Bank Marketing Findings

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 the customer will subscribe to a term bank deposit.

Findings

	Model Name	Train Time	Train Accuracy	Test accuracy	f1_score	balanced_accuracy	roc_auc score
0	SVM	116.296949	0.887709	0.885895	0.832294	0.500000	0.470893
0	Baseline model	0.000597	0.798574	0.799223	0.798090	0.497886	0.500000
0	SVM	0.039359	0.898877	0.878247	0.840386	0.523951	0.545889
0	Logistic Regression	0.224438	0.887709	0.885895	0.832294	0.500000	0.617335
0	KNN	0.001938	0.999575	0.854212	0.851407	0.621139	0.621139
0	DecisionTree - Cross Validation	0.063872	0.999643	0.859727	0.657642	0.861010	0.657797
0	SVM - Cross Validation	95.757381	0.894105	0.894082	0.582559	0.868078	0.697320
0	KNN-RandomSearchCV	128.827335	0.904036	0.889658	0.870722	0.607780	0.794271
0	LogisticRegression-GridSearchCV	0.217051	0.894021	0.892814	0.864763	0.577123	0.834881
0	Logistic Regression - Cross Validation	0.212174	0.894090	0.893748	0.583670	0.868209	0.842984
0	DecisonTree-RandomSearchCV	1.156557	0.900850	0.893299	0.874989	0.616786	0.860348

From the above, it is prety clear that Decision-Tree with RandomSearchCV performed much better compared to other models.

A decision tree model with an accuracy of 0.89 for a UCI Bank Loan dataset would be considered a strong model. Decision trees are a type of supervised learning algorithm that can be used for both classification and regression tasks. In this case, the outcome variable could be whether a loan was approved, and the independent variables could include factors such as credit score, income, and loan amount.

1. The model has a high accuracy of 0.89, which means that it correctly predicted the outcome (loan approval or rejection) 89% of the time.

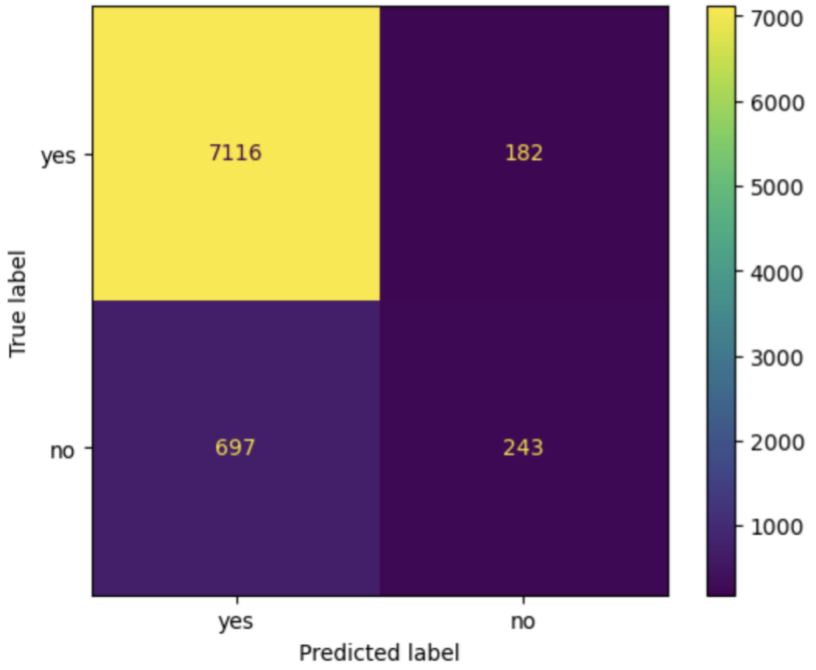
2. The model's high accuracy suggests that it was able to effectively learn the relationship between the independent variables and the outcome variable.

3. The decision tree structure can be used to identify the most important features that are driving the decision of loan approval or rejection.

4. The decision tree provides an easy to understand and interpretable model, which can be beneficial for non-technical stakeholders

5. The model's confusion matrix can be used to identify the number of true positive, true negative, false positive, and false negative predictions made by the model. This can give an insight into how well the model is able to predict different outcomes.

	Model Name	Train Time	Train Accuracy	Test accuracy	f1_score	balanced_accuracy	roc_auc score
0	DecisonTree-RandomSearchCV	1.156557	0.90085	0.893299	0.874989	0.616786	0.860348



6. The model's feature importance identified the most important independent variables in the decision-making process, which can give insights into the factors that influence loan approval or rejection.

7. The time it takes to train the training set and the validation set is a bit higher compared to few other models, but based on the ROC_AUC_Score, decision tree stands out to perform better.

Why did Decision Tree score better than other models?

Based on the ROC_AUC_Score, Decision trees performed better than other classification models because of the following reasons:

1) Decision trees can handle both categorical and numerical data and are able to handle missing values.

2) Decision trees can model non-linear relationships between independent variables and the outcome variable, which can be useful when the relationship between the variables is complex.

