Assignment 4.1

September 28, 2024

[1]:
```
import pandas as pd
import numpy as np

df = pd.read_csv(
    "/Users/gabrielmancillas/Desktop/ADS 505-01/Mod 04/Assignment 4.1/
 ↪Fundraising.csv"
)
df.head()
```

[1]:

	Row Id	Row Id.	zipconvert_2	zipconvert_3	zipconvert_4	zipconvert_5	\
0	1	17	0	1	0	0	
1	2	25	1	0	0	0	
2	3	29	0	0	0	1	
3	4	38	0	0	0	1	
4	5	40	0	1	0	0	

	homeowner dummy	NUMCHLD	INCOME	gender dummy	…	IC15	NUMPROM	\
0	1	1	5	1	…	1	74	
1	1	1	1	0	…	4	46	
2	0	2	5	1	…	13	32	
3	1	1	3	0	…	4	94	
4	1	1	4	0	…	7	20	

	RAMNTALL	MAXRAMNT	LASTGIFT	totalmonths	TIMELAG	AVGGIFT	TARGET_B	\
0	102.0	6.0	5.0	29	3	4.857143	1	
1	94.0	12.0	12.0	34	6	9.400000	1	
2	30.0	10.0	5.0	29	7	4.285714	1	
3	177.0	10.0	8.0	30	3	7.080000	0	
4	23.0	11.0	11.0	30	6	7.666667	0	

	TARGET_D
0	5.0
1	10.0
2	5.0
3	0.0
4	0.0

[5 rows x 24 columns]

[2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# 1. Understand the Data
print("First 5 rows of the dataset:")
print(df.head())

print("\nData types and missing values:")
print(df.info())

print("\nSummary statistics:")
print(df.describe())

# 2. Handle Missing Values
print("\nMissing values per column:")
print(df.isnull().sum())

# Fill missing values or drop columns/rows with missing values
# Example: Fill missing values with the mean
df.fillna(df.mean(), inplace=True)
```

```
First 5 rows of the dataset:
   Row Id  Row Id.  zipconvert_2  zipconvert_3  zipconvert_4  zipconvert_5  \
0       1       17             0             1             0             0
1       2       25             1             0             0             0
2       3       29             0             0             0             1
3       4       38             0             0             0             1
4       5       40             0             1             0             0

   homeowner dummy  NUMCHLD  INCOME  gender dummy  …  IC15  NUMPROM  \
0                1        1       5             1  …     1       74
1                1        1       1             0  …     4       46
2                0        2       5             1  …    13       32
3                1        1       3             0  …     4       94
4                1        1       4             0  …     7       20

   RAMNTALL  MAXRAMNT  LASTGIFT  totalmonths  TIMELAG   AVGGIFT  TARGET_B  \
0     102.0       6.0       5.0           29        3  4.857143         1
1      94.0      12.0      12.0           34        6  9.400000         1
2      30.0      10.0       5.0           29        7  4.285714         1
3     177.0      10.0       8.0           30        3  7.080000         0
4      23.0      11.0      11.0           30        6  7.666667         0

   TARGET_D
0       5.0
1      10.0
```

```
2       5.0
3       0.0
4       0.0

[5 rows x 24 columns]

Data types and missing values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3120 entries, 0 to 3119
Data columns (total 24 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Row Id           3120 non-null   int64
 1   Row Id.          3120 non-null   int64
 2   zipconvert_2     3120 non-null   int64
 3   zipconvert_3     3120 non-null   int64
 4   zipconvert_4     3120 non-null   int64
 5   zipconvert_5     3120 non-null   int64
 6   homeowner dummy  3120 non-null   int64
 7   NUMCHLD          3120 non-null   int64
 8   INCOME           3120 non-null   int64
 9   gender dummy     3120 non-null   int64
 10  WEALTH           3120 non-null   int64
 11  HV               3120 non-null   int64
 12  Icmed            3120 non-null   int64
 13  Icavg            3120 non-null   int64
 14  IC15             3120 non-null   int64
 15  NUMPROM          3120 non-null   int64
 16  RAMNTALL         3120 non-null   float64
 17  MAXRAMNT         3120 non-null   float64
 18  LASTGIFT         3120 non-null   float64
 19  totalmonths      3120 non-null   int64
 20  TIMELAG          3120 non-null   int64
 21  AVGGIFT          3120 non-null   float64
 22  TARGET_B         3120 non-null   int64
 23  TARGET_D         3120 non-null   float64
dtypes: float64(5), int64(19)
memory usage: 585.1 KB
None

