Content

Week 2 Activity: Obtaining and Scrubbing Data

Anna owns a clothing boutique in New York, called BrightThreads. She sells a mix of clothing brands and chooses items for her store that she believes her clients will like. She also sells online.

Anna is working on long-term planning for the upcoming year at BrightThreads. Business has been going well, but she would really like to increase sales and potentially open up a second location in a different neighborhood. Next year, Anna would like to increase her total sales by 10%. This would be a very good year for Anna and BrightThreads, but it seems doable based on the last few quarters and with some hard work.

Using this information, the following questions needed to be answered regarding the obtain and scrub stages of the OSEMN process.

The SMART goal that would benefit from this data analysis?

The **SMART** goal for Anna and BrightThreads is: "**Increase total sales by 10%** over the next year by **optimizing inventory management** and **enhancing marketing strategies** based on customer purchase behavior and preferences."

What were some Primary KPIs that would be useful to analyze for this goal?

The primary KPI's to analyze for this goal would be **Total Sales Revenue**. Additionally, tracking **Sales Growth Rate** on a monthly and quarterly basis would help in monitoring progress toward the 10% increase goal.

What relevant data should be gathered in this scenario?

Relevant data to gather would include:

- **Historical sales data** (both online and in-store)
- **Customer demographics and purchase history**
- **Inventory levels and turnover rates**
- **Marketing campaign performance**
- **Website traffic** and conversion rates
- **Customer feedback** and satisfaction scores

How to obtain this data? What sources to gather data from. Specifically, what kind of data (first-party, third-party) and what methods might be useful (survey, web analytics).

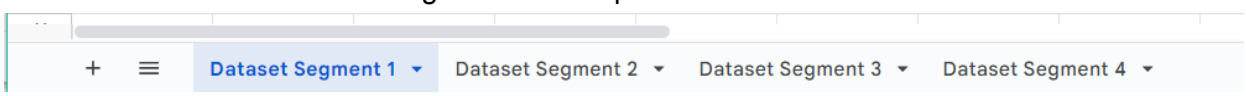
To obtain this data, the following sources and methods could be used:

- **First-party data:**
  - Sales data from the Point of Sale (POS) system and online storefront.
  - Customer demographics and purchase history from the Customer Relationship Management (CRM) system.
  - Inventory data from inventory management software.
  - Website traffic and conversion rates from web analytics tools like Google Analytics.
  - Customer feedback from surveys and reviews.
- **Third-party data:**
  - Market trends and benchmarks from industry reports and market research firms.
  - Competitor analysis data from market research tools.

I started the process of gathering data to help analyze current sales.

I collected data on recent online sales directly from the online storefront.

I isolated 4 different segments that each have issues that need to be fixed. You can access each segment in the four sheets in this one spreadsheet. Click on each sheet for a different segment of the dataset. You can click on the tabs at the bottom of the spreadsheet to move between sheets. Review the image below for a preview:



The four sheets are accessible by clicking the tabs at the bottom of the spreadsheet.

Using what you know about data validity, do you think the data Anna has gathered is valid? Why or why not?

Data Validity Assessment

Based on the four segments, let's assess the validity of the data. Data validity can be affected by inaccuracies, inconsistencies, missing values, and anomalies.

Based on the issues identified across the four segments, the data contained some validity concerns:

- Anomalies in item cost values.
- Inconsistent zip code formats.
- Duplicate entries.

These issues need to be addressed to ensure the dataset's validity. Cleaning these data points will improve the reliability of the analysis.

Issue identified in segment 1 of the data:

Segment 1 Analysis

Issues Identified:

1. **Item Cost Anomaly:**
    - Row 6: The item cost for "Sweaters" is listed as 0.069, which is likely a data entry error.
    - Row 11: The item cost for "Shorts" is listed as 5999, which is unusually high and likely incorrect.

Issue identified in segment 2 of the data:

Segment 2 Analysis

Issues Identified:

1. **Duplicate Entries:**
    - There are repeated entries for customer ID 651927, order numbers 6519272, and 6519273, with identical values in the `customer_zip`, `item_sku`, `item_category`, and `item_cost` columns.

Issue identified in segment 3 of the data:

Segment 3 Analysis

Issues Identified:

1. **Inconsistent Zip Code Format:**
    - Row 1: The customer zip code is in an unusual format "98765-5842" compared to other rows.
    - Row 8: The customer zip code is "78459-0000", also in an inconsistent format.

