Question 1: PhysicalSound

PhysicalSound is a direct-to-consumer e-commerce website based in Great Britain that sells physical music media such as vinyl, cassettes, & CDs. Here's an analysis to aid with business expansion:

Business Understanding:

The core objective of the business problem is to increase customer spending on the website by deciding who it wants to attract to the website, along with potentially better estimating future revenue.

The objective can be measured by analyzing the provided dataset, which includes the sales data from the month of July 2024. This can help understand what factors cause a customer to spend more on the website.

Through descriptive statistical analysis and machine learning models, we will be achieving the analytical goal that underlines our business objective.

The opportunity statement is as follows:

Achieved Result	Almost £80,000 in sales in the month of July, 2024
Disturbing Event	Company expansion
Desired Result	Better estimate future revenue by deciding who it wants to attract to the website
Key Question	How to increase customer spending on the website?
Criteria	Future sales & profitability

Data:

The data contains 2000 observations in 6 columns. To summarize customer behavior, here is the average customer that shopped at PhysicalSound in July:

Age	Past Spend	Time Spent	Spend
34 years	£13.02	81.21 seconds	£40.15

The total sales for the month of July amounts to **£79,209** (26 incomplete entries). As observed in the data, customers over the age of 34 spent, on an average, around £10 more than those 34 and under. In addition, customers who spent more time on the website than the average outspent others by around £9.

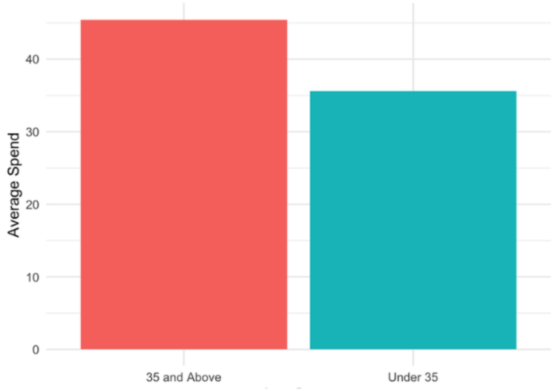
Although the past spends values show a similar trend, a deeper look at it is required. Analyzing the mean values just tells one half of the story, the other is made evident when looking at the spending behavior as new or old customers. A past spend of 0 indicates a new customer. Around 83% of PhysicalSound's customers are previous customers, indicating excellent customer retention. These customers are also the bigger spenders, outspending newer customers by 25% (ages 34 and under) & 17.5% (ages above 34).

When it comes to the final 2 criteria, there are no significant impacts that can be seen through statistics, the data across all 4 types of advertisement channels remains similar and consistent. For the voucher data, it is important to note that less than 30% customers use vouchers, but it does not show a significant difference in customer behavior.

This sums up the observable statistics that can be seen in the dataset without any manipulation.

