Bank Marketing Data Analysis:

The code is all the way at the bottom. (Help from ChatGPT was used and collaborated with Steve Choi)

For the data set: Import to python was used from the site the library is listed in bold in the list of dependencies

1) Importing Data and Libraries:

Dependeicies:
/opt/homebrew/bin/python3 -m pip install pandas\n
/opt/homebrew/bin/python3 -m pip install matplotlib\n
/opt/homebrew/bin/python3 -m pip install seaborn\n
/**opt/homebrew/bin/python3 -m pip install ucimlrepo**

```
rigpeawangchuk@Rigpeas-MacBook-Pro ~ % /opt/homebrew/bin/python3 /Users/rigpeawangchuk/Desktop/try.py
{'uci_id': 222, 'name': 'Bank Marketing', 'repository_url': 'https://archive.ics.uci.edu/dataset/222/bank+marketing', 'data_url': 'https://archive.ics.uci.edu/static/public/222
/data.csv', 'abstract': 'The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the cli
ent will subscribe a term deposit (variable y).', 'area': 'Business', 'tasks': ['Classification'], 'characteristics': ['Multivariate'], 'num_instances': 45211, 'num_features':
16, 'feature_types': ['Categorical', 'Integer'], 'demographics': ['Age', 'Occupation', 'Marital Status', 'Education Level'], 'target_col': ['y'], 'index_col': None, 'has_missin
g_values': 'yes', 'missing_values_symbol': 'NaN', 'year_of_dataset_creation': 2014, 'last_updated': 'Fri Aug 18 2023', 'dataset_doi': '10.24432/C5K306', 'creators': ['S. Moro',
 'P. Rita', 'P. Cortez'], 'intro_paper': {'title': 'A data-driven approach to predict the success of bank telemarketing', 'authors': 'Sérgio Moro, P. Cortez, P. Rita', 'publish
ed_in': 'Decision Support Systems', 'year': 2014, 'url': 'https://www.semanticscholar.org/paper/cab86052882d126d43f72108c6cb41b295cc8a9e', 'doi': '10.1016/j.dss.2014.03.001'},
'additional_info': {'summary': "The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Ofte
n, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. \n\nThere are four da
tasets: \n1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al.
, 2014]\n2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.\n3) bank-full.csv with all examples and 17 inputs, ordered by date (o
lder version of this dataset with less inputs). \n4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).
 \nThe smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM). \n\nThe classification goal is to predict if the client wi
ll subscribe (yes/no) a term deposit (variable y).", 'purpose': None, 'funded_by': None, 'instances_represent': None, 'recommended_data_splits': None, 'sensitive_data': None, '
preprocessing_description': None, 'variable_info': 'Input variables:\n   # bank client data:\n   1 - age (numeric)\n   2 - job : type of job (categorical: "admin.","unknown","u
nemployed","management","housemaid","entrepreneur","student",\n                                                                    "blue-collar","self-employed","retired","technician","services") \n   3 -
marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)\n   4 - education (categorical: "unknown","secondary","primary
","tertiary")\n   5 - default: has credit in default? (binary: "yes","no")\n   6 - balance: average yearly balance, in euros (numeric) \n   7 - housing: has housing loan? (bina
ry: "yes","no")\n   8 - loan: has personal loan? (binary: "yes","no")\n   # related with the last contact of the current campaign:\n   9 - contact: contact communication type (
categorical: "unknown","telephone","cellular") \n  10 - day: last contact day of the month (numeric)\n  11 - month: last contact month of year (categorical: "jan", "feb", "mar"
, ..., "nov", "dec")\n  12 - duration: last contact duration, in seconds (numeric)\n   # other attributes:\n  13 - campaign: number of contacts performed during this campaign a
nd for this client (numeric, includes last contact)\n  14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means
 client was not previously contacted)\n  15 - previous: number of contacts performed before this campaign and for this client (numeric)\n  16 - poutcome: outcome of the previou
s marketing campaign (categorical: "unknown","other","failure","success")\n\n  Output variable (desired target):\n  17 - y - has the client subscribed a term deposit? (binary:
"yes","no")\n', 'citation': None}}
          name     role         type      demographic                                                description  units missing_values
0          age  Feature      Integer              Age                                                       None   None             no
1          job  Feature  Categorical       Occupation  type of job (categorical: 'admin.','blue-colla...   None             no
2      marital  Feature  Categorical   Marital Status  marital status (categorical: 'divorced','marri...   None             no
3    education  Feature  Categorical  Education Level  (categorical: 'basic.4y','basic.6y','basic.9y'...   None             no
4      default  Feature       Binary             None                                has credit in default?   None             no
5      balance  Feature      Integer             None                                average yearly balance  euros             no
6      housing  Feature       Binary             None                                   has housing loan?   None             no
7         loan  Feature       Binary             None                                   has personal loan?  None             no
8      contact  Feature  Categorical             None  contact communication type (categorical: 'cell...   None            yes
9  day_of_week  Feature         Date             None                         last contact day of the week  None             no
10       month  Feature         Date             None  last contact month of year (categorical: 'jan'...   None             no
11    duration  Feature      Integer             None  last contact duration, in seconds (numeric). ...   None             no
12    campaign  Feature      Integer             None  number of contacts performed during this campa...   None             no
13       pdays  Feature      Integer             None  number of days that passed by after the client...   None            yes
14    previous  Feature      Integer             None  number of contacts performed before this campa...   None             no
15    poutcome  Feature  Categorical             None  outcome of the previous marketing campaign (ca...   None            yes
16           y   Target       Binary             None                has the client subscribed a term deposit?   None             no
```

