EDA

Adrian Taggruber

## Load Data

import pandas as pd  
dataset = pd.read\_csv("smoker.csv")

## Inspect structure

dataset.shape

(10000, 3)

dataset.info()

<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 10000 entries, 0 to 9999  
Data columns (total 3 columns):  
 # Column Non-Null Count Dtype  
--- ------ -------------- -----  
 0 smoker 10000 non-null int64  
 1 treatment 10000 non-null int64  
 2 outcome 10000 non-null int64  
dtypes: int64(3)  
memory usage: 234.5 KB

## Inspect Value

dataset.head()

|  | smoker | treatment | outcome |
| --- | --- | --- | --- |
| 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 |
| 2 | 2 | 1 | 5 |
| 3 | 3 | 1 | 67 |
| 4 | 4 | 0 | 8 |

dataset.tail()

|  | smoker | treatment | outcome |
| --- | --- | --- | --- |
| 9995 | 1 | 1 | 0 |
| 9996 | 0 | 0 | 1 |
| 9997 | 0 | 0 | 0 |
| 9998 | 1 | 1 | 0 |
| 9999 | 1 | 1 | 0 |

## Statistics

round(dataset.describe(), 2)

|  | smoker | treatment | outcome |
| --- | --- | --- | --- |
| count | 10000.00 | 10000.00 | 10000.00 |
| mean | 0.31 | 0.30 | 0.28 |
| std | 0.47 | 0.48 | 0.82 |
| min | 0.00 | 0.00 | 0.00 |
| 25% | 0.00 | 0.00 | 0.00 |
| 50% | 0.00 | 0.00 | 0.00 |
| 75% | 1.00 | 1.00 | 1.00 |
| max | 5.00 | 8.00 | 67.00 |

## Histogram

import matplotlib.pyplot as plt  
dataset["smoker"].hist()  
plt.show()

