In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats
```

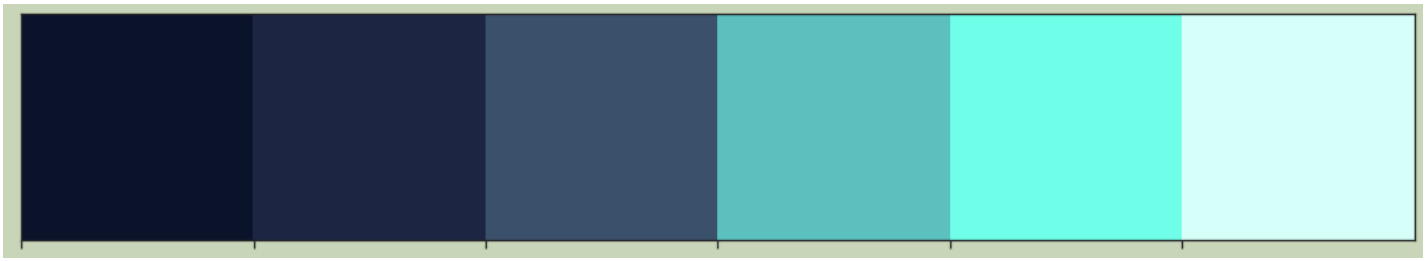
# Setting Visualization Theme

In [2]:
```
pd.set_option('display.max_rows', 500)
pd.set_option('display.max_columns', 20)

colors = ["#0B132B","#1C2541","#3A506B","#5BC0BE","#6FFFE9", "#d6fff9"]

sns.set(palette=colors, style='ticks',
        rc={'axes.facecolor':'#f0f7f4', 'figure.facecolor':'#C8D5B9', 'figure.figsize':(16
            'font.family':'Product Sans', 'patch.linewidth': 0.0, 'font.size':12}, )

sns.palplot(colors, size=3)
```



# Reading Data

In [3]:
```
df = pd.read_csv('data/clean.csv')
df.head()
```

Out[3]:

| | age | gender | year | course | gwa | openness_to_experience | conscientiousness | extroversion | agreeableness | ne |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 21 | 1 | 3 | 23 | 1.21 | 0.500000 | 0.62500 | 0.567568 | 0.423077 | |
