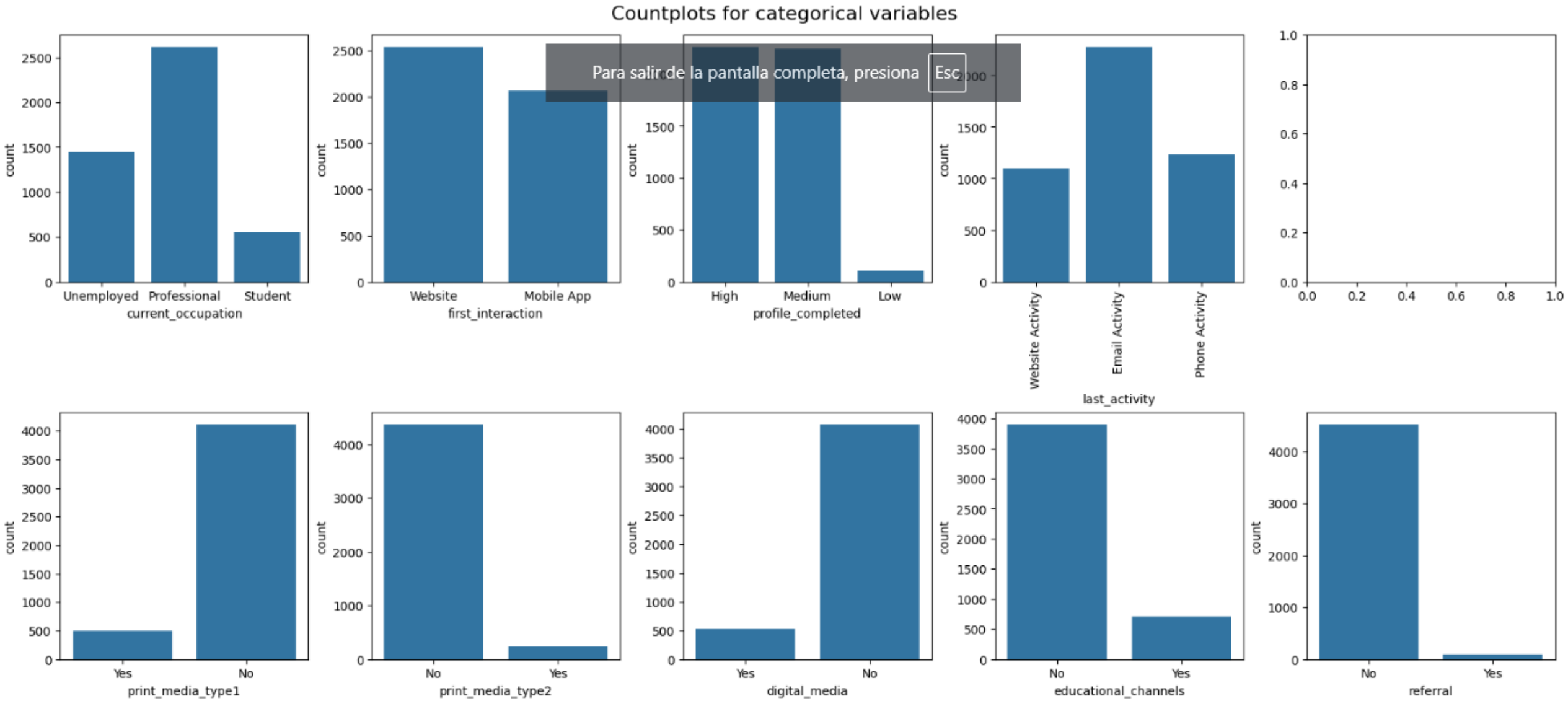
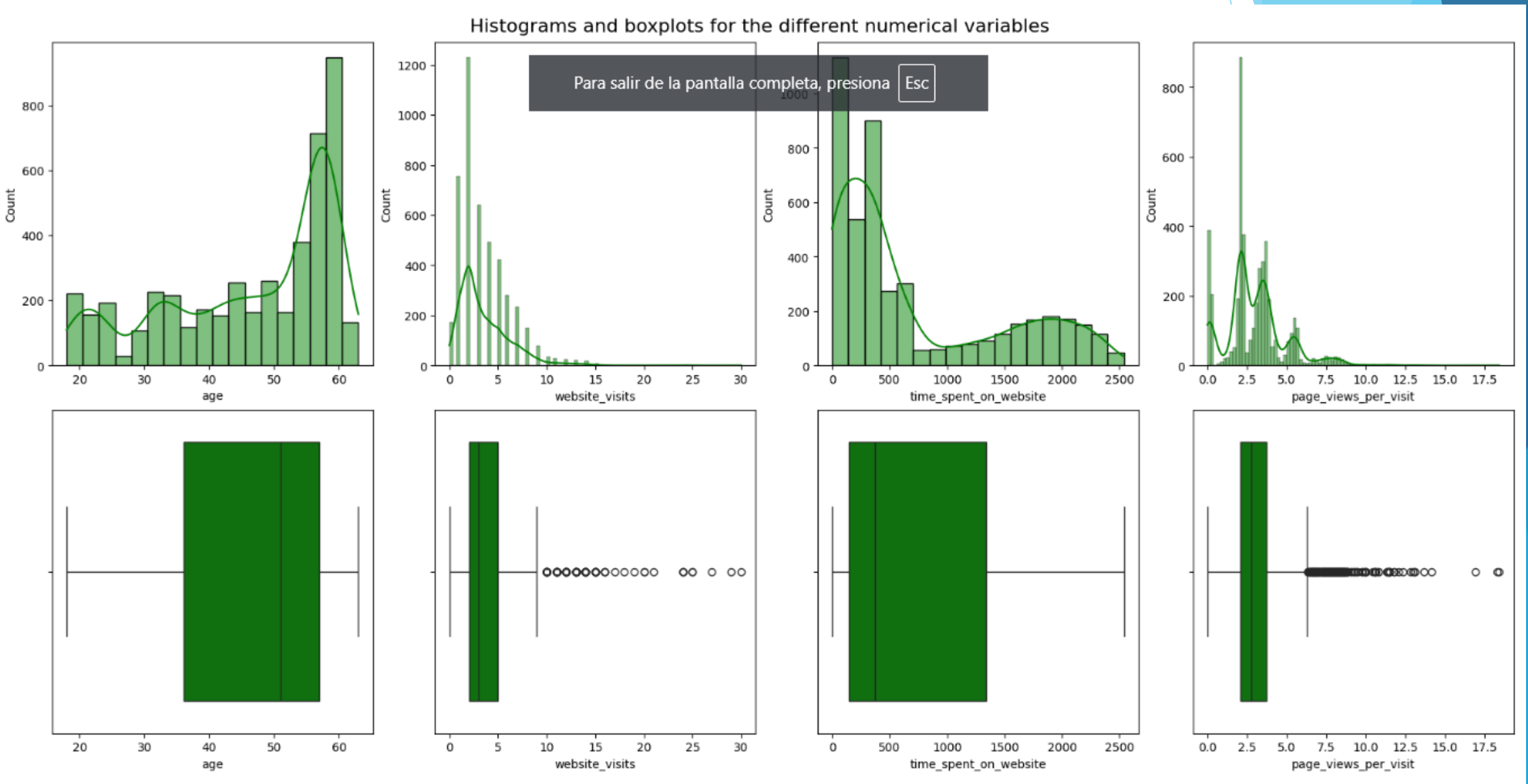
