Visualization Analysis

Assignment 2

Due: Friday 10 March 2023

Name : Mohith Krishna Behata

Student Id: 12589089

Reference Material:

- (All) VAD Chapters 1 & 2
- Lecture Notes
- Visualization Videos

All questions are worth 5 points.

Problem 1:

In the R package {datasets} there is a data set called "mtcars". Using the vocabulary and ideas of Chapter 2 of VAD, what type dataset is it? What are its items? What are the attributes of each item? What type is each attribute?

The "mtcars" dataset in R's {datasets} package is a tabular dataset. Tabular dataset means it consists of rows and columns where each row represent an observation or item and each column represents an attribute or variable. The items in the dataset are cars, and there are 32 different types of cars in the dataset and the attributes are various characteristics of each car.

The various characteristics of car are,

- mpg: Miles per gallon
- cyl: Number of cylinders in the engine
- disp: Displacement in cubic inches
- hp: horsepower
- drat: Rear axle ratio
- wt: Weight in thousands of pounds of the car
- qsec: 1/4 mile time in seconds
- vs: Engine type (0 = V-shaped, 1 = straight)
- am: Transmission type (0 = automatic, 1 = manual)
- gear: Number of forward gears
- carb: Number of carburetors

All attributes of each item in the dataset are numerical values with some attributes being continuous the example for these continuous attributes are “cyl”, “vs”, “am” and “gear”. But some attributes such as “vs”, “am”, and “gear” attributes can also be considered categorical since they only take a limited number of possible values. The attribute “model” represents the name of each car which can be considered as string variable.

The "mtcars" dataset is a dataset that contains 32 rows (one for each car) and 11 columns (one for each attribute) including 1 categorical and 10 being numerical attributes (both continuous and discrete).

Problem 2:

In your own life, build a simple table. Pick something of which you have several items (DVD, coffee cups, etc)? Identify 5 attributes and their associated type. Collect some data and give me the dataset (a table with at least 5 rows and 5 columns).

Game Name	Developer Name	Release Year	Genre	Rating
The Legend of Zelda: Breath of the Wild	Nintendo	2017	Action-adventure, Fantasy	9.5
Grand Theft Auto V	Rockstar North	2013	Action-adventure, Open-World	9.0
Half-Life 2	Valve Corporation	2004	First-Person Shooter, Science Fiction	9.3
Red Dead Redemption 2	Rockstar Games	2018	Action, Adventure, Western	9.8
The Last of Us	Naughty Dog	2013	Action, Adventure, Horror	9.3
God of War	Santa Monica Studio	2018	Action-adventure	9.5
Portal 2	Valve Corporation	2011	Puzzle-platformer	9.5
The Witcher 3: Wild Hunt	CD Projekt Red	2015	Action RPG	9.3

The Table above has items which are some of the games I have played in the past and still continuing to play and the table has attributes which are Game Name, Developer Name, Release Year, Genre and Rating. The data collected for each attribute are,

- Game Name: This attribute describes the tittle of the game (ex: “Half-Life 2, Portal 2).

- Developer Name: This attribute that describes the Name of the Company or Organization who has released the game (ex: “Naughty Dog, “Valve Corporation”).
- Release Year: This attribute describes the year in which the game was released (ex: 2004,2015)
- Genre: This attribute describes the genre which the game belongs to (ex: Action-adventure, Action RPG).
- Rating: This attribute describes the rating of the game (ex: 9.5,9.3).

The type of each attribute is,

- Game Name: Nominal.
- Developer Name: Nominal.
- Release year: Continuous.
- Genre: Categorical.
- Rating: Continuous.

Problem 3:

Explain “salience” in a visualization context. Discuss the salience of each of the visualizations given below:

In Visualization, Salience refers to the relative importance or prominence of visual elements with in a graphic. It can also be referred as the degree to which a particular visual element or feature attract the attention and stands out from other elements These features that differentiate them from their neighbors can be anything like colors, sizes, etc. The lesser the amount of salience, the more there are chances for the system to overlook them.

