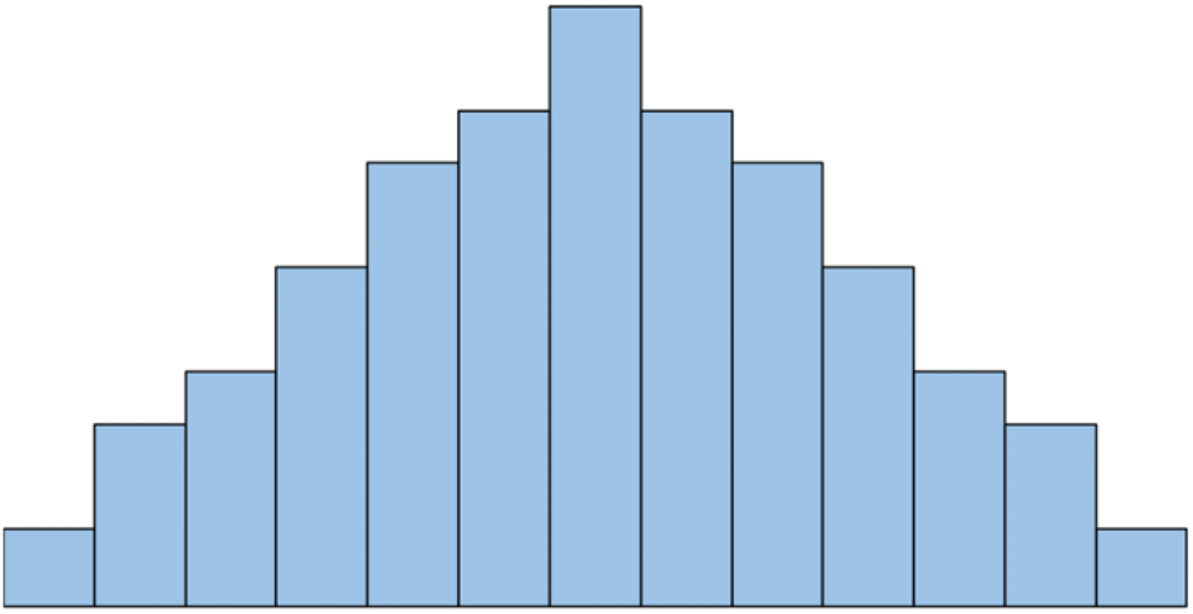
Ans1(a)1 : Bimodal graph

explain:

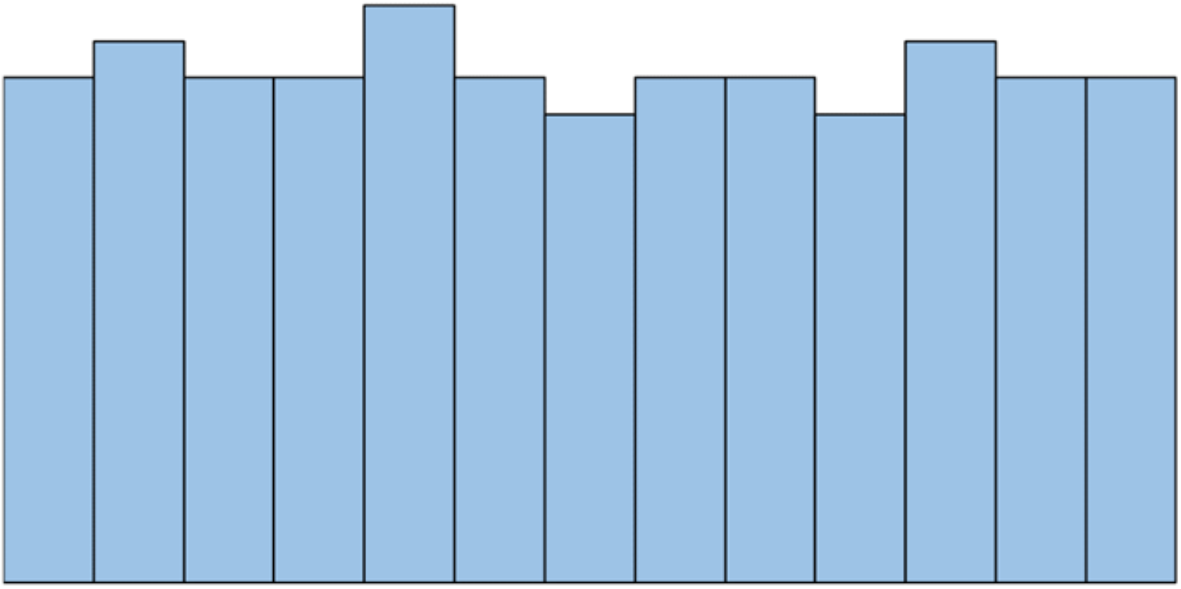
1. Bell-Shaped

A histogram is bell-shaped if it resembles a “bell” curve and has one single peak in the middle of the distribution. The most common real-life example of this type of distribution is the normal distribution.



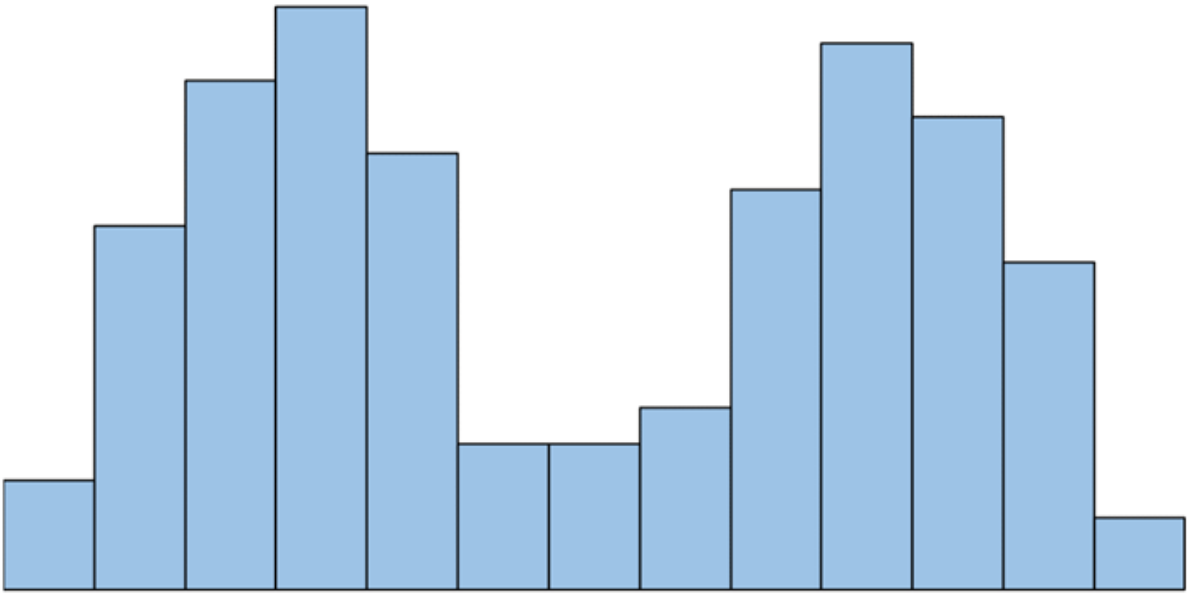
2. Uniform

A histogram is described as “uniform” if every value in a dataset occurs roughly the same number of times. This type of histogram often looks like a rectangle with no clear peaks.



