TASK

Data Visualisation - Approach and Techniques

Visit our website

Introduction

WELCOME TO THE DATA VISUALISATION - APPROACH AND TECHNIQUES TASK!

As science and technology advance, the amount of information and data we possess inevitably increases significantly. With such large amounts of data, how do we effectively find patterns and new information within our data and convey our findings? Data visualisation helps us with this.

Data visualisation is also commonly referred to as information visualisation or infoViz. It is essential because it allows easy communication of data. Imagine looking at your raw datasets in CSV format and trying to find something useful in the data. Then imagine later trying to explain your findings to someone else through text! It is often much better to represent data with visualisation.

DATA TYPES

Before we get into some common visualisation techniques, we need to understand the different types of statistical data. In Data Science, we tend to work with two main categories of variables: categorical and continuous.

Categorical Variables

Categorical variables are also known as discrete or qualitative variables. Examples of categorical variables are race, sex, age group, and educational level. According to Laerd (n.d.), these variables can be further divided into the following categories: *nominal*, *ordinal* or *dichotomous*.

- **Nominal variables** have two or more categories, which do not have a specific or predefined order. For example, properties could be classified as houses, condos or bungalows. Therefore, the variable that holds the property type is a nominal variable. For example, the state or province a person lives in would also be a nominal variable.

- **Dichotomous variables** are nominal variables which have only two categories or levels. For example, we could use a dichotomous variable to describe whether a person is a pensioner or not. In this case, the categories would be “True” or “False”.

- **Ordinal variables** are nominal variables, but the categories can be ordered or ranked. For example, if you were asked to rate your satisfaction with this course, your responses could be "Completely satisfied", "Mostly satisfied", "A little dissatisfied", or "Very dissatisfied".

Continuous variables

A continuous variable can take on infinitely many uncountable numerical values, including integers and floating points (decimals). They are also known as quantitative variables. Continuous variables can be further categorised as either interval or ratio variables:

- **Interval variables** are "variables for which their central characteristic is that they can be measured along a continuum and they have a numerical value (for example, temperature measured in degrees Celsius or Fahrenheit)" (Laerd, n.d.).

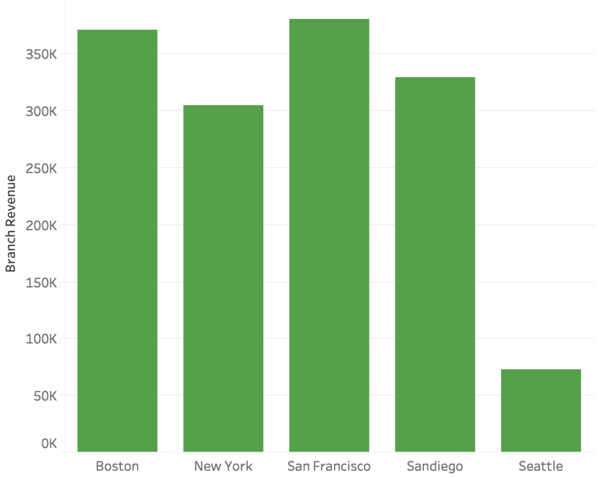
- **Ratio variables** are interval variables, but a 0 value means there is none of that particular variable. For example, "weight" is a ratio variable because if a variable measuring the sugar stock at a bakery was equal to 0 kgs it would mean there is no sugar in stock. Temperature (measured in degrees Celsius) would not be a ratio variable because 0°C does not mean that there is no temperature.

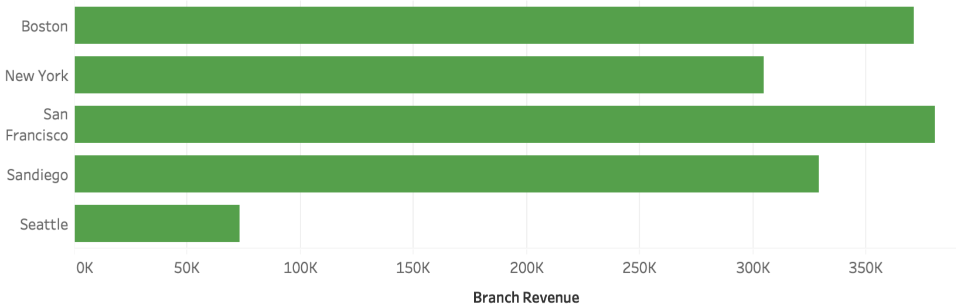
DATA VISUALISATION TECHNIQUES

You need to choose appropriate visualisations depending on what you want from your dataset. Here are some basic visualisation techniques:

Bar Chart

A bar chart/graph is also known as a column chart. It uses horizontal or vertical bars to show and compare values across categories. One axis shows categories, and the other axis shows the discrete value scale.





