CERTIFICATION - DATA ANALYST - PRODUCT SALES

Data Analyst Professional Practical Exam Submission

📝 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

The dataset consists of 15,000 rows and 8 columns before cleaning and validation. I have validated all the columns against the criteria in the dataset table:

- **week**: without missing values, same as description. No cleaning is needed.
- **sales_method**: 5 unique values without missing data. Has different names but with the same meaning: 'Email + Call' and 'em + call'; 'Email' and 'email'. Requires data standardisation.
- **customer_id**: 15,000 unique IDs, same as description. No cleaning is needed.
- **nb_sold**: numeric values, same as the description. No cleaning is needed.
- **revenue**: numeric values, same as the description. The column has around 7% of missing entries
- **years_as_customer**: has 2 year numbers that are more than the company has existed (more than 41 years)
- **nb_site_visits**: numeric values, same as the description.
- **state**: 50 unique states. Same as description.

So in summary, sales_method needs standardisation, revenue's missing data needs to be filled, and years_as_customer's two entries should be verified by the sales department for correctness.

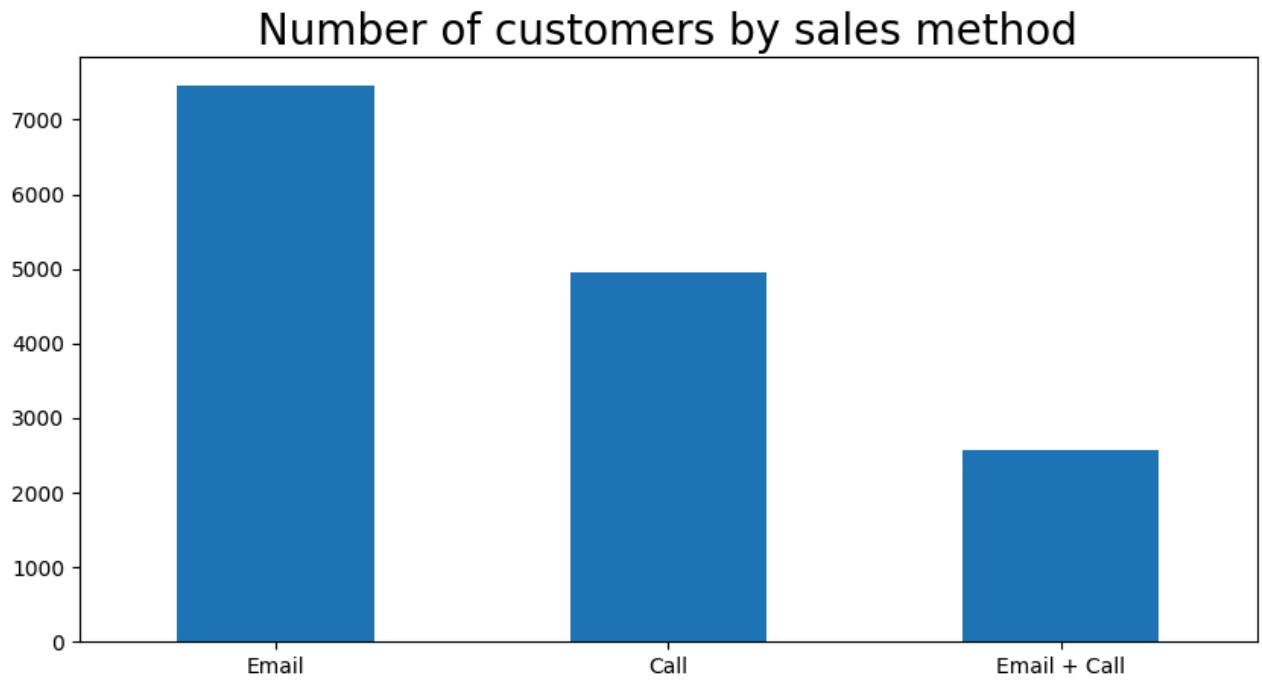
What I have cleaned

1. I have standardised the sales_method column to three unique values: 'Email', 'Email + Call' and 'Call'
2. I have filled missing values in the revenue column with the median (because it is robust against outliers)
3. I replaced two entries for the years_as_customer column with the number of years the company has existed (41 years)

