data-filling-by-linear-regression

October 18, 2024

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

[ ]:
```
df=pd.read_csv("/content/bike_buyers_2.csv")
df
```

[ ]:

	Unnamed: 0.1	Unnamed: 0	ID	Marital Status	Gender	Income	Children	Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age	Purchased Bike
0	0	0	12496	Married	Female	40000.0	1	Bachelors	Skilled Manual	Yes	0	0-1 Miles	Europe	42.0	No
1	1	1	24107	Married	Male	30000.0	3	Partial College	Clerical	Yes	1	0-1 Miles	Europe	43.0	No
2	2	2	14177	Married	Male	80000.0	5	Partial College	Professional	No	2	2-5 Miles	Europe	60.0	No
3	3	3	24381	Single	Male	70000.0	0	Bachelors	Professional	Yes	1	5-10 Miles	Pacific	41.0	Yes
4	4	4	25597	Single	Male	30000.0	0	Bachelors	Clerical	No	0				
..	…	…	…	…	…	…	…	…	…	…	…				
995	995	995	23731	Married	Male	60000.0	2	High School	Professional	Yes	2				
996	996	996	28672	Single	Male	70000.0	4	Graduate Degree	Professional	Yes	0				
997	997	997	11809	Married	Male	60000.0	2	Bachelors	Skilled Manual	Yes	0				
998	998	998	19664	Single	Male	100000.0	3	Bachelors	Management	No	3				
999	999	999	12121	Single	Male	60000.0	3	High School	Professional	Yes	2				

```
4          0-1 Miles         Europe  36.0            Yes
..               …            …   …             …
995        2-5 Miles  North America  54.0            Yes
996        2-5 Miles  North America  35.0            Yes
997        0-1 Miles  North America  38.0            Yes
998        1-2 Miles  North America  38.0             No
999        10+ Miles  North America  53.0            Yes

[1000 rows x 15 columns]
```

[ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Unnamed: 0.1      1000 non-null   int64
 1   Unnamed: 0        1000 non-null   int64
 2   ID                1000 non-null   int64
 3   Marital Status    1000 non-null   object
 4   Gender            1000 non-null   object
 5   Income            1000 non-null   float64
 6   Children          1000 non-null   int64
 7   Education         1000 non-null   object
 8   Occupation        1000 non-null   object
 9   Home Owner        1000 non-null   object
 10  Cars              1000 non-null   int64
 11  Commute Distance  1000 non-null   object
 12  Region            1000 non-null   object
 13  Age               992 non-null    float64
 14  Purchased Bike    1000 non-null   object
dtypes: float64(2), int64(5), object(8)
memory usage: 117.3+ KB
```

[ ]:
```
df.describe()
```

[ ]:
```
       Unnamed: 0.1   Unnamed: 0            ID         Income     Children  \
count   1000.000000  1000.000000   1000.000000    1000.000000  1000.000000
mean     499.500000   499.500000  19965.992000   56267.605634     1.911000
std      288.819436   288.819436   5347.333948   30974.380206     1.620403
min        0.000000     0.000000  11000.000000   10000.000000     0.000000
25%      249.750000   249.750000  15290.750000   30000.000000     0.000000
50%      499.500000   499.500000  19744.000000   60000.000000     2.000000
75%      749.250000   749.250000  24470.750000   70000.000000     3.000000
max      999.000000   999.000000  29447.000000  170000.000000     5.000000
```

```
              Cars         Age
count  1000.000000  992.000000
mean      1.460000   44.181452
std       1.117877   11.362007
min       0.000000   25.000000
25%       1.000000   35.000000
50%       1.000000   43.000000
75%       2.000000   52.000000
max       4.000000   89.000000
```

[ ]:
```
df.head()
```

[ ]:
```
   Unnamed: 0.1  Unnamed: 0     ID Marital Status  Gender   Income  Children  \
0             0           0  12496        Married  Female  40000.0         1
1             1           1  24107        Married    Male  30000.0         3
2             2           2  14177        Married    Male  80000.0         5
3             3           3  24381         Single    Male  70000.0         0
4             4           4  25597         Single    Male  30000.0         0

         Education      Occupation Home Owner  Cars Commute Distance   Region  \
0        Bachelors  Skilled Manual        Yes     0        0-1 Miles   Europe
1  Partial College        Clerical        Yes     1        0-1 Miles   Europe
2  Partial College    Professional         No     2        2-5 Miles   Europe
3        Bachelors    Professional        Yes     1       5-10 Miles  Pacific
4        Bachelors        Clerical         No     0        0-1 Miles   Europe

