1.1 Data Attributes

(To be done over Week #2)

(1) For the type of data below, state which would be considered continuous or discrete data types:

(a) Number of patients in the wards discrete
(b) The blood pressure of the patient continuous
(c) The pulse of the patient discrete
(d) The size of the patient's tumour continuous
(e) The emergency room waiting time rounded to the nearest minute discrete

(2) Which of the **NOIR** (Nominal, Ordinal, Interval, Ratio) scale would describe each of the data attributes below:

(a) Rank in the Singapore Army Ordinal
(b) Number of students registered for CZ4124 Ratio
(c) Shoe sizes Interval
(d) Jersey numbers in the Singapore national football team Nominal

N N I R R O Ratio

	A	B	C	D	E	F	G
1	Student ID	Gender	Year of Birth	Attendance (weeks)	Score (%)	Grade	Completion Time (sec)
2	1	M	2002	12	85.5	A+	1600.31
3	2	F	2004	2	20	F	600.13
4	3	F	2003	13	55.5	C+	1800.00
5	4	M	1999	10	73.5	B+	1800.00
6	5	M	2000	11	65	B	1500.23
7	6	F	2004	8	70	B+	1700.00
8	7	F	2005	13	90.5	A+	900.34

Cont

Table 1 – Excel Table of Students Particulars and their performance in an online test

(3) Which **NOIR** scale measure best represent each of the data category (column) in the Table 1?

(4) Which measure(s) in the Table 1 (if any) can be considered a continuous data variable? Provide suitable justification for your answer.

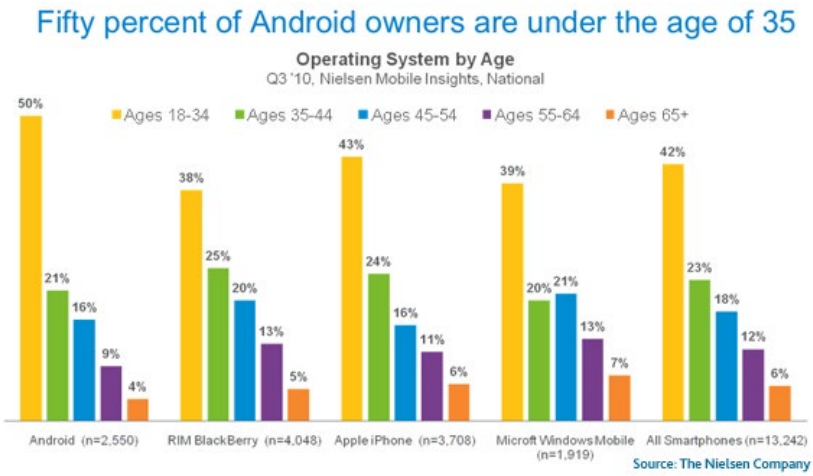


Figure 1 – Mobile OS by Age (From https://www.nielsen.com/us/en/insights/article/2010/mobile-snapshot-smartphones-now-28-of-u-s-cellphone-market/)

(5) How many data dimensions have been simultaneously visualised in the clustered bar (column) chart shown in Figure 1? Draw a possible table (*partial table will do*) for this dataset and state which **NOIR** scale measure best represent each data dimension in your table.

1.2 Data Wrangling

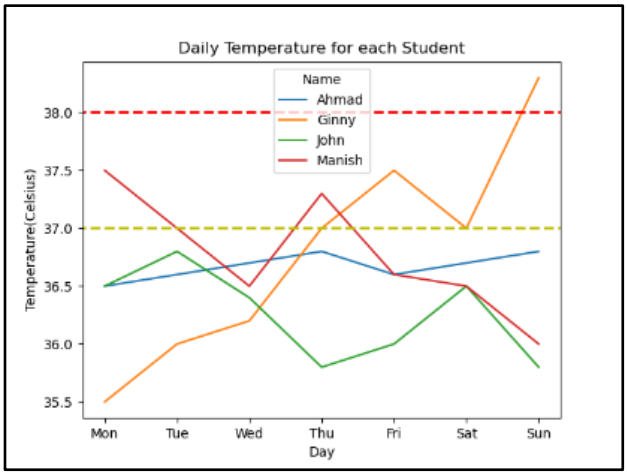
(1) **Reshaping dataframes**. Figure 2(a) shows the table that can be extracted from the datafile "*Daily_Temperature.csv*" (available in NTULearn). Using the template file "*1-2-Daily Temperature.py*", plot the line chart shown in Figure 2(c). In order to get to the stage where you can directly plot the line chart shown, the dataframe created by the table shown in Figure 2(a) must be reshaped to that shown in Figure 2(b). Use the pandas **`melt`** and **`pivot`** functions to reshape the dataframe.

Name	Mon	Tue	Wed	Thu	Fri	Sat	Sun
John	36.5	36.8	36.4	35.8	36.0	36.5	35.8
Ahmad	36.5	36.6	36.7	36.8	36.6	36.7	36.8
Ginny	35.5	36.0	36.2	37.0	37.5	37.0	38.3
Manish	37.5	37.0	36.5	37.3	36.6	36.5	36.0

(a)

Name	Ahmad	Ginny	John	Manish
Day				
Mon	36.5	35.5	36.5	37.5
Tue	36.6	36.0	36.8	37.0
Wed	36.7	36.2	36.4	36.5
Thu	36.8	37.0	35.8	37.3
Fri	36.6	37.5	36.0	36.6
Sat	36.7	37.0	36.5	36.5
Sun	36.8	38.3	35.8	36.0

(b)



(c)

Figure 2 – (a) Daily temperature table from "*Daily_Temperature.csv*". (b) The required dataframe format to plot the line chart shown in (c).

(2) **Data wrangling exercise**. Figures 3(a) and 3(b) show the tables in the datafiles "*Temp_week_1.csv*" and "*Temp_week_2.csv*" respectively (available in NTULearn). Apply the appropriate data wrangling techniques (in pandas) to produce the two different line charts shown in Figures 3(c) and 3(d). Fill in the missing temperature data using the weekly mean temperature.

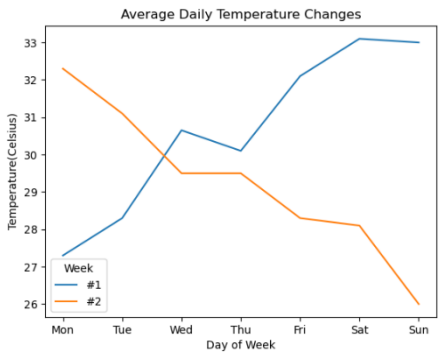
Challenge: If you can, fill in the missing data with a smoother interpolation from the two temporally nearest available data points.

(a)

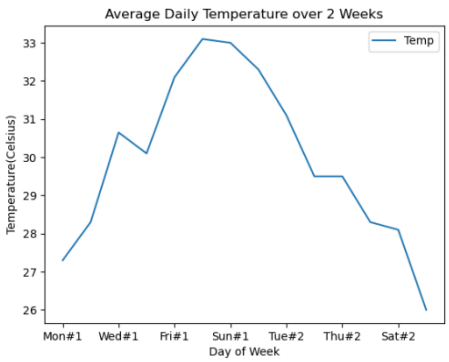
Day	Week	Temp
Mon	#1	27.3
Tue	#1	28.3
Wed	#1	
Thu	#1	30.1
Fri	#1	32.1
Sat	#1	33.1
Sun	#1	33.0

(b)

Day	Week	Temp
Mon	#2	32.3
Tue	#2	31.1
Wed	#2	29.5
Thu	#2	29.5
Fri	#2	28.3
Sat	#2	28.1
Sun	#2	26.0



(c)



(d)

Figure 3 – (a) Temperature tables for week #1 and (b) week #2. (c) Weekly temperature changes plotted as two different lines. (d) Single line plot of temperature changes over 2 weeks.

Appendix A – Listing of 1-2-Daily Temperature Python source file

```
# -*- coding: utf-8 -*-
"""
Created on Mon Aug  2 13:17:07 2021

CZ4124 Data Visualisation (Tutorial 1 Template)
@author: Put your name here

"""

import matplotlib.pyplot as plt
import pandas as pd

PlotWithPandas = True  # you can plot either with Pandas or Matplotlib
#---------------------------------------------------------------------
# Read in Daily Temperature datafile into dataframe Temp
#---------------------------------------------------------------------
Temp  =  pd.read_csv('C:/Datasets/Daily_Temperature.csv')    #  change  to  your
directory
print('\nTable read in\n',Temp,'\n')

#---------------------------------------------------------------------
# Use MELT to covert to long form and apply column names
#---------------------------------------------------------------------
# put in your code here


#---------------------------------------------------------------------
# Use PIVOT to convert to wide form with Names in each column
#---------------------------------------------------------------------
# put in your code here


#-------------------------------
# Use Pandas to plot line chart
#-------------------------------
if(PlotWithPandas):
    print('\nPlotting with Pandas')
    # put your Pandas plotting code here


#---------------------------------
# Use Matplotlib to plot line chart
#---------------------------------
else:
    print('\nPlotting with Matplotlib')
    # Alternatively, you can do the plot using Matplotlib or
    # any other Python plotting library
```

2.1 Visual Marks and Variables

(To be done over Week #3)

(1) What are visual marks for items? List the three visual marks proposed by Jacque Bertin and describe the main differences in their basic property.

(2) Name three perceptual channels (visual variables) proposed by Bertin and describe how they would alter the appearance of a line mark.

(3) In the context of Bertin's visual marks and variables, describe what visual marks are employed in the visualisation shown in Figure 1 and what perceptual channels are employed to encode the different various information. Discuss the potential limitations of the perceptual channels employed.

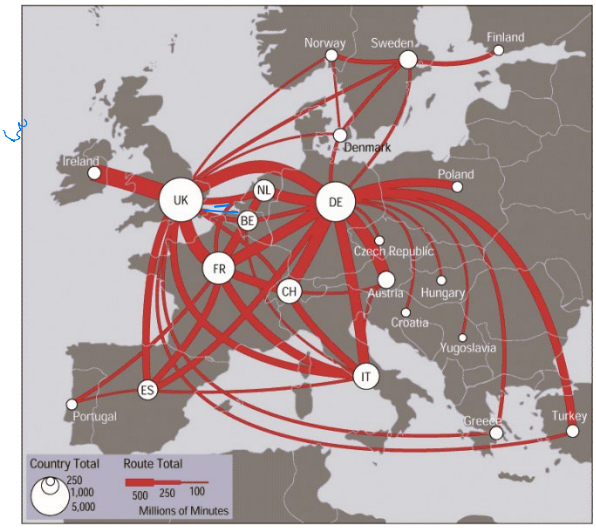


Figure 1 – Telecommunication Traffic Flow Map, © 2000 - TeleGeography, Inc.

Image taken from https://mappa.mundi.net/maps/maps_014/ . For more details on the visualisation, check out the link.

2.2 Visual Encoding

(1) Using your proposed visual encoding design, sketch a possible chart that will allow the effective visualisation of each of data tables show in Figures 2(a) to (c). The suggested visual mark for each chart has been indicated but you are free to choose an alternative. Consider carefully the attributes listed in each of the data table when designing your visual encoding and state any assumptions you made regarding the nature of these attributes that has influenced the design.

(2) What potential visual quality problems may arise due to nature of the given data attributes? How do you proposed this problem can be addressed or reduced?

Lines

Living Members from Class of '45

	Names	Gender	Age
1			
2			
3			
4			
5			
6			

(a)

Points

Orientation T-shirts - Comparing Halls

	No.	Gender	Weight	Height	Hall (A,B,C,D)	T-shirt Size
1						
2						
3						
:						
:						
700						

(b)

Areas

COVID-19 Infection in 2020

	Dates	Total Cases	Warded	ICU
1				
2				
3				
:				
:				
365				

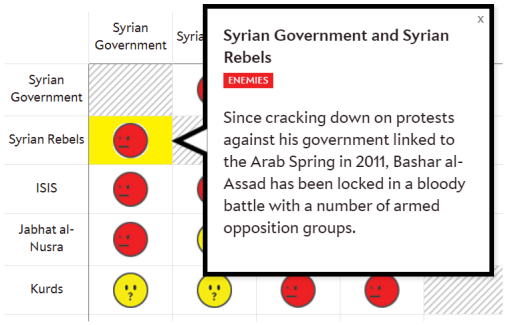
(c)

Figure 2 – Data table examples and the suggested visual mark that can use to design the chart.

2.3 Deconstruction Exercises

(1) **Middle East Friendship Chart**. Figure 3(a) shows an interactive visualisation done by Joshua Keating and Chris Kirk (@Slate.com). It depicts the relationships between the active players involved in the civil war in Syria. If you click on any coloured emoji icon, a text description is displayed next to the icon describing the nature of the relationship between the two parties, as shown in Figure 3(b).

(a)



(b)

Figure 3 – Middle East Friendship Chart

(From http://www.slate.com/blogs/the_world_/2014/07/17/the_middle_east_friendship_chart.html)

a) Draw the data table for the relational data model used to create the visualisation shown in Figure 3. You can give appropriate names to the different attributes in your columns.

b) What visual marks and variables (channel) are used to encode the different data depicted? Color, Position, Shape

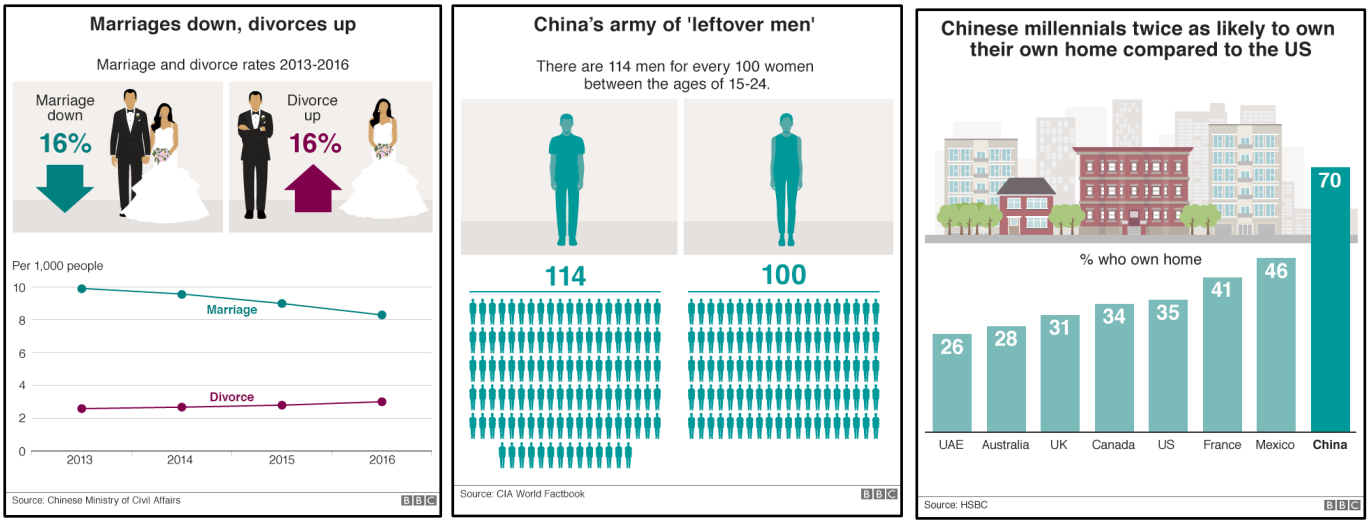
c) Describe if redundant coding has used in the visualisation and if so, discuss if it was done effectively.

d) Describe if the choice of colours used in visual encoding is appropriate.

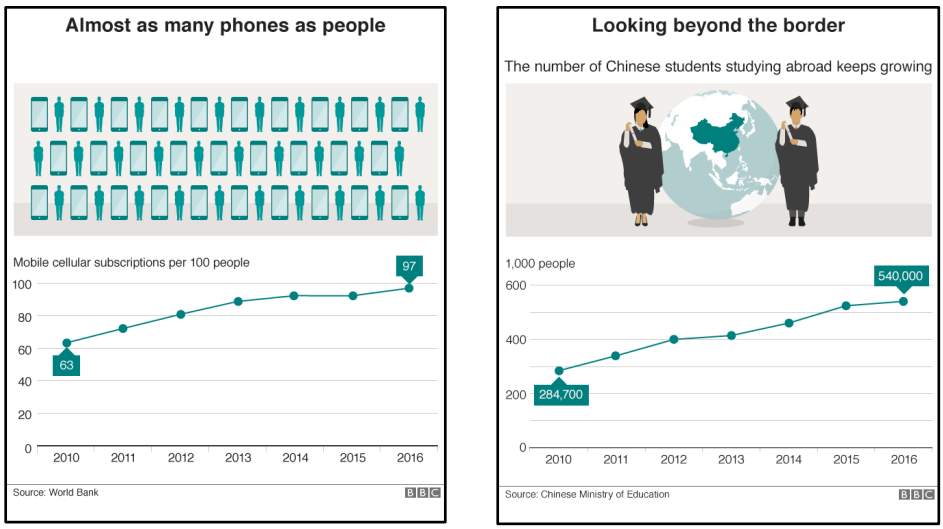
2.4 Critical Eye

(1) **BBC News – Fortunes of Chinese Family**. On 15 Oct 2017, BBC News published an article entitled **Five charts about the fortunes of the Chinese family**. In this article, 5 charts were presented on various statistical facts about the state of the Chinese family in China. These charts are shown in Figure 4.

