PGPDSE FT Project – Finance

Analysis on bank Marketing

Submitted By: Gayatri Jitendra Patil

Aim:

You aim to Determinate/Analysis factors for the subscription and no subscription. Using the ITP and NPV techniques find the below questions.

PROBLEM STATEMENT:

The problem is that the Bank Marketing campaigns of a Portuguese banking institution need to identify the factors that cause the customers to tend to take the subscription, as well as Bank Marketing campaigns of a Portuguese banking institution need to identify the reasons behind the customer which make them not take the subscription.

Industry Review:

There are two main approaches for enterprises to promote products and/or services: through mass campaigns, targeting general indiscriminate public or directed marketing, targeting a specific set of contacts (Ling and Li 1998). Nowadays, in a global competitive world, positive responses to mass campaigns are typically very low, less than 1%, according to the same study. Alternatively, directed marketing focus on targets that assumable will be keener to that specific product/service, making this kind of campaigns more attractive due to its efficiency (Ou et al. 2003). Nevertheless, directed marketing has some drawbacks, for instance it may trigger a negative attitude towards banks due to the intrusion of privacy (Page and Luding 2003). It should be stressed that due to internal competition and current financial crisis, there are huge pressures for European banks to increase a financial asset. To solve this issue, one adopted strategy is offer attractive long-term deposit applications with good interest rates, in particular by using directed marketing campaigns. Also, the same drivers are pressing for a reduction in costs and time. Thus, there is a need for an improvement in efficiency: lesser contacts should be done, but an approximately number of successes (clients subscribing the deposit) should be kept.

Literature Survey:

Since profitability and efficiency are traditional measures of performance, several authors (e.g., Aiello and Bonanno 2016; Athanasoglou et al. 2008; Defung et al. 2016; Dietrich and Wanzenried 2011; Ding et al. 2017; Djalilov and Piesse 2016; García-Herrero et al. 2009; Guru et al. 2002; Knezevic and Dobromirov 2016; Sufian and Kamarudin 2015) concluded that the performance determinants of Banking institutions can be divided into two categories—internal and external—and can be subdivided into industry-specific factors and macroeconomic

factors. The first category concerns factors that are specific to banks; that is, they are variables that are controlled by their management, which reflect the different management policies and decisions and, consequently, dictate their performance (Djalilov and Piesse 2016; Guru et al. 2002). External determinants are factors that derive from the country's economic and legal environment and, therefore, nothing related to how the bank is being managed (Dietrich and Wanzenried 2011; Ding et al. 2017; Djalilov and Piesse 2016). Thus, internal factors are part of, among others, bank asset structure, bank asset quality, bank capital, bank operational efficiency, bank revenue diversification, bank annual deposit growth, and bank size. In the same way, among the external factors, we can highlight the property, the fact that the banks are quoted or not, inflation, or economic growth. These variables are those used in our estimation model, and to that extent, it is precisely on these that the literature review will focus.

Dataset and Domain:

DATA DICTIONARY:

1 - age (numeric)

2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur", "student","bluecollar","selfemployed","retired","technician","services")

3 - marital : marital status (categorical: "married", "divorced", "single" ; note: "divorced" means divorced or widowed)

4 - education (categorical: "unknown", "secondary", "primary", "tertiary")

5 - Default: has credit in default? (binary: "yes", "no")

6 - balance: average yearly balance, in euros (numeric)

7 - Housing: has a housing loan? (binary: "yes", "no")

8 - Loan: has personal loan? (binary: "yes", "no")

- **related to the last contact of the current campaign:<\b>**

9 - contact: contact communication type (categorical: "unknown", "telephone", "cellular")

10 - day: last contact day of the month (numeric)
11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - duration: last contact duration, in seconds (numeric)

- <b>other attributes :< /b>

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes the last contact

14 - P-days: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical:
"unknown", "other", "failure", "success")

- **output variable (desired target):**

17 - y - has the client subscribed to a term deposit? (binary: "yes", "no")

DATASET DESCRIPTION:

Bank Marketing: The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required to assess if the product (bank term deposit) would be (or not) subscribed.

