Statistical Data Analysis 4

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats

#Excericse 1


columns = ['age', 'sex', 'on thyroxine', 'query on thyroxine', 'on antithyroid medic
true = ['t']
false = ['f']
data1 = pd.read_csv('allbp.data', header=None, sep=',', names=columns, index_col=Fal
data1 = data1.replace('?', np.NaN)
categories = ['sex', 'on thyroxine', 'query on thyroxine', 'on antithyroid medicatio
countables = ['age', 'TSH', 'T3', 'TT4', 'T4U', 'FTI', 'TBG']
for col in categories:
    data1[col] = data1[col].astype('category')

for col in countables:
    data1[col] = data1[col].astype('float64')
print(data1)
missingdata1 = data1.loc[:, data1.isnull().any()]
print(missingdata1)

print("There are 30 different variables and 2800 observations. Age, sex, TSH, T3, TT
```

```
       age sex on thyroxine query on thyroxine on antithyroid medication  \
0     41.0   F        False              False                       False
1     23.0   F        False              False                       False
2     46.0   M        False              False                       False
3     70.0   F         True              False                       False
4     70.0   F        False              False                       False
...    ...  ..          ...                ...                         ...
2795  70.0   M        False              False                       False
2796  73.0   M        False               True                       False
2797  75.0   M        False              False                       False
2798  60.0   F        False              False                       False
2799  81.0   F        False              False                       False

       sick pregnant thyroid surgery I131 treatment query hypothyroid  ...  \
0     False    False           False          False             False  ...
1     False    False           False          False             False  ...
2     False    False           False          False             False  ...
3     False    False           False          False             False  ...
4     False    False           False          False             False  ...
...     ...      ...             ...            ...               ...  ...
2795  False    False           False          False             False  ...
2796  False    False           False          False             False  ...
2797  False    False           False          False             False  ...
2798  False    False           False          False             False  ...
2799  False    False           False          False             False  ...

     T3 measured   T3 TT4 measured    TT4 T4U measured   T4U FTI measured  \
0            True  2.5         True  125.0         True  1.14         True
1            True  2.0         True  102.0        False   NaN        False
2           False  NaN         True  109.0         True  0.91         True
3            True  1.9         True  175.0        False   NaN        False
4            True  1.2         True   61.0         True  0.87         True
...           ...  ...          ...    ...          ...   ...          ...
2795        False  NaN         True  155.0         True  1.05         True
2796         True  0.7         True   63.0         True  0.88         True
2797        False  NaN         True  147.0         True  0.80         True
```

```
2798       False  NaN         True  100.0         True  0.83         True
2799        True  1.5         True  114.0         True  0.99         True

        FTI TBG measured  TBG
0     109.0        False  NaN
1       NaN        False  NaN
2     120.0        False  NaN
3       NaN        False  NaN
4      70.0        False  NaN
...     ...          ...  ...
2795  148.0        False  NaN
2796   72.0        False  NaN
2797  183.0        False  NaN
2798  121.0        False  NaN
2799  115.0        False  NaN

[2800 rows x 28 columns]
       age sex   TSH   T3    TT4   T4U    FTI  TBG
0     41.0   F  1.30  2.5  125.0  1.14  109.0  NaN
1     23.0   F  4.10  2.0  102.0   NaN    NaN  NaN
2     46.0   M  0.98  NaN  109.0  0.91  120.0  NaN
3     70.0   F  0.16  1.9  175.0   NaN    NaN  NaN
4     70.0   F  0.72  1.2   61.0  0.87   70.0  NaN
...    ...  ..   ...  ...    ...   ...    ...  ...
2795  70.0   M  2.70  NaN  155.0  1.05  148.0  NaN
2796  73.0   M   NaN  0.7   63.0  0.88   72.0  NaN
2797  75.0   M   NaN  NaN  147.0  0.80  183.0  NaN
2798  60.0   F  1.40  NaN  100.0  0.83  121.0  NaN
2799  81.0   F  1.20  1.5  114.0  0.99  115.0  NaN

[2800 rows x 8 columns]
There are 30 different variables and 2800 observations. Age, sex, TSH, T3, TT4, T4U,
FTI, TBG have missing values. The number of missing variables is: 4556
```

In [2]:

```
#Excercise 2


data1['age']

booleans = categories = ['on thyroxine', 'query on thyroxine', 'on antithyroid medic
for col in countables:
    print('mean', col, ':', np.mean(data1[col]))
    print('standard deviation', col, ':', np.std(data1[col]))

for col in booleans:
    print('frequency of yes in', col, ':', (data1[col].values == True).sum())
    print('relative frequency of yes in', col, ':', (data1[col].values == True).sum(
```

```
mean age : 51.8442300821722
standard deviation age : 20.45750468142629
mean TSH : 4.672150238473764
standard deviation TSH : 21.445189678356876
mean T3 : 2.0249661399548584
standard deviation T3 : 0.8244142519987397
mean TT4 : 109.07240061162081
standard deviation TT4 : 35.38567761354566
mean T4U : 0.9979121054734302
standard deviation T4U : 0.1943516472529468
mean FTI : 110.78798403193613
standard deviation FTI : 32.87742145674531
mean TBG : nan
standard deviation TBG : nan
frequency of yes in on thyroxine : 330
relative frequency of yes in on thyroxine : 0.11785714285714285
frequency of yes in query on thyroxine : 40
relative frequency of yes in query on thyroxine : 0.014285714285714285
frequency of yes in on antithyroid medication : 34
```