Look at these charts and the visual encoding used in each case. Critic these visuals (infographics). Comment if they were done well (i.e. the information intended to be conveyed was effectively communicated with the choice of visual encoding). Suggest if any of these charts could be improved visually to allow it to express the intended message more clearly and quickly.



(a) (b) (c)



(d) (e)

Figure 4 – Five charts from a BBC News article (Five charts about the fortunes of the Chinese family)
(Image from https://www.bbc.com/news/world-asia-china-41424041)

Optional Challenge (Give this a try)

(2) **Euro Cup 2008**. Duch, Waitzman and Amaral analysed three knockout-phase matches of the Spanish team using the vast amount of statistical information that was published online by UEFA. Using techniques from social network analysis, they produced the visualisation shown in Figure 5. Based on further descriptions given in their PLOS ONE open-access paper accessible at https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0010937, answer the following questions:

a) Identify as many types of data variables as possible that have been visually represented in the visualisation shown in Figure 4. For each, state its scale of measure (N, O or Q).

b) Describe what perceptual channel (visual variable) has been employed to encode the values of each of these data variables.

c) Describe if the choice of colours used in visual encoding is appropriate. If not, suggest how it could be improved.

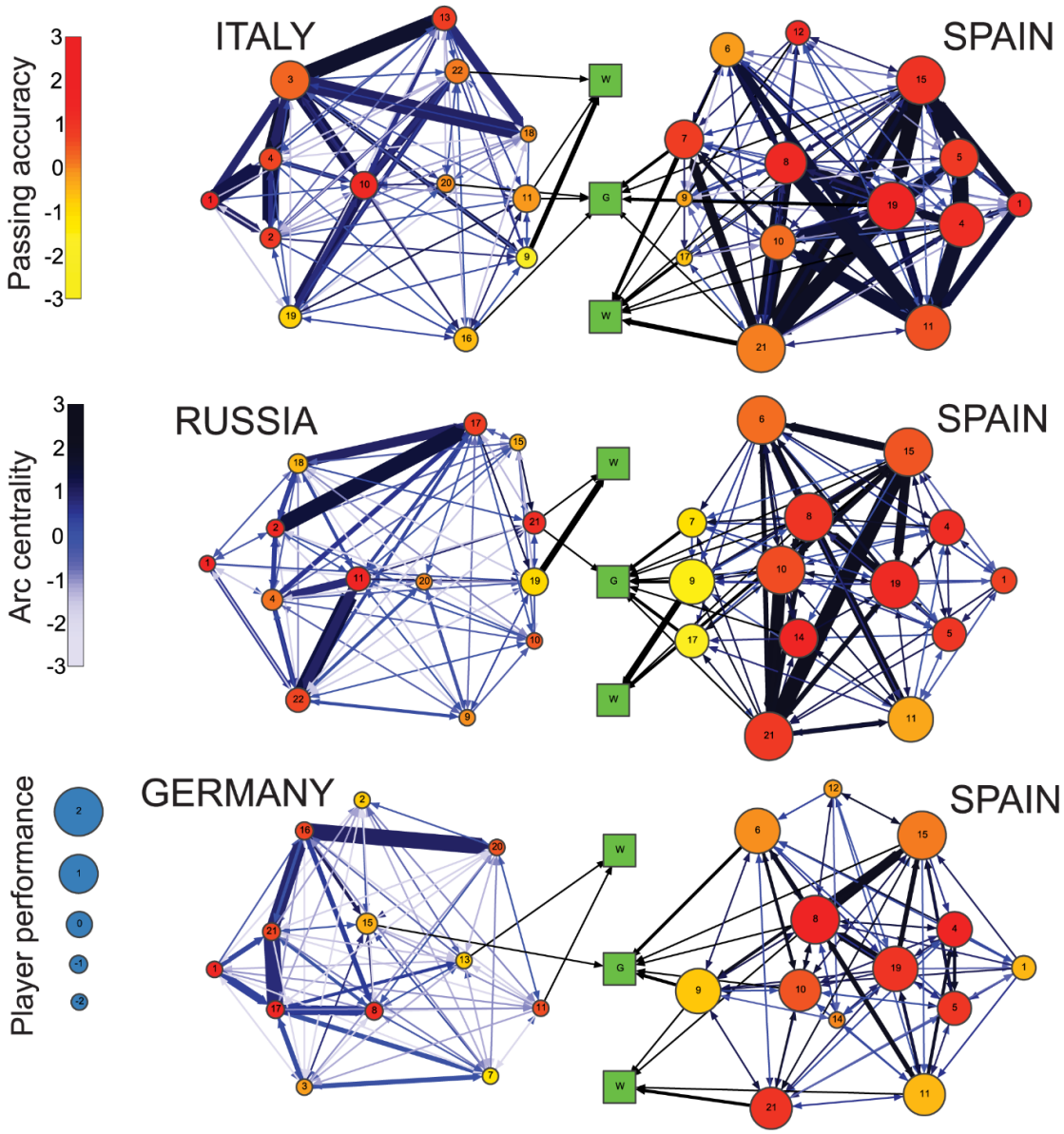


Figure 5 – Visualisation of the three knockout-phase matches of the Spanish team.
(Image from https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0010937.g005)

2.1 TUT 2

1 - Points
- lines
- areas

2 - position
- size ✓ for line
- value
- Texture
- colour
- orientation
- shape

3 Points
lines
area
position
size ✓
Texture
Colour

2.2

1)

Lines

Living Members from Class of '45

	Names	Gender	Age
1			
2			
3			
4			
5			
6			

(a)

Points

Orientation T-shirts - Comparing Halls

	No.	Gender	Weight	Height	Hall (A,B,C,D)	T-shirt Size
1						
2						
3						
:						
:						
700						

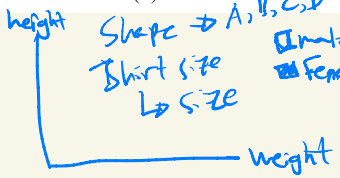
(b)

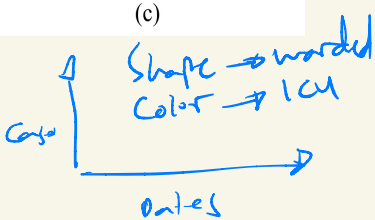
Areas

COVID-19 Infection in 2020

	Dates	Total Cases	Warded	ICU
1				
2				
3				
:				
:				
365				

(c)





3.1 Line and Bar Charts

(To be done over Weeks #4 & #5)

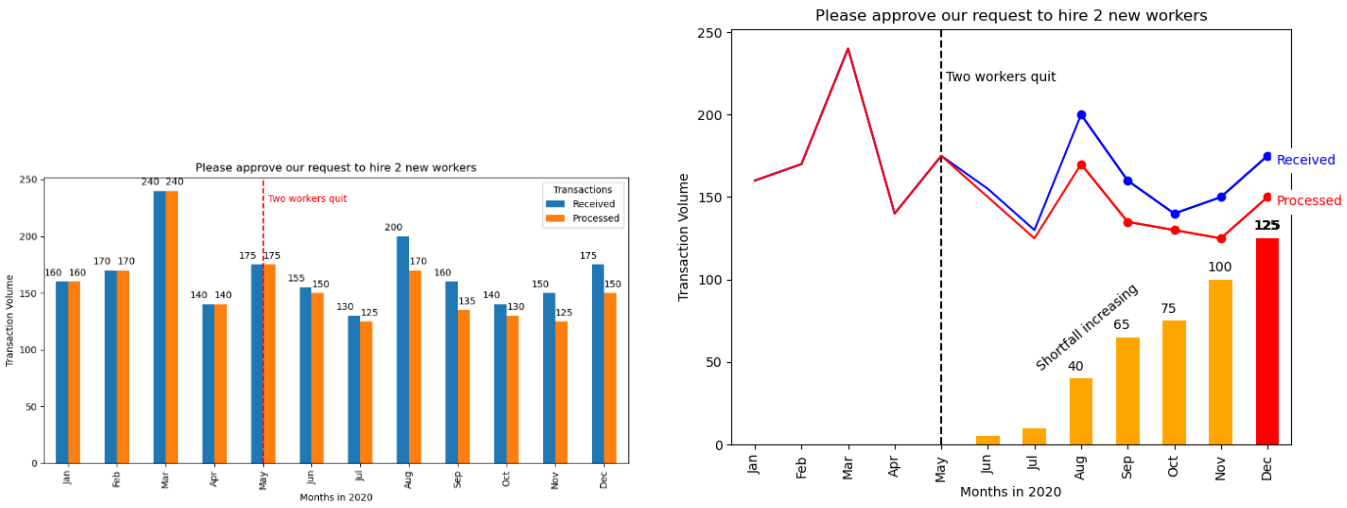
(1) Table 1 shows the transaction volume received and processed over the last 12 months of Jan to Dec. This data table can be found in the file "*hire2FTE.csv*". Using the Matplotlib and/or Seaborn Python visualisation library, create the cluster bar chart shown in Figure 1(a). Do note that you may have to perform some data wrangling (as in Tutorial #1) to prepare your dataframe for plotting the bar chart shown.

You are to annotate the pertinent information shown, which include the following:

a) Line marker showing two workers quit in the month of May.
b) The actual transaction volumes received and processed for each month.
c) The title of the plot is "**Please approve our request to hire 2 new workers**"
d) The *x* and *y* axis labels.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Transactions	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2	Received	160	170	240	140	175	155	130	200	160	140	150	175
3	Processed	160	170	240	140	175	150	125	170	135	130	125	150

Table 1 – The data table in the csv data file "hire2FTE"



(a) (b)

Figure 1 – (a) Cluster bar chart showing the monthly volumes of transaction receive and processed. (b) An equivalent combination of line and bar plots showing the same information.

(2) The purpose of your data visualisation is to present a case to your upper management that you need to quickly hire two additional full-time employees (FTE) as your team's ability to keep up with transaction volume received is falling way behind. Comment on the main reason why the bar chart in Figure 1(a) is not suitable for this purpose.

(3) Create the two red and blue line plots shown in Figure 1(b), along with the various marker and text annotations shown. Comment on their effectiveness, including choice of colours, marker placement, text annotation, highlights, etc. Suggest ways to further improve (if any).

(4) **Challenge** (Optional): Superimpose the increasing shortfall bar plots over your two line plots. The shortfall is computed by cumulating the monthly differences between the transactions received and processed. Include the annotations and highlights as shown in Figure 1(b) and comment on their effectiveness.

3.2 The Perils of Line Smoothing

(1) Smoothing data series is a method to remove local variations (noise) in data so that a more general trend can be observed in the data variations. However, there are situations when such smoothing operations actual distort the truth about what the data is actually telling us. First, study the profit figures in Table 2 and answer the following questions:

a) Which department achieved the highest ever annual profit over the last 7 years and in which year was this? X (7.5 in 2017)

b) Did Dept Z ever had higher annual profits compared to Dept X? If so, over which period was this? Yes (2021)

c) Describe the difference in profits differences between Dept Y and Z during the earlier days of this company (i.e. years 2015 and 2016). 0.1

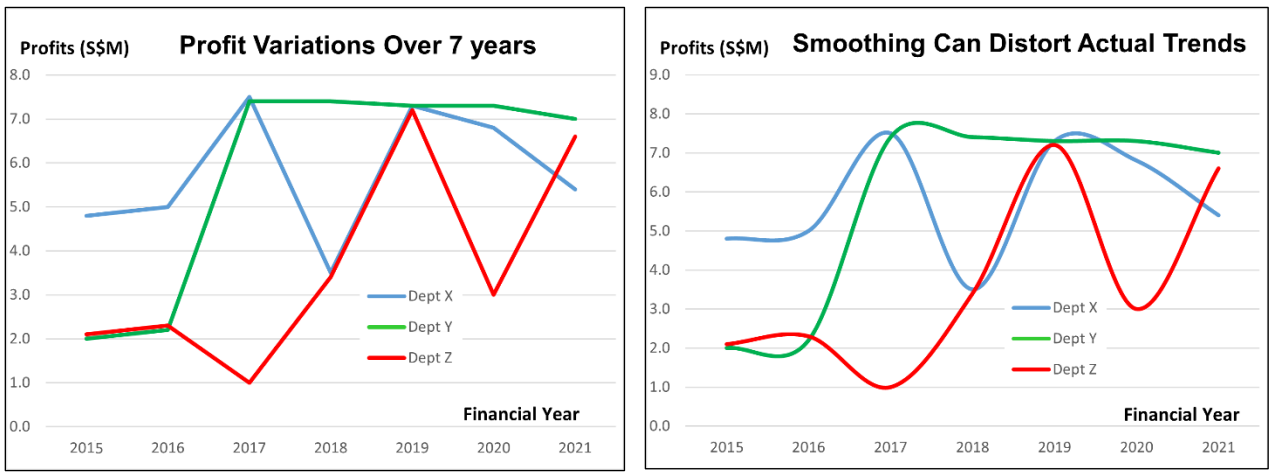
d) Year 2017 to 2019 were years of change. Did any department managed to maintain their profits? Only dept Y keeping steady between 7.3 to 7.4, no change at all in 2017 to 2018

e) Describe the relative performance of the three departments in year 2019.

Year	Profits in last 7 years (S$ M)		
	Dept X	Dept Y	Dept Z
2015	4.8	2.0	2.1
2016	5.0	2.2	2.3
2017	7.5	7.4	1.0
2018	3.5	7.4	3.4
2019	7.3	7.3	7.2
2020	6.8	7.3	3.0
2021	5.4	7.0	6.6

Table 2 – Department Profits over 7 years.

(2) Second, study the two line charts in Figure 2 carefully and described all the misleading trends created due to the application of the Excel smoothed line function in Figure 2(b)?



(a) (b)

Figure 2 – Excel line chart plots of the departmental profit variations in Table 1. (a) The three line plots of the actual profits of each departments X, Y and Z. (b) The line chart plotted using the "Smoothed line" option in Excel Line Chart.

(3) **Challenge** (Optional):

Using the Excel file "*dept profits.xlsx*" provided, create a clustered bar chart shown in Figure 2(d) to help you compare the annual profits between the three departments over the period of 2015 to 2021. You can do this quickly in Excel.

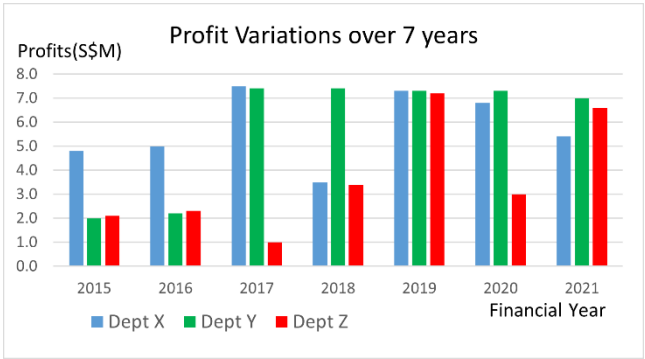


Figure 2(d)

3.3 Critical Eye

(1) **Communicating Singapore's COVID-19 Situation**. Figure 3 shows an infographic taken from The Straits Times. It describes a snapshot of the COVID-19 infection situation in Singapore from the start of May till the day of publication on 8 June 2021.

a) Why do you think both line and bar charts were used in visualising the virus infection data shown in the upper portion of Figure 3?

b) Three lines of different colours were used to depict the total new cases, community cases and unlinked cases. Would a stacked area chart be more appropriate since these data values by their very nature, will never overlap? Give reasons for your answer.

c) What about the bar chart in the background? Would it be better to replace it with an area chart of the same colour?

d) Why was a marker line drawn over the line charts at May 16? What can you say about the perceptual channel used on this marker line?

e) What story do you think is being communicated with the inclusion of this marker at May 16?

f) Why was a donut chart used to visualize the unlinked cases?

g) Could the donut chart be replaced with a pie chart? If so, what changes would you make to the visual design.

h) Should each slice of the donut chart be encoded with a different colour to improve visual clarity? Is there any particular reason why only two colours have been used in the donut chart?

