Group Name: BRAVO

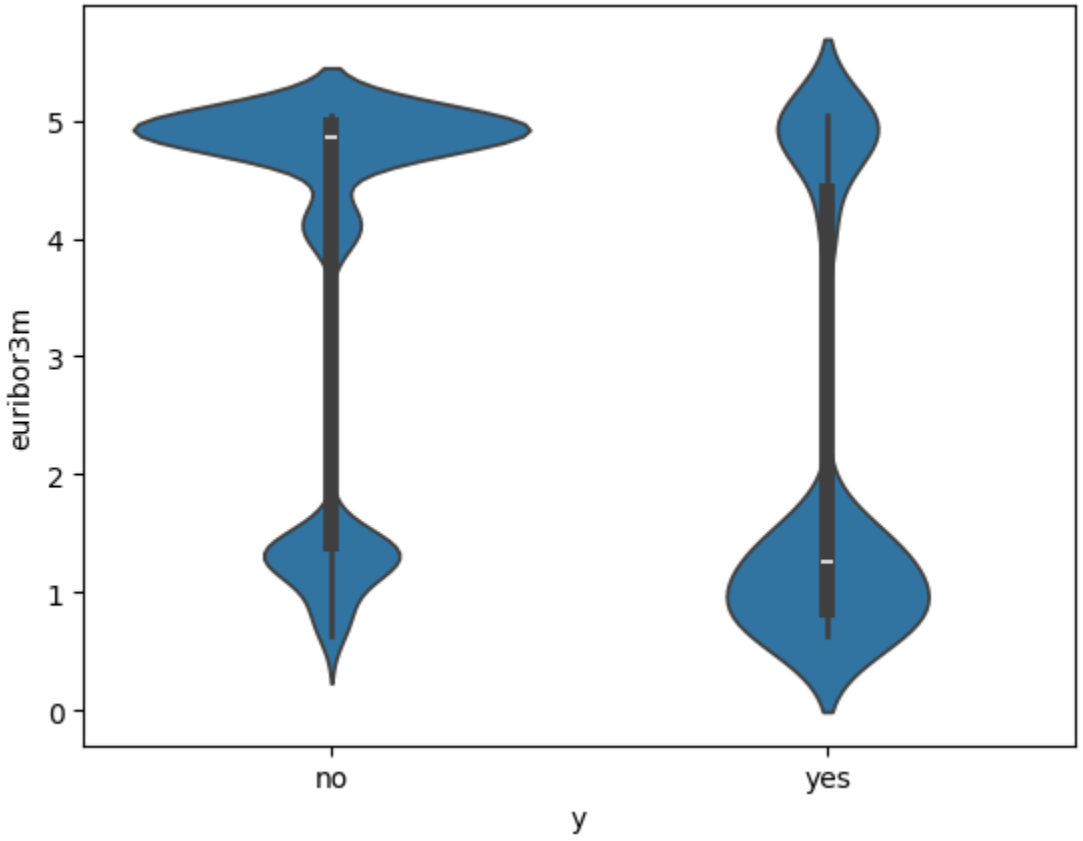
Name	Email (registered with Data Glacier)	Country	College/Company	Specialization
Jackson Taylor	jacksonian.r.taylor@gmail.com	United States	Santa Clara University	Data Science
Balamurugan Purushothaman	balamurugan2001viruda@gmail.com	United Kingdom	University of Liverpool	Data Science
Nazrin Thanikattil Rafeeque	101nazrin@gmail.com	United Kingdom	University of Hertfordshire	Data Science
Gunjan Varyani	gunjanvaryani916@gmail.com	United States	University of the Cumberlands	Data Science

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 helps 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).

The steps to solving this task include outlining the project, the initial data understanding and strategies to solve data problems, data cleansing and transformation, exploratory data analysis code, exploratory data analysis presentation and model recommendation, model selection and building, and presenting the final solution and code.

1. euribor3m vs Term Deposit Subscription(y) (violin plot):



Lower values of euribor3m (Euro Interbank Offered Rate for a three-month maturity with a daily indicator) are associated with increased number of subscribers.

The customer base seems like they have gravitated to valuing the benefits of buying term deposits when the euribor3m is low. There can be benefits to this like safety, stability, and capital preservation.

However, low euribor3m may not be the true cause of increased subscribers.

Although it may look like low euribor3m is contributing to subscriptions, it is possible that euribor3m just happened to be low when there have been successful campaigns and the success of the campaign itself is independent from the low euribor3m. The gradual change of euribor3m over time makes this a possibility. It can remain similar for a long time over the course of many campaigns.

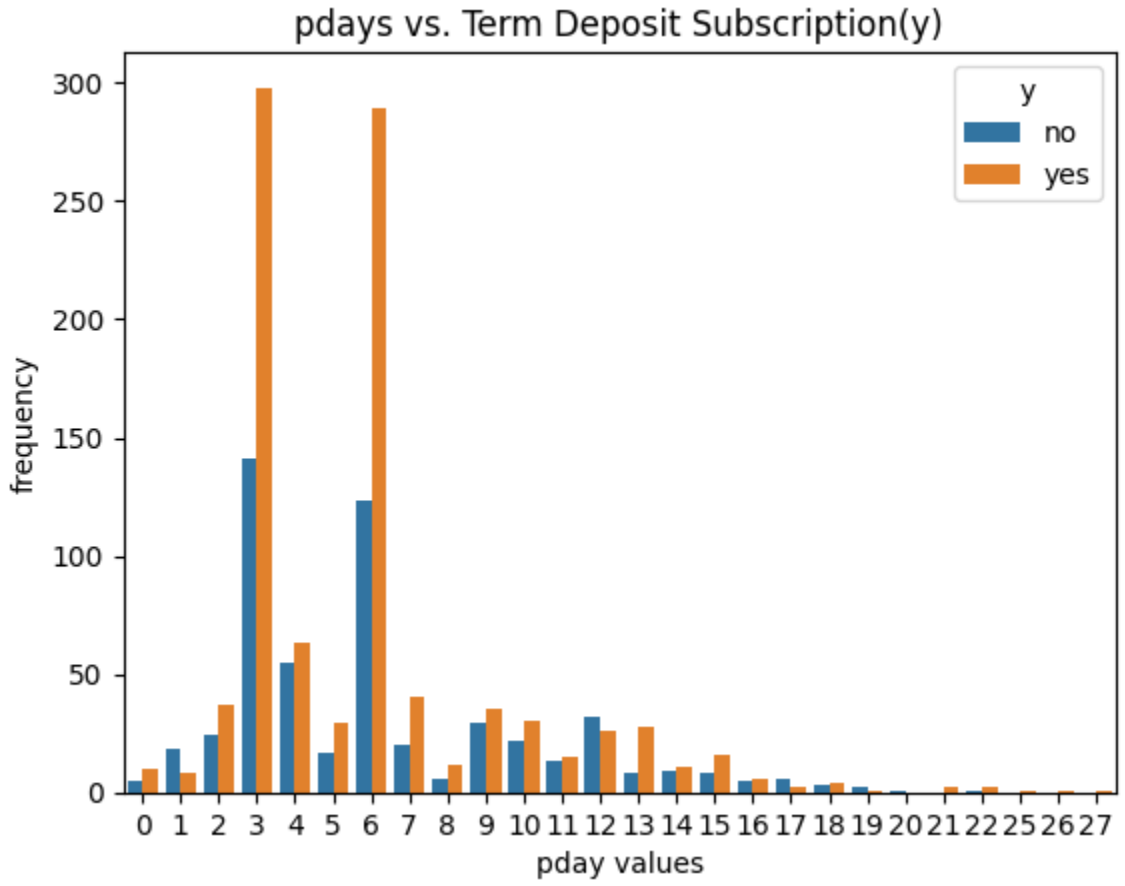
We recommend proceeding with an experiment. When there is a drop in euribor3m from one period to the next does the client base react with a justifiably increase in subscriptions or does it remain the same?

To see the trend, this experiment must take place on many different occasions, so the reaction is less attributed to time variables.

Since we don't have times of subscriptions available in the dataset, we cannot test this, but it should be simple to test if ABC bank has time windows for the data collected.

If it is found that lower euribor3m is the cause of increased subscriptions, then ABC bank should make sure they are prepared to launch more aggressive campaigns when they see a downward trend in euribor3m.

2. pdays vs. Term Deposit Subscription(y):



Observing when pdays is not equal to 999 reveals critical information about the sub-distribution.

The vast majority of clients were not contacted (999) in the previous campaign but this should not disregard the other data that is smaller in comparison.

At pday values 13, 6, and 3, these levels have the highest ratio of subscribers to all clients.

Note: this disregards the buyer proportions when there are only clients who bought for a certain level (ratio of subscribers to all clients = 1). It is assumed that anomalies occurred at these levels due to small sample variance.

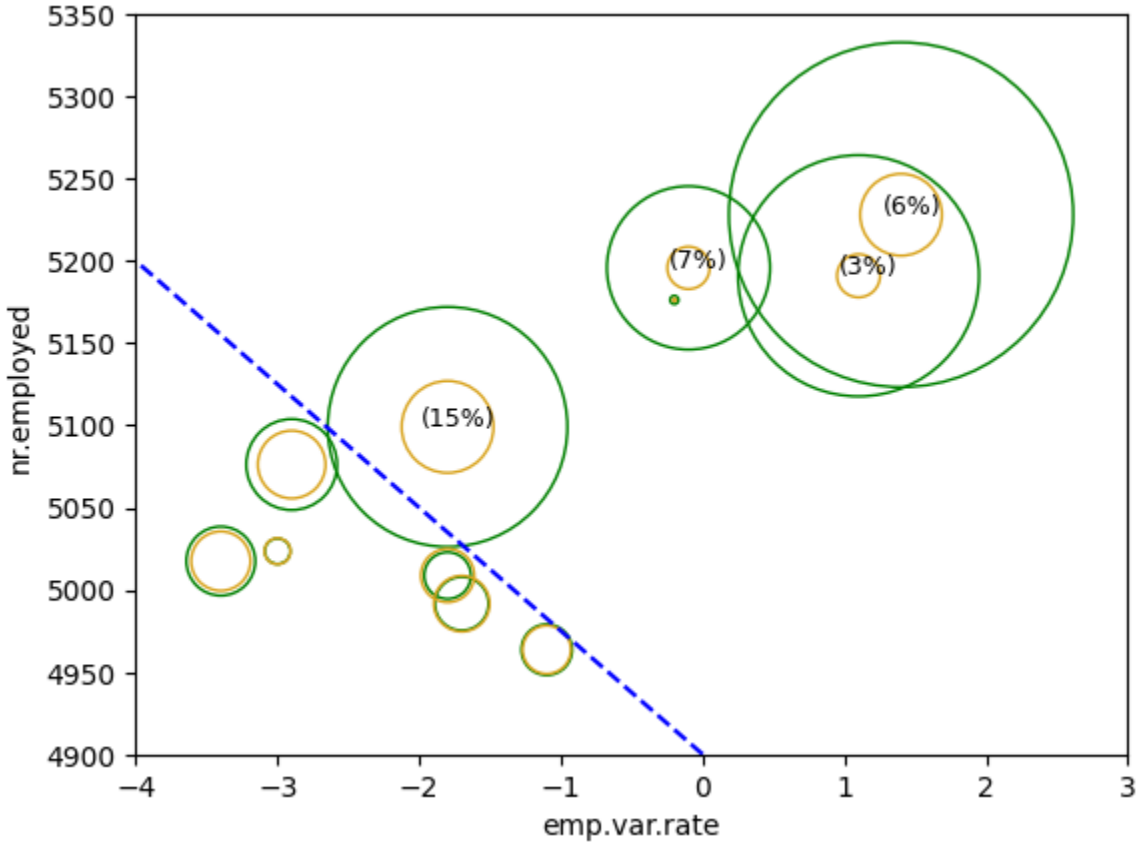
With this in mind, it seems likely that the high purchase percentage for pday value of 13 is also due to variances and small samples.

However, days 3 and 6 have substantially higher client frequencies than the rest which eliminates the potential for variance to be the factors in their purchasing success.

Since they are high, it seems likely that ABC bank has chosen those levels to recontact and determine if the client was interested or not. They may have even planned a follow-up meeting with the customer from their initial contact. Part of the success of these levels is probably due to the customers' interest in a follow up contact and ABC often choosing days that are 3 and 6 days apart from the initial contact.

The recommendation to ABC bank is to continue focusing on those levels (3 and 6) to contact clients. With additional data it should also be confirmed if the high purchasing success in pday = 13 is an anomaly or not. Likewise, the pdays with small samples should be re-sampled.

3. This section utilizes the scatter plot of emp.var.rate and nr.employed. Each point is represented by two bubbles.



There are two bubbles for each point representing subscribers and non-subscribers. The size of the green bubbles represents the number of clients who did not buy the product. The size of the gold bubbles represents the number of clients who bought the product.

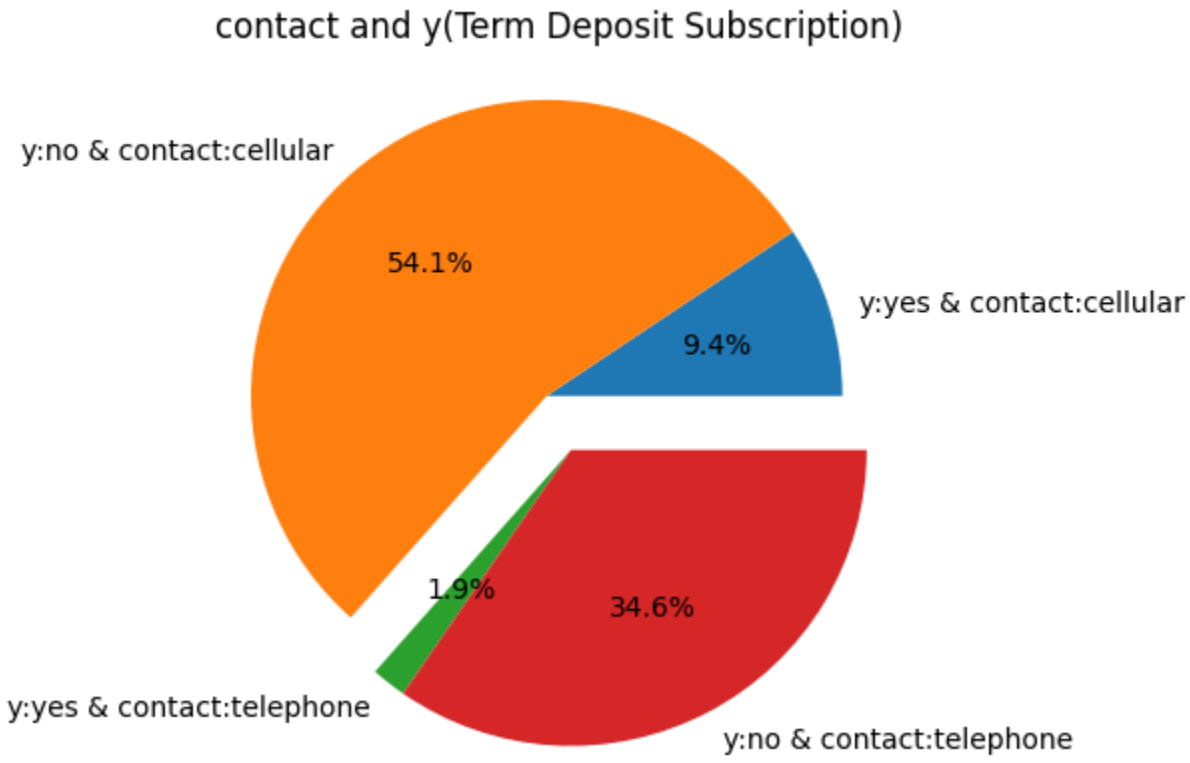
Some bubbles have percentages which indicate the percentage of how much the yellow bubble is of the green bubble: (clients who bought/clients who didn't)*100

It can be observed that there are not many combinations of emp.var.rate and nr.employed but each group has a large representation. Each point represents a point in time. Similar to the reason in the first section, the difference in percentages could be due to successes in the company's history independent from the combination of emp.var.rate and nr.employed. Considering this, take the following insight with light skepticism.

The graph displays a discernable boundary between combinations of emp.var.rate and nr.employed. On one side, all but one of the points have a low purchase percentage and a large number of observations. On the other side, the points have more equal percentage of subscribers to non-subscribers and less observations.

This generally means that it has been found when emp.var.rate and nr.employed are low, the purchase percentage is higher. It may be a good idea to maximize campaign efforts under these circumstances.

4. Pie chart of contact and y(Term Deposit Subscription):

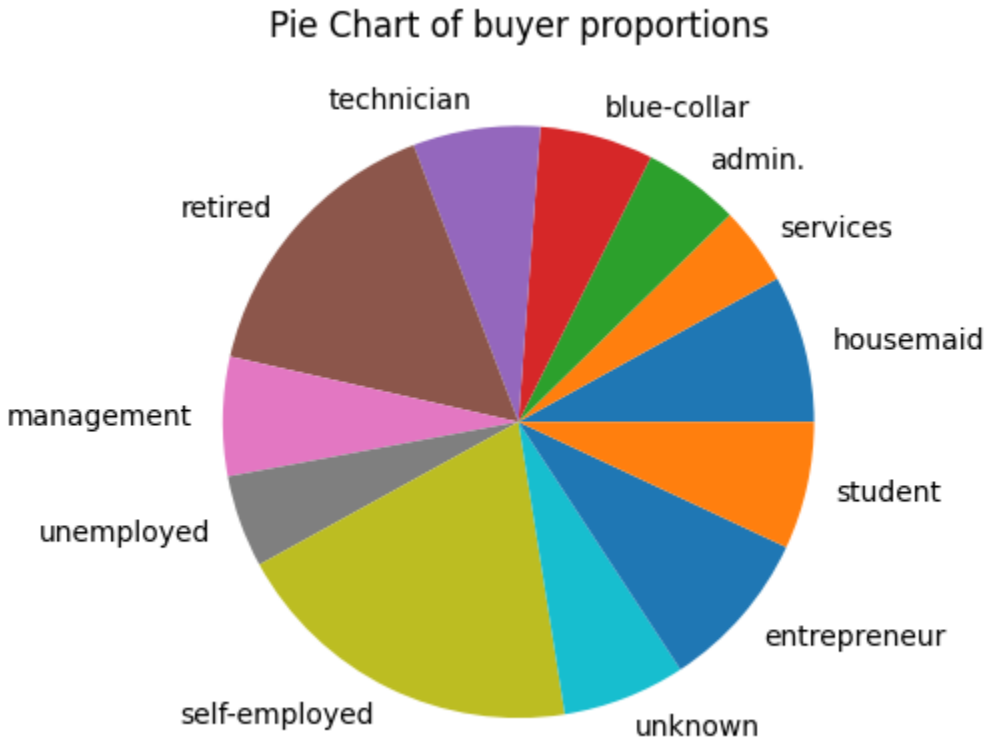


The pie chart shows that is clearly better to market to those with cellphones because they have a stronger rate of purchasing the product.

This could be the result of the accessibility that a phone offers. It makes it more convenient for clients to engage in banking services.

Targeting those with cellphones can be done through selective advertisement, especially if ABC bank has a mobile app that clients need to use to make transactions, check balance ect.

5. Pie Chart of buyer proportions:



The value of each slice of the pie is clients who subscribed in a job category divided by total clients in that job category.

Putting this to use in a business scenario would be trying to target the groups that represent substantial pieces of the pie.

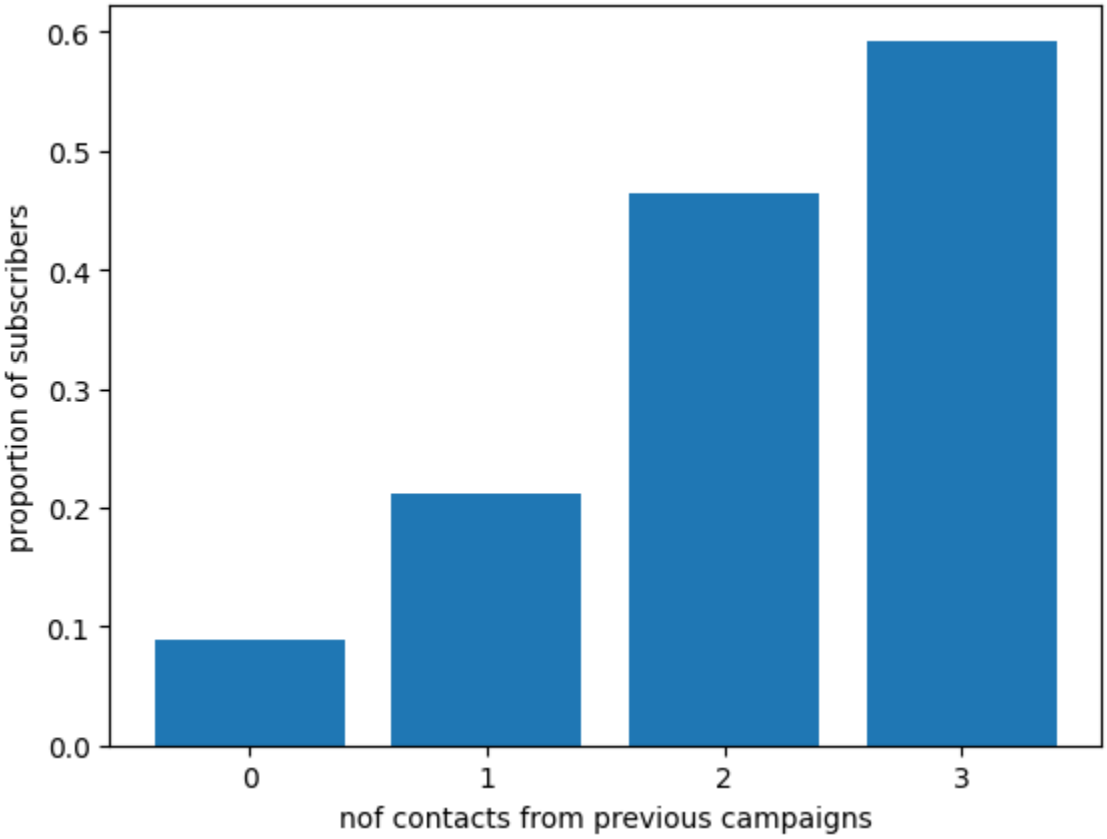
The standout sections appear to be clients who are retired and clients who are self-employed.

Marketing to certain jobs can be an easier task than many of the marketing tasks that could be construed from this dataset. Data is everywhere which could contain people's job title or heavily allude to it.

Once a population with jobs that are part of the key titles has been identified, it is time to target them. Targeting advertising can happen on google, job sites, and social media. There are also things like email marketing and propositions.

It can be relatively straightforward to advertise to certain job groups, if you have the right strategies and tools in place.