The concept of salience is very important because it can change the perspective on how a viewer perceives or interprets the data that they view. If the data that is being viewed is not prominent enough it can lead to misinterpretation, leading the viewer to make incorrect conclusions. There can also be a scenario where certain elements can become too dominant, they may distract the viewer from viewing or understanding the main message the visualization carries.

The one who creates visualization can use various visual prompts such as colors, sizes, shapes etc. to control the salience of different elements in the visualizations. Example on how salience can be controlled in a visualization are using bold colors or Bold fonts can make them more prominent. The goal of the visualization should be to create a balance to make important information standout and avoid visual clutter which makes the viewer confused.

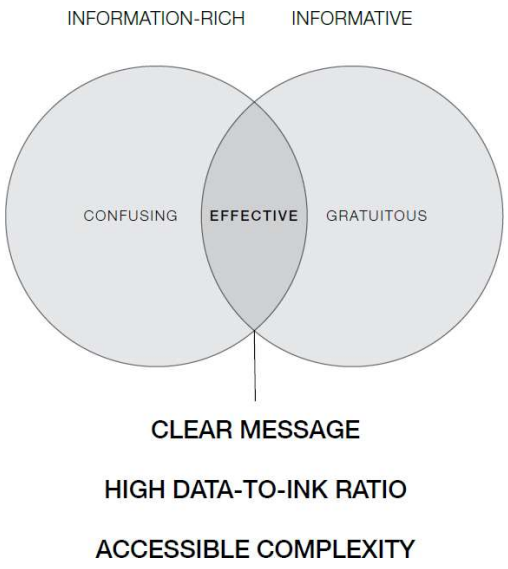
As we can see in the first image, the dark point here clearly distinguishes from the other light points which indicates the importance of dark point from the other points. The dark point clearly draws the attention of the viewer in the first image. The other salient items can be the attached circles if we consider proximity as a factor. Overall, I can say that the image has high salience.

From the Second Image, here in the image there are lot of points and all of them are dark in color but one aspect that can be considered to differentiate this points from each other are by their radius. The other salient items here can be the circles that are closer/attached to each other (3 such cases in the image) making them stand out compared to the rest of their neighbors. Overall, it has low salience.

Finally, from the third image, the lines are distinguished based on length and thickness'. The most salient item is the one bar/line which has a longer length and thickness compared to the rest of the neighbors, making it unique and standing out from the rest. Few of the bars/lines are thicker than the rest and some are having more length compared to other. Overall, it has low salience.

Problem 4:

Using complete sentences, explain the meaning of each word or phrase as applied to visualization. Interpret the diagram.



Meaning of each word or phrase applied in the visualization are:

INFORMATION-RICH: An information-rich visualization contains a large amount of data or information.

INFORMATIVE: An informative visualization effectively communicates that data to the viewer.

EFFECTIVE: Successful in attaining a desired or expected results.

CLEAR MESSAGE: An effective visualization, therefore, would strike a balance between these two concepts - it would contain enough information to be meaningful and valuable, while also presenting that information in a clear and understandable way.

CONFUSING (Information-rich but not informative): These types of visualizations may contain a large amount of data or information but are not effectively communicating that information to the viewer. They

may be overwhelming, confusing, or difficult to understand.

GRATUITOUS (Informative but not information-rich): These types of visualizations effectively communicate information to the viewer but may not contain a significant amount of data or information. They may be simple or straightforward, but still convey a meaningful message or insight.

HIGH DATA-TO-INK RATIO: Means that the visualization effectively communicates the data without using unnecessary ink or graphical elements. This results in a clean, easy-to-read visualization that focuses the viewer's attention on the important data points.

