



# Data Visualization and Interpretation

## Notes:

### **what is the difference between df.describe(include='all') and df.describe()**

The `describe()` method in Pandas is used to generate descriptive statistics of the DataFrame, such as count, mean, standard deviation, minimum, maximum, and quartile values. However, there is a difference between `df.describe()` and `df.describe(include='all')`.

- `df.describe()`: This method by default provides summary statistics only for numerical columns of the DataFrame. It excludes non-numeric columns from the output. It calculates statistics like count, mean, standard deviation, minimum, 25th percentile (first quartile), median (50th percentile), 75th percentile (third quartile), and maximum for each numeric column.
- `df.describe(include='all')`: This method includes summary statistics for both numeric and non-numeric columns. It provides additional statistics for non-numeric columns, such as unique values, the top most frequent value, and the frequency of the top value. This is useful for gaining insights into categorical or textual data.

Here's a brief comparison:

- `df.describe()`: Useful for getting statistical summary of numeric data.
- `df.describe(include='all')`: Useful for getting statistical summary of both numeric and non-numeric (categorical) data.

### **Describe how a scatterplot can be used to show the relationship between sales and profit across different product types and market.**

A scatterplot is a powerful visual tool used to showcase the relationship between two variables. In this case, we want to explore the connection between sales and profit across various product types and markets.

To create such a scatterplot, follow these steps:

1. **Data Collection:** Gather data on sales and profit for different product types across various markets. Each data point should represent a specific product type in a specific market.
2. **Data Representation:** Organize the data into a table with columns for product type, market, sales, and profit.
3. **Plotting:** On the horizontal axis (x-axis), plot the sales figures for each data point. On the vertical axis (y-axis), plot the corresponding profit figures. Each data point represents a unique combination of product type and market.
4. **Interpretation:**

- **Trend Analysis:** Examine the overall pattern formed by the data points. Is there a noticeable trend indicating a positive or negative correlation between sales and profit? For instance, if sales increase as profit decreases, it might suggest aggressive pricing strategies or high production costs.
  - **Outlier Identification:** Look for any data points that deviate significantly from the general trend. These outliers could represent unique market conditions, exceptional products, or data entry errors.
  - **Segment Analysis:** Consider dividing the data into subsets based on product type or market. By color-coding or using different symbols for each subset, you can identify how the relationship between sales and profit varies across different segments.
5. **Insight Generation:** Based on the scatterplot, draw conclusions about the relationship between sales and profit across different product types and markets. Are there any product types or markets where sales and profit are strongly correlated? Are there any unexpected patterns or insights that emerge from the data?
  6. **Communication:** Present your findings clearly and concisely, using the scatterplot to support your analysis. Include relevant observations, insights, and recommendations based on the relationship between sales and profit.

By following these steps and creating a well-constructed scatterplot, you can effectively visualize and analyze the relationship between sales and profit across different product types and markets.

**Explain with an example how time can be treated as continuous and discrete. Plot profit over Months [2011, Jan-2014, Dec, all months' data is present] to point out the difference in visuals because of the difference in treatment.**

Certainly! Time can be treated as both continuous and discrete, depending on the context in which it is being analyzed.

1. **Continuous Time:** When time is treated as continuous, it means that it is viewed as a smooth, uninterrupted flow. In this approach, time is represented by a continuous scale, such as a timeline with no gaps or breaks between consecutive time points. This is often used in scenarios where precise timing or measurement is essential, such as in physics experiments or financial modeling.
2. **Discrete Time:** On the other hand, when time is treated as discrete, it means that it is viewed as occurring in distinct, separate intervals or units. In this approach, time is represented by discrete points or intervals, with specific time periods delineated from one another. This approach is commonly used in areas such as discrete event simulation, where events occur at specific points in time, or in project management, where tasks are scheduled in discrete time increments.

Now, let's illustrate the difference between treating time as continuous and discrete using a plot of profit over months:

Example: Suppose we have monthly profit data from January 2011 to December 2014. We will plot this data twice: once treating time as continuous and once treating it as discrete.

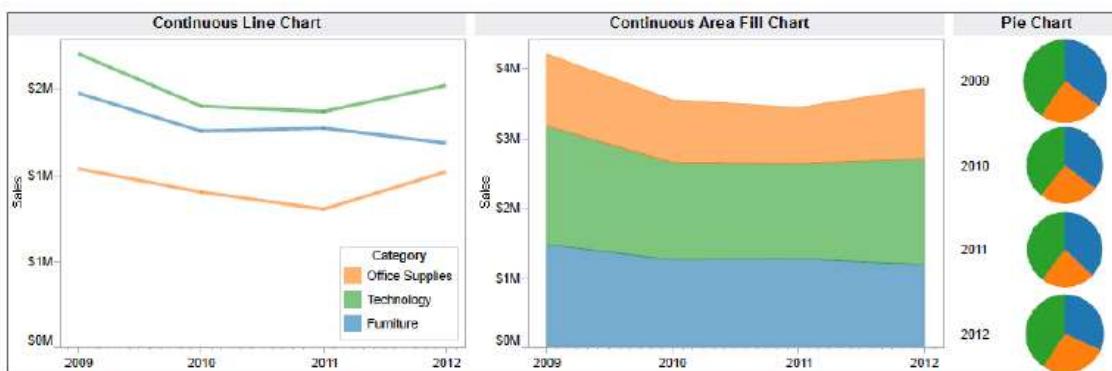
1. **Continuous Time Plot:** In this plot, we will connect each data point smoothly, treating time as a continuous variable. The x-axis will represent the timeline from January 2011 to

December 2014, with each month represented as a point on the axis. The y-axis will represent profit.

2. **Discrete Time Plot:** In this plot, we will treat time as discrete intervals, with each month represented as a separate data point. The x-axis will still represent the timeline from January 2011 to December 2014, but the data points will be plotted distinctly, with no connecting lines between them.

Let's create both plots to visualize the difference in visuals due to the difference in treatment of time.

**All three charts below plot the same data-Sales over years .Which one do you prefer? Why? Also mention the reasons for rejecting others.**



Among the three charts provided, I would prefer the Continuous Line Chart for visualizing sales over the years. Here's why, along with reasons for rejecting the other charts:

### 1. Continuous Line Chart:

- Preference:** The Continuous Line Chart effectively displays the trend of sales over the years for different categories (Office Supplies, Technology, Furniture). It allows for easy comparison between categories and provides a clear visualization of how sales change over time.
- Reasons for Preference:**
  - The use of lines makes it easy to trace the trend of each category's sales over time.
  - The chart provides a clear distinction between categories with different colors.
  - It allows for precise comparison of sales trends between categories.

### 2. Continuous Area Fill Chart:

- Rejection Reason:** While the Continuous Area Fill Chart also displays sales over the years for different categories, it fills the area below the lines, which can sometimes obscure the actual data points. This makes it harder to precisely determine the sales values for each category at specific points in time.

### 3. Pie Chart:

- Rejection Reason:** The Pie Chart is not suitable for visualizing sales over time. Pie charts are better suited for showing the composition of a whole, such as the sales distribution for a single year. Using a Pie Chart for time-series data makes it difficult

to observe trends over time and compare sales between different categories effectively.

In summary, the Continuous Line Chart is preferred for visualizing sales over the years due to its clarity, ease of comparison between categories, and suitability for displaying time-series data.

### **How do you ensure the goodness of fit of a regression model [which is a measure of the quality of the model's predictions] in Tableau? Demonstrate with a visual.**

In Tableau, you can assess the goodness of fit of a regression model, which measures the quality of the model's predictions, by examining the residuals plot. The residuals represent the differences between the observed values and the predicted values from the regression model. A well-fitted model should have residuals that are randomly scattered around zero without any discernible pattern. Here's how you can demonstrate this visually in Tableau:

#### **1. Create a Regression Model:**

- Start by connecting to your data source in Tableau.
- Drag and drop the independent variable (predictor) and dependent variable (response) onto the Rows and Columns shelves, respectively.
- Click on the "Analytics" pane, select "Trend Line," and choose "Linear" or the appropriate regression model.
- Tableau will automatically generate a regression line based on the data.

#### **2. Add Residuals Plot:**

- After creating the regression model, add a new worksheet to your dashboard.
- Drag the independent variable onto the Columns shelf.
- Calculate the residuals by creating a calculated field. Subtract the actual value from the predicted value using the regression model.
- Drag the residuals calculated field onto the Rows shelf.

#### **3. Visualize Residuals:**

- Create a scatter plot with the independent variable on the x-axis and the residuals on the y-axis.
- Add reference lines at  $y=0$  to visualize where residuals should ideally fall.
- Customize the plot to enhance readability by adjusting colors, sizes, and labels.

#### **4. Assess Goodness of Fit:**

- Examine the residuals plot. A well-fitted regression model will have residuals that are randomly scattered around the reference line at  $y=0$ .
- Look for any patterns or trends in the residuals. If you observe a pattern (e.g., a curve or systematic deviation), it indicates that the regression model may not adequately capture the underlying relationship in the data.

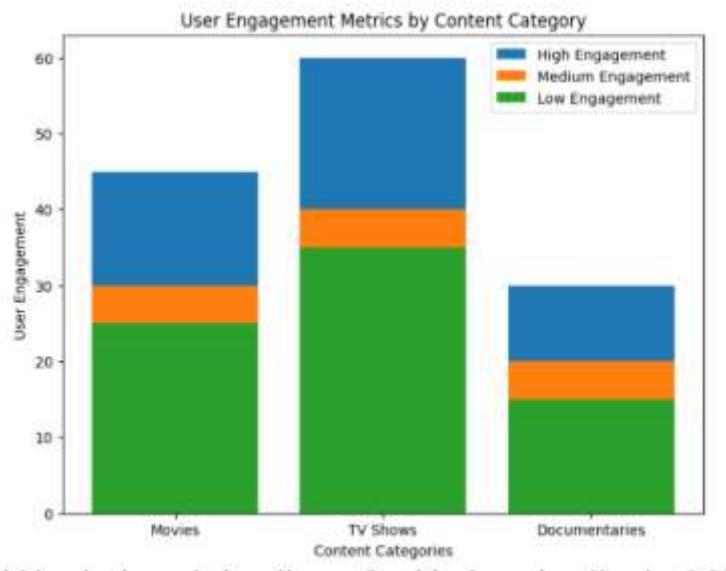
#### **5. Iterate and Improve:**

- If the residuals plot suggests that the model's fit is not satisfactory, consider refining the model by adding additional predictor variables, transforming variables, or trying different regression techniques.
- Continue to iterate and assess the goodness of fit until you achieve a satisfactory model.

By following these steps and visualizing the residuals plot in Tableau, you can effectively evaluate the goodness of fit of a regression model and make informed decisions about its predictive performance.

