Python case study

November 5, 2022

[130]:
```
%matplotlib inline
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
```

The excel file `data.xls` contains data for a series on experiments on cell migration and proliferation. Use what you have learned during the course to analyze the dataset. Show which treatments in which cell line have an effect on cell migratoin and/or proliferation. Collect your findings in a well organized jupyter notebook, which is the final product of this case study. The notebook should contain: * different kinds of plots, for example scatter plots, strip or swarm plots, histograms, and boxplots; * multiple plots in one figure using subplots * a written discussion of the plots; just code and figures is not enough

Make sure that your plots contain error bars wherever that is applicable and plots are fully anotated (e.g. titles, axes labels, legends).

Hints: * **Google is your friend** * Load the data from `data.xls` using `df = pd.read_excel('data.xls')`. * Markdown syntax * Jupyter notebook cheat sheet * Pandas cheat sheet * Matplotlib plotting examples * Matplotlib cheat sheet * Seaborn plotting examples * Seaborn cheat sheet * If you feel up to it, have a look at running statistical tests: https://www.scipy-lectures.org/packages/statistics/index.html.

0.1 Import Data and Inspect Data

[131]:
```
df = pd.read_excel('data.xls')
```

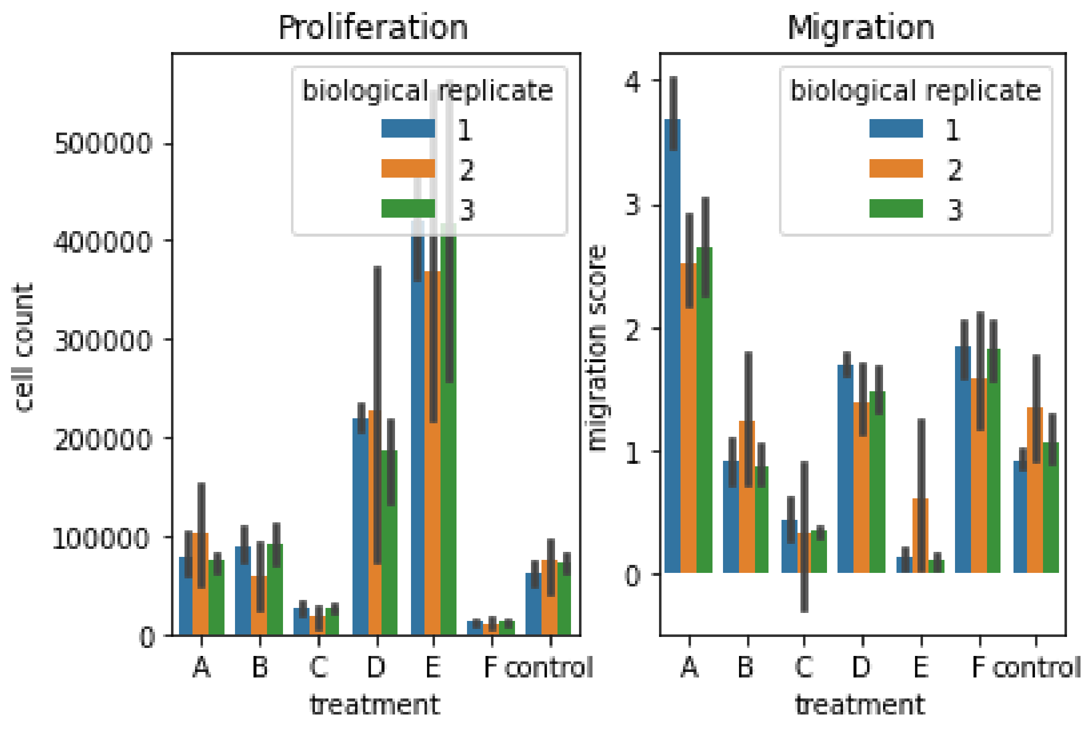
[132]:
```
df.describe()
```

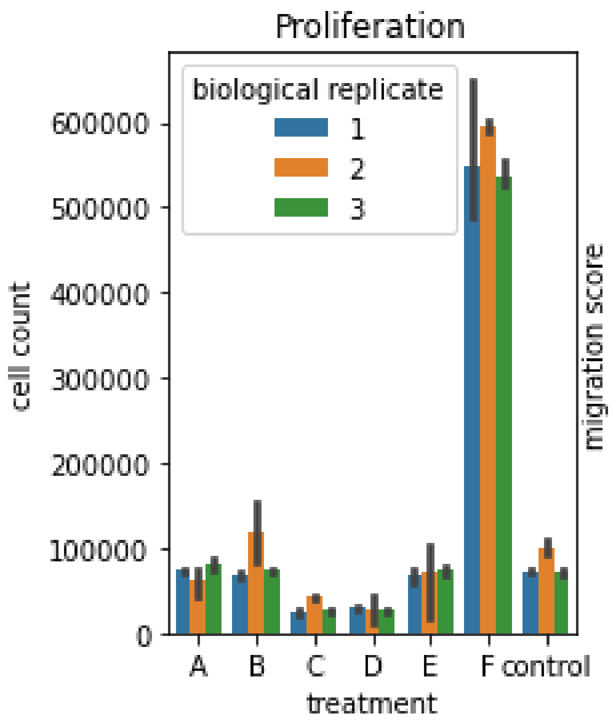
[132]:
```
          cell count  migration score  biological replicate  \
count     315.000000       315.000000            315.000000
mean    98029.471710         1.522320              2.000000
std    111412.765609         1.833875              0.817796
min    -25652.958231        -0.447149              1.000000
25%     42858.984183         0.568913              1.000000
50%     72331.955007         1.001566              2.000000
75%     90270.495558         1.787471              3.000000
max    649911.199232        11.346342              3.000000
```

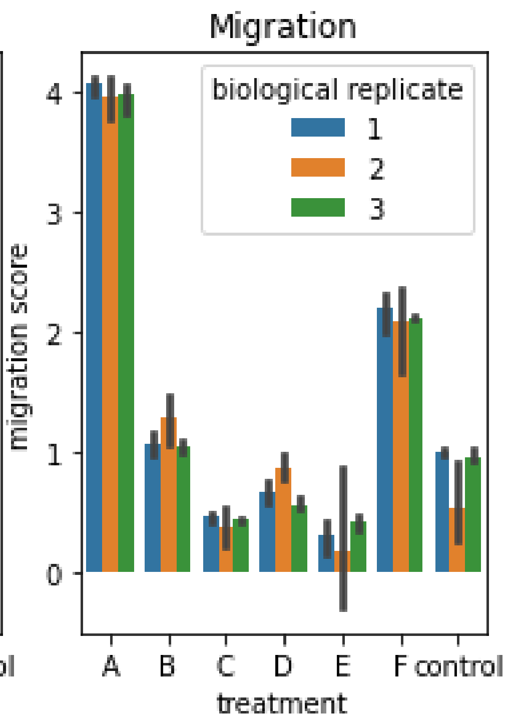
```
       technical replicate
count             315.000000
mean                2.000000
std                 0.817796
min                 1.000000
25%                 1.000000
50%                 2.000000
75%                 3.000000
max                 3.000000
```

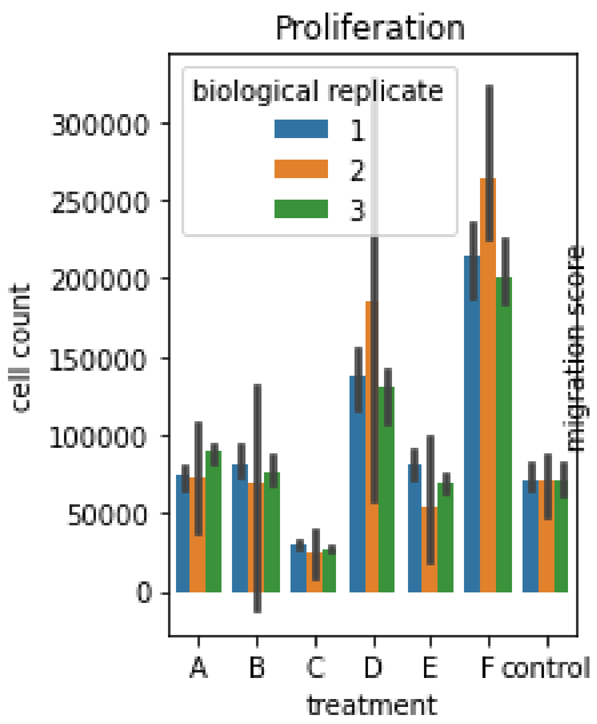
0.2 Group dataframe by same conditions to isolate the technical replicates

