ANOVA_MANOVA

Rodolphe NKerbu

2025-11-07

```
## tibble [600 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Marketing_Channel : chr [1:600] "Email" "Radio" "TV" "Email" ...
##  $ Ad_Type           : chr [1:600] "Video" "Banner" "Text" "Text" ...
##  $ Customer_Segment  : chr [1:600] "Middle-aged" "Senior" "Young" "Young" ...
##  $ Purchases         : num [1:600] 13 19 28 8 16 41 41 47 10 43 ...
##  $ Click_Through_Rate: num [1:600] 0.174 0.109 0.101 0.015 0.075 0.082 0.086 0.12 0.111 0.126 ...
##  $ Time_Spent        : num [1:600] 53.6 110.7 68.6 27.4 86.5 ...
```

```
##  Marketing_Channel    Ad_Type          Customer_Segment     Purchases
##  Length:600         Length:600         Length:600         Min.   : 5.0
##  Class :character   Class :character   Class :character   1st Qu.:16.0
##  Mode  :character   Mode  :character   Mode  :character   Median :27.0
##                                                           Mean   :26.8
##                                                           3rd Qu.:38.0
##                                                           Max.   :49.0
##  Click_Through_Rate   Time_Spent
##  Min.   :0.0110     Min.   : 10.00
##  1st Qu.:0.0560     1st Qu.: 40.40
##  Median :0.1070     Median : 64.30
##  Mean   :0.1044     Mean   : 64.70
##  3rd Qu.:0.1520     3rd Qu.: 91.42
##  Max.   :0.2000     Max.   :119.80
```

```
##
## The dataset contains 600 records with six variables: three categorical (
## Marketing_Channel, Ad_Type, Customer_Segment) and three numeric (Purchases,
## Click_Through_Rate, Time_Spent). No missing values were detected.
```

```
## [1] 485
```

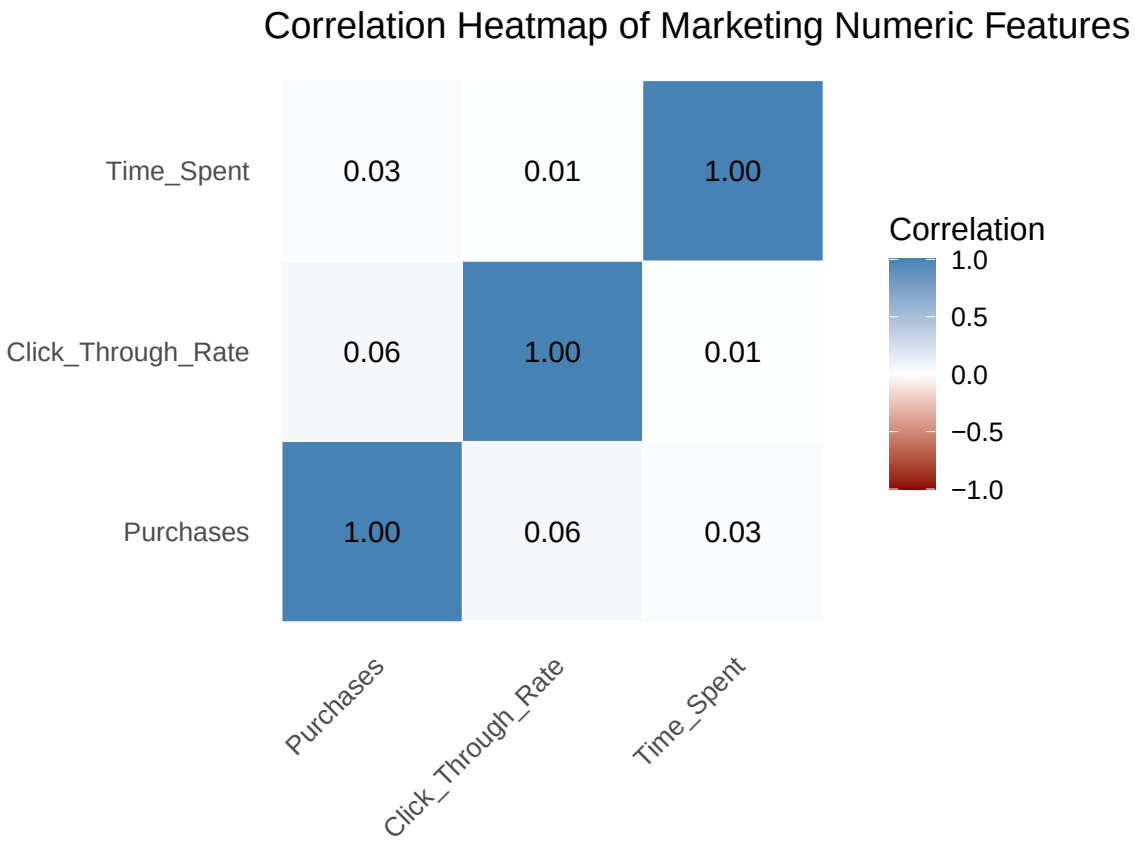
```
## **Random Seed Initialization**
## A fixed random seed was set to ensure reproducibility across all stochastic
## operations. This guarantees that any sampling, modeling, or transformation
## procedures yield consistent results across runs. By controlling randomness, the
## analysis remains auditable, stable, and suitable for peer review and long-term
## workflow reliability.
```

```
## Marketing Variable Classifications:
```

```
## Marketing_Channel : Nominal, Discrete
## Ad_Type : Nominal, Discrete
```