    Age Purchased Bike
0  42.0             No
1  43.0             No
2  60.0             No
3  41.0            Yes
4  36.0            Yes
```

[ ]:
```
df.tail()
```

[ ]:
```
     Unnamed: 0.1  Unnamed: 0     ID Marital Status Gender    Income  \
995           995         995  23731        Married   Male   60000.0
996           996         996  28672         Single   Male   70000.0
997           997         997  11809        Married   Male   60000.0
998           998         998  19664         Single   Male  100000.0
999           999         999  12121         Single   Male   60000.0

     Children        Education      Occupation Home Owner  Cars  \
995         2      High School    Professional        Yes     2
996         4  Graduate Degree    Professional        Yes     0
997         2        Bachelors  Skilled Manual        Yes     0
998         3        Bachelors      Management         No     3
```

```
999         3      High School    Professional        Yes     2

    Commute Distance         Region   Age Purchased Bike
995        2-5 Miles  North America  54.0            Yes
996        2-5 Miles  North America  35.0            Yes
997        0-1 Miles  North America  38.0            Yes
998        1-2 Miles  North America  38.0             No
999        10+ Miles  North America  53.0            Yes
```

[ ]:
```
df.shape
```

[ ]:
```
(1000, 15)
```

[ ]:
```
df.isnull()
```

[ ]:
```
     Unnamed: 0.1  Unnamed: 0     ID  Marital Status  Gender  Income  \
0           False       False  False           False   False   False
1           False       False  False           False   False   False
2           False       False  False           False   False   False
3           False       False  False           False   False   False
4           False       False  False           False   False   False
..            …         …    …             …     …     …
995         False       False  False           False   False   False
996         False       False  False           False   False   False
997         False       False  False           False   False   False
998         False       False  False           False   False   False
999         False       False  False           False   False   False

     Children  Education  Occupation  Home Owner   Cars  Commute Distance  \
0       False      False       False       False  False             False
1       False      False       False       False  False             False
2       False      False       False       False  False             False
3       False      False       False       False  False             False
4       False      False       False       False  False             False
..        …        …         …         …    …               …
995     False      False       False       False  False             False
996     False      False       False       False  False             False
997     False      False       False       False  False             False
998     False      False       False       False  False             False
999     False      False       False       False  False             False

     Region    Age  Purchased Bike
0     False  False           False
1     False  False           False
2     False  False           False
3     False  False           False
4     False  False           False
```

```
..      …    …              …
995   False  False          False
996   False  False          False
997   False  False          False
998   False  False          False
999   False  False          False

[1000 rows x 15 columns]
```

[ ]:
```
df.isnull().sum()
```

[ ]:
```
Unnamed: 0.1        0
Unnamed: 0          0
ID                  0
Marital Status      0
Gender              0
Income              0
Children            0
Education           0
Occupation          0
Home Owner          0
Cars                0
Commute Distance    0
Region              0
Age                 8
Purchased Bike      0
dtype: int64
```

[ ]:
```
df.isnull().mean()
```

[ ]:
```
Unnamed: 0.1        0.000
Unnamed: 0          0.000
ID                  0.000
Marital Status      0.000
Gender              0.000
Income              0.000
Children            0.000
Education           0.000
Occupation          0.000
Home Owner          0.000
Cars                0.000
Commute Distance    0.000
Region              0.000
Age                 0.008
Purchased Bike      0.000
dtype: float64
```

[ ]:
```
df1=df.copy()
```

[ ]:
```
df1
```

[ ]:
```
     Unnamed: 0.1  Unnamed: 0     ID Marital Status  Gender    Income  \
0               0           0  12496        Married  Female   40000.0
1               1           1  24107        Married    Male   30000.0
2               2           2  14177        Married    Male   80000.0
3               3           3  24381         Single    Male   70000.0
4               4           4  25597         Single    Male   30000.0
..            ...         ...    ...            ...     ...       ...
995           995         995  23731        Married    Male   60000.0
996           996         996  28672         Single    Male   70000.0
997           997         997  11809        Married    Male   60000.0
998           998         998  19664         Single    Male  100000.0
999           999         999  12121         Single    Male   60000.0

