Data Analyst Professional Practical Exam Submission

2023-04-28

Data Validation

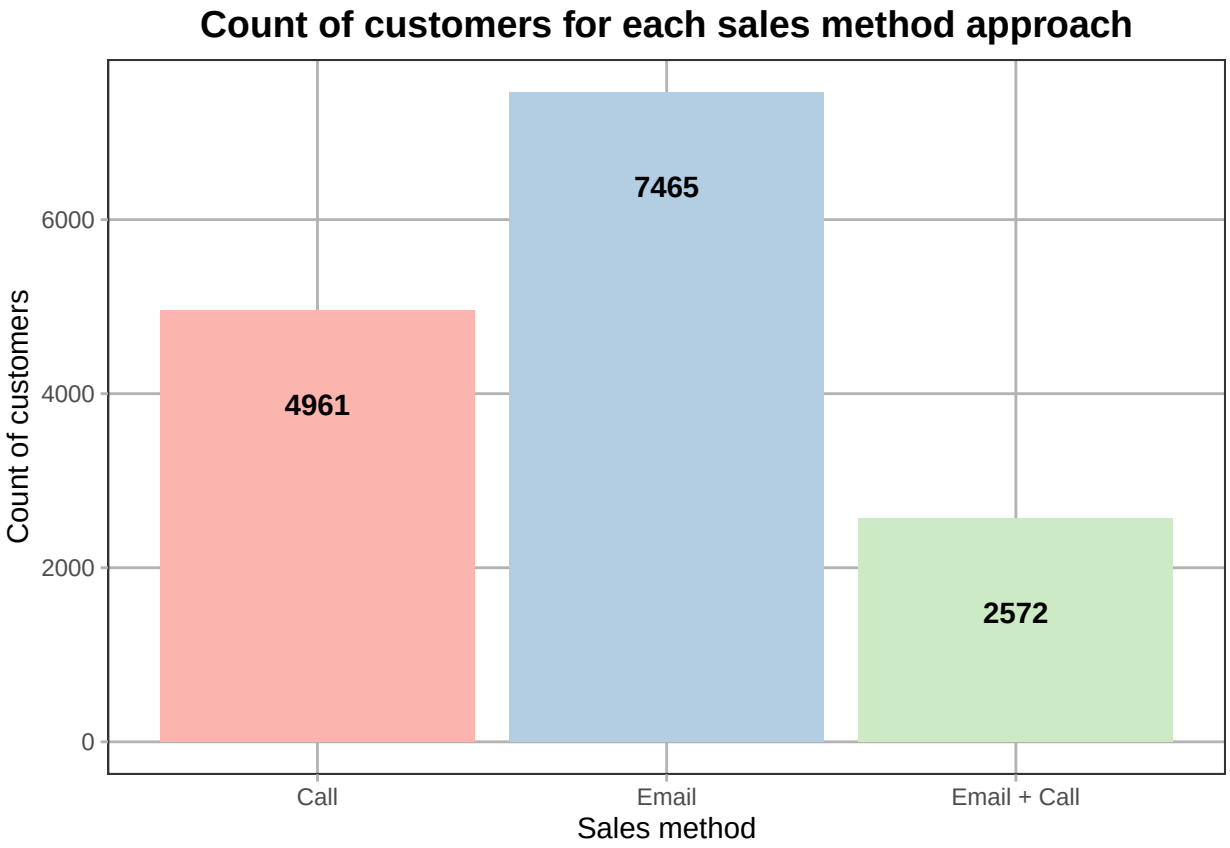
The product sales dataset contains **15000 observations and 8 variables** before performing cleaning and validation of the data. I have checked every variable to ensure they fit the provided criteria of the data:

