# PREDICTIVE ANALYTICS: FACEBOOK METRICS

Alfonso Berumen, MS, MBA

E: alfonso.berumen@gmail.com

LinkedIn: https://www.linkedin.com/in/alfonsoberumen/

SatRday-Los Angeles: APRIL 6, 2019

# FIRST STEPS WHILE I TALK ABOUT MYSELF

- **Download the data and script file (GitHub): [https://github.com/ladataanalytics/SatRday](https://github.com/ladataanalytics/SatRday)**
- **Install packages:**

```
##############
#packages
##############
#for some prepping
#install.packages("dplyr")
library(dplyr)
## Warning: package 'dplyr' was built under R version 3.5.2
##
## 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
#for EDA
#install.packages("ggplot2")
library(ggplot2)

#for error metrics
#install.packages("MLmetrics")
library(MLmetrics)
##
## Attaching package: 'MLmetrics'
## The following object is masked from 'package:base':
##
##     Recall
#install.packages("stargazer")
library(stargazer)
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
#for a pretty plot
#install.packages("lattice")
library(lattice)

#for a regression tress
#install.packages("rpart")
library(rpart)
#install.packages("rattle")
library(rattle)
## Rattle: A free graphical interface for data science with R.
## Version 5.2.0 Copyright (c) 2006-2018 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
#install.packages("rpart.plot")
library(rpart.plot)
#install.packages("RColorBrewer")
library(RColorBrewer)
```

# DATA

- **Facebook metrics Data Set:**
  - Source: https://archive.ics.uci.edu/ml/datasets/Facebook+metrics#

- **Description**: posts published during the year of 2014 on the Facebook page of a cosmetics brand
  - 500 of the 790 rows and part of the features that were analyzed for an academic research publication (Moro et al., 2016)

- **Citation**: Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. Journal of Business Research, 69(9), 3341-3351.

# DATA DESCRIPTION (MORO ET AL., 2016)

**Table 1**
Features from the compiled data set

| Feature | Type of information | Source | Data type |
|---|---|---|---|
| Posted | Identification | Facebook | Date/time |
| Permanent link | Identification | Facebook | Text |
| Post ID | | | |
| Post message | Content | Facebook | Text |
| Type | Categorization | Facebook | Factor: {Link, Photo, Status, Video } |
| Category | Categorization | Facebook page managers | Factor: {action, product, inspiration } |
| Paid | Categorization | Facebook | Factor: {yes, no } |
| Page total likes | Performance | Facebook | Numeric |
| Lifetime post total reach | | | |
| Lifetime post total impressions | | | |
| Lifetime engaged users | | | |
| Lifetime post consumers | | | |
| Lifetime post consumptions | | | |
| Lifetime post impressions by people who have liked your page | | | |
| Lifetime post reach by people who like your page | | | |
| Lifetime people who have liked your page and engaged with your post | | | |
| Comments | Performance | Facebook | Numeric |
| Likes | | | |
| Shares | | | |
| Total interactions | Performance | Computed | Numeric |

# DATA DESCRIPTION (MORO ET AL., 2016)

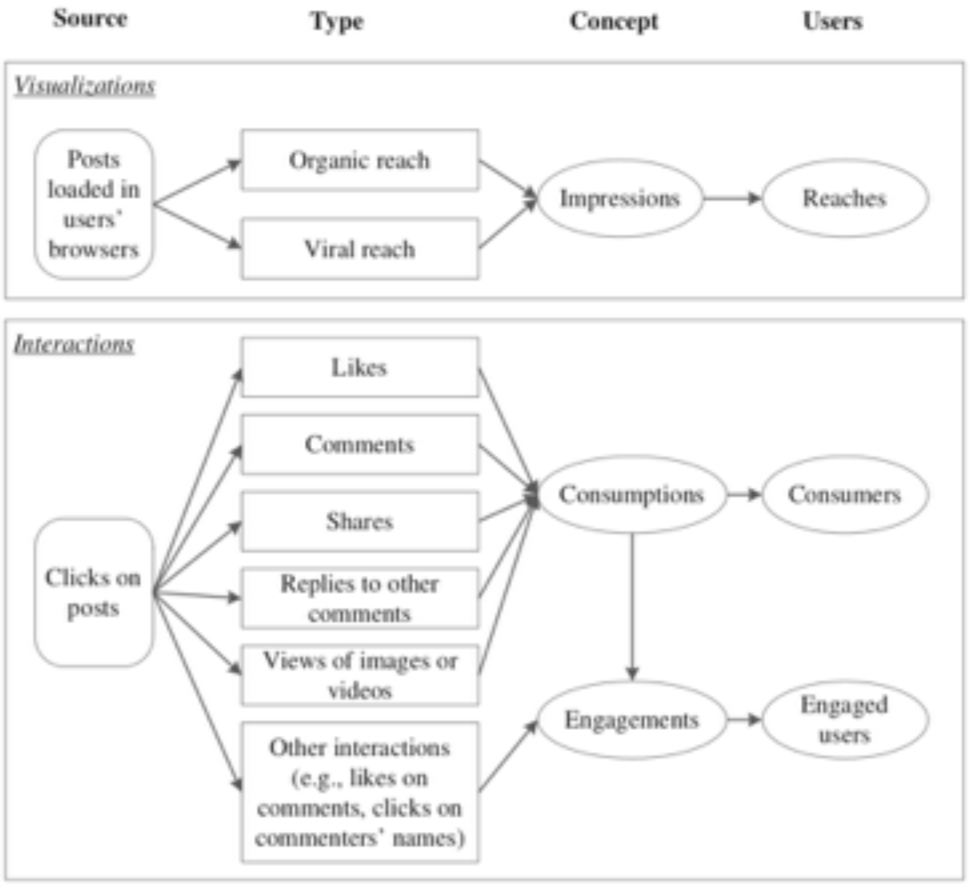



**Fig. 1.** Conceptual map on Facebook's performance metrics.
More detailed information can be obtained from:

https://developers.facebook.com/docs/graph-api/reference/v2.5/insights
http://www.agorapulse.com/blog/facebook-post-consumers-and-post-consumption

# WHAT ARE WE MODELLING TODAY?

**Table 2**
List of output features to be modeled

| Feature | Description[a] |
|---|---|
| Lifetime post total reach | The number of people who saw a page post (unique users). |
| Lifetime post total impressions | Impressions are the number of times a post from a page is displayed, whether the post is clicked or not. People may see multiple impressions of the same post. For example, someone might see a Page update in News Feed once, and then a second time if a friend shares it. |
| Lifetime engaged users | The number of people who clicked anywhere in a post (unique users). |
| Lifetime post consumers | The number of people who clicked anywhere in a post. |
| Lifetime post consumptions | The number of clicks anywhere in a post. |
| Lifetime post impressions by people who have liked a page | Total number of impressions just from people who have liked a page. |
| Lifetime post reach by people who like a page | The number of people who saw a page post because they have liked that page (unique users). |
| Lifetime people who have liked a page and engaged with a post | The number of people who have liked a Page and clicked anywhere in a post (Unique users). |
| Comments | Number of comments on the publication. |
| Likes | Number of "Likes" on the publication. |
| Shares | Number of times the publication was shared. |
| Total interactions | The sum of "likes," "comments," and "shares" of the post. |

[a] Descriptions extracted from:

- http://www.agorapulse.com/blog/facebook-reach-metrics-ultimate-guide
- https://www.facebook.com/help/274400362581037

# DATA DESCRIPTION (MORO ET AL., 2016)

List of input features used for modeling

| Feature | Description |
| --- | --- |
| Category | Manual content characterization: action (special offers and contests), product (direct advertisement, explicit brand content), and inspiration (non-explicit brand related content). |
| Page total likes | Number of people who have liked the company's page. |
| Type | Type of content (Link, Photo, Status, Video). |
| Post month | Month the post was published (January, February, March, ..., December). |
| Post hour | Hour the post was published (0, 1, 2, 3, 4, ..., 23). |
| Post weekday | Weekday the post was published (Sunday, Monday, ..., Saturday). |
| Paid | If the company paid to Facebook for advertising (yes, no). |