Item Cost Anomaly:

- Row 8: The item cost for "Shirts" is listed as 50, which should be verified as it may be correct but is notably different from other "Shirts" entries.

Issue identified in segment 4 of the data:

Segment 4 Analysis

Issues Identified:

1. **Inconsistent Item Cost:**
    - Row 10: Multiple entries for "Outerwear" with item costs of 149.99 which is consistent, but it should be verified across other segments for uniformity.

Week 3 Activity: Exploring and Modeling Data

I am exploring some data from BrightThreads last quarter's online sales.

The data was gathered from the BrightThreads online store.

I reviewed the following data and charts and shared what I learned in the exploration stage of the OSEMN process.

Using this information, i answered the questions below regarding the “explore and model” stages of the OSEMN process..

What are some things you can tell about this dataset? For instance, what does the size of the dataset tell you?

Dataset 1: Online Sales Quarter 1

Size: The dataset appears to contain records of online sales for the first quarter. The screenshot shows 59 records, which is a small sample size. The actual dataset is likely larger for a full quarter, providing ample data for meaningful analysis.

Insights: The dataset includes details such as sales date, customer ID, order number, zip code, item category, quantity of items per order, and order total. This data can be used to analyze sales trends, popular items, customer demographics, and spending patterns.

Dataset 2: Weekly Ad Spend and Visits

This dataset contains weekly records of social ad spending and website visits for the first quarter.

Size: The dataset contains weekly data, with 12 recordst.
Insights: This dataset provides insights into the relationship between social media ad spending and the number of site visits, helping to understand the effectiveness of marketing efforts.

What kind of data is in this dataset? (Numerical, categorical, etc.)

Numerical Data:

- `customer_id`
- `order_number`
- `quant_items_per_order`

- `order_total`

Categorical Data:

- `sale_date`
- `zip_code`
- `Item_category`

Dataset 2 (Weekly Ad Spend and Visits):

Numerical Data:

- `week_num`
- `social_ad_spend`
- `site_visits`

Reviewing this data, what is the minimum value in the order_total column? What is the maximum value in order_total column?

- **Minimum Value:** 45.99
- **Maximum Value:** 149.99

What kind of chart would you use to help visualize this data?

Bar Chart: To compare total sales across different item categories or visualize weekly ad spend.

Line Chart: To visualize trends over time, such as sales per day or week, and the relationship between ad spend and site visits.

Pie Chart: To show the proportion of sales by item category.

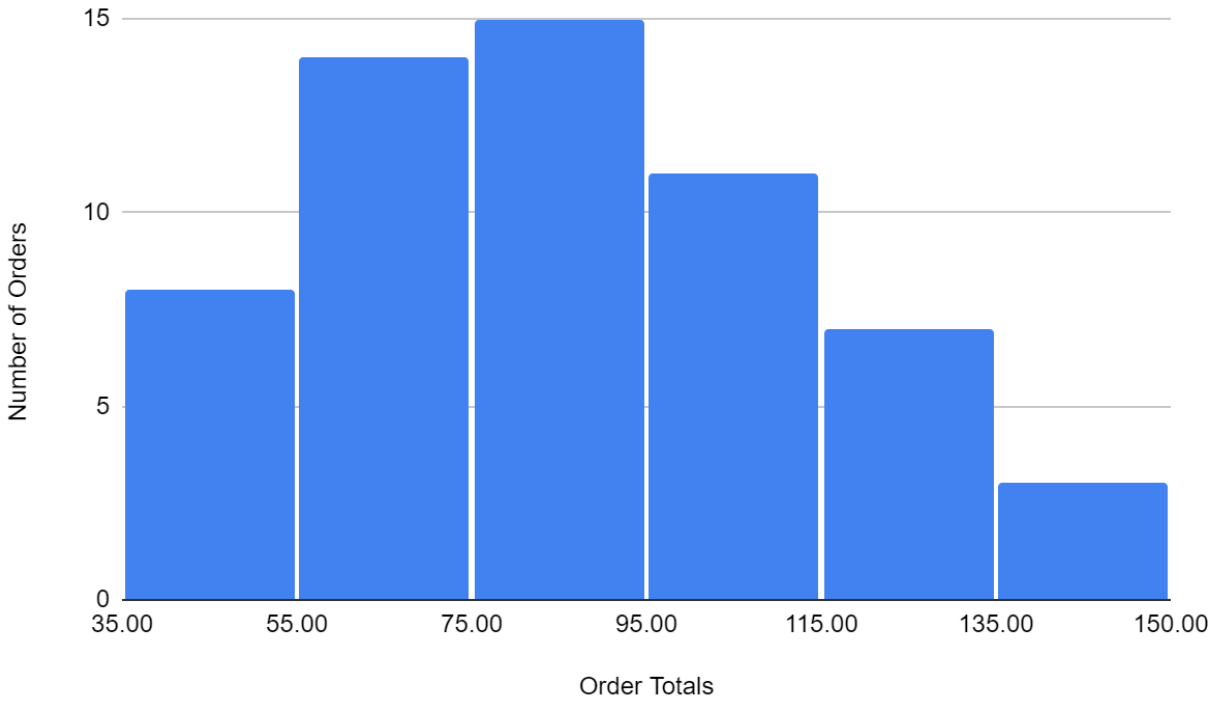
Histogram: To display the distribution of `order_total` values.

