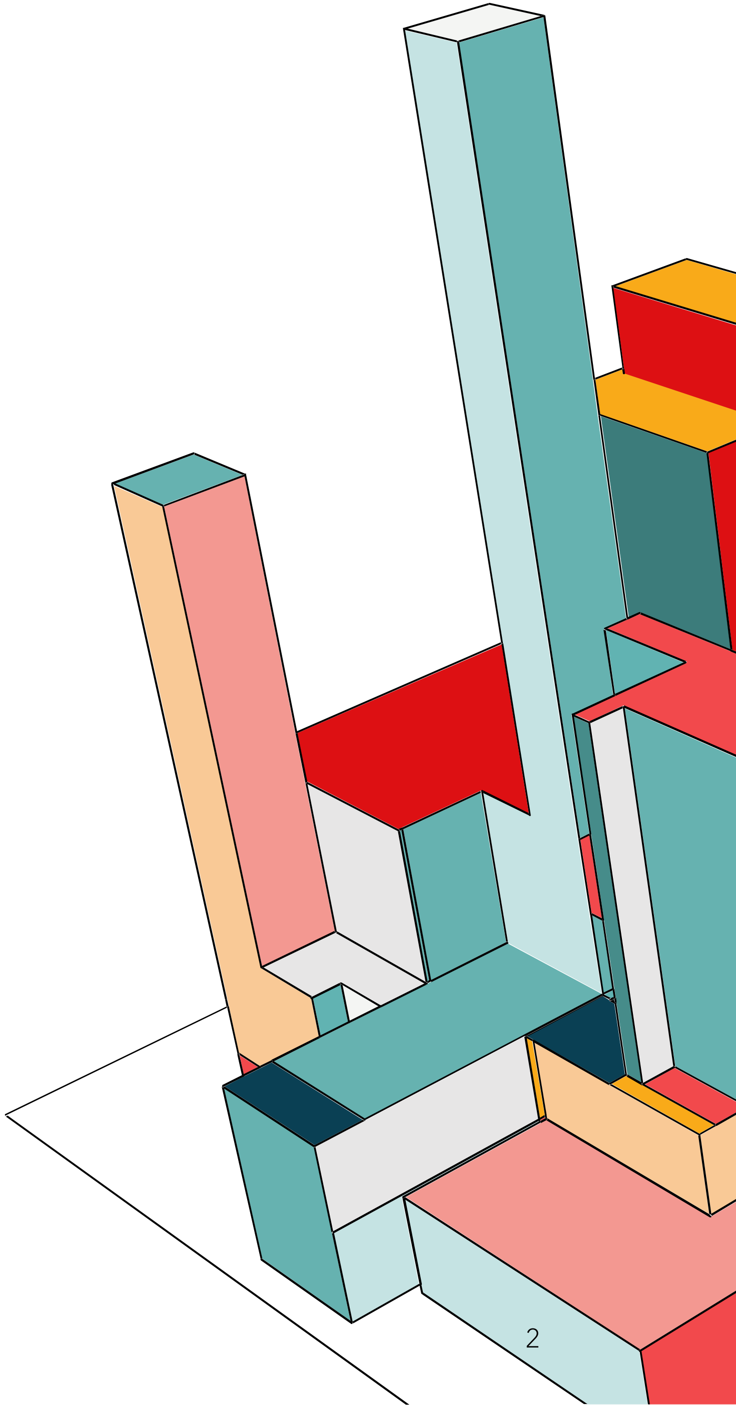
LEADS SCORING CASE STUDY

- Akshay Athawale
- Puru Sood
- Prathamesh Videkar

PROBLEM STATEMENT

- X EDUCATION SELLS ONLINE COURSES TO INDUSTRY PROFESSIONALS.
- X EDUCATION GETS A LOT OF LEADS, ITS LEAD CONVERSION RATE IS VERY POOR. FOR EXAMPLE, IF, SAY, THEY ACQUIRE 100 LEADS IN A DAY, ONLY ABOUT 30 OF THEM ARE CONVERTED.
- TO MAKE THIS PROCESS MORE EFFICIENT, THE COMPANY WISHES TO IDENTIFY THE MOST POTENTIAL LEADS, ALSO KNOWN AS ‘HOT LEADS’.
- IF THEY SUCCESSFULLY IDENTIFY THIS SET OF LEADS, THE LEAD CONVERSION RATE SHOULD GO UP AS THE SALES TEAM WILL NOW BE FOCUSING MORE ON COMMUNICATING WITH THE POTENTIAL LEADS RATHER THAN MAKING CALLS TO EVERYONE.



SOLUTION METHODOLOGY

Exploratory Data Analysis (EDA)

- Univariate data analysis: value count, distribution of variables, etc.
- Bivariate data analysis: correlation coefficients & pattern between the variables etc.
- Feature Scaling & Dummy variables & encoding of the data.
- Classification technique: logistic regression is used for model making & prediction.
- Validation of the model.
- Model presentation.
- Conclusions & Recommendations

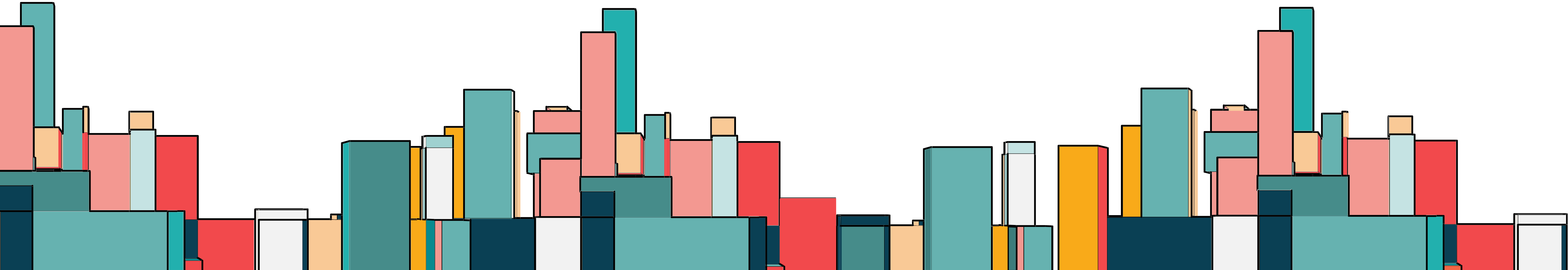
SOLUTION METHODOLOGY

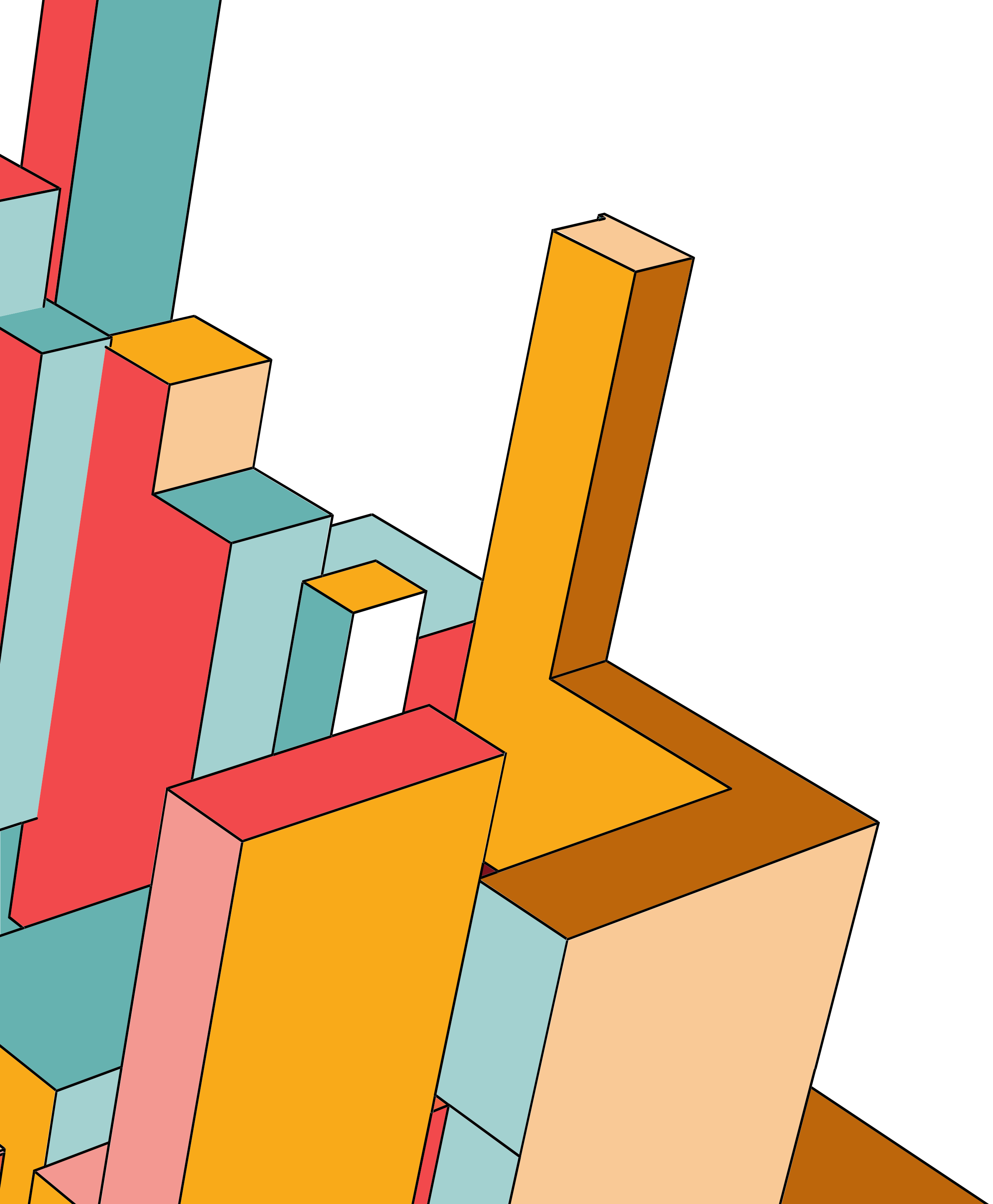
Data cleaning and data manipulation.