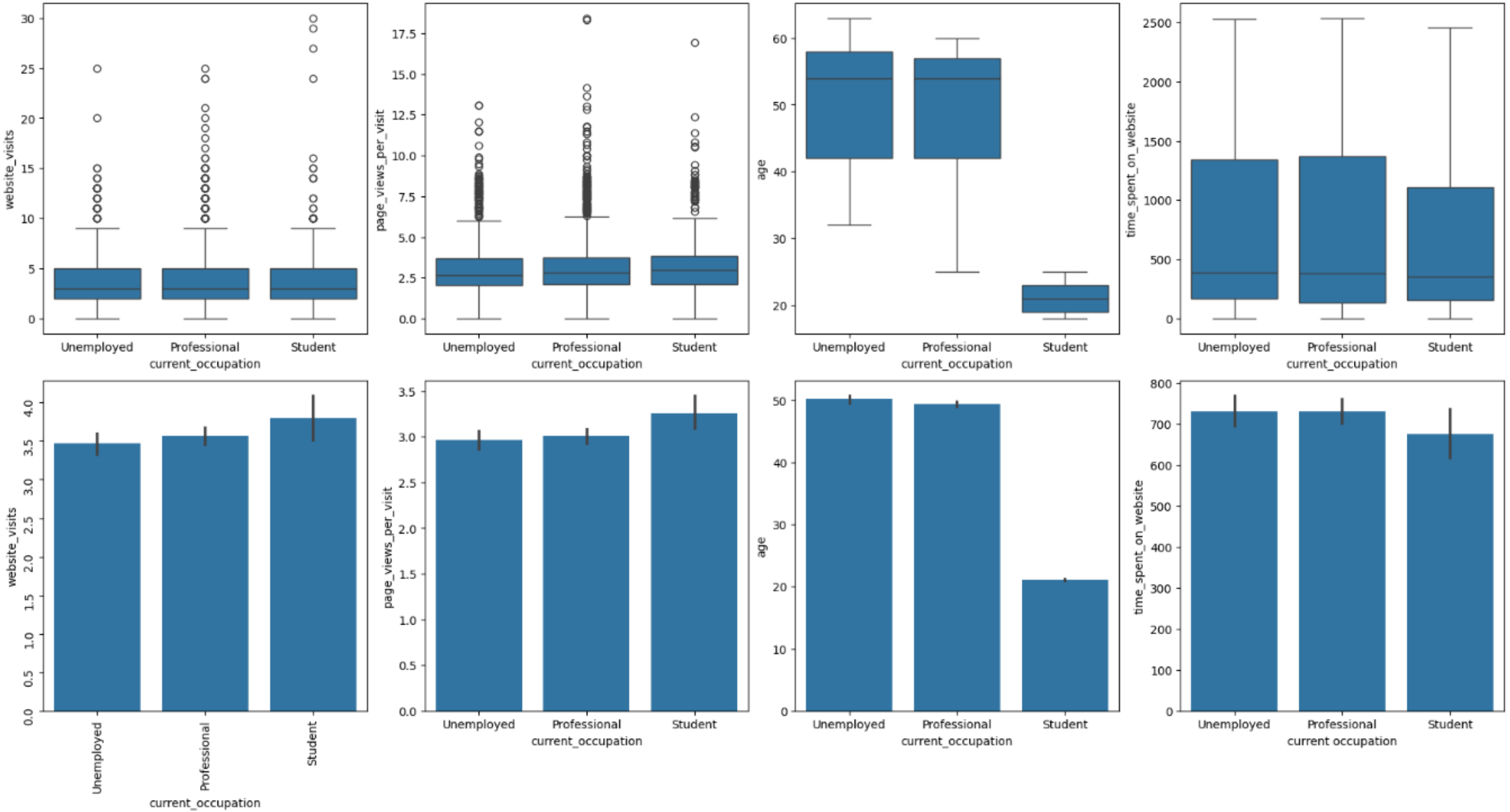
