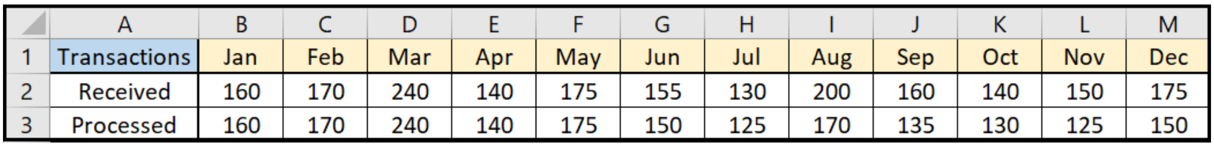
3.1 Line and Bar Charts

(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.

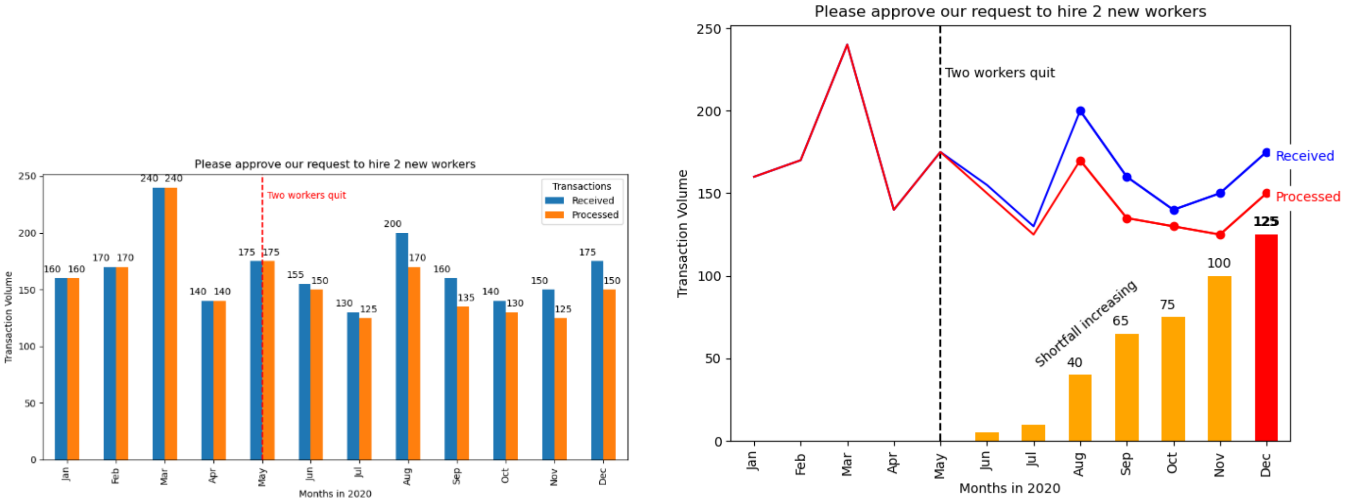
You are to annotate the pertinent information shown, which include the following:

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?

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

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).

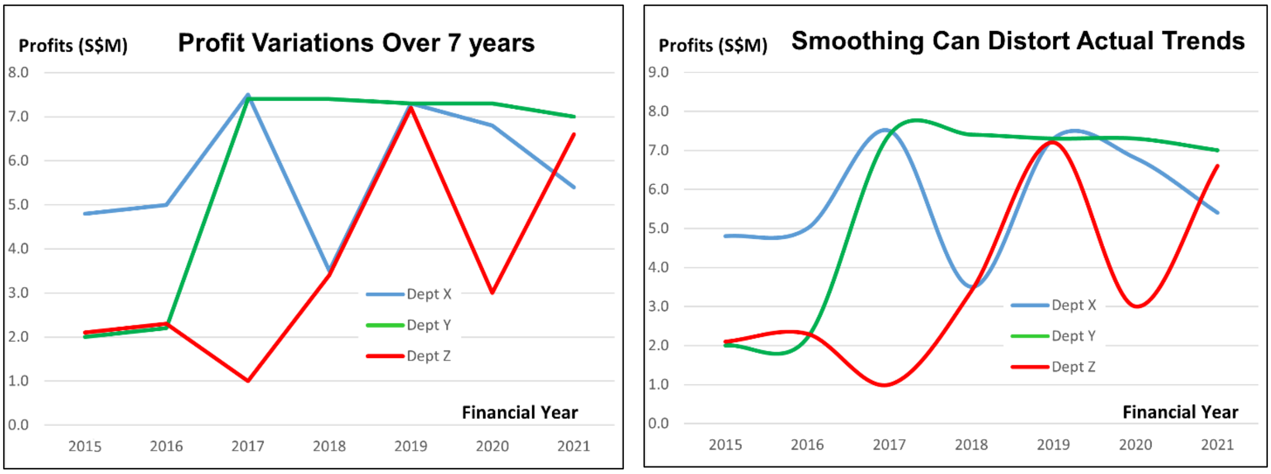
d) Year 2017 to 2019 were years of change. Did any department managed to maintain their profits?