- Check & handle duplicate data
- Check & handle NA values and missing values.
- Drop columns, if it contains a large number of missing values and are not useful for the analysis.
- Imputation of the values, if necessary.
- Check & handle outliers in data

BUSINESS OBJECTIVE

- X EDUCATION WONTS TO KNOW MOST PROMISING LEADS.
- FOR THAT THEY WANT TO BUILD A MODEL WHICH IDENTIFIES THE HOT LEADS.
- DEPLOYMENT OF THE MODEL FOR THE FUTURE USE.

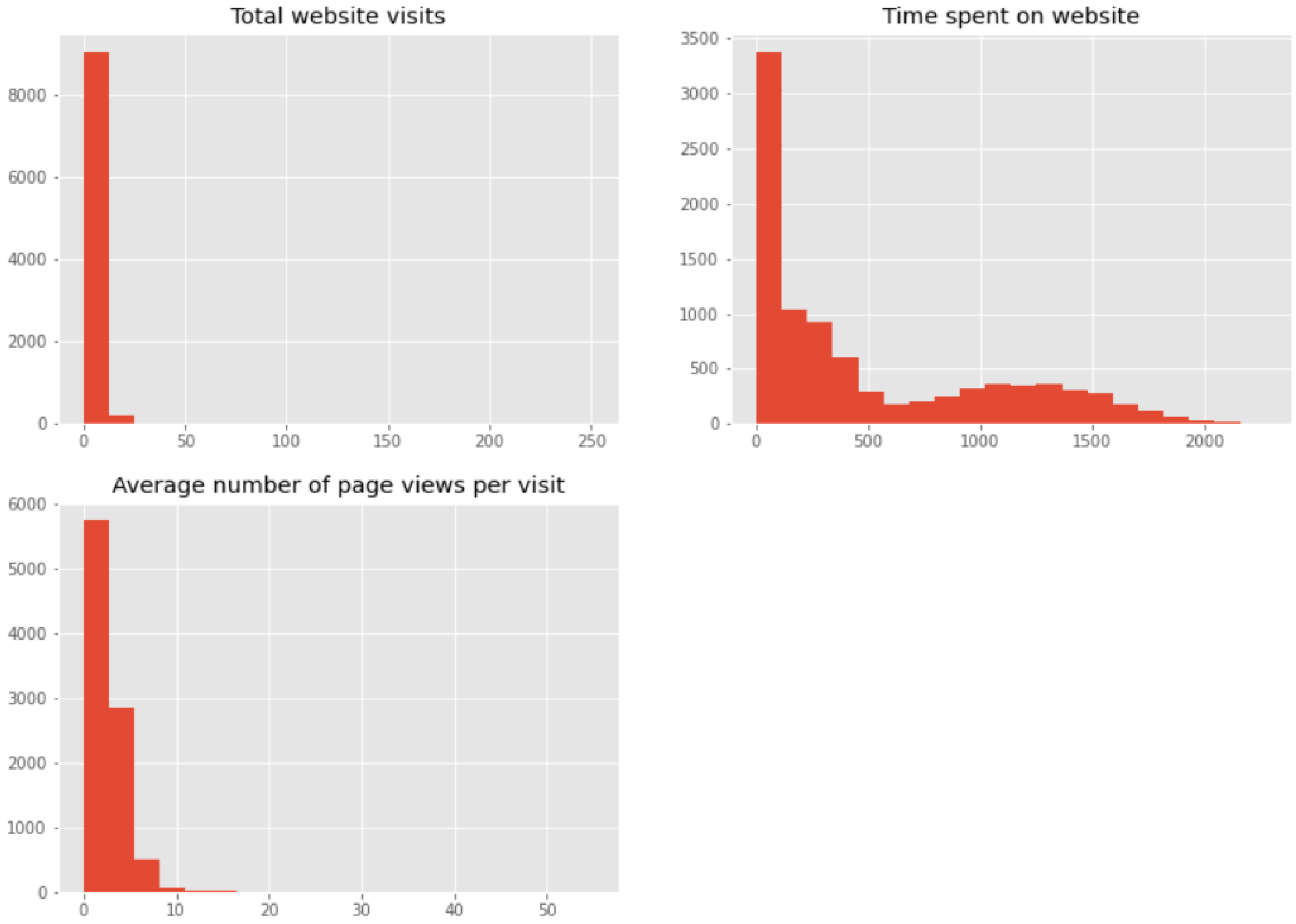




DATA MANIPULATION

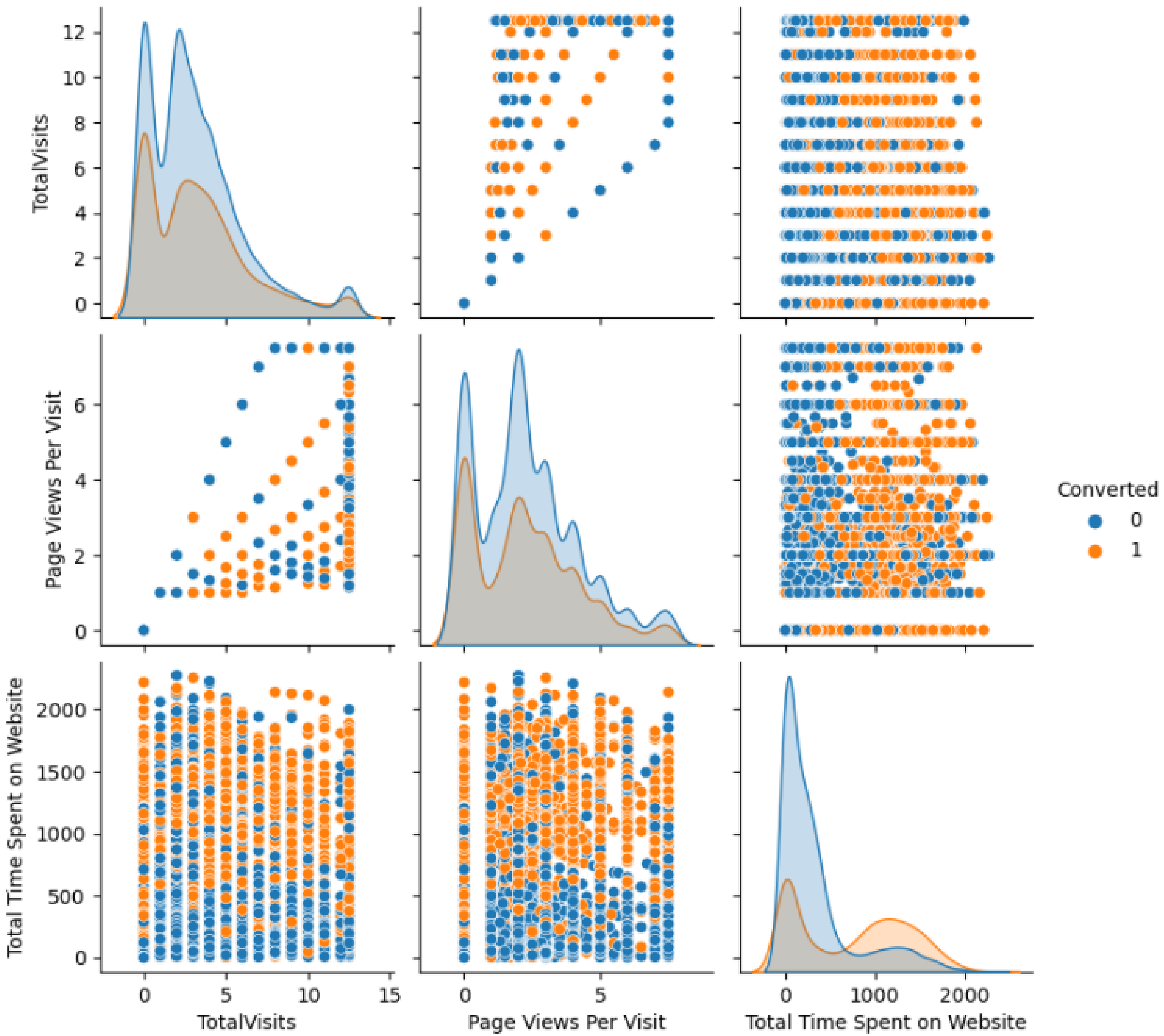
- Total Number of Rows=37,Total Number of Columns =9240.
- Single value features like "Magazine" , "ReceiveMoreUpdates About Our Courses" , "Update my supply"
- Chain Content" , "Get updates on DM Content" , "I agree to pay the amount through cheque" etc. have been dropped
- Removing the "ProspectID" & "Lead Number" which are not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which have enough variance, which have dropped, the features are: "Do Not Call" , "What matters most to you in choosing course" , "Search" , "Newspaper, Article" , "XEducation Forums" , "Newspaper" , "DigitalAdvertisement" etc.
- Dropping the column shaving more than 35% as missing values such as 'How did you hear about X Education' & 'Lead Profile'.