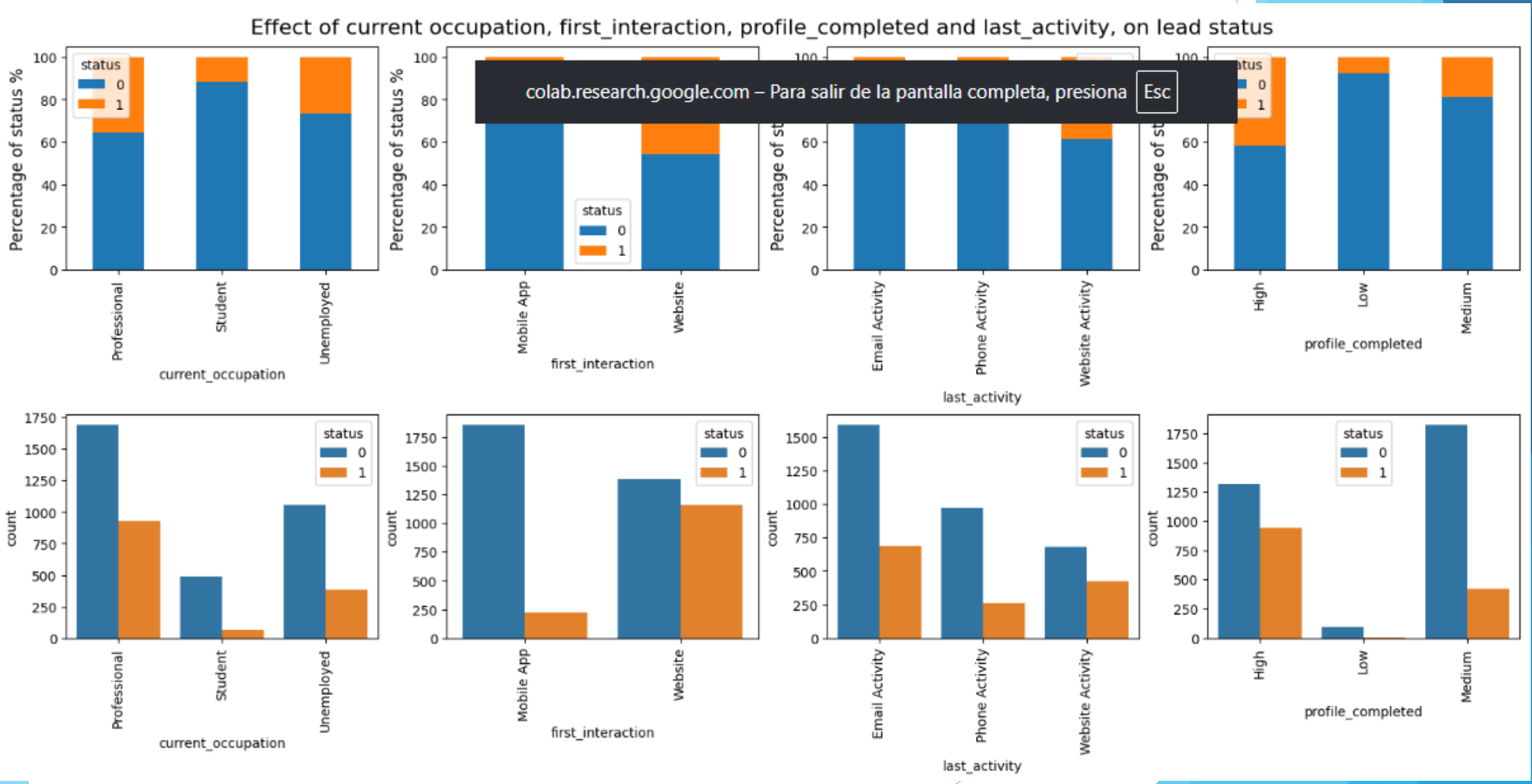
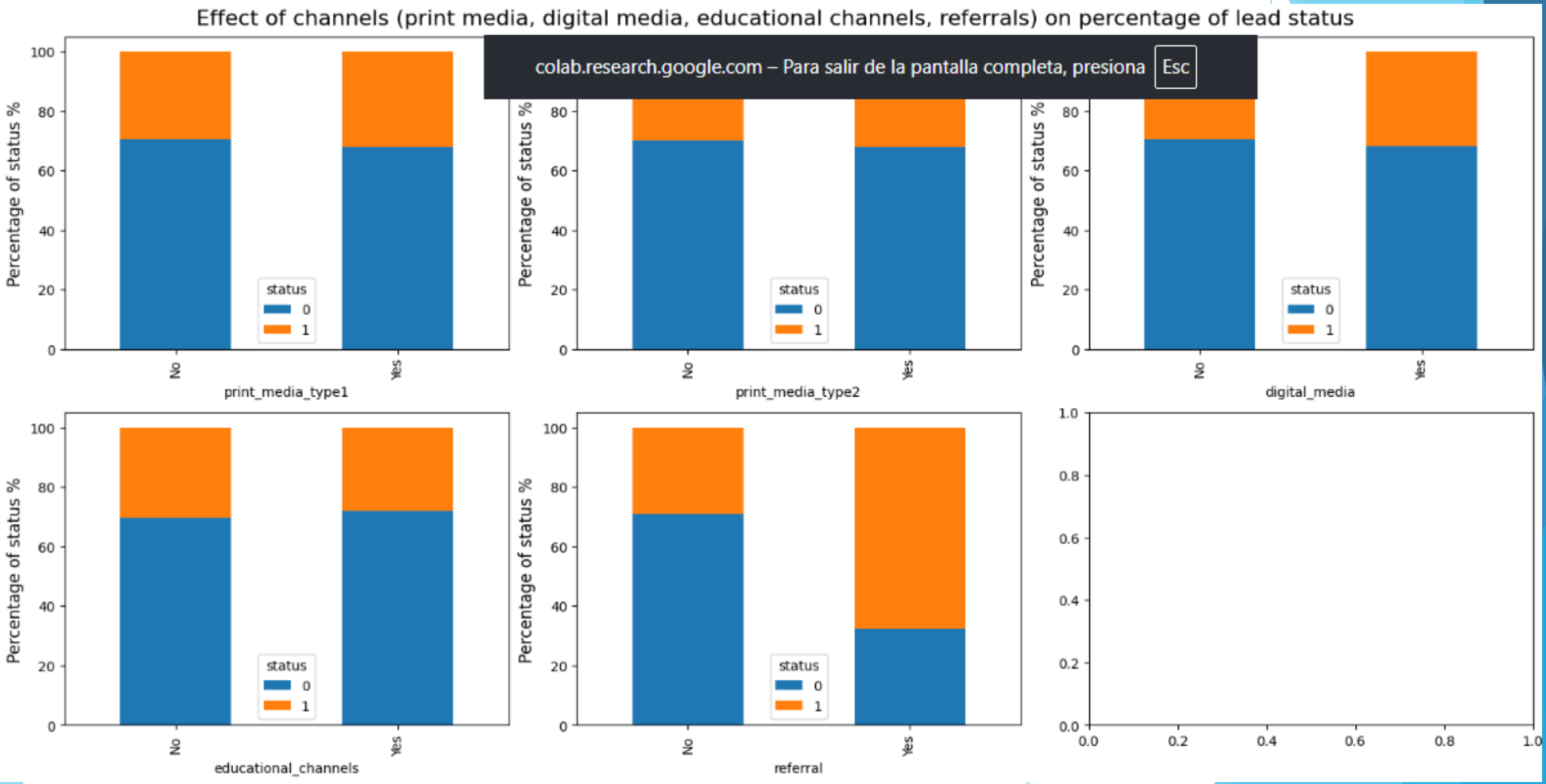