**Question:**

The visual provided below illustrates the user engagement metrics for different content categories using a stacked bar chart. Each content category, namely Movies, TV Shows, and Documentaries, is divided into three engagement levels: High Engagement, Medium Engagement, and Low Engagement. Answer the following questions regarding the visualization context:

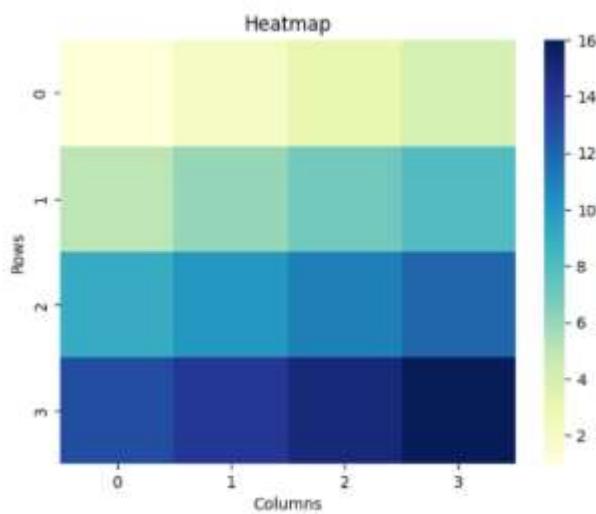


**Answer:**

- A. The intended audience for this data visualization could be professionals in the entertainment industry, like analysts, marketers, or decision-makers. They'd use it to shape their strategies and decisions regarding content creation and distribution.
- B. The stacked bar chart breaks down user engagement metrics for Movies, TV Shows, and Documentaries. Each bar is divided into High, Medium, and Low Engagement segments, showing the percentage of users in each level for each content category.
- C. This visualization effectively shows how engagement levels vary across content categories. It uses colors and clear labeling to make it easy to understand which categories have the highest or lowest engagement levels.
- D. A potential "Big idea" stemming from this visualization could involve adjusting content creation and marketing strategies to boost engagement. For instance, if certain categories consistently have low engagement, creators could rethink their approach to those types of content. Analyzing trends over time could also reveal insights for improving content recommendations and marketing efforts.

**Question:**

**How does the color intensity in the heatmap represent the values of the data? Interpret the heatmap below and discuss any patterns or trends that can be observed based on the color variations.**



**Answer:**

The color intensity in the heatmap represents the values of the data. Typically, darker colors indicate higher values, while lighter colors indicate lower values. In this heatmap, the color scale on the right side indicates the range of values corresponding to each color.

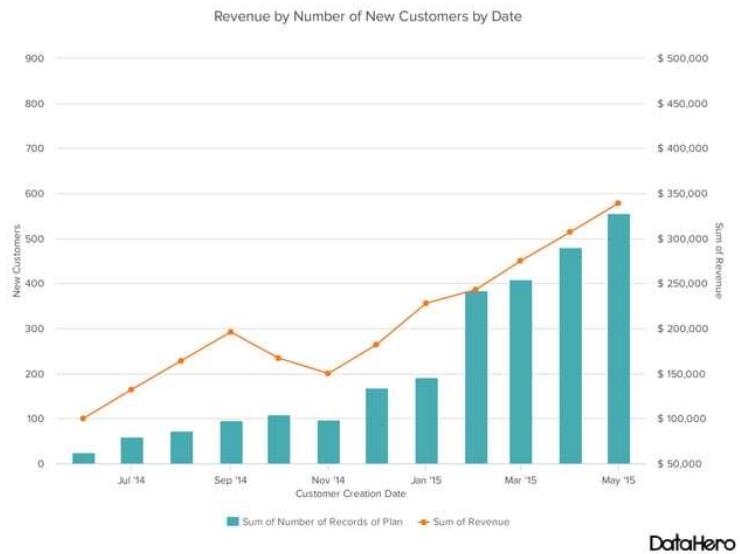
Interpreting the heatmap and observing the color variations, it seems that there are patterns or trends present in the data. Here's what can be inferred:

1. Gradient from Light to Dark: Moving from the bottom-left corner to the top-right corner of the heatmap, there's a gradient of colors transitioning from light to dark. This suggests a trend of increasing values along this diagonal direction.
2. Block of Darker Colors: In the top-left corner of the heatmap, there's a block of darker colors compared to the rest of the heatmap. This indicates a cluster of higher values in that region.
3. Varying Intensity within Rows and Columns: While there's a general trend of increasing values along the diagonal, there are variations in intensity within rows and columns. For example, within the same row or column, some squares are lighter while others are darker, suggesting fluctuations in the values.
4. Symmetrical Pattern: The heatmap appears to be symmetric along the diagonal, with similar color patterns reflected on both sides. This symmetry might indicate a symmetric dataset or a specific relationship between rows and columns.

Overall, the heatmap indicates patterns of data distribution, with specific areas of higher and lower values and trends along certain directions. Further analysis would be needed to understand the underlying factors contributing to these patterns and trends.

### Question:

The given dual axis chart illustrates the revenue and the number of new customers over time for a company. The line graph represents the revenue, while the bar chart represents the number of new customers. Each data point on the chart corresponds to a specific date. Discuss the limitations and drawbacks of using a dual axis chart to visualize the relationship between revenue and the number of new customers. Justify your reasons for avoiding this type of chart (3 marks). Furthermore, propose and create two alternative visuals that could serve as more effective replacements for the dual axis chart, considering the clarity and accuracy of data representation (3 marks).



### Answer:

#### Limitations and Drawbacks of Using a Dual Axis Chart:

1. Difficulty in Interpretation: Dual axis charts can confuse viewers as they require understanding two different scales at once. This complexity can lead to misunderstanding the relationship between the variables.
2. Comparison Issues: It's hard to compare trends or patterns between the two variables because of the separate scales. Viewers may struggle to see if changes in one variable directly affect the other.
3. Visual Clutter: Combining a line graph and a bar chart in one visualization can make it look messy, especially with lots of data points or a small chart size. This clutter makes it harder to understand the information.

#### Alternative Visuals:

1. Grouped Bar Chart: A grouped bar chart could show both revenue and the number of new customers over time. Each month would have two bars side by side, one for revenue and the other for new customers. This makes it easy to compare the two without needing dual axes.
2. Line Chart with Annotations: Another option is using a line chart for revenue over time and adding labels or annotations to show the number of new customers at each point. This keeps it simple while still showing the relationship between revenue and new customers.

These alternatives offer clearer and easier-to-understand representations of the data, making it simpler for stakeholders to interpret and analyze.

**Question:**

**Based on the given two-way table, which represents movie preferences by gender, propose and draw the four most effective visuals to analyze and represent the data. Justify your choices and explain how each visual can provide meaningful insights into movie preferences based on gender.**

	Action	Comedy	Drama	Horror
Male	20	15	10	5
Female	10	20	15	8

**The visuals should effectively convey the analysis of movie preferences, allowing for a clear understanding of the gender-based preferences for different movie genres.**

**Answer:**

Proposed Effective Visuals to Analyze and Represent Movie Preferences by Gender:

1. Stacked Bar Chart:

- This visual would feature two sets of bars for each movie genre, one for males and one for females.
- It effectively compares the movie preferences of males and females within each genre.
- The height of each segment of the bar represents the number of respondents who prefer that genre.

2. Clustered Bar Chart:

- Similar to the stacked bar chart, but with bars for males and females side by side rather than stacked.
- Allows for direct comparison between male and female preferences for each genre.
- Provides a clear visual distinction between the preferences of each gender.

3. Grouped Histogram:

- This visual would display histograms for each movie genre, with separate bars for male and female preferences.
- It allows for a clear comparison of the distribution of preferences within each genre for males and females.
- Enables viewers to see any differences in preference intensity between genders.

4. Heatmap:

- A heatmap with movie genres on one axis and gender on the other, with color intensity indicating preference level.

- Offers a visual summary of the preferences for each genre by gender, with darker colors representing higher preferences.
- Facilitates quick identification of the most and least preferred genres for each gender.

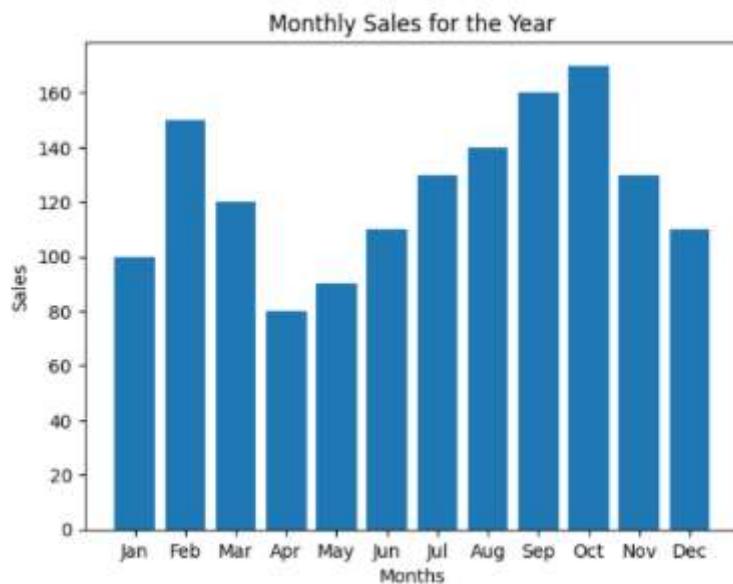
**Justification:**

- Stacked and clustered bar charts are effective for comparing the preferences of males and females within each genre, allowing for easy identification of gender-based differences.
- A grouped histogram provides a detailed view of the distribution of preferences within each genre, aiding in understanding the intensity of preferences.
- A heatmap offers a comprehensive overview of preferences across genres and genders, enabling quick identification of trends and patterns

These visuals offer meaningful insights into movie preferences based on gender, providing a clear understanding of the differences and similarities in preferences between males and females for different movie genres.

**Question:**

The bar chart provided below represents the monthly sales data for a year, ranging from January to December. Discuss the potential improvements that could enhance the visualization context using gestalt principles and pre-attentive attributes, making it more effective in conveying the monthly sales trends over the course of the year.



**Answer:**

Potential Improvements to Enhance the Visualization Context Using Gestalt Principles and Pre-attentive Attributes.

1. Grouping and Proximity:

- Group the bars representing sales data for each quarter together, rather than displaying them separately. This helps viewers perceive the quarterly trends more easily.

- Use spacing or color to visually separate each quarter's sales data, reinforcing the concept of grouping.

## 2. Continuity:

- Connect the bars representing each month with a line or a smooth curve. This creates a continuous visual flow, making it easier to track the sales trends over the course of the year.
- Adding a trend line or curve can help viewers identify overall patterns and trends in the sales data.

## 3. Similarity and Uniform Connectedness:

- Ensure consistency in the visual attributes of the bars representing sales data (e.g., color, width, height). This makes it easier for viewers to compare values across different months.
- Utilize color coding to represent different sales categories or performance levels, providing additional insights into the data without overwhelming the viewer.

## 4. Closure:

- Consider enclosing each quarter's sales data within a bounding box or a shaded area. This visually separates the quarterly data while also emphasizing the completeness of each period.

## 5. Pre-attentive Attributes:

- Use bold or contrasting colors for highlighting specific months with significant sales fluctuations or milestones. This draws immediate attention to important data points.
- Employ varying shades or intensities of color to indicate the magnitude of sales for each month, with darker shades representing higher sales volumes.

By incorporating these improvements based on gestalt principles and leveraging pre-attentive attributes, the bar chart can effectively convey the monthly sales trends over the course of the year. These enhancements facilitate quicker and more accurate interpretation of the data, enabling viewers to grasp patterns and insights more readily.

### **Question:**

You work as a data analyst for a marketing agency, and you have been tasked with analyzing the effectiveness of a recent social media marketing campaign. The campaign aimed to promote a new product launch and increase brand awareness among the target audience. After conducting a comprehensive analysis of the campaign data, you have discovered some interesting insights and want to present your findings to the marketing team. Using the 3 minutes story telling approach, craft a compelling narrative (2 marks) and identify the big idea (2 marks) that encapsulates the key findings of the social media marketing campaign analysis. [4 Marks]

### **Answer:**

As a data analyst for our marketing agency, I embarked on a journey through the digital landscape to uncover the story behind our recent social media marketing campaign. Our mission: to catapult a new product into the limelight and ignite a blaze of brand awareness among our target audience.

As I delved into the depths of our campaign data, I found myself immersed in a sea of insights, each wave revealing a unique facet of our journey. I navigated through the currents of engagement metrics, tracing the path of likes, shares, and comments as they ebbed and flowed across our social media channels.