     Children        Education      Occupation Home Owner  Cars  \
0           1        Bachelors  Skilled Manual        Yes     0
1           3  Partial College        Clerical        Yes     1
2           5  Partial College    Professional         No     2
3           0        Bachelors    Professional        Yes     1
4           0        Bachelors        Clerical         No     0
..        ...              ...             ...        ...   ...
995         2      High School    Professional        Yes     2
996         4  Graduate Degree    Professional        Yes     0
997         2        Bachelors  Skilled Manual        Yes     0
998         3        Bachelors      Management         No     3
999         3      High School    Professional        Yes     2

    Commute Distance         Region   Age Purchased Bike
0          0-1 Miles         Europe  42.0             No
1          0-1 Miles         Europe  43.0             No
2          2-5 Miles         Europe  60.0             No
3         5-10 Miles        Pacific  41.0            Yes
4          0-1 Miles         Europe  36.0            Yes
..               ...            ...   ...            ...
995        2-5 Miles  North America  54.0            Yes
996        2-5 Miles  North America  35.0            Yes
997        0-1 Miles  North America  38.0            Yes
998        1-2 Miles  North America  38.0             No
999        10+ Miles  North America  53.0            Yes

[1000 rows x 15 columns]
```

1 LINEAR REGRESSION

[ ]:
```
from sklearn.linear_model import LinearRegression
lr=LinearRegression()
```

[ ]:
```
train = df1[df1["Age"].notnull()]
test = df1[df1["Age"].isnull()]
```

[ ]:
```
train.head()
```

[ ]:
```
   Unnamed: 0.1  Unnamed: 0     ID Marital Status  Gender   Income  Children  \
0             0           0  12496        Married  Female  40000.0         1
1             1           1  24107        Married    Male  30000.0         3
2             2           2  14177        Married    Male  80000.0         5
3             3           3  24381         Single    Male  70000.0         0
4             4           4  25597         Single    Male  30000.0         0

         Education      Occupation Home Owner  Cars Commute Distance   Region  \
0        Bachelors  Skilled Manual        Yes     0        0-1 Miles   Europe
1  Partial College        Clerical        Yes     1        0-1 Miles   Europe
2  Partial College    Professional         No     2        2-5 Miles   Europe
3        Bachelors    Professional        Yes     1       5-10 Miles  Pacific
4        Bachelors        Clerical         No     0        0-1 Miles   Europe

    Age Purchased Bike
0  42.0             No
1  43.0             No
2  60.0             No
3  41.0            Yes
4  36.0            Yes
```

[ ]:
```
test.head()
```

[ ]:
```
     Unnamed: 0.1  Unnamed: 0     ID Marital Status  Gender       Income  \
9               9           9  19280        Married    Male  56267.60563
98             98          98  19441        Married    Male  40000.00000
225           225         225  14135        Married    Male  20000.00000
371           371         371  22918         Single    Male  80000.00000
554           554         554  18580        Married  Female  60000.00000

     Children        Education    Occupation Home Owner  Cars  \
9           2  Partial College        Manual        Yes     1
98          0  Graduate Degree      Clerical        Yes     0
225         1  Partial College        Manual        Yes     0
371         5  Graduate Degree    Management        Yes     3
554         2  Graduate Degree  Professional        Yes     0
```

```
    Commute Distance         Region  Age Purchased Bike
9          0-1 Miles         Europe  NaN             Yes
98         0-1 Miles         Europe  NaN             Yes
225        1-2 Miles         Europe  NaN              No
371        0-1 Miles        Pacific  NaN              No
554        2-5 Miles  North America  NaN             Yes
```

[ ]:
```
y=train["Age"]
```

[ ]:
```
x=train.drop(["Age","Marital Status","Gender","Education","Occupation","Home␣
↪Owner","Region","Purchased Bike","Commute Distance","Unnamed: 0.1","Unnamed:␣
↪0"],axis=1)
```

[ ]:
```
x.head()
```

[ ]:
```
      ID   Income  Children  Cars
0  12496  40000.0         1     0
1  24107  30000.0         3     1
2  14177  80000.0         5     2
3  24381  70000.0         0     1
4  25597  30000.0         0     0
```

[ ]:
```
lr.fit(x,y)
```

[ ]:
```
LinearRegression()
```

[ ]:
```
test
```

[ ]:
```
     Unnamed: 0.1  Unnamed: 0     ID Marital Status  Gender       Income  \
9               9           9  19280        Married    Male  56267.60563
98             98          98  19441        Married    Male  40000.00000
225           225         225  14135        Married    Male  20000.00000
371           371         371  22918         Single    Male  80000.00000
554           554         554  18580        Married  Female  60000.00000
688           688         688  11699         Single    Male  60000.00000
770           770         770  17699        Married    Male  60000.00000
986           986         986  23704         Single    Male  40000.00000

