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

a) Find mean, median, mode and describe

In [4]:
```
df1=pd.read_csv("cleaned_fiat500.csv")
df2=pd.read_csv("cleaned_2015")
df3=pd.read_csv("3_Fitness-1 - 3_Fitness-1.csv")
df4=pd.read_csv("uber - uber.csv")
df5=pd.read_csv("4_drug200 - 4_drug200.csv")
```

In [7]:
```
print(df1.columns)
print(df2.columns)
print(df3.columns)
print(df4.columns)
print(df5.columns)
```

```
Index(['Unnamed: 0', 'ID', 'model', 'engine_power', 'age_in_days', 'km',
       'previous_owners', 'lat', 'lon', 'price', 'Unnamed: 9', 'Unnamed: 10'],
      dtype='object')
Index(['Unnamed: 0', 'Country', 'Region', 'Happiness Rank', 'Happiness Score',
       'Standard Error', 'Economy (GDP per Capita)', 'Family',
       'Health (Life Expectancy)', 'Freedom', 'Trust (Government Corruption)',
       'Generosity', 'Dystopia Residual'],
      dtype='object')
Index(['Row Labels', 'Sum of Jan', 'Sum of Feb', 'Sum of Mar',
       'Sum of Total Sales'],
      dtype='object')
Index(['Unnamed: 0', 'key', 'fare_amount', 'pickup_datetime',
       'pickup_longitude', 'pickup_latitude', 'dropoff_longitude',
       'dropoff_latitude', 'passenger_count'],
      dtype='object')
Index(['Age', 'Sex', 'BP', 'Cholesterol', 'Na_to_K', 'Drug'], dtype='object')
```

In [11]:
```
print(np.mean(df1["age_in_days"]))
print(np.mean(df2["Happiness Score"]))
print(np.mean(df3["Sum of Total Sales"]))
print(np.mean(df4["passenger_count"]))
print(np.mean(df5["Age"]))
```

```
1650.9056603773586
5.3757341772151905
255.55555555555554
1.684535
44.315
```

In [13]:
```
print(np.median(df1["age_in_days"]))
print(np.median(df2["Happiness Score"]))
print(np.median(df3["Sum of Total Sales"]))
print(np.median(df4["passenger_count"]))
print(np.median(df5["Age"]))
```

```
1035.0
5.2325
167.0
1.0
45.0
```

In [32]:
```
print(df1.mode().iloc[0],end='\n\n')
print(df2.mode().iloc[0],end='\n\n')
print(df3.mode().iloc[0],end='\n\n')
print(df4.mode().iloc[0],end='\n\n')
print(df5.mode().iloc[0],end='\n\n')
```

```
Unnamed: 0                 0
ID                       1.0
model                 lounge
engine_power            51.0
age_in_days            366.0
km                   17000.0
previous_owners          1.0
lat                 41.903221
lon                  12.49565
price                10500.0
Unnamed: 9               NaN
Unnamed: 10              NaN
Name: 0, dtype: object
```

```
Unnamed: 0                                      0
Country                                Afghanistan
Region                          Sub-Saharan Africa
Happiness Rank                                82.0
Happiness Score                              5.192
Standard Error                             0.03751
Economy (GDP per Capita)                       0.0
Family                                         0.0
Health (Life Expectancy)                   0.92356
Freedom                                        0.0
Trust (Government Corruption)              0.32524
Generosity                                     0.0
Dystopia Residual                          0.32858
Name: 0, dtype: object
```

```
Row Labels                   A
Sum of Jan              100.00%
Sum of Feb               10.57%
Sum of Mar              100.00%
Sum of Total Sales           75
Name: 0, dtype: object
```

```
Unnamed: 0                                    1
key                         2009-02-12 12:46:00
fare_amount                                 6.5
pickup_datetime      2009-02-12 12:46:00 UTC
pickup_longitude                            0.0
pickup_latitude                             0.0
dropoff_longitude                           0.0
dropoff_latitude                            0.0
passenger_count                             1.0
Name: 0, dtype: object
```

```
Age              47.0
Sex                 M
BP               HIGH
Cholesterol      HIGH
Na_to_K        12.006
Drug            drugY
Name: 0, dtype: object
```

In [40]:
```
print(df1.describe(),end="\n\n\n"*4)
print(df2.describe(),end="\n\n\n"*4)
print(df3.describe(),end="\n\n\n"*4)
print(df4.describe(),end="\n\n\n"*4)
print(df5.describe(),end="\n\n\n"*4)
```

```
min             75.000000
25%            127.000000
50%            167.000000
75%            171.000000
max           1150.000000



         Unnamed: 0    fare_amount  pickup_longitude  pickup_latitude  \
count  2.000000e+05  200000.000000     200000.000000    200000.000000
mean   2.771250e+07      11.359955        -72.527638        39.935885
std    1.601382e+07       9.901776         11.437787         7.720539
```

b) Find sum(), cumsum(), count, min and max values

In [45]:
```
print(df1["km"].sum())
print(df2["Happiness Rank"].sum())
print(df3["Sum of Total Sales"].sum())
print(df4["fare_amount"].sum())
print(df5["Na_to_K"].sum())
```

```
82068790.0
12560
2300
2271991.0500000003
3216.897
```