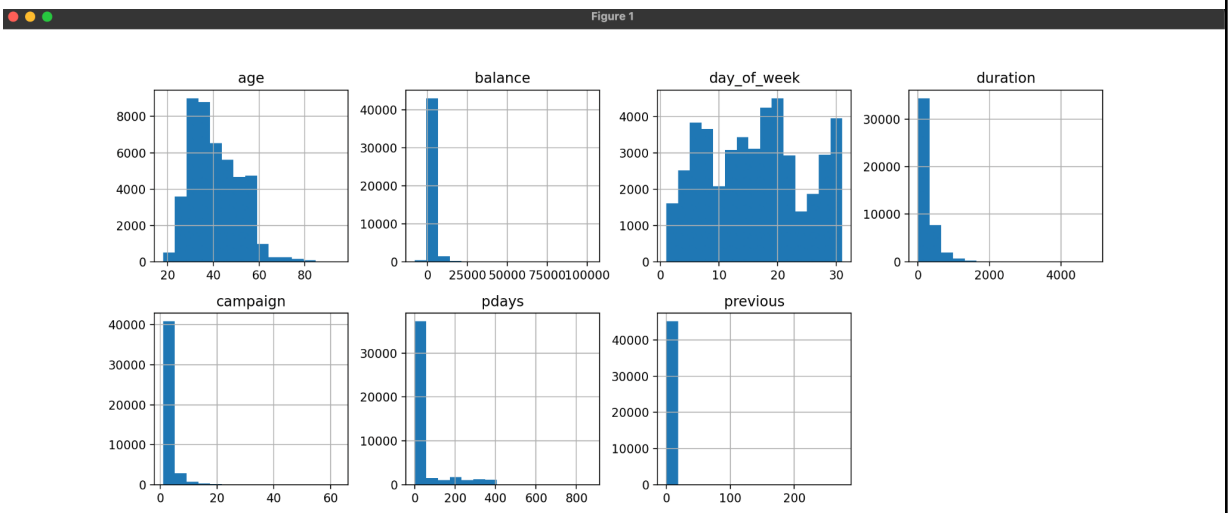
2) Exploratory Data Analysis (EDA)

a) Exploring the data set:

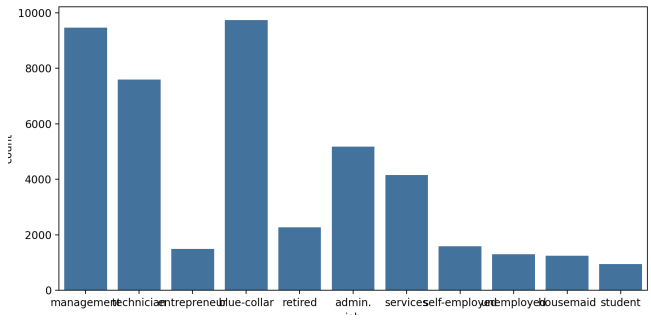
```
                age        balance   day_of_week      duration      campaign         pdays      previous
count  45211.000000   45211.000000  45211.000000  45211.000000  45211.000000  45211.000000  45211.000000
mean      40.936210    1362.272058     15.806419    258.163080      2.763841     40.197828      0.580323
std       10.618762    3044.765829      8.322476    257.527812      3.098021    100.128746      2.303441
min       18.000000   -8019.000000      1.000000      0.000000      1.000000     -1.000000      0.000000
25%       33.000000      72.000000      8.000000    103.000000      1.000000     -1.000000      0.000000
50%       39.000000     448.000000     16.000000    180.000000      2.000000     -1.000000      0.000000
75%       48.000000    1428.000000     21.000000    319.000000      3.000000     -1.000000      0.000000
max       95.000000  102127.000000     31.000000   4918.000000     63.000000    871.000000    275.000000

age              0
job            288
marital          0
education     1857
default          0
balance          0
housing          0
loan             0
contact      13020
day_of_week      0
month            0
duration         0
campaign         0
pdays            0
previous         0
poutcome     36959
dtype: int64

age             int64
job            object
marital        object
education      object
default        object
balance         int64
housing        object
loan           object
contact        object
day_of_week     int64
month          object
duration        int64
campaign        int64
pdays           int64
previous        int64
poutcome       object
dtype: object
```

b) Graphing relationships

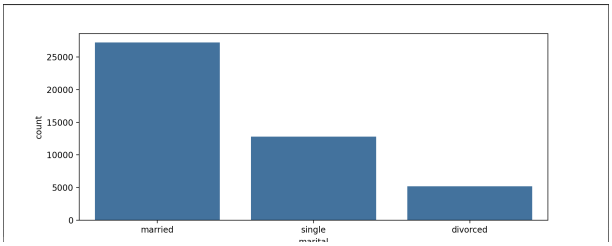
Numeric data:



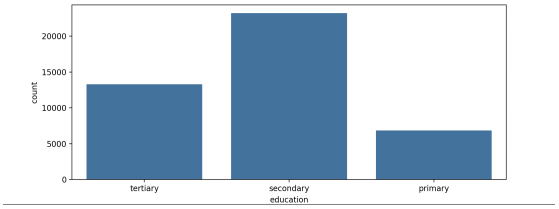
Non-numeric data:
-Type of person's income(student, retire, etc)



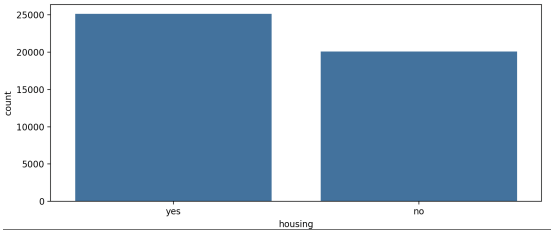
-Marital status



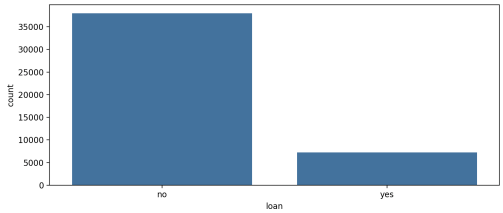
-Education level



-Housing



Loat:



There are also other other non-numeric data in in the database however, for the sake of our assignment these examples suffice.