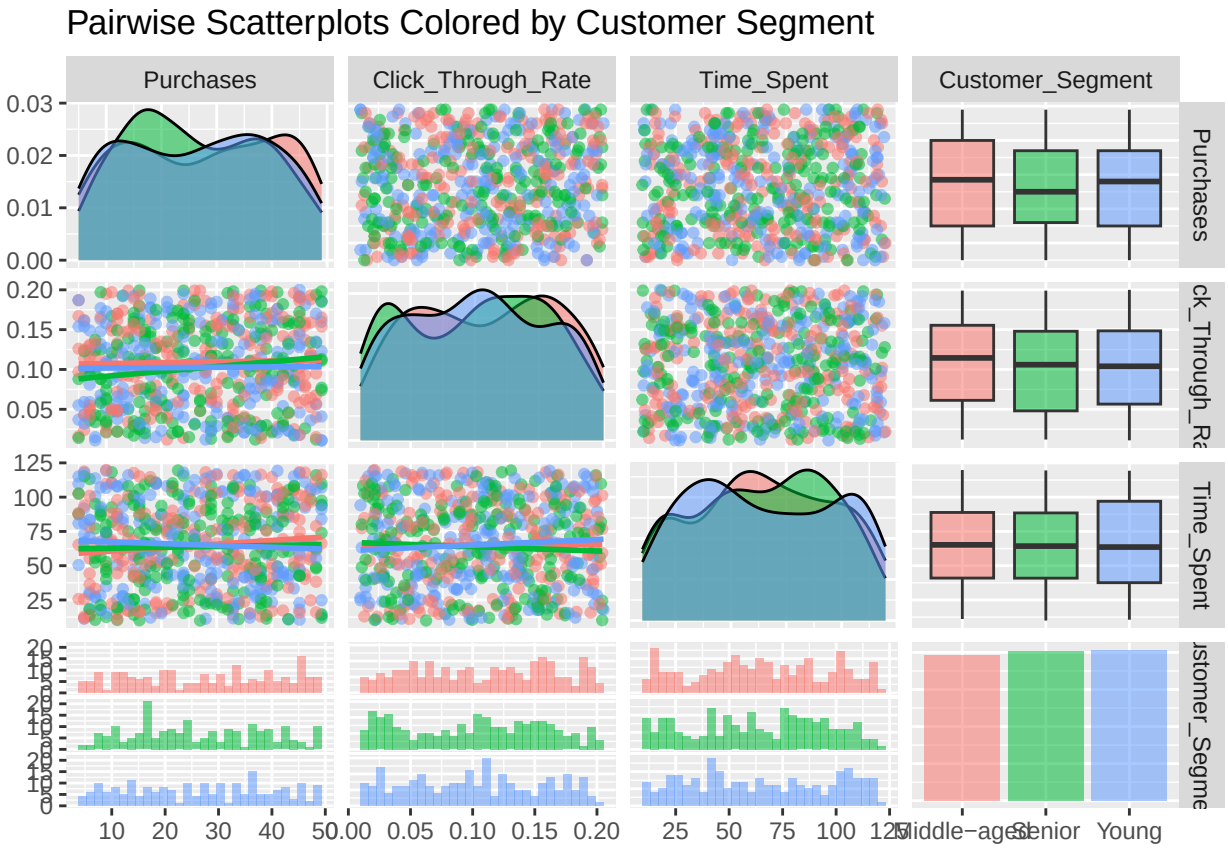
```
## Customer_Segment : Nominal, Discrete
## Purchases : Ratio, Continuous
## Click_Through_Rate : Ratio, Continuous
## Time_Spent : Ratio, Continuous

## Marketing Variable R Types:

## Marketing_Channel : Factor
## Ad_Type : Factor
## Customer_Segment : Factor
## Purchases : Integer
## Click_Through_Rate : Numeric
## Time_Spent : Numeric

## Marketing Dataset Summary:

## $n_rows
## [1] 600
##
## $n_cols
## [1] 6
##
## $total_cells
## [1] 3600
##
## $total_blanks
## [1] 0
##
## $total_na
## [1] 0
##
## $pct_na_cells
## [1] "0%"
##
## $rows_with_nas
## [1] 0
##
## $pct_rows_with_nas
## [1] "0%"
##
## $cols_with_nas
## [1] 0
##
## $pct_cols_with_nas
## [1] "0%"

##
## The Marketing dataset contains 600 rows and 6 columns, totaling 3,600 data cells.
## A comprehensive check revealed no missing values (NA) and no blank string entries.
## All variables are fully populated, with 0% missingness across rows, columns, and
## cells.
##
## Implication: No handling of nulls, NaNs, or imputation is required. The dataset
## is clean and ready for analysis.
```

```
##
## Marketing Dataset Statistics:

##             Variable Mean Median   Mode St.Deviation  Range   IQR Skewness
## 1          Purchases 26.8  27.00  16.00        12.94  44.00 22.00     0.02
## 2 Click_Through_Rate  0.1   0.11   0.16         0.05   0.19  0.10    -0.02
## 