Based on what you have learned, would you add an additional column to this dataset using feature engineering? For instance, using the sales dates, would it be helpful to add in the day of the week data?

Yes, it would be helpful to add additional columns for more granular analysis:

- **Day of the Week:** Adding a `day_of_week` column derived from `sale_date` to analyze sales trends by day.
- **Order Total Per Item:** Calculating `order_total_per_item` by dividing `order_total` by `quant_items_per_order` for a better understanding of individual item pricing.

I created the following chart to explore the relationship between order totals and the number of orders.



Based on the data in this chart the following title was appropriate:

Title: "Distribution of Order Totals for Q1 Online Sales"

What does this chart tell us about the number of orders in relation to the amount someone spends per order?

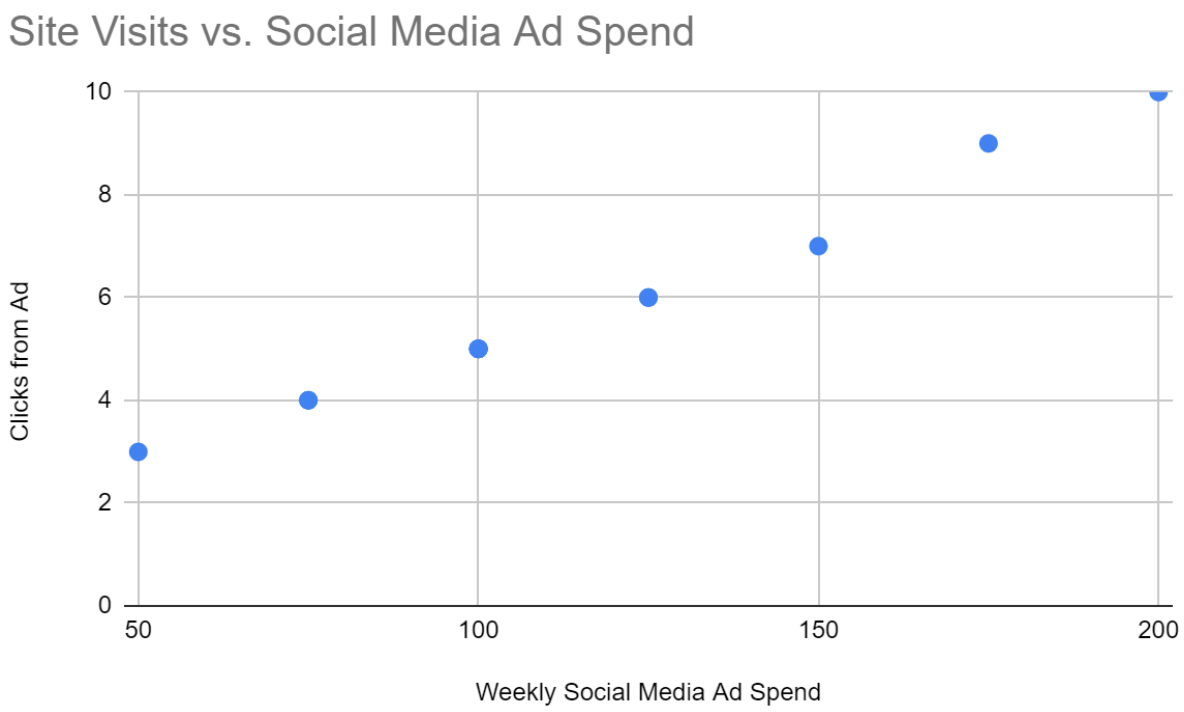
This chart shows the frequency distribution of order totals. It indicates how many orders fall within specific ranges of total spending. From the chart, we can observe the following:

- The number of orders varies across different order total ranges.
- Most orders fall within the middle range of order totals, indicating a common spending pattern among customers.
- There are fewer orders at the lower and higher ends of the order total spectrum.