Countplots for categorical variables
Para salir de la pantalla completa, presiona Esc
count
current_occupation
Unemployed
Professional
Student
first_interaction
Website
Mobile App
profile_completed
High
Medium
Low
last_activity
Website Activity
Email Activity
Phone Activity
print_media_type1
Yes
No
print_media_type2
No
Yes
digital_media
Yes
No
educational_channels
No
Yes
referral
No
Yes


Histograms and boxplots for the different numerical variables
Para salir de la pantalla completa, presiona Esc
Count
age
website_visits
time_spent_on_website
page_views_per_visit

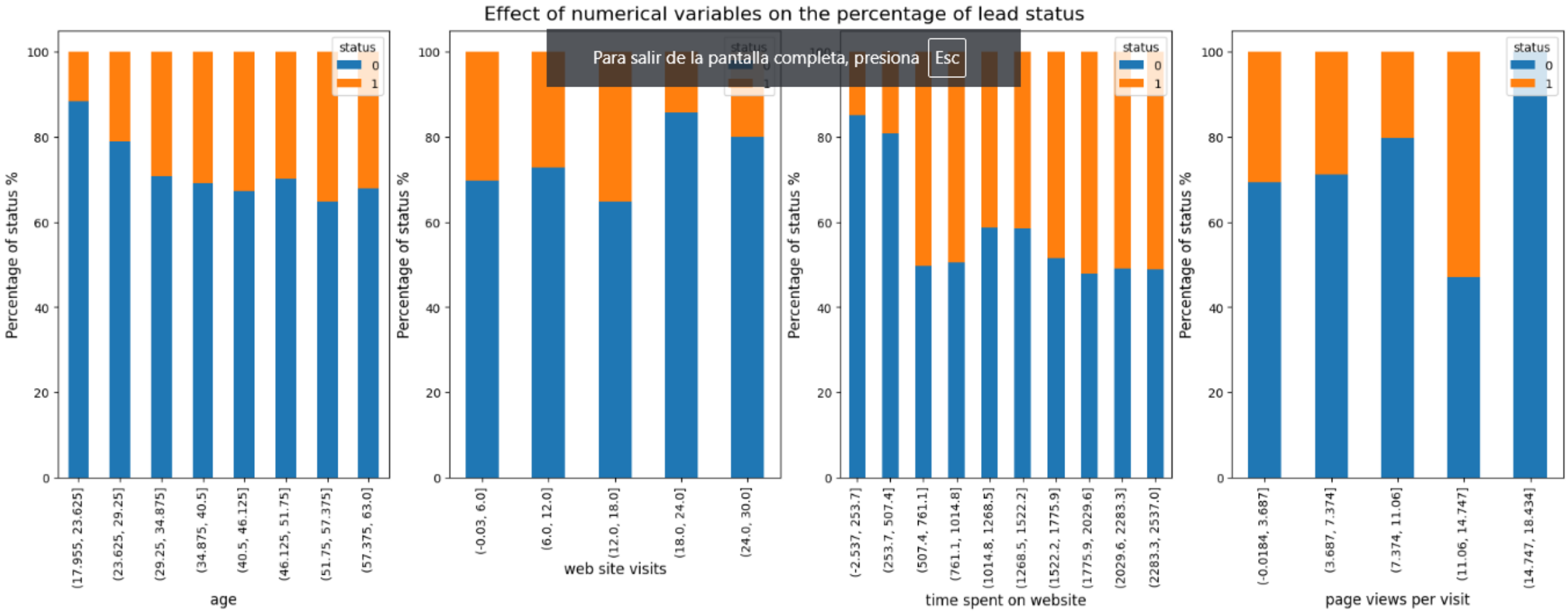
Bivariate boxplots, for numerical variables (y axis) vs current occupation (categorical variable in x axis)