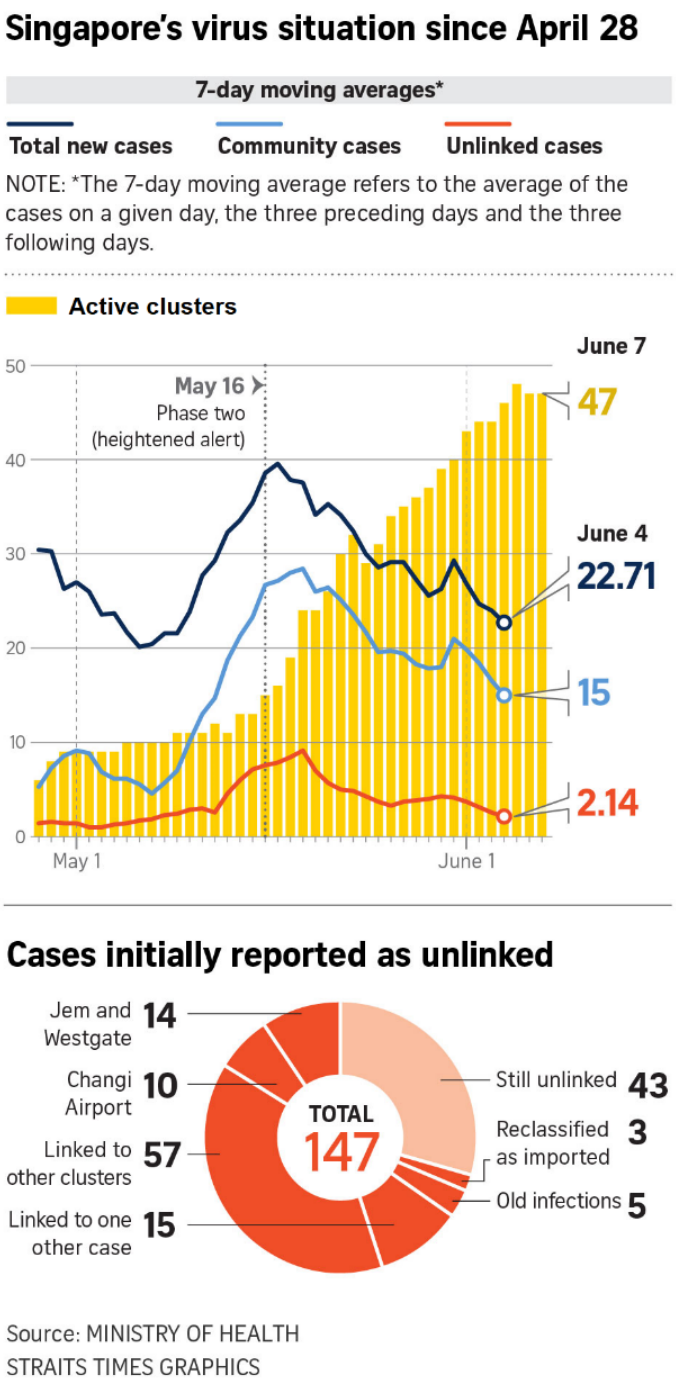


Figure 3 – COVID-19 Infection Situation in Singapore

(Image modified from The Straits Times on 8 June 2021 at https://www.straitstimes.com/singapore/health/unlinked-covid-19-cases-connection-to-clusters-found-within-days-in-singapore)

TUT 3

3.1 Line and Bar Charts
(1) Plotting Line Chart

Plot the two lines &dot markers

```
import pandas as pd
import matplotlib.pyplot as plt

# Read in csv file and separate Rec and Proc
Transaction = pd.read_csv('C:/Datasets/Hire2FTE.csv')
Receive = Transaction.loc[0,'Jan':'Dec']
Process = Transaction.loc[1,'Jan':'Dec']
MonthDiff = Receive-Process
ShortFall = MonthDiff

# plot two lines for Receive and Process
plt.plot(Receive,color='blue')          # plot Receive line in blue
plt.plot(Process,color='red')           # plot Process line in red

# draw dot markers
plt.plot(Receive['Aug':'Dec'], marker='o', color='blue')   # blue dots
plt.plot(Process['Aug':'Dec'], marker='o', color='red')    # red dots
```

highlight here

label 2 lines (blue n red)

3.1 Line and Bar Charts
(1) Plotting Line Chart

Label title, axes and shortfall bars

```
import pandas as pd
import matplotlib.pyplot as plt

# Read in csv file and separate Rec and Proc
Transaction = pd.read_csv('C:/Datasets/Hire2FTE.csv')
Receive = Transaction.loc[0,'Jan':'Dec']
Process = Transaction.loc[1,'Jan':'Dec']
MonthDiff = Receive-Process
ShortFall = MonthDiff

# rotate shortfall annotated text
plt.text(6,45,'Shortfall increasing',color='black',rotation=40)

# add axes labels and title
plt.title('Please approve our request to hire 2 new workers')   # plot title
plt.xlabel('Months in 2020')                                    # x-axis label
plt.ylabel('Transaction Volume')                                # y-axis label
```

3.1 Line and Bar Charts
(1) Plotting Line Chart

Compute and plot shortfall bars

```
import pandas as pd
import matplotlib.pyplot as plt

# Read in csv file and separate Rec and Proc
Transaction = pd.read_csv('C:/Datasets/Hire2FTE.csv')
Receive = Transaction.loc[0,'Jan':'Dec']
Process = Transaction.loc[1,'Jan':'Dec']
MonthDiff = Receive-Process
ShortFall = MonthDiff

# Cumulate the monthly difference into ShortFall
Total = 0,
red
for Index in MonthDiff:
    Total = Total + Index
    ShortFall.iloc[n] = Total
    n = n + 1

# draw shortfall bar plot using Pandas
Fig1 = ShortFall.plot.bar(color='orange')
```

3.1 Line and Bar Charts
(1) Plotting Line Chart

Highlight last shortfall bar in red

```
import pandas as pd
import matplotlib.pyplot as plt

# Read in csv file and separate Rec and Proc
Transaction = pd.read_csv('C:/Datasets/Hire2FTE.csv')
Receive = Transaction.loc[0,'Jan':'Dec']
Process = Transaction.loc[1,'Jan':'Dec']
MonthDiff = Receive-Process
ShortFall = MonthDiff

# highlight the last shortfall bar
LastShortFall = ShortFall.copy()
LastShortFall['Dec'] = 0
Fig1 = ShortFall.plot.bar(color='red')          # plot all red
Fig1 = LastShortFall.plot.bar(color='orange')   # turn Jan-Nov back to orange
```

3.1 Line and Bar Charts
(1) Plotting Line Chart

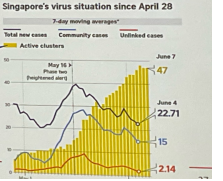
Annotate values of shortfall bars

```
import pandas as pd
import matplotlib.pyplot as plt

# Read in csv file and separate Rec and Proc
Transaction = pd.read_csv('C:/Datasets/Hire2FTE.csv')
Receive = Transaction.loc[0,'Jan':'Dec']
Process = Transaction.loc[1,'Jan':'Dec']
MonthDiff = Receive-Process
ShortFall = MonthDiff

# label shortfall bar numbers
LabelCount = 0
TextYPos = LabelCount + 0.7
LabelMonth = ShortFall['Aug':'Dec']
for Index in LabelMonth:
    TextYPos = Index + 5
    TextXPos = TextXPos + 1
    plt.text(TextXPos,TextYPos,Index,color='black')
plt.text(TextXPos,TextYPos,Index,color='black',fontweight='bold')
```

3.3 Critical Eye
(1) Communicating SG's COVID Situation

a) Why do you think both line and bar charts were used in visualising the virus infection data?

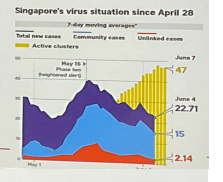
- The bar and line describe **different types of data.** Cluster = minimum of three cases
- The bar describes the number of active clusters.
- The line describes number of individual cases.
- Using both helps make this distinction clear and **avoid confusion.**



3.3 Critical Eye
(1) Communicating SG's COVID Situation

b) Lines with 3 different colours used. Would a stacked area chart be better (these data values will never overlap)? Say why?

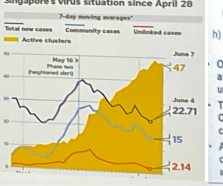
- **Stacked area chart would not be appropriate.**
- **It would occlude the active cluster information presented on the bar chart.**



3.3 Critical Eye
(1) Communicating SG's COVID Situation

c) What about the bar chart in the background? Would it be better to replace it with an area chart of the same colour?

- **Possible but not as effective as bar chart.**
- **Individual bars capture exact value for each day** and the abrupt changes in data values much better.
- Texture of bar chart gives good contrast to the 3 line plots.



3.3 Critical Eye
(1) Communicating SG's COVID Situation

d) Why was marker line drawn at May 16? What can you say about the perceptual channel used on this marker line?

- **The line at May 16 marks an important event** (start of the imposition of Phase 2-Heightened alert).
- **This line is dotted** (texture perceptual channel) **instead of solid so it is visually distinct** from the three lines use for the line chart.

3.3 Critical Eye
(1) Communicating SG's COVID Situation

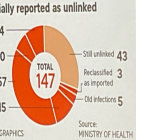
e) What story do you think is being communicated with the inclusion of this marker at May 16?

- By drawing this time marker over the line charts, it communicates the **effectiveness** of the heightened alert.
- Line charts show **clear changing trend** of falling infection cases soon after the line at May 16.

3.3 Critical Eye
(1) Communicating SG's COVID Situation

f) Why was a donut chart used to visualize the unlinked cases?

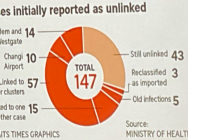
- The donut chart is used as this visualization communicated the **parts of the whole** (i.e. the different types of unlinked cases within the 147 total cases).



3.3 Critical Eye
(1) Communicating SG's COVID Situation

g) Could the donut chart be replaced with a pie chart? If so, what changes would you make to the visual design?

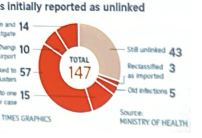
- The donut chart is better as the 147 total can be communicated clearly in the empty centre of the donut chart.
- If must use pie chart, some space must be used to annotate TOTAL 147.
- Cannot expect reader to do the maths by adding up all the numbers.



3.3 Critical Eye
(1) Communicating SG's COVID Situation

h) Should each donut slice have a different colour to improve visual clarity? Any reason for only two colours here?

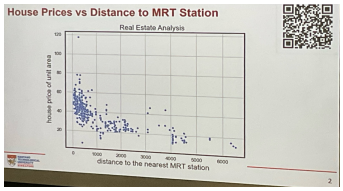
- **Only two colours is deliberate here as the significance is on "still unlinked" cases.**
- **This group is critical in curbing COVID-19 spread. Using more colours diminishes this message.**
- **Also add little additional visual clarity** (white segments markers and text annotation are already clearly visible).



! more effective to use darker colour to highlight key segment

3.4 Analysing Relationships

(1) **What Affects Property Prices**? The data file "*Real estate.csv*" provides data on 414 property transactions. The transaction price for each property sold is quoted based on cost per unit area, which means it gives the general value of the property and is not affected by its size. Associated with each transaction is also listed the following characteristics of the property that is of concern to us, namely:



a) Unit area price of the property ("**house price of unit area**")
b) Age of the property ("**house age**")
c) Distance from a MRT station ("**distance to the nearest MRT station**")
d) Number of convenient stores nearby ("**number of convenience stores**")

Using an appropriate chart and relationship model fitting technique (e.g. linear regression, polynomial regression, lowess, etc) visualise the general relationships between the various variables list above.

Real estate dataset from: https://www.kaggle.com/quantbruce/real-estate-price-prediction?select=Real+estate.csv

(2) **Distance to MRT** - What can you say about the general price of a property with respect to its distance from the nearest MRT station? What do you think is the reason for this relationship? Is this relationship a simple linear one or is it more complex? Tell this story with evidence from your visualizations.

Food for thought: https://www.propertyguru.com.sg/property-guides/mrt-effect-on-property-prices-39498 (accessed in June'21)

(3) **Age of the property** - What can you say about the general price of a property with respect to its age? Does old really means cheap? What is the reason for the relationship (Hint: Does the issue in part (2) has something to do with this? Make your case with evidence from your visualizations.

(4) **Number of convenience stores** – Do note that the data on **number of convenience stores** near the property is discrete and ordinal, as such it should be treated differently from the other data categories available in the data set (Hint: Plot the number of convenience stores on the x-axis so you can do multiple horizontal distribution visualisations using appropriate Seaborn plots).

(a) What can you say about the general price of a property with respect to the number of convenience stores in the neighbourhood? Make your case with evidence from your visualizations.

(b) What kinds of neighbourhood has many convenience stores? Do you think you can use the data in *Real estate* to do this analysis? If you can, in the light of this new analysis, what can you say about the case you made earlier in part 4a?

(5) **Faceting** - Relationships influencing a primary variable like the house price is multi-faceted and making hasty conclusions with a single visualisation should be avoided without first visualising and analysing the relationships between the other variables.

(a) Explore Seaborn's many powerful methods to create grids of multiple plots to explore the various relationship (checkout: https://seaborn.pydata.org/tutorial/axis_grids.html).

(b) What do you think is the primary factor influencing the general price of a property based on the data available in *Real estate.csv*? Provide the visualisations to support your view.

3.5 Analysing Distributions

(1) **Factors Affecting Students Academic Performance**. The data file "*Students performance.csv*" provides data on the math, reading and writing scores of 1000 students from five different ethnic groups A to E. Other information like gender, parent's education and whether they have completed their test preparation course are also given. Analyse these data with appropriate distribution plots. In some cases, you will need to **wrangle the data** into suitable format before you can use the various Seaborn distribution plots to visualise and compare the distributions.

Students performance dataset from: https://www.kaggle.com/spscientist/students-performance-in-exams

(2) **Gender Differences** - What can you say about the relative average performance differences between male and female students across the three different subjects? What subjects do girls generally do better in and what subjects do the boys do better in? Show this analysis using appropriate comparative distribution plots (*Hint*: consider using multiple violin plot, with its left/right split feature).

(3) **Toughest Subject** – Which of the three subjects did the entire cohort performed the worst in? Show this analysis using appropriate comparative distribution plots and annotate the median value for each of the subjects into your plot (*Hint*: consider using multiple box plots).

(4) **Test Preparation Course** – Does completing the test preparation course help students to do better in the three different subjects? Which subject shows the most improvement when the preparation course is completed? Show this analysis using appropriate comparative distribution plots and annotate the median values for both completion and non-completion of preparation course for each of the subjects into your plot to make it easy to visualize the different medians.

(5) **Performance across Ethnic Groups** – How did the different ethnic groups A to E performed in their tests? Which group did the best overall in all three subjects? (*Hint*: consider using a series of overlapping kernel density estimate plots).

(6) **Subject Performance across Ethnic Groups** – Were there performances differences in the three different subjects across the five different groups A to E (e.g. did one group do the best in math and another group did the best in reading, or did one group did the best in all three subjects)? (*Hint*: consider using a series of coloured box plots).

(7) **Influence of Parent's Education on Students' Performance** – Did the parent's educational background have an influence on their children's performance in general (i.e. all their subjects combined)? If they did, which of these educational backgrounds seems to result in the best performance and which resulted in the worst?

(8) **Influence of Parent's Education on Gender** – If the parent's educational background have an influence of the students' performance, was this influence more pronounce for the male or female students?

Optional Challenge

(Additional optional dataset to explore)

3.6 Analysing the Data

(1) **What Happened to the Pandemic in Malaysia?** A data file entitled "*owid MYS.csv*" has been provided for you, which is a subset of COVID data related to the COVID pandemic cases in Malaysia extracted on 9 June 2021 from the Our World in Data website at https://github.com/owid/covid-19-data/tree/master/public/data. Study the data table and locate the column labelled "**new_cases**".

(2) Using the Seaborn and any other relevant Python libraries, wrangle the data table given to you and plot the bar chart for the **new_cases** over the 500 days starting on the date '**25/1/2020**'.

(3) Superimpose over the bar chart an appropriate smoothed line so that you can visualize the general changing trends of number of new confirmed COVID-19 cases over the 500 days.

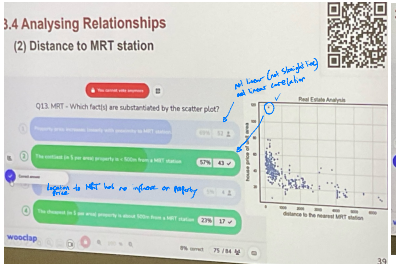
(4) Based on some of the relevant COVID-19 related news and web links (and any others you think relevant), annotate on your plot (i.e. using appropriate visual markers and text) any relevant events that you think may have influence the changing infection trends in your chart.

