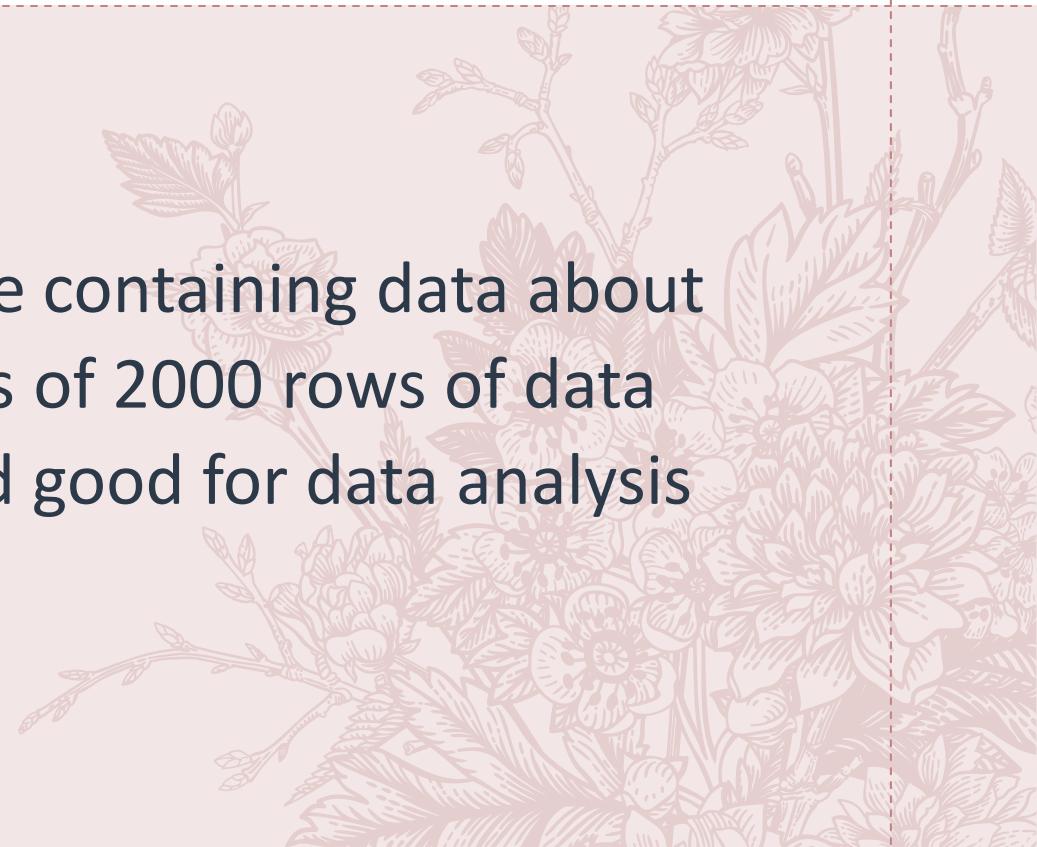


Daily Coffee Revenue

Nechama Luria & Rusi Rothschild

Data source

My partner and I used a dataset from Kaggle containing data about coffee shop daily revenue. This dataset consists of 2000 rows of data from myriad coffee shops, making it robust and good for data analysis and visualization.



7 features in the coffee shop dataset

1. **Number of Customers Per Day**- The total number of customers visiting the coffee shop on any given day.
2. **Average Order Value (\$)**- The average amount spent by each customer during their visit.
3. **Operating Hours Per Day**- The total number of hours the coffee shop is open for business each day.
4. **Number of Employees**- The number of employees working on a given day.
5. **Marketing Spent Per Day (\$)**- The amount of money spent on marketing campaigns or promotions on any given day.
6. **Location Foot Traffic (people/hour)**- The number of people passing by the coffee shop per hour.
7. **Daily Revenue (\$)**- This is the dependent variable representing the total revenue generated by the coffee shop each day.

Problem Statement

Which variable has the greatest impact on **Daily Revenue**? Identifying the key factors that influence **Daily Revenue** helps businesses make informed decisions to improve performance. By understanding which variables have the strongest impact, businesses can optimize their strategies and allocate resources effectively. This insight allows for better planning in the effort maximize profitability.



Importing and cleaning the data

1. Clean the data using python.
2. Preview the first 5 rows of this large dataset to get a feel of the dataset.
3. Check the length of the dataset; how many rows are we dealing with?
4. Check for any null values and proceed with dropping that data.
5. Check for any duplicate values and proceed with deleting those repetitious values.
6. Use the info function to provide a summary of the pandas dataframe that we are working with.
7. Use the describe function to provide a statistical summary of the numeric columns in the dataset.