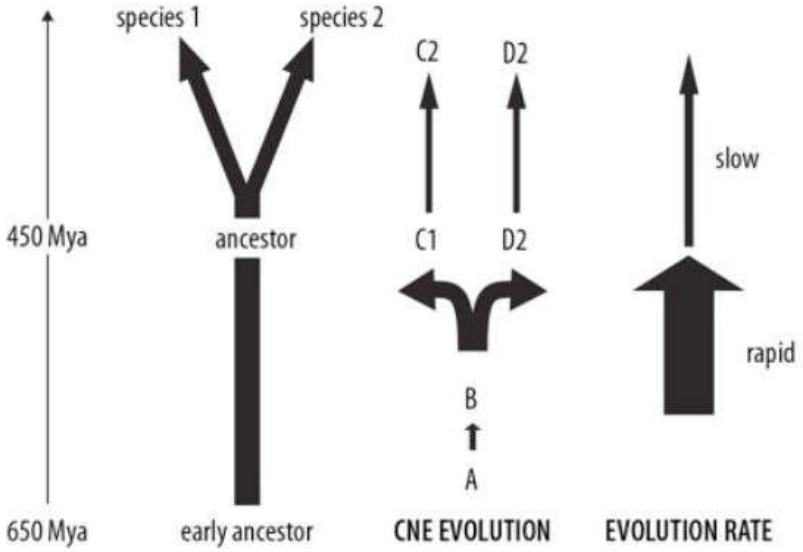
ACCESSIBLE COMPLEXITY: It's is a term used in data visualization to describe the concept of presenting complex information in a way that is easy to understand and accessible to a broad audience.

Now, about the interpretation of the diagram, it is said that visualization can be very confusing if it has a lot of information. This means that there can be a lot of additional content that is not informative. The main goal here is to have enough information to convey the actual message. This makes it really effective. The main motive here is to ensure that there is the right amount of data, enough to build the desired visual image that the viewer can understand.

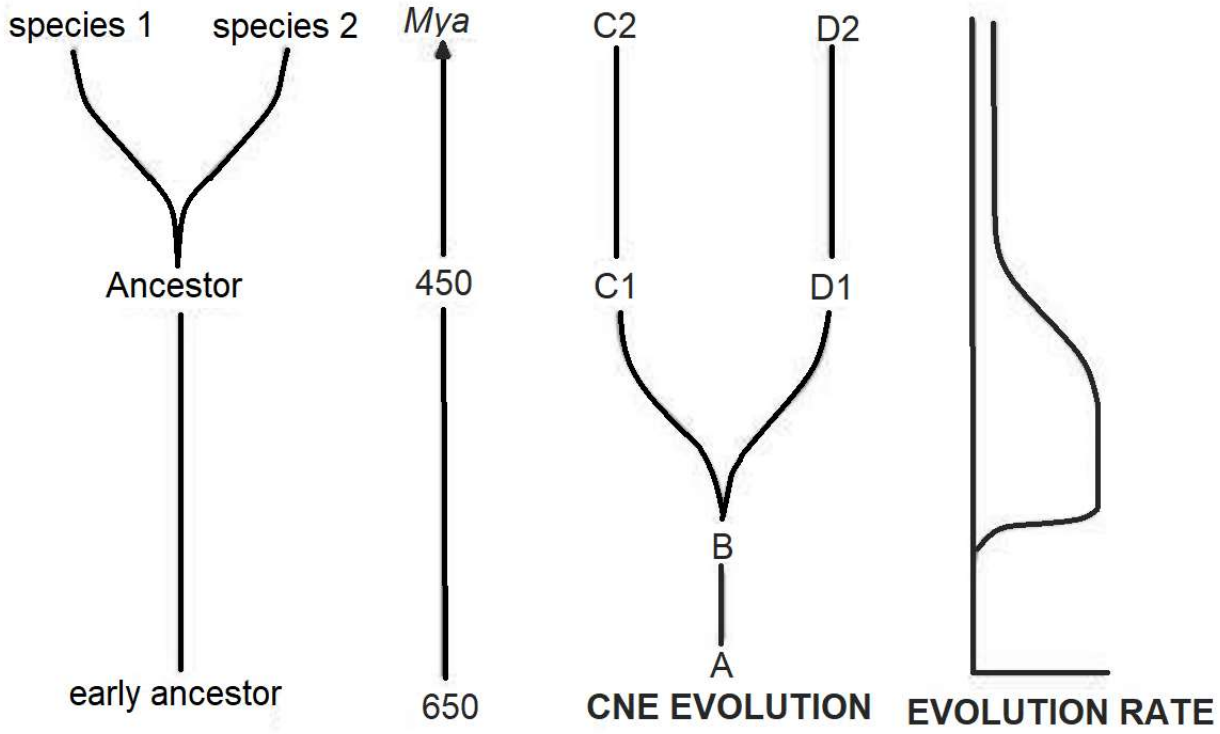
Problem 5:

The diagram given below is from the paper: "McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, et al. (2006) Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Res 16: 451-465."