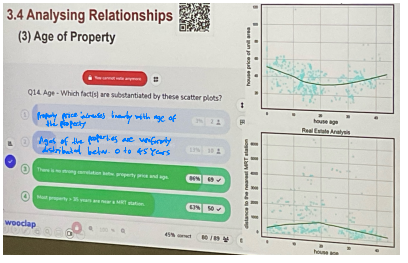
References to some COVID-19 relevant events in Malaysia:

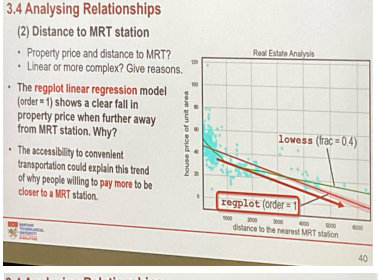
General info - https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Malaysia
27 Feb 2020 - https://en.wikipedia.org/wiki/2020_Tablighi_Jamaat_COVID-19_hotspot_in_Malaysia
26 Sep 2020 - https://www.straitstimes.com/asia/se-asia/malaysias-pm-muhyiddin-admits-sabah-state-polls-in-sept-caused-current-covid-19-wave
23 Dec 2020 - https://www.straitstimes.com/asia/se-asia/malaysia-has-identified-new-covid-19-strain-similar-to-one-found-in-3-other-nations
16 Jan 2021 - https://asia.nikkei.com/Spotlight/Coronavirus/Malaysia-s-Top-Glove-reports-COVID-19-outbreak-at-four-factories
13 Apr 2021 - https://www.straitstimes.com/asia/se-asia/malaysia-relaxes-some-coronavirus-curbs-as-fasting-month-arrives
2 May 2021 - https://www.bangkokpost.com/world/2109227/malaysia-reports-first-case-of-indian-covid-19-variant

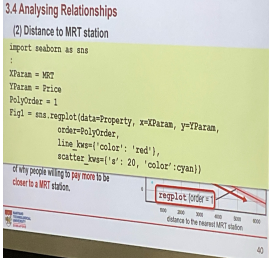
Note: Some of the web links may be out of date and no longer accessible. You can infer the related event from the title of the web link.

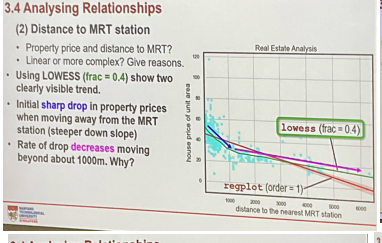
3.4

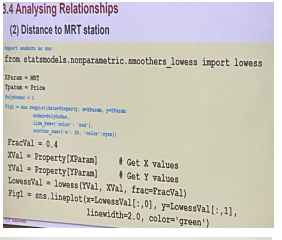

3.4 Analysing Relationships
(2) Distance to MRT station
Q13. MRT - Which fact(s) are substantiated by the scatter plot?
Real Estate Analysis
house price of unit area
distance to the nearest MRT station
wooclap

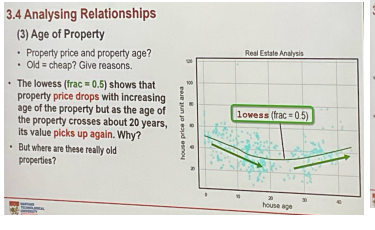

3.4 Analysing Relationships
(3) Age of Property
Q14. Age - Which fact(s) are substantiated by these scatter plots?
house price of unit area
house age
distance to the nearest MRT station
wooclap

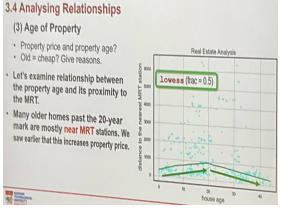

3.4 Analysing Relationships
(2) Distance to MRT station
Property price and distance to MRT?
Linear or more complex? Give reasons.
The regplot linear regression model (order = 1) shows a clear fall in property price when further away from MRT station. Why?
The accessibility to convenient transportation could explain this trend of why people willing to pay more to be closer to a MRT station.
Real Estate Analysis
house price of unit area
lowess (frac = 0.4)
regplot (order = 1)
distance to the nearest MRT station
40

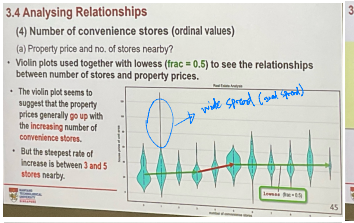

3.4 Analysing Relationships
(2) Distance to MRT station
import seaborn as sns
XParam = MRT
YParam = Price
PolyOrder = 1
Fig1 = sns.regplot(data=Property, x=XParam, y=YParam,
order=PolyOrder,
line_kws={'color': 'red'},
scatter_kws={'s': 20, 'color':cyan})
of why people willing to pay more to be closer to a MRT station.
regplot (order = 1)
distance to the nearest MRT station
40

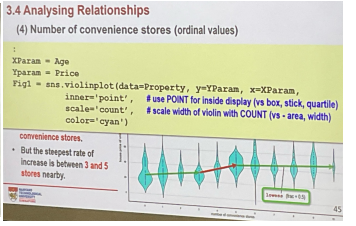

3.4 Analysing Relationships
(2) Distance to MRT station
Property price and distance to MRT?
Linear or more complex? Give reasons.
Using LOWESS (frac = 0.4) show two clearly visible trend.
Initial sharp drop in property prices when moving away from the MRT station (steeper down slope)
Rate of drop decreases moving beyond about 1000m. Why?
Real Estate Analysis
house price of unit area
lowess (frac = 0.4)
regplot (order = 1)
distance to the nearest MRT station

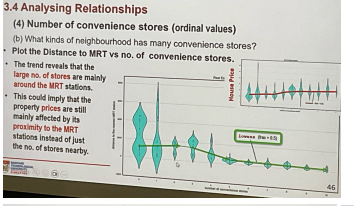

3.4 Analysing Relationships
(2) Distance to MRT station
from statsmodels.nonparametric.smoothers_lowess import lowess
XParam = MRT
YParam = Price
FracVal = 0.4
XVal = Property[XParam] # Get X values
YVal = Property[YParam] # Get Y values
LowessVal = lowess(YVal, XVal, frac=FracVal)
Fig1 = sns.lineplot(x=LowessVal[:,0], y=LowessVal[:,1],
linewidth=2.0, color='green')

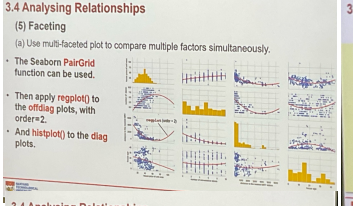

3.4 Analysing Relationships
(3) Age of Property
Property price and property age?
Old = cheap? Give reasons.
The lowess (frac = 0.5) shows that property price drops with increasing age of the property but as the age of the property crosses about 20 years, its value picks up again. Why?
But where are these really old properties?
Real Estate Analysis
house price of unit area
lowess (frac = 0.5)
house age

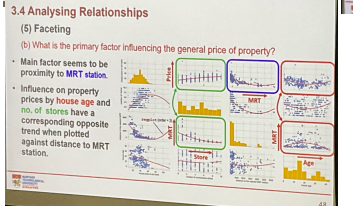

3.4 Analysing Relationships
(3) Age of Property
Property price and property age?
Old = cheap? Give reasons.
Let's examine relationship between the property age and its proximity to the MRT.
Many older homes past the 20-year mark are mostly near MRT stations. We saw earlier that this increases property price.
Real Estate Analysis
lowess (frac = 0.5)
house age

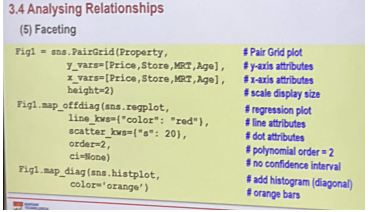

3.4 Analysing Relationships
(4) Number of convenience stores (ordinal values)
(a) Property price and no. of stores nearby?
Violin plots used together with lowess (frac = 0.5) to see the relationships between number of stores and property prices.
The violin plot seems to suggest that the property prices generally go up with the increasing number of convenience stores.
But the steepest rate of increase is between 3 and 5 stores nearby.
inside spread (quartiles)
lowess (frac = 0.5)
45


3.4 Analysing Relationships
(4) Number of convenience stores (ordinal values)
XParam = Age
Yparam = Price
Fig1 = sns.violinplot(data=Property, y=YParam, x=XParam,
inner='point', # use POINT for inside display (vs box, stick, quartile)
scale='count', # scale width of violin with COUNT (vs - area, width)
color='cyan')
convenience stores.
But the steepest rate of increase is between 3 and 5 stores nearby.
lowess (frac = 0.5)
45


3.4 Analysing Relationships
(4) Number of convenience stores (ordinal values)
(b) What kinds of neighbourhood has many convenience stores?
Plot the Distance to MRT vs no. of convenience stores.
The trend reveals that the large no. of stores are mainly around the MRT stations.
This could imply that the property prices are still mainly affected by its proximity to the MRT stations instead of just the no. of stores nearby.
House Price
lowess (frac = 0.5)
46


3.4 Analysing Relationships
(5) Faceting
(a) Use multi-faceted plot to compare multiple factors simultaneously.
The Seaborn PairGrid function can be used.
Then apply regplot() to the offdiag plots, with order=2.
And histplot() to the diag plots.


3.4 Analysing Relationships
(5) Faceting
(b) What is the primary factor influencing the general price of property?
Main factor seems to be proximity to MRT station.
Influence on property prices by house age and no. of stores have a corresponding opposite trend when plotted against distance to MRT station.
Price
MRT
Stores
Age
48


3.4 Analysing Relationships
(5) Faceting
Fig1 = sns.PairGrid(Property,
y_vars=[Price,Store,MRT,Age],
x_vars=[Price,Store,MRT,Age],
height=2) # Pair Grid plot # y-axis attributes # x-axis attributes # scale display size
Fig1.map_offdiag(sns.regplot,
line_kws={"color": "red"},
scatter_kws={"s": 20},
order=2,
ci=None) # regression plot # line attributes # dot attributes # polynomial order = 2 # no confidence interval
Fig1.map_diag(sns.histplot,
color='orange') # add histogram (diagonal) # orange bars

3.5

3.5 Analysing Distributions

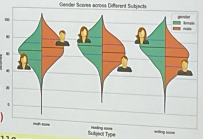
(2) Gender Differences

Relative performance between male and females?

Solution:

Use split violin plots.

- Males do a little better in Maths
- Female students do better in language subjects (reading & writing)



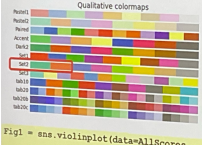

```
Fig1 = sns.violinplot(data=AllScores,
                      y='scores',x='subjects',
                      hue='gender', split=True,
                      inner='quartile', palette='Set2')
```

Matplotlib palette - https://matplotlib.org/stable/tutorials/colors/colormaps.html

50

3.5 Analysing Distributions

Qualitative colormaps




```
Fig1 = sns.violinplot(data=AllScores,
                      y='scores',x='subjects',
                      hue='gender', split=True,
                      inner='quartile', palette='Set2')
```

Matplotlib palette - https://matplotlib.org/stable/tutorials/colors/colormaps.html

50

3.5 Analysing Distributions

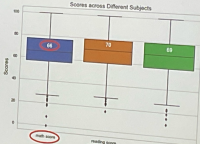
(3) Toughest subject

Which subject did the students did worst in?

Solution:

Use box plot.

Maths seems to be the most difficult subject for the average student.




```
Fig1 = sns.boxplot(y='scores', x='subjects', data=AllScores)
```

3.5 Analysing Distributions

(3) Toughest subject

Which subject did the students did worst in?

Solution: math


```
medians = AllScores.groupby(['subjects'])['scores'].median()
vert_offset = AllScores['scores'].median() * 0.05 # offset from median
for xtick in Fig1.get_xticks():
    Fig1.text(xtick,
              medians[xtick] + vert_offset,
              medians[xtick],
              horizontalalignment='center',
              size='large',
              colorm='w',
              weight='semibold')
```

x position
y position
text to print
text alignment
text size
text colour
semi BOLD font

3.5 Analysing Distributions

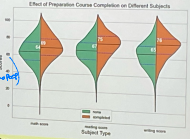
(4) Test Preparation Course

Did the prep course help?

Solution:

Use violin plot. (can use boxplot too but not here because)

Completing test preparation course seems to help improve average scores in all subjects. Most in writing (+11%)




```
Fig1 = sns.violinplot(data=AllScores,
                      y='scores',x='subjects',
                      hue='test preparation course', split=True,
                      inner='quartile', palette='Set2')
```

3.5 Analysing Distributions

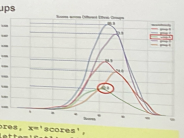
(5) Performance of Ethnic groups

Did performance vary among the 5 groups?

Solution:

Use KDE plot.

- Groups D and E did the best.
- Group A did the worst (62%)




```
Fig1 = sns.kdeplot(data=AllScores, x='scores',
      hue='race/ethnicity', palette='Set1', linewidth=2)
```

3.5 Analysing Distributions

(6) Subject Performance of Ethnic groups

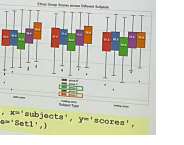
Did subject performance vary among the 5 groups?

Solution:

Use Box plot.

Group A did the worst in all three subjects.

Group E did the best in all three subjects (especially Maths)




```
Fig1 = sns.boxplot(data=AllScores, x='subjects', y='scores',
      hue='race/ethnicity', palette='Set1',)
```

3.5 Analysing Distributions

(7) Parent education on student scores

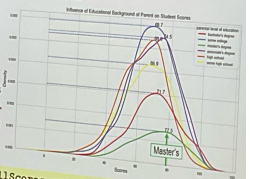
Did parental education level influence student's score?

Solution:

Use KDE plot.

Parent's education level does seem to have some influence.

- Master's degree did the best
- high school did worst




```
Fig1 = sns.kdeplot(data=AllScores,x='scores',
      hue='parental level of education',
      palette='Set1', linewidth = 3)
```

3.5 Analysing Distributions

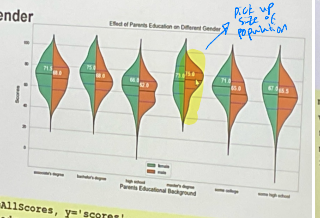
(8) Parents Ed and Gender

Did parent's education influence gender differently?

Solution:

Use split violin plot.

pick up size of population

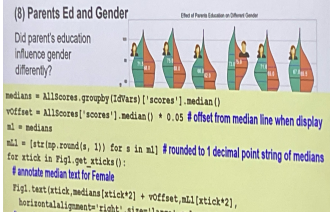



```
Fig1 = sns.violinplot(data=AllScores, y='scores',
      x='parental level of education', hue='gender', split=True,
      inner='quartile',palette='Set2', order=[ "associate's degree","bachelor's
      degree","high school","master's degree","some college","some high school"] )
```

3.5 Analysing Distributions

(8) Parents Ed and Gender

Did parent's education influence gender differently?




```
medians = AllScores.groupby(ldVars)['scores'].median()
vOffset = AllScores['scores'].median() * 0.05 # offset from median line when display
ml = medians
mll = [str(np.round(s, 1)) for s in ml] # rounded to 1 decimal point string of medians
for xtick in Fig1.get_xticks():
  # annotate median text for Female
  Fig1.text(xtick,medians[xtick*2] + vOffset,mll[xtick*2],
    horizontalalignment='right',size='large',color='w',weight='semibold')
  # annotate median text for Male
  Fig1.text(xtick,medians[(xtick*2)+1] + vOffset,mll[(xtick*2)+1],
    horizontalalignment='left',size='large',color='w',weight='semibold')
```

4.1 Weber's Law

(To be Done Before Week #6 and #7)

(1) Using Powerpoint *(or any application where you can create grey level squares with any intensity values*), create eight reasonable-sized and equally-spaced squares with incrementing Red, Green, Blue (R, G, B) intensity values as shown in Figure 1.

(2) With the eight grey level squares displayed within a white background, state which two squares have intensity levels that are most difficult to tell apart.

(3) Now change your background to totally black, that is, the grey value of (0,0,0). Now see if the same two squares that you had problem distinguishing are still the same ones. If not, which two squares do you now find most difficult to distinguish? Can you explain why the background intensity can influence your ability to discriminate the grey value intensity?

