Cruisebound

October 13, 2024

[1]:

```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv('bank-full.csv', delimiter=';')

# Display the first few rows
print(df.head())
```

```
   age           job  marital  education default  balance housing loan  \
0   58    management  married   tertiary      no     2143     yes   no
1   44    technician   single  secondary      no       29     yes   no
2   33  entrepreneur  married  secondary      no        2     yes  yes
3   47   blue-collar  married    unknown      no     1506     yes   no
4   33       unknown   single    unknown      no        1      no   no

   contact  day month  duration  campaign  pdays  previous poutcome   y
0  unknown    5   may       261         1     -1         0  unknown  no
1  unknown    5   may       151         1     -1         0  unknown  no
2  unknown    5   may        76         1     -1         0  unknown  no
3  unknown    5   may        92         1     -1         0  unknown  no
4  unknown    5   may       198         1     -1         0  unknown  no
```

[9]:

```
# Get basic info about the data
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

[8]:

```python
# Check for missing values
print(df.isnull().sum())
```

```
age          0
job          0
marital      0
education    0
default      0
balance      0
housing      0
loan         0
contact      0
day          0
month        0
duration     0
campaign     0
pdays        0
previous     0
poutcome     0
y            0
dtype: int64
```

[3]:

```python
# Summary statistics
print(df.describe())
```

```
                age        balance           day      duration      campaign  \
count  45211.000000   45211.000000  45211.000000  45211.000000  45211.000000
mean      40.936210    1362.272058     15.806419    258.163080      2.763841
std       10.618762    3044.765829      8.322476    257.527812      3.098021
min       18.000000   -8019.000000      1.000000      0.000000      1.000000
25%       33.000000      72.000000      8.000000    103.000000      1.000000
50%       39.000000     448.000000     16.000000    180.000000      2.000000
75%       48.000000    1428.000000     21.000000    319.000000      3.000000
```

```
max       95.000000  102127.000000     31.000000   4918.000000     63.000000

              pdays      previous
count  45211.000000  45211.000000
mean      40.197828      0.580323
std      100.128746      2.303441
min       -1.000000      0.000000
25%       -1.000000      0.000000
50%       -1.000000      0.000000
75%       -1.000000      0.000000
max      871.000000    275.000000
y
no     88.30152
yes    11.69848
Name: proportion, dtype: float64
```

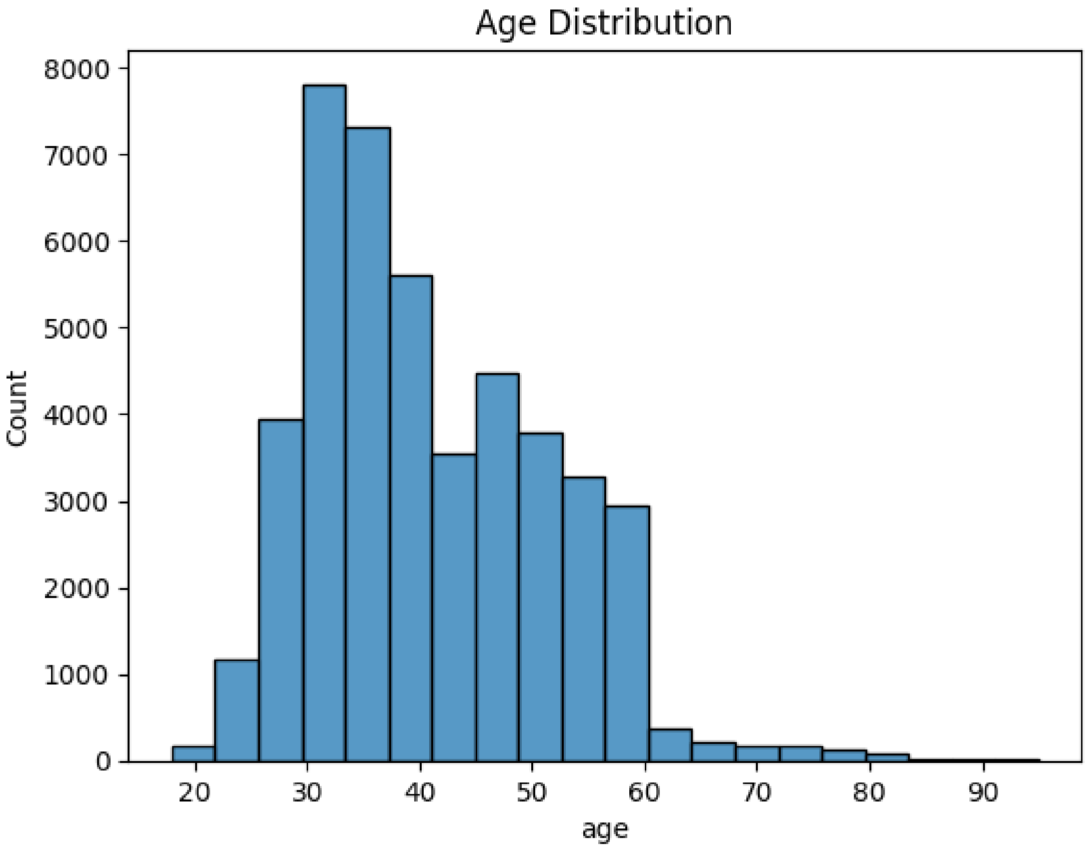
[10]:

```python
# Count the number of subscriptions to term deposits (target variable 'y')
print(df['y'].value_counts(normalize=True) * 100)  # Percentage of yes/no
```

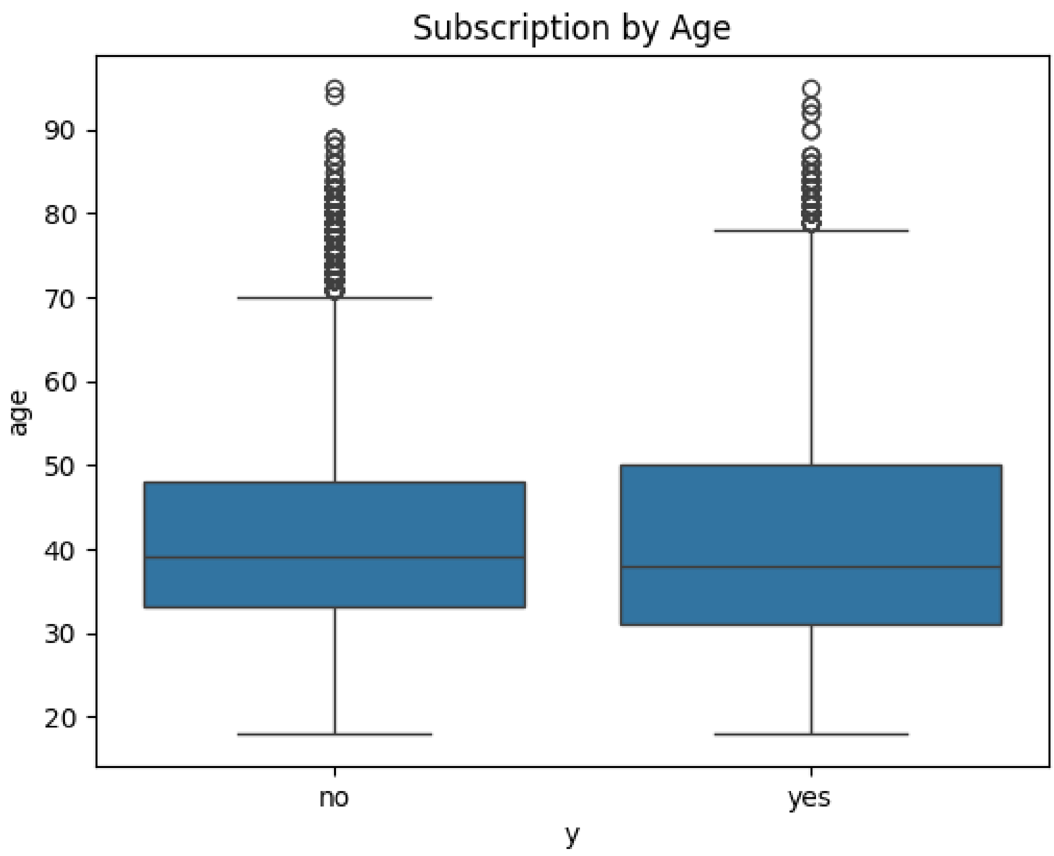
```
y
no     88.30152
yes    11.69848
Name: proportion, dtype: float64
```

[11]:

```python
sns.histplot(df['age'], bins=20)
plt.title('Age Distribution')
plt.show()
```

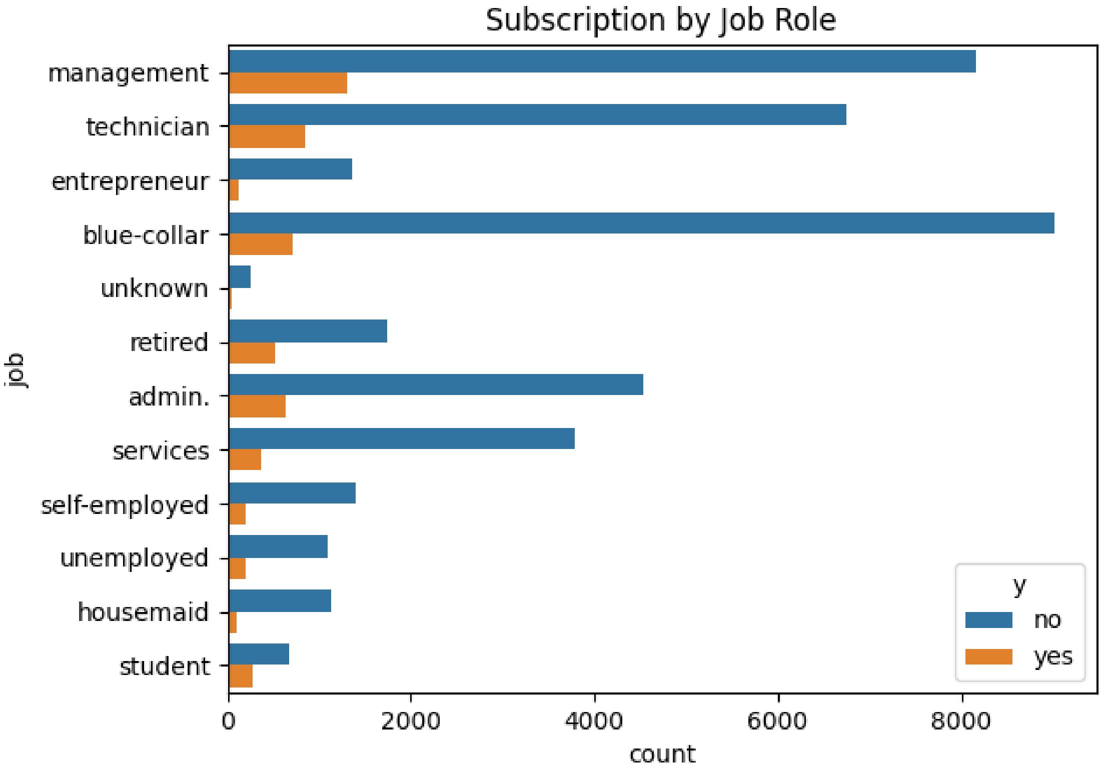


[12]:
```
sns.boxplot(x='y', y='age', data=df)
plt.title('Subscription by Age')