- **Week**: numeric values from 1 to 6 without any missing values. Matches the description so no cleaning is needed;
- **Sales Method**: 3 categories of sales methods without missing values. There were some deviations in category names that were fixed to fit the provided criteria (capitalized first letter of the category name). Observations with incorrect name "em + call" were changed to the name "Email + Call". Similarly, the category name "email" was assimilated into the name "Email";
- **Customer ID**: variable's data type is character and all the observations are 36 characters long. There are no duplicates, so these are indeed unique identifiers. Matches the description so no cleaning is needed;
- **Number of new products sold**: correct data type (integer) without missing values, which matches the data description. No cleaning is needed.
- **Revenue**: numeric values with **1074 missing values**. The amount of missing values makes up for 7,16% of all the observations in the dataset. Since the percentage of the missing values is more than 5%, I have decided to impute the missing values. I have analyzed that revenue results between customers mostly differ based on the number of new products sold and the applied sales method. To achieve higher accuracy after imputing the missing values, I have separated the original dataset in three subsets, based on sales method category. I then filled each subset's missing values by group's (variable nb_sold) mean. Lastly, I binded fixed subsets back into one dataset. After these actions, there are no missing values in the revenue variable and the values match the data description;
- **Years as a customer**: since the company was founded in 1984 and it is now 2023, the logical conclusion to make here is that the maximum years as a customer can be 39 years. After checking for values higher than the possible maximum years as a customer, I have discovered 2 incorrect observations. Because the number of incorrect observations is not significant compared to the size of the whole dataset, I have taken these observations out from the dataset. After confirming that the variable has no missing values, I conclude that it now matches the description;
- **Number of site visits**: correct data type (integer) without missing values, which matches the data description. No cleaning is needed;
- **State**: correct data type (character) without missing values, which matches the data description. I have checked that there are no spelling mistakes in the state names and all values follow the same format. No cleaning is needed.

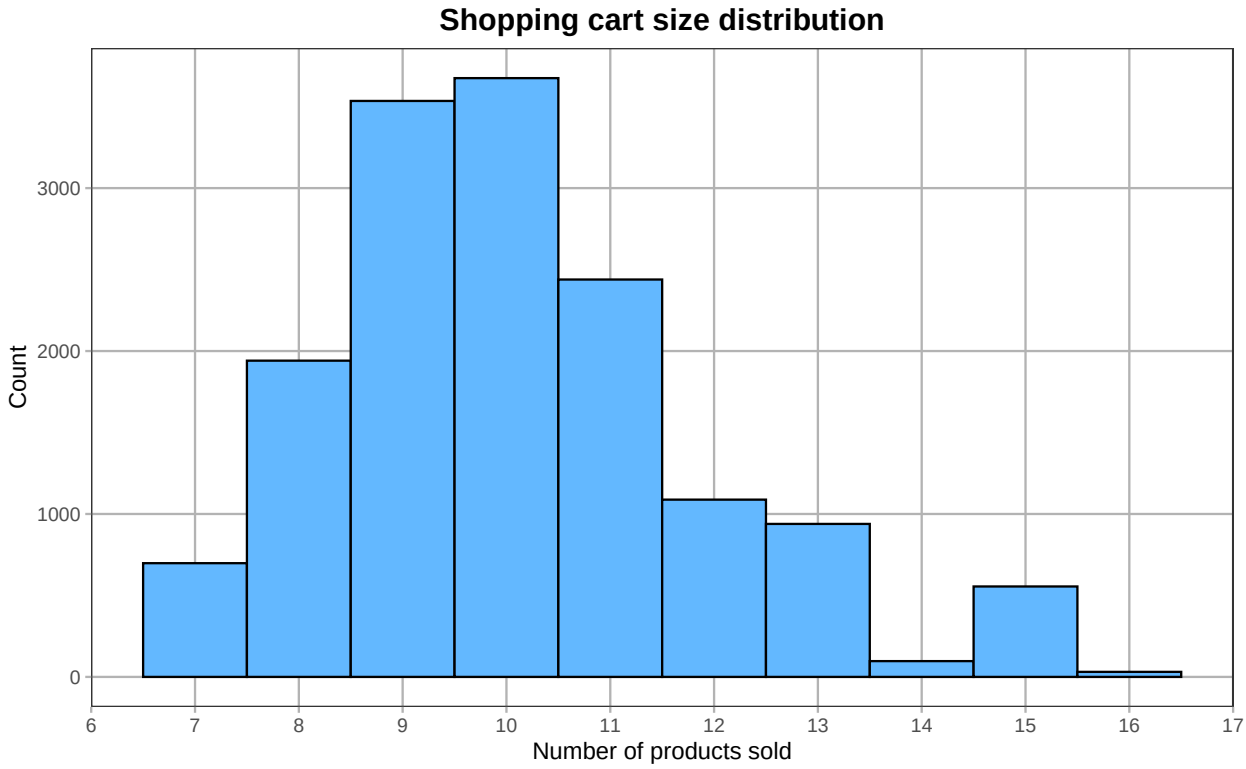
After validating and cleaning the data, the resulting dataset contains **14998 observations and 8 variables** without missing values and fully matching the data description.