Summary statistics:
            Row Id       Row Id.  zipconvert_2  zipconvert_3  zipconvert_4  \
count  3120.000000   3120.000000   3120.000000   3120.000000   3120.000000
mean   1560.500000  11615.770833      0.214423      0.185256      0.214423
std     900.810746   6698.678131      0.410487      0.388568      0.410487
min       1.000000     17.000000      0.000000      0.000000      0.000000
25%     780.750000   5820.750000      0.000000      0.000000      0.000000
50%    1560.500000  11735.500000      0.000000      0.000000      0.000000
```

```
75%    2340.250000  17435.750000      0.000000      0.000000      0.000000
max    3120.000000  23293.000000      1.000000      1.000000      1.000000

       zipconvert_5  homeowner dummy      NUMCHLD       INCOME  gender dummy  \
count   3120.000000      3120.000000  3120.000000  3120.000000   3120.000000
mean       0.384615         0.770192     1.069231     3.893910      0.609295
std        0.486582         0.420777     0.347688     1.636186      0.487987
min        0.000000         0.000000     1.000000     1.000000      0.000000
25%        0.000000         1.000000     1.000000     3.000000      0.000000
50%        0.000000         1.000000     1.000000     4.000000      1.000000
75%        1.000000         1.000000     1.000000     5.000000      1.000000
max        1.000000         1.000000     5.000000     7.000000      1.000000

       …          IC15      NUMPROM     RAMNTALL     MAXRAMNT     LASTGIFT  \
count  …  3120.000000  3120.000000  3120.000000  3120.000000  3120.000000
mean   …    14.702885    49.089423   110.399875    16.651397    13.522917
std    …    12.079882    22.717130   147.299933    22.223521    10.581439
min    …     0.000000    11.000000    15.000000     5.000000     0.000000
25%    …     5.000000    29.000000    45.000000    10.000000     7.000000
50%    …    12.000000    48.000000    81.000000    15.000000    10.000000
75%    …    21.000000    65.000000   134.625000    20.000000    16.000000
max    …    90.000000   157.000000  5674.900000  1000.000000   219.000000

       totalmonths      TIMELAG      AVGGIFT    TARGET_B     TARGET_D
count  3120.000000  3120.000000  3120.000000  3120.00000  3120.000000
mean     31.136859     6.861859    10.690713     0.50000     6.499612
std       4.132952     5.561209     7.443980     0.50008    10.597849
min      17.000000     0.000000     2.138889     0.00000     0.000000
25%      29.000000     3.000000     6.356092     0.00000     0.000000
50%      31.000000     5.000000     9.000000     0.50000     0.500000
75%      34.000000     9.000000    12.811652     1.00000    10.000000
max      37.000000    77.000000   122.166667     1.00000   200.000000

[8 rows x 24 columns]

Missing values per column:
Row Id             0
Row Id.            0
zipconvert_2       0
zipconvert_3       0
zipconvert_4       0
zipconvert_5       0
homeowner dummy    0
NUMCHLD            0
INCOME             0
gender dummy       0
WEALTH             0
HV                 0
```

```
Icmed              0
Icavg              0
IC15               0
NUMPROM            0
RAMNTALL           0
MAXRAMNT           0
LASTGIFT           0
totalmonths        0
TIMELAG            0
AVGGIFT            0
TARGET_B           0
TARGET_D           0
dtype: int64
```

[3]:
```python
# 3. Data Types and Conversion
# Convert 'booking_created' to datetime if necessary
if 'booking_created' in df.columns:
    df['booking_created'] = pd.to_datetime(df['booking_created'])