Best used for data that is:

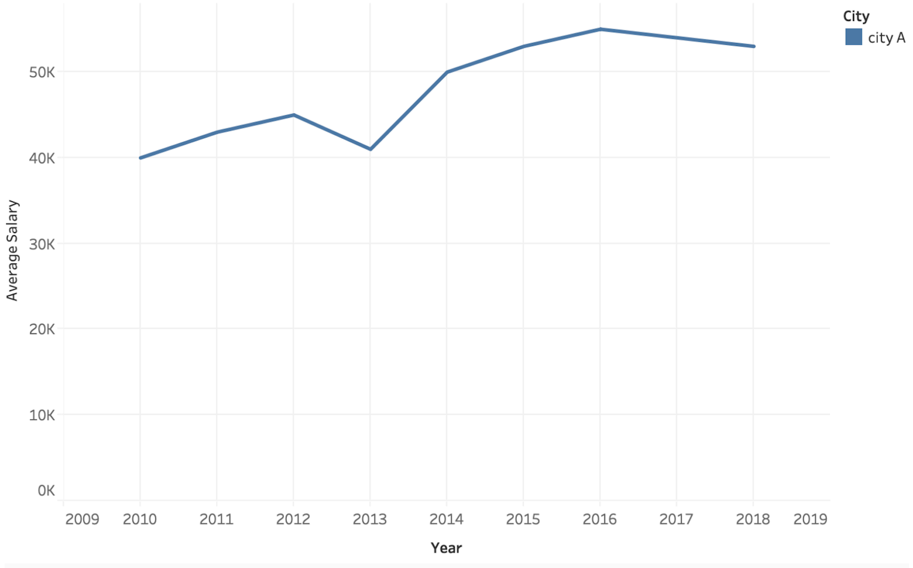
- Discrete
- Categorical

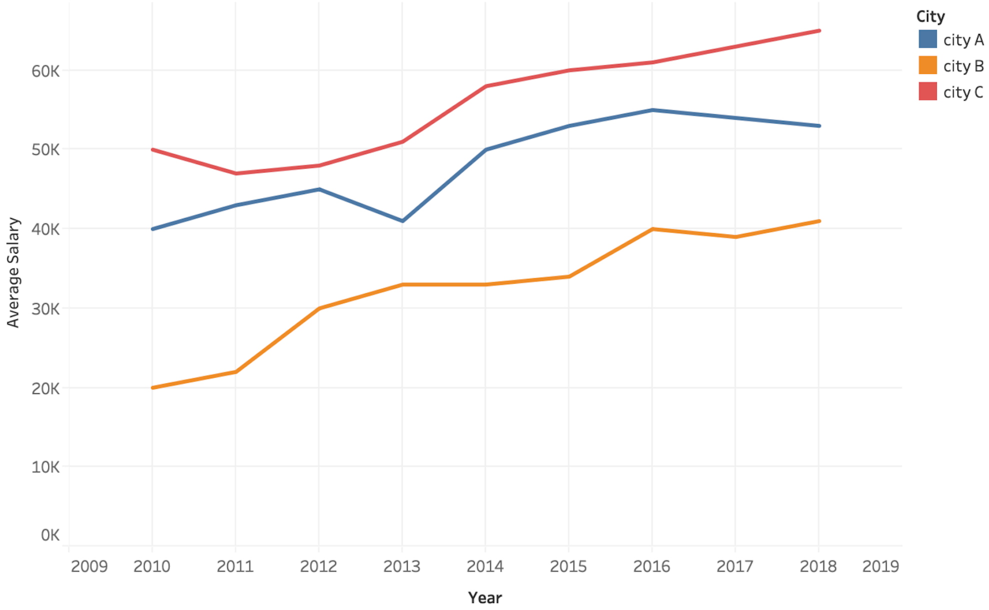
Things to keep in mind:

- Doesn't work well with too many categories of data (too many bars are hard to look at). For instance, if you're showing rainfall in mm for each day of a year, you would end up with 365 bars!
- You can colour-code the categories for patterns to stand out.

Line graph

A line graph shows values over a time period. Multiple line graphs can show and compare many different categories over a time period. Typically, one axis shows data values, while the other axis depicts time.





Best used for data that is:

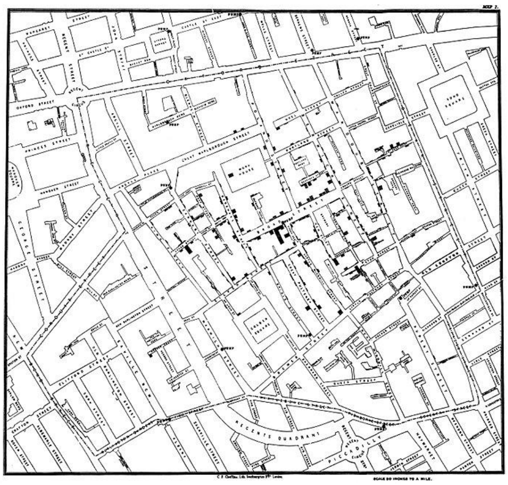
- Linear or continuous
- Time series

Things to keep in mind:

- Lines can be colour coded
- Time is often on the x-axis

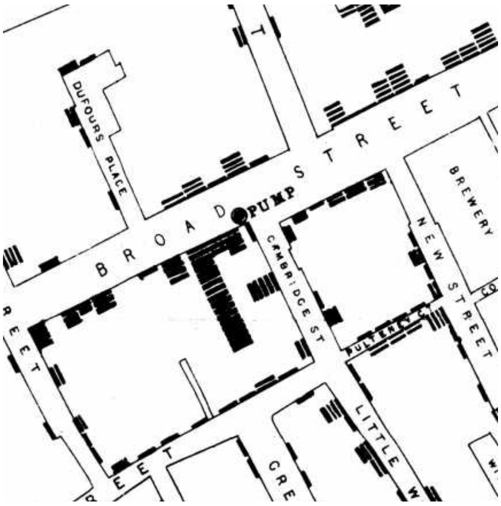
A note from our coding mentor

Masood



Did you know that one of the most famous data visualisations in history is a map by Dr. John Snow? (I'm not talking about the Game of Thrones character!)

In London, during the mid-18th-century, people were suffering from cholera outbreaks. Many thought it was spread by air, but he hypothesised that it was spread through contaminated water.

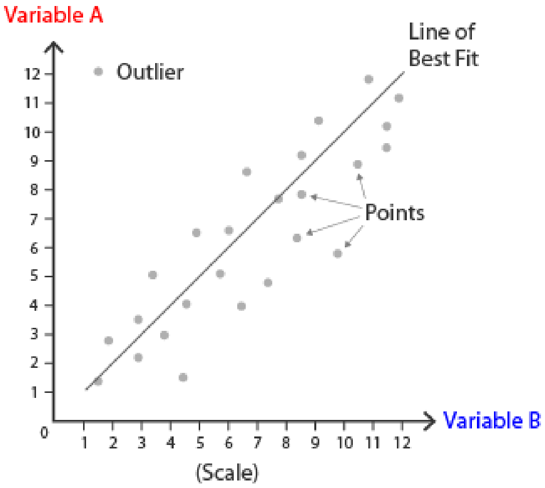


He drew out a map showing the deaths from cholera by streets and houses. There was a prominent cluster on Broad Street, where a pump was located.

After the local authorities removed it, the number of deaths reduced dramatically. This visualisation not only saved many lives but also changed the way we looked at diseases.

Scatterplot

A scatterplot is also known as a scatter graph, x-y plot or point graph. It uses a collection of data points placed on a cartesian plane (x-y axis). Through this visualisation, you can identify a trend or relationship between two variables.



Best used for data that is characterised by:

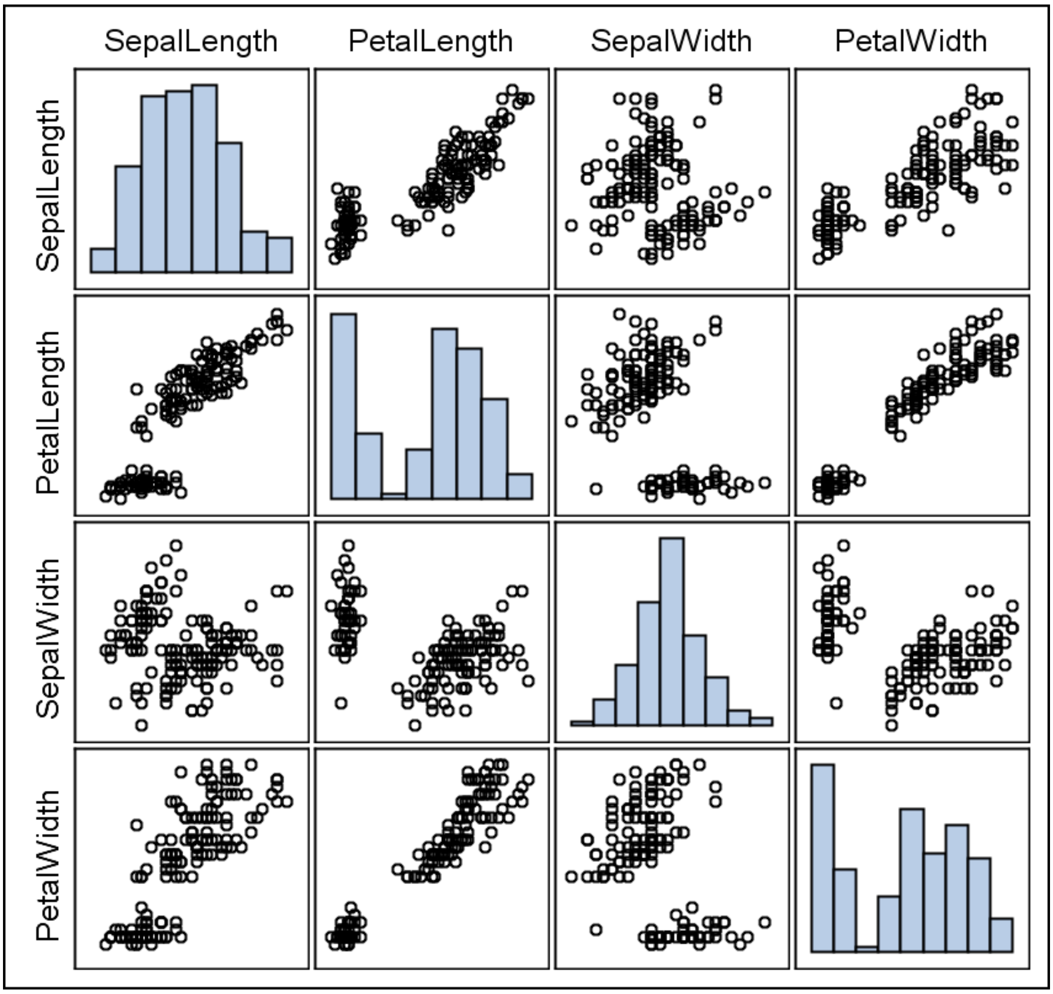
- Numerical values
- Two variables (X, Y)

