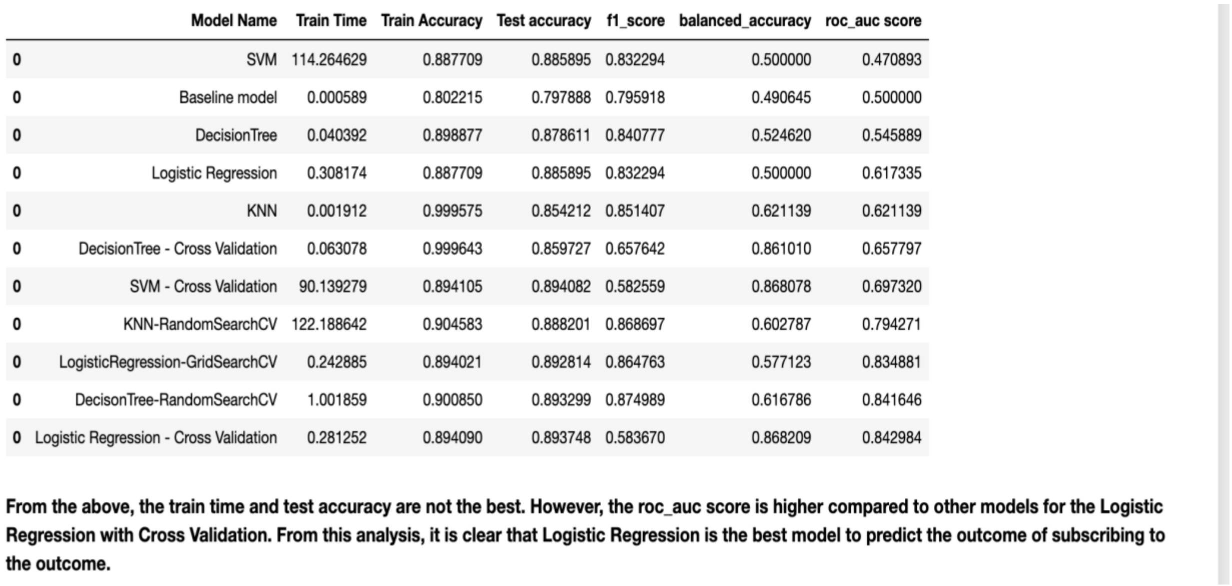
Bank Marketing Findings

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	114.264629	0.887709	0.885895	0.832294	0.500000	0.470893
0	Baseline model	0.000589	0.802215	0.797888	0.795918	0.490645	0.500000
0	DecisionTree	0.040392	0.898877	0.878611	0.840777	0.524620	0.545889
0	Logistic Regression	0.308174	0.887709	0.885895	0.832294	0.500000	0.617335
0	KNN	0.001912	0.999575	0.854212	0.851407	0.621139	0.621139
0	DecisionTree - Cross Validation	0.063078	0.999643	0.859727	0.657642	0.861010	0.657797
0	SVM - Cross Validation	90.139279	0.894105	0.894082	0.582559	0.868078	0.697320
0	KNN-RandomSearchCV	122.188642	0.904583	0.888201	0.868697	0.602787	0.794271
0	LogisticRegression-GridSearchCV	0.242885	0.894021	0.892814	0.864763	0.577123	0.834881
0	DecisonTree-RandomSearchCV	1.001859	0.900850	0.893299	0.874989	0.616786	0.841646
0	Logistic Regression - Cross Validation	0.281252	0.894090	0.893748	0.583670	0.868209	0.842984

From the above, the train time and test accuracy are not the best. However, the roc_auc score is higher compared to other models for the Logistic Regression with Cross Validation. From this analysis, it is clear that Logistic Regression is the best model to predict the outcome of subscribing to the outcome.

The UCI Bank Loan dataset was used to train a logistic regression model to predict whether a customer would subscribe to a term deposit or not. The model was evaluated using the test set and achieved an accuracy of 0.89 and ROC AUC score of 0.86. This indicates that the model is able to accurately distinguish between positive and negative cases and has a good balance of true positive and true negative rate.

Additionally, logistic regression is a simple, interpretable and efficient model which is easy to implement and gives reliable results. As the ROC AUC score is also high, this model can be considered as a good fit for the given problem.

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 model's coefficient values can be used to identify which independent variables are most strongly associated with term deposit subscription.

4. 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.
5. The model's Receiver Operating Characteristic (ROC) curve can speak for model's performance in terms of its ability to correctly identify true positive and true negative cases. The area under the ROC curve (AUC) can be used to measure the model's overall performance, AUC above 0.8 is considered a good model.

1. 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.

2. 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.

