

EDA PRESENTATION



```
[ ] data.dropna(inplace = True)
```

```
[ ] data.drop_duplicates(inplace = True)
```

WE START BY DROPPING NA VALUES
AND DUPLICATES

```
] data.head()
```

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

```
data.columns
```

```
Index(['age', 'job', 'marital', 'education', 'default', 'balance', 'housing',  
      'loan', 'contact', 'day', 'month', 'duration', 'campaign', 'pdays',  
      'previous', 'poutcome', 'y'],  
      dtype='object')
```

THIS IS HOW OUR DATA'S FIRST 5 ENTRY
LOOKS LIKE AND COLUMN NAMES

```
] data.corr()
```

| | 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 |

HERE ARE THE CORRELATION VALUES
BETWEEN COLUMNS

| | age | balance | day | duration | campaign | pdays | previous |
|-------|-------------|--------------|-------------|-------------|-------------|-------------|-------------|
| count | 4521.000000 | 4521.000000 | 4521.000000 | 4521.000000 | 4521.000000 | 4521.000000 | 4521.000000 |
| mean | 41.170095 | 1422.657819 | 15.915284 | 263.961292 | 2.793630 | 39.766645 | 0.542579 |
| std | 10.576211 | 3009.638142 | 8.247667 | 259.856633 | 3.109807 | 100.121124 | 1.693562 |
| min | 19.000000 | -3313.000000 | 1.000000 | 4.000000 | 1.000000 | -1.000000 | 0.000000 |
| 25% | 33.000000 | 69.000000 | 9.000000 | 104.000000 | 1.000000 | -1.000000 | 0.000000 |
| 50% | 39.000000 | 444.000000 | 16.000000 | 185.000000 | 2.000000 | -1.000000 | 0.000000 |
| 75% | 49.000000 | 1480.000000 | 21.000000 | 329.000000 | 3.000000 | -1.000000 | 0.000000 |
| max | 87.000000 | 71188.000000 | 31.000000 | 3025.000000 | 50.000000 | 871.000000 | 25.000000 |

HERE IS THE STATISTICAL VALUES OF
OUR DATA

- ▶ We will build a linear regression model since our columns are correlated and linear regression will fit good

