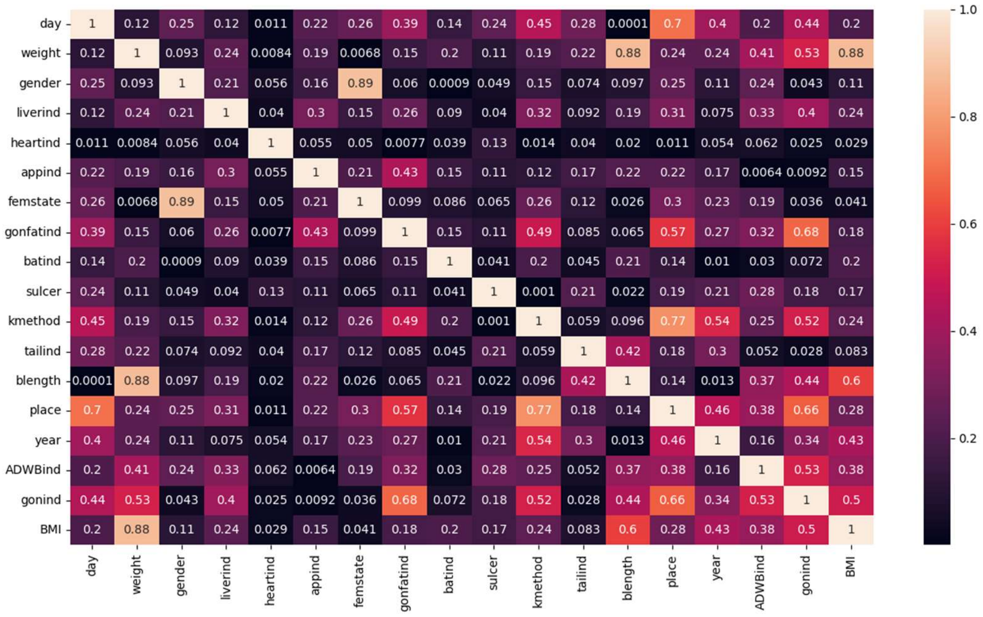
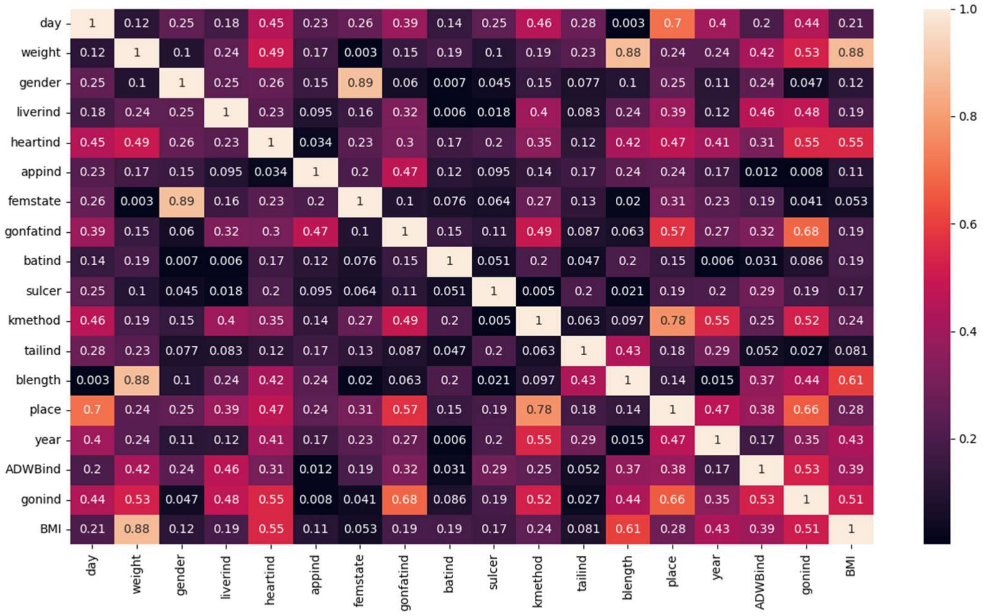
Task1

a) i) The strongest correlation involving categorical features is between **gender** and **femstate**, where the correlation is about **0.89**.
ii) The strongest correlation involving temporal features is between **day** and **place**, where the correlation is about **0.7**.
iii) The strongest correlation involving other numerical features is between **weight** and **blength**, where the correlation is about **0.88**.



