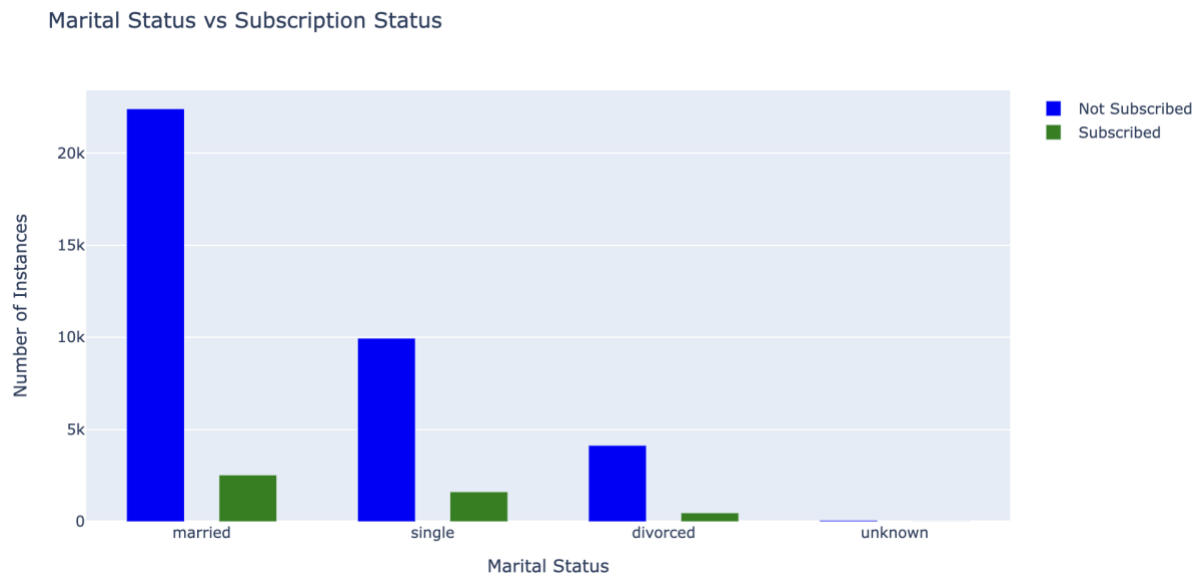
Bank Marketing

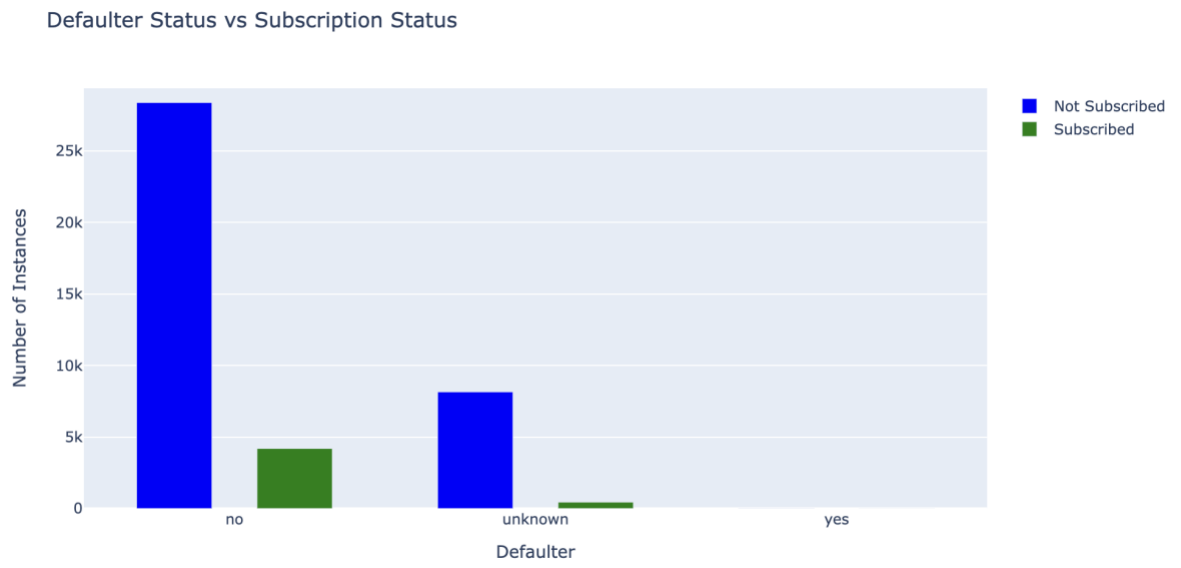
This project has as its objective the analysis of the “Bank Marketing” dataset, containing data relating to marketing campaigns based on telephone calls of a Portuguese bank. Our goal is to predict, through binary classification, whether or not the customer will subscribe to a term bank deposit.