```
relative frequency of yes in on antithyroid medication : 0.012142857142857143
frequency of yes in sick : 110
relative frequency of yes in sick : 0.039285714285714285
frequency of yes in pregnant : 41
relative frequency of yes in pregnant : 0.014642857142857143
frequency of yes in thyroid surgery : 39
relative frequency of yes in thyroid surgery : 0.013928571428571429
frequency of yes in I131 treatment : 48
relative frequency of yes in I131 treatment : 0.017142857142857144
frequency of yes in query hypothyroid : 163
relative frequency of yes in query hypothyroid : 0.05821428571428571
frequency of yes in query hyperthyroid : 173
relative frequency of yes in query hyperthyroid : 0.061785714285714284
frequency of yes in lithium : 14
relative frequency of yes in lithium : 0.005
frequency of yes in goitre : 25
relative frequency of yes in goitre : 0.008928571428571428
frequency of yes in tumor : 71
relative frequency of yes in tumor : 0.025357142857142856
frequency of yes in hypopituitary : 1
relative frequency of yes in hypopituitary : 0.00035714285714285714
frequency of yes in psych : 135
relative frequency of yes in psych : 0.048214285714285716
frequency of yes in TSH measured : 2516
relative frequency of yes in TSH measured : 0.8985714285714286
frequency of yes in T3 measured : 2215
relative frequency of yes in T3 measured : 0.7910714285714285
frequency of yes in TT4 measured : 2616
relative frequency of yes in TT4 measured : 0.9342857142857143
frequency of yes in T4U measured : 2503
relative frequency of yes in T4U measured : 0.8939285714285714
frequency of yes in FTI measured : 2505
relative frequency of yes in FTI measured : 0.8946428571428572
frequency of yes in TBG measured : 0
relative frequency of yes in TBG measured : 0.0
```

In [3]:

```
#Excercise 3

data2 = pd.read_csv('GDS5037.soft', skiprows=160, skipfooter=1)
data2
```

```
<ipython-input-3-53f7d8fe7534>:3: ParserWarning: Falling back to the 'python' engine
because the 'c' engine does not support skipfooter; you can avoid this warning by sp
ecifying engine='python'.
  data2 = pd.read_csv('GDS5037.soft', skiprows=160, skipfooter=1)
```

Out[3]:

	ID_REF\tIDENTIFIER\tGSM1068478\tGSM1068479\tGSM1068481\tGSM1068482\tGSM1068483\tGS
0	
1	
2	
3	
4	
...	
41103	
41104	
41105	
41106	
41107	

41108 rows × 1 columns

In [12]:

```
#Excerise 4
espoo = pd.read_csv('44-espoo.csv')
helsinki = pd.read_csv('44-helsinki.csv')

data3 = pd.merge(espoo, helsinki)
print('Amount of observation days in total:', data3['date'].unique().size)

gallen = data3[['date', 'Gallen-Kallela']].dropna()
vayla = data3[['date', 'Länsiväylä']].dropna()
tuulenkuja = data3[['date', 'Länsituulenkuja']].dropna()
esplanadi = data3[['date', 'Eteläesplanadi']].dropna()
kaivo = data3[['date', 'Kaivokatu']].dropna()
kuusi = data3[['date','Kuusisaarentie']].dropna()
meri = data3[['date', 'Merikannontie']].dropna()

print('Amount of observation days for Gallen-Kallela', gallen['date'].unique().size)
print('Amount of observation days for Länsiväylä', vayla['date'].unique().size)
print('Amount of observation days for Länsituulenkuja', tuulenkuja['date'].unique().
print('Amount of observation days for Eteläesplandi', esplanadi['date'].unique().siz
print('Amount of observation days for Kaivokatu', kaivo['date'].unique().size)
print('Amount of observation days for Kuusisaarentie', kuusi['date'].unique().size)
print('Amount of observation days for Merikannontie', meri['date'].unique().size)

allstreets = data3.dropna()
streets = data3.drop(['date'], axis=1)
nodate = allstreets.drop(['date'], axis=1)
names = ['Gallen-Kallela', 'Länsiväylä', 'Länsituulenkuja', 'Eteläesplanadi', 'Kaivo
print('Amount of days all streets were observed:', allstreets['date'].unique().size)

for col in names:
    print(col, ':', streets[col].sum())


print('Merikannontie seems to be the most popular')

for col in names:
    print(col, ':', nodate[col].sum())

print('When removing all Nan values the must busiest street seems to be Kaivokatu. R
```

```
Amount of observation days in total: 2714
Amount of observation days for Gallen-Kallela 1459
Amount of observation days for Länsiväylä 1460
Amount of observation days for Länsituulenkuja 2209
Amount of observation days for Eteläesplandi 2652
Amount of observation days for Kaivokatu 1680
Amount of observation days for Kuusisaarentie 2456
Amount of observation days for Merikannontie 2511
Amount of days all streets were observed: 1400
Gallen-Kallela : 1107151.0
Länsiväylä : 1324149.0
Länsituulenkuja : 2037502.0
Eteläesplanadi : 3254794.0
Kaivokatu : 3557794.0
Kuusisaarentie : 2645829.0
Merikannontie : 4491369.0
Merikannontie seems to be the most popular
Gallen-Kallela : 1054062.0
Länsiväylä : 1262780.0
Länsituulenkuja : 1471640.0
Eteläesplanadi : 1531870.0
```

```
Kaivokatu : 3016688.0
Kuusisaarentie : 1500513.0
Merikannontie : 2445090.0
When removing all Nan values the must busiest street seems to be Kaivokatu. Reason f
or this might be because on days when Merikannontie had more people not all streets
had as much. Might be related to like city events
```