



TASK

Data Visualisation I

Visit our website

Introduction

WELCOME TO THE DATA VISUALISATION I TASK!

As science and technology advances, it is inevitable that the amount of information and data we possess 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 important 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.



Get in touch
Connect for support

Remember that with our courses, you're not alone! You can contact your mentor to get support on any aspect of your course.

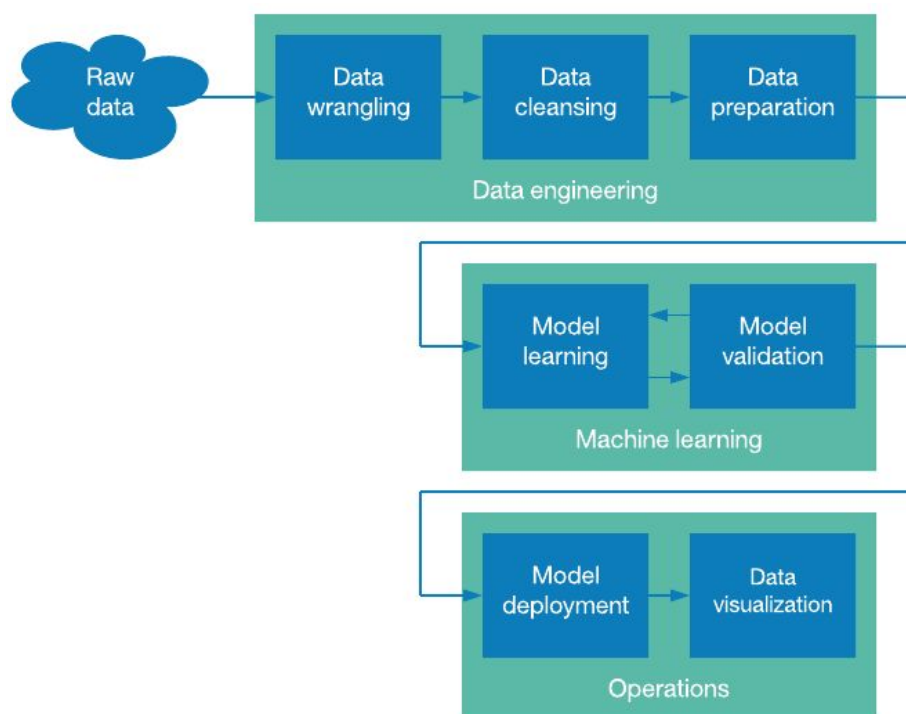
The best way to get help is to login to www.hyperiondev.com/portal to start a chat with your mentor. You can also schedule a call or get support via email.

Your mentor is happy to offer you support that is tailored to your individual career or education needs. Do not hesitate to ask a question or for additional support!

THE DATA SCIENCE PIPELINE

Until now, the focus of the tasks in this level has had to do with obtaining the data that we will use for data science. We now know how to read and work with data stored in various sources such as databases, Web APIs, JSON, XML and CSV files.

However, accessing the data is just the first step for data scientists. There are various tasks that data scientists perform to analyse data in a meaningful way. Different people describe these steps differently. The steps don't have to be carried out in a strict order. However, the diagram below broadly shows the steps that are usually involved in the data science pipeline:



(Jones, 2018)

Data wrangling is defined as “the process by which you identify, collect, merge, and preprocess one or more data sets in preparation for data cleansing” (Jones, 2018). Once you’ve wrangled the data into a format with which you can work, you **clean the data** by fixing any errors in the data set. This could involve dealing with missing data, correcting formatting issues and fixing, or removing, incorrect data. You can then prepare the data for machine learning. You will tackle machine learning in the next level. However, in later tasks within this level, you will learn techniques for **data preparation**, such as normalisation.

