# Linear Models 3: Tree Graphs Lab

Dr. Luisa Barbanti and Dr. Matteo Tanadini

Applied Machine Learning and Predictive Modelling 1, FS25 (HSLU)

## Contents

# 1 Load package

```
## Load packages
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
import seaborn as sns
```

# 2 Getting data

```
# Getting data

# Load the data
d_trees = pd.read_csv("../../Datasets/TreesChamagne2017_Lab_modified.csv",
                      sep = ';', decimal = ',')

# rename variables because "." causes problems in R
d_trees.rename(columns = {'growth.rate': 'growth_rate'}, inplace = True)
d_trees.rename(columns = {'diversity.site': 'diversity_site'}, inplace = True)
d_trees.rename(columns = {'density.site': 'density_site'}, inplace = True)

# Inspect the data
print(d_trees.info())
```

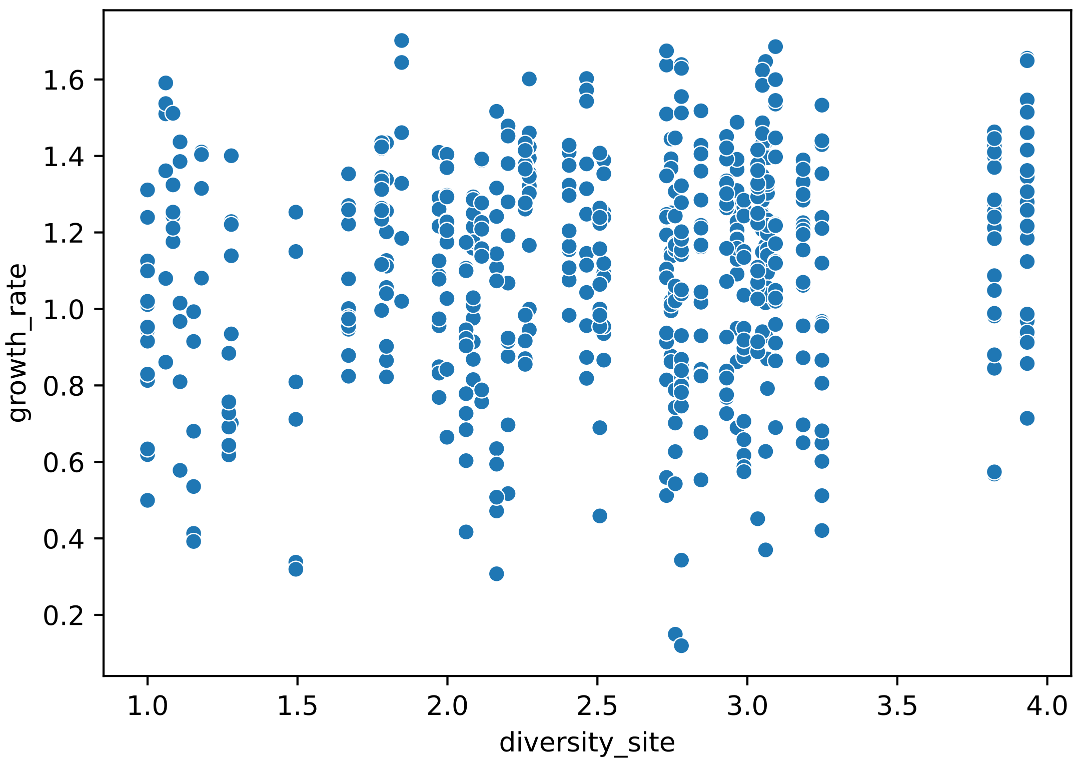
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 557 entries, 0 to 556
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   growth_rate         557 non-null    float64
 1   species             557 non-null    object 
 2   site                557 non-null    int64  
 3   Density.tree.Class  557 non-null    object 
 4   age                 557 non-null    float64
 5   size                557 non-null    float64
 6   density_site        557 non-null    float64
 7   density.tree        557 non-null    float64
 8   diversity.tree      557 non-null    float64
 9   diversity_site      557 non-null    float64
 10  sp.richness         557 non-null    int64  
 11  SiteID              557 non-null    int64  
dtypes: float64(7), int64(3), object(2)
memory usage: 52.3+ KB
None
```

