Week 10: Deliverables – Exploratory Data Analysis Report

Group Name: AI Boys

Date: 12/09/2022

Group Member ID	Name	Email	Country	College	Specialization
1	Islom Pulatov	islompulatov115@gmail.com	Poland	Epicode Global	Data Science
2	Ammar Sidhu	ammarsidhu@outlook.com	Canada	University of Toronto	Data Science

Project: Bank Marketing (Campaign)

Github Repo: https://github.com/Islompulatov/Bank_marketing

Problem Description:

- ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Business Understanding:

- The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
- The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

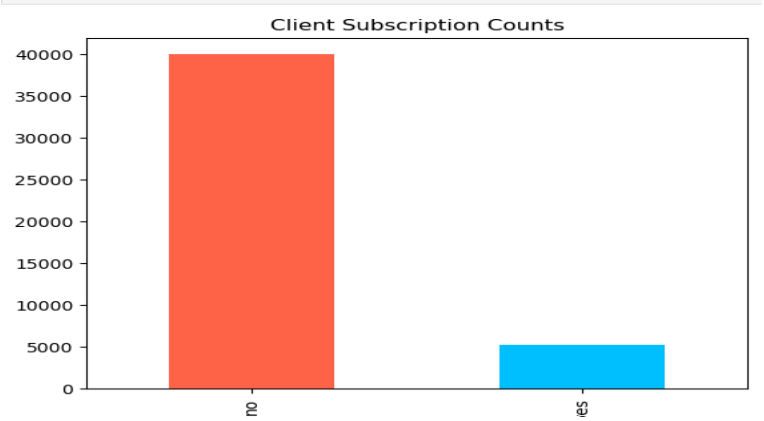
Data Cleaning:

- Found **no missing values** in the dataset
- Found **no duplicate rows** in the dataset
- Would have handled missing values with two different approaches if present:
  (1) **Dropped** Missing Values (Ammar)
  (2) **Impute** Missing Values with **Median** (Islom)
- Would have **dropped duplicate rows** in the dataset if present
- Handled outliers for the **7 Numerical Features** with two different approaches:
  (1) **Do not drop** outliers (Islom)
  (2) **Drop** outliers based on feature and context of the outliers (Ammar)

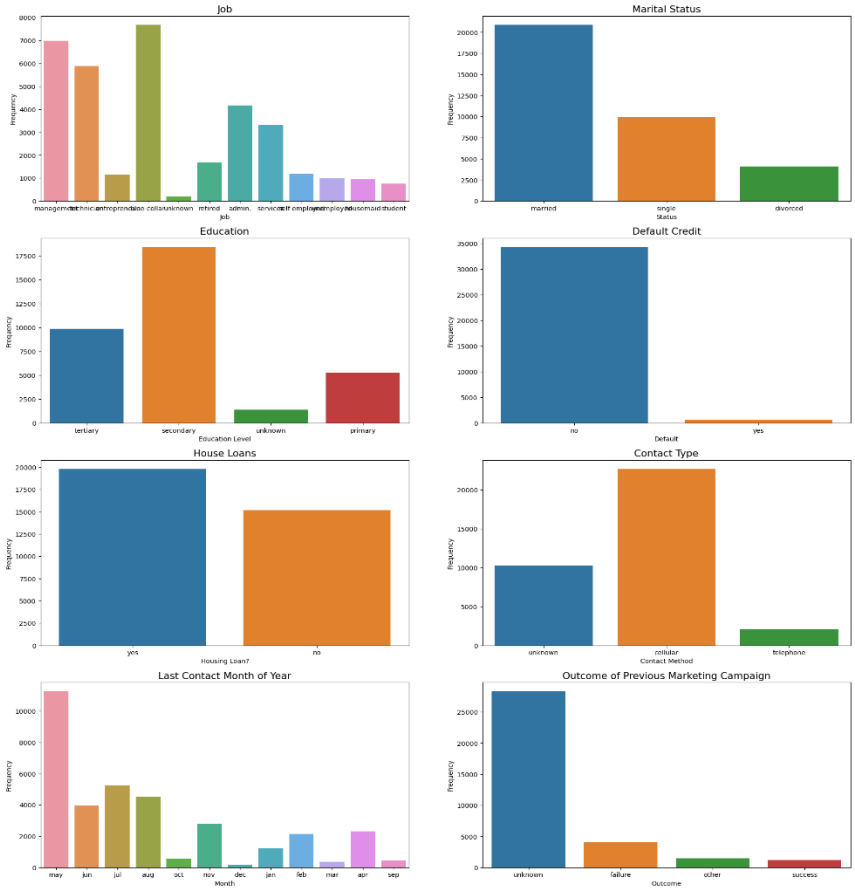
- ‘Unknown’ class for categorical variables were handled in two different ways:
  (1) As a unique class so not a missing value (Ammar)
  (2) Treated as a missing value and drop the corresponding row from data frame (Islom)