£79,209

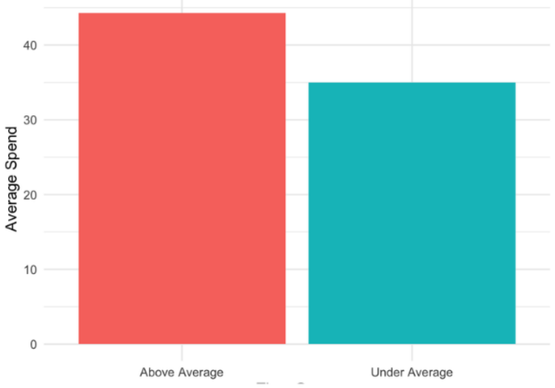
sales in July



average spend per age category

83%

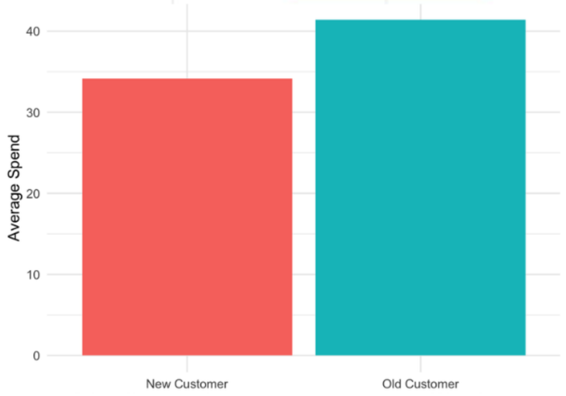
orders placed by old customers



average spend per time category

£10

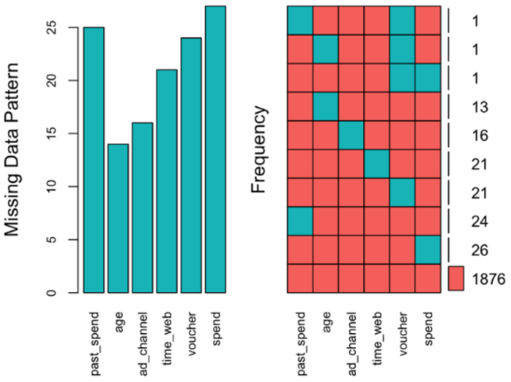
difference in average order value between an old and new customer



average spend per customer category

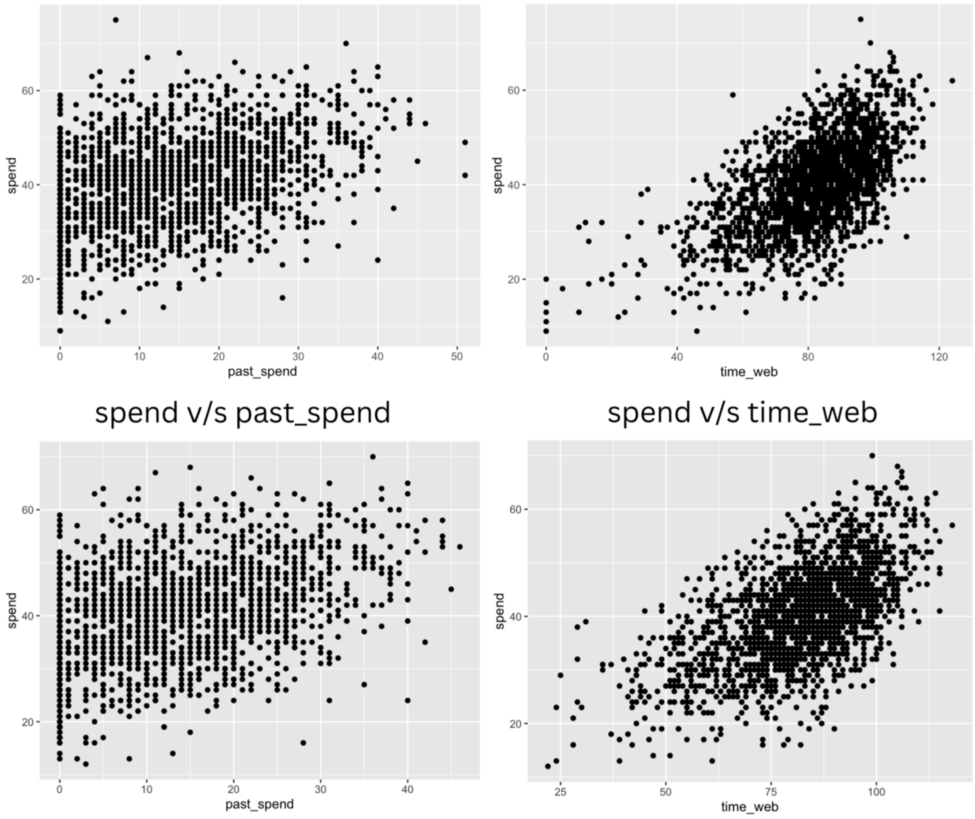
124

missing data entries



no. of missing data per variable

Digging deeper, we can see the major issue with the dataset. Out of the 2000 orders, 124 orders, or 6.2%, had incomplete data. Upon inspection for all the 6 variables, we fail to establish any reason for the missing values and establish that the values are Missing Completely at Random (MCAR). To tackle this problem, we use a mix of deletion and KNN to handle the missing data. Of the missing values, over a fifth are the target variable of spend. We delete all observations where the spend is missing, as it is unreliable to modify a dependent target variable. Deleting just about 1% of our total data will not be harmful.