	day	weight	gender	liverind	heartind	appind	femstate	gonfatind	batind	sulcer	kmethod	tailind	blength	place	year	ADWBind	gonind	BMI
day	1	0.12	0.25	0.12	0.011	0.22	0.26	0.39	0.14	0.24	0.45	0.28	0.0001	0.7	0.4	0.2	0.44	0.2
weight	0.12	1	0.093	0.24	0.0084	0.19	0.0068	0.15	0.2	0.11	0.19	0.22	0.88	0.24	0.24	0.41	0.53	0.88
gender	0.25	0.093	1	0.21	0.056	0.16	0.89	0.06	0.0009	0.049	0.15	0.074	0.097	0.25	0.11	0.24	0.043	0.11
liverind	0.12	0.24	0.21	1	0.04	0.3	0.15	0.26	0.09	0.04	0.32	0.092	0.19	0.31	0.075	0.33	0.4	0.24
heartind	0.011	0.0084	0.056	0.04	1	0.055	0.05	0.0077	0.039	0.13	0.014	0.04	0.02	0.011	0.054	0.062	0.025	0.029
appind	0.22	0.19	0.16	0.3	0.055	1	0.21	0.43	0.15	0.11	0.12	0.17	0.22	0.22	0.17	0.0064	0.0092	0.15
femstate	0.26	0.0068	0.89	0.15	0.05	0.21	1	0.099	0.086	0.065	0.26	0.12	0.026	0.3	0.23	0.19	0.036	0.041
gonfatind	0.39	0.15	0.06	0.26	0.0077	0.43	0.099	1	0.15	0.11	0.49	0.085	0.065	0.57	0.27	0.32	0.68	0.18
batind	0.14	0.2	0.0009	0.09	0.039	0.15	0.086	0.15	1	0.041	0.2	0.045	0.21	0.14	0.01	0.03	0.072	0.2
sulcer	0.24	0.11	0.049	0.04	0.13	0.11	0.065	0.11	0.041	1	0.001	0.21	0.022	0.19	0.21	0.28	0.18	0.17
kmethod	0.45	0.19	0.15	0.32	0.014	0.12	0.26	0.49	0.2	0.001	1	0.059	0.096	0.77	0.54	0.25	0.52	0.24
tailind	0.28	0.22	0.074	0.092	0.04	0.17	0.12	0.085	0.045	0.21	0.059	1	0.42	0.18	0.3	0.052	0.028	0.083
blength	0.0001	0.88	0.097	0.19	0.02	0.22	0.026	0.065	0.21	0.022	0.096	0.42	1	0.14	0.013	0.37	0.44	0.6
place	0.7	0.24	0.25	0.31	0.011	0.22	0.3	0.57	0.14	0.19	0.77	0.18	0.14	1	0.46	0.38	0.66	0.28
year	0.4	0.24	0.11	0.075	0.054	0.17	0.23	0.27	0.01	0.21	0.54	0.3	0.013	0.46	1	0.16	0.34	0.43
ADWBind	0.2	0.41	0.24	0.33	0.062	0.0064	0.19	0.32	0.03	0.28	0.25	0.052	0.37	0.38	0.16	1	0.53	0.38
gonind	0.44	0.53	0.043	0.4	0.025	0.0092	0.036	0.68	0.072	0.18	0.52	0.028	0.44	0.66	0.34	0.53	1	0.5
BMI	0.2	0.88	0.11	0.24	0.029	0.15	0.041	0.18	0.2	0.17	0.24	0.083	0.6	0.28	0.43	0.38	0.5	1

b) Also, with **liverind,** but especially with **heartind** the largest correlations increased significantly. Before the largest correlation involving heartind was 0.13, but after removing outliers the largest correlations were about 0.55. The reason for the huge difference are abnormally large or small values for the two features with the outliers in the dataset that could be due to some very special cases with the outliers or mistakes in measuring or recording the data. Ie. rat120 has a 100-times higher heartind than some of the other rats.