     Children        Education      Occupation Home Owner  Cars  \
9           2  Partial College          Manual        Yes     1
98          0  Graduate Degree        Clerical        Yes     0
225         1  Partial College          Manual        Yes     0
371         5  Graduate Degree      Management        Yes     3
554         2  Graduate Degree    Professional        Yes     0
688         2        Bachelors  Skilled Manual         No     2
770         1  Graduate Degree  Skilled Manual         No     0
986         5      High School    Professional        Yes     4
```

```
    Commute Distance         Region  Age Purchased Bike
9          0-1 Miles         Europe  NaN             Yes
98         0-1 Miles         Europe  NaN             Yes
225        1-2 Miles         Europe  NaN              No
371        0-1 Miles        Pacific  NaN              No
554        2-5 Miles  North America  NaN             Yes
688        0-1 Miles  North America  NaN              No
770        0-1 Miles  North America  NaN              No
986        10+ Miles  North America  NaN             Yes
```

[ ]:
```
x_new=test.drop(["Age","Marital Status","Gender","Education","Occupation","Home␣
↪Owner","Region","Purchased Bike","Commute Distance","Unnamed: 0.1","Unnamed:␣
↪0"],axis=1)
```

[ ]:
```
x_new.head()
```

[ ]:
```
        ID       Income  Children  Cars
9    19280  56267.60563         2     1
98   19441  40000.00000         0     0
225  14135  20000.00000         1     0
371  22918  80000.00000         5     3
554  18580  60000.00000         2     0
```

[ ]:
```
pred=lr.predict(x_new)
```

[ ]:
```
pred
```

[ ]:
```
array([44.38300953, 36.68285017, 40.55618041, 55.82211518, 44.06780402,
       45.39772594, 40.54533003, 55.88701149])
```

[ ]:
```
test["Age"]=pred
```

```
<ipython-input-35-be88c5391501>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  test["Age"]=pred
```

[ ]:
```
test
```

[ ]:
```
     Unnamed: 0.1  Unnamed: 0     ID Marital Status  Gender       Income  \
9               9           9  19280        Married    Male  56267.60563
98             98          98  19441        Married    Male  40000.00000
225           225         225  14135        Married    Male  20000.00000
```

```
371           371         371  22918         Single    Male  80000.00000
554           554         554  18580        Married  Female  60000.00000
688           688         688  11699         Single    Male  60000.00000
770           770         770  17699        Married    Male  60000.00000
986           986         986  23704         Single    Male  40000.00000

     Children        Education      Occupation Home Owner  Cars  \
9           2  Partial College          Manual        Yes     1
98          0  Graduate Degree        Clerical        Yes     0
225         1  Partial College          Manual        Yes     0
371         5  Graduate Degree      Management        Yes     3
554         2  Graduate Degree    Professional        Yes     0
688         2        Bachelors  Skilled Manual         No     2
770         1  Graduate Degree  Skilled Manual         No     0
986         5      High School    Professional        Yes     4

    Commute Distance         Region        Age Purchased Bike
9          0-1 Miles         Europe  44.383010            Yes
98         0-1 Miles         Europe  36.682850            Yes
225        1-2 Miles         Europe  40.556180             No
371        0-1 Miles        Pacific  55.822115             No
554        2-5 Miles  North America  44.067804            Yes
688        0-1 Miles  North America  45.397726             No
770        0-1 Miles  North America  40.545330             No
986        10+ Miles  North America  55.887011            Yes
```

[ ]:
```
train
```

[ ]:
```
     Unnamed: 0.1  Unnamed: 0     ID Marital Status  Gender    Income  \
0               0           0  12496        Married  Female   40000.0
1               1           1  24107        Married    Male   30000.0
2               2           2  14177        Married    Male   80000.0
3               3           3  24381         Single    Male   70000.0
4               4           4  25597         Single    Male   30000.0
..            …         …    …             …      …       …
995           995         995  23731        Married    Male   60000.0
996           996         996  28672         Single    Male   70000.0
997           997         997  11809        Married    Male   60000.0
998           998         998  19664         Single    Male  100000.0
999           999         999  12121         Single    Male   60000.0

     Children        Education      Occupation Home Owner  Cars  \
0           1        Bachelors  Skilled Manual        Yes     0
1           3  Partial College        Clerical        Yes     1
2           5  Partial College    Professional         No     2
3           0        Bachelors    Professional        Yes     1
4           0        Bachelors        Clerical         No     0
```