Evidently, the first job undertaken by a data scientist is to gain a clear understanding of the data that they are working with. The data scientist needs to

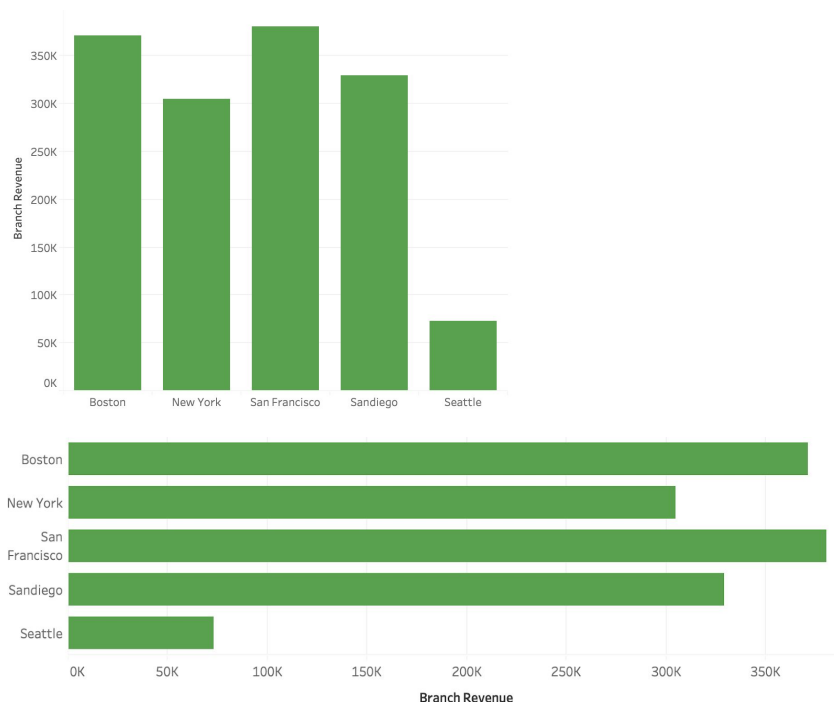
understand what data they have, what format it is in, what format it needs to be converted to, what is wrong with the data and how to fix problems. Various visualisations (such as graphs, tables and charts) are used to gain this insight into the data.

Therefore, data visualisation is important at both the end of the data science pipeline (to present the findings of a study) and in the initial phase (data engineering phase) of the data science pipeline to help gain an understanding of data. Thus, before we get to tasks that deal with data wrangling, data cleansing and data preparation, the next few tasks first deal with techniques used to visualise data. Then, in later tasks, you will learn many techniques needed for data wrangling and cleansing.

DATA VISUALISATION TECHNIQUES

Depending on what you want to find from your dataset, you need to choose appropriate visualisations. Here are some of the basic visualisation techniques:

- **Bar Chart:** A bar chart/graph is also known as a column chart. It uses either 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 (see the note from your mentor)

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.



A note from our coding mentor **Sarah**

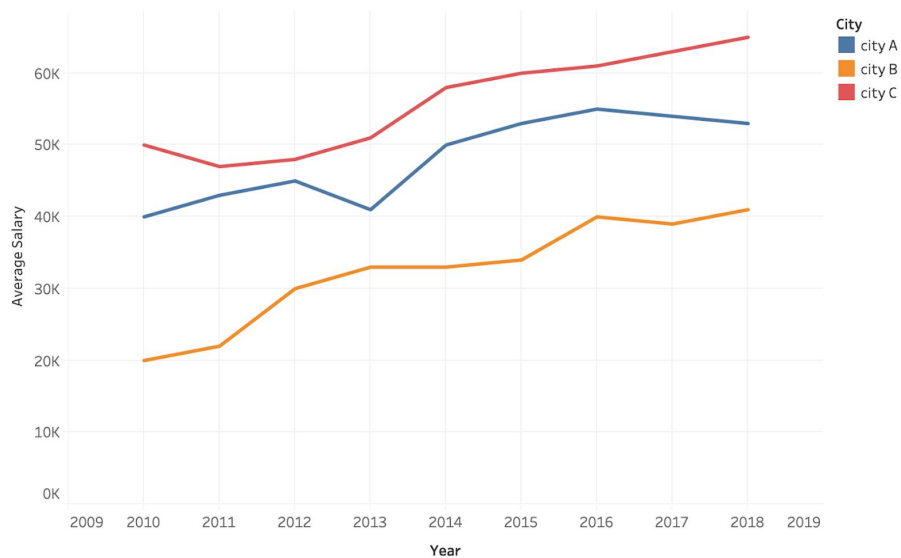
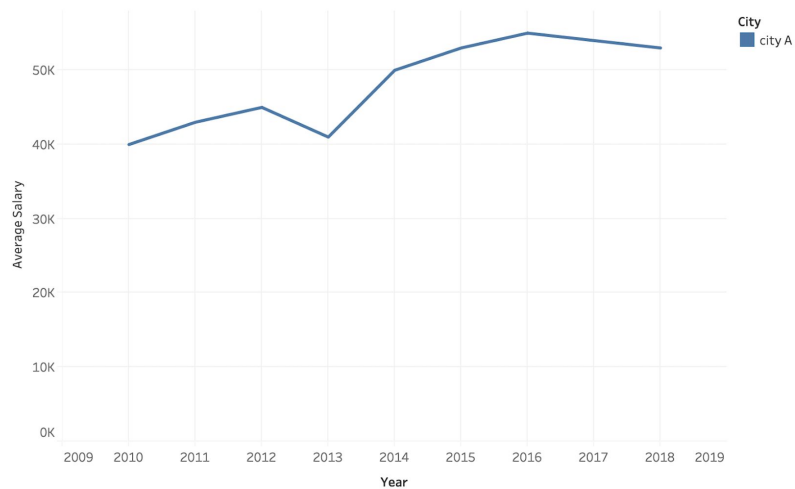
In data science, we work with different types of statistical data. Data can be numerical (the data can be measured using numbers), categorical (data represent characteristics such as a person's age, gender, education level, etc.) or ordinal (numerical and categorical data mixed).