3. Bimodal

A histogram is described as “bimodal” if it has two distinct peaks. We often say that this type of distribution has multiple modes – that is, multiple values occur most frequently in the dataset.

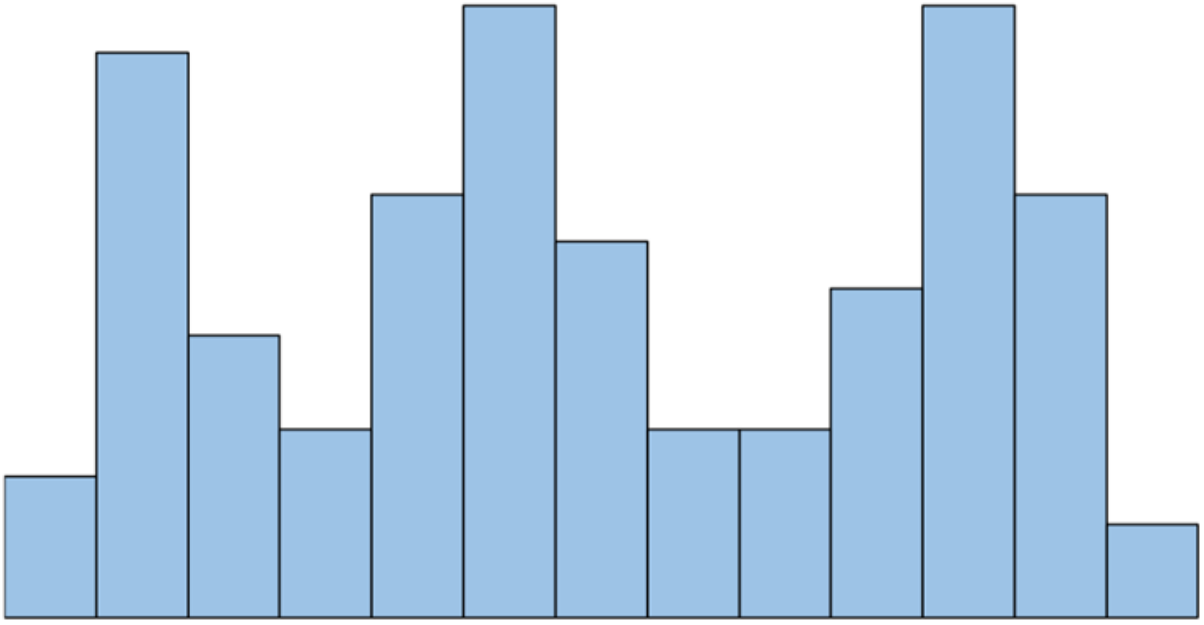


Related: What is a Bimodal Distribution?

4. Multimodal

A histogram is described as “multimodal” if it has more than two distinct peaks.

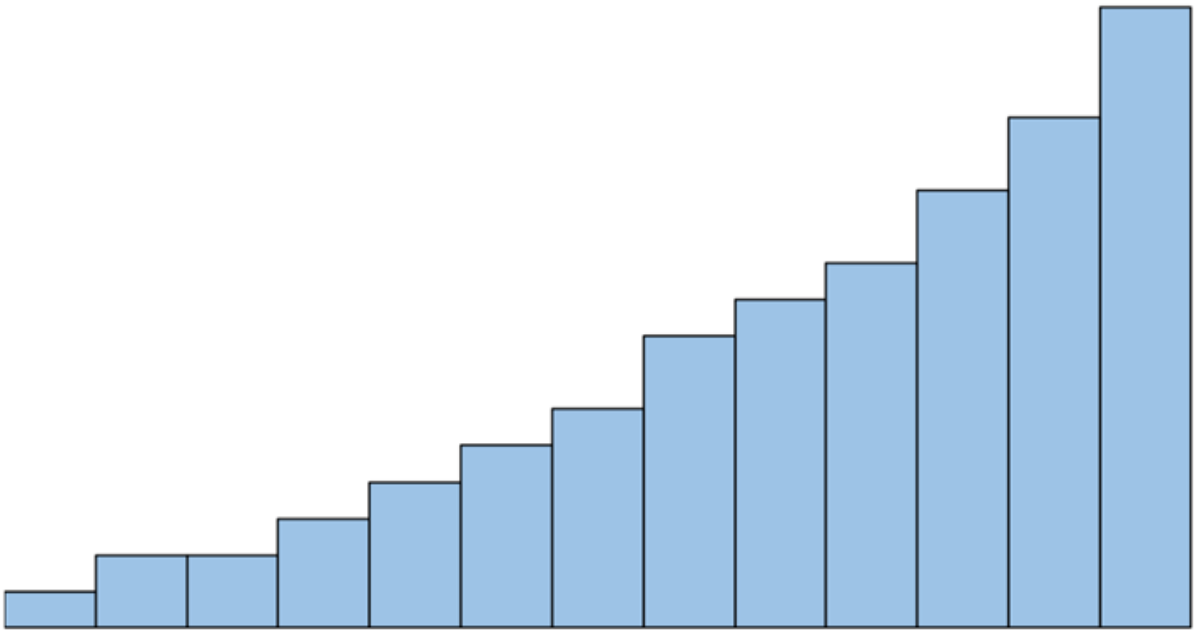
Related: What is a Multimodal Distribution?



5. Left Skewed

A histogram is left skewed if it has a “tail” on the left side of the distribution. Sometimes this type of distribution is also called “negatively” skewed.