| **1** | 22 | 0 | 3 | 23 | 1.43 | 0.583333 | 0.87500 | 0.162162 | 0.230769 | |
| **2** | 21 | 1 | 3 | 23 | 1.30 | 0.583333 | 0.75000 | 0.918919 | 0.615385 | |
| **3** | 22 | 1 | 3 | 10 | 1.75 | 0.541667 | 0.59375 | 0.459459 | 0.538462 | |
| **4** | 20 | 0 | 2 | 42 | 1.75 | 0.166667 | 0.28125 | 0.405405 | 0.384615 | |

# Checking for Normality

Shapiro-Wilk Test - The Shapiro-Wilk test is a way to tell if a random sample comes from a normal distribution. The test gives you a W value; small values indicate your sample is not normally distributed (you can reject the null hypothesis that your population is normally distributed if your values are under a certain threshold).

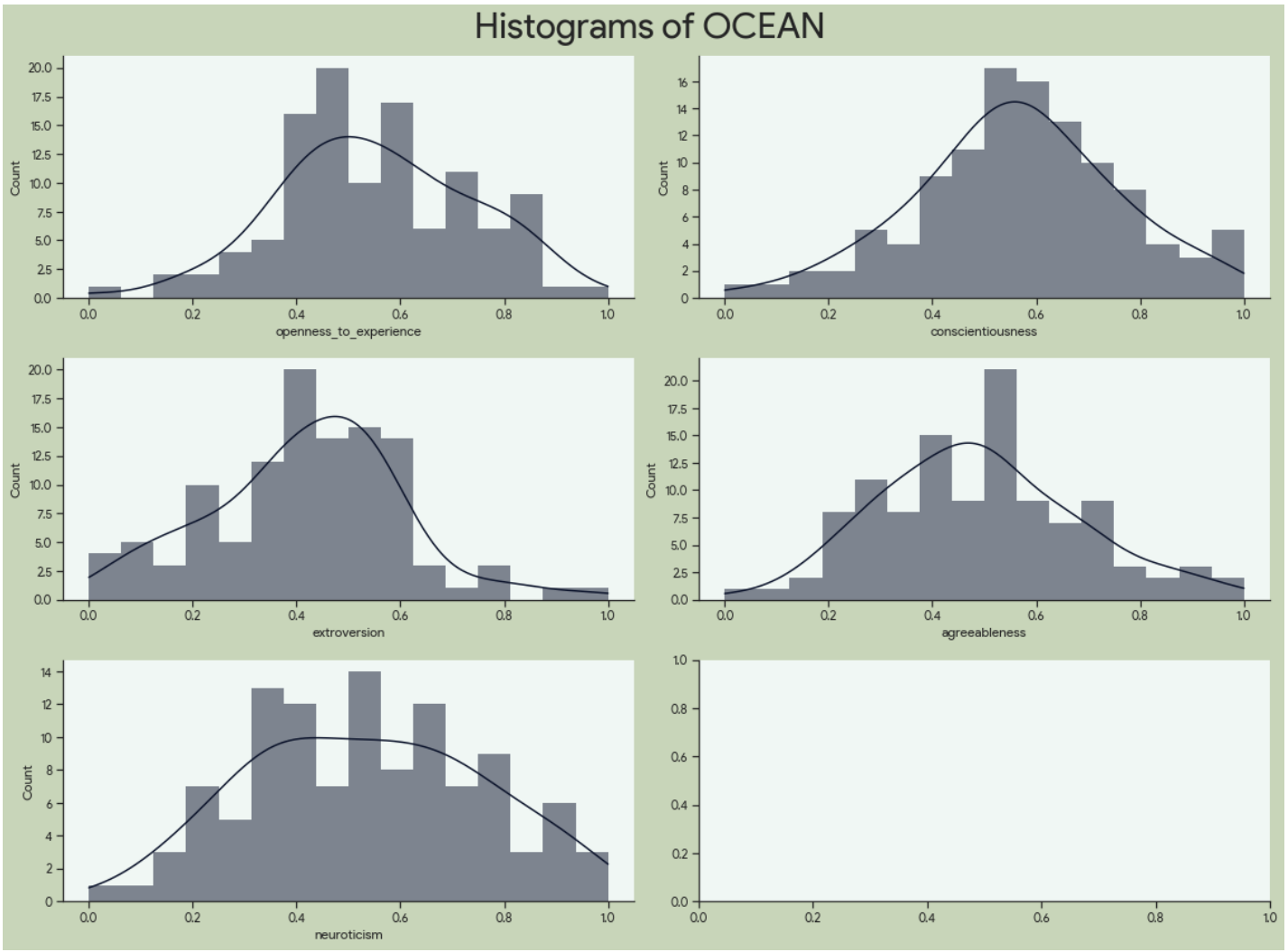
1. alpha: 0.05
2. H0: Data is normally distributed
3. H1: Data is not normally distributed

In [4]:

```
fig, ax = plt.subplots(3,2)
ax = ax.flatten()
fig.suptitle('Histograms of OCEAN', fontsize=32)
i=0

for x in df.columns[5:]:
    sns.histplot(df[x], kde=True, bins=16, ax=ax[i])
    i += 1

fig.tight_layout()
sns.despine()
plt.savefig('figures/Histograms of OCEAN.jpg')
```




In [5]:

```
for x in df.columns[5:]:
    stat, p = stats.shapiro(df[x])
    if p > 0.05:
        print(f'feature-{x} : looks Gaussian (fail to reject H0)')
    else:
        print(f'feature-{x} : does not look Gaussian (reject H0)')
```

```
feature-openness_to_experience : looks Gaussian (fail to reject H0)
feature-conscientiousness : looks Gaussian (fail to reject H0)
feature-extroversion : does not look Gaussian (reject H0)
feature-agreeableness : looks Gaussian (fail to reject H0)
feature-neuroticism : looks Gaussian (fail to reject H0)
```

# ANOVA

ANOVA, which stands for Analysis of Variance, is a statistical test used to analyze the difference between the means of more than two groups. Checking the OCEAN features if they share the same kind of information or they are entirely different from one another.

1. H0: Groups means are equal (no variation in means of groups)
2. H1: At least, one group mean is different from other groups

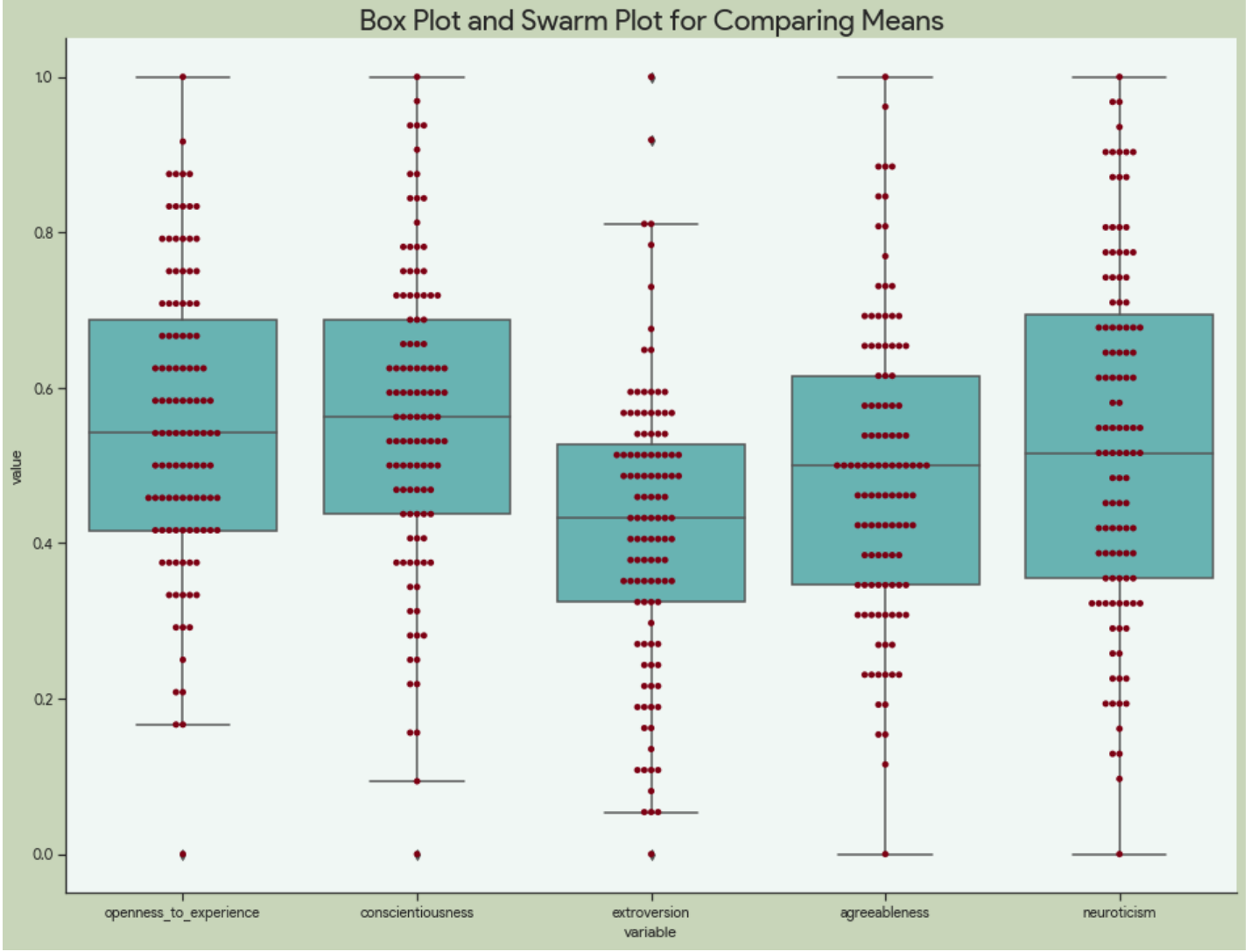
In [6]:
```
melted =df[['openness_to_experience','conscientiousness','extroversion','agreeableness','r
melted
```

Out[6]:

| | variable | value |
|---|---|---|
| **0** | openness_to_experience | 0.500000 |
| **1** | openness_to_experience | 0.583333 |
| **2** | openness_to_experience | 0.583333 |
| **3** | openness_to_experience | 0.541667 |
| **4** | openness_to_experience | 0.166667 |
| **...** | ... | ... |
| **550** | neuroticism | 0.483871 |
| **551** | neuroticism | 0.806452 |
| **552** | neuroticism | 0.354839 |
| **553** | neuroticism | 0.516129 |
| **554** | neuroticism | 0.741935 |

555 rows × 2 columns

In [7]:
```
plt.title('Box Plot and Swarm Plot for Comparing Means', fontsize=22)
sns.boxplot(x='variable', y='value', data=melted, color=colors[3])
sns.swarmplot(x="variable", y="value", data=melted, color='#7d0013')
sns.despine()
plt.savefig('figures/anova.jpg')
```




In [8]:

```
crit, p = stats.f_oneway(df['openness_to_experience'], df['conscientiousness'], df['extrov
                df['agreeableness'], df['neuroticism'])
if p < 0.05:
    print('Reject the null hypothesis: There are variation in groups')
else:
    print('Failed to reject the null hypothesis: No variation in groups')
```

```
Reject the null hypothesis: There are variation in groups
```

# Chi SQUARE

This is a test for the independence of different categories of a population.

1. H0: No Variation in groups
2. H1: There are variation in groups

In [9]:

```
tab = pd.crosstab(df['gender'], df['year'])
crit, p, dof, exp = stats.chi2_contingency(tab)

if p < 0.05:
    print('Reject the null hypothesis: There are variation in groups')
else:
    print('Failed to reject the null hypothesis: No variation in groups')
```