BUSINESS IMPORTANCE:

Finance is important in strengthening business finance because it helps companies take risks and grow. Businesses could do what they wanted without any financial support in the past. But now, with the increased use of technology and globalization, businesses are becoming more reliant on money to accomplish their goals. Finance is also important in strengthening business finance because it allows companies to take risks and grow.

With the increased use of technology, people are also becoming more reliant on money to accomplish their goals. Finance is important in strengthening because it allows companies to take risks and grow.

Financial statements are a great way to monitor the performance of a company. They provide information about the company's financial standing and how they are doing financially.In order to run a successful business, one requires money for numerous goals, such as:

- To acquire new machinery/equipment
- Recruiting and training staff
- Increasing the company's commercial opportunities
- To keep inventory increases

Business Questions Discussion:

Check Point 1:

1. Import data sets and Load data set

Hint: Make use of with statement and write a function to load the data if you are not able to load through the panda's method.

```
In [3]:  1  import numpy as np
         2  import pandas as pd
         3  import csv
         4
         5  df=pd.read_csv('bank1.csv',sep=';')
         6  df.head()
```

Out[3]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y	Unnamed 1
0	30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no	NaN
1	33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no	NaN
2	35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no	NaN
3	30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no	NaN
4	59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no	NaN

2. Make the data proper to make use of data for analysis

    A.   Identify the Features data types before entering into the analysis

```
In [2]:  1  df.drop( 'Unnamed: 17',axis=1,inplace=True)
```

```
In [4]:  1  df['poutcome'] = df['poutcome'].replace('unknown', np.nan)
         2  for i in df.columns:
         3      df[i]=df[i].replace('unknown', np.nan)
```

A. Identify the Features data types before entering into the analysis

```
In [5]:  1  df.info()
              1   job        4483 non-null   object
              2   marital    4521 non-null   object
              3   education  4334 non-null   object
              4   default    4521 non-null   object
              5   balance    4521 non-null   int64
              6   housing    4521 non-null   object
              7   loan       4521 non-null   object
              8   contact    3197 non-null   object
              9   day        4521 non-null   int64
              10  month      4521 non-null   object
              11  duration   4521 non-null   int64
              12  campaign   4521 non-null   int64
              13  pdays      4521 non-null   int64
              14  previous   4521 non-null   int64
              15  poutcome   816 non-null    object
              16  y          4521 non-null   object
              17  Unnamed: 17  0 non-null    float64
         dtypes: float64(1), int64(7), object(10)
         memory usage: 635.9+ KB
```

    B.   Convert the data types which are wrongly identified according to the business (domain).
         Kindly use the User Defined function and loop to convert the data types once.
  Answer: - data is already according identified according to the business (domain).
    C.   Find and Remove missing if any. Use visualization to find the missing values or Use general method to find the missing values.

    D.   Find duplicates (if necessary)
         Answer: - there is no duplicate row in dataset

```
In [7]:  1 df[df.duplicated()]
```

Out[7]:

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y	Unnamed: 17

```
In [8]:  1 df.shape
```

Out[8]: (4521, 18)

```
In [9]:  1 df.isna().sum()
```

Out[9]:
```
age              0
job             38
marital          0
education      187
default          0
balance          0
housing          0
loan             0
contact       1324
day              0
month            0
duration         0
campaign         0
pdays            0
previous         0
poutcome      3705
y                0
Unnamed: 17   4521
dtype: int64
```

```
In [10]:  1 import matplotlib.pyplot as plt
          2 import seaborn as sns
          3 plt.figure(figsize=(10,6))
          4 sns.heatmap(df.isna().transpose(),
          5             cmap="YlGnBu",
          6             cbar_kws={'label': 'Missing Data'})
          7 # plt.savefig("heatmap1.jpg")
          8 plt.show()
          9
```
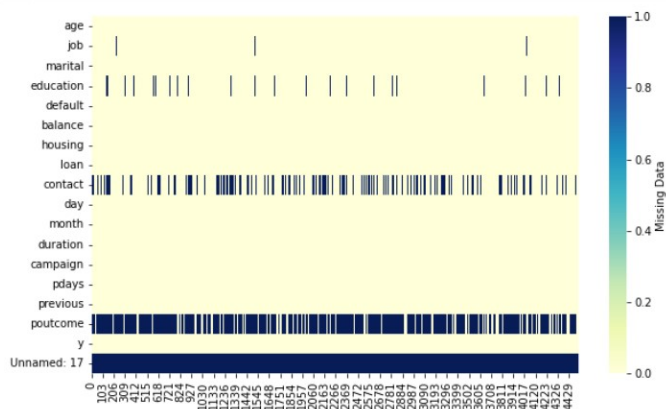