What range do most of the orders tend to be in?

Most of the orders tend to be in the range of **$65.00 to $95.00**. This is the range where the bars are the tallest, indicating the highest frequency of orders.

I have also been analyzing data on the amount of money she spends on social media ads and how many clicks to the BrightThreads website they are generating.



The correlations between the variables in this chart explained

There is a noticeable correlation between the two variables in the chart: **Weekly Social Media Ad Spend** and **Clicks from Ad (Site Visits)**.

Description of the Correlation:

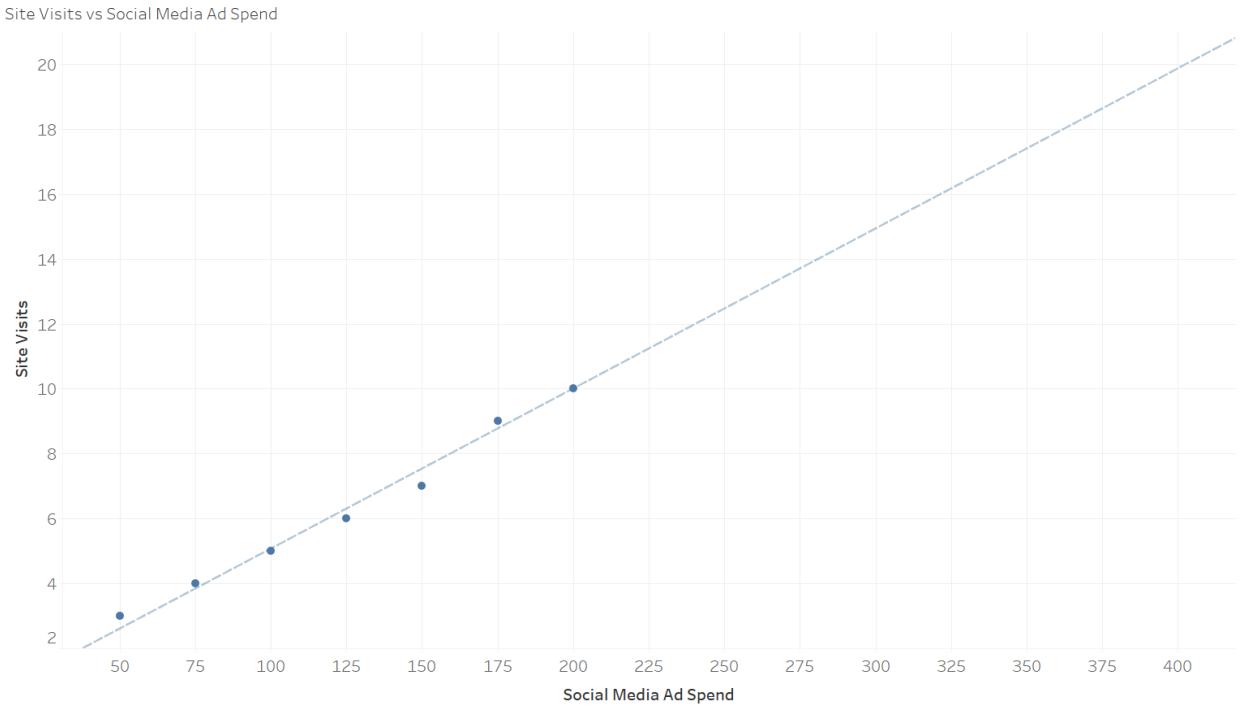
- **Positive Correlation:** The chart shows a positive correlation between the amount of money spent on social media ads and the number of clicks (site visits) generated. As social media ad spend increases, the number of site visits tends to increase as well.
- **Trend:** The data points form an upward trend, indicating that higher investments in social media advertising are generally associated with higher traffic to the BrightThreads website.