Redraw it and explain why you made each change you did.



Redrawn Images:



In the given image, only the first picture representing the time period in Mya (Million years ago) requires arrows since it represents only the directionality or progression over time.

Evolution rate visualization does not provide meaningful information based on the width of the arrow.

It provides no meaningful information about evolution rate based on the width of the arrows.

Additionally, since we are using the label, there is no need to change the arrow's width to convey the rate of evolution. To illustrate the slow evolution rate at 450 Mya, instead of using labels, we use a vertical line graph.

The new visualization will have a high data-to-ink ratio compared to the original.

Problem 6:

Give the book's definition of a visualization and compare it to the other definition I gave you in lecture. Compare and contrast.

Definition-1: (Book's Definition)

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Definition-2: (Lecture Definition)

There is nothing better than a picture for making you think of questions you forgot to ask.

Definition-3 (Lecture Definition)

Information visualization is a compact graphical presentation and user interface for manipulating large numbers of items possibly extracted from far larger datasets. Enables users to make discoveries, decisions, or Explanations about patterns (trend, cluster, gap, outlier...), groups of items, or individual items.

The goal of information visualization is to be able to manipulate large amounts of data, possibly from a much larger dataset, with a compact graphical presentation and user interface. Makes it possible for users to make discoveries, decisions, or explanations about patterns (trends, clusters, gaps, outliers...), groups of items, or individuals.

Each definition relates to the concept of information visualization, which involves using visual representations to understand data. Visualization can also help people perform tasks more effectively or gain insights from data, according to all of them.

Each definition has a different focus and scope, however in the first definition Computer-based visualization systems and their purpose are defined in the first definition, whereas visual representations can inspire curiosity and lead to questions in the second one Information visualization is further defined as a compact graphical presentation that uses a user interface designed to manipulate large datasets.

Unlike the other two definitions, the third emphasizes the potential uses of information visualization, such as finding patterns, making decisions, or explaining individual items.

As a result, while all three definitions pertain to information visualization, they differ in focus, scope, and level of detail.

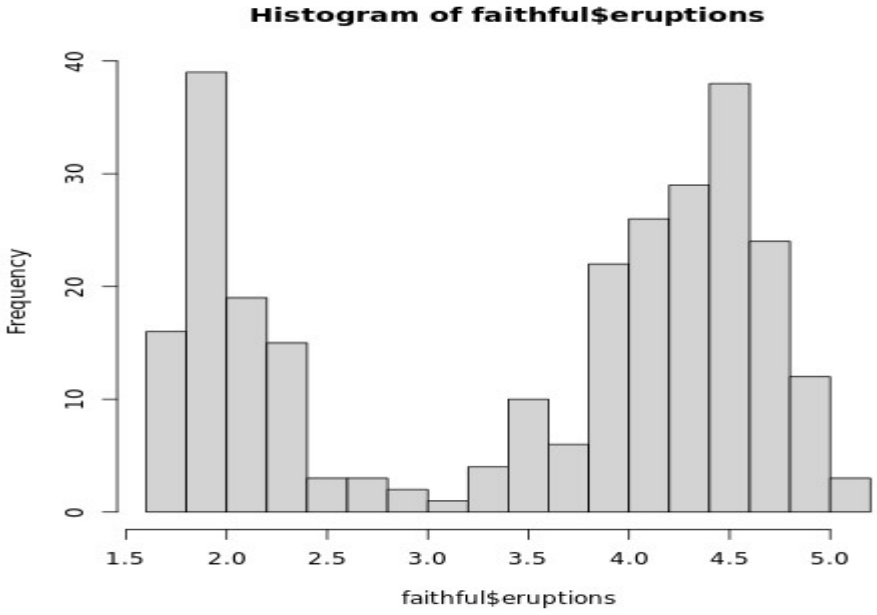
Problem 7:

Using R give an example that uses datasets to show mere statistical analysis of data can be deceptive whereas visualization quickly reveals all is not what it seems to be. (Hint: you can find this data in R with a solution provided you use the right search terms. Rely on the internet as the help feature in the current release of RStudio does not work.)