Check Point 2:

3. **Find the average balance of the customer who belongs to the subscribed customer and non-subscribed customer and also use a related plot to show them in visualization.**

```
In [11]:   1  df.groupby('y')['balance'].mean()
           2
```

```
Out[11]:  y
          no      1403.211750
          yes     1571.955854
          Name: balance, dtype: float64
```

```
In [12]:   1  # fig, axes = plt.subplots(figsize=(15,10),dpi = 500)
           2  df.groupby('y')['balance'].mean().plot(kind='bar')
           3  # plt.savefig("y vs balance bar.jpg")
           4  plt.show()
```
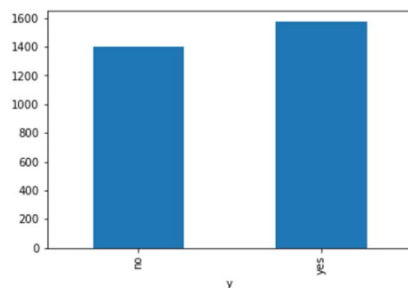

4. **Use a pie plot to find the distribution (frequency) of the education. Make sure to add labels and show the percentage of each education distribution.**

```
In [13]:   1  df['education'].value_counts()
```

```
Out[13]:  secondary    2306
          tertiary     1350
          primary       678
          Name: education, dtype: int64
```

```
In [14]:   1  fig, axes = plt.subplots(figsize=(15,5),dpi = 50)
           2  axes.pie(df['education'].value_counts(), autopct = '%.1f%%', radius = 1.2, labels = ['secondary','tertiary','primary'])
           3  # plt.savefig("pieplot1.jpg")
           4  plt.show()
```
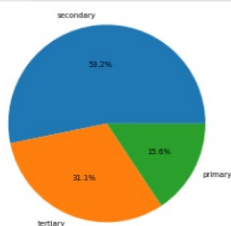

5. **Create a function that should be able to create a new feature (Variable) called season using the month column.**

```
In [15]:  1  df['month'].unique()
Out[15]: array(['oct', 'may', 'apr', 'jun', 'feb', 'aug', 'jan', 'jul', 'nov',
         'sep', 'mar', 'dec'], dtype=object)

In [16]:  1  condition= [(df['month']== 'mar'), (df['month']== 'apr'), (df['month']== 'may'),(df['month']== 'jun'),
          2          (df['month']== 'jul'),(df['month']== 'aug'),(df['month']== 'sep'),(df['month']== 'oct'),
          3          (df['month']== 'nov'),(df['month']== 'dec'),(df['month']== 'jan'),(df['month']== 'feb')]
          4  values=['summer','summer','summer','summer','rainy','rainy','rainy','rainy','winter','winter','winter','winter']
          5  df['season']=np.select(condition, values)
```
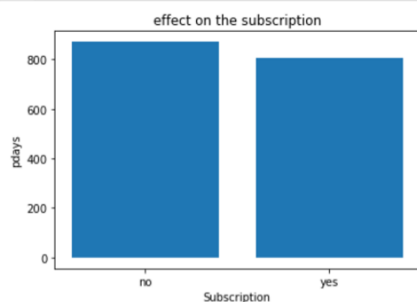
6. **Use the count plot with a variable that you created in the above question and also the Y variable to find the class distribution.**

```
In [15]:  1  df['month'].unique()
Out[15]: array(['oct', 'may', 'apr', 'jun', 'feb', 'aug', 'jan', 'jul', 'nov',
         'sep', 'mar', 'dec'], dtype=object)

In [16]:  1  condition= [(df['month']== 'mar'), (df['month']== 'apr'), (df['month']== 'may'),(df['month']== 'jun'),
          2          (df['month']== 'jul'),(df['month']== 'aug'),(df['month']== 'sep'),(df['month']== 'oct'),
          3          (df['month']== 'nov'),(df['month']== 'dec'),(df['month']== 'jan'),(df['month']== 'feb')]
          4  values=['summer','summer','summer','summer','rainy','rainy','rainy','rainy','winter','winter','winter','winter']
          5  df['season']=np.select(condition, values)
```