EDA Results:

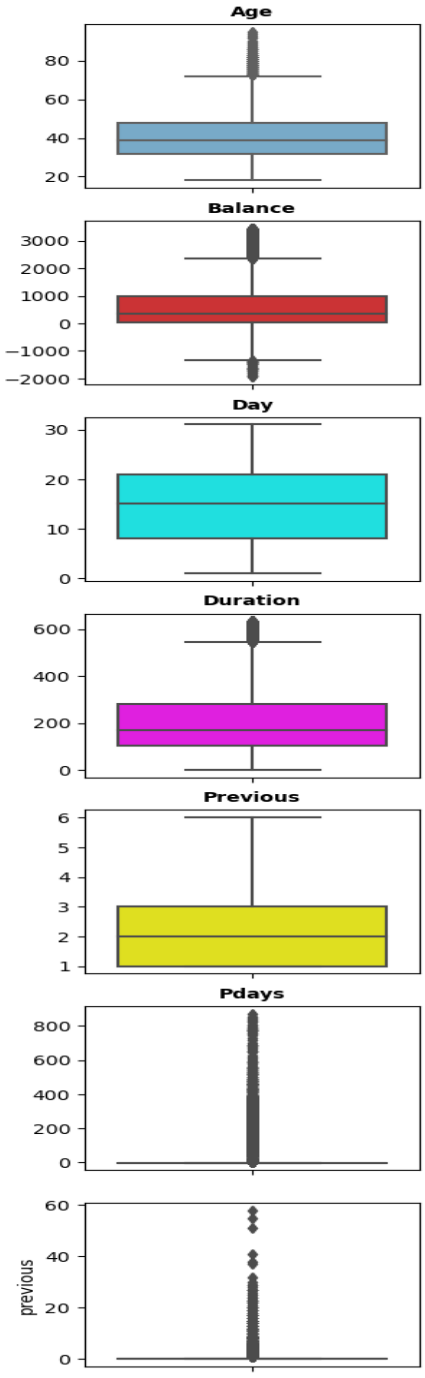
(1) Visualizing Target Feature



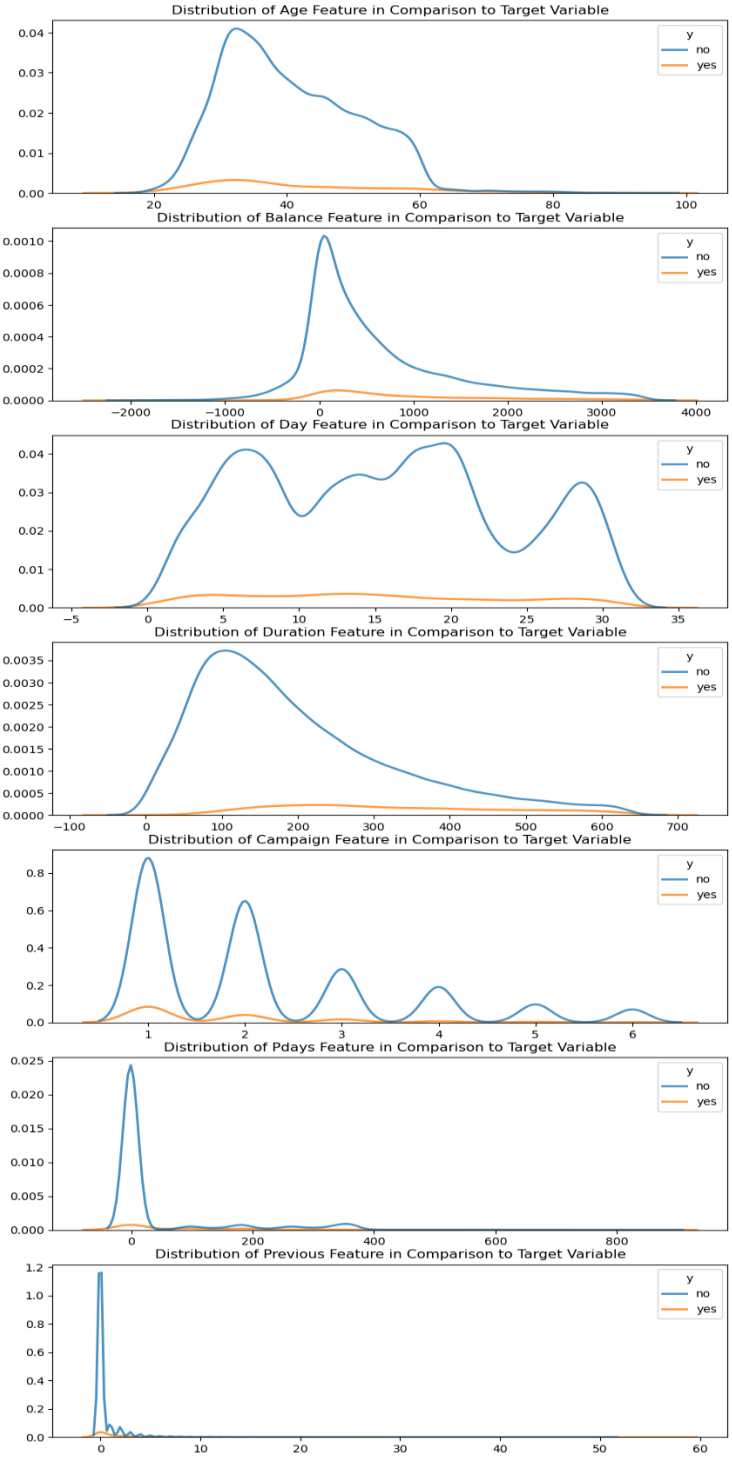
(2) Bar Plots for Categorical Features



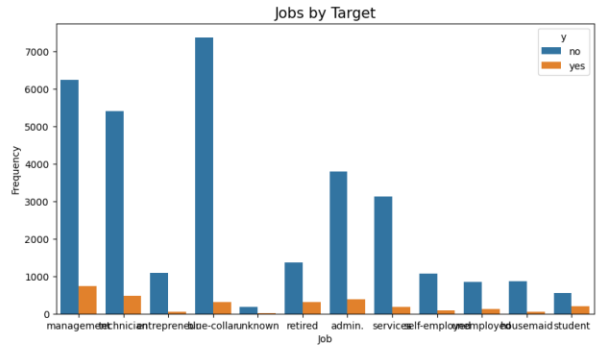