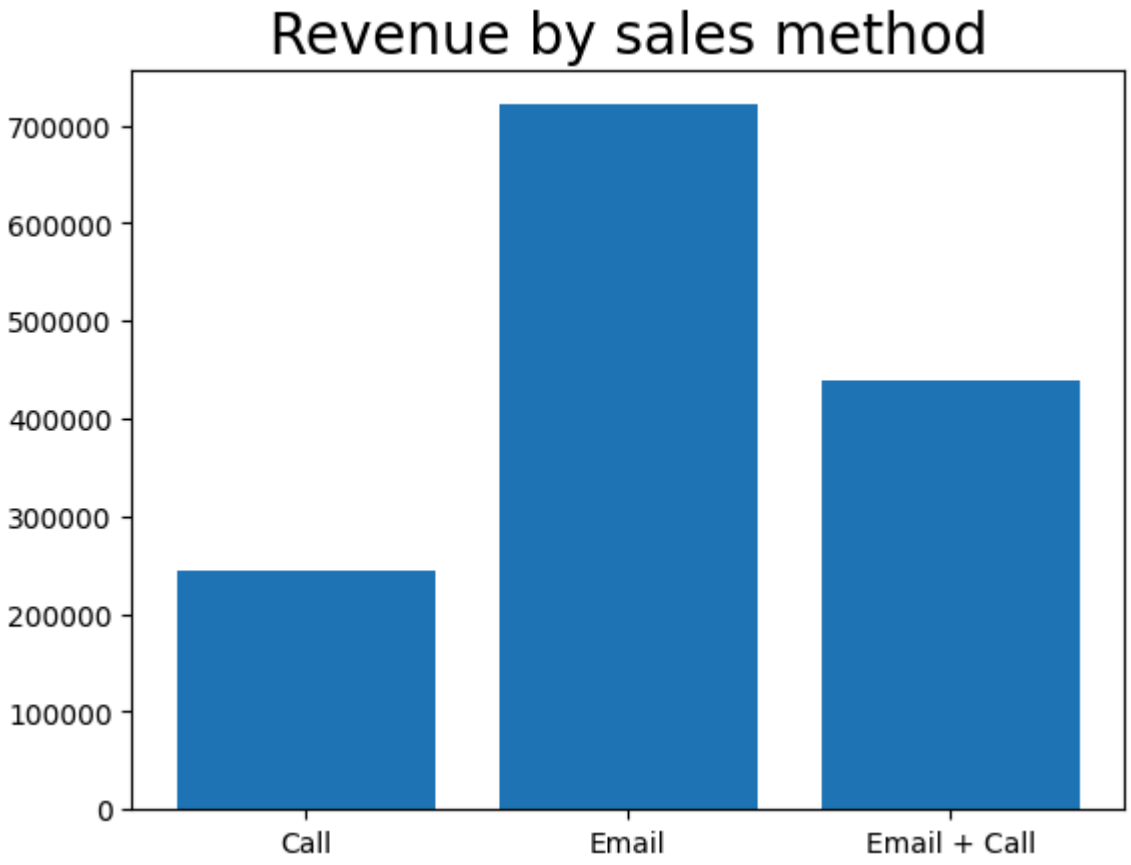
After the data validation, the dataset contains 15,000 rows and 8 columns without missing values.

Analysis

How many customers were there for each approach?

