CERTIFICATION - DATA ANALYST - PRODUCT SALES

datalab

# Data Analyst Professional Practical Exam Submission

**You can use any tool that you want to do your analysis and create visualizations. Use this template to write up your summary for submission.**

You can use any markdown formatting you wish. If you are not familiar with Markdown, read the Markdown Guide before you start.

### Task List

Your written report should include written text summaries and graphics of the following:

- Data validation:
  - Describe validation and cleaning steps for every column in the data
- Exploratory Analysis:
  - Include two different graphics showing single variables only to demonstrate the characteristics of data
  - Include at least one graphic showing two or more variables to represent the relationship between features
  - Describe your findings
- Definition of a metric for the business to monitor
  - How should the business use the metric to monitor the business problem
  - Can you estimate initial value(s) for the metric based on the current data
- Final summary including recommendations that the business should undertake

*Start writing report here..*

# Data Validation

| | | sales... | customer_id | | | years_as_cus... | nb_site_... | s. |
|---|---|---|---|---|---|---|---|---|
| 0 | 2 | Email | 2e72d641-95ac-497b-bbf8-4861764a7097 | 10 | | 0 | 24 | Ariz |
| 1 | 6 | Email + Call | 3998a98d-70f5-44f7-942e-789bb8ad2fe7 | 15 | 225.47 | 1 | 28 | Kans |
| 2 | 5 | Call | d1de9884-8059-4065-b10f-86eef57e4a44 | 11 | 52.55 | 6 | 26 | Wisc |
| 3 | 4 | Email | 78aa75a4-ffeb-4817-b1d0-2f030783c5d7 | 11 | | 3 | 25 | Indic |
| 4 | 3 | Email | 10e6d446-10a5-42e5-8210-1b5438f70922 | 9 | 90.49 | 0 | 28 | Illino |