6. Use the count plot with a variable that you created in the above question and also the Y variable to find the class distribution.

```
In [54]:  1  sns.countplot(data=df,x='season',hue='y')
          2  plt.show()
```
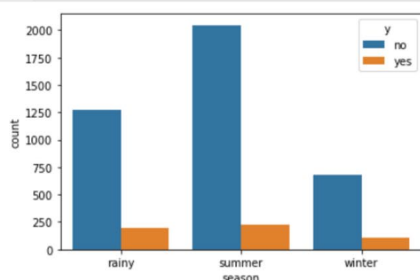
7. **Use the Pdays feature and find does it cause any effect on the subscription of the term using the bar plot.**

```
In [17]:  1
          2  plt.bar(df['y'],df['pdays'])
          3  plt.title(" effect on the subscription ")
          4  plt.xlabel("Subscription")
          5  plt.ylabel("pdays")
          6
          7  plt.show()
```

8. **Replace the -1 as Nan values for the P-days store.**

```
In [18]:    1  df['pdays'].fillna(-1)
            2  df['pdays'].value_counts().head(1)
Out[18]: -1    3705
         Name: pdays, dtype: int64
```

9. **Once you are done with question number 8, do the same analysis as question number 7. And observe the difference between question number 7 and question number 9.**

```
In [19]:    1  plt.bar(df['y'],df['pdays'])
            2  plt.title(" effect on the subscription ")
            3  plt.xlabel("Subscription")
            4  plt.ylabel("pdays")
            5
            6  plt.show()
```
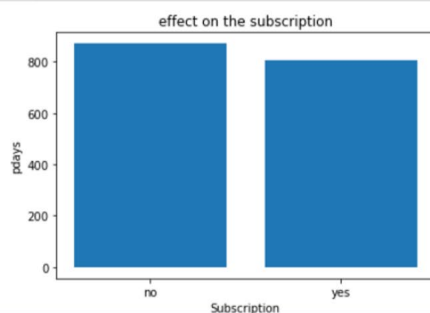

Inference:- there is no difference in both plot with and without null value

10. Does the customer take the term subscription who has less than 0 balance?
Hint: Use any kind of plot which would you the related information to this question.

```
In [72]:    1  # pd.crosstab(df['y'],df[df['balance']<0]['balance']).plot(kind='bar',legend=None)
            2  sns.barplot(y=df['y'],x=df[df['balance']<0]['balance'])
            3  plt.show()
```
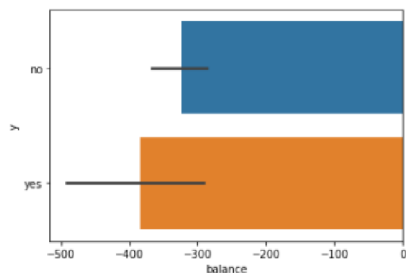
11. Use Pivot table to find the maximum balance for each type of job.

```
In [21]:   1  pivot1=pd.pivot_table(df,index='job',values='balance',aggfunc='max')
           2  pivot1
```

Out[21]:

job	balance
admin.	22171
blue-collar	16353
entrepreneur	42045
housemaid	26965
management	27359
retired	71188
self-employed	16430
services	26394
student	11555
technician	27733
unemployed	9019

12. Use the Age, balance, and Y column to plot the scatter plot and find what kind of relationship Age and balance had, and See the points which belong 0 and 1 class and how they are distributed.

```
In [22]:   1  sns.scatterplot(data=df,y='age',x='balance',hue='y',)
           2  plt.show()
```
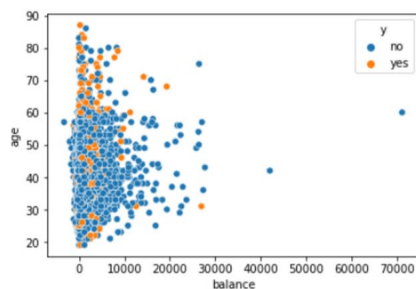