How the new product sales differ depending on the sales method?

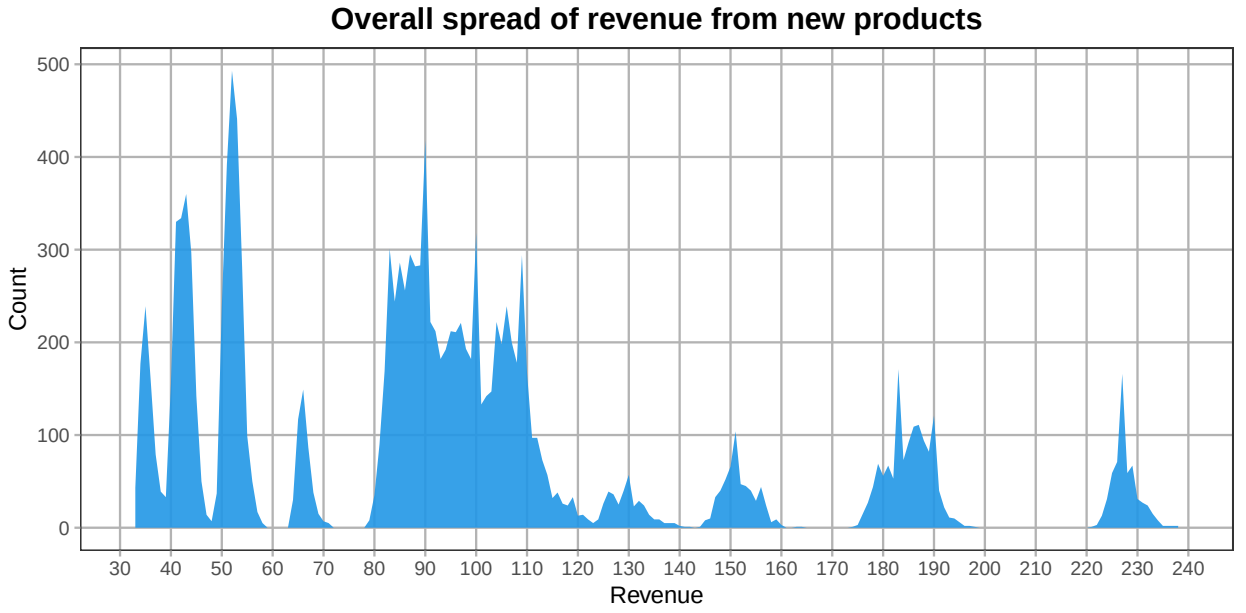
Looking at the data on new product sales during the last 6 weeks, we can see that most clients (almost **50%**) were contacted by email. As making calls required considerably more time for sales team members, they amounted to about **33%** of all new product sales. The third sales method accounted for the least amount of new product sales during the last 6 weeks. We aim to conclude if there is a need to increase the frequency of applying the Email + Call sales method.



The graph below shows the spread of customer basket sizes, which is the units per transaction metric. We can see that the most frequent basket size falls between 9 and 10 items per sale. The overall average basket size for the last 6 weeks is **10 units**.