```
print(d_trees.head())
```

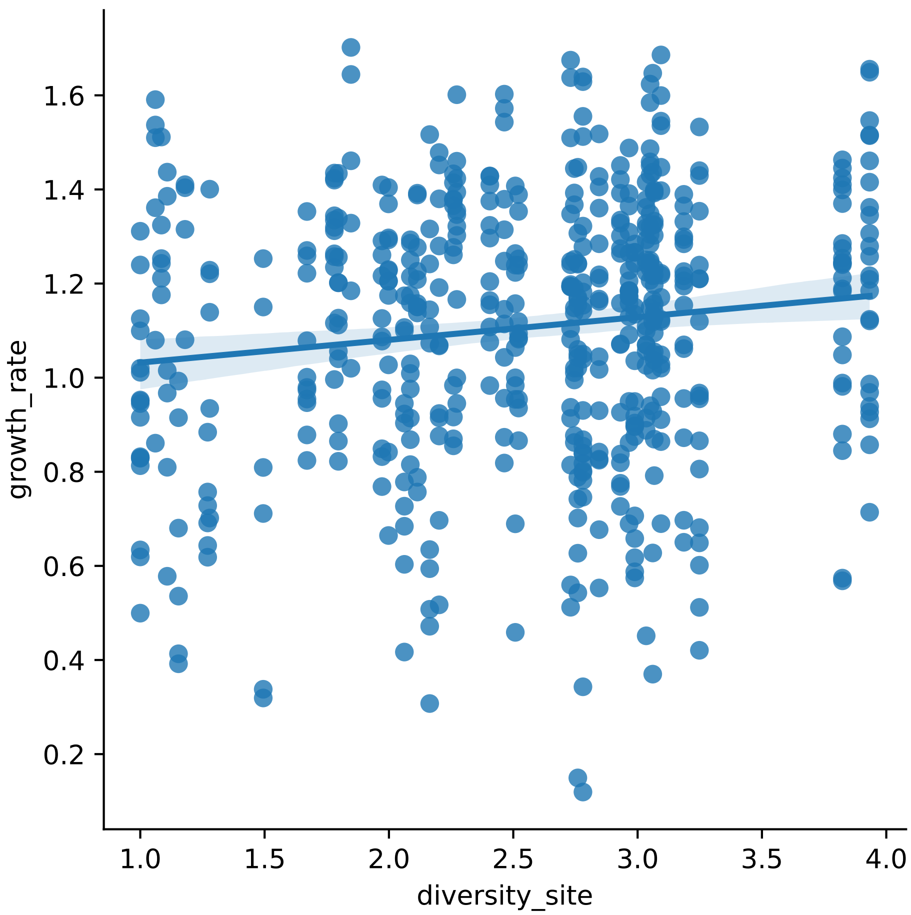
```
   growth_rate species  site  ... diversity_site  sp.richness  SiteID
```

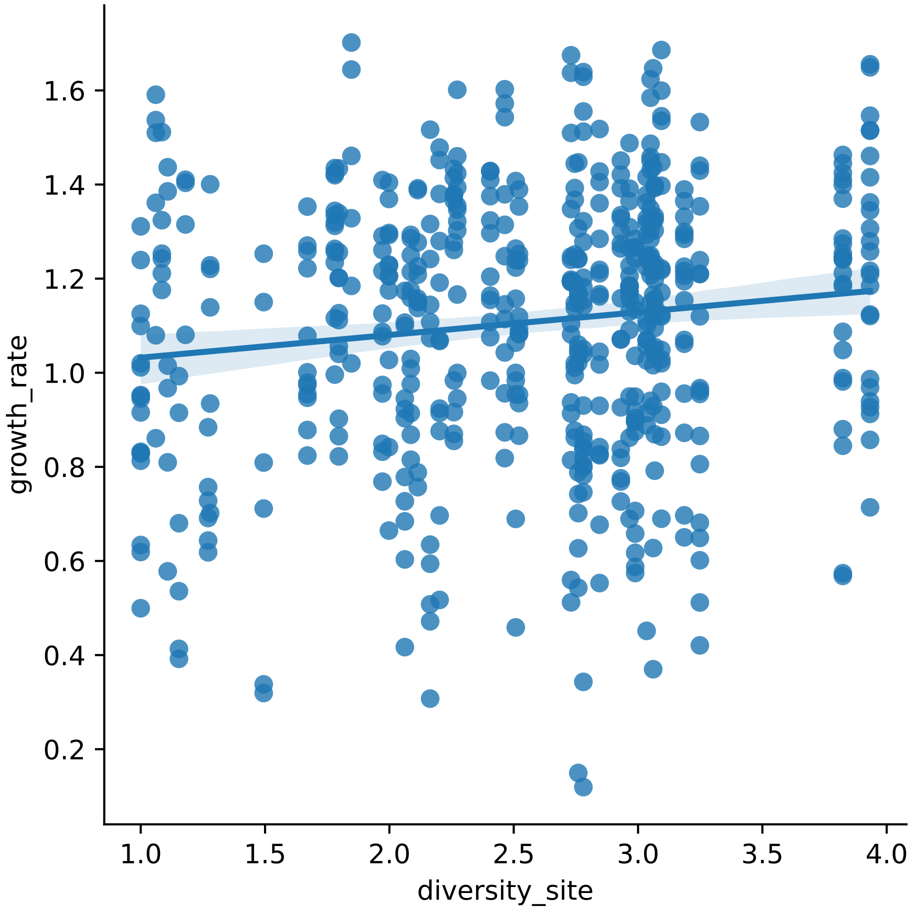
```
0     0.701705   Beech     1  ...       1.279284           1       1
1     1.138995   Beech     1  ...       1.279284           1       1
2     1.394101   Beech    12  ...       2.272922           2      12
3     0.999519  Spruce    12  ...       2.272922           2      12
4     1.354924  Spruce    12  ...       2.272922           2      12

[5 rows x 12 columns]
```

```
plt.clf()
# ----graphDiversitySite----
sns.scatterplot(x = 'diversity_site', y = 'growth_rate', data = d_trees)
```

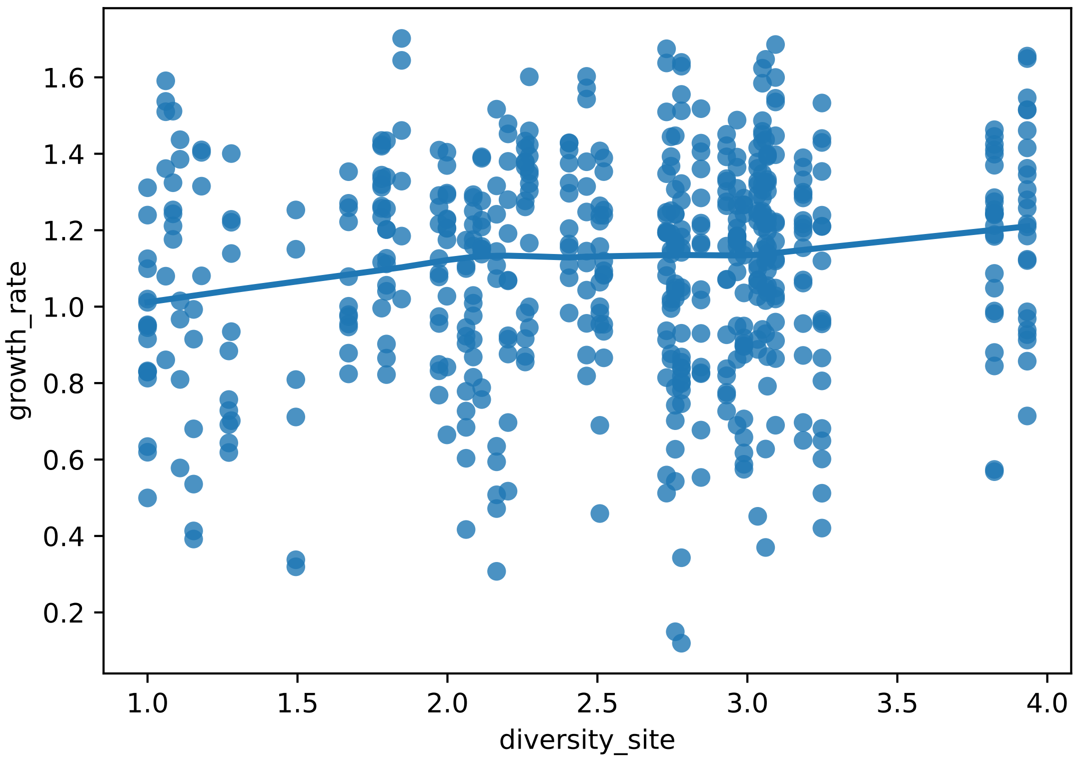



