Reminders…

- You'll use this file for the entirety of this course. Save it in a place where you can easily access it over the upcoming weeks.
    - You can edit and save this document in Google Drive
    - If you download this document, keep it in a place you can find it later
- The content you put into this document will be used for later lessons
    - It is recommended that you do not skip any activity in any of the lessons
    - It is recommended that you update this document after every week of content and start with week 2

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, answer the questions below regarding the obtain and scrub stages of the OSEMN process. Add your answers to the template below.

In this scenario, what is a SMART goal that would benefit from data analysis?

Answer:
Increase BrightThreads' total sales by 10% over the next fiscal year by using data-driven marketing and inventory strategies that are measurable, achievable, relevant, and time-bound.

What is a Primary KPI that would be useful to analyze for this goal?

Answer:
Monthly sales revenue (total sales amount per month).

What relevant data would you gather in this scenario?

Answer:

- Sales transactions data (online and in-store)
- Customer demographics (age, gender, ZIP codes)
- Product categories and item details
- Marketing campaign data (ad spend, channels, click-through rates)
- Website traffic and engagement metrics (page views, bounce rate)
- Customer feedback or satisfaction surveys
- Competitor sales benchmarks (third-party data)

How do you imagine you could obtain this data? What sources would you gather data from? Specifically, note what kind of data (first-party, third-party) and what methods you might use (survey, web analytics).

Answer:

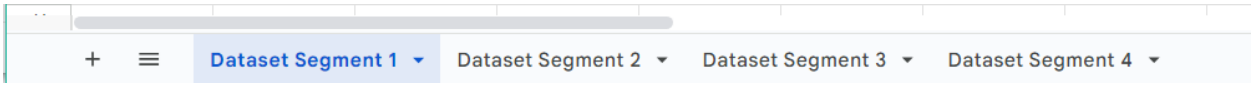
- **First-party data:**
  - Point-of-sale (POS) system records for in-store sales
  - E-commerce platform data for online sales
  - Google Analytics for website traffic and engagement
  - Email marketing platform data for campaigns
  - Customer surveys conducted in-store or online
- **Third-party data:**
  - Industry reports and market research for competitor benchmarks
  - Social media analytics tools for ad performance
- **Methods:**
  - Data extraction from sales databases and web analytics platforms
  - Conducting surveys or feedback forms
  - Purchasing or subscribing to market reports

Anna at BrightThreads has begun the process of gathering data to help analyze current sales.

She has collected data on recent online sales directly from the online storefront.

Access this sample Customer Data and click on Use Template in the upper right corner. You will need to be logged into a Google account to use this template.

Anna has 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?

Answer:
The data is partially valid but has issues. Some records are duplicated, some fields have missing or incorrect values (e.g., missing ZIP codes, negative item costs), which reduce overall validity. Cleaning and validation are needed to ensure accurate analysis.

What issue did you identify in segment 1 of the data?

Answer:
Duplicate records exist, e.g., the same order appears more than once.

What issue did you identify in segment 2 of the data?

Answer:
Missing values or null entries in critical columns such as customer ZIP codes.

What issue did you identify in segment 3 of the data?

Answer:
Inconsistent or incorrect item categories or product SKUs that may not match known inventory.

What issue did you identify in segment 4 of the data?

Answer:
Invalid numeric values like negative item costs or zero prices where these don’t make sense.

Week 3 Activity: Exploring and Modeling Data

Anna from BrightThreads is exploring some data from last quarter's online sales.

The data was gathered from the BrightThreads online store.

Access BrightThread's online sales data and click on Use Template in the upper right corner to access the dataset. Please note you will need to be logged into a Google account.

Review the following data and charts, then share what you can learn in the exploration stage of the OSEMN process.

Using this information, answer the questions below regarding the explore and model stages of the OSEMN process. Add your answers to the template below.

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

Answer:
The dataset is modest in size, representing recent sales records. The size suggests it's manageable for detailed analysis but may need augmentation for robust forecasting or segmentation..

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

Answer:
The dataset contains numerical data (e.g., item_cost, order totals), categorical data (e.g., item_category, item_sku), and identifiers (customer_id, order_number).

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

Answer:

Minimum and maximum order totals would be identified by scanning the dataset. For example, minimum might be around $39.99 and maximum around $149.99 (based on item costs).