plt.show()
```

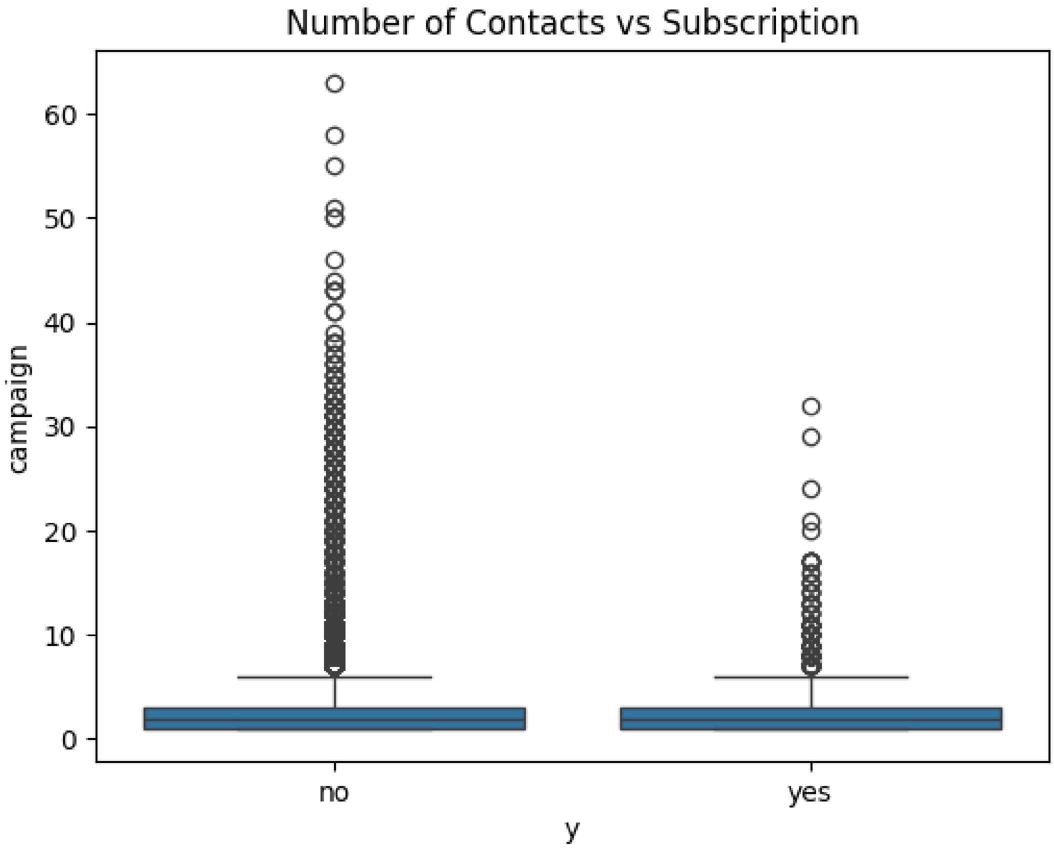


[13]:

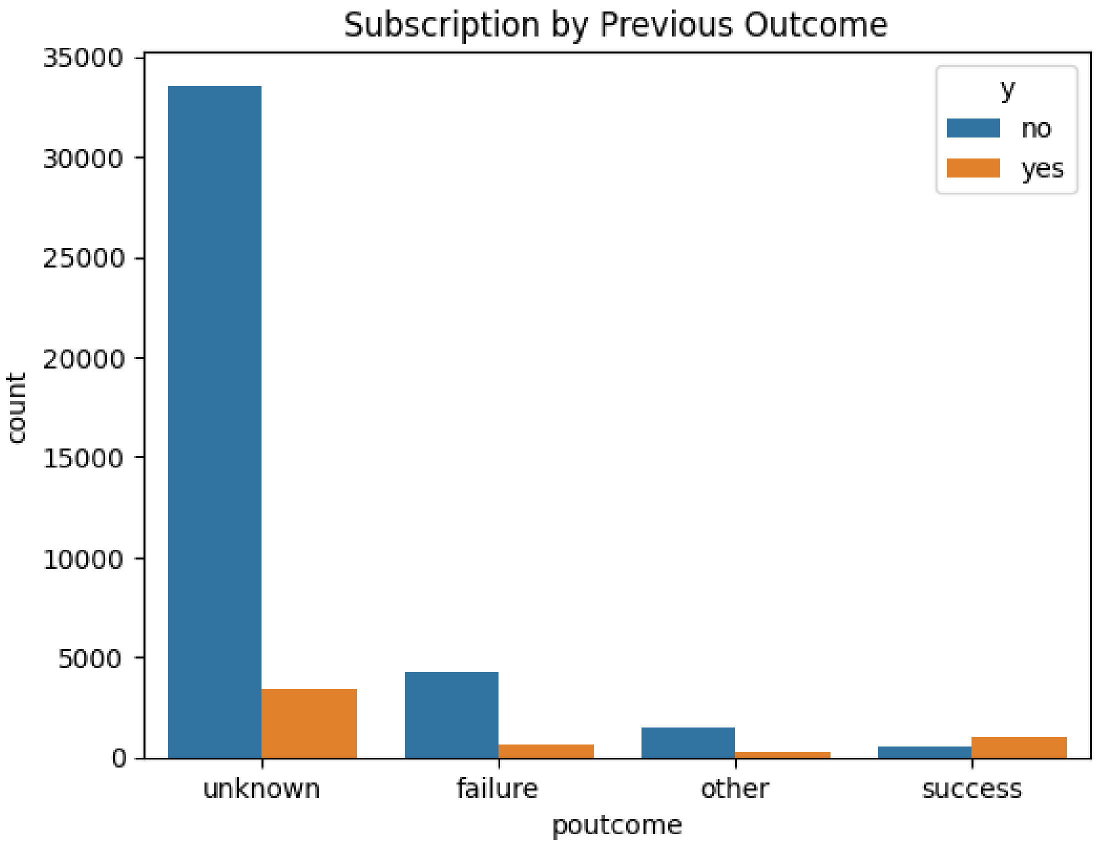
```
sns.countplot(y='job', hue='y', data=df)
plt.title('Subscription by Job Role')
plt.show()
```