Findings

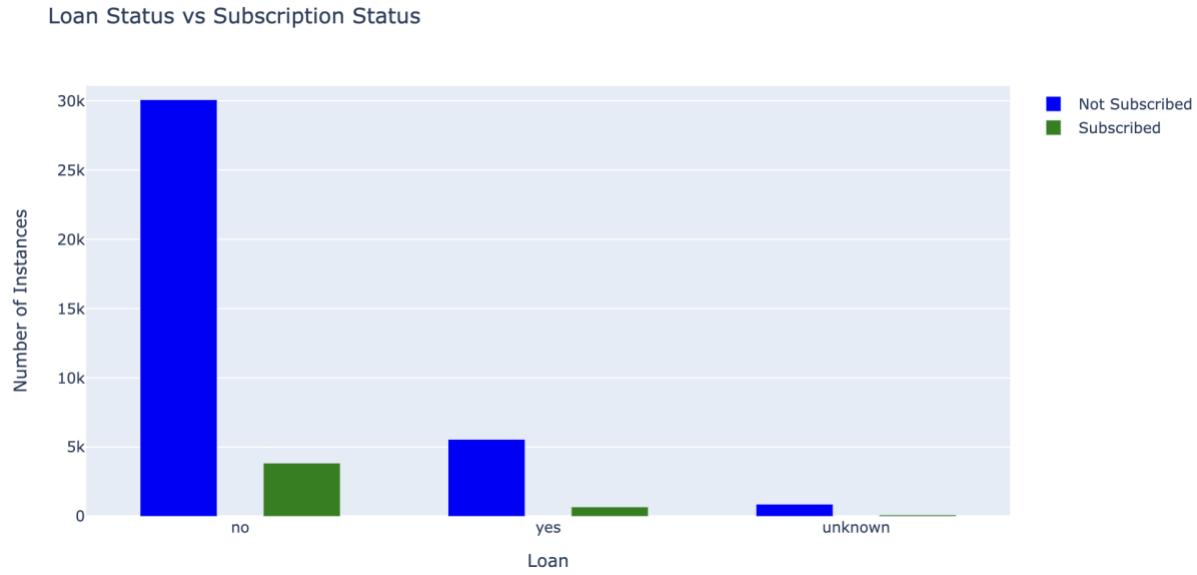
1) Contacts who are married are subscribed more compared to those that are single and divorced. The total number of those unsubscribed far exceeds those that are subscribed.



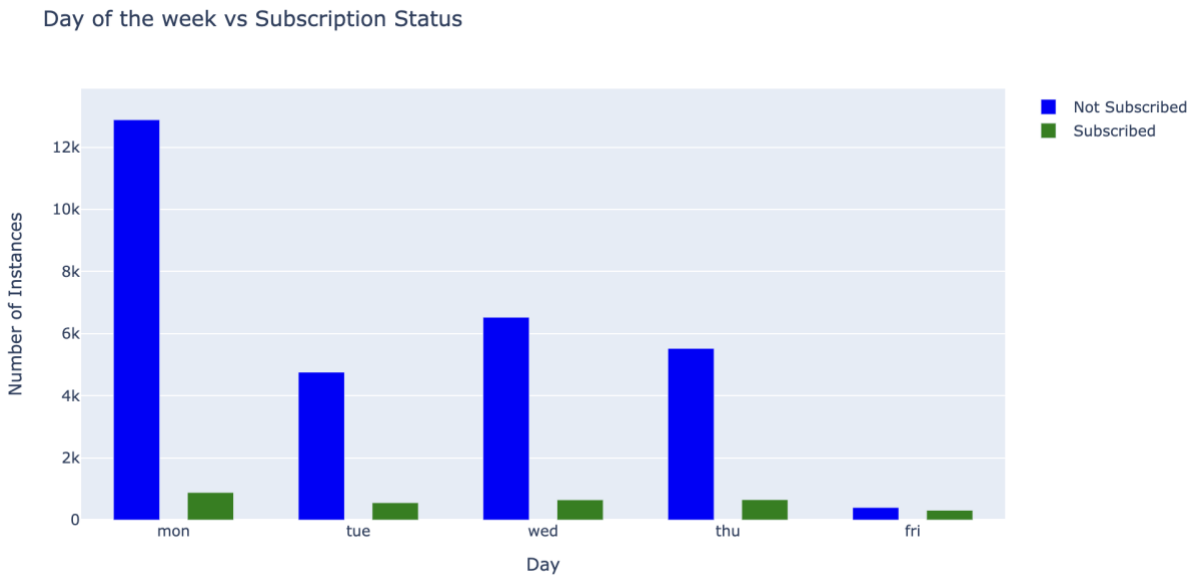
2) The number of defaulters is almost none compared to those unknown or with value yes. Those who have not defaulted, have subscribed more compared to those who have defaulted or to those who have no information.