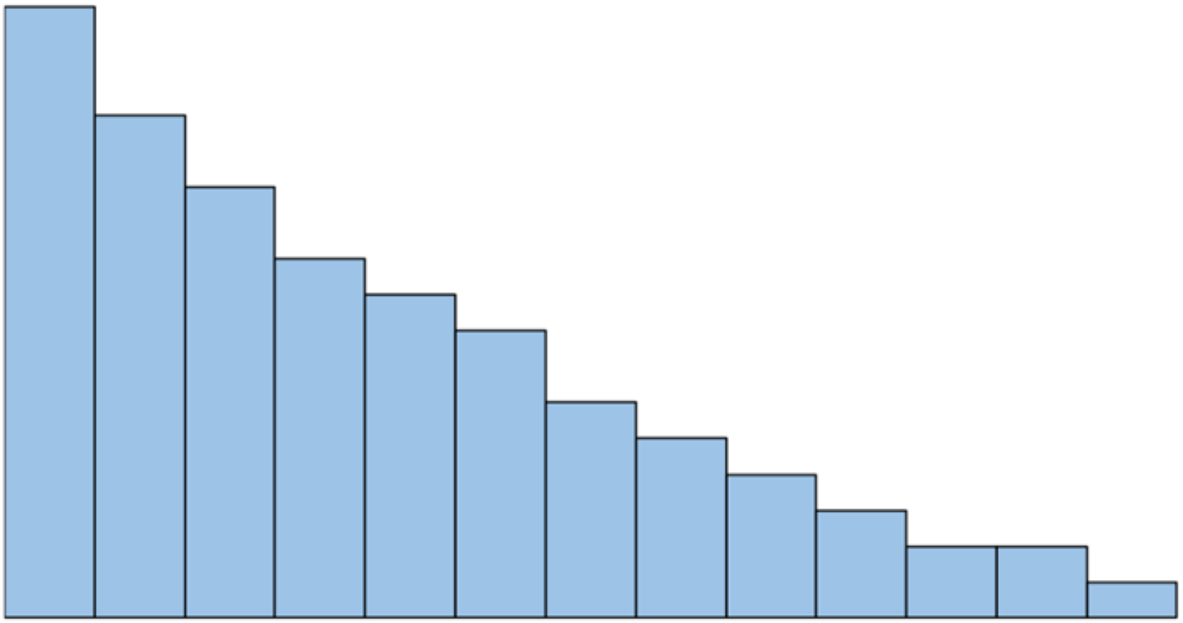
Related: 5 Examples of Negatively Skewed Distributions



6. Right Skewed

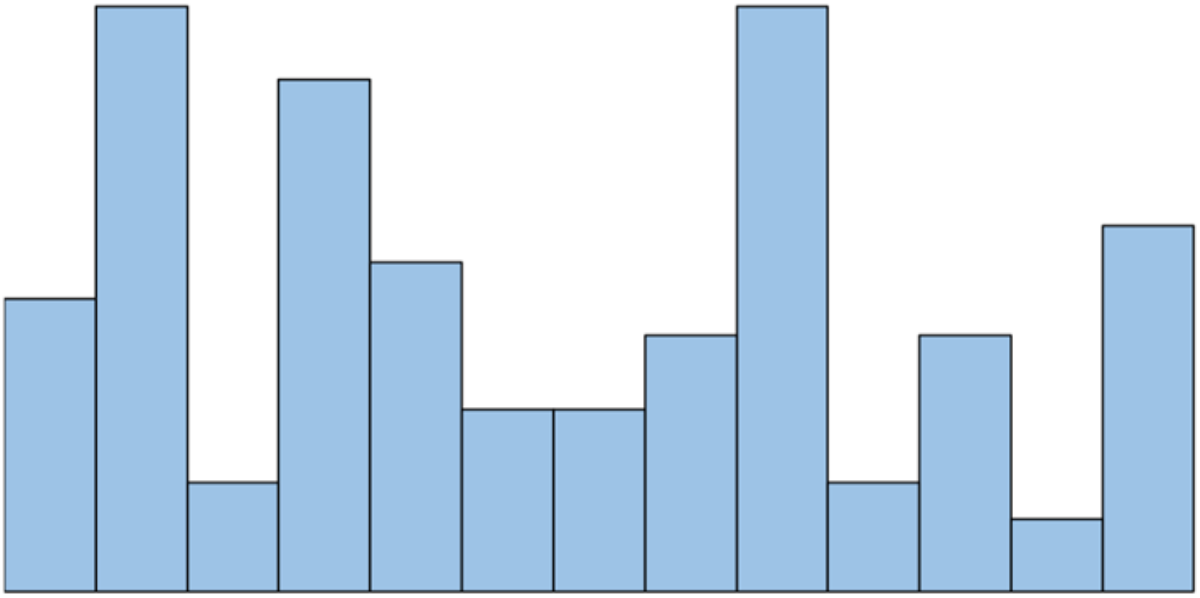
A histogram is right skewed if it has a “tail” on the right side of the distribution. Sometimes this type of distribution is also called “positively” skewed.

Related: 5 Examples of Positively Skewed Distributions



7. Random

The shape of a distribution can be described as “random” if there is no clear pattern in the data at all.



Ans1a2:

Toothbrush Color

Red

Blue

Green

Yellow

Ans1a3:
D

Ans1b1:

Geospatial graph, marks = Point, attribute = Category

Attributes:

- Attributes are the characteristics or properties of the data that you want to visualize. In the context of geospatial data, attributes often include information such as location (latitude and longitude), magnitude, category, or any other relevant data.
- For example, if you are visualizing earthquake data on a map, the attributes might include the earthquake's location (latitude and longitude), magnitude, date, and depth.