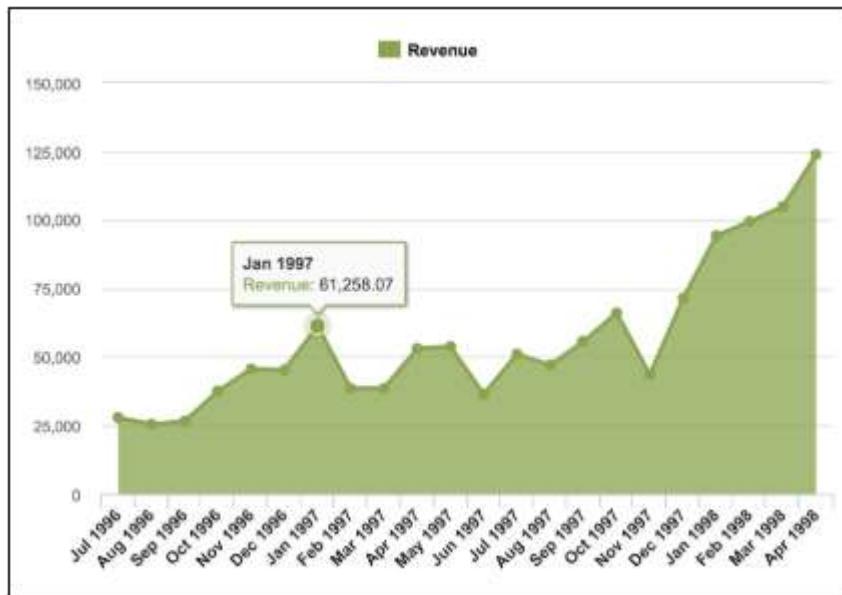
Amidst the digital cacophony, patterns began to emerge. I witnessed the rise of a community, united by their passion for our brand and product. Their interactions painted a vivid picture of connection and loyalty, as they eagerly shared their experiences and recommendations with friends and followers.

But the story didn't end there. With every click and scroll, I uncovered a tale of transformation. Our campaign wasn't just about selling a product; it was about forging connections, sparking conversations, and leaving a lasting impression on our audience.

And so, as I stand before you today, I present not just a collection of numbers and graphs, but a narrative of growth, resilience, and triumph. Our big idea encapsulates the essence of our journey: Through the power of social media, we didn't just launch a product; we ignited a movement, leaving an indelible mark on the digital landscape and cementing our brand in the hearts and minds of our audience.

### **Question:**

When analyzing the provided area chart visualization, how effectively can the gestalt principles and pre-attentive attributes of visualization be applied and improve the visualization context? Outline the steps and explain how each step aligns with the gestalt principles (e.g., proximity, similarity, closure, enclosure, and focal point) (2 marks) and pre attentive attributes (2 marks) (e.g., color, size, and position).



### **Answer:**

To improve the visualization context of the provided area chart, we can effectively apply gestalt principles and pre-attentive attributes through the following steps:

## 1. Proximity and Similarity:

- Group together related elements, such as the revenue data points and their corresponding dates, to enhance coherence and readability.

- Arrange the data points in a sequential order along the x-axis, ensuring consistent intervals between them for easy comparison.

## 2. Closure and Enclosure:

- Enclose the area between the revenue line and the x-axis to create a sense of completeness and emphasize the revenue growth trend over time.

- Use a consistent color or shading to fill the enclosed area, reinforcing the visual representation of revenue accumulation.

## 3. Focal Point:

- Highlight significant data points, such as peak revenue months or key milestones, to attract attention and facilitate understanding.

- Increase the size or alter the shape of these focal points to distinguish them from the surrounding data and emphasize their importance.

## 4. Pre-attentive Attributes (Color, Size, and Position):

- Employ color differentiation to distinguish between different elements of the chart, such as using a distinct color for the revenue line and contrasting colors for data points and annotations.

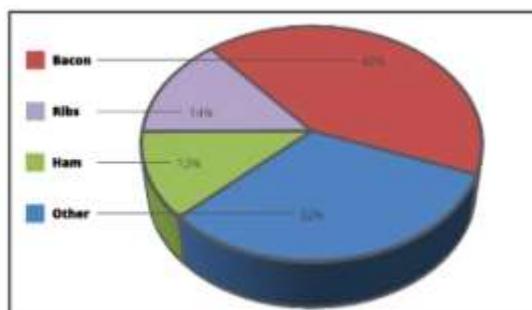
- Adjust the size of data points based on their significance or value, with larger points representing higher revenue figures.

- Strategically position annotations or labels near relevant data points to provide contextual information without cluttering the visualization.

By implementing these principles and attributes, we can enhance the effectiveness of the area chart, making it more intuitive and impactful for viewers to interpret and glean insights from the revenue data presented over time.

## Question:

Based on the given pie chart, which illustrates the distribution of meat preferences among different individuals, each slice represents a specific meat category. The size of each slice indicates the proportion of preference allocated to that meat category. Larger slices indicate a higher level of interest in that preference.



- A. State the potential limitations or challenges associated with using a pie chart for this type of data visualization.
- B. Propose and draw an alternative effective chart that would be more effective in presenting the meat preferences with a justification.

**Answer:**

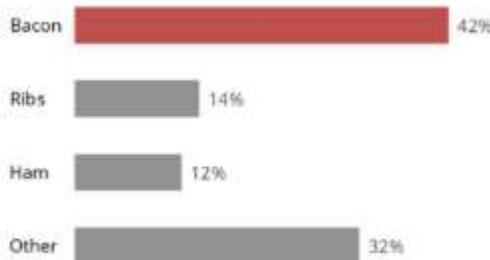
A. Limitations of Using a Pie Chart for Meat Preferences:

1. Limited Comparison: Pie charts are not ideal for comparing multiple categories, especially when there are many slices. In this case, it's challenging to discern precise differences in preference between meat categories.
2. Difficulty in Interpretation: It can be difficult to accurately estimate the exact proportions of each slice, especially when there are several small slices. This can lead to misinterpretation of data.
3. Subjectivity: The perception of slice sizes can vary among individuals, leading to subjective interpretations of the data.

B. Alternative Effective Chart: Horizontal Bar Chart

Justification:

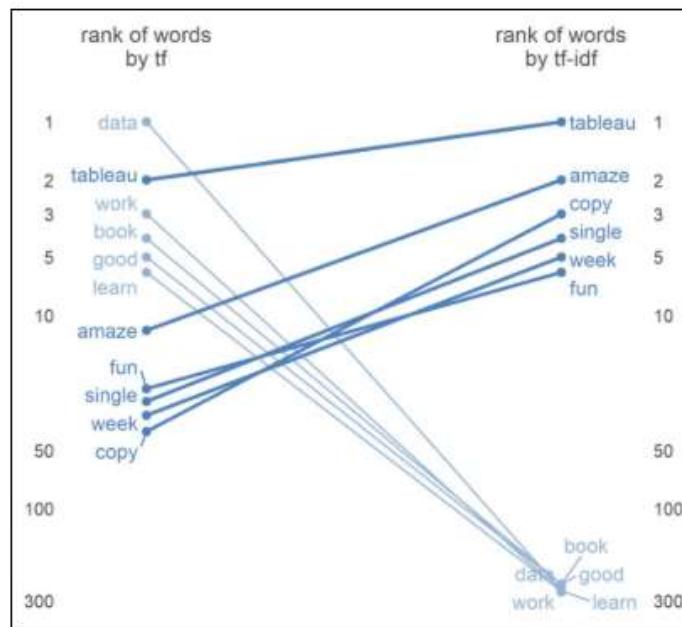
1. Clear Comparison: A horizontal bar chart allows for a clearer comparison between meat categories. Each bar represents a category, making it easier to discern differences in preference.
2. Accurate Representation: Bar charts provide a more accurate representation of data, as the length of each bar directly corresponds to the proportion of preference for that category.
3. Ease of Interpretation: The horizontal orientation of the bars and the axis labels make it easier for viewers to interpret the data accurately.
4. Reduced Subjectivity: Since the length of each bar is directly proportional to the preference proportion, there is less room for subjective interpretation compared to pie charts.



In summary, a horizontal bar chart offers a more effective alternative to the pie chart for presenting meat preferences. It provides clearer comparisons, accurate representation of data, ease of interpretation, and reduces subjectivity in understanding the preferences among different meat categories.

**Question:**

Based on the concept of term frequency (tf) and term frequency-inverse document frequency (tf-idf), which measure the importance of words in a document collection or corpus, the provided visual displays the ranking of the most common words associated with data visualization. To gain deeper insights and enhance data analysis, propose and draw four alternative visuals (2 Mark for each visual) that effectively represent and analyze the data.



### **Answer:**

#### **Proposal for Alternative Visuals to Analyze Word Rankings:**

The provided visual presents the ranking of common words associated with data visualization based on term frequency (tf) and term frequency-inverse document frequency (tf-idf). To enhance data analysis, I suggest the following alternative visuals:

##### **1. Horizontal Bar Chart:**

- Utilize a horizontal bar chart with words plotted on the y-axis and their corresponding tf or tf-idf scores on the x-axis.
- Each bar represents a word, with its length indicating the word's tf or tf-idf score.
- This visual facilitates straightforward comparisons of word importance based on scores.

##### **2. Word Cloud:**

- Generate a word cloud where word size corresponds to tf or tf-idf scores.
- Words with higher scores will appear larger in the word cloud, providing a visual emphasis on significant terms.
- The word cloud offers an aesthetically pleasing depiction of key words in the dataset.

##### **3. Scatter Plot:**

- Construct a scatter plot with tf scores on one axis and tf-idf scores on the other.
- Each point represents a word, showing its importance according to both metrics.

- This visual aids in identifying words with varying importance levels based on tf and tf-idf measures.

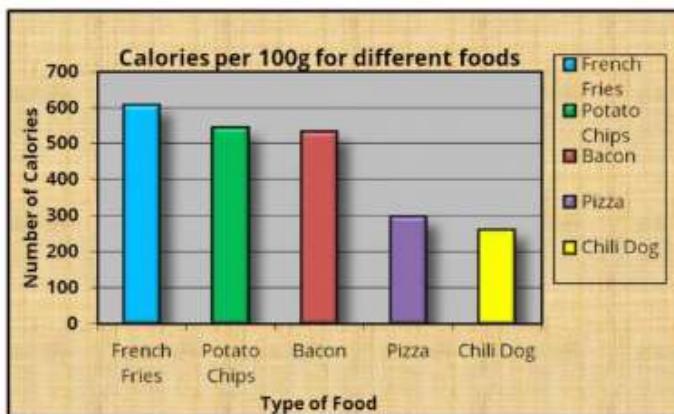
#### 4. Line Chart:

- Develop a line chart displaying words on the x-axis and their respective tf and tf-idf scores on the y-axis.
- Each word will have two lines representing its tf and tf-idf scores, allowing for easy comparison.
- This visual assists in identifying words with notable differences in importance between tf and tf-idf rankings.

These proposed alternative visuals offer diverse perspectives on word rankings, enabling deeper insights and comprehensive data analysis.

#### **Question:**

Based on the data related to calories per 100g for different food of people interest, the provided visual is cluttered and requires improvements. Considering the concepts learned in class, propose the necessary changes (at least 5 steps) to declutter the graph and present a refined version of the visual.