Rows: 5 Expand Table

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15000 entries, 0 to 14999
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   week               15000 non-null  int64
 1   sales_method       15000 non-null  object
 2   customer_id        15000 non-null  object
 3   nb_sold            15000 non-null  int64
 4   revenue            13926 non-null  float64
 5   years_as_customer  15000 non-null  int64
 6   nb_site_visits     15000 non-null  int64
 7   state              15000 non-null  object
dtypes: float64(1), int64(4), object(3)
memory usage: 937.6+ KB
None
```

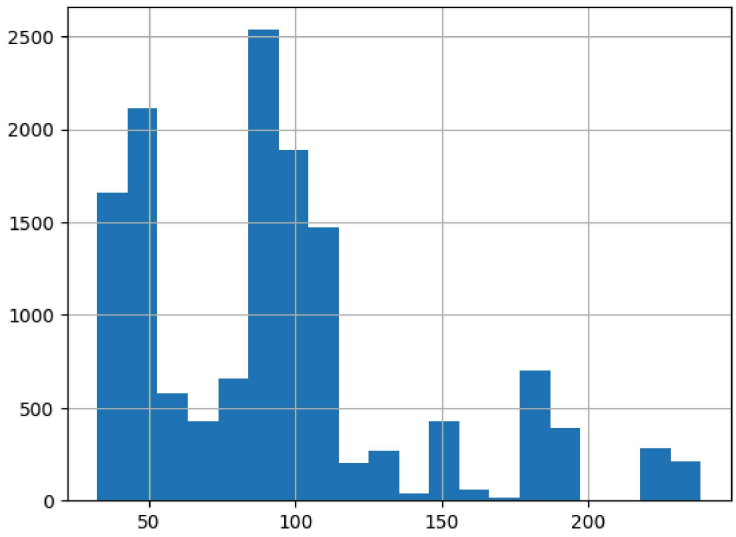
| i... | week | nb_sold | revenue | years_as_customer | nb_site_visits |
|---|---|---|---|---|---|
| count | 15000 | 15000 | 13926 | 15000 | 15000 |
| mean | 3.0982666667 | 10.0846666667 | 93.9349425535 | 4.9659333333 | 24.9908666667 |
| std | 1.6564198071 | 1.8122133327 | 47.4353122457 | 5.0449515589 | 3.5009142152 |
| min | 1 | 7 | 32.54 | 0 | 12 |
| 25% | 2 | 9 | 52.47 | 1 | 23 |
| 50% | 3 | 10 | 89.5 | 3 | 25 |
| 75% | 5 | 11 | 107.3275 | 7 | 27 |
| max | 6 | 16 | 238.32 | 63 | 41 |

Rows: 8 Expand Table

## Dataset Summary

Missing Values: The 'revenue' column has missing values and requires handling.

Summary Statistics: The displayed summary statistics offer insights into data distribution for analysis.




## Data Analysis and Imputation

The histogram plot of the null value column showed that the data was skewed. As a result, we opted to use the median to fill the missing values in the revenue column.

According to the data description the revenue varies depending on which products were sold, and we don't have any product type related column so we can't group the revenue column by any column for the imputation.

| | | sales... | customer_id | | | years_as_cus... | nb_site_... | st... |
|---|---|---|---|---|---|---|---|---|
| 13741 | 2 | Email | 18919515-a618-430c-9a05-2c7d8fea96af | 10 | 97.22 | 63 | 24 | Calif |
| 13800 | 4 | Call | 2ea97d34-571d-4e1b-95be-fea1c404649f | 10 | 50.47 | 47 | 27 | Calif |

Rows: 2 Expand Table

### Unique Values Summary

- Sales Method Case: The 'sales_method' column contains unique values in both title case and lower case. Standardizing the case may be necessary for consistency.
- Email and Call Duplication: The 'email + call' unique value also appears as 'em + call,' indicating a need for data cleaning to address this duplication.
- Years As a Customer: The years as a customer column had customers with years as a customer older than the company's age of 39 years as was stated in the data description, those rows needs to be drop.

### Data Quality Check

- Upon inspection for duplicated values, empty string, and negative values, the analysis revealed that the sales dataframe exhibits none of these issues.
- But We still need to convert the 'sales_method' column to categorical data type for efficiency.

### Data Validation Analysis Summary

At the conclusion of the data validation analysis, the following improvements have been made:

- Revenue Imputation Challenge: Lack of product-type related columns for grouping revenue.
- Missing Values: Addressed missing values in 'revenue' column.
- Unique Values Cleaning: Standardized 'sales_method' case and resolved email and call duplication.
- Data Skewness: Detected skewness in revenue column; used median for missing value imputation in the revenue column.
- Years as Customer: Removed rows with customer years exceeding the company's 39-year age.
- Data Quality Check: Ensured no duplicates, empty strings, or negative values.
- Efficiency Improvement: Converted 'sales_method' column to categorical data type.

## Exploratory Analysis

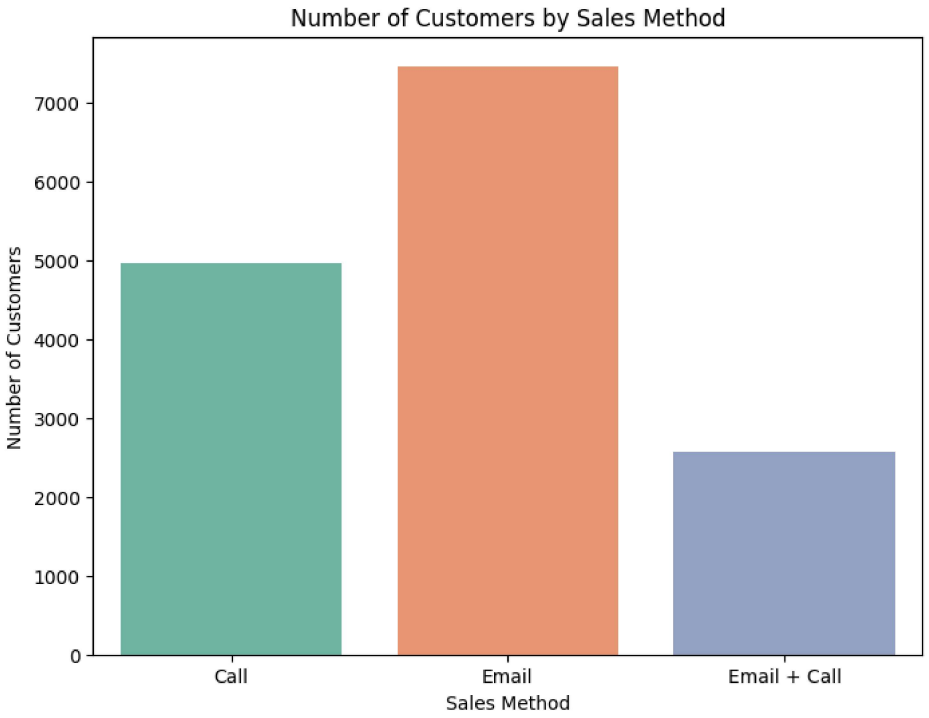
| index | |
|---|---|
| week_outlier | 0 |
| nb_sold_outlier | 586 |
| revenue_outlier | 1031 |
| years_as_customer_outlier | 533 |
| nb_site_visits_outlier | 227 |

Rows: 5 Expand Table

### The numbers represent the count of outliers detected for each respective column:

- For "nb_sold," there are 586 outliers.
- For "revenue," there are 1031 outliers.
- For "years_as_customer," there are 531 outliers.
- For "nb_site_visits," there are 227 outliers.

### Number of Customers for Each Sales Method



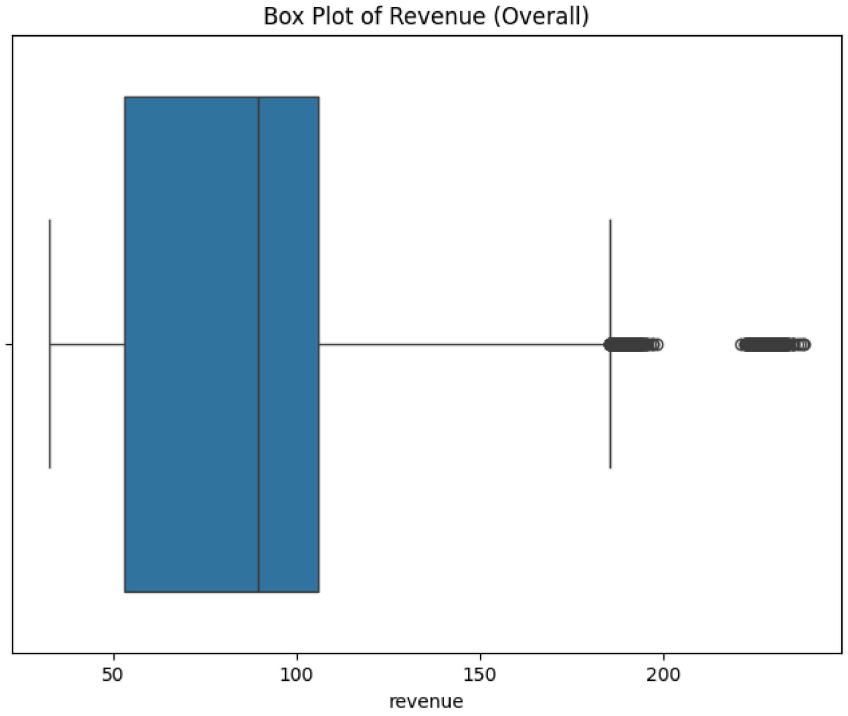

| sales... | cus... |
|---|---|
| Call | 4962 |
| Email | 7466 |
| Email + Call | 2572 |

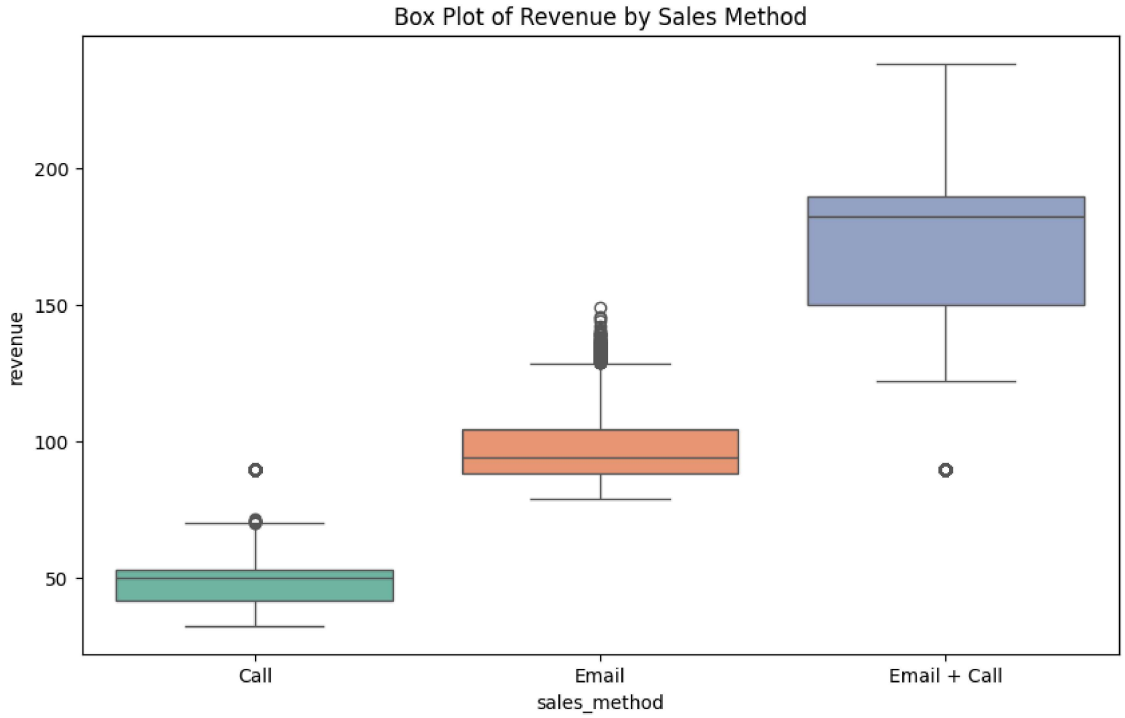
Rows: 3 Expand Table

**Based on the bar plot results:**

- For the "Call" approach, there were 4,962 customers.
- For the "Email" approach, there were 7,466 customers.
- For the "Email + Call" approach, there were 2,572 customers.
- These numbers represent the count of customers for each sales approach, as determined from the bar plot

## Spread of the Overall Revenue, and for Each Sales Method




Summary Statistics for Spread of Revenue Overall:

| | revenue |
|---|---|
| count | 15000 |
| mean | 93.6174006667 |
| std | 45.7197752341 |
| min | 32.54 |
| 25% | 53.04 |
| 50% | 89.5 |
| 75% | 106.07 |
| max | 238.32 |

Rows: 8

Expand Table

Summary Statistics for Spread of Revenue for each Method:

| sales... | | mean | std | | | | | |
|---|---|---|---|---|---|---|---|---|
| Call | 4962 | 49.12595526 | 11.5390396422 | 32.54 | 41.63 | 49.935 | 52.9775 | 89.5 |
| Email | 7466 | 96.5719032949 | 10.9748454702 | 78.83 | 88.39 | 94.275 | 104.46 | 148.97 |
| Email + Call | 2572 | 170.8756570762 | 42.0841626072 | 89.5 | 149.8225 | 182.135 | 189.535 | 238.32 |

Rows: 3 Expand Table

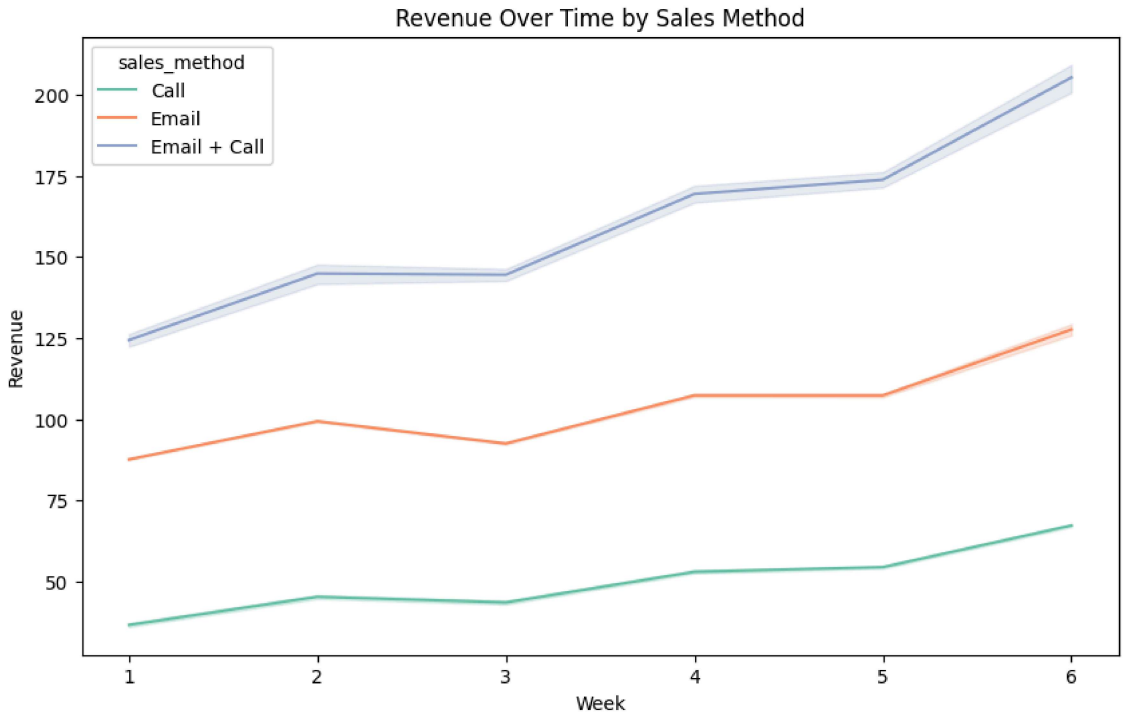
### Summary of Overall Spread of Revenue.

- Mean (Average): The typical revenue from sales is around $93.62.
- Maximum: The highest revenue observed in the dataset is approximately $238.32.
- Minimum: The lowest revenue observed in the dataset is about $32.54.
- Spread (Standard Deviation): Revenue values can vary, with most falling within approximately $45.72 of the average. This tells us that while some sales generate more revenue, others generate less.
- 25th Percentile: About 25% of sales have revenue below $53.04, indicating that a quarter of the sales generate less revenue.
- Median (Middle Value): The middle revenue value is about $89.50. Half of the sales generate more revenue than this, and half generate less.
- 75th Percentile: About 75% of sales have revenue below $106.07, indicating that most sales fall below this threshold.

### Summary of Revenue Spread by Sales Method

- "Email + Call" generates the highest average revenue but with higher variability, indicating that it can result in both higher and lower revenue outcomes.
- "Email" generates moderate average revenue with moderate variability.
- "Call" generates lower average revenue with lower variability.

## Difference in Revenue over Time for each Sales Methods



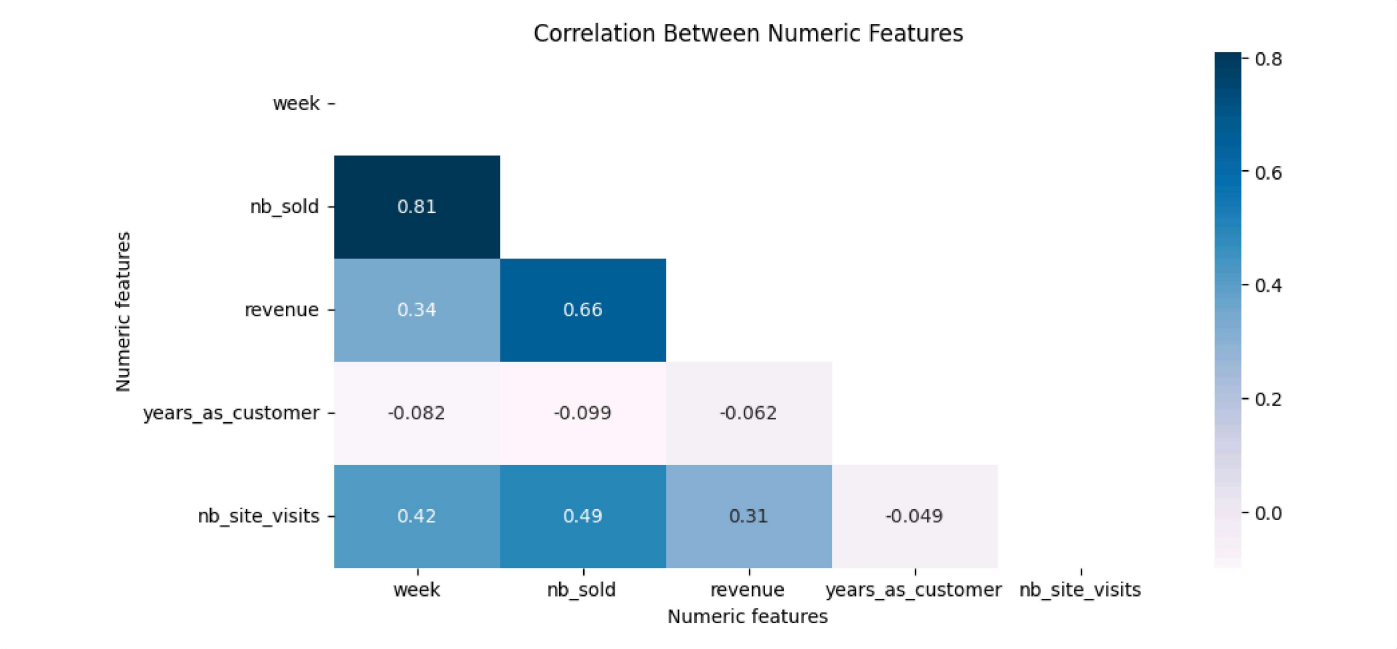
### Analysis of Revenue Change Over Time

Based on the line plot provided, which shows the change in revenue over time for each sales method, here is the answer to the question: "Was there a change in revenue over time?"

- For the "Email" sales method, revenue also shows an increasing trend over the six weeks, starting at 99.94 in the first week and rising to 132.01 in the sixth week.
- Similarly, for the "Email + Call" sales method, revenue exhibits a consistent upward trend, starting at 128.87 in the first week and reaching 225.47 in the sixth week.
- For the "Call" sales method, revenue appears to increase gradually over the six weeks, starting at 34.98 in the first week and reaching 65.01 in the sixth week. This suggests a positive trend in revenue for this method.

Based solely on the revenue trends observed in the provided data, here's a recommendation:

Email + Call Sales Method: This method generated the highest average revenue of approximately $170.88, and it exhibited a consistent upward trend in revenue over the six-week period. While it has higher variability, it also has the potential for higher revenue outcomes.




### Pairplot and Heatmap Intepretation

From the pair plot we can visualize the relationship between each features using the sales method as the hue, to vidualize the effect of each sales method on the features relationship. below is the explanation of each feature relationship;

- Week vs. nb_sold: Strong positive correlation (0.81), suggesting that as weeks progress, more new products tend to be sold.
- Week vs. Revenue: Moderate positive correlation (0.34), indicating that, on average, revenue increases with time, but less strongly than nb_sold.
- Week vs. nb_site_visits: Moderate positive correlation (0.42), suggesting an increase in site visits over time.
- Week vs. Years as Customer: Weak negative correlation (-0.08), implying that as weeks progress, average customer tenure slightly decreases.
- nb_sold vs. Revenue: Strong positive correlation (0.66), showing that higher new product sales are associated with increased revenue.
- Revenue vs. Years as Customer: Weak negative correlation (-0.06), indicating that longer-term customers may generate slightly less revenue.
- nb_site_visits vs. Revenue: Moderate positive correlation (0.31), suggesting that more site visits are associated with increased revenue, although less strongly than with nb_sold.
- nb_site_visits vs. nb_sold: Strong positive correlation (0.49), meaning more site visits are linked to higher new product sales.