Geospatial Graph

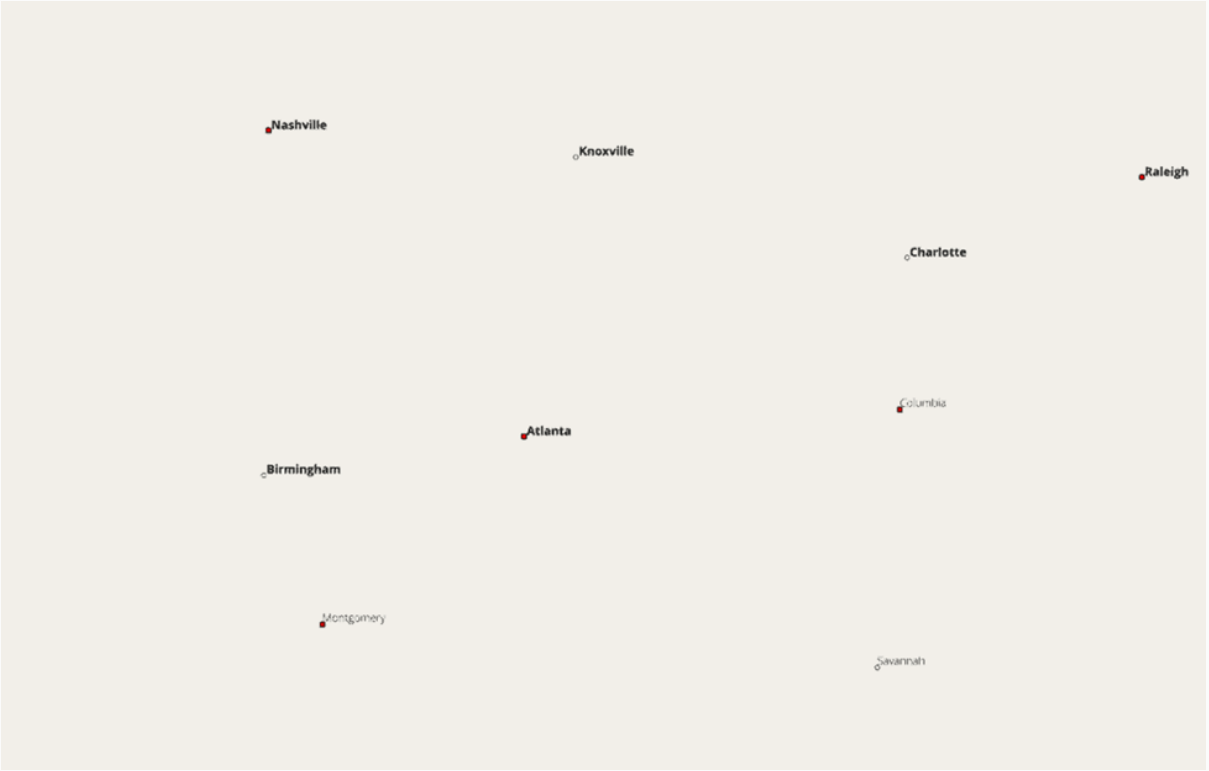
Explain:

Point Data

Point data is most commonly used to represent nonadjacent features and to represent discrete data points. Points have zero dimensions, therefore you can measure neither length or area with this dataset.

Examples would be schools, points of interest, bridge and culvert locations. Point features are also used to represent abstract points. For instance, point locations could represent city locations or place names.

In GIS, point data can be used to show the geographic location of cities. Map: Caitlin Dempsey using Natural Earth Data.

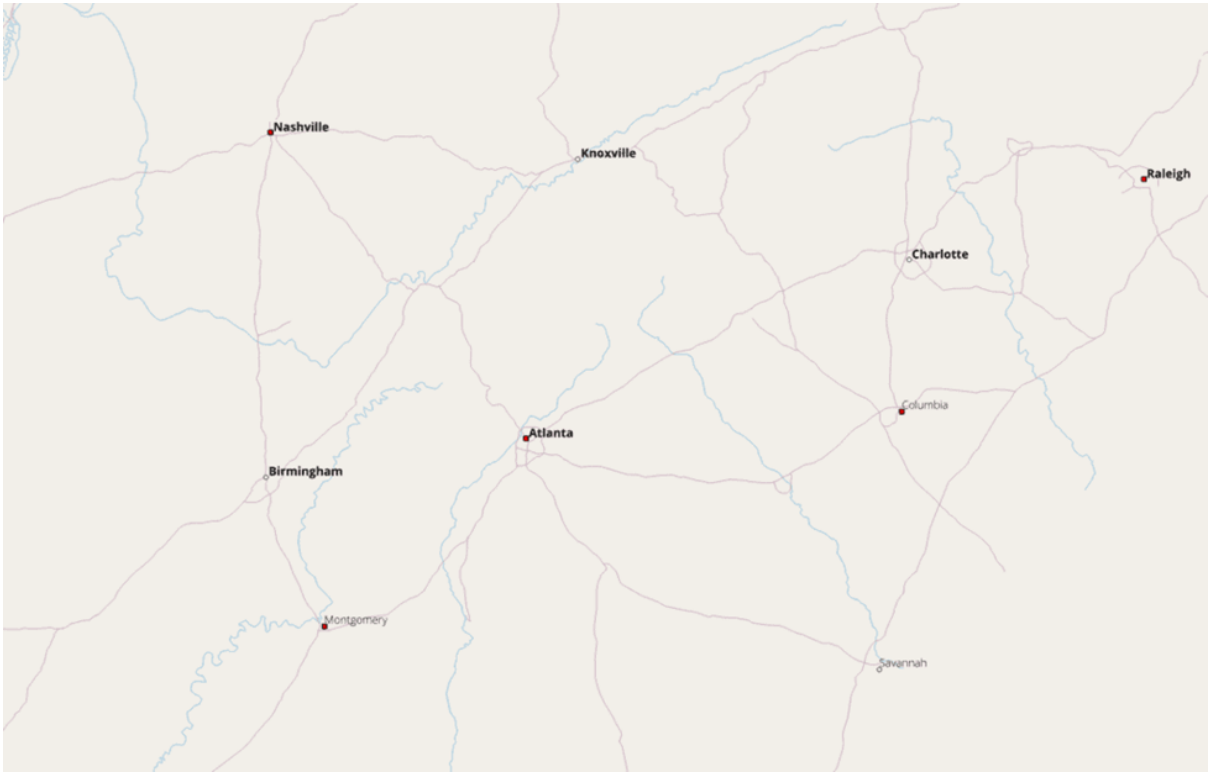


Line Data

Line (or arc) data is used to represent linear features. Common examples would be rivers, trails, and streets. Line features only have one dimension and therefore can only be used to measure length. Line features have a starting and ending point. Common examples would be road centerlines and hydrology.