Screenshots of code

```
[ ]: import pandas as pd
data = pd.read_csv("coffee_shop_revenue.csv")
```

```
[ ] data.head()
```

	Number_of_Customers_Per_Day	Average_Order_Value	Operating_Hours_Per_Day	Number_of_Employees	Marketing_Spend_Per_Day	Location_Foot_Traffic	Daily_Revenue
0	152	6.74	14	4	106.62	97	1547.81
1	485	4.50	12	8	57.83	744	2084.68
2	398	9.09	6	6	91.76	636	3118.39
3	320	8.48	17	4	462.63	770	2912.20
4	156	7.44	17	2	412.52	232	1663.42

```
[ ] len(data)
```

```
→ 2000
```

```
▶ print(data.isnull().sum())
```

	Number_of_Customers_Per_Day	Average_Order_Value	Operating_Hours_Per_Day	Number_of_Employees	Marketing_Spend_Per_Day	Location_Foot_Traffic	Daily_Revenue
count	0	0	0	0	0	0	0
mean							
std							
min							
25%							
50%							
75%							
max							
dtype:	int64						

```
[ ] print(data.duplicated().sum())
```

```
→ 0
```

```
[ ] data.info()
```

```
→ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Number_of_Customers_Per_Day  2000 non-null   int64  
 1   Average_Order_Value        2000 non-null   float64 
 2   Operating_Hours_Per_Day   2000 non-null   int64  
 3   Number_of_Employees       2000 non-null   int64  
 4   Marketing_Spend_Per_Day  2000 non-null   float64 
 5   Location_Foot_Traffic    2000 non-null   int64  
 6   Daily_Revenue             2000 non-null   float64  
dtypes: float64(3), int64(4)
memory usage: 109.5 KB
```

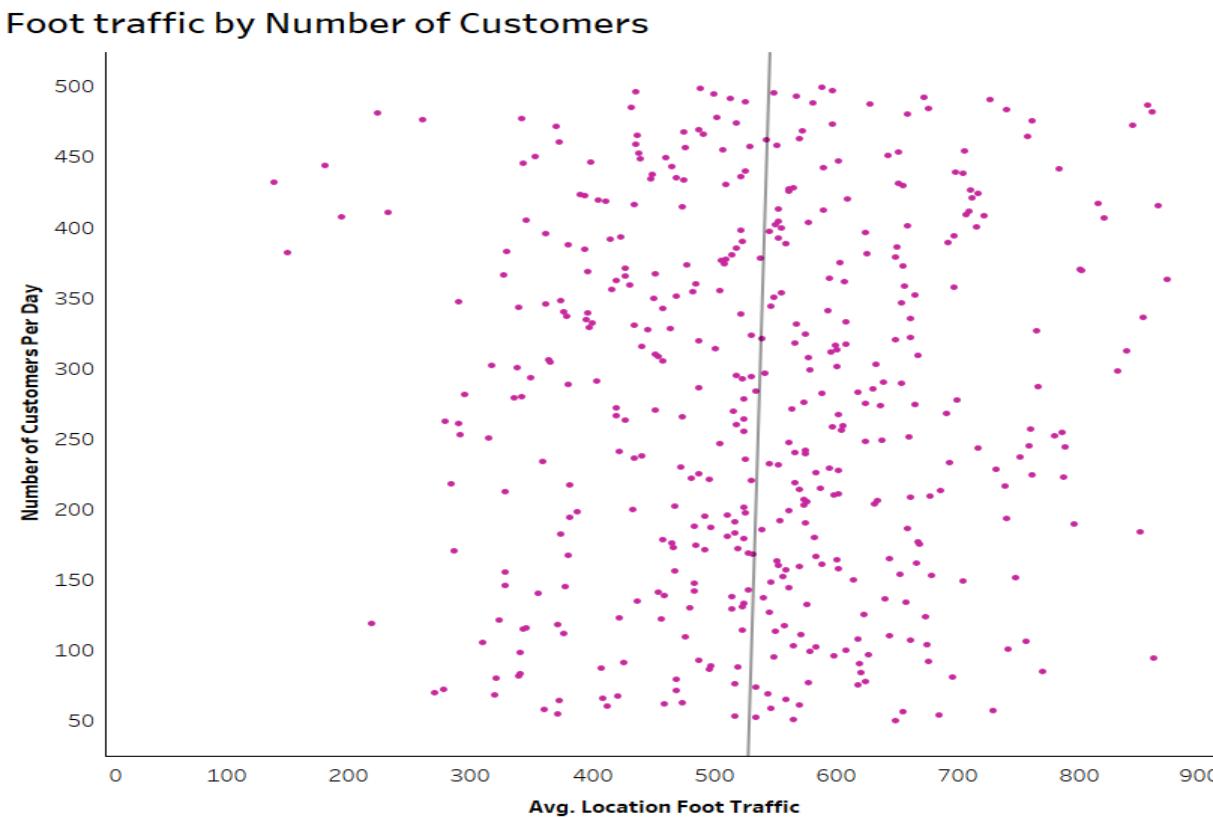
```
▶ print(data.describe(include='all'))
```

