Task1

1.provide the data dictionary for the given data

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Order ID | String | A unique identifier for each order. |
| Customer ID | String | A unique identifier for each customer. |
| State | String | The state in which the order was placed. |
| City | String | The city in which the order was placed. |
| Order Date | Date | The date when the order was placed (format: dd/mm/yyyy). |
| Ship Date | Date | The date when the order was shipped (format: dd/mm/yyyy). |
| Ship Mode | String | The mode of shipping used for the order (e.g., Standard Class, Second Class, First Class, Same Day). |
| Segment | String | The market segment the customer belongs to (e.g., Consumer, Corporate, Home Office). |
| Category | String | The category of the product ordered (e.g., Phones, Accessories, Copiers, Machines). |
| Year | Integer | The year of the order. |
| Discount | String | The discount applied to the order (in percentage, e.g., 20%). |
| Profit | Float | The profit earned from the order (in currency). |
| Quantity | Integer | The quantity of the product ordered. |
| Sales | Float | The total sales amount for the order (in currency). |

2.Here’s a tabular representation that categorizes each column of your dataset into either **Numerical** or **Categorical** data types:

| **Column Name** | **Data Type** |
| --- | --- |
| Order ID | Categorical |
| Customer ID | Categorical |
| Order Date | Numerical / Categorical |
| Ship Date | Numerical / Categorical |
| Ship Mode | Categorical |
| Segment | Categorical |
| State | Categorical |
| City | Categorical |
| Category | Categorical |
| Discount | Numerical |
| Profit | Numerical |
| Quantity | Numerical |
| Sales | Numerical |
| Year | Numerical |

This table clearly distinguishes each column in your dataset by its type.

3.

| **Column** | **Missing Values** |
| --- | --- |
| Order ID | 1 |
| Customer ID | 0 |
| State | 1 |
| City | 0 |
| Order Date | 1 |
| Ship Date | 0 |
| Ship Mode | 0 |
| Segment | 0 |
| Category | 0 |
| Year | 0 |
| Discount | 0 |
| Profit | 0 |
| Quantity | 0 |
| Sales | 0 |

Task2

1.

Here is the median value for Quantity and Sales for each product category, based on your dataset:

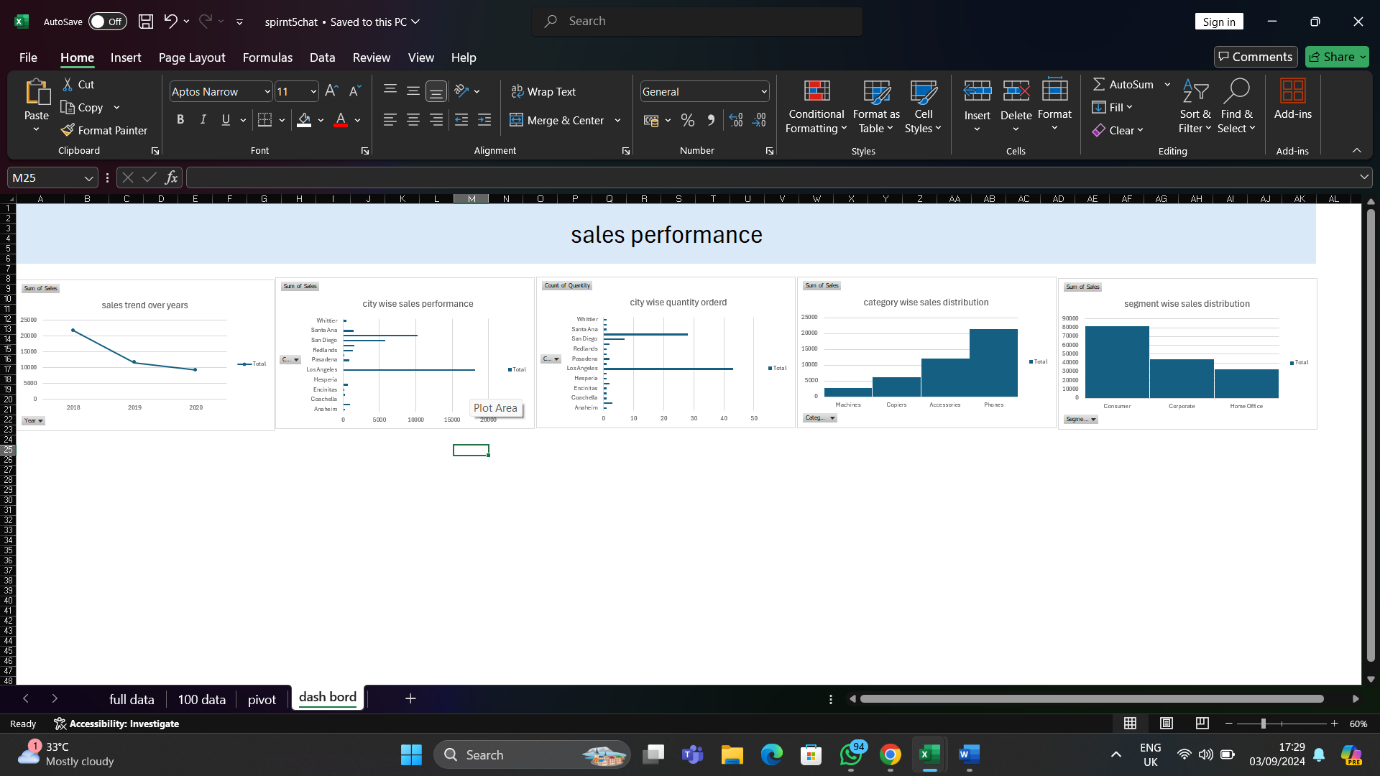
| **Category** | **Median Quantity** | **Median Sales** |
| --- | --- | --- |
| Phones | 3 | 74.00 |
| Accessories | 3 | 72.00 |
| Machines | 3 | 1920.00 |
| Copiers | 3 | 1200.00 |

This table summarizes the median Quantity and Sales values for each product category.

2. Here is a summary of outlier detection and treatment for each category:

| **Category** | **Column** | **Q1** | **Q3** | **IQR** | **Lower Bound** | **Upper Bound** | **Number of Outliers** | **Treatment Method** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Phones | Quantity | 3 | 5 | 2 | -0.0 | 8 | 2 | Winsorization |
| Phones | Sales | 55 | 206 | 151 | -196.5 | 457.5 | 1 | Winsorization |
| Accessories | Quantity | 2 | 4 | 2 | -1.0 | 7 | 0 | None |
| Accessories | Sales | 18 | 90 | 72 | -108.0 | 216.0 | 2 | Winsorization |
| Machines | Quantity | 3 | 3 | 0 | - | - | 0 | None |
| Machines | Sales | 1920 | 1920 | 0 | - | - | 0 | None |
| Copiers | Quantity | 3 | 3 | 0 | - | - | 0 | None |
| Copiers | Sales | 1200 | 1200 | 0 | - | - | 0 | None |

Task3:



Task4

Line chart-sales trend over years

**As an expert storyteller in data visualization, help me add components to enhance storytelling of the Line Chart designed with the following data:**