6. Proportions of subscribers vs previous(nof contacts from previous campaigns):



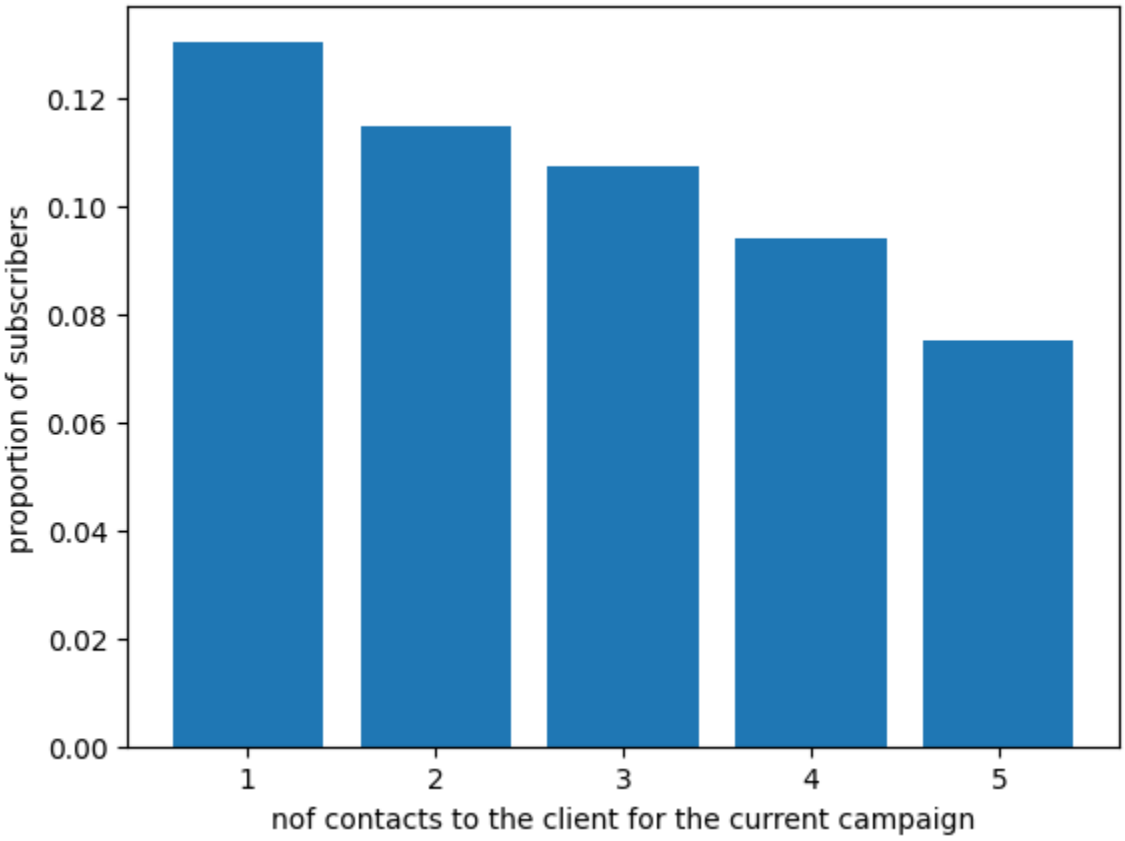
The more contacts performed in previous campaigns for this client the more likely they will buy it again.

If the client was not interested, they would not make the contact happen multiple time over again in previous campaigns.

Only (0-3) are shown because there are anomalies at the far end due to variance and small sample space.

ABC bank should make sure to focus on clients who have been contacted the most times during previous campaigns. If the data is in hand, this is a trivial matter.

7. Proportions of subscribers vs campaign(nof contacts to the client for the current campaign):



The first contact leads to the most success compared to any other contact level.

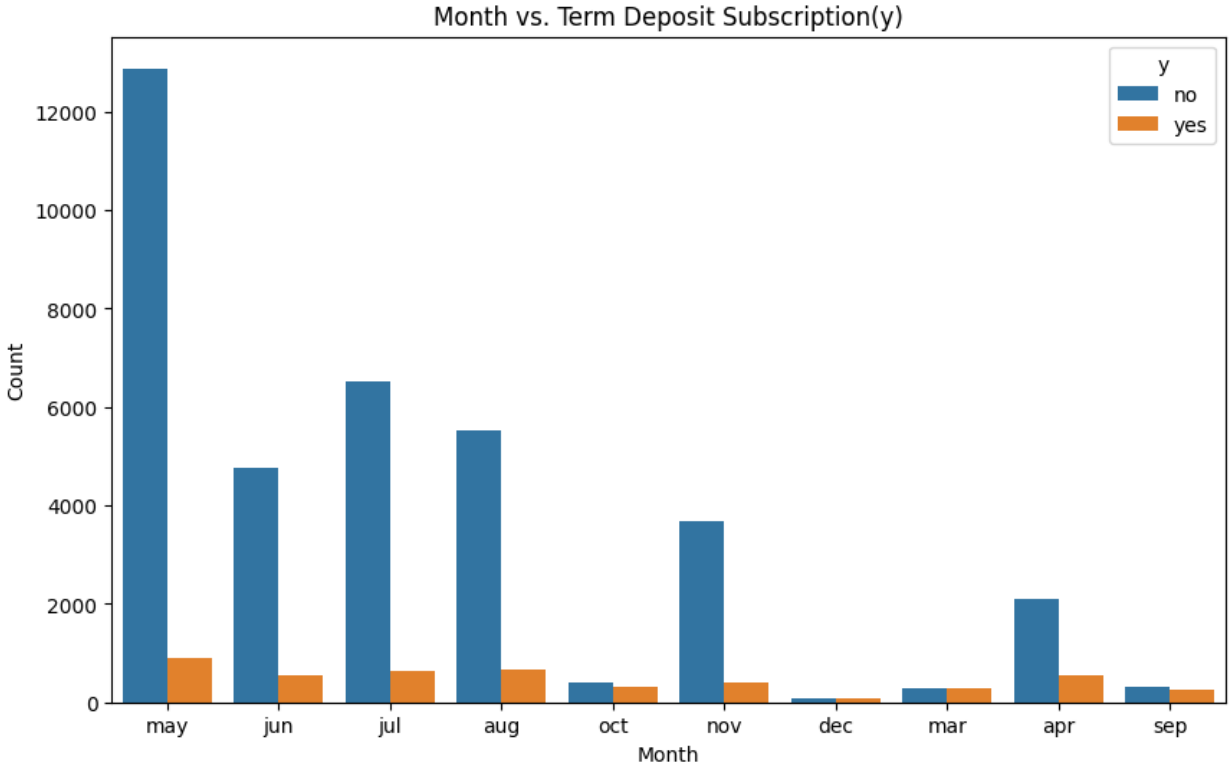
The second contact will often have less success because under many conditions if the client didn't purchase on the first call than there is less chance of them buying the next call. On the other hand, it could be an ongoing conversation that took more contacts.