	Number_of_Customers_Per_Day	Average_Order_Value	Operating_Hours_Per_Day	Number_of_Employees	Marketing_Spend_Per_Day	Location_Foot_Traffic	Daily_Revenue
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	274.296000	6.261215	11.667000	7.947000	252.614160	534.893500	1917.325940
std	129.441933	2.175832	3.438608	3.742218	141.136004	271.662295	976.202746
min	50.000000	2.500000	6.000000	2.000000	10.120000	50.000000	-58.950000
25%	164.000000	4.410000	9.000000	5.000000	130.125000	302.000000	1140.085000
50%	275.000000	6.300000	12.000000	8.000000	250.995000	540.000000	1770.775000
75%	386.000000	8.120000	15.000000	11.000000	375.352500	767.000000	2530.455000
max	499.000000	10.000000	17.000000	14.000000	499.740000	999.000000	5114.600000

Key questions we aim to answer during our data analysis

#1. Does more location foot traffic correlate to a higher number of customers per day?

Since we are looking to visualize the relationship between two continuous numerical variables, we decided to use a scatter plot. This will help us identify any correlations between the two measures. Scatter plots are also extremely helpful in detecting any outliers.



Analysis

The trend line suggests a very slight positive correlation between foot traffic and number of customers per day. This indicates that locations with higher foot traffic tend to have more customers, but the relationship is not very strong.

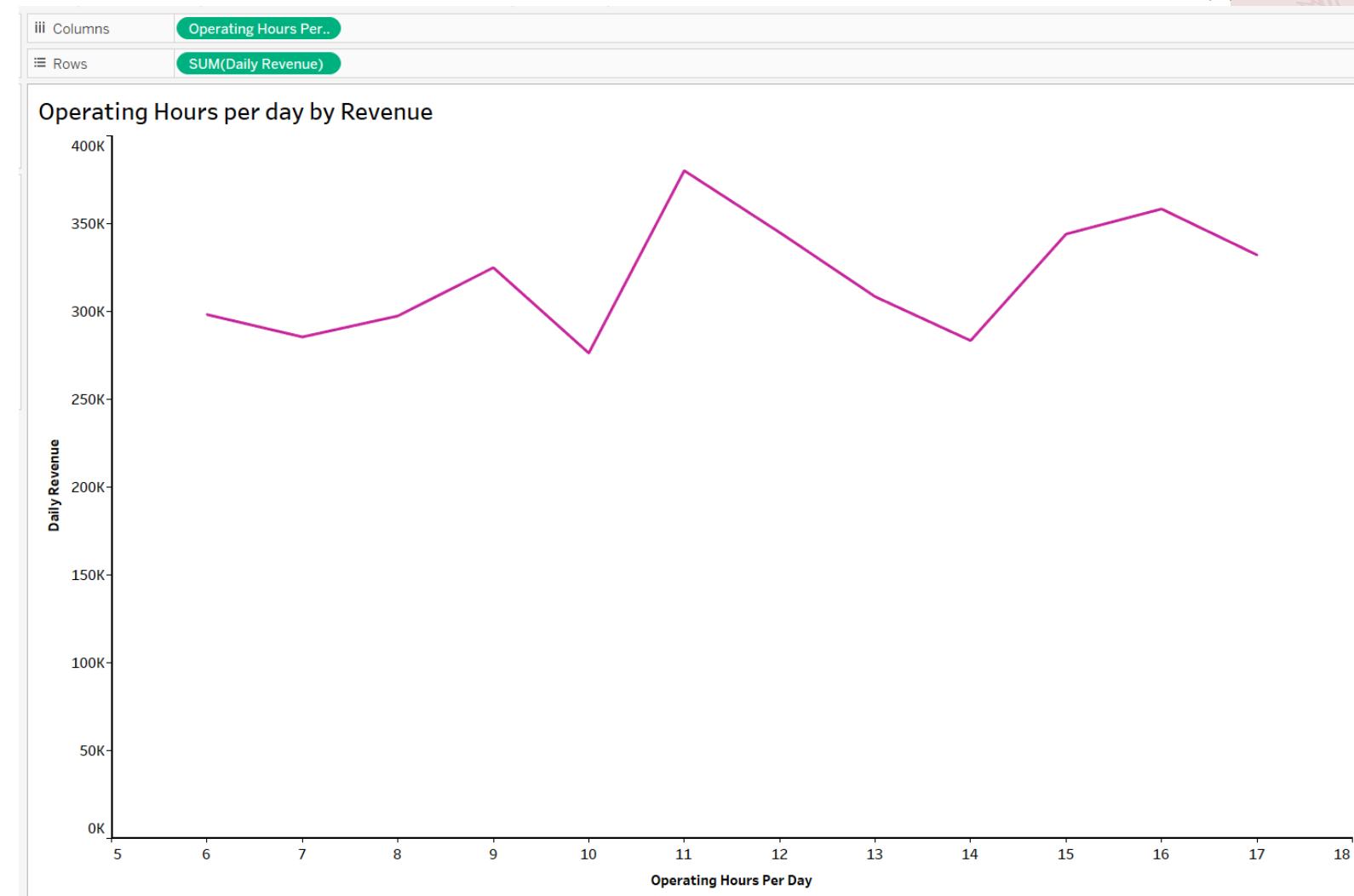
We also notice that there is a high variability; a wide spread of data. The data points are scattered about widely. This indicates that other factors besides foot traffic influence customer count. Further analysis will help us know what variables have a more significant effect on the customer count.