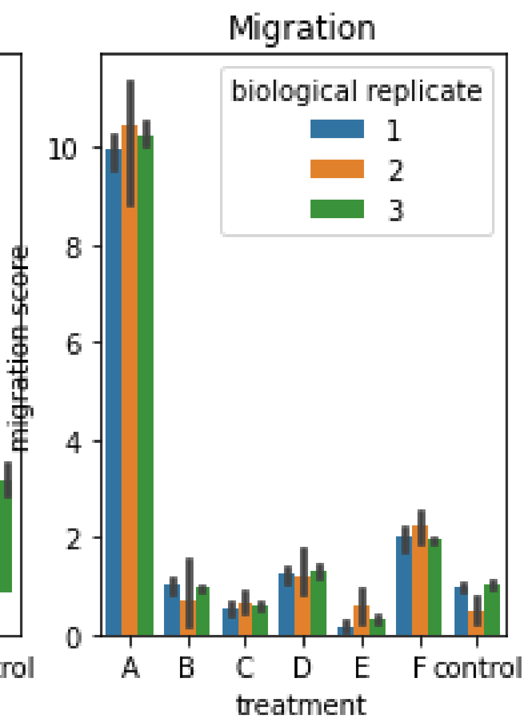
[133]:
```
for i in df['line'].unique():
    fig, ax = plt.subplots(1,2)
    sns.barplot(data=df[df['line']==i], y='cell count', x='treatment',␣
  ↪hue='biological replicate', ax=ax[0])
    sns.barplot(data=df[df['line']==i], y='migration score', x='treatment',␣
  ↪hue='biological replicate', ax=ax[1])
    ax[0].set_title('Proliferation')
    ax[1].set_title('Migration')
    plt.show()
```

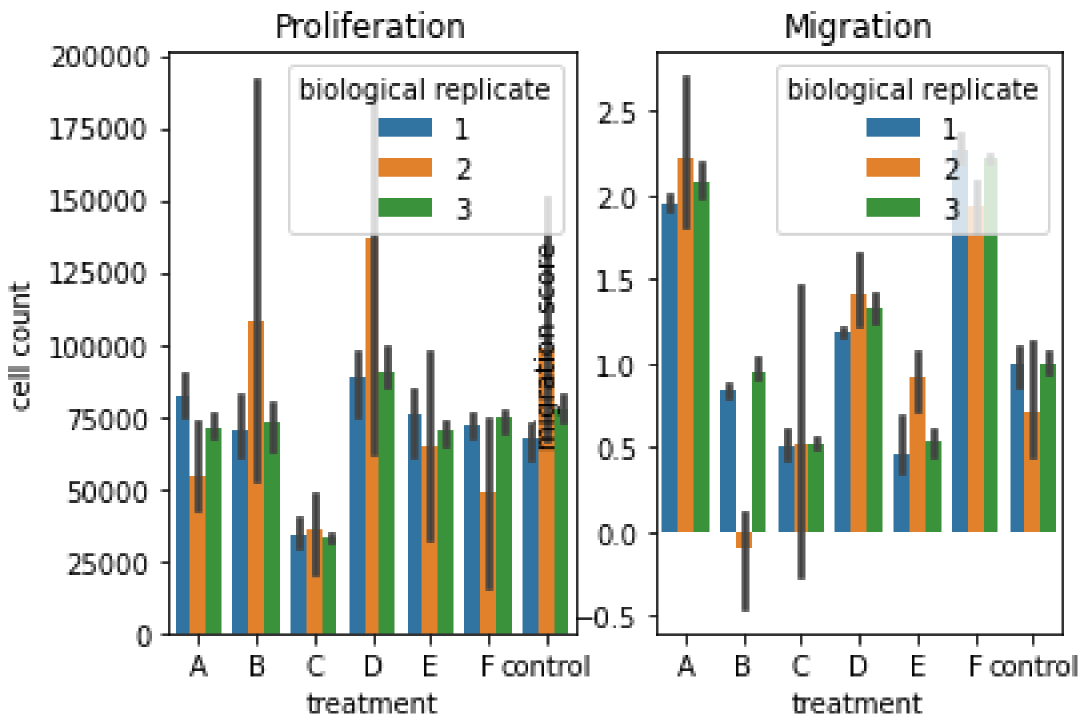


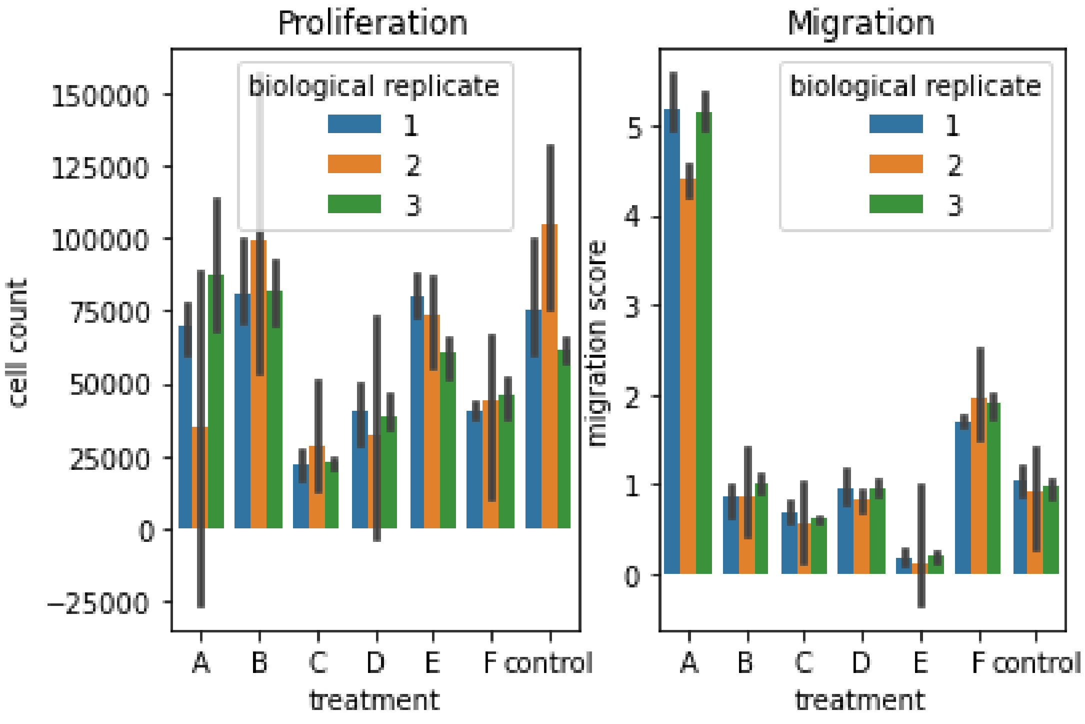

Proliferation
biological replicate
1
2
3
cell count
600000
500000
400000
300000
200000
100000
0
A
B
C
D
E
F
control
treatment


Migration
biological replicate
1
2
3
migration score
4
3
2
1
0
A
B
C
D
E
F
control
treatment


Proliferation
biological replicate
1
2
3
cell count
300000
250000
200000
150000
100000
50000
0
A
B
C
D
E
F
control
treatment


Migration
biological replicate
1
2
3
migration score
10
8
6
4
2
0
A
B
C
D
E
F
control
treatment


Proliferation
cell count
200000
175000
150000
125000
100000
75000
50000
25000
0
A
B
C
D
E
F
control
treatment
biological replicate
1
2
3
Migration
migration score
2.5
2.0
1.5
1.0
0.5
0.0
−0.5