```
..         …               …                …         …   …
995          2      High School    Professional        Yes     2
996          4  Graduate Degree    Professional        Yes     0
997          2        Bachelors  Skilled Manual        Yes     0
998          3        Bachelors      Management         No     3
999          3      High School    Professional        Yes     2

    Commute Distance         Region   Age Purchased Bike
0          0-1 Miles         Europe  42.0              No
1          0-1 Miles         Europe  43.0              No
2          2-5 Miles         Europe  60.0              No
3         5-10 Miles        Pacific  41.0             Yes
4          0-1 Miles         Europe  36.0             Yes
..               …            …   …              …
995        2-5 Miles  North America  54.0             Yes
996        2-5 Miles  North America  35.0             Yes
997        0-1 Miles  North America  38.0             Yes
998        1-2 Miles  North America  38.0              No
999        10+ Miles  North America  53.0             Yes

[992 rows x 15 columns]
```

[ ]:
```
final_data=pd.concat([train,test],axis=0)
```

[ ]:
```
final_data.tail()
```

[ ]:
```
     Unnamed: 0.1  Unnamed: 0     ID Marital Status  Gender   Income  \
371           371         371  22918         Single    Male  80000.0
554           554         554  18580        Married  Female  60000.0
688           688         688  11699         Single    Male  60000.0
770           770         770  17699        Married    Male  60000.0
986           986         986  23704         Single    Male  40000.0

     Children        Education      Occupation Home Owner  Cars  \
371         5  Graduate Degree      Management        Yes     3
554         2  Graduate Degree    Professional        Yes     0
688         2        Bachelors  Skilled Manual         No     2
770         1  Graduate Degree  Skilled Manual         No     0
986         5      High School    Professional        Yes     4

    Commute Distance         Region        Age Purchased Bike
371        0-1 Miles        Pacific  55.822115              No
554        2-5 Miles  North America  44.067804             Yes
688        0-1 Miles  North America  45.397726              No
770        0-1 Miles  North America  40.545330              No
986        10+ Miles  North America  55.887011             Yes
```

```
[ ]: final_data.reset_index(drop=True,inplace=True)
     final_data.head()
```

```
[ ]:    Unnamed: 0.1  Unnamed: 0     ID Marital Status  Gender   Income  Children  \
    0             0           0  12496        Married  Female  40000.0         1
    1             1           1  24107        Married    Male  30000.0         3
    2             2           2  14177        Married    Male  80000.0         5
    3             3           3  24381         Single    Male  70000.0         0
    4             4           4  25597         Single    Male  30000.0         0

            Education      Occupation Home Owner  Cars Commute Distance   Region  \
    0       Bachelors  Skilled Manual        Yes     0        0-1 Miles   Europe
    1  Partial College        Clerical        Yes     1        0-1 Miles   Europe
    2  Partial College    Professional         No     2        2-5 Miles   Europe
    3       Bachelors    Professional        Yes     1       5-10 Miles  Pacific
    4       Bachelors        Clerical         No     0        0-1 Miles   Europe

       Age Purchased Bike
    0  42.0             No
    1  43.0             No
    2  60.0             No
    3  41.0            Yes
    4  36.0            Yes
```

```
[ ]: final_data.tail()
```

```
[ ]:      Unnamed: 0.1  Unnamed: 0     ID Marital Status  Gender   Income  \
    995           371         371  22918         Single    Male  80000.0
    996           554         554  18580        Married  Female  60000.0
    997           688         688  11699         Single    Male  60000.0
    998           770         770  17699        Married    Male  60000.0
    999           986         986  23704         Single    Male  40000.0

         Children        Education      Occupation Home Owner  Cars  \
    995         5  Graduate Degree      Management        Yes     3
    996         2  Graduate Degree    Professional        Yes     0
    997         2        Bachelors  Skilled Manual         No     2
    998         1  Graduate Degree  Skilled Manual         No     0
    999         5      High School    Professional        Yes     4

        Commute Distance         Region        Age Purchased Bike
    995        0-1 Miles        Pacific  55.822115             No
    996        2-5 Miles  North America  44.067804            Yes
    997        0-1 Miles  North America  45.397726             No
    998        0-1 Miles  North America  40.545330             No
    999        10+ Miles  North America  55.887011            Yes
```

[ ]:
```
# Now Age has 0 null value.
final_data["Age"].isnull().mean()
```

[ ]:
```
0.0
```

2 THANK YOU