# Deliverable – Week 10

**Group Name:** Tony's group

**Team:**


**Name:** Tony Nguyen

**Email:** tonynguyen9707@gmail.com

**Country:** United States

**College:** University of Florida

**Specialization:** Data Analyst


## 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 help 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).

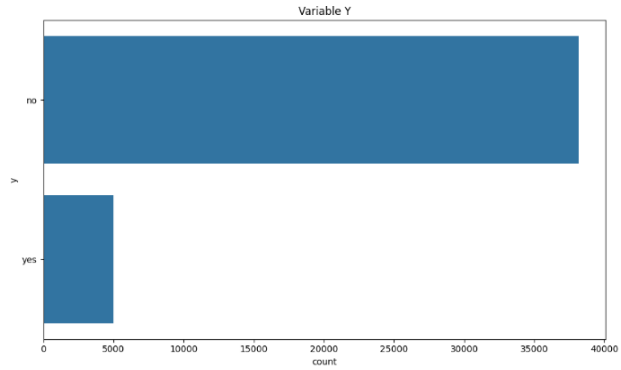
## GitHub Link

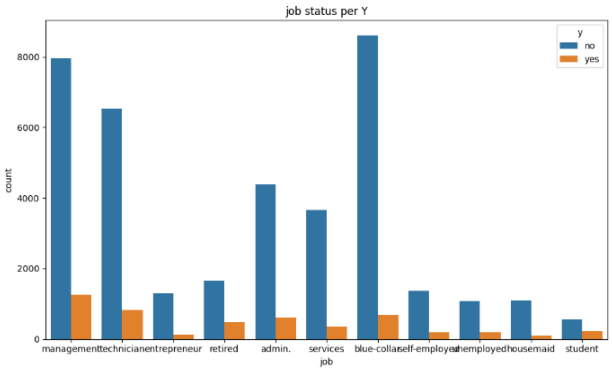
https://github.com/ttn20/DataGlacierInternship/tree/Bank-Marketing-Final-Project

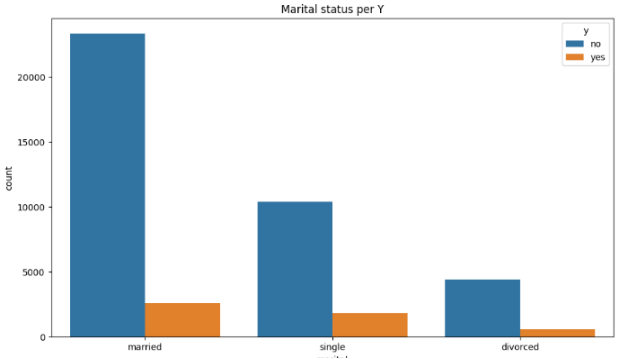
## Exploratory Data Analysis

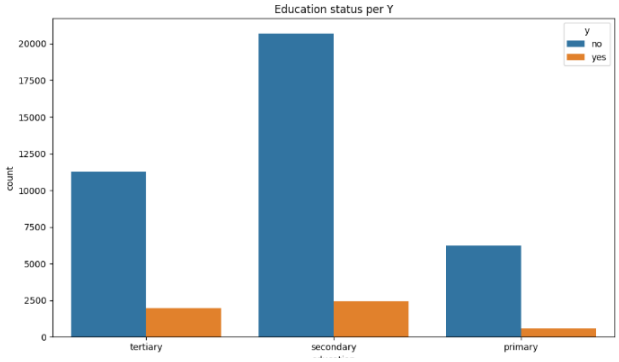
To see full code of EDA: eda.ipynb

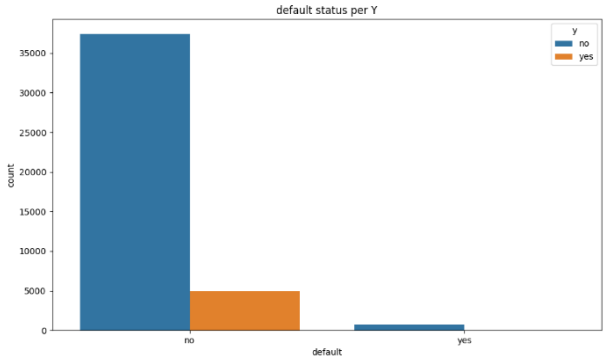
Categorical Data Graphs:

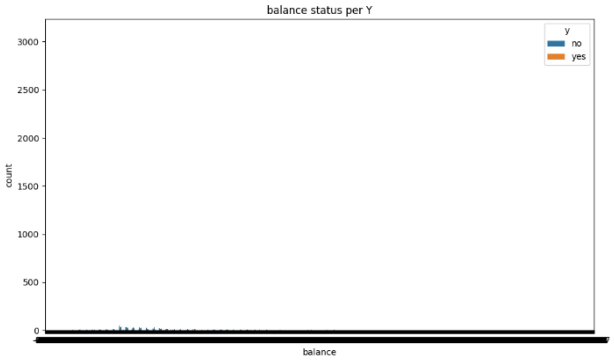

Variable Y
y
no
yes
0
5000
10000
15000
20000
25000
30000
35000
40000
count

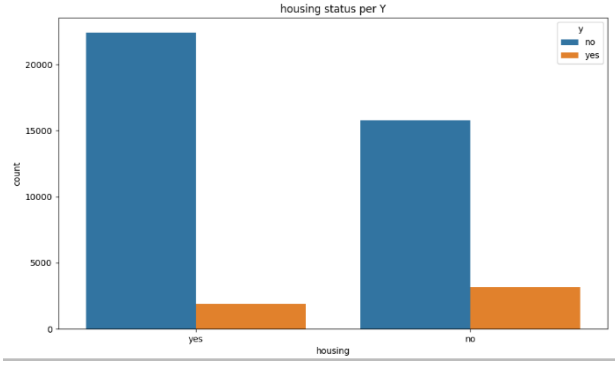

job status per Y
y
no
yes
count
0
2000
4000
6000
8000
management
technician
entrepreneur
retired
admin.
services
blue-collar
self-employed
unemployed
housemaid
student
job

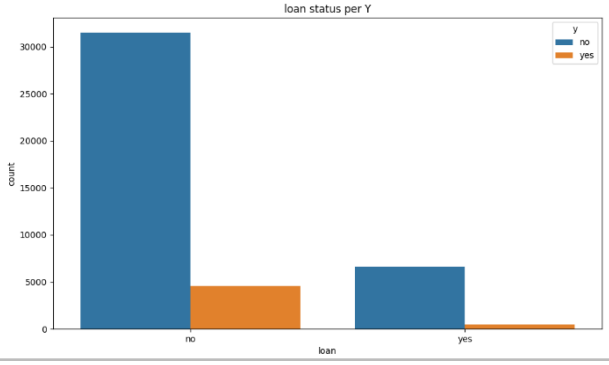

Marital status per Y
y
no
yes
count
0
5000
10000
15000
20000
married
single
divorced

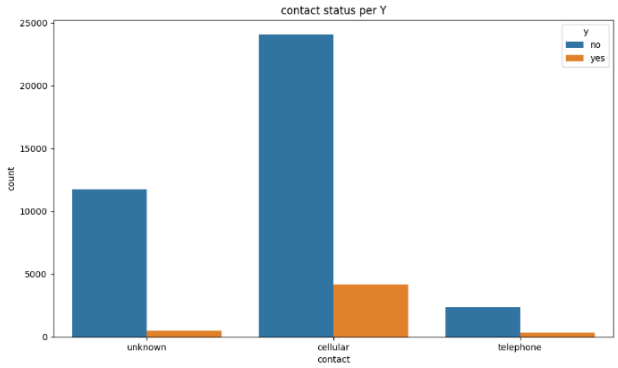

Education status per Y
y
no
yes
count
0
2500
5000
7500
10000
12500
15000
17500
20000
tertiary
secondary
primary