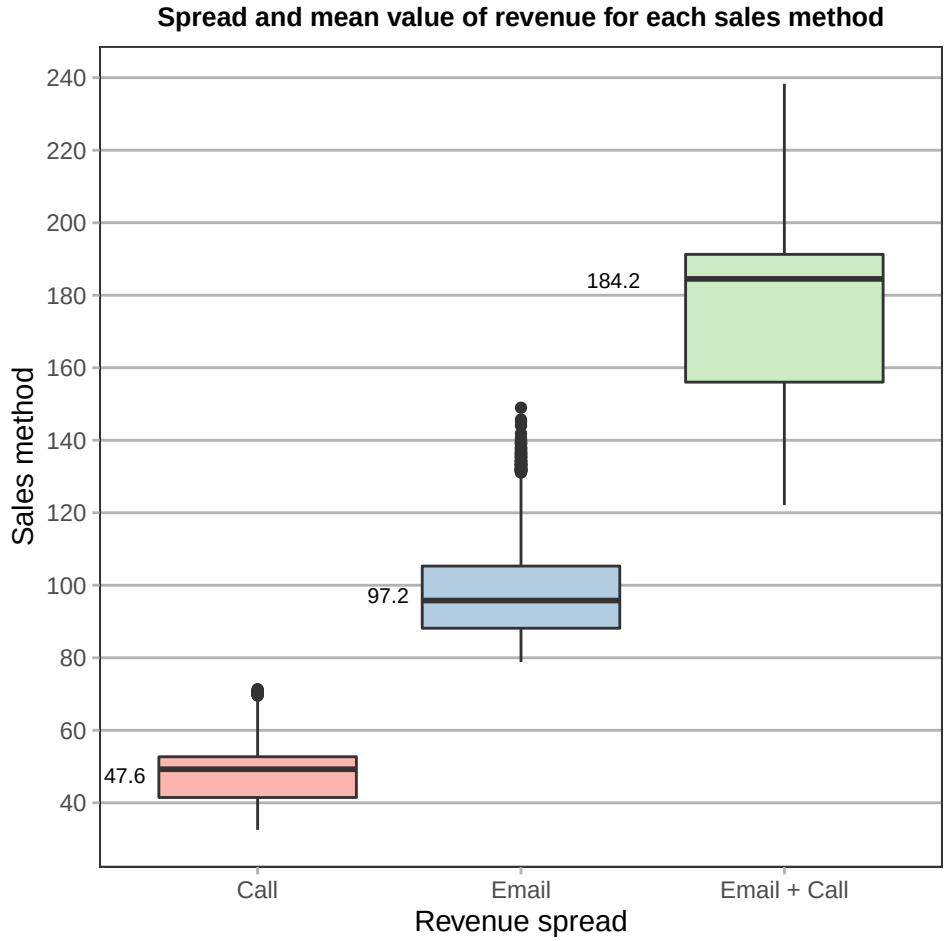


While looking at the spread of revenue for the new product line sales, we can notice a wide range going from **32.54 USD** to **238.32 USD**. After assessing visually, we can conclude that the sales revenue usually falls between approximately 80 USD and 110 USD.