(3) Boxplots Continuous Features (after data cleaning report work)

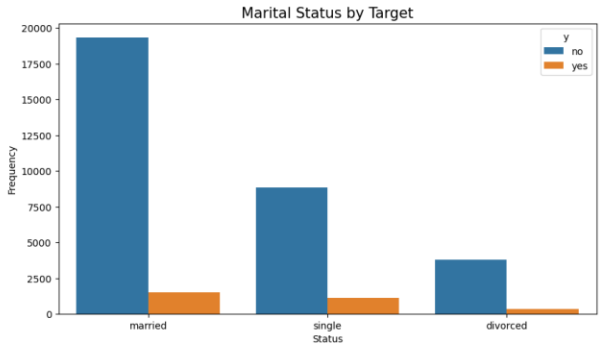


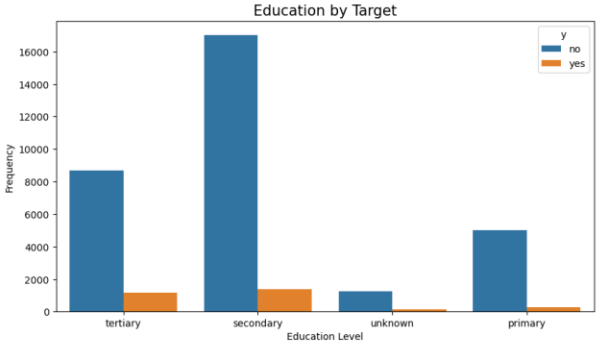
(4) Distribution Plots of Continuous Features by Target