## Metric for the Business to Monitor

- Metric: "Weekly Revenue"
- Definition: Measures weekly income generated from sales.
- Reason for Choosing: The weekly revenue was chosen as the metric to monitor because it provides a direct measure of the business's financial performance, allowing for real-time tracking of income generated from sales, which is a critical indicator of business success.

## Metric Initial Value Based on the Current Data

**Initial Value is $292,858**

Initial Weekly Revenue = Revenue from "Call" Sales (week1) + Revenue from "Email" Sales (week1) + Revenue from "Email + Call" Sales (week1)

Summing the revenue from the three sales methods provides an aggregate measure of the business's total weekly revenue. Choosing the first week as the initial value establishes a baseline for tracking revenue growth and performance over time.

### Summary and Recommendations:

- Based on the analysis of the provided data, the following recommendations are suggested:
- The business should focus on monitoring the "Weekly Revenue Growth Rate" as a key performance metric to track changes in revenue over time.
- Implementing a system for real-time tracking and reporting of this metric is essential for making informed decisions.
- The analysis indicates that all three sales methods ("Call," "Email," and "Email + Call") have shown positive revenue growth trends. However, the "Email + Call" sales method has exhibited the highest average revenue and consistent upward growth.
- Therefore, it is recommended that the business continues to emphasize and optimize the "Email + Call" sales method, as it has the potential for higher revenue outcomes.
- It's important to consider that this recommendation is based on revenue trends alone and should be complemented by a holistic assessment of other factors, such as cost-effectiveness, customer satisfaction, and overall business objectives.

### ✅ When you have finished...

- Publish your Workspace using the option on the left
- Check the published version of your report:
  - Can you see everything you want us to grade?
  - Are all the graphics visible?
- Review the grading rubric. Have you included everything that will be graded?
- Head back to the **Certification Dashboard** to submit your practical exam report and record your presentation