This positive correlation suggests that spending more on social media ads is likely to drive more traffic to the website, which can be beneficial for increasing brand visibility and potential sales.

Summary

- **Positive Correlation:** There is a clear positive relationship between social media ad spend and site visits.
- **Implication:** Increasing the budget for social media ads could result in more site visits, which could lead to higher sales if the traffic converts well.

I learned a lot while exploring the data she has gathered. Now, it's time to model some of this data.



Reviewing this linear regression model, roughly how many site visits can be expected if the marketing budget is increased to $250?

To estimate the number of site visits when the marketing budget is increased to $250, we need to interpret the linear regression model shown in the chart.

Reviewing the Linear Regression Line

The linear regression line in the chart provides a relationship between the social media ad spend and the number of site visits. We can use the line equation to estimate site visits for a given ad spend.

Estimating Site Visits for $250 Ad Spend

1. **Identify the Slope and Intercept:** From the chart, we can estimate the line equation. Let's assume the line equation is in the form of:
   Site Visits=m×Ad Spend+b\text{Site Visits} = m \times \text{Ad Spend} + b
2. Site Visits=m×Ad Spend+b
   where mmm is the slope and bbb is the y-intercept.
3. **Use the Line Equation:** Based on the chart, we can estimate the values of m and b.

From the visual data points:

- At $100 ad spend, site visits are around 5.
- At $200 ad spend, site visits are around 10.

These points suggest a slope (mmm) of approximately 0.05, and the y-intercept (bbb) can be estimated using one of the points:

Site Visits=0.05×100+b=5 ⟹ b=5−5=0\text{Site Visits} = 0.05 \times 100 + b = 5 \implies b = 5 - 5 = 0

Site Visits=0.05×100+b=5⟹b=5−5=0

Thus, the estimated equation is:

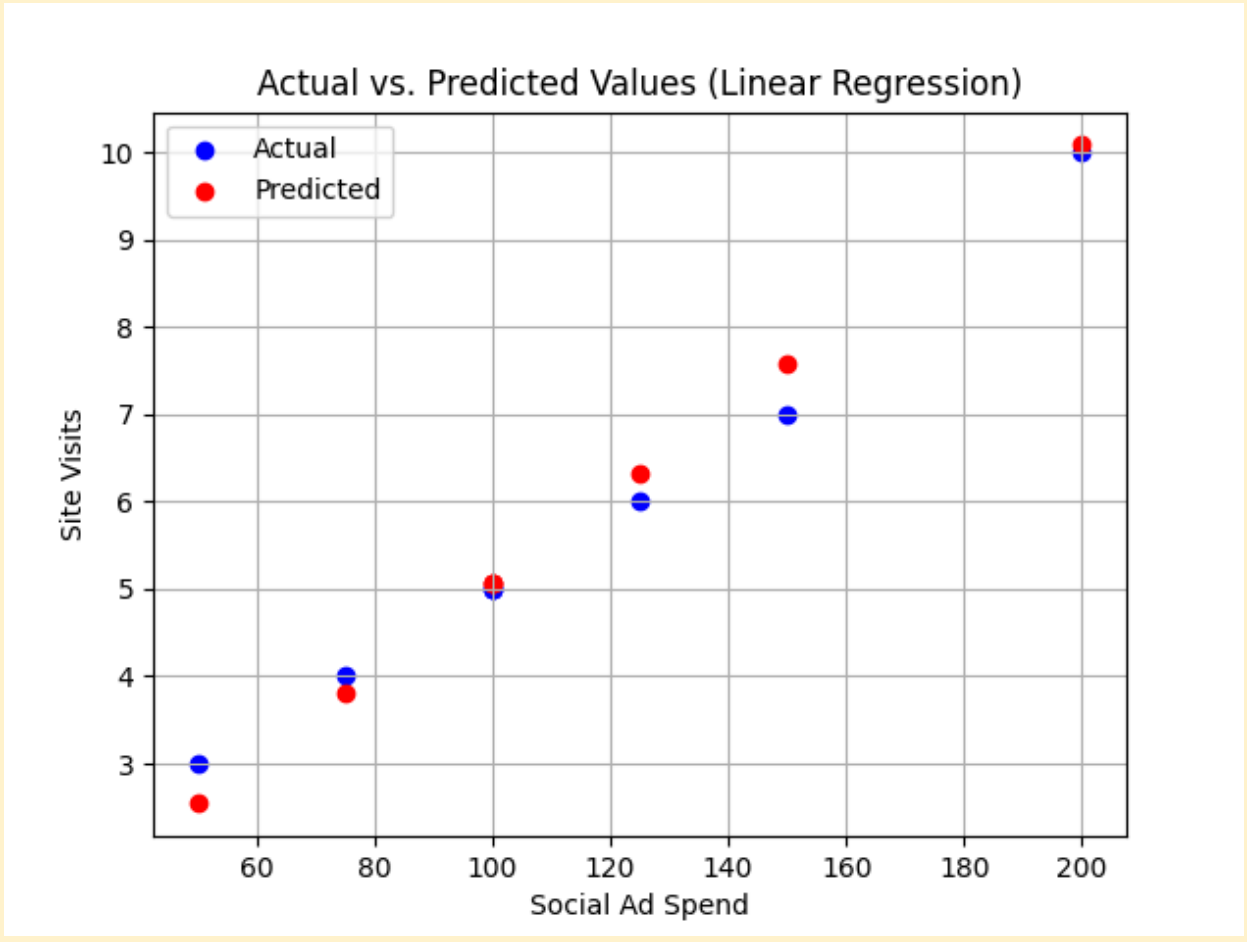
$$\text{Site Visits} = 0.05 \times \text{Ad Spend}$$