Numerical data can be either:

- *Discrete: data can be counted (e.g. the number of students in a class, number of people in a city, etc.), or*
- *Continuous: data can't be counted, but can be measured (e.g. a person's blood pressure, weight, height, etc.)*

For more information, see the short description of types of statistical data [here](#).

-
- **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.



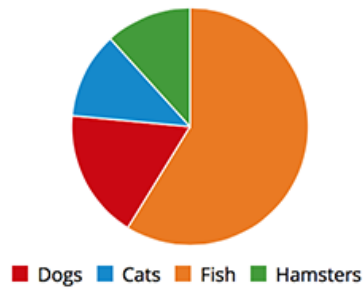
Best used for data that is:

- Linear/continuous
- Time series

Things to keep in mind:

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

- **Pie chart:** A pie chart illustrates the proportion of the quantity it represents. The total sum of all segments equals 100%. Each segment needs to be a different colour or pattern to differentiate it. It is good practice to have no more than 10 different segments.



Best used for data that is:

- Easy to categorise in distinct categories

- **Pyramid Chart:** A pyramid chart is a hierarchical visualisation that shows the relationships through the stacks. It can be bottom heavy and narrow at the top, or top-heavy and narrow at the bottom as an inverse pyramid.



Best used for data that is:

- Different with sub-categories
- Workable with non-specific numerical values

- **Word Cloud:** A word cloud displays weighted text - that is, the size of each text is weighed according to the number of times this word has appeared. It can be categorised by colour, which makes it all the easier to convey the text or findings.



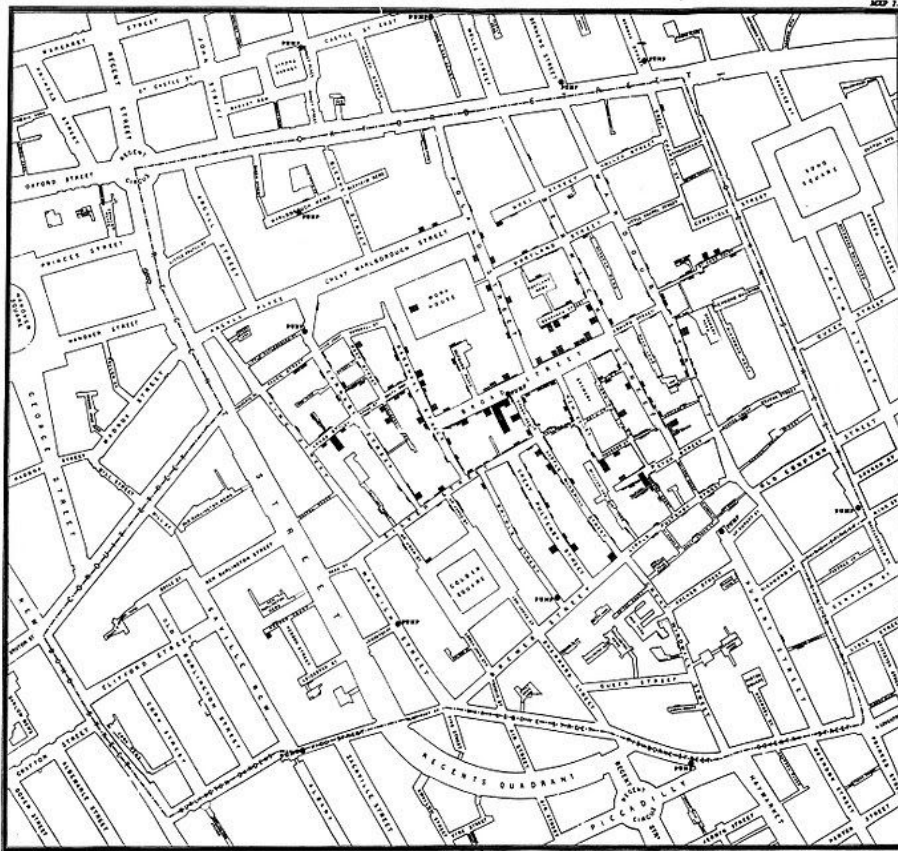
Best used for data that is:

- Text-based.



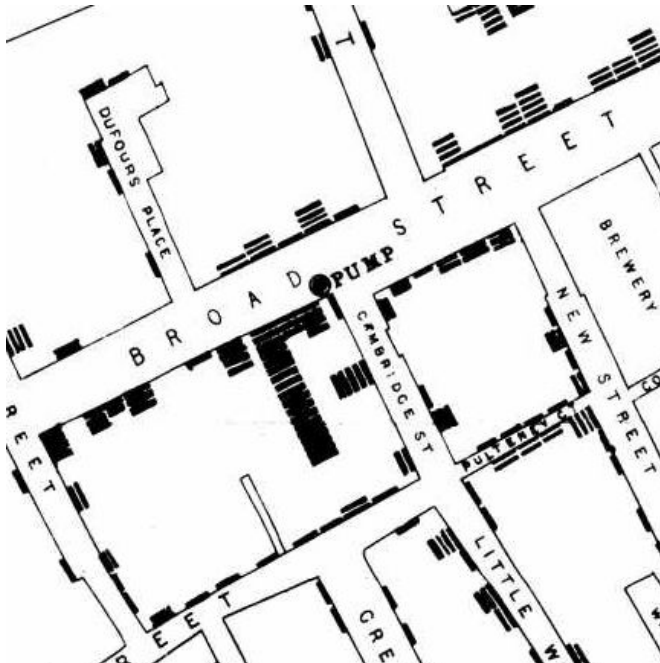
A note from our coding mentor **Masood**

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 an obvious 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.

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. The process of cleaning your data so that you are able to 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! If you cannot think of a very specific objective, you can go for a broader goal. 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 easier 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 is going to look, always keep in mind 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 simpler 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.

EXAMPLE: APPROACH TO DATA VISUALISATION

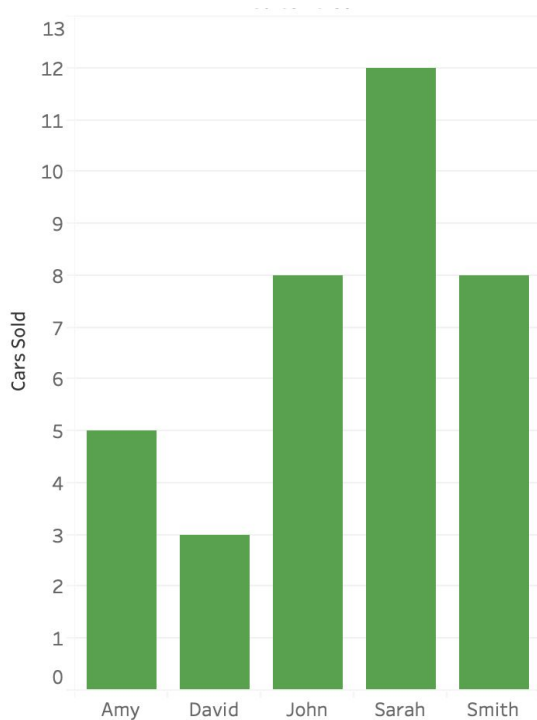
1. Start with a processed and cleaned dataset

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: In this dataset aimed at the car dealership, what would be useful to find? 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.