1 and 2 so
if change background to black then 6 and 7

(4) If you only had these eight discrete grey values to encode a set of ordinal values from your dataset, what background would you use on your chart to maximise your available distinguishable discrete luminance (intensity) attribute.

discriminate value of #5 / (112, 112, 112)

(5) How would you change the given set of discrete grey values to improve the discriminability of these eight intensity attributes if you would like to use a white background?

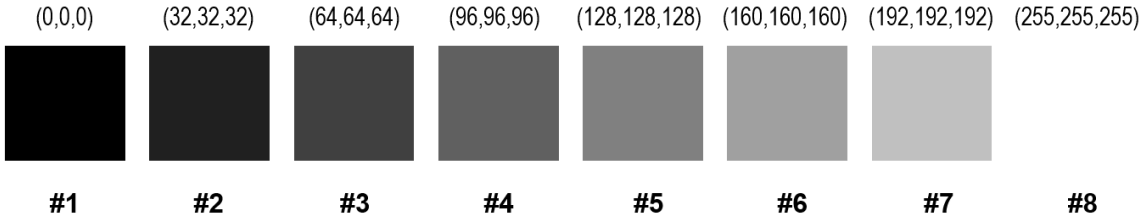


Figure 1 – Eight grey valued squares with equal intensity increments. The white square (#8) cannot be seen as it has the same intensity value as the white background.

4.2 Effective Visual Story

(1) Table 1 shows the average measured levels of a fictitious hormone called *Vitalis* in a population based on age group, gender and BMI. The visual story you want to communicate is that the **only group** with increasing levels of Vitalis as they **age** are **females** with **BMI < 25**.

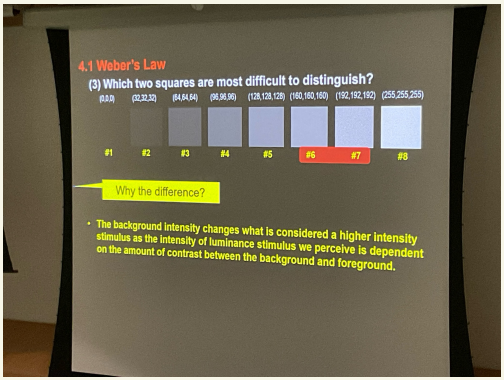
Design an appropriate chart to make this story **stand out visually** while providing all the information shown in the table.

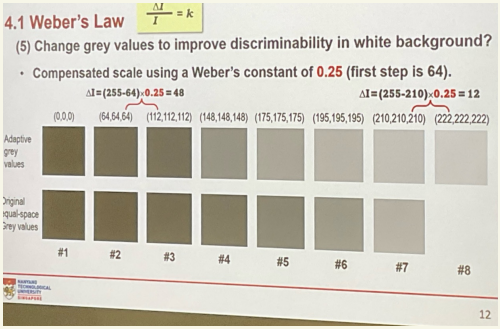
	Levels of the hormone Vitalis			
	Males		Females	
Body Mass Index (BMI)	Under 60 years	60 years or over	Under 60 years	60 years or over
Under 25	255	230	380	550
25 or over	440	325	720	500

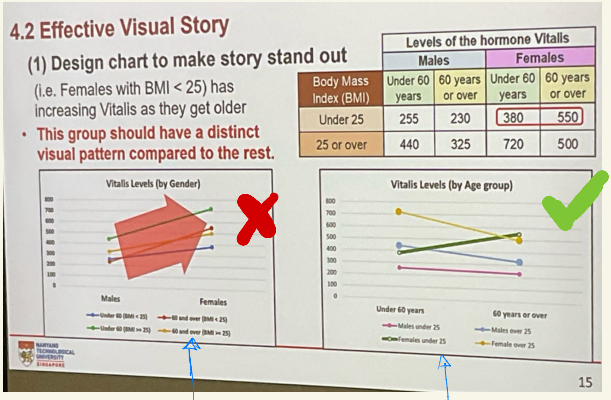
Table 1 – Average Vitalis levels measured in the sampled population under the various categories.

Note: The Excel file "*Hormone Vitalis – Excel.xlsx*" with Table 1 is given to you. You can generate your chart in Excel or use "*Hormone Vitalis – CSV.csv*" and do the plot in Matplotlib.

4)







if male and female

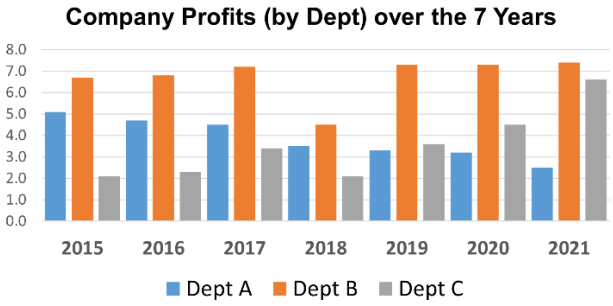
stand out more

4.3 Applying Gestalt Principles

(1) The table in Figure 2(a) shows the annual profit of each department (A, B and C) in an imaginary company. The data table, along with its bar chart shown in Figure 2(b) can be found in the Excel file "*Changing Profits – Excel.xlsx*". An equivalent *CSV* version of the dataset is also provided (if you want to do the visualization in Python).

	Profits in last 7 years (S$ M)		
Year	Dept A	Dept B	Dept C
2015	5.1	6.7	2.1
2016	4.7	6.8	2.3
2017	4.5	7.2	3.4
2018	3.5	4.5	2.1
2019	3.3	7.3	3.6
2020	3.2	7.3	4.5
2021	2.5	7.4	6.6

(a)



(b)

Figure 2 – (a) The data table showing how each department contributes to the company's changing profit each year (2015 to 2021). (b) A bar chart showing each department's annual profit.

(2) **Presentation by Department Head** – As the head of department (Dept C), you are making a **Powerpoint** (PPT) presentation to your Chief Executive Officer (CEO) to request for more resources to maintain the current strong growth in your department. Apply appropriate Gestalt principles to improve the design of the presentation shown in Figure 2(b).

Note: *You can use more than one chart in your PPT (if you wish), add animation, sequence overlay, etc to make build your case and communicate your purpose clearly and effectively.*

In your presentation design, communicate the following:

a) Firstly, you must show the annual profits for all three department over the 7 years but make sure the presentation is primarily focused on your department's performance over these years.

b) Show clearly the changing trend of the company's overall profits from one year to the next.

c) Highlight the really challenging year for the company and your department.

d) Communicate clearly how well your department has been performing over the last 7 years, especially after that very challenging year for the company.

e) Show how your Dept C's percentage contribution to the company's overall profit has been growing over the years.

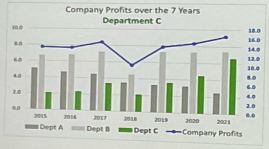
(3) **Presentation by CEO** – As the CEO of the company, you are making a **Powerpoint** (PPT) presentation to Board of Directors to re-distribute the limited resources in the company. You need to make a case that the poorest performing department will need to be restructured to re-allocated resources to the department with the strong growth trajectory. In your visualisation, apply appropriate Gestalt principles to visually communicate the following:

a) Compare the performance of the three departments over the last 7 years with the goal of highlighting the profit trend of the poorest performing department (Dept A) and the department with the strong growth potential (Dept C).

b) The changing percentage contributions of each department to the company's overall profits.

4.3 Applying Gestalt Principles

(2) HoD (Dept C) to request for more resources to maintain the current strong growth



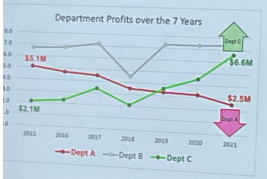
- **Proximity** - is used to group the departments together into the different years so it is easy to see the changing profits from year to year.
- **Similarity** - used in reverse to highlight Dept C's data as it is the only one in colour (Green connoted growth)
- **Connectedness** - An additional line chart with a secondary axis is used to show the company's overall profits.
- **Similarity** - in colour used to link right vertical axis to line plot.

Gestalt Principles used to improve the graph above?

17

4.3 Applying Gestalt Principles

(3a) CEO to Board – Redistribute resources, who is up and down…



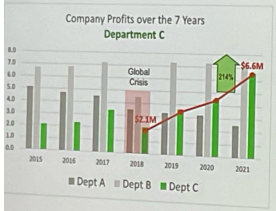
- **Connectedness** - A line chart shows changing trend. A growing green colour is assigned to Dept C and a warning red is assigned to Dept A. Non-relevant Dept B de-emphasised with a grey colour.
- **Similarity & Proximity** - Annotation of relevant figures (using similar colour as the lines) are placed at near lines shows the declining and increasing profits of Dept B and C respectively over the last 7 years

Gestalt Principles used to improve the graph above?

20

4.3 Applying Gestalt Principles

(2c-d) Highlight challenging year and performance after 2018 ….



- **Enclosure** - is used to highlight the challenging year of 2018. Global crisis is annotated over that year to explain why it is highlighted.
- **Connectedness** - A strong line is used to highlight Dept C's increasing profits.
- **Proximity** - Actual values are annotated at 2018 and at 2021 to show the amount of increase over the last four years).

Gestalt Principles used to improve the graph above?

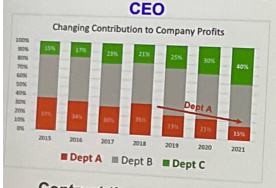
18

4.3 Applying Gestalt Principles

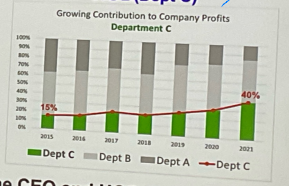
(3b) CEO to Board – Percentage contributions

only HoD department

CEO



HOD(Dept C)

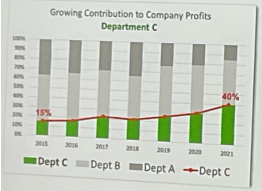


Contrast the two charts from the CEO and HOD (Dept C)

21

4.3 Applying Gestalt Principles

(2e) Highlight Dept C's percentage contribution over the years….



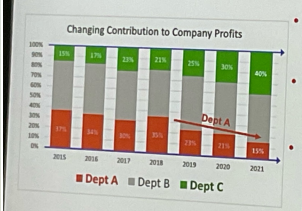
- **Similarity** - Percentage stacked bar chart to show the % contribution. Again, the same consistent green colour used to make Dept C's profit data stand out.
- **Connectedness** - Strong red line is highlight positive profit contribution trajectory of Dept C profits.
- **Proximity** - Actual % contributions are annotated on the bar charts (start and end) for further clarity of progress made so far.

Gestalt Principles used to improve the graph above?

4.3 Applying Gestalt Principles

(3b) CEO to Board – Percentage contributions

must say which gestalt principles to improve



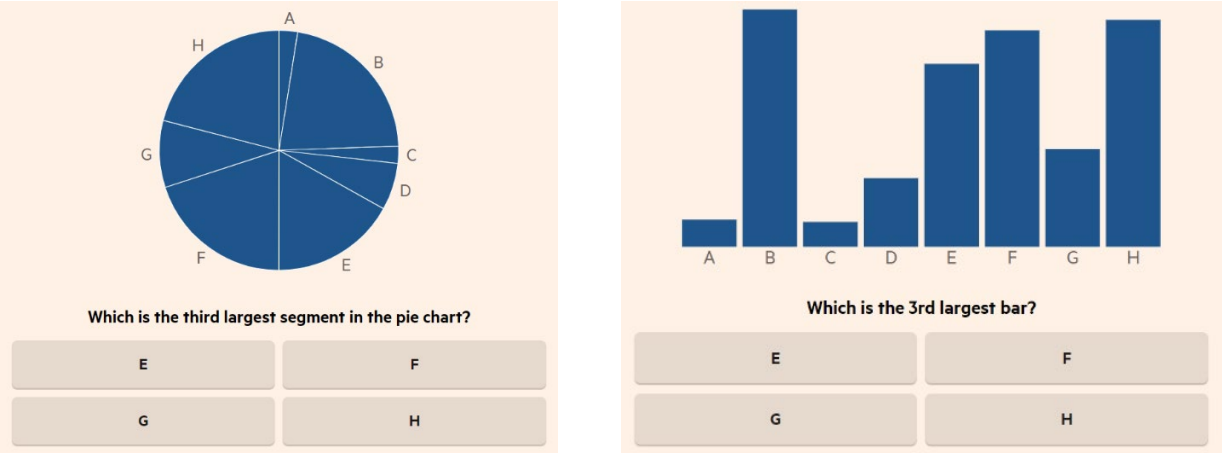
- **A stacked bar chart is used to show the changing percentage contribution of each department.**
- **Proximity** - A downward sloping red arrow near Dept A most recent profit data show recent declining profits.
- **Continuity & Similarity** - Colour similarity associates the data to each dept. Dept A & C are placed at the top & bottom ends as the principle of continuity with reference to horizontal axes will accentuate trends over the years.

What Gestalt Principles was used?

21

4.4 Making it Easy for Others

(1) Financial Times provides an interesting online quiz entitled "*The science behind good charts*", which was found at https://ig.ft.com/science-of-charts/ (**Note**: subscription may be needed for access to FT website and link may be outdated). This online quiz helps you appreciate the feeling when you are at the receiving end of poorly designed data visualisations.



(a) (b)

Figure 3 – (a) Pie Chart and (b) Bar chart taken from (https://ig.ft.com/science-of-charts/)

(2) Figure 3 shows two questions taken from that FT quiz. Redesign the pie chart and the bar chart to make it "really-really" easy for someone taking the quiz to answer the question correctly.

(3) Show what further improvements to the visualisation you would incorporate if your purpose is to get the audience to focus **specifically** and **precisely** on that third largest value in that dataset.

(4) Using the Excel file "*FT Pie and Bar – Excel.xlsx*" or the CSV file "*FT Pie and Bar – CSV.csv*" provided, implement your improved version of the pie and bar charts in parts (2) and (3).

(5) Describe what Gestalt principles were employed in improving your visualization.

(6) Which of the two types of charts would you use if your purpose is to show how close entity **F** is to the next two highest rivals?

(7) Are the exact differences between entity **F** and its top two rival entities **B** and **H** easy to visually estimate? If not, why is this so?

(8) How can you make the differences between entity **F** and its top two rivals more apparent by:

a) using the Gestalt principle of continuity.

b) using the Gestalt principle of enclosure.

(9) Implement these improvements on the chart you have selected.

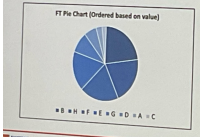
Note: You do not have to develop codes to implement every one of your improvement ideas. You can generate the basic pie or bar chart (using Excel or Python) and use appropriate applications with drawing tools (e.g. in Powerpoint, Word, Paint, etc) to add the additional improvement visuals into your base chart (exported out as an image file). You should spend more time thinking about the design instead.

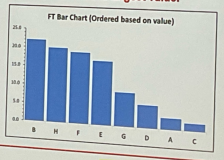
4.4 Making It Easy for Others

(2) Redesign to make it really easy to get it right

What Gestalt principle was employed?

- The Gestalt principle of **simplicity** can be applied by **sorting** the pie segments and bars in a largest to smallest order to help the viewer see which is the 3rd largest value.





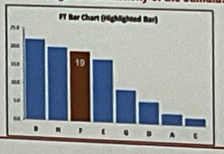
Pie Chart

Bar Chart

4.4 Making It Easy for Others

(7) Are the exact differences between entity F and its top two rival entities B and H easy to visually estimate? Why?

- Not easy to estimate the difference between F and B & H. The length disparity we are trying to visualise is much smaller than the overall length of the bar F. **Weber's Law** states that the increment must be larger if the intensity of the stimulus is large.

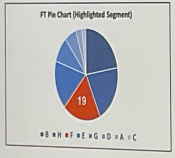


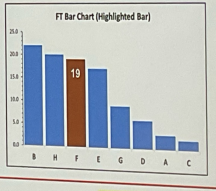
4.4 Making It Easy for Others

(3-4-5) How to get audience to focus on the 3rd largest value?

What Gestalt principle was employed?

- The principle of similarity (dissimilarity) is used when highlighting (with a distinctively different colour) the particular third largest data value.





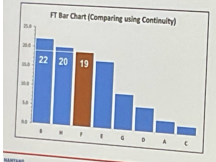
Pie Chart

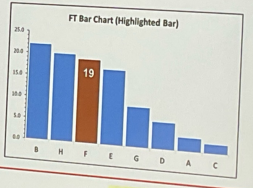
Bar Chart

4.4 Making It Easy for Others

(8) Make differences betw. F and B & H apparent using **continuity**?

- Principle of continuity uses a white line to draw the reference height of entity F into the top two rivals.





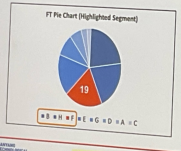
Highlighted Chart

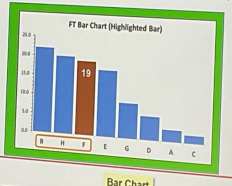
Bar Chart

4.4 Making It Easy for Others

(6) Which chart to show how close entity F is to the next two highest rivals?

