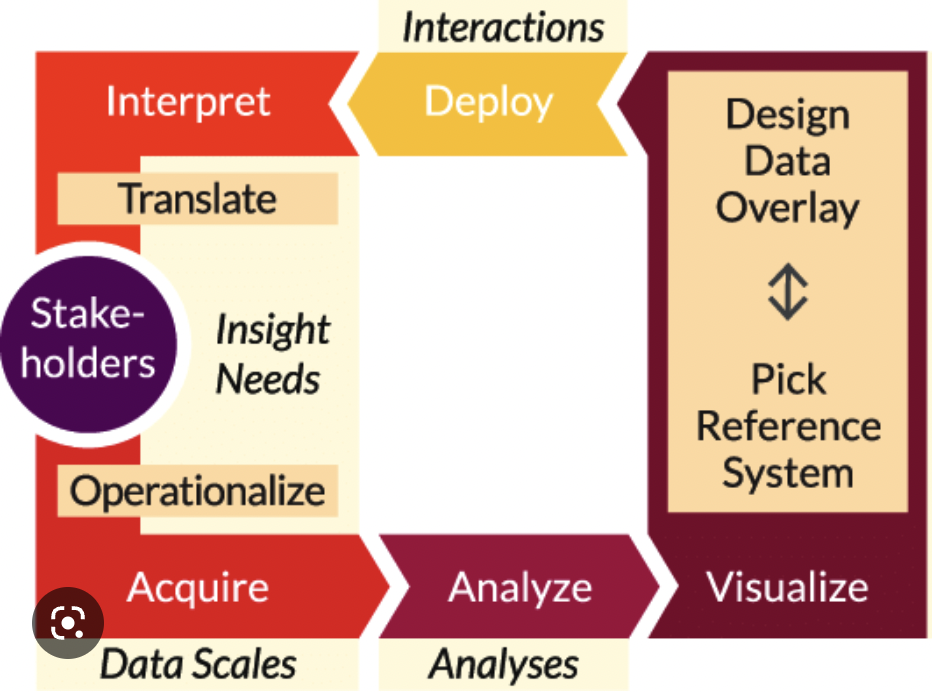
**CONTENTS**

| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
| --- | --- | --- |
| **1** | **INTRODUCTION** |  |
| **2** | **STAKE HOLDER NEEDS** | 1 |
| **3** | **TOOLS AND REQUIREMENTS** |  |
| **4** | **ANALYSIS** | 4 |
|  | 4.1 DATASET | 4 |
|  | 4.2 INSIGHT | 8 |
|  | 4.3 PROCESSING  4.4 WAYS TO VISUALIZE |  |
| **5** | **VISUALIZATION** | 12 |
| **6** | **EXECUTIVE SUMMARY** | 18 |

**INTRODUCTION**

This project is the implementation of the general processes and steps involved in visualization.



**FIGURE 1 Visualisation lifecycle**

To begin a data visualization project, it is important to gather the stakeholders' needs, objectives, or goals. Once the requirements are clear, the next step is to examine the dataset to uncover observations, or findings. This process is akin to searching for valuable information within a dataset, much like hunting for pearls in an oyster. After identifying insights, the analyst can generate outcomes that satisfy the stakeholders' needs. To help stakeholders understand the key findings, it is necessary to depict, present, or represent the data using appropriate charts. Finally, the analyst can deliver conclusions, inferences, or meaning from the data to the stakeholders based on their needs.

**STAKE HOLDER NEEDS**

The needs of the stake holder is as given below.

| At our Market place, we also care a lot about production per hour (pph). We have provided a set of data, and would like you to use Tableau to: |
| --- |
| i. Make a table showing the pph (items produced per hour worked), breaking out each location within each market, monthly. |
| ii. Some execs like to look just at Product A production, some prefer to include Product B, so please make a parameter in a drop-down that lets them choose to look at Product A only in this table or include Product B too. |
| iii. In the tooltip on this table, show what percent of that month and year’s total production (across all locations in the company) that this location accounted for. |
| iv. In the tooltip on this table, show the year-to-date for pph for each store. |
| v. In the tooltip on this table, show the year-over-year for that month |
| vi. In the tooltip on this table, show the year-over-year comparison of the year-to-date pph. |
| vii. Make it possible to collapse/roll up the stores to the markets and calculate the pph numbers (and tooltip) correctly still. |
| viii. Tell us something about this data in an interesting visual way in a separate viz. |

**TABLE 1 Requirements**

The stakeholders are interested in understanding the production per hour (pph) for each location within each market, broken down by month. They want to be able to view this information in a table format, with the ability to choose whether to include data for Product B or only view Product A.