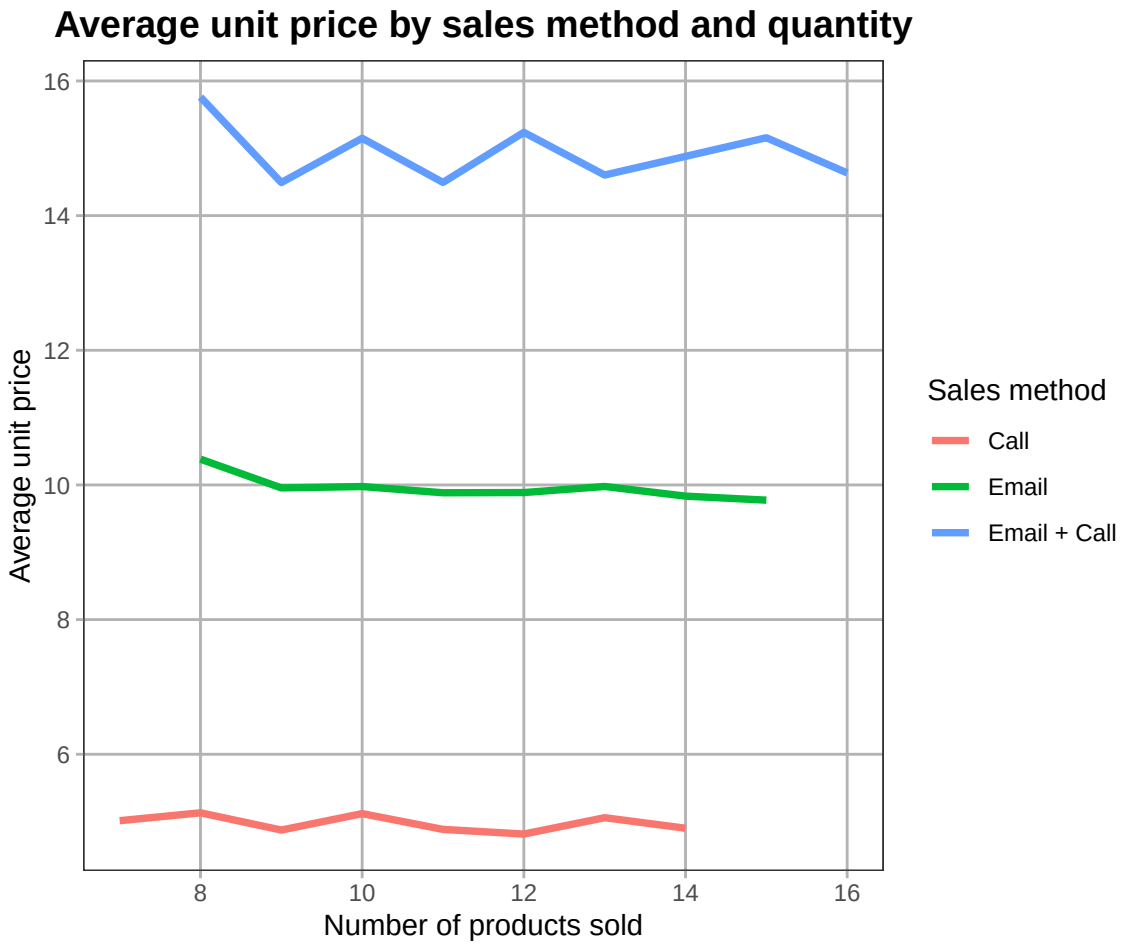


A more detailed look at the revenue spread reveals that it is highly related to the applied sales method. The graph shows that reaching clients and selling products via calls produces the smallest revenue range and an average sale of **47.6 USD**. While the easiest sales method results in an average of **97.2 USD** per sale,

contacting clients via email and calls generates twice the average. We can conclude that applying the third sales method ('Email and call') is exceptionally more effective than the other two.



We calculate and compare average unit prices to rule out the assumption that better revenue results for the third sales method could be related to higher quantities of products sold during the testing period. The graph shows us that by contacting via Email and Call, sales team members were able to convince customers to purchase higher valued products from the new line. The low average unit price for sales made by calls could be explained by the assumption that customers are more likely to purchase lower-priced items quickly and without getting additional visual information.

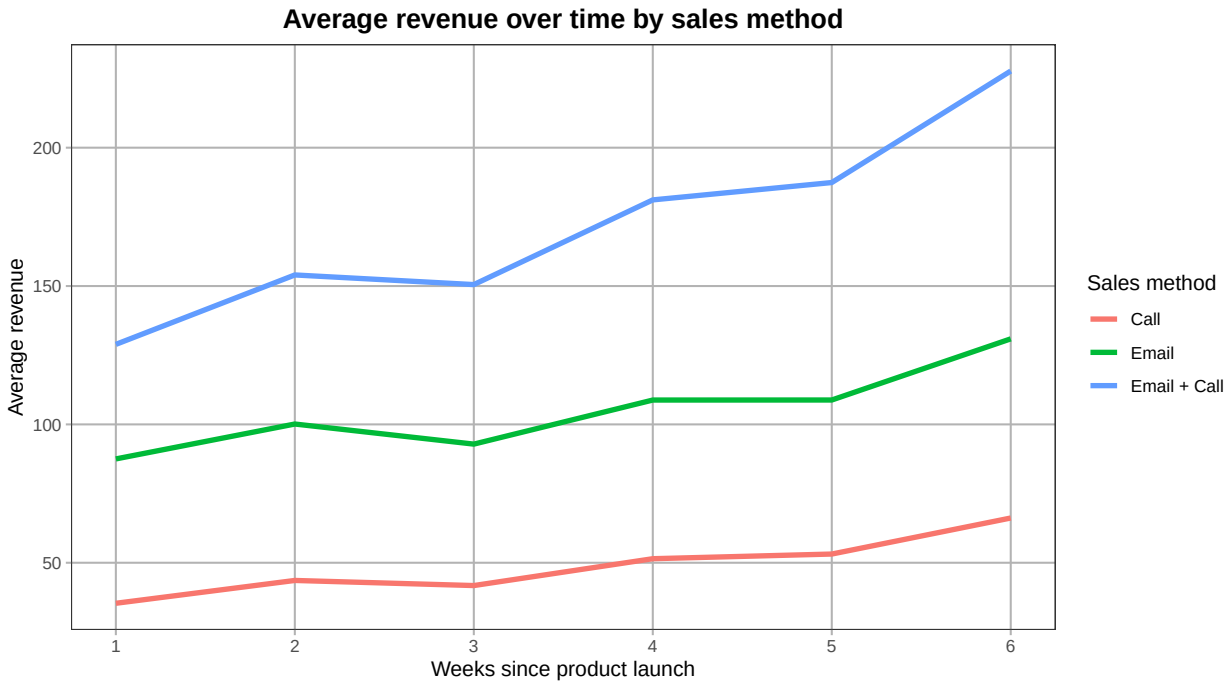


After analyzing the data, we can make a reasonable conclusion that the third sales method is significantly more effective and is worth the resources it requires. Not only was the overall revenue from the last 6 weeks higher with this sales method, but it also resulted in selling more valued items to the customers.

New product line revenue over time

The analyzed testing period stretches out 6 weeks, which enables us to evaluate how customers reacted to different sales methods. Since the percentages of customers contacted by different sales methods were very disproportionate, we continue to analyze average sales revenue to evaluate differences over time.

The plot below shows that the average revenue shows an increasing trend for all sales methods. Nevertheless, the 'Email + call' sales method demonstrates the steepest revenue growth after the new product launch. Again, we can confirm that the revenue averages were highest for sales made with the third method. We must also consider that different sales methods involve contacting the customer at different times of the testing period. All the sales methods display a similar upward trend, without any remarkable changes at particular weeks. Because of this, we conclude that the time the customer was contacted again does not significantly influence the sales method's effectiveness.

