📝 Task List

Written report includes written text summaries and graphics of the following:

- Data validation:
  - Validation and cleaning steps for every column in the data
- Exploratory Analysis:
  - Included two different graphics showing single variables only to demonstrate the characteristics of data
  - Included at least one graphic showing two or more variables to represent the relationship between features
  - Described my 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

Summary report

Four steps overview :

1. Data validation
2. Exploratory Analysis
3. KPIs and benchmarks
4. Conclusion and recommendations

I. Data validation

Describe validation and cleaning steps for every column in the data

1. I used `df.info()` function to display basic information about the dataframe

- there are `8` columns with `15000` rows
- `revenue` variable is currently set as `object` and will need to be changed to `numeric` as part of data type validation
- noticed "NA" values under `revenue` column
- checked for overall `data consistency` like `data type` and `logic`

2. Next, I've used `print(df.isna().sum())` to check for missing values

- `revenue` has 1074 values
- all other variables have 0
- untimatly decided to convert revenue to numeric and keep null values due to insignificant portion of the dataset only 7%

3. Categorical values validation Discovered that `sales_method` variable had two mislabled categories

- used `.unique()` method to identify unique categories
- used `sales mapping` techique to combine all categories into `three` main
- confirmed `state` variable has correct values `print(df['state'].unique())`

4. Checked for dublicated values

- Confirmed no duplicate rows in the dataset

5. Data type validation

- converted `revenue` variable to `float64` using `pd.to_numeric(df['revenue'], errors='coerce')` technique

6. Checked for outliers and inconsistent data

- checked for negative values, ensuring data integrity in revenue or nb_sold
- used `df[df['nb_sold'] < 0]` technique
- used `boxplot` vizualization to check for outliers

7. Explored patterns in missing values

- analyzed the `revenue` variable in relation to `state` and `sales_method` to identify any patterns in the missing values
- found no significant correlations or relevant patterns, leading to the decision to retain the missing values in the dataset

8. Feature correlation analysis:

- performed a correlation analysis to understand the relationships between numerical variables.
- used heatmap to visualize these correlations

Conclusions

a) During my investigation of the `product_sales` dataset, I found 8 columns and 15,000 rows of data spaning over six weeks since the initial product launch.

b) I checked for **missing** values and found 7% of the `revenue` variable was missing data. I decided to **keep** these rows instead of dropping due to a small percentage (7%).

c) Later I checked the **data types** and found that `revenue` was stored as an object. I **converted it to** `float64` to allow for statistical analysis.

d) Furthermore, I discovered that `sales_method` variable, which had three main categories and two additional ones which seemed mislabeled. To standardize the data, I merged these two additional categories with the main three using a **mapping dictionary**. This ensured consistent grouping of sales methods into defined categories.

II. EDA

a) Distribution of sales over time :

I was interested in seeing how sales were distributed over time. To perform this, I used the `week` variable and calculated the `value_counts()` to determine the number of sales occurences per week. I found that week 1 had the highest rate at 25%, followed by weeks 2 through 5 with around 17% each. The lowest sales occured in week 6, of around 8%.

*number of transactions

b) Number of new products sold :

Next, I was curious to learn the typical number of new products sold, therefore I used `nb_sold` variable using a boxplot. This visualization helped me understand the median, quartiles, and range of data. I learned that the median number of new products sold is 10. This indicates that 50% of the time, 10 or fewer new products were sold. As we can see in the boxplot, the overall range of the data spans from approximately 7 to 16 units sold. This means that the minimum number of new products sold was around 7, while the maximum was around 16. The interquartile range (IQR), represented by the box, spans from approximately 9 to 11 units sold, indicating that the middle 50% of the data falls within this range, while the most common number of products sold is between 9 and 11.

c) Uncovering a relationshp between variables :

I used a correlation matrix to see the strenght of relationships between variables in our sales dataset. I found two that stood out like `number_of_product_sold` and `week` sales were made since the product launch. I observed strong correlation of `.81` indicating more sales were made in the later weeks. Additionally, I discovered a moderate correlation of `.49` between `number_of_product_sold` and `number_of_site_visits`, suggesting that these increase together. Finally, `week` and `number_of_site_visits` displayed moderate correlation of `.42`.

I used a correlation matrix to assess the strength of relationships between variables in our sales dataset. Two stood out: the `number_of_product_sold` and the `week` in which sales were made since the product launch. I observed a strong correlation of `.81`, indicating that more sales were made in the later weeks. Additionally, I found a moderate correlation of `.49` between the `number_of_product_sold` and the `number_of_site_visits`, suggesting that these tend to increase together. Finally, the `week` and the `number_of_site_visits` showed a moderate correlation of `.42`.

It was interesting to learn that `number of years as a customer` does not have a strong relationship with other variables, indicating that customer loyalty may be less relevant than expected.

d) Linear regression :

Based on the observed correlations, I decided to perform a linear regression to visualize sales over time and explore the relationship between `number_of_items_sold` and the `week` of the product launch. The results provide compelling evidence that the number of items sold increases in the later weeks.

e) Sales method by state :

I identified five top performing states (California, Florida, Illinois, New York, and Texas) for the number of items sold by sales method. I found that emailing customers is the dominanting tactic across these states, however, calling is the second best, and in California is still results in large number of sales. This suggests that a more **personal approach** through calls is effective in these states and could be further explored. The Email + Call method plays a smaller role across states, but still contributes moderately, especially in Florida and Illinois. However, the separate use of Email and Call seems more effective than their combined use, indicating that a sequential approach (first emailing, followed by calls) might perform better.

f) Sales method by site visits :

Data suggests that understanding the number of visits a customer makes on the site (KYC) can help in determining the most effective sales method approach.

