Data Science Intern at Data Glacier

Project: Bank Marketing (Campaign)

Week 10: Deliverables

- **Group Name:** LimitLess Team
- **Team member's details:**
  1. **Name:** Chitra S Chaudhari
     - **Email:** chitraksonawane@gmail.com
     - **Country:** USA
     - **Company:** Create & Learn LLC, GrandCircus Detroit
     - **Specialization:** Data Science
  2. **Name**: Nuri Öztürk
     - **Email:** ozturknuri8@gmail.com
     - **Country:** Turkey
     - **Specialization:** Data Science
- **Batch Code:** LISUM 17
- **Date:**
- **Submitted to** Data Glacier
- **Github Repo link:** https://github.com/ChitraChaudhari/Bank-Marketing-campaign-

Problem description

ABC Bank wants to be able to predict which clients are most likely to subscribe to a term deposit. In this way, the bank wants to save time and money by running the marketing campaign more effectively and successfully.

Data understanding

In this project, we have been given 4 datasets:

- Bank
- Bank-full
- Bank-additional
- Bank-additional-full

To gain a deeper understanding of the dataset, we will do in-depth exploratory data analysis. This will allow us to acquire a clear picture of the dataset, relation of features, and properties.

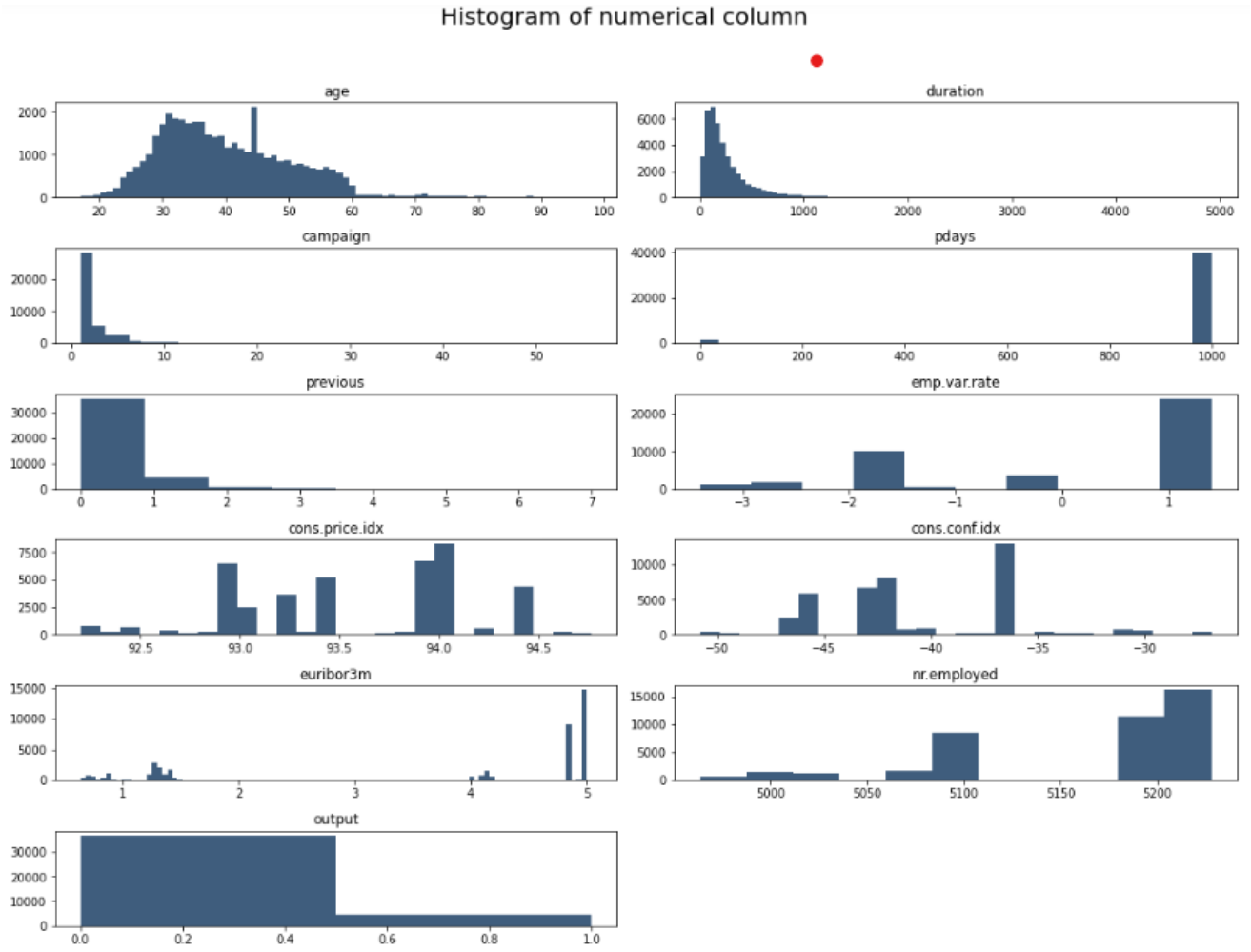
Attribute information:

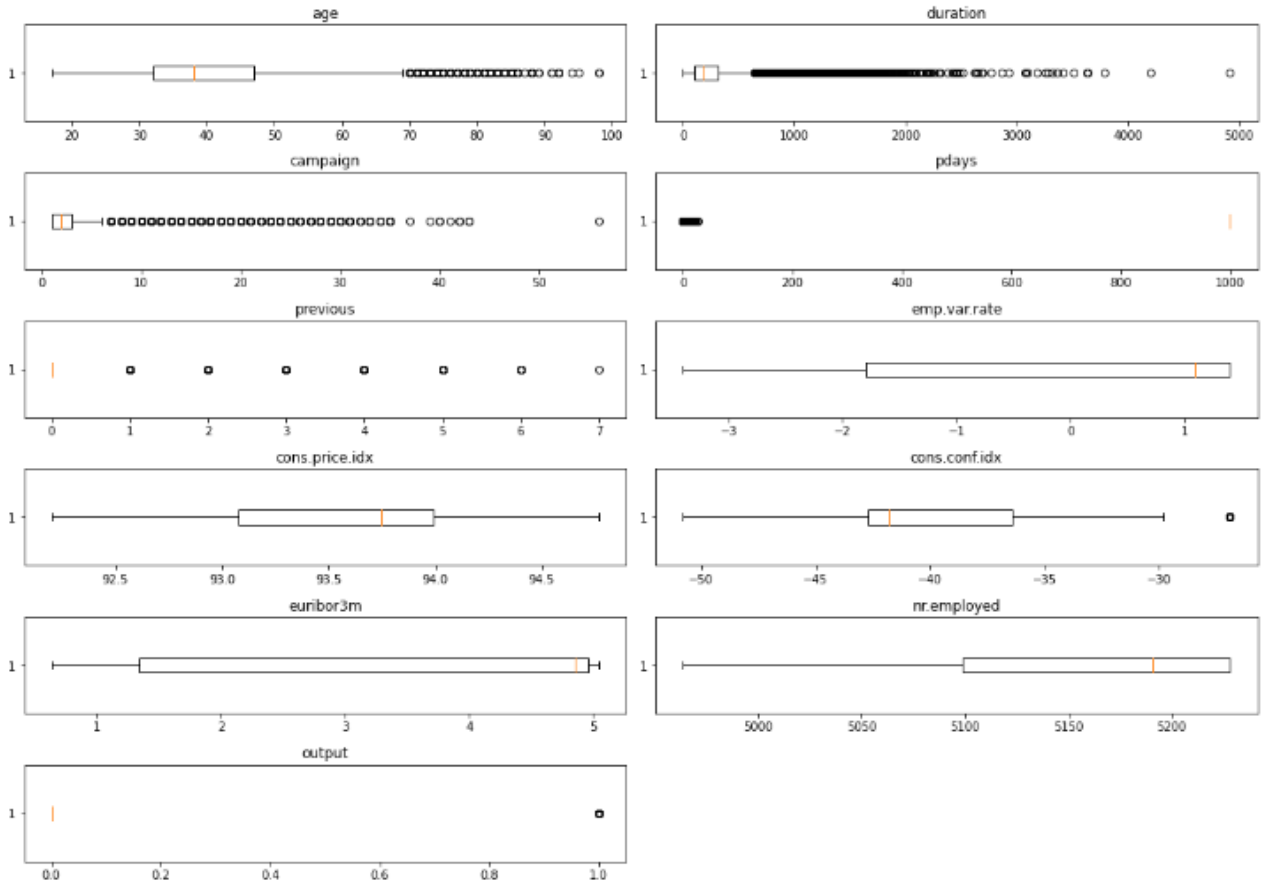
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   age             41188 non-null  int64  
 1   job             41188 non-null  object 
 2   marital         41188 non-null  object 
 3   education       41188 non-null  object 
 4   default         41188 non-null  object 
 5   housing         41188 non-null  object 
 6   loan            41188 non-null  object 
 7   contact         41188 non-null  object 
 8   month           41188 non-null  object 
 9   day_of_week     41188 non-null  object 
 10  duration        41188 non-null  int64  
 11  campaign        41188 non-null  int64  
 12  pdays           41188 non-null  int64  
 13  previous        41188 non-null  int64  
 14  poutcome        41188 non-null  object 
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object 
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