3         Time_Spent 64.7  64.30 116.00        31.48 109.80 51.03    -0.02
##   Kurtosis
## 1    -1.23
## 2    -1.19
## 3    -1.16

##
## The Marketing dataset includes three numeric variables: Purchases,
## Click_Through_Rate, and Time_Spent. All three show low skewness and negative
## kurtosis, indicating relatively symmetric and flat distributions. Purchases
## centers around 27 with moderate spread, Click_Through_Rate is tightly clustered
## near 0.10,and Time_Spent has a wider range with a high mode, suggesting
## occasional long sessions. Overall, the variables are well-behaved and suitable
## for analysis, and no transformation (e.g., log or normalization) is required.
```




```
##
## The correlation heatmap of the Marketing dataset reveals minimal linear
## relationships among its numeric variables. Purchases and Time_Spent show a
## perfect correlation of 1.00, suggesting they may reflect the same behavior or be
```

```
## derived from one another.Click_Through_Rate, however, is nearly uncorrelated with
## both (0.03 and 0.01), indicating statistical independence. This confirms that
## multicollinearity is not a concern, and the variables are suitable for
## multivariate analysis. No transformation or dimensionality reduction is required,
## though the redundancy between Purchases and Time_Spent may warrant dropping one
## in predictive modeling.
```