EDA

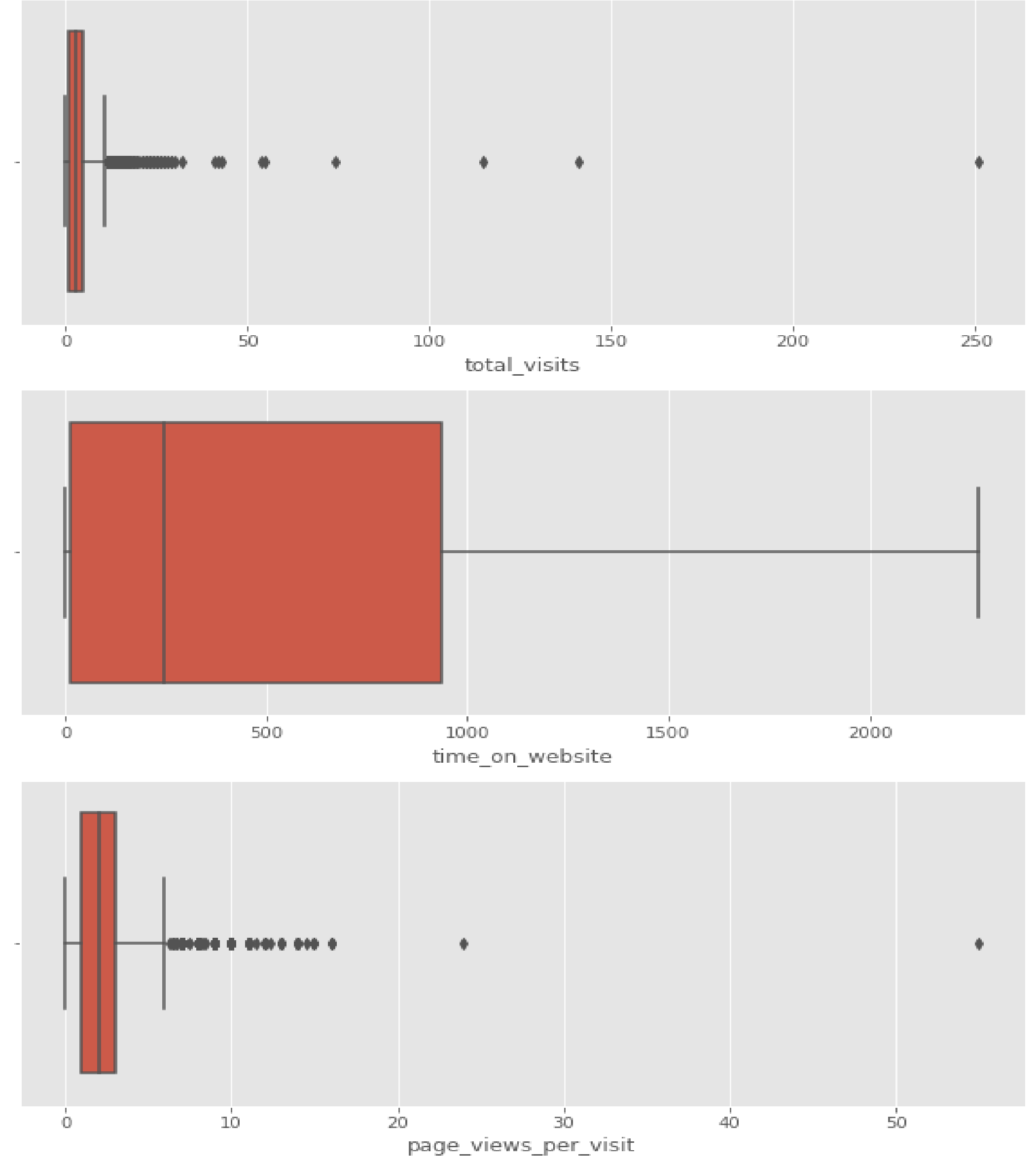


Observation

High peaks and skewed data. There might be a possibility of outliers.


TotalVisits
Page Views Per Visit
Total Time Spent on Website
Converted
● 0
● 1
0
2
4
6
8
10
12
0
5
10
15
0
2
4
6
0
500
1000
1500
2000

CHECK FOR OUTLIERS

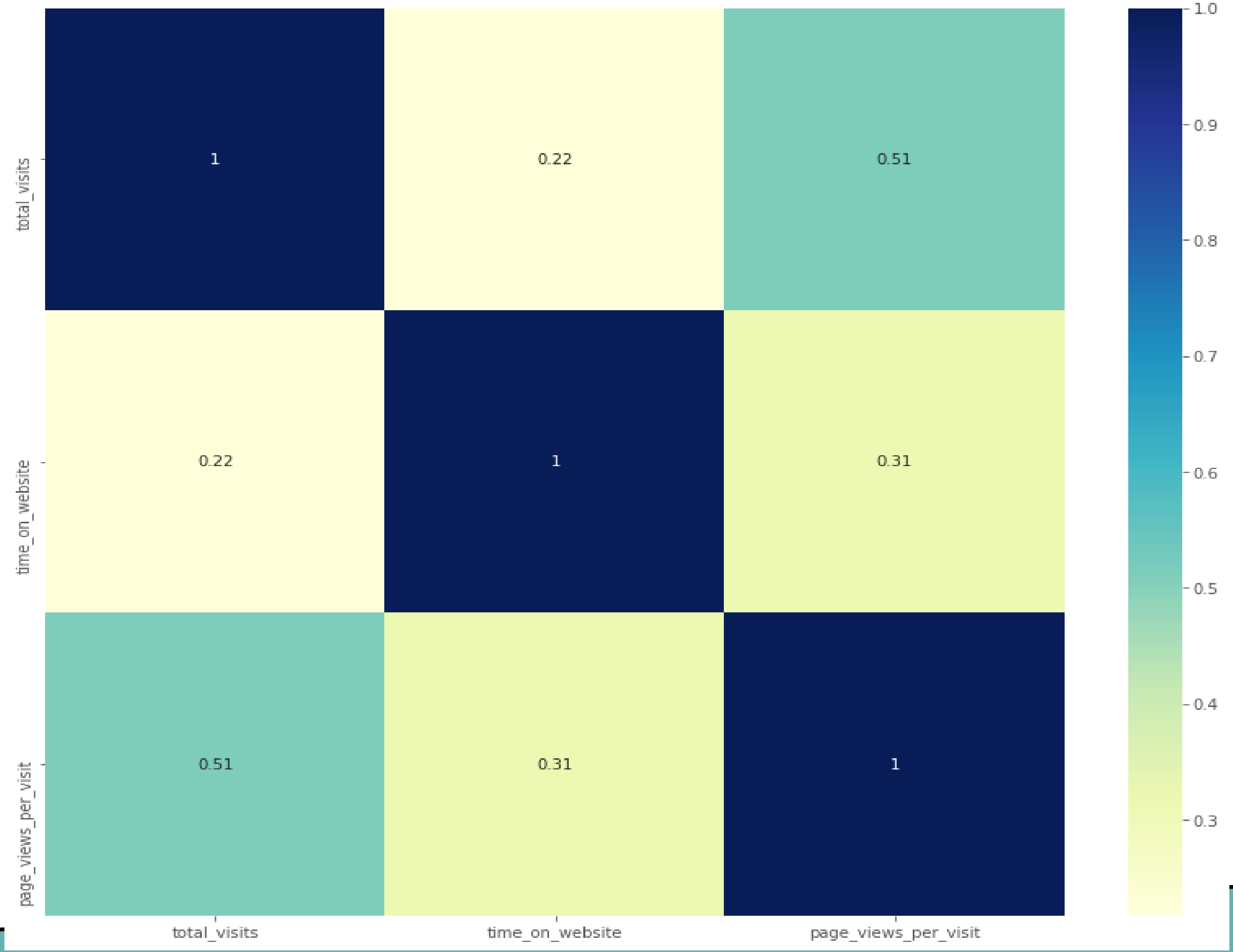


Observation

Looking at both the box plots there are upper bound outliers in both total_visits & page_views_per_visit columns.

We can also see that the data can be capped at 99 percentile.