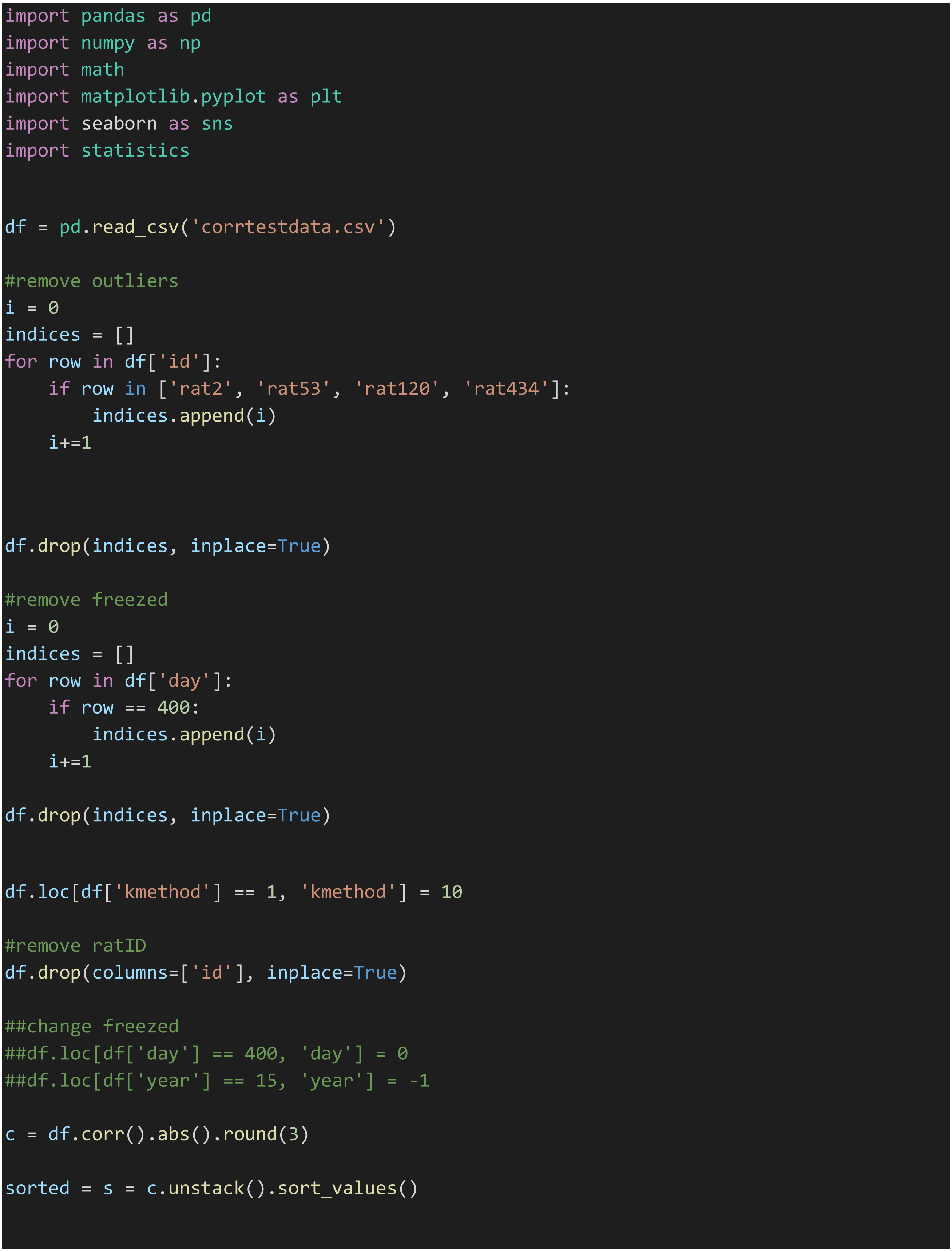
	day	weight	gender	liverind	heartind	appind	femstate	gonfatind	batind	sulcer	kmethod	tailind	blength	place	year	ADWBind	gonind	BMI
day	1	0.12	0.25	0.18	0.45	0.23	0.26	0.39	0.14	0.25	0.46	0.28	0.003	0.7	0.4	0.2	0.44	0.21
weight	0.12	1	0.1	0.24	0.49	0.17	0.003	0.15	0.19	0.1	0.19	0.23	0.88	0.24	0.24	0.42	0.53	0.88
gender	0.25	0.1	1	0.25	0.26	0.15	0.89	0.06	0.007	0.045	0.15	0.077	0.1	0.25	0.11	0.24	0.047	0.12
liverind	0.18	0.24	0.25	1	0.23	0.095	0.16	0.32	0.006	0.018	0.4	0.083	0.24	0.39	0.12	0.46	0.48	0.19
heartind	0.45	0.49	0.26	0.23	1	0.034	0.23	0.3	0.17	0.2	0.35	0.12	0.42	0.47	0.41	0.31	0.55	0.55
appind	0.23	0.17	0.15	0.095	0.034	1	0.2	0.47	0.12	0.095	0.14	0.17	0.24	0.24	0.17	0.012	0.008	0.11
femstate	0.26	0.003	0.89	0.16	0.23	0.2	1	0.1	0.076	0.064	0.27	0.13	0.02	0.31	0.23	0.19	0.041	0.053
gonfatind	0.39	0.15	0.06	0.32	0.3	0.47	0.1	1	0.15	0.11	0.49	0.087	0.063	0.57	0.27	0.32	0.68	0.19
batind	0.14	0.19	0.007	0.006	0.17	0.12	0.076	0.15	1	0.051	0.2	0.047	0.2	0.15	0.006	0.031	0.086	0.19
sulcer	0.25	0.1	0.045	0.018	0.2	0.095	0.064	0.11	0.051	1	0.005	0.2	0.021	0.19	0.2	0.29	0.19	0.17
kmethod	0.46	0.19	0.15	0.4	0.35	0.14	0.27	0.49	0.2	0.005	1	0.063	0.097	0.78	0.55	0.25	0.52	0.24
tailind	0.28	0.23	0.077	0.083	0.12	0.17	0.13	0.087	0.047	0.2	0.063	1	0.43	0.18	0.29	0.052	0.027	0.081
blength	0.003	0.88	0.1	0.24	0.42	0.24	0.02	0.063	0.2	0.021	0.097	0.43	1	0.14	0.015	0.37	0.44	0.61
place	0.7	0.24	0.25	0.39	0.47	0.24	0.31	0.57	0.15	0.19	0.78	0.18	0.14	1	0.47	0.38	0.66	0.28
year	0.4	0.24	0.11	0.12	0.41	0.17	0.23	0.27	0.006	0.2	0.55	0.29	0.015	0.47	1	0.17	0.35	0.43
ADWBind	0.2	0.42	0.24	0.46	0.31	0.012	0.19	0.32	0.031	0.29	0.25	0.052	0.37	0.38	0.17	1	0.53	0.39
gonind	0.44	0.53	0.047	0.48	0.55	0.008	0.041	0.68	0.086	0.19	0.52	0.027	0.44	0.66	0.35	0.53	1	0.51
BMI	0.21	0.88	0.12	0.19	0.55	0.11	0.053	0.19	0.19	0.17	0.24	0.081	0.61	0.28	0.43	0.39	0.51	1