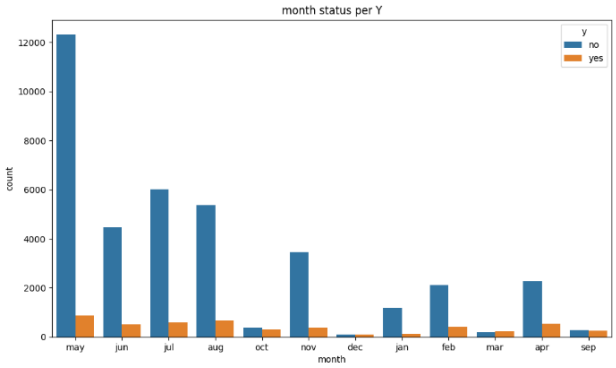

default status per Y
y
no
yes
count
0
5000
10000
15000
20000
25000
30000
35000
no
yes
default

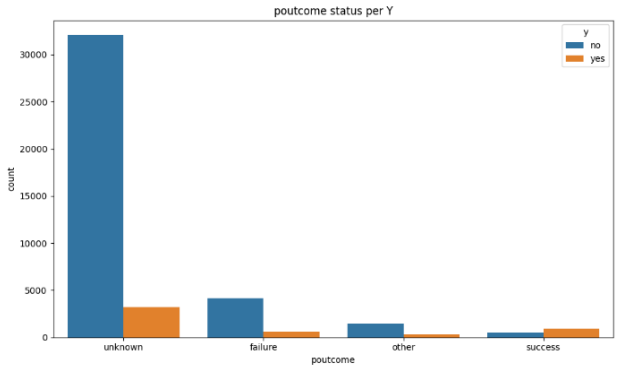

balance status per Y
y
no
yes
count
0
500
1000
1500
2000
2500
3000
balance


housing status per Y
y
no
yes
count
0
5000
10000
15000
20000
yes
no
housing

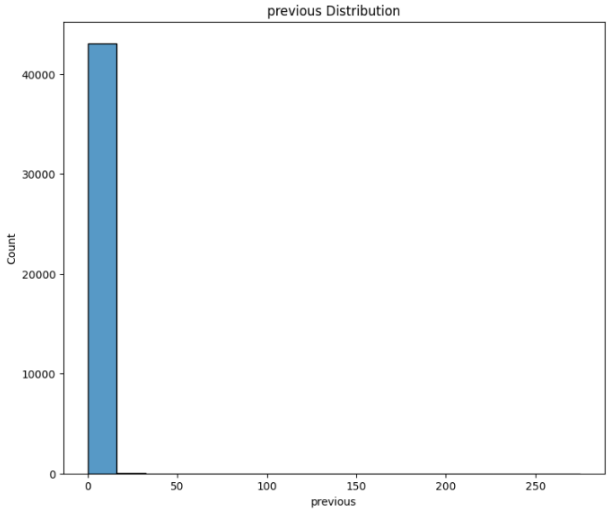

loan status per Y
y
no
yes
count
0
5000
10000
15000
20000
25000
30000
no
yes
loan