Here I have taken the data set "faithful", which has the data on the old faithful geyser in Yellow stone national park.
Firstly I looked at the summary of the eruptions using the R code
summary(faithful$eruptions*)* which gives us an output with data as follows
Min-1.60
1st Qu- 2.163
Median-4.000
Mean-3.488
3rd Qu- 4.454
Max-5.100

This means that the avg duration of eruption is close to 3 and a half minutes, with the min to max duration ranging from 1.6 to 5.1.

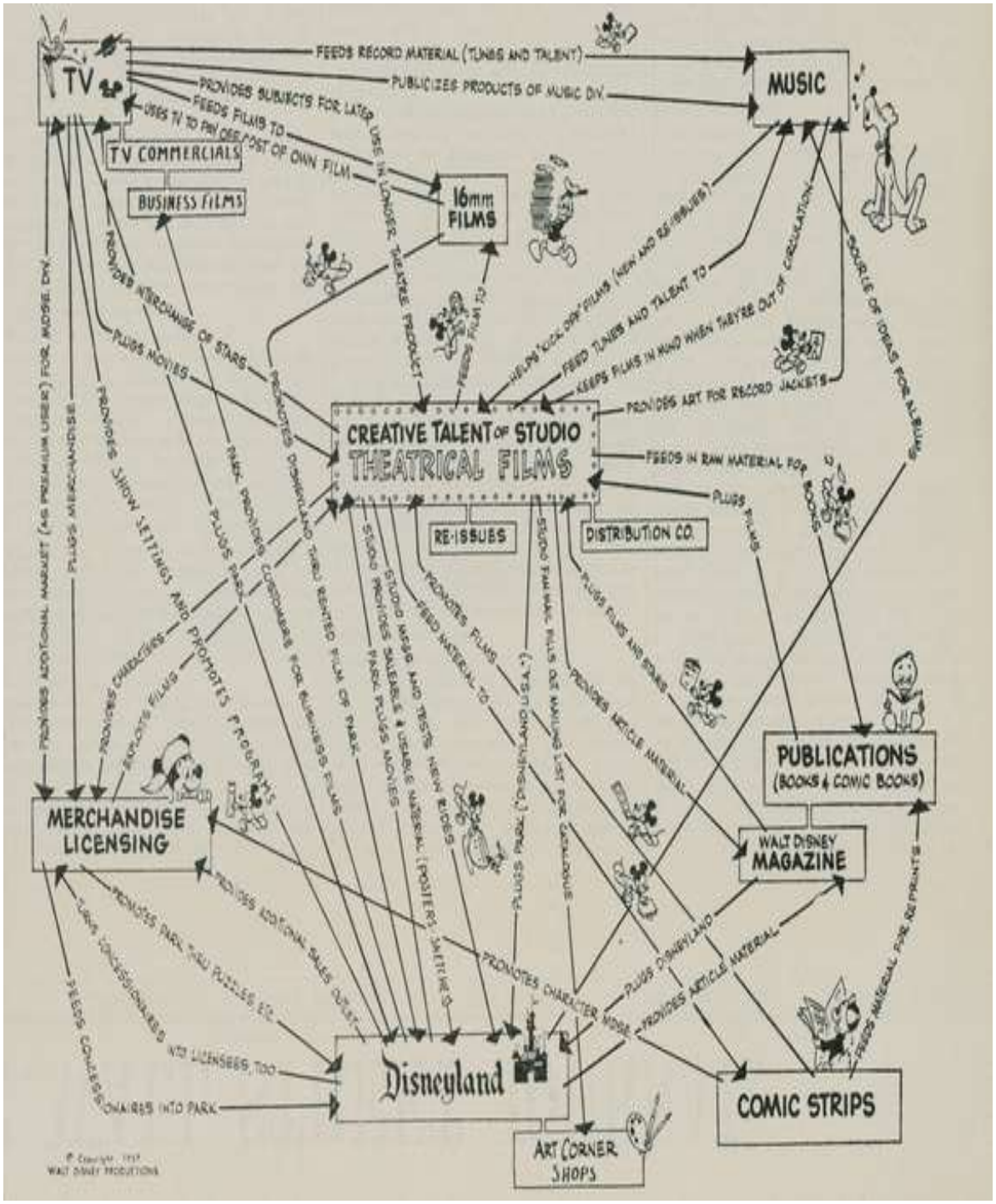
Now, I plotted a histogram for the same data. The R code for this is:
hist(faithful$eruptions, breaks=20)
The histogram is as follows:



Here we can clearly see that a group of eruptions are ranging from 1.5 to 2.5 and another one ranges from 3.5 to 4.5. We would have clearly missed out on this pattern if we had stuck with mere statistical data alone.

Problem 8:

The following is a directed graph diagram I found in Harvard Business Review. It was drawn by the Disney Corporation in 1957. By the definitions and concepts, you have stated on the first page, is the following an information visualization? If it is, what type of data does it encode?



That's right, it's an information visualization. Known as Synergy Map, it depicts Disney's core strategy in a napkin sketch.

Each box represents a different segment or division of Walt Disney's entertainment empire. Labels on each box identify the segment or division it represents, for example, "Animation," "Theme Parks," or "Merchandise."

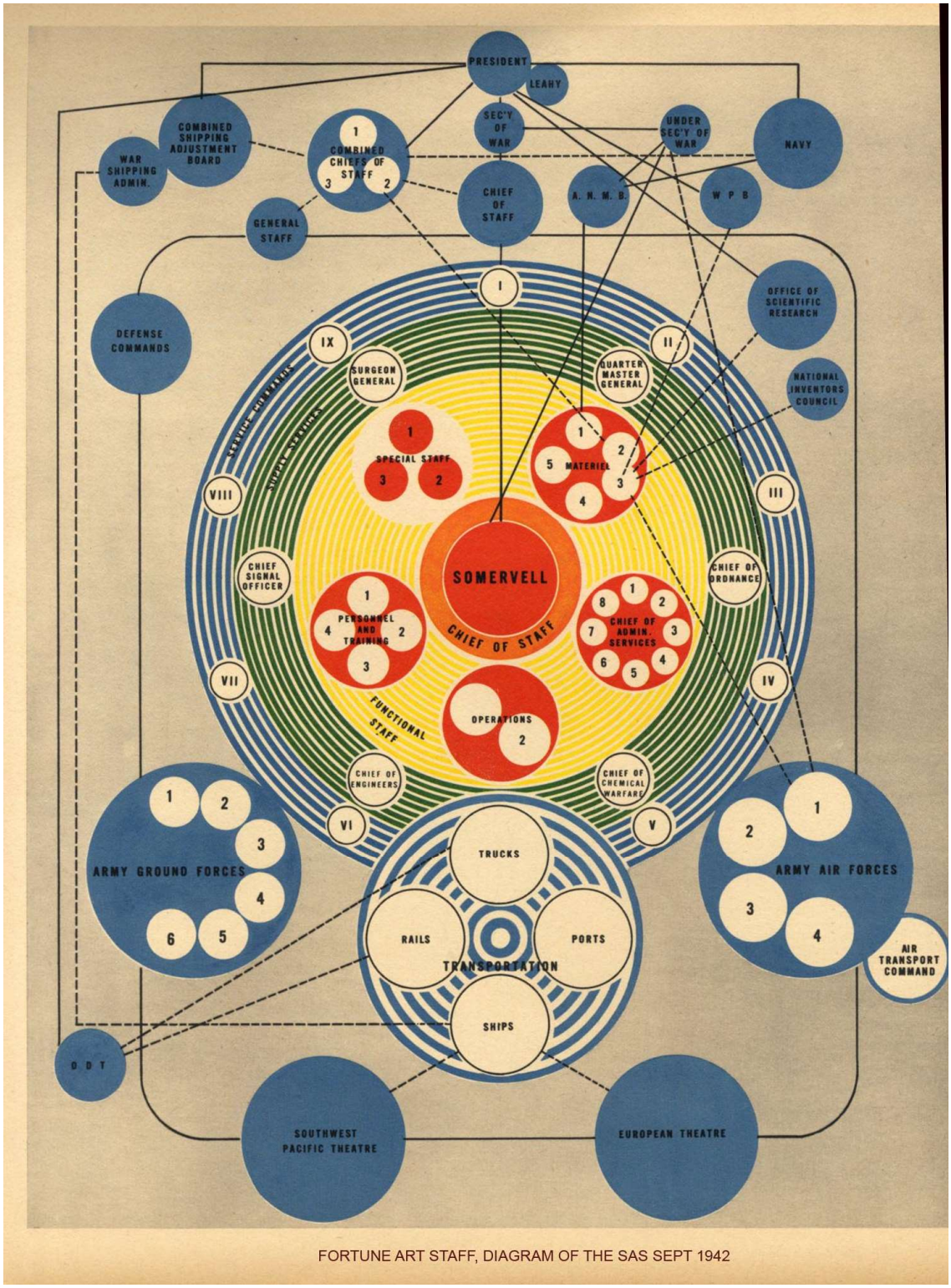
Walt Disney's entertainment empire is represented by arrows that illustrate the relationships and dependencies between different parts of Walt Disney's entertainment empire. Synergies occur when one segment of the business supports or enhances another. For example if we see Disney's animated films can drive attendance at its theme parks, which feature popular characters and attractions based on the films, as shown by the arrow between "Animation" and "Theme Parks".