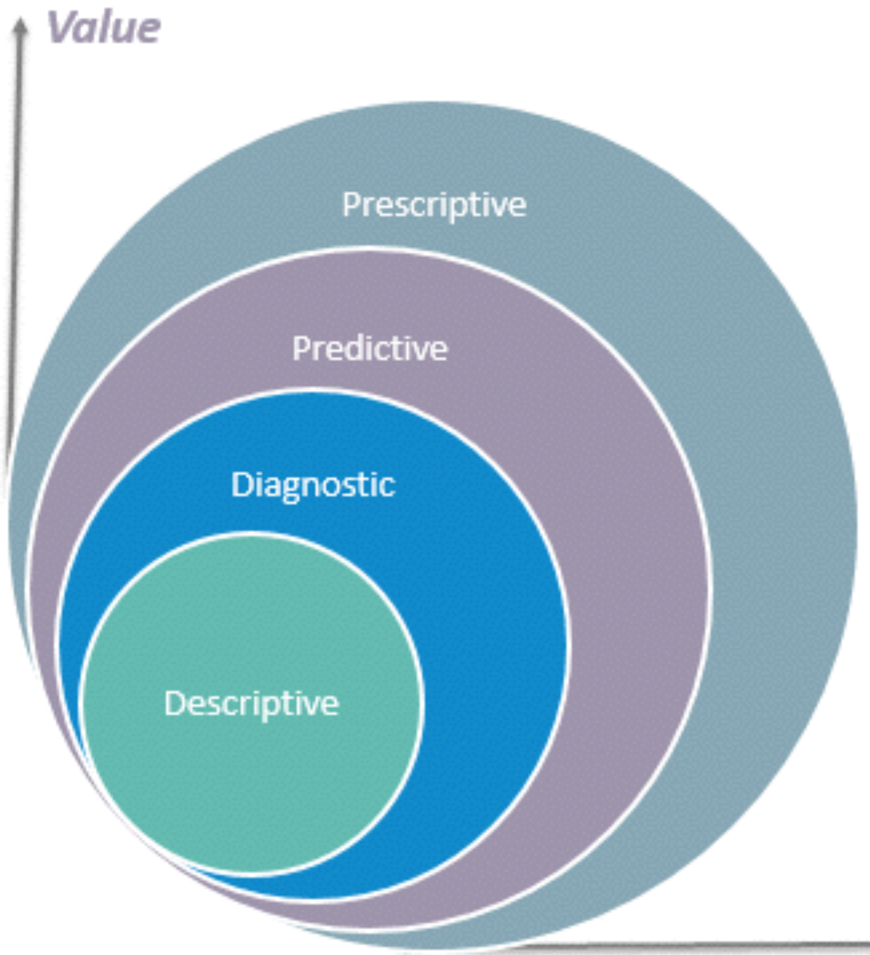
# DATA

- **LET'S TAKE A LOOK AT THE DATA IN R**

# PREDICTIVE ANALYTICS

# WHAT IS PREDICTIVE ANALYTICS?

## 4 types of Data Analytics




### What is the data telling you?

**Descriptive:** ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

**Diagnostic:** ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

**Predictive:** ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

**Prescriptive:** ***What do I need to do?***

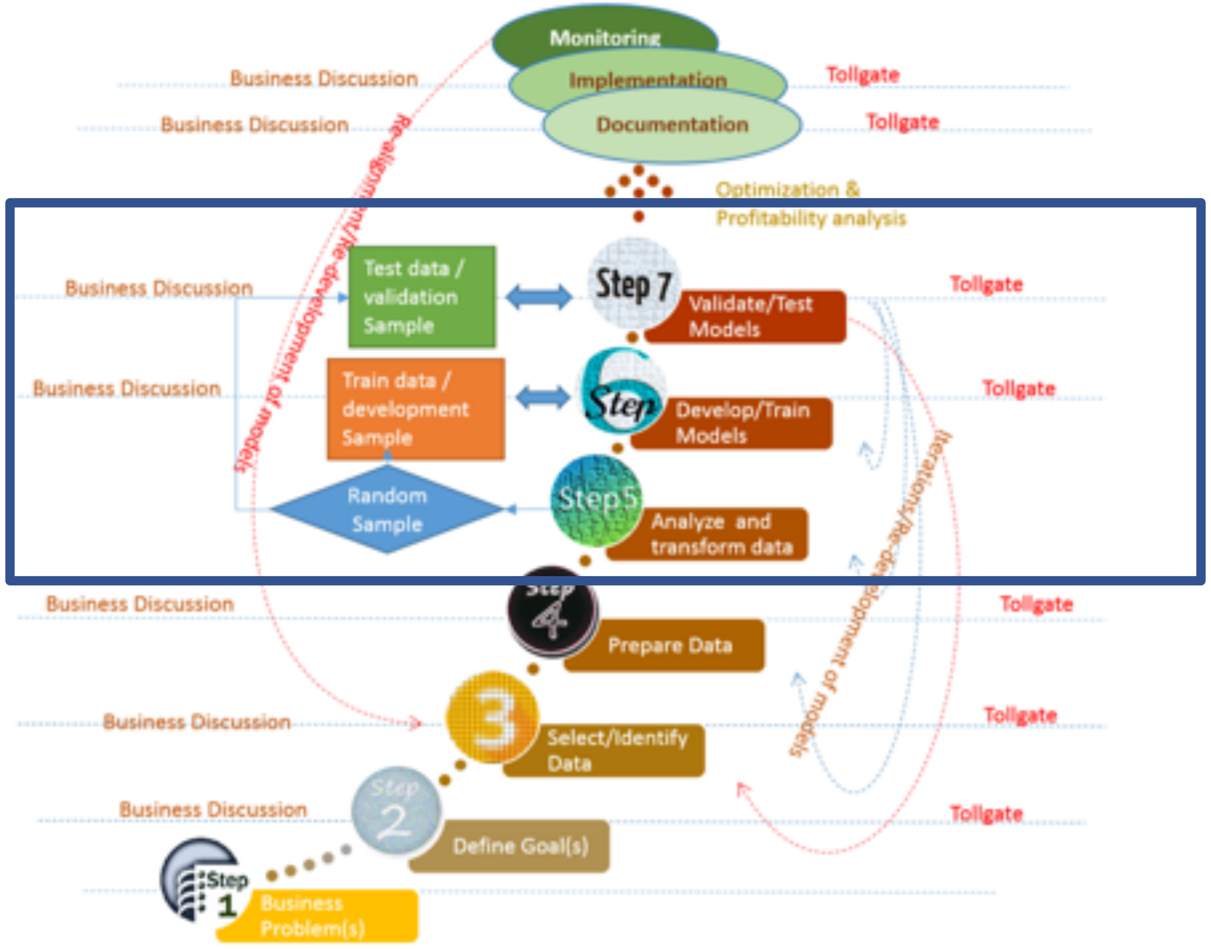
- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

*Complexity*

Principa
www.principa.co.za

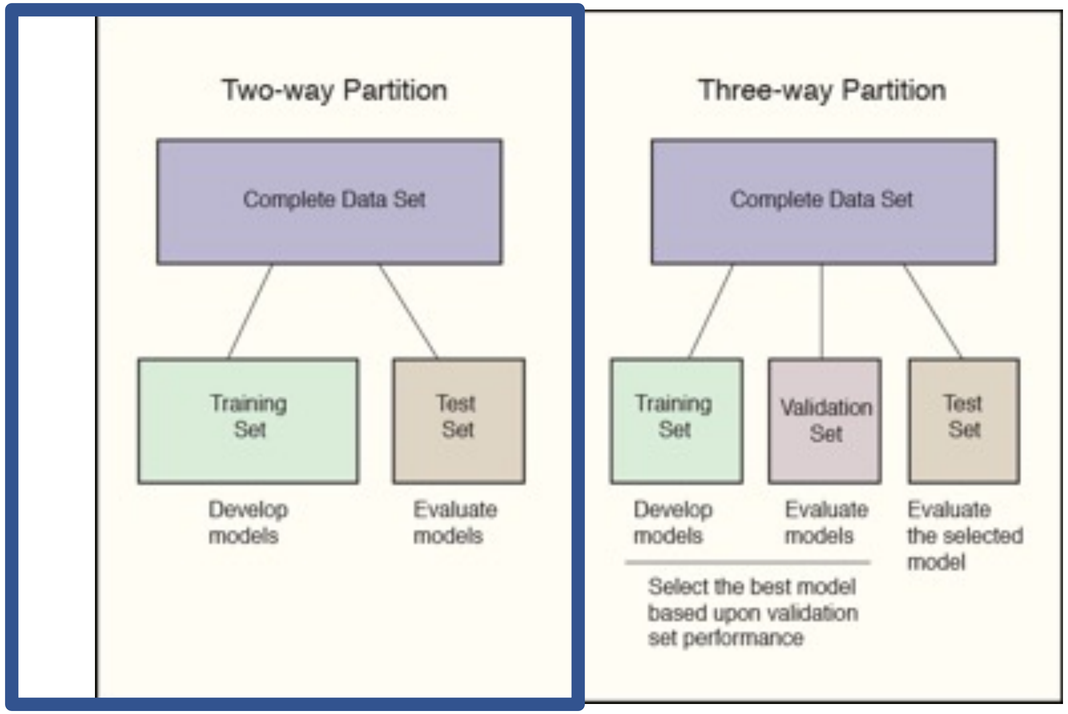
- Source: https://insights.principa.co.za/4-types-of-data-analytics-descriptive-diagnostic-predictive-prescriptive

# GENERAL PROCESS




- Source: *Ariful Mondal; https://rstudio-pubs-static.s3.amazonaws.com/223423_8ca6fccca1e44939be3f85ecbfa9598f.html*

# MODEL BUILDING: PARTITIONING AND EVALUATION (PREDICTION)



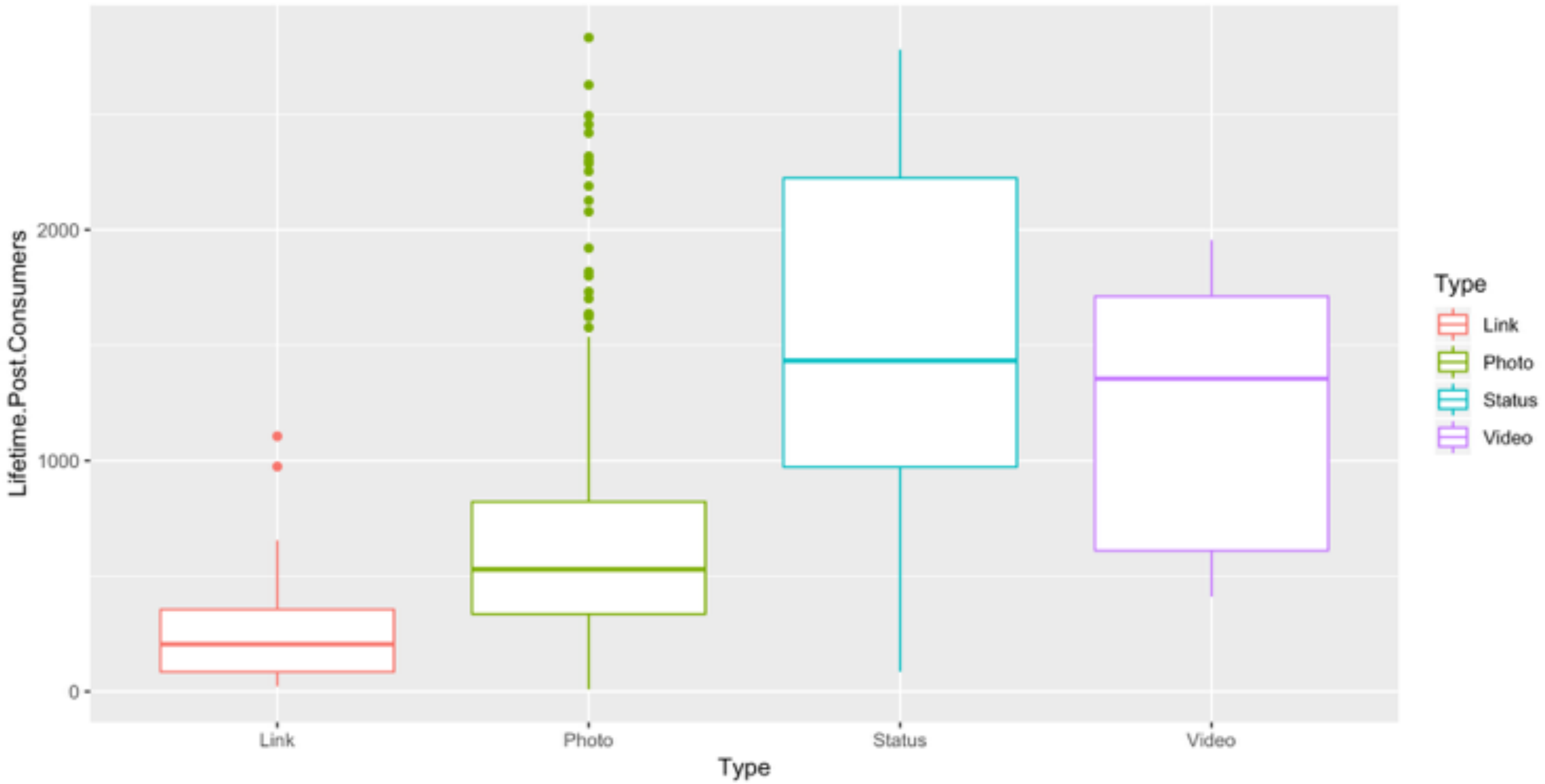

- Source: *Modeling Techniques in Predictive Analytics with Python and R: A Guide to Data Science*, Thomas W. Miller

# QUICK EDA / “PREPARE DATA”

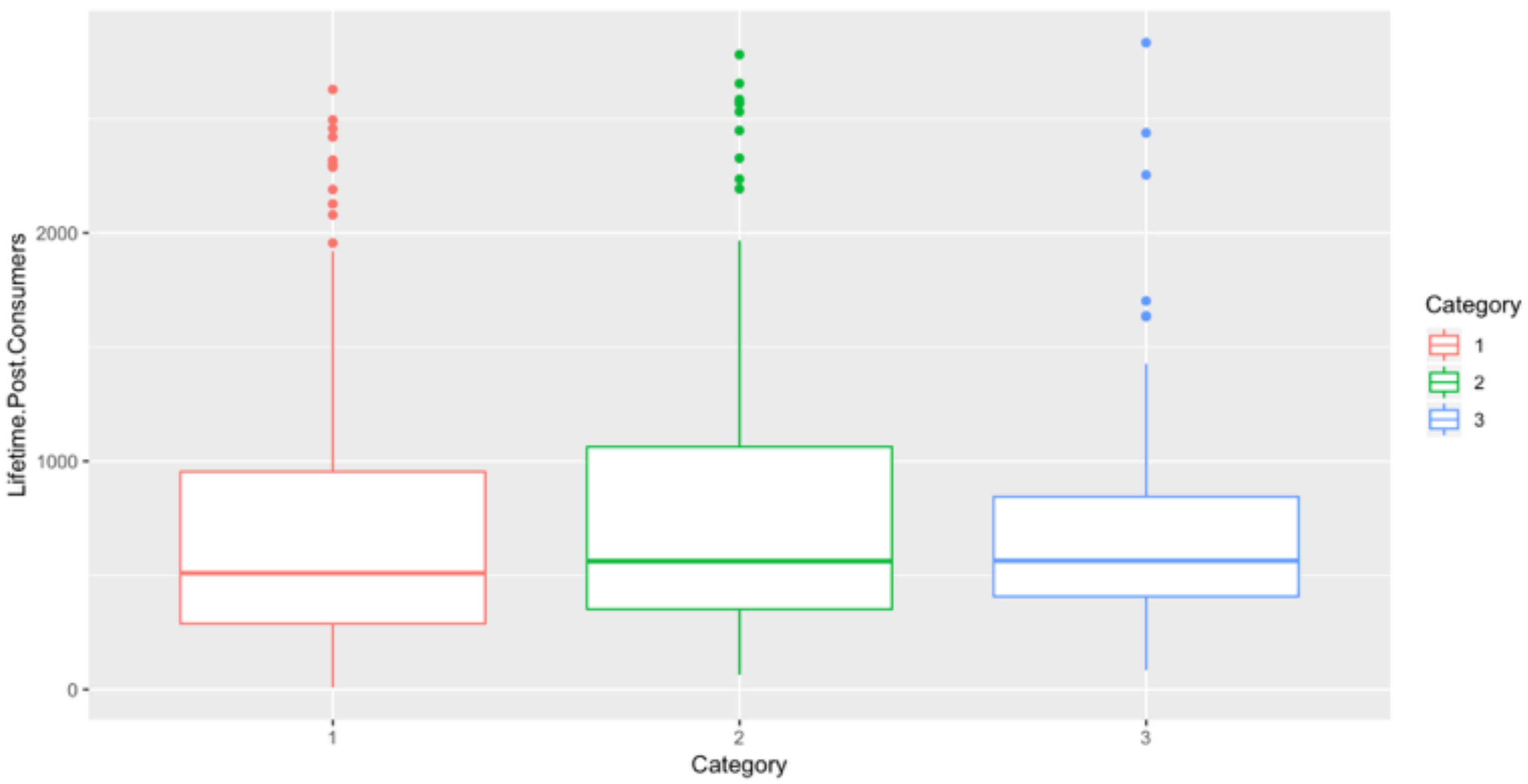
# EDA

- REMEMBER OUR TARGET IS LIFETIME CONSUMERS
- EXCLUDED 11 RECORDS BASED ON TARGET -> OVER 3,000 LIFETIME POST CONSUMERS)
- LET’S DO SOME EDA IN R