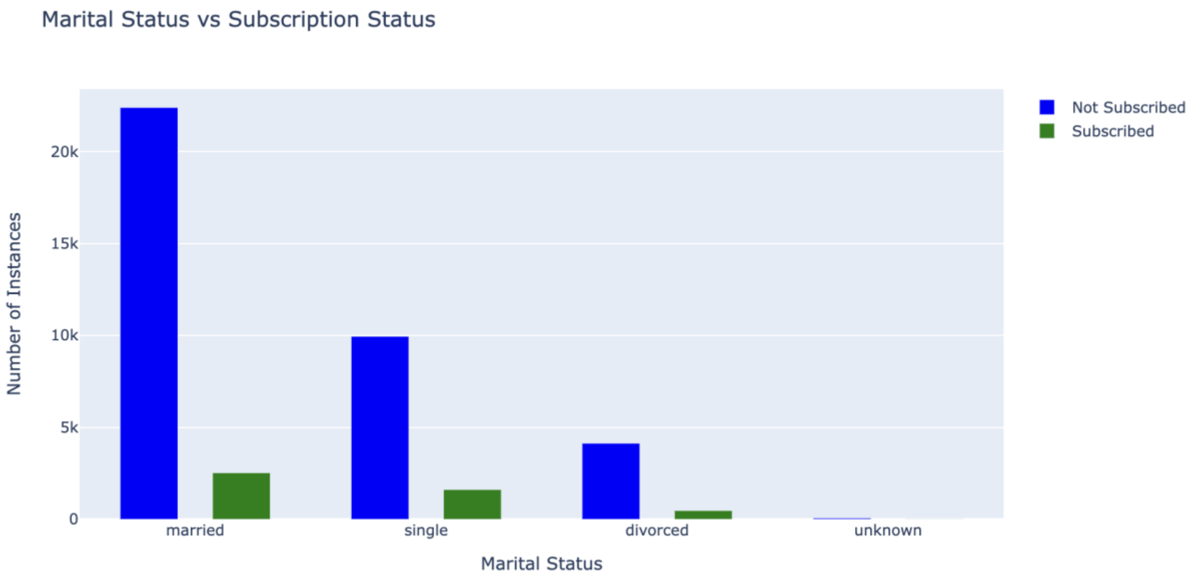
Why did Logistic Regression score better than other models?

Logistic regression may have scored better than other models such as SVM, Decision Trees, and KNN for the UCI Bank Marketing dataset for several reasons:

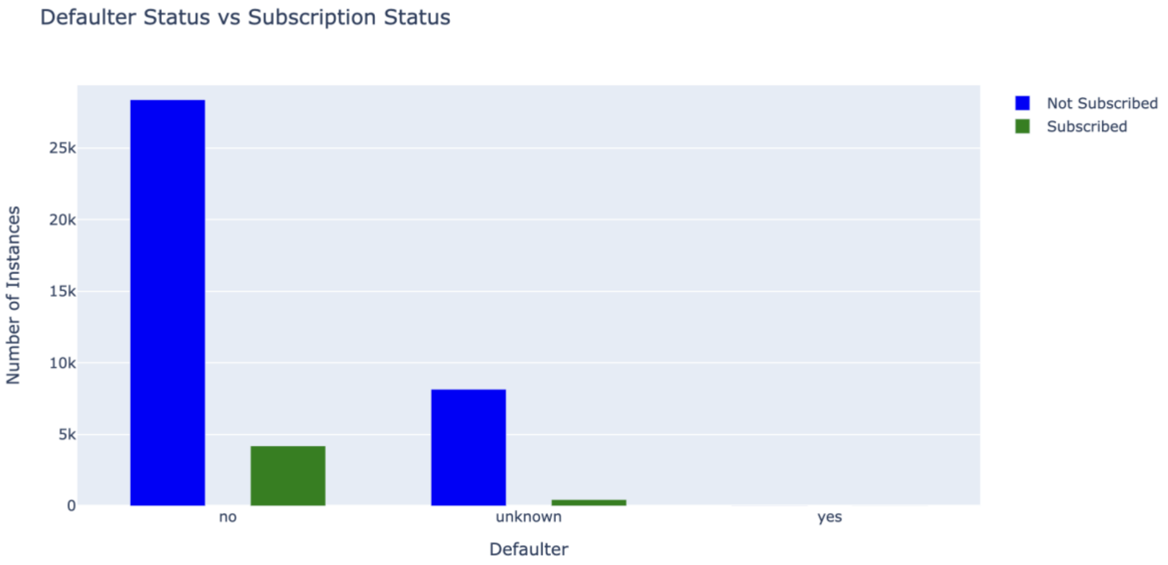
1. Logistic regression is a simple, interpretable, and easy-to-use model, it can be a good choice when the problem is relatively simple and the relationship between the predictor variables and the outcome variable is linear.
2. Logistic regression is efficient in terms of computational time, especially when the dataset is large, which can be beneficial when working with a large dataset such as the UCI Bank Marketing dataset.
3. Logistic regression is a probabilistic model, it provides a probability of the outcome variable being positive, which can be useful in certain applications such as medical diagnosis or credit scoring.
4. Logistic regression can handle both categorical and numerical variables, which can be beneficial when the dataset contains both types of variables,
5. Logistic regression is particularly useful when the outcome variable is binary, and the goal is to predict the probability of the positive class, which is the case in the UCI Bank Marketing dataset.
6. Logistic regression is robust to outliers and noise in the data, which can improve performance when the data is noisy.