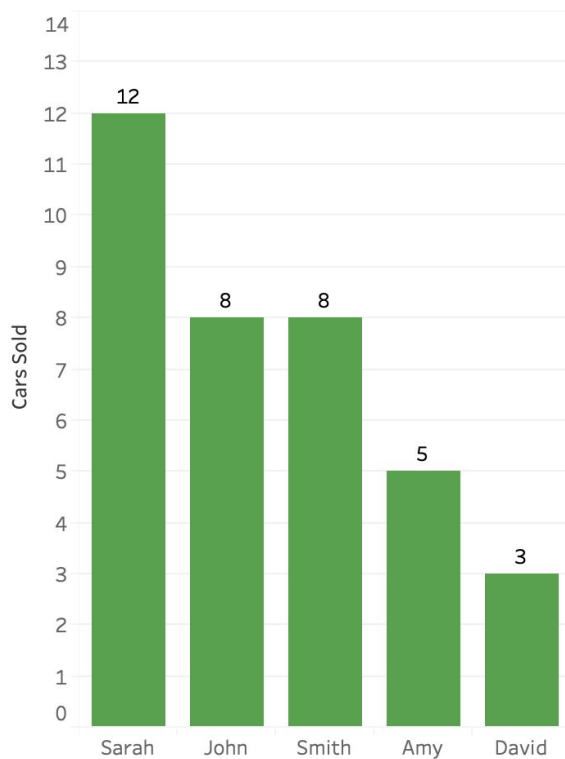


5 & 6. Note down findings & perfect your visualisation: We can see that Sarah has sold the most cars and David has sold the least. From this visualisation, the car dealership can:

- reward the best salesperson (Sarah)
- Provide more training for (David and Amy)

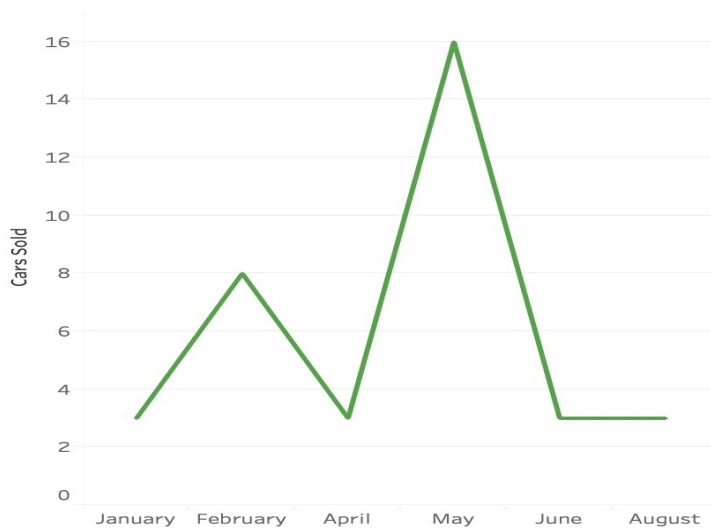
If we were to make the bar graph more aesthetically pleasing and easier to understand, we can implement a few changes:

- 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 arrive at further conclusions: It seems that 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 we had - aren't the visualisations much better to understand!

Compulsory Task 1

- Create a file named **task1.txt** and describe at least two scenarios, referring to real-world data, where using data visualisation would significantly ease the process of data analysis, inference, or presentation of findings.

Compulsory Task 2

Follow these steps:

- Inside this folder, find 2 datasets:

first_name	last_name	student_id	gender	math_score	english_score	science_score	Mother	Father	Address
Theresa	Imore	1	F	89	78	50	Jane Claremint	Michael Scofield	90 East Buckingham
Brook	Gratrex	2	M	67	56	65	Joesephine Brown	Lincon Burrows	60 Fairway Ave.
Jerrold	Isenor	3	M	98	67	42	Claire Mint	Ballack Benet	749 West Kingston Street
Jo	Pretsel	4	M	56	72	87	Alexis Mahone	Duke Harry	Lawrence Township, NJ

ID	Gender	Age	Income
1	Male	Under 30	Low
2	Male	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	Low
5	Female	30 or more	High

- Create a file named **task2.txt** and explain what type of visualisation (s) you would create from each of the above two datasets and why?

Completed the task(s)?

Ask your mentor to review your work!

[Review work](#)



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:

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>**