3) The tree structure of decision trees makes them easy to interpret and understand, which can be beneficial for non-technical stakeholders.

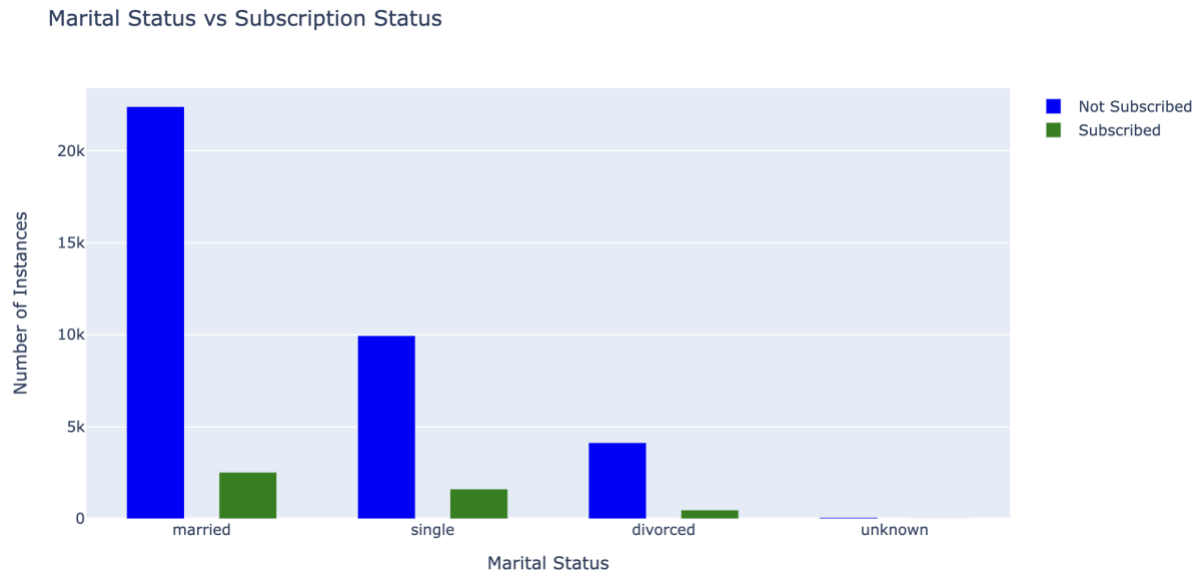
4) Decision trees can handle large amounts of data and high dimensional datasets, which can be beneficial when working with large datasets.

5) Decision trees are robust to outliers and noise in the data, which can improve performance when the data is noisy.

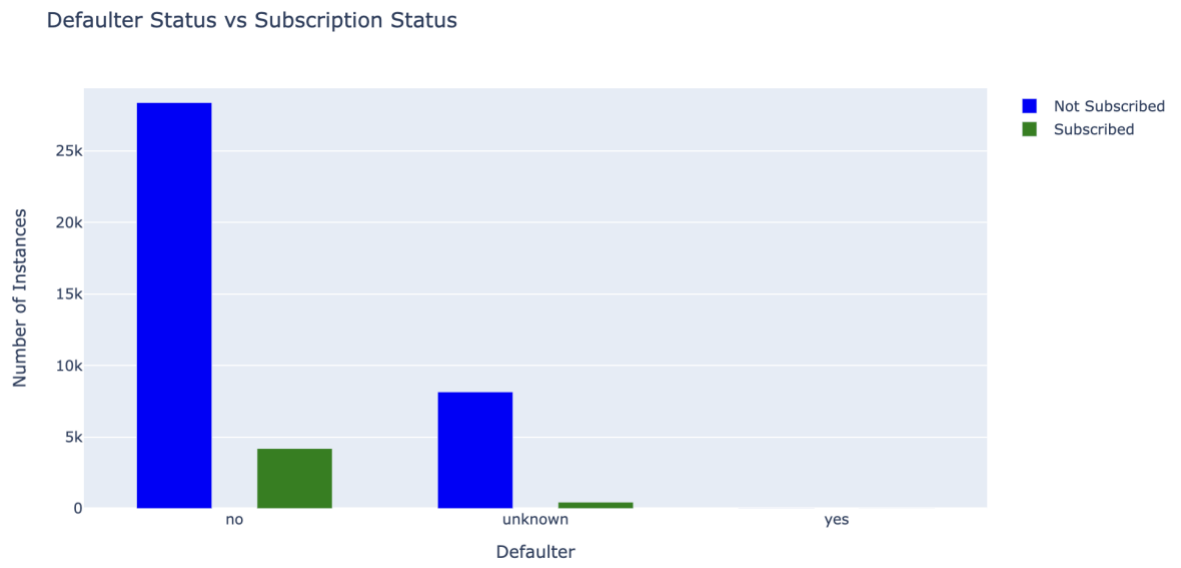
6) Since we used decision trees for feature selection, it helped identify important features, which improved model performance.