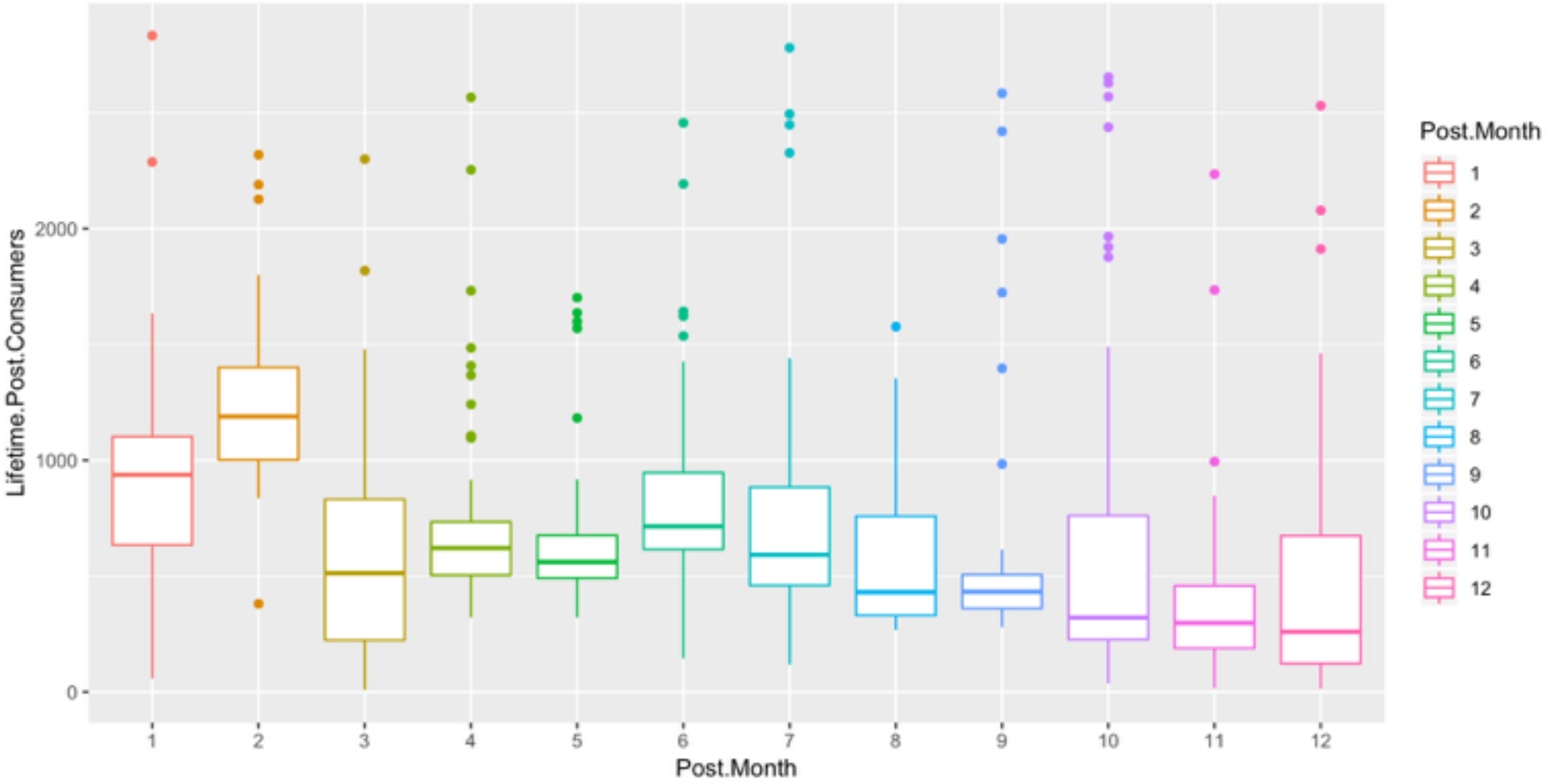
# EDA: TARGET AND FEATURES - TYPE



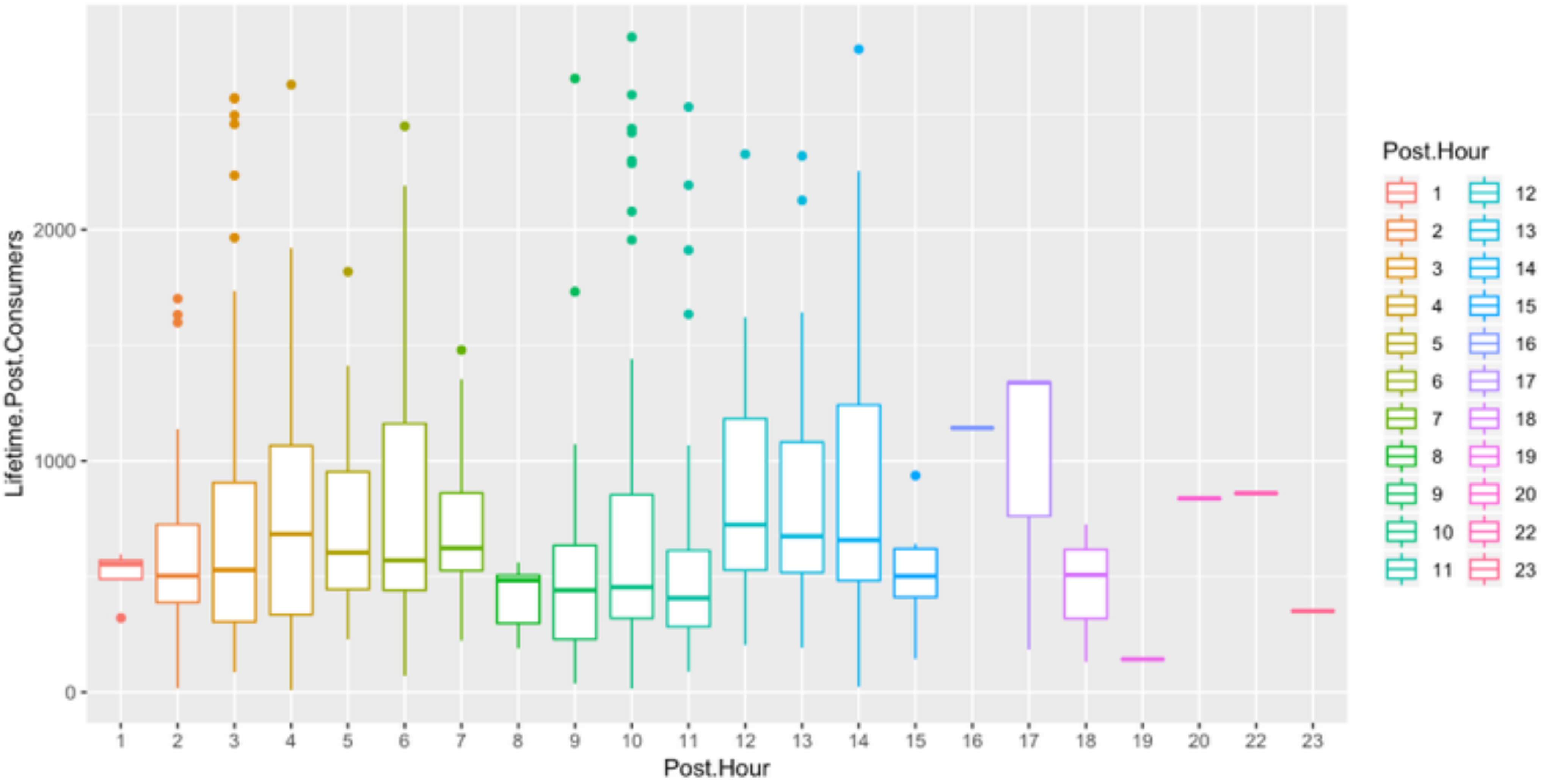
# EDA: TARGET AND FEATURES - CATEGORY



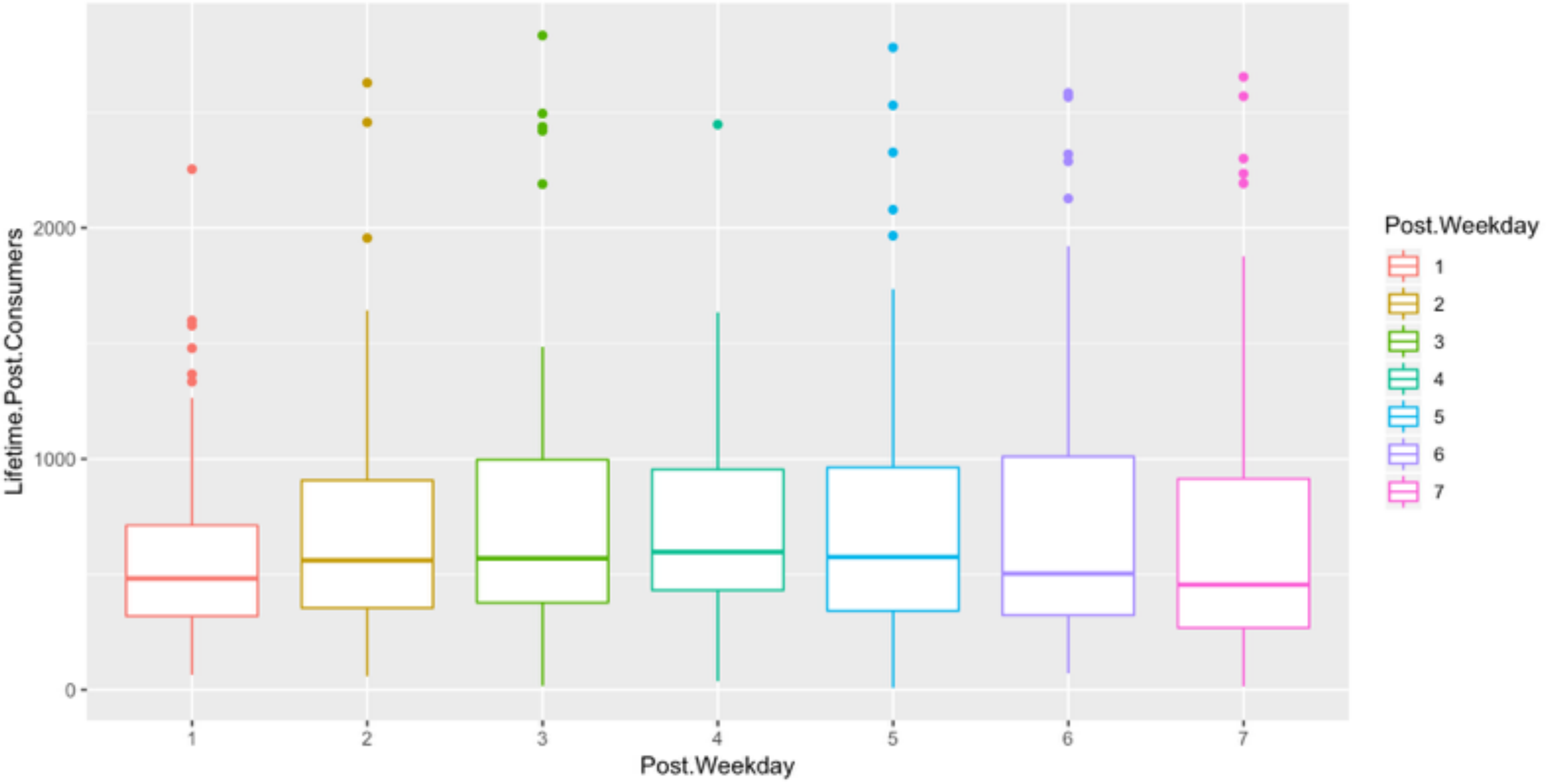
# EDA: TARGET AND FEATURES - MONTH



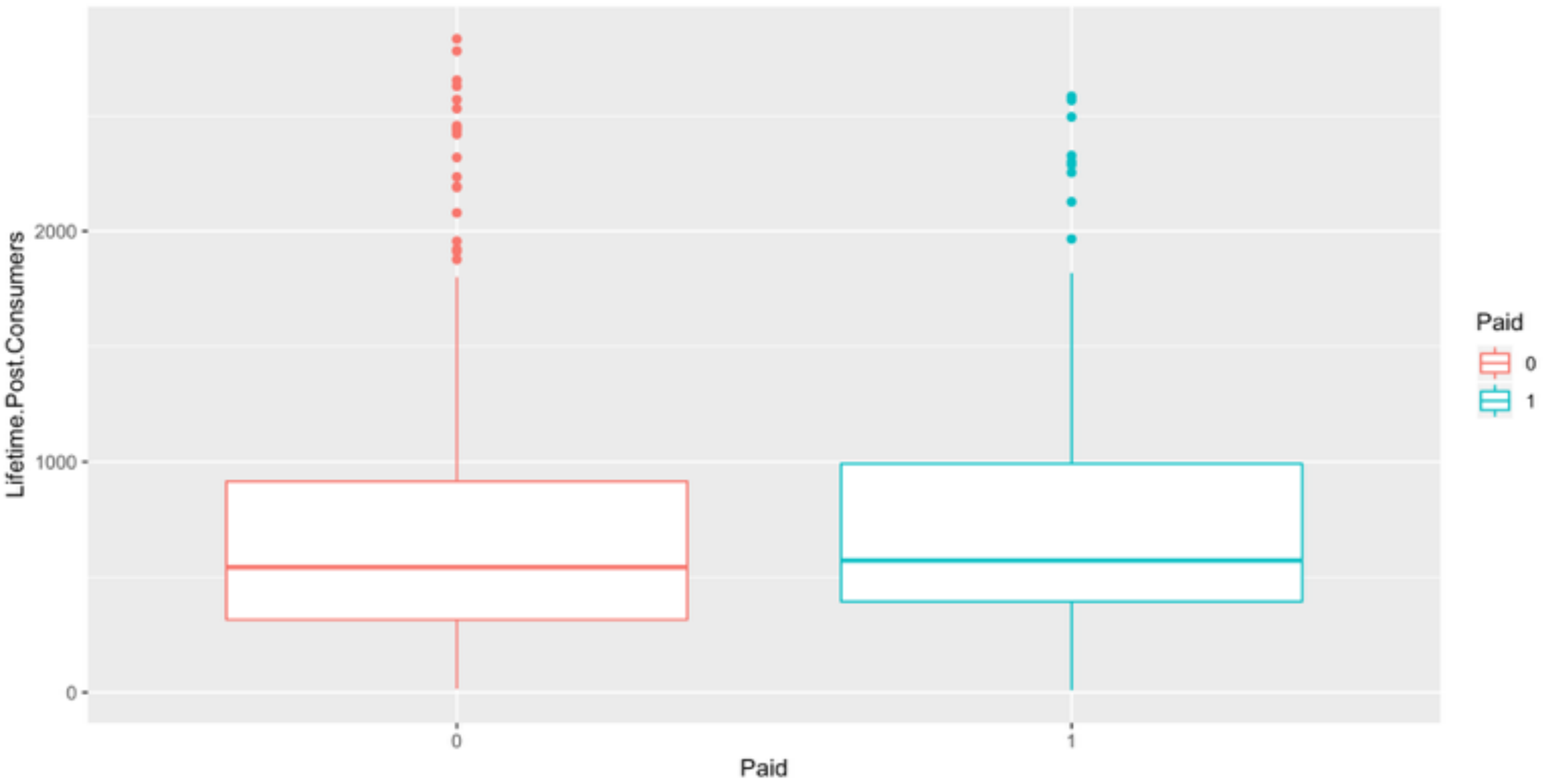
# EDA: TARGET AND FEATURES - HOUR



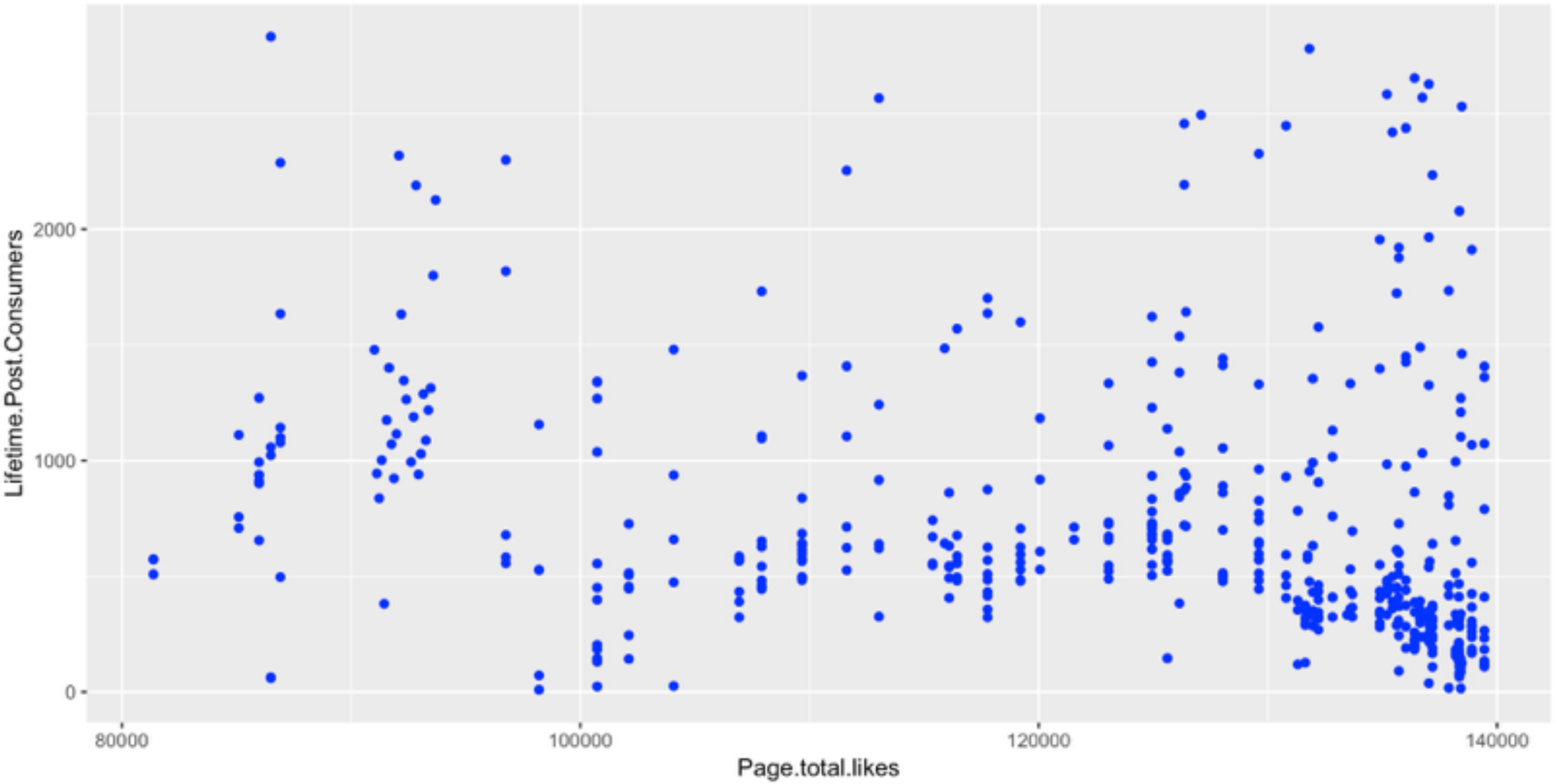
# EDA: TARGET AND FEATURES - WEEKDAY



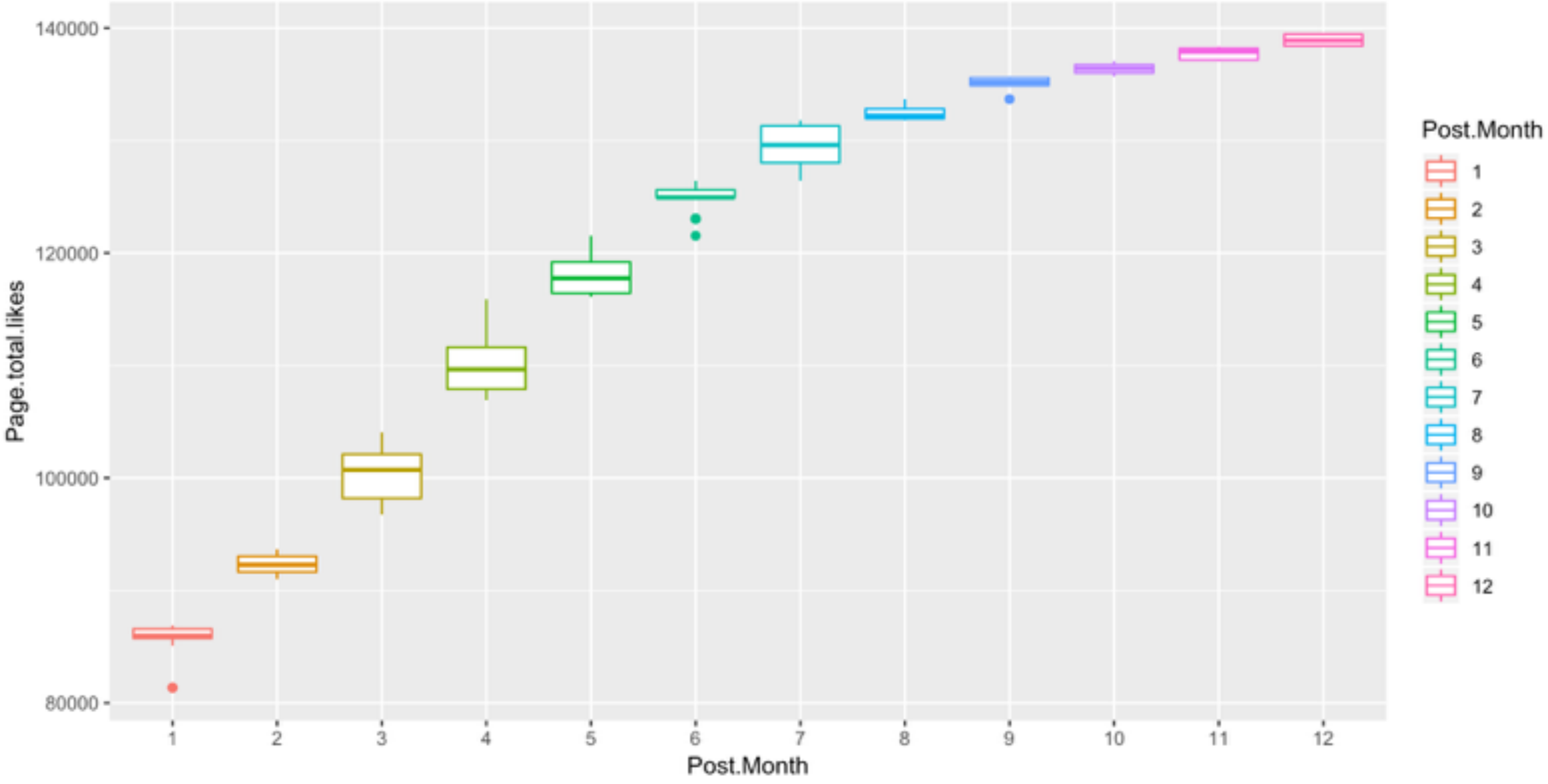
# EDA: TARGET AND FEATURES – PAID



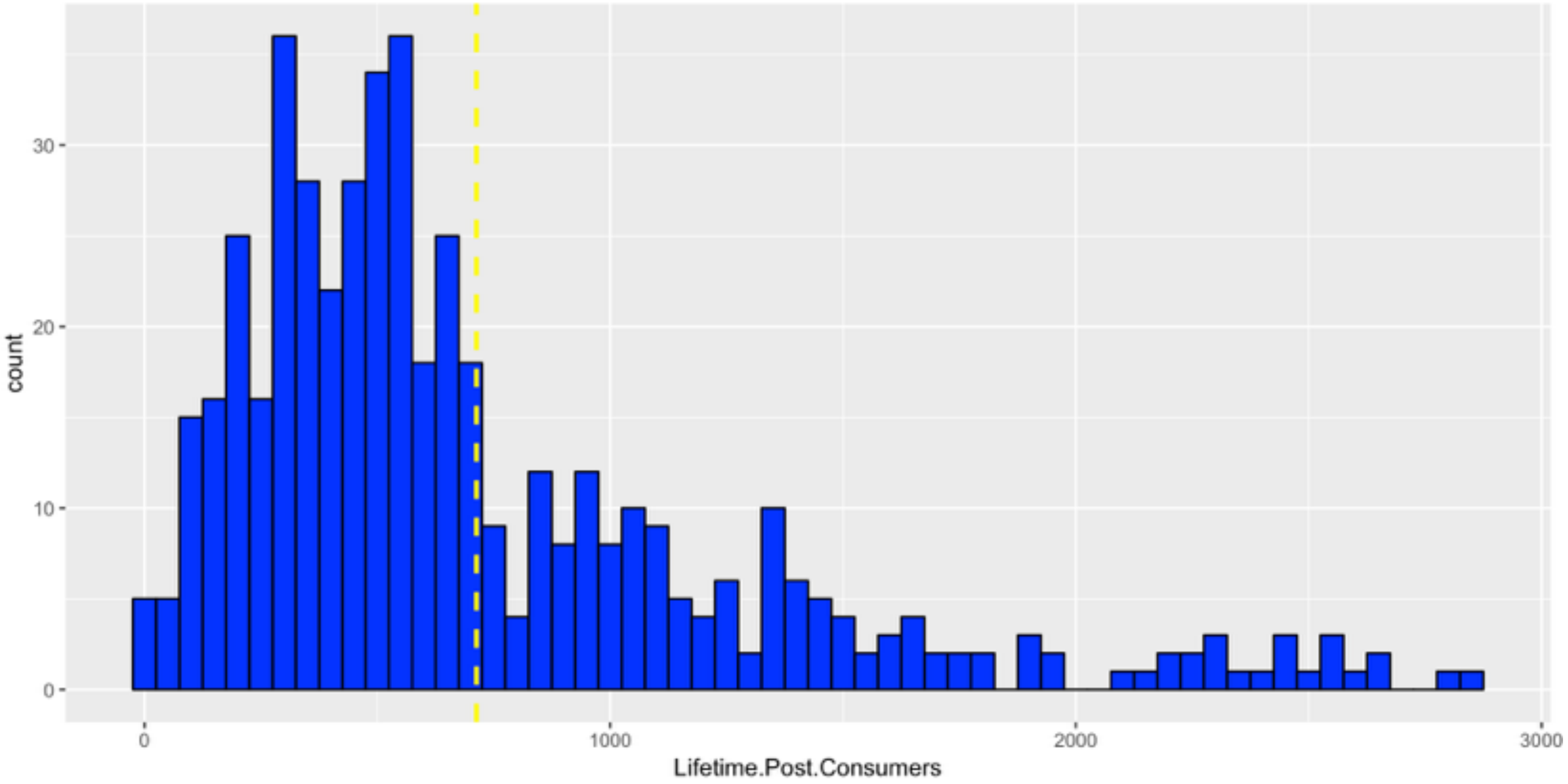
# EDA: TARGET AND FEATURES – PAGE TOTAL LIKES



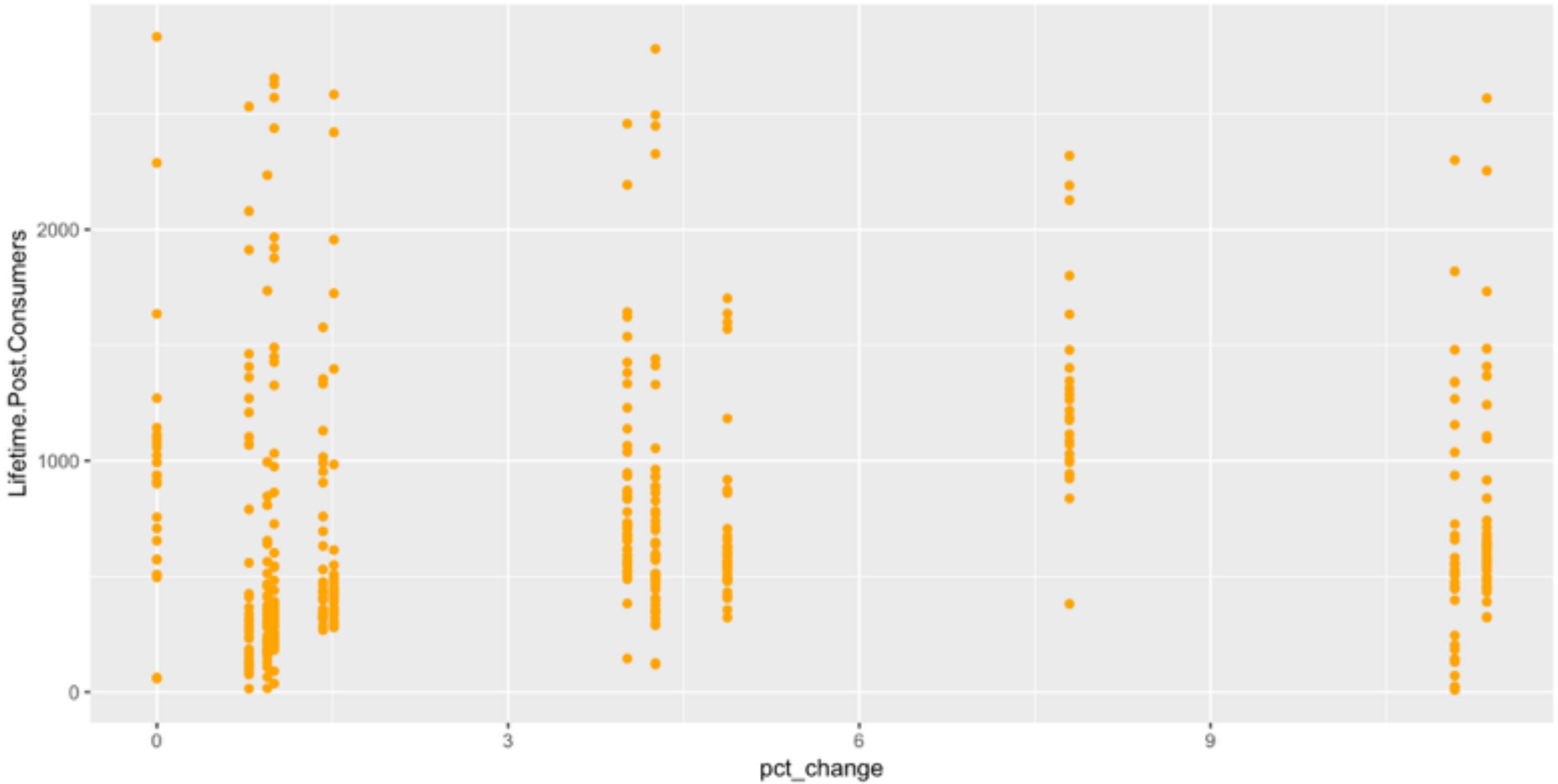
# EDA: FEATURES – MONTH AND PAGE TOTAL LIKES



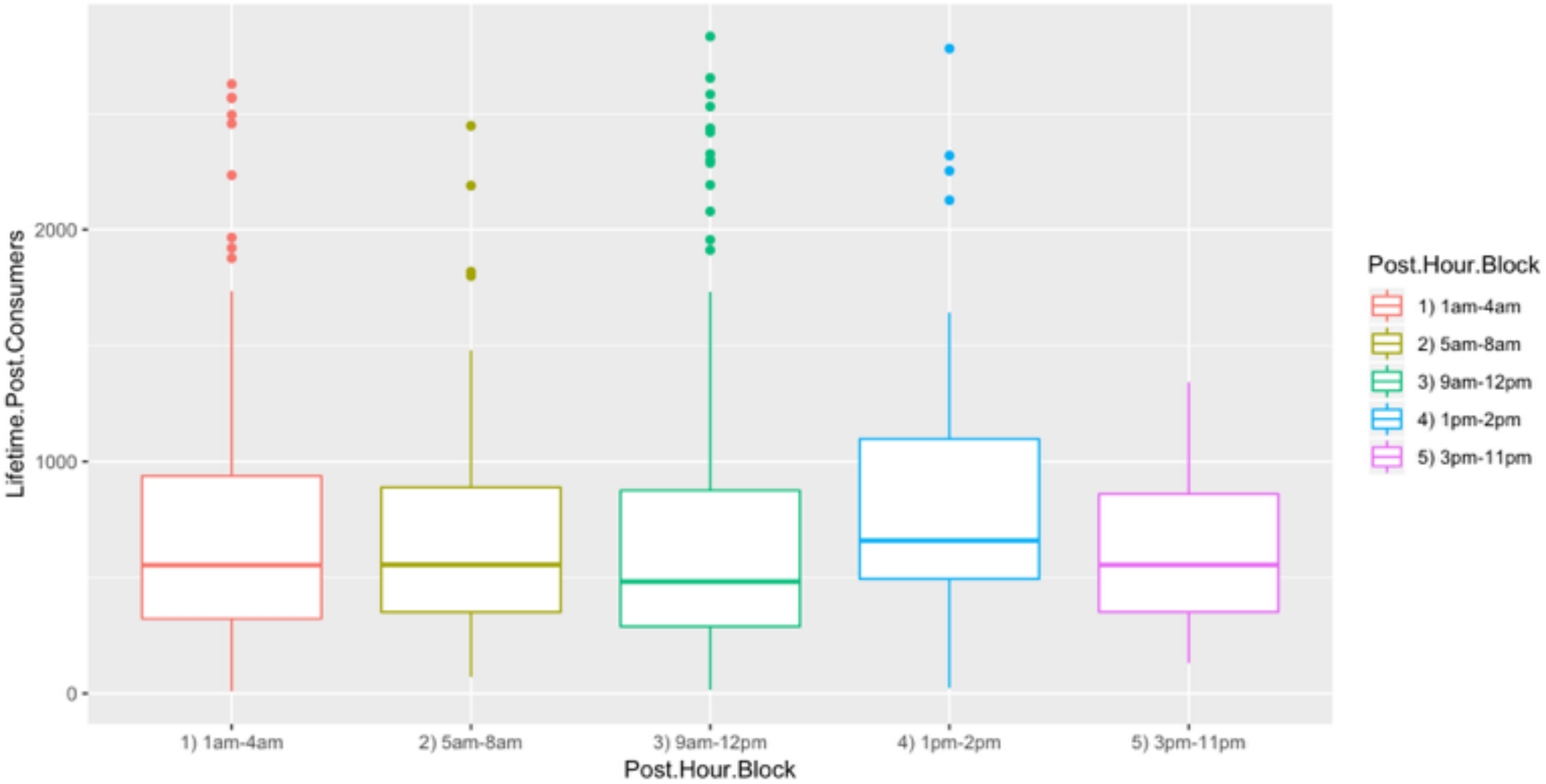
# EDA: TARGET



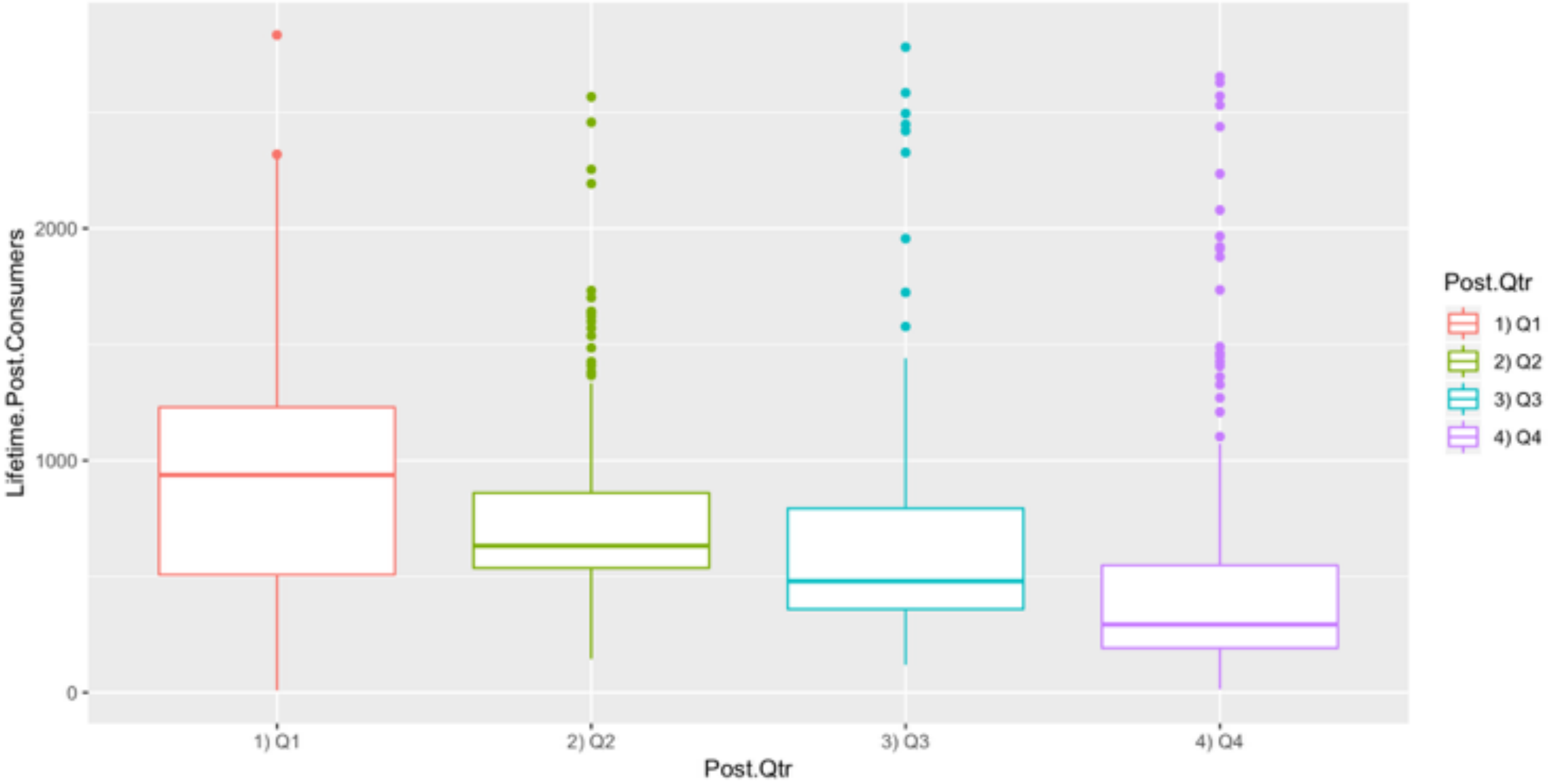
# EDA: ADDED FEATURES – MONTHLY PERCENT CHANGE IN PAGE LIKES



# EDA: POST HOUR BLOCK



# EDA: POST QTR


Lifetime.Post.Consumers
2000
1000
0
1) Q1
2) Q2
3) Q3
4) Q4
Post.Qtr
Post.Qtr
1) Q1
2) Q2
3) Q3
4) Q4

# MODELLING / “DEVELOP/TRAIN MODELS”

# DATA USED FOR MODELLING

- FINAL DATA SET:
  - 484 POSTS
  - TARGET: LIFETIME CONSUMERS
  - FEATURES: 10 VARIABLES, 3 OF WHICH WERE CREATED FROM THE ORIGINAL 7
- TRAINING/TEST SPLIT: 80% TRAINING/20% TEST
- LET’S GO BACK INTO R AND DO SOME MODELLING

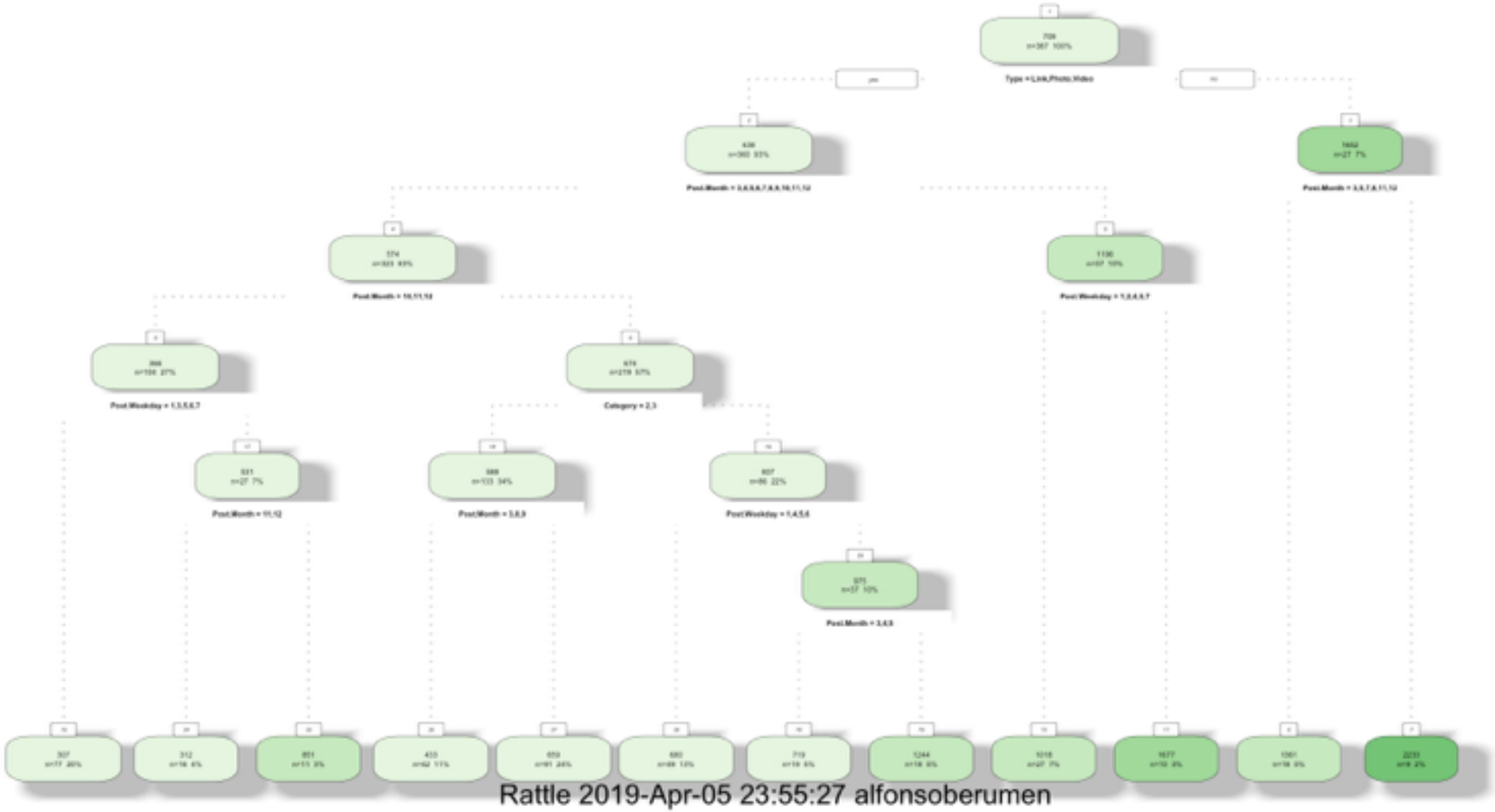
# MULTIPLE LINEAR REGRESSION: RESULTS

```
reg_model4 <- lm(Lifetime.Post.Consumers~
              Category+
              Type+
              Post.Month+
              Post.Weekday+
              Post.Hour.Block+
              Paid,
             data = facebook.train)
summary(reg_model4)
##
## Call:
## lm(formula = Lifetime.Post.Consumers ~ Category + Type + Post.Month +
##     Post.Weekday + Post.Hour.Block + Paid, data = facebook.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1715.97  -225.78   -80.89    96.28  1999.91
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                838.44     174.76   4.798 2.36e-06 ***