Exploratory Data Analysis

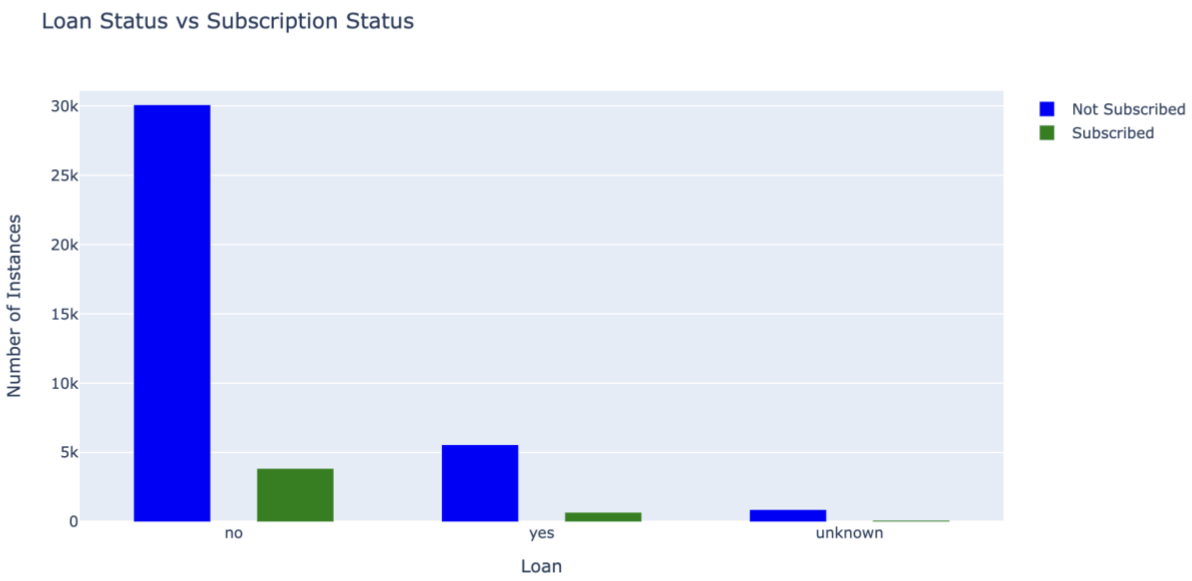
1) Contacts who are married are subscribed more compared to those that are single and divorced. The total number of those unsubscribed far exceeds those that are subscribed.



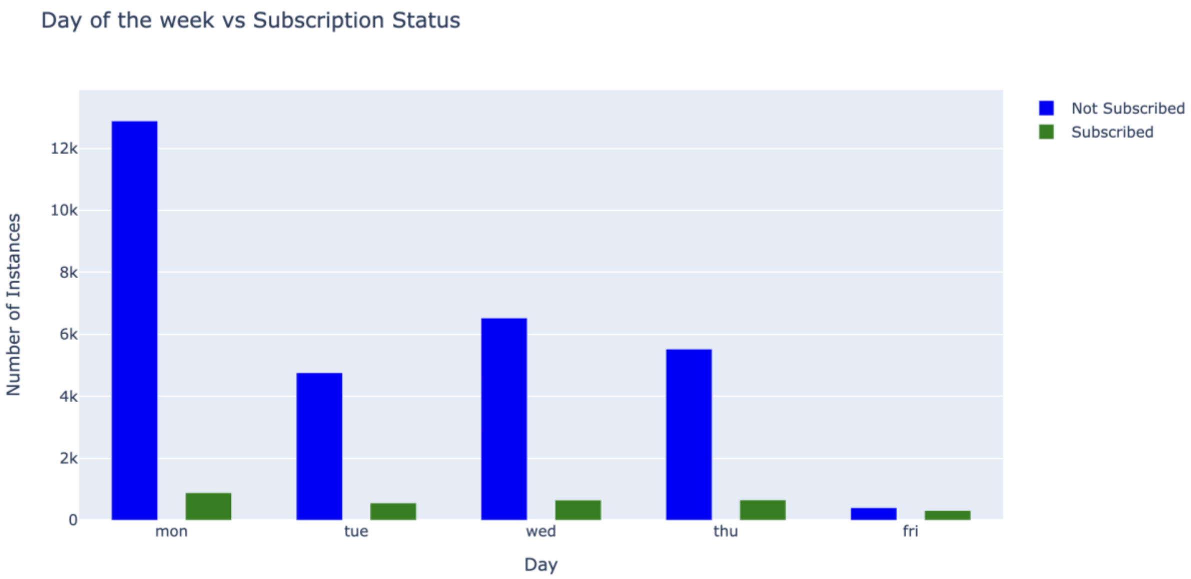
2) The number of defaulters is almost none compared to those unknown or with value yes. Those who have not defaulted, have subscribed more compared to those who have defaulted or to those who have no information.