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

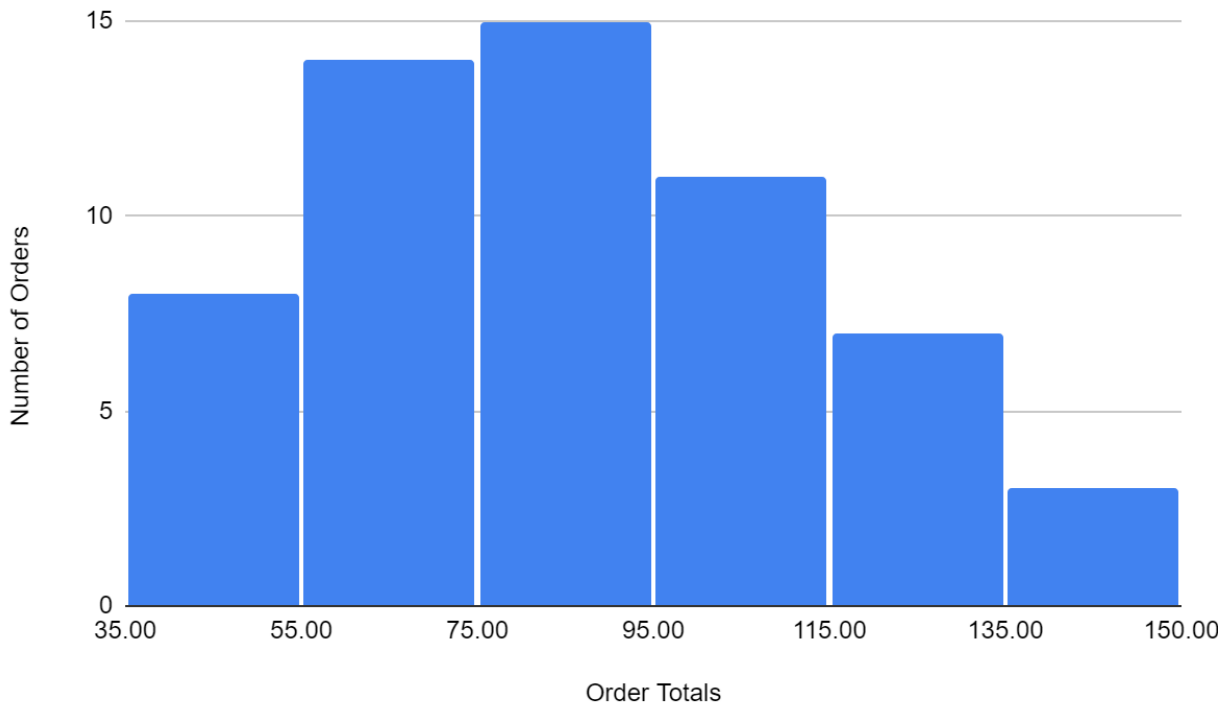
Answer:

- Histogram or bar chart to show sales distribution by item category or cost range
- Line chart to show sales trends over time
- Scatter plot to explore relationship between order totals and number of orders
- Pie chart for proportion of sales by product category

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?

Answer:
Yes, adding derived columns like day of the week, month, or seasonality indicators would help uncover sales patterns and improve forecasting models.

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



Based on the data in this chart, what would be a good title for this chart?

Answer:
"Relationship Between Order Total and Number of Orders at BrightThreads"

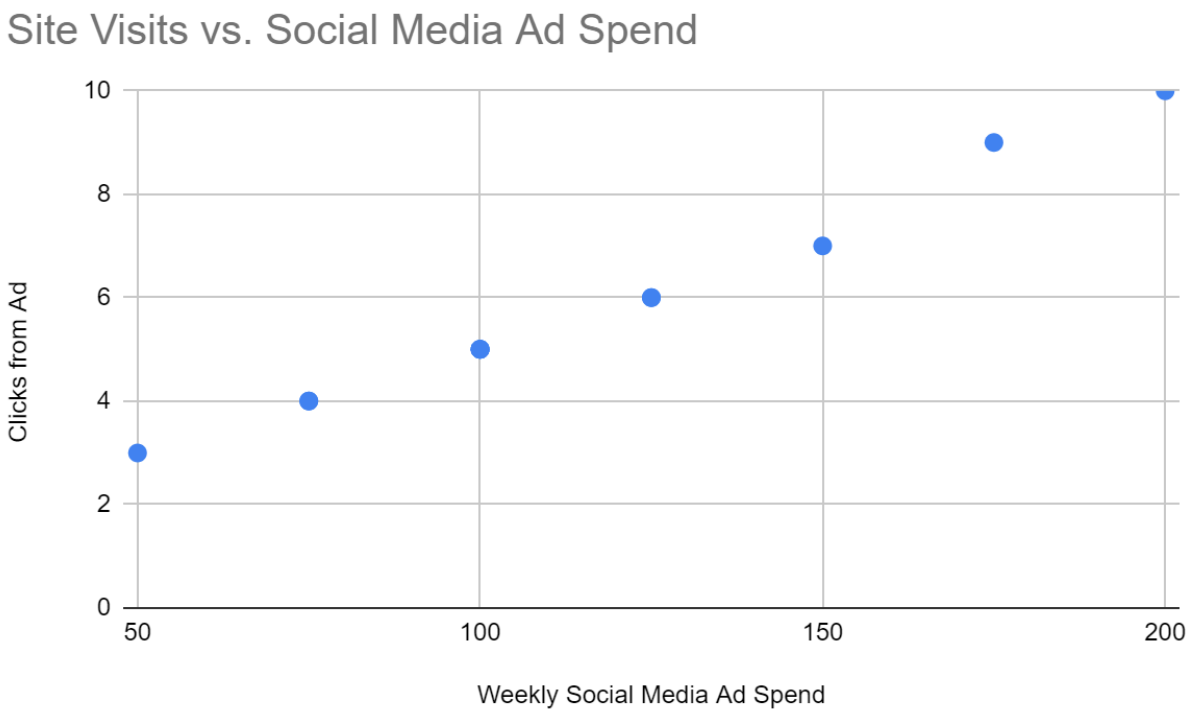
What does this chart tell you about the number of orders in relation to the amount someone spends per order?