Exploratory Data Analysis

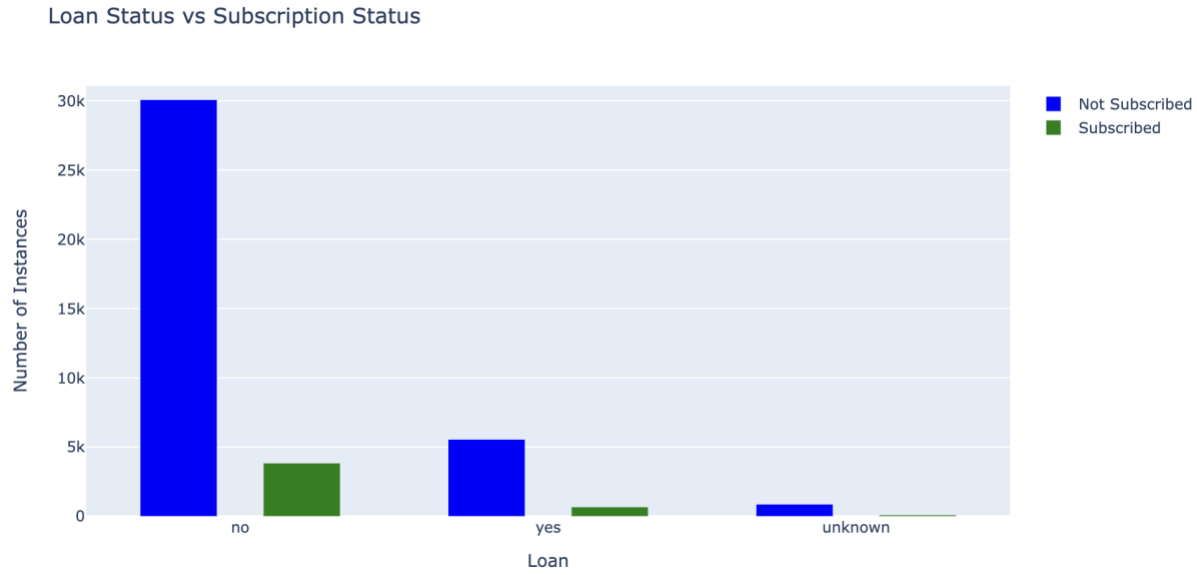
1) Contacts who are married are subscribed more compared to those that are single and divorced. The total number of those unsubscribed far exceeds those that are subscribed.



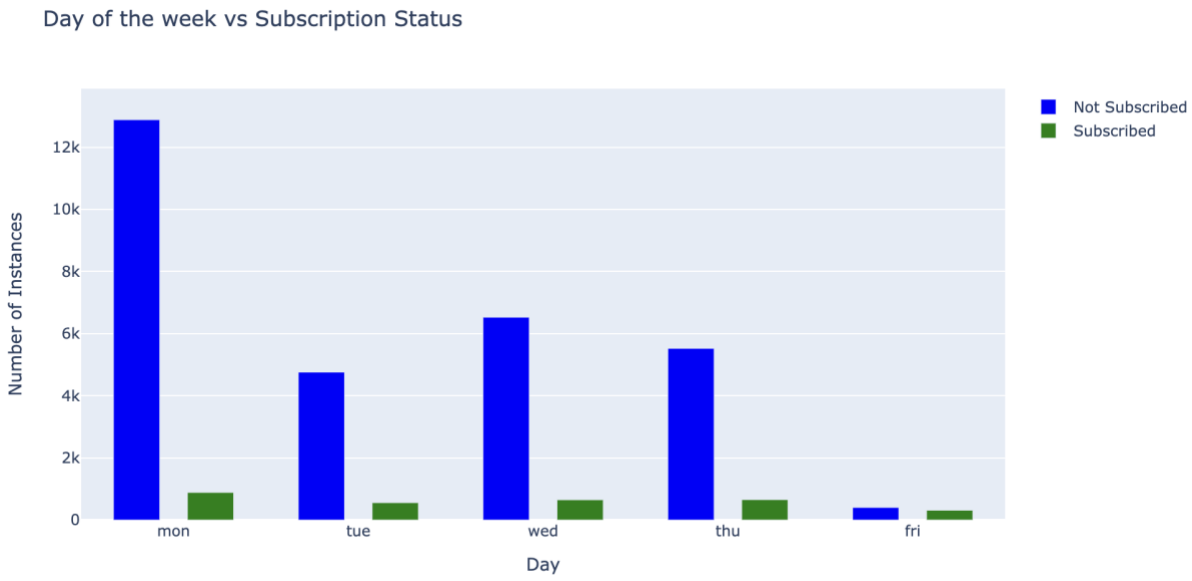
2) The number of defaulters is almost none compared to those unknown or with value yes. Those who have not defaulted, have subscribed more compared to those who have defaulted or to those who have no information.