HEATMAP

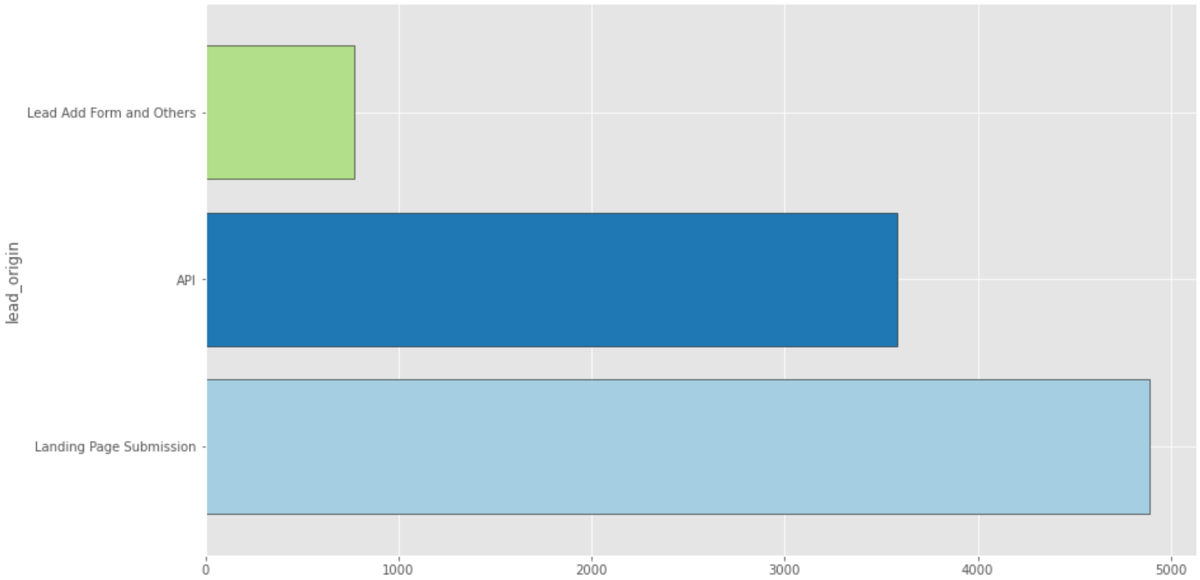


Observation

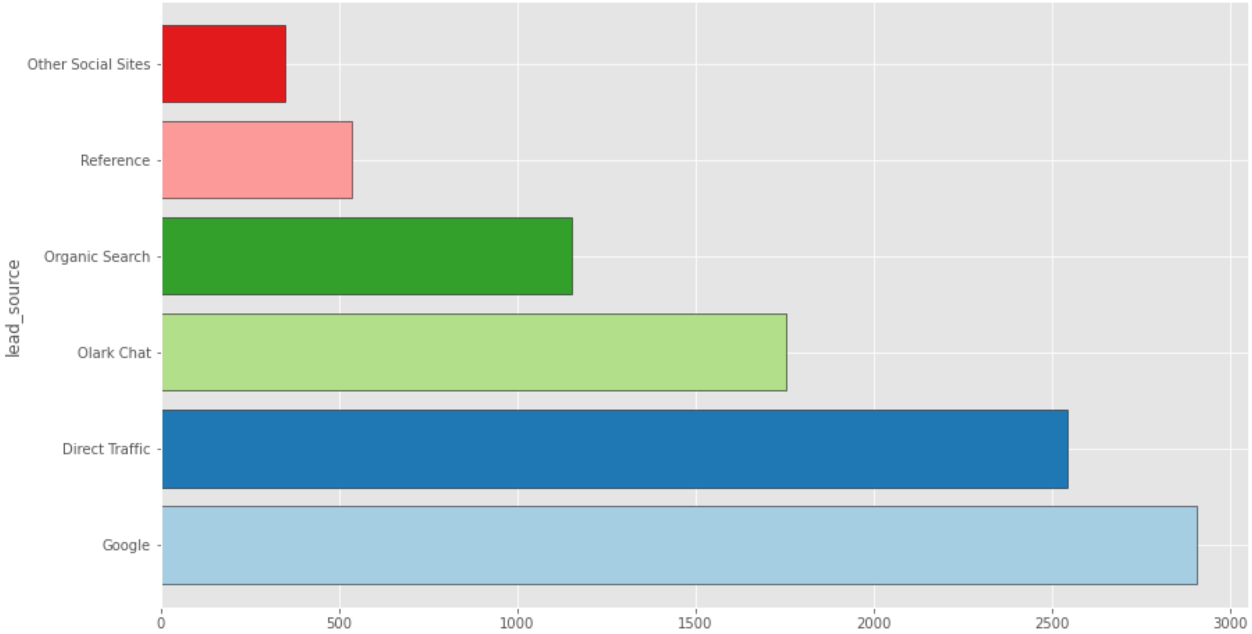
No significant correlation such that columns can be dropped

OTHER GRAPHS

LEAD ORIGIN

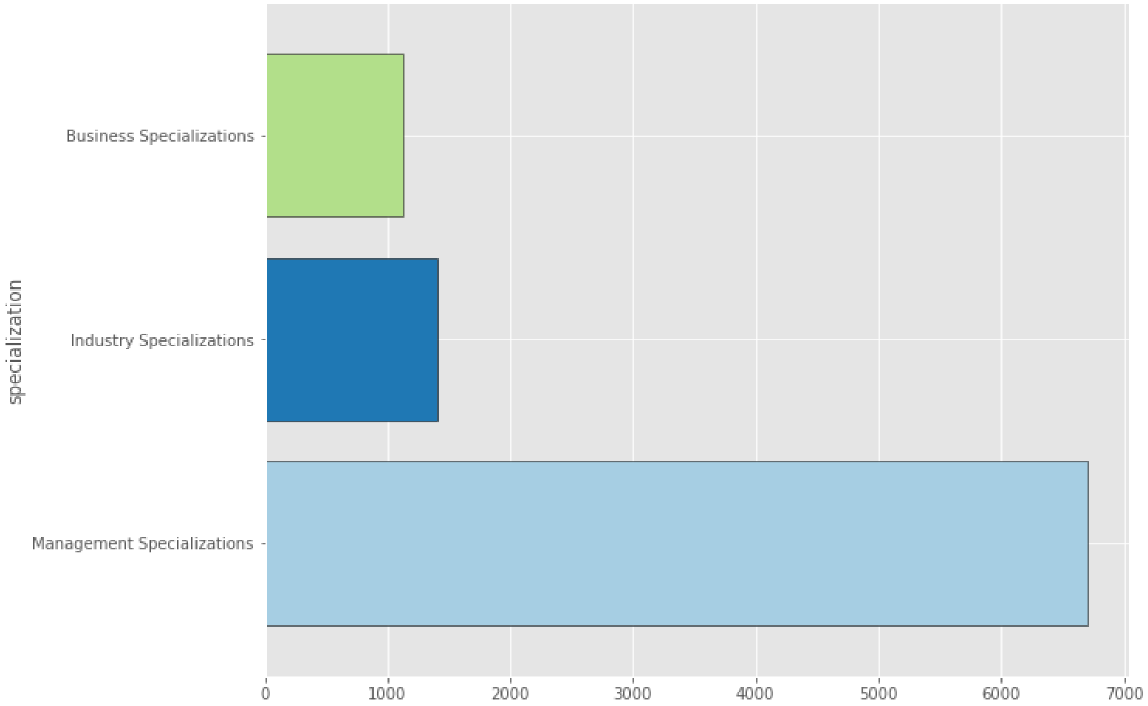


LEAD SOURCE

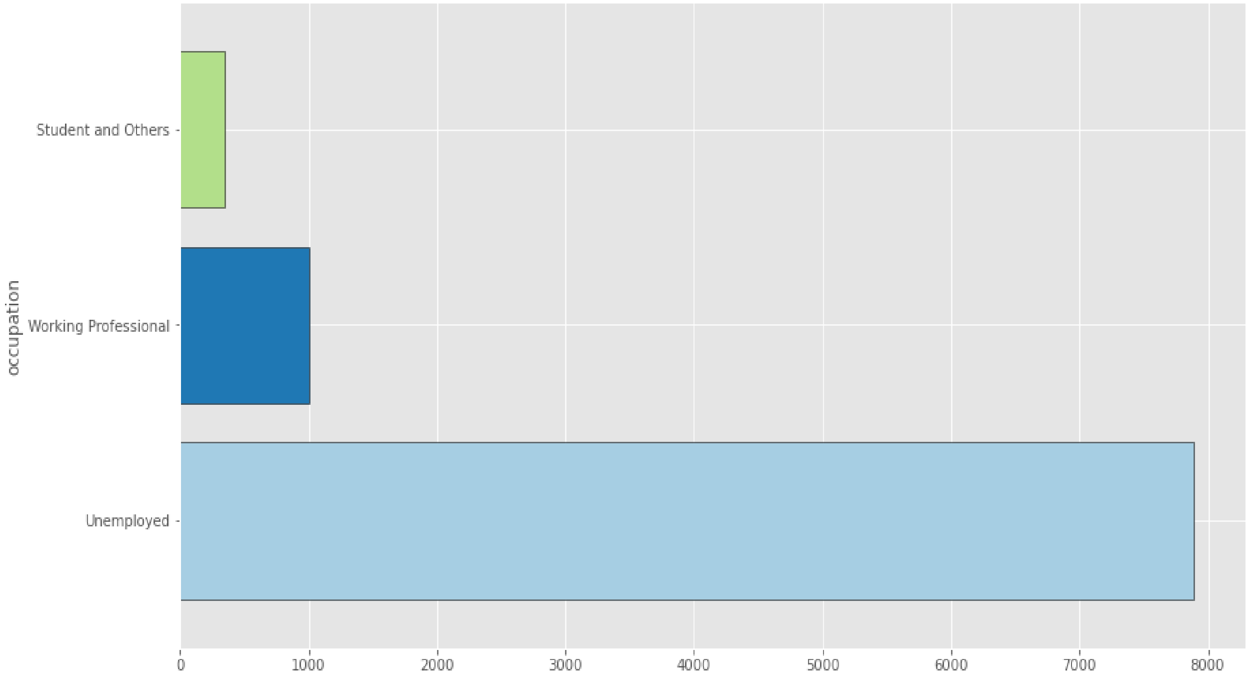


OTHER GRAPHS

SPECIALIZATION

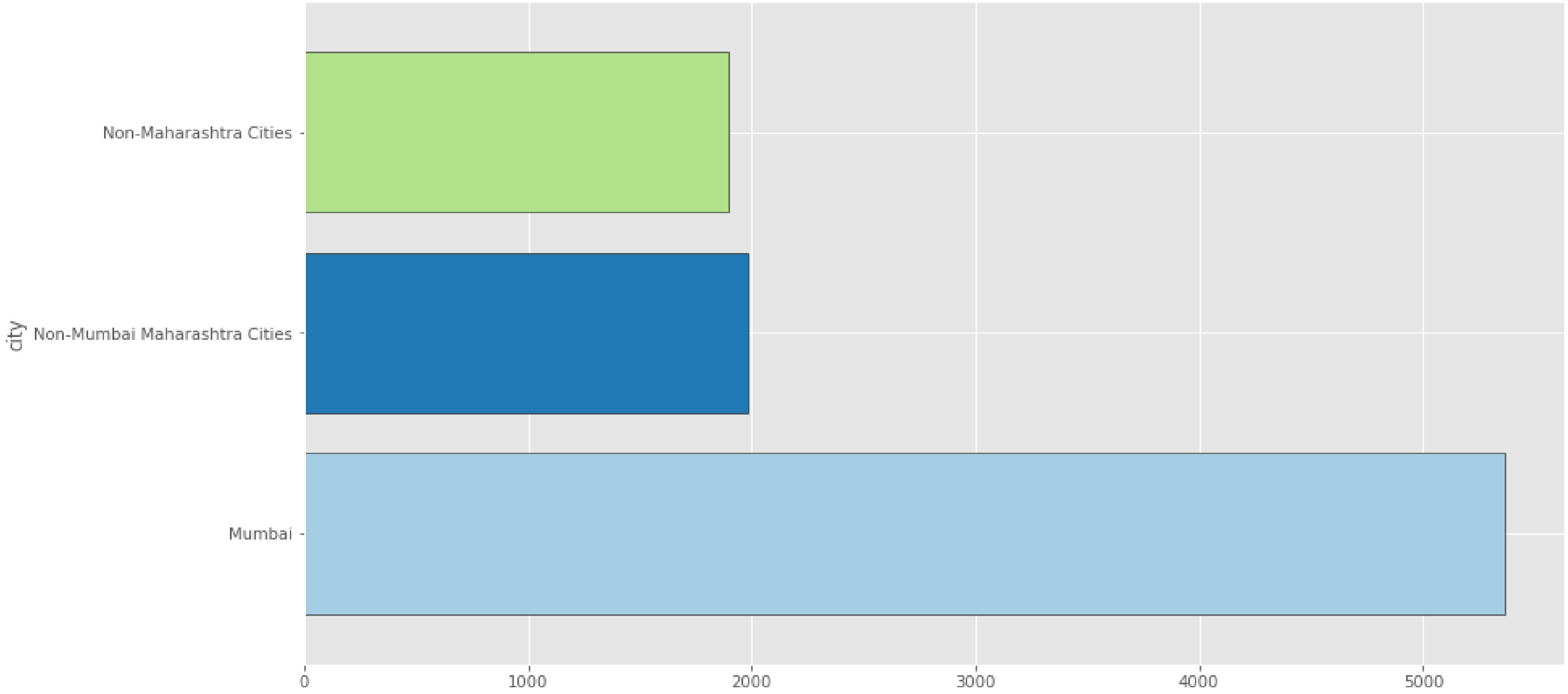


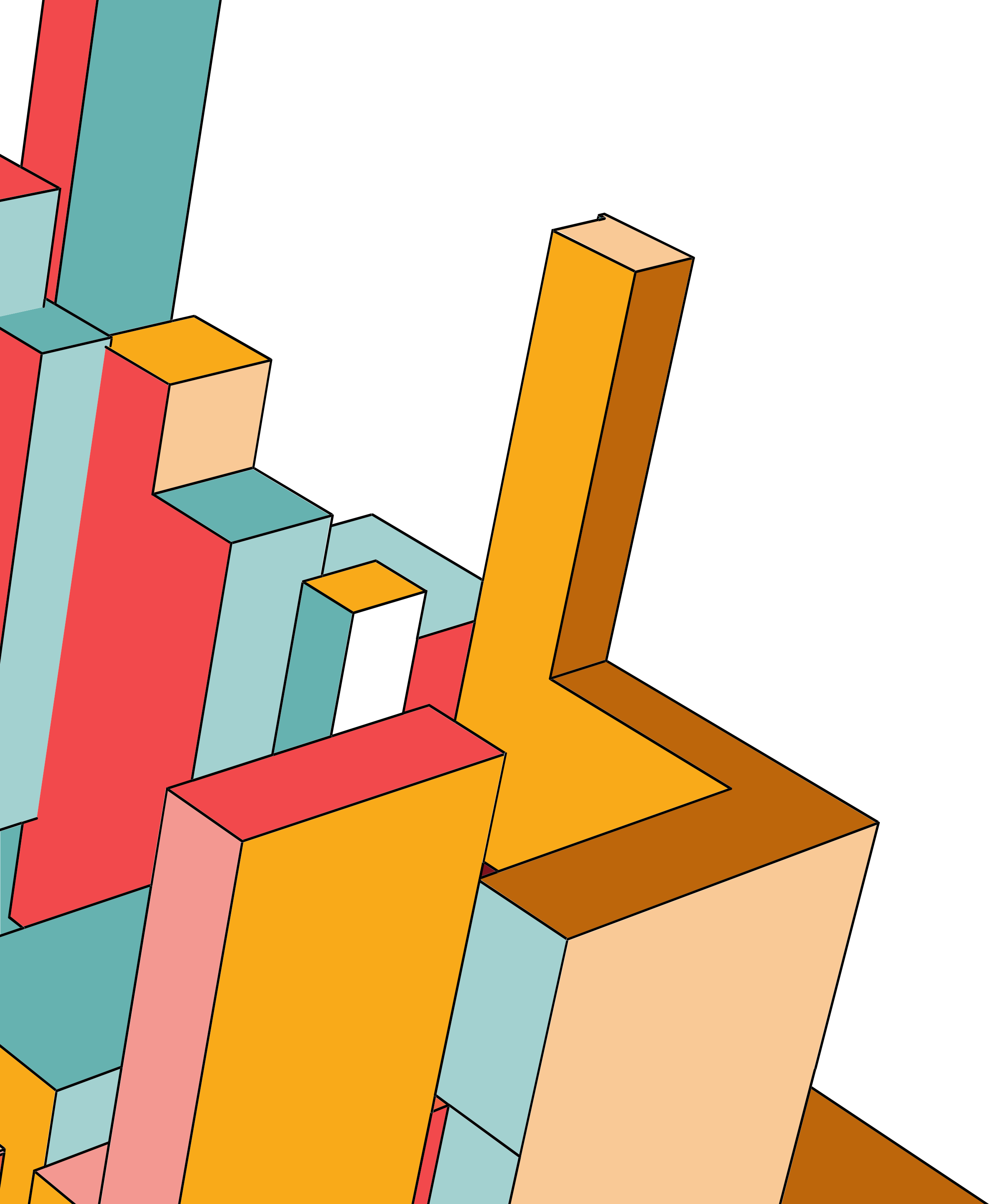
OCCUPATION



OTHER GRAPHS

CITY

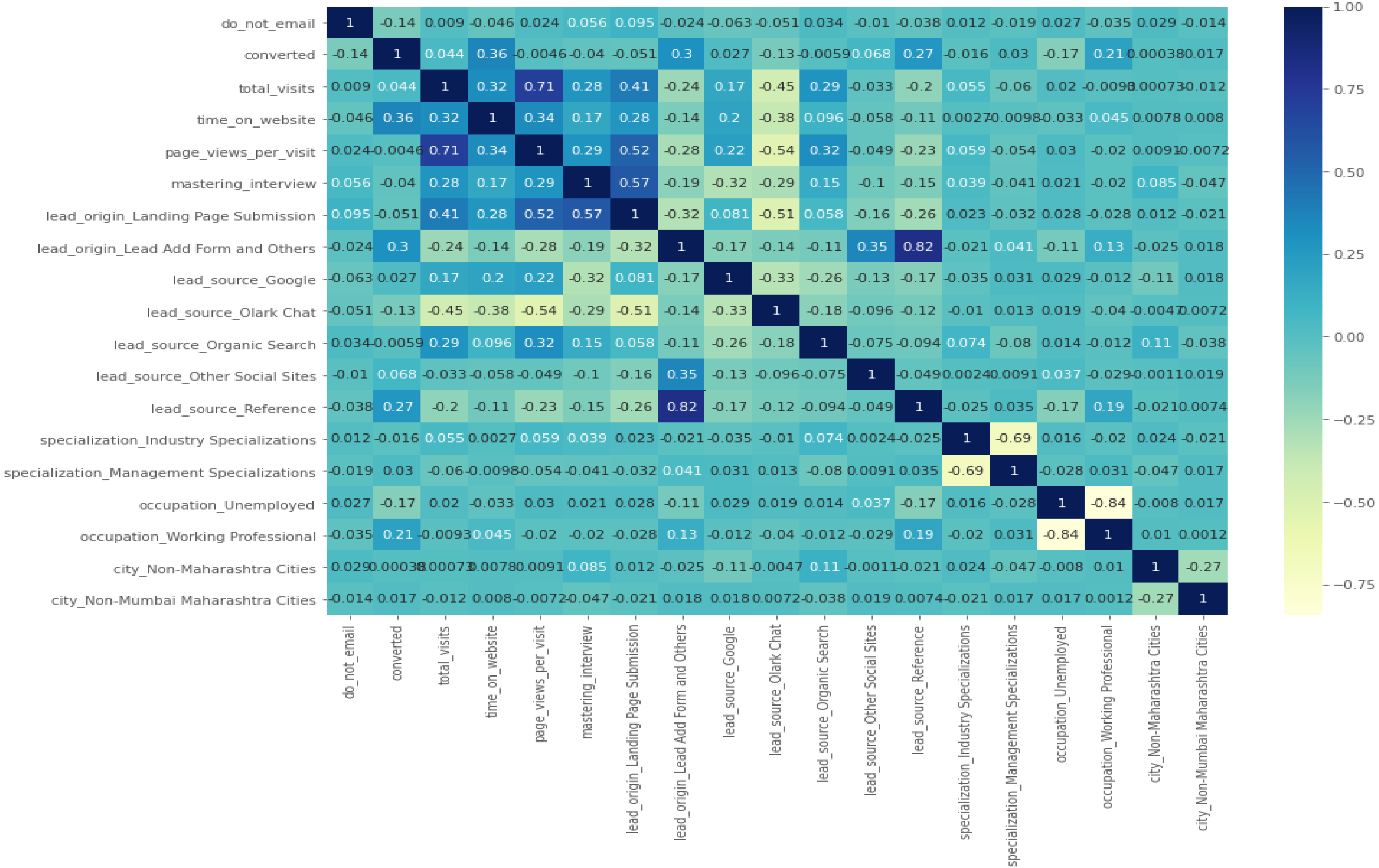




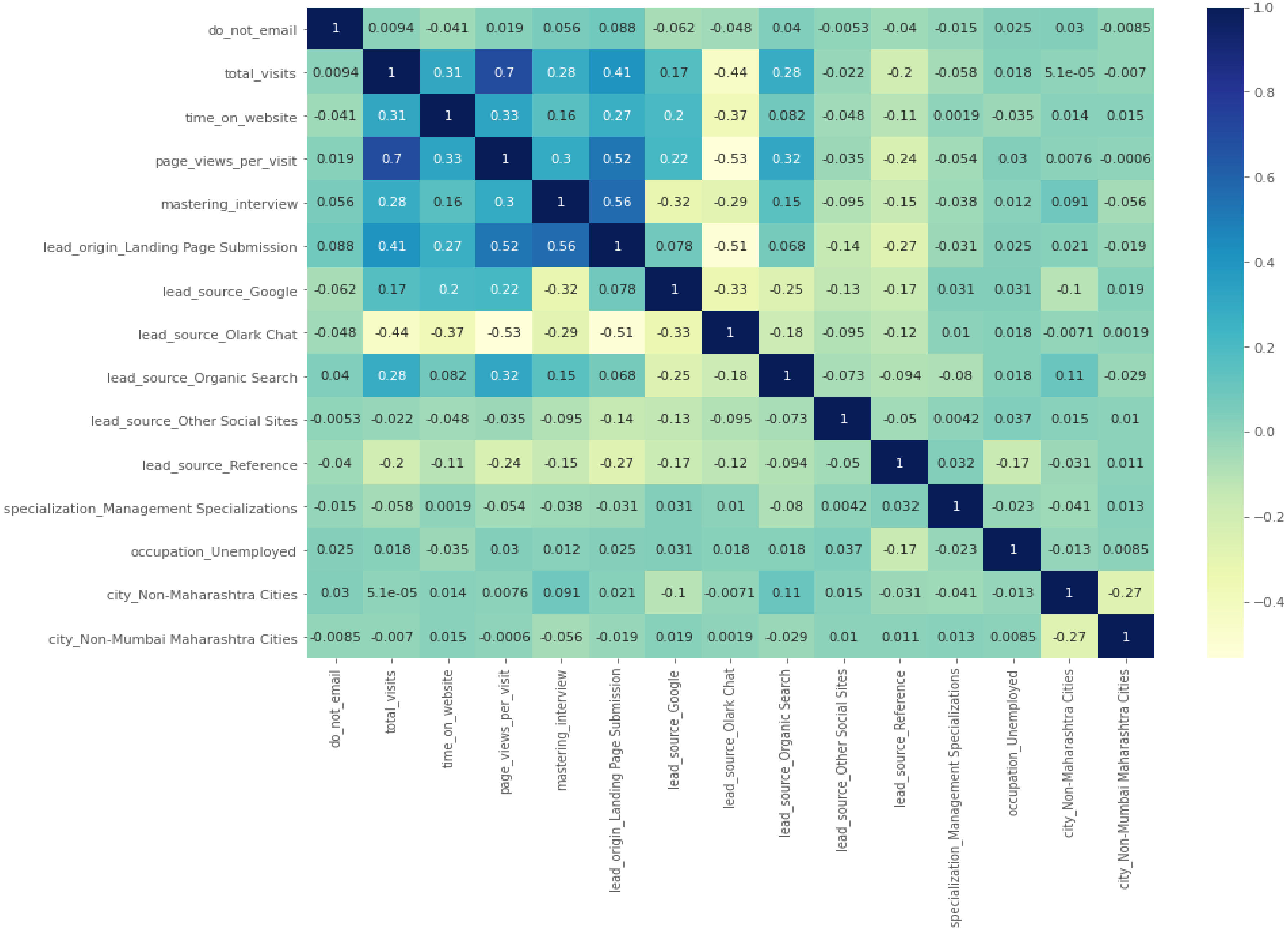
DATA PREPERATIONS

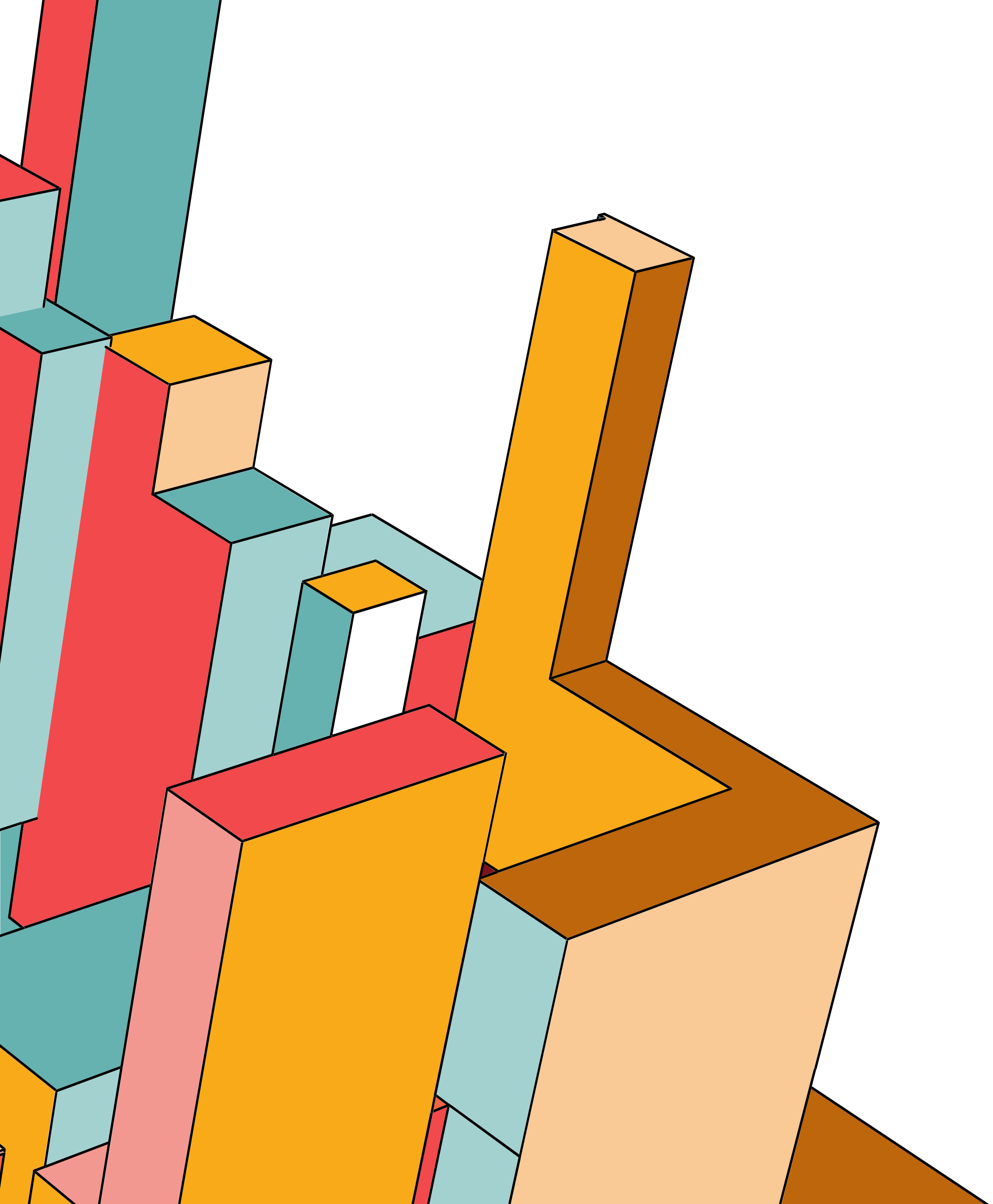