EDA

Descriptive analysis:

We checked all four data sets for Nan values and duplicates. There are only 12 duplicate rows in bank-full-addition, and we dropped the duplicate rows. To understand numerical features and see their distributions, we created histogram charts and box plot graphs.





The box plot for the ‘age’, ‘duration’, and ‘campaign’ feature shows there is a significant number of outliers in distribution.

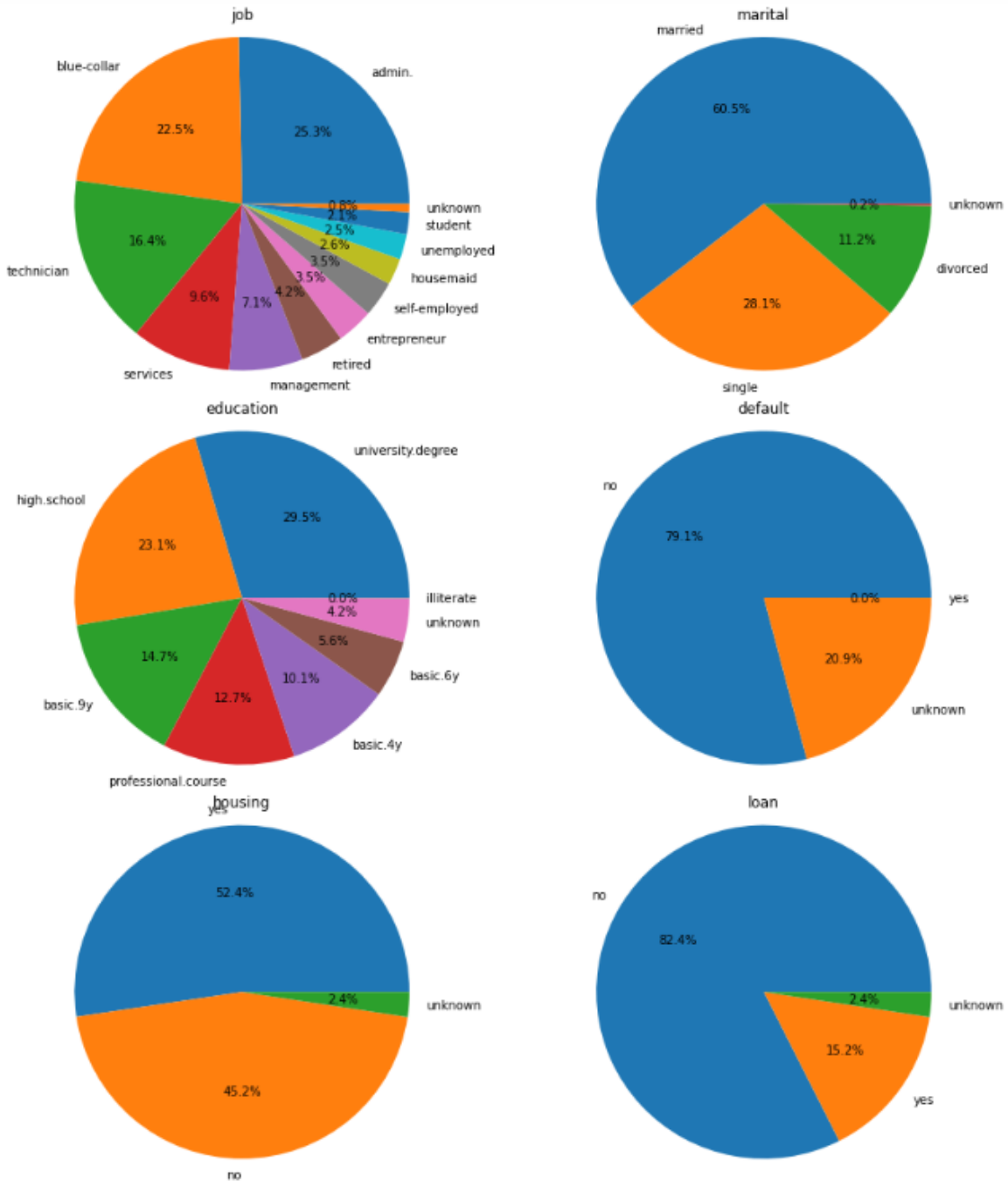
For outliers treatment, we can totally drop the outliers from the dataset but that could lead to the loss of some important data, which can help in giving predictions from the model. We can further look at the statistical distribution of these variables to decide further.