There appears to be a dense grouping of points between 400 and 700 location foot traffic, and 100 and 300 customers. There are a few outliers, where customer counts are significantly higher or lower than expected for a given foot traffic.

Operating Hours per day by Revenue

#2. How much of an effect does operation hours per day have on the daily revenue? If the shop stays open for more hours, does that necessarily mean that the shop will make more money?

We placed operating hours per day in the columns and the sum of daily revenue in the rows. Using these to numerical measures, we created a line graph. To enhance clarity, we set the axis labels and titles to bold black for a sharp, professional look. We also removed grid lines and styled the line in pink to match the workbook's color scheme.



Analysis

The graph reveals that revenue peaks around 11-12 operating hours per day before declining and then rising again at 15-16 hours per day. While there is a slight upward trend, revenue fluctuates rather than increasing steadily with more operating hours, indicating that more hours do not always lead to higher revenue. This suggests that additional factors influence revenue beyond just operating hours.

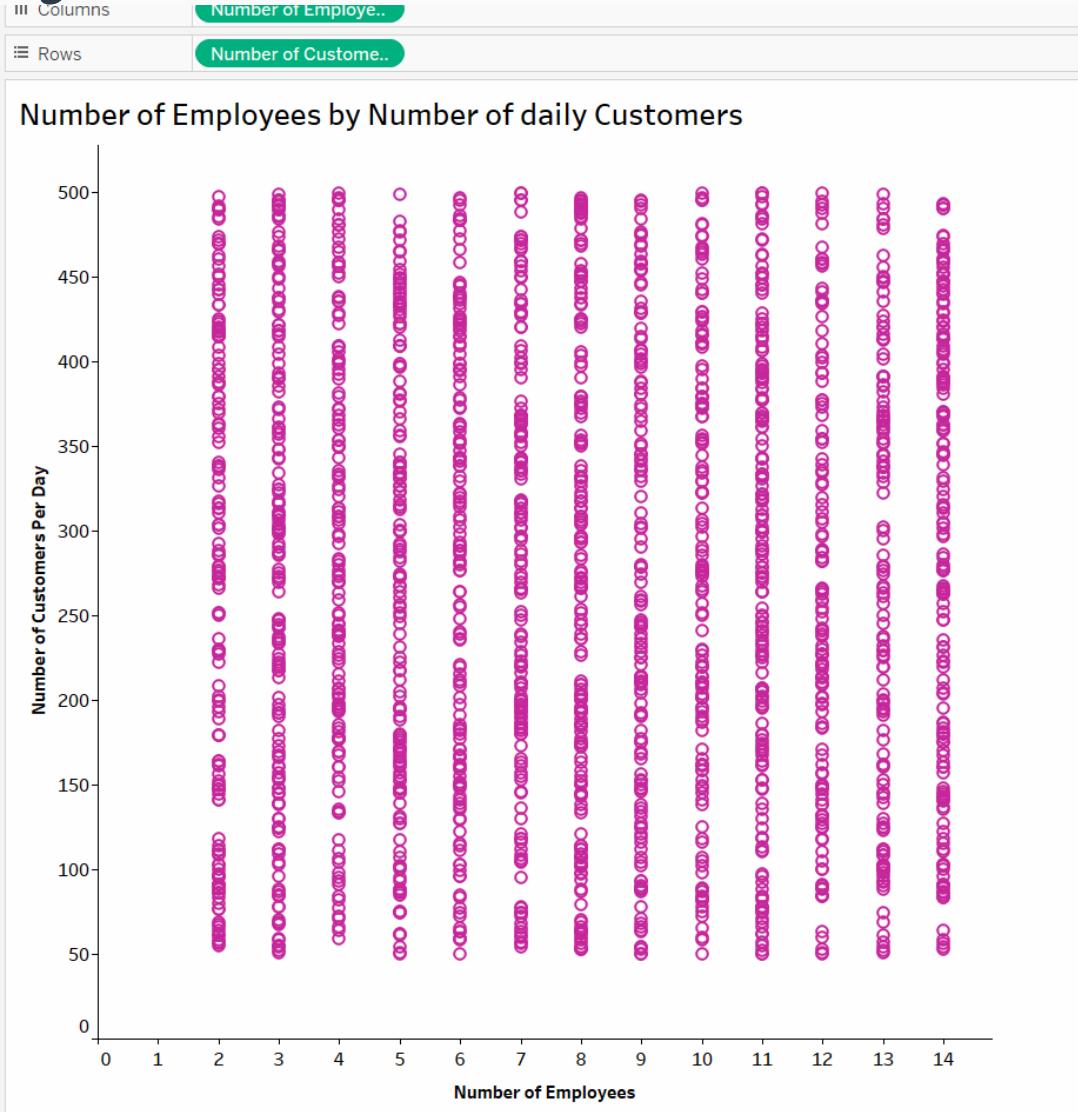
Although this visualization may initially seem unhelpful, it actually provides valuable insights for businesses looking to increase revenue. While it might appear that staying open longer leads to higher earnings, the data reveals that other factors have a stronger impact on revenue.