## Category2                 -149.49      70.02  -2.135 0.033443 *
## Category3                 -218.21      60.99  -3.577 0.000394 ***
## TypePhoto                  388.38     133.98   2.899 0.003975 **
## TypeStatus                1574.36     163.49   9.630  < 2e-16 ***
## TypeVideo                  920.49     268.19   3.432 0.000668 ***
## Post.Month2                134.92     153.23   0.880 0.379182
## Post.Month3               -561.42     148.12  -3.790 0.000176 ***
## Post.Month4               -397.12     134.11  -2.961 0.003269 **
## Post.Month5               -546.38     139.77  -3.909 0.000111 ***
## Post.Month6               -357.45     130.62  -2.737 0.006518 **
## Post.Month7               -456.91     132.95  -3.437 0.000658 ***
## Post.Month8               -595.97     143.07  -4.165 3.90e-05 ***
## Post.Month9               -604.41     137.64  -4.391 1.48e-05 ***
## Post.Month10              -652.35     130.37  -5.004 8.82e-07 ***
## Post.Month11              -969.46     136.95  -7.079 7.70e-12 ***
## Post.Month12              -811.06     138.16  -5.870 9.91e-09 ***
## Post.Weekday2               53.63      87.88   0.610 0.542101
## Post.Weekday3               87.00      95.54   0.911 0.363116
## Post.Weekday4               45.51      89.17   0.510 0.610048
## Post.Weekday5              -24.89      90.64  -0.275 0.783755
## Post.Weekday6              100.71      87.95   1.145 0.252923
## Post.Weekday7               56.25      84.42   0.666 0.505661
## Post.Hour.Block2) 5am-8am  -64.59      85.56  -0.755 0.450821
## Post.Hour.Block3) 9am-12pm -27.68      56.02  -0.494 0.621563
## Post.Hour.Block4) 1pm-2pm   23.13      81.70   0.283 0.777267
## Post.Hour.Block5) 3pm-11pm -194.21    143.81  -1.350 0.177708
## Paid1                       70.37      53.38   1.318 0.188285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 452.8 on 359 degrees of freedom
## Multiple R-squared:  0.4343, Adjusted R-squared:  0.3917
## F-statistic: 10.21 on 27 and 359 DF,  p-value: < 2.2e-16
#PRETTY CLOSE FIT TO FIRST MODEL
#AND->DON'T HAVE TO DEAL WITH HOUR COUNT ISSUE
```