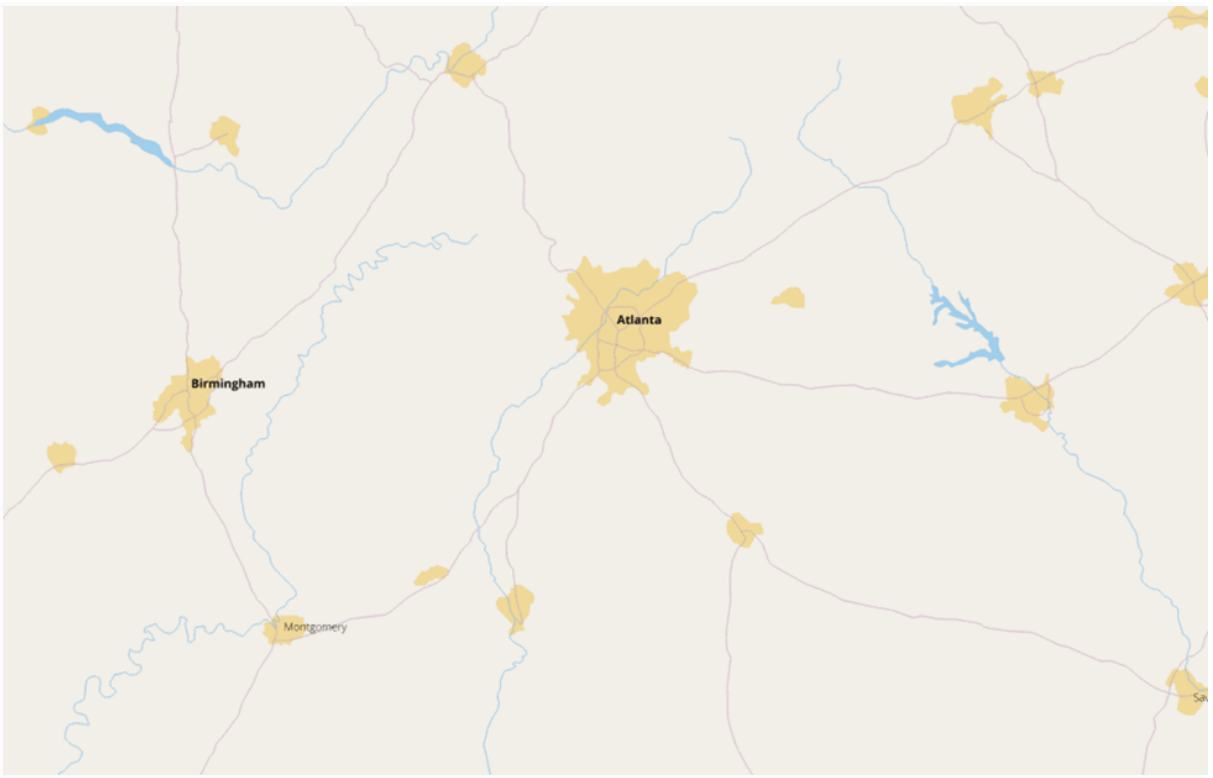
Symbology most commonly used to distinguish arc features from one another are line types (solid lines versus dashed lines) and combinations using colors and line thicknesses. In the example below roads are distinguished from the stream network by designating the roads as a solid black line and the hydrology a dashed blue line.

In this map, roads and waterways are shown as line data. Map using Natural Earth Data.



Polygon Data

Polygons are used to represent areas such as the boundary of a city (on a large scale map), lake, or forest. Polygon features are two dimensional and therefore can be used to measure the area and perimeter of a geographic feature.
Polygon features are most commonly distinguished using either a thematic mapping symbology (color schemes), patterns, or in the case of numeric gradation, a color gradation scheme could be used.



With maps presented at a larger scale, city locations are represented as a polygon to show the extent of each city. Map made with Natural Earth Data.

Both line and point feature data represent polygon data at a much smaller scale. They help reduce clutter by simplifying data locations.

As the features are zoomed in to, the point location of a school is more realistically represented by a series of building footprints showing the physical location of the campus.

Line features of a street centerline file only represent the physical location of the street. If a higher degree of spatial resolution is needed, a street curbwidth file would be used to show the width of the road as well as any features such as medians and right-of-ways (or sidewalks).

Ans1b2:

Not sure: 1) although we are not showing the enough information but we can increase the size of the age groups.

we can Use geospatial heatmaps, it will help to prevent the data privacy(age groups in our case), which were mentioned in the graph by states.

Ans1b3:

RISKS:

Identification Risk: The combination of detailed patient information (date of diagnosis, blood type, gender, ethnicity) with geographic data (county) increases the risk of re-identification, allowing individuals to be identified.

1. **Sensitive Attribute Disclosure:** The release of detailed attributes such as blood type, gender, and ethnicity may lead to sensitive information disclosure, potentially stigmatizing certain groups or individuals.

How can we prevent these:

1) we can use broader groups instead of direct/ specific details like:

- Instead of reporting individual blood types (e.g., A+, B-, AB+), you could use broader blood type categories, such as:

- Group A: A+, A-,
- Group B: B+, B-,
- Group AB: AB+, AB-,
- Group O: O+, O-

- Instead of reporting detailed ethnic categories, you could use more general classifications:

- · Asian
- · Black/African American
- · White/Caucasian
- · Hispanic/Latino

- ❖ If reporting the exact date of diagnosis poses privacy risks, consider using broader time intervals, such as reporting by months or quarters instead of specific dates.

For example:

- Q1: January, February, March
- Q2: April, May, June

Q 2 a)

(i) Data Volume and Growth:

What is the estimated volume of inventory and sales data generated daily across all 10 gift shops?

Do you anticipate significant growth in data volume in the future?

Data Retrieval Requirements:

How frequently will the central system need to retrieve data from individual stores?

Are there specific performance requirements for data retrieval?

Data Security and Compliance:

What security measures are crucial for safeguarding sensitive inventory and sales data?

Are there any regulatory or compliance requirements that need to be considered in the storage solution?

(ii)

Data Storage Approach:

For a project of this nature, a suitable data storage approach would be a relational database management system (RDBMS). RDBMS, such as MySQL, PostgreSQL, or Microsoft

SQL Server, allows for structured organization of data, efficient querying, and maintaining relationships between different data entities.

Reason:

Structured Data: Inventory and sales data are highly structured with clear relationships (e.g., product id linked to sales transactions).

ACID Properties: RDBMS ensures data integrity and consistency with ACID properties (Atomicity, Consistency, Isolation, Durability).

Query Performance: RDBMS is optimized for complex queries, essential for extracting valuable insights from sales data.

(iii)

Inclusion of Website Logs and Social Media Content:

No, it wouldn't change the recommendation.

Extendibility of RDBMS: Many modern RDBMS systems support the integration of semi-structured or unstructured data through extensions, allowing the inclusion of website logs and social media content.

Data Integration Platforms: Additional tools or platforms can be employed to process and integrate website logs and social media content seamlessly with the existing RDBMS.

By utilizing an RDBMS and potentially extending it with complementary tools, the client can maintain a centralized and organized data storage system capable of handling both traditional inventory and sales data as well as newer data types from online interactions.

Q 2(b)
(i) Three Examples of Simple Metadata for a Favorite Item of Clothing:

Color: Navy Blue
Material: Cotton
Size: Medium
(ii) Metadata Classification and Explanation:

Color (Descriptive):
Classification: Descriptive
Explanation: Describes a visual characteristic of the item, providing information about its appearance. It aids in quickly identifying the item and is descriptive in nature.

Material (Descriptive):
Classification: Descriptive
Explanation: Describes the fabric from which the clothing item is made. It provides insights into the item's texture, feel, and care instructions, making it a descriptive metadata element.

Size (Structural):
Classification: Structural
Explanation: Specifies a structural attribute of the clothing item, indicating its size. Size is a fundamental characteristic that helps organize and categorize clothing items based on their physical dimensions.