- A bar chart should be used if we want to compare data values more precisely.





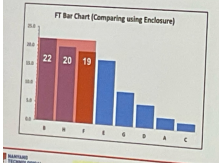
Pie Chart

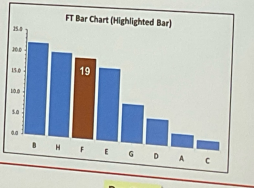
Bar Chart

4.4 Making It Easy for Others

(8) Make differences betw. F and B & H apparent using **enclosure**?

- The principle of enclosure uses the boundary of the enclosure to make the differences of the bar heights more visible.





Highlighted Chart

Bar Chart

4.5 Selecting Colour Palettes

(1) Describe what are categorical (qualitative), sequential and diverging colour palettes; and under what condition would it be appropriate to use each of them. Using the Python Seaborn functions, create an example of each one of these palettes with 9 discrete and distinctive colours.

(2) **Children Growth Rate** - The CSV file "*Children Growth Rate – CSV.csv*" charts the average yearly growth of boys and girls. There are interesting observations about the relative heights of boys and girls as they age from birth (0 years) to when they are 20 years old. Using **two stacked bars**, similar to that shown in Figure 4, and an **appropriate colour palette**, create a visualisation that will facilitate the comparison of growth rates between the gender. What did you observe?

Note: The children average growth rate dataset was modified from data obtained from *Disabled World* website at https://www.disabled-world.com/calculators-charts/height-weight-teens.php.

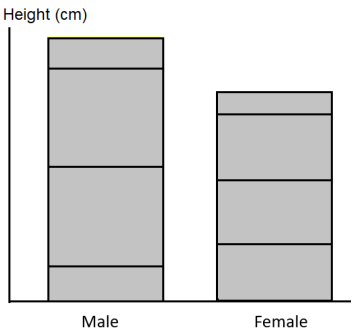


Figure 4 – A Sample Stacked Bar Chart

4.6 Less Colour but More Clarity

(1) Figure 5 shows a data visualisation that presents the results from a survey regarding the work-life balance of various professional groups. List down what you think are poorly designed aspects of the visualisation that compromised the **accessibility** of the information presented.

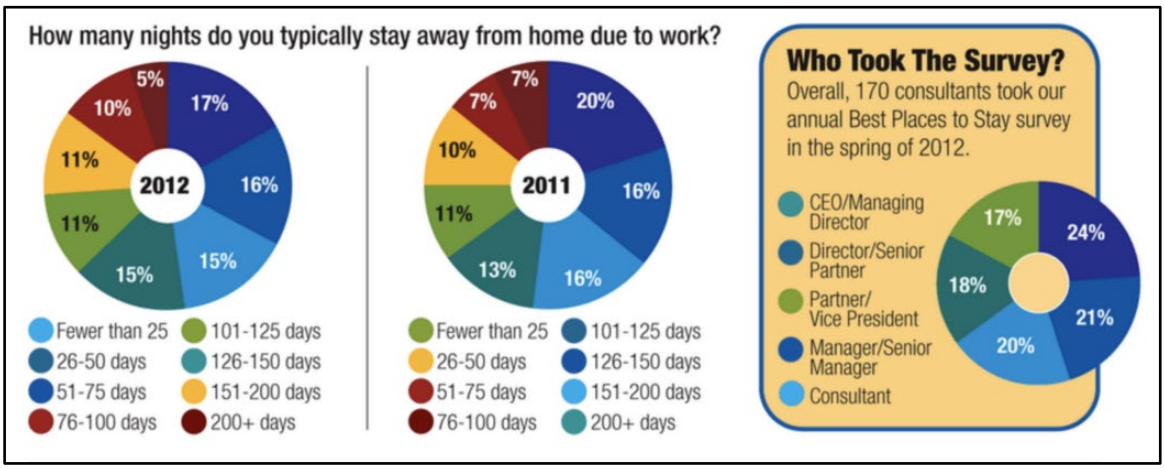


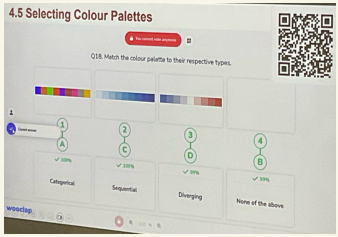
Figure 5 – A poorly designed visual that was featured in https://towardsdatascience.com/color-in-data-visualization-less-how-more-why-348514a3c4d8.

(2) **No constrain redesign** – You are given no constraints in terms of chart type, number of charts, number of colours, use of annotations, etc. Redesign an improved version of the visual shown in Figure 5. Your new visualisation design must incorporate and show all the data and legend information present in Figure 5.

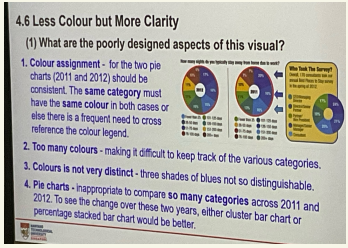
(3) **One-colour only redesign** – Complete the redesign as in part (2) with no constraints except that now you can only use one colour (this one colour excludes the black for your text and borders and the white for your background).

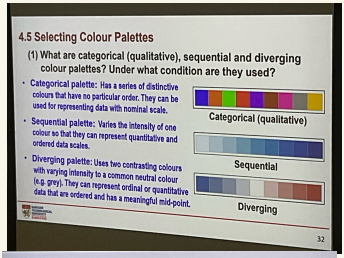
Note: Those interested to create the visual can use the data in "*Away at Work – Excel.xlsx*" file.

45

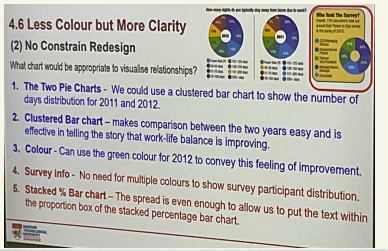


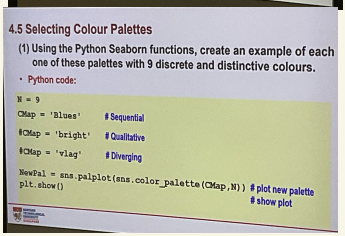
46.

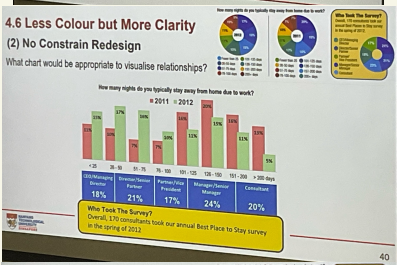


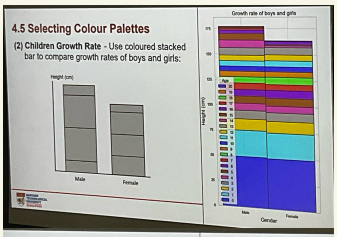


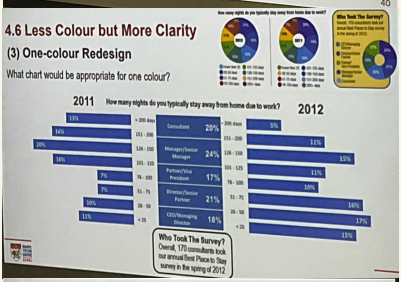
Solution

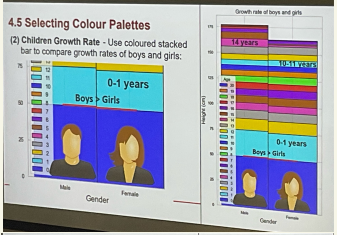


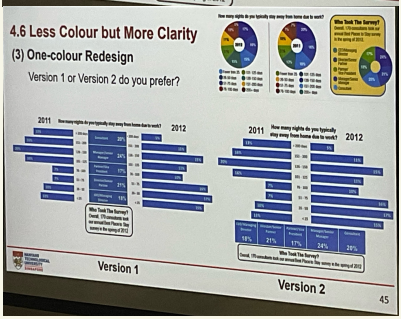


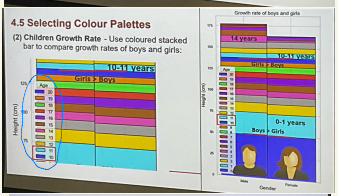


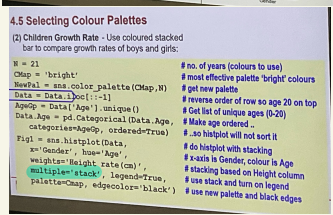










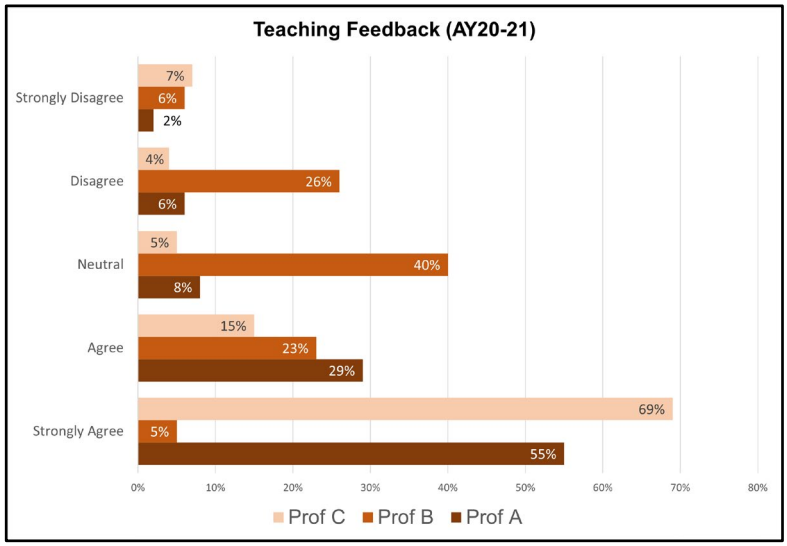


4.7 Critical Eye and Creative Fix

(4) **Teaching Feedback Performance**. Figure 6(b) shows a fictitious clustered bar chart. It was created from the data table in Figure 6(a) of collated questionnaire responses tabulated as a percentage (%) of all responses received for each Professor **A**, **B** and **C** during a recent student teaching feedback survey. The survey was done based on a 5-point Likert scale, with the most favourable response stated as "Strongly Agree" and the least as "Strongly Disagree".

Likert Rating	Prof A	Prof B	Prof C
Strongly Agree	55%	5%	69%
Agree	29%	23%	15%
Neutral	8%	40%	5%
Disagree	6%	26%	4%
Strongly Disagree	2%	6%	7%

(a)



(b)

Figure 6 – (a) Data table of teaching feedback score percentages over a 5-point Likert scale for three professors. (b) A clustered bar chart comparing the teaching performances of the 3 professors.

(5) Is the data visualisation shown in Figure 6(b) effective in comparing and communicating the relative teaching performances of the three Profs **A**, **B** and **C** based on their respective distribution of 5-point Likert-based responses? Your comments should touch on the following:

a) The choice of chart for the purposes stated above.

b) The choice of colours used in the chart.

c) Relevant Gestalt principles that have or have not been exploited to make the visualisation more effective.

(6) Carefully consider the nature of the data and the intended purposes of the visualisation, then design an improved version. Implement your new design using the data from the Excel file "*Teaching Feedback - Excel.xlsx*" or the CSV file "*Teaching Feedback – CSV.csv*" provided.

(7) Describe the rationale for your improved visualisation design, which should cover the following:

a) Reasons for your choice of the particular chart type.

b) Choice of colours or colour palette used in the chart.

c) Relevant Gestalt principles you have employed in your improved design.

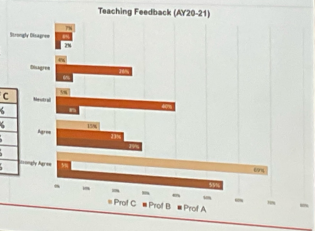
(8) What can you infer when comparing the relative teaching feedback distributions of Profs **A**, **B** and **C**?

4.7 Critical Eye and Creative Fix

(1) Teaching Feedback Performance

- Data tabulated as a percentage (%) of all responses received.
- Survey done based on a 5-point Likert scale.

Likert Rating	Prof A	Prof B	Prof C
Strongly Agree	55%	5%	69%
Agree	29%	23%	15%
Neutral	8%	40%	5%
Disagree	6%	26%	4%
Strongly Disagree	2%	6%	7%



4.7 Critical Eye and Creative Fix

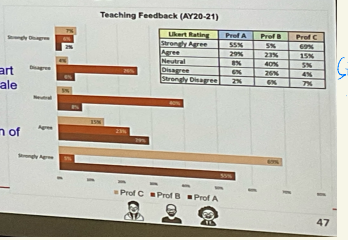
(2) Effectiveness in comparing teaching performance?

a) Choice of chart?
b) Choice of colour?
c) Gestalt principles?

a) Choice of chart - cluster bar chart is **inappropriate** as the Likert scale distribution is a **percentage**.

Stacked percentage bar chart better. It will show the distribution of the 5 categories for each Prof. making the comparison much easier.

Stacked bc is better



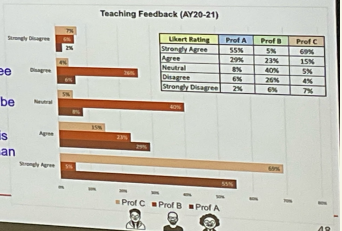
4.7 Critical Eye and Creative Fix

(2) Effectiveness in comparing teaching performance?

a) Choice of chart?
b) Choice of colour?
c) Gestalt principles?

b) Choice of colour - Since the three Profs are nominal categories, a **more distinctive palette** would be better.

The gradation of brown colours is closer to a **sequential** palette than a qualitative palette.



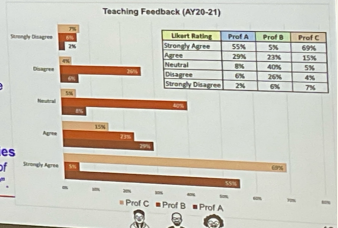
4.7 Critical Eye and Creative Fix

(2) Effectiveness in comparing teaching performance?

a) Choice of chart?
b) Choice of colour?
c) Gestalt principles?

c) Gestalt principles - the use of proximity groups the Likert scale categories. Easier to compare scores in each category.

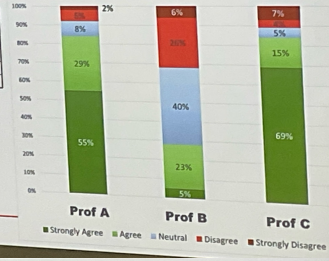
But, proximity also makes it harder to view **across categories** to ask questions like "*how much of the scores are positive, i.e. SA and A?*".



4.7 Critical Eye and Creative Fix

(3) Possible New Design

Likert Rating	Prof A	Prof B	Prof C
Strongly Agree	55%	5%	69%
Agree	29%	23%	15%
Neutral	8%	40%	5%
Disagree	6%	26%	4%
Strongly Disagree	2%	6%	7%



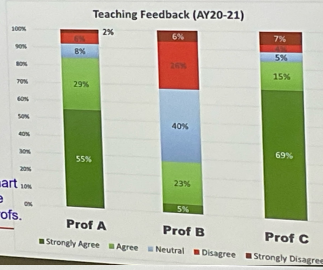
4.7 Critical Eye and Creative Fix

(4) Rationale for Design

a) Choice of chart?
b) Choice of colour?
c) Gestalt principles?

a) Choice of chart - Percentage stacked bar chart is more appropriate as the Likert scale rating distribution is given as a **percentage**.

Uniform height enforced by chart makes it is easy to compare the rating distribution across the Profs.



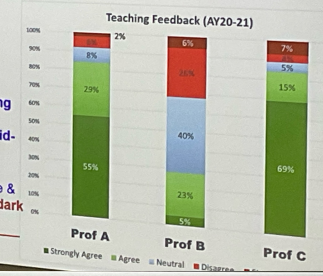
4.7 Critical Eye and Creative Fix

(4) Rationale for Design

a) Choice of chart?
b) Choice of colour?
c) Gestalt principles?

b) Choice of colour - Use **diverging** palette as Likert scores have negative, positive and neutral **mid-point**.

Colour match the negative and positive emotions. **Green** for +ve & **reddish** for the -ve, with a very **dark shade** of red for **SD**



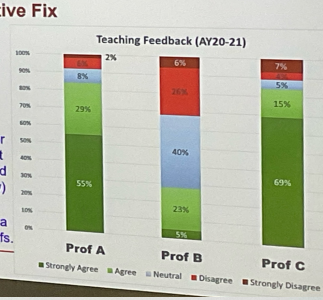
4.7 Critical Eye and Creative Fix

(4) Rationale for Design

a) Choice of chart?
b) Choice of colour?
c) Gestalt principles?

c) Gestalt principles - principle of **similarity** is used when colour for each Likert category is consistent across the three stacked bars and they are made distinct (dissimilarity) from other categories.