Not useful for:

- Categorical data
- Time series

Scatterplot matrix

Unlike scatterplot visualisations, a scatterplot matrix helps you determine relationships between multiple variables. A histogram of each variable can included on the diagonal. Histograms are a visualisation tool used to understand the distribution of numerical data.



(**Source**: Matange, 2012)

Best used for data that is characterised by:

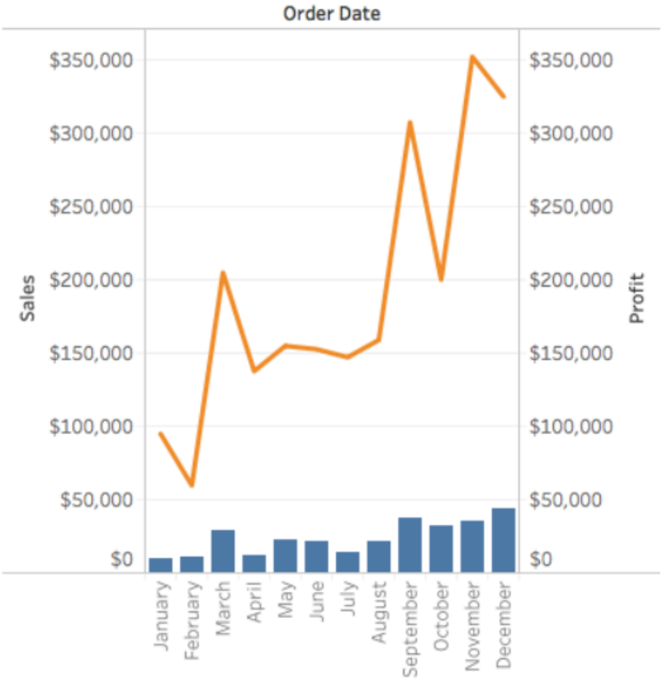
- Numerical values
- Multiple variables

Not useful for:

- Time series

Double-axis chart

A double-axis chart is also known as a dual-axis chart. A double-axis chart has two y-axes, thus the graph will have x, y1, y2 axes. It can be both a bar chart with a line graph or multiple lines on a visualisation. Be careful about having a messy visualisation because of the dual-axis.



(**Source**: *Tableau)*

Best used for data that is:

- Time series
- Discrete
- Continuous

Extra resource

Data visualisation, also known as infographics, can be overwhelming! But there are many resources, as well as information, worth consulting. **Here** is a catalogue of some of the latest data visualisation techniques.

APPROACH TO DATA VISUALISATION

1. Start with a processed and cleaned dataset

If you have raw data (not processed and not cleaned), it would likely contain issues such as missing data, unstructured data, incorrect input or unnecessary information. Thus, it is important to clean your data before visualisation. Cleaningyour data so that you can use it for analysis and visualisation is sometimes referred to as data wrangling. Sometimes you may have to take different parts of your dataset to create different visualisations.

2. Know your dataset

Knowing where your dataset comes from allows you to have a better understanding of the data background and what you want to find. Knowing it well can also help you to organise and plan your visualisations more appropriately. You may become familiar with your dataset during the process of cleaning and wrangling (see the previous step).

3. Determine what you want to find

The purpose of having data is to find information that can be helpful to achieve our goals. Think about what you are trying to find. For example, if you are working with data from a school class, you may start with a question like, “Which students have the lowest scores?”. The objective here would be to ascertain which students are performing most poorly so that you may provide extra support or tutorial classes to these students.

Often, you don’t exactly know what to look for, or it may happen that you find something in your dataset that you weren’t expecting! You can go for a broader goal if you cannot think of a specific objective. For instance, in the example we just considered, you can begin your analysis by aiming to look at student scores in general (not necessarily the lowest scores).

4. Create data visualisations

Depending on your data types and what you want to find, plan out the types of visualisations that you want to create. Some people often start with a quick sketch of the visualisations.

Note: You may go back and forth between step 4, 5, and 6. Often you may find yourself interchanging steps 5 and 6.

5. Refine your visualisation

If you have decided on the visualisations you are doing, or if you already have existing visualisations, you would likely want to refine them further. This means making the visualisations more aesthetically pleasing and more accessible for the target audience to understand. Some examples of improving your visualisation would be: changing the colours, making the borders thicker, enlarging your titles/fonts, etc. You may want to solicit other people's opinions since aesthetics can be subjective.