Whatever the age is. There are no any such relations but the balance maximum balance is 20000 And only subscription taken people are less than 10000

13. Use the violin plot and also the box plot to find the distribution of the balance for each class of the Y column. And try to tell why we have a Violin plot and Box plot both.

```
In [23]:  1  fig, axes = plt.subplots(1,2, figsize=(20, 10))
          2  fig.suptitle('violin plot and box plot for balance distribution of class in column y')
          3  sns.violinplot(ax=axes[0], data=df, x='y', y='balance')
          4  sns.boxplot(ax=axes[1], data=df, x='y', y='balance')
          5  plt.show()
```
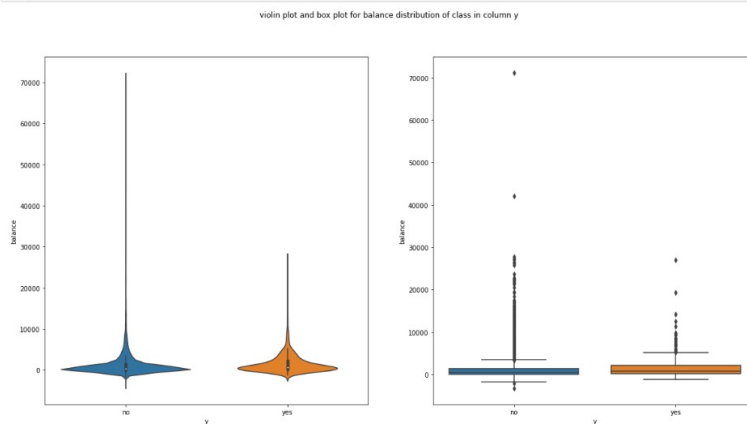

A violin plot is more informative than a plain box plot. While a box plot only shows summary statistics such as mean/median and interquartile ranges, the violin plot shows the full distribution of the data. The difference is particularly useful when the data distribution is multimodal (more than one peak). In this case a violin plot shows the presence of different peaks, their position and relative amplitude

14.   Use a pie plot to know the Proportion (distribution) of the defaulters and non-defaulters.
Note: Try to explore more parameters that are there in the pie-plot method.

```
In [24]:  1  plt.pie(df['default'].value_counts(), autopct = '%.1f%%', radius = 1.2, labels = df['default'].unique(),
          2          explode = [0,0.7],shadow=True)
          3  plt.show()
```
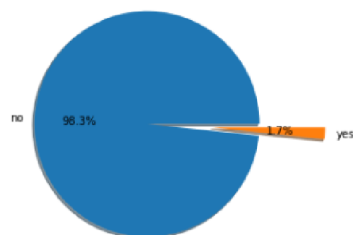

15.   Use Box plot and strip plot to know the distribution of the P-days with respect to Y
classes and differentiate both plots.

```
In [25]:  1  fig, axes = plt.subplots(1,2, figsize=(20, 10))
          2
          3  fig.suptitle('box plot and strip plot for distribution of the pdays with subcription')
          4
          5  sns.boxplot(ax=axes[0], data=df, x='y', y='pdays')
          6  sns.stripplot(ax=axes[1], data=df, x='y', y='pdays')
          7  plt.show()
```
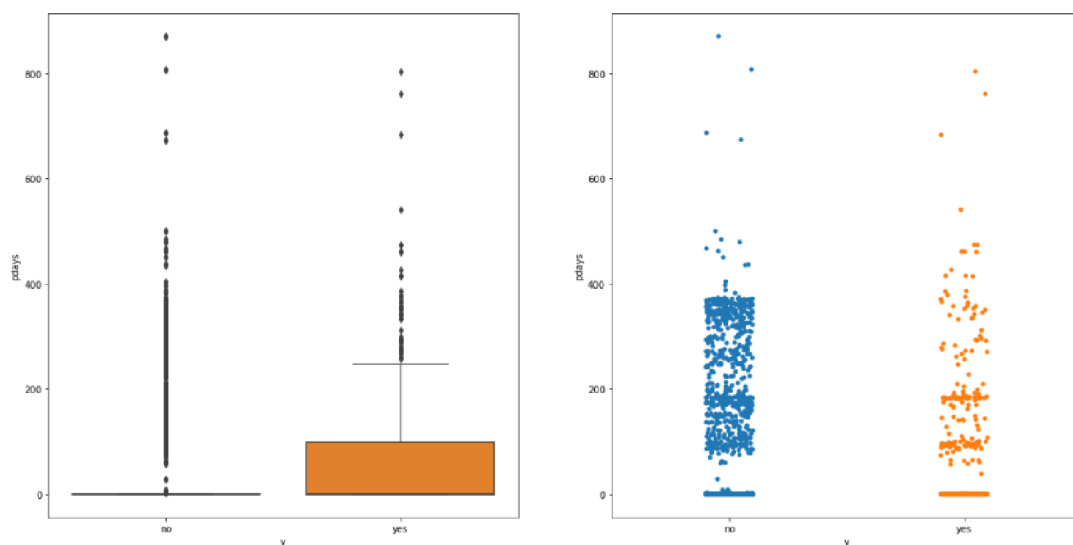