# 4. Descriptive Statistics
print("\nUpdated summary statistics after handling missing values:")
print(df.describe())

# 5. Data Visualization
# Distribution of numerical features
df.hist(bins=30, figsize=(20, 15))
plt.show()

# Correlation heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

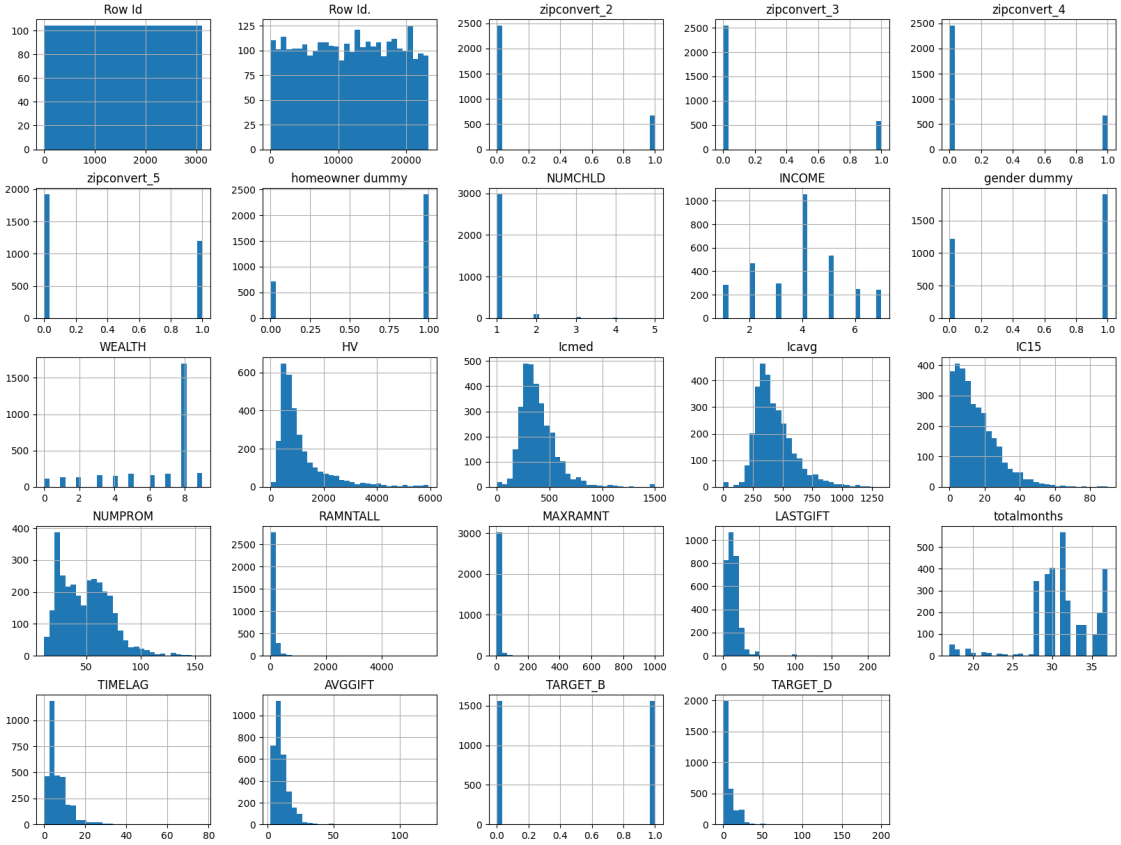
```
Updated summary statistics after handling missing values:
            Row Id       Row Id.  zipconvert_2  zipconvert_3  zipconvert_4  \
count  3120.000000   3120.000000   3120.000000   3120.000000   3120.000000
mean   1560.500000  11615.770833      0.214423      0.185256      0.214423
std     900.810746   6698.678131      0.410487      0.388568      0.410487
min       1.000000     17.000000      0.000000      0.000000      0.000000
25%     780.750000   5820.750000      0.000000      0.000000      0.000000
50%    1560.500000  11735.500000      0.000000      0.000000      0.000000
75%    2340.250000  17435.750000      0.000000      0.000000      0.000000
max    3120.000000  23293.000000      1.000000      1.000000      1.000000

       zipconvert_5  homeowner dummy      NUMCHLD       INCOME  gender dummy  \
count   3120.000000      3120.000000  3120.000000  3120.000000   3120.000000
```