We recommend to focus on contacting those who have not been contacted for the current campaign. However, if there is an ongoing conversation that takes more than one contact, those should be prioritized as well.

The higher buy rate for the first contact indicates that a single contact to talk about and buy the product is normally how long it takes for the client to make a decision.

Calling again and again still leads to more customers buying the product, but the priority should be customers that are mid conversation or have not been contacted before.

8. Month vs. Term Deposit Subscription(y):



The observations show that the season of May has the most clients. Why does the company want to keep campaigning in May if the purchase rate is the lowest?

ABC bank could have some reason for doing dedicating campaigns in May.

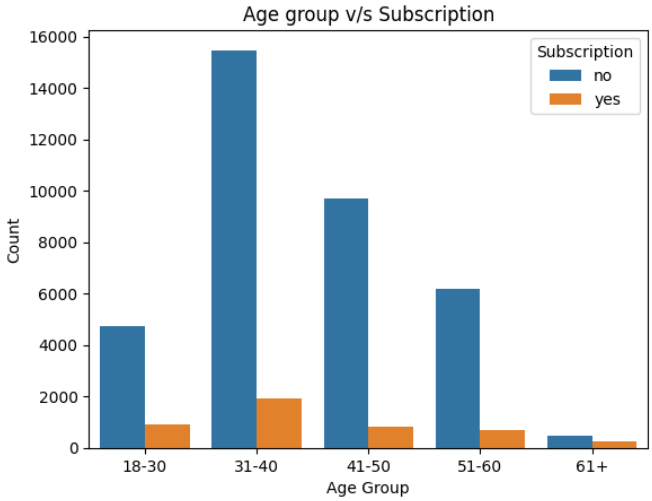
However, if ABC bank has overlooked this in analysis it may be reasonable to change its model.

At the very least, the following months of June, July, and August have better returns per customer and there are months that are better still.

Why not dedicate more campaigns to those months?

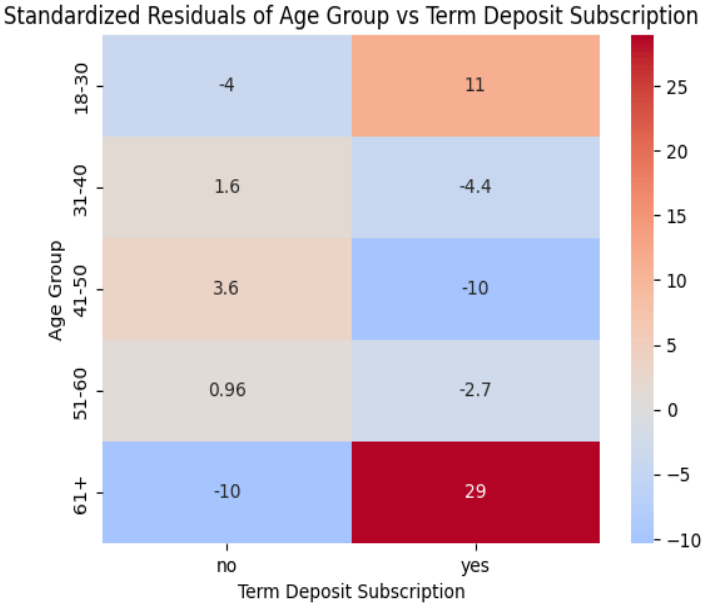
9. Age Range v/s Term deposit subscription:

To understand the likelihood of subscribing to term deposits, initially, the age attribute is converted to categorical values of 5 different ranges. On analysing the relationship between age range and subscription using a bar plot, it is evident that the age group 31- 40 have the highest number of term deposit subscribers whereas the age range of 61+ has the least number of subscriptions.



