# 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](#_kxie68fkxlc8)

[Week 3 Activity: Exploring and Modeling Data](#_lcvb3omhatri)

[Week 4 Activity: iNterpreting Data](#_uxdv22fdhkto)

# 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?

| The smart goal for Anna is to set the goal to increase her total sales by 10% at the end of next year. |
| --- |

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

| The Primary KPI would be sales of clothes at each point of time in the year. And the support for the KPI will be the Type of clothe that is in demand. |
| --- |

What relevant data would you gather in this scenario?

| Types of clothe , sales of clothe , How it is selling (Online or In store ) , Area of demand , Brand of the clothe. |
| --- |

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 primary data can be collected from first party sources like the Store registers and website logs(database) , the survey of feedback can also be considered from user about review of clothe. |
| --- |

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](https://docs.google.com/spreadsheets/d/12pYhNUBH4D96rbnEH5s1eUTVr6Bp_Qytj4AZ5KKMsDE/edit?usp=sharing) 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 is validate as it is coming from the website purchase database and have the order records that are made in real time. |
| --- |

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

| The Data is duplicated. Identified using order\_id and order\_number. |
| --- |

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

| The issue is occurred with column named as “customer\_zip” because of the different format’s are used for zip-codes need to change to common format. |
| --- |

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

| Missing key information for some orders such as the item\_sku and order\_numbers. |
| --- |

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

| The wrong values of price occur at row 6 and row 11. |
| --- |

# 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](https://docs.google.com/spreadsheets/d/1ZOJDKQL1PfQP1Y3-D2kE9eC-Ll3w3JghQEOhl_RQMRg/template/preview) 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 Dataset provided contains the two sheets. 1st sheet tells about the Online Sales Data which contains Language of data as date , numerical data( id’s , prices , bills) and textual/Categorical data as item category. The 2nd sheet tells about the spending on the advertisement of products which covers numeric values. |
| --- |

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

| Data as date , numerical data( id’s , prices , bills) and textual/Categorical data as item category. |
| --- |

Reviewing this data, what is the minimum value in the order\_total column? What is the maximum value in order\_total column?

| Minimum is 39.99 and Maximum is 149.99 . |
| --- |

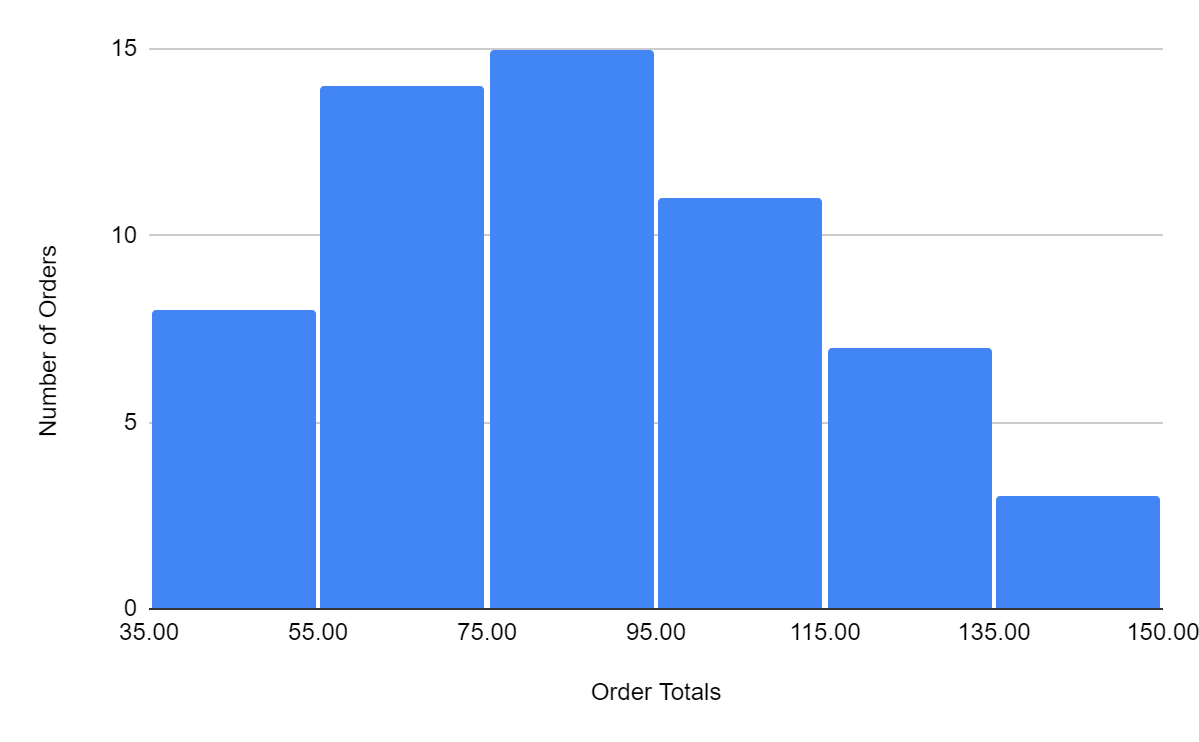
What kind of chart would you use to help visualize this data?