```
plt.clf()
# ----graphDiversitySiteWithRegrLine----
sns.lmplot(x = 'diversity_site', y = 'growth_rate', data = d_trees, ci = 95)
```

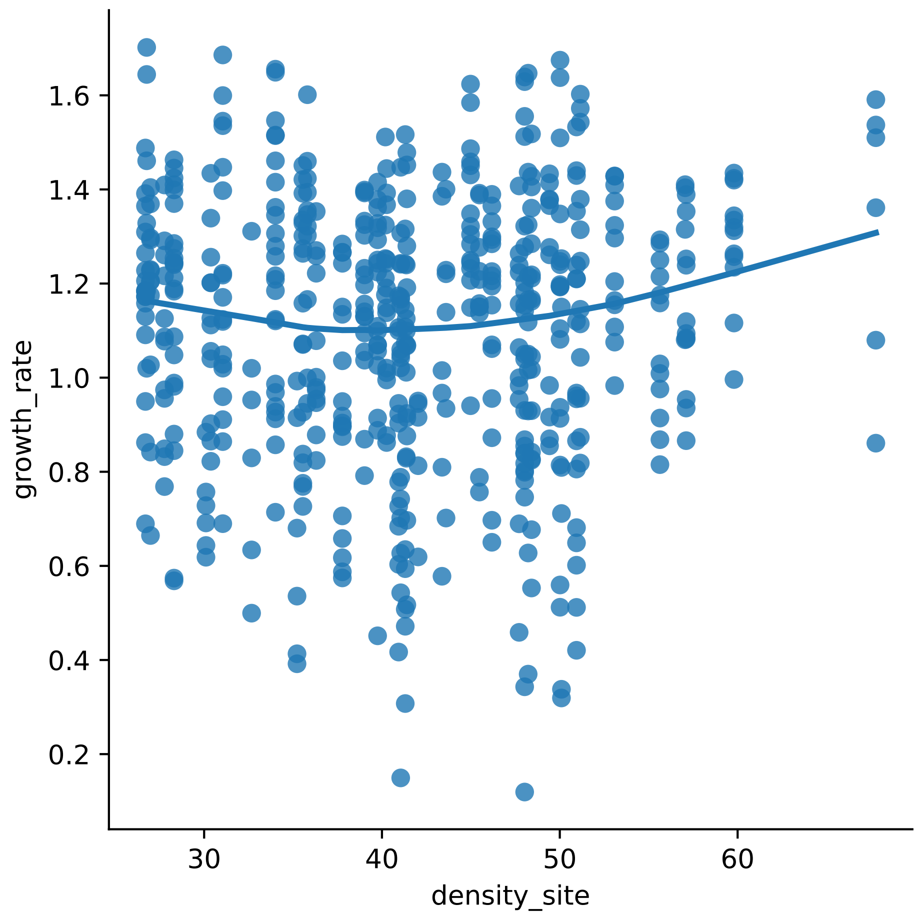

growth_rate
diversity_site
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
1.0
1.5
2.0
2.5
3.0
3.5
4.0



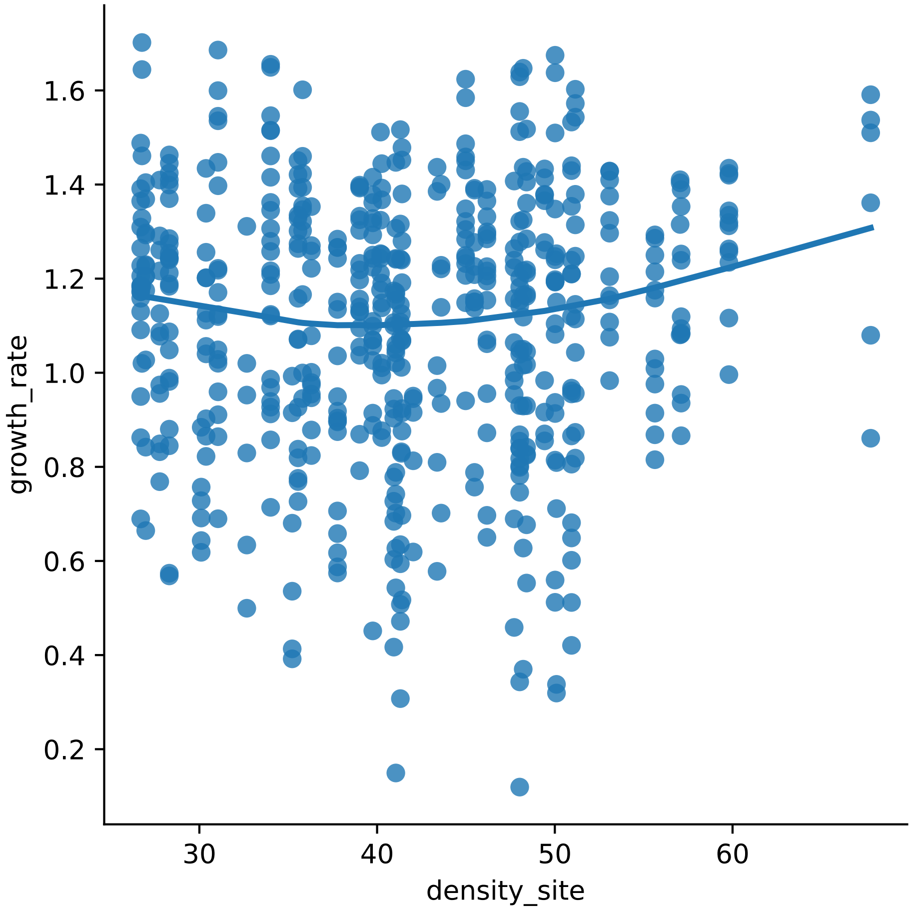

