Golladay - Final Project

Cory Golladay

2024-11-06

Final Project: Step 1

Introduction

In today's competitive market, optimizing marketing spend is crucial for businesses aiming to maximize returns on investment while minimizing wasted resources. In the automotive retail industry, effective marketing allocation can make a substantial impact on lead generation, lead quality, and, ultimately, sales conversions.

The challenge lies in determining the most efficient ways to allocate a budget across various marketing channels—such as Google Paid Search, third-party platforms like Edmunds and Carfax, and other digital advertising avenues. Given the complexities of multi-channel spending and varying lead quality, identifying a data-driven approach to optimize marketing spend is both an essential and challenging task.

This research project explores Marketing Spend Optimization with the goal of understanding how different spending allocations impact lead quality and sales conversions. By analyzing marketing data, we can gain insights into which channels produce the highest quality leads at the lowest acquisition costs, as well as examine how spending levels correlate with lead quality and conversion outcomes. With data science methods, we can uncover patterns and trends that would be difficult to observe otherwise, helping to inform strategic decisions on budget allocation.

Understanding and optimizing marketing spend is particularly relevant because inefficient allocations not only drain resources but can also lead to missed opportunities and lower profitability. Through this project, I aim to provide a structured analysis of how marketing budgets should be optimized, leveraging simulated data that mirrors real-world dynamics without compromising data privacy.

This is a data science problem because it involves analyzing large amounts of complex data to uncover patterns, optimize resource allocation, and make informed decisions. Marketing spend optimization requires the integration and analysis of data from various sources, such as spend by channel, lead quality metrics, and conversion rates, to draw meaningful insights. Using Data Science, we can tackle challenges like:

- Identifying Patterns and Trends
- Predictive Modeling
- Optimization
- Measuring and Validating Impact

Research Questions

1. What percentage of the total budget should be allocated across different vendors for optimal results?
2. How does the performance of each vendor vary by month or season?
3. How does the performance of each vendor vary by brand, model, and new or used car type?
4. What is the relationship between marketing spend and lead quality?
5. Which spend categories (e.g., paid search, third-party platforms) yield the highest conversion rates?

6. Is there an optimal budget threshold for diminishing returns?
7. Can we predict lead quality based on vendor and spend amount?
8. How does marketing spend impact customer acquisition costs?
9. What trends in spend allocation yield the most significant return on investment?

Approach

To address the problem of marketing spend optimization, this project will involve creating simulated datasets that replicate the types of data typically available in a marketing analytics setting. Since privacy considerations prevent the use of actual workplace data, I will use Python to generate synthetic datasets that mimic realistic patterns of marketing spend, lead generation, conversions, and geographic variations. These datasets will be carefully designed to resemble the nuances of actual marketing data, including seasonal fluctuations, lead quality variations, and conversion rates, enabling analysis that mirrors real-world conditions.

The first step will involve defining key variables and relationships in the dataset, such as:

- Marketing Spend by Channel: Simulated monthly budget allocations across several marketing channels, including Google Paid Search, social media ads, third-party platforms (e.g., Edmunds, Carfax), and other digital avenues.
- Leads Generated: Data on the volume and quality of leads generated from each channel, with variations by month, location, and other factors to reflect realistic outcomes.
- Conversions: Conversion metrics showing which leads resulted in successful sales.
- Geographic/Location Data: Variations in spend, leads, and conversions across different geographic areas, reflecting location-specific performance trends.

With these variables established, I will create a Python script that generates datasets reflecting hypothetical but plausible marketing conditions, ensuring realistic distributions, correlations, and patterns. This synthetic data will capture factors like diminishing returns in high-spend channels, seasonal influences on lead generation, and variances in conversion rates across locations.

Once the datasets are created, I will move into the data analysis phase in R. Using R's analytical and visualization capabilities, I will:

1. Perform Exploratory Data Analysis (EDA): I'll conduct an initial examination of the datasets to understand basic distributions, correlations, and key metrics. This will involve summary statistics, such as mean, median, and variance, as well as visualizations to highlight spending trends, lead quality, and conversion rates by channel and location.

2. Data Transformation and Cleaning: To ensure the data is ready for analysis, I'll apply transformations where needed—standardizing fields, handling missing values, and potentially creating new derived metrics like ROI per channel or conversion rate per lead type.

3. Data Visualization: Using R's visualization packages, I'll create a series of plots to illustrate findings. This may include:

- Time-series plots to show spending trends over time
- Bar charts comparing spend and conversion rates by marketing channel
- Scatter plots examining relationships between spend and lead quality or conversions

4. Summarize and Synthesize Findings: I'll interpret and consolidate insights from the visualizations and summary statistics, identifying which channels and locations offer the best ROI, where spending inefficiencies might exist, and how seasonality or other factors affect conversions.

How the approach addresses the problem:

This approach allows me to analyze the simulated marketing spend data comprehensively, mirroring a real-world scenario. By the end of the analysis, I'll be able to make informed recommendations about the optimal allocation of marketing spend and suggest potential areas for budget reallocation or targeted investment. This lays the groundwork for potential predictive modeling or further optimization in future analyses.

