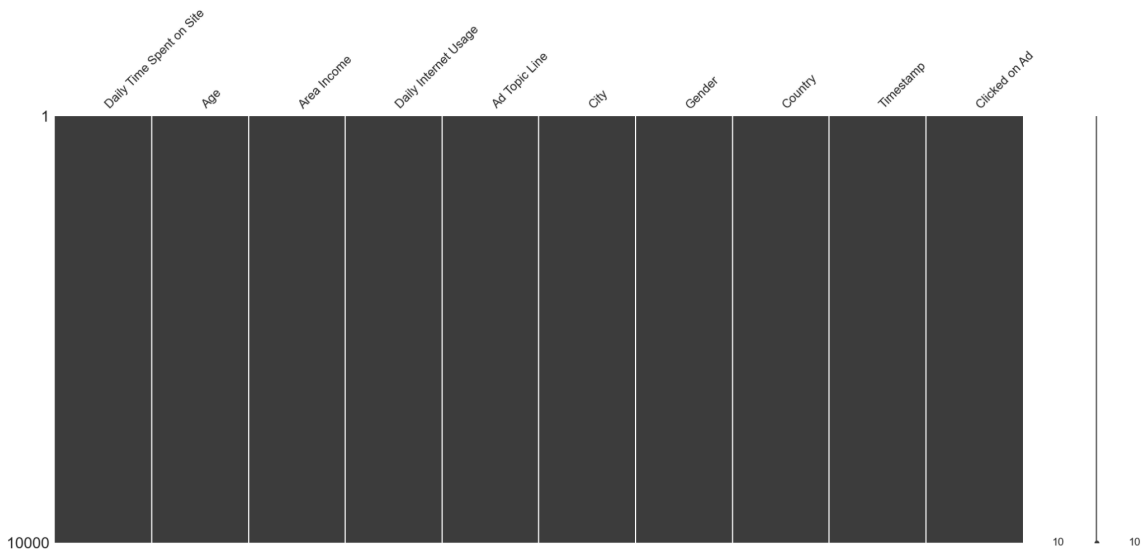
Ad-clicks EDA report

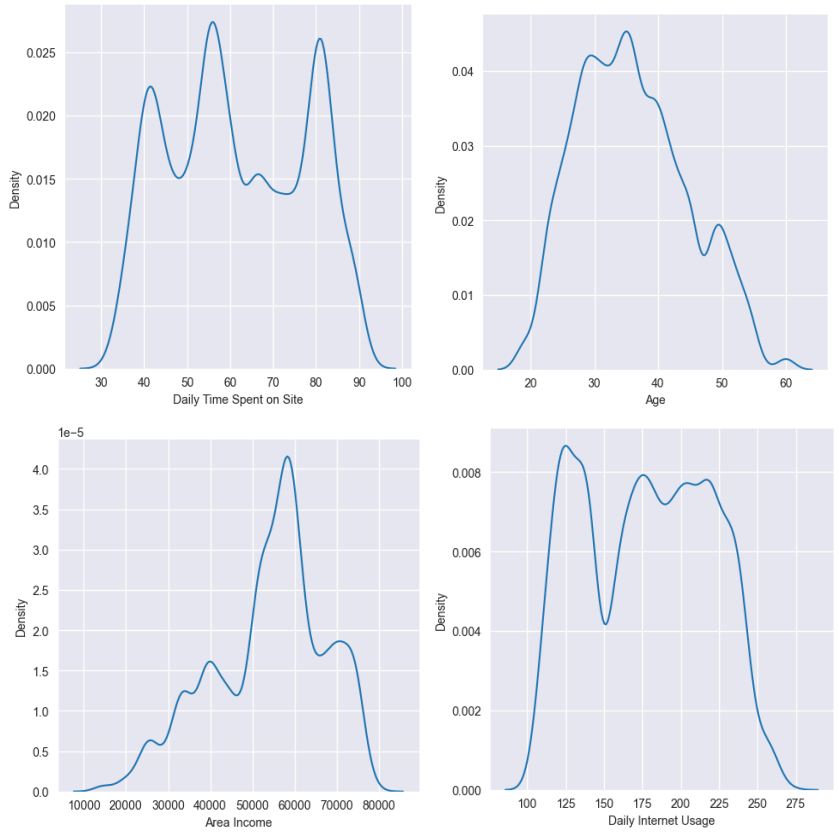
1- Missing values



According to the above matrix plot, the data has no missed values. So we don't need to worry about handling this part.