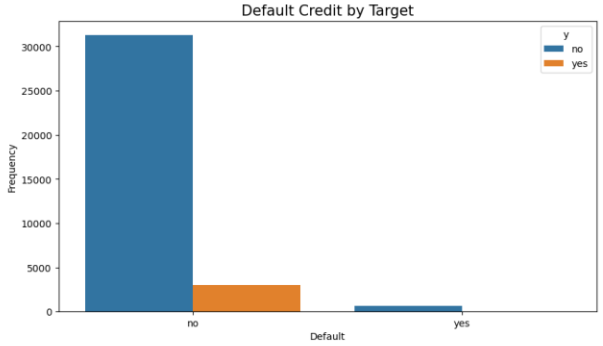


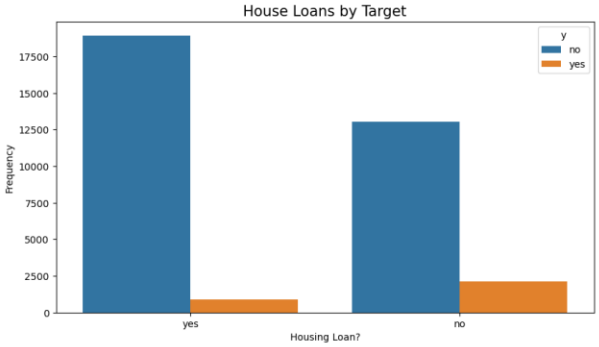
(5) Bar Graphs for Categorical Features by Target