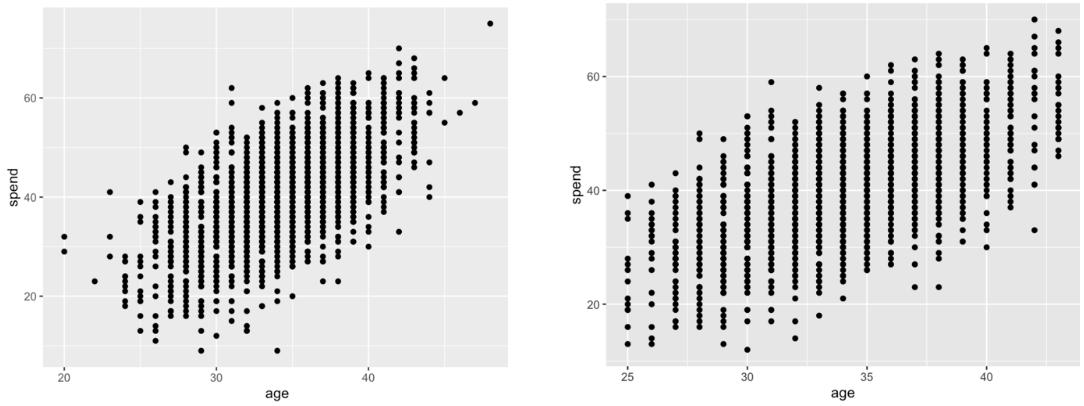


spend v/s past_spend

spend v/s time_web

These graphs (top) introduce a new problem for us, the observable outliers. Outliers are data points which are unique but very different to the general data. They sometimes confuse the models that we train and can influence the result of the model unfavorably towards itself, hence we eliminate outliers from our model’s input data.

The removal (bottom) of outliers also aids in linear regression, making the data more linear, and the model more robust.



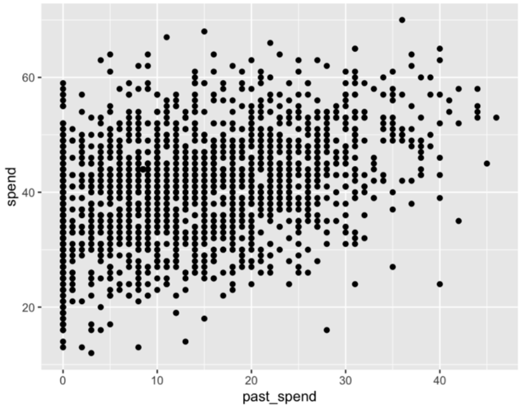
spend v/s age

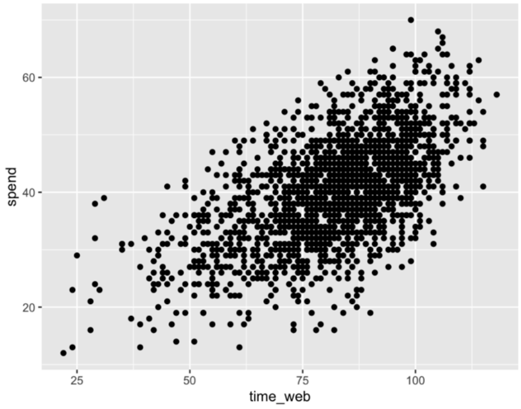
Some of the values omitted in our analysis include:

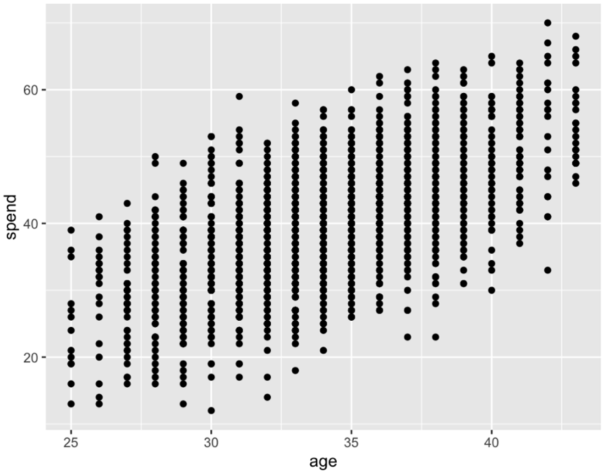
- The oldest customer, who also spent £75, the highest order total in July.
- The youngest customers of age 20.
- Two spenders with previous order totals of £51.
- Customers with time_web value of 0.