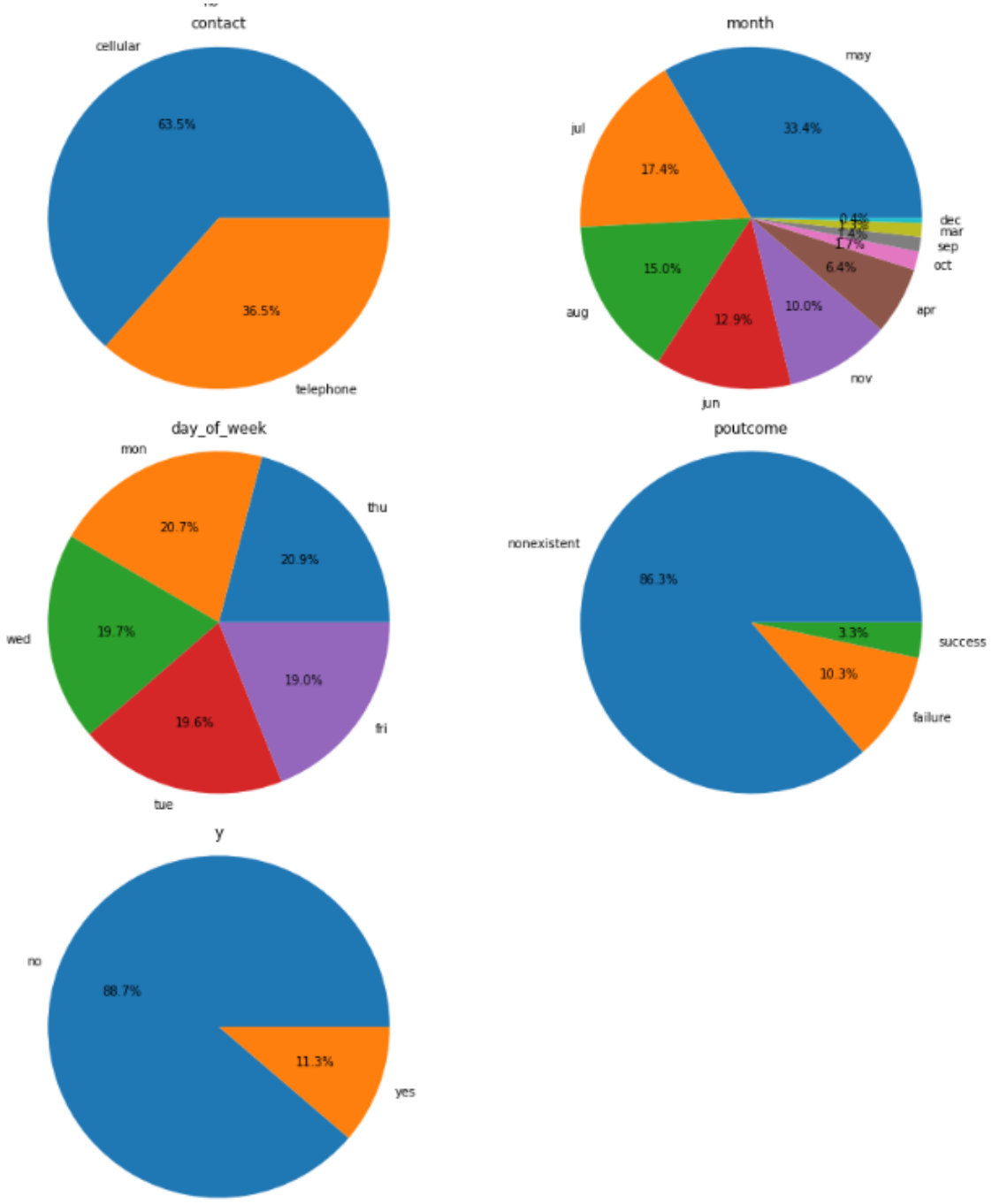
	age	duration	campaign
count	41188.00000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593
std	10.42125	259.279249	2.770014
min	17.00000	0.000000	1.000000
25%	32.00000	102.000000	1.000000
50%	38.00000	180.000000	2.000000
75%	47.00000	319.000000	3.000000
max	98.00000	4918.000000	56.000000

As mentioned in the data description we will be dropping the duration column entirely. The maximum values for ‘age’ and ‘campaign’ are 98 and 56 respectively, and they don't seem unrealistic. Therefore the outliers in the data distribution of these features don't need to be dropped.

Categorical Attributes:

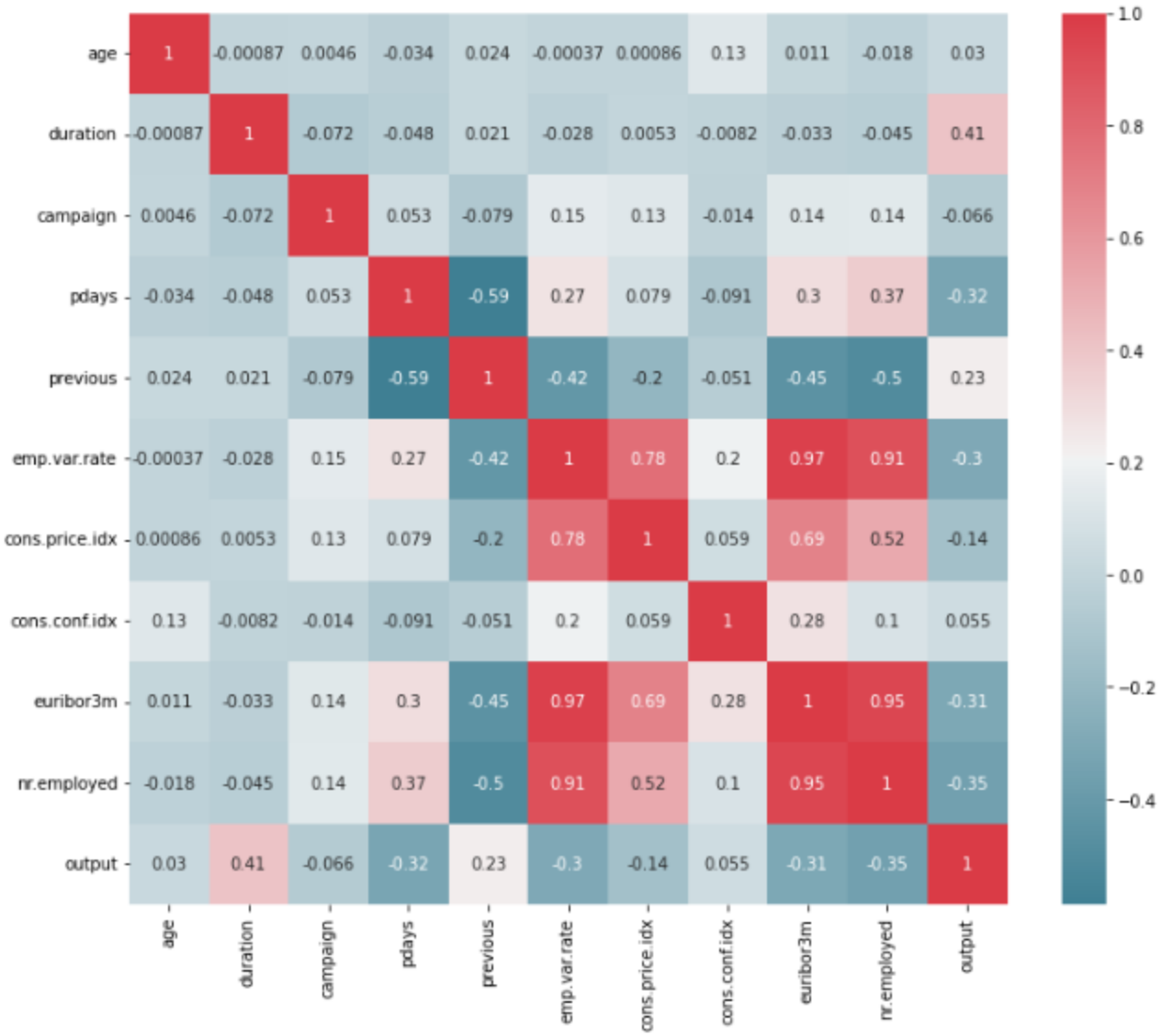
To understand the distribution of categorical variables pie charts are created as seen below.