2- Data distribution

I checked the numerical data distribution in earlier steps of the task, to have a quite general plot of data, and also outliers, and figure out how the numerical data can affect the click-on ads.

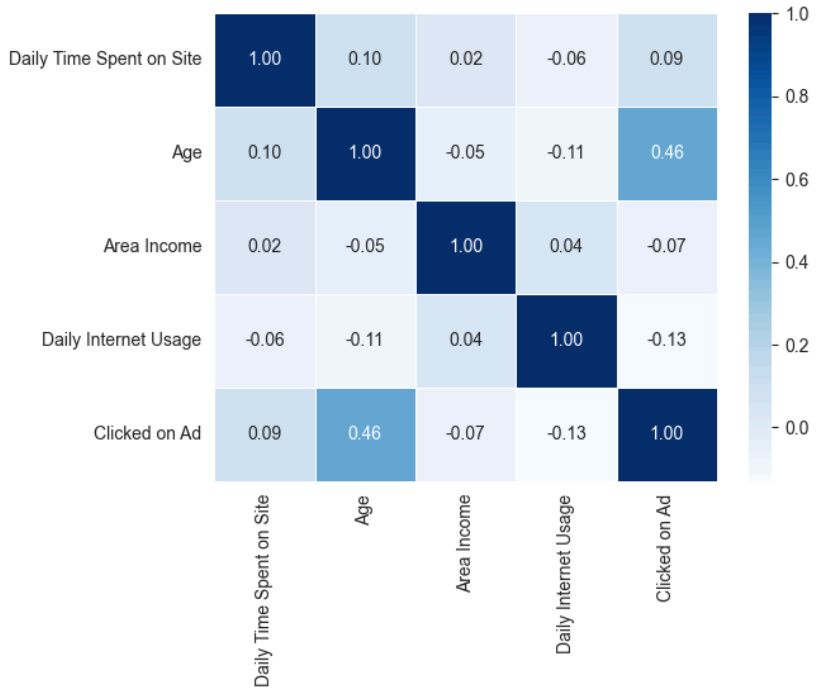


Each plot shows the density of a numerical column(defined on the X-axis). As we can see, the scale of the data, is significantly different, so we may need to scale these features in the feature engineering phase.

2- Correlations

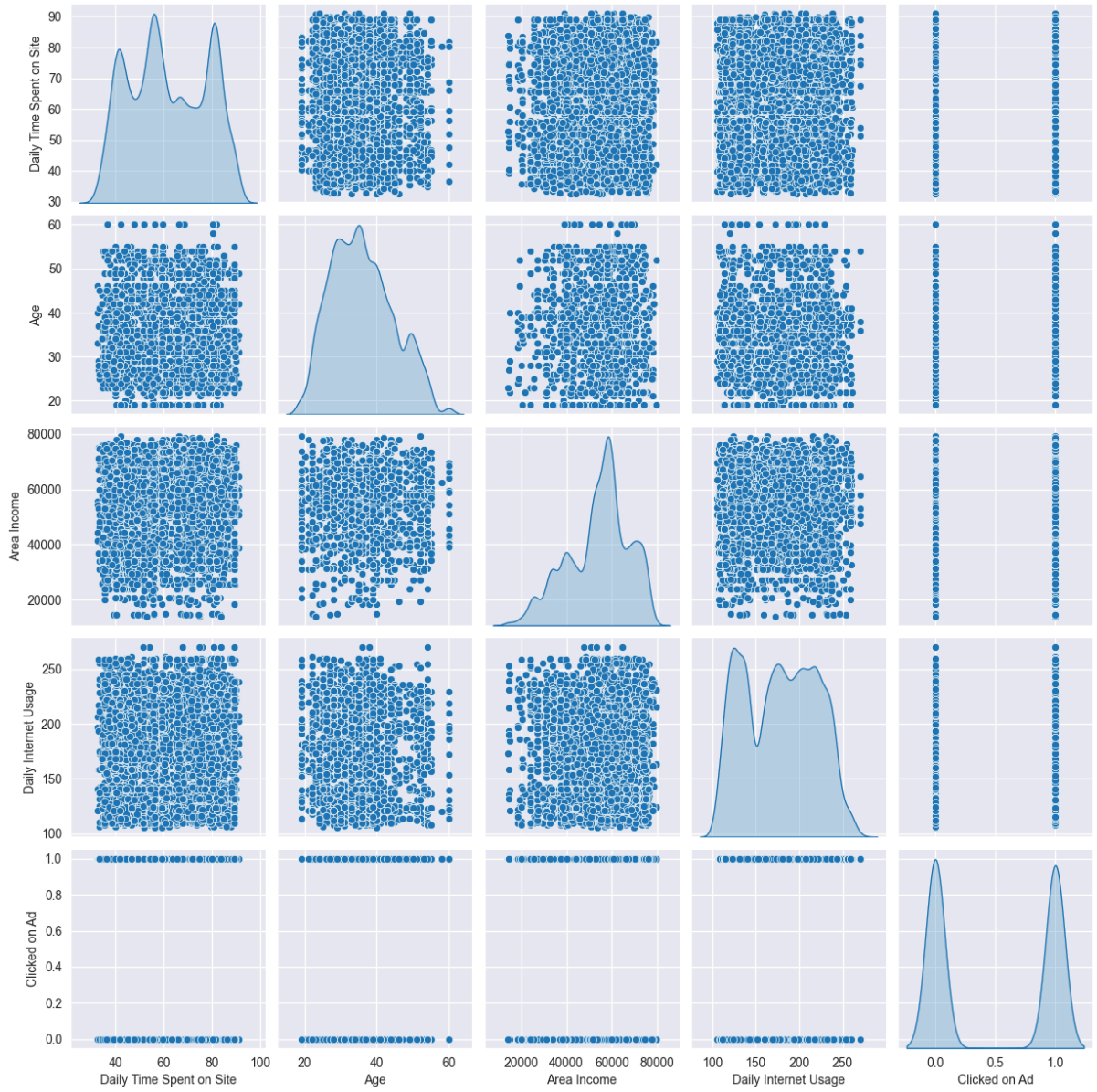
Now let us investigate how features relate to each other.