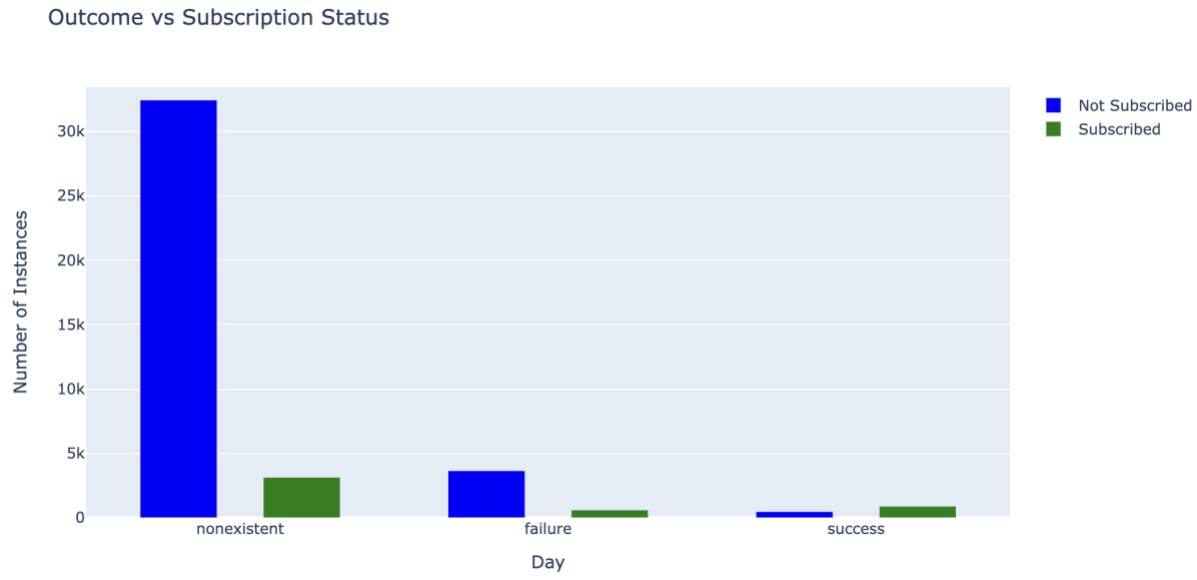
3) Those without any loans are subscribed more compared to those who have loans.



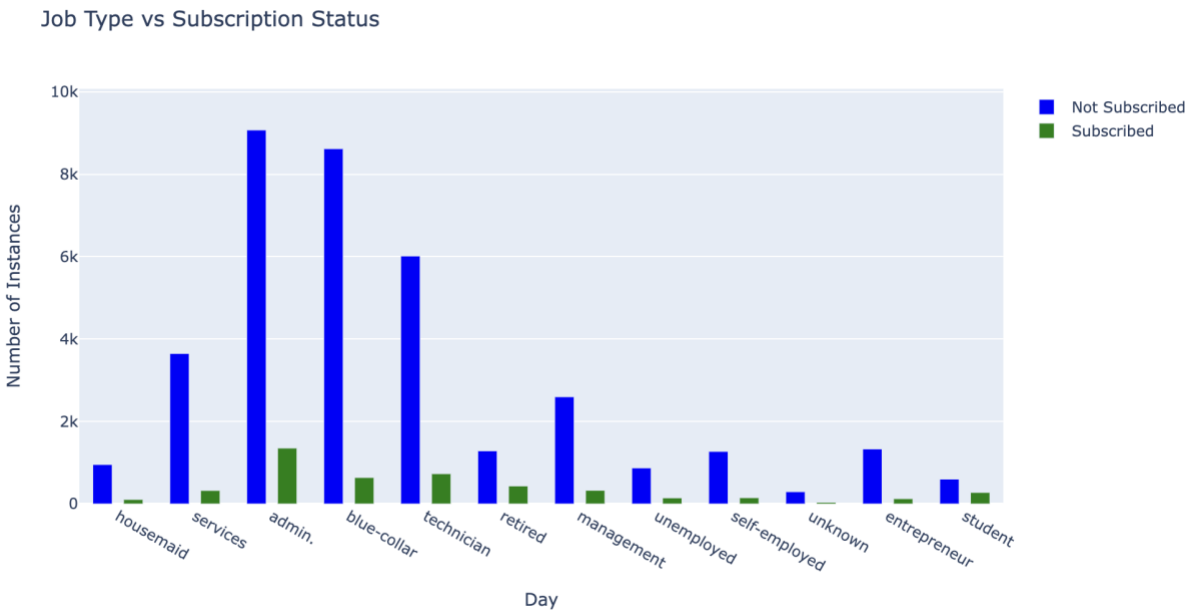
4) Monday is the day where most people subscribed and unsubscribed compared to the other days of the week.