(iii) Using a Standard for Collecting Metadata:

Change in Quality of Metadata:

Consistency: A metadata standard ensures uniformity and consistency in how data is described and structured. For example, if there's a standardized format for recording color, material, and size, it ensures that everyone uses the same terminology and format.
Potential Difficulty with Enforcing a Metadata Standard:

Resistance to Adoption: One potential difficulty is getting all students to consistently adhere to the metadata standard. Students may have different preferences or habits when describing their favorite clothing items, and enforcing a standardized approach might face resistance.

Q 2(c)
Considerations for Special Handling:
Legal Bases for Processing:

Organizations must identify and document a lawful basis for processing personal data. Consent, contractual necessity, legal obligations, and legitimate interests are common legal bases.
Data Subject Rights:

GDPR grants individuals various rights, including the right to access, rectify, erase, and restrict the processing of their personal data. Organizations must facilitate the exercise of these rights.
Data Protection Impact Assessment (DPIA):

Certain processing activities that involve a high risk to individuals' rights and freedoms may require a DPIA to assess and mitigate potential risks.
Data Breach Notification:

GDPR mandates the notification of data breaches to the relevant supervisory authority and, in certain cases, to affected individuals.

Q 3 (i)
As we dont have dataset, for any similar question check for below errors:
Formatting Issues:

- Formatting Is Irregular between Different Tables/Columns
- Extra Whitespace (select for “ABC” or “ABC ”)
- Irregular Capitalization
- Inconsistent Delimiters (commas or tabs)
- Irregular NULL Format (“” or NULL or NA)
- Invalid Characters (ascii or unicode)
- Weird or Incompatible Datetimes (Day-Month or Month-Day, timezones)
- Operating System Incompatibilities (eg, ‘\n’ or ‘\r\n’)
- Wrong Software Versions

Content Issues:

- Duplicate Entries
- Multiple Entries for a Single Entity
- Missing Entries
- NULLs (what do they mean?)
- Huge Outliers
- Out-of-Date Data
- Artificial Entries (eg, €999,999 salary)
- Irregular Spacings (eg, gaps in time series)
- Incorrect or inconsistent units (metres or inches, dollars or euros))

Q3(ii):

Missing Values:

Possible Introduction: During data collection or data entry.
Pipeline Phase: Data Collection or Data Entry.
Assumptions: Assumed that all data would be collected or entered, but due to system failures, human errors, or other issues, some values are missing.
Outliers:

Possible Introduction: Measurement errors, data entry errors, or genuine extreme values.

Pipeline Phase: Data Collection, Data Entry, or Data Transformation.
Assumptions: Extreme values may be introduced due to measurement errors, transcription errors, or represent true outliers in the dataset.
Inconsistent Data Types:

Possible Introduction: Data entry errors or inconsistencies in data formats.
Pipeline Phase: Data Collection or Data Entry.
Assumptions: Assumed a consistent data format, but errors occurred during data entry or collection, leading to variations in data types.
Duplicates:

Possible Introduction: Data entry errors, system glitches, or unintentional replication during data merging.
Pipeline Phase: Data Entry, Data Transformation, or Data Merging.
Assumptions: Duplicate entries may be introduced due to errors during data entry, processing, or merging.

Q 3 (c) :
Data Collection Phase:

Method: Implement rigorous validation checks during data collection.
Reason: This helps catch errors and inconsistencies at the source, preventing them from entering the dataset. Validation checks can include range checks, format checks, and consistency checks to ensure data quality from the outset.

Data Entry Phase:

Method: Use data entry controls and validation rules.
Reason: By implementing controls and rules during data entry, you reduce the likelihood of typos, missing values, or inconsistent data types. This ensures that the data entered adheres to predefined standards.

Data Transformation Phase:

Method: Apply outlier detection methods and data cleaning techniques.
Reason: Outlier detection methods, such as z-scores or interquartile range (IQR), can help identify and handle extreme values. Cleaning techniques, such as imputation for missing values or standardization of data types, contribute to a more uniform and accurate dataset.
Data Merging and Integration Phase:

Method: Use unique identifiers and conduct thorough data reconciliation.
Reason: Unique identifiers help ensure that records from different sources are correctly matched during data merging. Data reconciliation involves checking for consistency between datasets to identify and resolve discrepancies.
Overall Quality Assurance:

Method: Establish a data quality framework and conduct regular audits.

Reason: A comprehensive data quality framework involves defining standards, conducting audits, and implementing processes for continuous improvement. Regular audits help identify and address issues proactively, ensuring sustained data quality.
Documentation and Metadata:

Method: Document assumptions, transformations, and cleaning processes. Maintain comprehensive metadata.
Reason: Transparent documentation allows others (and your future self) to understand the data processing steps, assumptions, and decisions made during the analysis. It contributes to the reproducibility of results and facilitates collaboration.

4(a):

Gathering:

3. Activity: Liaising with DCU Registry to get datasets from the student registration and results systems.

Tool/Application: SQL for querying databases or API requests to retrieve relevant datasets.

7.Activity: Conducting student surveys to answer the key questions about their experience.

Tool/Application: Online survey tools like Qualtrics, Google Forms, or SurveyMonkey.

Processing:

2. Activity: Removing incorrect entries from the student datasets.

Tool/Application: Python with pandas for data cleaning and manipulation

5.Activity: Anonymising student comments that include identifying details.

Tool/Application: Python for text processing and redaction, or tools like OpenRefine.

Analysing:

4.Activity: Calculating the average satisfaction levels based on the sentiment ratings.

Tool/Application: Python with pandas or Excel for data analysis and calculation of average satisfaction levels.

6.Activity: Converting student words into sentiment ratings and correlating with the field of study.