2-1- Correlation heatmap



The plot shows a significant correlation between Age and clicked-on ads, which means the older persons, are more likely to click on the ads. It refers to the unfamiliarity of the majority of older persons to social media advertisements. So it's an important feature for us since it has a significant with target features. Another feature that has less effect on the target is Daily spent time on the internet, which is quite reasonable, the people who are more on the internet, its more pruned to click on the ads, accidentally or on purpose!
But daily internet usage has a negative correlation with the target. So three main features are **linearly** correlated with the target.

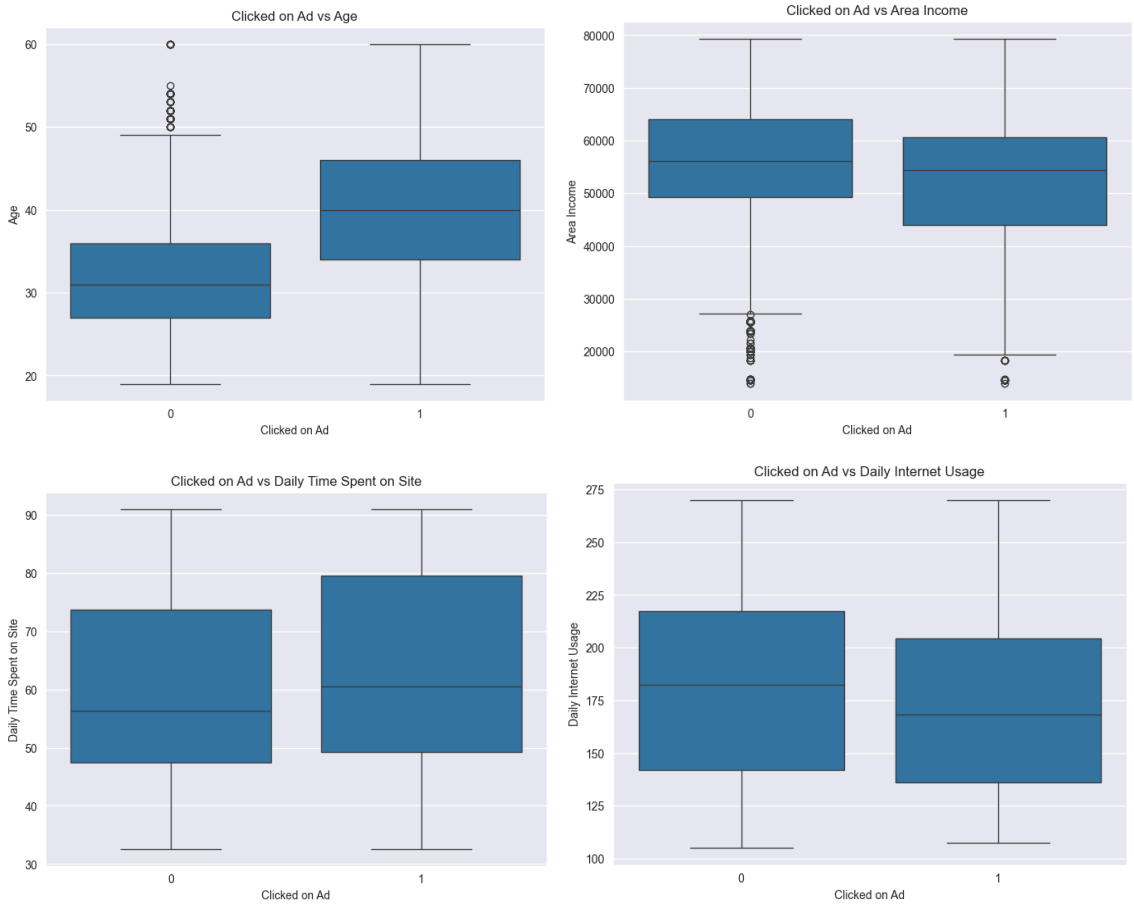
2-2- Correlation matrix