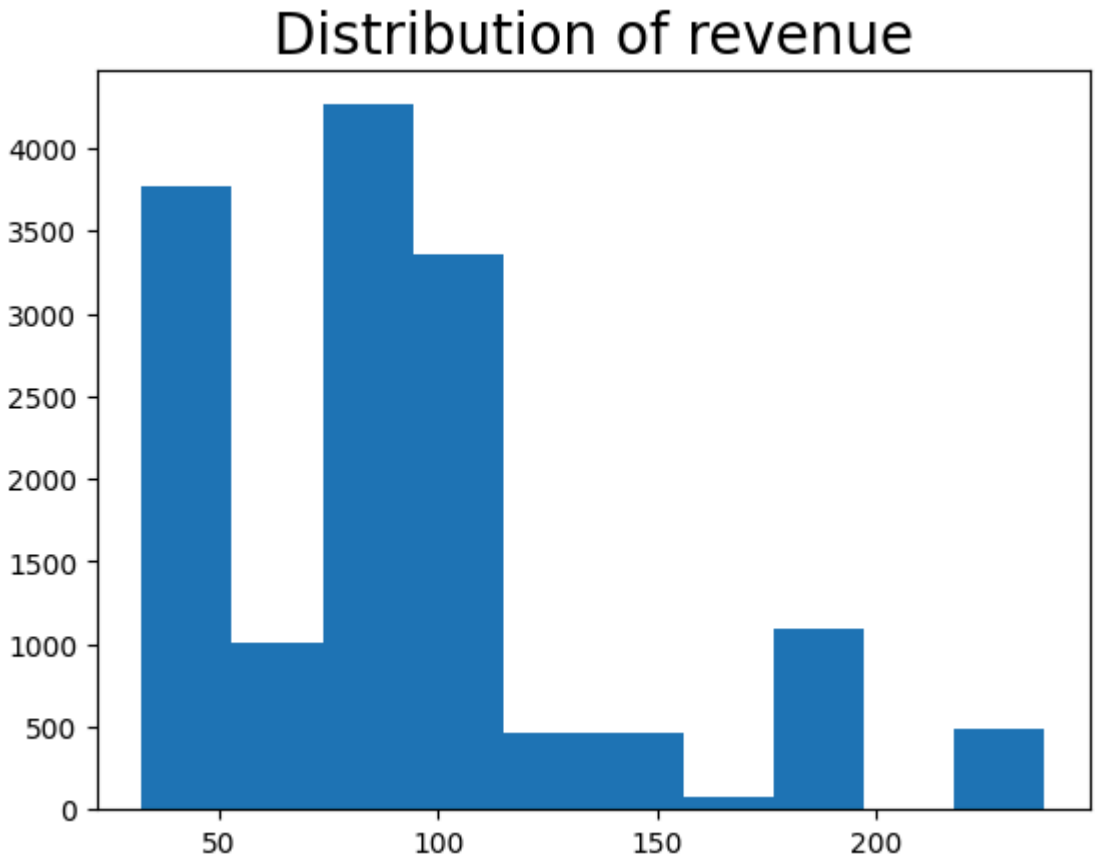


Almost half of all customers were contacted through email. It requires the least effort from the sales team, so there are more companies (potential customers) which have been contacted through emails, leading to a higher customer amount with this method.

