Choice	Verdict	Explanation
Household income above SGD 20,000 may be considered as outliers (abnormally high) in this data.	Correct	That's quite far indeed. It's higher than (Q3 + 1.5 * IQR), where Inter-Quartile Range IQR = Q3 – Q1. That may be considered outlier as per standard norms.
Household income above SGD 10,500 may be considered as outliers (abnormally high) in this data.	Wrong	Nope. There are of course 25% of houses above third quartile (Q3), and they are not all outliers.
Household income less than SGD 1,000 may be considered as outliers (abnormally low) in this data.	Wrong	Not quite. It's still within the Q1 to (Q1 – 1.5 * IQR) interval on the lower side, and can’t be termed outliers.
We can't say that any of the other answers are true unless we know the average household income.	Wrong	Some justification about outliers can be drawn from the quartile gaps. Think again.

Reference Module 2 Topic 2 : Uni-Variate Visualization Slide 5

Question 8

Suppose you want to know where you stand in terms of annual income in Singapore. Which one of the following statements would you support in this context? Select all that you think are right.

Answer Choice	Verdict	Explanation
Knowing just the average income does not help a lot, as the average may be affected by really high/low outliers.	Correct	True. Outliers affect the average or the mean quite heavily, and hence it is not a robust indicator. It's useful, but to an extent.
It's good to know the median income in Singapore to judge which percentage of the demography I belong to.	Correct	True, as the median will definitely tell you if you are in the higher 50% or the lower 50% of the population. Nothing more though. So, useful to some extent.
Histogram/KDE of income in Singapore is a much richer source of information compared to a quartile box plot.	Correct	No wonder -- the box plot only tells you which quartile you belong to. However, the histogram shows a detailed distribution from where you can find out which box/band you belong to. So, it is more informative.
It's good to know the average income in Singapore to judge where I stand with respect to the population.	Wrong	The average or the mean is NOT a robust indicator -- it is highly influenced by high/low outliers.

Reference Module 2 Topic 2 : Uni-Variate Visualization No specific slide. The overall concept matters.

Quizzes : Module 2 Part 2

Exploratory Data Analysis

In the LAMS Sequence, you have learned the theory behind this module. It is also expected that you have attempted the quizzes embedded within the LAMS Sequence, and have used the “unlimited attempts” opportunity to score 100%. Here are the quiz questions, consolidated with their answers and corresponding feedback. This is for your after-LAMS revision.

Question 1

Suppose the cost of having dinner (for two) at Clark Quay follows a Normal Distribution with Mean = SGD 50 and Standard Deviation (SD) = SGD 15. What is the probability that you will pay LESS than or equal to SGD 65 if you go for dinner at Clark Quay with a friend?

Answer Choice	Verdict	Explanation
Around 0.84 or around 84% chance	Correct	Correct answer. You are looking for probability “equal to or below” Mean + SD = SGD 65, which is around 0.84.
Around 0.5 or around 50% chance	Wrong	Nope. Check the distribution again. 50% should be the probability for less than the Mean = SGD 50.
Around 0.68 or around 68% chance	Wrong	Nope. Be careful. 68% is between Mean - SD to Mean + SD. The question asks something different.
Around 0.16 or around 16% chance	Wrong	Quite close. You are thinking in the right direction. 16% is for being "equal to or above" SGD 65 or Mean + SD.

Reference Module 2 Extra Topic : Normal Distribution No specific slide. The overall concept matters.

Question 2

As a follow up of the previous question (Mean = SGD 50 and SD = SGD 15), what is the probability that you will pay more than SGD 100 for a dinner for two at Clark Quay?

Answer Choice	Verdict	Explanation
Less than 0.00135 or less than 0.135% chance	Correct	Correct answer. SGD 100 is more than Mean + 3 * SD, and it is only one side of the tail. Hence half of 0.0027.
Less than 0.0027 or less than 0.27% chance	Almost there …	Correct line of thought, but 0.27% is for both tails of the distribution. More than SGD 100 is only on one side. Think again.
Less than 0.05 or less than 5% chance	Close, but …	This is one of the correct answers, but 5% is a huge margin. I am sure you can narrow it down. Check again.
Probability 0 or 0% chance	Wrong	For a Normal Distribution, we can't have strict maximum. There is always a probability for paying more than SGD 100.

Reference Module 2 Extra Topic : Normal Distribution No specific slide. The overall concept matters.

Question 3

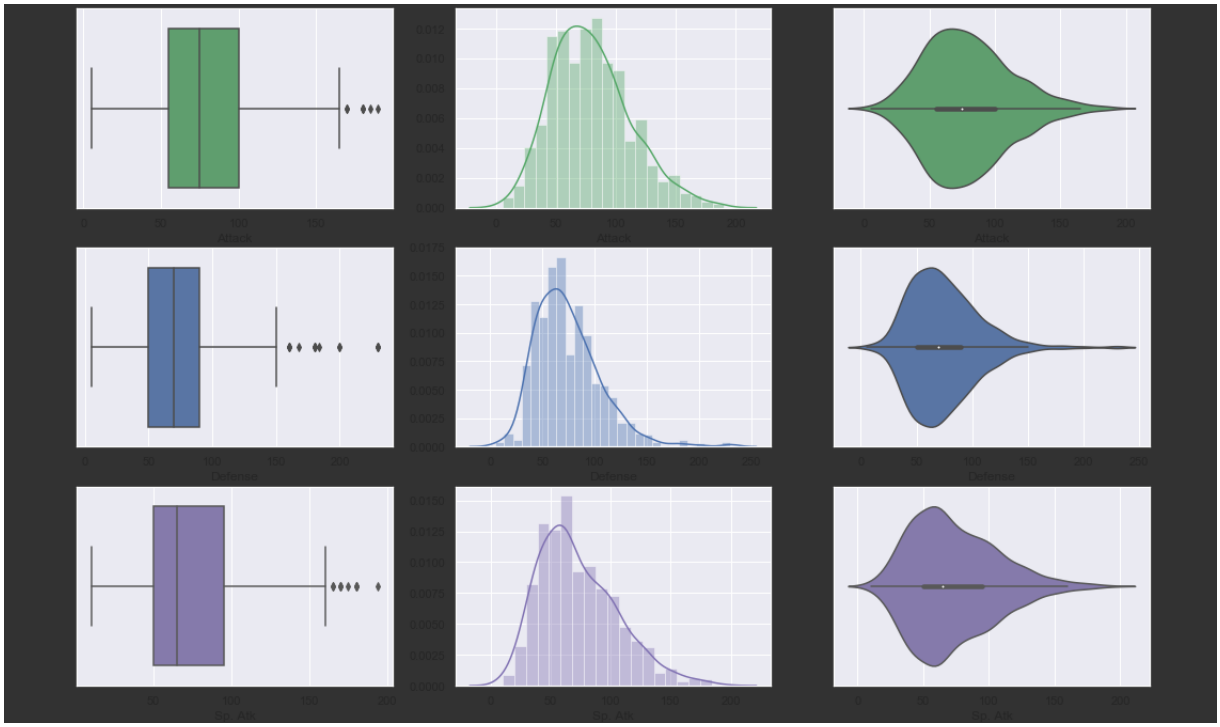
You know the Normal Distribution of cost of dinner (for two) at Clark Quay (Mean = SGD 50 and SD = SGD 15). What do you think 50% of the dinner options (or more), that is, more than or equal to half of the dinner choices at Clark Quay, cost for a dinner for two?