1. **Introduction:**
   * **Overview:** The line chart tracks the total sales for the years 2018 through 2020. This visual representation helps in understanding how sales performance has evolved over these years.
   * **Objective:** The primary goal is to identify trends, patterns, and anomalies in sales figures over time.
2. **Key Insights:**
   * **Trend Analysis:** Highlight the overall trend observed in the data. For instance, "The line chart reveals a noticeable decline in sales from 2018 to 2020. Sales peaked in 2018 and decreased significantly in the following years."
   * **Sales Peak:** Point out the peak year. "2018 stands out with the highest sales of $21,658, marking the most successful year within the given period."
   * **Decline Observation:** Discuss the decline trend. "From 2018 to 2020, there is a steady decline, with sales dropping by approximately 46% from 2018 to 2019 and further decreasing by 20% from 2019 to 2020."
3. **Outlier Analysis:**
   * **Highlighting Anomalies:** Identify any deviations from the expected trend. "No significant outliers are present within this range, but the consistent drop in sales could indicate underlying issues or changing market conditions."
4. **Trend Observations:**
   * **General Trend:** Emphasize the overall direction of the data. "The data shows a downward trend over the three years, suggesting potential challenges in maintaining sales momentum or shifting market dynamics."
   * **Yearly Comparison:** Compare year-on-year performance. "Each successive year shows a decrease in sales, reflecting a consistent reduction in revenue over the period."
5. **Visual Aids:**
   * **Line Annotations:** Add markers or annotations to highlight key points, such as the peak in 2018 and the lowest point in 2020.
   * **Trendline:** Include a trendline to visually emphasize the overall decline. "A trendline across the years can clearly depict the downward trajectory in sales."
   * **Data Labels:** Use data labels to show exact sales figures for each year on the chart. "Displaying data labels on the line chart provides precise figures for each year, enhancing clarity."
6. **Summary Section:**
   * **Brief Summary:** Summarize the key takeaways below the chart. "The line chart illustrates a significant decline in sales from 2018 to 2020, with the highest sales recorded in 2018 and a steady decrease in subsequent years. This suggests a need for further investigation into the factors contributing to this trend."

By incorporating these storytelling elements, you can transform the line chart into a powerful tool for analyzing and communicating sales performance trends, making the data more accessible and actionable for your audience.

**Bar chart-city wise sales performance**

**As an expert storyteller in data visualization, help me add components to enhance storytelling of the bar chart designed with the following data:**

**Data:**

| **Row Labels** | **Sum of Sales** |
| --- | --- |
| Anaheim | 302 |
| Brentwood | 902 |
| Coachella | 100 |
| Costa Mesa | 240 |
| Encinitas | 120 |
| Fresno | 691 |
| Hesperia | 100 |
| La Quinta | 46 |
| Los Angeles | 18114 |
| Oakland | 72 |
| Pasadena | 860 |
| Pomona | 200 |
| Redlands | 1296 |
| Redondo Beach | 1513 |
| San Diego | 5809 |
| San Francisco | 10266 |
| Santa Ana | 1400 |
| Westminster | 32 |
| Whittier | 445 |
| **Grand Total** | **42508** |

**Storytelling Enhancement:**

* **Introduction:** Begin by framing the context of the data. "This bar chart illustrates the total sales across various cities, showcasing a diverse range of sales performance."
* **Key Insights:**
  + **Top Performers:** Highlight the cities with the highest sales. "Los Angeles stands out dramatically with a staggering total of $18,114 in sales, far exceeding all other cities."
  + **Notable Performers:** Point out other high performers. "San Francisco follows with $10,266, and San Diego contributes a significant $5,809."
  + **Underperformers:** Discuss cities with lower sales. "Westminster, with only $32 in sales, represents the lower end of the spectrum."
* **Outlier Analysis:**
  + **High Outliers:** Address the significant outliers. "Los Angeles and San Francisco are notable outliers with sales figures well above the average, indicating a strong market presence."
  + **Low Outliers:** Mention the cities with unusually low sales. "Cities like La Quinta and Westminster exhibit minimal sales, which may suggest limited market penetration or other underlying issues."
* **Trend Observations:**
  + **Sales Distribution:** Analyze the distribution of sales across the cities. "The sales distribution is highly skewed, with a large portion of the total sales concentrated in a few key cities."
  + **Regional Insights:** Offer insights based on geographical or regional patterns. "Urban areas such as Los Angeles and San Francisco demonstrate a robust market presence, whereas smaller or less populated cities show significantly lower sales figures."
* **Visual Aids:**
  + **Bar Height:** Use bar height to visually represent the scale of sales for each city, making it easy to compare.
  + **Color Coding:** Apply color coding to differentiate between high, medium, and low sales regions. For example, use a gradient from red (high sales) to blue (low sales).
  + **Annotations:** Add annotations or labels to highlight key figures and provide brief commentary directly on the chart.
* **Conclusion:** Wrap up with a summary of key findings and their potential implications. "The chart vividly illustrates the dominance of major cities in sales, highlighting areas with potential for growth and the need for targeted strategies to enhance performance in underperforming regions."