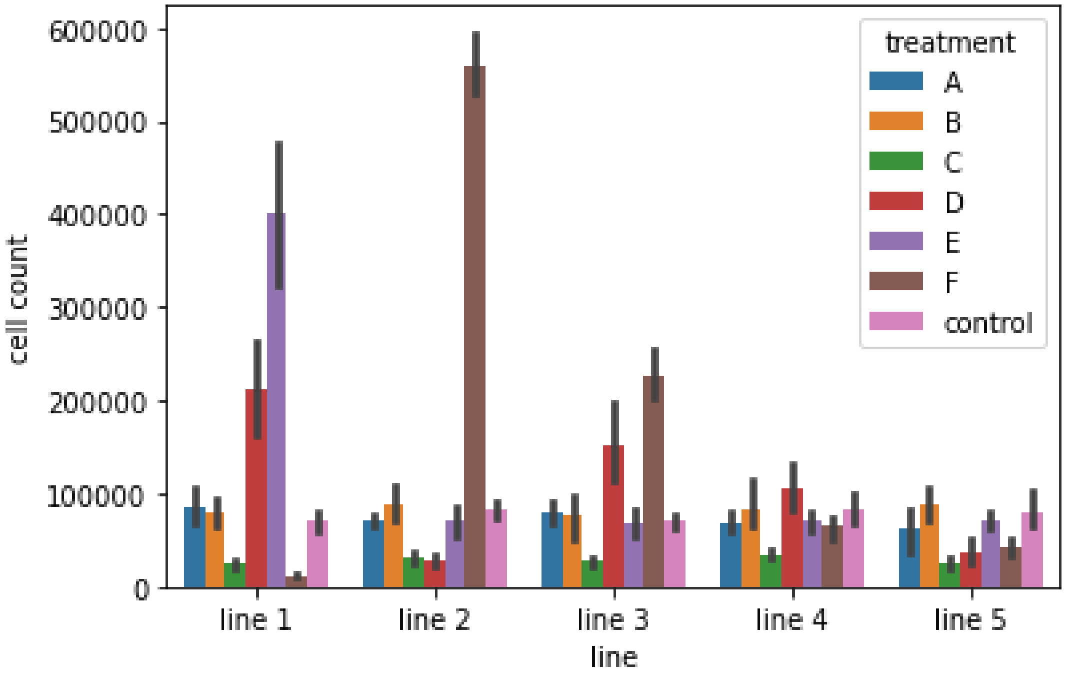

Proliferation
cell count
150000
125000
100000
75000
50000
25000
0
−25000
A
B
C
D
E
F
control
treatment
biological replicate
1
2
3
Migration
migration score
5
4
3
2
1
0

There seem to be quite some technical errors, especially in the proliferation data. However, since I am not sure of how the data was gathered and I have not been given a logbook I do not know if these points are definite outliers. It won't pose a problem since most of the means of technical replicates tend to be similar across the biological replicates.

0.3 Averaging the biological replicates

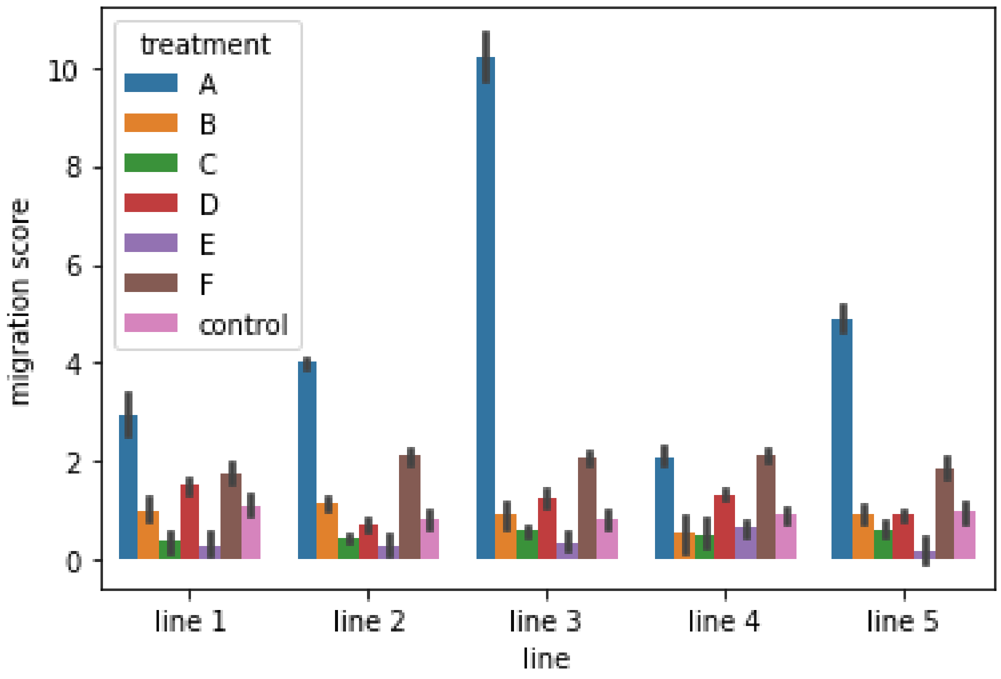
[134]:
```
sns.barplot(data=df, x='line', y='cell count', hue='treatment')
```

[134]:
```
<AxesSubplot:xlabel='line', ylabel='cell count'>
```



[135]:
```
sns.barplot(data=df, x='line', y='migration score', hue='treatment')
```

[135]:
```
<AxesSubplot:xlabel='line', ylabel='migration score'>
```



0.4 Normalizing data

The axis do have quite large values, it would be nice if we could normalize to a control of 0 of a reference cell line. We'll choose cell line 1 to be the reference of all cell lines.