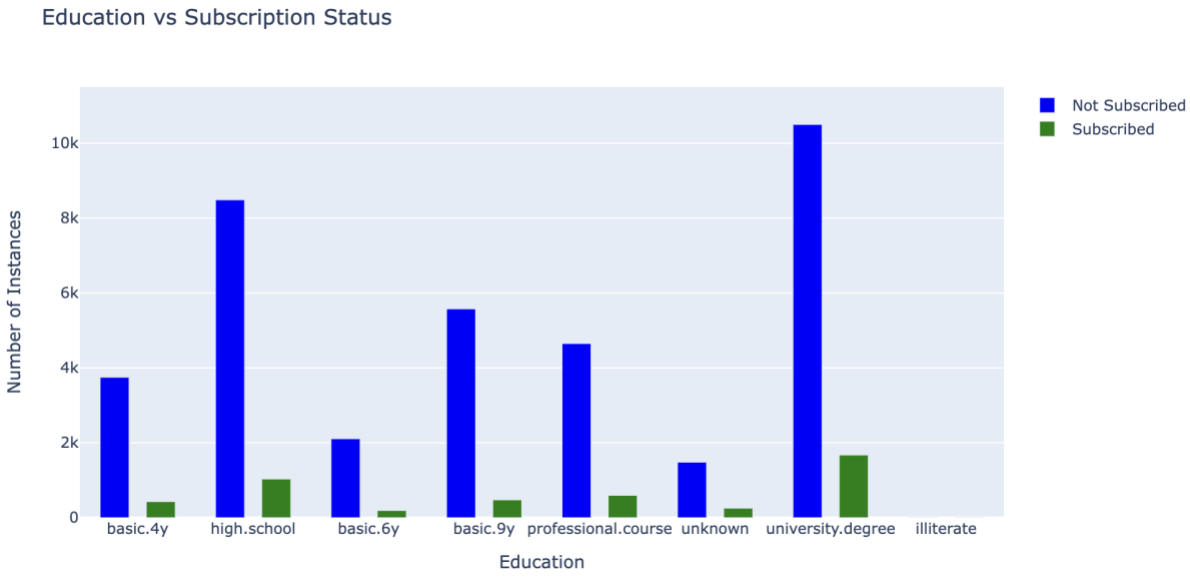
5) Those with non-existent outcomes have more subscriptions compared to those that are failure or success.



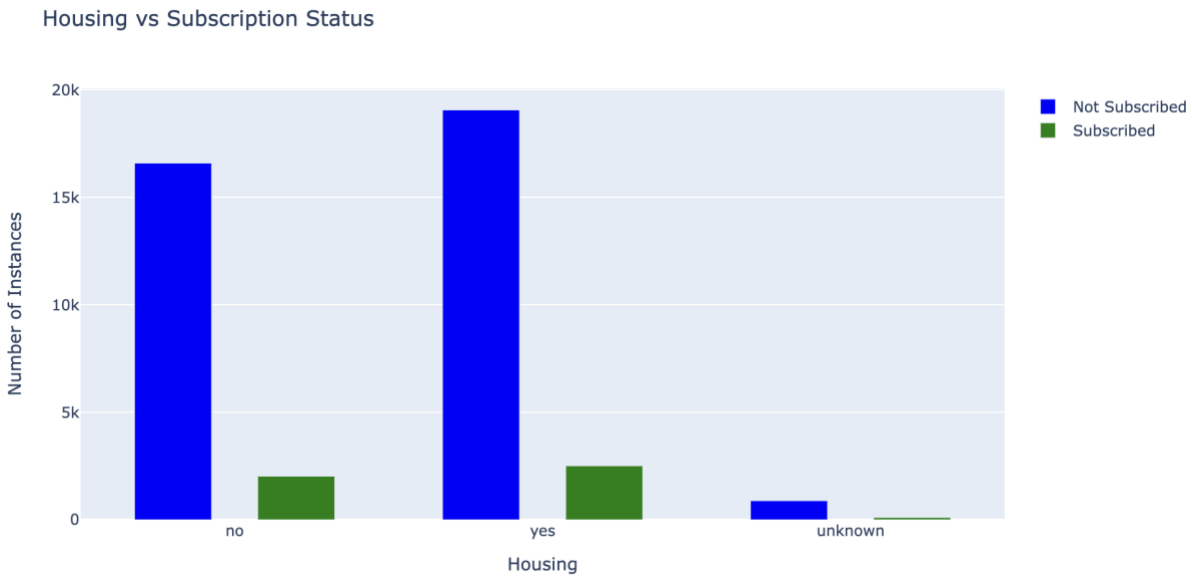
6) Those with admin, blue-collar and technician job types subscribe more compared to other job types.