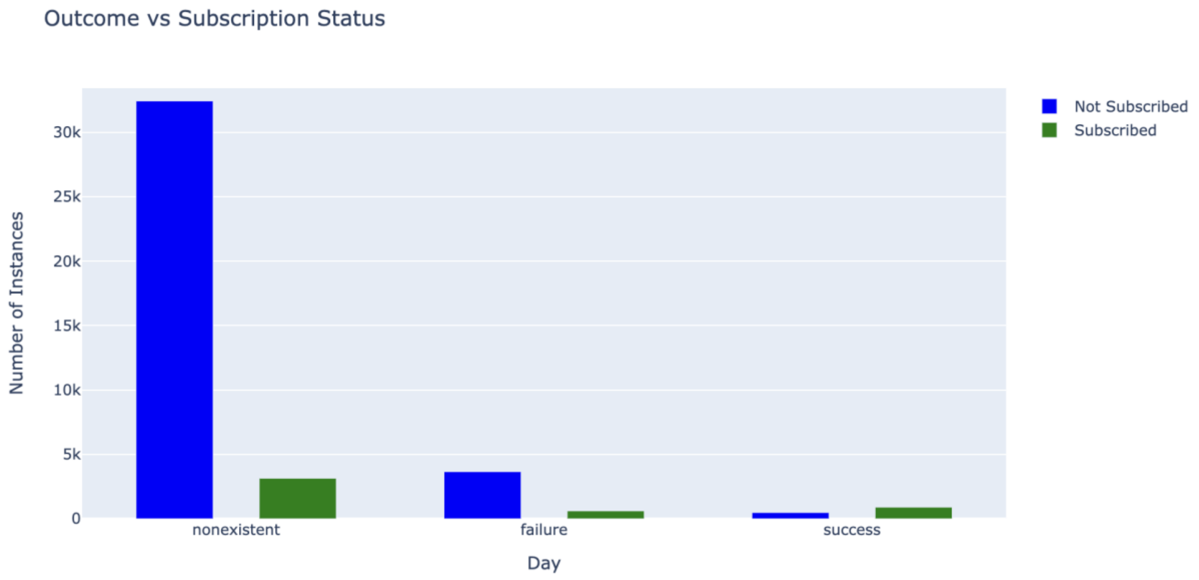
3) Those without any loans are subscribed more compared to those who have loans.



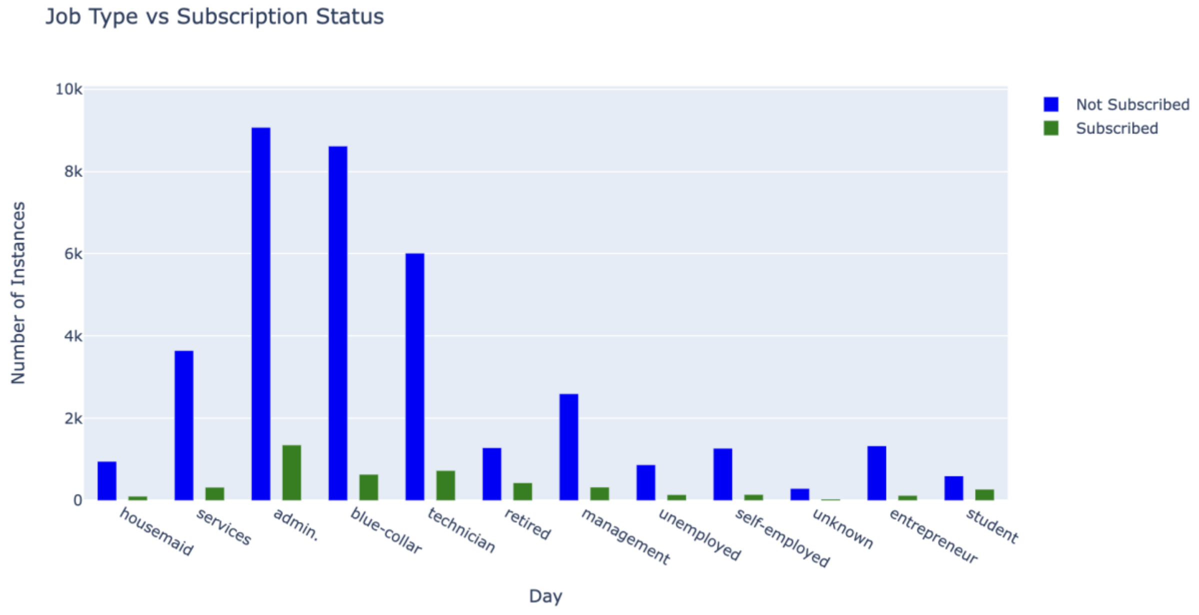
4) Monday is the day where most people subscribed and unsubscribed compared to the other days of the week.



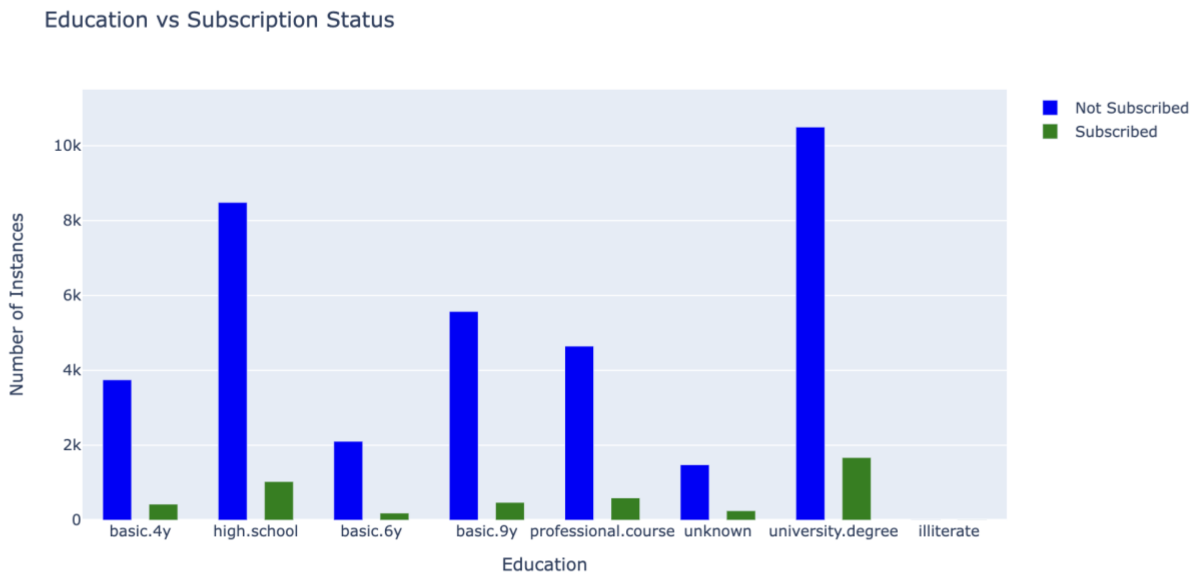
5) Those with non-existent outcomes have more subscriptions compared to those that are failure or success.