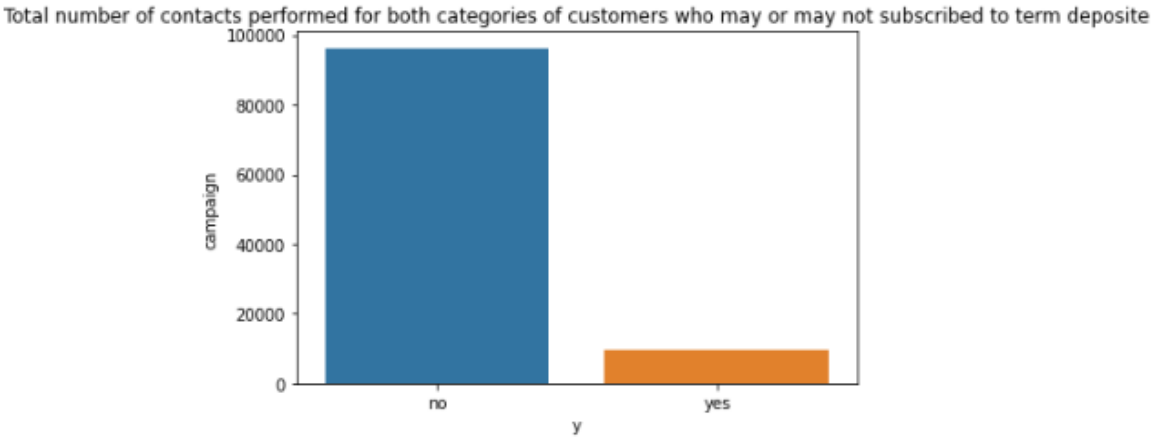
Correlation Analysis

Correlation or bivariate analysis helps understand the relationship between two attributes and shows how they are related.

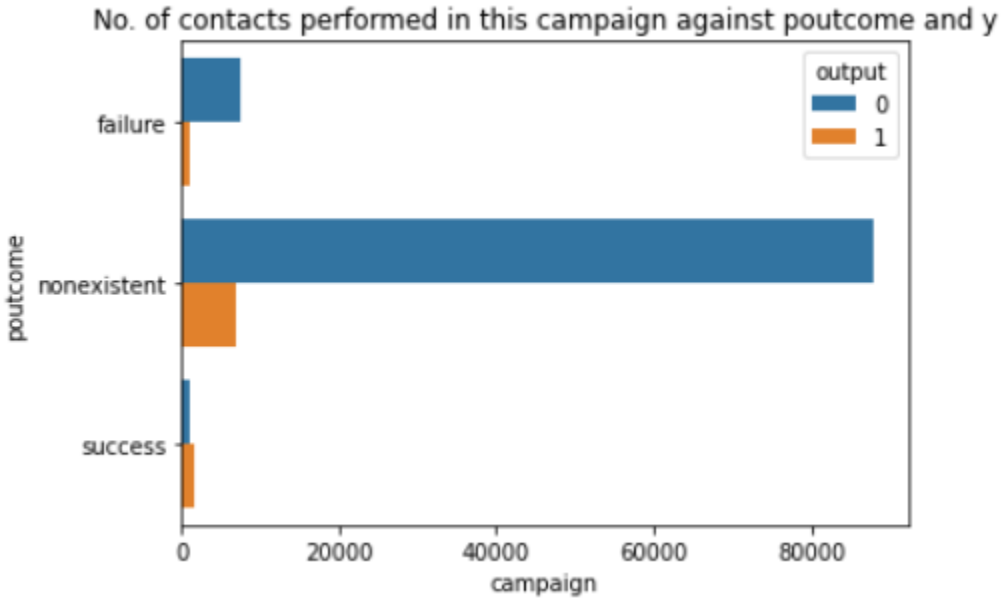


'Campaign' and 'y'

'campaign' feature, i.e. number of contacts made to the customer who has/hasn't subscribed to the term deposit plan. We analyzed the campaign feature against the target variable ‘y’. The number of contacts made in this campaign for those not subscribing to a term deposit is 96234, almost 86700 more than those opting for a term deposit plan.



The outcome of the previous marketing campaign will surely affect the decision of the customers on whether to opt or not for the term deposit scheme offered by the bank. The result of the previous campaign shows that the outcome was non-existent for 35000 customers. On the other hand, the number of failures and successes of the previous campaign is close to 12% of the total number of customers. We also analyzed the previous campaign outcome against the number of contacts made in this campaign.

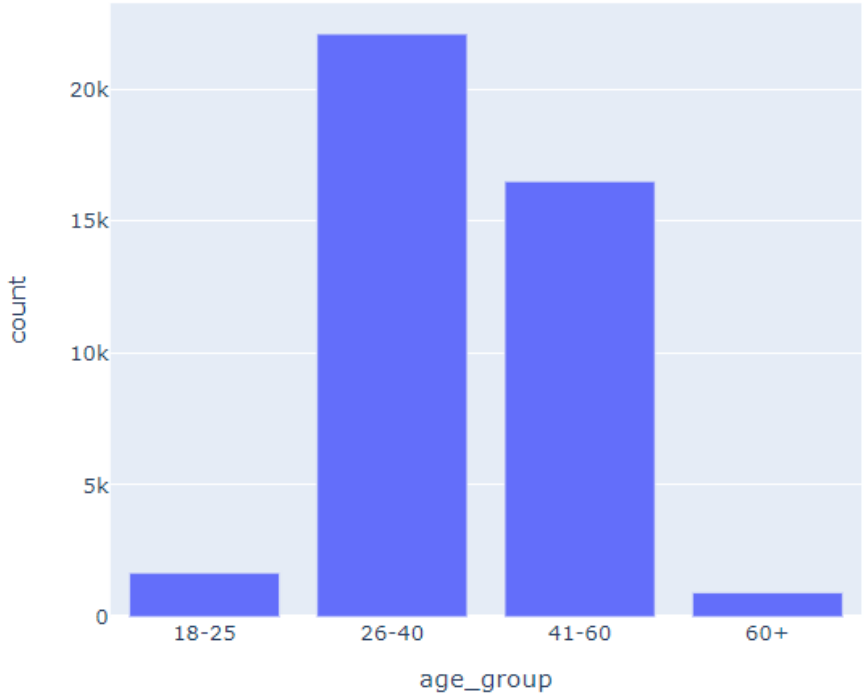


The above plot shows that more than 80000 contacts are made for customers who have not subscribed to the term deposit plan. And also the numbers for whom poutcome was failure or success were less comparatively.