```
plt.clf()
# ----graphDiversitySiteWithSmoother----
sns.regplot(x = 'diversity_site', y = 'growth_rate', data = d_trees,
            lowess = True)
## CI are not supported when lowess = True
```



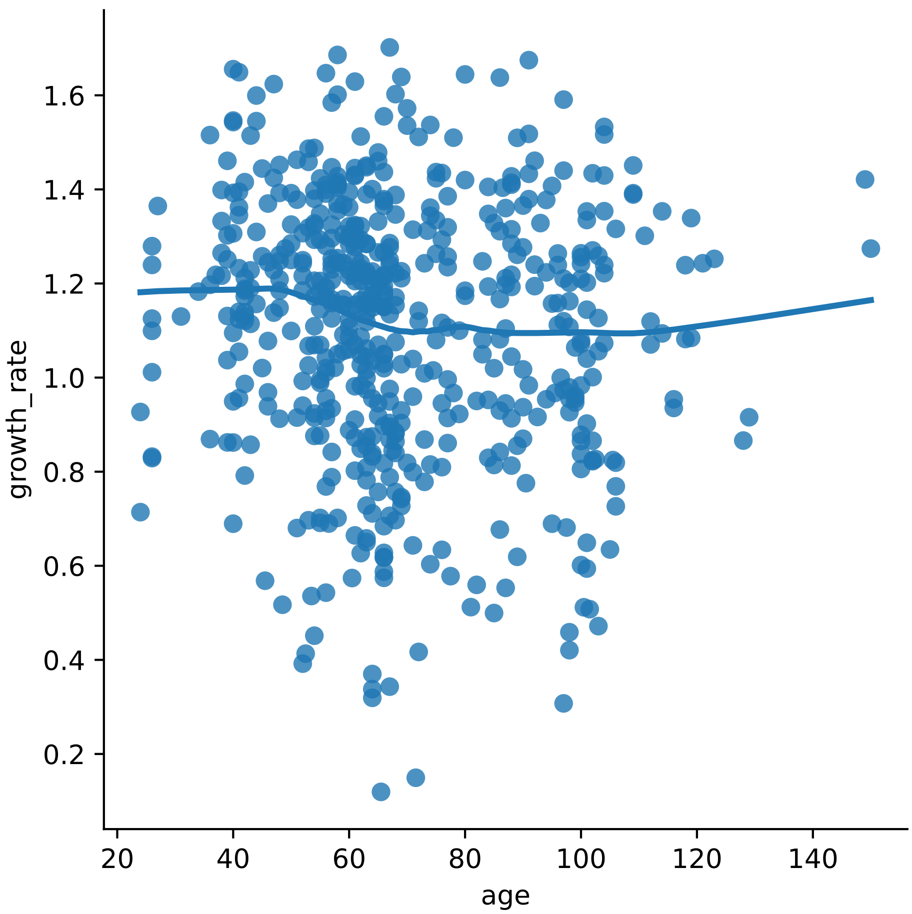

