Assignment 1

Gengcong Yan - 1009903

CS-E4650 Methods of Data mining

October 3, 2021

1 Task 1

1.1 a)

I calculate all pairwise correlations between features, excluding only the rat id. The functions below I implemented are for calculation.

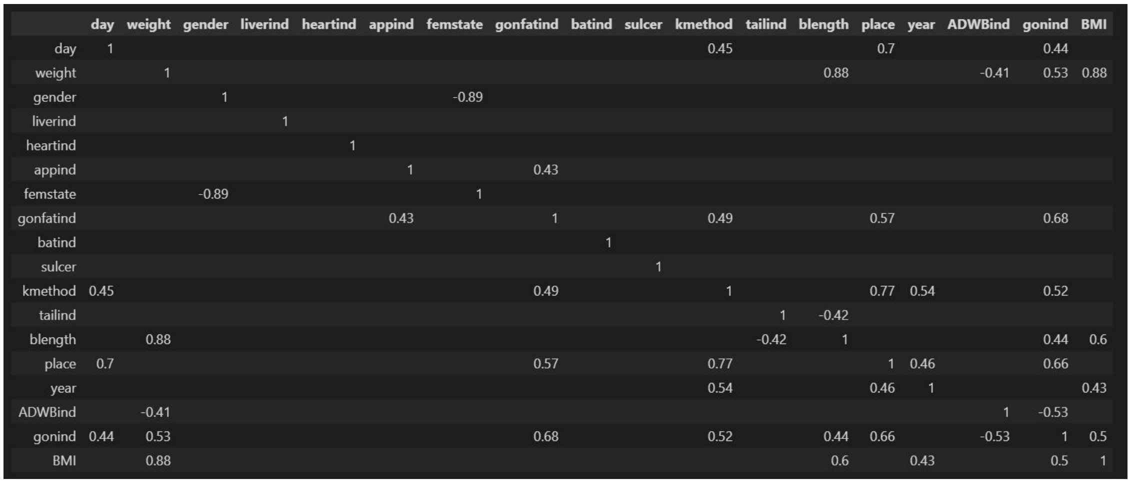
```python
def check_strong(x):
    if np.abs(x)>0.4:
        return x
    else:
        return ""

def get_correlation(data):
    p_corr=data.corr(method='pearson')
    p_corr=p_corr.round(2)
    p_corr
    corr_strong=p_corr.applymap(check_strong)
    return corr_strong
```

The correlation result is shown in fig 1 after deleting the values less than 0.40. We can see some strong correlations in the form of features pairs below:

1. femsate-(gender), kmethod-(place)
2. day-(kmethod,place,gonind), year-(kmethod,place,BMI)
3. weight-(blength,ADWBind,gonind,BMI),gonfatind-(appind,kmethod,place,gonind)

1.2 b)

After removing outlier rats, rat2, rat53, rat120, and rat434, and calculate correlations again.we can see liveind starts to show strong correlations with ADWBind and gonind with the value 0.46 and -0.48. Heartind also starts to show strong correlations with many features



	day	weight	gender	liverind	heartind	appind	femstate	gonfatind	batind	sulcer	kmethod	tailind	blength	place	year	ADWBind	gonind	BMI
day	1										0.45			0.7			0.44	
weight		1											0.88			-0.41	0.53	0.88
gender			1				-0.89											
liverind				1														
heartind					1													
appind						1		0.43										
femstate			-0.89				1											
gonfatind						0.43		1			0.49			0.57			0.68	
batind									1									
sulcer										1								
kmethod	0.45							0.49			1			0.77	0.54		0.52	
tailind												1	-0.42					
blength		0.88										-0.42	1				0.44	0.6
place	0.7							0.57			0.77			1	0.46		0.66	
year											0.54			0.46	1			0.43
ADWBind		-0.41														1	-0.53	
gonind	0.44	0.53						0.68			0.52		0.44	0.66		-0.53	1	0.5
BMI		0.88											0.6		0.43		0.5	1

Figure 1: Strong Correlations Bewteen Features.

like blength and ADWBind. Because some of their values in liverind or heartind about our deleted outliers are apparently unreasonable. they either took up too much or too litter percent of whole body weights, which is impossible.

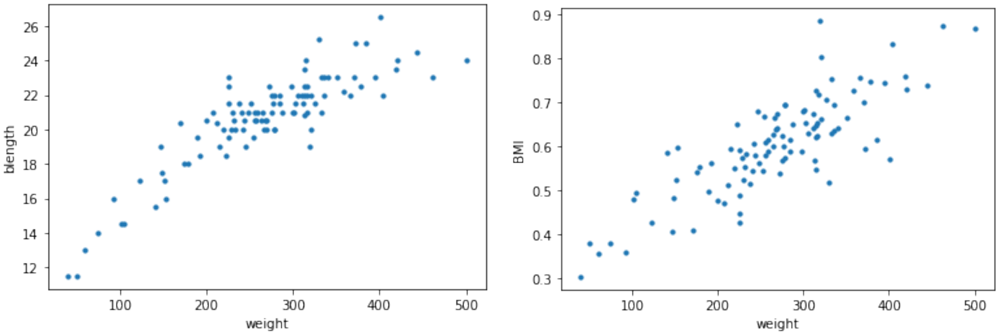
1.3 c)

```
freezer_index=data_dropb.loc[data_dropb["day"]==400].index
data_dropb.loc[freezer_index,"day"]=0
data_dropb.loc[freezer_index,"year"]=-1
corr3=get_correlation(data_dropb)
corr3

data_dropb=data_dropb.drop(freezer_index)
corr4=get_correlation(data_dropb)
corr4
```

We change the special codes for the freezer rats: day=0 and year=-1, remove them last to see correlations differences. the correlation of day and years with many other features are vanished and extremely negative. I think day and year don't have strong correlations with other features, those time related feature can't maintain obvious connection between the actual physiological characteristics.

1.4 e)

```
data_dropb["femstate"]=data_dropb["femstate"]+10
data_dropb["place"]=data_dropb["place"]+5
corr5=get_correlation(data_dropb)
corr5
```



(a) Weight-Blength

(b) Weight-BMI

Figure 2: Features Correlation Plots

We change special codes of categorical features femstate, kmethod and place by add numbers to the original values. And we can see the correlations related to them doesn't change at all. Because they are just categorical features in data, which mean the codes represent some certain category without meaning any Mathematical values. So you can change them to any code you like as long as it's the same number in one category.

1.5 f)

From the observation in correlations matrix, I think ***weight, blength,gonind,BMI, gonfatind*** are reliable features and shows strong linear trend to each other. When one of the metrics increases or decreases, the others will increase or decrease as well. Because weight and height are the most basic characteristics, all other indicators based on them show a strong correlation with them. From the scatter plot in Fig.2, we can observe the linear trend between different features like Weight, Blength and BMI.