This way, it is easier to compare a given rating across the three Profs.



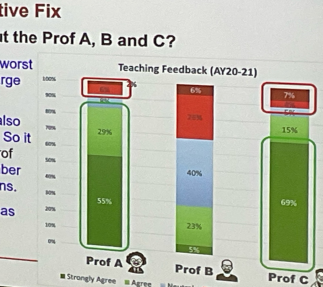
4.7 Critical Eye and Creative Fix

(5) What can you say about the Prof A, B and C?

Prof B - is clearly the one with the worst teaching performance due to the large % of **SD** and **D** ratings.

Prof C - has largest % of **SA**. But also more % of **SD** compared to Prof A. So it seems many students really like Prof C's teaching but a reasonable number shared very strong negative opinions.

Prof A - has same 84% of **SA** & **A** as Prof C but fewer students strongly dislike his or her style of teaching.



4.8 Psychological Principles for Effective Graphics

(1) Melting Antarctica Sea Ice

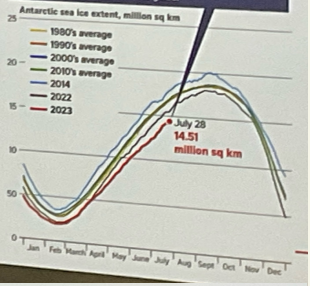
Uncharted territory

Every winter, Antarctica is locked in by a vast ring of sea ice. For decades, the area of winter sea ice had been remarkably steady, but in recent years, the extent of that ice has started to shrink alarmingly and scientists fear climate change is to blame.

NOTE: *US National Snow and Ice Data Centre's (NSIDC) scientists use the 1981 to 2010 average of sea ice concentration and extent to have a consistent basis for comparison to today's fluctuating conditions. The 30 years observed in this baseline provide enough data to even out short-term variability.

Sources: NSIDC, ARCTIC DATA ARCHIVE SYSTEM
STRAITS TIMES GRAPHICS

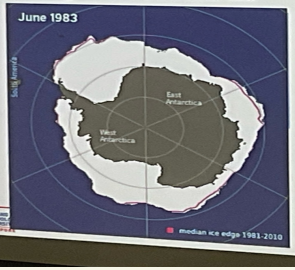
The sea ice extent for this time of year is by far the lowest since satellite measurements started more than 40 years ago and has become yet another indicator of the record-breaking extremes happening around the globe.



Critique example

4.8 Psychological Principles for Effective Graphics

(1) Melting Antarctica Sea Ice



The sea ice extent for this time of year is by far the lowest since satellite measurements started more than 40 years ago and has become yet another indicator of the record-breaking extremes happening around the globe.

Antarctic sea ice extent, million sq km
1980's average
1990's average
2000's average
2010's average
2014
2022
2023
July 28
14.51
million sq km
Jan
Feb
March
April
May
June
July
Aug
Sept
Oct
Nov
Dec

4.8 Psychological Principles for Effective Graphics

(1) Melting Antarctica Sea Ice

a) What is the main message the line chart is attempting to communicate?

Main message - the sea ice extent for this time of the year is the **lowest since the last 40 years** of satellite measurements.

This is to emphasize the **record-breaking extremes** in climate change that is happening around the world.

The sea ice extent for this time of year is by far the lowest since satellite measurements started more than 40 years ago and has become yet another indicator of the record-breaking extremes happening around the globe.

Antarctic sea ice extent, million sq km
1980's average
1990's average
2000's average
2010's average
2014
2022
2023
July 28
14.51
million sq km
Jan
Feb
March
April
May
June
July
Aug
Sept
Oct
Nov
Dec

.8 Psychological Principles for Effective Graphics

(1) Melting Antarctica Sea Ice

b) Identify psychological principles and describe how it was used to make visual more effective.

Goldilocks Principle & Principle of Capacity Limitation

1980's to 2010's data were **averaged** and plotted as only **four line plots** to **avoid clutter.**

This makes line plots of recent changes in the melting sea ice extent more clearly observable.

Antarctic sea ice extent, million sq km
1980's average
1990's average
2000's average
2010's average
2014
2022
2023
July 28
14.51
million sq km
Jan
Feb
March
April
May
June
July
Aug
Sept
Oct
Nov
Dec

More historical measures plotted as smooth averaged lines.

4.8 Psychological Principles for Effective Graphics

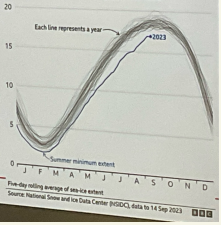
(1) Melting Antarctica Sea Ice

Data Visualisation by BBC News

What principle was employed?

Antarctica sea-ice far lower than usual

Daily sea-ice extent in million sq km, 1979-2023



4.8 Psychological Principles for Effective Graphics

(1) Melting Antarctica Sea Ice

Data Visualisation by BBC News

What principle was employed?

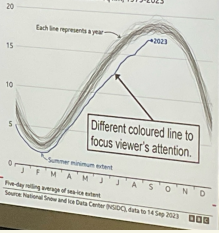
Principle of Salience

The **2023** line plot (most recent year) was drawn in **blue** to focus user's attention to the latest sea ice extent measurements.

All other lines in inconspicuous light grey

Antarctica sea-ice far lower than usual

Daily sea-ice extent in million sq km, 1979-2023



Different coloured line to focus viewer's attention.

4.8 Psychological Principles for Effective Graphics

(1) **Melting Antarctica Sea Ice**. Study the line chart in Figure 7, which shows the data of the Antarctica sea ice extents from 1980 to 2023.

a) What do you think is the main message the line chart is attempting to communicate to the reader?

b) Identify which of Stephen Kosslyn's eight psychological principles of effective graphics were employed by The Straits Times to create the infographics shown. You may list more than one. (see Lecture notes chapter 7). For each of the principles that you have identified, describe how it was used to convey the intended message more effectively.

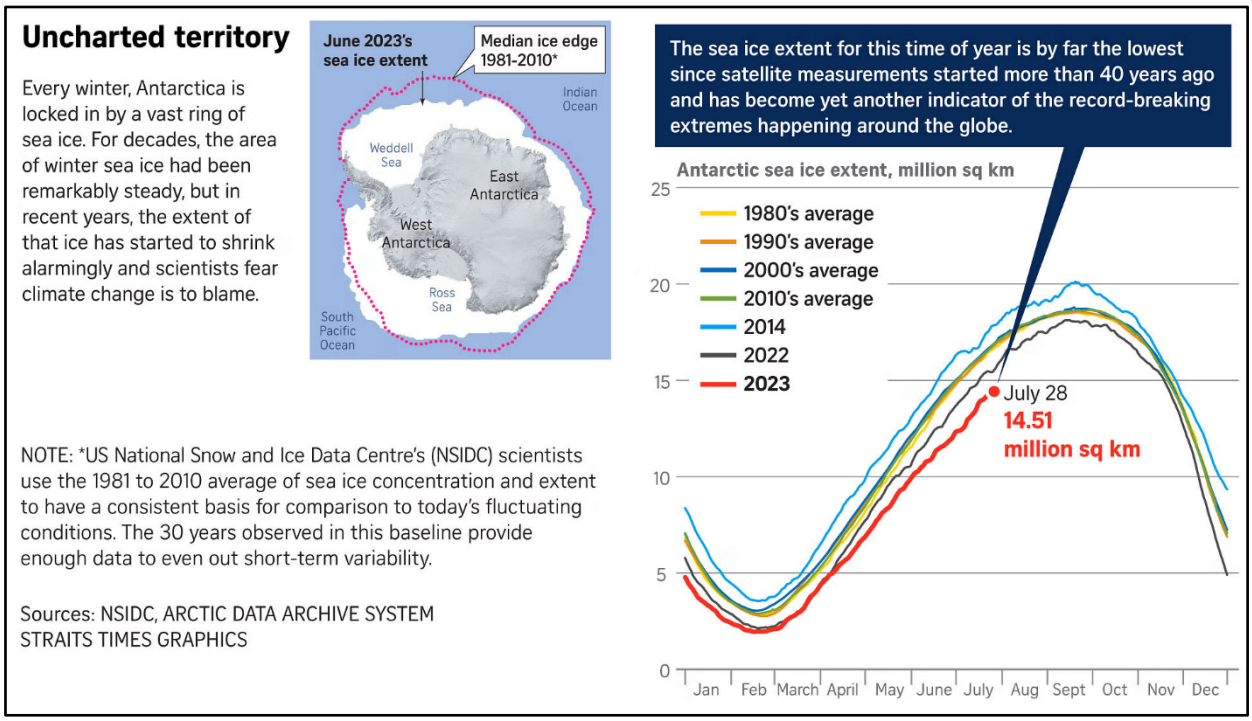


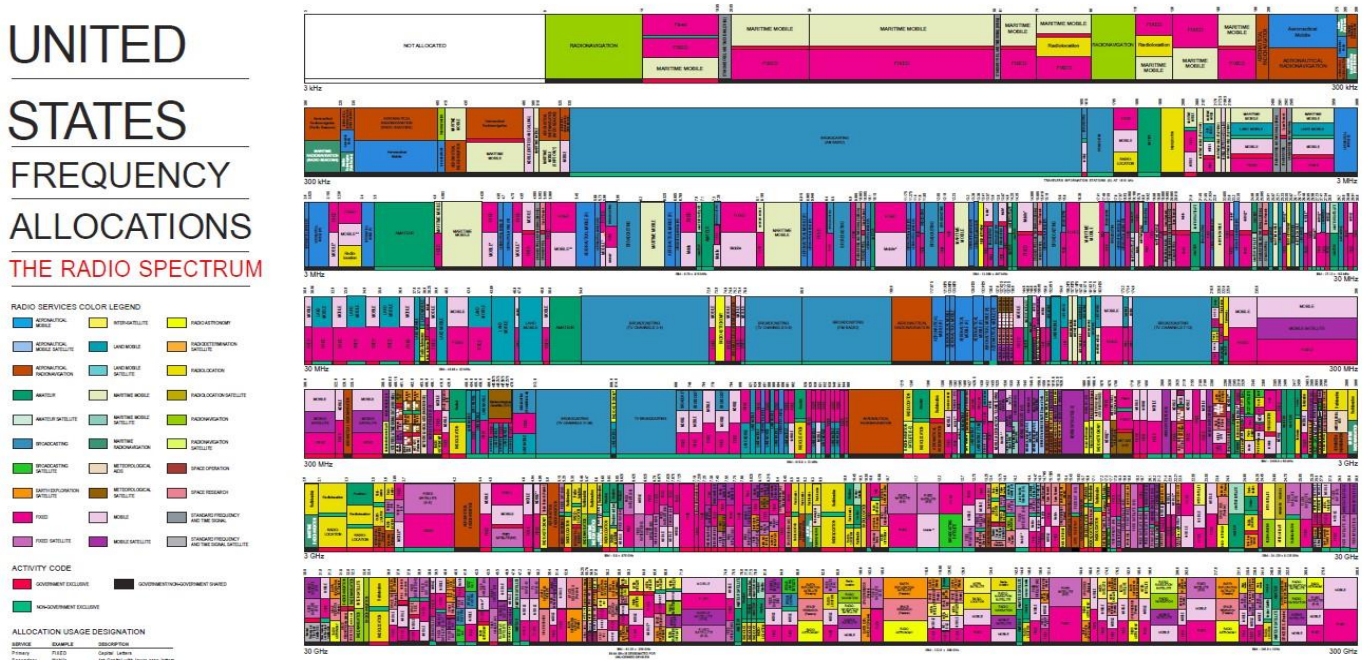
Figure 7 – The infographics illustrating the sea ice extent variations around the Antarctica from 1980 to 2023. This data visualisation was taken from an article in The Straits Times entitled "Antarctica's record low sea ice worries climate scientist", which was published on 30 July 2023. Link to article: https://www.straitstimes.com/world/antarctica-s-record-low-sea-ice-worries-climate-scientists.

part 2

SC 4024 Tutorial 5: Visualization Principles and Interactions

1. Color Encoding for Radio Spectrum Allocation

The following figure shows the visualization for radio spectrum allocation in the United States, which is collected from here: https://www.ntia.gov/files/ntia/publications/2003-allochrt.pdf

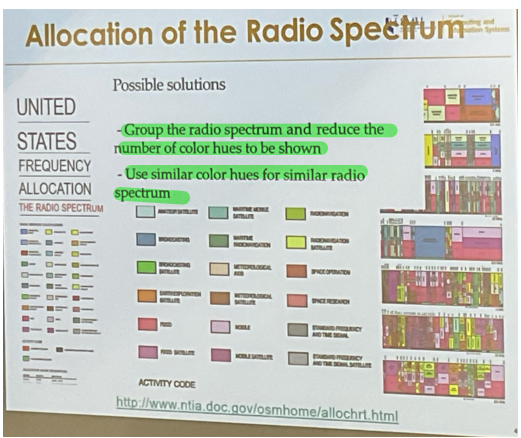


1.1 By referring to the visualization principles and best practices we have learned in this course, please **point out the major issues of the color encoding design** of the above visualization.

1.2 Suppose you are the visualization designer, do you have **any suggestions to address the above issues in terms of color encoding design**?

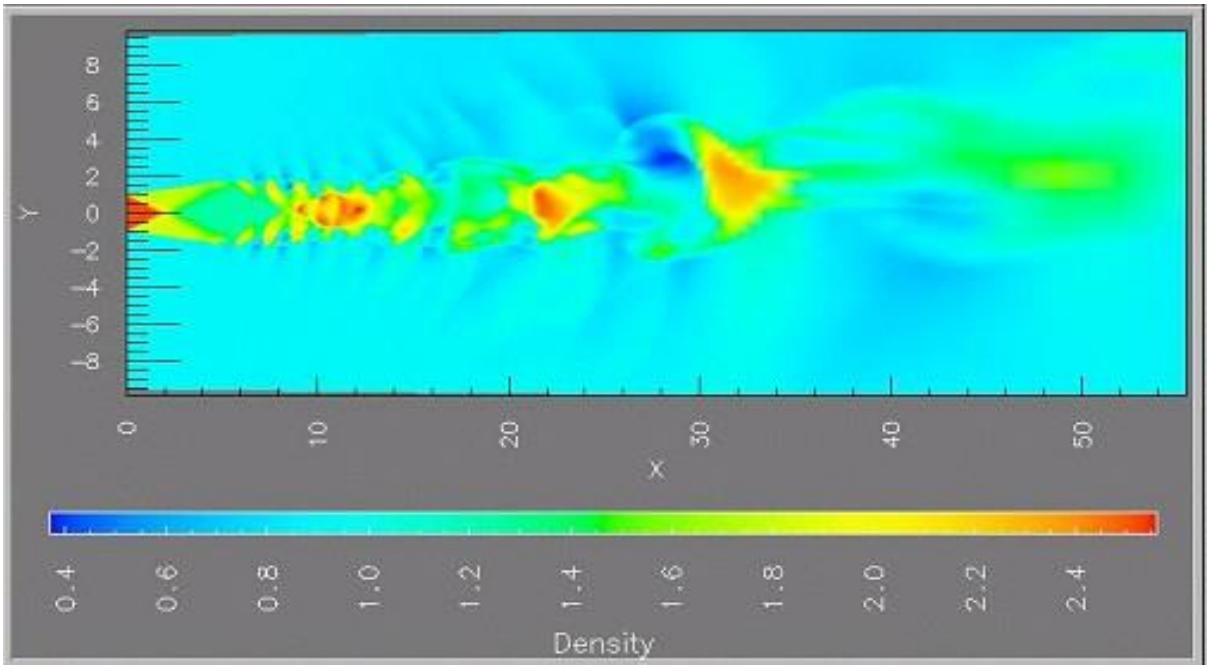
1.1) - too many color, can't remember the mapping
- some color will be too similar (perception issue)

1.2)



2. Rainbow Colormaps

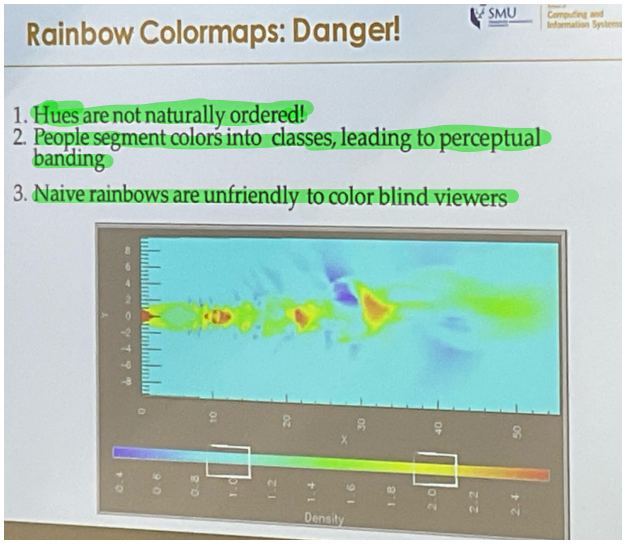
A rainbow colormap is based on the order of colors in the spectrum of visible light—the same colors that appear in a rainbow. Below is an example of rainbow colormap, which is used to display the salt density of the sea in a specific area.