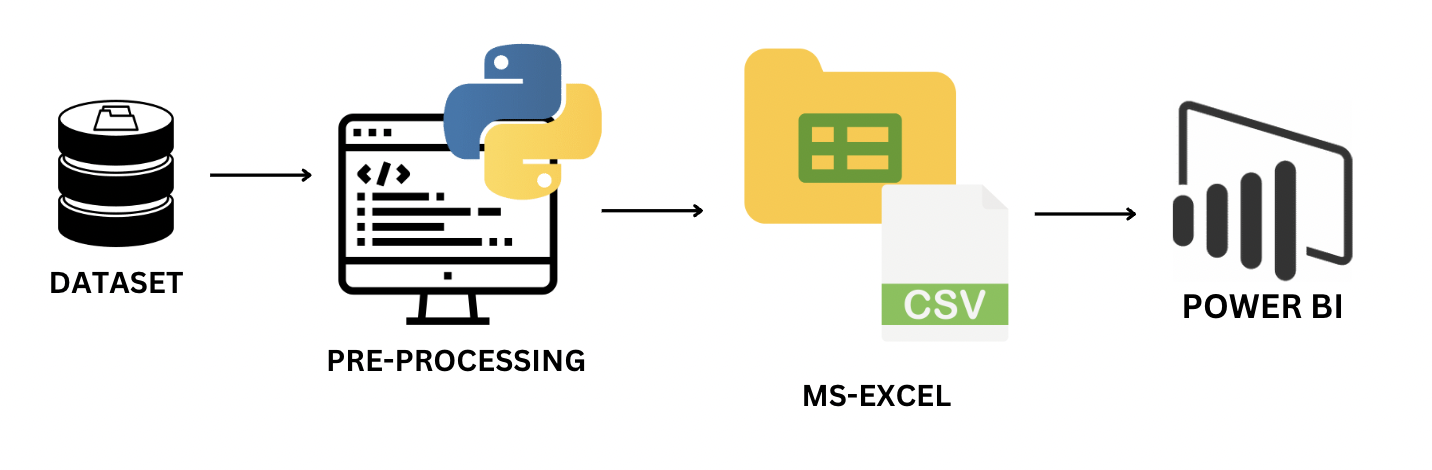
Additionally, the stakeholders are interested in understanding how each location's production compares to the overall company's production, both for the month and year-to-date. They want to be able to view this information in a tooltip when hovering over the data in the table.

The stakeholders also want to understand year-over-year changes in production for each location, both for the month and year-to-date, again viewable in the tooltip.

Finally, the stakeholders want the ability to collapse or roll up store-level data to the market level while still accurately calculating pph and displaying the relevant tooltip information.

At the end, it seems like the stakeholders are interested in gaining insights into production performance at the store and market level, and want to be able to view and interact with this data in a variety of ways. They also want to understand how each location's performance compares to the overall company's performance, and how this changes over time.