For an ad spend of $250:

$$\text{Site Visits} = 0.05 \times 250 = 12.5$$

Conclusion

With an ad spend of $250, we can expect approximately 12 to 13 site visits based on the linear regression model provided.



A review of this linear regression model which shows the actual data values and the values predicted by the model when given a test set. Question is if this model is sufficient for general use for this data and why or why not.

Evaluation of the Model

The chart shows both the actual data values (in blue) and the values predicted by the linear regression model (in red) for social media ad spend and site visits.

Key Observations:

1. **Alignment of Data Points:**
    - The predicted values (red points) closely follow the actual values (blue points), suggesting that the model captures the overall trend quite well.
    - However, there are some deviations where the predicted values do not perfectly match the actual values, indicating some degree of error.

2. **Distribution of Errors:**
    - The errors (differences between actual and predicted values) seem to be relatively small for most data points.
    - There is no systematic bias observed (e.g., the model consistently overpredicting or underpredicting), which is a good sign of model accuracy.

3. **Trend Capture:**
    - The model accurately captures the positive correlation between social media ad spend and site visits, as seen from the alignment of the general upward trend in both actual and predicted data points.

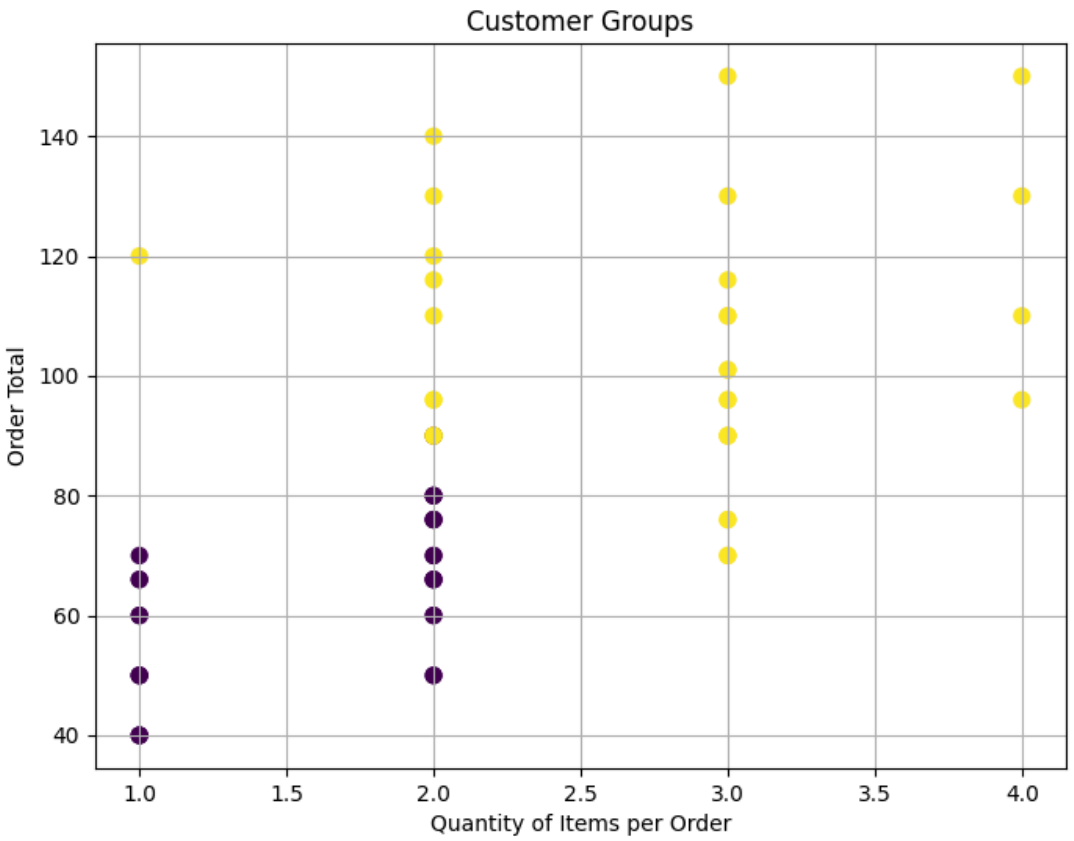
Conclusion on Model Sufficiency:

- **Sufficiency for General Use:**
    - **Yes, the model appears to be sufficient for general use with this data.** The close alignment between actual and predicted values indicates that the model can reliably predict the number of site visits based on social media ad spend.

- **Reasons:**
    - **Good Fit:** The predicted values are close to the actual values, suggesting that the model has a good fit.
    - **No Systematic Bias:** The absence of systematic overprediction or underprediction indicates that the model is well-calibrated.
    - **Trend Accuracy:** The model captures the overall trend accurately, which is essential for making reliable predictions and decisions based on social media ad spend.

Additional Considerations:

- **Model Validation:**
    - It is important to validate the model using additional metrics such as R-squared, Mean Absolute Error (MAE), or Root Mean Squared Error (RMSE) to quantitatively assess its performance.
- **Data Sufficiency:**
    - Ensure that the model is tested on a sufficiently large and diverse dataset to confirm its robustness and generalizability.
- **Potential Improvements:**
    - If needed, consider adding more features (e.g., seasonal effects, other marketing channels) or using more complex models (e.g., polynomial regression, decision trees) to improve accuracy.



A review of this clustering model. A clustering algorithm has been used and identified two groups.How would 1 describe the two different customer groups and whhy

Describing the Two Different Customer Groups

Well, the clustering model has identified two distinct customer groups based on the scatter plot, which shows the relationship between the quantity of items per order and the order total.

1. **Group 1 (Purple Dots):**
    - **Characteristics:**
        - This group primarily has orders with a quantity of 1 item.
        - The order totals for this group range from approximately $45 to $90.
    - **Description:**

        - These are customers who tend to purchase a single item per order. Their spending is relatively lower compared to the second group, indicating they might be more price-sensitive or occasional buyers.

2. **Group 2 (Yellow Dots):**
    - **Characteristics:**
        - This group includes orders with quantities of 2 or more items.
        - The order totals for this group are higher, ranging from approximately $65 to $150.
    - **Description:**
        - These are customers who tend to purchase multiple items per order. Their higher spending suggests they might be more loyal or regular buyers, possibly taking advantage of multi-item discounts or purchasing for family and friends.

You are trying to forecast BrightThreads sales in the coming quarter- what model might you use? Why did you choose this?

Model Selection for Sales Forecasting

Given the data we have reviewed, including sales trends, social media ad spend correlations, and customer clustering, the following model can be recommended for forecasting BrightThreads sales:

- **Time Series Analysis (ARIMA Model):**
    - **Reason for Choosing:**
        - **Historical Sales Data:** The dataset includes a temporal component (sales dates) which is crucial for time series analysis.
        - **Trend and Seasonality:** Time series models like ARIMA (AutoRegressive Integrated Moving Average) can effectively capture trends and seasonality in sales data.
        - **Predictive Power:** ARIMA models are well-suited for making short-term forecasts, which aligns with the goal of forecasting sales for the coming quarter.