When deciding how your visualisation will look, always remember who your audience is. For instance, if your audience is a group of data science experts, then you can choose complicated visualisations such as a scatterplot matrix, but if your audience consists of generalists, you should create a more straightforward visualisation such as a bar graph.

6. Note down your findings

Once you have your visualisations, you need to look for patterns, make inferences, dig deeper and conclude what you have found. The initial stages of this analysis can add to your original goals and research question, as mentioned in step 3.

APPROACH EXAMPLE

1. Start with a processed and cleaned dataset

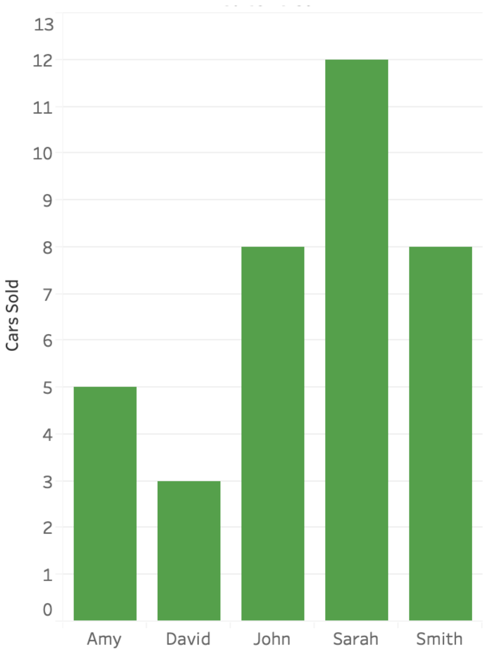
month (2018)	salesPerson	carsSold
February	Amy	2
February	Smith	6
April	Sarah	3
May	John	7
May	Sarah	9
January	Amy	3
June	John	1

June	Smith	2
August	Sarah	3

2. Know your dataset: above is the processed sales data of a small car dealer. You have chosen to look at the months of Jan - August in 2018, the sales team, and the number of cars sold.

3. Determine your objectives: what would be useful to find in this dataset aimed at the car dealership management team? Perhaps who has sold the most and least number of cars? Any pattern in the months?

4. Creating your visualisation: a bar graph can suitably represent this dataset.

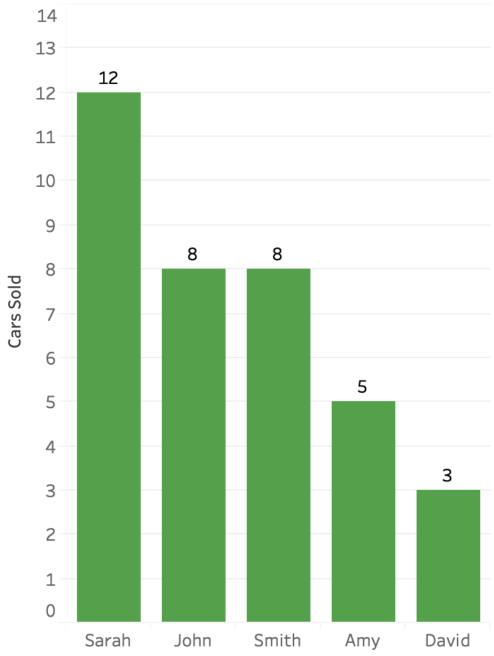


5 & 6. Note down findings & perfect your visualisation: We can see that Sarah has sold the most cars and David has sold the least. From this visualisation, the car dealership can:

- Reward the best salesperson (Sarah).
- Provide more training for the worst salespeople (David and Amy).

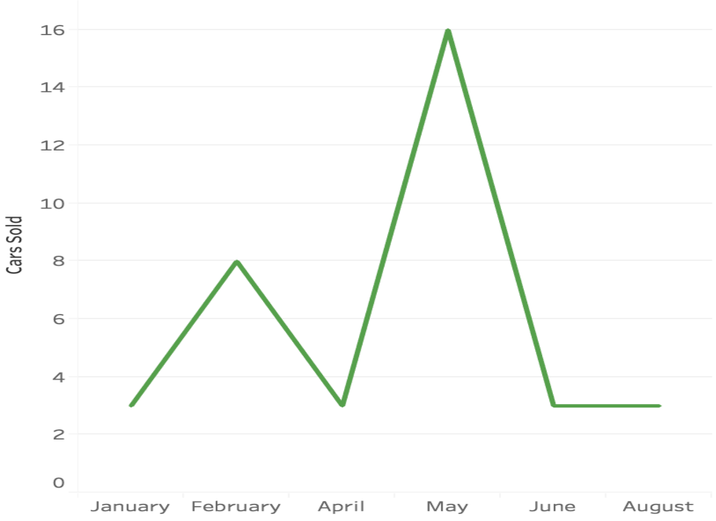
Suppose we were to make the bar graph more aesthetically pleasing and easier to understand. In that case, we can implement a few changes:

- Descending order: quickly shows the order
- Label the value of cars sold: immediately know how much each person sold



Or perhaps we can improve the graphs first before drawing conclusions, especially if it is a big dataset and a more complicated graph.