To check and validate the results from the above plot a Chi-square test has been performed assuming age group and term deposit subscriptions are independent(null hypothesis). The Chi test yielded a very high chi2 value of 664.315 implying a large difference between the observed and expected frequencies under the null hypothesis and a minimal p-value lesser than the significance level of 0.05 suggesting that the observed distribution of subscription status across age groups is unlikely to occur by chance.

The significance of each age range on term deposit subscription is better understood by calculating the standard residual from expected values derived from the Chi test.



From the heatmap for standard residual, it is evident that for age range 18 -30 and 61+ have positive standard residual.

Despite having the most subscribers, the 31-40 age group's standardized residuals demonstrate that their subscription rate is roughly in line with the entire statistics. This shows that there is no significant variation from the predicted behaviour for this age group.

The 18-30 and 61+ age categories have seen a higher-than-expected number of subscribers. This shows that these groups are more likely to subscribe to term deposits than expected.

In contrast, the 41-50 age group has received fewer subscriptions than predicted, indicating a potential lack of interest or other impediments to subscribing within this cohort.

Recommendations:

- Retain current marketing strategies for the age range 31-40 since they are performing as expected.
- Consider improving financial awareness programs and providing attractive entry-level term deposit options to leverage early financial planning practices for the age range 18-30.
- Emphasize the safety and stability of term deposits, addressing the post-retirement investment needs of age category 61+. This could improve the subscriptions in ABC Bank.
- For the age group 41-50 conduct research to understand the specific barriers faced by this age group from term deposit purchases.

Model Recommendations:

Random Forest (bagging ensemble method):

- Can handle class imbalance by adjusting class weights or using balanced class sampling. The target variable for our dataset (y) is highly imbalanced.
- Provides feature importance. In our EDA, there are a handful of features that seem to be less important. This was found with the tests done of many categorical variables. The random forest can allow us to remove features post build based on feature importance. If a feature is just producing noise, then the model can be rebuilt with those features removed.
- Less prone to overfitting and noise due to the averaging of multiple trees. There is a possibility of overfitting when a certain number of iterations are run for other classification methods.
- Can handle nonlinear data without need for feature transformation. This is critical because there are only a handful of linear relationships in our data. Random forests can accommodate that.
- Can work well with datasets that have a mix of numerical and categorical features with less preprocessing like scaling or one hot encoding. Our dataset is mixed with numeric and categorical features.

SVM (Support Vector Machine) (discriminative classifier):

- Can work well with both linear and nonlinear relationships. This is good because there are only a handful of linear relationships in our data. SVM can use different kernel functions to make complex decision boundaries.
- Works well with high dimensional space. Data is high dimensional when there is high value to combining many features rather than just concerning the individual value of features. This is the case with our dataset.
- Robust to outliers. Our data has many outliers, and we can test with the removal or without the removal of those outliers with SVM. SVM is resilient to this because it has regularization parameters.
- Less overfitting. Our dataset seems to have a solid number of irrelevant or less relevant features that could produce a lot of noise with other models. SVM is resilient to this because it has regularization parameters.
- Most effective in small to medium sized datasets. Our dataset and number of features are considered medium.
- Well suited for binary classification tasks. That is our model type. This should improve performance.

GBM (Gradient Boosting Machine) (ensemble model):

- GBM is a good choice of algorithm to train our model with. Our dataset has significant class imbalance within the target variable. This can be addressed by adjusting learning rates, using class weights, and employing boosting algorithms like AdaBoost or XGBoost.
- GBM calculates feature significance scores, which aid in identifying essential predictors and reducing noise from less significant information. This could improve model accuracy.
- This algorithm works by building sequential trees, where each tree corrects errors of the previous ones. This enables it to capture complex nonlinear relationships present in our dataset.
- The parameter tuning (nof trees, learning rate, depth) help to deal with different datatypes and structures more effectively. Our dataset has a combination of both numerical and categorical values. Hence, it would be ideal.
- However, careful parameter tuning is required to achieve a computationally inexpensive model.

Logistic Regression (linear model)

- LR is a more clear and straight forward algorithm. The coefficients of this model that illustrate the impact of each feature on the probability of subscribing to term deposit, aid in understanding and transparency.
- LR is inherently designed for binary outcomes due to the logistic function's capability to model probabilities between two classes. This feature matches well with our requirements so it could be an appropriate choice.
- LR is computationally efficient, ideal for managing and processing the large number of records in the dataset, enabling quick iterations.