Answer Choice	Verdict	Explanation
Between SGD 35 and SGD 65	Correct	Correct answer. The range is Mean - SD to Mean + SD, and thus more than 50% of the options (around 68%) should fall here.
Less than or equal to SGD 50	Correct	Correct answer. Exactly 50% or half of the options are less than or equal to the Mean in a Normal Distribution.
Between SGD 50 and SGD 80	Wrong	Nope. SGD 50 is the Mean, and SGD 80 is Mean + 2 * SD. This range contains less than 50% of the places. Think again.
Less than or equal to SGD 40	Wrong	Nope. We can't say this as SGD 40 is considerably (SGD 10) less than the Mean, or SGD 50. Thus less than 50% options lie below SGD 40.

Reference Module 2 Extra Topic : Normal Distribution No specific slide. The overall concept matters.

Question 4

Study the following three uni-variate distributions (green, blue, magenta). Which one has the highest Median out of the three? Assume that the scale (markers) of the x-axis in each of the following plots is similar -- so, don't worry about the x-values.

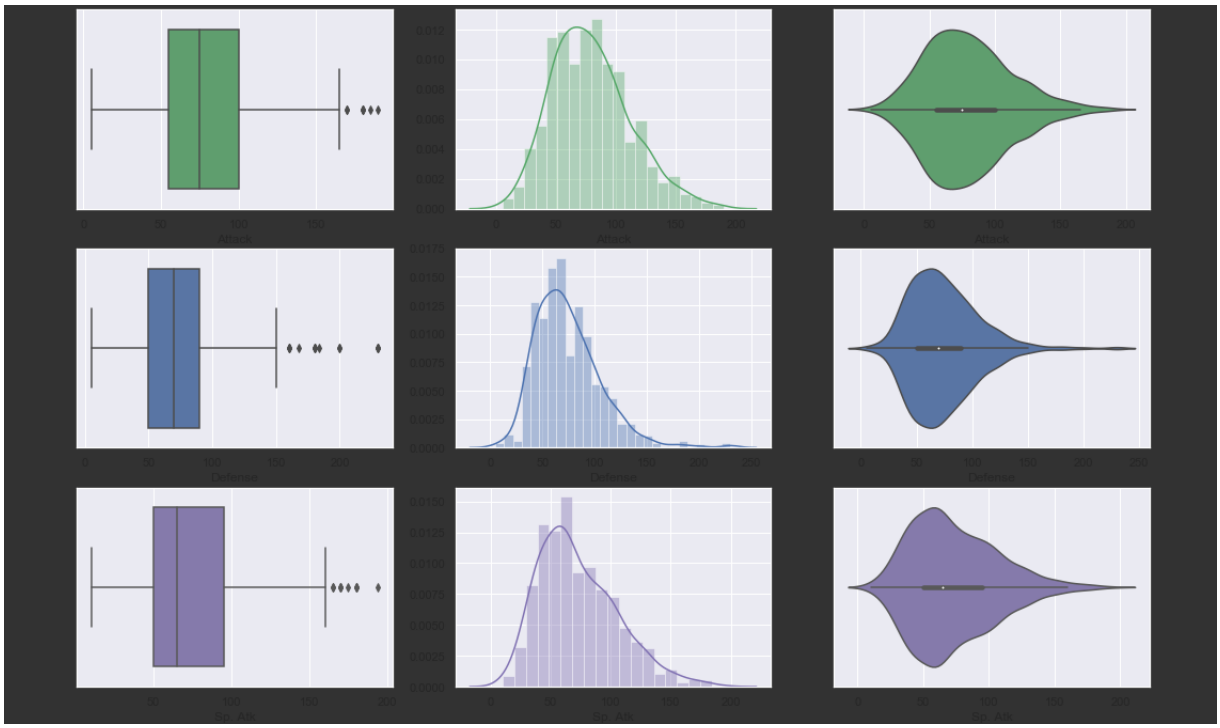


Answer Choice	Verdict	Explanation
The first distribution (green)	Correct	Correct answer. Medians are straight-forward to compare between distributions if the x-axis is the same.
The second distribution (blue)	Wrong	Nope. Check the distributions again. Median is the vertical line in the middle of the Box in a box-plot.
The third distribution (magenta)	Wrong	Nope. Check the distributions again. Median is the vertical line in the middle of the Box in a box-plot.
Impossible to determine as it depends on the outliers	Wrong	Nope. Median is generally not affected too much by outliers. Check the box-plots carefully, once again.

Reference Module 2 Topic 4 : Multi-Variate Exploration Slide 5 and Slide 6

Question 5

Study the following three distributions (green, blue, magenta) once again. Which one of these above three distributions is the most dissimilar to a Normal Distribution?