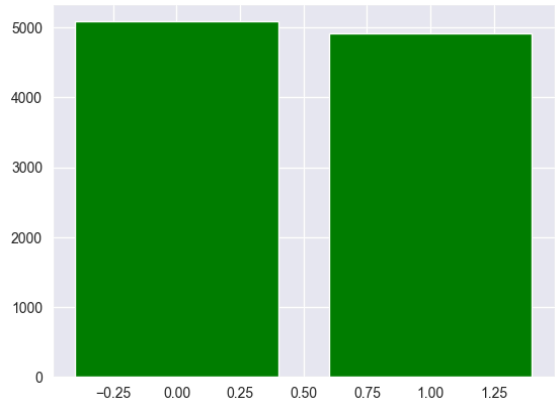
It's the same plot as the heatmap(semantically) but with a different view.

3- Target feature investigation

The most important part of the project is to find and make relevant features for classify the target value of each record. So it's important to figure out all useful information to make reasonable features. In this section, after checking the numerical features correlations using a heatmap, I tried to visualize the relations more accurately using **box plot.**

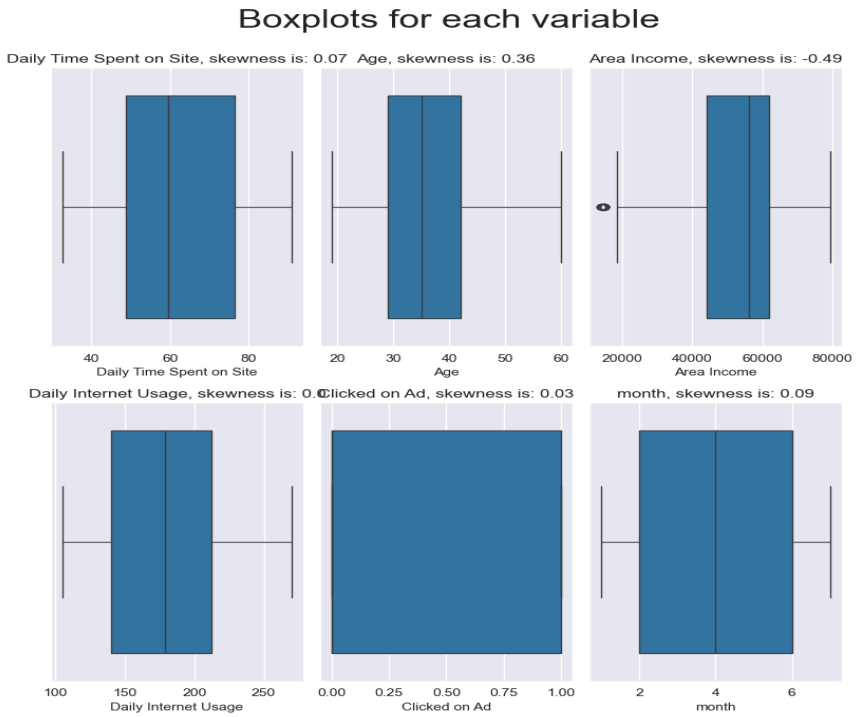


The top left figure shows that the Age is significantly related to the target. and also top right and downright figures clearly show that there is a negative relation between target feature.



