Week 3 | Examination assignments

Exploratory data analysis (EDA)

Erik Vanhatalo

2025-03-16

Contents

Regarding the Assignments

This document outlines the **individual** examination assignments for this week. We strongly recommend that you work through the training assignments for this week before progressing to the examination assignments.

Prepare and document solution before the examination session

You need to prepare your solutions before the examination session. At the examination session you will need to have your imported datafiles and prepared solutions easily available so that you will be able to perform the tasks you are given at the session.

Quarto - recommended way of documenting your solutions

There are some options for you to prepare, document and save your solutions. You can do it in a standard R script (.R file). which is easy and straightforward. However, we strongly recommend that you use Quarto because it makes it possible for you to combine executable code with your own notes. You can also create *MS Word*, *pdf* or *html* documents with your solutions that you can use as you prepare. During Week 1 there will be an introductory lecture featuring Quarto. Quarto is integrated in RStudio. **Whatever choice you make it is important that you are able to run your prepared solutions fast and easy at the examination sessions.**

1. Examination assignment 1

You will be working with the MS Excel dataset: **"dataset_Facebook".** These data are related to Facebook posts published during the year 2014 on the Facebook page of a renowned cosmetics brand. The dataset contains 500 of the 790 rows that were part of the original paper by Moro et al. Please refer to the description of the dataset from the UCI Machine Learning Repository and to the full paper by (Moro, Rita, and Vala 2016), for explanations of the variables/features. The paper also provides a background to the dataset, which is valuable for understanding. Herein, we provide two important tables from the paper (Table 2 and Table 3), see below.

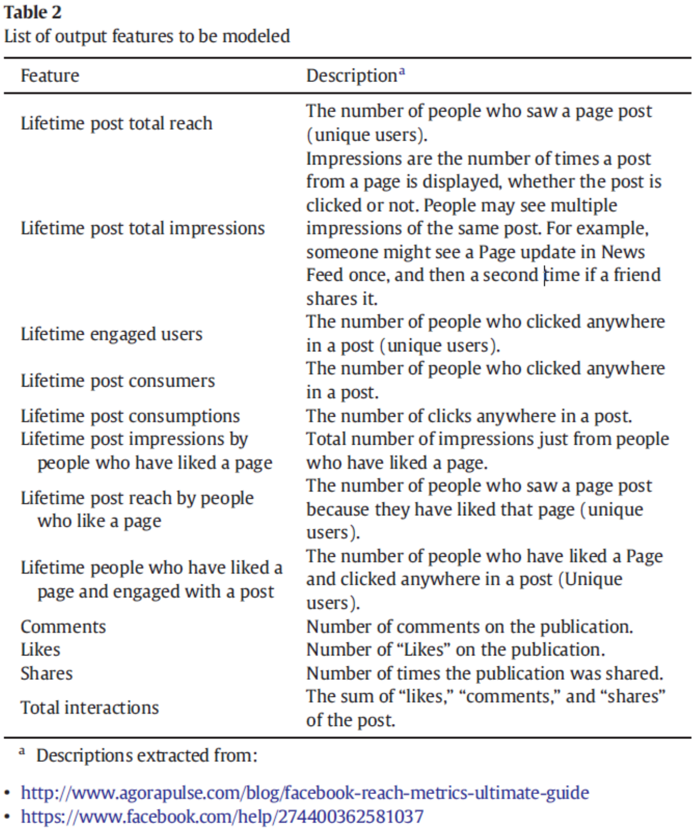

Table 2
List of output features to be modeled

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

Figure 1: Table 2 from Moro et al. (2016)

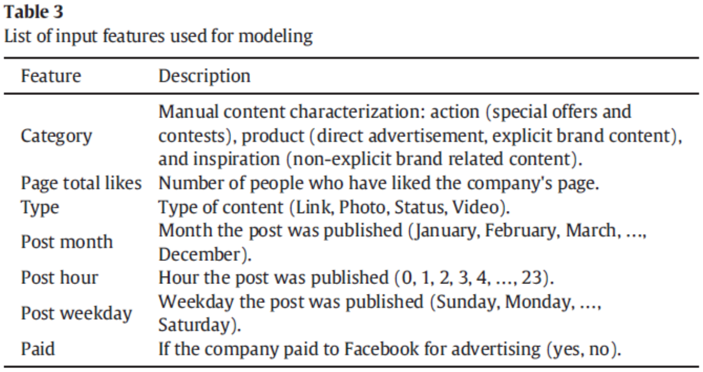

Table 3
List of input features used for modeling

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).

Figure 2: Table 3 from Moro et al. (2016)

1.1 List of tasks

Your initial task here is to wrangle the data and prepare it for exploratory data analysis. **Background:** There are two potential problems that you should focus on.

1. The variable names are long and include spaces, which makes it cumbersome to code in R.
2. Some of the categorical variables are given in numbers in the dataset.
    - 'Category', 'Post Month', 'Post Weekday', and 'Paid'. You may benefit from having them in categories that can be understood as well. Not least when producing plots.
    - The Category variable:
        - 1 = action (special offers and contests)
        - 2 = product (direct advertisement, explicit brand content)
        - 3 = inspiration (non-explicit brand related content)
    - Post Month:
        - 1 = January, …12 = December
    - Post Weekday variable:
        - 1 = Sunday, 2 = Monday, …, 7 = Saturday.
    - Paid:
        - 1 = Paid, 0 = Not paid.

Complete the following tasks or answer the following questions:

1. Import the dataset into R. Look over the class of all variables/features. Change class if needed. See the Moro et al. (2016) paper for guidance in this work.
2. Change variable names into concise names that you understand; but without spaces in the variable names.
3. Add additional variables to the dataset that have categories given in text (see above) instead of numbers. That is, keep old variables but add new ones to the dataset.

2. Examination Assignment 2 – Measures of location

and variability

You now **choose one of the numeric variables** (i.e., your renamed version of the variable).

Complete the following tasks or answer the following questions:

1. For the variable you chose, calculate the following estimates of location:
   - Mean, trimmed mean (10%), median, and create a boxplot for the variable.
2. For the variable you chose, are there any outliers? If so, how many?
3. For the variable you chose, calculate the following measures of variability:
   - Variance, standard deviation, mean absolute deviation from the median, range, and interquartile range.

3. Examination Assignment 3 – Data distributions

Choose one of the numeric variables (i.e., your renamed version of the variable). You can continue with the same variable from Assignment 2 or pick another one.

Complete the following tasks or answer the following questions:

1. Calculate the following quantiles of the variable you chose: 5%, 25%, 50%, 75%, and 95%.
2. Create a nice-looking histogram for the variable you chose with labels that can be understood for the x and y axes.
3. Create a nice-looking density plot for the variable you chose with labels that can be understood for the x and y axes.
4. Create a plot visualizing multiple boxplots so that you can compare the data distributions for the variable **'Total interactions'** (your renamed version of the variable) for the three different categories in the **'Category'** variable.*Tip: You will probably need to filter your observations for the total interactions variable to avoid impact from outliers on the visualization.*
5. Create a variant of the visualization that can be seen in **Figure 7** in the paper by Moro et al. (2016). But instead of a plot with mean values, you should use boxplots to illustrate the data distributions. Comment on the results.

6. Study **Figure 13** in the paper by Moro et al. (2016). Create a variant of this visualization using violin plots to show how ‘Lifetime Post Consumers’ vary among weekdays. Comment on the results.

4. Examination Assignment 4 – Exploring Categorical Data

It is important to be able to explore and visualize categorical data in a dataset. This task will let you show your ability to do so.

Complete the following tasks or answer the following questions:

1. Create **bar chart visualizations** to show which category of the variable **‘Type’** was the most commonly used post type, which **month** had the most posts, which **weekday** had the most posts, and posts were typically **paid or not?** *Hint, if you want to order your chart in descending or increasing order, or in other reasonable order – you may want to make sure that your variable classes are factors and then you can use ‘fct_infreq’ function. Another useful function is ‘fct_relevel’*

    - Which type was most commonly used?
    - Which month had the most posts?
    - Which weekday had the most posts?
    - Were most posts paid or not paid?

5. Examination Assignment 5 – Exploring Correlation

In this task you will show that you have the ability to calculate correlation coefficients and explore correlation among variables.

Complete the following tasks or answer the following questions:

1. Calculate the Pearson correlation coefficient between '**Lifetime Post total reach’** and **‘Lifetime Post total consumers.’** Interpret your calculated correlation coefficient. This means that you explain how the Pearson correlation coefficient should be interpreted.
2. Create a correlation plot that visualizes all correlation coefficients among the **eight (8)** variables/features that start with “Lifetime…”.

    - Interpret the results here. What conclusions can you draw about the correlation among these eight variables?

6. Examination Assignment 6 – Scatterplots and Contours

In this task, you will show that you can create nice scatterplots and contour plots for numeric data.

Complete the following tasks or answer the following questions:

1. Choose two numeric variables/features for which you find it interesting to explore a possible dependence through a scatterplot. Create a nice-looking scatterplot for these two variables you choose. Add a smoother to the plot. Comment on the result. *Hint: think about geom_smooth() within the 'ggplot2' package, which is part of the 'tidyverse' family of packages.*
2. For the same variables/features construct a scatterplot but add contours based on the density of points in the scatterplot. ***Note:*** We have too little data for this to be effective but do it anyway!
3. Create scatterplots between two numeric variables that you choose yourself but compare these scatterplots for the two categories of the 'Paid' variable in the same plot. Create another similar plot where you compare scatterplots for the three categories of the 'Category' variable. That is, can you visualize the same scatterplot between two variables but differentiate between the two levels of these categorical variabls in the same plot? Comment on the results. *Hint: If you still have NA values in your dataset, think about the 'drop_na' function within the 'tidyr' package, whihc is also part of the 'tidyverse' family of packages.*

7. The new assignments at the examination session?

If you have solved the above tasks, documented your solutions, and are able to run the code fast in the classroom you should be well-prepared to solve the additional assignment(s) that you will get at the examination session. In preparation for the examination assignments it is also important that you and have worked through the training assignments. If you have done all this, **do not worry!**

References

Moro, Sérgio, Paulo Rita, and Bernardo Vala. 2016. “Predicting Social Media Performance Metrics and Evaluation of the Impact on Brand Building: A Data Mining Approach.” *Journal of Business Research* 69 (9): 3341–51.