Tool/Application: Natural Language Processing (NLP) libraries in Python (such as NLTK or spaCy) for sentiment analysis, and statistical tools for correlation analysis

Presenting:

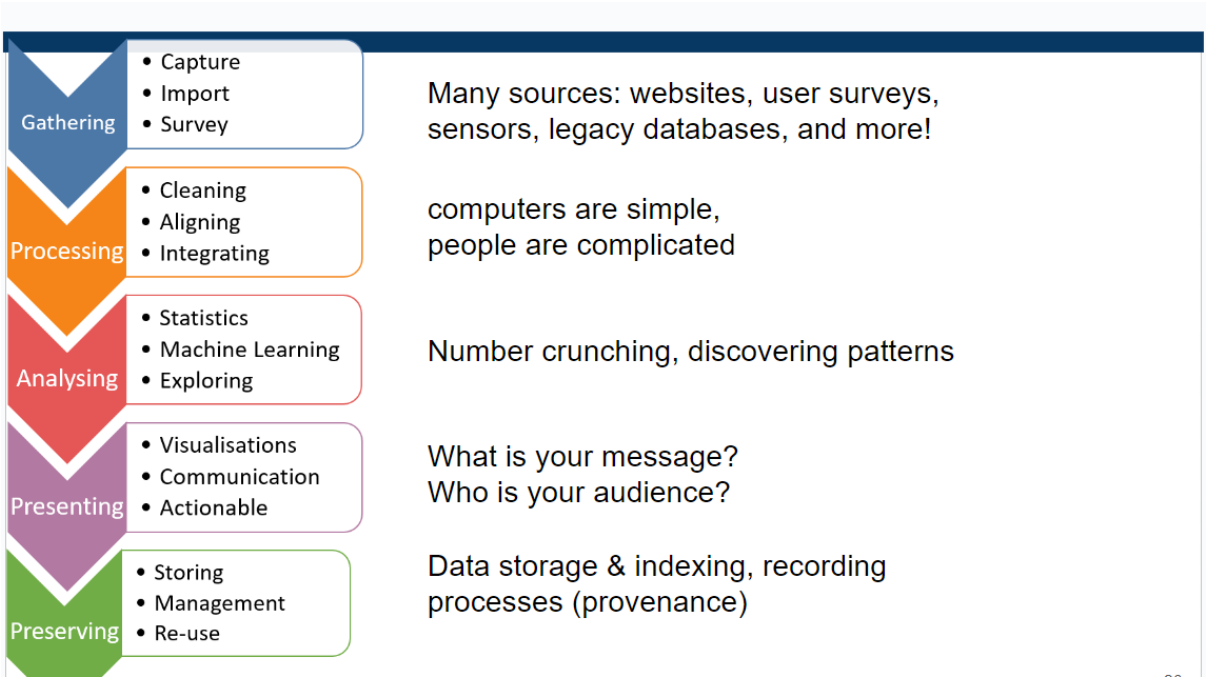
8.Activity: Creating a document to share with senior university management summarising the findings.

Tool/Application: Data visualization tools like Tableau, Power BI, or Matplotlib/Seaborn in Python for creating visualizations, and Microsoft Word or Google Docs for document creation.

Preserving:

1.Activity: Documenting the data formats used in the study and saving all of the created datasets.

Tool/Application: Markdown or a documentation tool (e.g., Jupyter Notebooks) to document data formats, and saving datasets in a secure and organized storage system (e.g., databases, cloud storage).



4(a)2.

One of the main weaknesses of the Generic Data Analytics Pipeline is that it does not include a specific stage for interpreting and communicating findings. While the pipeline does touch on presenting findings, it doesn't explicitly address the interpretation and explanation of results to stakeholders or the broader audience. This can lead to misinterpretations and a lack of understanding of the data's implications.

To address this weakness, the pipeline could be expanded to include a dedicated stage for interpretation and communication. This stage could involve the following steps:

Data Summarization: Clearly summarize the key findings and insights from the analysis.

Visualization: Create clear and informative visualizations to illustrate the findings.

Storytelling: Craft a compelling narrative that connects the data to the business problem or research question.

Targeted Communication: Tailor the communication of findings to the specific audience, considering their level of technical expertise and context.

Actionable Recommendations: Provide actionable insights and recommendations based on the analysis.

By incorporating these steps, the Generic Data Analytics Pipeline would provide a more comprehensive and effective approach to data analysis, ensuring that the insights gained from the data are not only extracted but also effectively communicated and acted upon.

Additionally, the pipeline could benefit from more emphasis on data ethics and privacy considerations. This includes ensuring data privacy compliance, protecting sensitive information, and being transparent about data handling practices.

By incorporating these important aspects, the Generic Data Analytics Pipeline would become a more robust and responsible framework for data analysis.

4 (b):
Data Collection PA. Rating of temperature comfort in offices (cold, cool, perfect, warm, hot):

Qualitative (because it represents categories or labels).
Ordinal (because there is a meaningful order or ranking among the categories, but the differences between them may not be uniform).
B. Number of times a character’s name is used in a TV show episode:

Quantitative (because it represents a numerical count).
Discrete (because the count of times is a whole number and cannot be a fraction).
C. Names of pets owned by all CA682 students:

Qualitative (because it represents categories or labels).
Nominal (because there is no inherent order or ranking among the pet names).
D. All winning times (in seconds) for men’s 100m sprint at the Olympic Games:

Quantitative (because it represents numerical measurements).

Continuous (because the winning times can take any value within a range, and there are infinitely many possible values between any two times).
Ratio (because there is a true zero point, i.e., a time of 0 seconds is meaningful, and ratios of times are meaningful).
Summary:

A: Qualitative, Ordinal
B: Quantitative, Discrete
C: Qualitative, Nominal
D: Quantitative, Continuous, Ratio

4 C.
"An individual’s step count data for a 1 year period from a personal smart device (e.g., a Fitbit)."

Three classical characteristics of big data:

Volume:

Explanation: Volume refers to the sheer size of the data. In the case of an individual's step count data for a 1-year period, the volume can be significant, especially if the data is recorded at a high frequency (e.g., per second or minute).
Assumption: If the step count data is collected frequently (e.g., every second or minute), and if it spans a large population, it could exhibit a large volume characteristic, making it a good example of big data in terms of volume.
Velocity:

Explanation: Velocity refers to the speed at which data is generated, processed, and analyzed. In the case of step count data, if it's recorded in real-time or near-real-time, it could exhibit high velocity.
Assumption: If the step count data is continuously recorded and updated in real-time, it satisfies the velocity characteristic, contributing to the definition of big data.
Variety:

Explanation: Variety refers to the diversity of data types and sources. For step count data, the variety might include timestamped step counts, heart rate data, GPS locations, and more.
Assumption: If the step count data includes various types of information, such as timestamps, heart rate, or location data, it exhibits the variety characteristic, contributing to the big data definition.
Overall Assessment:
The scenario of an individual's step count data for a 1-year period from a personal smart device is a good example of big data. It aligns with the classical characteristics of big data in terms of volume (large dataset size), velocity (real-time or near-real-time updates), and variety (multiple types of data). The assumption is that the data is collected frequently, covers a significant time period, and includes diverse types of information related to the individual's physical activity.

It's important to note that big data is not solely defined by volume, velocity, and variety; other characteristics such as veracity and value may also play a role in determining whether a dataset truly qualifies as big data for a particular use case.