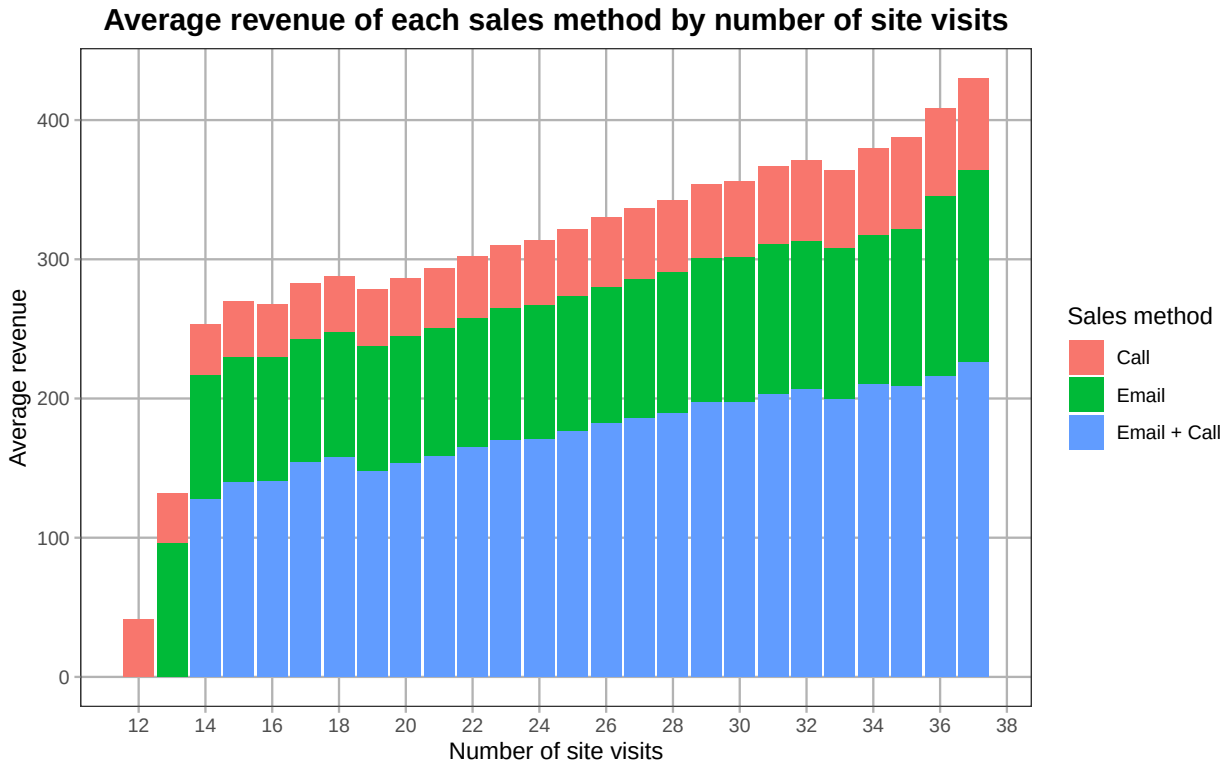


Overall, the upwards trend over the testing period portrays that customers take their time to make the purchasing decision. Therefore, customers should be given enough time and resources to think their decision through and learn as much information about the new product, as possible. It also seems that being proactive, and reminding customers about new products might result in them spending more on average.