The strip plots show all the data at that level of theory. The box plots show the extrema (whisker tails), interquartile range (box boundaries), and median (horizontal line). The violin plots show the probability density of the data

Business Questions Discussion (Identified by the students):

### 16. Correlation between all numerical columns

```
In [25]:  1  cor1=df.corr()
          2  cor1
```

Out[25]:

	age	balance	day	duration	campaign	pdays	previous
age	1.000000	0.083820	-0.017853	-0.002367	-0.005148	-0.008894	-0.003511
balance	0.083820	1.000000	-0.008677	-0.015950	-0.009976	0.009437	0.026196
day	-0.017853	-0.008677	1.000000	-0.024629	0.160706	-0.094352	-0.059114
duration	-0.002367	-0.015950	-0.024629	1.000000	-0.068382	0.010380	0.018080
campaign	-0.005148	-0.009976	0.160706	-0.068382	1.000000	-0.093137	-0.067833
pdays	-0.008894	0.009437	-0.094352	0.010380	-0.093137	1.000000	0.577562
previous	-0.003511	0.026196	-0.059114	0.018080	-0.067833	0.577562	1.000000

17. Count of deposit subscribe by people
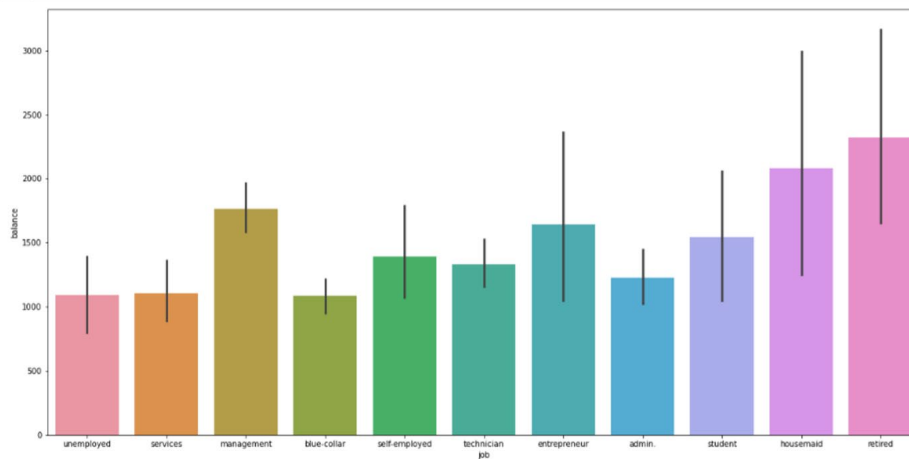
```
In [27]:    1  sns.countplot(x='y', data=df)
            2  plt.show()
```
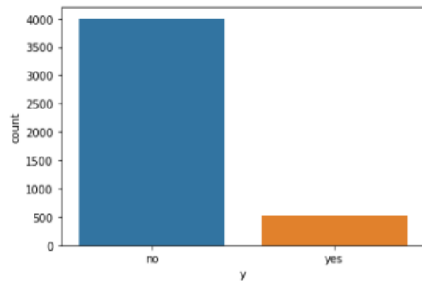