```
plt.clf()
# ----graphDensity----
sns.lmplot(x = 'density_site', y = 'growth_rate', data = d_trees, lowess = True)
```



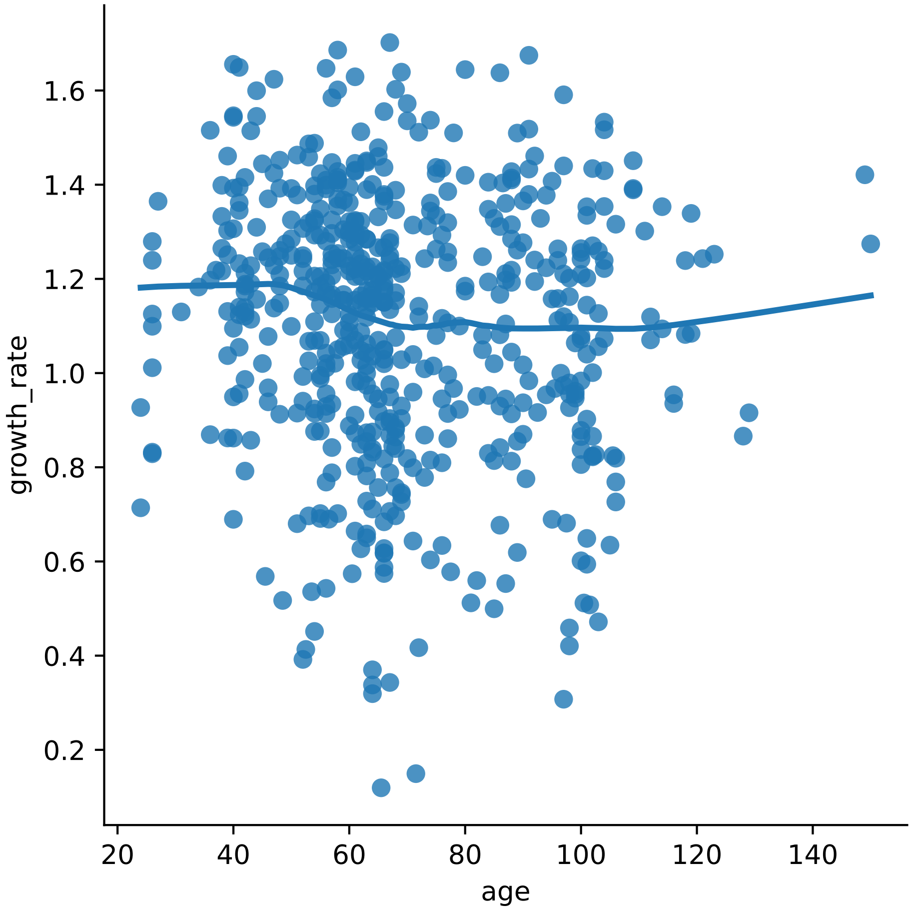
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
growth_rate
30
40
50
60
density_site



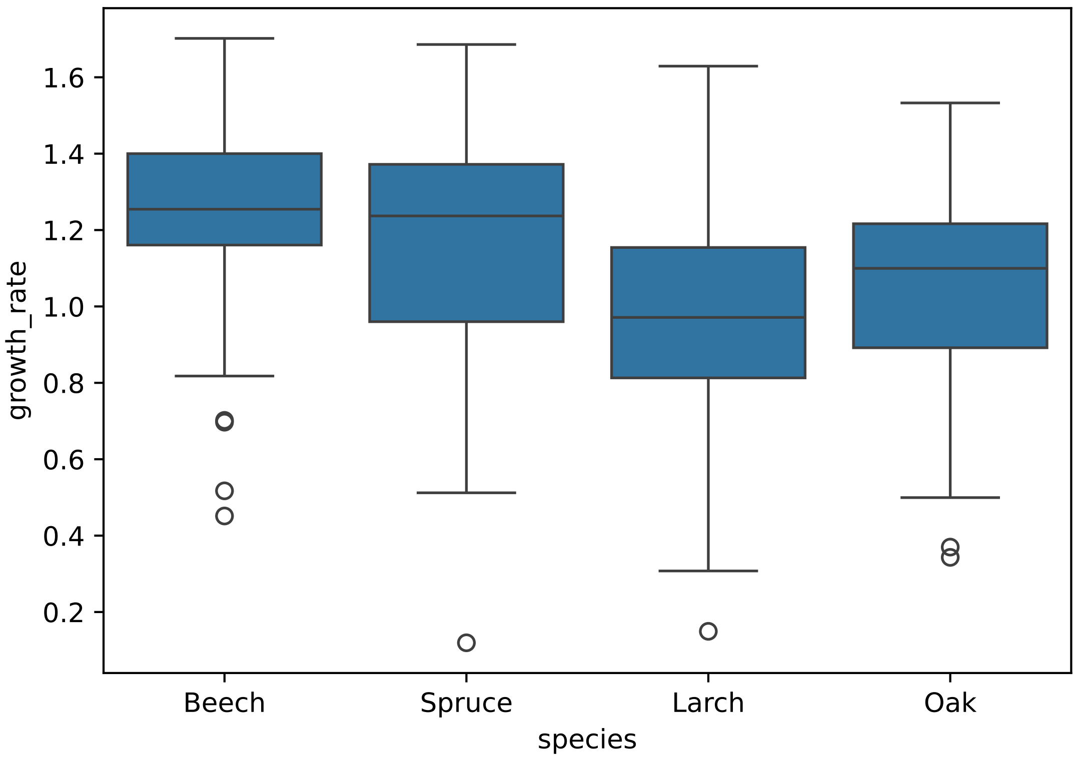