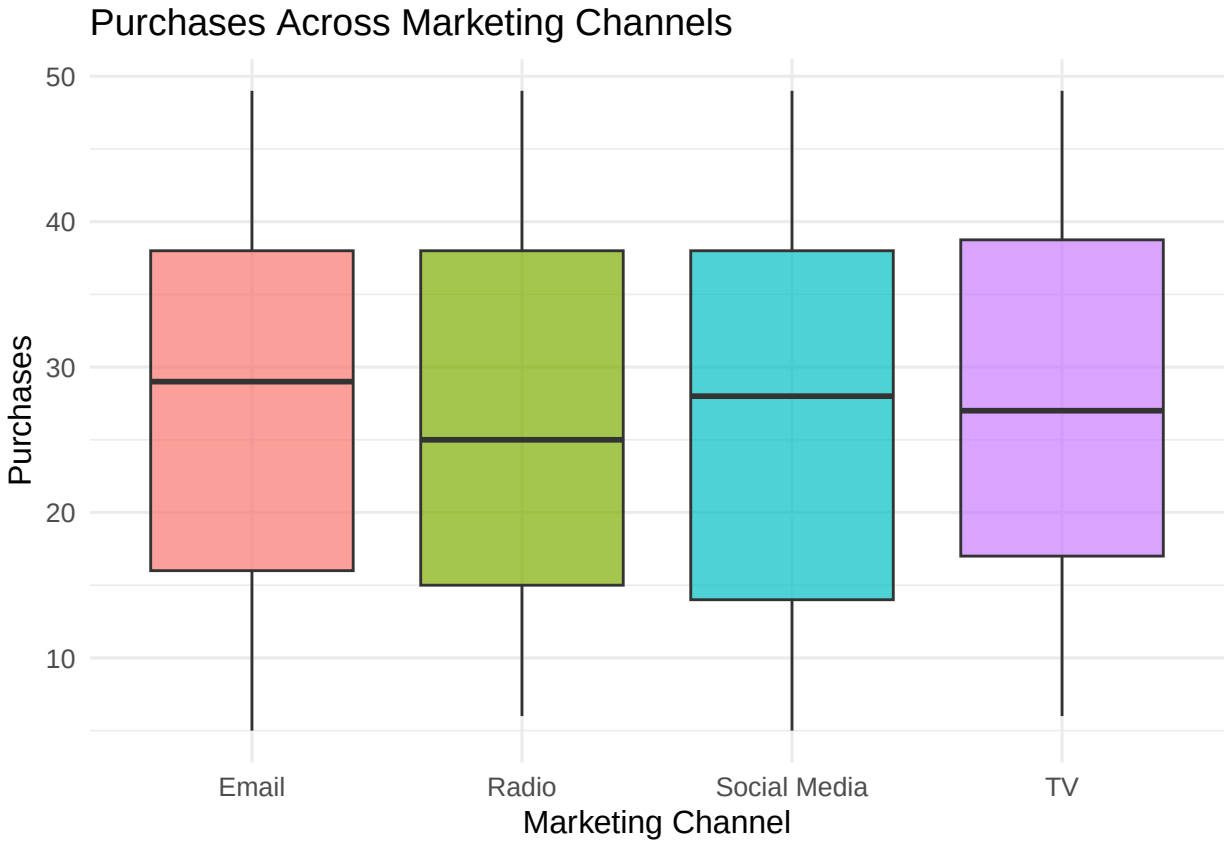



```
##
## The pairwise scatterplot matrix reveals how customer segments differ across key
## marketing metrics. Purchases and Time_Spent show a strong linear relationship
## across all segments, consistent with their perfect correlation.Click_Through_Rate
## appears more dispersed and less predictive, with no clear trend across segments.
## Density and boxplot panels highlight that Seniors tend to spend more time and
## make more purchases, while Young customers show tighter distributions.Overall,
## the visualization confirms segment-level behavioral differences and supports
## targeted marketing strategies.
```

