Week 4

LATEST SUBMISSION GRADE

90%

1. Imagine you have analysed some data and need to make a presentation to the board of directors. You realise that it would be unreasonable to expect all the board members to have a background in data analysis, so you have decided it would be best to make some visualisations to aid with your presentation. Which of the following ideas would best help with your visualisations? **0 / 1 point**

- ○ Include as much data as possible in each graph for the sake of brevity and to allow the audience to make quick comparisons.
- ◉ Use only bar, line, and pie graphs, since they are simple and commonly used.
- ○ Include all related details in each chart so that the audience can have a complete understanding of the underlying data as well as the method(s) used in the analysis.
- ○ The options above are all bad ideas.

! **Incorrect**

Incorrect. While bar, line and pie graphs are commonly used, we should always choose the type of graph based on the data we have and the point we are trying to make.

2. In practice task 1, we looked at a dataset of student marks and performed some simple data cleaning. In particular, we estimated some of the missing test scores for “Student 7” by using the available test scores and filtering for other students with similar marks. Which of the following is a valid reason for doing this, as opposed to (say) just averaging all available scores? **1 / 1 point**

- ○ We expect that the Overall Score for Student 7 would be a good indicator of their OLSAT score.
- ○ The marks of first graders (including student 7) could very likely have a different distribution to (say) kindergarteners.
- ○ District 2 (for Student 7) would have residents from a particular socio-economic background, which would indirectly affect test scores.
- ◉ All of the above.

✓ **Correct**

Correct.

3. Suppose you are working with a data generated by machines on a production line. The data also contains some manual inputs from engineers at the factory. As we have seen from the second video in this week's material and the associated reading, we need to clean this data to prepare it for analysis. Which of the following issues would we LEAST likely encounter when cleaning this data? **1 / 1 point**

- (●) Inconsistent formatting in the automatically generated data by the machinery.
- ( ) Outliers (errors) in the automatically generated data by the machinery.
- ( ) Missing values in the automatically generated data by the machinery.
- ( ) Inconsistent formatting in the manual input by the engineers.

✓ **Correct**
Correct. The machines would generally be set up to generate data in a consistent format.

4. In Practice Task 1 this week, we mentioned that sometimes data cleaning involves deleting a certain entry entirely. In general, which of the following would NOT be a valid reason to delete an entry? **1 / 1 point**

- ( ) The entry is not informative, i.e. it does not convey any useful information.
- ( ) The entry contains some data, but there are too many values missing.
- (●) When we plot the data, this entry does not conform to the trend that appears in the rest of the data.
- ( ) We are reasonably convinced that this entry is incorrect.

✓ **Correct**
Correct. There may be a valid reason why one particular entry does not conform to the rest of the data, and this may actually be worth further investigation.

5. In Practice Task 2 this week, we measured interest in global warming in different countries by looking at search data from Google Trends. Conveniently, the data was already relatively **1 / 1 point**

"clean". However, there may still be issues with our data as far as gauging a country's interest in global warming. Which of the following would be examples of such issues?

- ( ) Developing countries may rely on more traditional sources of news information, such as newspapers and television. The relatively few people that rely on the internet for their information may not be representative of the population.
- ( ) The data was scored from 0 to 100 based on a proportion of the maximum number of searches. There could for example be a country with very few people interested in global warming, but this small group may consistently search for news relating to global warming, leading to a relatively high score.

- Countries such as China restrict internet access for political reasons (in this instance, Google cannot be easily accessed in China). As such, the data for such countries would most likely be inaccurate.
- All of the above.

Correct
Correct.

6. Suppose you are analysing data on fish populations, and you are trying to predict fish populations in the near future. You have population measurements for trout, salmon, and tuna taken on various dates over the past 5 years. What sort of graph would be most suited to plot this data? **1 / 1 point**

- A heat map, with the same colour scale across the different types of fish.
- A line chart, with each type of fish plotted using a different colour.
- A bar chart with one bar for each type of fish.
- A pie chart, with each type of fish plotted using a different colour.

Correct
Correct.

Whilst not impossible, it would be difficult for a bar graph to show trends over time. Whilst not impossible, it would be difficult for a bar graph to show trends over time.

7. You are processing the results of a survey on consumer preferences. One of the questions asked people for their favourite flavour of ice cream. Which of the following visualisations would be best suited to plotting this data, and why? **1 / 1 point**

- A bar chart, since it easily compares the proportions of each ice cream flavour preferred.
- A heat map, since the use of colours give an intuitive way to interpret the number of people listing each flavour as their favourite.
- A pie chart, since it easily compares the proportions of each ice cream flavour preferred.
- A line chart, since it would directly show the trends in flavour preferences.

Correct
Correct. A pie chart is particularly useful for comparing proportions.

8. Imagine you have performed a detailed analysis on house prices in a particular city spanning the past decade. You are preparing a presentation for a group of real estate investors, and as part of your presentation, you have prepared some data on typical price growth in each suburb. What would be the most appropriate graph to visualise this data? **1 / 1 point**

- (●) A Heat map.
- ( ) A bar chart.
- ( ) A scatter plot.
- ( ) A pie chart.

✓ **Correct**
Correct. A heat map would be a very intuitive way to view this data. In particular, we could superimpose our data on a map of the city.

9. Which of the following data sets would be best suited to being plotted on a bar chart? **1 / 1 point**

- ( ) The spending in a given financial year for a particular company, by department.
- ( ) The number of visitors at each national park in Brazil, over the past year.
- (●) The number of customers at a restaurant each day, over the past 2 months.
- ( ) Voting patterns in Kenya, by electoral district.

✓ **Correct**
Incorrect. As this data is collected over time (the past 2 months), it would be far more appropriate to plot it with a line chart.

10. Which of the following data sets would be best suited to being plotted on a pie chart? **1 / 1 point**

- ( ) The number of visitors at each national park in Brazil, over the past year.
- (●) The spending in a given financial year for a particular company, by department.
- ( ) Voting patterns in Kenya, by electoral district.
- ( ) The number of customers at a restaurant each day, over the past 2 months.

✓ **Correct**
Correct. If we are analysing this data, we will very liekly be interested in the relative budget allocations of each department.