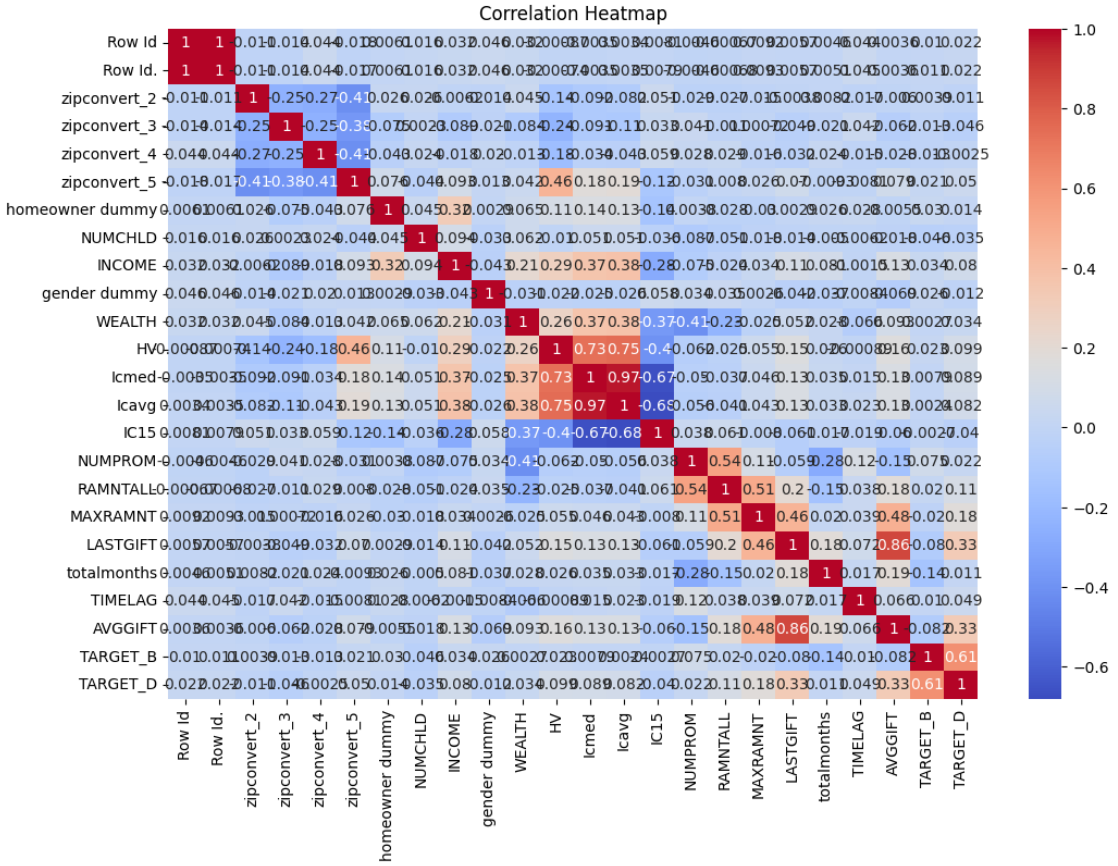
mean	0.384615	0.770192	1.069231	3.893910	0.609295
std	0.486582	0.420777	0.347688	1.636186	0.487987
min	0.000000	0.000000	1.000000	1.000000	0.000000
25%	0.000000	1.000000	1.000000	3.000000	0.000000
50%	0.000000	1.000000	1.000000	4.000000	1.000000
75%	1.000000	1.000000	1.000000	5.000000	1.000000
max	1.000000	1.000000	5.000000	7.000000	1.000000

	…	IC15	NUMPROM	RAMNTALL	MAXRAMNT	LASTGIFT	\
count	…	3120.000000	3120.000000	3120.000000	3120.000000	3120.000000	
mean	…	14.702885	49.089423	110.399875	16.651397	13.522917	
std	…	12.079882	22.717130	147.299933	22.223521	10.581439	
min	…	0.000000	11.000000	15.000000	5.000000	0.000000	
25%	…	5.000000	29.000000	45.000000	10.000000	7.000000	
50%	…	12.000000	48.000000	81.000000	15.000000	10.000000	
75%	…	21.000000	65.000000	134.625000	20.000000	16.000000	
max	…	90.000000	157.000000	5674.900000	1000.000000	219.000000	

	totalmonths	TIMELAG	AVGGIFT	TARGET_B	TARGET_D
count	3120.000000	3120.000000	3120.000000	3120.00000	3120.000000
mean	31.136859	6.861859	10.690713	0.50000	6.499612
std	4.132952	5.561209	7.443980	0.50008	10.597849
min	17.000000	0.000000	2.138889	0.00000	0.000000
25%	29.000000	3.000000	6.356092	0.00000	0.000000
50%	31.000000	5.000000	9.000000	0.50000	0.500000
75%	34.000000	9.000000	12.811652	1.00000	10.000000
max	37.000000	77.000000	122.166667	1.00000	200.000000

[8 rows x 24 columns]


Row Id
Row Id.
zipconvert_2
zipconvert_3
zipconvert_4
zipconvert_5
homeowner dummy
NUMCHLD
INCOME
gender dummy
WEALTH
HV
lcmed
lcavg
IC15
NUMPROM
RAMNTALL
MAXRAMNT
LASTGIFT
totalmonths
TIMELAG
AVGGIFT
TARGET_B
TARGET_D


Correlation Heatmap
Row Id
Row Id.
zipconvert_2
zipconvert_3
zipconvert_4
zipconvert_5
homeowner dummy
NUMCHLD
INCOME
gender dummy
WEALTH
HV
Icmed
Icavg
IC15
NUMPROM
RAMNTALL
MAXRAMNT
LASTGIFT
totalmonths
TIMELAG
AVGGIFT
TARGET_B
TARGET_D
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

[ ]: