

ADVERTISING ON FACEBOOK

► Facebook is an online social media and social networking service owned by American company Meta Platforms. Founded in 2004 by Mark Zuckerberg

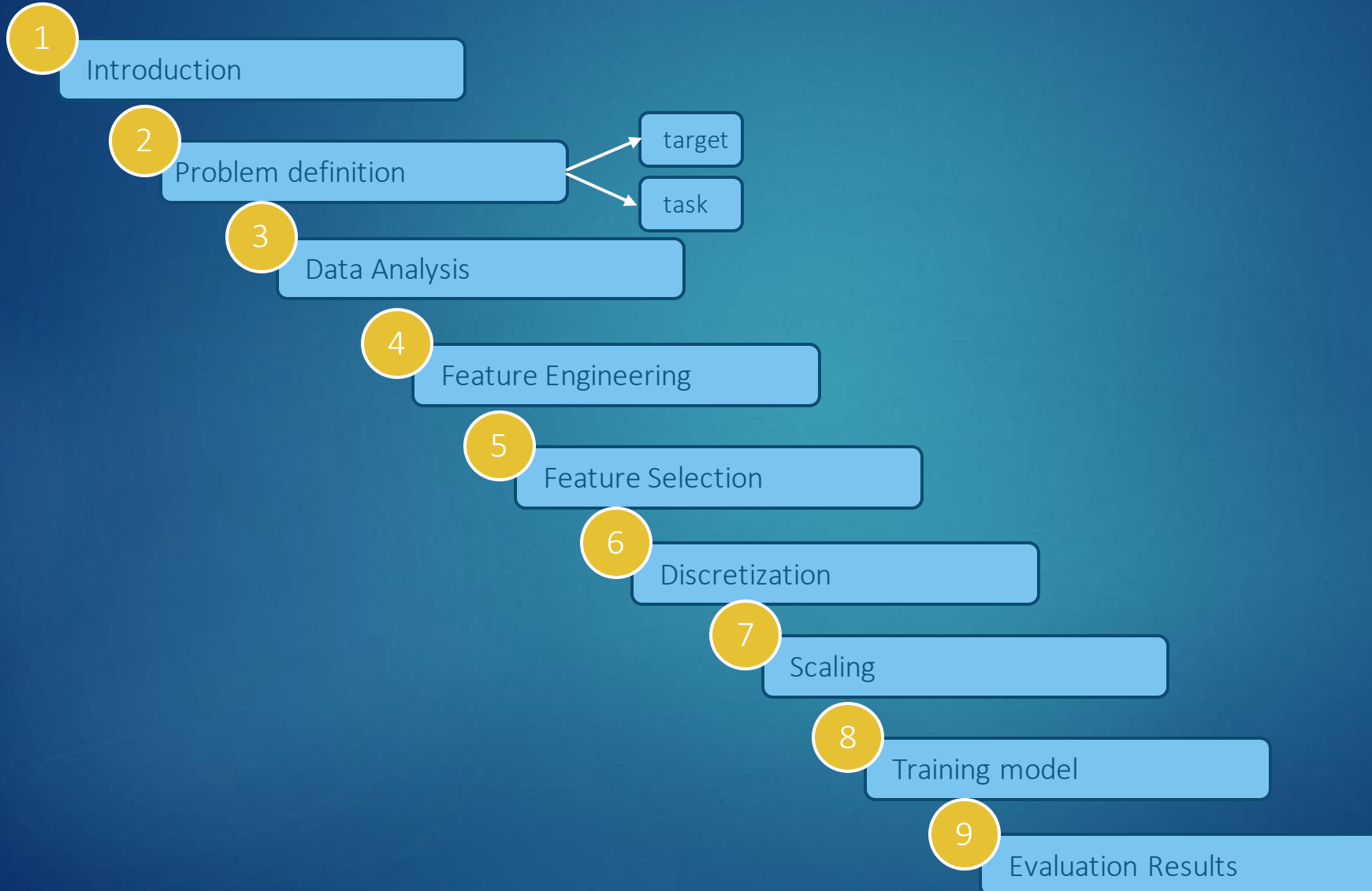




CONTENT

- ▶ TARGETED MARKETING ADS ON FACEBOOK
- ▶ EXPLORE AND VISUALIZE
- ▶ SPLIT DATASET ON TRAIN AND TEST SETS
- ▶ IMPUTATION MISSING VALUES
- ▶ MACHINE LEARNING MODEL:
(NAÏVE BAYES CLASSIFICATION, BAGGING CLASSIFICATION, K-NEIGHBORS CLASSIFICATION, RANDOM FOREST CLASSIFICATION, LOGISTIC REGRESSION, SUPER LEARNER)
- ▶ CONCLUSION

PROJECT STRUCTURE



Python data analysis libraries



Pandas for processing and analysis of structured data

Anaconda is a distribution of the Python to simplify package management and deployment



Numpy for used for mathematical calculations, statistics, matrix



Matplotlib for data visualization



Seaborn for data visualization



ANALYSIS OF DATA

► ANALYSIS DATASET

5 Independent variables 1 dependent variable

	Names	emails	Country	Time Spent on Site	Salary	Clicked
0	Martina Avila	cubilia.Curae.Phaseilus@quisaccumsanconvallis.edu	Bulgaria	25.649648	55330.06006	0
1	Harlan Barnes	eu.dolor@diam.co.uk	Belize	32.456107	NaN	1
2	Naomi Rodriguez	vulputate.mauris.sagittis@ametconsectetueradip...	Algeria	20.945978	41098.60826	0
3	Jade Cunningham	malesuada@dignissim.com	Cook Islands	54.039325	37143.35536	1
4	Cedric Leach	felis.ullamcorper.viverra@egetmollislectus.net	Brazil	34.249729	NaN	0

1 Names

2 emails

3 Country

4 Time Spent on Site

5 Salary

6 Clicked

208 country

MIN: 5 min, MAX: 60 min

MIN: \$20, MAX: \$100 000

TARGETED MARKETING ADS

ON FACEBOOK

You have been hired as a consultant to a start-up that is running a targeted marketing ads on Facebook. The company wants to analyze customer behavior by predicting which customer clicks on the advertisement.

Customer data is as follows:

Inputs:

- Name
- e-mail
- Country
- Time on Facebook
- Estimated Salary (derived from other parameters)

Outputs:

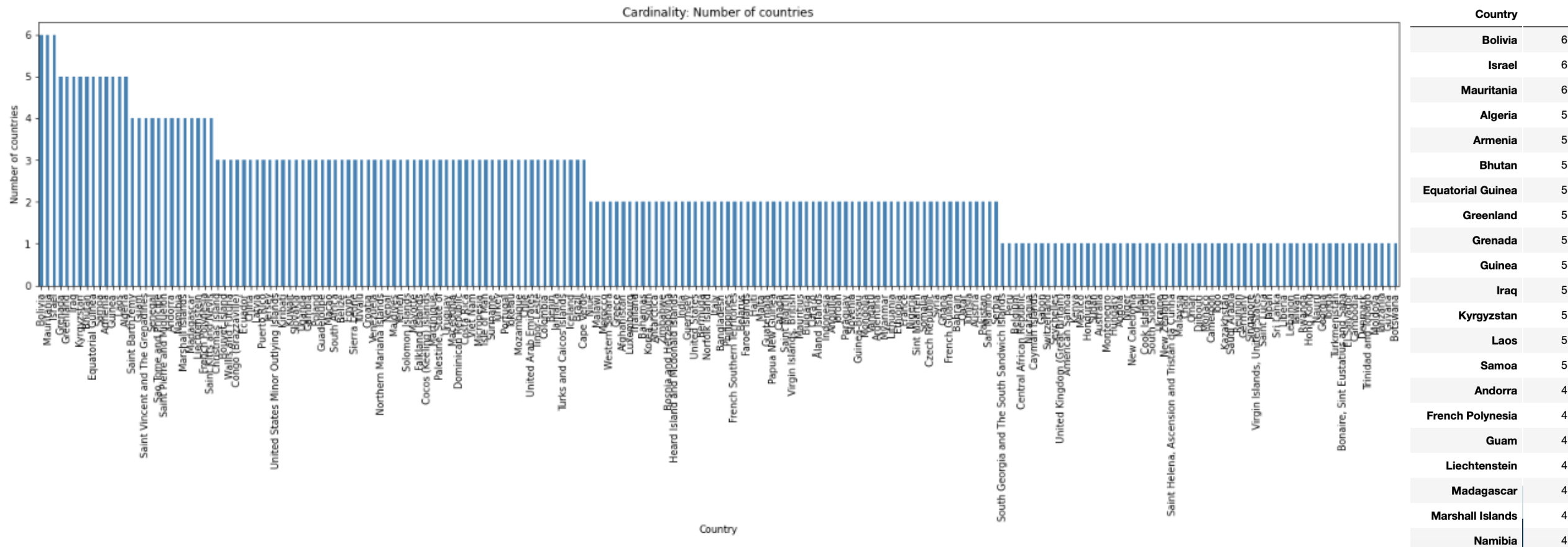
- Click (1: customer clicked on Ad, 0: customer did not click on Ad)

TOTAL: 499 rows, 6 columns

There are 1 binary variables : Clicked

There are 2 continuous variables: Time Spent on Site, Salary

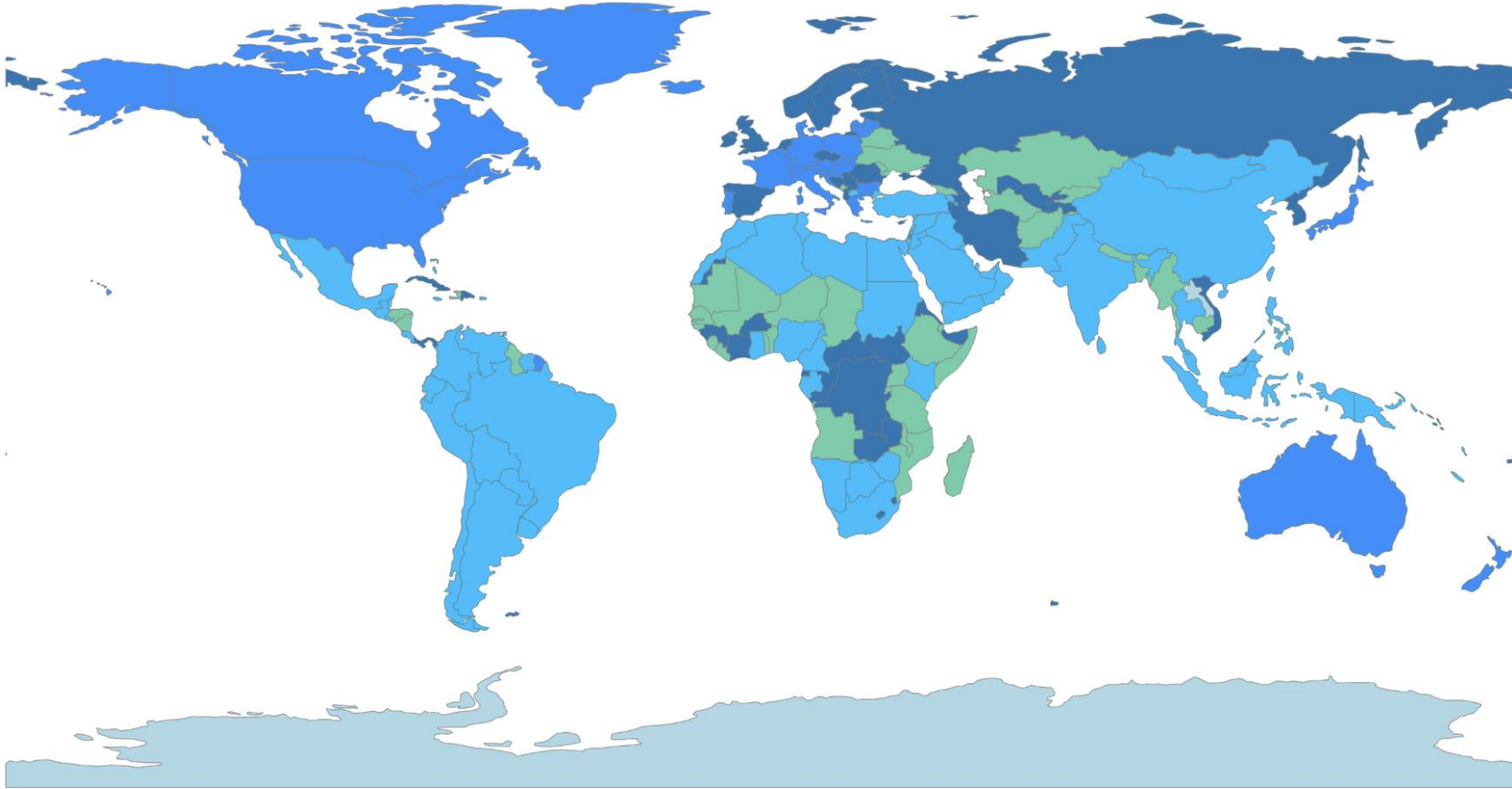
There are 3 categorical variables: Name, e-mail, Country



CARDINALITY OF COUNTRIES

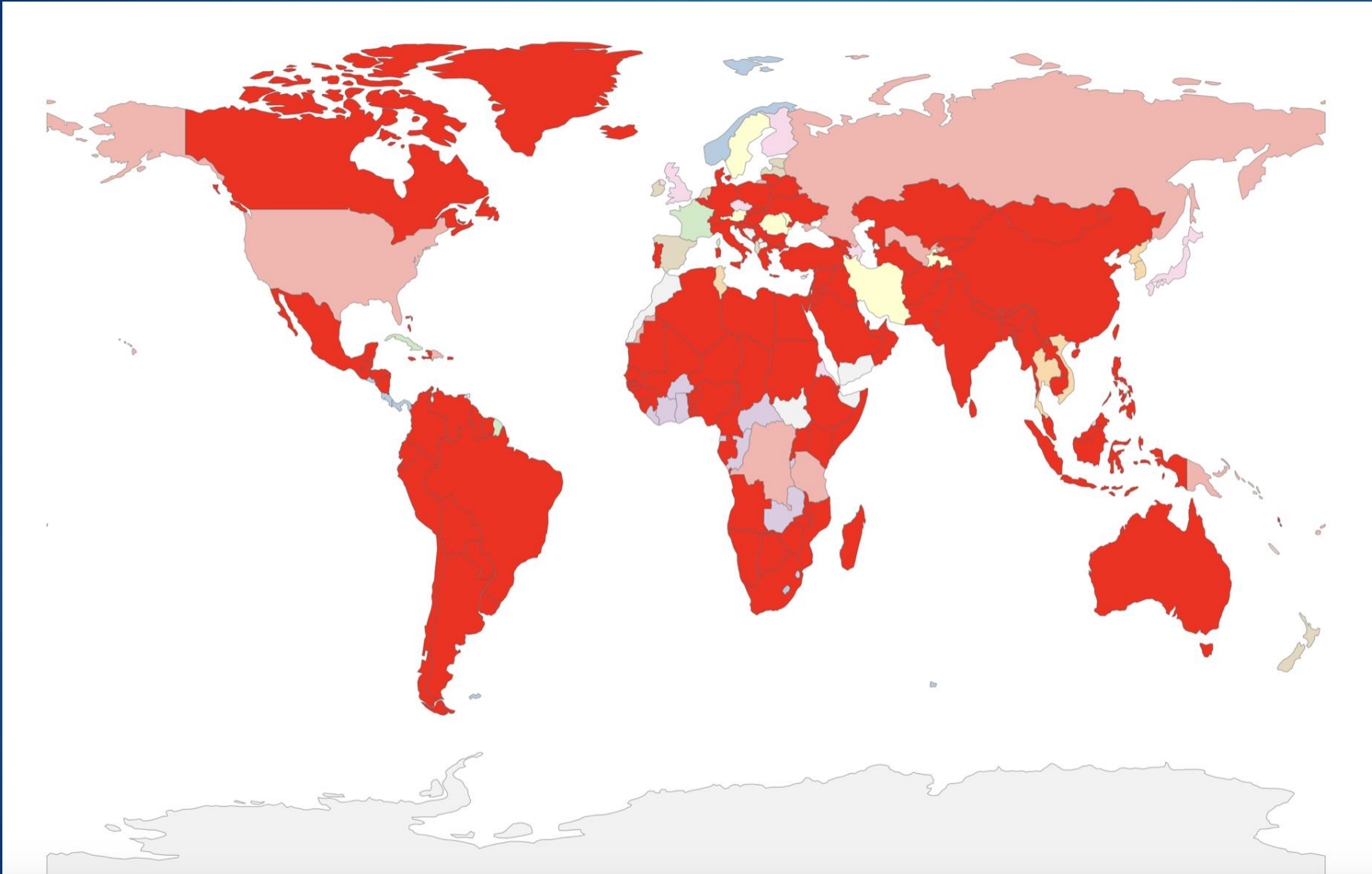
Countries: VietNam, Marshall Islands, Germany, Kyrgyzstan, Gambia, Reunion, Jordan, Slovakia, Kiribati, United States Minor Outlying Islands, Puerto Rico, Guinea, Guam, China, Ecuador, Anguilla, Kazakhstan, Jersey, Macao, Indonesia, Ghana, Cameroon, Myanmar, Central African Republic, Bolivia, Egypt, Tuvalu, Qatar, Venezuela, Togo, Nepal, Saudi Arabia, San Marino, etc.

COUNTRIES INTERESTED IN ADS



Countries from which people clicked on the advertisement on site. Dodgerblue shown developed countries, Skyblue – developing countries, Green – others countries, Darkblue – our dataset has no this countries.

COUNTRIES INTERESTED IN ADS

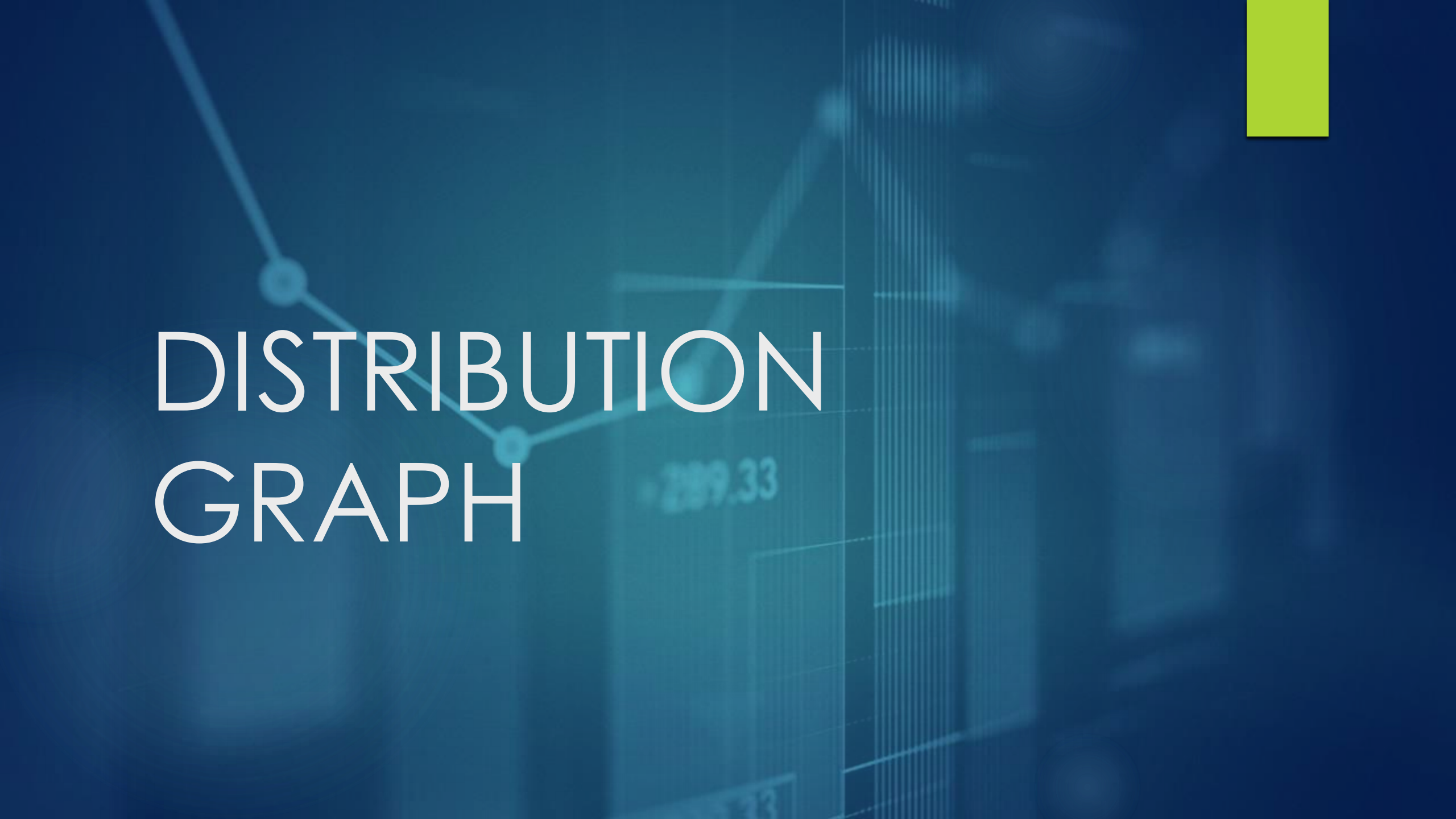


Countries from which people clicked on the advertisement on site is red.

TIME SPENT on SITE FROM MAXIMUM TO MINIMUM

- ▶ Sri Lanka
- ▶ Hong Kong
- ▶ Cook Islands,
- ▶ Israel,
- ▶ El Salvador
- ▶ Saint Bathemy
- ▶ United States Minor Outlying Islands
- ▶ Uganda*
- ▶ Bouvet Island
- ▶ Malta
- ▶ Bhutan, etc

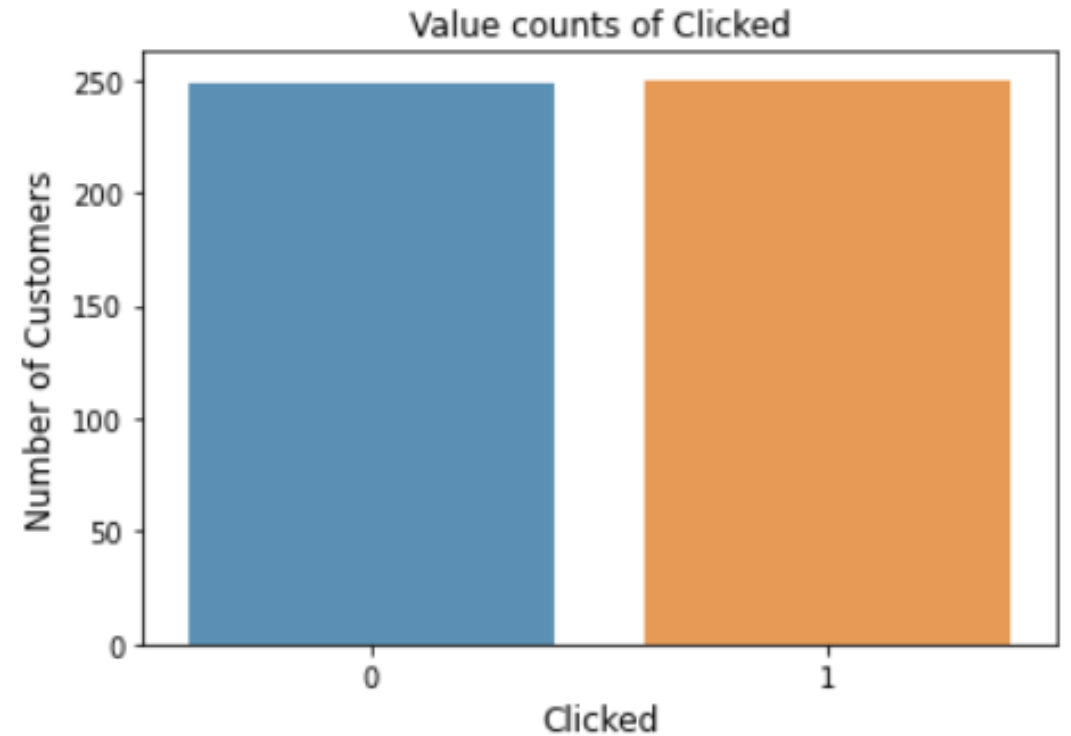
	Names	emails	Country	Time Spent on Site	Salary	Clicked
54	Rafael Peterson	NaN	Sri Lanka	60.000000	NaN	1
173	Cameron Cash	auctor@ipsum.org	Hong Kong	56.434684	60976.07092	1
3	Jade Cunningham	malesuada@dignissim.com	Cook Islands	54.039325	37143.35536	1
321	Mcclure, Avye L.	dictum@lorem.org	Israel	54.002969	44373.39815	1
302	Knapp, Quamar P.	Maecenas.libero.est@miacmattis.com	El Salvador	53.664216	53210.35911	1
247	MacKenzie O. Fowler	Aliquam.nec.enim@nec.co.uk	Saint Barth?emy	53.153079	43293.50716	1
75	Hedley Greene	eleifend@felis.org	United States Minor Outlying Islands	53.073907	65344.84072	1
262	Heather G. Goodwin	semper.egestas@maurissapien.co.uk	Uganda	53.049426	73043.34184	1
5	Carla Hester	mi@Aliquamerat.edu	NaN	52.009946	80642.65222	1
387	Munoz, Kennedy K.	dolor@nislelementumpurus.edu	Bouvet Island	51.513734	72735.56023	1
190	Ulric Robles	fringilla@ornare.edu	Malta	51.342165	45686.28719	1
205	Vaughan L. Mathis	eu@iaculis.org	Bhutan	50.985402	63695.04722	1
279	Dominic I. Faulkner	pharetra.Nam@sociisnatoque.org	Tonga	50.875878	53562.22630	1
163	Gannon Nguyen	et.rutrum.eu@congue.net	Kenya	50.272759	75305.25384	1
240	Timothy J. Terrell	orci@lobortis.com	Laos	50.227680	31707.31895	0
285	Kirestin F. Yang	eleifend.egestas.Sed@tempus.net	South Africa	50.044654	69461.32716	1
464	Kyle	NaN	Papua New Guinea	49.968733	71843.97055	1
415	Hyatt	erat.Vivamus@ligula.co.uk	Algeria	49.556217	66154.72594	1
18	Sloane Mann	at.augue@augue.net	Chad	48.870175	34774.44407	1
160	Hector Price	Aliquam.nisl@semegetmassa.co.uk	Martinique	48.861551	73141.85329	1
289	Madeson R. Salinas	Cum.sociis.natoque@acnullaIn.edu	Bonaire, Sint Eustatius and Saba	48.794413	55633.03835	1
94	Martina Fuentes	elit@nequeIn.com	Senegal	48.483032	66439.65064	1
338	Morales, Halla M.	elit.elit.fermentum@erosturpisnon.org	United States Minor Outlying Islands	48.349931	55006.64006	1

The background is a dark blue gradient. It features faint, semi-transparent line graphs and bar charts. A prominent line graph on the left has three data points connected by lines. To the right, there are several vertical bars of varying heights. A yellow rectangle is positioned in the top right corner. The text 'DISTRIBUTION GRAPH' is centered in a large, white, sans-serif font.

DISTRIBUTION GRAPH

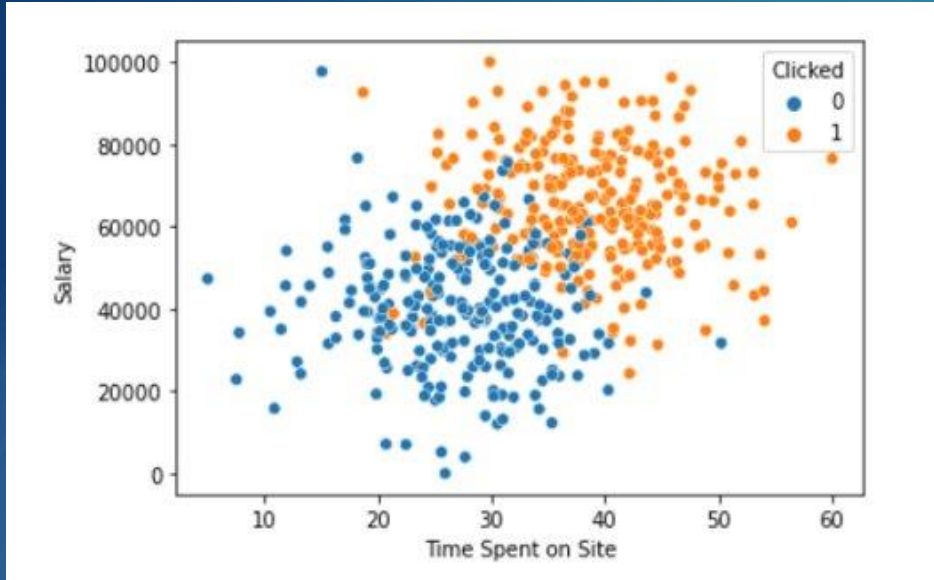
PROPORTION PERCENTAGE

- ▶ TOTAL 499
- ▶ NUMBER OF CUSTOMER WHO CLIKED ON AD = 250
- ▶ PERCENTAGE CLICKED = 50.1%
- ▶ DID NOT CLICK = 249
- ▶ PERCENTAGE WHO DID NOT CLICK = 49.89%

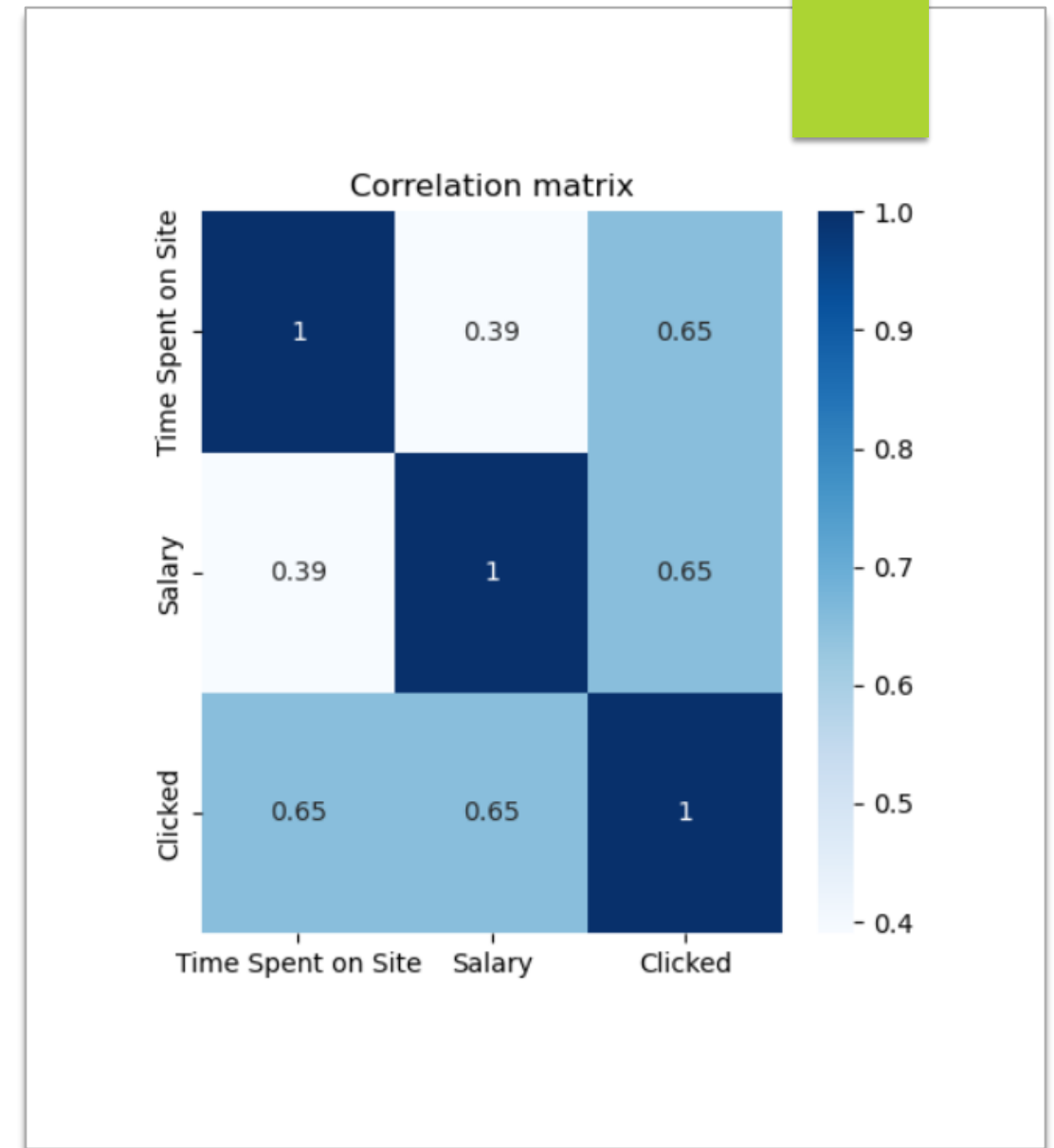


Click = 1 (orange)
No click = 0 (blue)

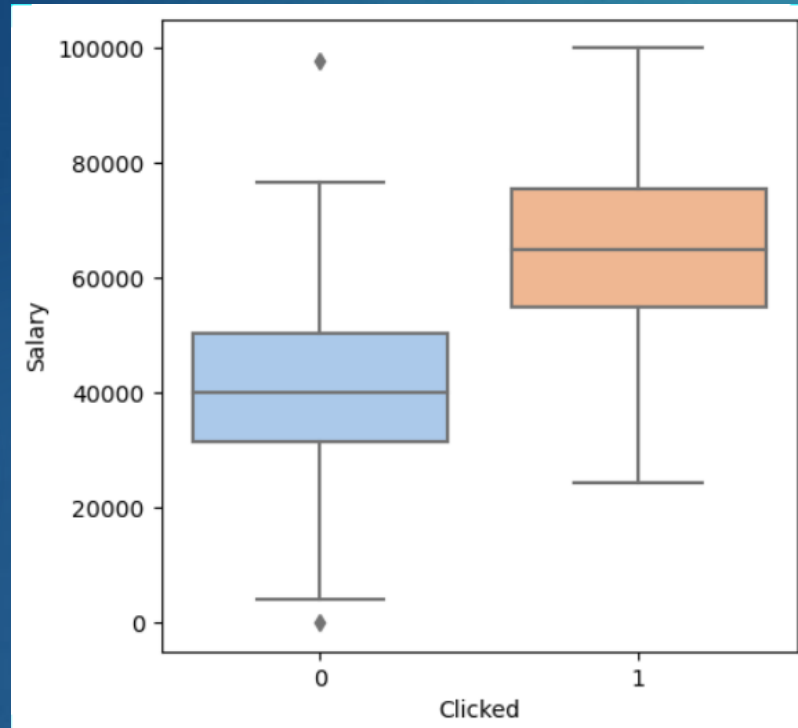
Distribution regarding salary and time spent on the site Facebook



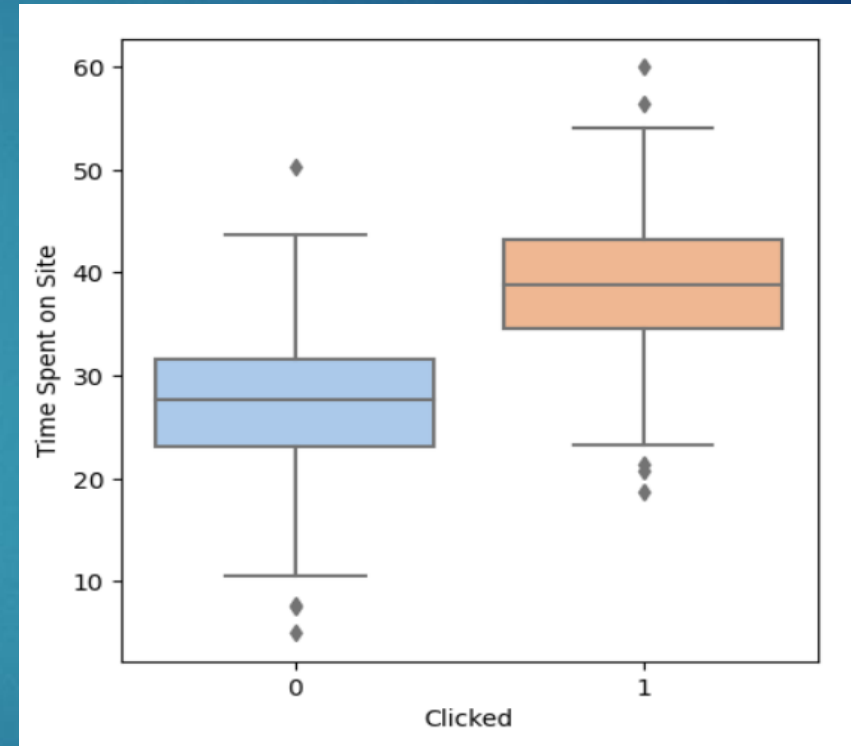
- Click = 1 (orange)
- No click = 0 (blue)



Distribution regarding Salary



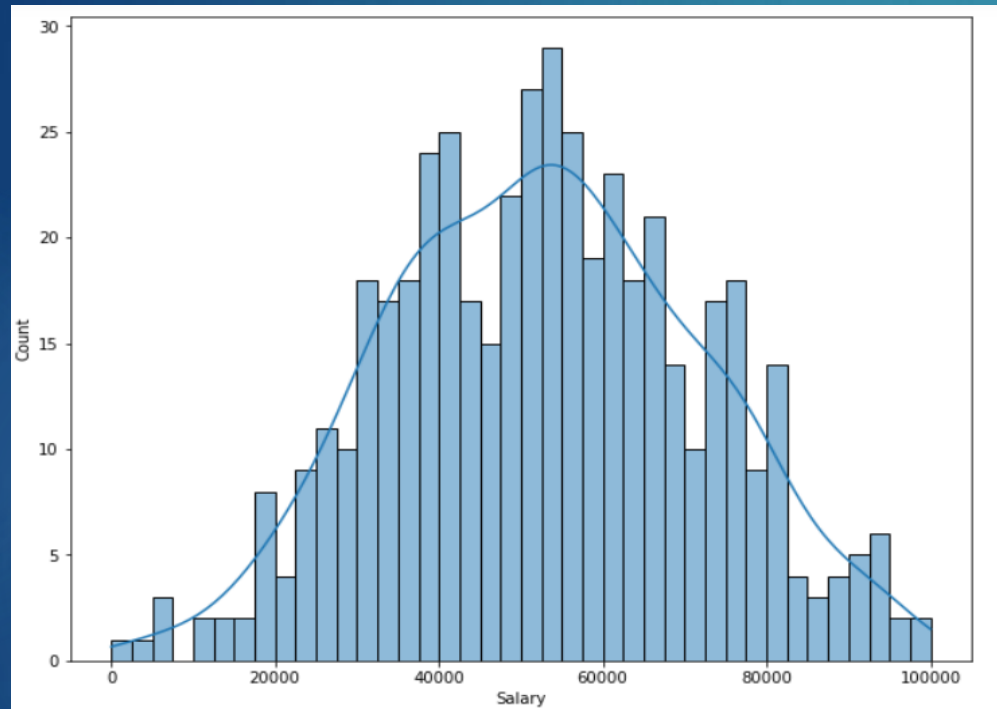
Distribution regarding Time spent on site



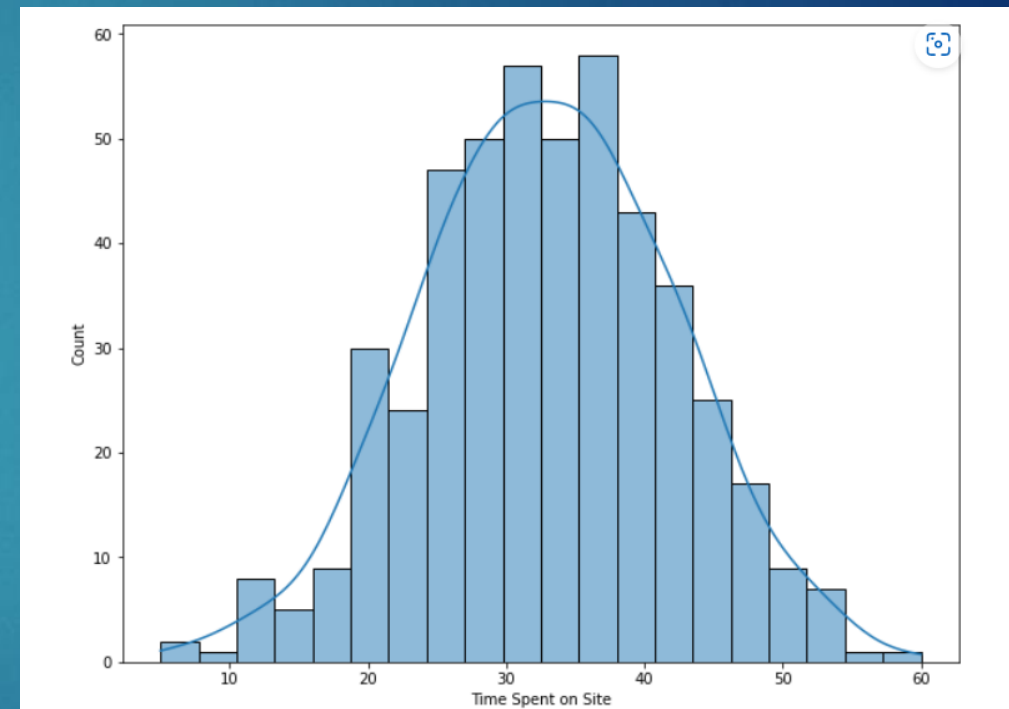
Click = 1 (orange)
No click = 0 (blue)

DISTRIBUTION GRAPH

Salary

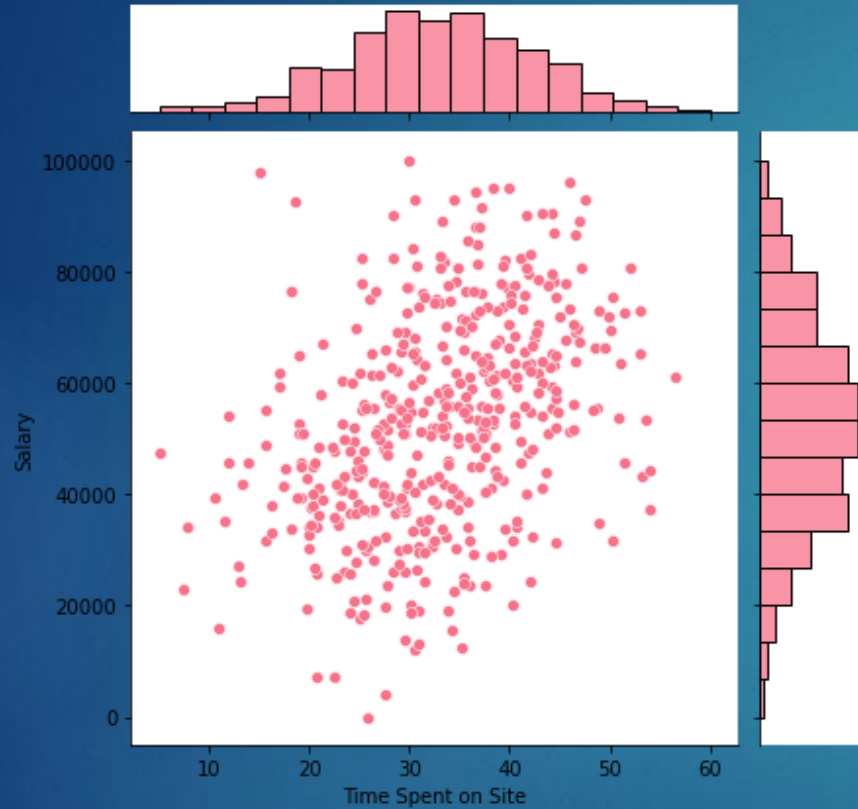


Time spent on the site



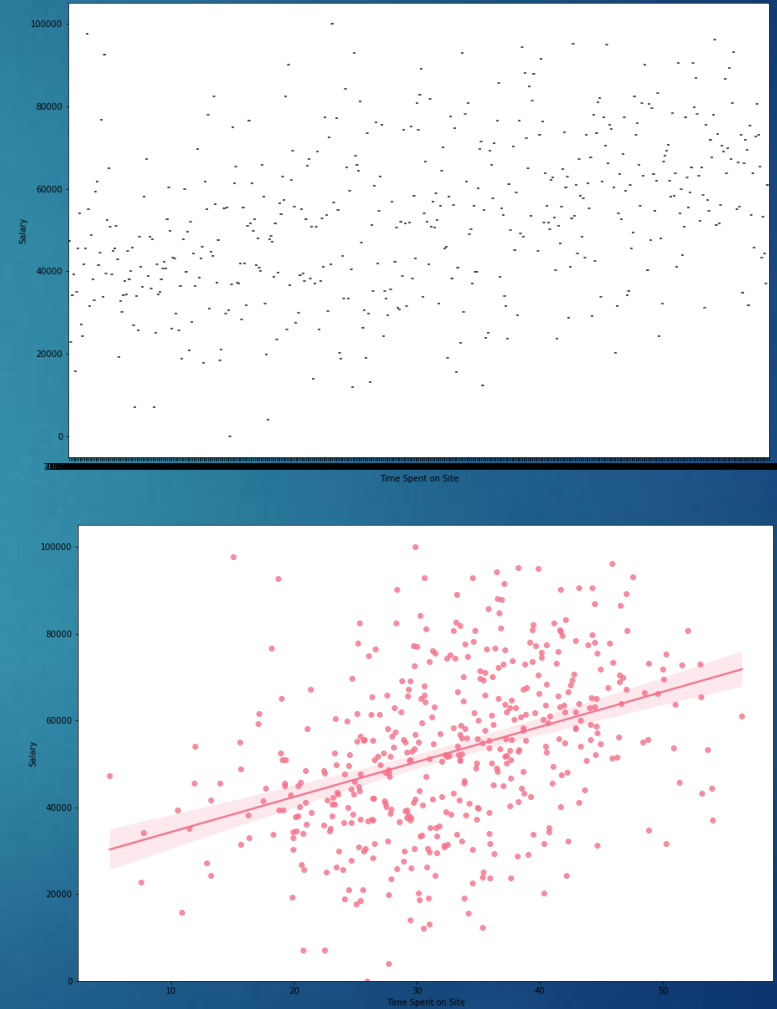
DISTRIBUTION GRAPH

Salary



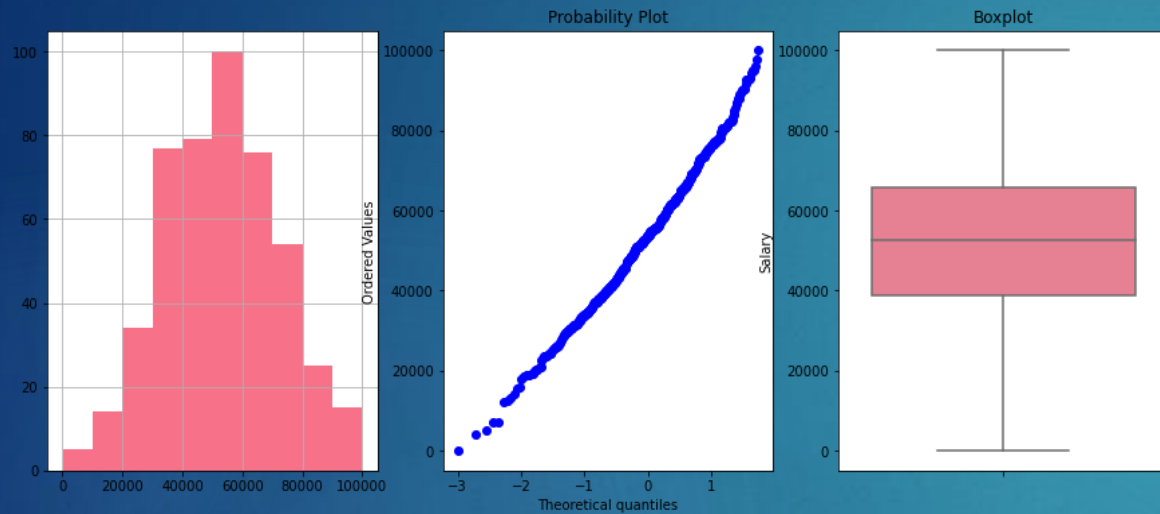
Time spent on the site

Salary

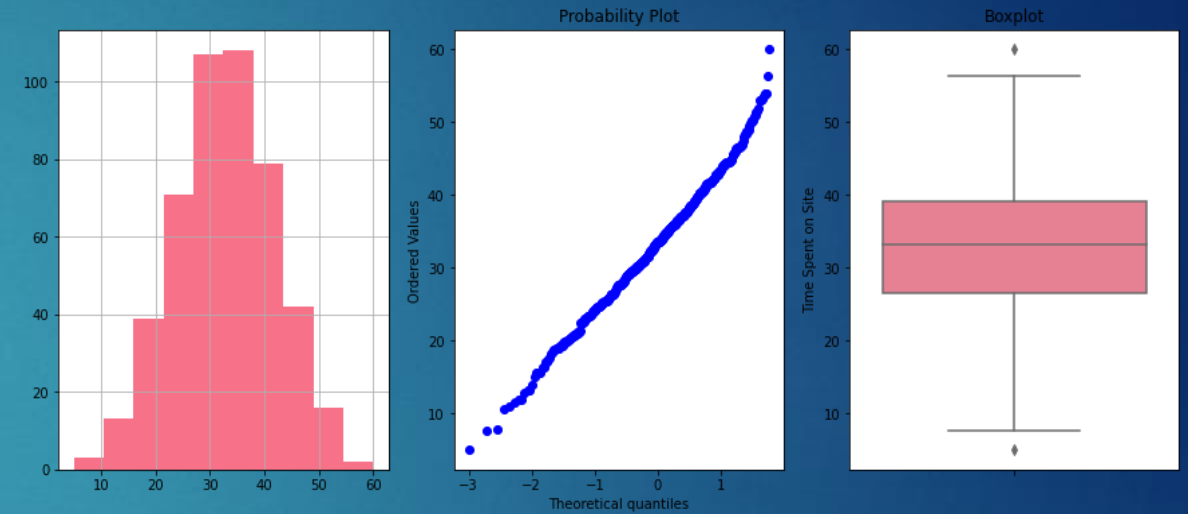


Time spent on the site

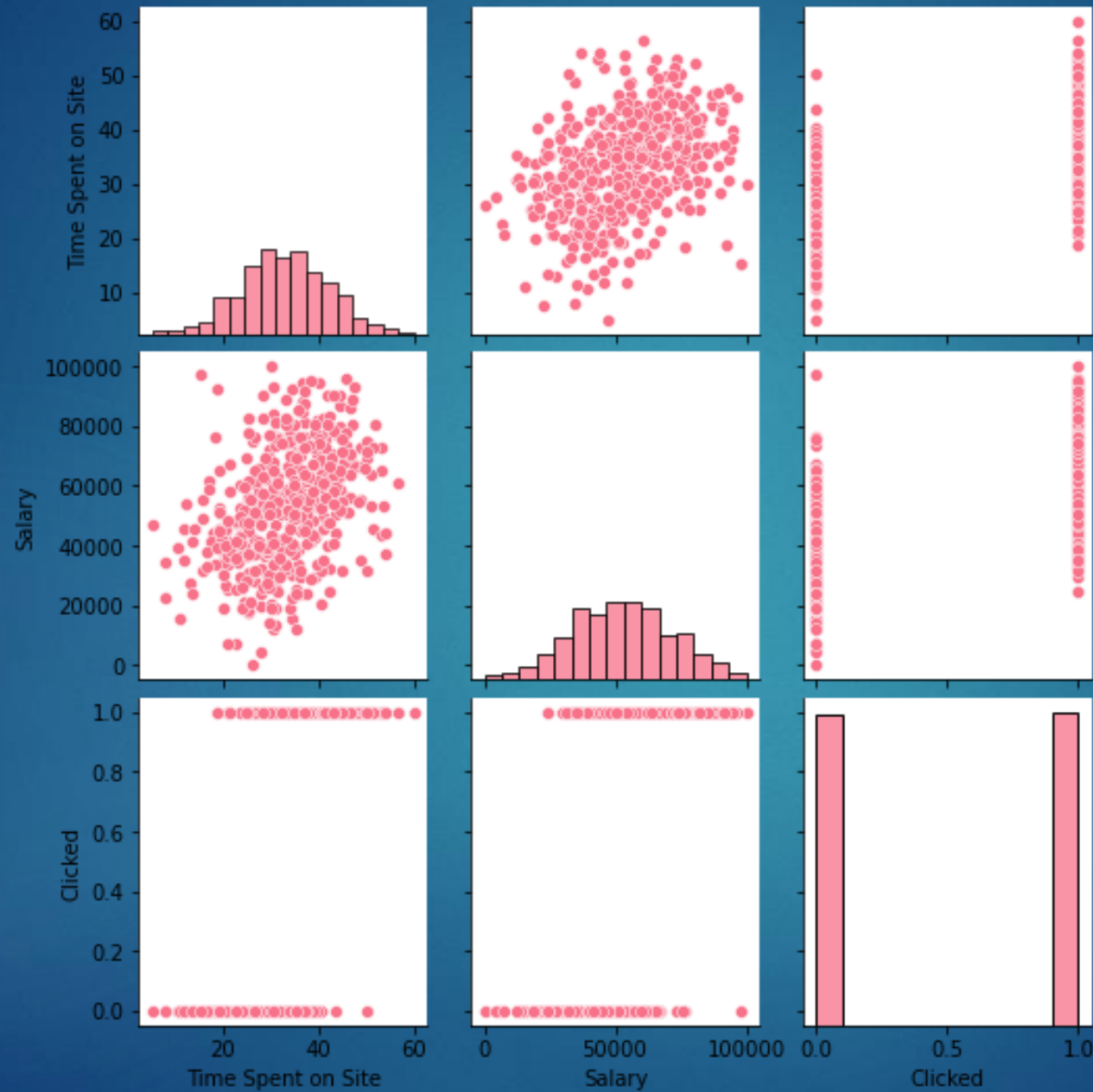
DISTRIBUTION, PROBABILITY PLOT, BOX PLOT



Salary



Time spent on the site



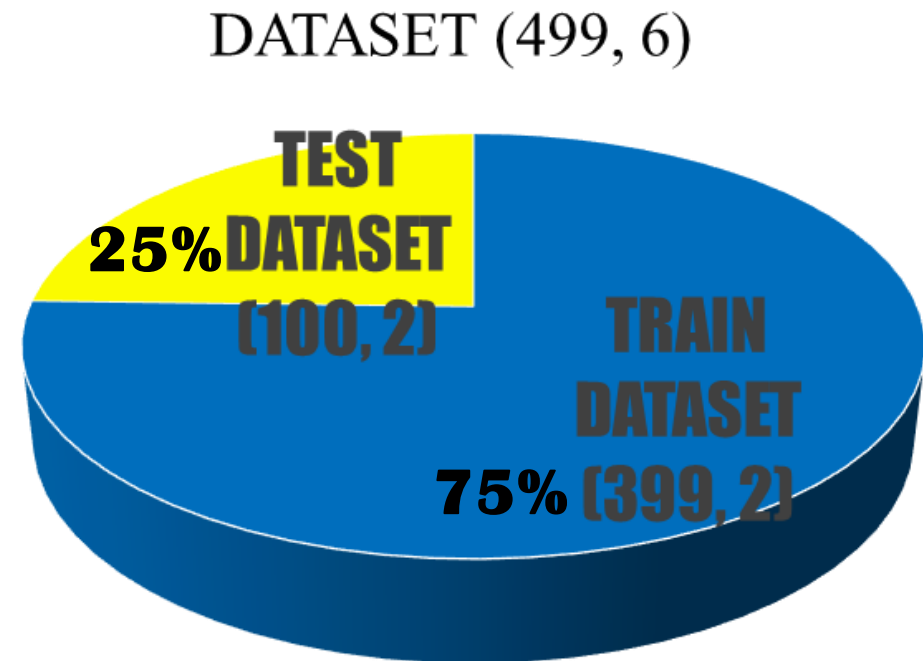
Time spent on the site

Salary

Clicked

SPLIT DATASET ON TRAIN AND TEST SETS

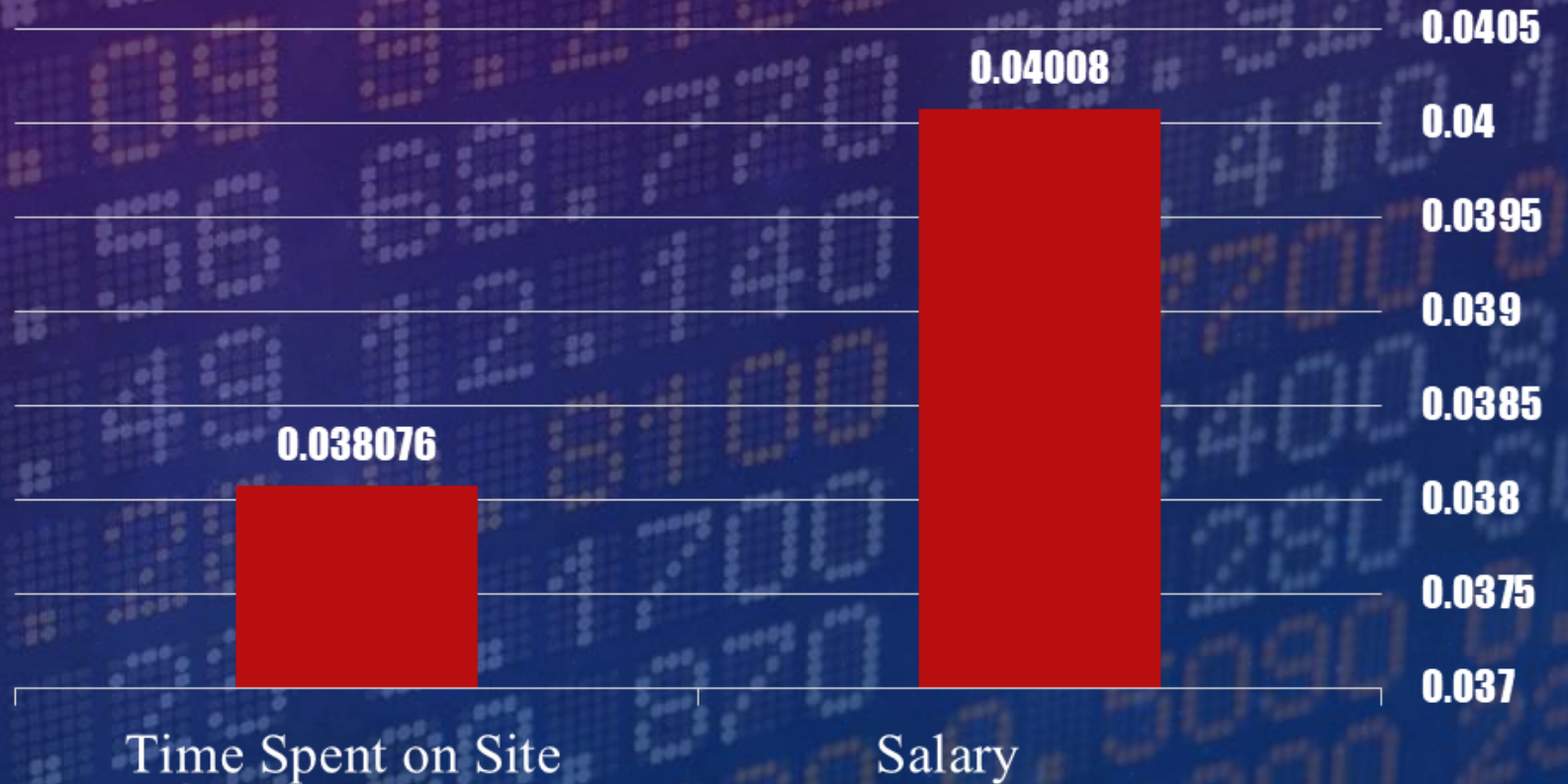
DROP
VARIABLES :
'Names',
'emails', 'Country'



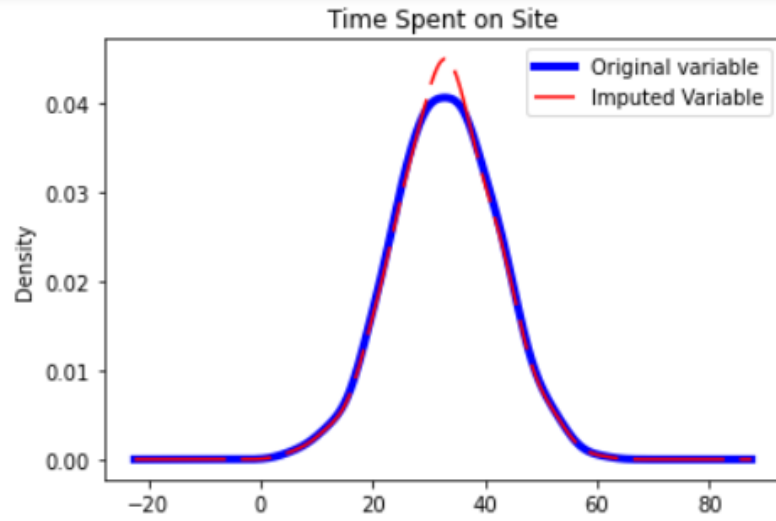
MISSING DATA IN DATASET

Time Spent on Site 0.038076

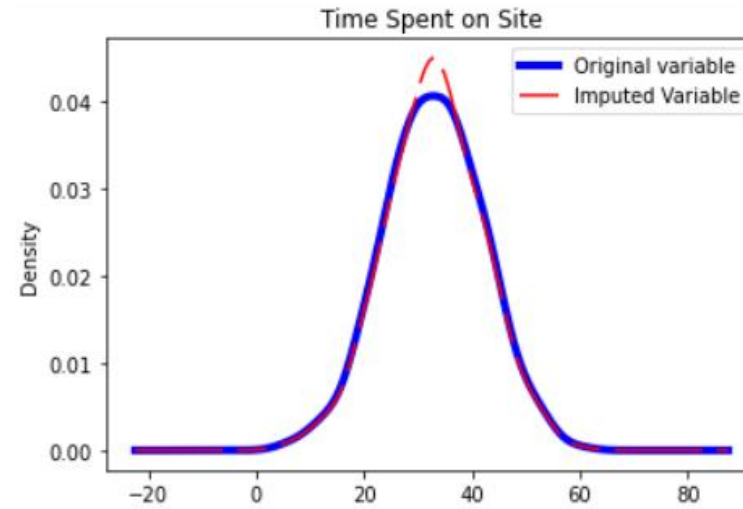
Salary 0.040080



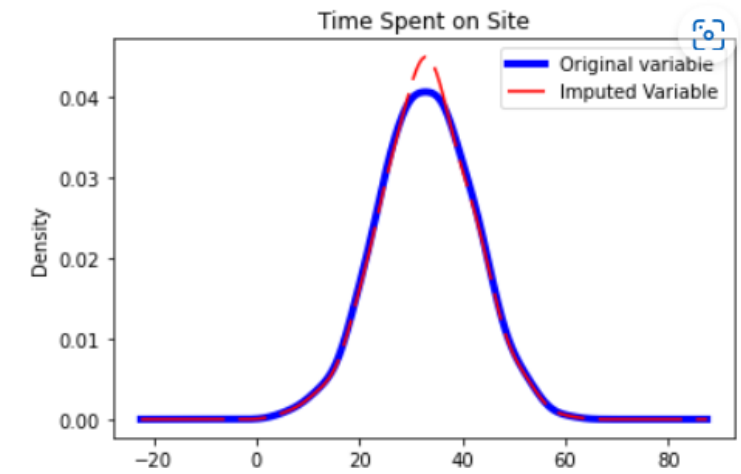
mean



median



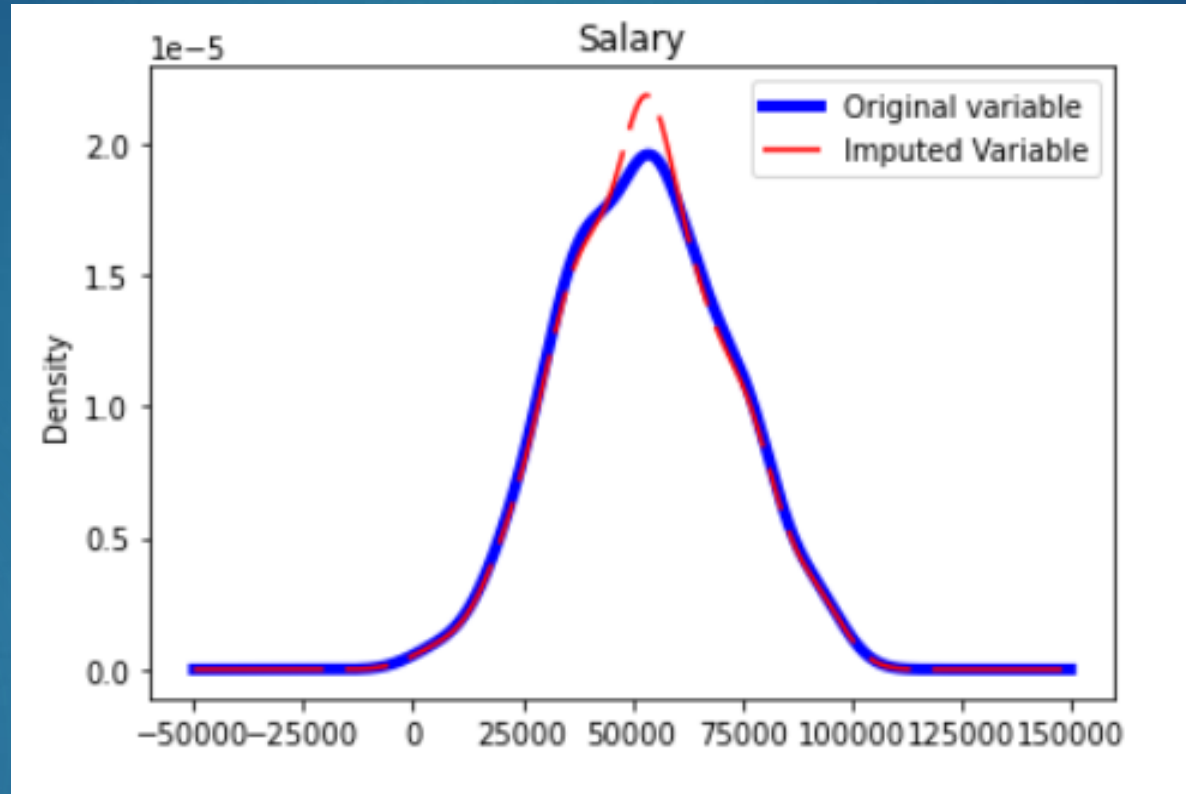
constant



Original variable variance: 83.7
Variance after imputation: 80.5

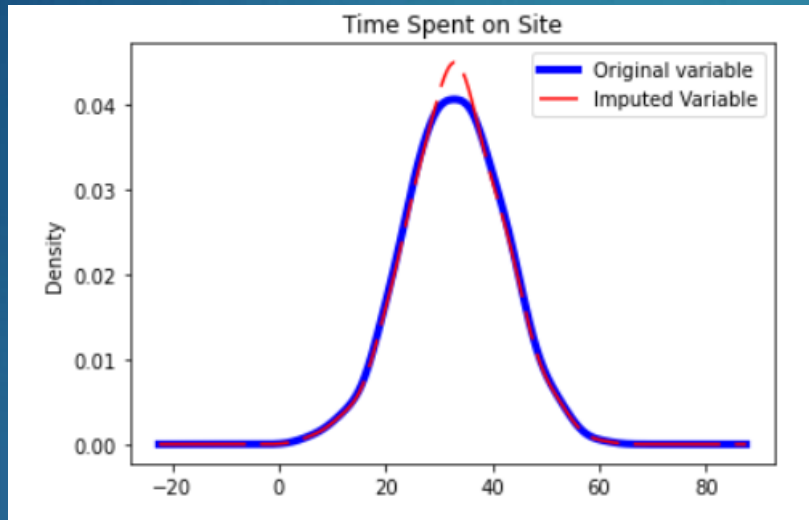
Before and After
Simple Imputer mean, median, constant

Before and After Simple Imputer mean

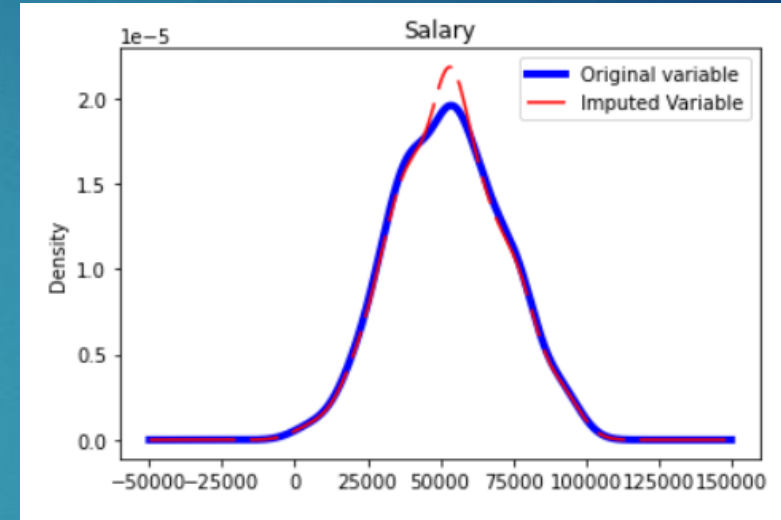


Original variable variance: 360410001.7
Variance after imputation: 345935704.4

Before and After KNN Imputer



Original variable variance: 83.7
Variance after imputation: 80.5



Original variable variance: 360410001.7
Variance after imputation: 345935704.4

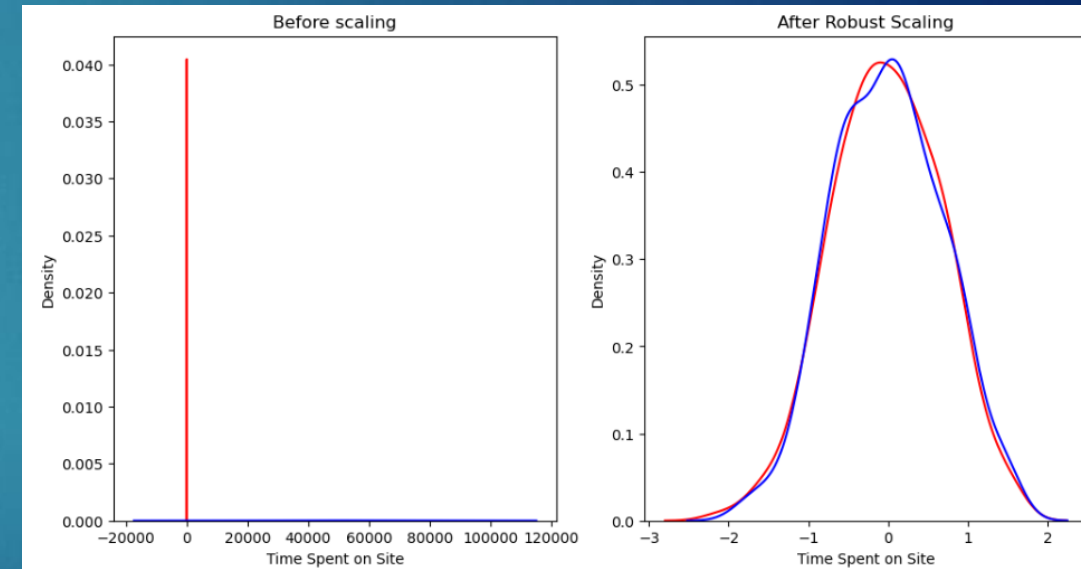
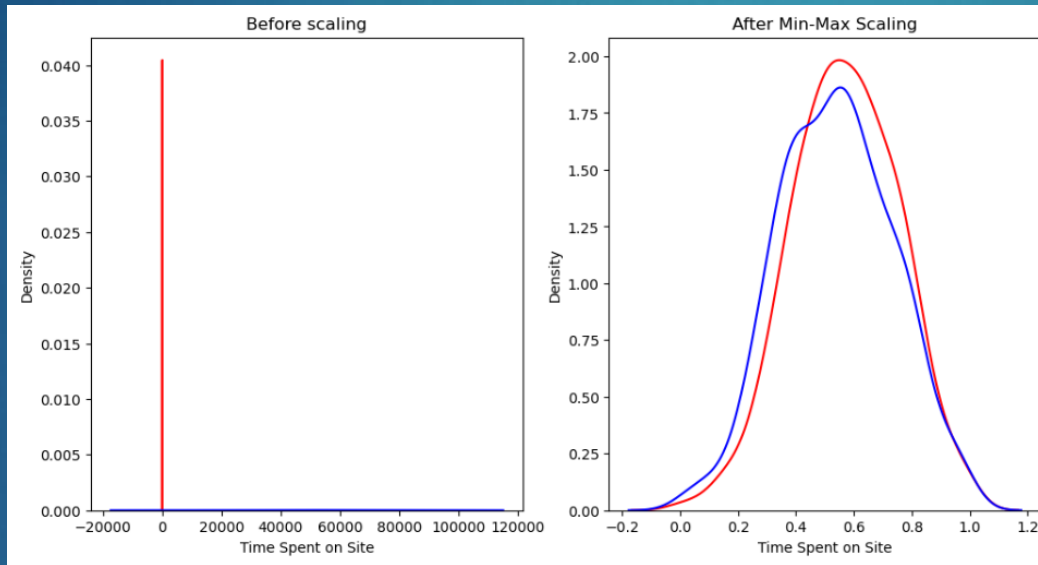
feature magnitude

Scaler Before and After

	Time Spent on Site	Salary	Clicked
count	480.000000	479.000000	499.000000
mean	32.897659	52774.276537	0.501002
std	9.149082	18984.467380	0.500501
min	5.000000	20.000000	0.000000
25%	26.432445	38888.117260	0.000000
50%	33.098900	52623.649480	1.000000
75%	39.193105	65782.900650	1.000000
max	60.000000	100000.000000	1.000000

Min Max Scaler

Robust Scaler



TRAINING MODELS



MACHINE LEARNING MODEL:

NAÏVE BAYES CLASSIFICATION

BAGGING CLASSIFICATION

K-NEIGHBORS CLASSIFICATION

RANDOM FOREST CLASSIFICATION

LOGISTIC REGRESSION

SUPER LEARNER

NAÏVE BAYES CLASSIFICATION

Naïve bayes clasifcation Grid_search best_parameters:

`{'GNB__priors': None, 'GNB__var_smoothing': 1e-08, 'discretizer__bins': 15, 'imputer_num__imputation_method': 'mean'}`

Train set
GaussianNB
roc-auc: 89%

Test set
GaussianNB
roc-auc: 86 %

BAGGING CLASSIFICATION

Bagging Classification Grid search Best parameters:

```
{'Bagging_Classifier_n_estimators': 10,  
'discretizer_bins': 7,  
'imputer_num_imputation_method': 'median'}
```

Train set
BaggingClassifier
roc-auc: 91%

Test set
BaggingClassifier
roc-auc: 85%

K-NEIGHBORS CLASSIFICATION

KNC Grid_search Best_parameters:

```
{'KNC__algorithm': 'ball_tree', 'KNC__leaf_size': 30, 'KNC__n_neighbors': 11, 'KNC__weights': 'uniform', 'discretizer__bins': 15, 'imputer_num__imputation_method': 'mean'}
```

Train set
KNC roc-
auc: 90%

Test set
KNC roc-
auc: 83%

RANDOM FOREST CLASSIFIER

RFC Grid_search Best_parameters:

```
{'RFC__criterion': 'entropy', 'RFC__max_depth': 5, 'RFC__max_features':  
'auto', 'RFC__n_estimators': 100, 'discretizer__bins': 15,  
'imputer_num__imputation_method': 'mean'}
```

Train set
RFC roc-
auc: 91%

Test set
RFC roc-
auc: 86%

LOGISTIC REGRESSION

LR Grid_search Best_parameters:

```
{'LR__C': 0.5, 'LR__max_iter': 100, 'LR__penalty': 'l1', 'LR__solver':  
'liblinear', 'discretizer__bins': 9, 'imputer_num__imputation_method':  
'median'}
```

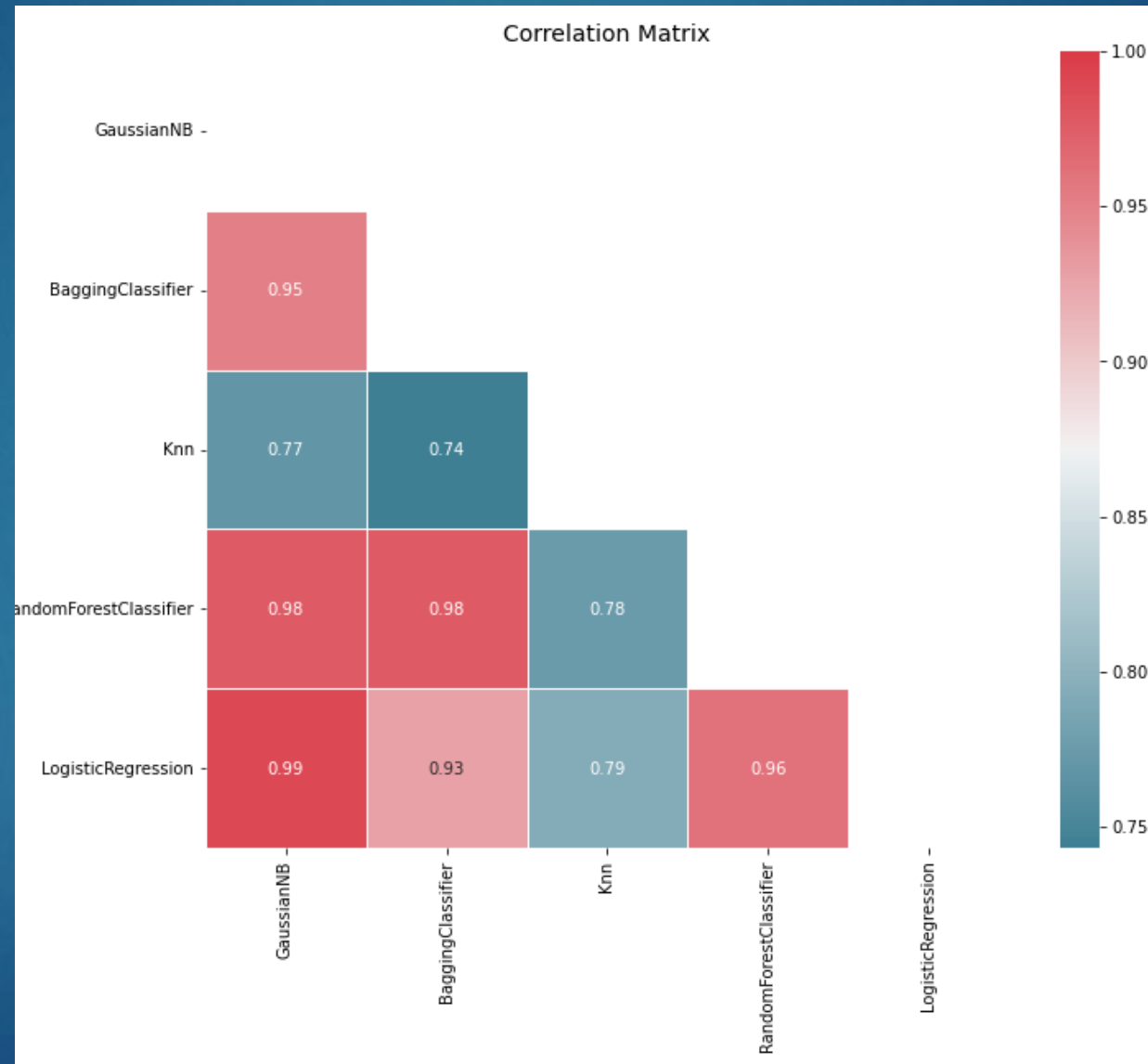
Train set LR
roc-auc:
89%

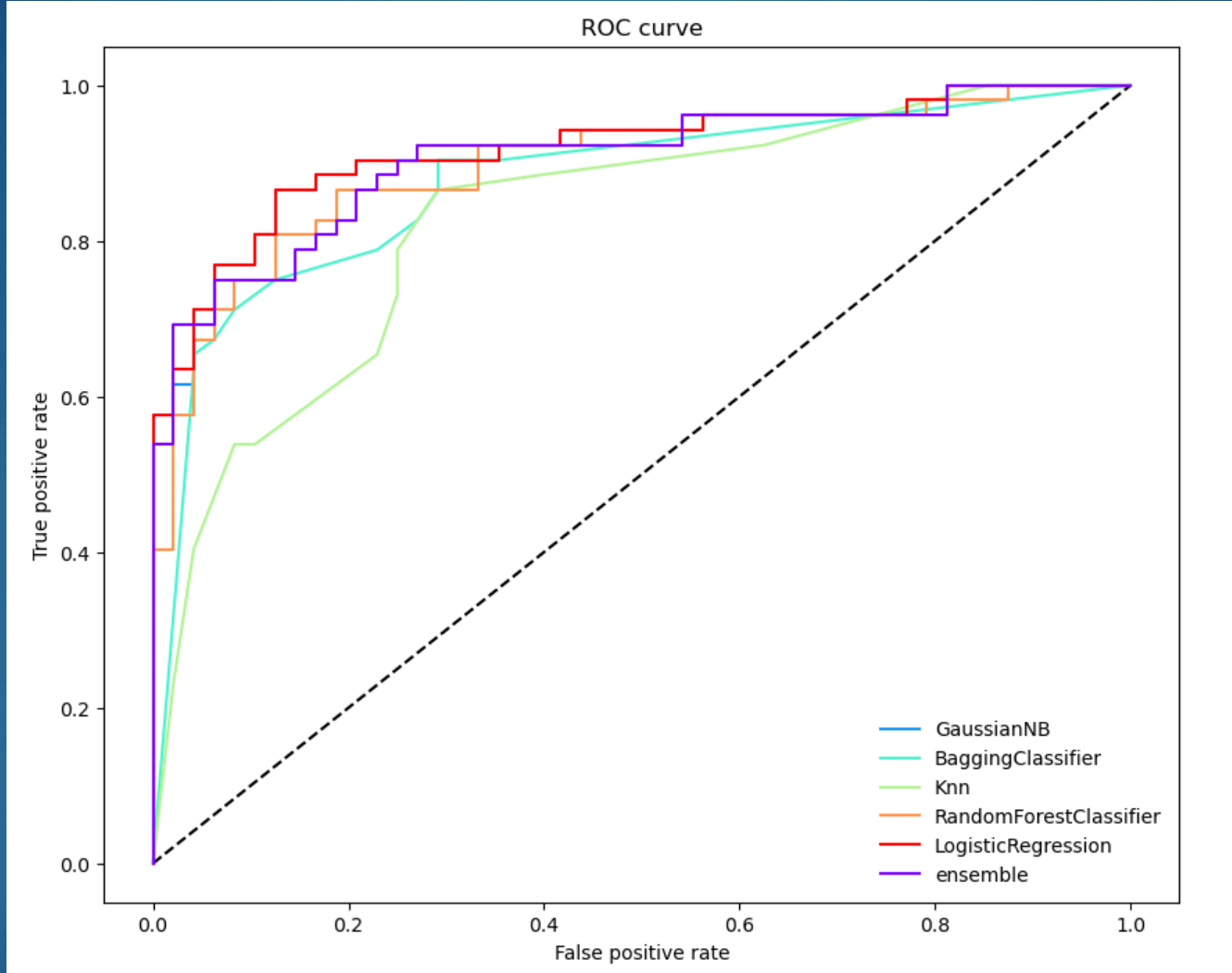
Test set LR
roc-auc:
87%

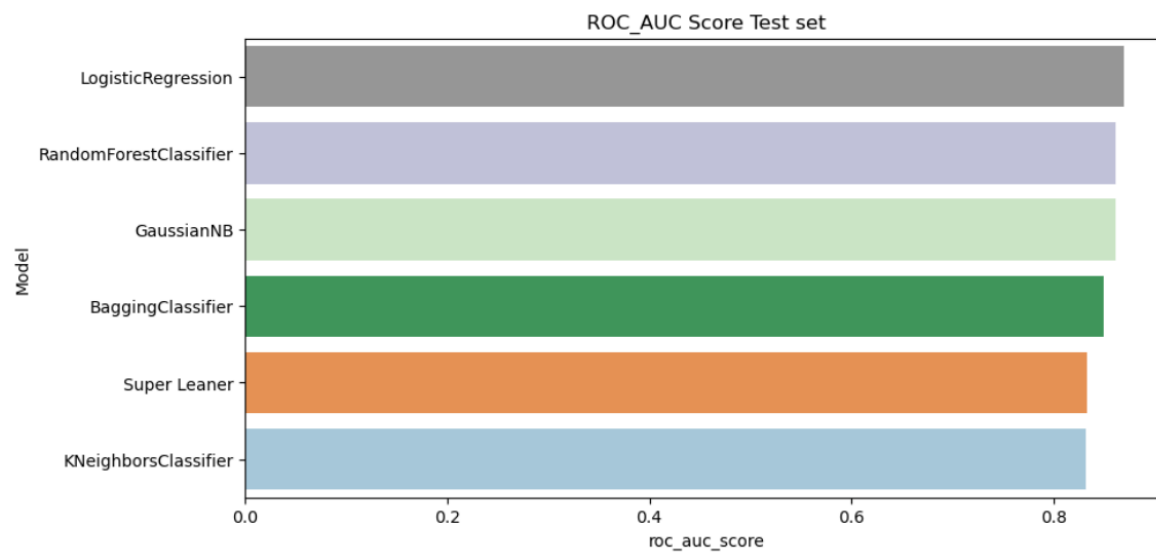
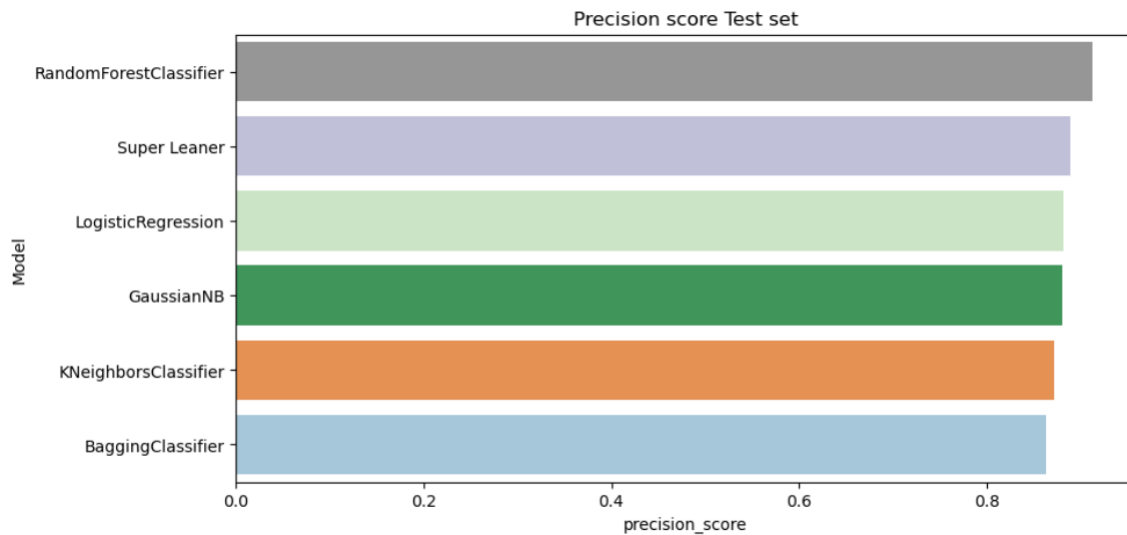
Super Learner

- ▶ Confusion matrix:
 - ▶ $\begin{bmatrix} 43 & 5 \\ 10 & 42 \end{bmatrix}$
- ▶ Precision score: 89%
- ▶ Recall score: 80%
- ▶ Accuracy score: 85%
- ▶ f1_score: 84%

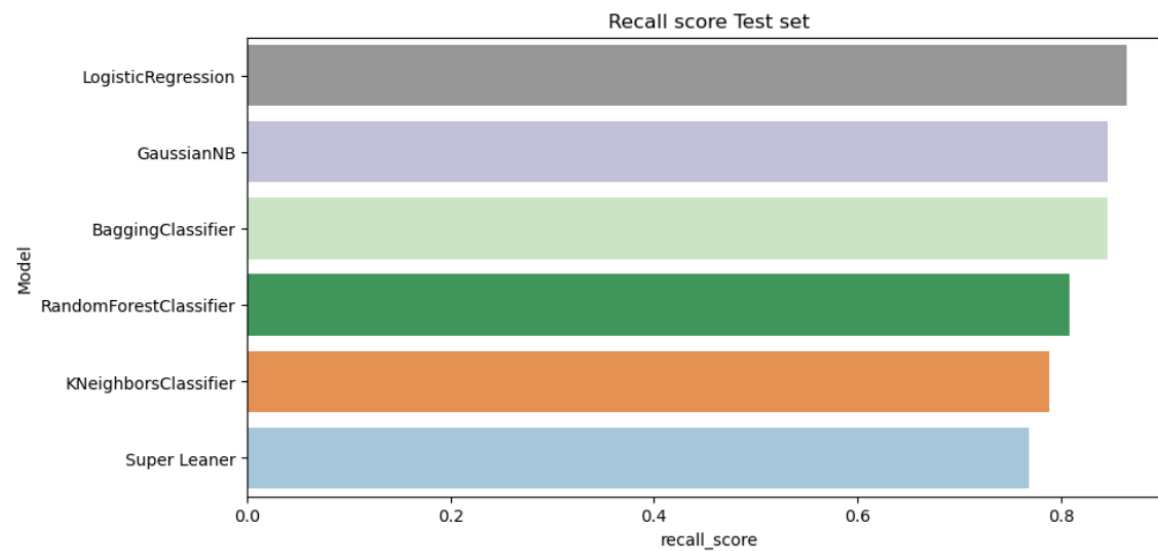
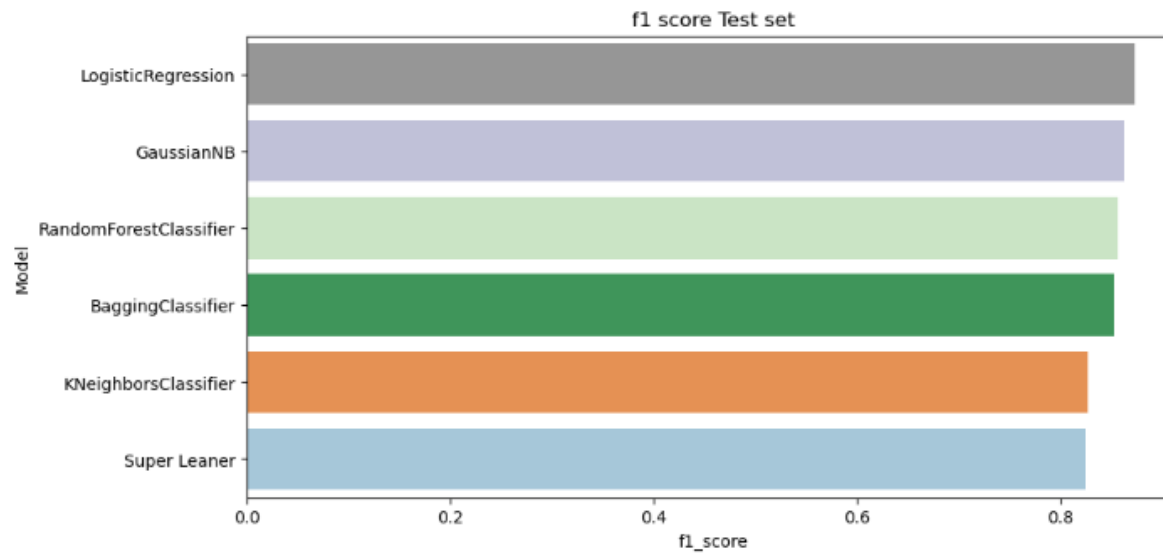
Correlation Matrix of Models







METRIC R RESULTS



METRIC RESULTS

CONCLUSION

► According to the results, the best model is

1. NAÏVE BAYES CLASSIFICATION

2. LOGISTIC REGRESSION

3. RANDOM FOREST CLASSIFICATION