Customer account, purchasing data, and engagement data from a supermarket chain's loyalty card program:

Volume:

Explanation: The volume of data in this scenario can be substantial, especially if the supermarket chain has a large customer base, and the loyalty card program records extensive details about customer transactions and engagement.
Assumption: If the loyalty card program captures data for a large number of customers and includes detailed transaction and engagement information, it can be considered as having a high volume, aligning with the big data characteristic.
Velocity:

Explanation: The velocity could be high if the loyalty card program records customer transactions in real-time or near-real-time.
Assumption: If the data is updated rapidly as customers make purchases or engage with the loyalty program, it satisfies the velocity characteristic of big data.
Variety:

Explanation: Variety refers to the diversity of data types, and in this scenario, it can include customer profiles, purchasing history, engagement metrics, and more.
Assumption: If the loyalty card program collects and stores diverse types of data, it satisfies the variety characteristic of big data.
Overall Assessment:
The supermarket loyalty card program's data appears to exhibit characteristics of big data, considering its potentially large volume, high velocity in capturing transactions, and variety in terms of the types of customer-related information collected.

All 8 episodes (video files) of the TV show "Stranger Things":

Volume:

Explanation: The volume in this case is relatively low compared to scenarios dealing with vast amounts of data. Storing 8 video files, even if they are large, might not constitute a "big" data volume.
Assumption: Assuming the size of the video files is reasonable and manageable, the volume characteristic might not align with the traditional definition of big data.
Velocity:

Explanation: Velocity refers to the speed at which data is generated, processed, and analyzed. In the case of video files, the velocity is generally not high because the content is static.

Assumption: Assuming that the TV show episodes are pre-recorded and not updated in real-time, the velocity characteristic may not align with the traditional definition of big data.
Variety:

Explanation: Variety involves the diversity of data types. In this scenario, the data type is video content.
Assumption: While there is variety in terms of content, it might not exhibit the same level of variety as datasets with multiple types of structured and unstructured data.
Overall Assessment:
The TV show "Stranger Things" episodes, while representing digital content, may not align with the classical characteristics of big data. The volume is relatively low, velocity is not a significant factor, and the variety is limited compared to datasets with diverse data types.

In both assessments, it's essential to consider the context and specific characteristics of the data in question. The definition of "big data" can vary based on the industry, use case, and technological context.

Q 5 a.

Here are three possible improvements that could be made to the pie chart you sent:

1. Remove the legend. Legends can be cluttered and distracting, especially in pie charts with many slices. In this case, the slice labels are large enough and clear enough that a legend is not necessary. Removing the legend would give the chart a cleaner and more modern look.

2. Use more contrasting colors. The current colors are all muted and similar in value, which makes it difficult to distinguish between the slices. Using more contrasting colors would make the chart more visually appealing and easier to read. For example, the largest slice (Health) could be colored red, while the smallest slice (Remaining Ministerial Vote Groups) could be colored blue.

3. Sort the slices by size. The current order of the slices is random, which makes it difficult to compare their sizes. Sorting the slices by size from largest to smallest would make the chart easier to read and would allow the viewer to quickly see which categories are the most significant.

Design rules and theories to support these improvements:

Simplicity: Pie charts should be as simple as possible, with clear and concise labels and a limited number of colors. Removing the legend and using more contrasting colors would both help to simplify the chart.
Clarity: Pie charts should be easy to read and understand. Sorting the slices by size would make it easier for the viewer to compare the sizes of the slices and to identify the most significant categories.
Visual appeal: Pie charts should be visually appealing and engaging. Using more contrasting colors would make the chart more visually appealing, while removing the legend and sorting the slices by size would make the chart more readable and informative.

In addition to the above improvements, it would also be helpful to add a title to the chart and to label the pie slices with percentages, rather than just absolute values. This would make the chart more informative and easier to interpret.

Overall, these simple improvements would make the pie chart more visually appealing, easier to read, and more informative.

Q 5 (b)
A. Compare the performance of stocks in Microsoft, Apple, and Samsung over the last 5 years:

Appropriate Graph Type: Line Chart (Time Series Chart)
CHRTS Category: Time Series
Justification: A line chart is effective for showing trends and changes over time. In this case, you can plot the stock prices of Microsoft, Apple, and Samsung on the y-axis against the time (last 5 years) on the x-axis. Each stock would have its own line, allowing easy comparison of their performance trends. This type of chart helps visualize the overall trajectory of stock prices and identify patterns or changes over the selected time period.

B. Explore movie commercial performance for the IMDB top 50 by director based on cost to make and ticket sales:

Appropriate Graph Type: Scatter Plot
CHRTS Category: Distribution
Justification: A scatter plot is suitable for exploring the relationship between two continuous variables, such as cost to make and ticket sales. Each point on the plot represents a movie, with the x-axis representing the cost to make and the y-axis representing ticket sales. Different directors can be represented by different colors or shapes. This type of chart helps identify patterns, correlations, or outliers in the data. It's particularly useful for understanding the distribution of movies based on their commercial performance and how different directors compare in terms of cost and revenue.

Q 5 C:
(i) What is the main communication purpose and why?

The main communication purpose of the graphic is to reach decision makers. This is because decision makers are people who are concerned about their health and want to know more about their health. The graphic highlights the following key points:

81% of adults are searching for health care topics online
The top searched medical terms are diabetes, breast cancer, shingles, heart disease, and gallbladder
84% of women research health topics online
30% of healthcare consumers consider a hospital with an active online presence to be more "cutting edge"
These points are all relevant to decision makers who are interested in learning more about how to market their healthcare services online.

(ii) **What design choices or guidelines have been used to support this purpose?**

The following design choices and guidelines have been used to support the communication purpose:

Specific and shareable content: The graphic is specific to healthcare marketing, and it is designed to be easily shared on social media and other online platforms. This is likely to appeal to decision makers who are looking for information on how to market their healthcare services online.
Use of data and statistics: The graphic uses data and statistics to support its claims. This makes the graphic more credible and persuasive to decision makers.
Visual appeal: The graphic is visually appealing and easy to read. The use of bright colors, simple fonts, and clear labels makes the graphic easy to understand and remember.
Overall, the graphic is well-designed and effective in communicating its message to decision makers.

In addition to the above, the following design choices also support the communication purpose:

The graphic is prominently titled "3 Keys to Marketing Health Care Online." This clearly communicates the main purpose of the graphic to the viewer.
The graphic uses a simple pie chart to illustrate the top searched medical terms. This is a clear and concise way to present this data, and it makes it easy for the viewer to see which medical terms are the most popular.
The graphic uses a variety of colors and fonts to make it visually appealing and engaging. The colors are bright and eye-catching, and the fonts are easy to read.
The graphic includes a call to action at the bottom. This encourages the viewer to learn more about healthcare marketing by visiting the Fluency Media website.
Overall, the graphic is well-designed and effective in communicating its message to decision makers. It is clear, concise, visually appealing, and includes a call to action.