c) When changing the values of **day** and **year** for freezer rats some correlations changed a lot. Removing all freezer rats made the correlations normalize back to similar numbers as before. This is mainly due to relationships between different features. Ie. the largest **place** values (7-9) all belong to freezer rats. We can say that "the larger the day the larger the place", but if we change the day of freezer rats from 400 to 0, the correlation changes. Removing all freezer rats almost normalizes the situation since there are only 34 of them and they only represent a small portion of the data.

d) .

e) A similar correlation as in part c can be found between **place** and **kmethod**. Ie. when kmethod is 1, 317 rats have a place value of 2. That is 85% of the rats that have 1 as kmethod. By changing each kmethod that has value 2 to 10, we can reduce the correlation between place and kmethod from 0.74 down to 0.03.

f) The most reliable correlations are the strongest ones. Between **femstate** and **gender, weight** and **blength** and between **weight** and **BMI**, the correlations are almost 0.9 and in most cases it applies that the higher weight the higher blength and BMI, so these correlations for example show strong linear trends.



```python
import pandas as pd
import numpy as np
import math
import matplotlib.pyplot as plt
import seaborn as sns
import statistics


df = pd.read_csv('corrtestdata.csv')

#remove outliers
i = 0
indices = []
for row in df['id']:
    if row in ['rat2', 'rat53', 'rat120', 'rat434']:
        indices.append(i)
    i+=1



df.drop(indices, inplace=True)

#remove freezed
i = 0
indices = []
for row in df['day']:
    if row == 400:
        indices.append(i)
    i+=1

df.drop(indices, inplace=True)


df.loc[df['kmethod'] == 1, 'kmethod'] = 10

#remove ratID
df.drop(columns=['id'], inplace=True)

##change freezed
##df.loc[df['day'] == 400, 'day'] = 0
##df.loc[df['year'] == 15, 'year'] = -1

c = df.corr().abs().round(3)

sorted = s = c.unstack().sort_values()
```

```
plt.figure(figsize = (15, 15))
sns.heatmap(c, annot = True)

##plt.matshow(df.corr())
plt.show()
```