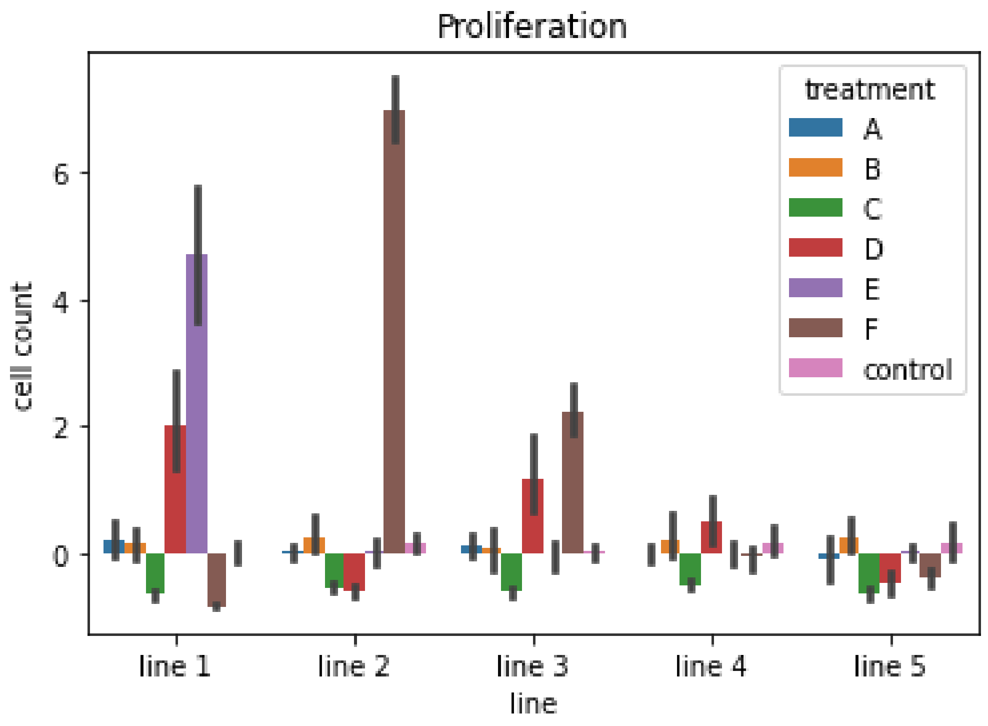
[136]:
```
Control_value = df[(df['line']=='line 1') & (df['treatment']=='control')].mean()
```

```
/tmp/ipykernel_2294/1213248773.py:1: FutureWarning: Dropping of nuisance columns
in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future
version this will raise TypeError.  Select only valid columns before calling the
reduction.
  Control_value = df[(df['line']=='line 1') &
(df['treatment']=='control')].mean()
```

[137]:
```
df['cell count'] = df['cell count']/Control_value['cell count']-1
df['migration score'] = df['migration score']/Control_value['migration score']-1
```

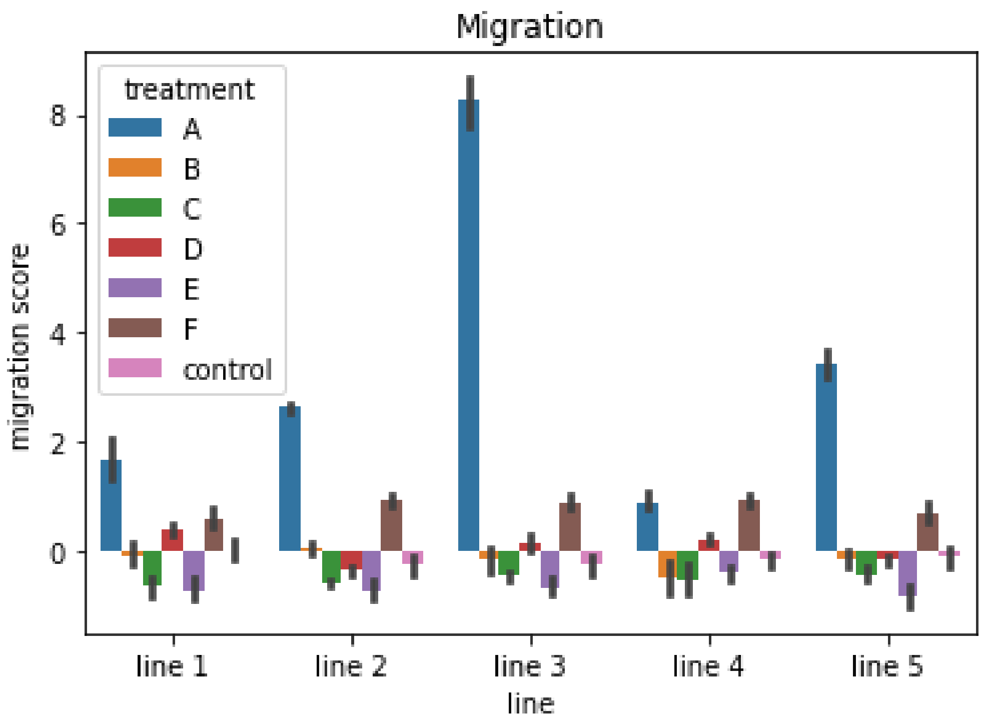
[148]:
```
sns.barplot(data=df, x='line', y='cell count', hue='treatment')
plt.title("Proliferation")
```