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

	Profits in last 7 years (S$ M)		
Year	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.

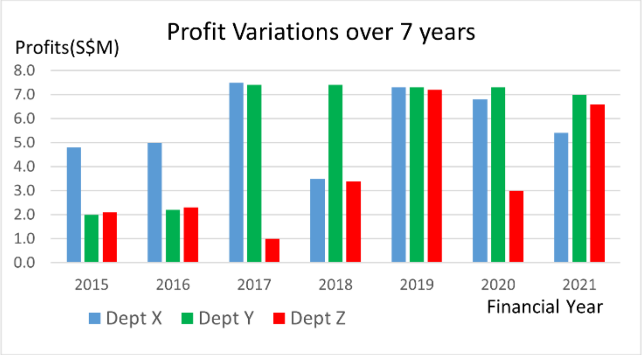


Figure 2(d)

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?

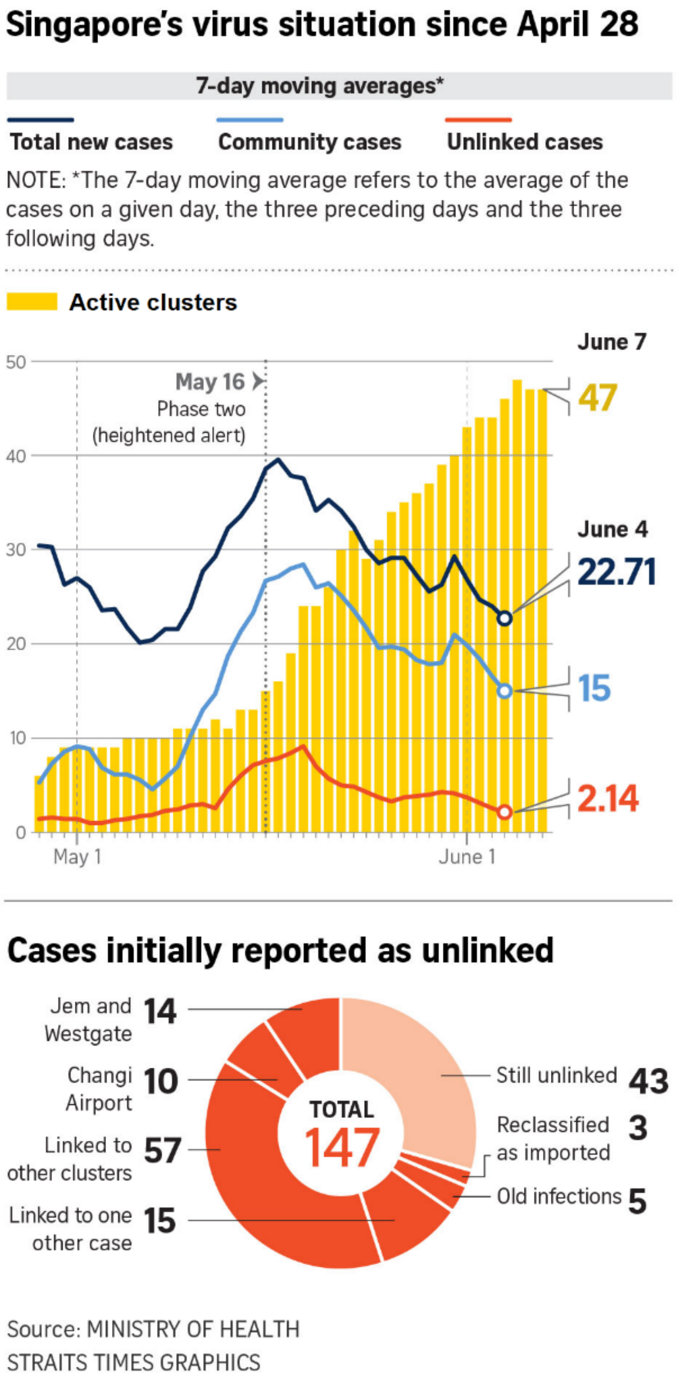
d) Why was a marker line drawn over the line charts at May 16? What can you say about the perceptual channel used on this marker line?

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?



NOTE: *The 7-day moving average refers to the average of the cases on a given day, the three preceding days and the three following days.

Source: MINISTRY OF HEALTH
STRAITS TIMES GRAPHICS

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)

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.