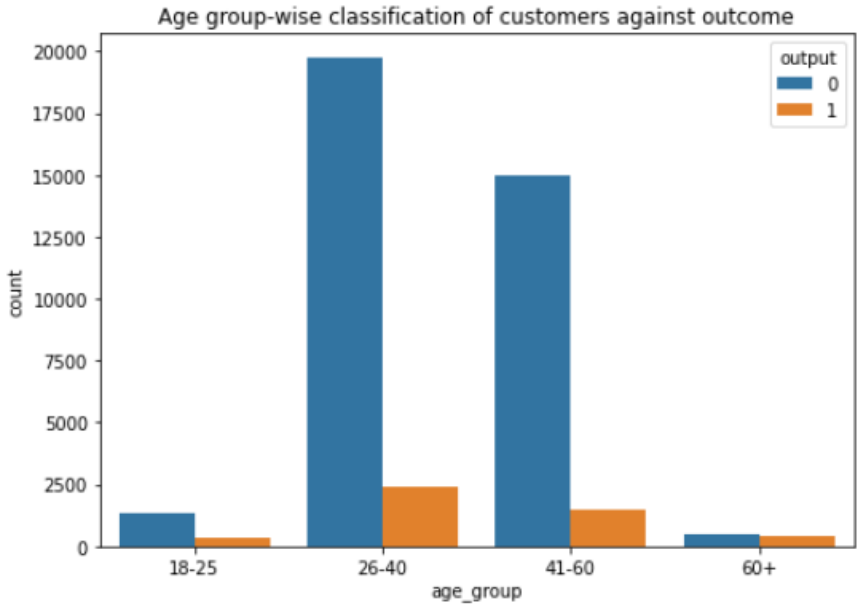
Also, we can see how the number of contacts performed during this campaign combined with other features affects our outcome.

Age Feature:

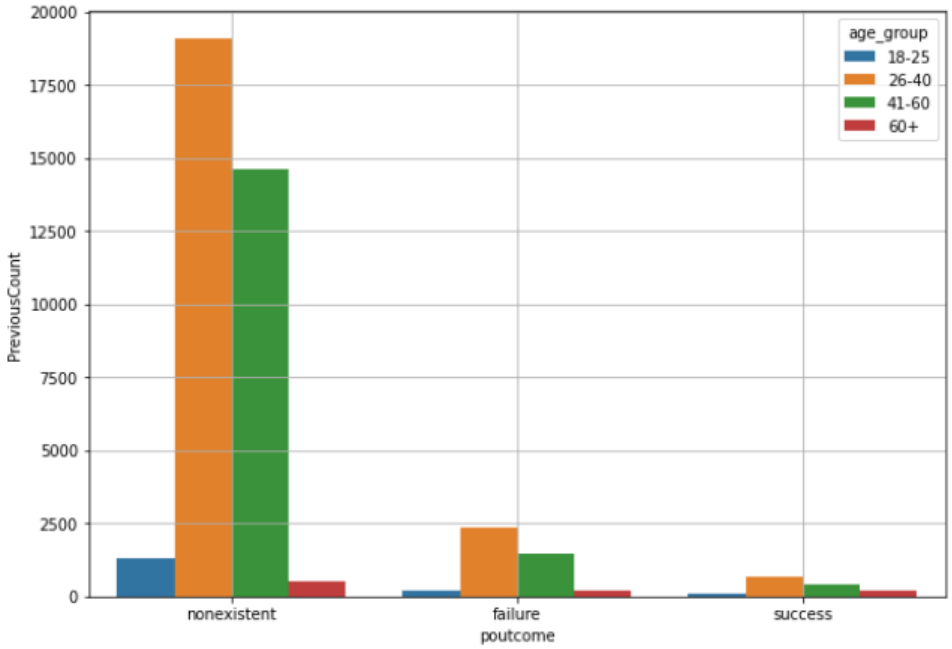
In the ‘age’ feature we have customers with ages ranging from 17 to 98 years old. For analysis purpose, we created age groups as 17-25,26-40,41-60, and 60+



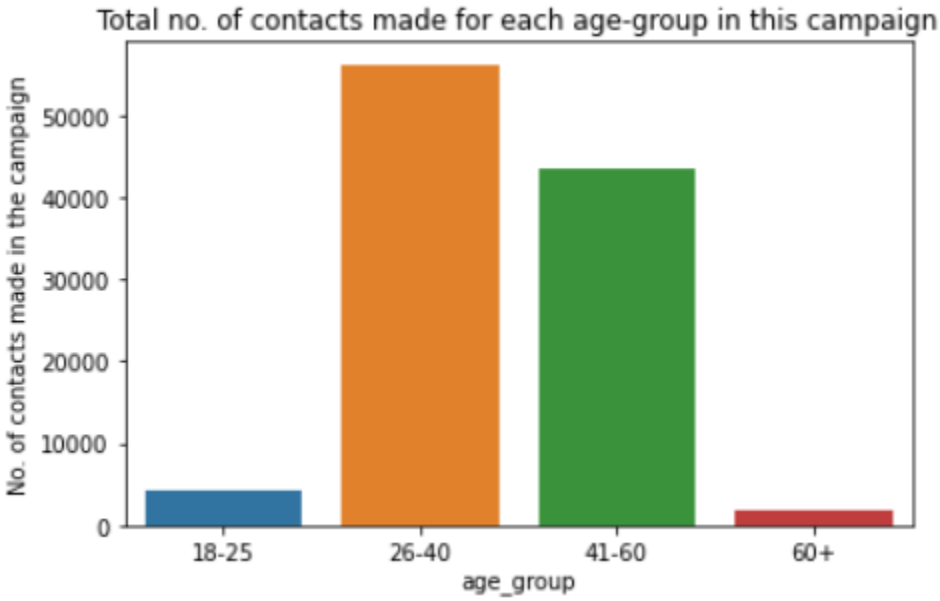
This newly created feature is further analyzed with the target variable.



Approximately 20000 customers from the 26-40 age group and 15000 customers from the 41-60 age group haven't subscribed to the term plan.



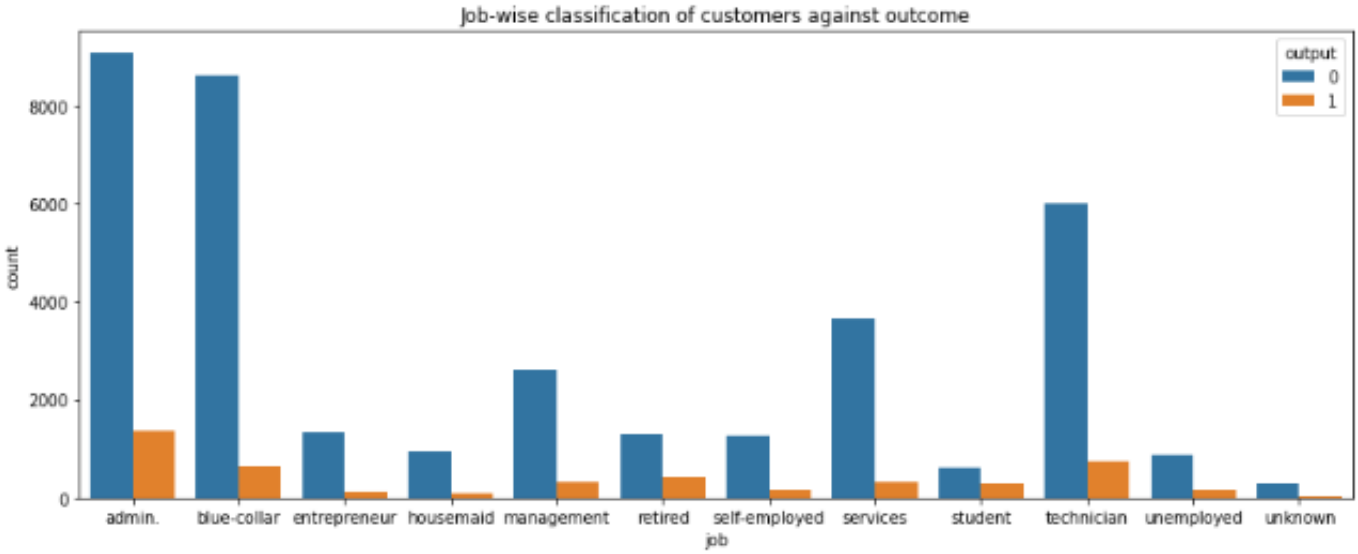
When the previous campaign's outcome was also checked against the age groups and we can see the almost same trend with customers' choices.