Answer Choice	Verdict	Explanation
The third distribution (magenta).	Correct	Correct answer. This one is the farthest from Normal, that is, the most skewed distribution. Look at the Median.
The second distribution (blue)	Wrong	Not so much. It is definitely not Normal, but not too far (skewed) either. If we drop the outliers, it may actually be Normal. There's another one more dissimilar to Normal.
The first distribution (green)	Wrong	Not so much. It is definitely not Normal, but not too far (skewed) either. There's another one more dissimilar to Normal.

Reference Module 2 Topic 4 : Multi-Variate Exploration Slide 5 and Slide 6

Question 6

Study the joint-plots, and arrange them in order of Correlation -- highest to lowest.

Possibly > 0.7 | Possibly ~ 0.5 | Possibly < 0.3 | Possibly < -0.7

Note that correlation can be both positive and negative (+1 to -1). Absolute value depicts the dependence.

Reference Module 2 Topic 3 : Bi-Variate Exploration Slide 9 and Slide 10

Question 7

Suppose that the "Time students take to complete this LAMS Sequence" and the "Marks students score in this LAMS Quiz" have a correlation of 0.8. What can you infer from this?

Answer Choice	Verdict	Explanation
The data for LAMS Quiz Scores have a linear relationship with the data for Time taken to complete LAMS Sequence.	Correct	Correct answer. The linear relationship is the most you can infer from a high correlation. No "causality" can be claimed.
To score more marks in the LAMS Quiz, one must take a long time to complete the LAMS Sequence.	Wrong	Nope. This is a common misconception. High correlation DOES NOT mean that one variable "causes" the other.
If You scored a high mark in the LAMS Quiz, it is clear that You must have taken a long time to complete the Sequence.	Wrong	Nope. This is a common misconception. You, specifically, may be an outlier or an anomaly in the dataset, and may not follow the norm. High correlation only says that you are "likely" to have taken a long time, but it can't be claimed with certainty.

Reference Module 2 Topic 3 : Bi-Variate Exploration Slide 11 and Slide 12

Question 8

Suppose we look at Sale Prices (SalePrice) of houses in Singapore, and find that it has 0.7 correlation with the General Living Area (GrLivArea) of the houses. What can we infer?

Answer Choice	Verdict	Explanation
General Living Area (GrLivArea) may be an important variable in "predicting" Sale Prices (SalePrice) of houses in Singapore.	Correct	Correct answer. Indeed, we will try to use GrLivArea to predict the SalePrice. We will use the technique of Linear Regression.
General Living Area (GrLivArea) has no linear relationship with Sale Prices (SalePrice) of houses in Singapore.	Wrong	Nope. The correlation 0.7 is quite high, denoting a significant linear relationship between GrLivArea and SalePrice.
Increase in General Living Area (GrLivArea) causes the Sale Prices (SalePrice) of houses in Singapore to go higher.	Wrong	Nope. Once again, "causality" is not implied by correlation. However, we can try to use the high correlation to predict.

Reference Module 2 Topic 3 : Bi-Variate Exploration Slide 11 and Slide 12

Question 9

In a multi-variate dataset, we find 30 numeric variables -- one of them is the SalePrice, and the others other general living area, lot area, basement area, garage area, etc. 10 out of the 30 variables have strong positive correlation (above 0.6) with SalePrice, 5 out of the 30 have strong negative correlation (below -0.6) with SalePrice, and others have weak correlation (between 0.2 to -0.2) with SalePrice. What would your next step be?

Answer Choice	Verdict	Explanation
Consider the 10 variables with strong positive correlation (above 0.6) important for predicting SalePrice.	Correct	Correct answer. Strong positive correlation denotes strong linear relationship, hence important as predictors.
Consider the 5 variables with strong negative correlation (below -0.6) important for predicting SalePrice.	Correct	Correct answer. Strong negative correlation denotes strong linear relationship, hence important as predictors.
Can't decide which variables are important to predict SalePrice, and hence, will consider all variables equal.	Wrong	Nope. At least in case of a Linear Regression, strong positive or negative correlation helps in prediction.
Can't decide which variables are important to predict SalePrice, as strong correlation does not imply causality.	Wrong	Nope. You are right that strong correlation does not imply causality, but it sure helps in predicting SalePrice. :-)

Reference Module 2 Topic 4 : Multi-Variate Exploration Slide 8 and Slide 9

Question 10

True or False : Strong correlation of SalePrice with a categorical variable (like Type of the House) also denotes strong relationship.

Answer Choice	Verdict	Explanation
False	Correct	Correct answer. First, you have to decide what the definition of "Correlation" is in case of a categorical variable. Then, you have to see if the definition makes sense in terms of a strong relationship. You can't imply strong relationship based on a wrong definition of correlation.
True	Wrong	Nope. First, you have to decide what the definition of "Correlation" is in case of a categorical variable. Then, you have to see if the definition makes sense in terms of a strong relationship. You can't imply strong relationship based on a wrong definition of correlation.

Reference Module 2 Topic 3 : Bi-Variate Exploration Slide 11 and Slide 12

Quizzes : Module 3 Part 2

Linear Regression

In the LAMS Sequence, you have learned the theory behind this module. It is also expected that you have attempted the quizzes embedded within the LAMS Sequence, and have used the "unlimited attempts" opportunity to score 100%. Here are the quiz questions, consolidated with their answers and corresponding feedback. This is for your after-LAMS revision.

Question 1

If two variables have a high positive correlation (above 0.7, say), what can we say about the variables? Multiple options may be correct.