**TOOLS AND REQUIREMENT**

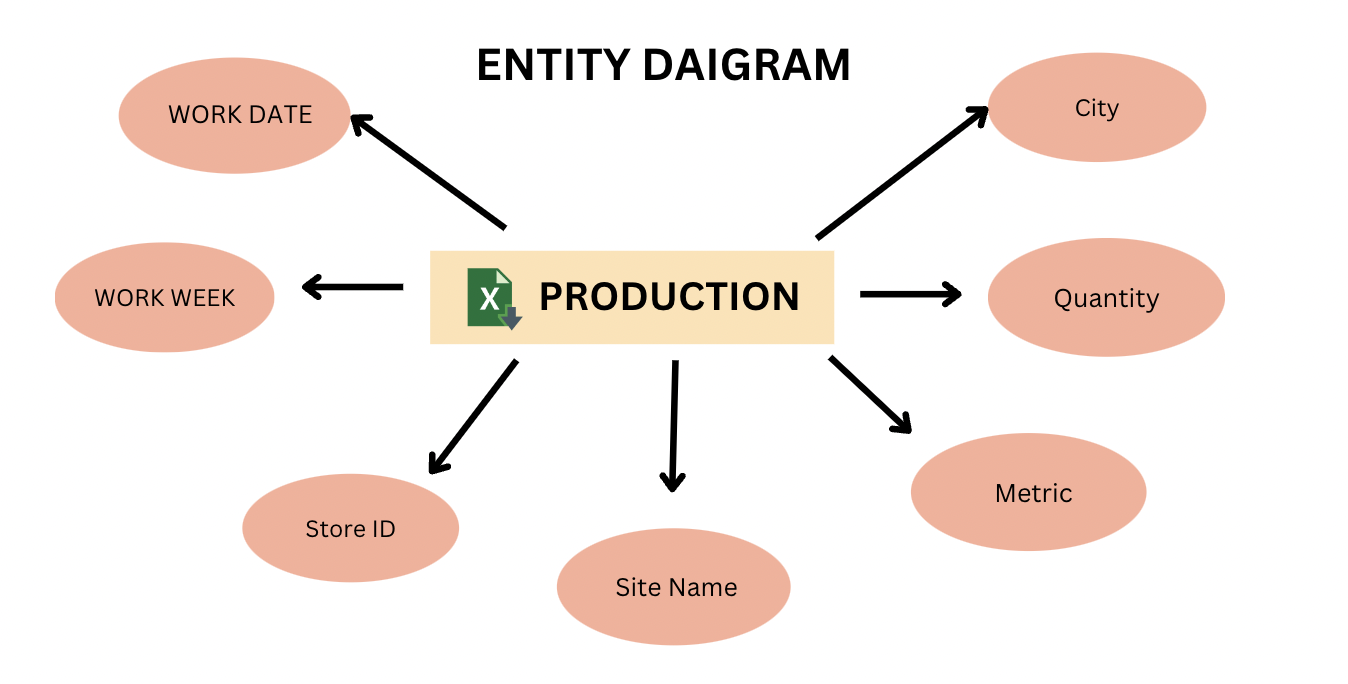
The needs of the stake holder is as given below.

**FIGURE 2 Tools Used**

* Python - a popular programming language for data analysis and processing. It can be used to perform various tasks, including data cleaning, data transformation, and statistical analysis.
* CSV file - a file format used for storing and organizing data in a spreadsheet. It is a simple and widely used format that can be easily read and processed.
* Excel - a spreadsheet application developed by Microsoft. It is a popular tool for storing and organizing data in a tabular format.
* Power BI - as the name suggests is Buisness Intelligence tool for business analytics provided by Microsoft. It is a powerful tool for data visualization and analysis that can be used to create interactive dashboards and reports. Power BI can connect to various data sources, including Excel spreadsheets and CSV files, and can be used to create custom visualizations based on your data.

The requirements for this project would be to have a dataset in a CSV format, a working knowledge of Python to process and clean the dataset, Excel to store and organize the cleaned data, and Power BI to create interactive visualizations based on the data. The goal of the project would be to gain insights from the data and communicate those insights effectively to others using visualizations.

**ANALYSIS**

**DATASET**

**FIGURE 2 Entity diagram**

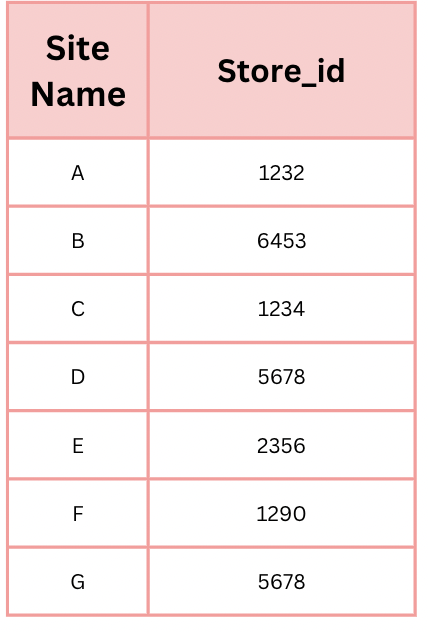
The dataset as the diagram suggests contains the given columns.