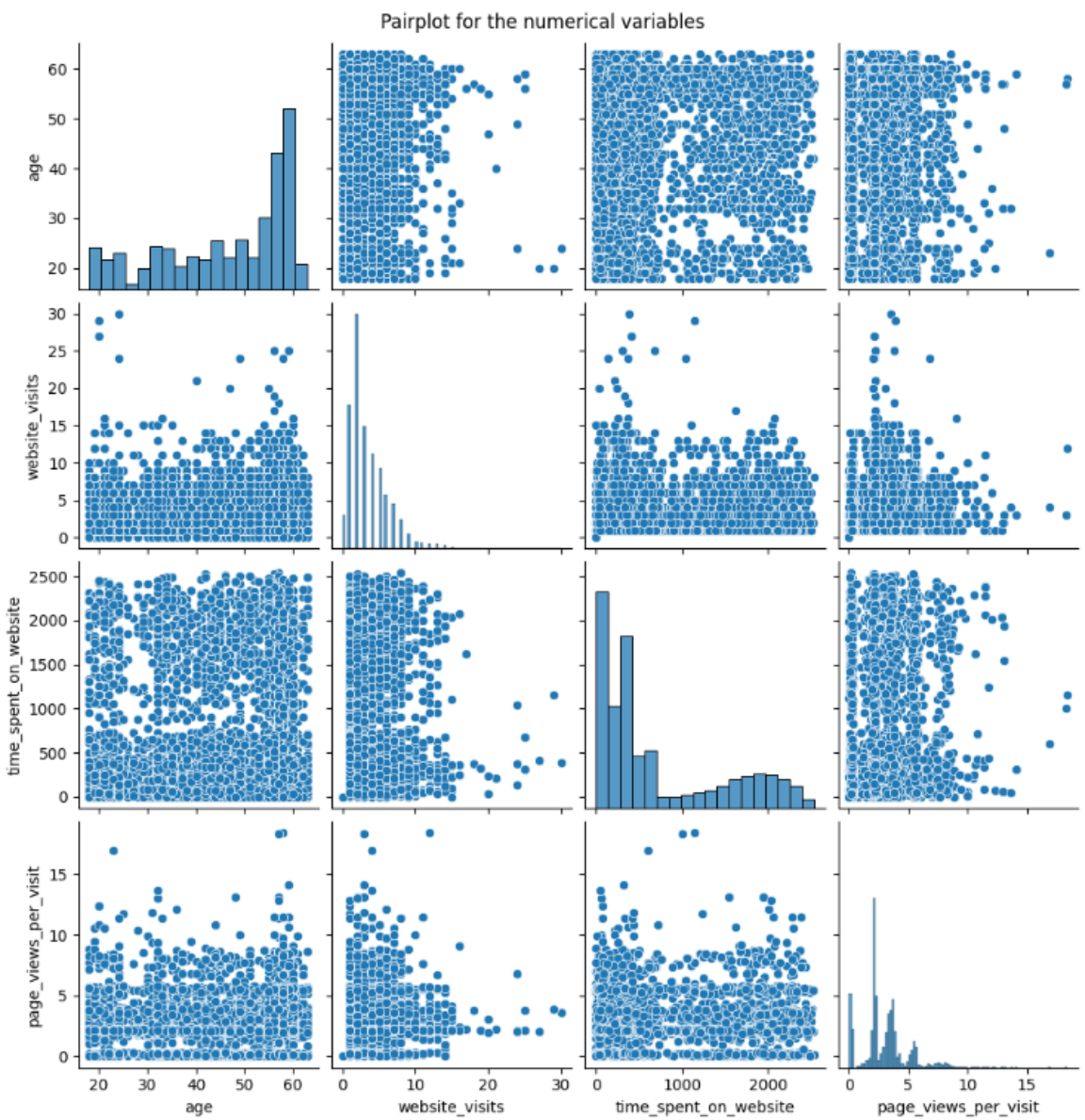

Effect of current occupation, first_interaction, profile_completed and last_activity, on lead status
Percentage of status %
count
status
0
1
Professional
Student
Unemployed
current_occupation
Mobile App
Website
first_interaction
Email Activity
Phone Activity
Website Activity
last_activity
High
Low
Medium
profile_completed
colab.research.google.com – Para salir de la pantalla completa, presiona Esc