```
[14]: sns.boxplot(x='y', y='campaign', data=df)
      plt.title('Number of Contacts vs Subscription')
      plt.show()
```

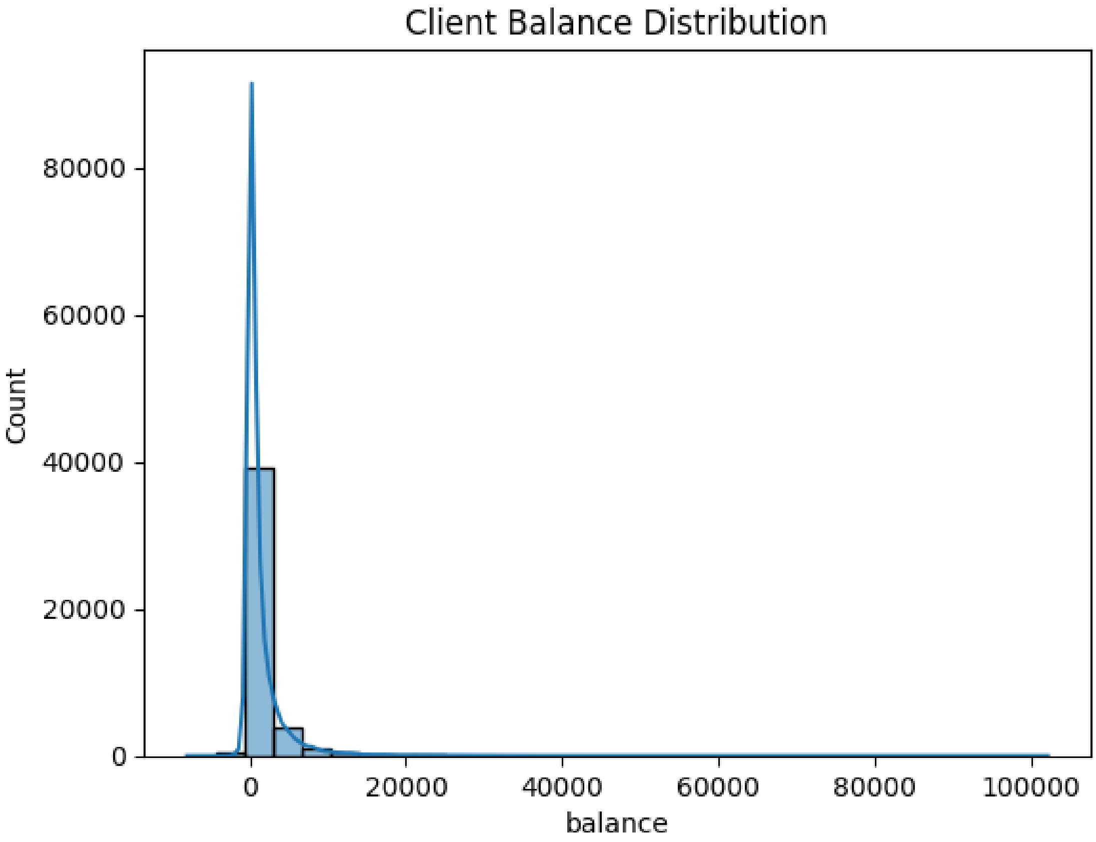


[15]:
```
sns.countplot(x='poutcome', hue='y', data=df)
plt.title('Subscription by Previous Outcome')
plt.show()
```

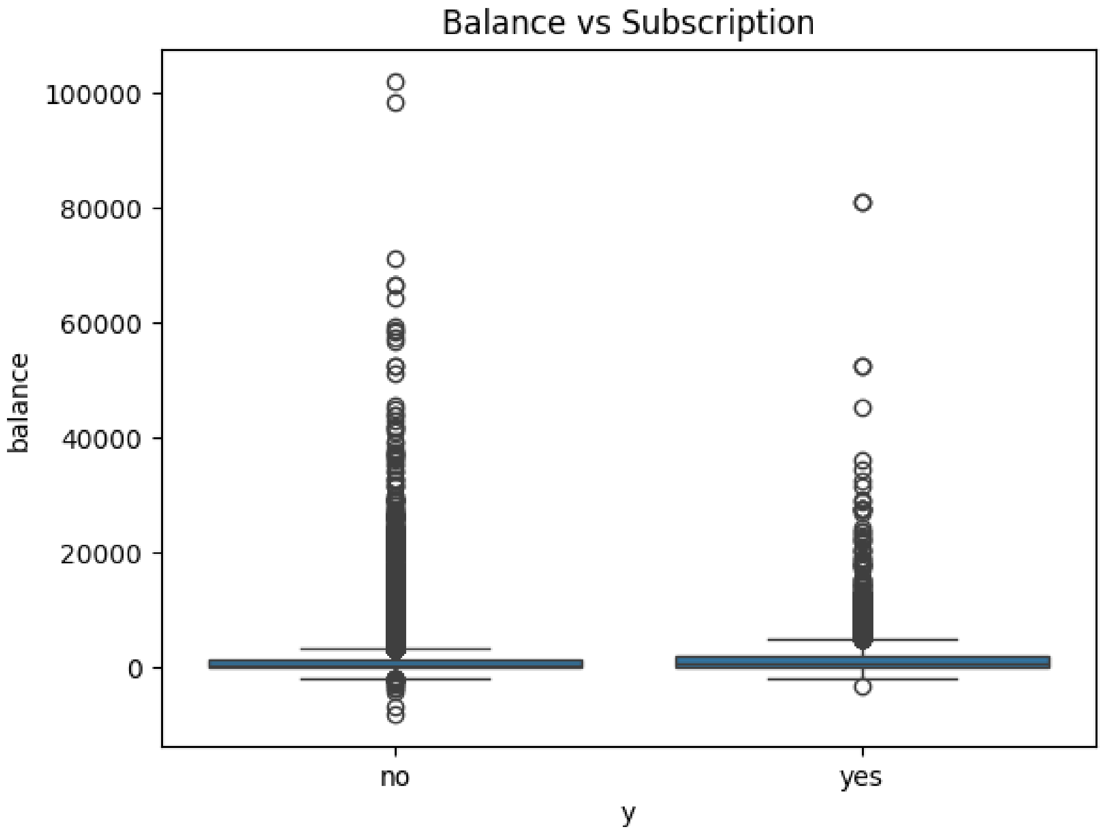


[16]:

```
sns.histplot(df['balance'], bins=30, kde=True)
plt.title('Client Balance Distribution')
plt.show()
```



```
[17]: sns.boxplot(x='y', y='balance', data=df)
      plt.title('Balance vs Subscription')
      plt.show()
```


Balance vs Subscription
balance
100000
80000
60000
40000
20000
0
no
yes
y

```
[ ]:
```