- It's generally best to **CALL** when a customer makes 15 to 18 visits
- **Emailing*** results in highers sales for those who visit the site 19 to 30 times
- And **Email + Call** is best for those who visit the site 31+ times

g) Average sales amount :

The Average Order Value (AOV) is 94€. Most sales range between 52€ and 107€, but the relatively high standard deviation of 47€ indicates significant variability in transaction amounts. This variability may suggest the presence of different customer segments or variations in the types of purchases made.

h) Customer segmentation :

Sales vary depending on several factors that we've seen above, such as site visits and sales method. Specifically, **Email + Call** `sales_method` results in highest sales amount on average.

III. KPIs and benchmarks

The key metric for the business to monitor should include :

- Average order value (AOV) : Currently 94€
- Average number of items sold per order : Currently 10 items

These **two** metrics provide insight into both the monetary value and the volume of sales, allowing the business to track overall sales performance and customer purchasing behavior.

1. **How the business should use these metrics to monitor the business problem?**

Average Order Value (AOV) :

The AOV helps the business understand how much revenue is generated from each transaction. By tracking AOV over time, the business can evaluate the effectiveness of its pricing strategies, upselling efforts, and overall customer value. The business should aim to increase AOV through targeted strategies, such as promoting higher-value products, offering discounts for larger orders, or bundling products.

Number of items sold per order :

Monitoring the average number of items sold per order reveals customer purchasing patterns. It indicates how successfully sales reps are upselling or cross-selling products. Sales reps should be incentivized to aim for selling 9 to 11 items per order to maintain or improve current sales levels. This range aligns with the current average but offers room for optimization, particularly through targeted sales efforts or promotional campaigns.

2. **Estimated initial values based on current data**

- Current Average Order Value (AOV): 94€
- Current Average Number of Items Sold per Order: 10 items

By using these initial benchmarks, the Pens and Printers business can establish a baseline and set goals for improving AOV and the number of items sold per order. Over time, tracking these metrics will help the business identify trends, measure the impact of sales strategies, and optimize sales performance.

IV. Conclusion and recommendations

1. **Sales method :**

Emailing customers remains the most effective sales method across the board, but incorporating a `Call` strategy could enhance sales, particularly in states like California and Texas where it already performs well. A combined or sequential strategy (starting with `Email`, followed by `Calls`) may lead to better results, particularly in regions where `Call` sales are already significant.

a. **State-specific insights :**

- California: Both `Email` and `Call` methods are highly successful, indicating the potential for even greater results through a multi-channel approach.
- Florida and New York: Similar to California, `Email` is dominant, but `Call` also plays a key role. There may be potential to improve sales further by optimizing the `Call` strategy.
- Illinois and Texas: In these states, while `Email` leads, the gap between Email and Call is smaller, suggesting that personalized outreach through `calls` might resonate more with customers here.

b. **Site visitor frequency insights :**

- For customers who visit the site 15 to 18 times, a phone `Call` tends to be the most successful approach.
- `Email` generates higher sales for those visiting 19 to 30 times.
- For customers with 31 or more visits, a combination of `Email + Call` proves to be the most effective.

c. **Week-specific insights :**

- For customers who are in the 1st to 4th weeks since product launch, an `Email` tends to be the most successful approach.
- `Call` generates higher sales for those in the 5th week.
- For customers in the 6th week, a combination of `Email + Call` proves to be the most effective.

2. **Top-performing states for revenue and sales :**

Analysis revealed that the top-performing states (California, Florida, Illinois, New York, and Texas) and that `emailing` is the dominant sales tactic, followed by `calling`. Interestingly, California shows strong sales from `calls`, suggesting a personal approach is effective. While the combined `Email + Call` method contributes moderately, *separate* `email` and `call` strategies seem more effective overall, hinting that a sequential approach might yield better results.

3. **Customer loyalty :**

Surprisingly, customer tenure shows **no strong relationship** with key metrics like `revenue`, `items sold`, or `site visits`. This suggests that customer loyalty, as measured by tenure, may not be a reliable predictor of sales or engagement. This could indicate transactional customer relationships, evolving customer needs, or a need for improved loyalty programs to incentivize long-term customers.

Sales rep deck

We need to know:

- How many customers were there for each approach?
- What does the spread of the revenue look like overall? And for each method?
- Was there any difference in revenue over time for each of the methods?
- Based on the data, which method would you recommend we continue to use? Some of these methods take more time from the team so they may not be the best for us to use if the results are similar.
- Are there any other differences between the customers in each group?

Number of customers for each approach :

Revenue spread overall and for each sales method :

Revenue over time :

Recommend prioritizing Email, as well as sequencial Email + Call due to it's efficiency and effectiveness :

Customer engagements, state, and weeks since the product launch :