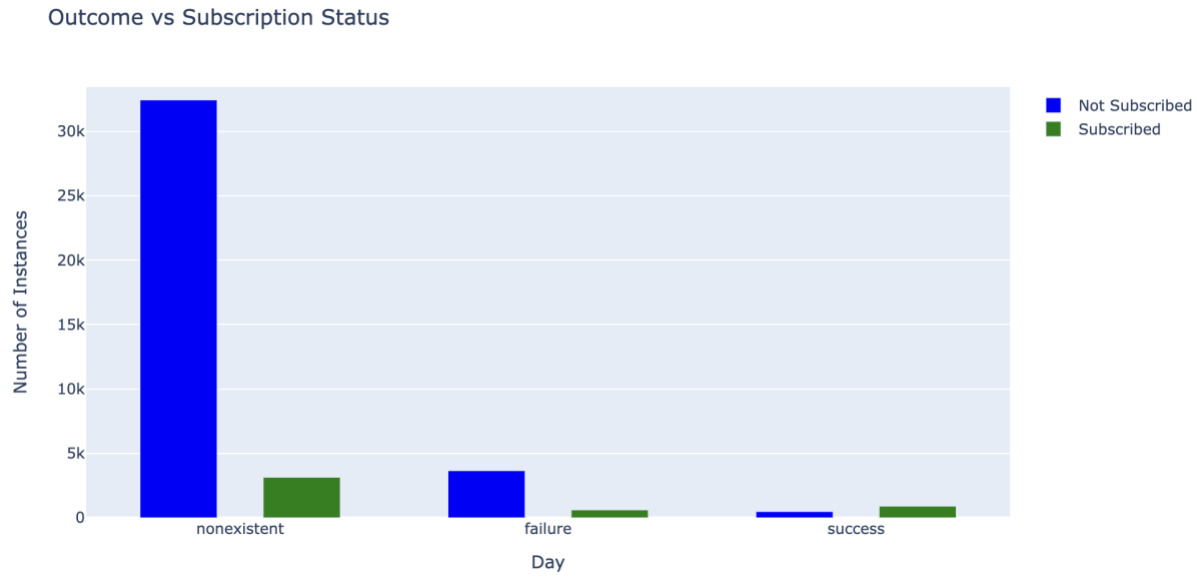
3) Those without any loans are subscribed more compared to those who have loans.



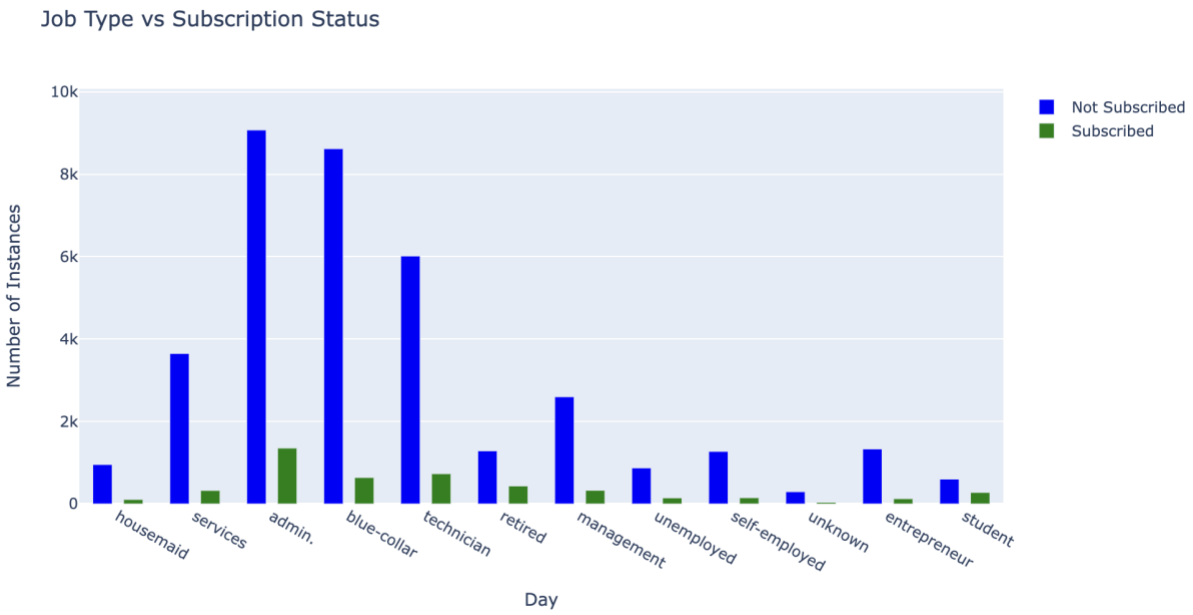
4) Monday is the day where most people subscribed and unsubscribed compared to the other days of the week.



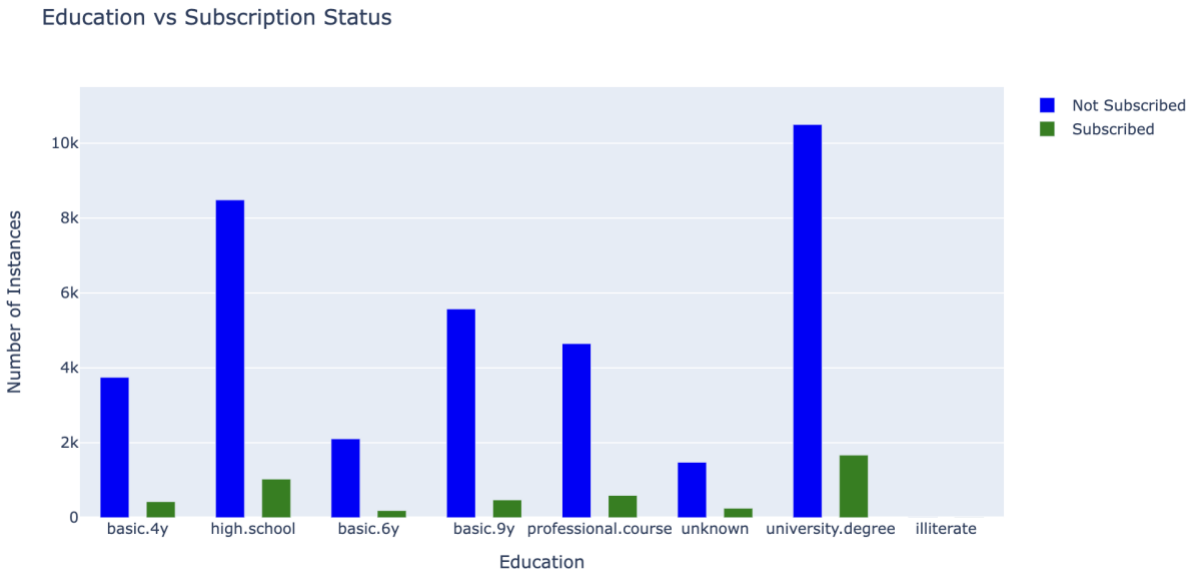
5) Those with non-existent outcomes have more subscriptions compared to those that are failure or success.



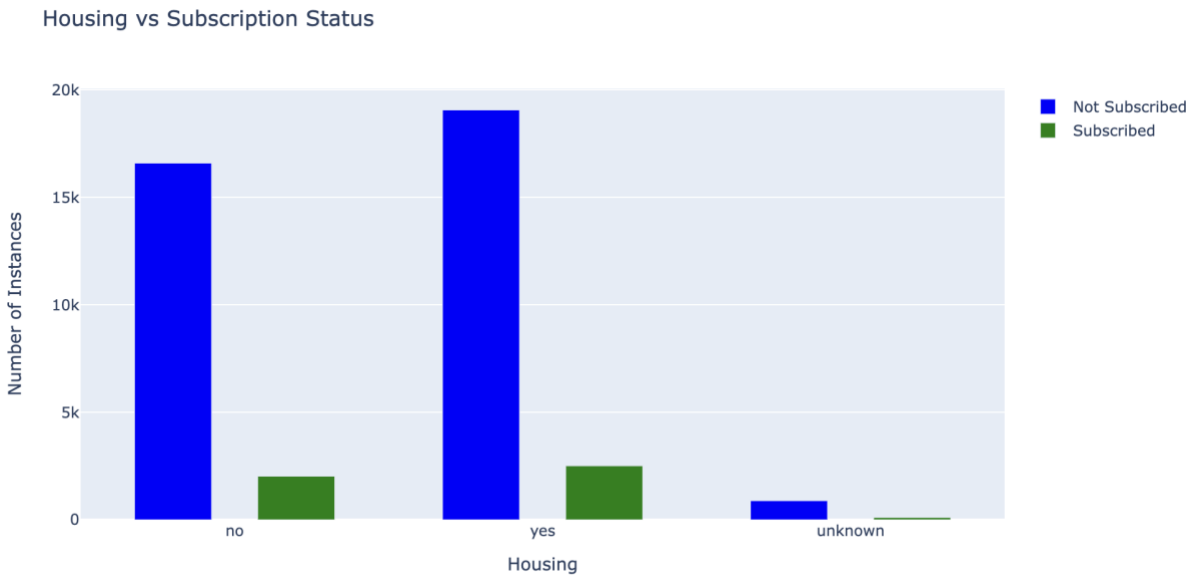
6) Those with admin, blue-collar and technician job types subscribe more compared to other job types.