* The WORK DATE column contains the date when the work was done in the format as YYYY-MM-DD.
* The WORK WEEK consisted of the week in which the work was done in the same format as before.
* The Store ID consisted of numeric 4-digit data with 7 unique values such as [1232,6453,1234,5678,2356,1290,7890].
* The Site Name consisted of a string with 7 unique values such as [Store A, Store B, Store C, Store D,Store E,Store F].
* The Metric column represents a string with 3 unique values ['Product A Produced', 'Hours Worked', 'Product B Produced']
* The Quantity column represents numerical values.
* The City column represents a string with 5 unique values ['San Deigo', 'San Francisco', 'Irvine','LA', 'San Mateo']

**INSIGHTS**

There are basically five major insights that were found out from the data that was given:

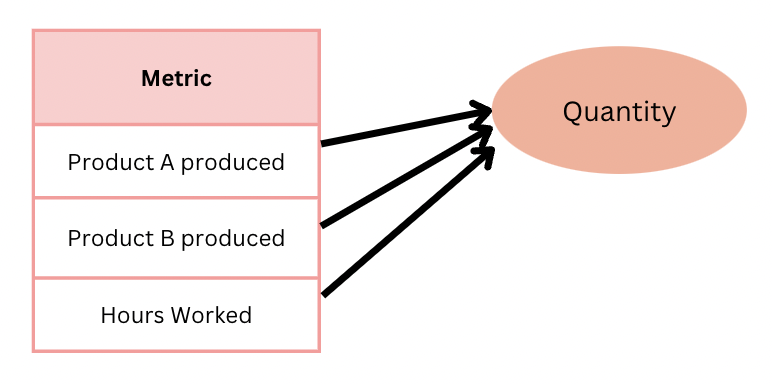
* WORK DATE column not in format. In power BI we can formulate the data only when it is in the format DD/MM/YYYY or if it consists ‘/’ instead of ‘-’. Thus there is a need to change the format.
* Ignorance of WORK WEEK column as it may not lead to any good interpretation for the stake holder.
* Site Name and Store ID represent the same value.



**TABLE 2 Site Name Vs Store ID**

For each Site Name the data set has a respective Store ID that is used.Thus using both the columns for visualization could be a repetition , therefore we can ignore any one column in the dataset. It is recommended to ignore Store ID as it is numeric it would difficult for the stake holders to interpret thus we can use the string data Site Name for the visualization.

* There is an unique relationship between the metric and the quantity in the given dataset.



**FIGURE 3 Metric vs Quantity**

The quantity here represents the value for each metric that has been given.

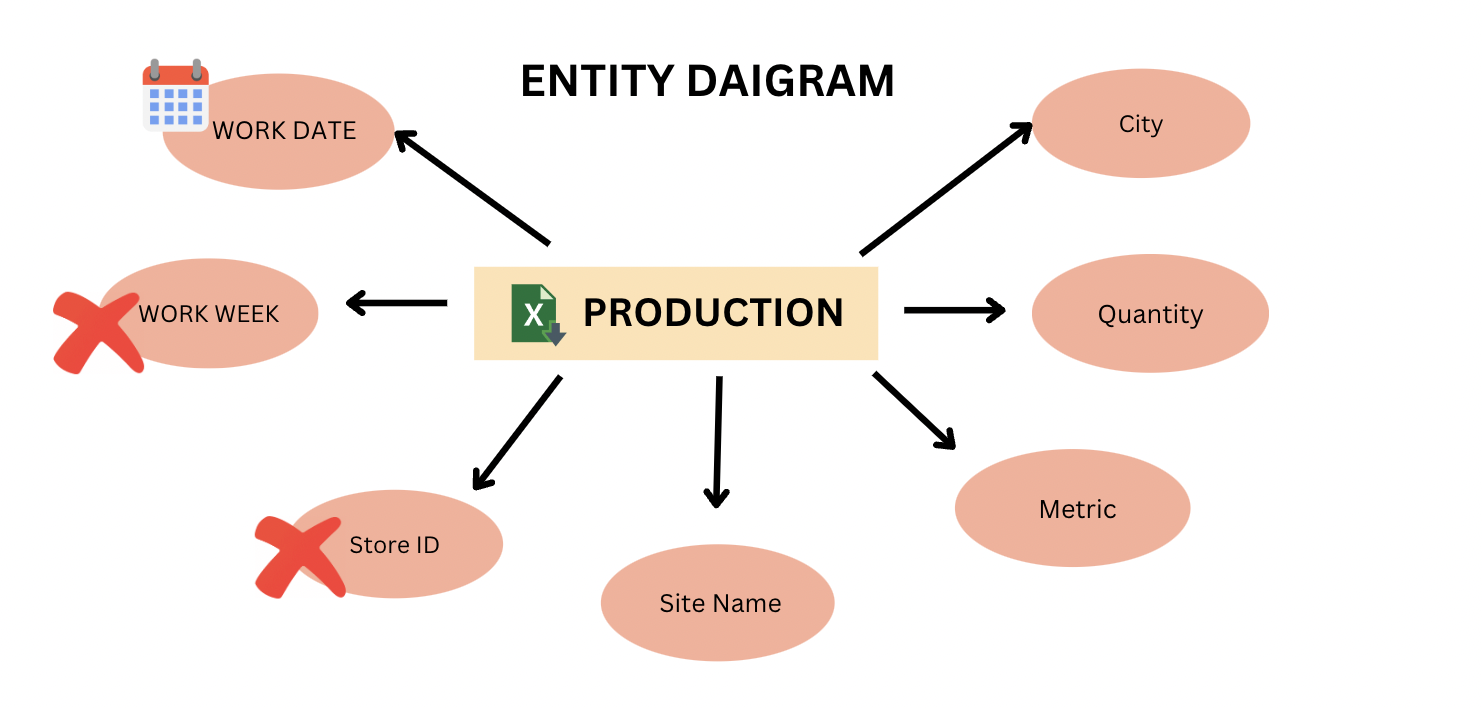
* One of the most important understanding from the dataset is the relationship between the Site Name and the City.