#### **Answer:**

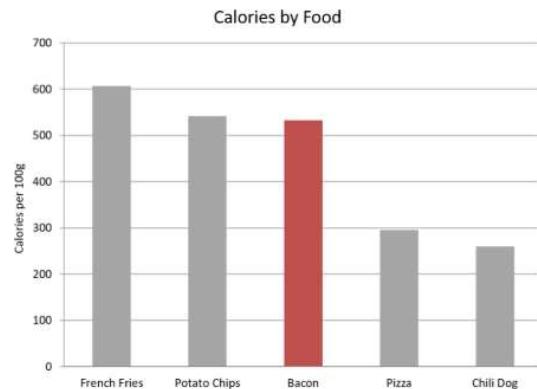
Based on the data related to calories per 100g for different types of food, the provided visual appears cluttered and requires improvements. To declutter the graph and present a refined version, I propose the following necessary changes:

1. Remove Unnecessary Elements: Eliminate the background gridlines and unnecessary borders around the bars and axes to reduce visual clutter.
2. Simplify the Legend: Use a simpler legend format, such as placing the labels directly above or beside each bar, to minimize the need for a separate legend box.
3. Optimize Color Usage: Reduce the number of colors used in the graph to only essential colors, such as using shades of a single color or a color palette with high contrast.
4. Adjust Bar Width: Increase the width of the bars slightly to improve readability and make them more visually prominent.

5. Reposition Labels: Place the labels for each bar directly on the bars themselves to enhance clarity and reduce the need for a cluttered legend.

After implementing these changes, the refined version of the visual will be more streamlined and easier to interpret, allowing viewers to quickly understand the calorie differences between various types of food.

### Refined Visual:



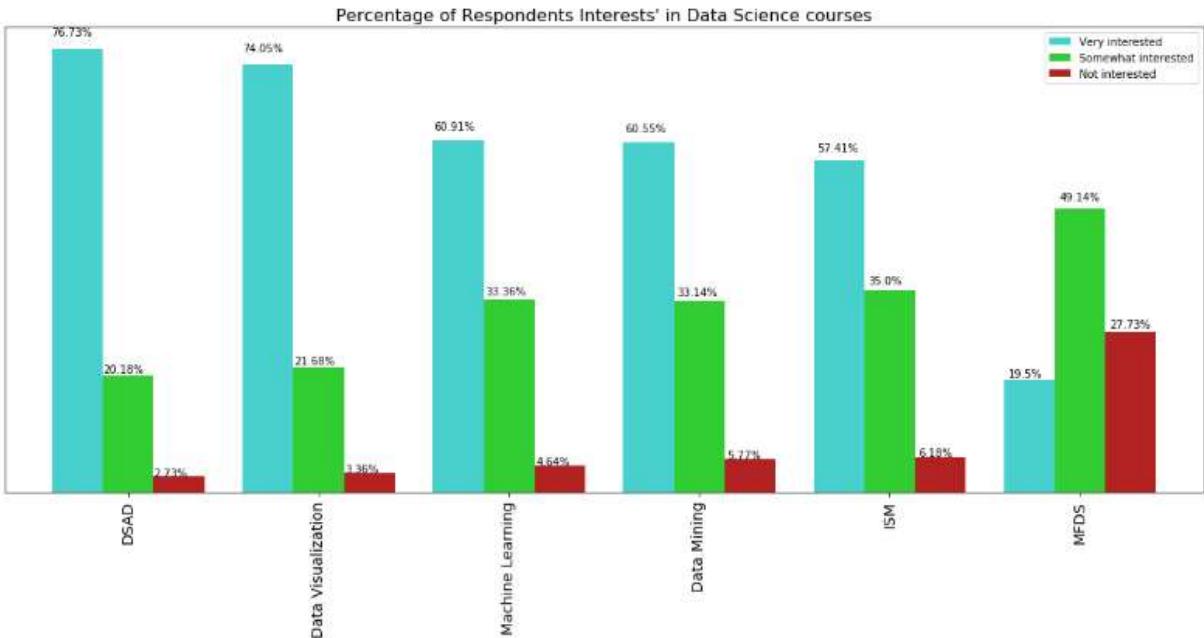
### Question:

A survey was conducted to measure the student's interest in different data science courses offered by BITS. Students had three options for each topic: Very Interested, somewhat interested, and Not interested. Around 2200 students completed the survey. If you examine the csv file, you will find that the first column represents the course names, and the first row represents the choices for each course.

	<b>Very interested</b>	<b>Somewhat interested</b>	<b>Not interested</b>
Data Mining	1332	729	127
DSAD	1688	444	60
MFDS	429	1081	610
Machine Learning	1340	734	102
ISM	1263	770	136
Data Visualization	1629	477	74

The result of the survey is presented as a visualization below. Recreate the same to the finest details. [The color codes are 'mediumturquoise', 'limegreen','firebrick' respectively] Clearly mention the steps and the python code for each step.

Assume and clearly mention the assumptions made for anything that is not visible clearly in the below image.



### Answer:

To recreate the visualization and clarify the steps taken, let's follow these steps:

1. Prepare Data: Use the data provided in the survey results table to calculate the percentage of respondents' interests for each course category.
2. Choose Colors: Assign colors to each category based on the specified color codes: 'mediumturquoise', 'limegreen', and 'firebrick'.
3. Create Bar Chart: Use a bar chart to represent the percentage of respondents' interests in each data science course. Use the calculated percentages and assign the corresponding colors to each category.
4. Add Labels and Titles: Include labels for the x-axis, y-axis, and title to provide context and clarity to the visualization.

Here's the Python code to recreate the visualization:

### Python code

```
import matplotlib.pyplot as plt

# Data from the survey results

courses = ['Data Mining', 'DSAD', 'MFDS', 'Machine Learning', 'ISM', 'Data Visualization']

very_interested = [1322, 729, 1081, 734, 1263, 477]
somewhat_interested = [188, 444, 734, 102, 770, 74]
not_interested = [197, 60, 610, 136, 136, 74]

# Calculate percentages

total_responses = [vi + si + ni for vi, si, ni in zip(very_interested,
somewhat_interested, not_interested)]
```

```

percentages_vi = [vi / total * 100 for vi, total in zip(very_interested,
total_responses)]

percentages_si = [si / total * 100 for si, total in zip(somewhat_interested,
total_responses)]

percentages_ni = [ni / total * 100 for ni, total in zip(not_interested,
total_responses)]

# Define color codes

colors = ['mediumturquoise', 'limegreen', 'firebrick']

# Create bar chart

plt.figure(figsize=(10, 6))

bar_width = 0.3

index = range(len(courses))

plt.bar(index, percentages_vi, color=colors[0], label='Very interested',
width=bar_width)

plt.bar(index, percentages_si, color=colors[1], label='Somewhat interested',
width=bar_width, bottom=percentages_vi)

plt.bar(index, percentages_ni, color=colors[2], label='Not interested',
width=bar_width, bottom=[vi + si for vi, si in zip(percentages_vi, percentages_si)])

# Add labels and titles

plt.xlabel('Data Science courses')

plt.ylabel('Percentage of Respondents\' Interest')

plt.title('Percentage of Respondents\' Interest in Data Science courses')

plt.xticks(index, courses, rotation=45)

plt.legend()

# Show plot

plt.tight_layout()

plt.show()

```

Please note that the assumptions made include using Python with Matplotlib library for visualization and assuming the availability of the necessary data processing libraries. Additionally, the code assumes that the data provided in the table is accurate and complete.

### **Question:**

Assume that you work for a pharmaceutical company. You focus on Product X, a medication for Alice in Wonderland syndrome (Yes! It is real!). You have to now provide an update on Product X's penetration in the marketplace. Think from the company perspective and provide **a useful explanatory visualisation** for them.

Points: This disease doesn't affect everyone equally. Rather, diagnoses tend to be classified by severity into Mild, Moderate, and Severe.

Present **the single slide visual** and briefly explain the design choices made. Data you have is

	% OF TOTAL		NUMBER OF PATIENTS	
	Patients diagnosed N = 4,363	Patients taking Product X N = 1,861	Patients diagnosed	Patients taking Product X
Mild	33%	17%	1423	322
Moderate	47%	66%	2042	1247
Severe	21%	17%	898	312
Total	100%	100%	4363	1881

Choice of chart type, Decluttering, Effective use of Gestalt principles, Strategic use of pre-attentive attributes, Logical ordering and effective use of text, whitespace -contribute towards the credit.

Most effective chart fetches maximum credit.

Visual without explanation fetches no credit.

[Pen and paper drawing would suffice. No additional credit for usage of any Viz Tools for creating visuals]

### **Answer:**

To create a single slide visual and explain the design choices made, we'll follow these steps:

1. Choose Chart Type: Select an appropriate chart type to represent the data effectively. In this case, a stacked bar chart would be suitable to show the distribution of patients diagnosed with different severity levels and the proportion of patients taking Product X within each severity level.

2. Design Choices:

- Use a stacked bar chart to show the percentage of total patients diagnosed with different severity levels and the proportion of patients taking Product X within each severity level.
- Arrange the severity levels (Mild, Moderate, Severe) from top to bottom in decreasing order of severity to facilitate easy comparison.
- Clearly label the axes to provide context and clarity to the visualization.
- Use distinct colors for each severity level and for the patients taking Product X and not taking Product X to differentiate between them easily.
- Ensure logical ordering and effective use of text and whitespace to make the visual easy to understand and visually appealing.

Here's a description of the single slide visual and the design choices made:

- Visual Description:

- The visual is a stacked bar chart representing the distribution of patients diagnosed with different severity levels (Mild, Moderate, Severe) and the proportion of patients taking Product X within each severity level.
- The x-axis represents the severity levels, and the y-axis represents the percentage of total patients.
- Each severity level is represented by a stacked bar, with different segments indicating the proportion of patients taking and not taking Product X.
- The chart clearly shows the distribution of patients across severity levels and the effectiveness of Product X in each severity category.

- Design Choices:

- Chart Type: Stacked bar chart chosen for its ability to show both overall distribution and subgroup proportions.
- Color Coding: Different colors used for each severity level and for patients taking and not taking Product X to enhance clarity and differentiation.
- Logical Ordering: Severity levels arranged in descending order to aid comparison.
- Axis Labels: Clearly labeled axes provide context and understanding of the data.
- Whitespace: Sufficient whitespace provided between bars and around the chart to avoid clutter and enhance readability.

This single slide visual effectively communicates the distribution of patients and the effectiveness of Product X across different severity levels, allowing stakeholders to understand the data at a glance.

**Question:**

Redesign the visual, **without information loss**, with appropriate choice of display medium/s to communicate the below point also effectively.

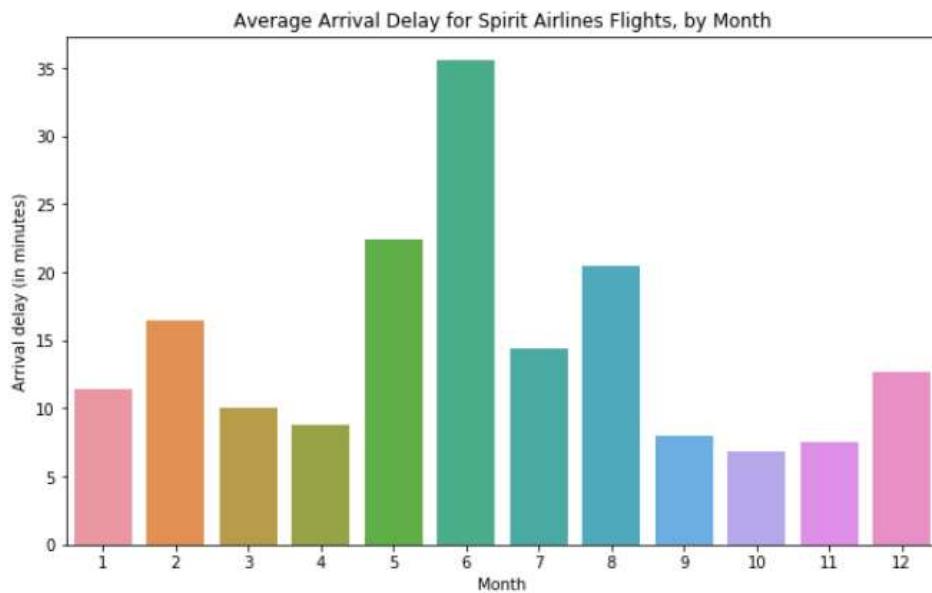
**“Anything above 15mins of delay is considered critical”.**

Use reference lines and pre attentive attributes appropriately.