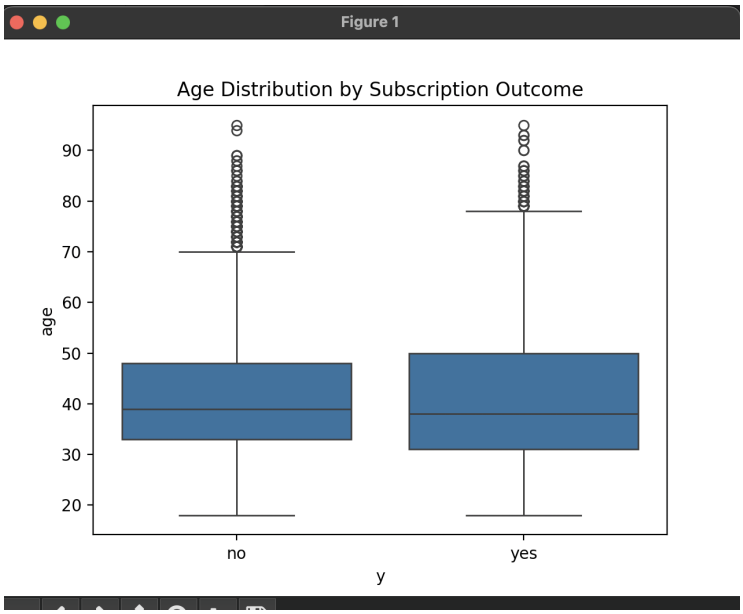
Hypothesis Formulation

Judging from the data set derived in the above snippet:

1) “There is a correlation between the age of clients and the likelihood of subscribing to a term deposit.”
2) “The duration of the last contact with a client has an impact on the likelihood of them subscribing to a term deposit.”

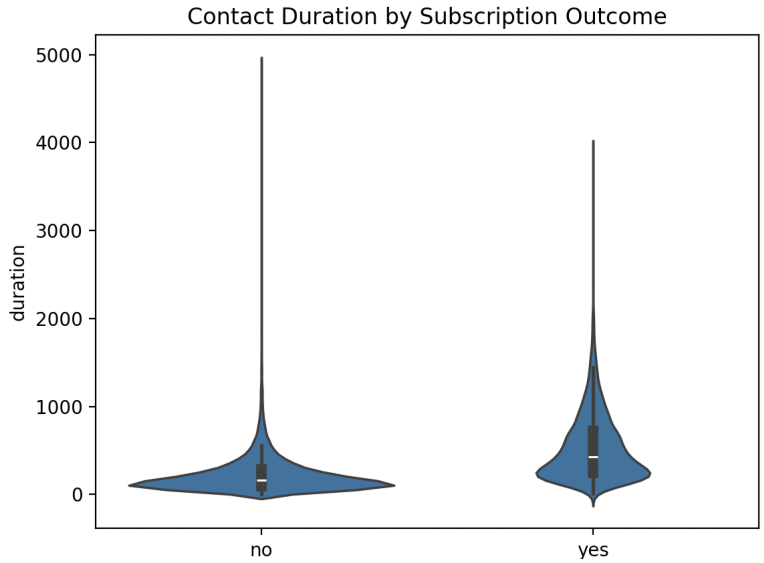
Hypothesis Testing

Hypothesis 1:



Explanation: We see that the boxes are quiet similar therefore there seems to be little to no correlation. For a simplistic analysis, this graph disproves our hypothesis.

Hypothesis 2:



Shorter calls seem to be less likely to result in a subscription. Which supports our second hypothesis. mne

```
# Importing libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from ucimlrepo import fetch_ucirepo
```

```
# fetch dataset
bank_marketing = fetch_ucirepo(id=222)
 # data (as pandas dataframes)
X = bank_marketing.data.features
y = bank_marketing.data.targets
 # metadata
print(bank_marketing.metadata)
 # variable information
print(bank_marketing.variables)

#Describe dataset

print("Data analysis:")
print("Head:")
print()
print(X.head())
print()
print(X.describe())
print()
print(X.isnull().sum())
print()
print(X.dtypes)
print()

#Seperate in to numerical and qulatative data sets
numeric_columns = X.select_dtypes(include=['int64', 'float64']).columns
X[numeric_columns].hist(bins=15, figsize=(15, 6), layout=(2, 4))
plt.show()
categorical_columns = X.select_dtypes(include=['object']).columns
for col in categorical_columns:
   plt.figure(figsize=(10,4))
   sns.countplot(x=col, data=X)
   plt.show()

y = y.squeeze()

X['age'] = X['age'].squeeze()
X['duration'] = X['duration'].squeeze()

#Hypo 1 testing
sns.boxplot(x=y, y=X['age'])
plt.title('Age Distribution by Subscription Outcome')
plt.show()

#Hypo 2 testing
sns.violinplot(x=y, y=X['duration'])
plt.title('Contact Duration by Subscription Outcome')
plt.show()
```