But what about a time trend in sales? Sometimes you may want to use different visualisations on the same dataset to discover or convey more information.



If you create a line graph with the same dataset, you can draw further conclusions: May was the best month for sales, while June and August were slower. Through this information, the car dealership may:

- Create more promotions starting June
- Hire more staff during May

Compare the bar graph and line graph to the table data – aren't the visualisations much easier to understand?

Practical Task 1

Examine the graphs below and use some research and background knowledge to make conclusions based on your observations. For each graph, answer the questions associated with the graph in a document titled **data_viz.** Convert your answer document to a PDF before submitting it.

1. The following bar graph shows the gender wage gap in 26 countries based on data collected by the **OECD**. The gender wage gap is calculated by finding the difference between male and female median wages and dividing it by male median wages. It is represented as a percentage in this graph.

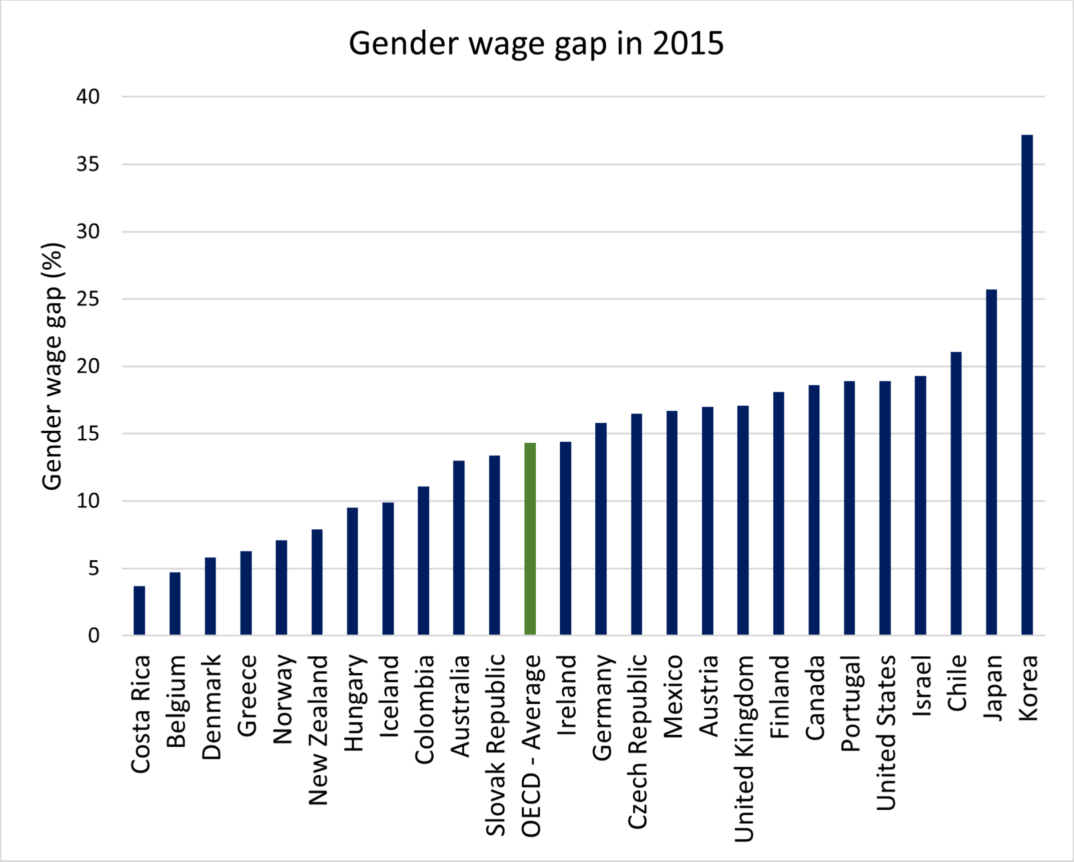


Figure 1: Gender wage gap in 2015 (**Source**)

- Which three countries have the lowest gender wage gap?
- Which three countries have the highest gender wage gap?

- Do some research on the country with the lowest gender wage gap and comment on why you think it succeeded in achieving a low gender wage gap in 2015 (max. 150 words).

2. The following line graph shows the sale of isopropanol from May 2019 to March 2020 in the United States of America. The sales are measured using US cents per weight (lb) of the product (US CTS/lb). Focus on the general trend of the three lines on the graph rather than what each of the lines refers to specifically when answering the questions.