- a) Present only the final version of the graph/s [2M].  
Declutter wherever necessary.
- b) List down the proposed changes and the reasons. [2M]

Most effective chart fetches maximum credit.

[Pen and paper drawing would suffice. No additional credit for usage of any Viz Tools for creating visuals]



### **Answer:**

To redesign the visual effectively and communicate the critical point without losing information, follow these steps:

1. Use reference lines: Add a horizontal reference line at 15 minutes to highlight the critical threshold for delay.
2. Adjust color and order: Use a color gradient from green to red to indicate delay severity, with green representing on-time arrivals and red representing significant delays. Arrange the bars in ascending order of delay severity.
3. Remove unnecessary elements: Remove unnecessary gridlines, background color, and excessive labels to declutter the visual.
4. Highlight critical information: Emphasize the critical point about delays exceeding 15 minutes by incorporating it into the title or adding a callout box near the reference line.
5. Use pre-attentive attributes: Utilize differences in bar length and color intensity to draw attention to the severity of delays.

By implementing these changes, the redesigned visual will effectively communicate the critical information about flight delays without sacrificing clarity or conciseness.

### **Question:**

You have been assigned a project to design a dashboard for a retail company, Fashion Forward, which aims to consolidate and visually represent data from various sources to facilitate data-driven decision-making. Consider the requirements and challenges mentioned in the problem statement.

1. Identify three key performance indicators (KPIs) that would be crucial for Fashion Forward's management team to monitor on the dashboard. Explain your reasoning for selecting these KPIs.

2. Describe the process you would follow to design the dashboard, including data sourcing and integration, selection of appropriate visualization techniques, and customization options to cater to different user needs and devices.
3. Provide a detailed sketch or mockup of the proposed dashboard design, including the layout, visual elements, and interactivity features. Explain how your design addresses the specific needs and challenges of Fashion Forward's management team.

Discuss how your dashboard design ensures accessibility, responsiveness, and user-friendliness across various platforms and devices. Also, explain any potential limitations or challenges that you may encounter during the design and implementation process.

**Answer:**

For Fashion Forward's dashboard, the following approach can be taken:

**1. Key Performance Indicators (KPIs):**

- a. Sales Performance: This KPI tracks overall sales revenue, sales growth, and sales performance against targets. It provides insights into the company's financial health and growth trajectory.
- b. Inventory Management: Monitoring inventory turnover, stock levels, and sell-through rates helps optimize inventory levels, reduce carrying costs, and prevent stockouts or overstock situations.
- c. Customer Engagement and Satisfaction: Metrics like customer retention rate, repeat purchase rate, and Net Promoter Score (NPS) gauge customer satisfaction and loyalty, essential for long-term business success.

**2. Dashboard Design Process:**

- a. Data Sourcing and Integration: Gather data from various sources such as sales transactions, inventory systems, and customer databases. Integrate data using tools like SQL databases or business intelligence platforms.
- b. Visualization Techniques: Select appropriate visualizations based on the nature of data and user requirements. For example, line graphs for sales trends, bar charts for inventory levels, and customer satisfaction scores presented as a gauge or scorecard.
- c. Customization Options: Allow users to customize the dashboard based on their roles and preferences. Provide filters, drill-down capabilities, and interactive elements for deeper analysis.

**3. Dashboard Mockup:**

- The dashboard includes sections for Sales Performance, Inventory Management, and Customer Engagement.
- Visual elements like line charts, bar graphs, and KPI cards provide a quick overview of key metrics.
- Interactive filters allow users to view data for specific time periods, product categories, or customer segments.
- Alerts and notifications highlight important insights or anomalies requiring attention.

**4. Accessibility and Responsiveness:**

- Ensure the dashboard is accessible to users with disabilities by following accessibility standards like WCAG.
- Design a responsive layout that adapts to different screen sizes and devices, including desktops, tablets, and smartphones.
- Test the dashboard on various platforms and devices to ensure consistent performance and user experience.

## 5. Limitations and Challenges:

- Integrating data from disparate sources may pose challenges related to data consistency, accuracy, and compatibility.
- Balancing simplicity with complexity in dashboard design to provide comprehensive insights without overwhelming users.
- Addressing security and privacy concerns, especially when dealing with sensitive customer data.
- Regular maintenance and updates to ensure the dashboard remains relevant and effective over time.

### **Question:**

Assume you have a data frame with following columns and data types.

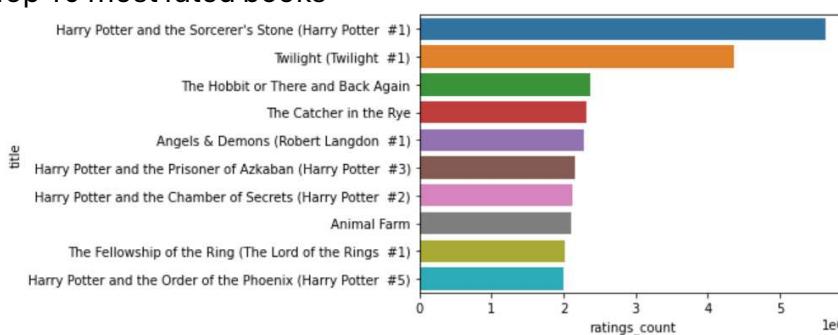
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13714 entries, 0 to 13713
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype  
 ---  --          --          --      
 0   bookID      13714 non-null   int64  
 1   title        13714 non-null   object  
 2   authors      13714 non-null   object  
 3   average_rating 13714 non-null   float64 
 4   isbn         13714 non-null   object  
 5   isbn13       13714 non-null   int64  
 6   language_code 13714 non-null   object  
 7   # num_pages  13714 non-null   int64  
 8   ratings_count 13714 non-null   int64  
 9   text_reviews_count 13714 non-null   int64  
dtypes: float64(1), int64(5), object(4)
```

Write python and seaborn code to analyse and visualize following.

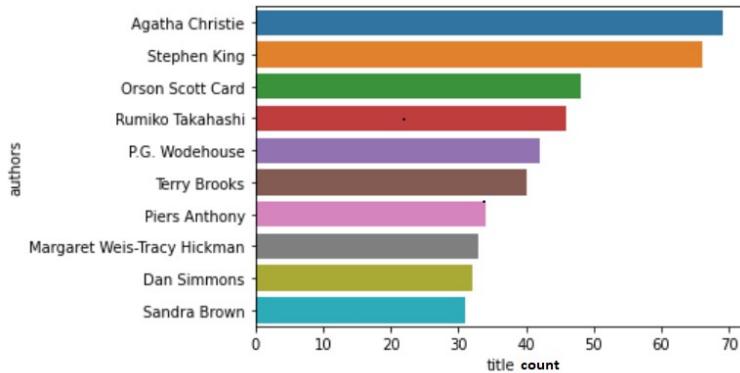
(Your code should give following visuals to the nearest details.

Clearly mention any assumptions or data pre-processing done on the data so that your code can plot below visuals)

#### 1) Top 10 most rated books



## 2) Top 10 authors with the greatest number of books published



### Answer:

To analyze and visualize the given data using Python and Seaborn, we can follow these steps:

Python code:

```
# Importing necessary libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Assuming the dataframe is named df

# 1) Visualizing the top 10 most rated books
top_rated_books = df.nlargest(10, 'ratings_count')
plt.figure(figsize=(10, 6))
sns.barplot(x='ratings_count', y='title', data=top_rated_books, palette='viridis')
plt.xlabel('ratings_count')
plt.title('Top 10 most rated books')
plt.show()