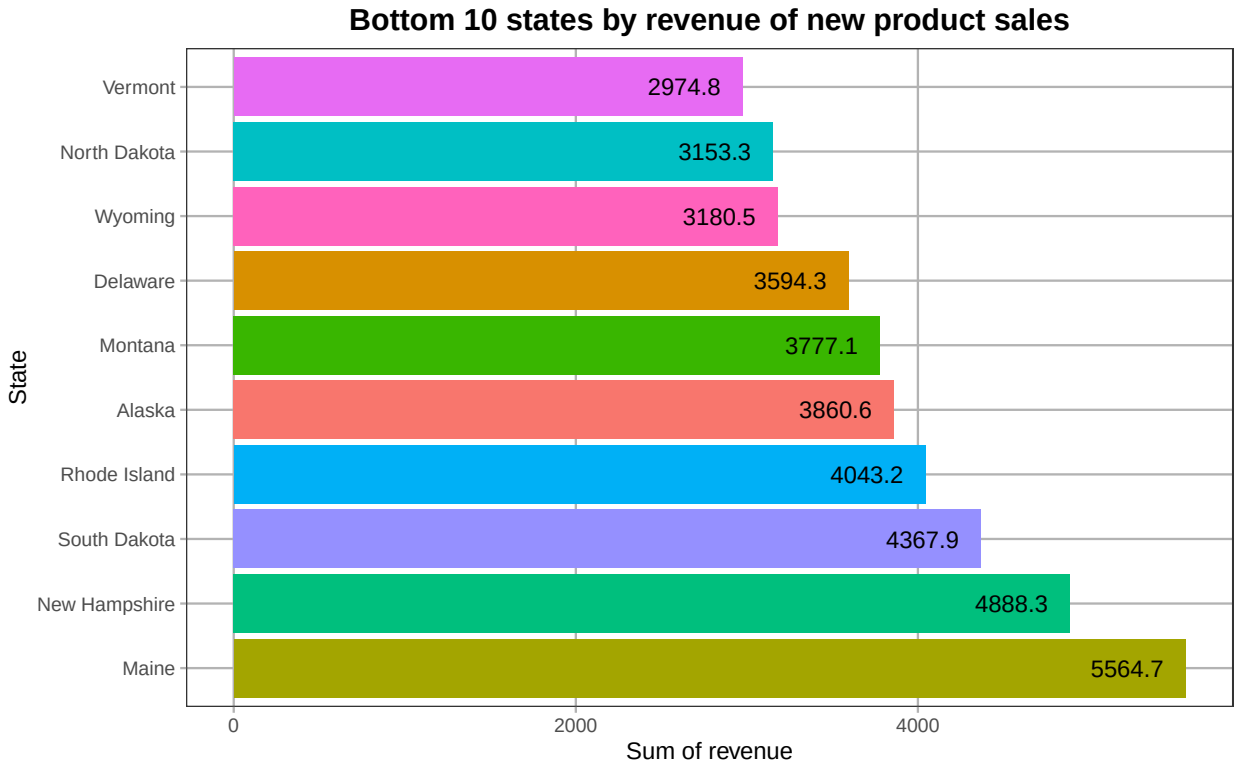
Other findings

Another helpful information our data contains is the number of site visits. Understanding how the generated average revenue relates to the web traffic is essential for building a strong online presence and securing sales. As we can see from the graph below, the average revenue increased with the number of visits. Based on this, we can make a few conclusions:

- The more information customers get about new products, the more they are convinced to buy them;
- Customers want to base higher valued purchases on as much information about the company and the products as possible;
- For some customers, the information on the company's site might be the main deciding point. Therefore, we must ensure it is user-friendly, informative, and attention-grabbing.



Since we are looking for ways to improve our sales approach, we should also consider how well our sales perform in different regions. When comparing the states by the overall revenue generated, we can exclude 10 states with the lowest revenue. Since each state resulted in almost 29k USD of revenue on average, there is room for improvement for the states where the sales underperformed.



Business Metrics

Since our goal is to choose the most effective sales approach, I would recommend using the **average unit price** as our main metric and the spread and average revenue as the reinforcing metrics. The first metric will continue to show us how much revenue on average is generated, considering the amount of products sold. The other metrics will provide insight into what is the overall result from the chosen sales method.

Based on the last 6 weeks data, the average unit price is **9.22 USD**, average revenue is **95.7 USD** and overall revenue falls between **32.5 USD** and **238.3 USD**. These metrics differ greatly between the tested sales methods:

- **Email**: average unit price - **10 USD**; average revenue - **97.2 USD**; interval from **78.8 USD** to **149 USD**.
- **Call**: average unit price - **5.01 USD**; average revenue - **47.6 USD**; interval from **32.5 USD** to **71.4 USD**.
- **Email and Call**: average unit price - **15.1 USD**; average revenue - **184 USD**; interval from **122 USD** to **238 USD**.

Recommendations

Moving forward, I would recommend focusing on the steps below:

- The data analysis and the calculated base business metrics strongly suggest we should be focusing on using the 'Email + Call' as our primary sales method;

- I wouldn't recommend continuing to carry out sales only by phone as it proved to be an ineffective sales method. Compared to the other two sales methods, the 'Call' method produces significantly lower results and is highly time-consuming for the sales team;
- Update and improve the company's site. Data shows that customers visited the company's site quite frequently, and more site visits might have influenced higher revenue on average. Because of this, we should update our site to be more user-friendly, functional, and eye-catching;
- Focus more attention on the bottom performing states:
  - Analyze what may be the reasons for lower revenue in the bottom 10 states;
  - Update the sales approach according to found reasons;
  - Improve the marketing approach in these states to increase brand image and attract new customers
- Improve data for better quality analysis:
  - Fix the missing values issue - ensure there are no missing revenue (or other) values;
  - Include data about how much time was spent on each customer - could be used for creating and following another useful business metric.