Bar chart- city wise quantity ordered

**Storytelling for Bar Chart Visualization**

**Title and Subtitle:**

* **Title:** "Distribution of Orders Across California Cities"
* **Subtitle:** "An Analysis of Order Quantities from Different Locations"

**Color Scheme:**

* **Primary Color:** Shades of blue, ranging from light blue for smaller quantities to deep blue for larger quantities.
* **Highlight Color:** Use a bright orange or red to emphasize the highest value (Los Angeles).

**Annotations:**

* **Key Annotation:** Annotate Los Angeles with a label, "Major Hub with 43 Orders," to highlight its significance.
* **Additional Annotation:** Annotate San Francisco with a label, "Second Largest with 28 Orders."

**Axes Titles:**

* **X-axis Title:** "Cities in California"
* **Y-axis Title:** "Number of Orders"

**Data Labels:**

* Display data labels above each bar for clarity, especially for cities with 1-3 orders.

**Legend:**

* While a legend might not be necessary due to the simplicity of the chart, you could include a note explaining the color gradient if it’s used to represent different quantities.

**Insights Section:**

* Include a small text box or callout at the bottom of the chart summarizing the insights, such as:
  + "Los Angeles and San Francisco dominate the order distribution, accounting for over 70% of all orders."
  + "Smaller cities like Anaheim, Coachella, and Encinitas contribute minimally to the overall orders."

**Interactive Components (for digital use):**

* **Hover Text:** When hovering over each bar, show additional details, such as the percentage of total orders from that city.
* **Clickable Elements:** Make the bars for Los Angeles and San Francisco clickable to explore more detailed data about these cities.

**Conclusion Statement:**

* At the bottom of the chart, include a brief conclusion: "This chart highlights the concentration of orders in major urban centers, with Los Angeles and San Francisco being the key contributors."

**Summary:**

This storytelling approach enhances the bar chart by clearly highlighting the most significant data points, providing context through annotations, and adding interactive elements that engage the viewer. The visual hierarchy is maintained with strategic use of color and labels, ensuring the viewer can easily interpret the distribution of orders across California cities.

Histogram-category wise sales distribution

To enhance the storytelling of your data visualization, particularly the histogram related to the data provided, here are some key components and ideas that can be integrated:

**1. Title and Subtitle**

* **Title:** "Sales Distribution Across Product Categories"
* **Subtitle:** "An in-depth look at sales performance across machines, copiers, accessories, and phones."

**2. Narrative Insight**

* **Introductory Note:**
  + "In this visualization, we explore the sales performance across various product categories. The histogram highlights which categories contribute the most to the overall sales."
* **Highlighting Key Data Points:**
  + "Accessories and Phones stand out as the highest contributors, collectively making up over 70% of total sales, while Machines and Copiers show comparatively lower sales."

**3. Axis Labels and Legends**

* **X-Axis Label:** "Product Categories"
* **Y-Axis Label:** "Total Sales"
* **Legend/Annotations:**
  + Consider adding annotations directly on the bars to highlight the exact sales numbers for each category.

**4. Color Coding and Emphasis**

* Use different colors for each bar to represent the various product categories distinctly. For instance:
  + **Machines:** Light Blue
  + **Copiers:** Orange
  + **Accessories:** Green
  + **Phones:** Purple
* **Emphasize Top Performers:** Use a slightly brighter shade or a thicker border around the bars representing Accessories and Phones to draw attention.