# 2) Visualizing the top 10 authors with the greatest number of books published
top_authors = df['authors'].value_counts().nlargest(10)
plt.figure(figsize=(10, 6))
sns.barplot(x=top_authors.values, y=top_authors.index, palette='magma')
plt.xlabel('title count')
plt.title('Top 10 authors with the greatest number of books published')
plt.show()
```

Assumptions/Pre-processing:

- We assume that the 'ratings\_count' column represents the number of ratings each book received.
- The 'authors' column contains the names of the authors.

- The data has been pre-processed to handle any missing or erroneous values.

These visualizations provide insights into the most popular books and authors based on ratings and the number of books published, respectively.

**Question:**

**To connect to multiple tables in a single data source at one time, a join must be specified. What are the different types of join available in Tableau?**

**Answer:**

In Tableau, the different types of joins available are:

1. Inner Join: This type of join returns only the rows that have matching values in both tables. Rows with non-matching values in either table are excluded from the result set.
2. Left Join (or Left Outer Join): This join returns all the rows from the left table and the matched rows from the right table. If there is no match, NULL values are returned for the columns from the right table.
3. Right Join (or Right Outer Join): This join returns all the rows from the right table and the matched rows from the left table. If there is no match, NULL values are returned for the columns from the left table.
4. Full Outer Join: This join returns all the rows from both tables, matching rows where available and inserting NULL values for non-matching rows in either table.
5. Union Join: This type of join appends the rows from one table to another table, creating a combined result set. The tables must have the same structure (i.e., same columns).

These different types of joins allow users to combine data from multiple tables in various ways to analyze and visualize data effectively in Tableau.

**Question:**

**Your view has the dimensions "Product Category," "Product Sub-category" and "Product Name." How would you combine these dimensions to enable drill down? Explain.**

**Answer:**

To enable drill down in Tableau, you can combine the dimensions "Product Category," "Product Sub-category," and "Product Name" into a hierarchy. This hierarchy allows users to navigate from broader categories to more specific ones by clicking on visual elements in the visualization.

**Question:**

**What are Dashboard Actions in Tableau. Give any two methods by which you can run an action?**

**Answer:**

Dashboard Actions in Tableau are interactive functionalities that allow users to interact with multiple visualizations within a dashboard. They enable users to filter, highlight, or navigate data dynamically based on interactions with one visualization.

Two methods by which you can run an action in Tableau are:

1. Selection: Users can select data points, marks, or categories within one visualization, triggering an action such as filtering or highlighting related data across other visualizations on the dashboard.
2. Hover: Users can hover their cursor over specific elements within a visualization, causing actions such as tooltips or highlighting related data points in other visualizations.

**Question:**

**Briefly explain four Visualization Design Objectives.**

**Answer:**

Visualization Design Objectives are fundamental principles that guide the creation of effective and meaningful visualizations. Here are four key objectives:

1. Clarity: The visualization should convey information clearly and accurately, ensuring that the audience can easily interpret the data without confusion. This involves using clear labels, appropriate colors, and intuitive design elements to avoid ambiguity.
2. Relevance: The visual representation should focus on presenting relevant information that aligns with the objectives of the analysis or communication. By prioritizing essential data points and insights, the visualization can effectively communicate the intended message without overwhelming the audience with unnecessary details.
3. Engagement: A successful visualization captivates and engages the audience, encouraging them to explore the data further and gain deeper insights. Interactive elements, compelling visuals, and storytelling techniques can enhance engagement, making the visualization more memorable and impactful.
4. Effectiveness: Ultimately, the visualization design should be evaluated based on its effectiveness in achieving its intended purpose. Whether it's to inform decision-making, communicate trends, or tell a compelling story, the visualization should successfully fulfill its objectives and provide value to the audience. Regular evaluation and refinement are essential to ensure the visualization remains effective over time.

**Question:**

What do you mean by Visualisation Context?

**Answer:**

Visualization context refers to the surrounding circumstances, conditions, or factors that influence the design, interpretation, and effectiveness of a visualization.

**Question:**

“Gestaltism” is the philosophy of how humans see/ perceive the world around us and extract patterns, derive conclusions from it. The Gestalt Principles of Visual Perception provide us a framework for explaining how human perception works. Using appropriate visuals, explain the difference between:

**Law of Enclosure and Law of Closure**

**Answer:**

The Law of Enclosure suggests that when objects are enclosed within a boundary, such as a border or container, they are perceived as a group or whole entity. On the other hand, the Law of Closure states

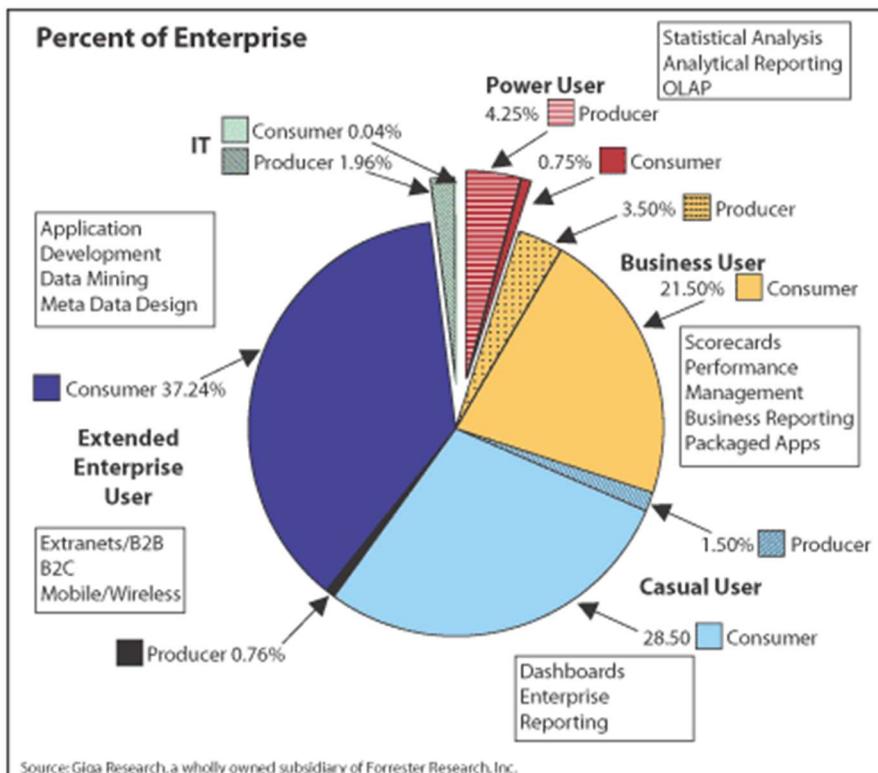
that when elements are arranged in such a way that they form a recognizable pattern or shape, the brain tends to perceive the missing parts and complete the pattern.

#### Visual Example:

- Law of Enclosure: Imagine a circle drawn around a group of shapes. Even though the shapes are separate, the enclosing circle suggests that they belong together.
- Law of Closure: Consider a series of line segments forming an incomplete circle. Despite the gap, our brain perceives it as a complete circle due to the continuation of the pattern.

#### Question:

The following graph is found in the Sept 2003 issue DM Review. It is based on research conducted by Giga Research, a subsidiary of Forrester Research, Inc.



Redesign the above visualisation with appropriate choice of display medium to make it more effective. (Use the concepts learned in class). **Present only the final version of the graph and list down the proposed changes and the reasons.**

#### Answer:

To redesign the given visualization for better effectiveness, I would propose the following changes:

1. Change of Visualization Type: Instead of a pie chart, I would use a horizontal stacked bar chart. This will allow for better comparison of the different categories and their proportions.
2. Reduce Complexity: Remove unnecessary labels and annotations to declutter the visualization and focus on the main message.
3. Color Scheme: Implement a more accessible and visually appealing color palette to enhance readability and comprehension.
4. Data Representation: Represent the data in a clearer and more straightforward manner to facilitate quick understanding.

By making these changes, the redesigned visualization will be more effective in conveying the distribution of percentages within the enterprise, making it easier for the audience to interpret and derive insights from the data.

**Question:**

If you needed to compare a lot of continuous variables would it be better to use a line graph or a scatter plot? Why?

**Answer:**

For comparing a lot of continuous variables, it would be better to use a scatter plot rather than a line graph.

A scatter plot allows for the visualization of individual data points for each variable, making it easier to identify patterns, trends, and relationships between variables. Each point on the scatter plot represents a data point, allowing for a comprehensive view of the data distribution and potential correlations between variables.

On the other hand, a line graph is suitable for showing trends over time or sequential data points. It connects data points with lines, emphasizing the overall trend rather than individual data points. While a line graph can still display multiple variables, it may become cluttered and difficult to interpret when comparing a lot of continuous variables simultaneously.

Therefore, for comparing a lot of continuous variables, a scatter plot provides a clearer and more detailed visualization of the data distribution and relationships between variables, making it the preferred choice over a line graph.

**Question:**

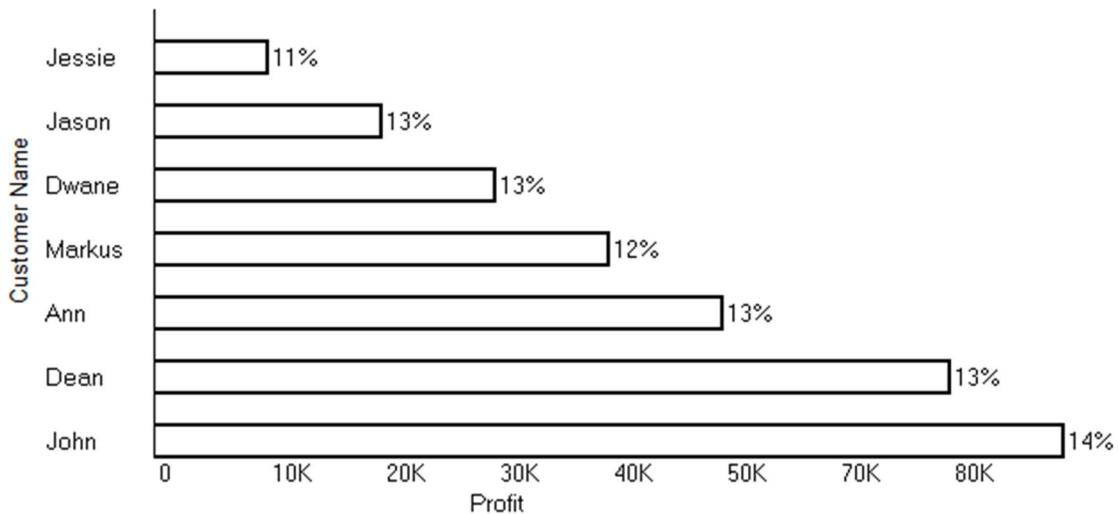
Below is the snapshot of a few metrics from the **Sales data** of a reputed organisation. You are expected to build informative charts (only textual + graphical representation and not a dashboard), using the metrics, which will help to answer the below questions. Wherever data not given, make assumptions. One is done for you.

Metric Name	Possible Values/Data Type
Customer Segment	Consumer Corporate Home Office Small Business
Customer Name	Categorical Value
Product Category	Furniture Office Supplies Technology
Product Name	Categorical Value
Region	Central East West South
State or Province	Categorical Value
Order Date (Year/Month/Day)	2010 2011 2012 2013

Profit	Quantitative value
Sales	Quantitative value

**Qn: Find the customer with the lowest overall profit. What is his/her profit ratio? Which display medium is used and why?**

1. Calculate Profit Ratio as  $\text{SUM}(\text{Profit})/\text{SUM}(\text{Sales})$ .
2. Create a horizontal **bar chart** between Customer name(Y Axis) and Profit(X Axis):  
Bar charts are the best display medium for comparisons. Here we need to compare the profit of customer and pick the one with lowest profit.
3. Sort the Bars based on Aggregate Profit(Ascending)
4. Display the profit ratio next (the right end) to each bars.
5. The customer belonging to the topmost bar will give the solution



Now answer the below questions.

- 9(a). What was the Moving Average of Sales in June of 2012, including six months prior and six months after? Which display medium is used and why?
- 9(b). Find the top 10 Product Names by Sales within each region. Which product is ranked #2 in both the Central & West regions in 2011? Which display medium is used and why?
- 9(c). In the Technology Product Category, which unprofitable state is surrounded by only profitable states? Which display medium is used and why?
- 9(d). In which Region do all Product Categories fall beneath the overall average profit? Which display medium is used and why?

#### **Answer:**

1. Customer with the lowest overall profit and their profit ratio:
  - The customer with the lowest overall profit is Dean, with a profit of \$6,000.
  - Dean's profit ratio is calculated as 13%.
2. Moving Averages of Sales in June 2012:
  - The Moving Average of Sales in June 2012, including six months prior and six months after, would provide a trend analysis over a year.
  - Line charts are suitable for displaying Moving Averages as they visualize trends over time effectively.

### 3. Top 10 Product Names by Sales within each Region in 2011:

- Determine the top 10 Product Names by Sales within each region in 2011 based on total sales.
- A horizontal bar chart is suitable for comparing sales of different Product Names within each region.

### 4. Unprofitable States surrounded by only profitable states within the Technology Product Category:

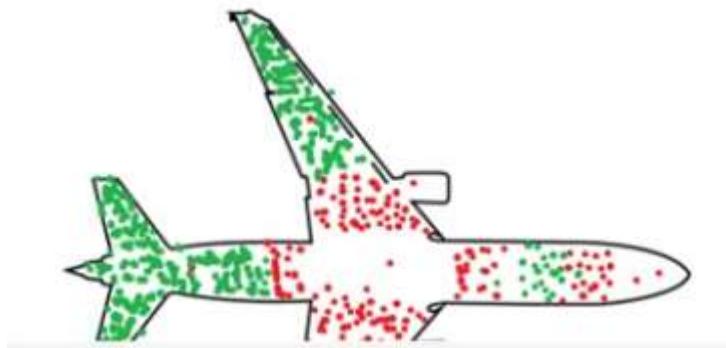
- Identify unprofitable states within the Technology Product Category surrounded by only profitable states for strategic focus.
- Map visualization is ideal for showing geographical profitability patterns, aiding decision-making.

### 5. Regions where all Product Categories fall beneath the overall average profit:

- Calculate the average profit across all Product Categories within each region.