```
plt.clf()
# ----graphAge----
sns.lmplot(x = 'age', y = 'growth_rate', data = d_trees, lowess = True)
```



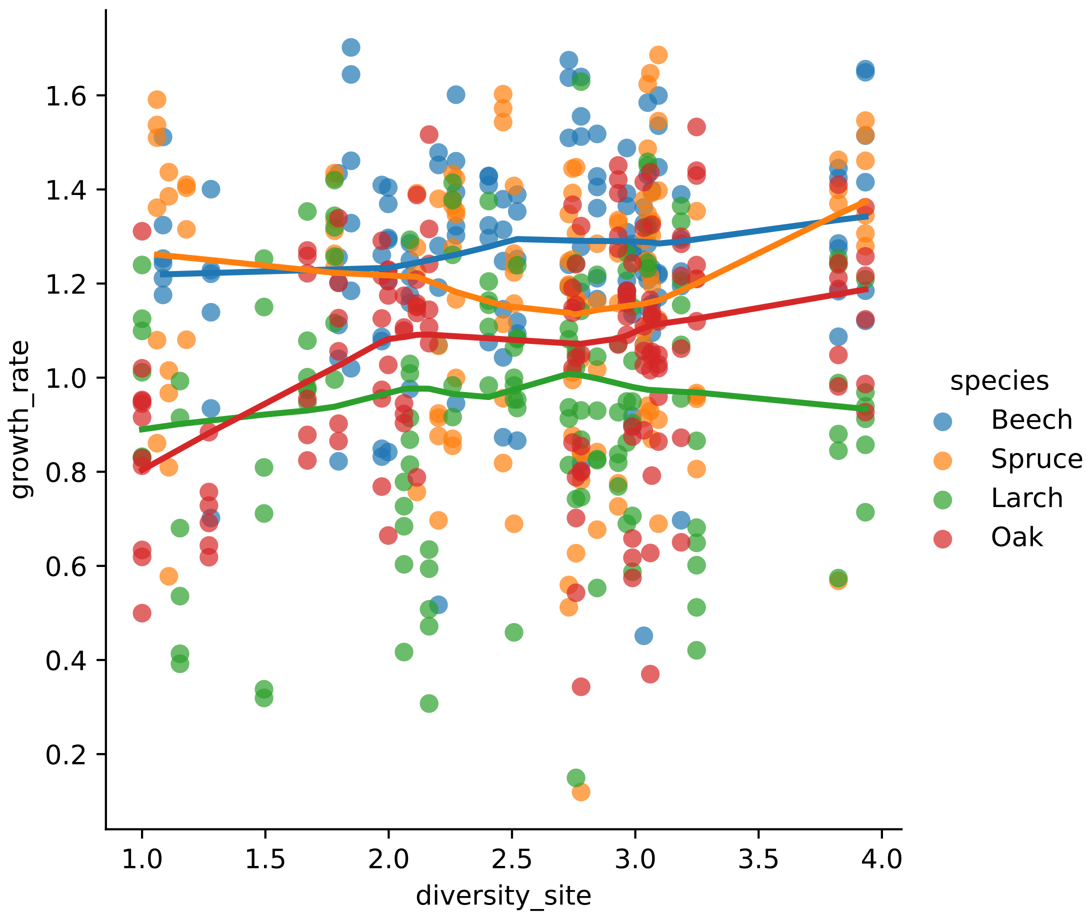
growth_rate
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
20
40
60
80
100
120
140
age

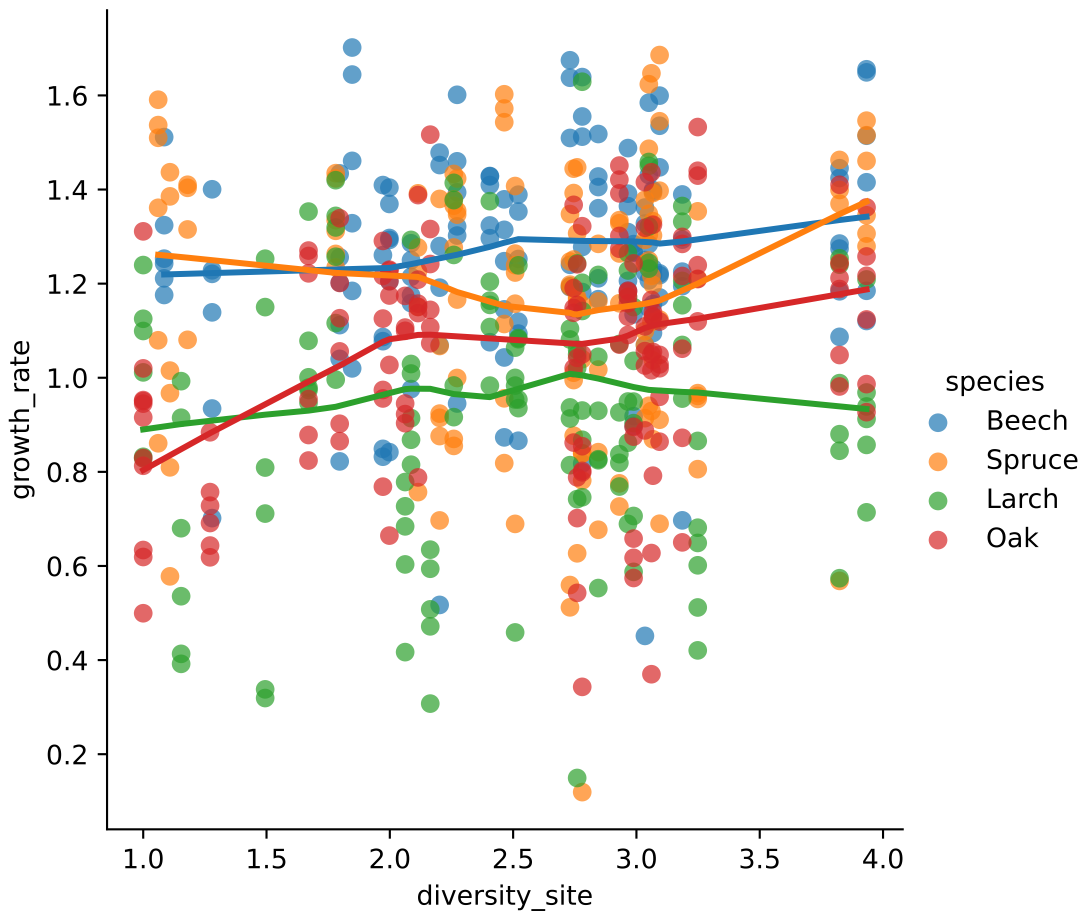



```
# ----GraphSpecies----
sns.boxplot(x = 'species', y = 'growth_rate', data = d_trees)
```

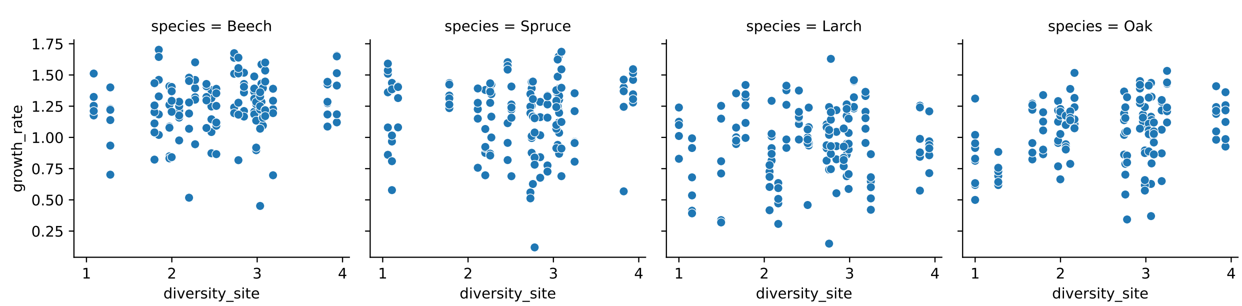