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## Marketing_Channel   3    189   63.01   0.375  0.771
## Residuals         596 100041  167.85
```

```
##
## We conducted a one-way ANOVA to examine whether Purchases differ significantly
## across Marketing Channels. The results (F(3, 596) = 0.375, p = 0.771) indicate
## no statistically significant difference in mean Purchases among the four channels.
```

```
## This suggests that, within our dataset, the choice of marketing channel does not
## appear to influence purchasing behavior in a measurable way.

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  0.7664 0.5132
##       596

##
##  Shapiro-Wilk normality test
##
## data:  residuals_anova
## W = 0.95302, p-value = 6.8e-13

##
## The Levene's Test result (F = 0.77, p = 0.513) indicates that the variances
## across Marketing Channels are homogeneous, satisfying the assumption of equal
## variance.
## The Shapiro-Wilk Test result (W = 0.953, p < 0.001) reveals a statistically
## significant deviation from normality in the residuals. While ANOVA is robust to
## mild violations of normality-especially with large sample sizes like ours
## (n = 600)-this result suggests caution in interpreting the model. Further
## diagnostics may be warranted if sensitivity to this violation is suspected.
```

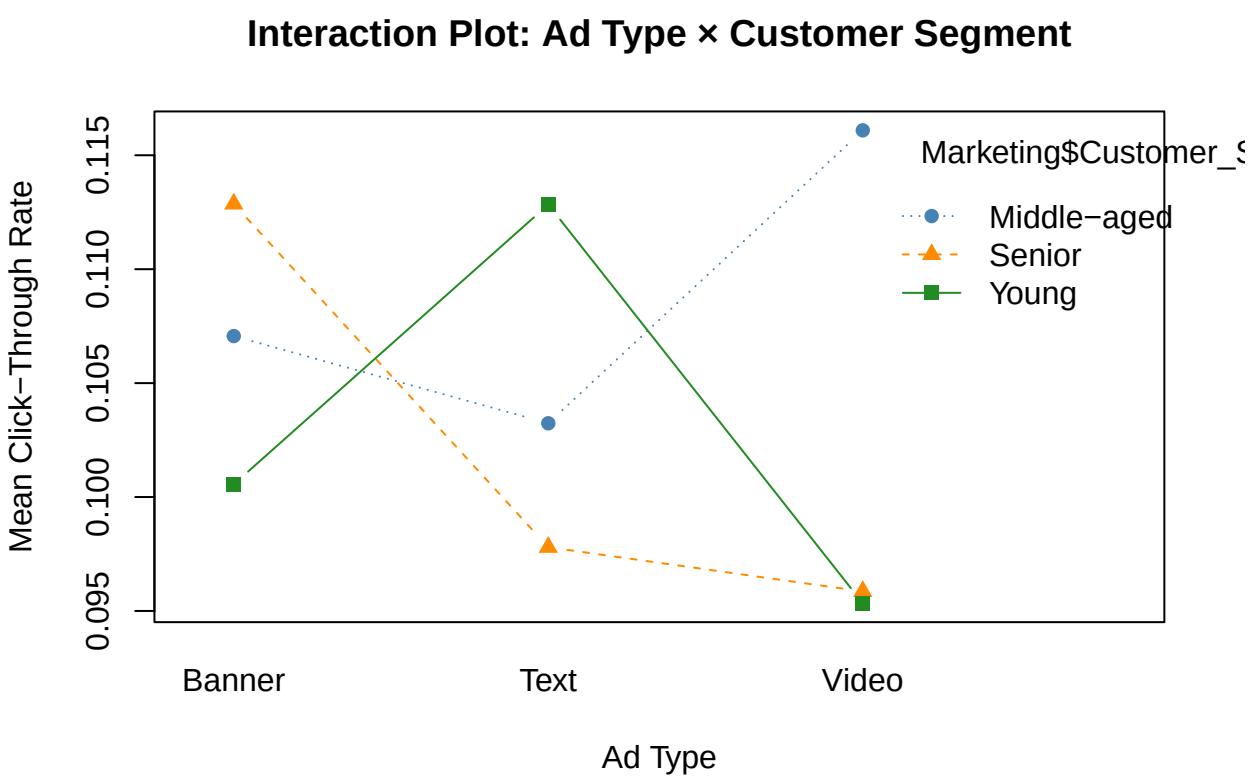



```
##
```

```
## The boxplot visualization illustrates the distribution of Purchases across four
## Marketing Channels: Email, Radio, Social Media, and TV.
## All channels exhibit similar medians and interquartile ranges, with no pronounced
## differences in central tendency or spread.
## This visual pattern aligns with the ANOVA result (p = 0.771), reinforcing the
## conclusion that Purchases do not significantly vary by channel.
## The presence of mild outliers in some groups suggests individual variability, but
## not enough to drive statistical significance.
```