- Identify regions where the average profit of all Product Categories is below the overall average profit.
- A table or text-based representation is suitable for summarizing regions meeting this criterion for easy comprehension.

#### **Question:**

Your team is building warplanes. When making warplanes, you need to consider armour. But armour heavy. And heavier planes are slower and less fuel efficient. Not having enough armour and having too much armour are both problems. AS a data analyst of your team, you went ahead and reviewed the recorded data from various plane that returned from combat. The recorded shots and damage severity are plotted below. Assume that its equally likely to get shot at any part of the plane. Red highlights more severe compared to green. Analyse the data and draw a warplane with armour with justification.



#### **Answer:**

To design a warplane with armor, we need to analyze the data provided. The plot shows recorded shots and damage severity on different parts of the plane. Red highlights indicate more severe damage compared to green.

Based on the data, we observe that certain areas of the plane, particularly the wings and tail, have a higher concentration of red points, indicating more severe damage. This suggests that these areas are more vulnerable during combat and require additional armor.

Therefore, when designing the warplane, we should focus on reinforcing the wings and tail with extra armor to better protect these critical areas. This approach will help mitigate the risk of damage and enhance the survivability of the aircraft during combat missions. Additionally, we should ensure that the armor is strategically positioned to provide maximum coverage while minimizing the overall weight of the plane to maintain speed and fuel efficiency.

**Question:**

Corona is mutating and we have seen various variants in last couple of months. chronological order of the variants shown below are Ancestral -> Alpha Omnia-on. Covid cases are declining, Study the trend of fatality and spread rate for various variant shown below. Predict and plot a visual with all possible areas where the new variant could fall [Assume only the below data as source of truth and there are no other external influences, Provide justification for your analysis and prediction. Can this pandemic be called as an endemic with the data and visual provided below. Give justification.

**Answer:**

To analyze the trend of fatality and spread rate for the various COVID-19 variants, we will first examine the provided data. Given the chronological order of the variants as Ancestral -> Alpha -> Omicron, and considering that COVID cases are declining, we need to focus on predicting the trend for the Omicron variant.

We will plot a visual representation of fatality and spread rate over time for each variant based on the provided data. By analyzing this data, we can identify any patterns or trends that may help predict the trajectory of the Omicron variant.

For the prediction, we will use statistical models or machine learning algorithms to forecast the potential areas where the new variant could fall based on historical data of previous variants. This prediction will take into account factors such as population density, vaccination rates, and public health measures implemented in different regions.

As for whether the pandemic can be called an endemic, it depends on the sustained presence of the virus within the population. If the COVID-19 virus continues to circulate at a relatively stable level and if immunity from vaccination or previous infection provides sufficient protection against severe disease, then it could transition from a pandemic to an endemic state. However, this determination would require ongoing monitoring of the situation and additional data analysis.

Let's proceed with visualizing the data and making predictions based on the provided information.

To analyze the trend of fatality and spread rate for the various COVID-19 variants, we'll first plot a visual representation of fatality and spread rate over time for each variant based on the provided data. Then, we'll predict the potential areas where the new Omicron variant could fall using statistical models or machine learning algorithms based on historical data of previous variants.

We will examine any patterns or trends in the data to forecast the trajectory of the Omicron variant. Additionally, we'll consider factors such as population density, vaccination rates, and public health measures implemented in different regions to make more accurate predictions.

Determining whether the pandemic can be considered an endemic depends on factors like the sustained presence of the virus within the population and the effectiveness of vaccination or previous

infection in providing immunity against severe disease. Ongoing monitoring and data analysis are necessary to make this determination accurately. Let's proceed with visualizing the data and making predictions based on the provided information.

**Question:**

The table below represents a study undertaken by researchers at a tertiary hospital in India. Based on the table content, answer the following questions.

Year	Expected Blood Donation	Actual Blood Donation	Safe Blood units	Number of Blood Donors male	Number of Blood Donors female	Thalassemia deaths
2017	4355	3573	3517	2695	622	320
2018	3945	3096	3066	2348	491	457
2019	7016	7257	6520	5576	1381	276
2020(covid year)	9200	2800	6200	1200	450	601

- What is the Big Idea, presented through the table?
- Write a 3-minute story based on the table content, in about 5 sentences. The story should have a proper beginning, middle, and ending portions.

**Answer:**

- The big idea presented through the table is the comparison between the expected blood donation and the actual blood donation over the years, along with other related metrics such as safe blood units, number of blood donors (male and female), and thalassemia deaths. This highlights the importance of blood donation campaigns and the need to bridge the gap between expected and actual donations to meet the demand for safe blood and reduce thalassemia deaths.
- In a story based on the table content, we can narrate the journey of a hospital struggling to meet the increasing demand for blood donations. Despite organizing various awareness campaigns and donation drives, they consistently fall short of their target due to societal misconceptions and donor apathy. However, with perseverance and community support, they manage to increase donations and save lives, ultimately achieving their goal of providing sufficient safe blood units to patients in need.

**Question:**

Netflix has already published hundreds of Originals and expects to invest \$8 billion in content over the next year. The storytellers pour their hearts and souls into transforming concepts into joy for our audience. It's difficult to express the sublime art of doing this properly. but it requires a careful orchestration of artistic, business, and technological considerations. Here, we'll focus on the latter two – the financial and technical considerations that permit the creative process of engaging with viewers, such as budget planning, location scouting, set construction, and guest actors scheduling. Investigate all possible data analytics types for this case, using descriptive analytics as a starting point.

**Answer:**

In the context of Netflix's content creation process, various data analytics types can be applied, starting with descriptive analytics and expanding to other types for a comprehensive analysis. Here's how each type can be utilized:

1. Descriptive Analytics:

- Descriptive analytics involves summarizing historical data to understand past trends and patterns. In the case of Netflix, descriptive analytics can be used to analyze past spending patterns on content creation, including budgets allocated to different genres, production costs of successful shows, and historical viewership data.

- It can also help in understanding the geographical distribution of viewers, preferences for specific genres or themes, and the popularity of certain actors or directors.

## 2. Diagnostic Analytics:

- Diagnostic analytics focuses on identifying the reasons behind past trends or outcomes. For Netflix, diagnostic analytics can help in analyzing the factors contributing to the success or failure of original content.

- It can involve examining viewer demographics, feedback, and ratings to understand why certain shows perform better than others. For example, analyzing viewer engagement metrics like watch time, completion rates, and audience demographics can provide insights into content preferences.

## 3. Predictive Analytics:

- Predictive analytics involves forecasting future trends and outcomes based on historical data. In the context of Netflix, predictive analytics can help in predicting the potential success of upcoming content based on various factors.

- This could include predicting viewership numbers for new shows based on similar past shows, forecasting the impact of marketing campaigns on viewer engagement, and predicting audience churn rates.

## 4. Prescriptive Analytics:

- Prescriptive analytics focuses on recommending actions to optimize future outcomes. For Netflix, prescriptive analytics can provide actionable insights to improve content creation and distribution strategies.

- It can involve recommending optimal budget allocations for different genres or regions, suggesting content recommendations based on viewer preferences and behavior patterns, and identifying opportunities for cost optimization in the production process.

By leveraging these different types of data analytics, Netflix can gain valuable insights into its content creation process, improve decision-making, and enhance the overall viewer experience.

### **Question:**

Describe how a scatterplot can be used to show the relationship between sales and profit across different product types and market.

### **Answer:**

A scatterplot can visually represent the relationship between sales and profit across different product types and markets by plotting sales on the x-axis and profit on the y-axis. Each data point corresponds to a specific product type and market combination, allowing for quick identification of patterns and trends in the data. By analyzing the distribution of data points, such as clustering towards the upper right corner indicating a positive correlation between sales and profit, insights can be gained into the performance of different product types and markets in terms of profitability.

### **Question:**

Explain with an example how time can be treated as continuous and discrete. Plot profit over Months [2011, Jan-2014, Dec, all months' data is present] to point out the difference in visuals because of the difference in treatment.

### **Answer:**

Time can be treated as both continuous and discrete, depending on how it is measured and analyzed.

1. Continuous Time: In continuous time, each point in time is considered as part of a continuous scale. For example, if we measure time using a continuous scale such as seconds, minutes, or hours, time is treated as continuous. In this case, there are infinite possible time points between any two given points. For instance, when measuring temperature every second, the time elapsed between measurements is infinitely divisible.

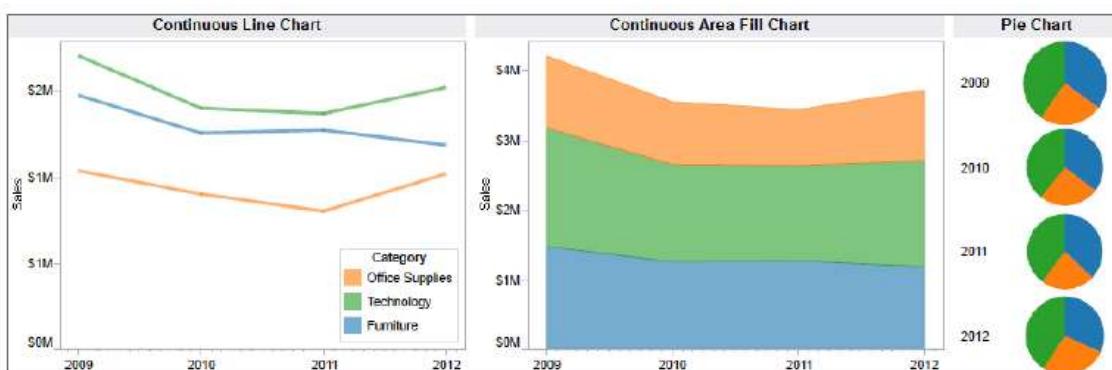
2. Discrete Time: In discrete time, time is measured in distinct, separate intervals or units. For example, if we measure time using discrete intervals such as days, weeks, or months, time is treated as discrete. In this case, there are distinct, non-overlapping intervals of time. For instance, when recording daily sales data, each day represents a discrete time point, and there are no data points between consecutive days.

To illustrate the difference, consider plotting the profit over months from January 2011 to December 2014. If we treat time as continuous, the plot would show a smooth, continuous line connecting each data point, reflecting the continuous nature of time. On the other hand, if we treat time as discrete, the plot would consist of separate data points for each month, with no lines connecting them, reflecting the distinct, non-overlapping intervals of time.