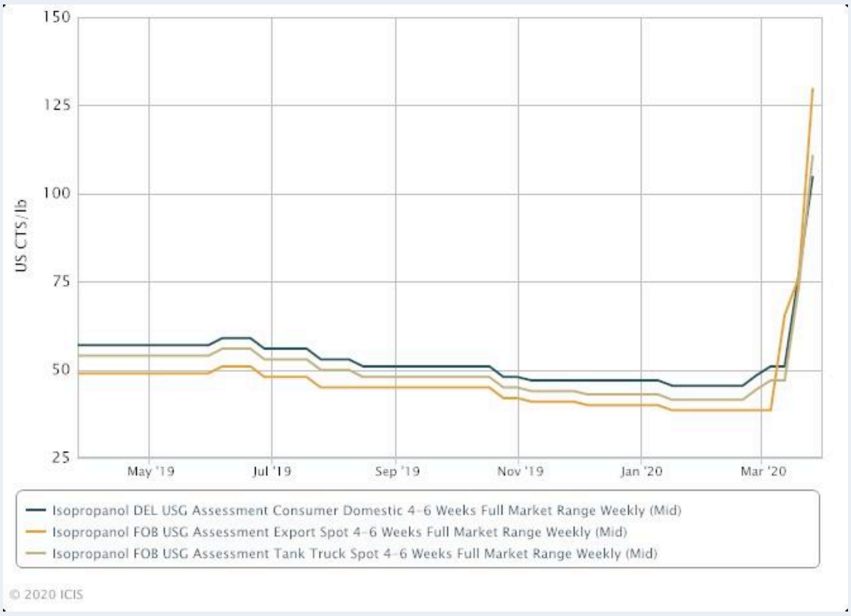


Figure 2: Isopropanol sales from May 2019 to March 2020 (**Source**)

- Explain what is happening in the graph during March 2020 with regards to isopropanol sales (max. 100 words).

- Describe a possible reason for the observation you made about isopropanol sales in March 2020 (max. 100 words). **Hint:** Isopropanol is the main ingredient in hand sanitiser.

3. Below, the bubble plot (a scatter plot with variable dot size) shows carbon dioxide (CO_2) emissions per person in tonnes vs the gross domestic product (GDP) per capita (average per person). No unit is given for the GDP per capita; however, the US dollar is typically used when comparing different countries (Callen, n.d.). Each dot represents a country. The colours of the dots refer to the continent to which the country belongs. The size of the dot refers to the size of the population in the country. The larger the dot, the larger the population.