[148]:
```
Text(0.5, 1.0, 'Proliferation')
```



```
[149]: sns.barplot(data=df, x='line', y='migration score', hue='treatment')
       plt.title("Migration")
```

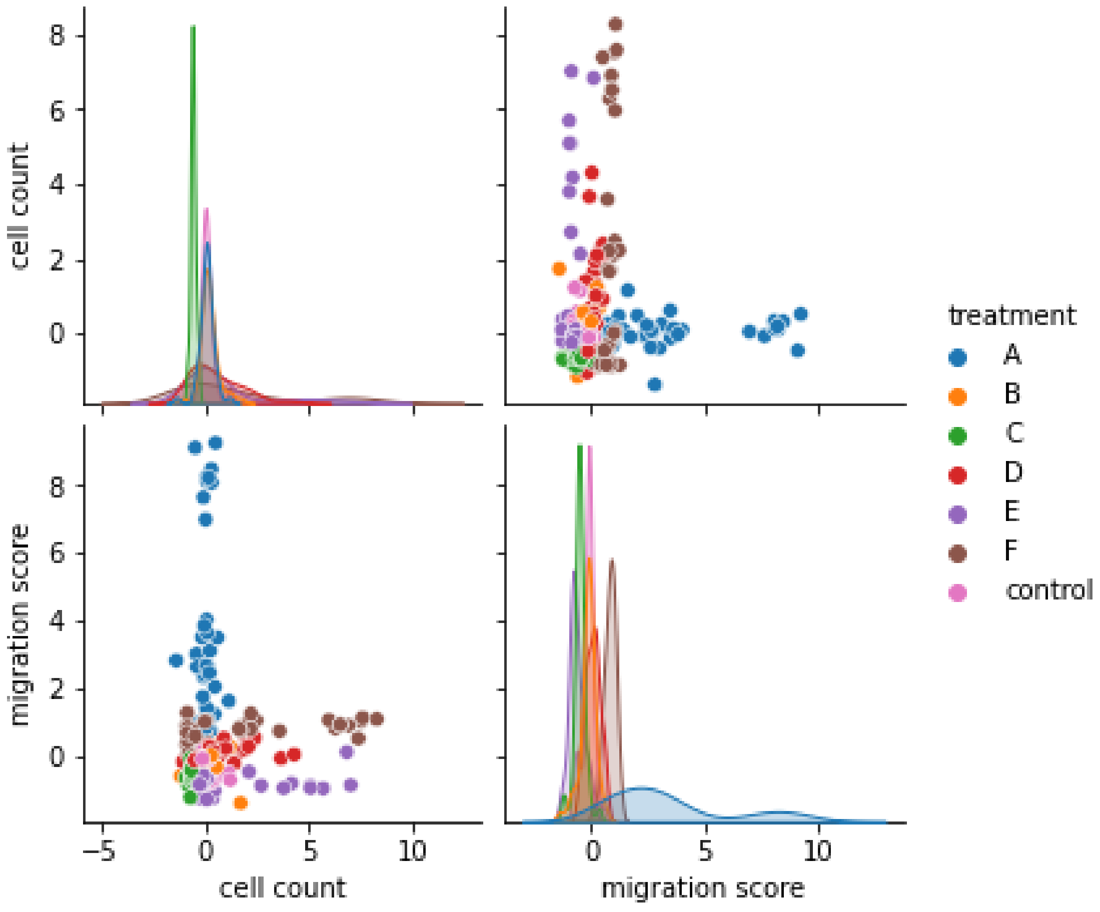
```
[149]: Text(0.5, 1.0, 'Migration')
```



These plots only show what treatments do on one of the properties of the cell. However, it would also be nice to show treatments having a correlated effect, e.g. both decrease/increase migration and proliferation