**TABLE 3 Site Name vs City**

As given in the table we could see that for the given City we have a respective Site Name that has been given. Except the city San deigo every other city has one site in it. Whereas San Deigo has 3 sites in it. Thus this would have a drastic impact on the production and hours worked.

**PROCESSING**

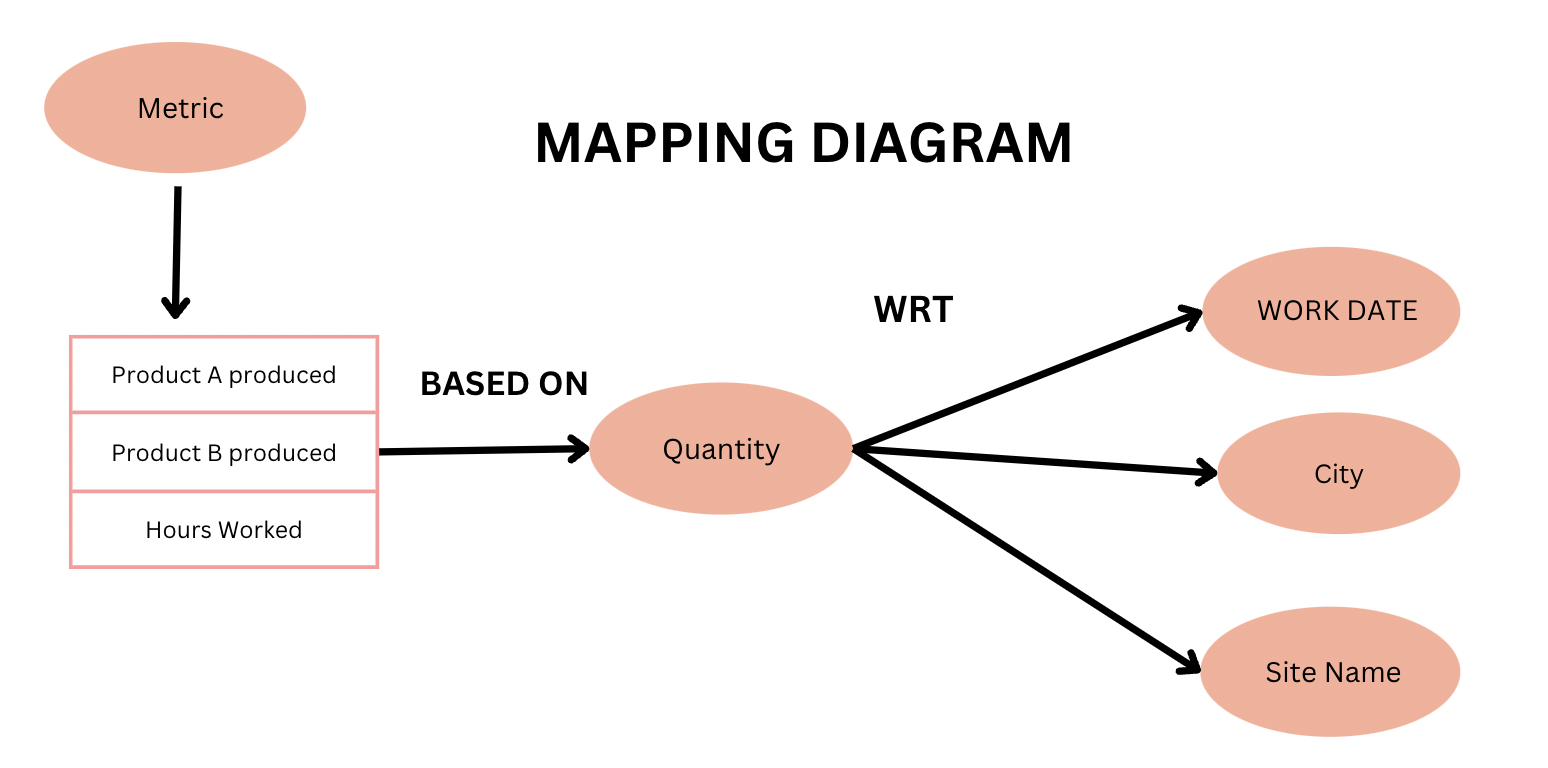


**FIGURE 4 Processed Dataset**

Thus the WORK DATE format was edited as per the format.

Then the WORK WEEK and Store ID column was ignored. It is not recommended to remove it from the dataset as it might help in any future visualization or for other purposes.

**WAYS TO VISUALISE**

****

**FIGURE 5 Mapping daigram**

The diagram represents on how the columns are compared with each other. First we take the column quanititty on one side. Where it can represent three other metrics such as Product A Produced, Product B produced, Hours Worked. Thus we compare the Quantity with each and every other column. Such as based on WORK DATE, City or Site Name on how much production was done or hours were worked.

**VISUALISATION**

Visualization is the graphical or pictorial representation of the insights from the findings of our data to make the stake holders understand the context. There are generally two types of analysis such as exploratory and explanatory analysis.Exploratory analysis is the understanding of the data and forming insights from it. Whereas, when it comes to explanatory analysis, being able to concisely articulate exactly who you want to communicate to and what you want to convey before you start to build content reduces iterations and helps ensure that the communication you build meets the intended purpose. Understanding and employing concepts like the 3‐minute story, the Big Idea, and storyboarding will enable you to clearly and succinctly tell your story and identify the desired flow. While pausing before actually building the communication might feel like it’s a step that slows you down, in fact it helps ensure that you have a solid understanding of what you want to do before you start creating content, which will save you time down the road.