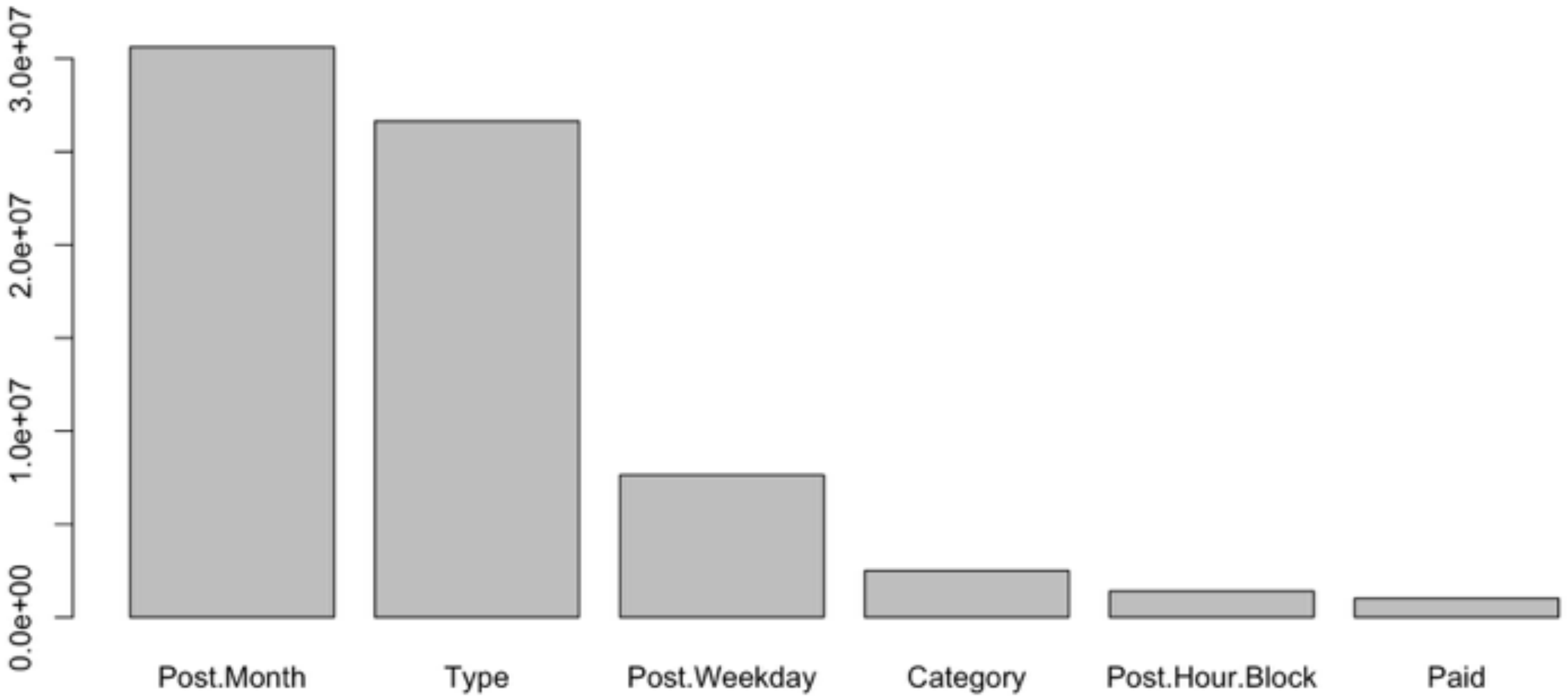
**Results**

| | *Dependent variable:* |
|---|---|
| | Lifetime.Post.Consumers |
| Category2 | -149.485** |
| | (70.020) |
| Category3 | -218.208*** |
| | (60.995) |
| TypePhoto | 388.382*** |
| | (133.975) |
| TypeStatus | 1,574.365*** |
| | (163.488) |
| TypeVideo | 920.485*** |
| | (268.190) |
| Post.Month2 | 134.918 |
| | (153.231) |
| Post.Month3 | -561.424*** |
| | (148.124) |
| Post.Month4 | -397.117*** |
| | (134.111) |
| Post.Month5 | -546.378*** |
| | (139.774) |
| Post.Month6 | -357.450*** |
| | (130.621) |
| Post.Month7 | -456.912*** |
| | (132.951) |
| Post.Month8 | -595.967*** |
| | (143.073) |
| Post.Month9 | -604.411*** |
| | (137.637) |
| Post.Month10 | -652.354*** |
| | (130.375) |
| Post.Month11 | -969.463*** |
| | (136.955) |
| Post.Month12 | -811.056*** |
| | (138.165) |
| Post.Weekday2 | 53.627 |
| | (87.882) |
| Post.Weekday3 | 87.001 |
| | (95.542) |
| Post.Weekday4 | 45.515 |
| | (89.165) |
| Post.Weekday5 | -24.893 |
| | (90.640) |
| Post.Weekday6 | 100.709 |
| | (87.947) |
| Post.Weekday7 | 56.248 |
| | (84.422) |
| Post.Hour.Block2) 5am-8am | -64.586 |
| | (85.558) |
| Post.Hour.Block3) 9am-12pm | -27.680 |
| | (56.024) |
| Post.Hour.Block4) 1pm-2pm | 23.129 |
| | (81.699) |
| Post.Hour.Block5) 3pm-11pm | -194.211 |
| | (143.807) |
| Paid1 | 70.370 |
| | (53.384) |
| Constant | 838.435*** |
| | (174.761) |
| Observations | 387 |
| $R^2$ | 0.434 |
| Adjusted $R^2$ | 0.392 |
| Residual Std. Error | 452.842 (df = 359) |
| F Statistic | 10.207*** (df = 27; 359) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

# TREES: VISUALS



# TREES: VISUALS – VARIABLE IMPORTANCE



ERROR

# ERROR: NUMERIC

- **Residuals**: $(y_i - \hat{y}_i)$ or ***observed minus prediction***

- **Mean Squared Error** (MSE): square the residuals, sum, then divide by n or ***average of*** $(\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i})^2$

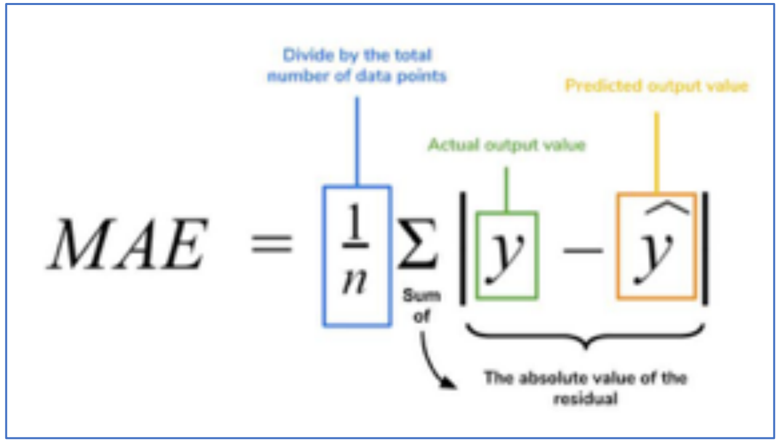
$$MSE = \frac{1}{n} \Sigma \underbrace{\left( y - \hat{y} \right)^2}_{\text{The square of the difference between actual and predicted}}$$

- **Root Mean Squared Error** (RMSE): square-root of MSE or **how far, on average, the residuals are from zero** or **average distance between the observed values and the predictions in the same units as our original y**

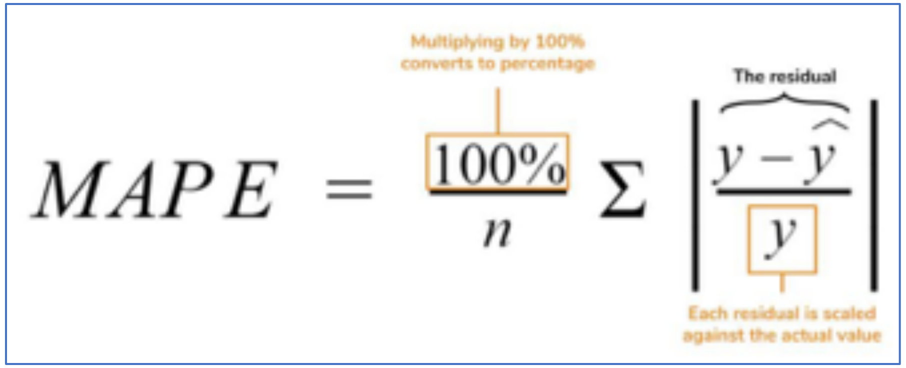
$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y - \hat{y})^2}\,.$$

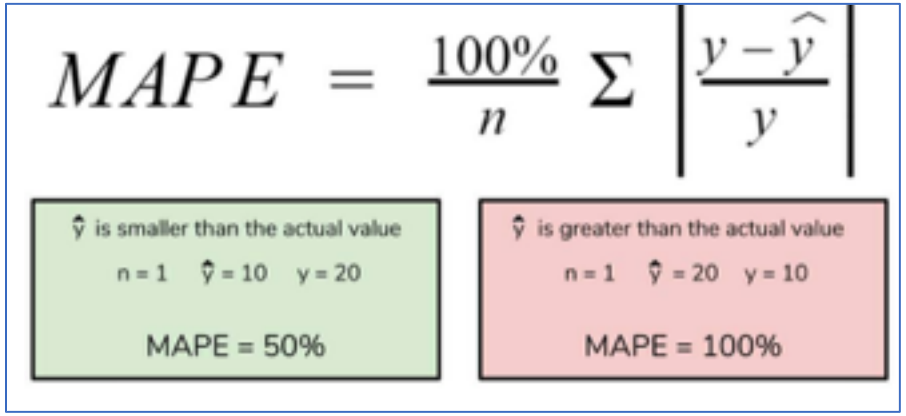
# ERROR: NUMERIC (CONTINUED)

- **Mean Absolute Error** (MAE): **the average of the absolute value of residual=**$(y_i - \hat{y}_i)$



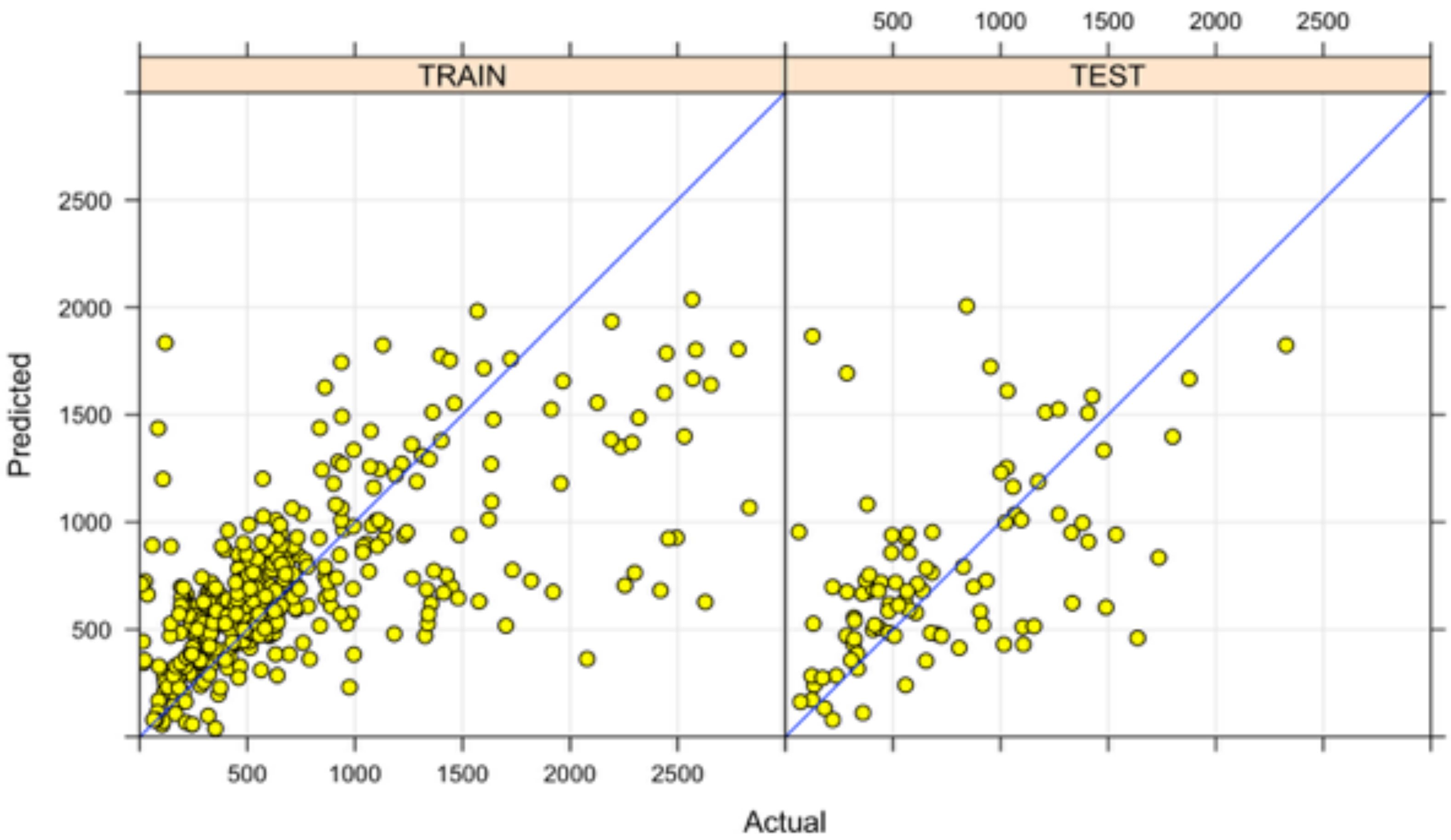

- **Mean Absolute Percentage Error** (MAPE): **accuracy as a %**



# PREDICTION ERROR - VISUAL

- **LET'S TAKE A DEEPER LOOK AT THE ERROR FROM OUR REGRESSION MODEL IN R**



# SUMMARY: COMPARE / “VALIDATE/TEST MODELS”

# SUMMARY OF PREDICTION ERROR

## ML REGRESSION

### TRAINING

```
#error summary
regression_results_training<-data.frame(regression_training_MSE,
                                        regression_training_RMSE,
                                        regression_training_MAE,
                                        regression_training_MAPE)
regression_results_training
##   regression_training_MSE regression_training_RMSE regression_training_MAE
## 1                190228.8                 436.1522                289.2652
##   regression_training_MAPE
## 1                 1.085442
```