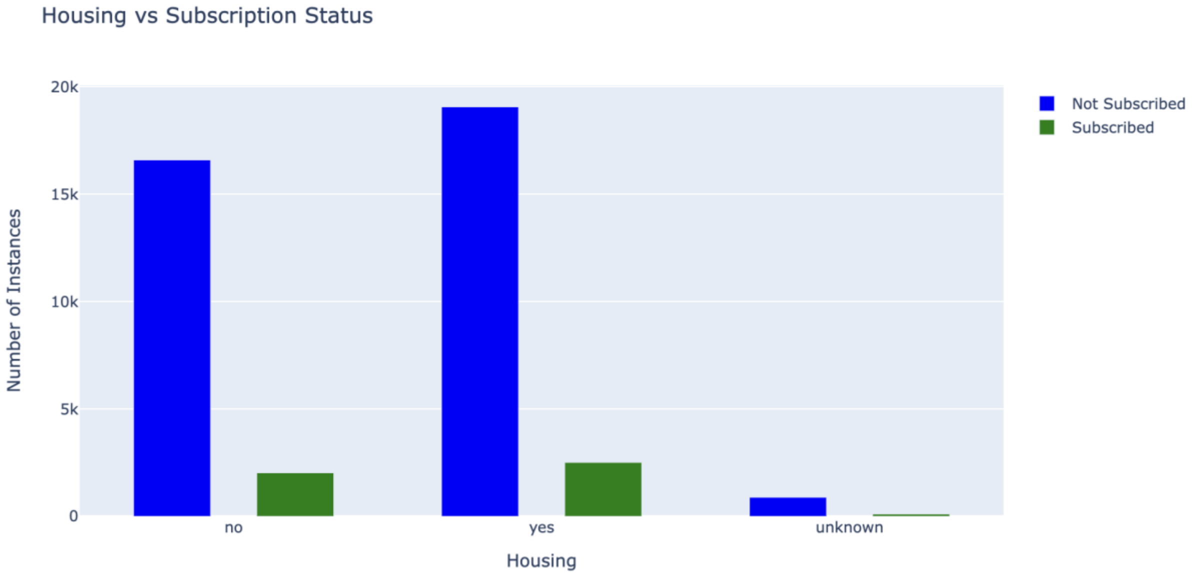
6) Those with admin, blue-collar and technician job types subscribe more compared to other job types.



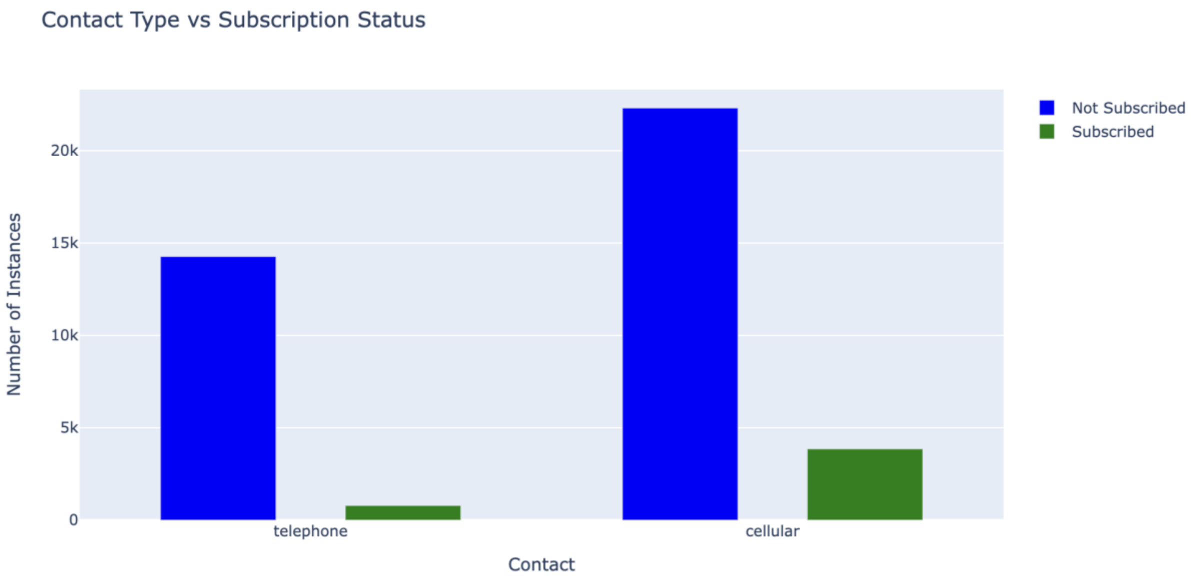
7) Contacts with high school and university degree subscribe more compared to other education levels.