- Numerical Variables are normalized
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 37

LOOKING AT CORRELATIONS



AFTER DROPPING HIGHLY CORRELATED DUMMY VARIABLES



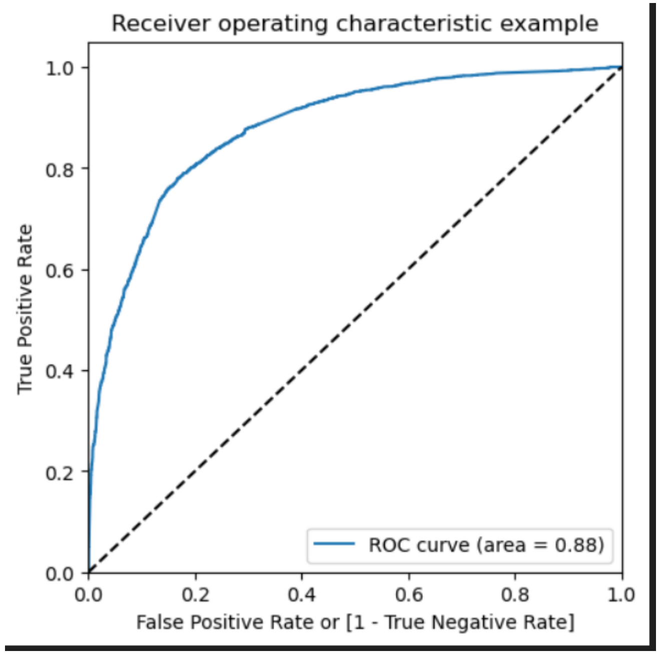


MODEL BUILDING

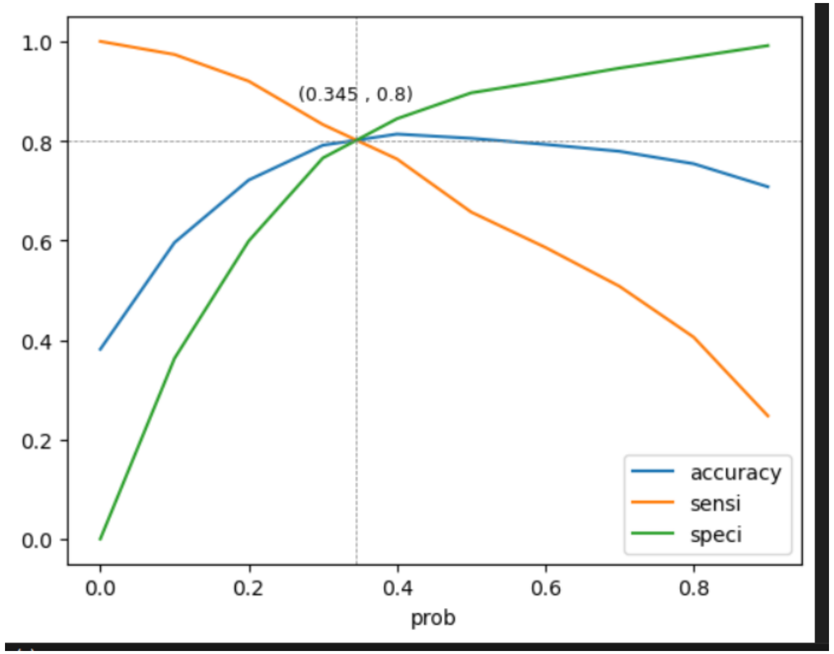
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p-value is greater than 0.05 and vi value is greater than 5