This was done through quantitative metrics (like distance through standard deviations) and logical reasoning.

The next step is to fill in the remaining values, instead of deleting all values and losing out on valuable data, a different approach is taken. Multiple Imputation is not ideal as the data is not MAR (missing at random due to observable data). Simple imputation introduces imbalance in the data by adding mean/mode to all missing variables, so KNN is opted for. KNN is a method which compares a missing value row, with other similar rows, and fills the missing value based on those neighbors. This ensures data consistency and prediction of similarity which will eventually be done through regression for the target variable as well.







As seen from the graphs, there is linearity in the data, with the spend increasing as age increases, this is consistent with our initial observation which showed that ages above 35 spend considerably more than ages below 35. Same is observed for time_web, with spend increasing as the customer's time on the website increases. The graph for past_spend does not exactly fit the assumptions that we check for in linearity. These assumptions include:

- Visible linearity in the data.
- Independent variation across the graph.
- Variance remains consistent.

Despite this, it still softly follows these rules, hence we have selected a linear regression model for the prediction of customers spend.

Linear regression is also good choice for our case as we are predicting a continuous variable, and it is great at quantifying the relationship between variables. This enables us in understanding what factors drive customers to spend more on the website.

Before running our model, we will compare the statistics of our original data with our final data.

Mean	**past_spend**	**spend**	**age**	**time_web**
Original	13.01	40.14	34.14	81.21
Modified	12.92	40.53	34.24	82.17

The data stays true to the original and fit to be used for our predictions.

To train a linear regression model, we split the data into two parts: training data and test data. For our purpose, the split is 85% to 15%, using nearly 1650 observations to predict the other 300.

Factors Influencing Customer Spending on the Website:

Based on the regression, several key factors influence customer spending on the website:

1. **Past Spending** (past_spend):
   - The coefficient for past_spend is **0.359** (p-value < 0.001), indicating that for every additional unit of past spending, current spending increases by approximately 35.9%. This suggests that past spending behavior strongly predicts future spending.
2. **Age** (age):
   - The coefficient for age is **1.475** (p-value < 0.001). Older customers tend to spend more, with a significant increase of $1.48 in spending for each additional year of age.
3. **Time Spent on the Website** (time_web):
   - The coefficient for time_web is **0.3511** (p-value < 0.001), indicating that customers who spend more time browsing the website tend to spend more money. This emphasizes the importance of engaging customers on the platform.
4. **Advertising Channel** (ad_channel):
   - This factor has a statistically insignificant coefficient. It suggests that ad channel types 2,3,4 have minimal impact compared to 1.
5. **Voucher Use** (voucher):
   - The coefficient for voucher is not statistically significant. Voucher use does not meaningfully predict spending.

Predictive Model for Customer Spending:

1. **Model Performance**:
   - The model explains **86.7%** of the variance in customer spending. This indicates the model is effective at capturing the factors influencing spending.
   - The root mean square error (RMSE) on the test data is **3.22,** meaning the model's predictions are, on average, within $3.22 of the actual spending. This is a reasonable margin of error, considering the context.

Recommendations for the Company:

1. **High-Impact Factors**:
   - **Engagement**: Encourage customers to spend more time on the platform through features like personalized recommendations as time_web is a strong predictor.
   - **Retention**: Focus on products for customers who have spent more in the past, as they are more likely to spend more in the future.
   - **Older customers** are associated with higher spending. Marketing efforts and website features tailored to this demographic may yield better results.

2. **Reevaluate Promotional Strategies**:
    - Voucher use and ad channel type have limited predictive power for spending. Consider better strategies to influence customers to increase spending through these variables.

Predictions for Additional Customers

Using the trained model, predictions for the 20 new customers from new_customer24.csv have been generated. These predictions provide an estimate of their spending based on the same factors used in the model. If the RMSE for the new customer predictions is similar to or lower than that of the test data, the model can be considered effective for expansion evaluation.

Order no.	**Prediction**
1	16.58222
2	45.28235
3	46.05561
4	46.71376
5	34.74997
6	44.57836
7	47.32167
8	54.41219
9	51.13611
10	28.51691
11	49.51831
12	24.44095
13	37.44566
14	40.69802
15	37.28730
16	49.48061
17	46.02768
18	48.42845
19	32.35340
20	55.64907