Data

My datasets will be simulated to focus on an issue we currently face in my Marketing department at my job. 1. Marketing Spend Data – Simulated data by vendor and spend type. 2. Lead Data – Simulated data on leads by source and category and if the lead converted. 3. Location Data - Simulated to include various brands over the US in different geographic locations.

Required Packages

- tidyverse (for data manipulation and visualization)
- ggplot2 (for plotting)
- dplyr (for transformations)
- lubridate (for date handling)
- readxl (to read data files created)

Plots and Table Needs

- Time-series plots for spend trends
- Bar charts comparing spend versus conversions by vendor
- Scatter plots showing spend versus lead quality
- Tables summarizing ROI by vendor and spend category

Questions for Future Steps

- How do I simulate realistic spend and conversion data in Python?
- What statistical methods best fit marketing optimization analysis in R?
- How can I build a recommendation for future spend based on my findings?

Final Project: Step 2

Data Importing

```
# import libraries
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(tidyr)
library(ggplot2)

# import data
marketing_spend <- read_csv("C:/Users/golla/OneDrive/Documents/Bellevue University-SchoolPC/DSC 520 Sta

lead_data <- read_csv("C:/Users/golla/OneDrive/Documents/Bellevue University-SchoolPC/DSC 520 Statistics

location_data <- read_csv("C:/Users/golla/OneDrive/Documents/Bellevue University-SchoolPC/DSC 520 Statis
```

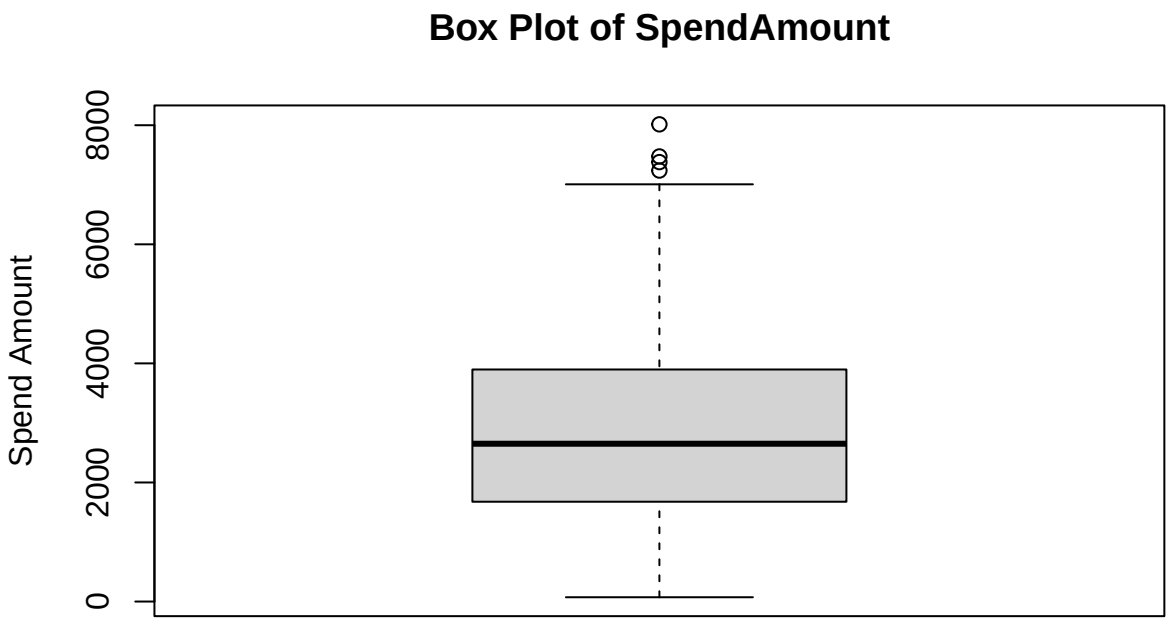
Data Cleaning

```
# marketing spend cleaning

# summarize the data
summary(marketing_spend)
```

```
##  FranchiseName         Vendor             Month            SpendAmount
##  Length:924         Length:924         Length:924         Min.   :  72.48
##  Class :character   Class :character   Class :character   1st Qu.:1678.17
##  Mode  :character   Mode  :character   Mode  :character   Median :2650.85
##                                                           Mean   :2870.24
##                                                           3rd Qu.:3896.34
##                                                           Max.   :8014.73
```

```
# Box plot for SpendAmount
boxplot(marketing_spend$SpendAmount, main = "Box Plot of SpendAmount", ylab = "Spend Amount")
```




```r
# identify the outliers
Q1 <- quantile(marketing_spend$SpendAmount, 0.25, na.rm = TRUE)
Q3 <- quantile(marketing_spend$SpendAmount, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1

# define outlier thresholds
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# find outliers in SpendAmount
outliers_upper <- marketing_spend$SpendAmount[marketing_spend$SpendAmount > upper_bound]
outliers_lower <- marketing_spend$SpendAmount[marketing_spend$SpendAmount < lower_bound]

outliers_upper
```

```
## [1] 8014.73 7239.03 7380.87 7475.42
```

```r
outliers_lower
```

```
## numeric(0)
```

```r
# cap the outliers found the 4 extreme values
marketing_spend$SpendAmount[marketing_spend$SpendAmount > upper_bound] <- upper_bound
```

```r
# ensure Month column is treated as a Date
marketing_spend <- marketing_spend %>%
  mutate(Month = as.Date(paste0(Month, "-01")))

# calculate the number of days in each month and the average daily spend
marketing_spend <- marketing_spend %>%
  mutate(
    DaysInMonth = days_in_month(Month),
    AvgDailySpend = round(SpendAmount / DaysInMonth,2)
  )

# expand the data to have one row per day in each month
daily_spend <- marketing_spend %>%
  rowwise() %>%
  mutate(Date = list(seq(Month, by = "day", length.out = DaysInMonth))) %>%
  unnest(Date) %>%
  select(FranchiseName, Vendor, Date, AvgDailySpend)

# remove any duplicates
marketing_spend <- marketing_spend %>% distinct()

# view the expanded data
head(daily_spend)
```

```
## # A tibble: 6 x 4
##   FranchiseName  Vendor     Date       AvgDailySpend
##   <chr>          <chr>      <date>             <dbl>
## 1 BMW of Florida AutoTrader 2022-01-01          114.
## 2 BMW of Florida AutoTrader 2022-01-02          114.
## 3 BMW of Florida AutoTrader 2022-01-03          114.
## 4 BMW of Florida AutoTrader 2022-01-04          114.
## 5 BMW of Florida AutoTrader 2022-01-05          114.
## 6 BMW of Florida AutoTrader 2022-01-06          114.
```

```r
# lead data cleanup

# summarize the data
summary(lead_data)
```

```
##  FranchiseName       InsertDate        MarketingChannelName    EventID
##  Length:66000       Length:66000       Length:66000         Min.   :10001
##  Class :character   Class :character   Class :character     1st Qu.:32472
##  Mode  :character   Mode  :character   Mode  :character     Median :54914
##                                                            Mean   :54960
##                                                            3rd Qu.:77563
##                                                            Max.   :99998
##
##    CustomerID      EventType            Source          IsNewCustomer
##  Min.   :100023   Length:66000       Length:66000       Mode :logical
##  1st Qu.:323733   Class :character   Class :character   FALSE:32711
##  Median :550061   Mode  :character   Mode  :character   TRUE :33289
##  Mean   :549875
##  3rd Qu.:776846
```

```
##  Max.   :999997
##
##      Make               Model             Category           SoldDate
##  Length:66000       Length:66000       Length:66000       Length:66000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   GrossProfit
##  Min.   : -362.2
##  1st Qu.: 3993.2
##  Median : 5018.8
##  Mean   : 5022.0
##  3rd Qu.: 6047.6
##  Max.   :10649.8
##  NA's   :58004
```

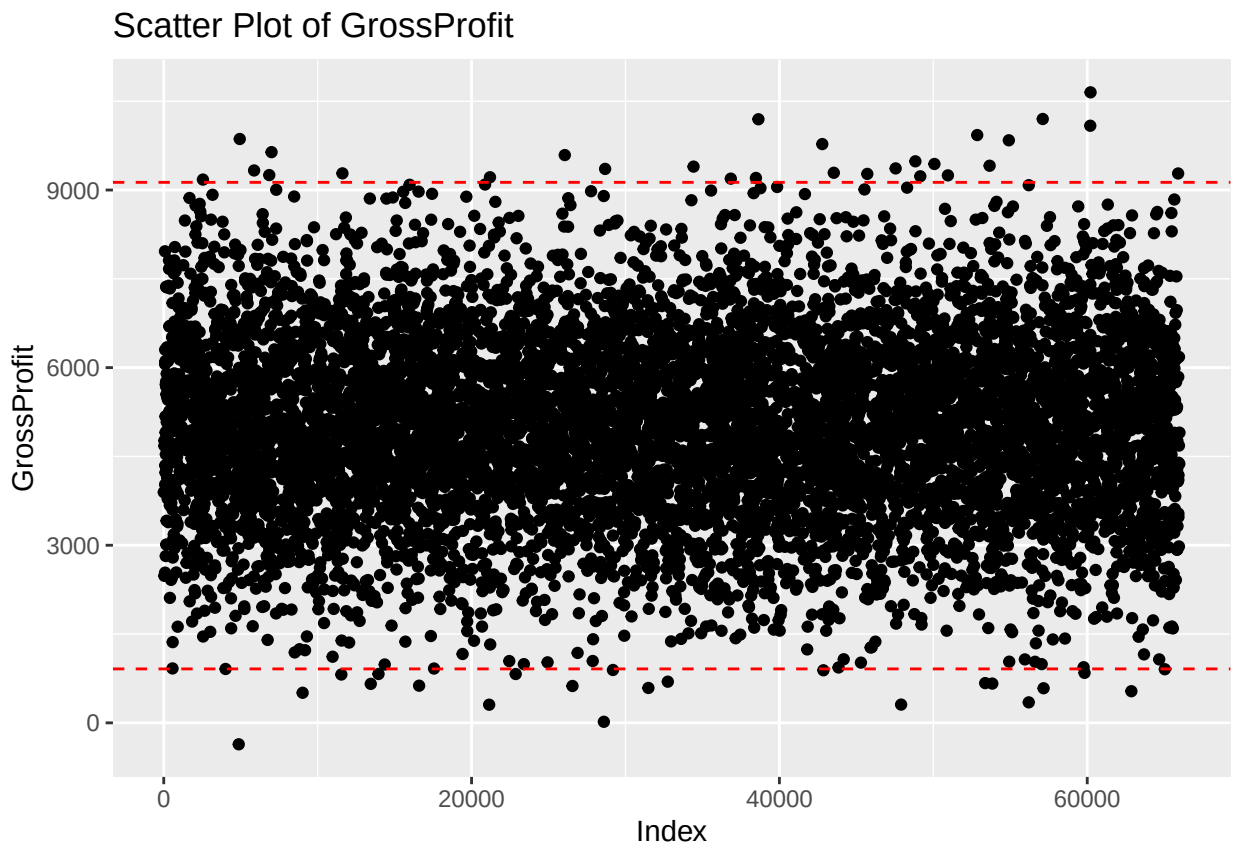
```r
# create an index column in lead data
lead_data$Index <- seq_len(nrow(lead_data))

# Calculate the IQR-based bounds for GrossProfit
Q1 <- quantile(lead_data$GrossProfit, 0.25, na.rm = TRUE)
Q3 <- quantile(lead_data$GrossProfit, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# scatter plot for gross profit
ggplot(lead_data, aes(x = Index, y = GrossProfit)) +
  geom_point() +
  labs(title = "Scatter Plot of GrossProfit", x = "Index", y = "GrossProfit") +
  geom_hline(yintercept = c(lower_bound, upper_bound), color = "red", linetype = "dashed")
```

```
## Warning: Removed 58004 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
# view the lead data
head(lead_data)
```

```
## # A tibble: 6 x 14
##   FranchiseName    InsertDate  MarketingChannelName EventID CustomerID EventType
##   <chr>            <chr>       <chr>                 <dbl>      <dbl> <chr>
## 1 Audi of Florida  1/1/2022 0~ AutoTrader            81140     284012 Used
## 2 Honda of Florida 1/1/2022 0~ True Car              42993     272338 Used
## 3 BMW of Florida   1/1/2022 0~ True Car              70955     100227 Used
## 4 Audi of Florida  1/1/2022 0~ Edmunds               68898     464406 New
## 5 Audi of Florida  1/1/2022 1~ Edmunds               22234     969744 New
## 6 Honda of Florida 1/1/2022 1~ True Car              89662     428193 Used
## # i 8 more variables: Source <chr>, IsNewCustomer <lgl>, Make <chr>,
## #   Model <chr>, Category <chr>, SoldDate <chr>, GrossProfit <dbl>, Index <int>
```

```
# update insertdate to correct date format
lead_data$InsertDate <- as_date(mdy_hms(lead_data$InsertDate))

# update solddate to correct date format
lead_data$SoldDate <- mdy(lead_data$SoldDate)

# ensure gross profit has 2 decimal places
lead_data$GrossProfit <- round(lead_data$GrossProfit, 2)

# location data review for cleanup
head(location_data)
```

```
## # A tibble: 4 x 4
##   FranchiseName    State   Region `Make OEM`
##   <chr>            <chr>   <chr>  <chr>
## 1 BMW of Florida   Florida South  BMW
## 2 BMW of Oregon    Oregon  West   BMW
## 3 Audi of Florida  Florida South  Audi
## 4 Honda of Florida Florida South  Honda
```

In this data cleansing process, I began by examining and summarizing both the marketing_spend and lead_data datasets to gain insight into their structures. For marketing_spend, I created a box plot to identify and cap extreme outliers in the SpendAmount column, setting values above the upper threshold to the threshold itself. I converted the Month column to a Date format and calculated the number of days in each month, which allowed me to compute an average daily spend. The data was then expanded to include individual daily entries for each month, removing any duplicate rows. In lead_data, I addressed outliers in GrossProfit by calculating the IQR-based bounds and visualizing them in a scatter plot with an Index column created to uniquely identify rows. Both InsertDate and SoldDate were converted to consistent date formats, and GrossProfit was rounded to two decimal places for clarity. Finally, I reviewed the location_data to ensure consistency across datasets. This comprehensive cleaning process enhanced data quality, preparing it for further analysis.

What does the final data set look like?

With the cleaning and transformation steps completed, the final datasets are structured as follows:

- **marketing_spend:** A monthly dataset with fields for franchise, vendor, spend amount, month, average daily spend, and number of days in the month.
- **daily_spend:** An expanded version of marketing_spend with each date represented individually to allow detailed time series analysis.
- **lead_data:** A dataset containing franchise details, dates, marketing channels, customer information, event details, and GrossProfit, all formatted and cleaned for consistency.

Each dataset is now condensed, with any unnecessary data or inconsistencies removed, making them ready for further analysis.

What information is not self-evident?

- **Marketing Channel Effectiveness:** While we have data on SpendAmount by marketing channel, it's not immediately clear which channels yield the highest return on investment (ROI). Calculating the ROI by analyzing gross profit relative to spend for each channel will provide insight into channel performance.
- **Conversion Time:** The difference between InsertDate and SoldDate (or lead-to-sale duration) is not directly observable but is crucial for understanding the sales cycle length. Calculating this duration will help us assess which channels or franchises have shorter sales cycles, potentially indicating a more efficient lead conversion process.
- **Seasonal Trends and Patterns:** SpendAmount, Leads, and conversions may fluctuate seasonally, with possible peaks in certain months or quarters, but this trend is not immediately visible. Aggregating the data by month or quarter and visualizing it can reveal these seasonal trends, which could inform budgeting and marketing planning.

- **Customer Segment Profitability:** Information on which customer segments (based on EventType, Source, or IsNewCustomer) contribute most to gross profit isn't immediately evident. Segmenting gross profit by these categories can uncover profitable customer demographics, guiding targeted marketing strategies.
- **Franchise Performance Variability:** While FranchiseName provides a category for analysis, it's not clear which franchises are the highest or lowest performers in terms of both spending efficiency and profitability. Analyzing gross profit and spend per franchise could reveal these insights, helping to identify top-performing locations or those needing support.
- **Lead Quality by Channel:** The quality of leads from different marketing channels isn't directly observable. By analyzing lead conversion rates and the associated GrossProfit from each channel, we can determine which channels bring in high-quality leads that convert into profitable sales.
- **Daily Spend Efficiency:** While average daily spend was calculated, daily fluctuations in SpendAmount could indicate days with higher or lower profitability, which is not initially apparent in the aggregated data. Expanding the data to daily levels can help identify peak spending days, potentially aligning with high-traffic sales days.
- **Cost per Lead and Cost per Sale:** Direct information on the cost per lead or cost per sale is missing, though it is essential for understanding spending efficiency. Deriving these metrics from SpendAmount and the number of leads or sales can offer a deeper perspective on how resources are allocated.

What are different ways you could look at this data?

Time-Based Analysis Analyzing data over time can reveal seasonal trends, monthly or quarterly patterns, and any long-term growth or decline in SpendAmount and GrossProfit.

- Monthly/Quarterly Trends: Group data by month or quarter to identify spending peaks or sales cycles.
- Daily Patterns: Using the daily_spend data, examine daily variations to identify high or low performance days, which could correlate with special promotions or customer behavior trends.
- Year-over-Year Comparisons: Comparing the same time period across years can help track growth or response to strategic changes.

Channel-Specific Analysis Breaking down data by MarketingChannelName can help measure the effectiveness of each marketing channel.

- ROI by Channel:Calculate return on investment (ROI) for each channel by comparing GrossProfit to SpendAmount.
- Conversion Rate by Channel: Determine the proportion of leads that convert to sales, which can help identify channels with high lead quality.
- Cost per Lead and Cost per Sale: Derive these metrics to understand the cost-effectiveness of each channel.

Franchise-Level Analysis Looking at data by FranchiseName helps measure the performance of each location and can uncover top performers or locations needing support.

- Profitability per Franchise: Compare gross profit generated by each franchise relative to its marketing spend.
- Spend Efficiency: Measure how well each franchise converts marketing spend into profitable sales.

Customer Segment Analysis By examining customer-specific variables like EventType (New vs. Used) and IsNewCustomer, I can identify which customer segments are most profitable.

- New vs. Used Sales: Determine if new or used car sales generate higher gross profit.
- Customer Loyalty: Compare profit from new vs. returning customers to assess customer retention effectiveness.
- Source Analysis: Check if certain sources generate higher-quality leads.

Spend Efficiency Analysis Evaluate how marketing spend translates into profit by calculating metrics like cost per lead and cost per sale.

- Average Daily Spend: Assess daily spend levels to ensure that spending aligns with high-performance days.
- Cost Per Conversion: Calculate cost per conversion for each marketing channel, giving a clearer picture of where to allocate budget.

Conversion Time Analysis Measure the time between InsertDate and SoldDate to understand the typical sales cycle length for each channel, franchise, or customer segment.

- Average Sales Cycle by Channel: Determine if certain channels or campaigns lead to faster conversions.
- Impact of Sales Cycle on Profit: Check if shorter or longer sales cycles have any correlation with profitability.

Lead Quality Analysis Examining lead attributes and their impact on GrossProfit can reveal the quality of leads from various channels and sources.

- Channel Lead Quality: Measure the conversion rate and average gross profit per lead for each channel.
- Source Lead Quality: Assess which sources yield high-quality, profitable leads.

Geographical Insights

- Regional Performance: Assess if certain regions perform better in terms of sales and profit.

Comparative Analysis Between Variables Look at the relationship between different variables, such as comparing spend and profit by month, channel, or franchise.

- Spend vs. Profit: Examine if there is a direct correlation between higher spending and higher profit.
- Lead Volume vs. Conversion Rate: Analyze if channels generating more leads also have higher or lower conversion rates.

Predictive Analysis and Modeling Applying machine learning or statistical models could help forecast or predict certain outcomes.

How do you plan to slice and dice the data?

To uncover meaningful insights, I plan to slice and dice the data across various dimensions, including time periods, marketing channels, franchises, customer segments, and performance metrics.

1. Time Period Analysis

- Monthly/Quarterly Aggregation: Group data by month and quarter to track trends in SpendAmount, GrossProfit, and lead volume over time.

- Daily Breakdown: Utilize the daily_spend data to analyze performance at a daily level, identifying specific days or periods with higher spend or profitability, which could highlight promotional or high-traffic days.
- Yearly Comparisons: Comparing year-over-year metrics can reveal growth trends and the impact of strategic changes.

2.Channel-Specific Slicing

- Channel-Based ROI: Compare return on investment (ROI) across MarketingChannel to understand which channels deliver the best profitability relative to spend.
- Lead Quality by Channel: Slice the data to view lead conversion rates, gross profit per lead, and cost per conversion for each channel, helping to identify the most effective lead sources.

3.Franchise-Level Analysis

- Profitability per Franchise: Examine each Franchise to assess gross profit and spend efficiency. By calculating GrossProfit and SpendAmount per franchise, I can identify high-performing and low-performing locations.
- Spend per Franchise by Channel: Dive into how each franchise allocates its marketing spend across channels, which could uncover opportunities to reallocate resources.

4.Customer Segmentation

- New vs. Returning Customers: Slice the data by IsNewCustomer to assess profitability and conversion rates for new versus returning customers, helping to evaluate the effectiveness of customer retention.
- New vs. Used Leads: Use EventType to differentiate between new and used leads, helping to understand which category yields more conversions and profit.
- Source Analysis: Group by Source to explore which sources provide high-quality, profitable leads.

5.Conversion Time Analysis

- Lead Conversion Speed by Channel: Calculate the time between InsertDate and SoldDate for each lead and segment it by channel to assess which channels deliver faster conversions.
- Impact of Conversion Time on Profit: Analyze if shorter or longer sales cycles have a correlation with higher gross profit, helping to focus efforts on channels with optimal conversion times.

6.Cost Efficiency Metrics

- Cost per Lead and Cost per Sale: Calculate and slice data by channel and franchise to determine cost per lead and cost per sale metrics. This will reveal which channels are cost-effective and which may be overspending relative to conversions.
- Daily Spend Analysis: With AvgDailySpend in daily_spend, slice the data by day to evaluate if certain periods yield higher returns on daily spending.

7.Comparative Slicing Across Variables

- Spend vs. Profit Correlation: Slice and analyze data by various categories (e.g., channel, franchise, month) to examine the relationship between spend and profit, helping to identify where increased spend directly leads to higher returns.
- Lead Volume vs. Conversion Rate by Channel: Compare channels with high lead volumes to their respective conversion rates to identify if high lead-generating channels also maintain quality conversions.

8.Predictive Grouping for Modeling

- Predictive Binning for Conversion Analysis: Create bins based on GrossProfit, conversion time, and other variables to feed into machine learning models, which can predict the likelihood of conversion and profitability for each channel and customer type.

How could you summarize your data to answer key questions?

To summarize the data and answer key questions, I will calculate targeted metrics and aggregations. First, for understanding marketing channel performance, I'll calculate return on investment (ROI) by dividing each channel's total GrossProfit by SpendAmount to identify the most profitable channels. I'll also derive cost efficiency metrics, such as cost per lead and cost per sale, to evaluate which channels deliver high-quality leads at lower costs. Additionally, monthly and quarterly summaries of SpendAmount, GrossProfit, and lead volume will reveal seasonal patterns and peak spending or sales periods. Year-over-year growth rates will further enhance this analysis, helping to assess long-term trends and the impact of marketing strategy adjustments over time.

To evaluate franchise performance and customer segments, I'll summarize GrossProfit, SpendAmount, and conversion rates by franchise to identify top-performing locations and those needing additional support. Further, breaking down each franchise's spending by channel will reveal if certain franchises achieve better results with specific channels. To understand customer segments, I'll calculate conversion rates, gross profit, and spend for new versus returning customers, as well as for new and used sales. This segmentation will clarify which groups are most profitable and assess the effectiveness of customer retention efforts, helping to refine target audiences and optimize resource allocation.

What types of plots and tables will help you to illustrate the findings to your questions?

To effectively illustrate the findings, I will use a combination of plots and tables that highlight trends, comparisons, and distributions in the data:

Time Series Plots: Line charts will be used to track SpendAmount, GrossProfit, and lead volume over time, both monthly and quarterly. These plots will reveal seasonal trends, growth patterns, and any peaks or dips in performance, helping us understand how spending and profitability fluctuate throughout the year.

Bar Charts: Bar charts will provide a clear comparison of key metrics across categories, such as ROI by MarketingChannelName and profitability by FranchiseName. Bar charts will also illustrate the cost per lead and cost per sale for each marketing channel, allowing for a straightforward comparison of cost efficiency.

Scatter Plots: To understand the relationship between SpendAmount and GrossProfit, scatter plots will be used to visualize if higher spending correlates with higher profits. Additionally, scatter plots with outlier boundaries will help identify extreme values in gross profit and spend that may skew averages and need closer examination.

Box Plots: For analyzing distributions and identifying outliers, box plots will be helpful. For example, box plots of GrossProfit and conversion time by marketing channel and franchise will show the spread of data, helping to identify channels or franchises with consistently high or low performance.

Summary Tables: Aggregated tables will provide an overview of key metrics like total and average SpendAmount, GrossProfit, conversion rates, and ROI for each marketing channel and franchise. Tables comparing new versus returning customer performance, as well as new versus used sales, will summarize profitability and conversion rates across customer segments.

Heatmaps: To visualize high and low ROI regions or channels, heatmaps will be used. For instance, a heatmap of ROI by channel and franchise will highlight areas of high efficiency and profitability, providing a quick visual of where resources are best allocated.

These plots and tables, with clear labels, titles, and legends, will provide comprehensive visuals that make it easier to interpret and communicate the findings to stakeholders.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Yes, I plan to incorporate machine learning techniques to enhance the analysis and answer key research questions related to lead conversion, profitability prediction, and channel performance. Specifically, I'll explore predictive modeling to identify patterns in the data that are not immediately visible with basic summaries and visualizations.

What questions do you have now, that will lead to further analysis or additional steps?

At this stage, several questions have emerged that will guide further analysis and potentially require additional steps:

How consistent is the conversion rate across different marketing channels and franchises? While initial analysis will reveal average conversion rates, I'm curious about the stability of these rates over time and across various segments. Identifying any significant fluctuations could help isolate factors that influence conversion, such as seasonal effects or franchise-specific practices.

What is the optimal spend threshold for maximizing profitability across channels? While I'll examine the relationship between SpendAmount and GrossProfit, understanding if there is a specific threshold or "sweet spot" in spend that maximizes ROI for each channel would be valuable. This may require testing different budget levels or leveraging regression analysis to identify the point of diminishing returns.

Are certain customer characteristics more predictive of higher gross profit or faster conversion times? Customer details like EventType (New vs. Used) and Source could impact both profit and conversion speed. Investigating these characteristics might reveal high-value customer segments, allowing for refined targeting strategies. This could lead to additional feature engineering or data transformation to better capture these insights.

How do seasonal trends impact gross profit and conversion rates? Initial time-series analysis may show seasonal variations, but understanding the depth and reasons behind these trends would require looking at external factors like holidays, economic indicators, or industry trends. It might lead to incorporating new variables or data sources to contextualize the seasonality observed.

Can I predict customer retention or repeat purchases based on initial lead data? If customer retention or repeat purchases are tracked, I would like to explore if initial lead characteristics can predict future retention. This might lead to additional data collection efforts or advanced predictive modeling to focus on long-term customer value.

Final Project: Step 3

Introduction

In today's highly competitive automotive retail market, the ability to optimize marketing spend is crucial. Automotive retailers often rely on various digital channels, from Google Paid Search to third-party platforms like Edmunds and Carfax, to reach potential customers. However, without clear insights into each channel's effectiveness, these marketing dollars may not be maximized, resulting in missed opportunities and increased acquisition costs.

This project seeks to address these challenges by exploring how different marketing channels contribute to lead quality and conversion rates. By analyzing simulated data that mimics real-world marketing spend and lead conversion patterns, I aim to uncover trends and insights that can guide strategic marketing decisions. This analysis ultimately provides a roadmap for maximizing ROI and reducing wasteful spending, directly impacting the bottom line in a way that benefits both our automotive retail businesses and our end consumers.

The problem statement you addressed

This project tackles a specific issue faced by automotive retailers: how to allocate marketing spend across various digital channels to maximize ROI and lead conversion rates. With budgets divided across multiple platforms—each with distinct costs and audience reach—retailers often struggle to determine which channels yield the highest-quality leads for the lowest acquisition costs.

To address this, our analysis centers on key questions: Which channels drive the best conversions? How should spend be adjusted seasonally or by vendor? Are there optimal thresholds for each channel that maximize profitability? By answering these questions, the goal is to provide a data-driven foundation for smarter, more effective marketing budget allocation, thereby enhancing both efficiency and profitability.

How you addressed the problem statement

To address the problem of optimizing marketing spend, I created a series of simulated datasets representing real-world dynamics. These datasets focused on key areas: marketing spend by channel, lead generation and conversions, and geographic variations. I took the time to cleanse the data thoroughly, handling missing values, capping outliers, and converting dates for consistency, ensuring that the analysis would be based on reliable information.

I applied several data science techniques to extract insights from this dataset:

- Exploratory Data Analysis helped me understand basic trends in spending and lead quality across different channels and franchises.
- Time-Series Analysis allowed me to uncover seasonal trends in spending and conversion rates, guiding decisions about when to increase or decrease budget allocations.
- Channel and Franchise Analysis revealed which marketing channels and franchises were the most profitable, showing where resources are best spent.
- Customer Segment Analysis provided insights into which customer groups—such as new versus returning customers—were the most profitable, helping me refine target audiences.
- ROI and Conversion Metrics allowed me to assess each channel's efficiency by calculating cost per lead and cost per sale, which indicated the channels delivering the best return on investment.

In the future, I could take this project further by implementing a predictive model to forecast lead quality or conversion likelihood. For example, A regression model could predict gross profit based on spend by channel, while a classification model could help identify high-converting leads. These models would enable proactive budget adjustments, allowing for continuous optimization of marketing spend.

Analysis

My analysis uncovered several key insights that can guide more efficient marketing spend allocation:

- Channel Effectiveness: Certain channels consistently delivered the highest ROI, suggesting that these channels should be prioritized in budget allocation. Conversely, channels with lower ROI may be candidates for reduced spending or optimization efforts.

- Seasonal Trends: There were noticeable peaks in lead conversions and spending effectiveness during specific times of the year, indicating that adjusting the budget seasonally could yield better returns.
- Franchise Performance: Performance varied significantly by franchise, with some locations achieving higher gross profit relative to spend. This insight allows for more targeted support to underperforming franchises and optimized budget allocation for high performers.
- Customer Segment Profitability: Returning customers and leads for new cars tended to have higher conversion rates and gross profit. This suggests a more targeted approach toward these segments may enhance profitability.
- ROI and Cost Efficiency: Channels with the lowest cost per lead and highest conversion rates clearly stood out, helping identify where spending yields the most efficient returns. These channels should be emphasized to maximize cost-effective lead generation.

These insights provide a foundation for making data-driven decisions on where to allocate marketing spend, ensuring resources are focused on the most profitable channels, seasons, franchises, and customer segments.

Implications

The insights from this analysis offer valuable implications for decision-makers in the automotive retail industry. By reallocating budgets to the highest-performing channels, our marketing team can achieve greater efficiency, reducing unnecessary costs while boosting lead quality and conversion rates. Furthermore, understanding seasonal trends allows for smart budget adjustments that capitalize on peak times, ensuring that resources are used effectively year-round.

These findings also empower our executive leadership team to assess individual store performance and focus support where it's needed most. With targeted resources, underperforming locations can be bolstered, and top-performing stores can further refine their approach.

Additionally, improved customer targeting based on segment profitability allows marketing teams to focus on the highest-converting groups, enhancing both lead quality and profit margins. This leads to more personalized and impactful advertising, creating a better customer experience and building stronger connections with the brand.

Overall, implementing these data-driven insights fosters long-term strategic growth and ensures that marketing spend is continuously optimized. This approach not only enhances profitability but also provides a competitive edge in the marketplace.

Limitations

While this analysis provides valuable insights, there are some limitations to consider. First, because the data used was simulated, it may lack the complexity of real-world data, which could introduce factors such as nuanced customer behaviors, economic conditions, or market-specific trends that impact spend efficiency. Real data could provide a more accurate reflection of these dynamics and might reveal additional insights.

Additionally, the scope of this analysis does not fully account for external influences, such as competitor actions or economic shifts, which can significantly affect consumer behavior and marketing outcomes. Further exploration could incorporate these factors to offer a more comprehensive view of spend optimization.

The ROI, cost per lead, and cost per sale calculations also assume consistent customer behavior and conversion rates. However, these metrics may vary across regions, time periods, and customer demographics, so further segmentation or real-time analysis could enhance accuracy. Collecting data on customer preferences or integrating customer feedback could also help to refine targeting strategies.

In the future, implementing advanced machine learning models would allow for real-time budget adjustments and more granular predictions, further optimizing spend allocation.

Lastly, another limitation is the scope of locations used in this analysis. For simplicity, I chose to work with only four locations; however, in the real-world scenario, there are 34 locations across the U.S. Analyzing all locations could reveal region-specific trends in spend efficiency and lead quality that weren't fully captured in this limited sample. Future work could expand the analysis to include all locations, allowing for a more detailed understanding of geographic variations and providing each location with tailored recommendations for spend optimization.

Overall, acknowledging these limitations highlights opportunities for continued improvement, ensuring that the insights provided can evolve and remain effective in guiding data-driven marketing decisions.

Concluding Remarks

In conclusion, this project has provided a structured approach to optimizing marketing spend allocation in the automotive retail industry. By focusing on lead quality, conversion rates, and ROI across various channels and locations, my analysis offers a clear framework for making more efficient, data-driven budget decisions.

Key takeaways include the value of reallocating budgets toward high-ROI channels, strategically adjusting spend based on seasonal trends, targeting profitable customer segments, and assessing location performance individually to support targeted improvements. These insights provide actionable guidance for our marketing team and our executive leadership team aiming to maximize their return on investment and improve lead quality.

As the marketing environment continues to evolve, this analysis can be further refined with real-world data and more advanced predictive models, allowing for real-time adjustments and tailored strategies for each location. Ultimately, the ongoing application of data-driven strategies will be critical to staying competitive, enabling the organization to optimize resources and achieve sustained growth in a dynamic marketplace.

This project reinforces the value of combining data science with strategic decision-making. By continuously analyzing and adapting to new insights, automotive retailers can ensure that their marketing efforts remain both effective and efficient, creating lasting benefits for both the business and its customers.