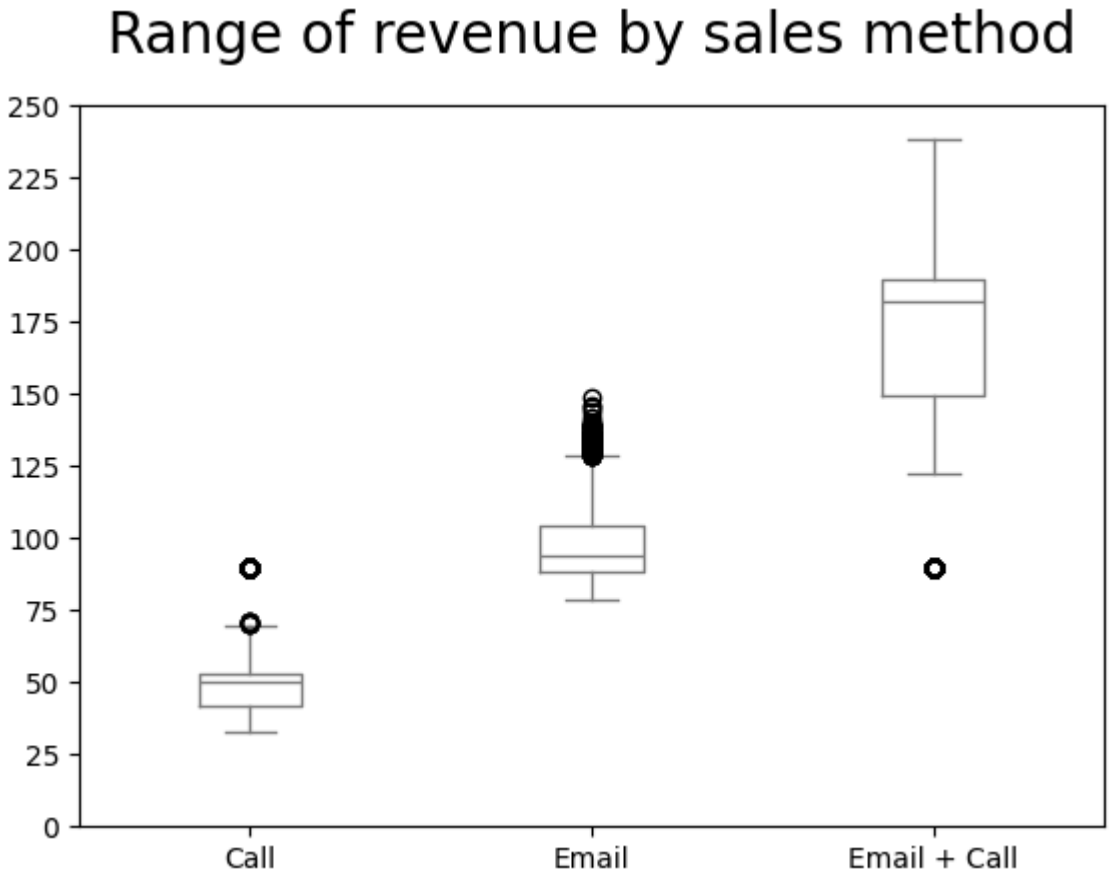


Email method also brought most of the revenue

What does the spread of the revenue look like overall? And for each method?

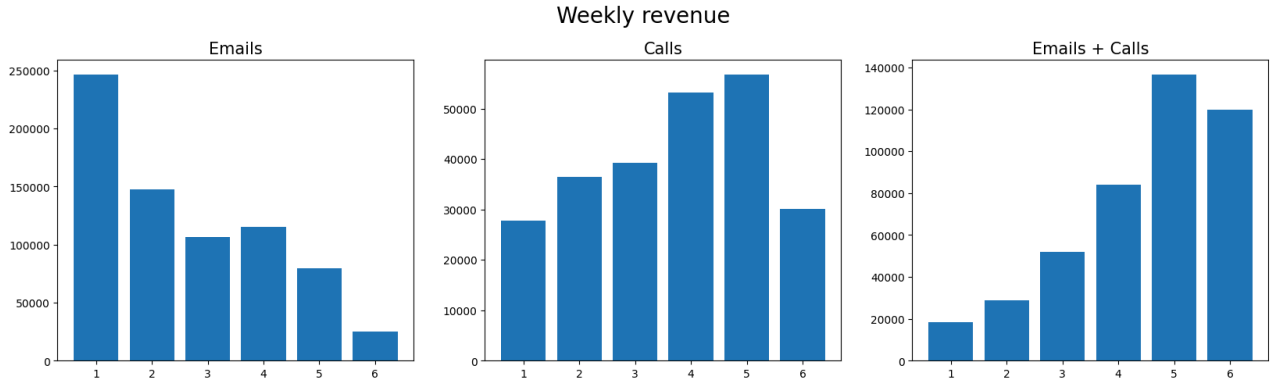


The majority of revenue is concentrated under 120 USD.



We can see that Emails are concentrated under 110, calls under 55, and emails + calls are more spread out and have higher values overall. This shows us that calling a customer does not lead to higher revenue compared to the Email method, despite greater effort by the sales team. But if the sales team emails and then calls a customer, this leads to increased revenue.

Was there any difference in revenue over time for each of the methods?



Emails' method has seen a decline in revenue over time, with the first week being at ≈ 246,000 USD and the 6th week at 25,000 USD (a 90% decline). The second week also showed a significant drop of 40%. Average week-to-week change was minus 31% It is a significant drop, which could be because the first email was sent at the new product line launch, which created a 'wow-effect'.