In [46]:
```
print(df1["km"].cumsum())
print(df2["Happiness Rank"].cumsum())
print(df3["Sum of Total Sales"].cumsum())
print(df4["fare_amount"].cumsum())
print(df5["Na_to_K"].cumsum())
```

```
0          25000.0
1          57500.0
2         199728.0
3         359728.0
4         466608.0
           ...
1532    81700303.0
1533    81815583.0
1534    81927583.0
1535    81988040.0
1536    82068790.0
Name: km, Length: 1537, dtype: float64
0          1
1          3
2          6
3         10
4         15
       ...
153    11934
154    12089
155    12245
156    12402
157    12560
Name: Happiness Rank, Length: 158, dtype: int64
0      75
1     235
2     336
3     463
4     642
5     809
6     980
7    1150
8    2300
Name: Sum of Total Sales, dtype: int64
0              7.50
1             15.20
2             28.10
3             33.40
4             49.40
            ...
199995    2271924.05
199996    2271931.55
199997    2271962.45
199998    2271976.95
199999    2271991.05
Name: fare_amount, Length: 200000, dtype: float64
0        25.355
1        38.448
2        48.562
3        56.360
4        74.403
         ...
195    3169.628
196    3181.634
197    3191.528
198    3205.548
```

```
199    3216.897
Name: Na_to_K, Length: 200, dtype: float64
```

In [47]:
```
print(df1["km"].count())
print(df2["Happiness Rank"].count())
print(df3["Sum of Total Sales"].count())
print(df4["fare_amount"].count())
print(df5["Na_to_K"].count())
```

```
1537
158
9
200000
200
```

In [49]:
```
print(df1["km"].min())
print(df2["Happiness Rank"].min())
print(df3["Sum of Total Sales"].min())
print(df4["fare_amount"].min())
print(df5["Na_to_K"].min())
```

```
1232.0
1
75
-52.0
6.269
```

In [50]:
```
print(df1["km"].max())
print(df2["Happiness Rank"].max())
print(df3["Sum of Total Sales"].max())
print(df4["fare_amount"].max())
print(df5["Na_to_K"].max())
```

```
235000.0
158
1150
499.0
38.247
```

c) Find covariance and correlation (spearman and pearsons)

In [54]:
```
from scipy.stats import pearsonr
from scipy.stats import spearmanr
from numpy import cov
```

In [61]:
```
print(df1.cov(),end=5*"\n")
print(df2.cov(),end=5*"\n")
print(df3.cov(),end=5*"\n")
print(df4.cov(),end=5*"\n")
print(df5.cov(),end=5*"\n")
```

```
pickup_latitude   -4.211348e+04     -0.648348        -72.098340
dropoff_longitude  5.668481e+04      1.167142        124.982650
dropoff_latitude   2.953191e+04     -0.741010        -65.774618
passenger_count    5.009811e+04      0.139296         -0.006569

                   pickup_latitude  dropoff_longitude  dropoff_latitude  \
Unnamed: 0           -42113.484730       56684.809960      29531.911251
fare_amount              -0.648348           1.167142         -0.741010
pickup_longitude        -72.098340         124.982650        -65.774618
pickup_latitude          59.606729         -78.465589         36.846061
dropoff_longitude       -78.465589         172.066387        -81.733638
dropoff_latitude         36.846061         -81.733638         46.169699
passenger_count          -0.016691           0.000598         -0.006209

                   passenger_count
Unnamed: 0            50098.109363
fare_amount               0.139296
pickup_longitude         -0.006569
pickup_latitude          -0.016691
dropoff_longitude         0.000598
```

In [66]:
```
print(pearsonr(df1["age_in_days"],df1["km"]))
print(pearsonr(df2["Happiness Rank"],df2["Happiness Score"]))
print(pearsonr(df4["fare_amount"],df4["passenger_count"]))
print(pearsonr(df5["Age"],df5["Na_to_K"]))
```

```
(0.8338906229249816, 0.0)
(-0.9921053148284925, 1.401375958157213e-142)
(0.010149925554531472, 5.644844770180446e-06)
(-0.06311949726772591, 0.3745756399034559)
```

In [68]:
```
print(spearmanr(df1["age_in_days"],df1["km"]))
print(spearmanr(df2["Happiness Rank"],df2["Happiness Score"]))
print(spearmanr(df4["fare_amount"],df4["passenger_count"]))
print(spearmanr(df5["Age"],df5["Na_to_K"]))
```

```
SpearmanrResult(correlation=0.8341055708983908, pvalue=0.0)
SpearmanrResult(correlation=-0.9999999999999999, pvalue=0.0)
SpearmanrResult(correlation=0.023295684126286974, pvalue=2.0202215346065764e-25)
SpearmanrResult(correlation=-0.047273882688479915, pvalue=0.5062200581387418)
```