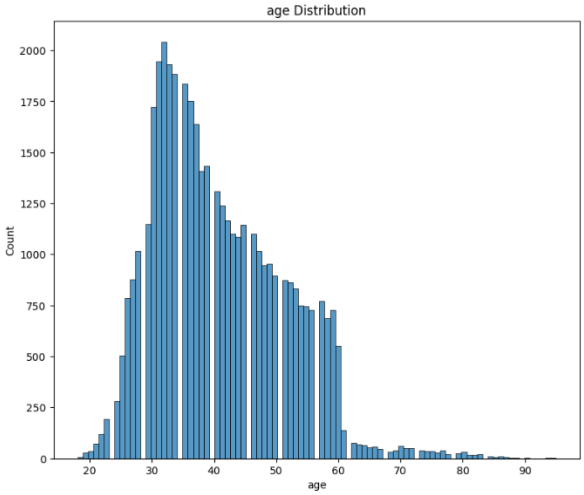


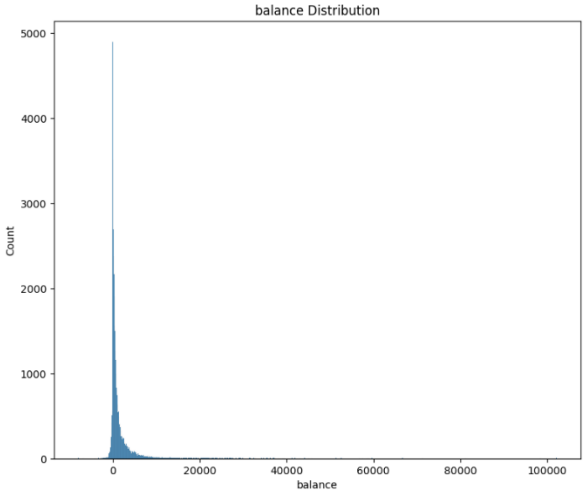



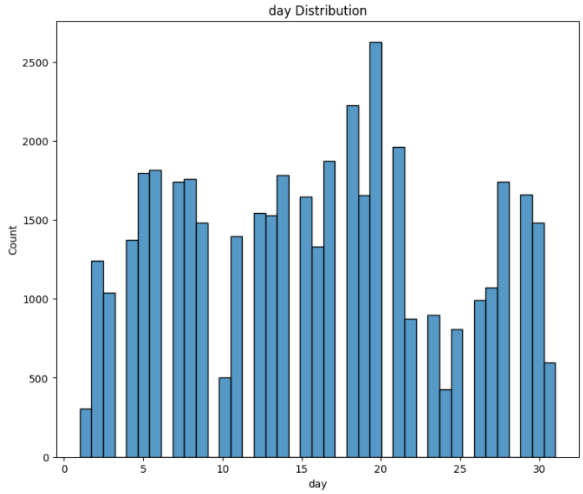



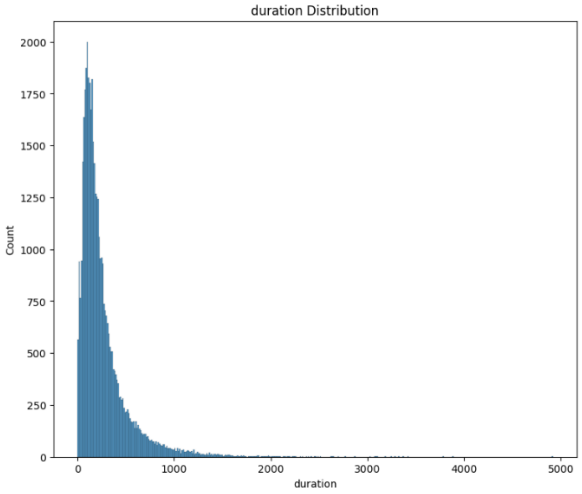

## Numerical Data Graphs:

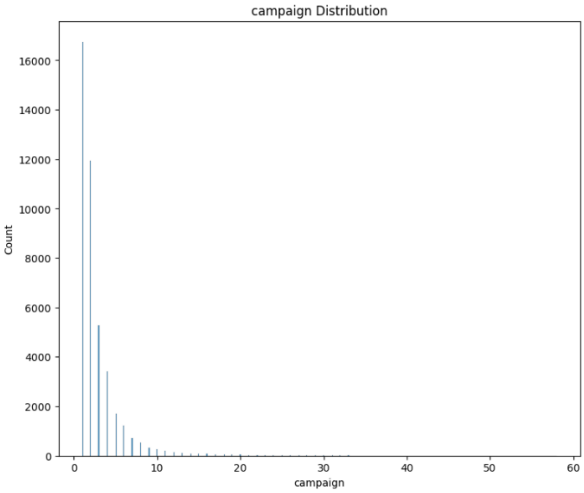


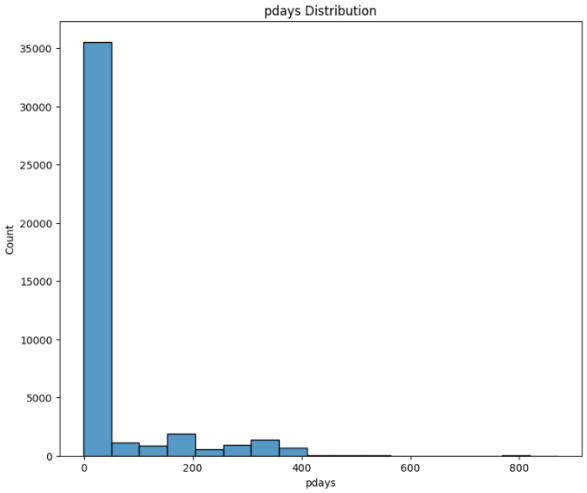

age Distribution
Count
age


balance Distribution
Count
balance

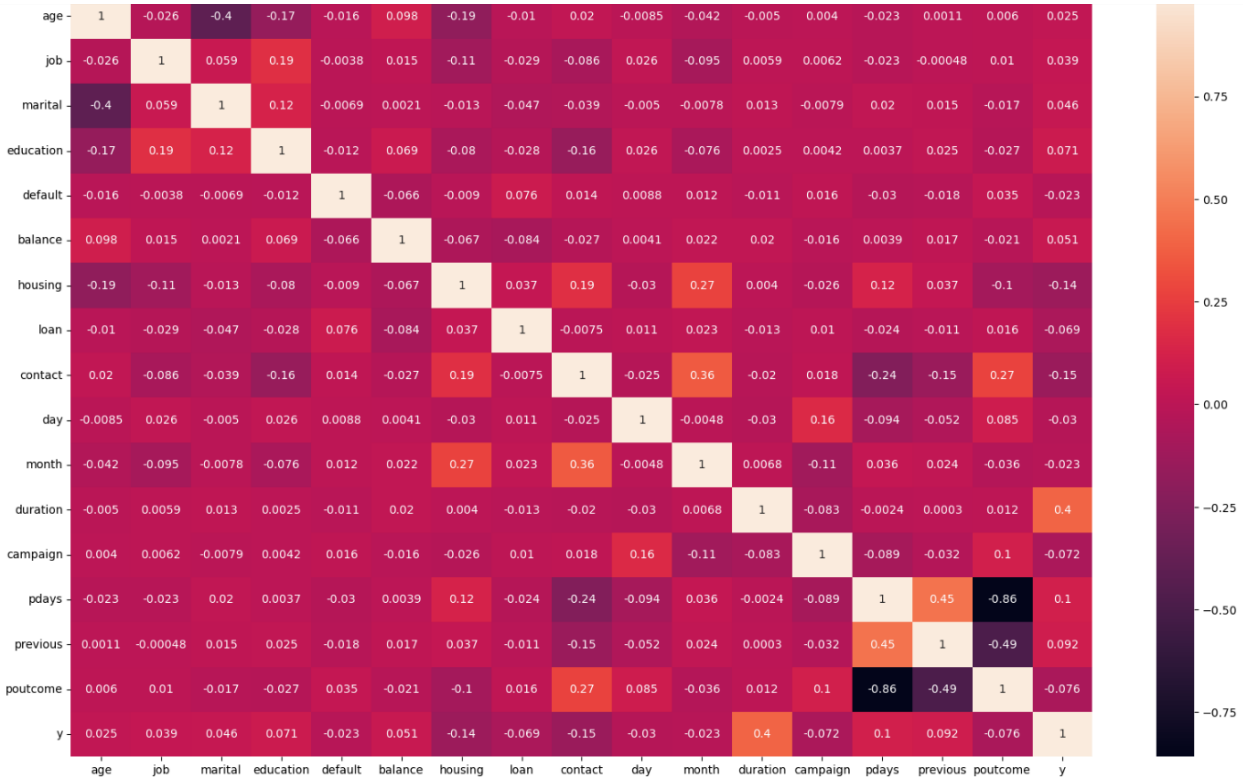

day Distribution
Count
day


duration Distribution
Count
duration


campaign Distribution
Count
campaign


pdays Distribution
Count
pdays

Correlation:




```
duration     0.157921
contact      0.021171
housing      0.019127
pdays        0.010291
previous     0.008421
poutcome     0.005784
campaign     0.005251
education    0.005030
loan         0.004734
balance      0.002641
marital      0.002093
job          0.001546
day          0.000915
age          0.000613
default      0.000540
month        0.000532
Name: y, dtype: float64
```