| I will convert the values of numerical columns such as ‘Order\_total’ to categorical and use the bar chart to compare the trend of type of clothes purchased against the category of the clothe. |
| --- |

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?

| I will like to add column to know in which month the sales are down based from the sales date to interpret which clothes are being sold on which month and how much quantity it is being sold. |
| --- |

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?

| Distribution of Order Totals based on number of orders. |
| --- |

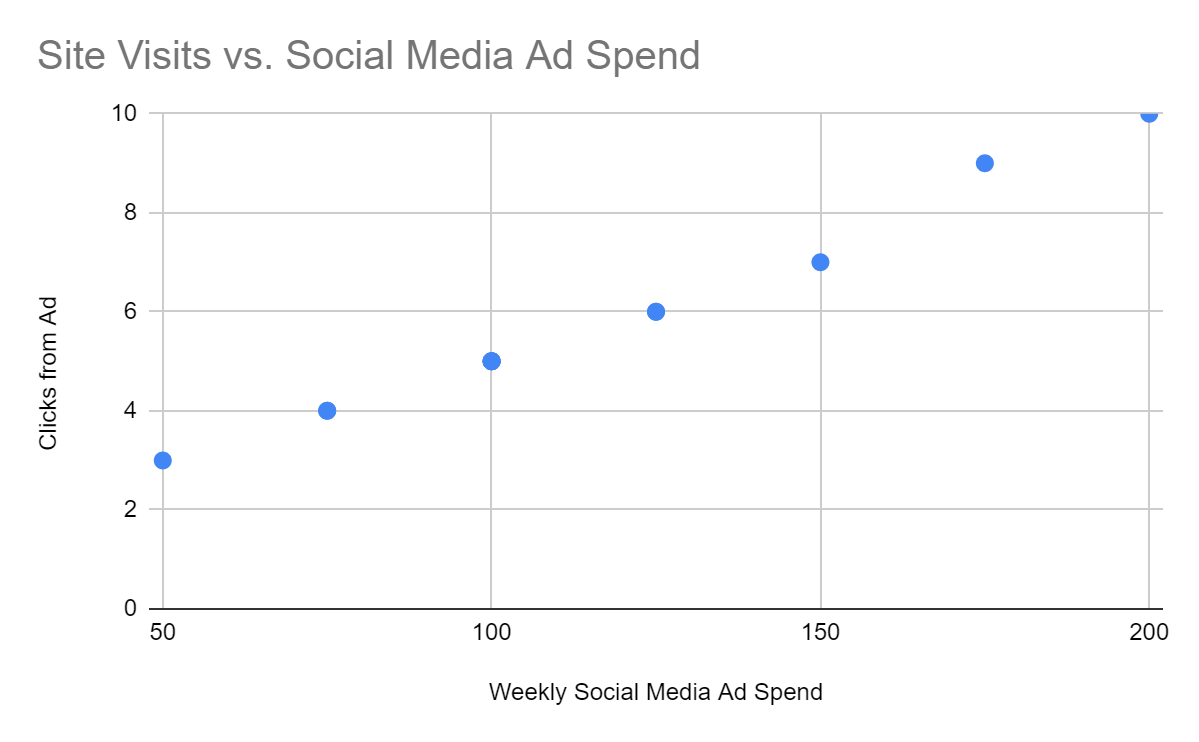
What does this chart tell you about the number of orders in relation to the amount someone spends per order?

| This chart showing normal distribution where the bell-shaped curve is found which tell us that the main cluster value i.e. the number of orders and order totals are revolve around the mean of the values. |
| --- |