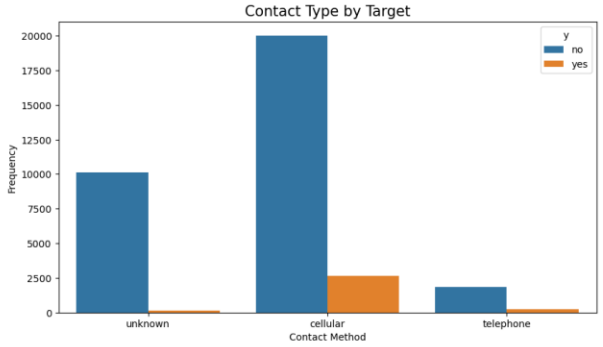


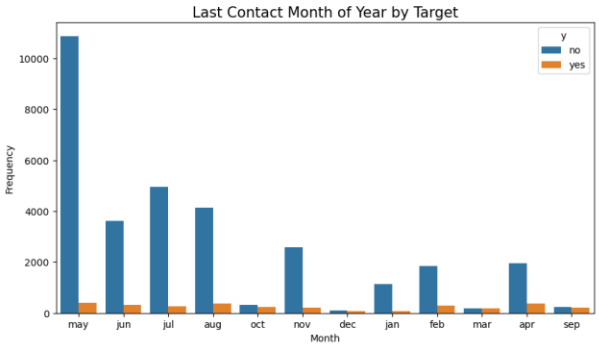


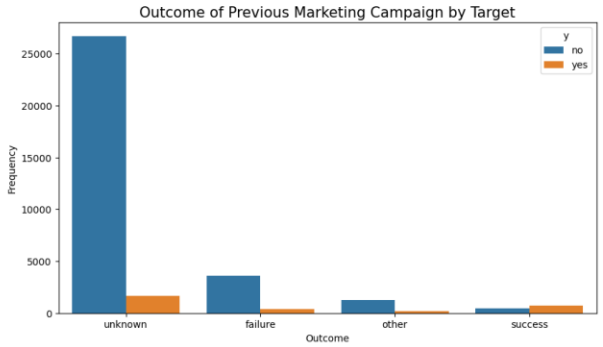








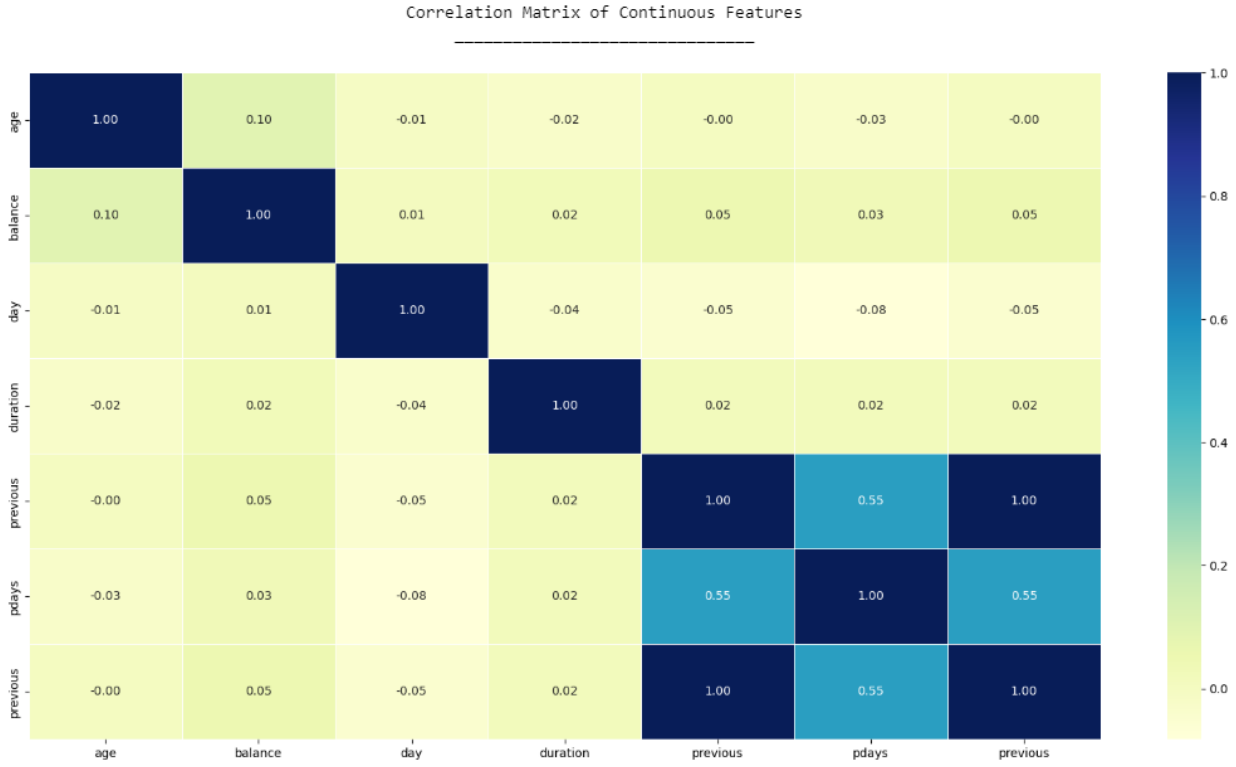




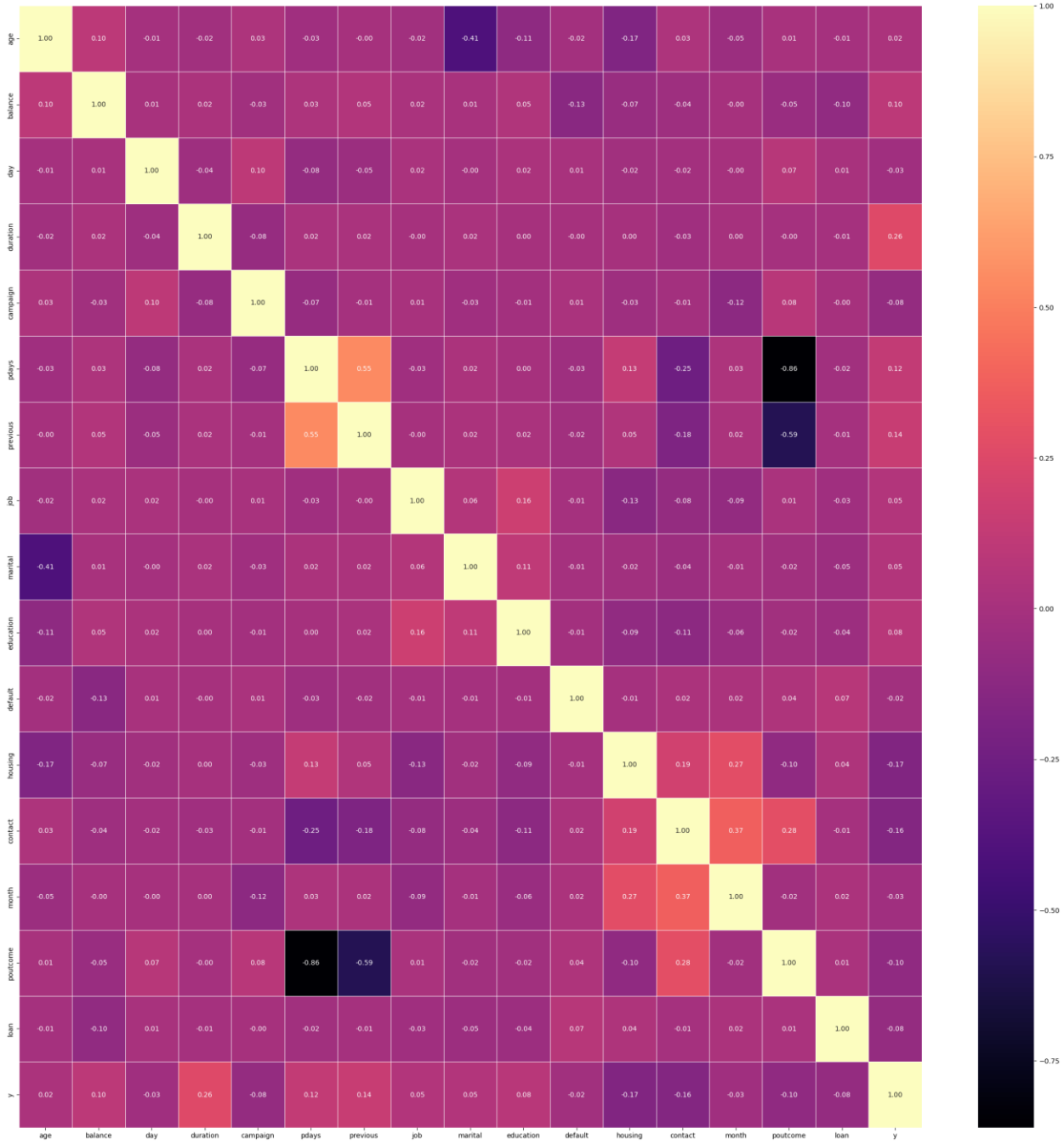
(6) Dataframe with One-Hot Encoded Categorical Features

	age	balance	day	duration	campaign	pdays	previous	job	marital	education	default	housing	contact	month	poutcome	loan	y
0	58	2143	5	261	1	-1	0	4	1	2	0	1	2	8	3	0	0
1	44	29	5	151	1	-1	0	9	2	1	0	1	2	8	3	0	0
2	33	2	5	76	1	-1	0	2	1	1	0	1	2	8	3	1	0
3	47	1506	5	92	1	-1	0	1	1	3	0	1	2	8	3	0	0
4	33	1	5	198	1	-1	0	11	2	3	0	0	2	8	3	0	0
5	35	231	5	139	1	-1	0	4	1	2	0	1	2	8	3	0	0
6	28	447	5	217	1	-1	0	4	2	2	0	1	2	8	3	1	0
7	42	2	5	380	1	-1	0	2	0	2	1	1	2	8	3	0	0
8	58	121	5	50	1	-1	0	5	1	0	0	1	2	8	3	0	0
9	43	593	5	55	1	-1	0	9	2	1	0	1	2	8	3	0	0

(7) Correlation Matrix of Continuous Features



(8) Correlation Matrix of all Features in Bank Marketing Features



Inference from Correlation Analysis:

- There is no multicolinearity between the numerical features.
- The only feature with a moderate correlation with the target - `y` is the `duration` feature.
- There is a strong correlation between the encoded `poutcome` feature, and the `pdays` feature.
- Some features are negatively correlated with each other.

Conclusions from EDA

- There are no NaN values and no duplicate values in the dataset.
- The target feature is imbalanced as there are more than 8x the customers who subscribed vs. who dd not.
- Except for the `duration` feature, all the other features have a low correlation with the target.
- Most customers are married, have loans, and work collar jobs.

Next Steps and Recommendations:

- Test and Train Ensemble and Boosting Classification Models with Cost-Sensitive Learning (Ammar).
- Test and Train Ensemble and Boosting Classification Models with SMOTE (Islom).
- Tune Hyperparameters of the Best Performing Models.
- Compare Model Performances and check to see if dropping ‘unknown’ entries had an impact on accuracies.