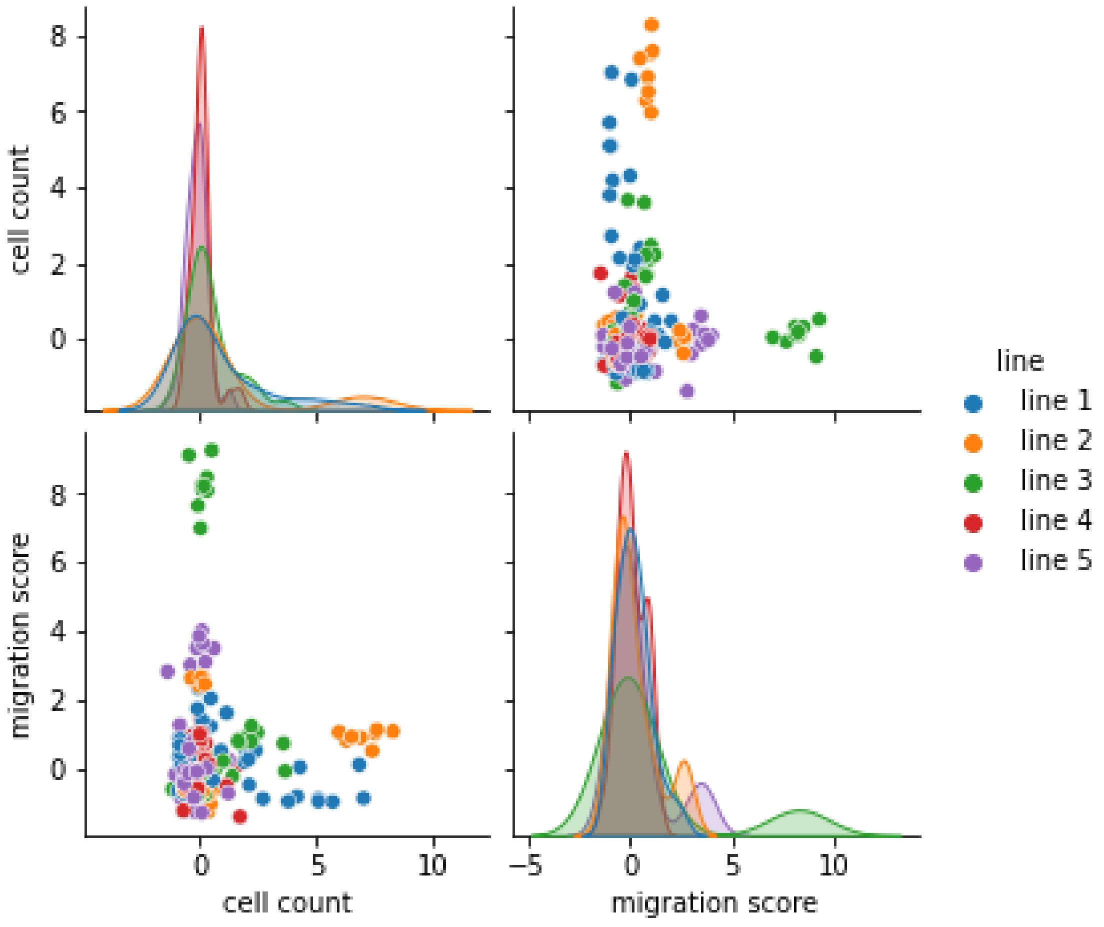
[151]:
```
pair_df = df[['cell count', 'migration score', 'treatment', 'line']]
sns.pairplot(pair_df, hue='treatment')
```

[151]:
```
<seaborn.axisgrid.PairGrid at 0x7f736e429520>
```



[152]:
```
sns.pairplot(pair_df, hue='line')
```

[152]:
```
<seaborn.axisgrid.PairGrid at 0x7f736ca02160>
```



There seems to be little correlation between migration and proliferation

0.5 Conclusion

We can conclude that treatment A increases migration quite more than other treatments in almost every cell line of which line 3 is affected the most. For the treatments that increase proliferation (D, E and F) they do it in a cell line-specific manner, sometimes increasing proliferation while sometimes decreasing it. Treatment C seems to have a similar decrease in proliferation across all cell types and also decreaeses the migration score in a similar manner across cell lines. This treatment C would be a promising therapeutic for the inhibition of proliferation and migration. However, it does decrease both properties ever so slightly, so optimilization of the treatment should be performed.

```
[ ]:
```