Mesirov Lab Data Preprocessing Assignment

August 26, 2021

[ ]:

```
import pandas as pd;
import os;
import sys;
!{sys.executable} -m pip install cmapPy;
import cmapPy;
from cmapPy.pandasGEXpress.parse import parse;
import matplotlib.pyplot as plt
```

[ ]:

```
#1
def process_gct(gct_file, summary=False):
    if summary:
        gct_dataframe = cmapPy.pandasGEXpress.parse.parse(gct_file)
        print(str(gct_dataframe.data_df.shape[0]) + ' rows and ' +␣
 ↪str(gct_dataframe.data_df.shape[1]) + ' columns were imported.')
        return gct_dataframe
    else:
        gct_dataframe = cmapPy.pandasGEXpress.parse.parse(gct_file)
        return gct_dataframe
```

[ ]:

```
#2A
gct_df = process_gct('BRCA_minimal_60x19.gct', True)
gct_df
```

[ ]:

```
#2B
gct_df = process_gct('BRCA_large_20783x40.gct')
gct_df
```

[ ]:

```
#3
hist1 = gct_df.data_df.hist(column='A7-A0DB-normal',bins=40)
hist2 = gct_df.data_df.hist(column='A7-A13E-normal', bins=40)
hist3 = gct_df.data_df.hist(column='BH-A0B3-primary', bins=40)
hist4 = gct_df.data_df.hist(column='BH-A0B5-primary', bins=40)
```

[ ]:

```
#4
gct_df_new = gct_df.data_df.copy()
gct_df_new['Mean'] = gct_df_new.mean(numeric_only=True, axis=1)
gct_df_new['Median'] = gct_df_new.median(numeric_only=True, axis=1)
```

```
gct_df_new['Standard Deviation'] = gct_df_new.std(numeric_only=True, axis=1)
gct_df_new
```

[ ]:

```
#5A
rows_to_keep = int(0.9*len(gct_df_new))
gct_df_new_filtered = gct_df_new.nsmallest(rows_to_keep, 'Standard Deviation')
gct_df_new_filtered
```

[ ]:

```
#5B
print(gct_df_new_filtered.loc[:, ~gct_df_new_filtered.columns.isin(['Mean',␣
 ↪'Median', 'Standard Deviation'])].mean())
print(gct_df_new_filtered.loc[:, ~gct_df_new_filtered.columns.isin(['Mean',␣
 ↪'Median', 'Standard Deviation'])].median())
new_hist1 = gct_df_new_filtered.hist(column='A7-A0DB-normal',bins=40)
new_hist2 = gct_df_new_filtered.hist(column='A7-A13E-normal', bins=40)
new_hist3 = gct_df_new_filtered.hist(column='BH-A0B3-primary', bins=40)
new_hist4 = gct_df_new_filtered.hist(column='BH-A0B5-primary', bins=40)
```