```
plt.clf()
# ----graphDiversityGrouped----
sns.lmplot(x = 'diversity_site', y = 'growth_rate', hue = 'species',
    data = d_trees, lowess = True, scatter_kws = {'alpha': 0.7}
)
```

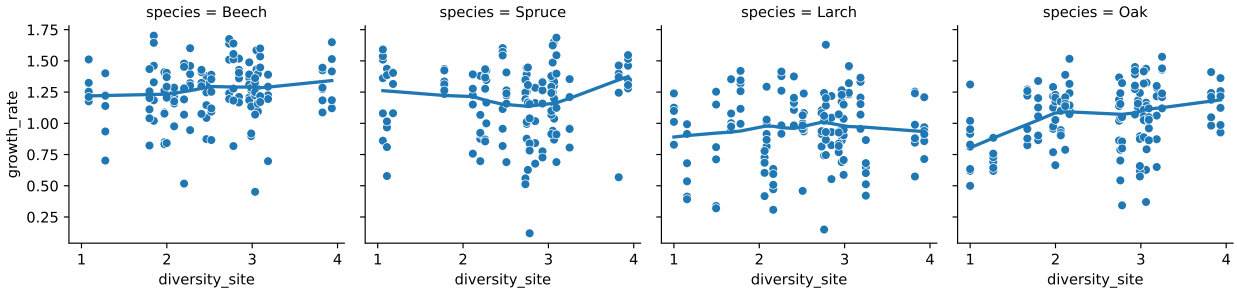

growth_rate
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
1.0
1.5
2.0
2.5
3.0
3.5
4.0
diversity_site
species
Beech
Spruce
Larch
Oak

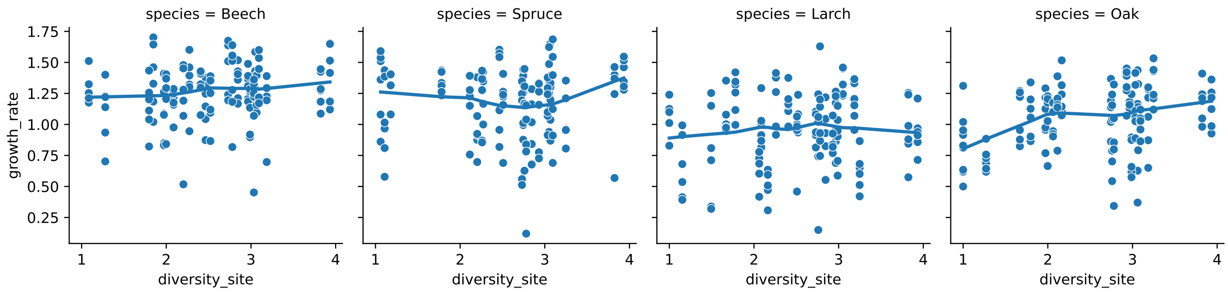



```python
plt.clf()
# ----graphDiversityPanelling----
g = sns.FacetGrid(d_trees, col = 'species')
g.map_dataframe(sns.scatterplot, x = 'diversity_site', y = 'growth_rate')
```

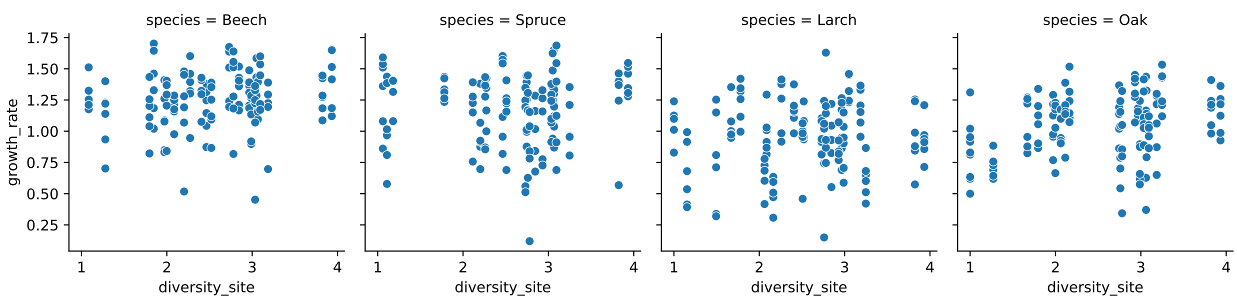



```python
g.map_dataframe(sns.regplot, x = 'diversity_site', y = 'growth_rate',
                scatter = False, ci = None, lowess = True)
```

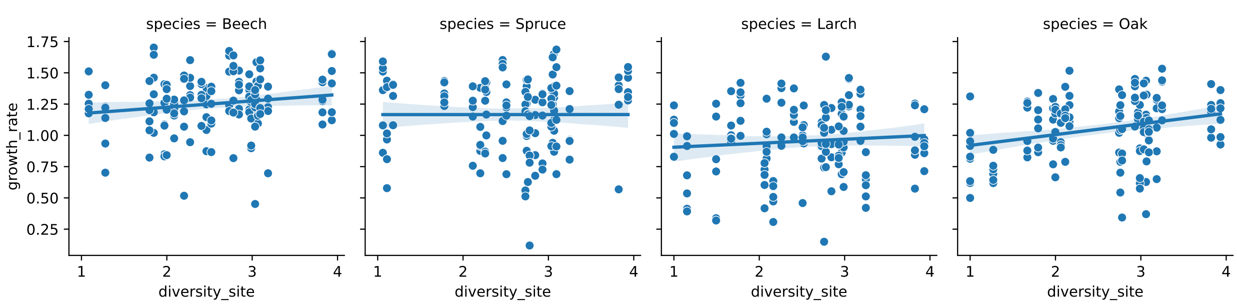


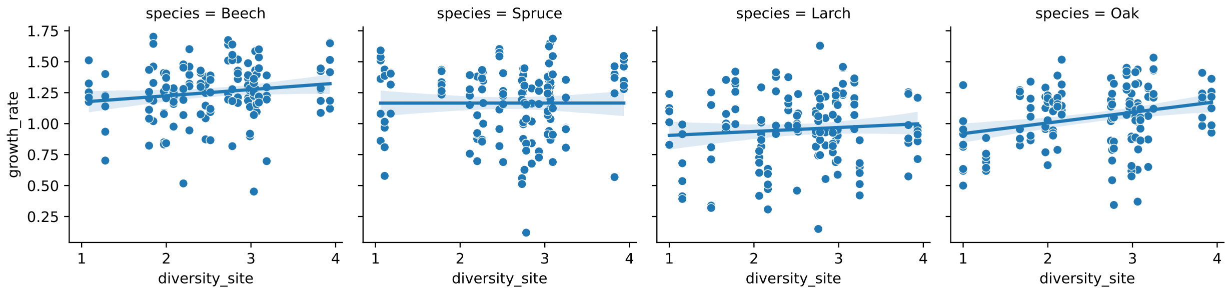




```
plt.clf()
# ----graphDiversityPanellingRegressionLines----
g = sns.FacetGrid(d_trees, col = 'species')
g.map_dataframe(sns.scatterplot, x = 'diversity_site', y = 'growth_rate')
```




```
g.map_dataframe(sns.regplot, x = 'diversity_site', y = 'growth_rate',
                scatter = False, ci = 95)
```