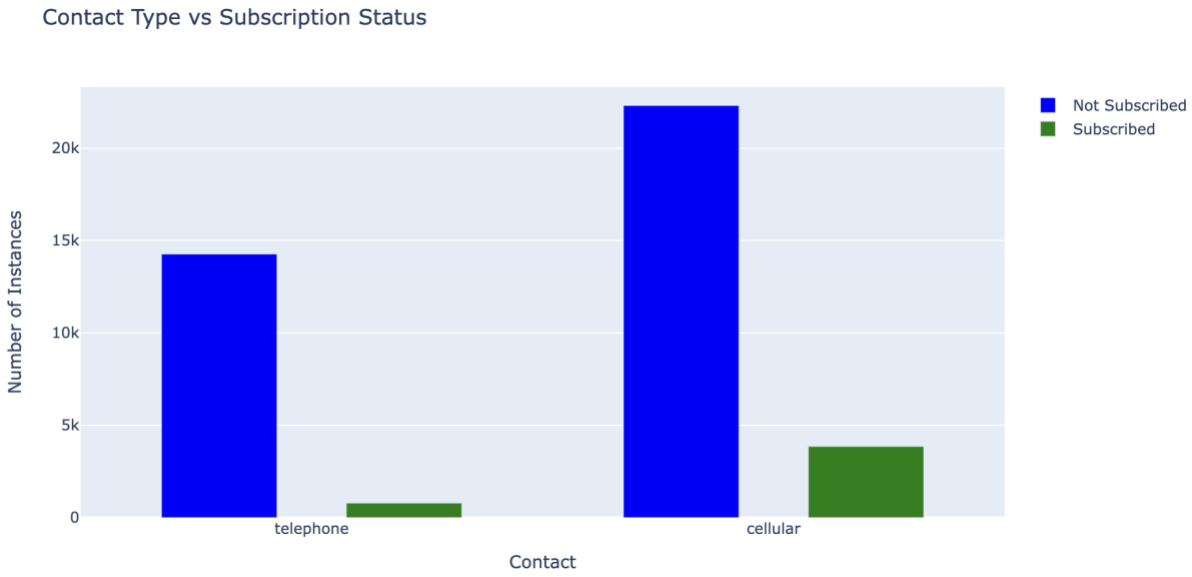
7) Contacts with high school and university degree subscribe more compared to other education levels.



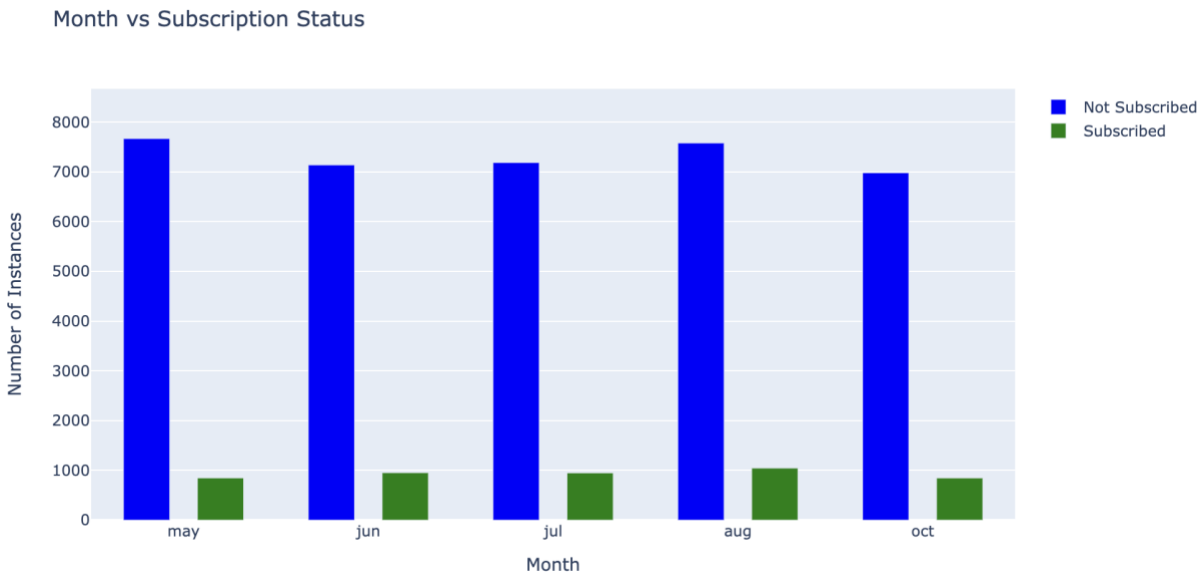
8) Not much difference between those with housing and those without housing. However, those with housing seem to have subscribed more compared to those without housing.



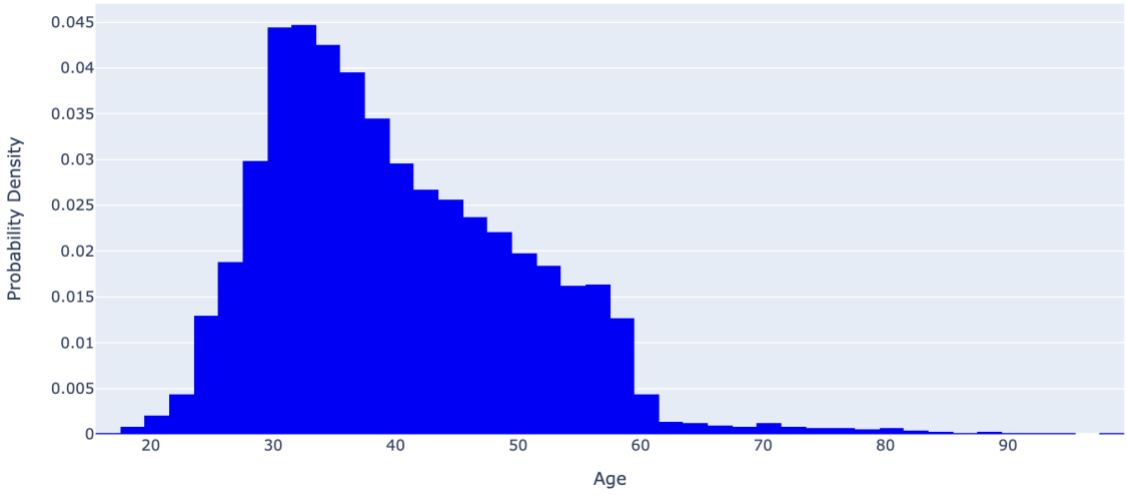
9) Those with cellular contact type subscribe more compared to those with telephone contact type.