18. Visualize the bar plot of 'job' and 'balance'

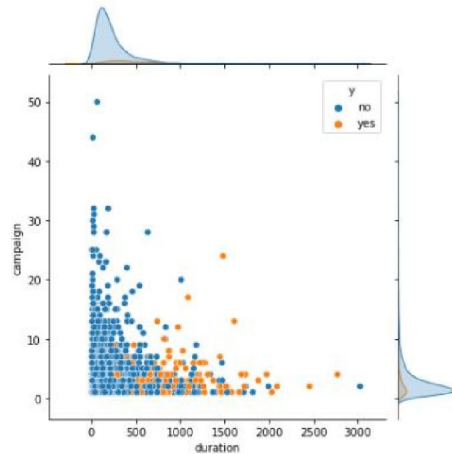18. Visualize the barplot of 'job' and 'balance'

```
In [62]:    1  plt.figure(figsize=(20,10))
            2  sns.barplot(data=df,y='balance',x='job')
            3  plt.show()
```
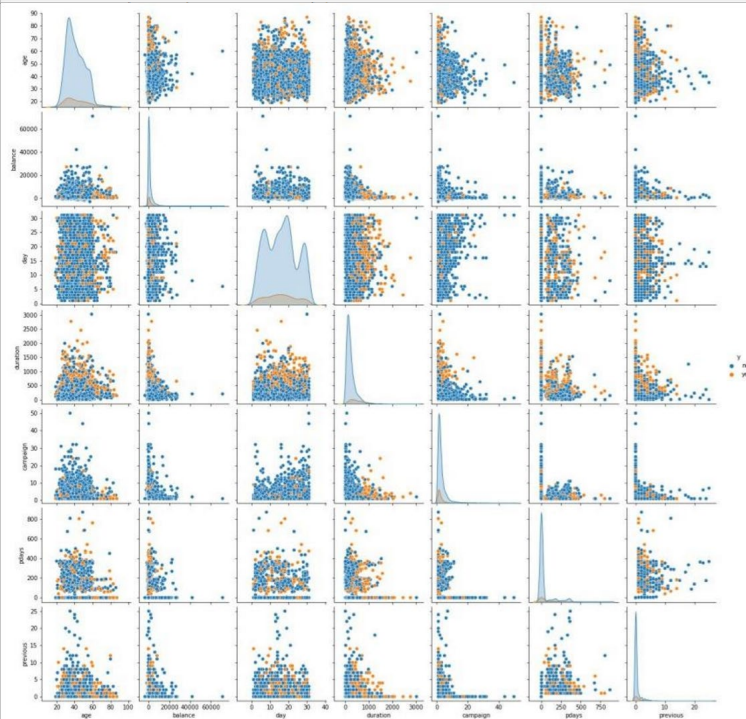

19. Visualize the relationship between 'duration' & 'campaign': with response result

```
In [50]:   1  sns.jointplot(x='duration',y='campaign',data=df,hue='y')
           2  plt.show()
```


20. Relation of all numeric columns with each other on y column

```
In [74]:   1  plt.figure(figsize=(15,8),dpi=500)
           2  sns.pairplot(df,hue='y')
           3  plt.savefig('pair.jpg')
           4  plt.show()
```

Conclusion:-

Write the lessons learned:-

Banks play very important roles in the economic development of nations as they, too a large extent, wield control over the supply of money in circulation and are the main stimuli of economic progress.

So we have learnt from this mini project that is how to prepare data inspection and cleaning

Inform then we have done so many questions about figure out the subscription loan bank account details and where it is male or female were the bank account is negative or positive and from this analytics we have use a many type of plot like bar plot ,pie plot ,scratch plot ,box plot and so on

We Understand that how the bank work with their customer and how they analytics there data

Skills Used:-

We have to use Jupyter notebook to develop this project

We have to use numpy, pandas built in functions and data manipulation skills

And we have to develop the skills to know the inferences of plots

There are intense use of matplotlib and seaborn library to plot the graphs

Domain understanding developed:-

The campaigns are useful to improve banking business

Communications are very useful to grow the relation with bank and account holders