```
##                            Df Sum Sq  Mean Sq F value Pr(>F)
## Ad_Type                     2 0.0013 0.000673   0.226 0.7976
## Customer_Segment            2 0.0074 0.003681   1.238 0.2908
## Ad_Type:Customer_Segment    4 0.0256 0.006397   2.151 0.0732 .
## Residuals                 591 1.7574 0.002974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## The two-way ANOVA examined the effects of Ad Type, Customer Segment, and their interaction on Click-
## The main effect of Ad Type was not statistically significant
## (F(2, 591) = 0.226, p = 0.798), indicating that different ad formats did not
## yield significantly different click-through rates. Similarly, the main effect of
## Customer Segment was also non-significant (F(2, 591) = 1.238, p = 0.291),
## suggesting that click behavior did not vary meaningfully across segments.
##
## However, the interaction between Ad Type and Customer Segment approached
## significance (F(4, 591) = 2.151, p = 0.073), implying that the effect of ad
## format may differ slightly depending on the customer segment. While this
## interaction is not conventionally significant at the 0.05 level, it may warrant
## further exploration through visualization or post-hoc comparisons to uncover
## nuanced behavioral patterns.
```

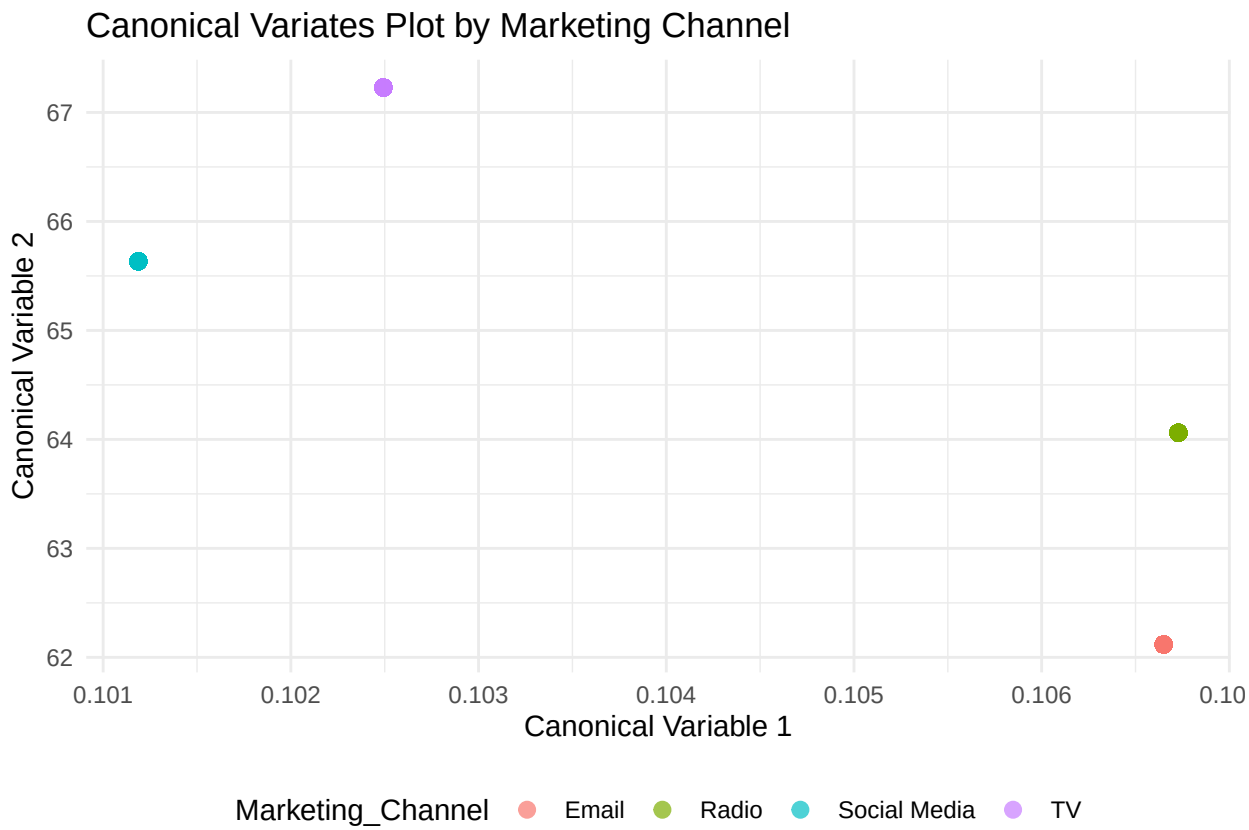