Answer Choice	Verdict	Explanation
There is a strong linear relationship between the variables.	Correct	Right. Strong positive correlation does suggest a strong linear relationship. Causality may not be clear though.
Knowing one of the variables may help us predict the values of the other.	Correct	Strong positive correlation suggests linear relationship, and such a relationship is useful in prediction.
There is no linear relationship between the variables.	Wrong	Nope. Strong positive correlation does suggest a strong linear relationship. There may not be causality though.
There is no non-linear relationship between the variables.	Wrong	Sorry. Can't infer anything about non-linear relations from correlation. It only tells us about linear relation.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 7 and Slide 8

General Comment : Correlation does tell us about linear relationship between variables, but not about causality. Correlation, in the classical sense, does not tell us anything about non-linear relationship between variables. High correlation means strong linear relationship, which is quite useful if we want to predict one variable using the other. This is something that we see frequently in case of Linear Regression. Also note that we are talking about Pearson Correlation Coefficient so far; you may want to check it out on Wikipedia or any book on Statistics. ;-)

Question 2

Suppose that you have 1000 observations in a dataset, and you plan to use 750 as your Train set. Which of the following is/are most appropriate in such a scenario?

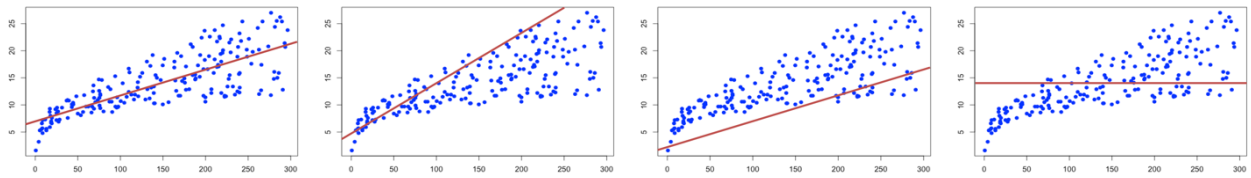
Answer Choice	Verdict	Explanation
The remaining 250 observations (or a subset of these) may be used as the Test set.	Correct	Correct. The remaining 250 observations have not been used for Training, and hence may be used for Test.
The 750 observations in Train set should be selected carefully -- may be chosen uniformly at random from the Dataset.	Correct	Correct. The Train set should be representative of the main Dataset. Uniform random selection (often) helps in achieving that.
It does not matter how we choose the 750 observations from the Dataset to get Train set. Choose the first 750 observations.	Wrong	Actually, it does matter. The Train set should be a representative of the Dataset. You do not know if the "first 750" data points are randomly distributed in the dataset or arranged in some predetermined fashion. Try selecting uniformly at random, it often helps reducing such bias.
It does not matter which observations from the main Dataset are taken for the Test set. Choose any 250 at random.	Wrong	Actually, it does matter. We should not use the same observations for Train and Test. So, the remaining 250 observations are the ones from which we should draw the Test set.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 5

General Comment : There are two crucial aspects in choosing the Train and Test sets. First, the Train set should be a representative of the main Dataset, and hence, should be chosen quite carefully. Uniform random selection from the main labeled dataset often helps achieve this. Second, the Test set and Train set should not have any overlap, as in that case, observations used for Train will again be used for Test, reducing the fairness of the whole learning process.

Question 3

Arrange the following linear models in decreasing order of how well they fit the dataset -- the Best Fit one to the Worst Fit one.



Minimum RSS Slightly large RSS Much higher RSS Maximum RSS (= TSS)

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 10 and Slide 11

General Comment : The best fit linear model will make the least Sum Square of Errors (RSS), while the worst fit linear model will make the most Sum Square Error (RSS). Minimum RSS is close to zero, while worst RSS is equal to TSS.

Question 4

Arrange the steps of Linear Regression, as follows, in order in which they are executed.

1. Guess the initial values of the "Parameters" for the hypothesized Linear Model.
2. Predict the values of the Response Variable for all observations in Train data.
3. Compute the Errors in Train data, compared to actual values of the Response.
4. Choose a specific Cost Function (like Sum Square of Errors) for Optimization.
5. Reassign or Tune the "Parameters" of the model to Optimize the Cost Function.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 11

Question 5

Which ones of the following are decent choices for Cost Function in Linear Regression? Multiple options may be correct.

Answer Choice	Verdict	Explanation
Sum Square of Errors / Residual Sum of Squares over the Train Set	Correct	Of course. This is the most commonly used Cost Function in Linear regression.
Absolute Sum of Errors / Residual Absolute Sum over the Train Set	Correct	Sure, this works too. Note that optimization is a little hard with the non-differentiable function. But still, it will work in theory.
Minimum Absolute Error / Minimum Absolute Residual in the Train Set	Wrong	No, this won't work. Think about it -- even a horribly fit line can pass through one of the points in the Train set. In that case, the minimum error will actually be Zero (0), resulting in the minimum possible Absolute Error. You need to consider all points.
Maximum Squared Error / Maximum Residual Square in the Train Set	Wrong	No, this won't work. Think about it -- all the weight for this Cost Function is placed on the point that is farthest from the line. Thus, the cost function is unnecessarily biased to the outliers, and not to the other points. You need to consider all points.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 11

General Comment : Think about the Cost Function as a cumulative measure of all Errors in the Train set. We also require the Linear Regression to work by "optimizing" the Cost Function, which often requires differentiating the cost function (remember calculus in terms of maximum or minimum of a function?). Put these two ideas together, and think again -- which one of the options are potential Cost Functions.

Question 6

The hypothesized linear model and the cost function in a linear regression problem are as follows. Which of the following are True? Multiple options can be correct.

Linear Model : *Response* = *a* x *Predictor* + *b*

Cost Function : *J* = Sum of (*Response* – *a* x *Predictor* - *b*)^2

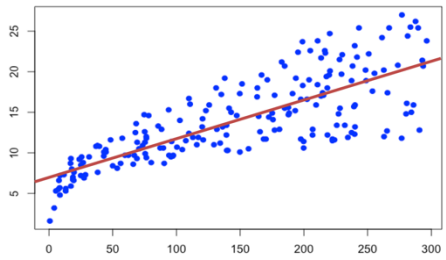
Answer Choice	Verdict	Explanation
a and *b* are the Parameters of the hypothesized Linear Model	Correct	Correct. This is by definition of the Linear Model. To estimate these parameters is the goal of Linear Regression.
Response and *Predictor* in the Cost Function *J* are available as fixed values from the Train set	Correct	Correct. Even though *Response* and *Predictor* look like variables, they are not. All values of these two items are available from the Train set, and we simply plug these values in the Cost Function before the optimization process starts.
a and *b* are the actual Variables in the Cost Function *J* after we put in the values from the Train set	Correct	Correct. When we say that optimizing the Cost Function is the goal of Linear Regression, we mean that the Cost Function is a function of the parameters *a* and *b*, and that we have to optimize (minimize) the bi-variate Cost Function *J*(*a*,*b*).
Cost Function *J* is a function of *Response* and *Predictor*, as the parameters *a* and *b* are guessed.	Wrong	The cost function is really a function of the parameters *a* and *b*. Even though *Response* and *Predictor* look like variables, they are not. All values of these two items are available from the Train set, and we simply plug these values in the Cost Function before the optimization process starts. When we start the optimization, we guess the values of *a* and *b* just to initiate the process.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 11

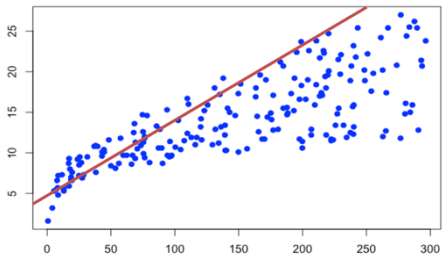
General Comment : Make sure you understand that the variables in a Cost Function are actually the "parameters" of the model, and that the Response and Predictors are all constant values, available from the Train set.

Question 7

The following linear models (A and B) were fit on the Train data shown in the figures. Which one would best predict the Test data?



Model A



Model B

Answer Choice	Verdict	Explanation
It is impossible to say which model will better predict on the Test data, as it depends on other factors.	Correct	Correct. If the Train data is similar to the Test data, Model A will do a good job of prediction, but there is no guarantee otherwise. In Machine Learning, we try hard to make sure that the Train data is similar to the Test data. However, if we do not know for sure that the Train data is similar to the Test data, we can't really say that the model fitting the Train data best will surely be the best predictor for the Test data. It really depends on a lot of other factors.
Linear Model A will best predict the Test data, as it best fits the Train data, as shown in the figure.	Wrong	This would be correct only if the Train data is similar to the Test data. In Machine Learning, we try hard to make that happen. However, if we do not know for sure that the Train data is similar to the Test data, we can't really say that the model fitting the Train data best will surely be the best predictor for the Test data.
Linear Model B will best predict the Test data, as it best fits the Train data, as shown in the figure.	Wrong	It is clear from the figures that Linear Model A best fits the Train data, if we consider standard Sum Squared Errors. Thus, the statement is not correct. However, we do not know for sure if Model B is a better predictor for Test data, as we do not know the Test data yet.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 13 and Slide 14

General Comment : Note that the Train data and Test data are mutually independent, and thus, it is hard to conclude something about the Test data unless we know that the Train data is really similar to the Test data. This is a big issue in Machine Learning, known as "Generalization of Model". Go ahead and Google for this issue. :-)

$MSE = \frac{RSS}{n}$

Question 8

Suppose that there are 100 observations in a Train set, and the Residual Sum of Squares (RSS) for a linear model is 745. What is the Mean Squared Error (MSE) of the linear model on the Train set?

Answer Choice	Verdict	Explanation
7.45	Correct	Correct. The MSE is RSS divided by the number of data points in the Train set, that is 745/100 = 7.45.
74500	Wrong	Nope. It is the other way round -- that is, MSE = RSS divided by the number of data points. Check again.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 14

General Comment : RSS is the Sum of Square of Errors, while MSE is the Mean of Square of Errors. Hence the relation.

Question 9

Suppose the MSE of a linear model on the Train set is 4.25, the number of observations in the Train set is 100, and the Variance of the Response variable in the Train set is 8.5. What is the value of R^2 on the Train set in this case?

Answer Choice	Verdict	Explanation
0.5	Correct	Correct. R^2 = 1 – (MSE/VAR) in the Test set. The information about 100 observations is redundant in this case.
-49	Wrong	Nope. R^2 = 1 – (MSE/VAR) in the Test set. The information about 100 observations is redundant in this case.
-0.5	Wrong	Can't be negative. Remember that R^2 can only be between 0 to 1.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression Slide 14

General Comment : Check Slide 14 (last but one) in this lesson to find out the relationship between R^2, MSE, VAR, RSS and TSS. Remember that R^2 can only be between 0 to 1. Hence, be careful and cross-check your calculation.

Extra note : Computing R^2 on Test Data is a little misleading. You compute R^2 on Train Data to judge the performance of your model, but evaluate the same performance on Test Data using MSE or RMSE, and not using R^2. If you obtained negative R^2 in Test Data during your Lab Exercises, that is possible, as VAR is computed on Train Data.

Question 10

What do you think happens if for some linear model, we get $R^2 = 1$ in the Train set? Is it good or bad for prediction?

Answer Choice	Verdict	Explanation
$R^2 = 1$ means the Residual Sum of Squares (RSS) is Zero, that is, the model "perfectly" fits Train set. However, there is no guarantee that it will best predict the Test set, as we do not know if it is similar to the Train set.	Correct	Correct. $R^2 = 1$ or RSS = 0 on the Train set just means that the model "perfectly" fits the Train set. There is no guarantee that it will best predict the Test set, as we do not know if it is similar to the Train set. In fact, we get a little worried if $R^2 = 1$ in practice -- it often means that we've "overfit" the Train set. Google "overfitting". :-)
$R^2 = 1$ means the Residual Sum of Squares (RSS) is Zero, that is, the model "perfectly" fits Train set. This is the ideal case, as we have the best fit model. Definitely, this model will be the best one to predict on the Test set.	Wrong	Well, you are correct on the first count. RSS is really 0 on the Train set. However, this just means that the model "perfectly" fits the Train set. There is no guarantee that it will best predict the Test set, as we do not know if it is similar to the Train set. In fact, we get a little worried if $R^2 = 1$ in practice -- it often means that we've "overfit" the Train set. Google "overfitting". :-)
$R^2 = 1$ means the Residual Sum of Squares (RSS) is Maximum (equal to TSS), that is, the model is the "worst" fit on the Train set. This is the worst case, and definitely, the model will be the worst one to predict on the Test set.	Wrong	Wrong. Check the formula for R^2 once again. $R^2 = 1$ means RSS = 0, not maximum.

Reference Module 3 Topic 2 : Uni-Variate Linear Regression No specific slide. It's slightly beyond.

General Comment : Check the formula for R^2 to obtain its relationship with RSS. This should be really easy. However, the second part is non-intuitive. Even if a model "perfectly" fits the Train set, there is no guarantee that it will best predict the Test set, as there is a chance of "overfitting". Go ahead and Google for this issue. ;-)

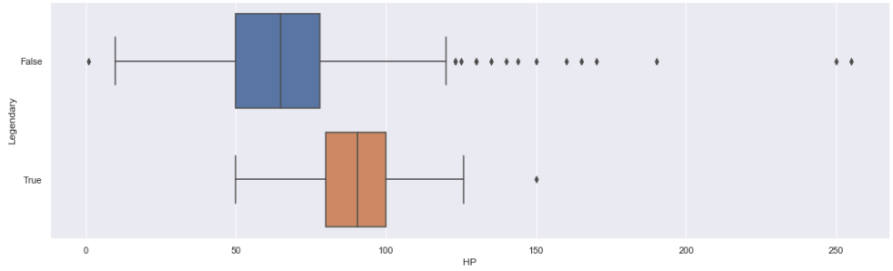
Quizzes : Module 4

Data-Driven Classification

In the LAMS Sequence, you have learned the theory behind this module. It is also expected that you have attempted the quizzes embedded within the LAMS Sequence, and have used the “unlimited attempts” opportunity to score 100%. Here are the quiz questions, consolidated with their answers and corresponding feedback. This is for your after-LAMS revision.

Question 1

Suppose that the boxplot of the HP (Hit Points) of Pokemons with respect to the categorical variable "Legendary" is as in the picture. Which of the following can you infer from this boxplot? Multiple options may be correct. Choose all options that seem right.



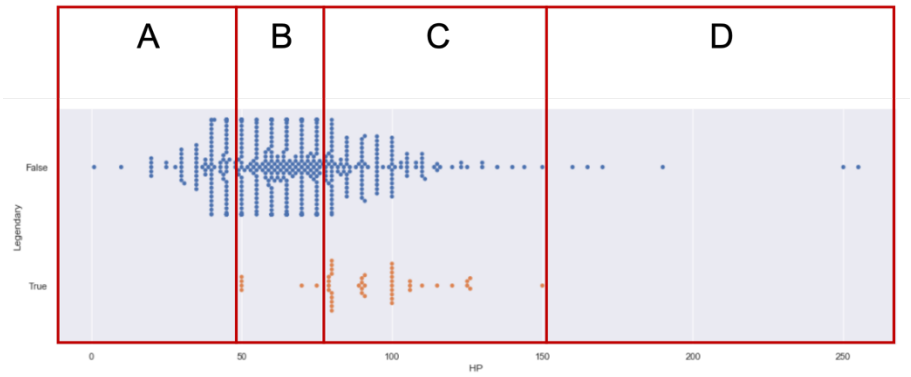
Answer Choice	Verdict	Explanation
HP (Hit Points) seems to be an important variable in predicting "Legendary".	Correct	True. The box plots for HP (Hit Points) are distinctly different for "Legendary" = True and "Legendary" = False. This indicates that HP may be an important differentiator between the two levels of the categorical variable "Legendary".
There is a strong relationship between the variables HP (Hit Points) and Legendary.	Correct	True. The box plots for HP (Hit Points) are distinctly different for "Legendary" = True and "Legendary" = False. If there was NO relation between the two, then the boxplots for HP across the various levels of "Legendary" would look quite similar.
There is a strong linear relationship between the variables HP (Hit Points) and Legendary.	Wrong	Nope. The boxplot does not suggest a linear relationship. In fact, a linear relationship is not even well-defined for a categorical variable against a continuous variable. Similarly, there is no notion of correlation between these two variables.

Reference Module 4 Topic 1 : Binary Classification Slide 5 and Slide 6

General Comment : Significant difference in the Box Plots of a continuous variable against multiple levels of a categorical variable tells us that the continuous variable is "important" in differentiating between the levels of the categorical variable. Hence, it is important for prediction.

Question 2

Suppose that the swarmplot of HP (Hit Points) against "Legendary" is as follows, with a specific partition made on the data. Which of the following statements are correct? Multiple answers may be correct. Choose all options that seem right.



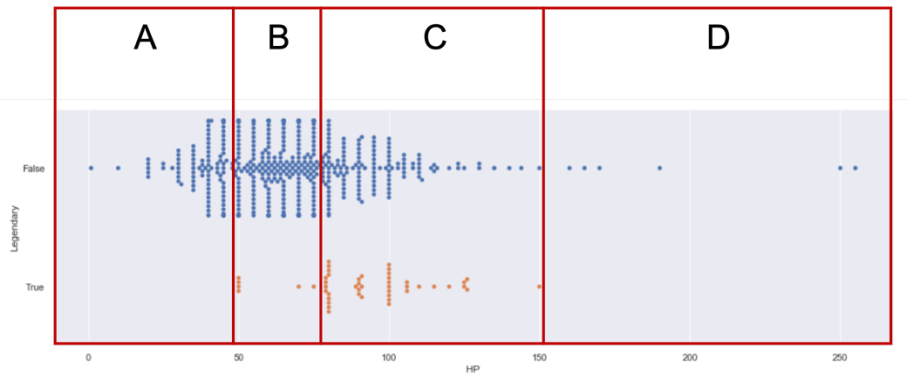
Answer Choice	Verdict	Explanation
We can be "completely confident" about Partition A -- it is "Legendary" = False.	Correct	True. There are only datapoints for "Legendary" = False in Partition A. Hence we can be "completely confident".
We can be "completely confident" about Partition D -- it is "Legendary" = False.	Correct	True. There are only datapoints for "Legendary" = False in Partition D. Hence we can be "completely confident".
We can be "completely confident" about Partition B -- it is "Legendary" = False.	Wrong	Wrong. There are datapoints for both "Legendary" = True and False in Partition B. Hence we can't be entirely confident.
We can be "completely confident" about Partition C -- it is "Legendary" = False.	Wrong	Wrong. There are datapoints for both "Legendary" = True and False in Partition C. Hence we can't be entirely confident.