Employee count by daily customers

#3. Does the number of employees affect the number of customers per day? Would people rather go to a store that is adequately staffed?

Again, since we are dealing with 2 numerical values, we will use a scatter plot to provide insight.



Analysis

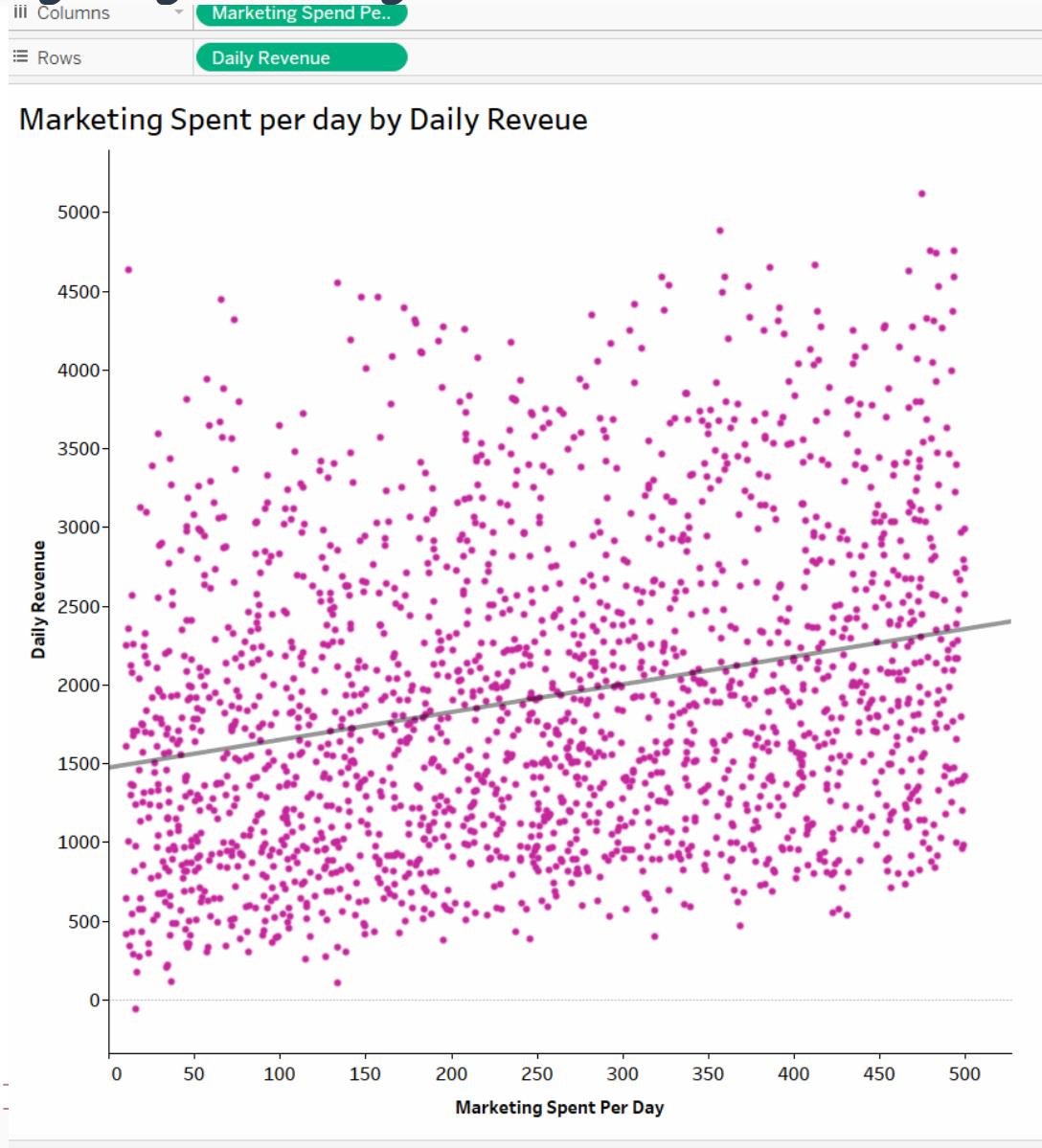
The distribution of data points, or lack thereof, suggests that having more employees does not strongly correlate with an increase in customers. While staffing levels vary, the customer count appears to fluctuate independently. This implies that other factors might be more influential in driving customer traffic rather than just the number of employees present.



Marketing Spent per day by Daily Revenue

#4. How does the amount of money spent on marketing affect the number of customers per day? Does the money spent on marketing strategies succeed in drawing in more customers?