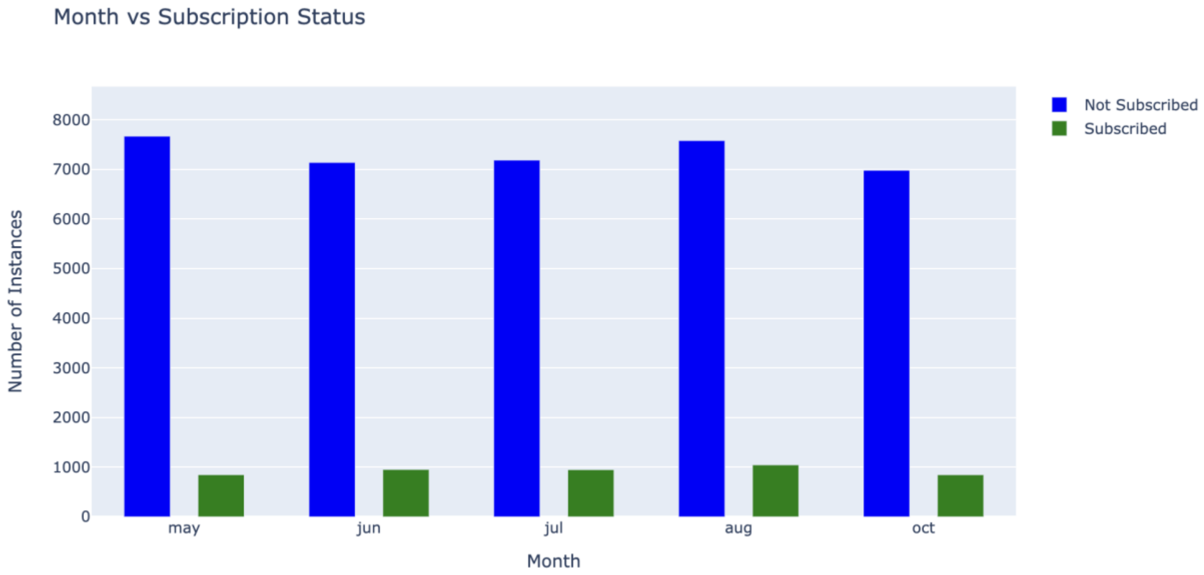
8) Not much difference between those with housing and those without housing. However, those with housing seem to have subscribed more compared to those without housing.



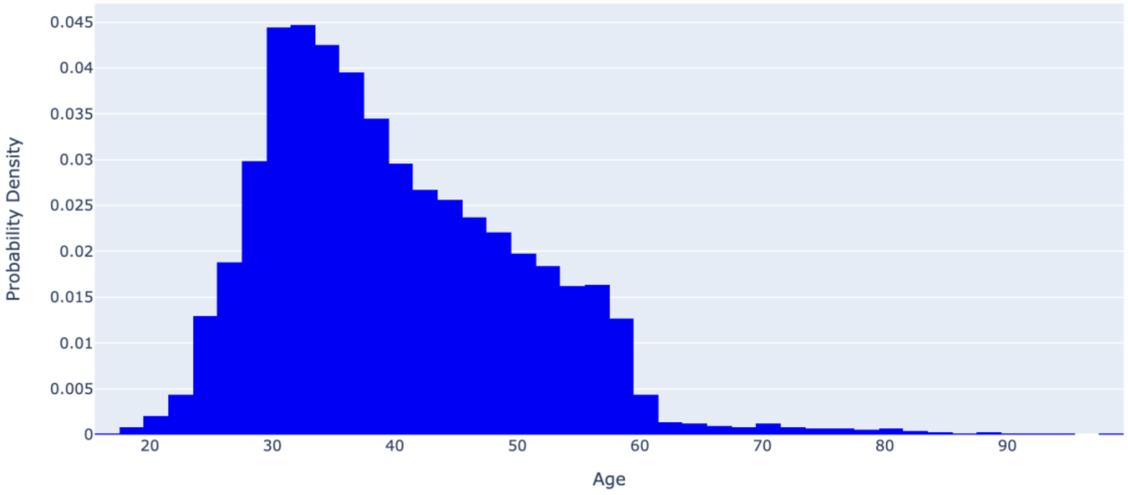
9) Those with cellular contact type subscribe more compared to those with telephone contact type.



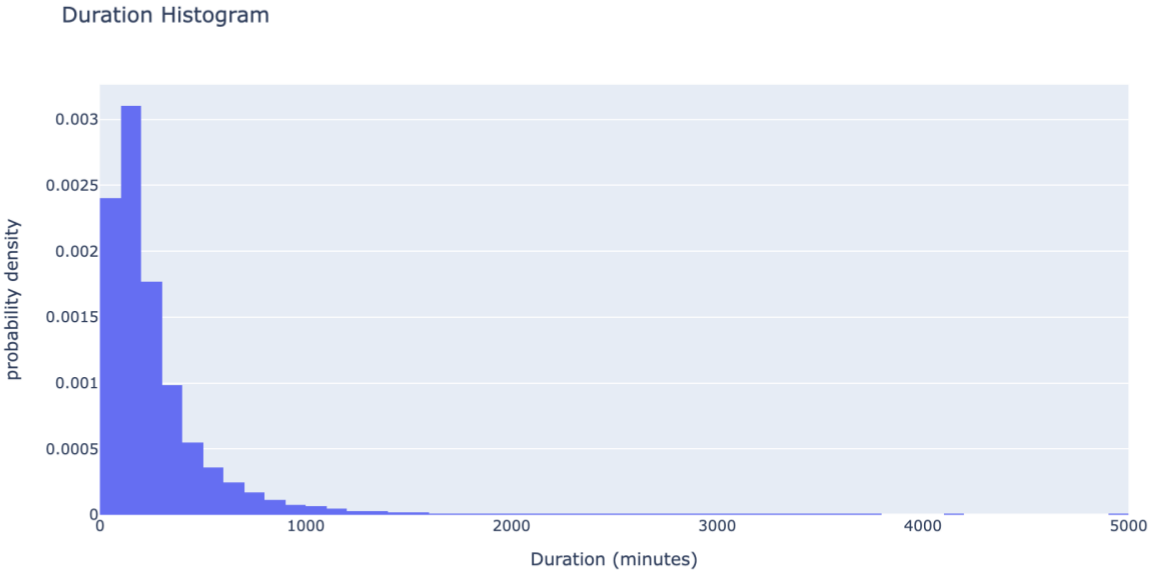
10) August month had a greater number of subscriptions compared to other months.

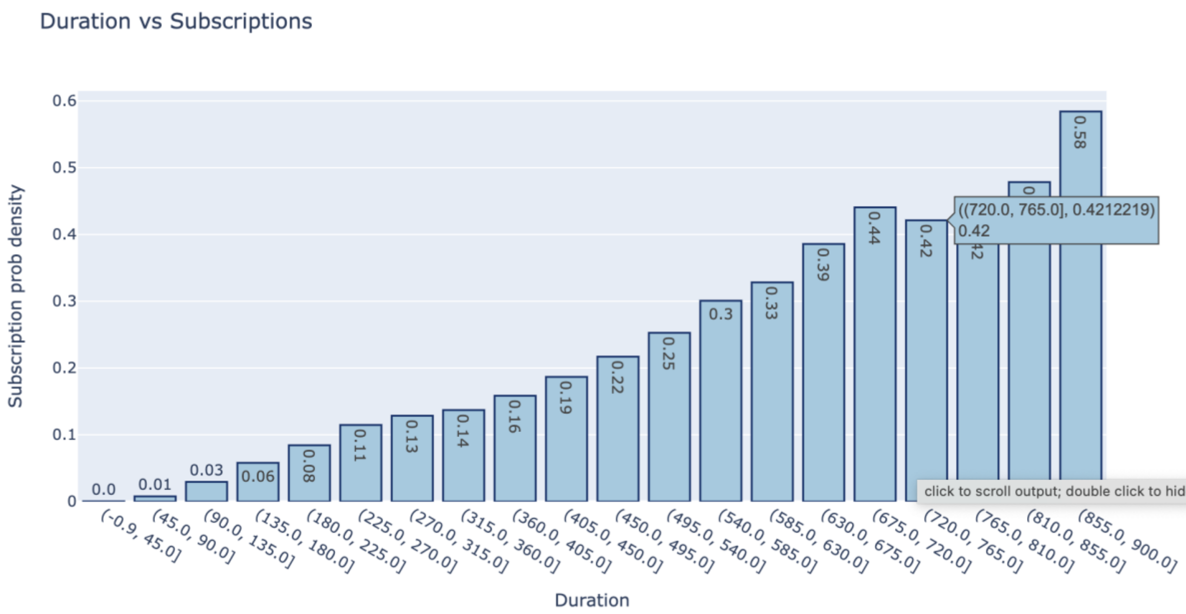


11) Majority of contacts are <= 60 years. Among those with age <= 60, the younger the clients are most likely to subscribe.