The Calls method has seen a steady growth (average 6.9%) with two major increases at week 2 (+31%) and week 4 (+35%). A major drop occurred at week 6 (-46%)

The emails + calls method showed the highest weekly average revenue increase of almost 50%.

Based on the data, which method would you recommend we continue to use?

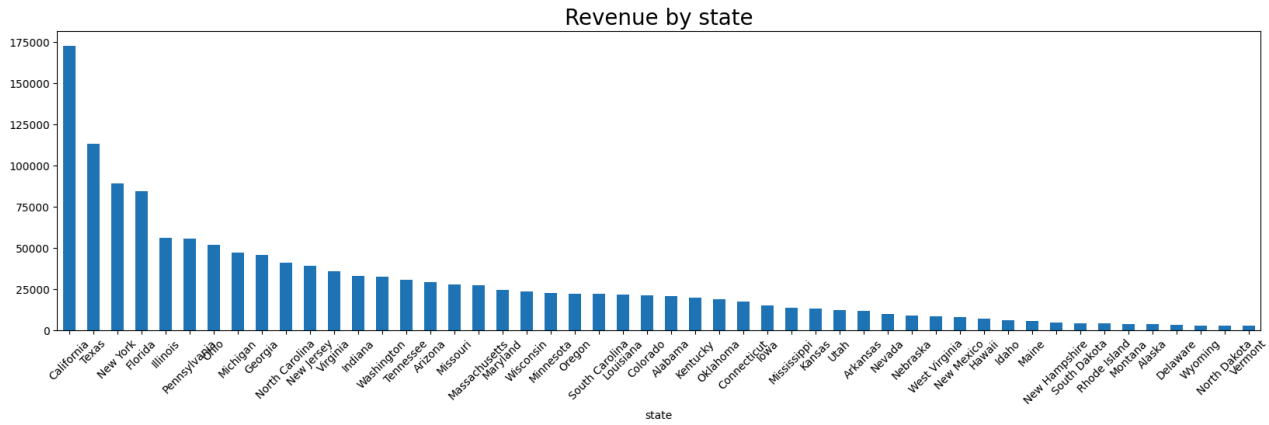
I would recommend using the Emails + Calls method as it requires less effort from the sales team and brings more revenue than the Calls or Email methods. Average revenue per method:

- Emails: 96.57$
- Calls: 49.12$
- Emails + Calls: 170.87$

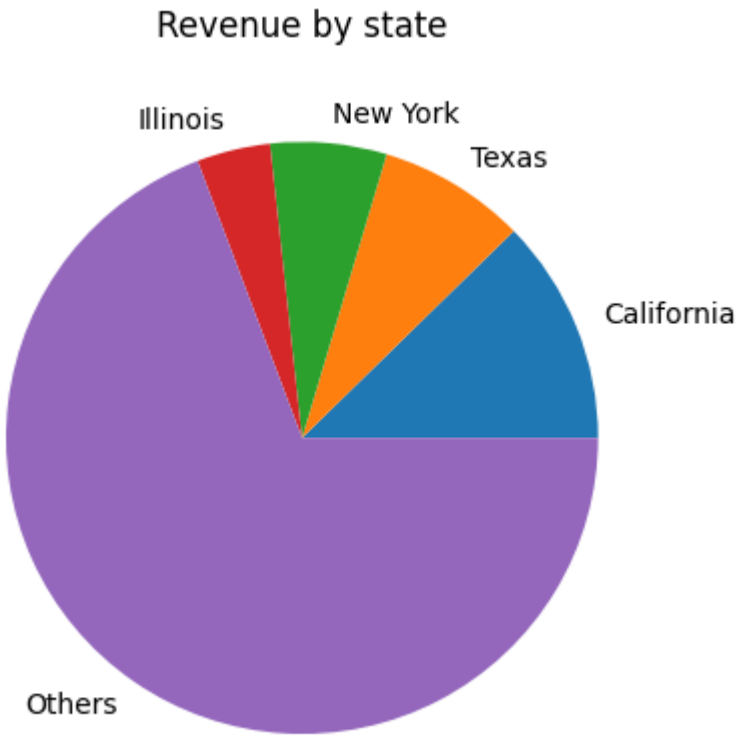
- First email may 'plant an idea' of the new product line within the customer's head. But if you don't remind yourself about it in time, it will be too late, as we saw in the Email method. Three weeks is a long delay after which customers are not willing to buy anymore.
- Calling a customer does not generate more profit in the end, despite higher effort

So, in summary, I would recommend emailing the customer and then calling them a week later. This approach is the best dynamic over time, with an average weekly increase of almost 50%.

