LEAD SCORING ASSIGNMENT

Student: Pham Thieu Quan (phamthieuquan.ivf@gmail.com)

STRUCTURE OF THIS ANALYTIC

PART A **UNDERSTANDING THE DATA AND DATA PREPARATION**

1. Data visualization
2. Check the dataset for completeness

PART B **DATA ANALYSIS**

1. Use statistical techniques to describe the data
2. Data Cleaning
3. Exploratory Data Analysis
4. Data Preparation for model

PART C **BUILDING A LINEAR REGRESSION MODEL**

1. Divide the data into training and testing sets
2. Use the training set to build a linear regression model
3. Evaluate the model's performance on the testing set

PART D **PRESENTING THE RESULTS**

1. Assignment-based Subjective Questions
2. Present Goals of the Case Study

PART A UNDERSTANDING THE DATA AND DATA PREPARATION

Problem Introduction

Challenge:

Low lead conversion rate (30%) despite a high number of leads.

Need to identify "Hot Leads" for efficient conversion improvement.

Desired Outcome:

Build a lead scoring model that assigns scores to leads, predicting their conversion likelihood.

Target lead conversion rate of 80%.

Data:

Information about leads, including their source, website interaction, demographics, etc.

Label indicating whether a lead converted (1) or not (0).

Actionable Points:

Analyze potential features for the model (e.g., website visits, time spent, source).

Recommend suitable machine learning algorithms (e.g., logistic regression, random forest).

Provide an example of building a lead scoring model using a specific technique.

Analyzing scenario
What to be concerned?

Data

- Data Quality: Check for missing values, outliers, inconsistencies, and biases. Missing values in crucial features like "source" or "specialization" might need careful handling.
- Feature Relevance: Evaluate the relevance of each feature to predicting lead conversion. Some features like "city" might have limited predictive power.
- Feature Relationships: Explore potential relationships between features. Multicollinearity (correlated features) can affect model performance.
- Class Imbalance: If the dataset has significantly more unconverted leads than converted ones, address class imbalance to avoid biased models.

Analyzing scenario
What to be concerned?

Model Building:

- Algorithm Selection: Choose an appropriate algorithm based on data characteristics and desired model properties. For example, logistic regression is good for interpretability, while random forest might capture non-linear relationships.
- Hyperparameter Tuning: Optimize model hyperparameters to improve performance. Overfitting, underfitting, and regularization are key concerns.
- Model Evaluation: Use appropriate metrics like AUC-ROC and confusion matrix to assess model performance. Evaluate on unseen data (test set) to avoid overfitting.
- Feature Importance: Understand which features contribute most to lead scoring to refine targeting strategies.

Analyzing scenario
What to be concerned?

Key Considerations:

- Business Context: Align the model with X Education's specific goals and constraints.
- Ethical Considerations: Be aware of potential biases in data and model outputs and address them responsibly.
- Privacy and Security: Ensure data privacy and security throughout the analysis and model deployment.
- Iterative Process: Data analysis and model building are iterative processes. Expect to revisit and refine your approach as you learn more.

Assignment-based Subjective Questions

PART A - UNDERSTANDING THE DATA AND DATA PREPARATION

What to be concerned?

1. Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?
2. What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?
3. X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.
4. Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Some observation

1. There is no duplicate data

2. Missing Values: A significant number of columns contain missing values. These will require handling, potentially through removal or imputation.

3. Duplicate Information: Prospect ID and Lead Number seem to serve the same purpose (unique identifiers). Consider keeping one and dropping the other to avoid redundancy.

4. Unfriendly Column Names: The current column names are excessively long. This can make data exploration and analysis more cumbersome. Modifying them for clarity is recommended.

5. Categorical Value Handling: Certain categorical columns contain the value "Select." This should be treated as equivalent to a missing value since it reflects the absence of a selection.

6. These observations highlight areas for cleaning and preprocessing the data before further analysis. By addressing these issues, I can prepare the data for more reliable and informative results.

Top Percentage of null values

country	26.63
specialization	36.58
source	78.46
occupation	29.11
course_selection_reason	29.32
tags	36.29
lead_quality	51.59
lead_profile	74.19
city	39.71
asymmetrique_activity_index	45.65
asymmetrique_profile_index	45.65
asymmetrique_activity_score	45.65
asymmetrique_profile_score	45.65

Some observation

1. A significant number of columns contain missing values, mainly in demographic information (country, city, specialization), lead scoring (lead_quality, lead_profile), and engagement (asymmetrique_* indices).
2. Some potentially crucial features like lead_source and specialization have moderate missingness, which might require careful imputation or exclusion depending on its impact on your analysis.
3. Columns with low missingness can be directly used for further analysis.

Recommendations

1. Analyze the distribution of missing values within each column (missing completely at random, missing not at random) to guide imputation strategies.
2. Explore the relationship between missing values and other variables to assess potential biases.
3. Decide whether to impute missing values based on the analysis and the importance of each feature for your goals.
4. Consider dropping columns with a very high percentage of missing values if imputation is not feasible or appropriate.

PART B - DATA ANALYSIS

Handle categorical columns with high number of missing values

- → Drop columns that have null values > 40% or Sales generated columns
- → The top 5 null columns remaining as follow

country	26.63
specialization	36.58
occupation	29.11
course_selection_reason	29.32
city	39.71

PART B - DATA ANALYSIS

Handle categorical columns with low number of missing values

1. Merge categories that have low representation
2. Impute the missing values

Handle Binary columns

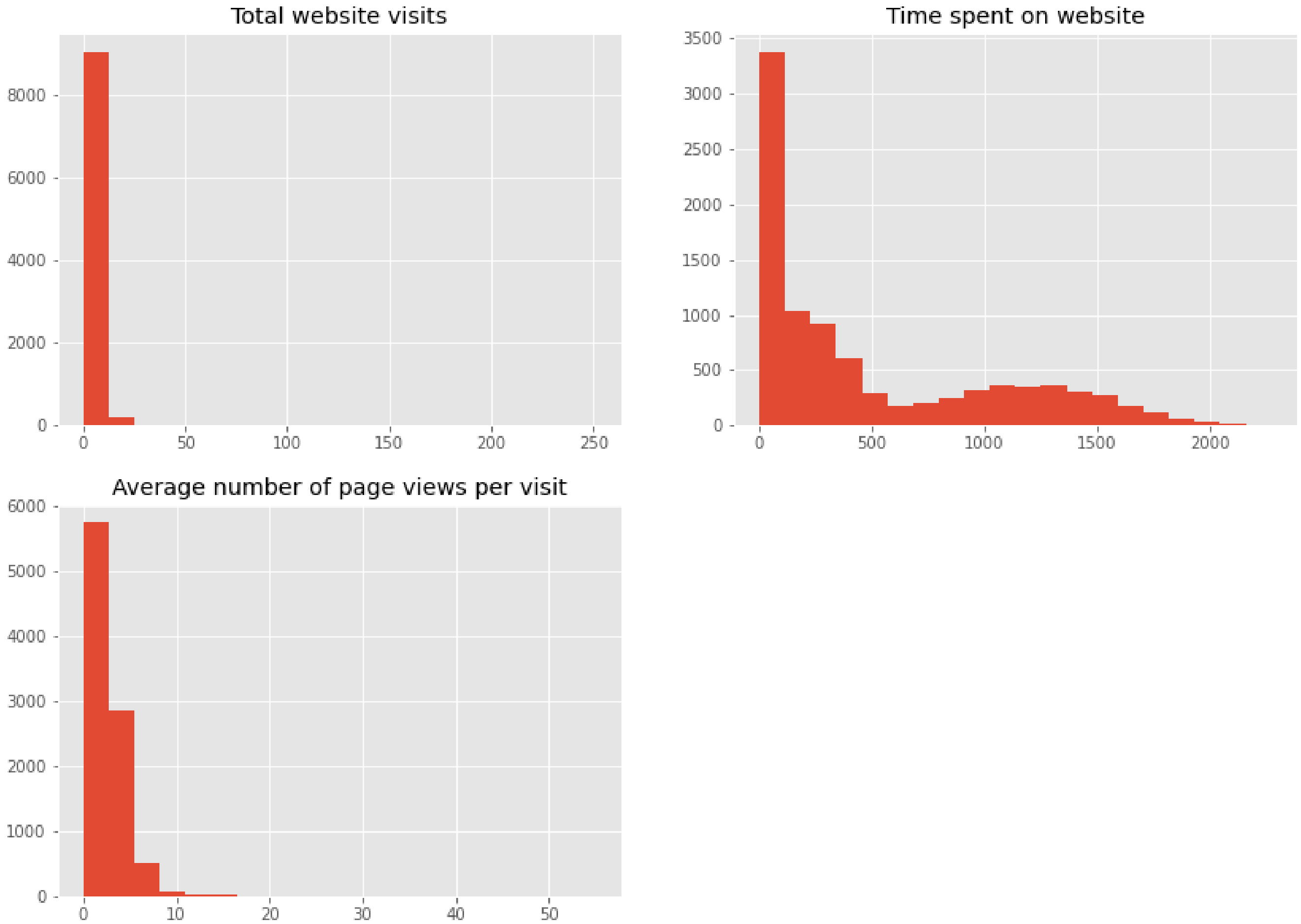
Drop those columns that have significant data imbalance
Drop all those columns that have only 1 unique entry

Handle Numerical columns

PART B - DATA ANALYSIS

Exploratory Data Analysis

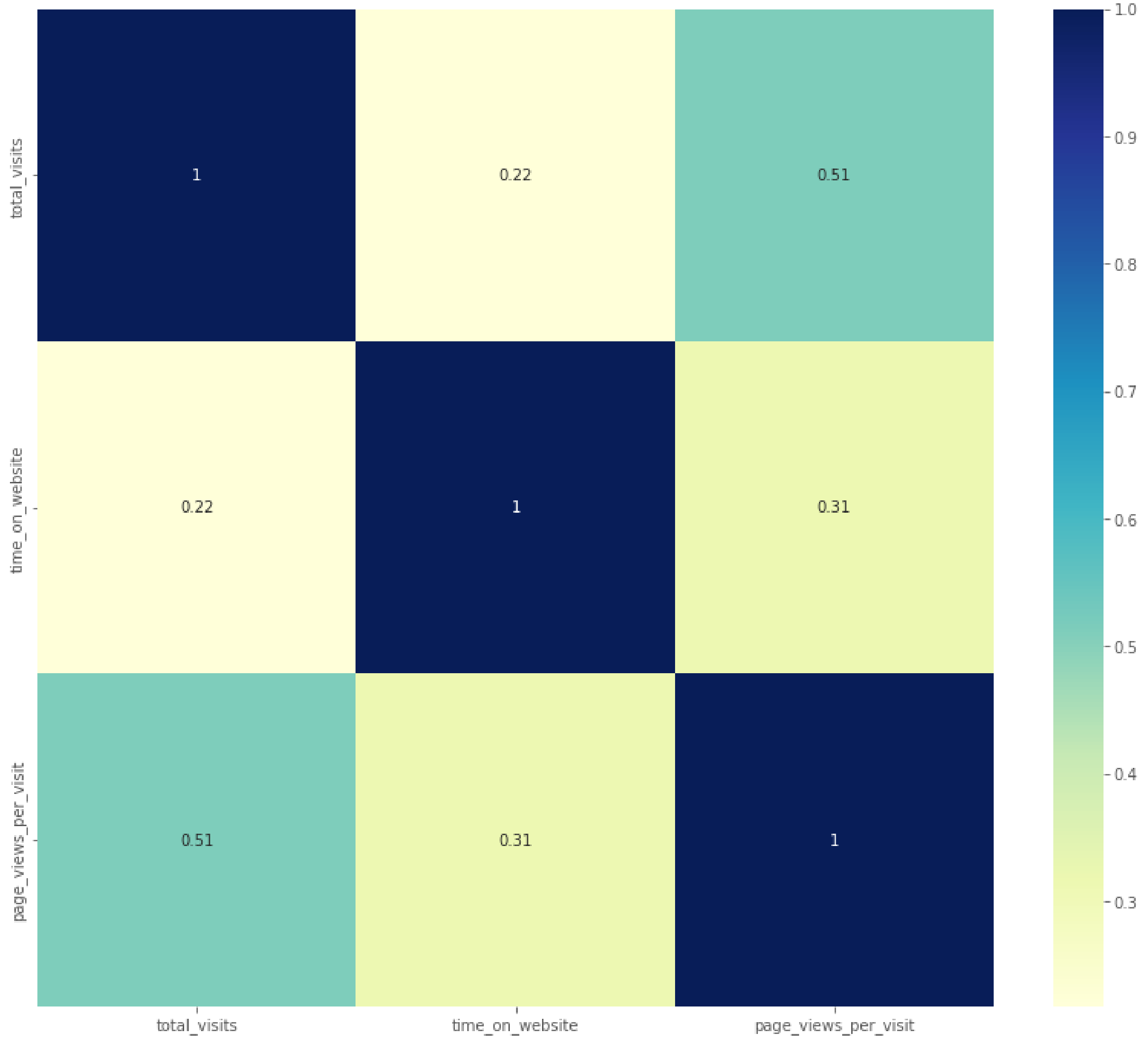
1.



PART B - DATA ANALYSIS

Exploratory Data Analysis

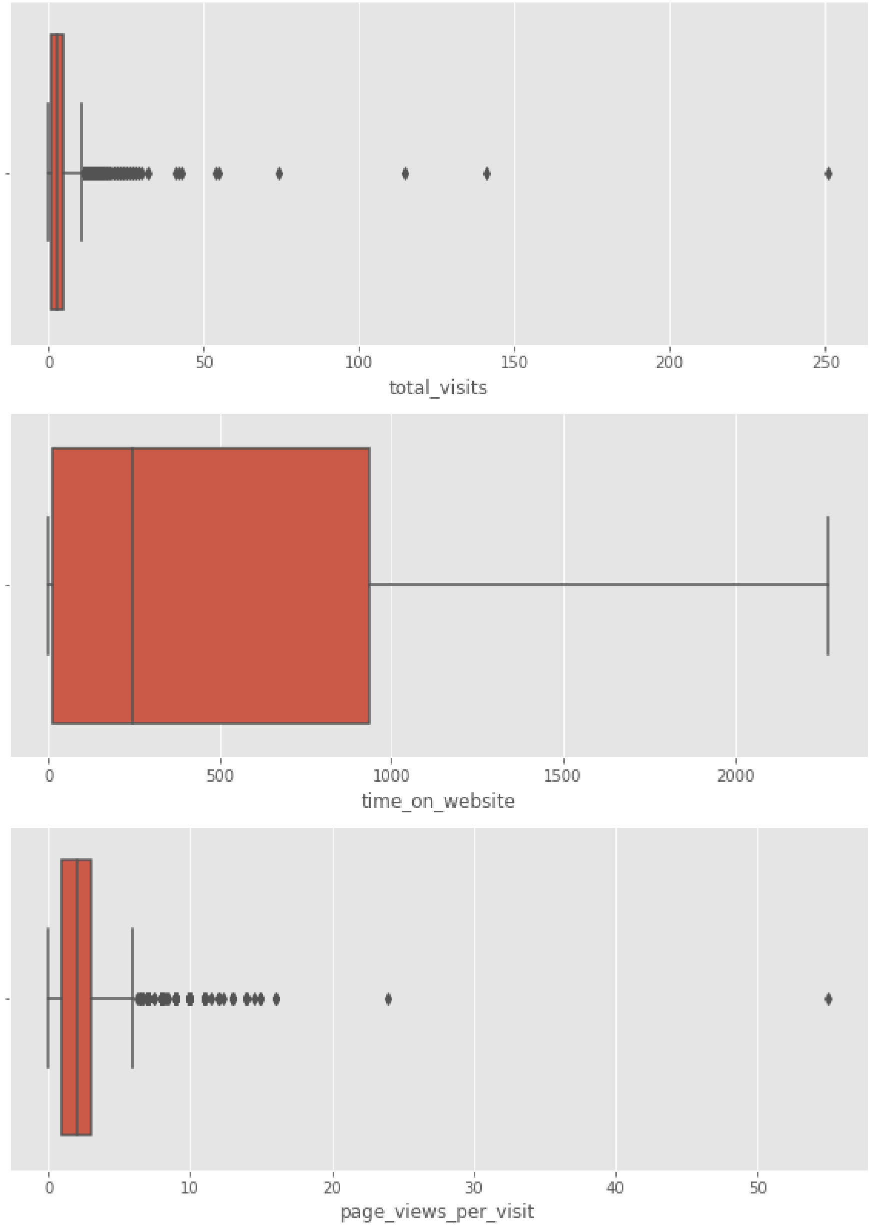
1.



PART B - DATA ANALYSIS

Exploratory Data Analysis

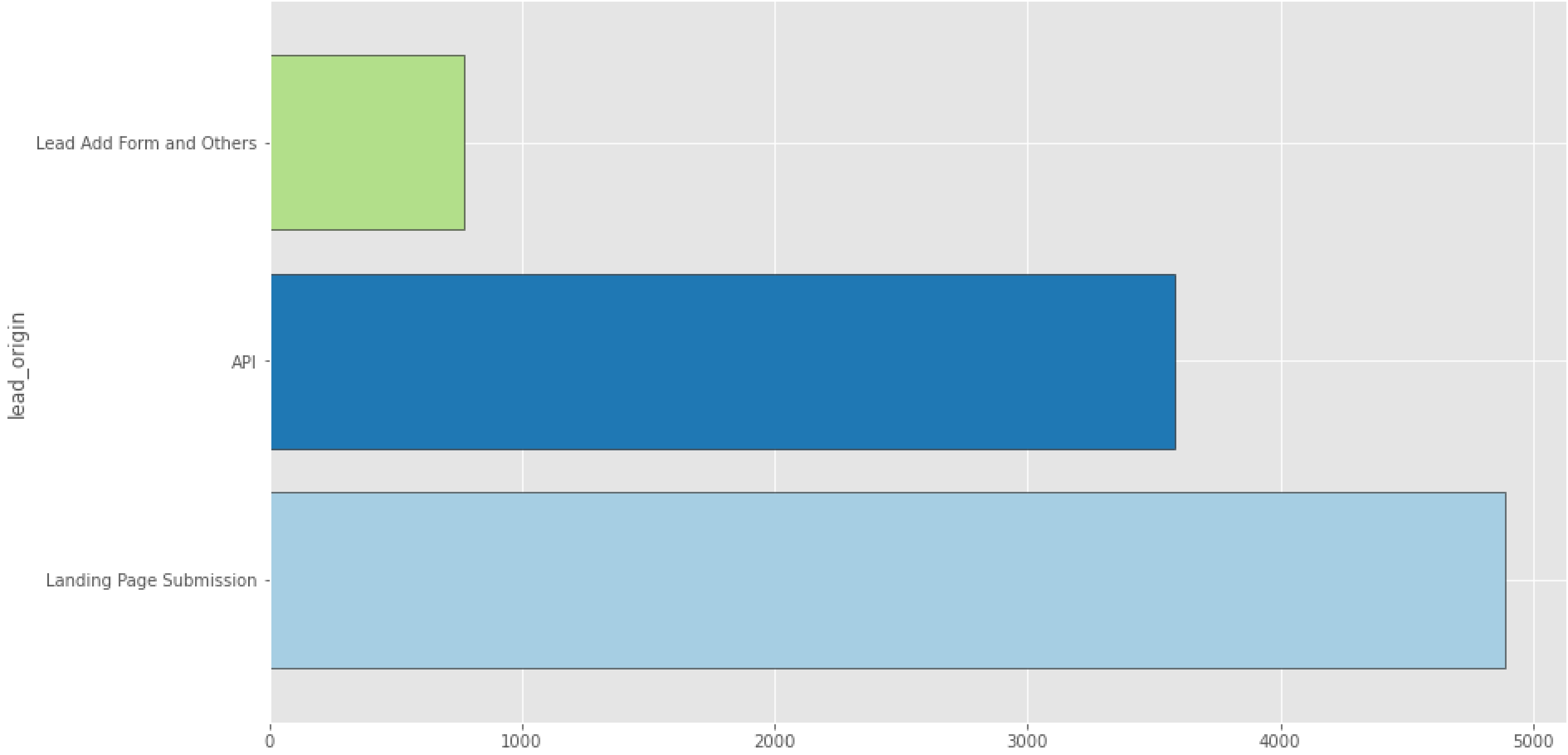
1.



PART B - DATA ANALYSIS

Exploratory Data Analysis

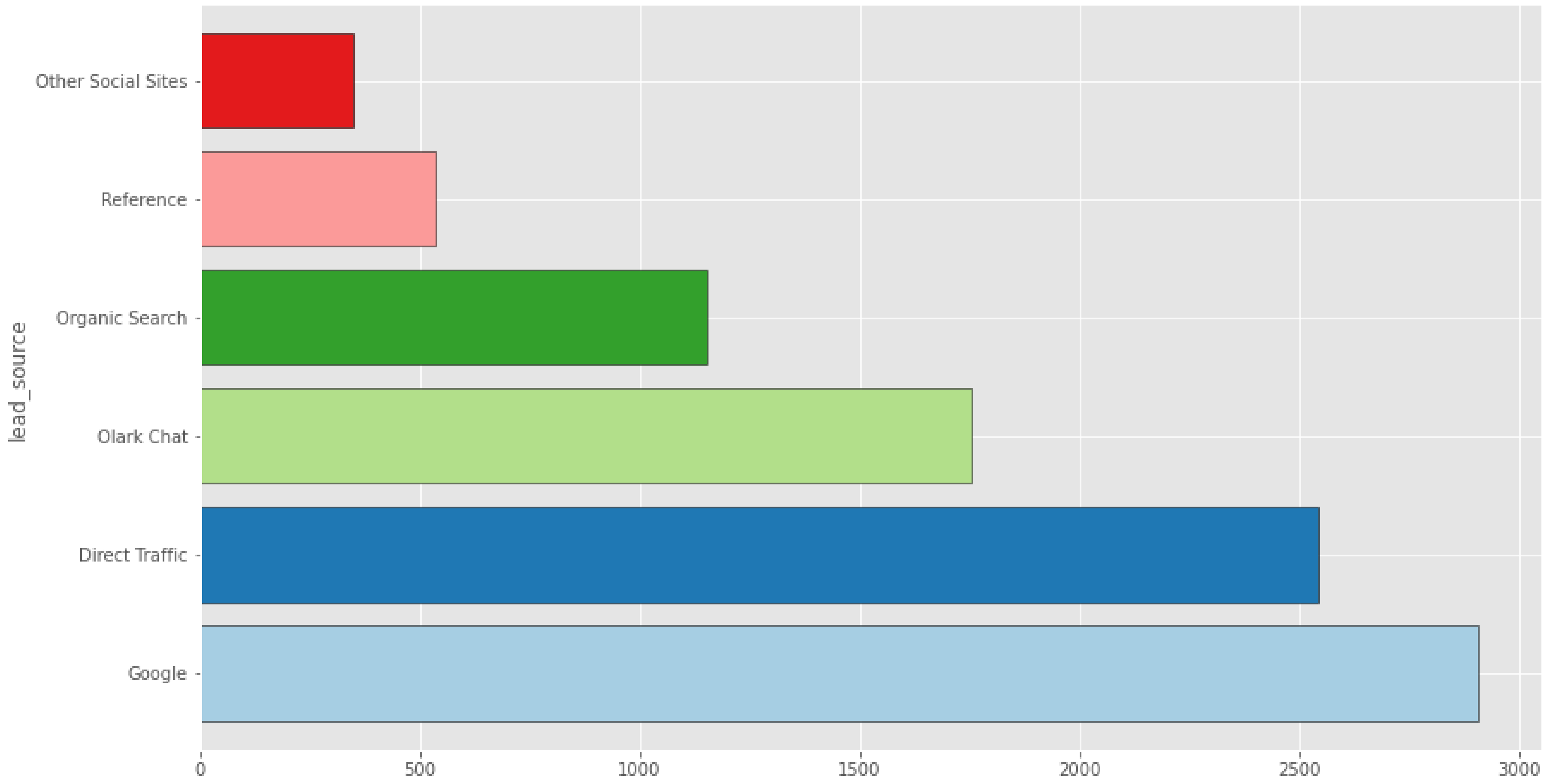
1.



PART B - DATA ANALYSIS

Exploratory Data Analysis

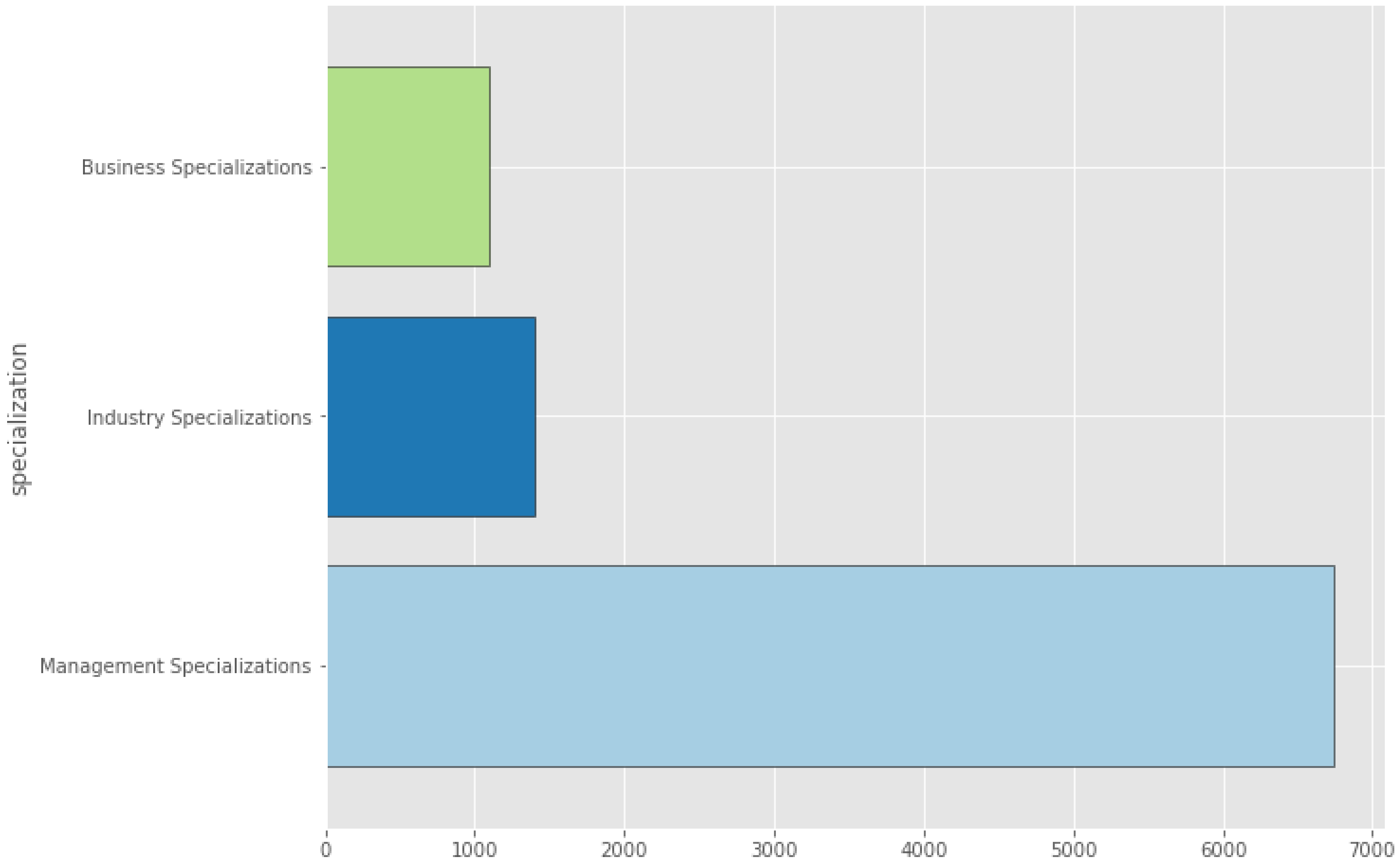
1.



PART B - DATA ANALYSIS

Exploratory Data Analysis

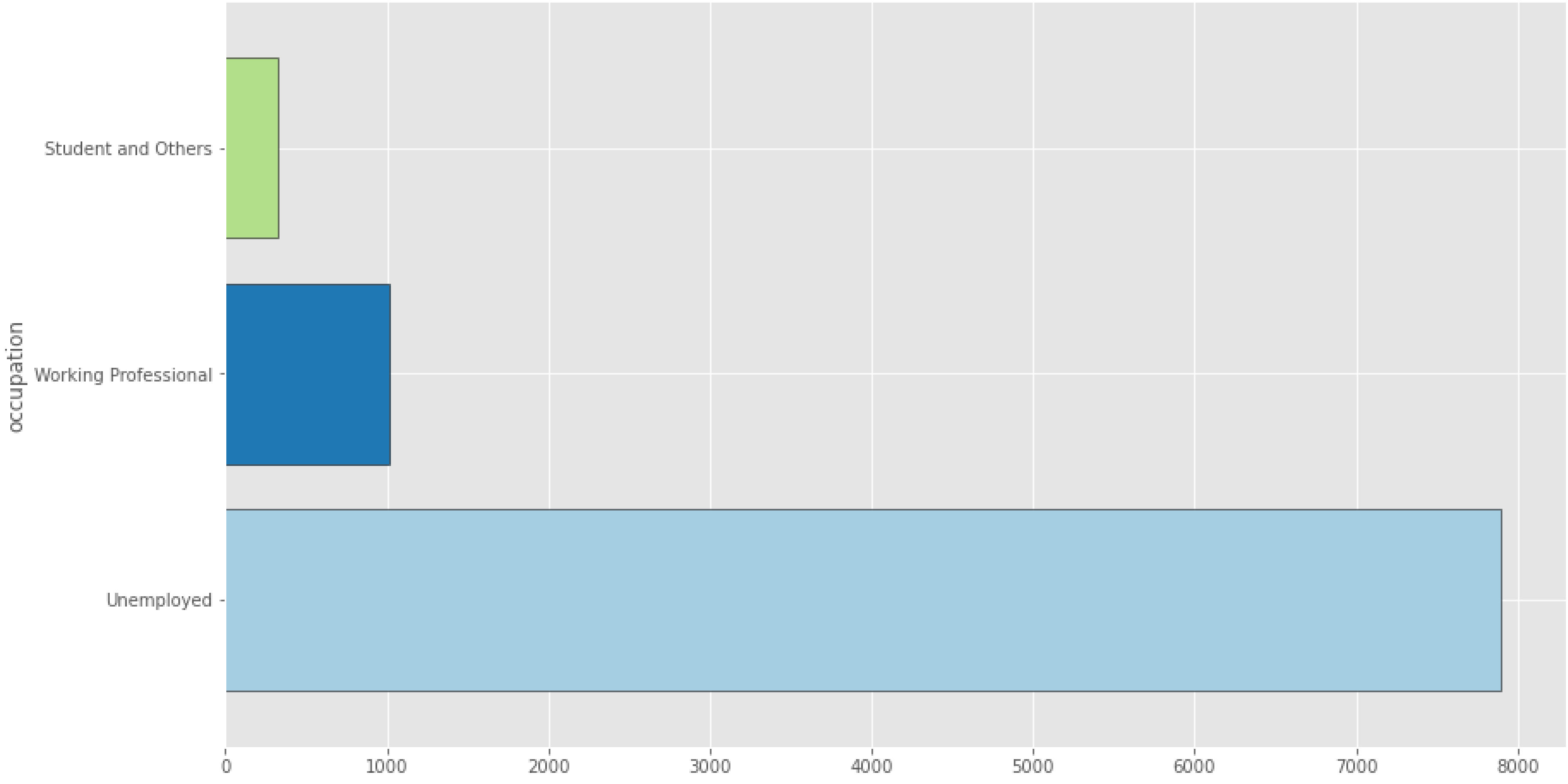
1.



PART B - DATA ANALYSIS

Exploratory Data Analysis

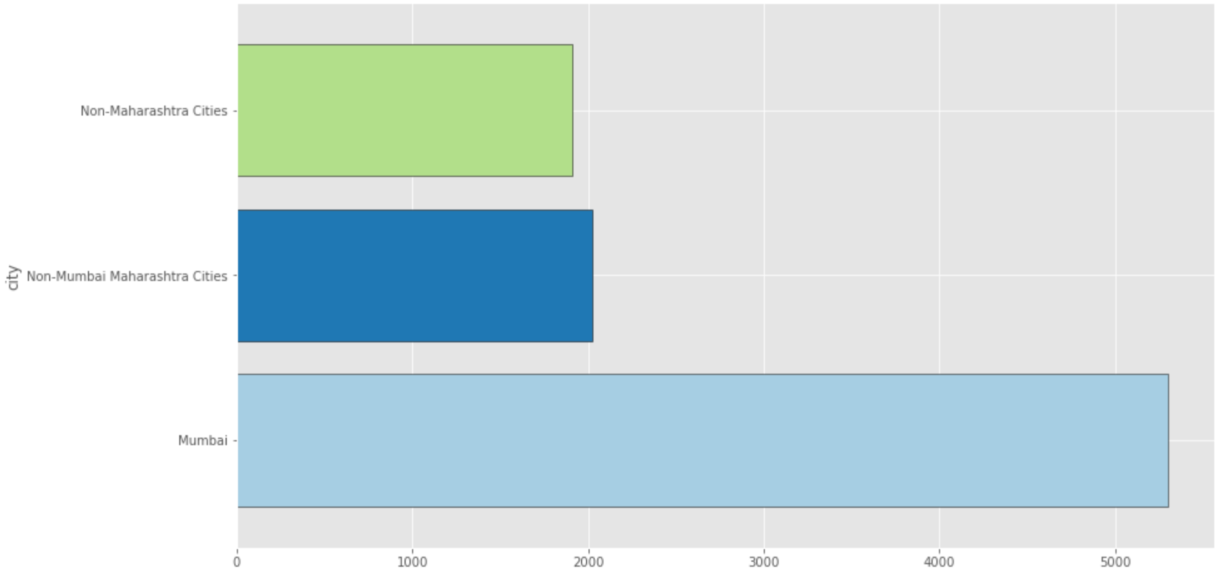
1.

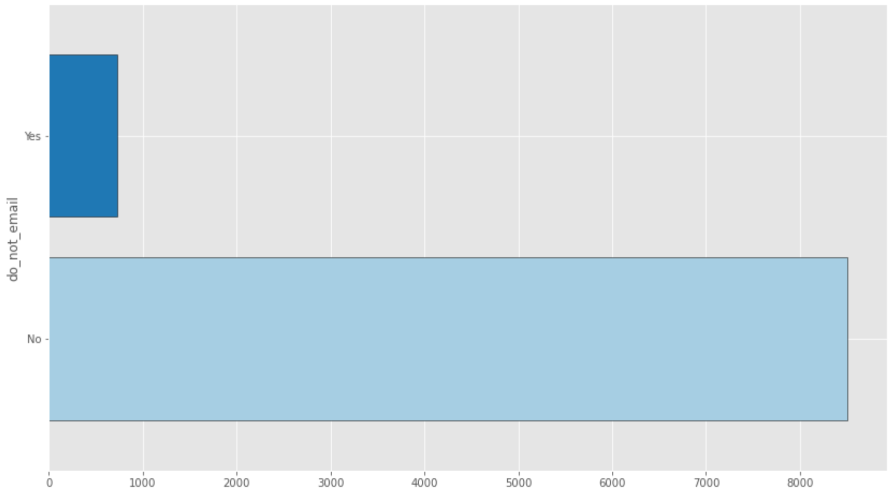


PART B - DATA ANALYSIS

Exploratory Data Analysis

1.





PART B - DATA ANALYSIS

Data Preparation

1. Unique values

```
lead_number = 9240
lead_origin = 3
lead_source = 6
do_not_email = 2
specialization = 3
occupation = 3
city = 3
mastering_interview = 2
```

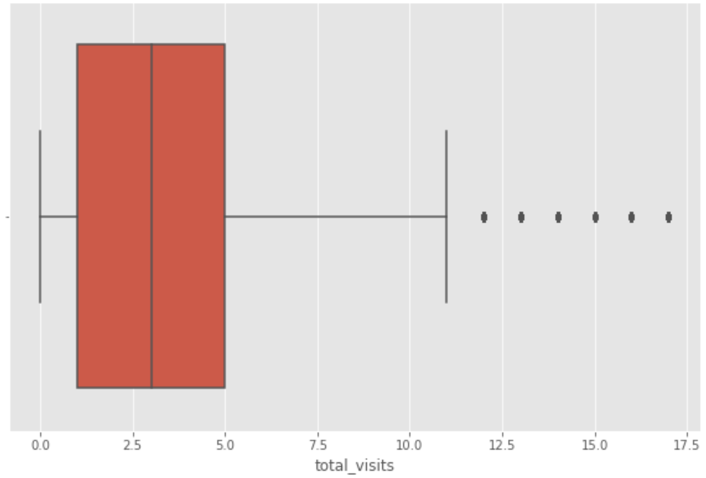
- Creating dummy variable for categorical columns
- Categorical columns are: lead_origin, lead_source, specialization, occupation, city

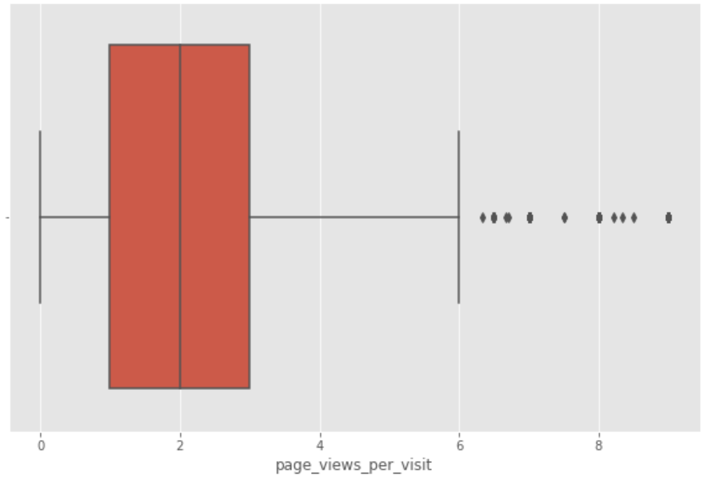
	total_visits	time_on_website	page_views_per_visit
count	9240.00	9240.00	9240.00
mean	3.44	487.70	2.36
std	4.82	548.02	2.15
min	0.00	0.00	0.00
25%	1.00	12.00	1.00
50%	3.00	248.00	2.00
75%	5.00	936.00	3.00
90%	7.00	1380.00	5.00
95%	10.00	1562.00	6.00
99%	17.00	1840.61	9.00
max	251.00	2272.00	55.00

PART B - DATA ANALYSIS

Exploratory Data Analysis

1.





Training and testing sets of ratio 7:3

PART C - BUILDING A LINEAR REGRESSION MODEL

Divide the data into training and testing sets

Training numeric feature

do_not_email	total_visits	time_on_website	page_views_per_visit	mastering_interview	lead_origin_Landing Page Submission	lead_source_Google	lead_source_Olark Chat	lead_source_Organic Search	lead_source_Other Social Sites	lead_source_Reference	specialization_Management Specializations	occupation_Unemployed	city_Non-Maharashtra Cities	city_Non-Mumbai Maharashtra Cities	
count	6468	6468	6468	6468	6468	6468	6468	6468	6468	6468	6468	6468	6468	6468	6468
mean	0.08	0	0	0	0.31	0.53	0.32	0.19	0.12	0.04	0.06	0.73	0.86	0.21	0.22
std	0.27	1.00	1.00	1.00	0.46	0.50	0.47	0.39	0.33	0.19	0.24	0.44	0.35	0.41	0.41
min	0	-1.02	-0.89	-1.18	0	0	0	0	0	0	0	0	0	0	0
25%	0	-0.72	-0.86	-0.67	0	0	0	0	0	0	0	0	1.00	0	0
50%	0	-0.10	-0.44	-0.16	0	1.00	0	0	0	0	0	1.00	1.00	0	0
75%	0	0.51	0.81	0.35	1.00	1.00	1.00	0	0	0	0	1.00	1.00	0	0
max	1.00	4.20	3.27	3.40	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

PART C - BUILDING A LINEAR REGRESSION MODEL

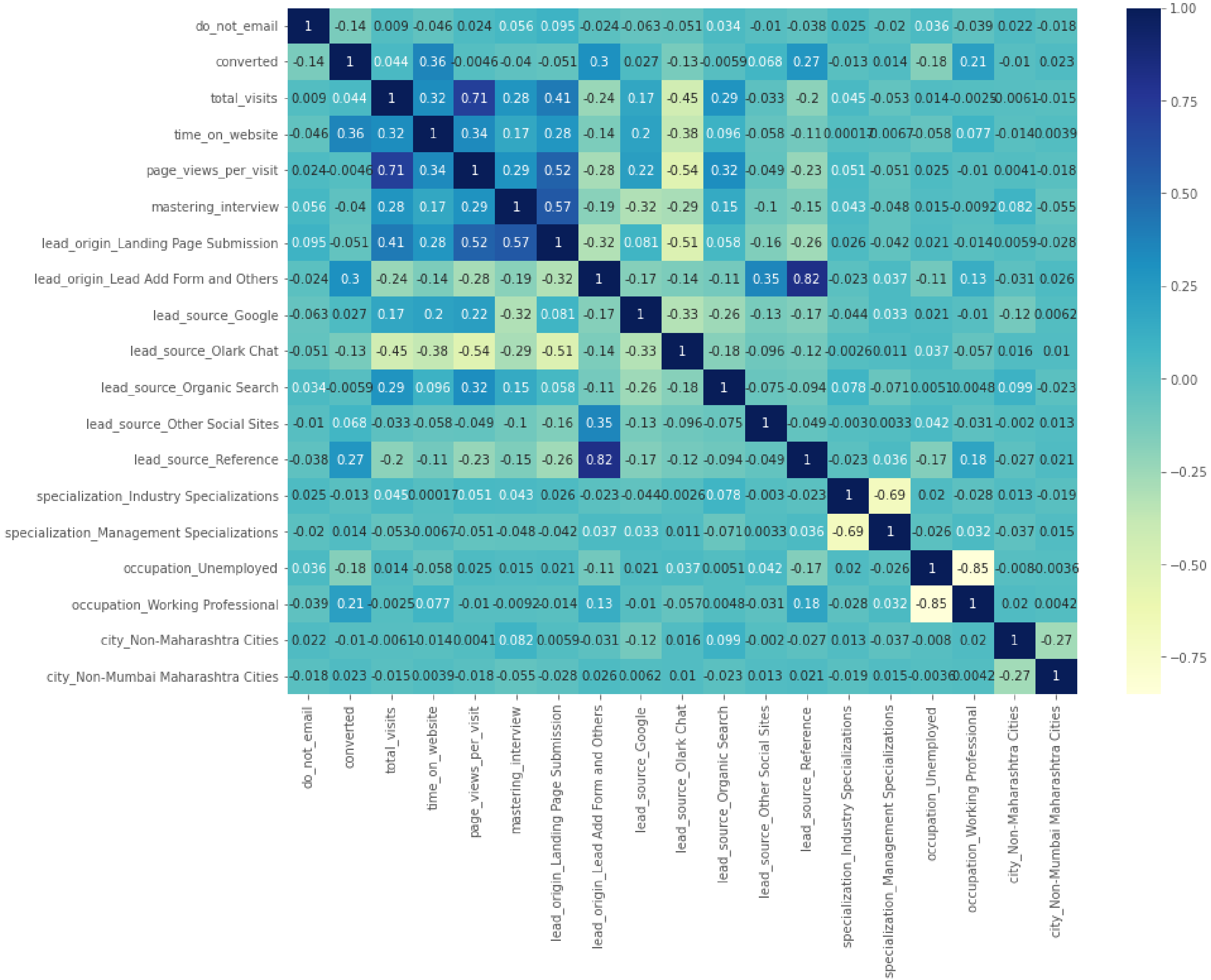
Model Building

Conversion rate 38.5

PART C - BUILDING A LINEAR REGRESSION MODEL

Exploratory Data Analysis

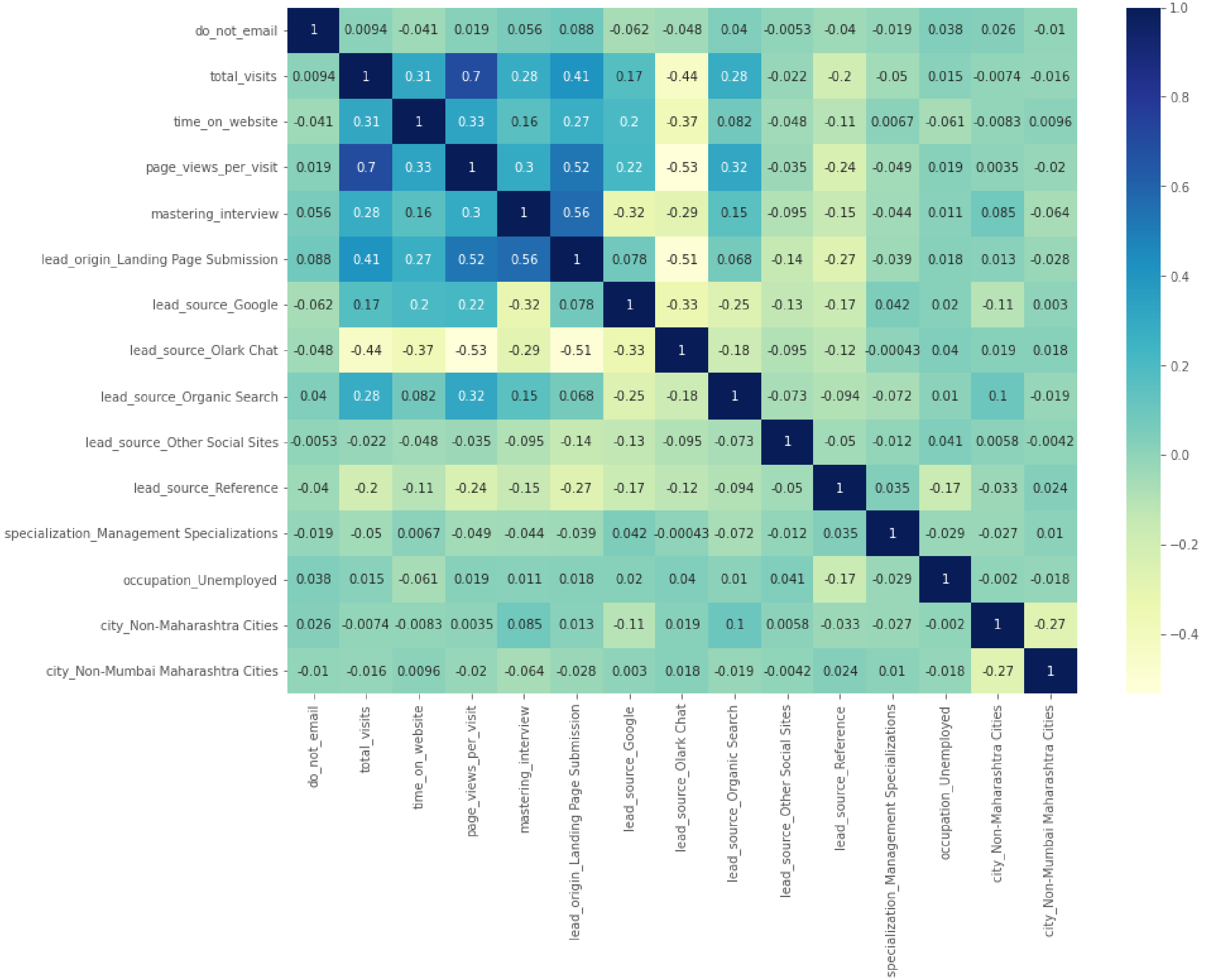
The correlation matrix
Before drop columns



PART C - BUILDING A LINEAR REGRESSION MODEL

Exploratory Data Analysis

The correlation matrix
After drop columns



PART C - BUILDING A LINEAR REGRESSION MODEL

Model Building

Generalized Linear Model Regression Results

Dep. Variable:	converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6452
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3307.4
Date:	Mon, 19 Feb 2024	**Deviance:**	6614.7
Time:	22:19:16	**Pearson chi2:**	6.67e+03
No. Iterations:	5	**Pseudo R-squ. (CS):**	0.2641
Covariance Type:	nonrobust		

PART C - BUILDING A LINEAR REGRESSION MODEL

Model Building

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3534	0.147	-2.399	0.016	-0.642	-0.065
do_not_email	-1.2038	0.144	-8.333	0.000	-1.487	-0.921
total_visits	0.1415	0.042	3.373	0.001	0.059	0.224
time_on_website	1.0413	0.036	29.257	0.000	0.972	1.111
page_views_per_visit	-0.1825	0.048	-3.775	0.000	-0.277	-0.088
mastering_interview	0.0007	0.094	0.007	0.994	-0.183	0.185
lead_origin_Landing Page Submission	9.495e-05	0.092	0.001	0.999	-0.181	0.181
lead_source_Google	0.3617	0.100	3.602	0.000	0.165	0.558
lead_source_Olark Chat	0.6850	0.137	5.016	0.000	0.417	0.953
lead_source_Organic Search	0.2099	0.116	1.811	0.070	-0.017	0.437
lead_source_Other Social Sites	1.6308	0.175	9.308	0.000	1.287	1.974
lead_source_Reference	3.9581	0.221	17.921	0.000	3.525	4.391
specialization_Management Specializations	0.0273	0.069	0.394	0.693	-0.108	0.163
occupation_Unemployed	-0.8496	0.086	-9.917	0.000	-1.018	-0.682
city_Non-Maharashtra Cities	0.0758	0.078	0.966	0.334	-0.078	0.230
city_Non-Mumbai Maharashtra Cities	0.0939	0.076	1.234	0.217	-0.055	0.243

PART D - PRESENTING THE RESULTS

1. Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?

→ This question focuses on overall model feature importance.

Assignment-based Subjective Questions

→ To identify the top 3 variables contributing most

1. Analyze Feature Ranking:

•**RFE:** which rank features based on their elimination order. The top-ranked features (with support_ = True) are considered more important.

•**GLM Coefficients:** Analyze the absolute value of the regression coefficients in the GLM results. Larger coefficients (positive or negative) indicate a stronger association with the target variable (conversion). Consider features with the highest absolute coefficients (excluding the intercept) as potentially important.

2. Combine Both Methods:

•**Cross-Validate Importance:** Compare the feature rankings from RFE and GLM coefficients. Features consistently appearing at the top in both methods are likely the most influential.

3. Domain Knowledge and Interpretation:

•**Consider the context:** While both ranking methods are valuable, use the understanding of the data and domain knowledge to interpret the results.

PART D - PRESENTING THE RESULTS

1. Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?

Assignment-based Subjective Questions

→ To identify the top 3 variables contributing most

•**RFE:** which rank features based on their elimination order. The top-ranked features (with support_ = True) are considered more important.

```
[('do_not_email', True, 1),
 ('total_visits', True, 1),
 ('time_on_website', True, 1),
 ('page_views_per_visit', True, 1),
 ('mastering_interview', False, 3),
 ('lead_origin_Landing Page Submission', True, 1),
 ('lead_source_Google', True, 1),
 ('lead_source_Olark Chat', True, 1),
 ('lead_source_Organic Search', True, 1),
 ('lead_source_Other Social Sites', True, 1),
 ('lead_source_Reference', True, 1),
 ('specialization_Management Specializations', False, 2),
 ('occupation_Unemployed', True, 1),
 ('city_Non-Maharashtra Cities', True, 1),
 ('city_Non-Mumbai Maharashtra Cities', True, 1)]
```

PART D - PRESENTING THE RESULTS

Assignment-based Subjective Questions

1. Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?

• **GLM Coefficients:** Analyze the absolute value of the regression coefficients in the GLM results. Larger coefficients (positive or negative) indicate a stronger association with the target variable (conversion). Consider features with the highest absolute coefficients (excluding the intercept) as potentially important.

ANSWER:

1. lead_source_Reference
2. lead_source_Other Social Sites
3. do_not_email

		coef	std err	z	P>|z|	[0.025	0.975]
	const	-0.3534	0.147	-2.399	0.016	-0.642	-0.065
#3	do_not_email	-1.2038	0.144	-8.333	0.000	-1.487	-0.921
	total_visits	0.1415	0.042	3.373	0.001	0.059	0.224
	time_on_website	1.0413	0.036	29.257	0.000	0.972	1.111
	page_views_per_visit	-0.1825	0.048	-3.775	0.000	-0.277	-0.088
	mastering_interview	0.0007	0.094	0.007	0.994	-0.183	0.185
	lead_origin_Landing Page Submission	9.495e-05	0.092	0.001	0.999	-0.181	0.181
	lead_source_Google	0.3617	0.100	3.602	0.000	0.165	0.558
	lead_source_Olark Chat	0.6850	0.137	5.016	0.000	0.417	0.953
	lead_source_Organic Search	0.2099	0.116	1.811	0.070	-0.017	0.437
#2	lead_source_Other Social Sites	1.6308	0.175	9.308	0.000	1.287	1.974
#1	lead_source_Reference	3.9581	0.221	17.921	0.000	3.525	4.391
	specialization_Management Specializations	0.0273	0.069	0.394	0.693	-0.108	0.163
	occupation_Unemployed	-0.8496	0.086	-9.917	0.000	-1.018	-0.682
	city_Non-Maharashtra Cities	0.0758	0.078	0.966	0.334	-0.078	0.230
	city_Non-Mumbai Maharashtra Cities	0.0939	0.076	1.234	0.217	-0.055	0.243

PART D - PRESENTING THE RESULTS

2. What are the top 3 categorical/dummy variables in the model which **should be focused** the most on in order to increase the probability of lead conversion?

This question specifically asks for the top 3 categorical/dummy variables, implying focus on discrete features. → This targets **actionable variables** where we can directly take steps to influence conversion

ANSWER:
Categorical/dummy + High Positive contribution are
1. lead_source_Reference
2. lead_source_Other Social Sites
3. lead_source_Olark Chat

→ Should be focused the most

Assignment-based Subjective Questions

			coef	std err	z	P>|z|	[0.025	0.975]
		const	-0.3534	0.147	-2.399	0.016	-0.642	-0.065
		do_not_email	-1.2038	0.144	-8.333	0.000	-1.487	-0.921
		total_visits	0.1415	0.042	3.373	0.001	0.059	0.224
		time_on_website	1.0413	0.036	29.257	0.000	0.972	1.111
		page_views_per_visit	-0.1825	0.048	-3.775	0.000	-0.277	-0.088
		mastering_interview	0.0007	0.094	0.007	0.994	-0.183	0.185
		lead_origin_Landing Page Submission	9.495e-05	0.092	0.001	0.999	-0.181	0.181
		lead_source_Google	0.3617	0.100	3.602	0.000	0.165	0.558
Categorical/dummy	#3	lead_source_Olark Chat	0.6850	0.137	5.016	0.000	0.417	0.953
		lead_source_Organic Search	0.2099	0.116	1.811	0.070	-0.017	0.437
Categorical/dummy	#2	lead_source_Other Social Sites	1.6308	0.175	9.308	0.000	1.287	1.974
Categorical/dummy	#1	lead_source_Reference	3.9581	0.221	17.921	0.000	3.525	4.391
		specialization_Management Specializations	0.0273	0.069	0.394	0.693	-0.108	0.163
		occupation_Unemployed	-0.8496	0.086	-9.917	0.000	-1.018	-0.682
		city_Non-Maharashtra Cities	0.0758	0.078	0.966	0.334	-0.078	0.230
		city_Non-Mumbai Maharashtra Cities	0.0939	0.076	1.234	0.217	-0.055	0.243

PART D - PRESENTING THE RESULTS

3. A period of 2 months every year during which they hire some interns. Suggest a good strategy they should employ at this stage.

Assignment-based Subjective Questions

ANSWER:

1. **Tier leads by predicted probability:** Instead of calling all "1" predictions, prioritize leads closest to 1 (highest conversion likelihood). Then, segment by "lead_source" ("Reference", "Other Social Sites", and "Olark Chat"). This prioritizes the most promising leads within segments with positive conversion influence.

2. **Leverage interns for initial outreach:** Utilize interns for email/short call outreach while reserving experienced salespeople for higher-priority or challenging leads. Train interns on effective scripts and personalized messaging based on segments.

3. **Track individual performance:** Monitor both lead conversion and intern success to identify individuals excelling in specific segments and provide targeted coaching.

4. **Multi-channel approach:** Don't solely rely on phone calls. Use emails, social media messages, or SMS personalized based on lead profiles and segment preferences.

5. **Limited-time incentives:** Offer exclusive discounts or promotions during the internship period to encourage immediate conversion.

6. **Address "do_not_email" preferences:** If a lead opted out of email, respect their choice and prioritize alternative channels like phone calls or SMS with clear value propositions.

PART D - PRESENTING THE RESULTS

4. The company reaches its target for a quarter before the deadline. Suggest a good strategy they should employ at this stage.

Assignment-based Subjective Questions

ANSWER:

1. **Recalibrate lead priority:** Instead of solely relying on predicted conversion probability, consider incorporating additional factors like:
   - **Engagement:** Prioritize leads actively interacting with emails, website content, or social media to gauge genuine interest.
   - **Lead source:** Focus on sources like "Reference" or "Olark Chat" with known positive conversion influence.
   - **Time since last contact:** Engage with leads who haven't been contacted recently to avoid oversaturation.

2. **Utilize scoring system:** Develop a scoring system that combines multiple factors (predicted probability, engagement, etc.) to rank leads and prioritize those most likely to benefit from contact without requiring aggressive phone calls.

THANK YOU

LET’S COLLABORATE AND BUILD TOGETHER!

Student: Pham Thieu Quan (phamthieuquan.ivf@gmail.com)