2.1 Does the above rainbow colormap look good to you? Why or why not? You are required to provide the detailed justifications.

2.2 If there are issues in the colormap of the above visualization, how should we improve it?

2.1



2.2 - use seq or diverging colormap

SC 4024 Tutorial 6: Exploratory Data Analysis

AY2024-2025 Semester 1

1. Fundamental Concepts of Correlation Analysis

For the following Pearson correlation coefficients (r):

0.32, -0.76, 0.13, -0.25, 0.04, -0.03, 0.01

a) Which is the strongest correlation? -0.76

b) Which is the weakest correlation? 0.01

2. Exploratory Data Analysis

During the lecture of Chapter 9.1, we take the course scores of students as an example to illustrate how we can conduct exploratory data analysis via data visualization and statistical analysis. Now you are provided with a dataset file called "students_performance.csv", and you are asked to conduct exploratory data analysis on it. Suppose you are interested in **checking if there is a difference between the math scores and reading scores of students from group A**, please answer the following questions:

as long as 1 of attr not follow normal distribution

Should not use Z test

2.1 By referring to what we have learned about exploratory data analysis and correlation analysis, briefly describe **what kind of steps** you will take to answer such a question, and what kind of **visualizations** and **statistical tests** you will use.

- Check dataset size
- Parametric / non-parametric test
- use Z-test

2.2 What are your **null hypothesis (H_0)** and **alternate hypothesis (H_1)** for such an analysis?

null → there is no diff in math n write
alte → ther is diff in math n write

P > 0.05 → normal

2.3 By following the steps in 2.1, you are asked to write Python code to create appropriate visualizations and conduct the corresponding relevant statistic tests. Note: You are requested to use the Python packages like **scipy.stats** and **statsmodels.stats**.

P < 0.05 → reject null hypothesis

P ≥ 0.05 fail to reject null hypothesis

2.4 Suppose you are asked to check there is a difference between the math scores and reading scores of **students from group B or group C**, what kind of changes do you need to do in your procedures of exploratory data analysis in 2.3?

null : there is no diff between my score n reading score 0.17 > 0.05

alter : there is a diff in math n writting score

SC 4024 Tutorial 7: Geospatial Data Visualization

AY2024-2025 Semester 1

1. Selecting Appropriate Geospatial Data Visualization

We have learned a series of geospatial data visualization approaches for different types of geospatial data. Suppose you are requested to visualize the following data, what kind of geospatial data visualization approaches will you use? Justify your design choice.

a) The annual sales of each McDonald's at different locations of Singapore in 2023. Point based → Symbol map × Choropleth

b) The average temperature of different countries in the EU. Choropleth

c) The aging population (age>65) of different states in the U.S. Cartomap

2. Symbol Map Implementation

You are provided with a CSV file called "*sg_universities.csv*" containing the detailed location and student size information of five universities in Singapore. You are requested to use ***d3.js*** (more specifically, ***d3.geo***) package to create a symbol map to visualize the student size of all the provided universities on a map. Below are detailed requirements:

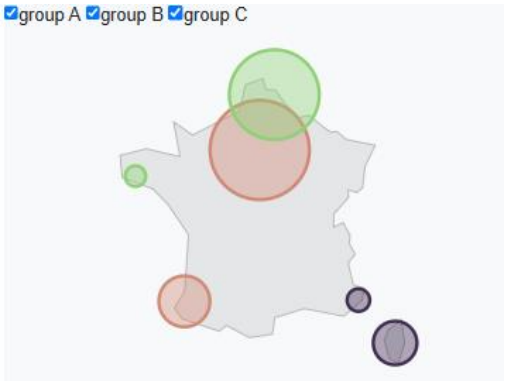
a) ***Visual encoding:*** use a circle to represent each university with its radius encoding the student size (i.e., use bubble map to visualize the geospatial data), and use different fill color for different universities;
b) ***Interaction:*** provide five checkbox to allow users to interactively select whether a university should be shown on the map or not;
c) Animation: when a university is selected or un-selected, use animation to show or remove the glyph (i.e., circle);
d) You are requested to use **D3 version 7** to implement the symbol map.

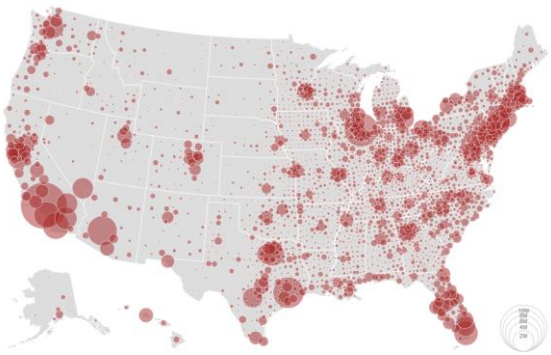
You are required to finish it before the tutorial on **Nov. 5th**.

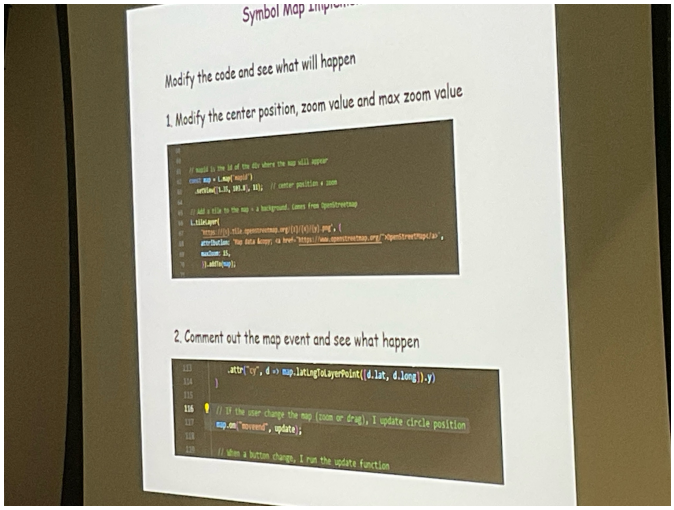
References:

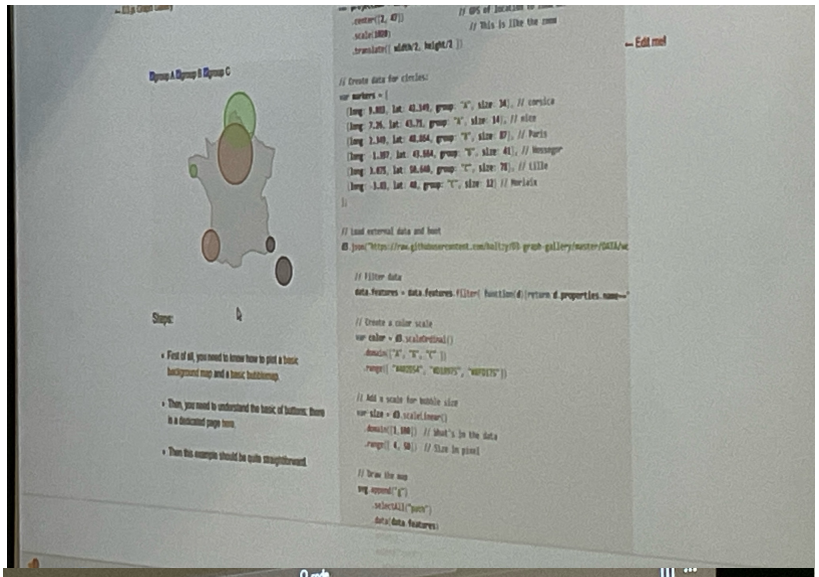
[1] Bubblemap with filter button: https://d3-graph-gallery.com/graph/bubblemap_buttonControl.html

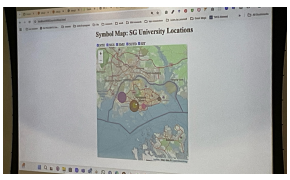
[2] Bubblemap: https://observablehq.com/@d3/bubble-map/2

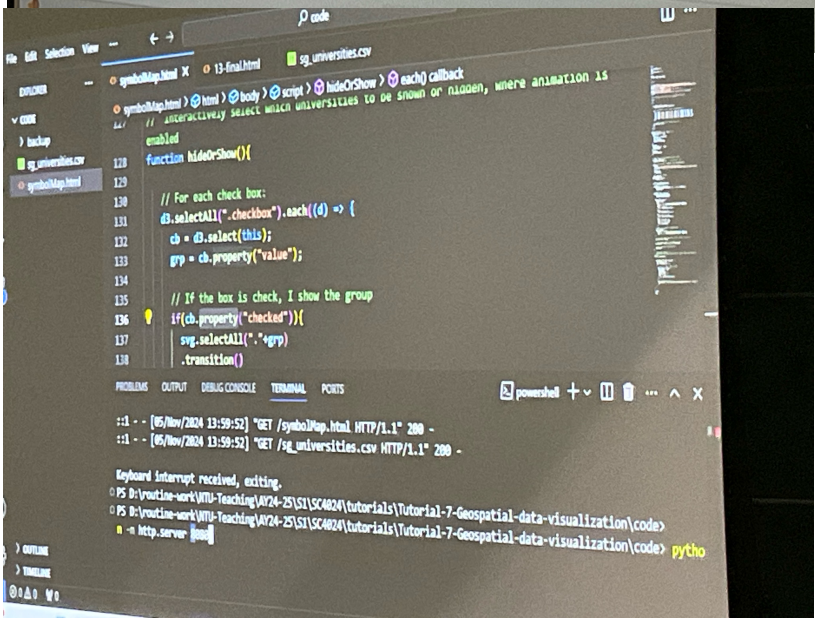












Tutorial 8

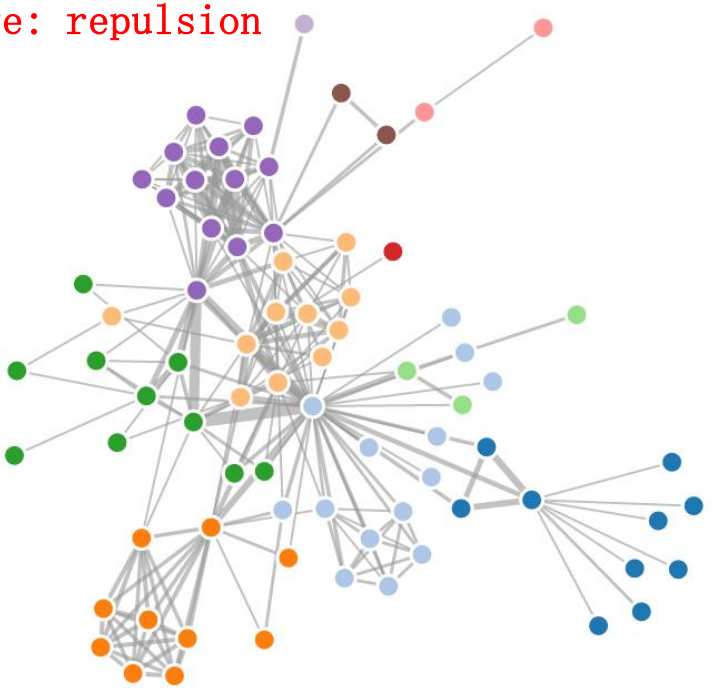
Graph Visualization

WANG Yong
College of Computing and Data Science
NTU

Force-directed Layout

- There can be multiple forces between each pair of nodes.
- The layout is determined by all the forces.

```
var chargeForce = d3.forceManyBody()
    .strength(0.1) //positive: attraction, negative: repulsion
    .distanceMax(500)
    .distanceMin(60);

// Creates a new circle collision force with
// the specified radius
var collideForce = d3.forceCollide()
     .strength(2)
     .radius(20);

var simulation = d3.forceSimulation()
     .force("link", d3.forceLink())
     .force("charge", chargeForce)
     .force('collide', collideForce)
     .force("center", d3.forceCenter(width / 2, height / 2));
```



Note: we are using d3@v7; different versions of d3 can have slightly-different implementations !

Useful References

Official API:
https://github.com/d3/d3/blob/main/API.md#forces-d3-force

https://github.com/d3/d3/blob/main/API.md#forces-d3-force

https://github.com/d3/d3-force/blob/v3.0.0/README.md#forceManyBody

Examples:
https://www.d3indepth.com/force-layout/

https://observablehq.com/@d3/force-directed-graph

Your Task

1. Download the skeleton code

2. General task: finish the code to draw a force-directed graph

```
57        .attr("stroke-width", function(d) {
58          // TO-DO: change the stroke-width as the squre root of the "value" of the current link
59          return 4;
60        });
```

```
70        .attr("fill", function(d) {
71          // TO-DO: revise the code to use different color to represent different groups
72          return 'grey';
73        })
```

Your Task

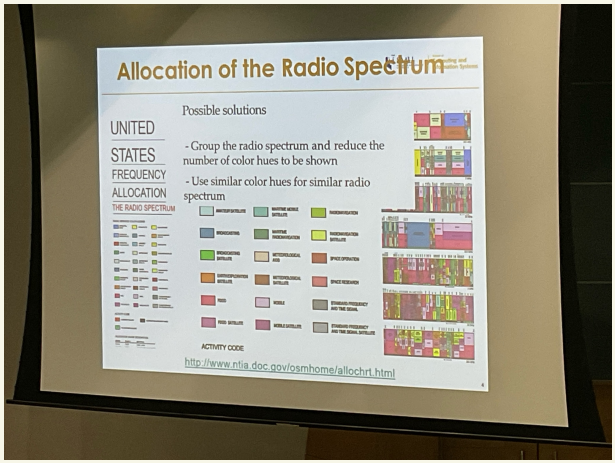
1. Download the skeleton code

2. General task: finish the code to draw a force-directed graph

3. Play with different parameter settings. Check the function of each parameter.

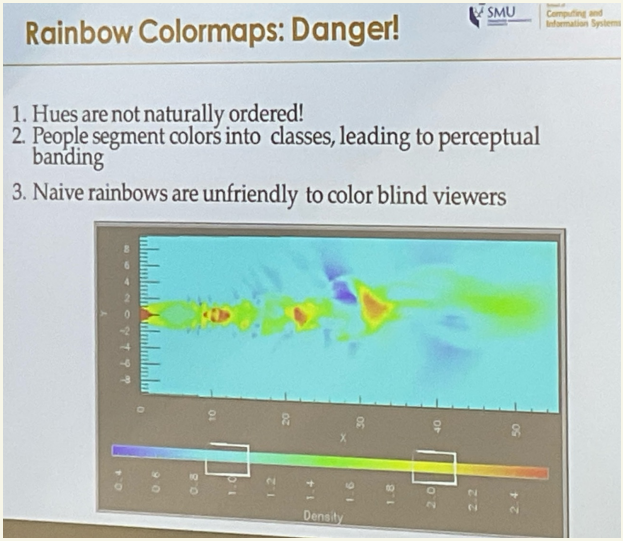
```
31  // TO-DO:
32  //    1) Try to modify the values of strength, distanceMax and distanceMin, radius and see what will happen
33  //    2) Try to use only "var chargeForce =  d3.forceManyBody()" and see what will happen
34  var chargeForce =  d3.forceManyBody().strength(0.1).distanceMax(500).distanceMin(60);
35  var collideForce = d3.forceCollide().strength(2).radius(20);
36
```

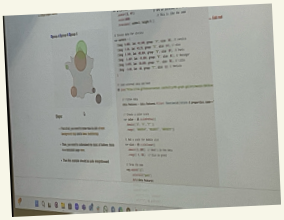
THANK YOU~

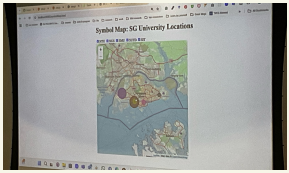


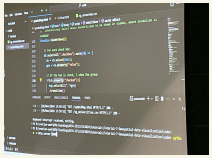
tut 7
1 a Symbol map
b Choropleth map
c Cartogram
tut 6 03 part 2

TUT 7 Q2









Strong Correlation is 0.76
Weakest correlation is 0.01
↳ see absolute