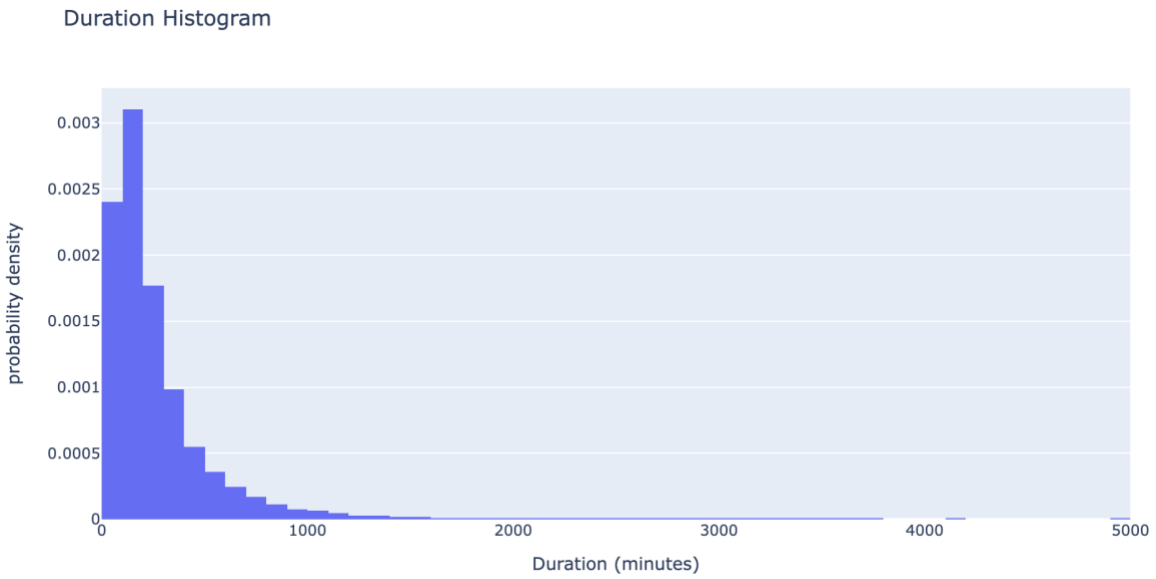
10) August month had a greater number of subscriptions compared to other months.

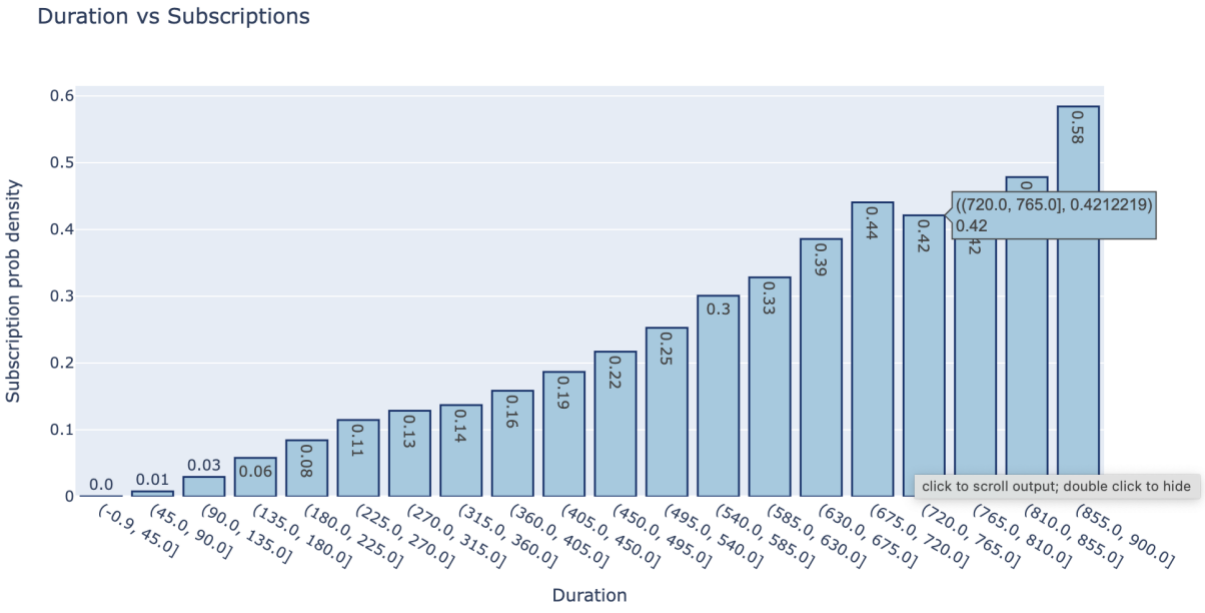


11) Majority of contacts are <= 60 years. Among those with age <= 60, the younger the clients are most likely to subscribe.