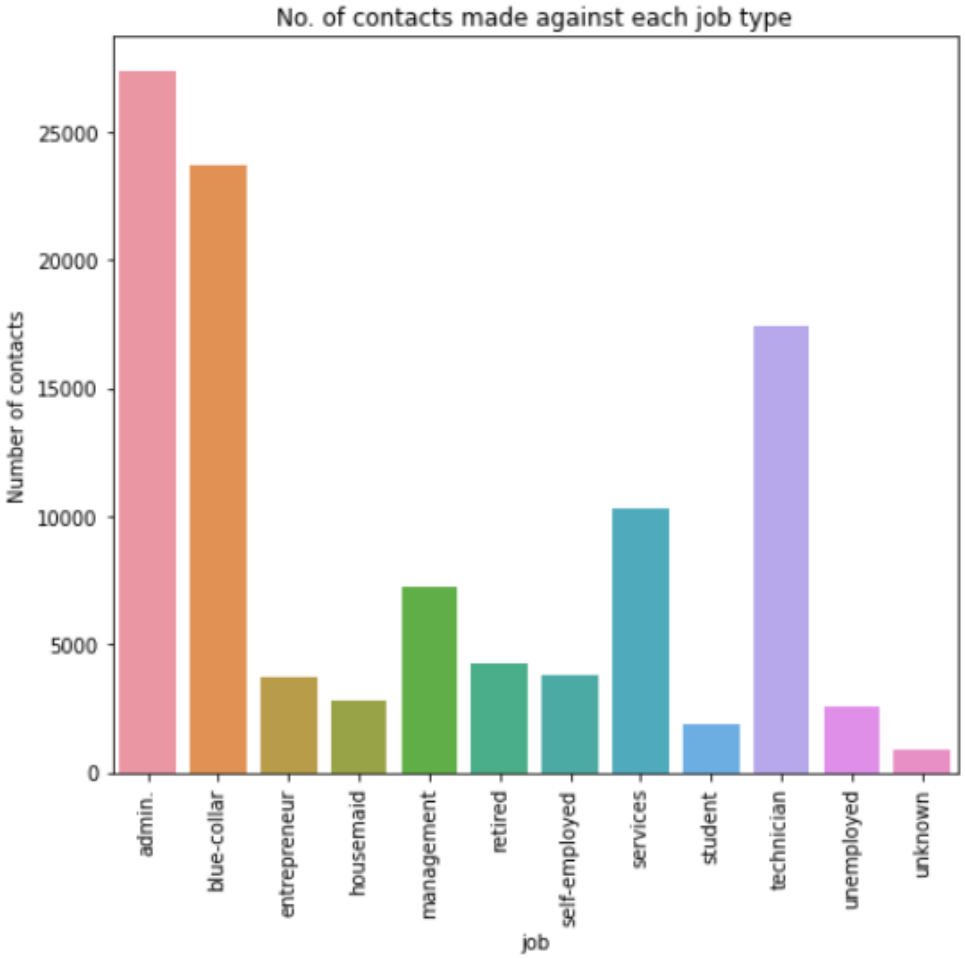
Definitely, as there are more customers from 26 to 60 years old, the marketing team is also focusing more on those two age groups.

Job feature:

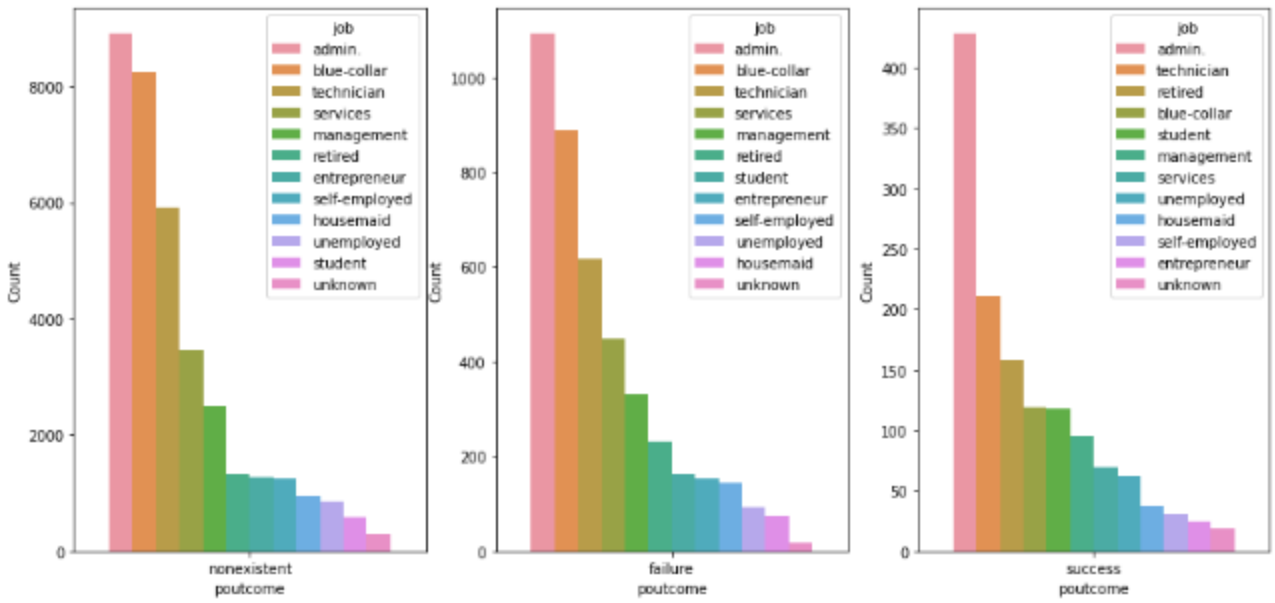
Then we analyzed job categories against the target variable.