Reference Module 4 Topic 1 : Binary Classification Slide 5 and Slide 6

General Comment : You are "completely confident" on the parts where all datapoints are of the same type. Otherwise, there is always a probability that the datapoints could be of one class or the other. In fact, finding "pure" partitions, with single-type datapoints, is quite rare in practice.

Question 3

Arrange the following partitions in decreasing order of their "confidence" for the variable "Legendary" to be False. That is, the partition with the maximum confidence for "Legendary" = False should be first, and decrease to the minimum confidence partition.



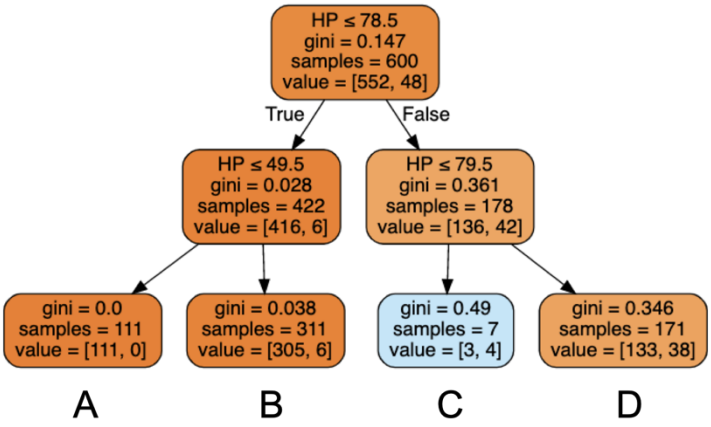
Partition A > Partition D > Partition B > Partition C

Reference Module 4 Topic 1 : Binary Classification Slide 6 and Slide 7

General Comment : The most confident partitions will have maximum datapoints of the same class, while the least confident ones will have a random mix. Partition A is higher in confidence, as it has more datapoints than Partition D.

Question 4

Arrange the leaf nodes of the following decision tree in decreasing order of their "confidence". That is, the most confident node for predicting "Legendary" should be first, and then decrease to the least confident node for predicting "Legendary".



lower gini, ↑ confidence

1. Node A, with Gini = 0 and Datapoints = [111, 0]
2. Node B, with Gini = 0.038 and Datapoints = [305, 6]
3. Node D, with Gini = 0.346 and Datapoints = [133, 38]
4. Node C, with Gini = 0.49 and Datapoints = [3, 4]

Reference Module 4 Topic 1 : Binary Classification Slide 7

General Comment : The lower the Gini Index, the more confident the node is in predicting "Legendary". Gini = 0 will mean all data points are of the same "Legendary" label. Thus, Gini Index tells you about the confidence in a node.

Question 5

Suppose that a specific node is a Decision Tree has the following distribution for "Legendary" : True = 300, False = 300. That is, there are 600 datapoints in that node (or partition), out of which 300 are "Legendary" = True, and 300 are "Legendary" = False. What is the Gini Index for this specific node of the tree?

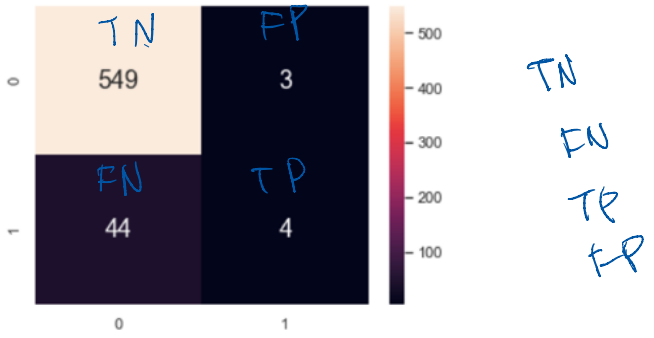
Answer	Explanation
0.5	Correct. Gini is 0.5 for a Binary Uniform distribution, with 50:50 ratio.

Reference Module 4 Topic 1 : Binary Classification Slide 7

General Comment : Note that for a perfectly uniform distribution, as in this case (300:300), the Gini Index is always 0.5. This in fact, tells you that the node has "least possible" confidence about prediction.

Question 6

Arrange the following quantities in decreasing order, as per the Confusion Matrix shown in the picture. That is, the highest quantity should come first, and then decrease to the lowest quantity coming at the end.



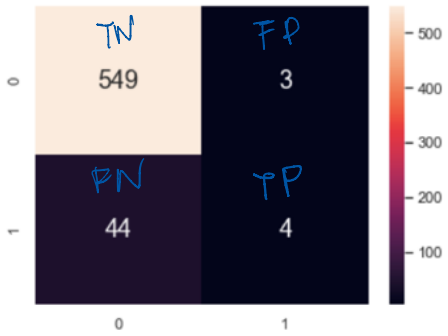
1. True Negatives -- "Legendary" = False (0) predicted as "Legendary" = False (0) 549
2. False Negatives -- "Legendary" = True (1) predicted as "Legendary" = False (0) 44
3. True Positives -- "Legendary" = True (1) predicted as "Legendary" = True (1) 4
4. False Positives -- "Legendary" = False (0) predicted as "Legendary" = True (1) 3

Reference Module 4 Topic 1 : Binary Classification Slide 9 and Slide 10

General Comment : Check definitions of True Positives, True Negatives, False Positives and False Negatives.

Question 7

Based on the confusion matrix as follows, calculate the False Positive Rate (FPR) in decimals. Submit your answer as a decimal number rounded off to four decimal places.