- Predictions on test data set
- Overall accuracy 81%

ROC CURVE

ROC CURVE FOR TRAIN SET



ACCURACY SENSITIVITY AND SPECIFICITY FOR VARIOUS PROBABILITIES

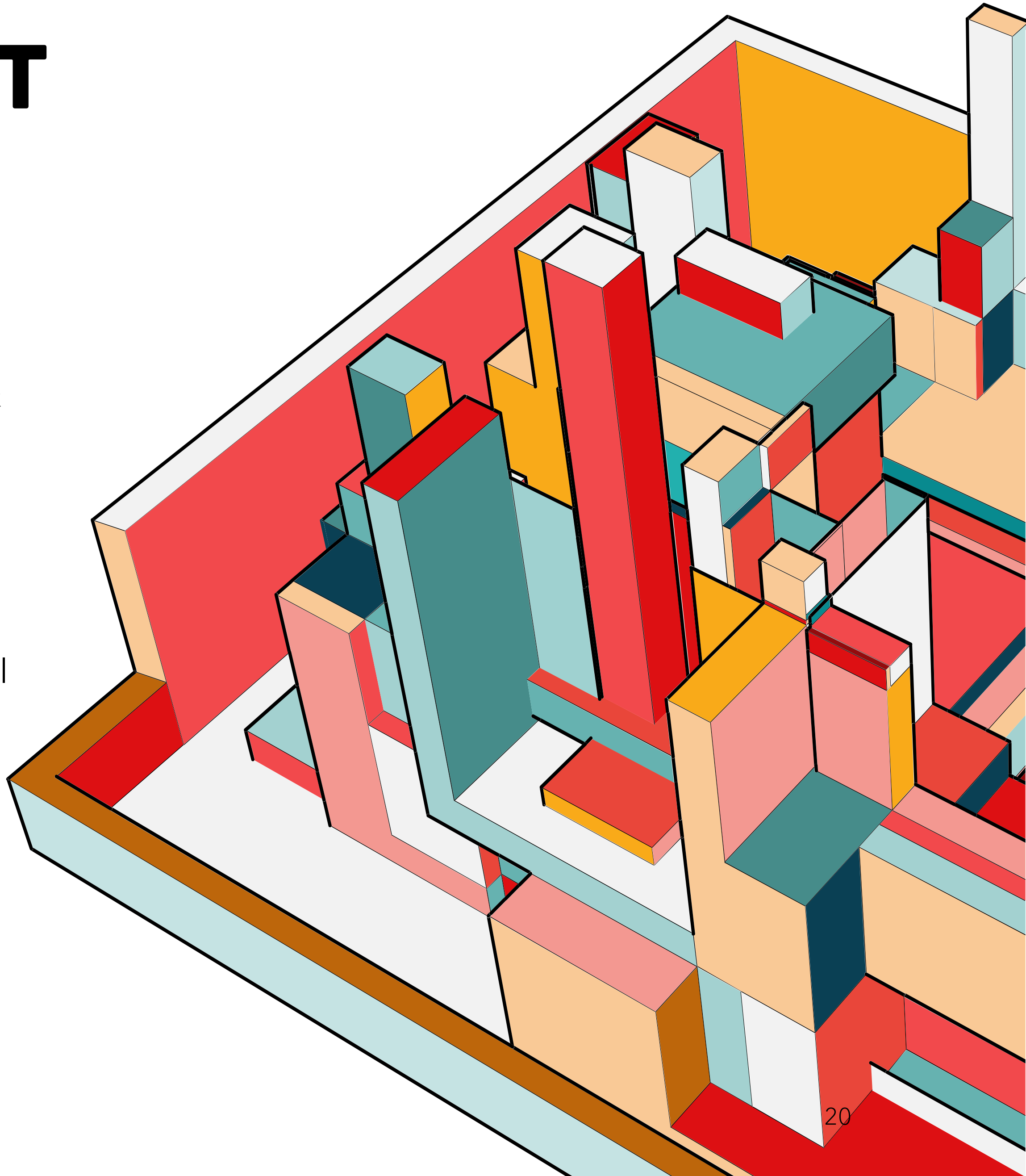


CONCLUSIONS ON ROC-CURVE

- Finding Optimal Cut off Point
- Optimal cut-off probability is that
- Probability where we get balanced sensitivity & specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.