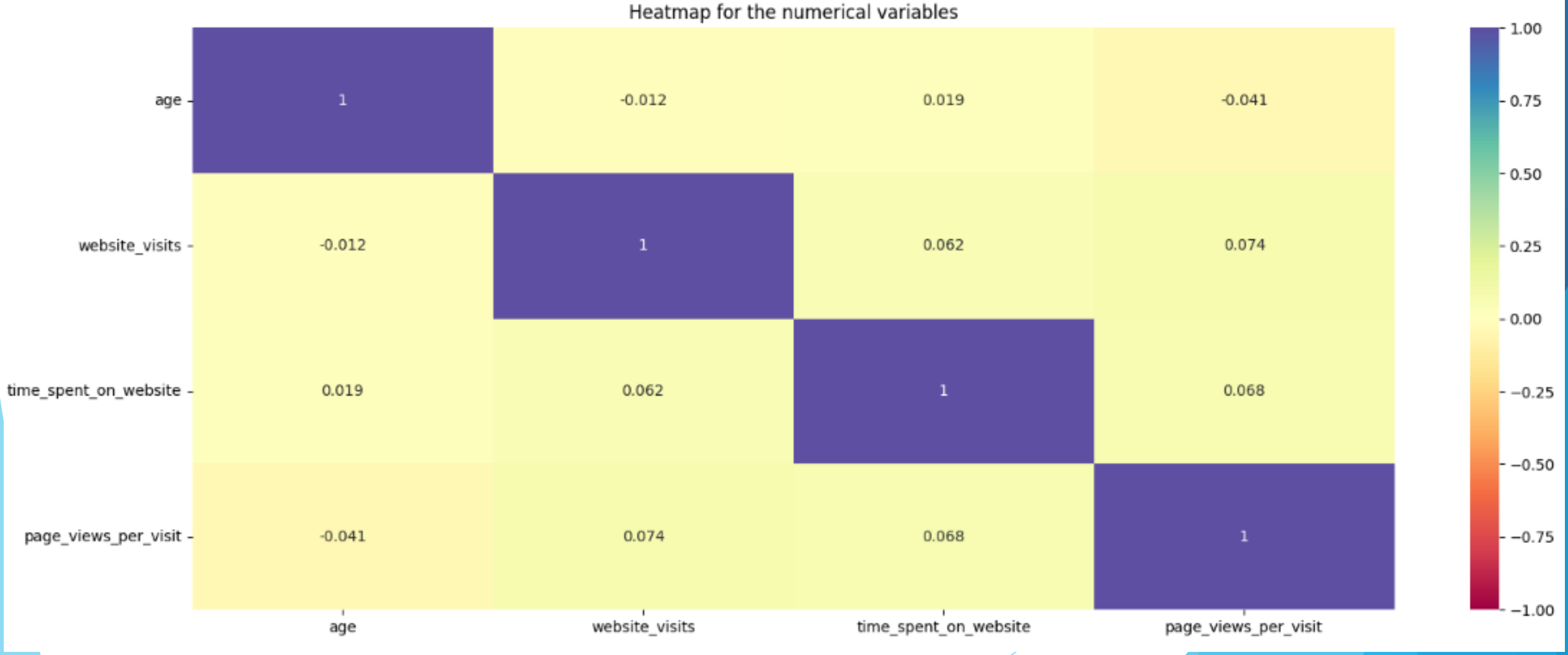

Effect of channels (print media, digital media, educational channels, referrals) on percentage of lead status
colab.research.google.com – Para salir de la pantalla completa, presiona Esc
Percentage of status %
status
0
1
No
Yes
print_media_type1
print_media_type2
digital_media
educational_channels
referral