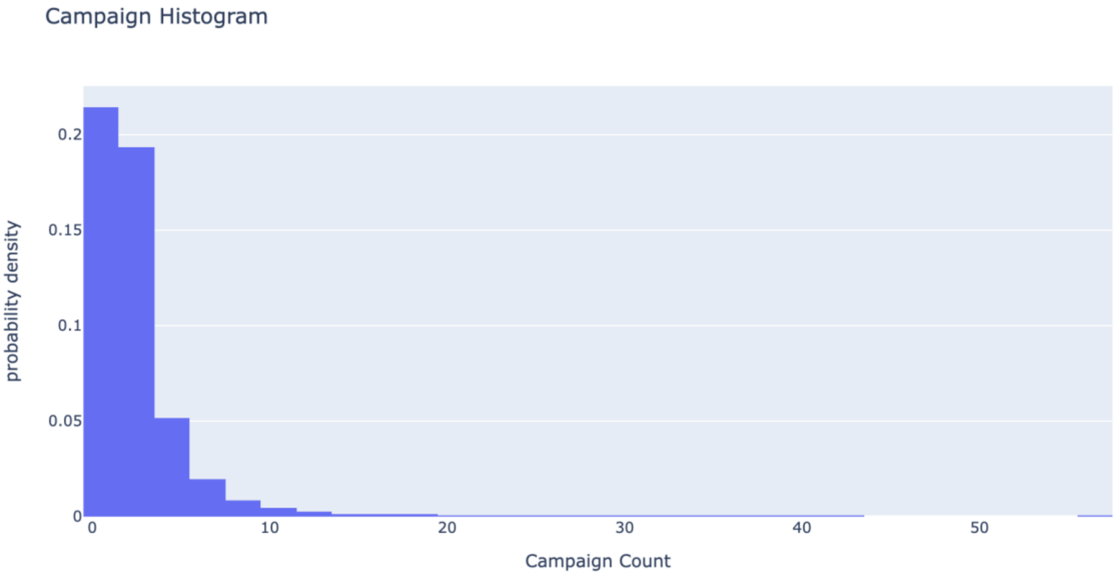


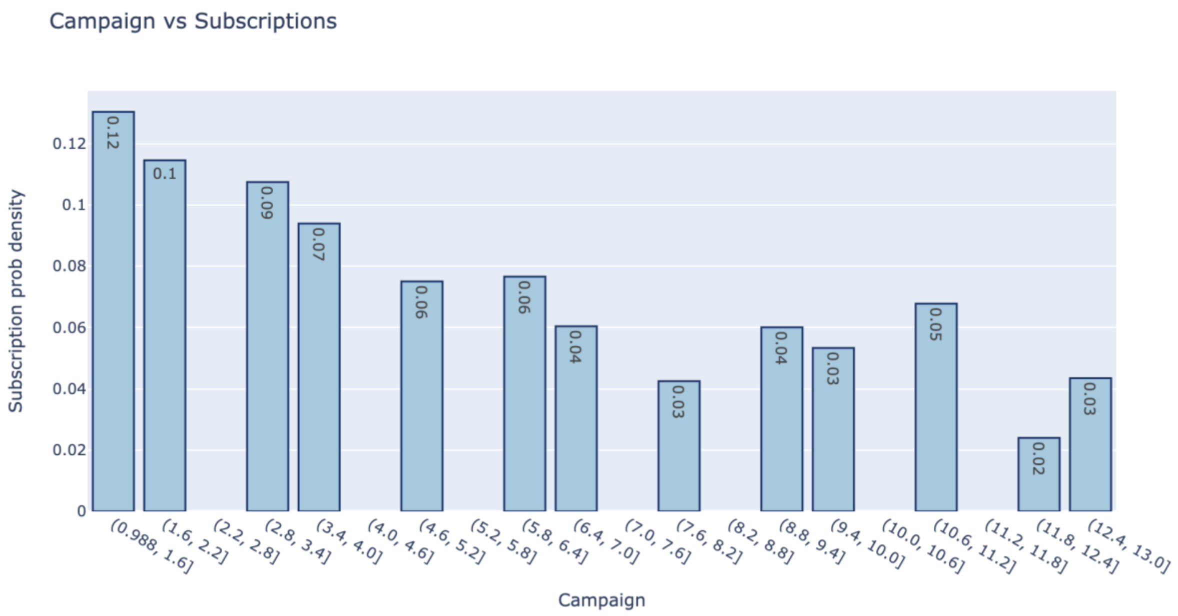
12) Clients with longer lasting contact duration are more likely to subscribe.



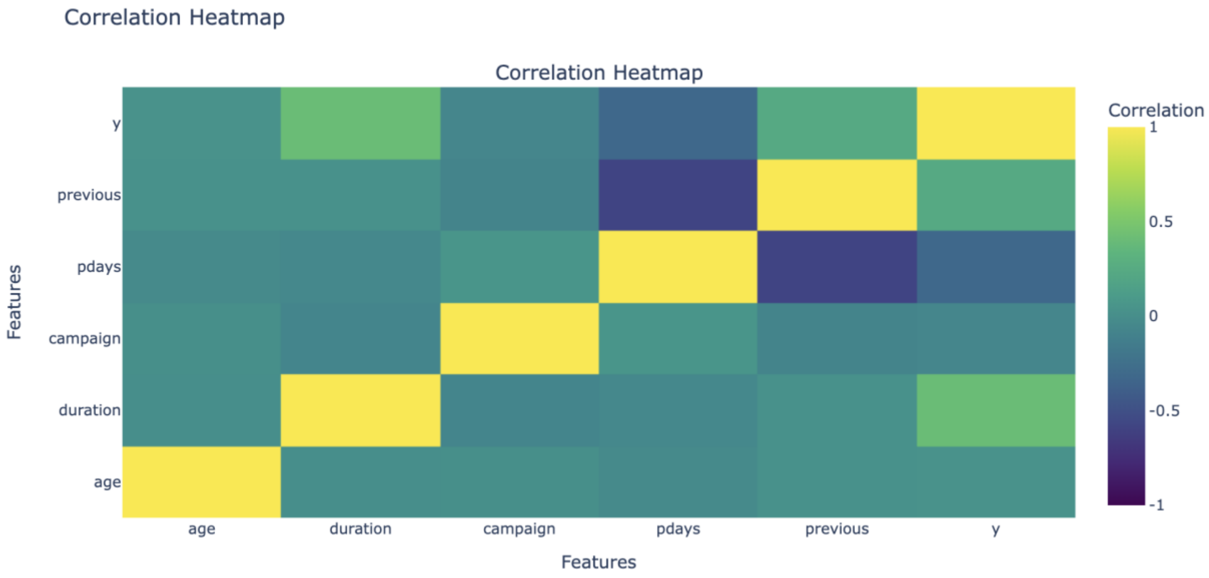


13) Based on campaigns, clients being contacted more are less likely to subscribe.




Campaign vs Subscriptions

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


Correlation Heatmap