PREDICTION ON TEST SET

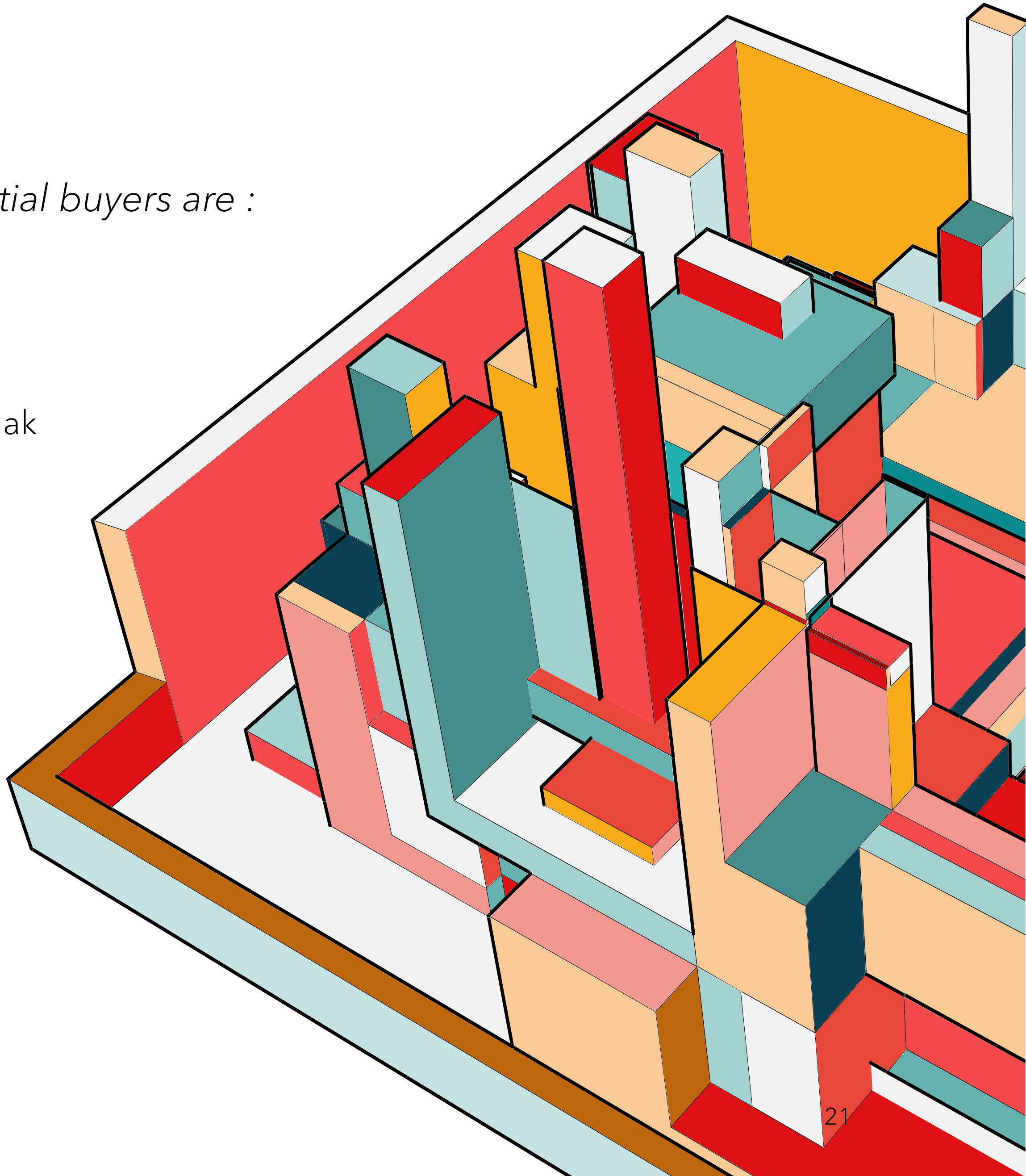
- Before predicting on the test set, we need to standardize the test set & need to have exact same columns present in our final train dataset.
- After this we did model evaluation i.e. finding the accuracy, precision, & recall
- The accuracy score we found was 0.82, precision 0.75, & recall 0.75 approximately.
- This shows that our test prediction is having accuracy, precision, & recall scores in an acceptable range

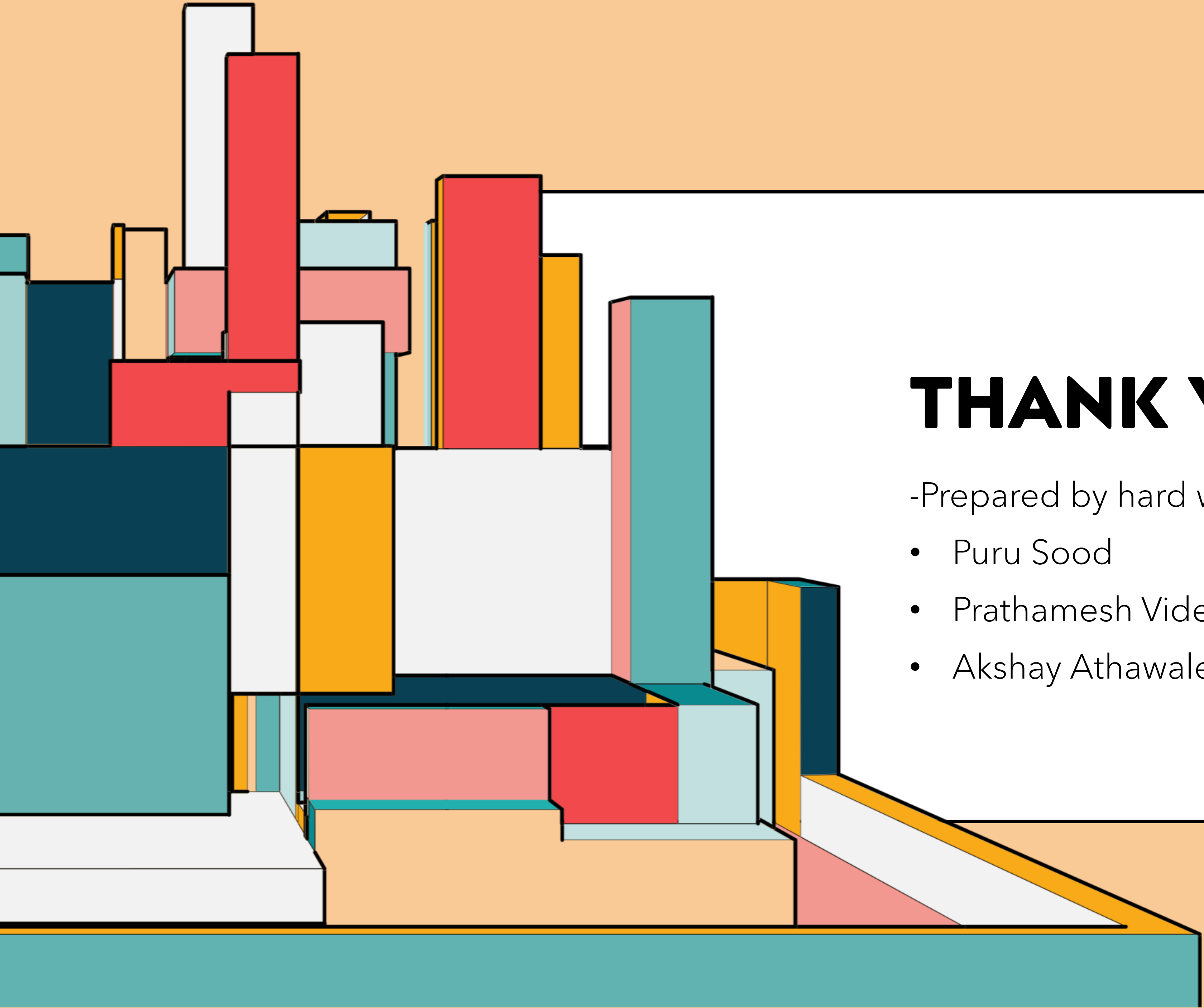


CONCLUSION

It was found that the variables that mattered the most in the potential buyers are :

- The total time spent on the Website.
- When the lead source was: Google Direct traffic Organic search Welingak website
- When the last activity was: SMS Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional





THANK YOU

-Prepared by hard work of

- Puru Sood
- Prathamesh Videkar
- Akshay Athawale