Customers in the professions of admin, blue-collar, and technician have majorly rejected the term deposit plan. These three jobs comprise approximately 64% of the customer base

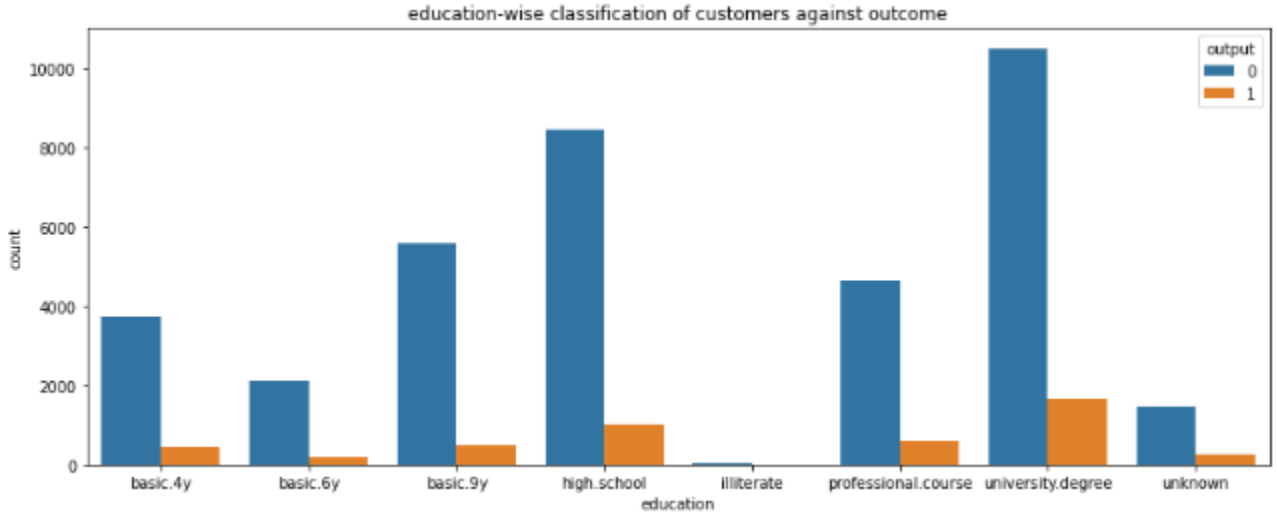


We found out that admin, blue collar, and technician job category customers are contacted over 20000 times.

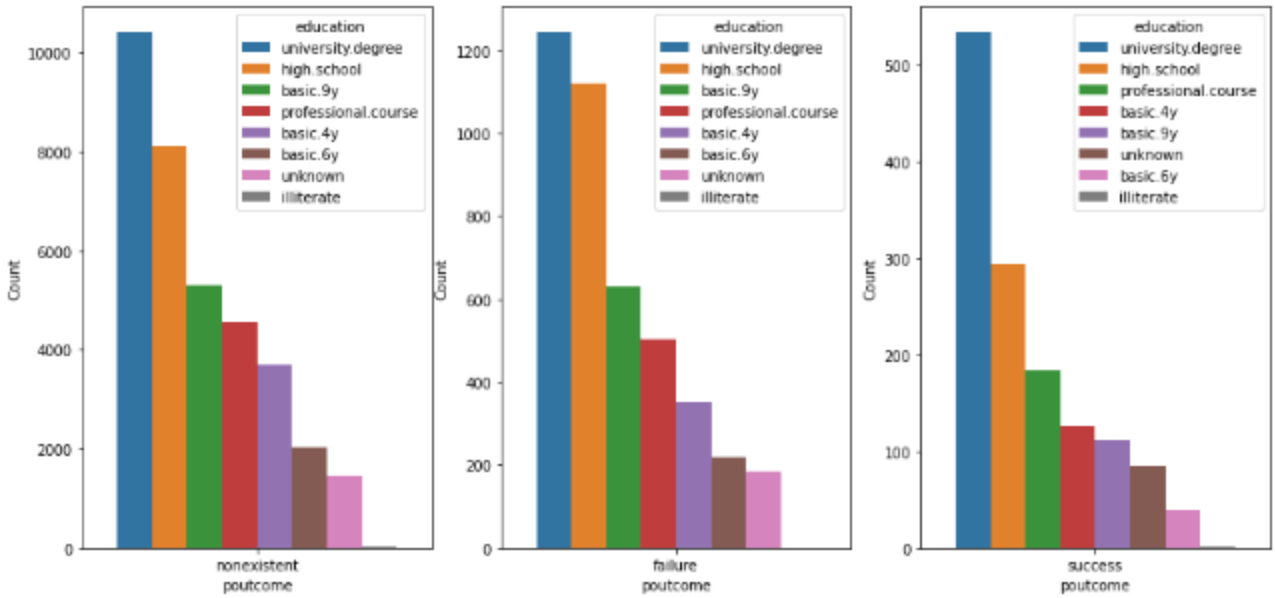


As seen in the above graph we analyzed the 'job' feature with the 'poutcome' variable. For the previous campaign, customers from the same job categories opted as seen in this campaign. For all non-existent outcomes, there are over 22000 customers in the roles of administrative, blue-collar, and technician job categories.