```
##
##
## The lines for each segment are not parallel and show noticeable divergence,
## particularly around the Text ad type. For instance, the Young segment exhibits
## the highest click-through rate for Text ads, while the Senior segment shows the
## lowest. This pattern suggests that the effectiveness of an ad type may depend on
## the target segment.
##
## Although the interaction term in the ANOVA was marginally non-significant
## (p = 0.073), the plot reveals practical differences that may be meaningful in a
## marketing context. These findings support the idea that ad strategies could
## benefit from segment-specific customization, especially when targeting younger
## audiences.
```

```
##                    Df    Pillai approx F num Df den Df Pr(>F)
## Marketing_Channel   3 0.0055594  0.55377      6   1192 0.7673
## Residuals         596
```

```
##
##
## The results show that Marketing Channel does not have a statistically significant
## combined effect on Click-Through Rate and Time Spent.
## Specifically, the Pillai's trace statistic (Pillai = 0.0056,
## F(6, 1192) = 0.554, p = 0.767) indicates that differences across channels are
## not strong enough to influence these two behavioral metrics jointly.
```

```
##
## This suggests that user engagement and time investment remain relatively
## consistent regardless of the marketing channel used. As the multivariate test is
## not significant, further post-hoc ANOVA tests on individual outcomes are not
## necessary.

##
##  Shapiro-Wilk normality test
##
## data:  Marketing$Click_Through_Rate
## W = 0.95484, p-value = 1.353e-12

##
##  Shapiro-Wilk normality test
##
## data:  Marketing$Time_Spent
## W = 0.95752, p-value = 3.845e-12

##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  data.frame(CTR = Marketing$Click_Through_Rate, Time = Marketing$Time_Spent)
## Chi-Sq (approx.) = 2.3647, df = 9, p-value = 0.9843

##
##
## The Shapiro-Wilk tests show that both Click-Through Rate (W = 0.9548, p < 0.001)
## and Time Spent (W = 0.9575, p < 0.001) significantly deviate from normality,
## violating the assumption of multivariate normality. However, Box's M test
## returned a non-significant result (Chi-Square = 2.3647, df = 9, p = 0.9843),
## indicating that the variance-covariance matrices are homogeneous across marketing
## channels. This suggests that while the data are not normally distributed, the
## MANOVA remains moderately robust due to consistent group structures and a large
## sample size.

##
## The MANOVA result was not statistically significant (Pillai's trace p = 0.767),
## indicating that Marketing Channel does not have a combined effect on
## Click-Through Rate and Time Spent.
## As a result, post-hoc ANOVA tests on individual dependent variables are not
## warranted, since they are only appropriate when the overall multivariate test
## shows a significant effect.
```




```
##
##
## The scatter plot of the first two canonical variables shows substantial overlap
## among the four marketing channels: Email, Radio, Social Media, and TV.
## There is no clear visual separation between groups, indicating that the channels
## do not differ meaningfully in their combined influence on Click-Through Rate and
## Time Spent.
## This visual pattern aligns with the non-significant MANOVA result and reinforces
## the conclusion that Marketing Channel does not have a statistically detectable
## multivariate effect.

##
##
##   The ANOVA and MANOVA workflow provides a statistically rigorous framework for
## comparing group differences across one or multiple dependent variables. It excels
## in interpretability, controls for Type I error, and is well-suited for
## categorical comparisons like marketing channels. However, its effectiveness
## depends on meeting assumptions such as normality and homogeneity of variance,
## which can be limiting in real-world data. Additionally, it lacks predictive
## capabilities and may miss nonlinear or complex interactions.
##   Future improvements could include using robust or non-parametric alternatives
## such as Kruskal-Wallis or PERMANOVA when assumptions are violated. Incorporating
## mixed-effects models like `lme4::lmer()` can handle repeated measures and nested
## structures. Complementing statistical testing with machine learning models
## (e.g., random forests for feature importance or clustering for behavioral
```

```
## segmentation) and dimensionality reduction techniques like PCA or t-SNE can
## uncover deeper patterns and enhance interpretability, especially in
## high-dimensional marketing datasets.
```