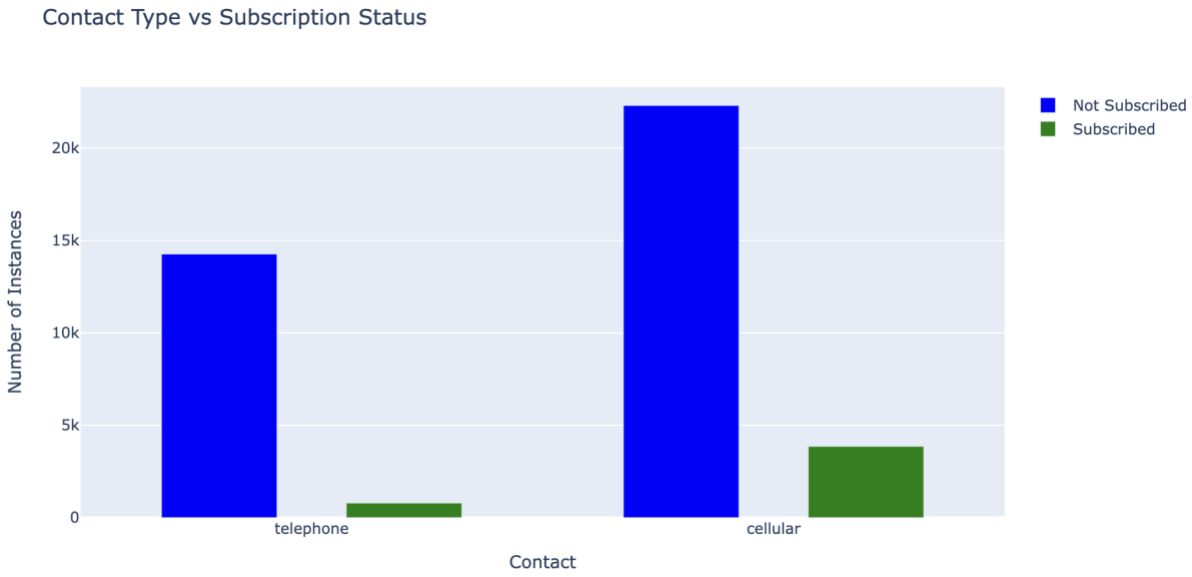
7) Contacts with high school and university degree subscribe more compared to other education levels.



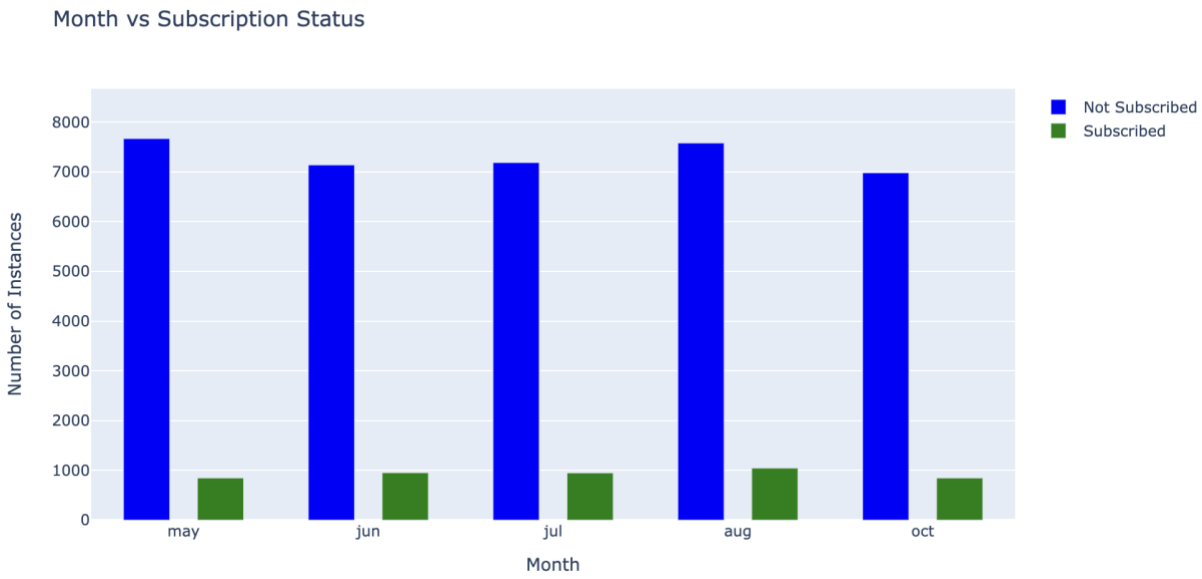
8) Not much difference between those with housing and those without housing. However, those with housing seem to have subscribed more compared to those without housing.