```
plt.clf()
# ----lm0----
lm_trees_0 = smf.ols(
    'growth_rate ~ species + age + density_site + diversity_site + species:age + species:density_site +
    data = d_trees
).fit()
print(lm_trees_0.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            growth_rate   R-squared:                       0.237
Model:                            OLS   Adj. R-squared:                  0.216
Method:                 Least Squares   F-statistic:                     11.22
Date:                Tue, 25 Feb 2025   Prob (F-statistic):           5.33e-24
Time:                        10:05:34   Log-Likelihood:                -6.1818
No. Observations:                 557   AIC:                             44.36
Df Residuals:                     541   BIC:                             113.5
Df Model:                          15
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                       1.1452      0.139      8.216      0.000       0.871       1.419
species[T.Larch]               -0.6053      0.206     -2.942      0.003      -1.009      -0.201
species[T.Oak]                 -0.5521      0.199     -2.770      0.006      -0.944      -0.161
species[T.Spruce]               0.0158      0.266      0.060      0.952      -0.506       0.538
age                             0.0002      0.001      0.177      0.860      -0.002       0.003
species[T.Larch]:age           -0.0036      0.002     -2.233      0.026      -0.007      -0.000
species[T.Oak]:age              0.0034      0.002      2.131      0.034       0.000       0.006
species[T.Spruce]:age          -0.0033      0.002     -1.837      0.067      -0.007       0.000
density_site                   -0.0009      0.002     -0.407      0.684      -0.005       0.004
species[T.Larch]:density_site   0.0119      0.004      3.275      0.001       0.005       0.019
species[T.Oak]:density_site    -0.0026      0.004     -0.682      0.495      -0.010       0.005
species[T.Spruce]:density_site  0.0053      0.004      1.305      0.192      -0.003       0.013
diversity_site                  0.0513      0.031      1.652      0.099      -0.010       0.112
species[T.Larch]:diversity_site 0.0205      0.043      0.475      0.635      -0.064       0.105
species[T.Oak]:diversity_site   0.0812      0.043      1.896      0.059      -0.003       0.165
species[T.Spruce]:diversity_site -0.0459    0.050     -0.922      0.357      -0.144       0.052
==============================================================================
Omnibus:                       37.526   Durbin-Watson:                   1.784
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               44.291
Skew:                          -0.621   Prob(JB):                     2.41e-10
```

```
Kurtosis:                       3.606   Cond. No.                     2.87e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.87e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```python
anova = sm.stats.anova_lm(lm_trees_0)
print(anova)
```

```
                           df     sum_sq   mean_sq          F        PR(>F)
species                   3.0   7.142050  2.380683  38.625859  1.248546e-22
age                       1.0   0.039165  0.039165   0.635437  4.257177e-01
species:age               3.0   0.606157  0.202052   3.278235  2.076423e-02
density_site              1.0   0.282238  0.282238   4.579225  3.280851e-02
species:density_site      3.0   0.541740  0.180580   2.929856  3.315005e-02
diversity_site            1.0   1.296585  1.296585  21.036700  5.603738e-06
species:diversity_site    3.0   0.466217  0.155406   2.521408  5.707840e-02
Residual                541.0  33.344233  0.061634        NaN           NaN
```

```python
# ----drop1InteractionLm0----
lm_trees_1 = smf.ols(
    'growth_rate ~ species + age + density_site + diversity_site + species:age + species:density_site',
```

```
    data = d_trees
).fit()
print(lm_trees_1.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:            growth_rate   R-squared:                       0.227
Model:                            OLS   Adj. R-squared:                  0.210
Method:                 Least Squares   F-statistic:                     13.28
Date:                Tue, 25 Feb 2025   Prob (F-statistic):           3.66e-24
Time:                        10:05:34   Log-Likelihood:                -10.049
No. Observations:                 557   AIC:                             46.10
Df Residuals:                     544   BIC:                             102.3
Df Model:                          12                                         
Covariance Type:            nonrobust                                         
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                          1.0788      0.113      9.518      0.000       0.856       1.302
species[T.Larch]                  -0.5427      0.153     -3.539      0.000      -0.844      -0.241
species[T.Oak]                    -0.3333      0.159     -2.098      0.036      -0.645      -0.021
species[T.Spruce]                 -0.2583      0.155     -1.662      0.097      -0.564       0.047
age                                0.0004      0.001      0.346      0.730      -0.002       0.003
species[T.Larch]:age              -0.0038      0.002     -2.379      0.018      -0.007      -0.001
species[T.Oak]:age                 0.0023      0.002      1.550      0.122      -0.001       0.005
species[T.Spruce]:age             -0.0033      0.002     -1.861      0.063      -0.007       0.000
density_site                      -0.0010      0.002     -0.427      0.669      -0.006       0.004
species[T.Larch]:density_site      0.0120      0.004      3.353      0.001       0.005       0.019
species[T.Oak]:density_site       -0.0010      0.004     -0.267      0.789      -0.008       0.006
species[T.Spruce]:density_site     0.0087      0.004      2.380      0.018       0.002       0.016
diversity_site                     0.0729      0.016      4.567      0.000       0.042       0.104
==============================================================================
Omnibus:                       37.580   Durbin-Watson:                   1.745
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               44.393
Skew:                          -0.621   Prob(JB):                     2.29e-10
Kurtosis:                       3.610   Cond. No.                     1.99e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.99e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
anova_1 = sm.stats.anova_lm(lm_trees_1)
print(anova_1)
```

```
                         df    sum_sq   mean_sq          F        PR(>F)
species                 3.0  7.142050  2.380683  38.304480  1.804728e-22
age                     1.0  0.039165  0.039165   0.630150  4.276464e-01
species:age             3.0  0.606157  0.202052   3.250959  2.153729e-02
density_site            1.0  0.282238  0.282238   4.541124  3.353759e-02
species:density_site    3.0  0.541740  0.180580   2.905479  3.424331e-02
```

| | | | | | |
|---|---|---|---|---|---|
| diversity_site | 1.0 | 1.296585 | 1.296585 | 20.861668 | 6.111838e-06 |
| Residual | 544.0 | 33.810449 | 0.062152 | NaN | NaN |