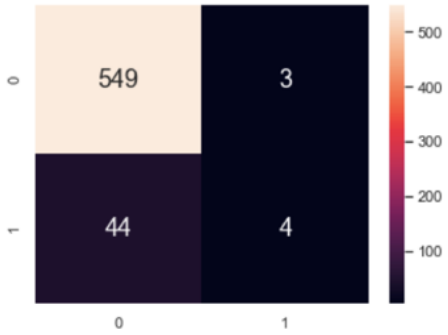
$\frac{FP}{FP+TN} = \frac{3}{3+549} \approx 0.0$

Answer	Explanation
0.0054	Correct. FPR = FP / (TN + FP) = 3 / (549 + 3) = 3 / 552 = 0.0054

Reference Module 4 Topic 1 : Binary Classification Slide 10

Question 8

Based on the confusion matrix as follows, calculate the False Negative Rate (FNR) in decimals. Submit your answer as a decimal number rounded off to four decimal places.



$\frac{FN}{FN+TP} = \frac{44}{44+4}$

Answer	Explanation
0.9167	Correct. FNR = FN / (TP + FN) = 44 / (4 + 44) = 44 / 48 = 0.9167

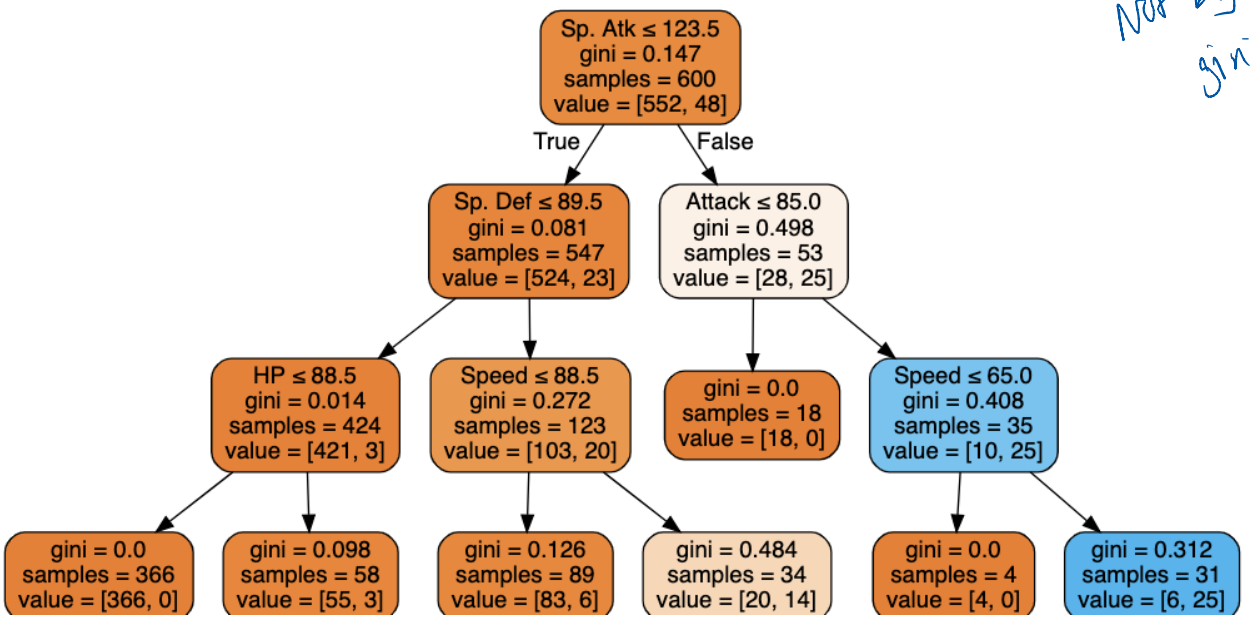
Reference Module 4 Topic 1 : Binary Classification Slide 10

General Comment : Just to recap quickly, here are the formulas : FPR = FP / (TN + FP) and FNR = FN / (TP + FN).

Question 9

Based on the Tree, arrange the variables in decreasing order of their "importance" in predicting the binary variable Legendary.

1]:



Not by gini

1. Sp. Atk – Special Attack of a Pokemon — as it occurs in the very first split of the tree
2. Sp. Def – Special Defence of a Pokemon — as it occurs right after, in the next split
3. Speed – Speed of a Pokemon — as it occurs in more nodes, and improves Gini better
4. HP – Hit Points for a Pokemon — as it occurs only in one node, and improves Gini slightly

Reference Module 4 Topic 1 : Binary Classification Slide 12 and Slide 13

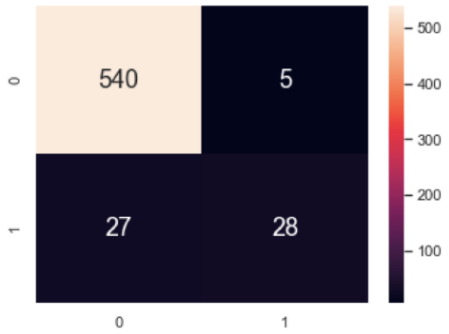
General Comment : The more the drop in Gini from a parent node to the children node, the better is the split. Now, if a variable helps in dropping the Gini from parent to children more than another variable, it is deemed more important.

The tree automatically decides the best variables by splitting the tree using them. Thus, the higher the variable occurs in a tree, the more important it is. We will use the Gini drop idea for variables that occur at the same level of a tree.

According to the logic, HP tackled an almost pure node [421, 3] and dropped to children [366, 0] and [55, 3]. This is not as significant as Speed, which tackled relatively more mixed nodes [103, 20] and [10, 25], to drop Gini much further.

Question 10

What is the classification accuracy of the Decision Tree that produces the following confusion matrix? Enter your answer as a decimal number, correct (rounded off) to four decimal places.



TP + TN / Total

Answer	Explanation
0.9467	Correct. Classification Accuracy = (TP + TN) / Total = (540 + 28) / (540 + 5 + 27 + 28)

Reference Module 4 Topic 1 : Binary Classification Slide 10

General Comment : Classification Accuracy is simply the fraction of correct predictions, that is, (TP + TN) / Total.