**EXECUTIVE SUMMARY**

**Key Findings:**

1. City-Wise Statistics:

* The best city based on production would be San Deigo with 55 K production of products A and B. The Production per hour is also the highest in San Diego with 0.46 by nearly working 116 K hours. The reason is due to the presence of three sites B, C, and E in San Diego whereas the other cities have only one site in each of them.
* The Worst city would be LA with the least production of 13 K, with the least PPh of 0.29 by nearly working for 46K hours.

1. Site-Wise Statistics:

* The best site based on production would be Site E in San Deigo with 25 K production, with the highest pph of 0.55 by nearly working for 44 K hours.
* The Worst city would be Site C in San Deigo with the least production of 11 K , with a mediocre pph of 0.38 by nearly working for 29K hours(maybe it might be a small site with little production as the
* pph is not very less).

1. Year-Wise Statistics:

* The production of Product A has a continuous peak and falls every quarter of a year stating an unstable wave. The prediction for 2017 is supposed to 4k.
* The production of Product B is supposed to have an marginal increase with the date with a prediction of 1.5 k
* There were stores that produced only either product A or product A and product B.
* The year 2016 has an increase in production than 2015. And 2017 is supposed to see a pitfall in production due to instability in the increase of production in 2016.

**Problem:**

To find out which site from which city performs the best to setup new sites for selling the production of product A and B nearby.

**Solution:**

Thus Site E, San Diego performed better than other sites and cities and the production was high in the year of 2016.

References

[1]storytelling with data -a data visualization guide for business professionals by cole nussbaumer knaflic