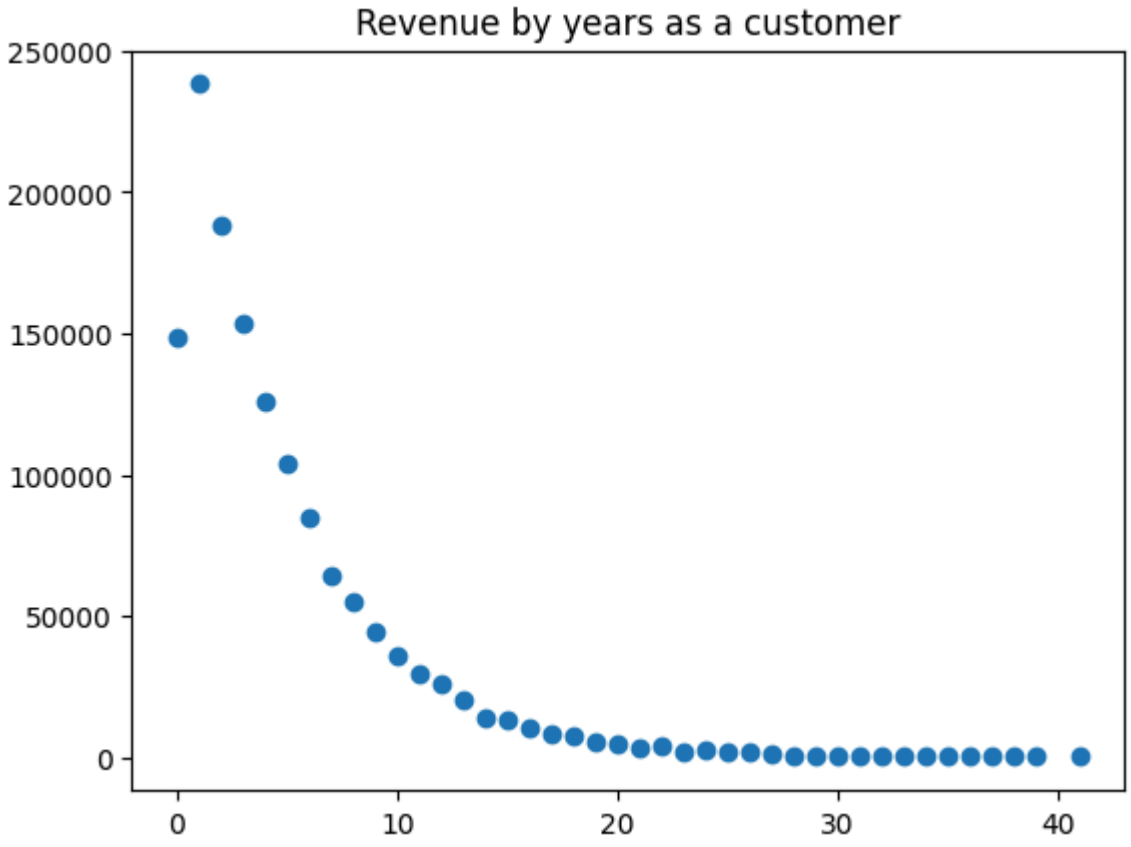
Revenue by state



As we can see, the top 4 states: California (12.31%), Texas (8.06%), New York (6.34%) and Illinois (4%) represent 30% of the total revenue for the whole product line. This could indicate higher demand for the products in these regions compared to other states.