### TEST

```
#error summary
regression_results_test<-data.frame(regression_test_MSE,
                                    regression_test_RMSE,
                                    regression_test_MAE,
                                    regression_test_MAPE)
regression_results_test
##   regression_test_MSE regression_test_RMSE regression_test_MAE
## 1            192217.5             438.4262            310.9937
##   regression_test_MAPE
## 1            0.7992681
```

## TREE

### TRAINING

```
#error summary
tree_results_training<-data.frame(tree_training_MSE,
                                  tree_training_RMSE,
                                  tree_training_MAE,
                                  tree_training_MAPE)
tree_results_training
##   tree_training_MSE tree_training_RMSE tree_training_MAE
## 1          170175.2            412.523          268.2479
##   tree_training_MAPE
## 1           1.070785
```

### TEST

```
#error summary
treee_results_test<-data.frame(tree_test_MSE,
                               tree_test_RMSE,
                               tree_test_MAE,
                               tree_test_MAPE)
treee_results_test
##   tree_test_MSE tree_test_RMSE tree_test_MAE tree_test_MAPE
## 1      264986.6       514.7685       358.415      0.8612332
```

# WHAT WAS PUBLISHED?

# USING SVM - MAPE: 27.2%

**Table 5**
Results for performance metrics predictions

| Performance metric | Mean absolute percentage error | Source of metric |
| --- | --- | --- |
| Lifetime people who have liked your page and engaged with your post | 26.9 | |
| Lifetime post consumers | 27.2 | |
| Lifetime engaged users | 28.8 | |
| Lifetime post consumptions | 33.1 | Interactions |
| Shares | 35.8 | |
| Lifetime post reach by people who like your page | 37.5 | Visualizations |
| Likes | 41.2 | Interactions |
| Lifetime post impressions by people who have liked your page | 47.8 | Visualizations |
| Lifetime post total reach | 49.6 | |
| Comments | 63.9 | Interactions |
| Lifetime post total impressions | 69.3 | Visualizations |

# IMPORTANT INPUTS

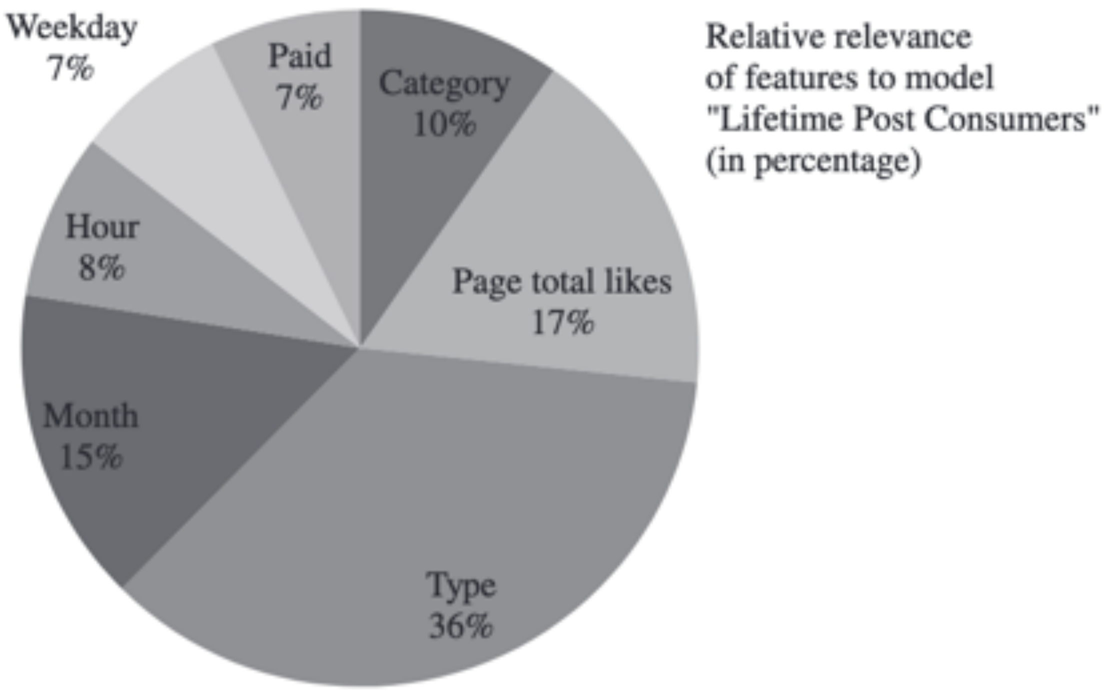



**Fig. 6.** Relevance of the input features for "Lifetime Post Consumers."

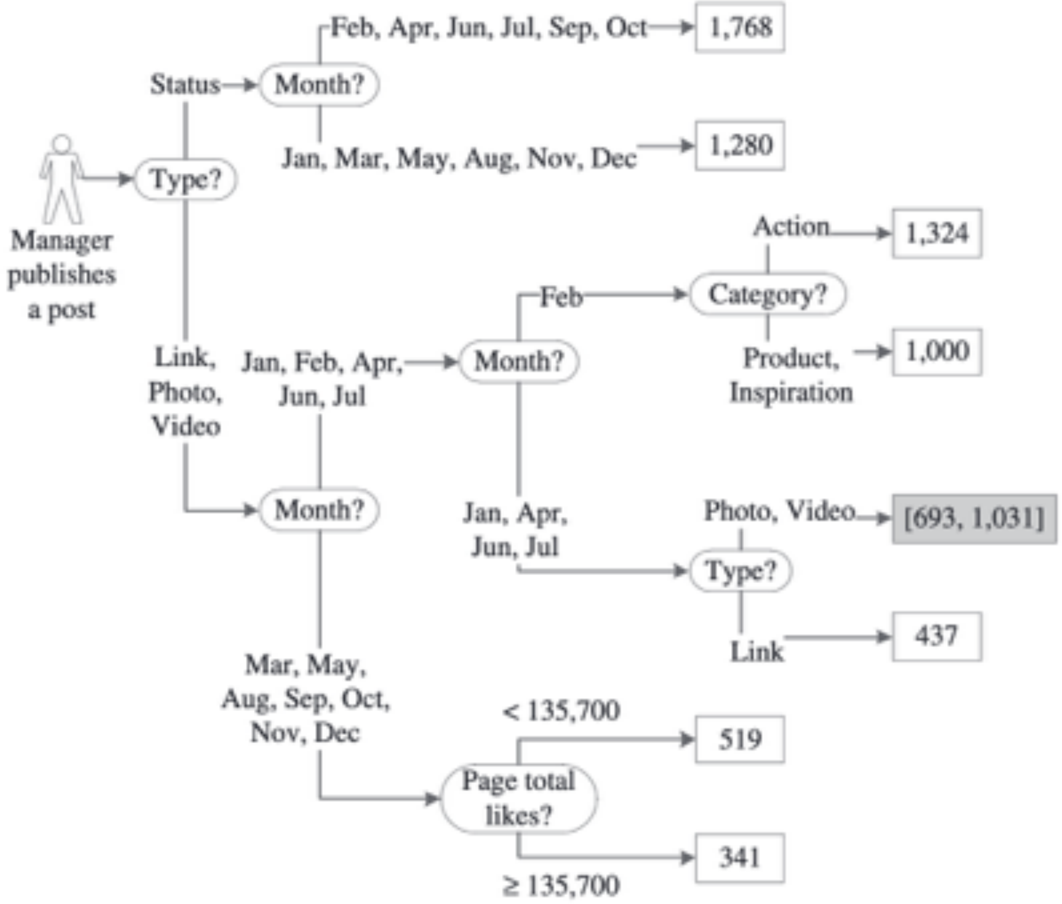



**Fig. 14.** Rules extracted from the support vector machine model.

# VALUE/IMPLICATIONS

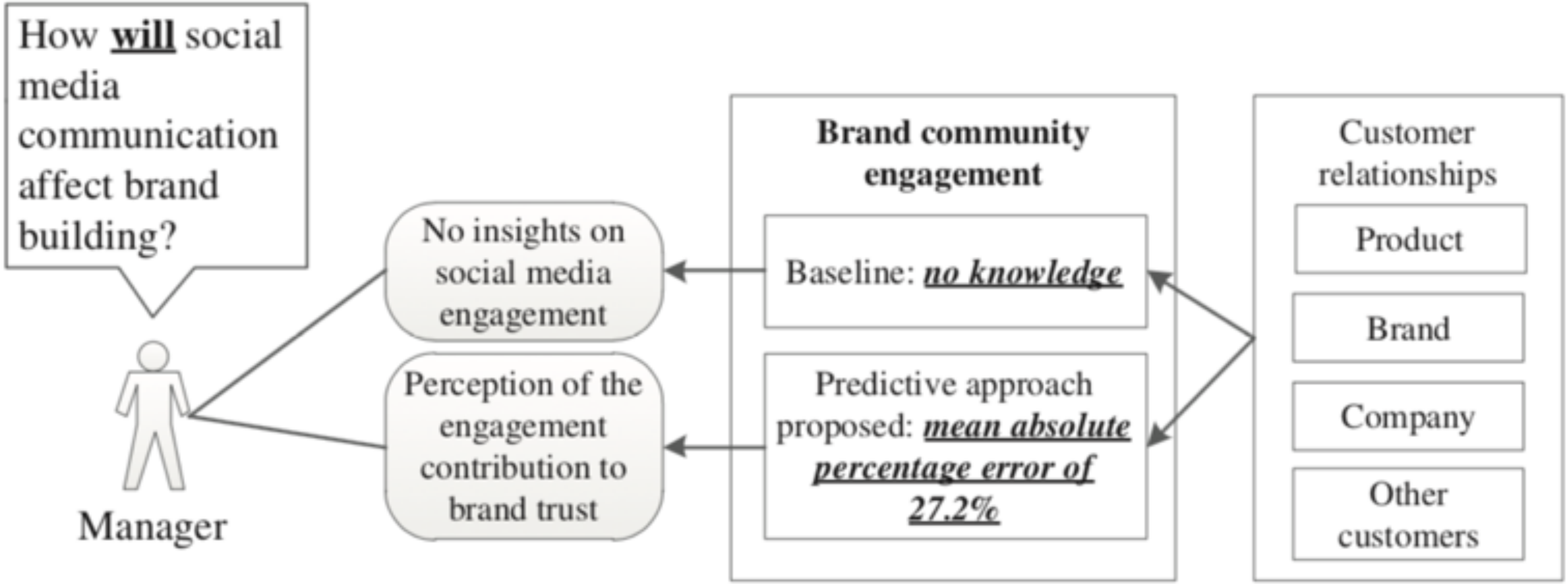



**Fig. 5.** Application of the model for "Lifetime Post Consumers" (adapted from Habibi et al., 2014).

# THANK YOU

# QUESTIONS?