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

| 75-95. |
| --- |

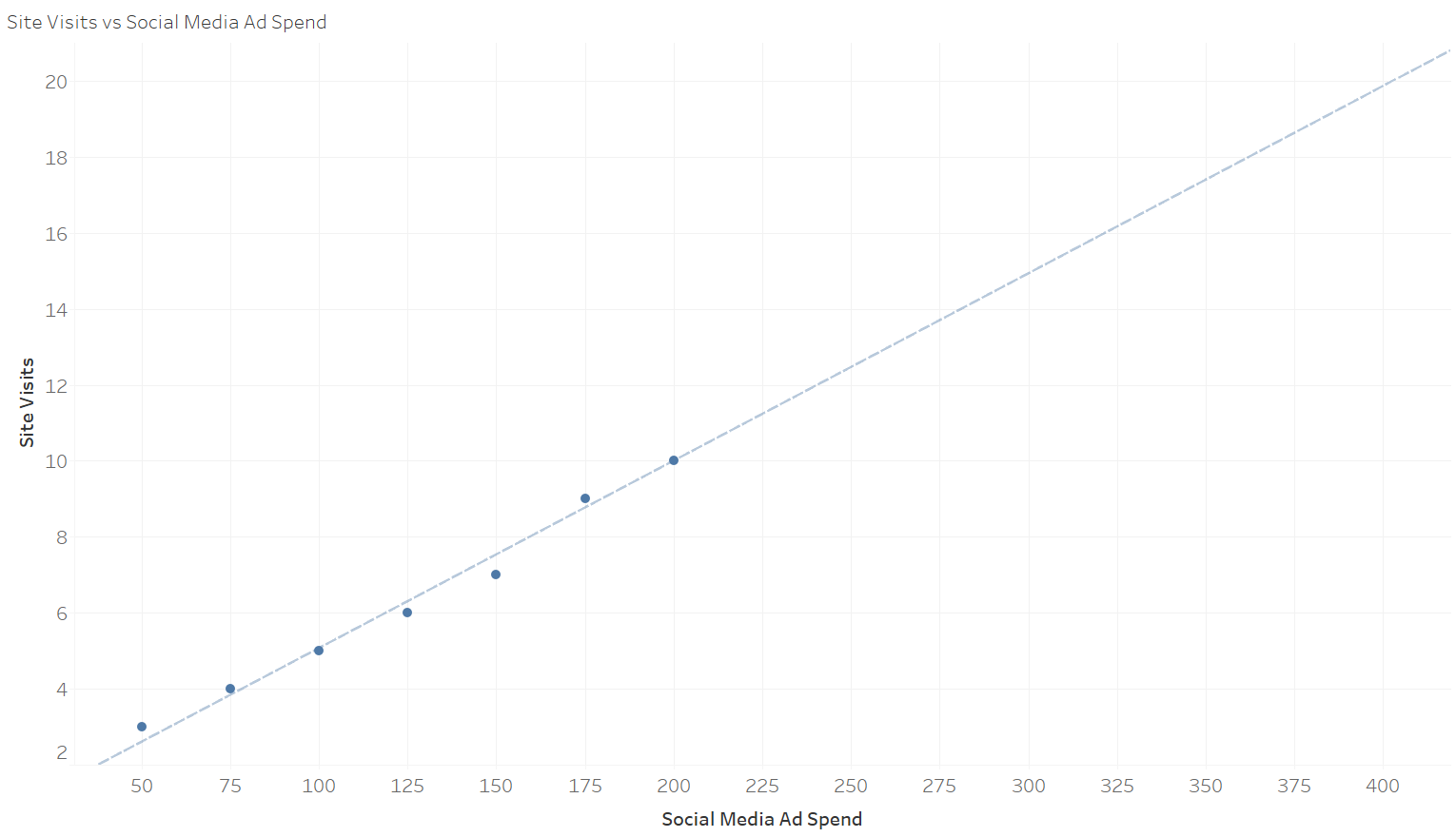
Anna has also been analyzing data on the amount of money she spends on social media ads and how many clicks to the BrightThreads website they are generating.