And also, the plot above shows the distribution of the target class, which shows that the classes are balanced.

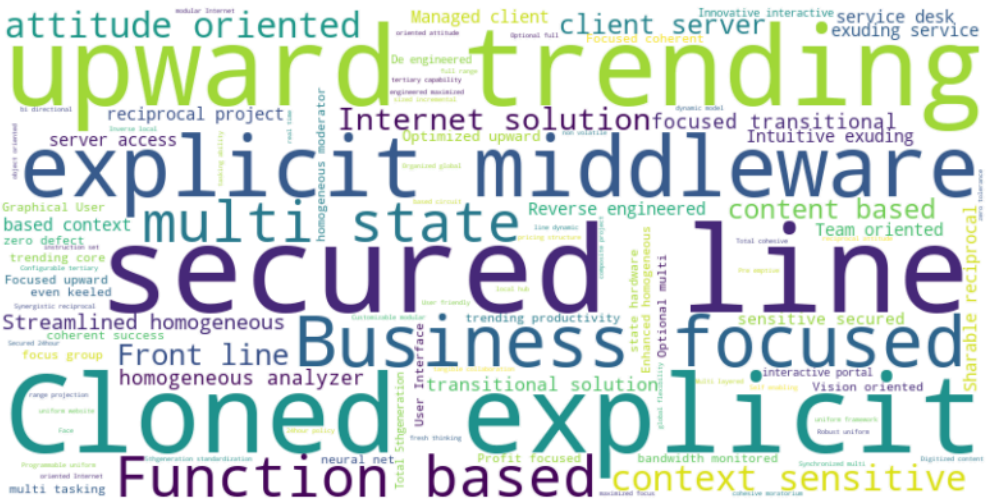
4- Outliers



The box plots above show the density of the numerical feature of data, as we can see, only area income has some outlier points, so we can just ignore them since the model that we choose is robust to the outliers.(Random forest classifier).

5- Textual features

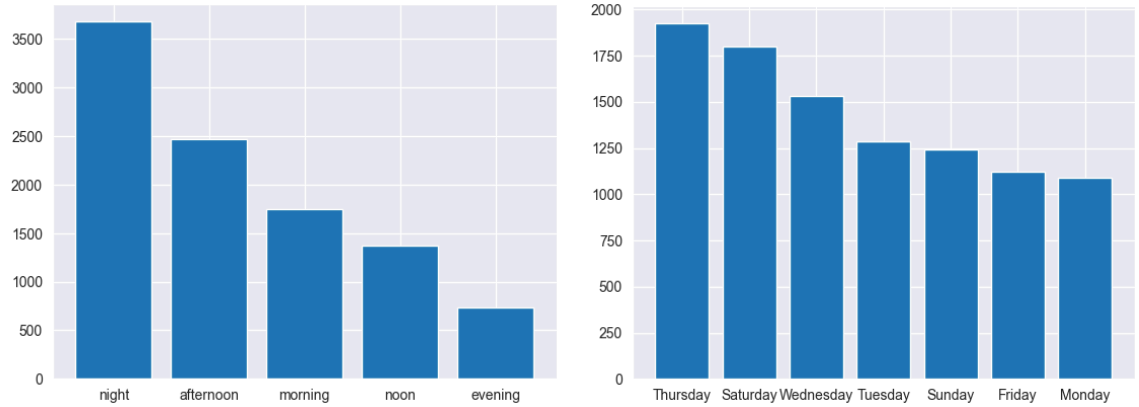
The dataset contains a textual column, which is the topic of the ads. It's quite reasonable to investigate these kinds of features while it's very important to make a good prediction.
Let's take a look at the word cloud of this column.



The plot above shows the most frequent words of the text data(larger font means more frequency of words)

6- TImestamp feature

We extract useful information from the timestamp feature. Which you can see in the plots below.



The left plot shows the number of active users in each part of the day, as you can see the most frequent time is night and the least is evening. And also the right plot shows the activity on the weekdays.