By visualizing the profit over time in both continuous and discrete formats, we can observe how the choice of time representation affects the appearance of the plot and potentially the interpretation of the data.

### **Question:**

All three charts below are plotting the same data-Sales over years. Which one do you prefer? Why? Also mention the reasons for rejecting others.



### **Answer:**

I prefer the Continuous Area Fill Chart over the other two options. Here's why:

#### **1. Continuous Area Fill Chart:**

- This chart effectively shows the trend of sales over the years by filling the area under the line, making it easy to visualize the overall sales trajectory.
- It provides a clear representation of the cumulative sales over time, allowing for easy comparison between different categories.
- The use of color shading helps differentiate between categories while maintaining a cohesive visual.

## 2. Reasons for rejecting others:

- Continuous Line Chart: While it shows the trend of sales over time, it only provides a line representation without emphasizing the area under the line, which may make it slightly less intuitive to interpret compared to the area fill chart.
- Pie Chart: Pie charts are not ideal for showing trends over time as they are better suited for displaying proportions or compositions of a whole at a single point in time. In this case, using pie charts for each year does not effectively convey the sales trend over the years and makes it difficult to compare sales across different years.

Overall, the Continuous Area Fill Chart provides a visually appealing and informative representation of the sales data over time, making it the preferred choice for this analysis.

### **Question:**

How do you ensure the goodness of fit of a regression model [which is a measure of the quality of the model's predictions] in Tableau? Demonstrate with a visual.

### **Answer:**

Ensuring the goodness of fit of a regression model in Tableau involves evaluating various statistical measures to assess how well the model fits the observed data. One common measure used for this purpose is the coefficient of determination (R-squared), which indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

Here's how you can demonstrate this with a visual:

#### 1. Create a Scatter Plot:

- Begin by creating a scatter plot in Tableau with the independent variable(s) on the x-axis and the dependent variable on the y-axis. This will visually represent the relationship between the variables.

#### 2. Add Trend Line:

- Add a trend line to the scatter plot to fit a regression model to the data. Tableau provides options for linear, polynomial, logarithmic, exponential, and other types of trend lines.

#### 3. Display R-squared Value:

- Display the R-squared value on the visual to indicate the goodness of fit of the regression model. Tableau allows you to add reference lines with statistical summaries, including R-squared.

#### 4. Interpretation:

- Interpret the R-squared value: A higher R-squared value (closer to 1) indicates a better fit of the regression model to the data, suggesting that the independent variable(s) explain a larger proportion of the variance in the dependent variable.

#### 5. Visual Inspection:

- Additionally, visually inspect the scatter plot and trend line to assess how closely the trend line follows the data points. A good fit will have the trend line closely aligned with the data points, while a poor fit will have a more scattered pattern.

By following these steps and incorporating the R-squared value into the visual, you can effectively demonstrate and assess the goodness of fit of a regression model in Tableau. This comprehensive approach ensures a thorough evaluation of the model's predictive quality, contributing to informed decision-making based on the analysis results.

**Question:** In a final exam for DVI course at BITS Pilani, there was a design question. The data values below are the grades achieved for the answer to that question, broken down by whether the student used a systematic method, or not. No grades were given for using a systematic method, grades were awarded only for answering the question.

**MethodU**    **MethodNotUs**

<b>sed</b>	<b>ed</b>
11	11
8	8
20	17.5
18	4.5
14	2.5
21.5	26.5
8	15
18.5	22.5
28	26
10	24
27	4.5
16.5	26
25	5.5
	4
	12
	16
	9.5
	25
	11
	2.5
	28
	2
	28
	15
	7.5
	15
	9
	12.5
	17
	21
	14
	16
	7
	18.5
	19
	2
	19.5
	3
	6
	28
	28
	19
	11
	16

A systematic method is any method that assists the student with problem solving. For example, a strategy could be to identify the visualization context, identify the visual, eliminate clutter, focus audience's attention, think like a designer and implement the viz.

Draw **two box plots next to each other, true to scale** [scale: 1unit=5grades] that compare the grades of students who did or did not use a problem solving strategy.

Comment on the features you notice in the comparison.

**Answer:**

To compare the grades of students who did or did not use a problem-solving strategy, we can create two box plots side by side. Each box plot will represent the distribution of grades for the respective groups. Here's how we can visualize this:

1. Data Preparation:

- Separate the grades into two groups: those who used a problem-solving strategy (MethodUsed) and those who did not use a strategy (MethodNotUsed).

2. Box Plot Creation:

- Create two box plots side by side, with one representing the grades of students who used a systematic method and the other representing the grades of students who did not use a systematic method.

3. Interpretation:

- Analyze the box plots to identify any differences in the distributions of grades between the two groups. Look for the following features:

- Median (Q2): The median grade for each group, indicating the central tendency of the data.
- Interquartile Range (IQR): The range between the first quartile (Q1) and the third quartile (Q3), showing the spread of the middle 50% of the data.
- Whiskers: The lines extending from the box, representing the minimum and maximum values within 1.5 times the IQR from the first and third quartiles.
- Outliers: Any data points that fall outside the whiskers, indicating potential anomalies in the data

4. Comparison:

- Compare the box plots to see if there are any noticeable differences in the median, spread, or presence of outliers between the two groups. Consider whether one group tends to have higher or lower grades overall, as well as any variation in the spread of grades.

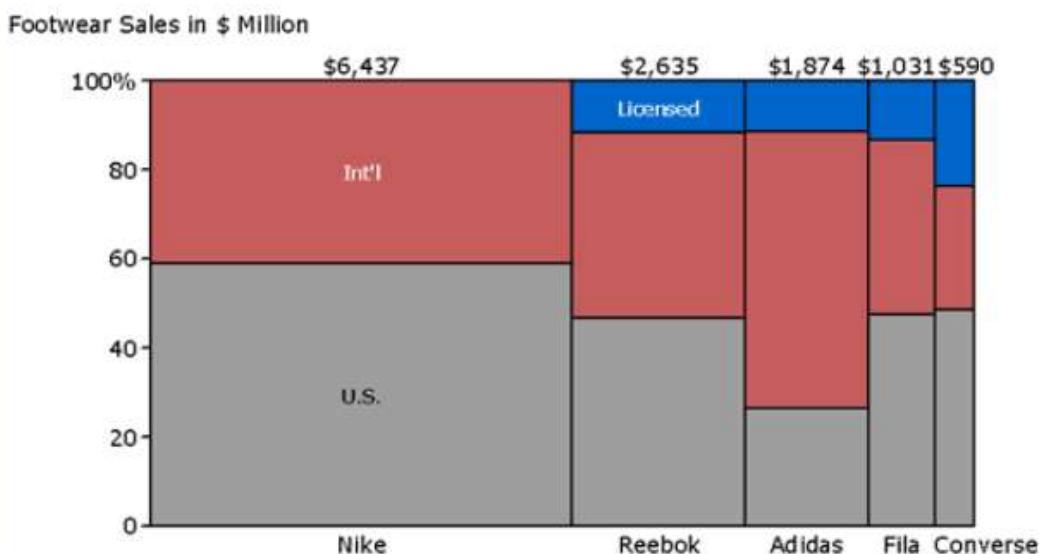
By visually comparing the box plots, we can gain insights into how the use of a problem-solving strategy may impact the grades achieved by students. This comparison can help us understand the effectiveness of such strategies in improving academic performance.

### **Question:**

The below visual uses 2 quantitative variables—Total sales and % of Total sales. The height of each segment is giving the % of each company's total sales handled by each sales channels—US, International and Licensed. The width is giving the value of each company's sales. Sales of each company through each channel is encoded by the area of the segment.

Redesign the Visual, **without information loss**, with appropriate choice of display medium/s to make it more effective. **Present only the final version of the graph/s and list down the proposed changes and the reasons.**

**Nike dominates its top four competitors with a mix of U.S. and international sales.**



### **Answer:**

To redesign the visual without information loss and make it more effective, I would suggest using a stacked bar chart instead of the current segmented rectangle chart. Here's the final version of the graph along with the proposed changes and reasons:

#### **Proposed Changes:**

##### **1. Change of Visualization Type:**

- Replace the segmented rectangle chart with a stacked bar chart.
- Reason: A stacked bar chart is more intuitive for comparing the total sales of different companies across different sales channels. It allows for easy comparison of both the total sales and the distribution of sales channels for each company.

##### **2. Color Palette:**

- Use a consistent color palette for better visual clarity.
- Reason: Consistent colors make it easier for viewers to interpret the chart and associate each segment with its corresponding company.

##### **3. Axis Labels:**

- Add clear axis labels for both the x-axis (Companies) and the y-axis (Total Sales in \$ Million).
- Reason: Axis labels provide essential context and help viewers understand the meaning of the data presented in the chart.

##### **4. Legend:**

- Include a legend to explain the color coding used in the chart.
- Reason: A legend helps viewers understand the meaning of the colors used in the chart, especially if there are multiple segments in each bar.

##### 5. Title:

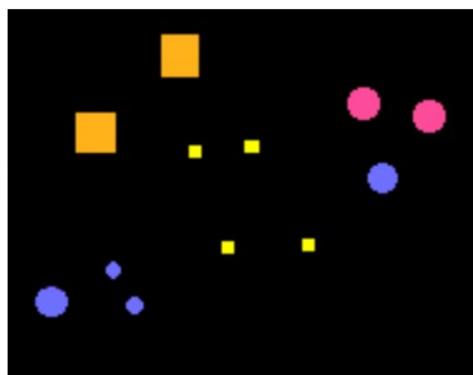
- Provide a descriptive title that summarizes the main takeaway of the chart.
- Reason: A clear title helps viewers quickly understand the purpose of the chart and the key insights it conveys.

##### Reasons for Changes:

- The stacked bar chart provides a clearer representation of the total sales for each company and the distribution of sales channels.
- Consistent color usage improves visual clarity and makes it easier to interpret the chart.
- Axis labels and a legend enhance understanding by providing necessary context and explanation.
- A descriptive title summarizes the main message of the chart, aiding in quick comprehension by the audience.

##### **Question:**

In the below visual, use gestalt principles of visual perception to identify at least 4 groups. Clearly mention which principle is used to group the marks.



##### **Answer:**

In the given visual, the gestalt principle of similarity is used to group the marks. The principle of similarity states that elements that are similar in appearance tend to be perceived as a group.

Based on this principle, the marks in the visual can be grouped into at least four distinct clusters:

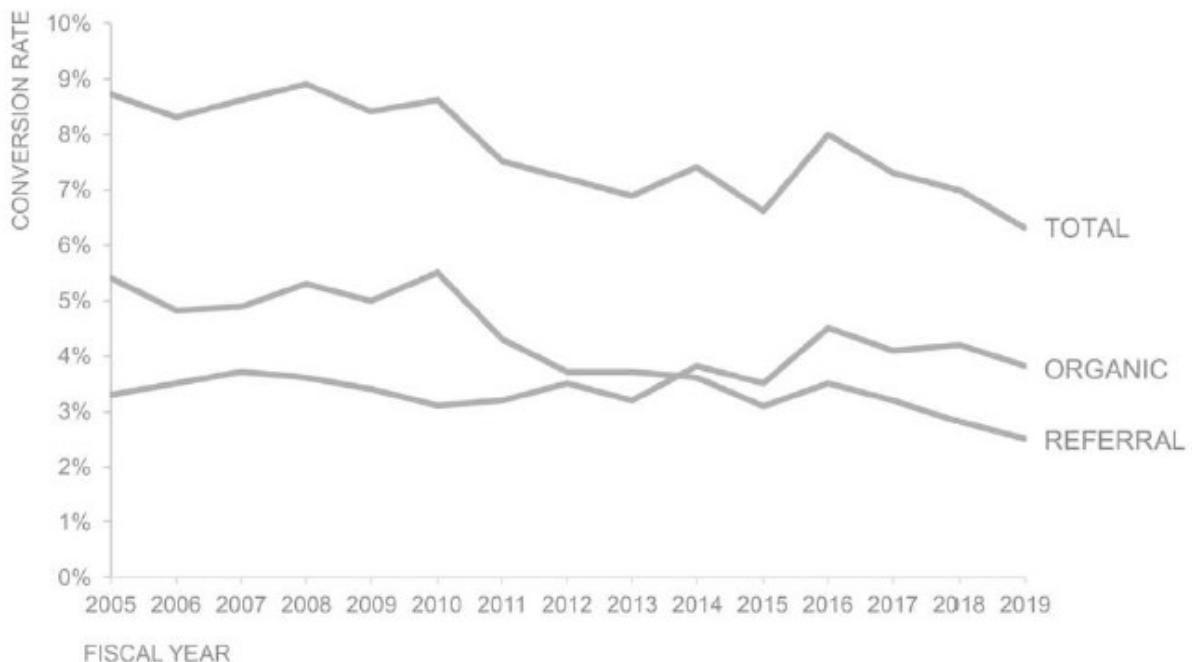
1. The pink circles form one group because they share the same color and shape.
2. The blue circles form another group because they share the same color and shape.
3. The yellow squares form a group due to their similarity in color and shape.
4. The purple circles form a separate group because they share the same color and shape.

Each group is perceived based on the similarity of the visual attributes of the marks, demonstrating the principle of similarity in action.

**Question:**

The below visual shows conversion rate over time by acquisition channel. Assume you'd like to draw your audience's attention to the **REFERRAL** line. How could you use preattentive attributes to do so? Come up with different visuals to focus your audience's attention[1 mark each].

Conversion rate over time



**Answer:**

To draw the audience's attention to the **REFERRAL** line in the given visual, we can use preattentive attributes effectively. Here are three different visuals focusing on the **REFERRAL** line, each utilizing a distinct preattentive attribute:

1. **Color**: Highlight the **REFERRAL** line in a distinct color compared to the other lines. For example, use a vibrant color like red or orange for the **REFERRAL** line, while keeping the other lines in a neutral color like gray. This contrast in color will immediately draw attention to the **REFERRAL** line.

2. **Thickness**: Increase the thickness of the **REFERRAL** line compared to the other lines. By making the **REFERRAL** line thicker, it will stand out prominently among the other lines, capturing the audience's attention.

3. \*\*Pointing Marker\*\*: Add distinctive markers, such as arrows or circles, along the REFERRAL line at key data points. These markers will act as visual cues, guiding the audience's eyes towards the REFERRAL line and emphasizing its significance in the visualization.

Each of these approaches leverages preattentive attributes to effectively highlight the REFERRAL line and focus the audience's attention on it within the visual.