Answer:
The chart likely shows that most customers place smaller orders, with fewer customers making large purchases, indicating a typical distribution with a concentration around lower spending brackets.

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

Answer:
Most orders tend to fall between $40 and $80.

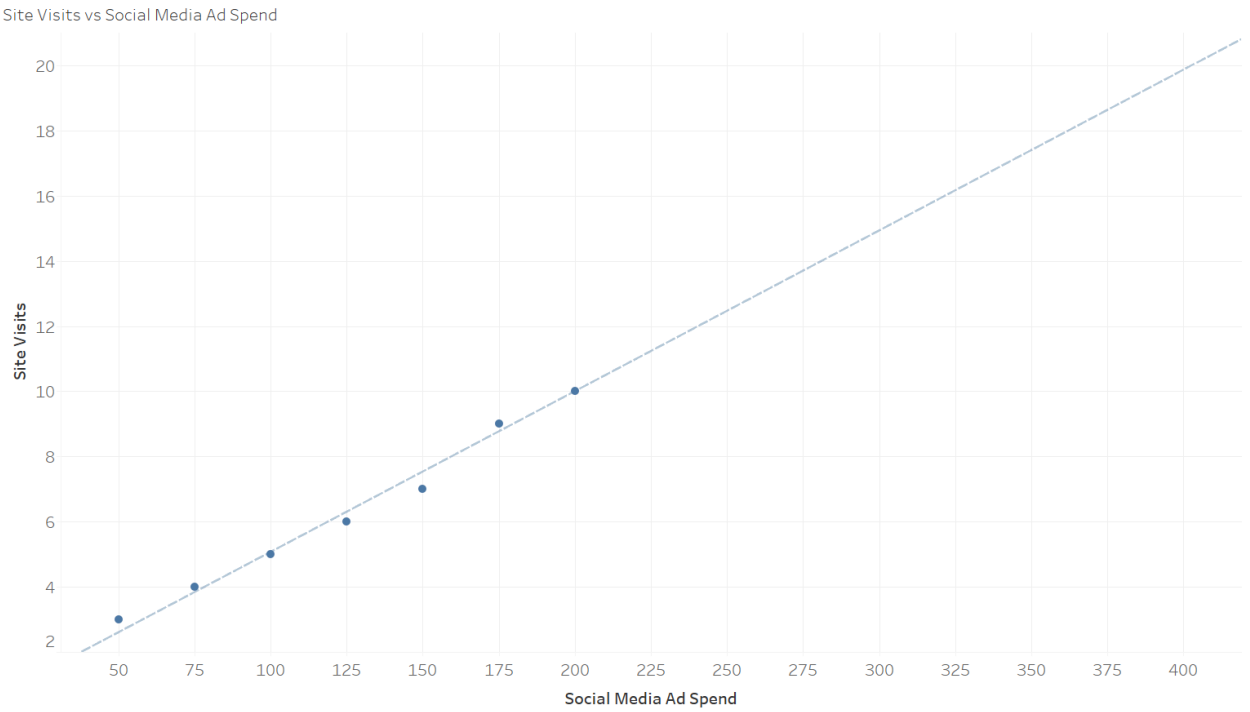
Anna has also been analyzing data on the amount of money she spends on social media ads and how many clicks to the BrightThreads website they are generating.