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

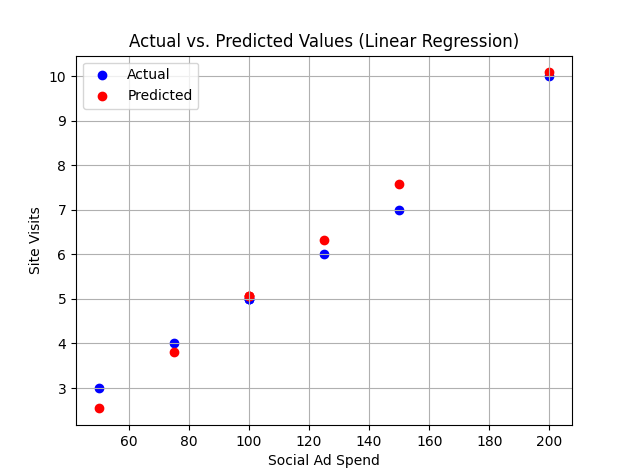
| This is the positive correlation which tells that the amount spend on adds are positively increasing the number of click generated from the Ad. |
| --- |

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



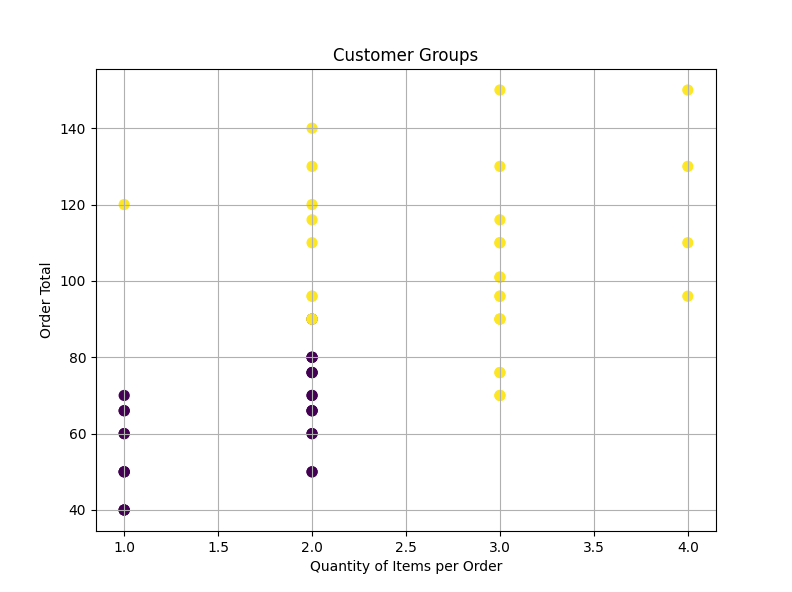
Reviewing this linear regression model, roughly how many site visits can be expected if the marketing budget is increased to $250?

| 13. |
| --- |

****

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 it can be sufficient for general use as it is promising the 95% satisfaction of the values that are predicted to the actual 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 are two groups formed as per the Quantity of the items and order total. The one group shows that the quantity with low values such as less than 2.0 produce around 80 order totals and the another tells that the values that are equal or greater quantity than 2 are tends to give the order total above the 80. |
| --- |

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

| As we have the past data which is scubbeld and cleaned with some feature engineering and the goal is to predict the sales in upcoming quarter i will prefer to use the linear regression model it needs simple and minimal data and also it will satisfy our goal to predict the sells of upcoming quarter with help of past data. |
| --- |

# 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](https://docs.google.com/presentation/d/1Fp3u1KWgR2D_FOzwJMwoRzYySTi4_KN3u_YlaKXsyUY/edit?usp=sharing).

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?

| To identify how can the BrightThreads increase there sales by 10% in upcoming year. |
| --- |

How does the date answer Anna’s questions?

| * Small changes have the potential to noticeably increase sales * Focusing on our most popular items can increase sales * Reallocating advertising spend to better performing social media channels can also increase sales |
| --- |

How can Anna apply this in a business context?

| * In the immediate future, social media ad dollars should be reallocated   + The ads themselves seem to be doing well, but we can deliver them more effectively * In the coming months, shift inventory to a smaller, tailored selection   + Our most popular items are bringing in a large proportion of our monthly sales * Reevaluate in 6 months time   + Check back in on the social media engagement numbers and sales figures |
| --- |

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

| From slide no 2 to 4. |
| --- |

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

| 6. |
| --- |

What slides in the presentation included visualization of the project?

| From slide no 7 to 12. |
| --- |

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

| From slide no 4 to 6. |
| --- |

What slides in the presentation offered recommendations after the project?

| Slide no 15 and 16. |
| --- |

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

| From the slide no 4 to slide no 14. |
| --- |

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

| Set-up : Slide 1 to 4 ( It revise the objective and tell the data sources) , Buildup : Slide 5 to 12( Explain the methods and Visulization) , Climax : Slide 13 and 14( Tells the answer to the question in objective) , Conclusion : Slide 15 and 16 ( Tells how to implement solution and give references and confidence). |
| --- |