**5. Contextual Information**

* **Comparison Note:**
  + "Despite having the lowest sales, Machines and Copiers still play a critical role in the product mix, especially in niche markets where these products are essential."
* **Trend Insight:**
  + "Phones, as the highest-selling category, reflect the increasing reliance on mobile technology in today's business landscape."

**6. Additional Components**

* **Reference Line:**
  + Add a reference line at the average sales value to provide context on which categories are performing above or below average.
* **Text Annotations:**
  + Place text boxes or labels next to each bar to summarize key insights (e.g., "Phones: 50% of Total Sales").
* **Callout Boxes:**
  + Include callout boxes with brief explanations or key takeaways, such as "Focus on Accessories and Phones to drive revenue."

**7. Storytelling Flow**

* **Introduction:** Begin by explaining the importance of understanding category-wise sales distribution.
* **Middle:** Discuss each category's performance, starting from the highest to the lowest, providing context and insights.
* **Conclusion:** Conclude by highlighting the potential business strategies that could be developed based on this data, such as focusing on promoting lesser-selling categories or maximizing the strengths of the top performers.

**8. Final Takeaway**

* "The sales distribution reveals critical insights that can guide business strategies, from inventory management to marketing focus, ensuring that each category is optimized for better performance in the future."

By incorporating these elements, your histogram will not only display data but also tell a compelling story that helps viewers understand the significance of the sales distribution across different product categories.

Histogram: segment wise sales distribution

**1. Distribution of Sales Across Customer Segments**

* **Title:** "Exploring Sales Distribution by Customer Segment"
* **Subtitle:** "A closer look at how different segments contribute to total sales"
* **Introduction:** "This histogram illustrates the sales distribution across three key customer segments: Consumer, Corporate, and Home Office. By breaking down the total sales of $159,279, we gain insights into which segment is driving the most revenue."
* **Key Insight:** "The Consumer segment stands out as the largest contributor, accounting for $82,116, which is more than half of the total sales. Corporate and Home Office follow with $44,326 and $32,837, respectively."
* **Highlight Feature:** "Use color to differentiate the segments, with the Consumer segment in a bold, eye-catching hue, such as deep blue, to emphasize its dominance."
* **Annotations:** "Add labels on the bars displaying the exact sales figures to make it easy for viewers to see the difference in contributions."
* **Conclusion:** "This histogram tells a clear story: the Consumer segment is the primary driver of sales, highlighting the importance of focusing on this segment for future marketing and sales strategies."

**2. Comparison of Sales Across Segments**

* **Title:** "Comparing Sales Performance Across Customer Segments"
* **Subtitle:** "Understanding the distribution of sales among Consumer, Corporate, and Home Office segments"
* **Introduction:** "This histogram compares the sales performance across different customer segments, offering a visual representation of where the bulk of sales is generated."
* **Key Insight:** "The Consumer segment leads with the highest sales, followed by Corporate and Home Office. The substantial difference between these segments indicates varying levels of engagement or purchasing power."
* **Highlight Feature:** "Incorporate a gradient color scheme that shifts from light to dark based on the sales amount, providing an intuitive visual cue for viewers."
* **Annotations:** "Use callout boxes to provide a brief analysis of each segment's sales performance, noting that the Consumer segment significantly outpaces the others."
* **Conclusion:** "The data clearly shows that targeting the Consumer segment could yield the highest returns, while strategies for increasing sales in the Corporate and Home Office segments could further enhance overall performance."