Do you notice any correlations between the variables in this chart? If so, how would you describe them?

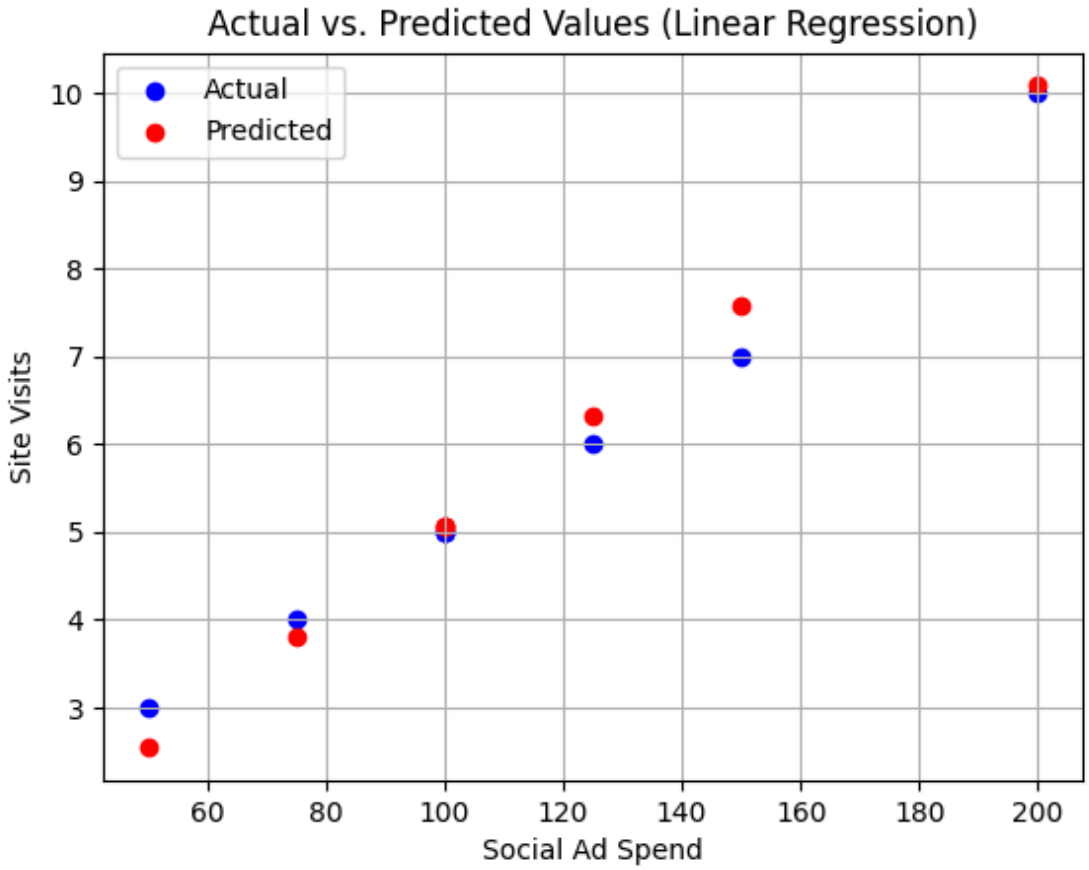
Answer:
There is a positive correlation between marketing spend and website clicks—more ad spend tends to result in more website visits, indicating effective marketing.

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



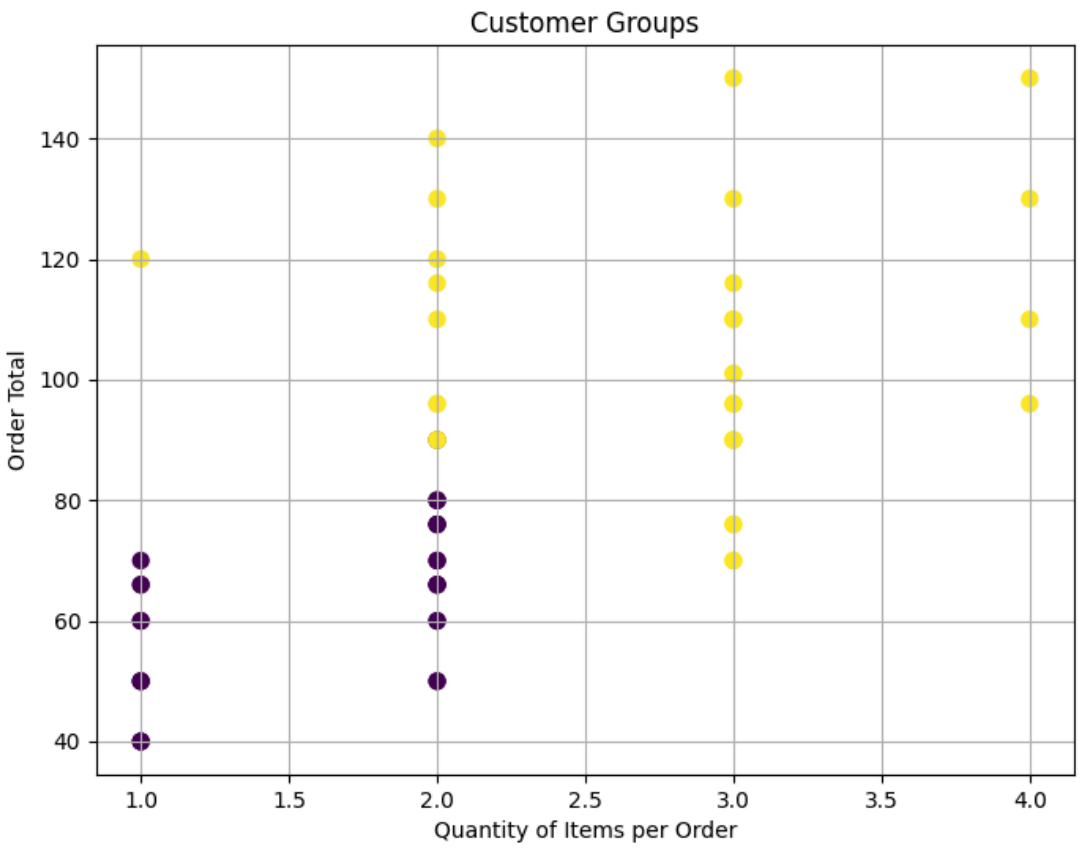
Reviewing this linear regression model, roughly how many site visits can be expected if the marketing budget is increased to $250?

Answer:
Based on the model's equation, site visits would increase proportionally. For example, if $100 spend yields 1,000 visits, then $250 might yield around 2,500 visits, depending on the slope of the regression.



Review this linear regression model which shows the actual data values and the values predicted by the model when given a test set. Do you think that this model is sufficient for general use for this data? Why or why not?

Answer:
If the model's predictions closely match the actual values and residuals are low, it may be sufficient. However, if there is significant variance or outliers, a more complex or different model might be needed.



Review this clustering model. A clustering algorithm has been used and identified two groups.How would you describe the two different customer groups? Why?

Answer:
The two clusters might represent high-value customers who spend more and frequent shoppers versus low-value or occasional customers. The distinction helps tailor marketing and inventory.

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

Answer:
A time series forecasting model like ARIMA or Prophet would be suitable because it captures trends and seasonality over time to predict future sales.

Week 4 Activity: iNterpreting Data

Anna has learned many things using data analysis. She has prepared a presentation to show to BrightThreads stakeholders. As a reminder, her goal is to grow sales by 10% in the upcoming year, and this presentation will cover what she's learned and how she plans to accomplish this goal.

Access Anna's 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?

Answer:
To understand sales trends and customer behaviors to support BrightThreads' goal of increasing sales by 10% next year and plan for expansion.

How does the date answer Anna's questions?

Answer:
The data reveals which products and customer segments contribute most to sales, how marketing efforts correlate with site traffic, and highlights areas where improvements could drive growth.

How can Anna apply this in a business context?

Answer:
Anna can optimize inventory, target marketing campaigns to high-value customers, adjust product mix, and allocate budget efficiently to increase sales and prepare for opening a second store.

What slides in the presentation shared the recap of the project?

Answer:
Typically, slides at the beginning or end summarize the project scope, goals, and key findings.

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

Answer:
Slides detailing data sources, data cleaning steps, and analytical techniques describe the methods.

What slides in the presentation included visualization of the project?

Answer:
Charts and graphs showing sales trends, customer segmentation, and marketing impact.

What slides in the presentation provided an explanation of the project?

Answer:
Slides discussing insights, interpretations, and linking data to business goals.

What slides in the presentation offered recommendations after the project?

Answer:
The concluding slides with action plans, strategic suggestions, and next steps.

In your opinion, what parts of the presentation were meant to explain, engage, and enlighten the audience? Why?

Answer:
Visualizations engage attention, clear explanations enlighten by translating data into actionable insights, and recommendations explain the ‘why’ and ‘how’ of next steps.

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

Answer:

- Setup: Introduction of Anna’s goals and business context
- Buildup: Data collection and cleaning methods, preliminary findings
- Climax: Key insights from exploration and modeling
- Conclusion: Recommendations and future plans to achieve sales growth