Content

Week 2 Activity: Obtaining and Scrubbing Data
Week 3 Activity: Exploring and Modeling Data
Week 4 Activity: iNterpreting Data

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?

In the following scenario the SMART goal for Anna would be to increase sales of BrightThreads by 10% by next year

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

The primary Key Performance Indicator for this goal would be the clothing items that are trending and are making the most sales which will help Anna increase her revenue.

What relevant data would you gather in this scenario?

The data required in the following scenario will be the number of loyal customers, which product is selling the most/least, the margin of profit on the investment that is being made,customer retention rate.

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).

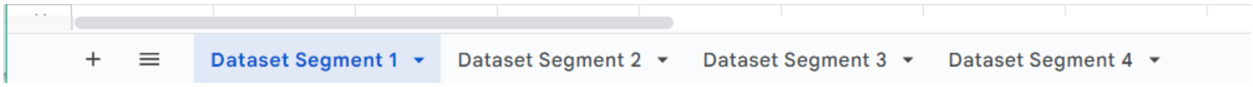
The most easiest way would be getting the Data from Anna's sales from online/offline store that is the first-party data, apart from that we could use third-party data like open source data sets like kaggle,google analytics to get the overall trend of the current market and help Anna make a better decision for her new stock she is planning to introduce in her store.

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?

The data anna has gathered is somewhat valid from her business perspective since she needs to know whether a certain clothing apparels will sell or not and for that you need the cost of sales and the cost of buying the apparel which is not available in the spreadsheet , in order to calculate the 10% increase in sales we will need to calculate the profit from the apparel that are being sold , she could have gathered additional information such as the time/date of the purchase to understand the market sales trend during the year and also whether the apparels were sold during a discount period or no, this could achieving maximum understanding of customer behaviour.

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

Rows 4,5 and 10,11 show duplicate rows (same details have been repeated)

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

Row 2,8 have different formats of zip-code compared to other zip-codes (zip codes are in detail)

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

Rows 4,11 having missing values

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

Row 6,11 have unrealistic cost of items that is 0.069 and 5999 which are extreme values compared to the other items sold in the store

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?

The spreadsheet has 2 datasets online sales data and the weekly ad spend and visits.
The data contains transaction of 58 sales with the customer details and the products bought by them and its type, and the total amount paid by the customer at checkout

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

There are 3 types of data present in the following dataset that is :

date column - Date type.

Customer id ,order no,zip code,quantity,order total - Numerical type.

Item category- Categorical type.

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

Min : 39.99

Max : 149.99

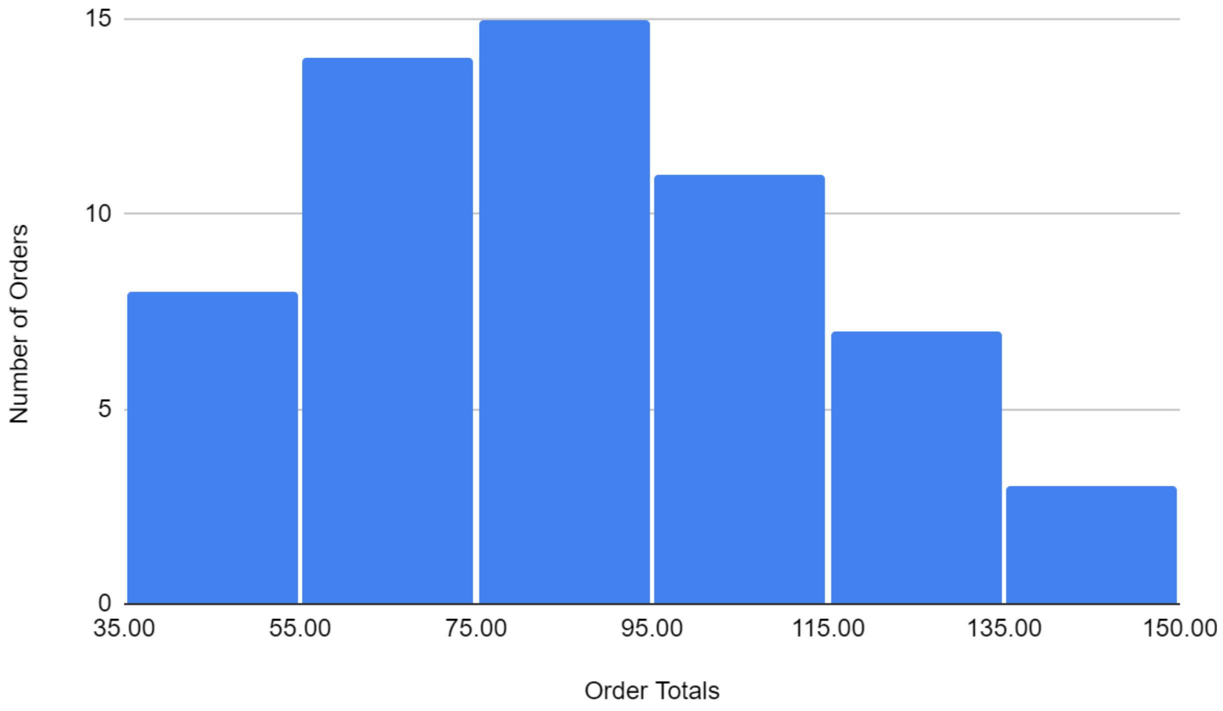
What kind of chart would you use to help visualize this data?

Bar chart which could show which type of products were sold the most, (low,medium,expensive items)

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, by adding an extra column of which day the orders were created we can understand which day of the week is the most suitable time to sell the products and also gives us an insight on how to time the customers purchase behavior , apart from that we can also add a column by binning the products according to their price from low to expensive to get an in depth understanding of what the customers prefer buying.

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?

Total Sales by orders with relation to price

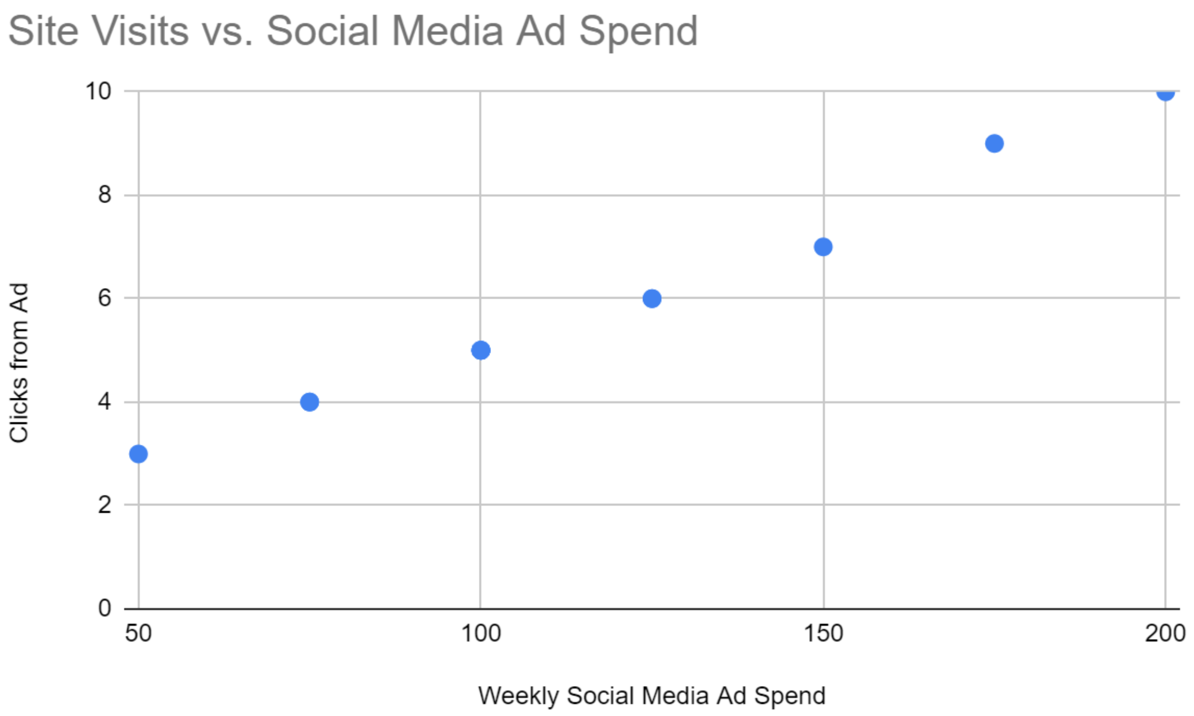
What does this chart tell you about the number of orders in relation to the amount someone spends per order?

When the number of orders is greater than 10 we can see that the order total always lies in the range of 55 to 115.

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

Most orders lies in the range of 55 to 95.

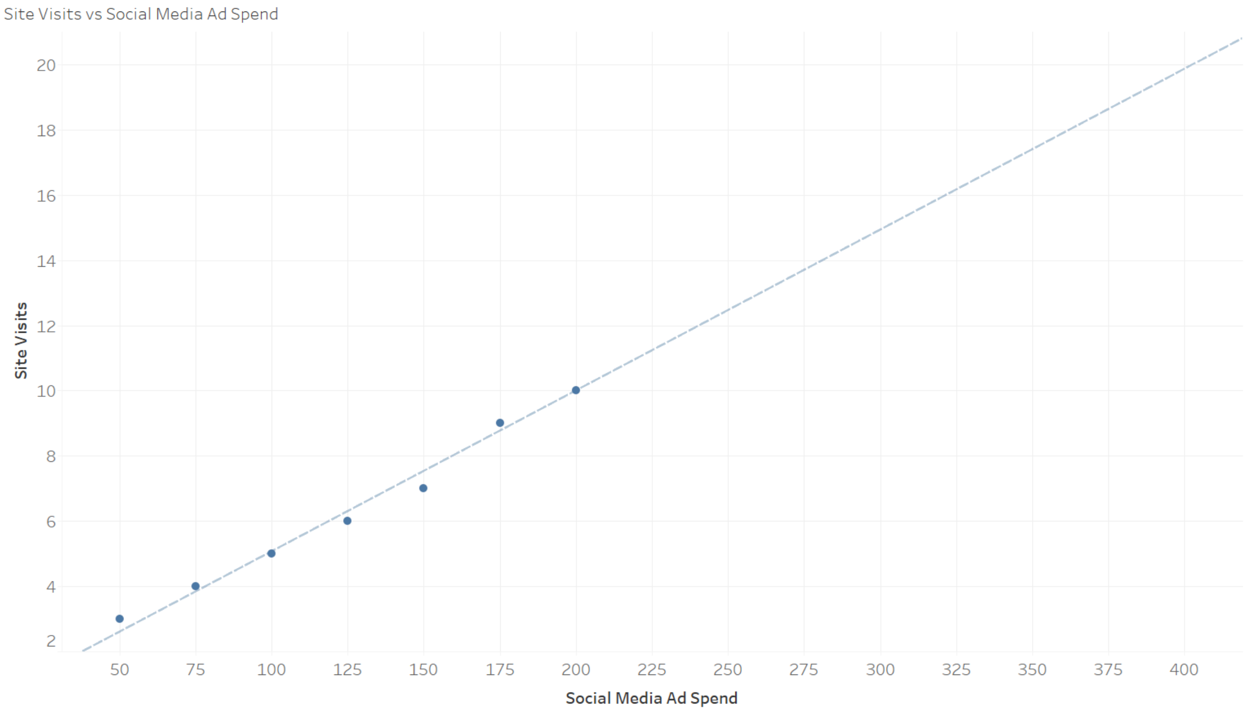
Anna has also been analyzing data on the amount of money she spends on social media ads and how many clicks to the BrightThreads website they are generating.



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

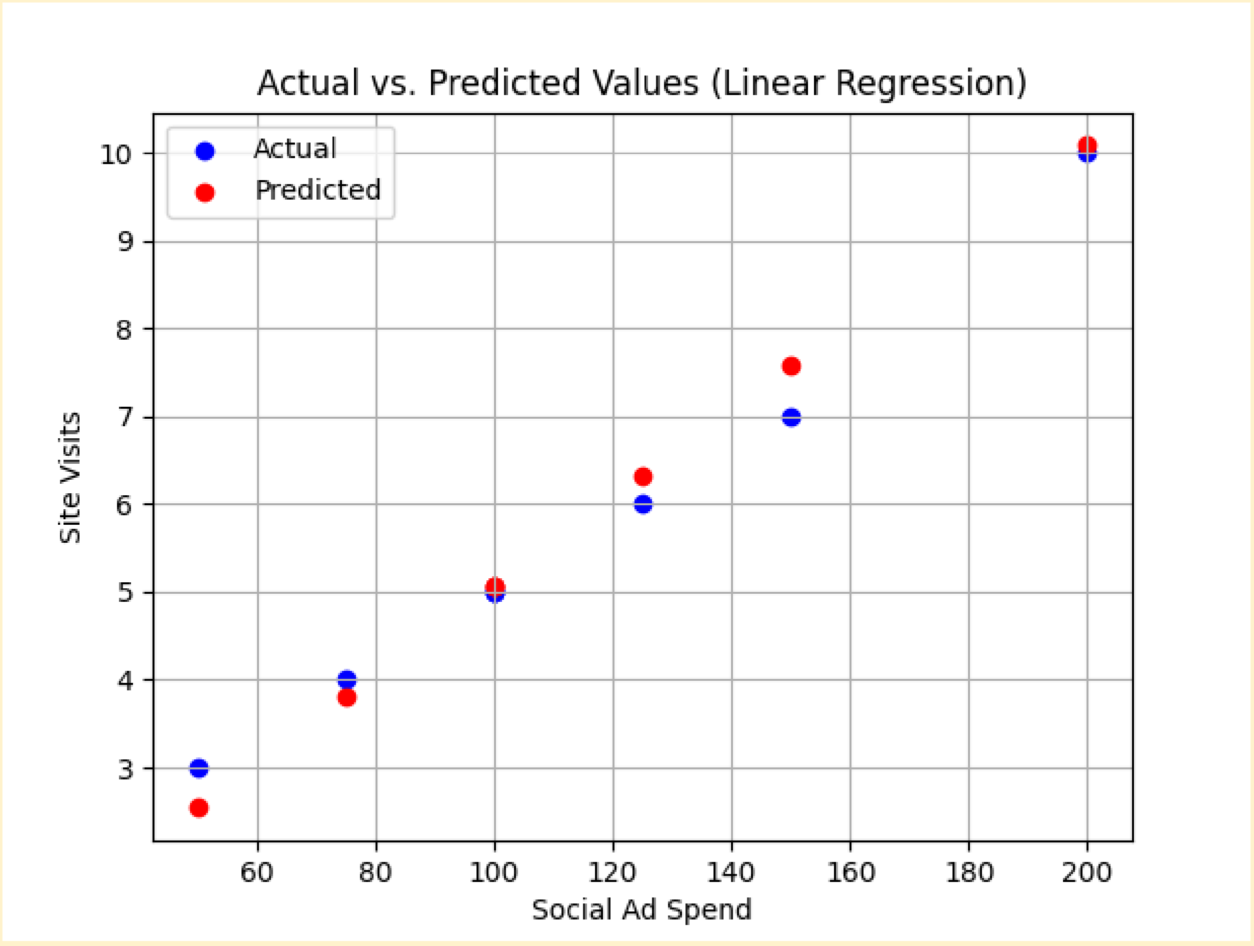
There is a positive correlation between clicks from ads to weekly social media ad spend
The more you spend on the weekly social media ad the more number of clicks from ads are generated.

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



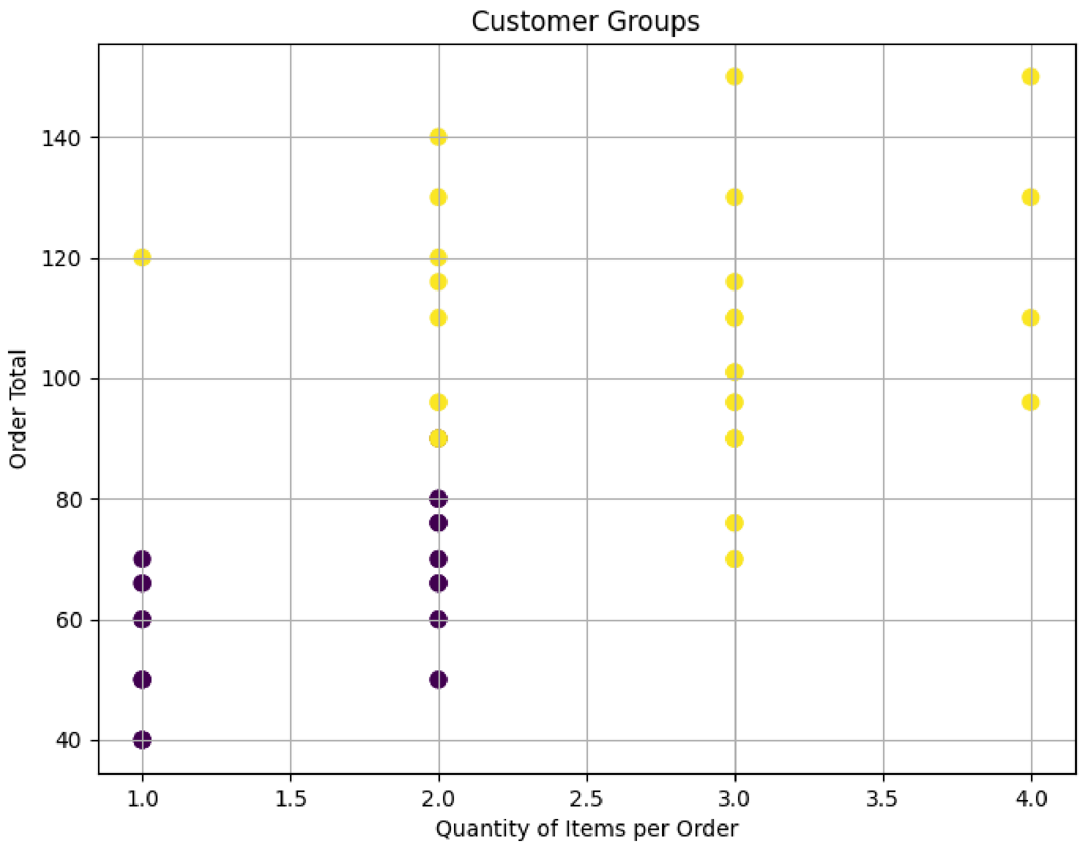
Reviewing this linear regression model, roughly how many site visits can be expected if the marketing budget is increased to $250?

Roughly between 10-14 (10 min and 14 max) site visits if the ad budget is increased by 250$



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?

Yes the model is sufficient enough for the general use of data as it is roughly near the same values as predicted and has very less deviation from the predicted values.



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

There 2 groups customer , we can assume from the clustering model that the purple group prefers buying items less than the order total of 80$ and also has a quantity of 2 ,whereas the yellow group has a minimum of 2 items per order and mostly has an order total of 80$ and above , we can classify them as 2 category of customers , one group who prefers spending on expensive products and one group prefers spending on cheaper products .

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

I would prefer using a linear regression model so that it will be easier to understand the slope or the sales curve of the store , it will help understand the number of customers and their purchasing power so that I can make your inventory suitable to the customers.and based on the positive and negative correlation i could be able to conclude whether the store is performing optimally or not.

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?

The objective is to grow sales by 10%,Grow footprint ,customers satisfaction in the upcoming year

How can Anna apply this in a business context?

Anna can find the potential bottlenecks in the sales which are caused due to human errors or store sales, finding out what is stunting their growth, which can help Anna persuade stakeholders to make the right decision

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

Slide No 6 covered the methods used in the project that is predictive models (linear regression) the slides also included the goal,story,team,tools.

What slides in the presentation included visualization of the project?

Slides 7,8,9,10,11,12 gave us a visual representation of Anna’s findings after OSEMN framework, it showed the sales of last 2 years how much they'll need to sell each quarter to hit 10%,top selling items,potential sales,current ad spendings and its potential.

What slides in the presentation offered recommendations after the project?

Slides 15 and 16 presented the potential changes that need to be made in order to hit the main objective of the organization.

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

Slides 1,2 - Setup
Slides 3 to 12 - Buildup
Slides 13,14 - Climax
Slides 15 - Conclusion