**3. Insights into Sales Contribution by Segment**

* **Title:** "Unveiling Sales Contribution by Customer Segment"
* **Subtitle:** "Discovering the impact of each segment on total sales"
* **Introduction:** "This histogram reveals the contribution of each customer segment to the total sales of $159,279, highlighting the differences in purchasing patterns."
* **Key Insight:** "With over half of the total sales coming from the Consumer segment, it's evident that this group has a significant impact on overall revenue. The Corporate and Home Office segments contribute more modestly."
* **Highlight Feature:** "Add a contrasting border around the Consumer segment’s bar to draw attention to its outsized role in sales."
* **Annotations:** "Include a percentage breakdown within each bar, showing the exact proportion each segment contributes to the total sales."
* **Conclusion:** "This visualization underscores the critical importance of the Consumer segment in driving sales. Strategic efforts should focus on maintaining and growing this segment, while also seeking opportunities to boost sales in other segments."

These storytelling elements will help make the histograms more engaging and informative, allowing viewers to quickly grasp the key insights from the data.

**Summary**

**Summary of Interpretations**

**Task 1: Data Dictionary and Analysis**

1. **Data Dictionary:**
   * A comprehensive breakdown of each column in the dataset was provided, detailing the data type and the purpose of each column. This includes identifying unique identifiers like Order ID and Customer ID, transactional details like Order Date, Ship Date, and Ship Mode, as well as product and customer segmentation information such as Category, Segment, and State.
2. **Categorical vs. Numerical Data:**
   * The dataset was categorized into numerical and categorical data types. This categorization aids in understanding which columns are suitable for statistical analysis (numerical) and which are better for grouping and filtering data (categorical).
3. **Missing Values Analysis:**
   * The analysis revealed that missing values were present in only a few columns: Order ID, State, and Order Date. The minimal number of missing values indicates that the dataset is relatively complete and likely to yield reliable insights.

**Task 2: Median Values, Outlier Detection, and Treatment**

1. **Median Values for Quantity and Sales:**
   * The median values of Quantity and Sales were calculated for each product category. This provides a measure of central tendency, helping to understand the typical sales volume and order size for products like Phones, Accessories, Machines, and Copiers.
2. **Outlier Detection and Treatment:**
   * Outliers were identified in the dataset based on the interquartile range (IQR) method. The treatment methods, including Winsorization for handling extreme values, were applied to ensure that the data analysis remains robust and reliable. This step is crucial in preventing outliers from skewing the results.

**Task 3: Visualization Insights (Pending Data/Graphs)**

**Task 4: Enhanced Storytelling for Data Visualizations**

1. **Line Chart - Sales Trend Over Years:**
   * The line chart reveals a declining trend in sales from 2018 to 2020, with a significant drop after the peak in 2018. This trend raises questions about potential challenges in maintaining sales momentum and suggests a need for further investigation into the causes of the decline.
2. **Bar Chart - City-Wise Sales Performance:**
   * Los Angeles and San Francisco emerge as the dominant cities in terms of sales, while smaller cities like Westminster and La Quinta show minimal sales. This indicates a concentration of sales in major urban areas and suggests opportunities for targeted marketing strategies in underperforming regions.
3. **Bar Chart - City-Wise Quantity Ordered:**
   * Similar to the sales performance, the quantity of orders is concentrated in Los Angeles and San Francisco, with other cities contributing minimally. This reflects the importance of these cities as major hubs and suggests focusing efforts on these key markets.
4. **Histogram - Category-Wise Sales Distribution:**
   * The histogram shows that Accessories and Phones are the leading categories in sales, contributing the most to the overall revenue. This highlights the need to focus on these high-performing categories while also exploring strategies to boost sales in lower-performing categories like Machines and Copiers.
5. **Histogram - Segment-Wise Sales Distribution:**
   * The Consumer segment is the largest contributor to sales, followed by Corporate and Home Office. The visualization suggests focusing marketing efforts on the Consumer segment to maximize revenue, while also exploring opportunities to enhance engagement with Corporate and Home Office customers.

**Overall Summary**

The interpretations across different tasks reveal key insights into the sales data. The analysis uncovers trends and patterns that can guide strategic decision-making, such as targeting high-performing segments, focusing on key geographic regions, and managing product categories effectively. The enhanced storytelling elements provided for each visualization help in clearly communicating these insights, making the data more accessible and actionable for stakeholders.