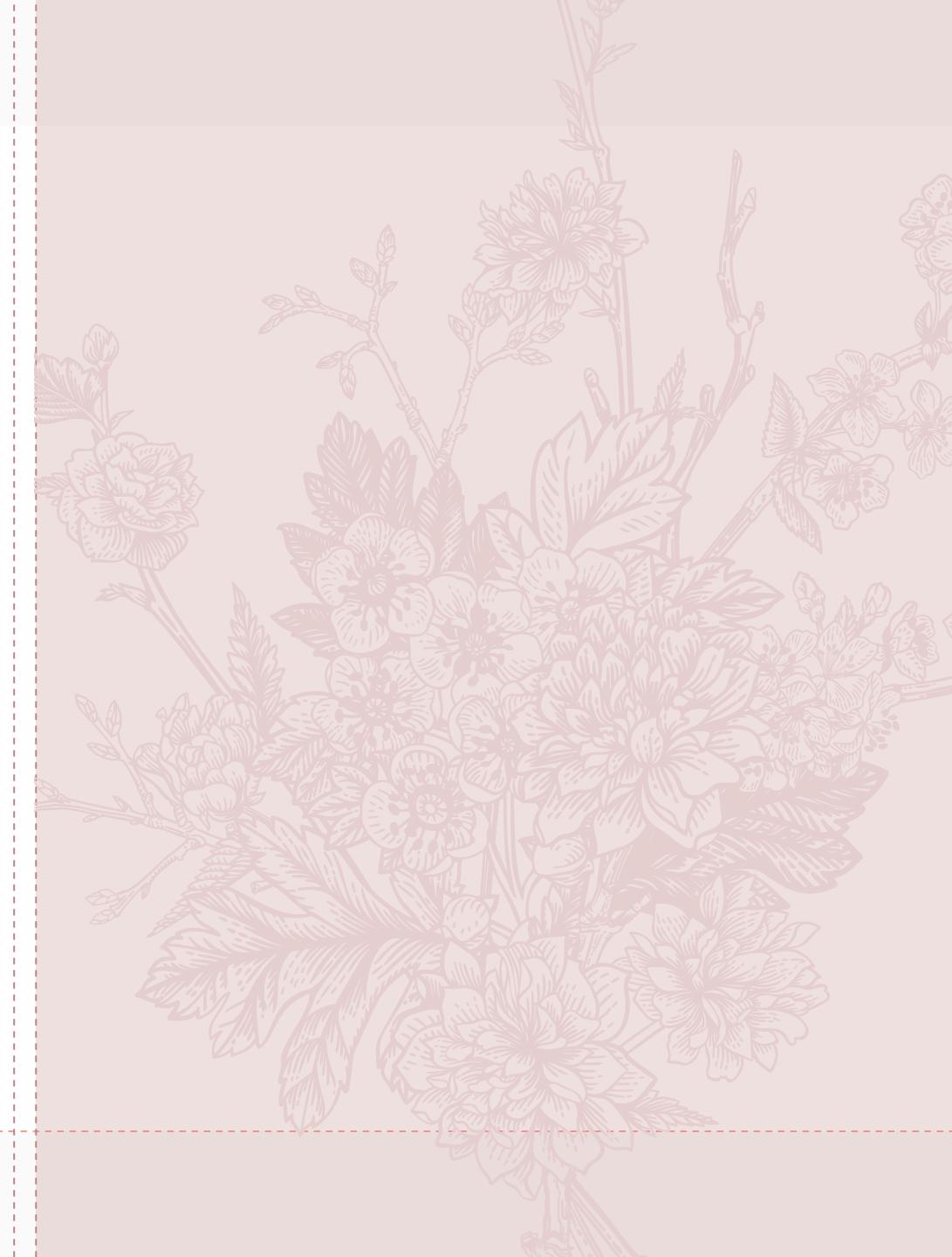
Again, since we are dealing with 2 numerical values, we will use a scatter plot to provide insight.



Analysis

I placed the marketing spend per day along the X-axis and the daily revenue along the Y-axis. The scatter plot indicates a weak positive correlation. The higher marketing spending generally leads to increased revenue, but the effect is not very strong.

