# Task

1. *What % of ‘email’ companies signed up on a Saturday or Sunday?* Answer: **32.36%**

Steps:

* Extract the day of the week from ‘free\_trial\_signup\_date’ column by using the formula: = TEXT(B2,"ddd")
* Create a ‘sign up on weekend’ column with Boolean value:

*0 = not sign up on weekend*

*1 = sign up on weekend*

I used the if(or ()) function: =IF(OR([@[Day of the week ]]="sun",[@[Day of the week ]]="sat"),1, 0)

Table

Description automatically generated

* Create a pivot table:
* Drag the google\_analytics\_medium column to Rows
* Add count of “Company ID” and sum of “Sign up on weekend” to the Values box

Graphical user interface, text

Description automatically generated

We’ll get a table that provide us with the info needed:

Here there are 169 “email” companies that sign up on the weekend over the total of 522 “email” companies.

So 169/522\*100% = **32.36%** “email” companies sign up on the weekend.

1. *Are app users or non app users more likely to currently be a subscriber?* **App users are more likely to currently be a subscriber.** Using pivot table again, we get:

Graphical user interface, application

Description automatically generated

The percentage of app users who are currently subscribers: 1642/1910\*100% = 85.97%

The percentage of non\_app users who are currently subscribers: 2615/4050\*100% = 64.57%

1. *Of companies that signed up on or after 1st July 2020, which medium had the highest average minutes activity on any device over the last 7 days?* ***Organic***

Here, I used MATCH & INDEX to take the column ‘Minutes\_activity\_last\_7\_days\_all\_devices” from the signup file to the company\_information file. The I changed the free\_trial\_signup\_date filter to *equal to or after* 01/07/2020:

Graphical user interface, text, application, chat or text message

Description automatically generated Graphical user interface, text, application

Description automatically generated

I created a pivot table and sort the values from largest to smallest. Here, undetected medium (none) had the highest average minutes activity on all devices over the last 7 days. Among the detected medium, **organic** had the highest average minutes activity on all devices over the last 7 days.

# **Investigation**

I'd like to create a classification model that can predict whether a company will subscribe to Company A using data from the datasets, such as whether they've ever contacted support, left a Trustpilot review, their minutes of activity, their free trial registration date, and so on. Before I do that, I'd like to look into the relationship between the target (subscription status) and the available independent variables.

Chart, bar chart

Description automatically generatedText

Description automatically generated with low confidence

1. Exploratory Data Analysis

The dashboard suggests that the independent variables are highly correlated to the target (subscribe/cancel). High engagement suggests subscription. Subscribers spend more than twice as much time on Company A than cancellers (4 minutes vs 2 minutes). Customer support engagement is likewise substantially connected, with 91.73 percent of subscribers contacting support and 97.5 percent of cancellers never contacting support. The percentage of subscribers using app is more than double that of cancellers (38% vs 16%). Subscribers are 20 times more likely than non-subscribers to leave a Trustpilot review. Subscribers discover Company A mostly through their own searches, with 86 percent coming from organic sources. As for cancellers, the arrived medium is equally distributed. The most popular day to sign up for free trial is Friday.

1. Data Pre-processing

I used Python on Jupiter Notebook to work on the model. 339 missing values in the column *subscription\_status* (the target) was removed, and the merged dataset was split into training and testing set by [stratified sampling](https://en.wikipedia.org/wiki/Stratified_sampling) based on subscription\_status. The categorical variable, *Google Analytics Medium, is encoded into dummy variables,* *Minutes\_activity\_last\_7\_days\_mobile* and *Minutes\_activity\_last\_7\_days\_desktop* are not considered because they will be correlated with another all-including variable, *Minutes\_activity\_last\_7\_days\_all\_devices.*

1. Model

[Random Forest](https://en.wikipedia.org/wiki/Random_forest) is the machine learning algorithms used to fit the data. It is chosen because of the higher accuracy through cross validation. Grid Search is used to find the best Random Forest model. Here, we can see that the best model returns 96,6% accuracy. The correlation matrix on the test set shows 98% of the subscribed companies are correctly predicted. The mean train score is only 1.6% larger than the mean test score, suggesting little overfitting.

Table

Description automatically generated Chart, treemap chart

Description automatically generatedText, letter

Description automatically generated

1. Assumptions and notes

I assumed that the *Minutes\_activity\_last\_7\_days* is only for free-trial users. Otherwise, it would be obvious that subscribers would be more active on the website than cancellers. Also, the screenshot shows the list of independent variables that have been used by the model and their weights in the model. Data for different years are not evenly collected (only 1 month of data for 2021, 4 months for 2019 and the whole 12 months for 2020). In this model, I assume that the month and year observations are distributed evenly in training and testing dataset. Variable year and month might need to be removed.