An arrow from "Theme Parks" to "Merchandise" illustrates how popular attractions at theme parks can drive sales of merchandise featuring those characters.

Disney was able to communicate his vision in an understandable and memorable way through this visualization. Businesses can better communicate complex ideas and strategies by using visual aids, including charts, graphs, and diagrams, which leads to better decision-making and improved results.

Problem 9:

The following is a picture of the SAS from 1942. It was drawn by the Fortune magazine Art Staff in September 1942. By the definitions and concepts, you have stated on the first page, is the following an information visualization? If it is, what type of data does it encode?



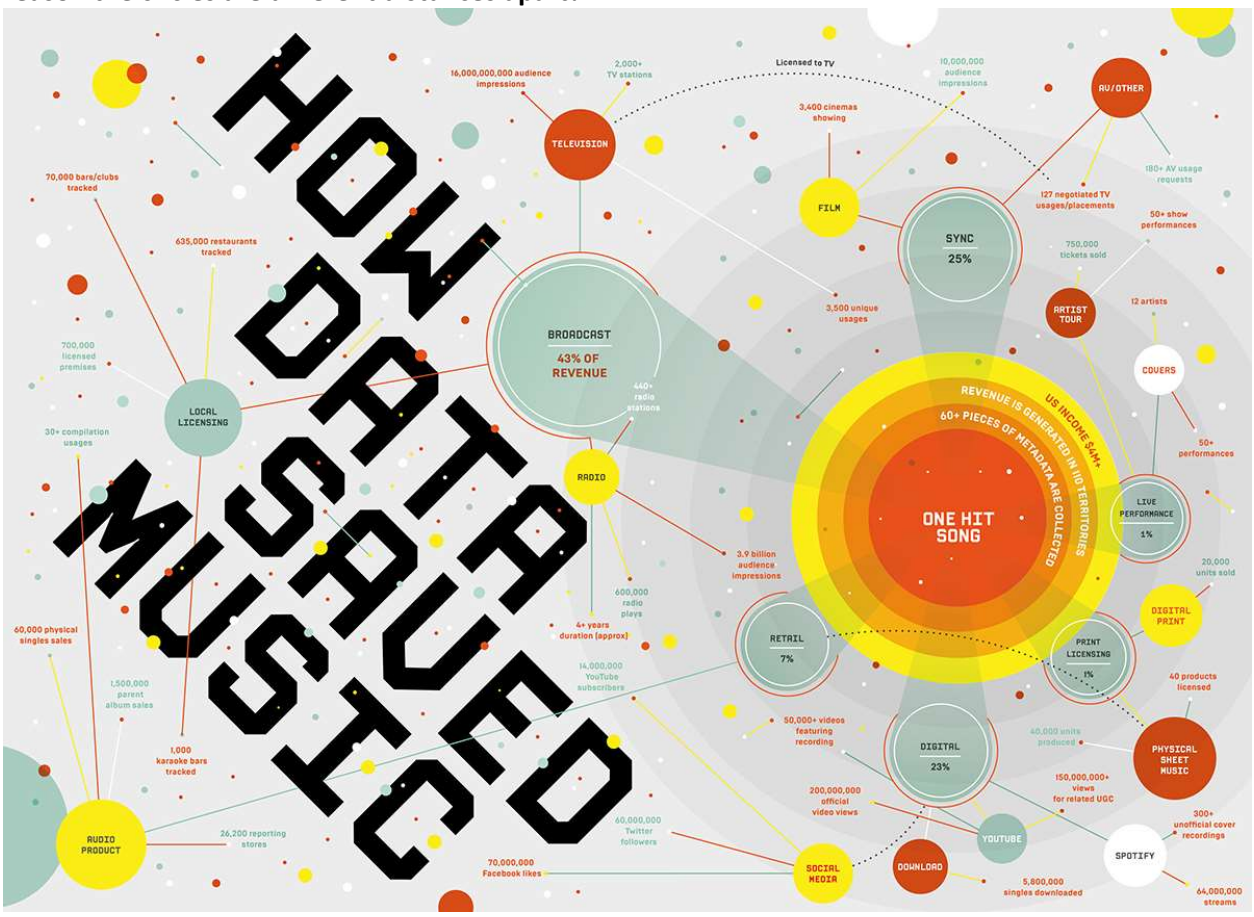
FORTUNE ART STAFF, DIAGRAM OF THE SAS SEPT 1942

Yes, it is information visualization. In the diagram, you can see the hierarchy of officials in the U.S. Army during World War I II There is a hierarchical structure, with the top-ranking official at the center and lower-ranking officials branching out from it. The majority of the visualization provides us with in-depth knowledge about the numerous categories and the connections between each of them, despite the fact that a few numbers are involved. Thus, I believe that the data represented here is categorical data.

The size of the circles and the thickness of the lines connecting them convey information about the number of people at each level of the hierarchy and the level of authority of each official, respectively. The dashed lines in the diagram represent advisory or support relationships rather than direct lines of authority.

Problem 10:

The following is a vis from WIRED UK from 2015. The creators claim it illustrates "*the relative importances of distribution streams and types of musical distribution in terms of generating revenue for record labels and artists*". By the definitions and concepts, you have stated on the first page, is the following an information visualization? This is a busy vis. What are the circles? Why are they different sizes? What are the arcs or edges? Why do the colors change? Is there a reason the circles are different distances apart?



Yes, it is Information Visualization. The above Visualization is designed to provide a clear and visually engaging way to show which revenue streams are the most significant for a hit song and how they contribute to the overall revenue generated by the song.

As we can see in the visualization that circles represent the different revenue streams and sub category in those revenue stream that single hit song earned.

Circles of different sizes represent the percentage of revenue earned by each revenue stream. For example the revenue generated by a song via broadcast that is 43% so it is visualized in a big circle where as the least revenue it generated was 1% which is why it was represented in small circle compared to the rest.

The arc or edges are used to show how different revenue streams can be connected to each other or other relevant items such as record labels or type of media etc.

The different colors are used to distinguish among the sub-categories show in the visualization.

The circles are spaced out to prevent the visualization from becoming too cluttered and make it easier for the viewer to see the connections between different elements.