Effect of numerical variables on percentage of lead status

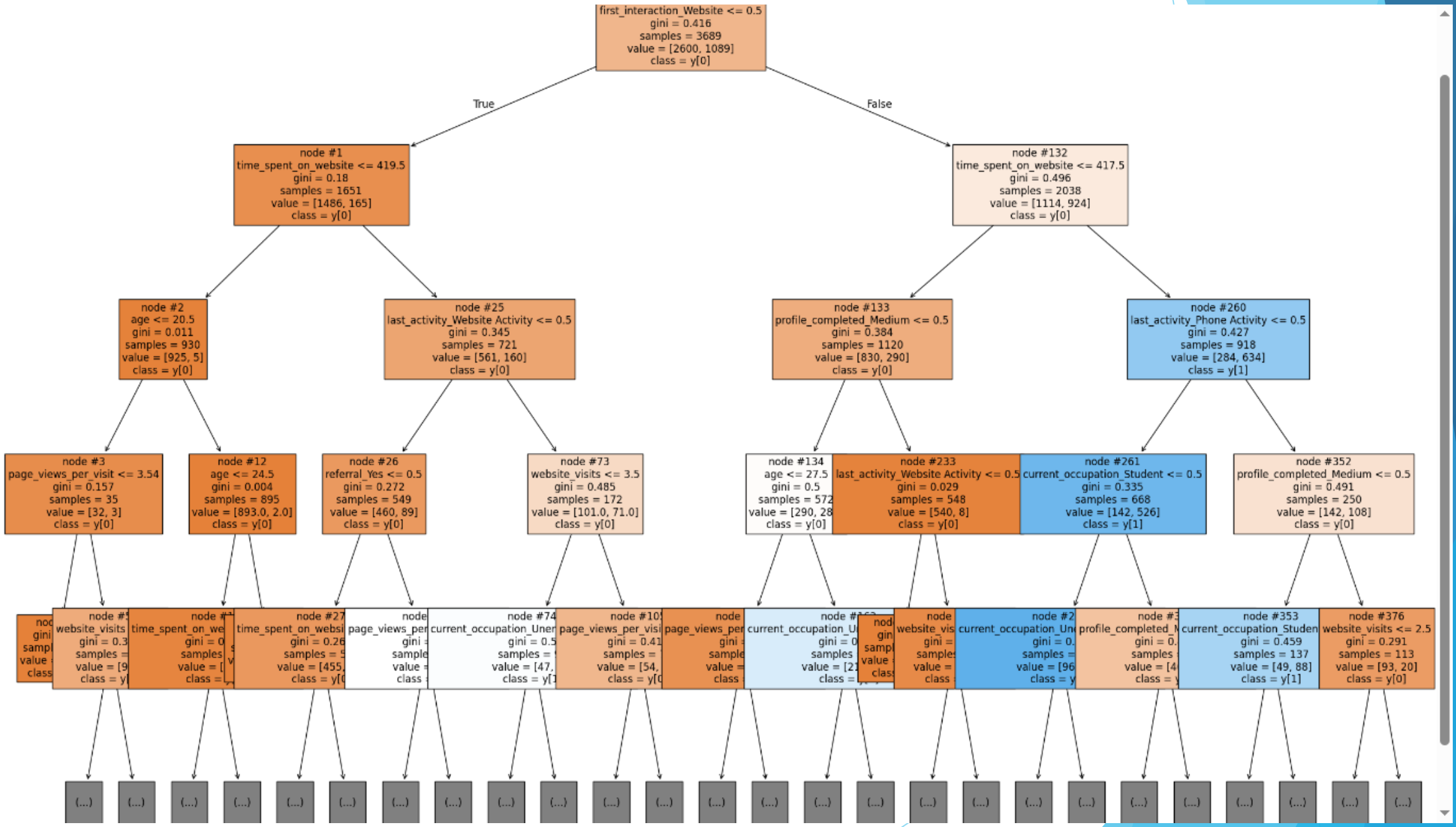


Pairplot for numerical variables




Heatmap for the numerical variables
age
website_visits
time_spent_on_website
page_views_per_visit
1
-0.012
0.019
-0.041
-0.012
1
0.062
0.074
0.019
0.062
1
0.068
-0.041
0.074
0.068
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Decision tree



Tuning of random forest

Tuning of the random forest:

Tuned Random Forest Regressor

Note: we optimize the parameters:\

- max_depth = max number of levels in each decision tree
- max_features = max number of features considered for splitting a node
- n_estimators = number of trees in the foreset

Running the code below takes some time.

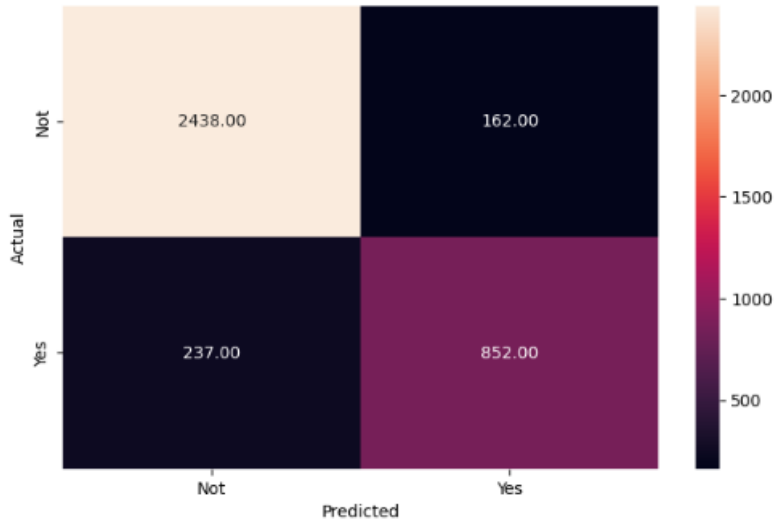
```
rf_tuned = RandomForestClassifier(random_state = 1)

# Grid of parameters to choose from
rf_parameters = {"n_estimators": [100, 110, 120],

    "max_depth": [5, 7, 15, 20, 30, 50],

    "max_features": [0.8, 1, 2]
            }

# Run the grid search
rf_grid_obj = GridSearchCV(rf_tuned, rf_parameters, scoring = 'neg_mean_squared_error', cv = 5)

rf_grid_obj = rf_grid_obj.fit(x_train, y_train)

# Set the rf_tuned_regressor to the best combination of parameters
rf_tuned_regressor = rf_grid_obj.best_estimator_

rf_tuned_regressor.fit(x_train, y_train)
```

```
RandomForestClassifier(max_depth=7, max_features=0.8, random_state=1)
```

Metrics of tuned random forest on training data:

```
# Model Performance on the train data:
y_pred_train_rf_tuned = rf_tuned_regressor.predict(x_train)
metrics_score(y_train, y_pred_train_rf_tuned)
```

	precision	recall	f1-score	support
0	0.91	0.94	0.92	2600
1	0.84	0.78	0.81	1089
accuracy			0.89	3689
macro avg	0.88	0.86	0.87	3689
weighted avg	0.89	0.89	0.89	3689