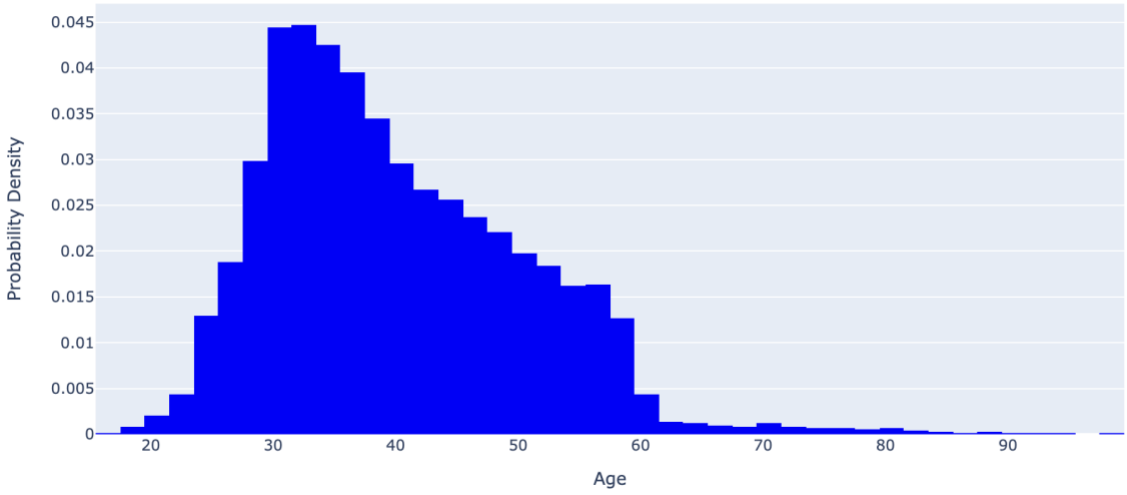
9) Those with cellular contact type subscribe more compared to those with telephone contact type.



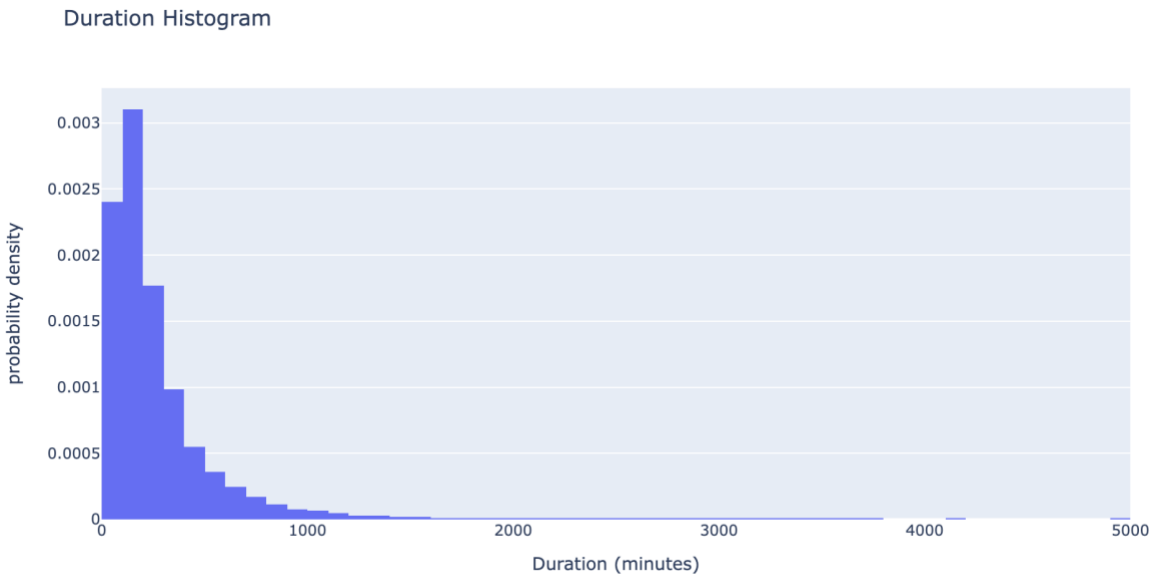
10) August month had a greater number of subscriptions compared to other months.