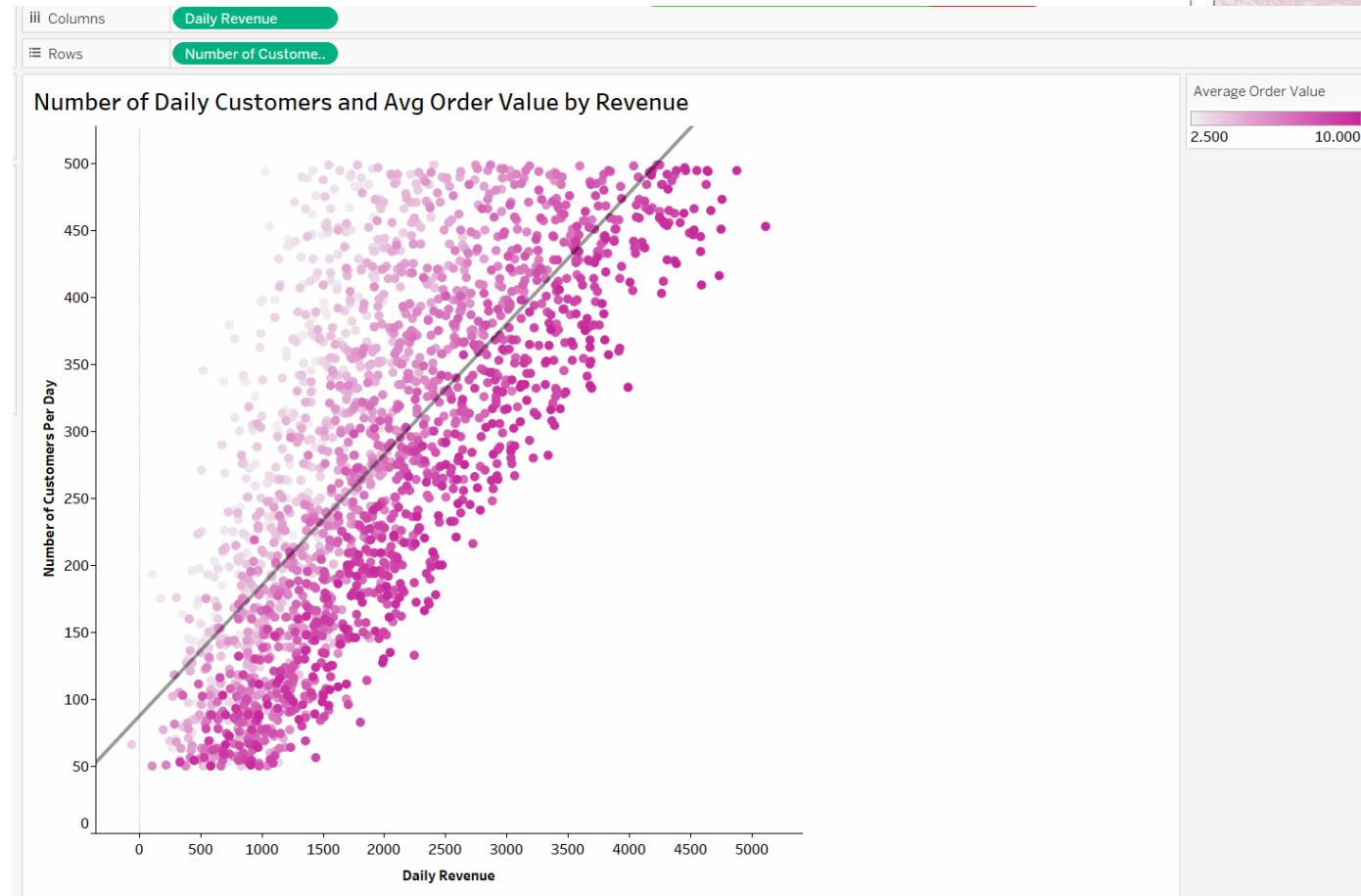
The data points are widely scattered, suggesting that other factors also play a role in revenue generation.



Number of Daily Customers and Average Order Value by Revenue

#5. How do the number of customers per day and the average order value correlate with the daily revenue? Do more customers breed more revenue, or does the average order value have a larger significance on the daily revenue?

Since we are dealing with 2 numeric values, we will use a scatter plot to answer this question. We placed Daily Revenue on the Columns shelf and Number of Customers Per Day on the Rows shelf, with Average Order Value assigned to the color bucket. This setup created a scatter plot, enhanced with a trend line for better visual analysis. To match our workbook's color scheme, we assigned pink to the data points, while using a contrasting black trend line to make it stand out and draw attention.



Analysis

The scatter plot and trendline show a clear positive relationship between Daily Revenue and Number of Customers—more customers lead to higher revenue. The color gradient also highlights that higher Average Order Values (darker shades) contribute to increased revenue. This is likely because higher revenue results not only from more customers, but also from a higher average order value.

While the correlation is strong, there is variation in the data. This is largely due to variability of the average order value, which is apparent when looking at the color gradient. Besides the average order value, there are also other also other factors that influence daily revenue.



Business Implications

- Focus on Customer Acquisition: Increasing the number of customers is a key driver of revenue.
- Increase Average Order Value: Strategies to encourage customers to spend more per order should be explored.
- Analyze High-Performing Days: Investigate the days with high revenue, high customer count, and high average order value to understand the factors contributing to their success.



Ethical Implications



1. Privacy and Data Sensitivity

Even though this dataset appears to focus on business operations, it may contain sensitive business performance data, such as revenue trends and customer traffic. If the dataset includes personally identifiable information (PII), sharing or analyzing it without proper authorization could violate confidentiality agreements.

Mitigation:

- Ensure that all data used for analysis is anonymized and does not expose sensitive business details.
- Obtain necessary permissions before using or sharing the data externally.

2. Misrepresentation and Bias

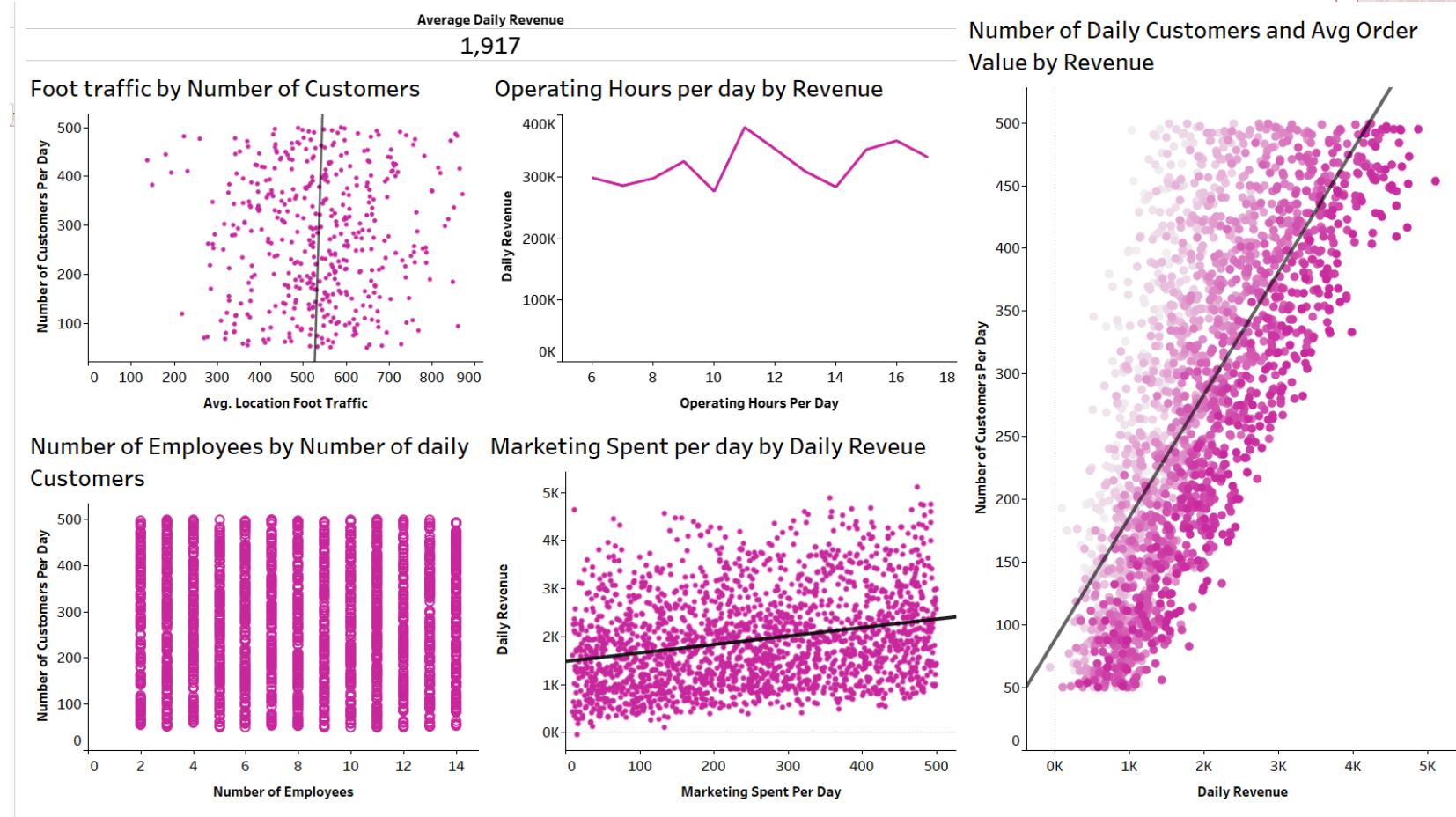
Data can sometimes be interpreted in misleading ways, especially if key factors influencing revenue (e.g., seasonal trends, location variations, or external economic factors) are not considered. If conclusions are drawn without acknowledging potential biases, they could lead to incorrect business decisions.

Mitigation:

- Provide context for any analysis, ensuring that all relevant factors are considered.
- Avoid cherry-picking data points that support a particular narrative while ignoring those that contradict it.
- Use visualizations and statistical methods to validate findings before making claims about causation

Dashboard

We started by creating a KPI named Average Daily Revenue, displaying its value prominently. Next, we incorporated five visualizations, making the fifth one the largest since it provides the most critical insights, highlighting the two variables with the greatest impact on revenue. Finally, we ensured that each visualization was interactive, allowing viewers to explore the data and extract meaningful insights from the dashboard.



In conclusion...

The **Number of Customers Per Day** and **Average Order Value** have the greatest influence on **Daily Revenue**, as more customers lead to more transactions, while a higher order value increases the total revenue per sale. Analyzing the correlation between these variables and revenue can help confirm their impact. Visualizing these relationships through scatter plots and trendlines provides a clear representation of their influence.