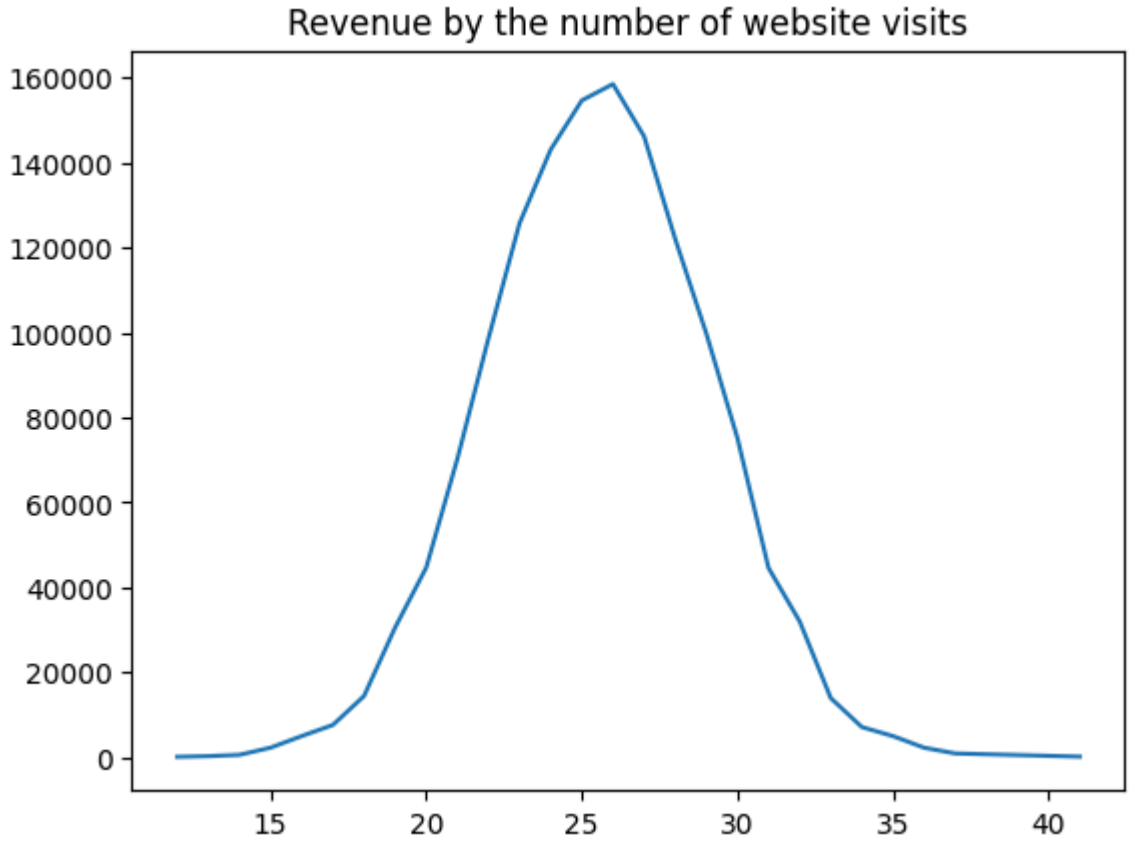


Does years as a customer influence revenue?



We can see a negative correlation between two variables. Long-standing clients are not interested in the new products offered by Pens and Printers.

Does the number of website visits correlate with the revenue?



The optimal number of visits is around 25. Clients with this number of website visits brought the majority of the revenue.

Business metrics

Since our goal is to see what sales methods were most effective, I would recommend we use total revenue by sales method and revenue over time for each sales method as our business metrics. If we see positive dynamics in the revenue for the Emails + Calls method (now it is 440,000$), it would indicate a very good sign to achieve our goal.

Recommendation

- Data for the last 6 weeks showed that despite email having the biggest share of the total revenue, it also showed the biggest drop in the week-to-week revenue metric. Email was effective at the product line launch, but for the long term, we should focus on the Emails + Calls method as it showed steady weekly surplus in the revenue and a higher average return in total.
- Geographical segmentation showed that the best performing states are California, Texas, New York and Illinois, which represent 30% of the total revenue for the whole product line. We should focus on these regions and make sure customers here are always satisfied.
- Old customers were not interested in the new products. We should target a relatively new customer base.
- The optimal number of website visits is around 25 for the last 6 months. We should prioritise customers with such a number of visits in our sales strategy.