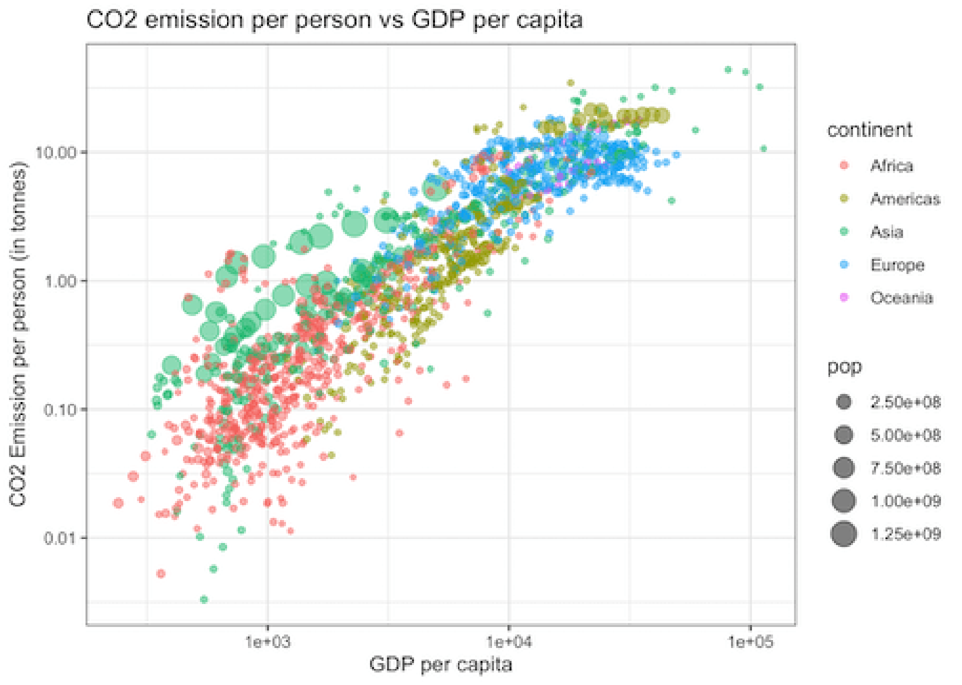


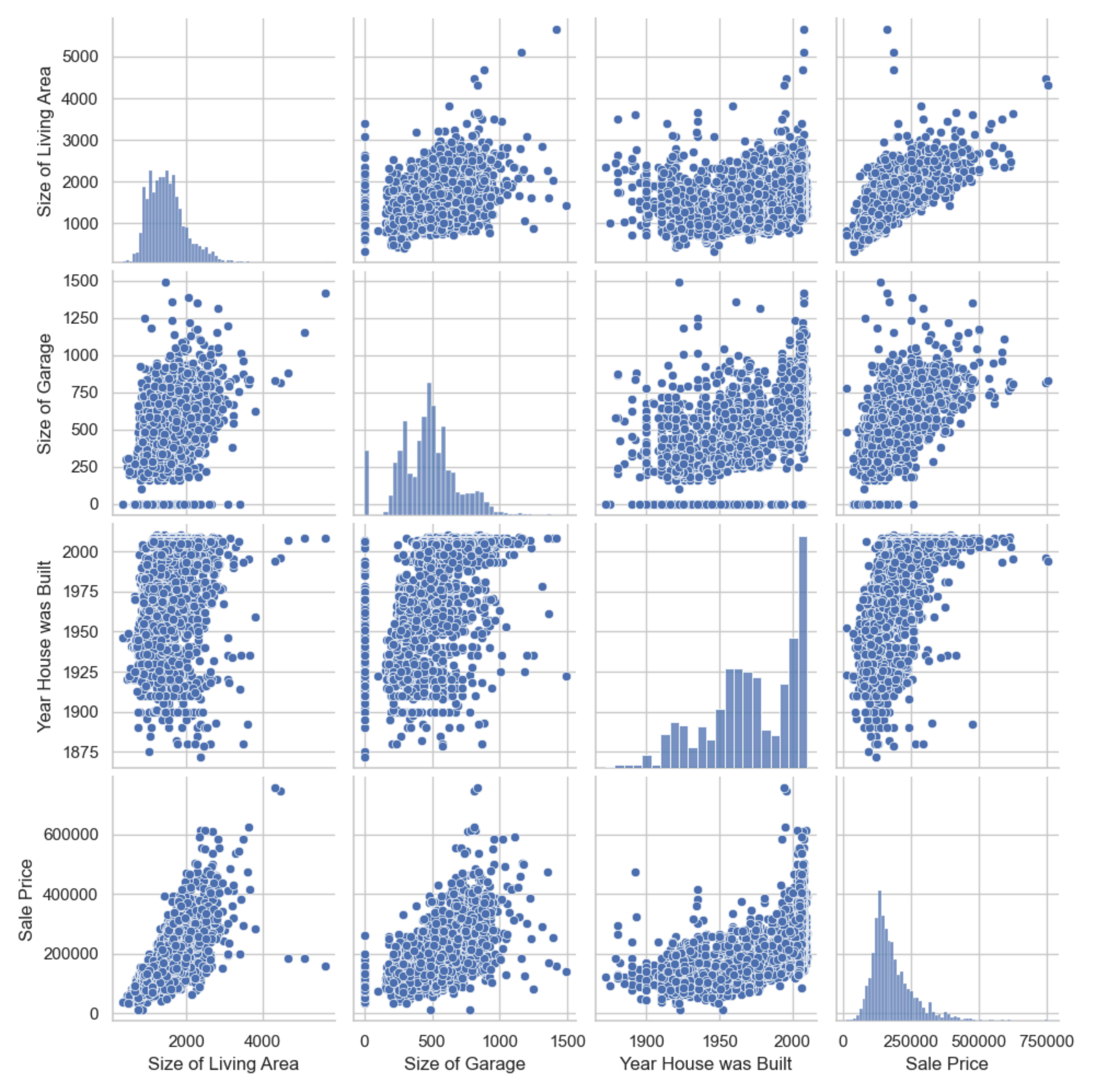
Figure 3: CO_2 emissions per person vs GDP per capita

- Discuss the relationship between CO_2 emissions per person and GDP per capita for each continent listed in the figure legend (max. 350 words).

Practical Task 2

The following scatterplot matrix is from the Ames Housing dataset. It contains data collected by the Ames City Assessor's Office describing 2930 property sales which occurred in Ames, Iowa between 2006 and 2010. The data includes the sale price ($), year the house was built, size of the garage (ft^2) and size of the living area (ft^2).

- Examine this graph and answer the questions that follow in your **data_viz** document.



- What do the graphs along the diagonal represent?
- Are most garages in Ames larger or smaller than 1000 ft^2?
- Are the most expensive houses in Ames built before or after 1950?
- Describe the relationship between ‘Size of Living Area’ and ‘Sale Price’.

Rate us

Share your thoughts

HyperionDev strives to provide internationally-excellent course content that helps you achieve your learning outcomes.

Think that the content of this task, or this course as a whole, can be improved? Do you think we've done a good job?

Click here to share your thoughts anonymously.

REFERENCES

Callen, T. (n.d.). Purchasing Power Parity: Weights Matter. Retrieved 08 March 2023, from **https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/Purchasing-Power-Parity-PPP#**

Jones, T. (2018). An introduction to data science. Retrieved 25 August 2020, from **https://www.ibm.com/developerworks/library/ba-intro-data-science-1/index.html**

Holtz, Y and Healy, C. (n.d.) The Gender Wage Gap. Retrieved 08 March 2023, from **https://www.data-to-viz.com/story/OneNumSevCatSubgroupOneObsPerGroup.html**

Koray, D. (2020). US IPA prices rise on unprecedented hand sanitizer demand. Retrieved 13 October 2022, from **https://www.icis.com/explore/resources/news/2020/03/26/10487220/us-ipa-prices-rise-on-unprecedented-hand-sanitizer-demand/**