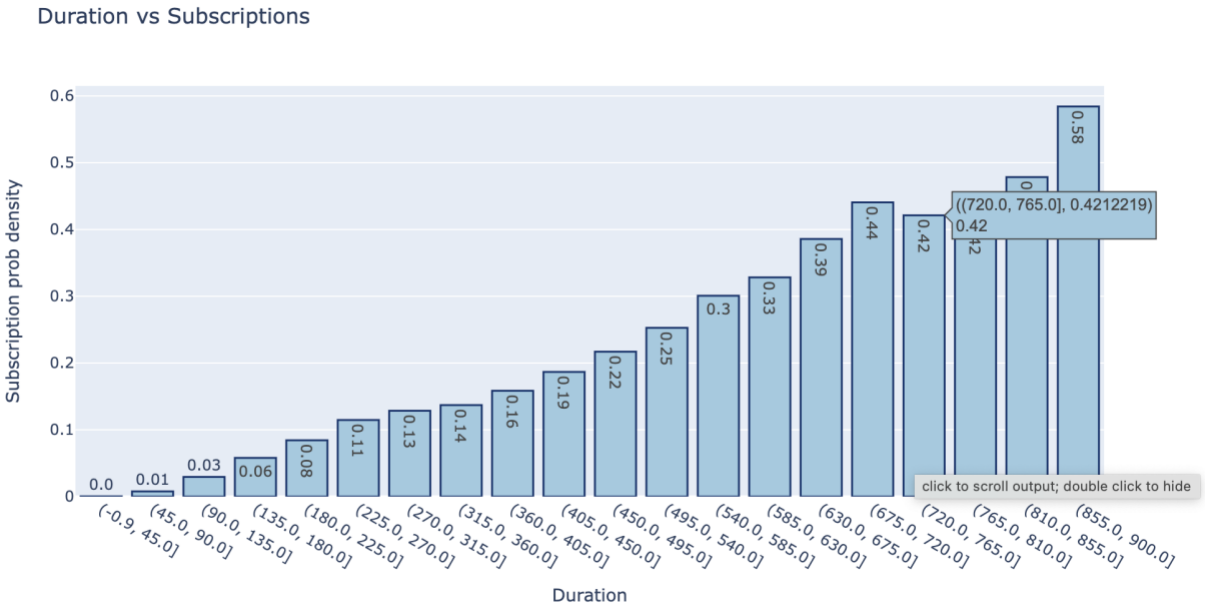


11) Majority of contacts are <= 60 years. Among those with age <= 60, the younger the clients are most likely to subscribe.