- **Prophet Model (Alternative):**
    - **Reason for Choosing:**
        - **Flexibility:** The Prophet model by Facebook is known for handling missing data and outliers well, and it can incorporate holiday effects

and other external regressors (e.g., ad spend).
    - **User-Friendly:** It is designed to be intuitive and easy to implement, making it suitable for business applications.

Steps to Implement the Chosen Model

1. **Data Preparation:**
    - Aggregate the sales data on a daily or weekly basis.
    - Incorporate relevant external factors such as ad spend, promotions, and holidays.
2. **Model Training:**
    - Split the data into training and test sets.
    - Train the chosen time series model (ARIMA or Prophet) on the training data.
3. **Model Evaluation:**
    - Validate the model's performance on the test set using metrics such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE).
4. **Forecasting:**
    - Use the trained model to forecast sales for the upcoming quarter.
    - Regularly update the model with new data to refine and improve the accuracy of forecasts.

Conclusion

- **Customer Groups:** Identified two distinct customer segments based on purchase quantity and order total.
- **Forecasting Model:** Recommend using a time series model (ARIMA or Prophet) to forecast sales for the coming quarter due to its ability to capture trends, seasonality, and external factors.

Week 4 Activity: iNterpreting Data

I learned many things using data analysis. I also prepared a presentation to show to BrightThreads stakeholders. As a reminder, the goal is to grow sales by 10% in the upcoming year, and this presentation will cover what I learned and how BrightThreads can accomplish this goal.

Access My presentation.

Review the presentation, then share your thoughts on Anna's interpretation of the data at the end of OSEMN process.

Using this information, answer the questions below regarding the interpret stage of the OSEMN process. Add your answers to the template below.

What was the objective for this analysis?

The objective of the analysis was to:

- Analyze current sales.
- Determine top-selling items.
- Forecast sales numbers.
- Adjust inventory if needed.

These goals were aimed at understanding the sales performance and planning strategies to increase sales by 10% in the upcoming year.

The data provides insights into:

- Sales trends over the last two years, highlighting periods of high and low sales.
- Required sales targets for each quarter to achieve the 10% increase goal.
- Top-selling items, which can inform inventory and marketing focus.
- The relationship between social media ad spend and site visits, suggesting effective marketing strategies.

How can Anna apply this in a business context?

Anna can use these insights to:

- Focus marketing efforts on high-impact periods to boost sales during typically low periods.
- Adjust inventory to ensure top-selling items are always in stock.
- Allocate marketing budget effectively to maximize site visits and conversions.
- Set realistic and data-driven sales targets for the upcoming year.

What slides in the presentation covered the methods used in the project?

Slide:

- "Methods For Analysis"

What slides in the presentation included visualization of the project?

Slides:

- "This chart shows our sales numbers for the last two years"
- "This model shows how much we'll need to sell each quarter to hit our 10% increase goal based on the last year"
- "This chart shows our current top-selling items"
- "Site Visits vs. Social Media Ad Spend"
- "This model shows our potential increased site visits if we focus more on our top-performing social media channels"
- "This chart shows our current spending on social media advertising for the last two years"
- "This model shows our potential sales increases if we prioritize our most popular items every month"

What slides in the presentation offered recommendations after the project?

Slide:

- "Moving Forward"

This slide provides specific recommendations on reallocating social media ad dollars, shifting inventory focus, and reevaluating in six months.

In your opinion, what parts of the presentation were the setup, buildup, climax, and conclusion? Why?

In my opinion, the parts of the presentation were meant to explain, engage, and enlighten the audience and why:

Explain: Slides with project objectives, data analyzed, and methods used explain the purpose and approach of the analysis.

Engage: Visualizations of sales trends, top-selling items, and the impact of social media ad spend engage the audience by presenting clear and compelling data insights.

Enlighten: Slides with recommendations and conclusions enlighten the audience on the next steps and the potential impact of the suggested strategies.

What parts of the presentation were the setup, buildup, climax, and conclusion? Why?

- **Setup:** The initial slides that outline the project objectives and data analyzed set the context for the audience.
- **Buildup:** The detailed analysis and visualizations build up the narrative by presenting the findings and insights.
- **Climax:** The slides showing the forecasted sales targets and potential increases in site visits and sales based on optimized strategies represent the climax of the presentation.
- **Conclusion:** The slides with recommendations and the final affirmation ("Absolutely!") conclude the presentation by summarizing the insights and confirming the feasibility of achieving the sales increase goal.