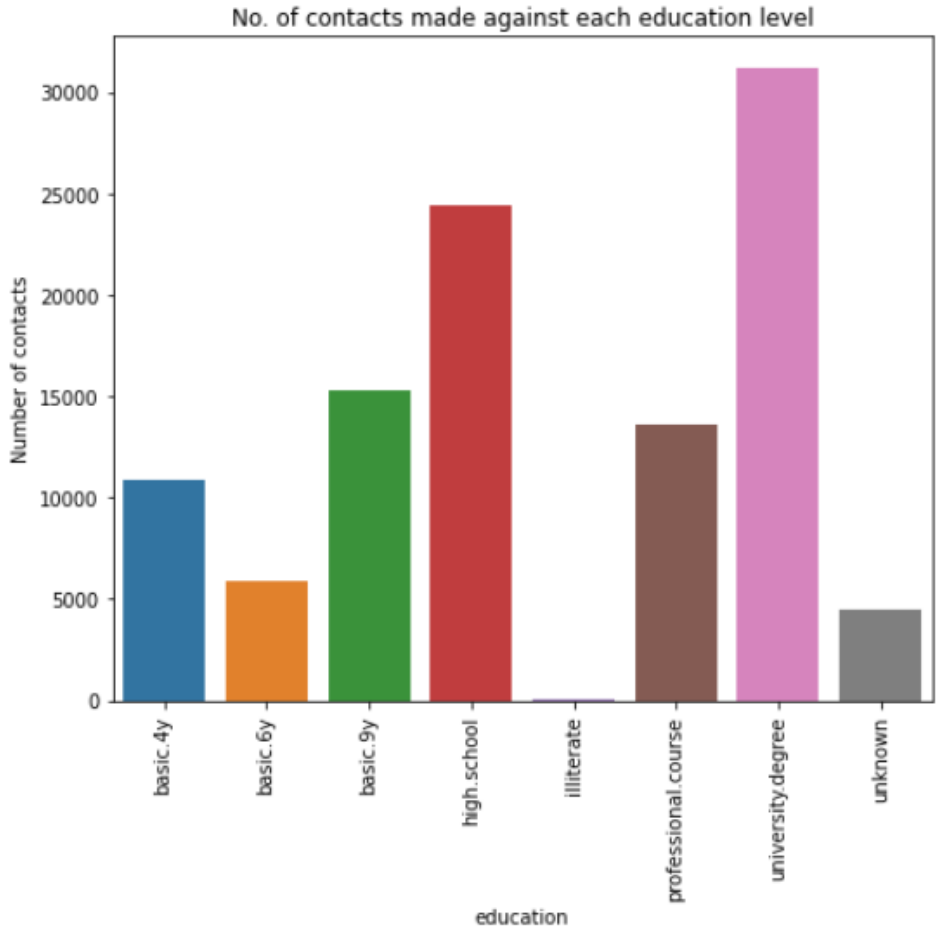
'Education' feature



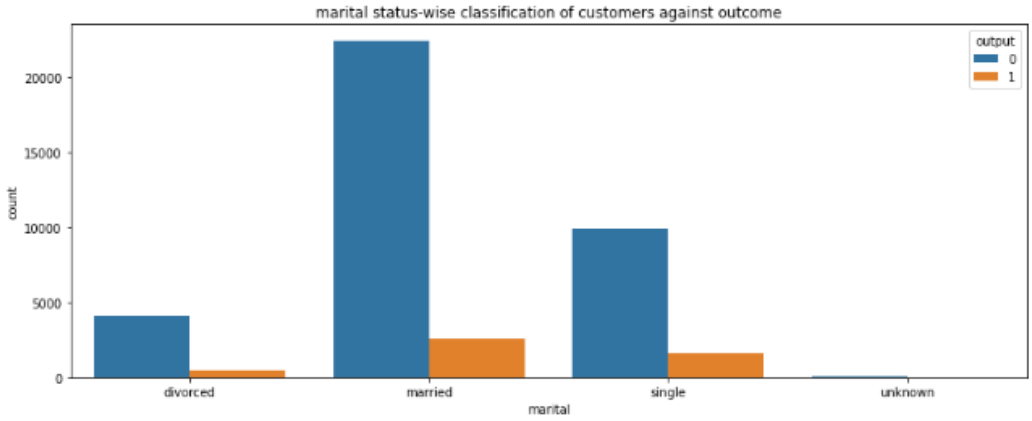
We have more than 50% of customers with University and high school education. To understand if they are opting more for a term deposit, we analyzed the education feature with the target variable y. We also have around 2000 customers with unknown education levels. We also see a similar trend during the last campaign.



The marketing team is also targeting the largest customer base education-wise.

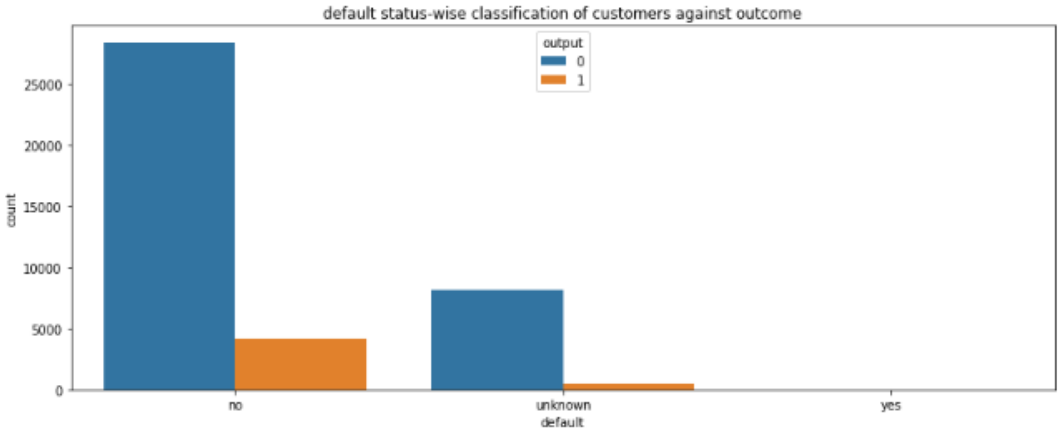


Marital feature:



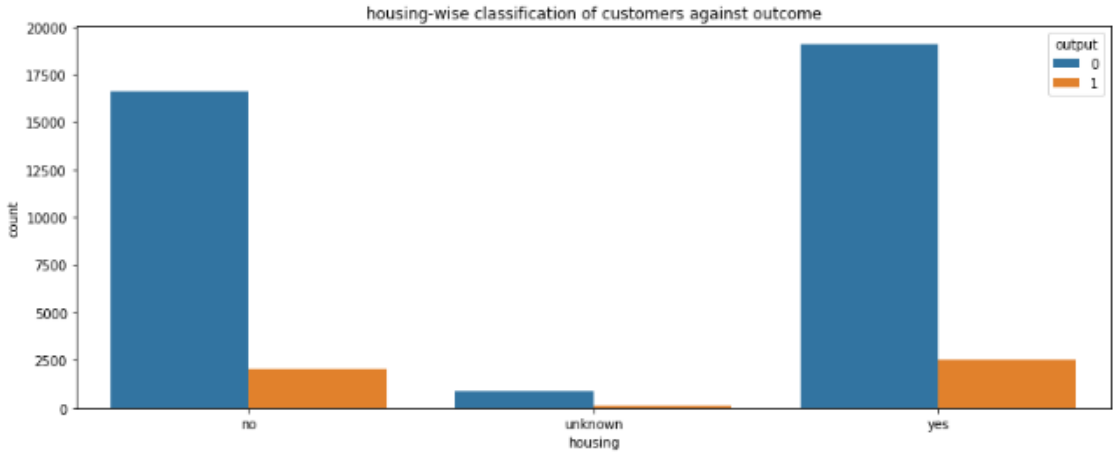
Majority of married and single customers have shown a rejection to term deposit plan.

Default feature:



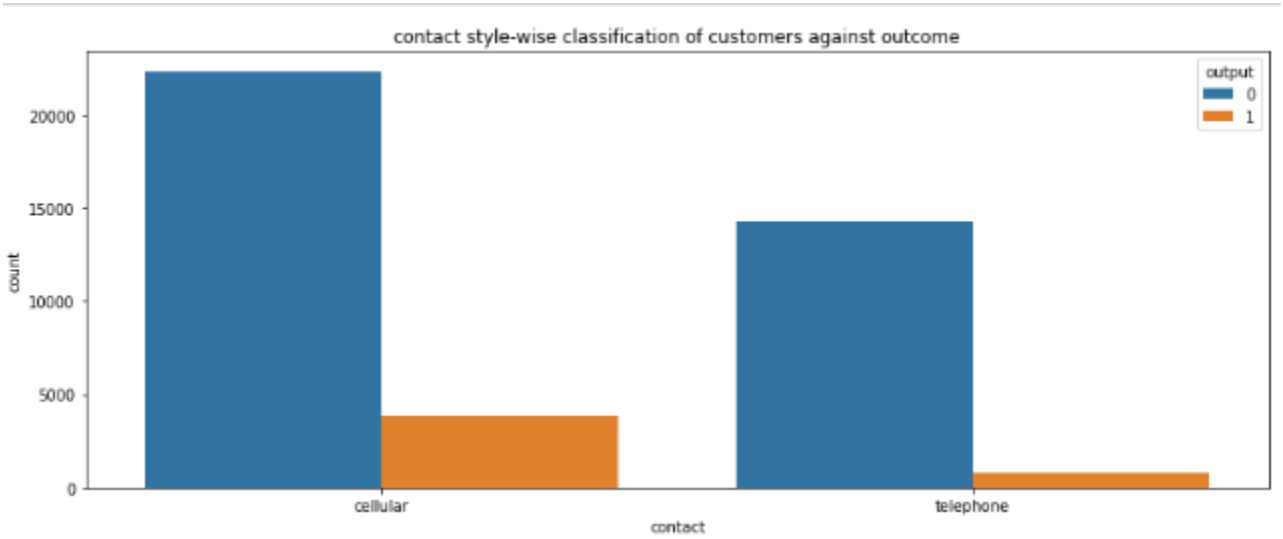
We have zero customers who have defaulted and arount 21% or less with unknown default status doesnt make sense.

Housing feature:



Contact Feature:

In order to check if way of contacting customer affect their choice to subscribe to term plan, we analyzed contact feature against the target variable y.



Looking at the numbers, approximately 63.5% customers have been contacted through cellular mode in this campaign, with rest through telephone.