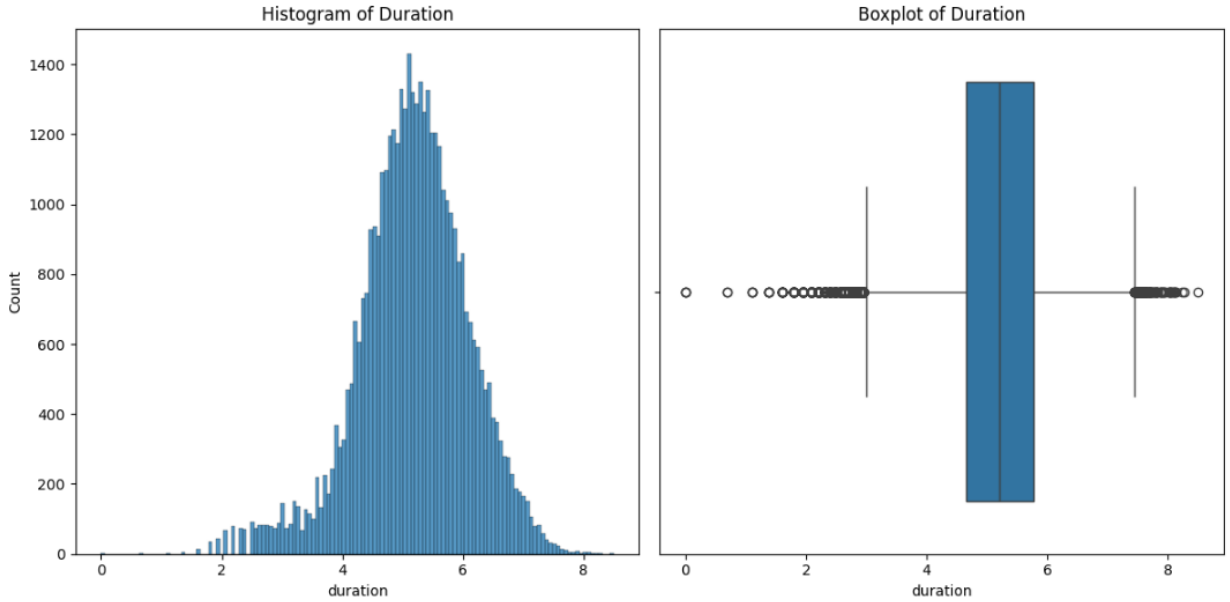
Duration is significantly more correlated than the other variables. From the Data Cleaning & Checking section, we learned that duration has outliers and is skewed. To fix that, use numpy.log1p to normatlize the data.

```
# Fixing outliers and skewness in duration
data["duration"] = np.log1p(data["duration"])

fig, axes = plt.subplots(1, 2, figsize=(12, 6))
sb.histplot(data=data, x='duration', ax=axes[0])
axes[0].set_title('Histogram of Duration')

sb.boxplot(data=data, x='duration', ax=axes[1])
axes[1].set_title('Boxplot of Duration')

plt.tight_layout()
plt.show()
```




To balance the y variable that we looked at earlier, we use oversampling.

```
# Balancing Y variable

data_majority = data[data['y'] == 0]
data_minority = data[data['y'] == 1]
data_minority_upsampled = resample(data_minority, replace=True, n_samples=len(data_majority), random_state=1)
data_balanced = pd.concat([data_majority, data_minority_upsampled])
data_balanced['y'].value_counts()
```

```
0    38172
1    38172
Name: y, dtype: int64
```

With these changes, future models with this data should be improved.

# Final Recommendations

To increase customers that make a deposit:

- Duration: Increase follow up contact (Limitation shown in pdays)
- Contact: Use cellular to contact customers.
- Housing: Focus on customers without a housing loan.
- pdays: Do not contact customers in an excessive amount.