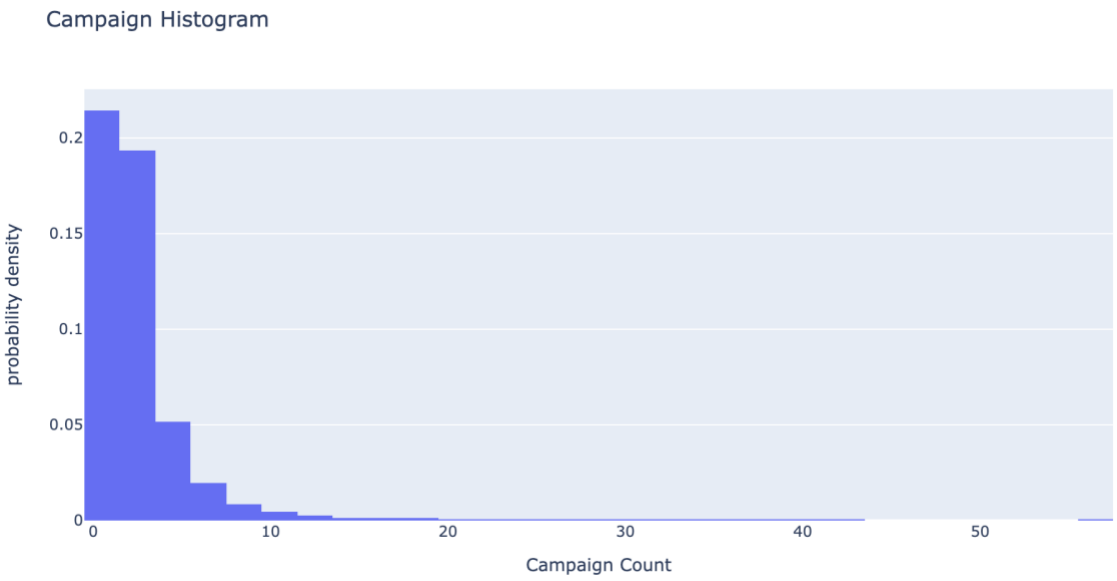


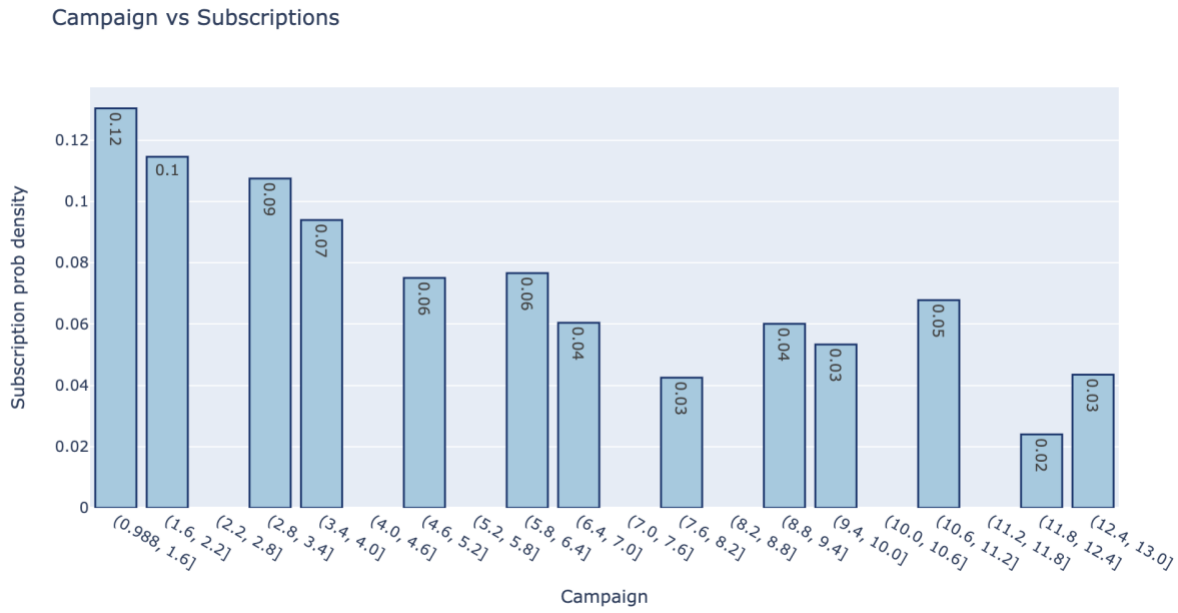
12) Clients with longer lasting contact duration are more likely to subscribe.





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