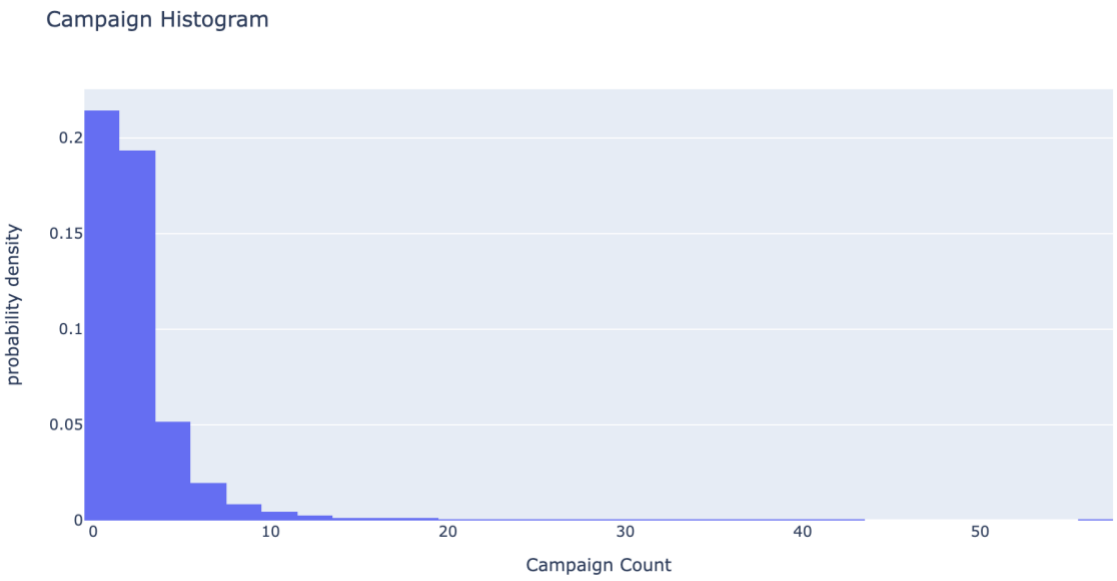


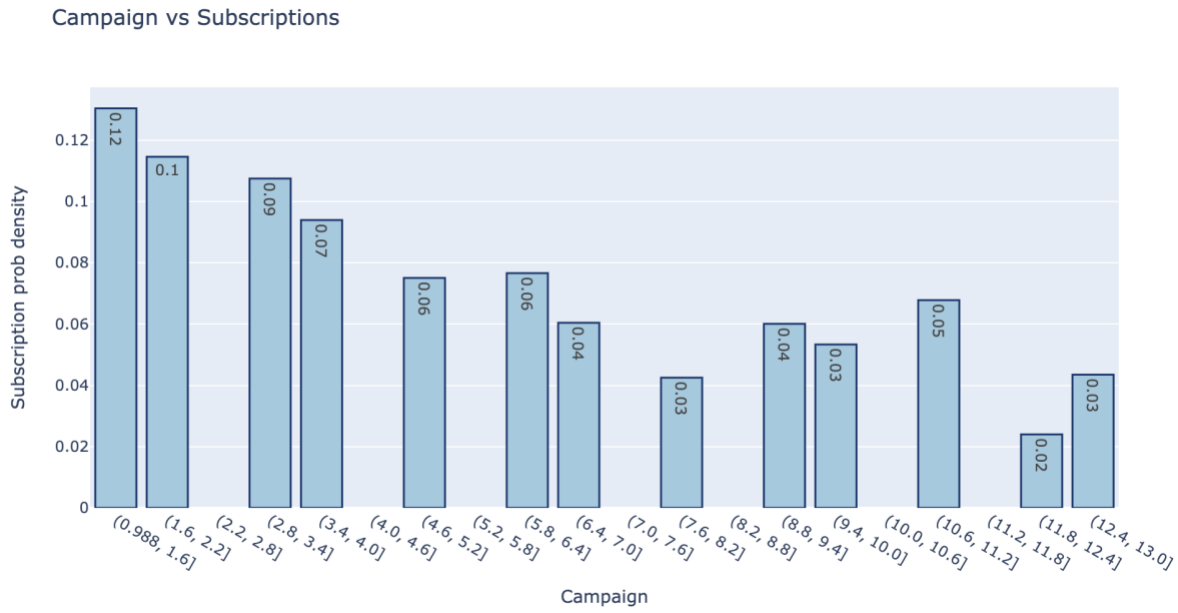
12) Clients with longer lasting contact duration are more likely to subscribe.





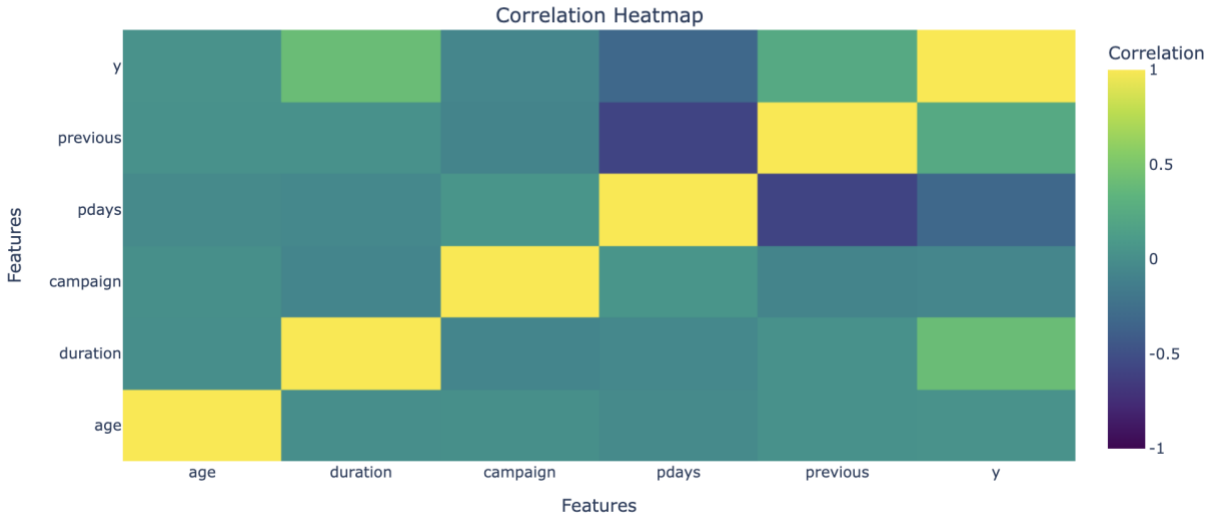
13) Based on campaigns, clients being contacted more are less likely to subscribe.





14) Pdays & Previous are correlated, while the rest are not considered correlated with each other.

Correlation Heatmap



Models

	Model Name	Train Time	Train Accuracy	Test accuracy	f1_score	balanced_accuracy	roc_auc score
0	SVM	113.990209	0.887709	0.885895	0.832294	0.500000	0.470893
0	SVM	0.040731	0.898877	0.878369	0.840459	0.524020	0.545889
0	Logistic Regression	0.263456	0.887709	0.885895	0.832294	0.500000	0.617335
0	KNN	0.001981	0.999575	0.854212	0.851407	0.621139	0.621139
0	DecisionTree - Cross Validation	0.066258	0.999643	0.859727	0.657642	0.861010	0.657797
0	SVM - Cross Validation	101.120975	0.894105	0.894082	0.582559	0.868078	0.697320
0	KNN-RandomSearchCV	132.429425	0.904036	0.889658	0.870722	0.607780	0.794271
0	LogisticRegression-GridSearchCV	0.489536	0.894021	0.892814	0.864763	0.577123	0.834881
0	DecisonTree-RandomSearchCV	1.255225	0.900880	0.893178	0.874695	0.615791	0.841610
0	Logistic Regression - Cross Validation	0.296021	0.894090	0.893748	0.583670	0.868209	0.842984

From the above, it is prety clear that Logistic Regression with Cross Validation performed much better compared to other models.