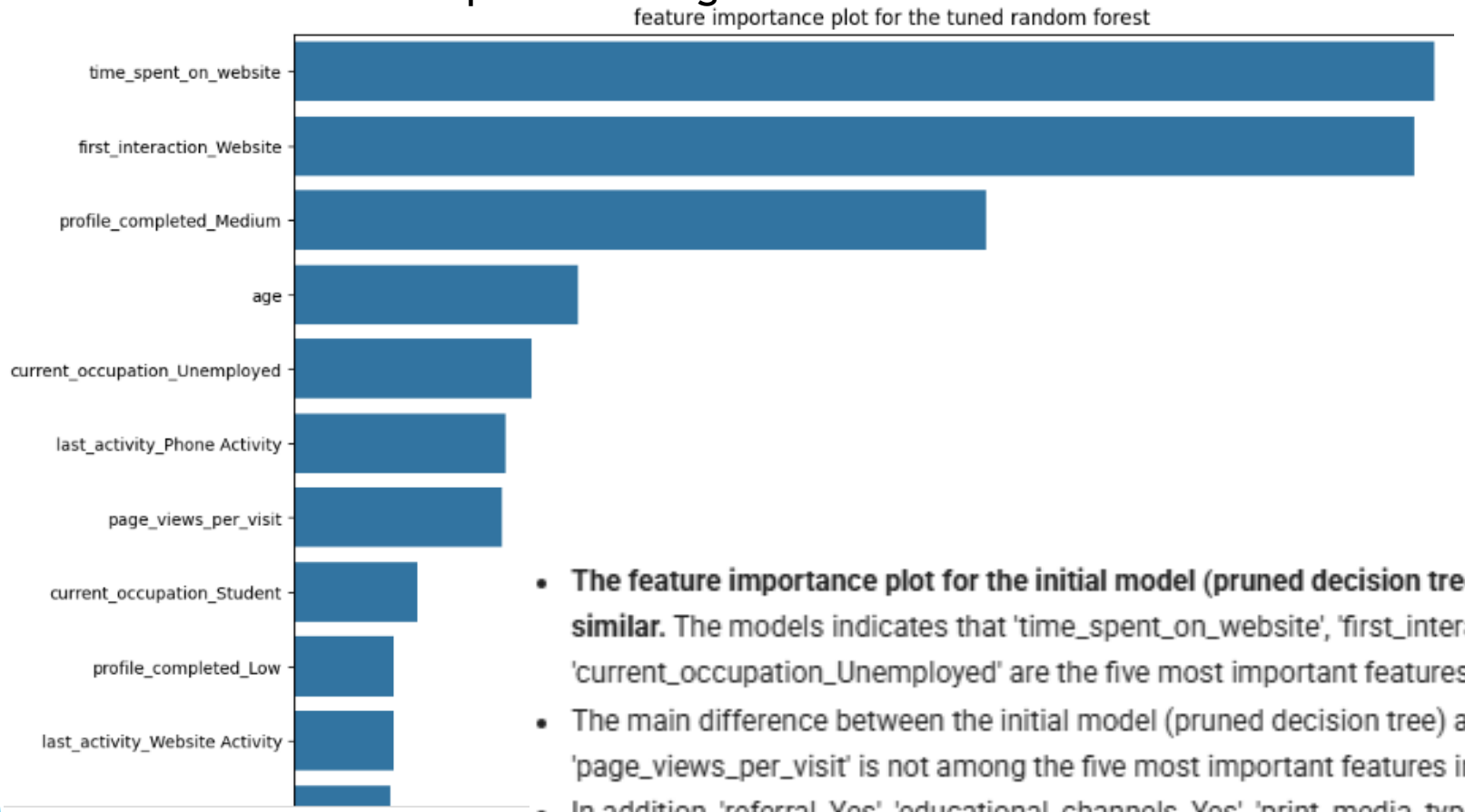


Metrics of tuned random forest on test data:

```
# Model Performance on the test data:
y_pred_test_rf_tuned = rf_tuned_regressor.predict(x_test)
metrics_score(y_test, y_pred_test_rf_tuned)
```

	precision	recall	f1-score	support
0	0.88	0.93	0.91	635
1	0.83	0.73	0.78	288
accuracy			0.87	923
macro avg	0.86	0.83	0.84	923
weighted avg	0.87	0.87	0.87	923

‘Feature importance’ figure for the tuned Random Forest model



- **The feature importance plot for the initial model (pruned decision tree) and that for tuned model (tuned random forest) are quite similar.** The models indicates that 'time_spent_on_website', 'first_interaction_Website', 'profile_completed_Medium', 'age', 'current_occupation_Unemployed' are the five most important features.
- The main difference between the initial model (pruned decision tree) and tuned model (tuned random forest), is that 'page_views_per_visit' is not among the five most important features in the tuned random forest.
- In addition, 'referral_Yes', 'educational_channels_Yes', 'print_media_type1_Yes', 'print_media_type2_Yes' have a low importance.
- We can say that leads with higher likelyhood to become paid customers are characterized by: spend more time on website, their first interaction is through website, have a High level of profile completion; they are older, and their current occupation is professional.
- The channels (print_media_type1, print_media_type2, digital_media, educational_channels, referral) have a low importance in the achievement of paid customers. Also, referral is the channel with higher importance.

Actionable Insights and Recommendations

Conclusions

- We have utilized the decision tree, pruned decision tree, random forest and tuned random forest.
- The tuning of random forest resulted in avoidance if overfitting.
- The obtained values of precision and recall compared between train and test are very similar, with a precision of 0.83-0.84 and a recall of 0.73-0.78.
- We have identified the key factors involved in the conversion of leads to paid customers, by means of the 'feature importance' diagram.
- It is possible to improve the tuning by including other parameters in the optimization, and by modifying the parameter values used as possible values in the optimization.

Recommendations:

- The variables 'time_spent_on_website' and 'first_interaction_Website' are the most important ones for identifying which leads are more likely to convert to paid customers.
- Then, to improve the conversion of unpaid to paind customers, the ExtraaLearn company should focus on: i) increasing the advertising that motivates people to look at the website; ii) reviewing the website (the information that is given to the readres) and improving it if possible.
- The variables 'current_occupation_Unemployed' has a significantly higher importance than 'current_occupation_Student'. Therefore, the ExtraaLearn company should focus its advertising and communication campaigns on unemployed people and Professional people rather than students.