```
Failed to reject the null hypothesis: No variation in groups
```

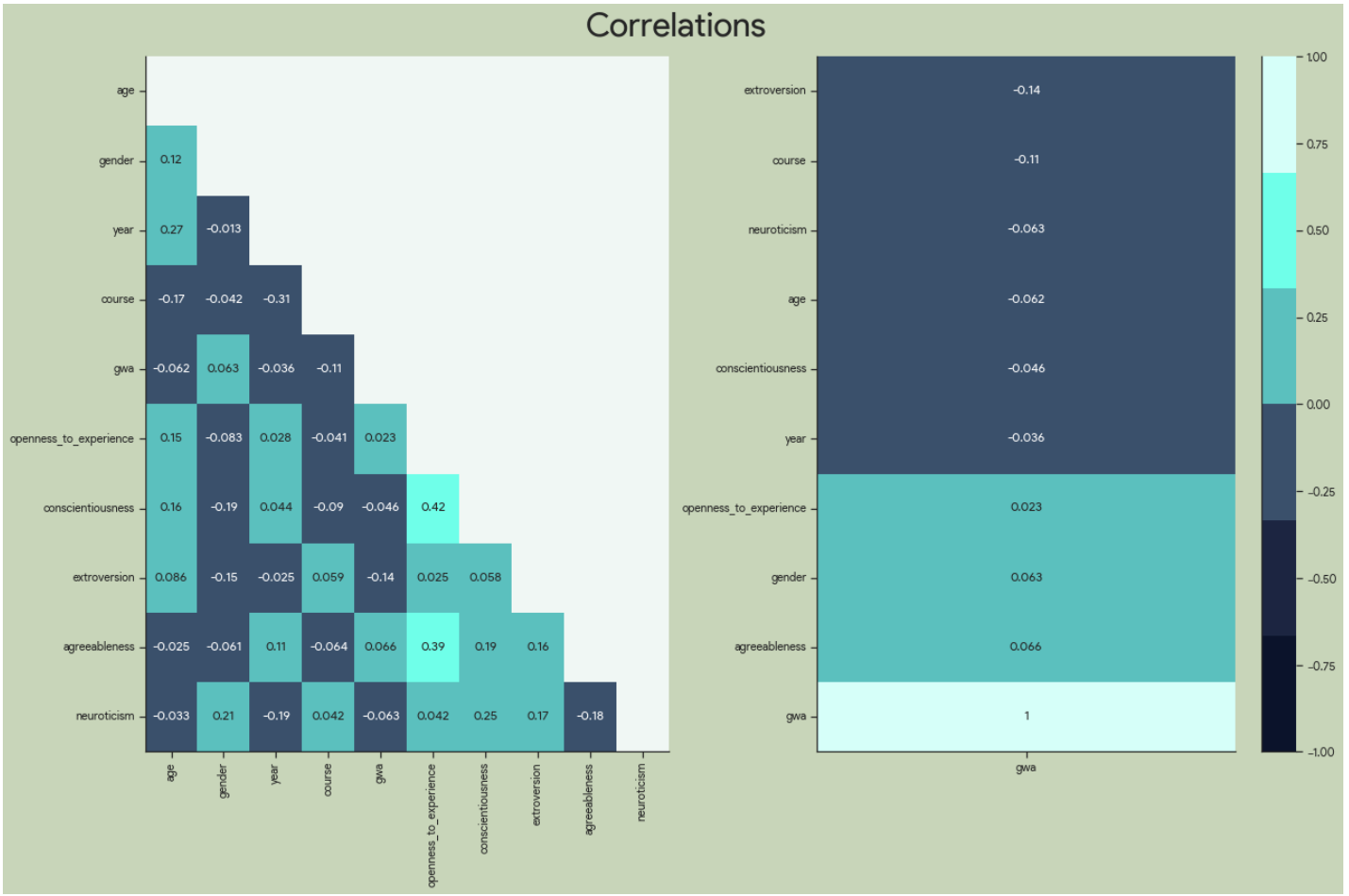
# Correlations between each variable

In [10]:
```
fig, ax = plt.subplots(1,2, figsize=(18,12))
ax = ax.flatten()
fig.suptitle('Correlations', fontsize=32)
i=0

mask = np.triu(df.corr())
sns.heatmap(df.corr(), mask=mask, annot=True, vmin=-1, vmax=1, cmap=colors, ax=ax[i], cbar
i += 1
sns.heatmap(df.corr()[['gwa']].sort_values('gwa'), annot=True, vmin=-1, vmax=1, cmap=color

fig.tight_layout()
sns.despine()

plt.savefig('figures/correlations.jpg')
```




# Insights

1. There seems to be a correlation between openness to experience and agreeableness.
2. There seems to be a correlation between openness to experience and conscientiousness.
3. There seems to be no correlation between the GWA and its demographics.
4. There seems to be no correlation between the GWA and OCEAN with extroversion being slightly correlated.

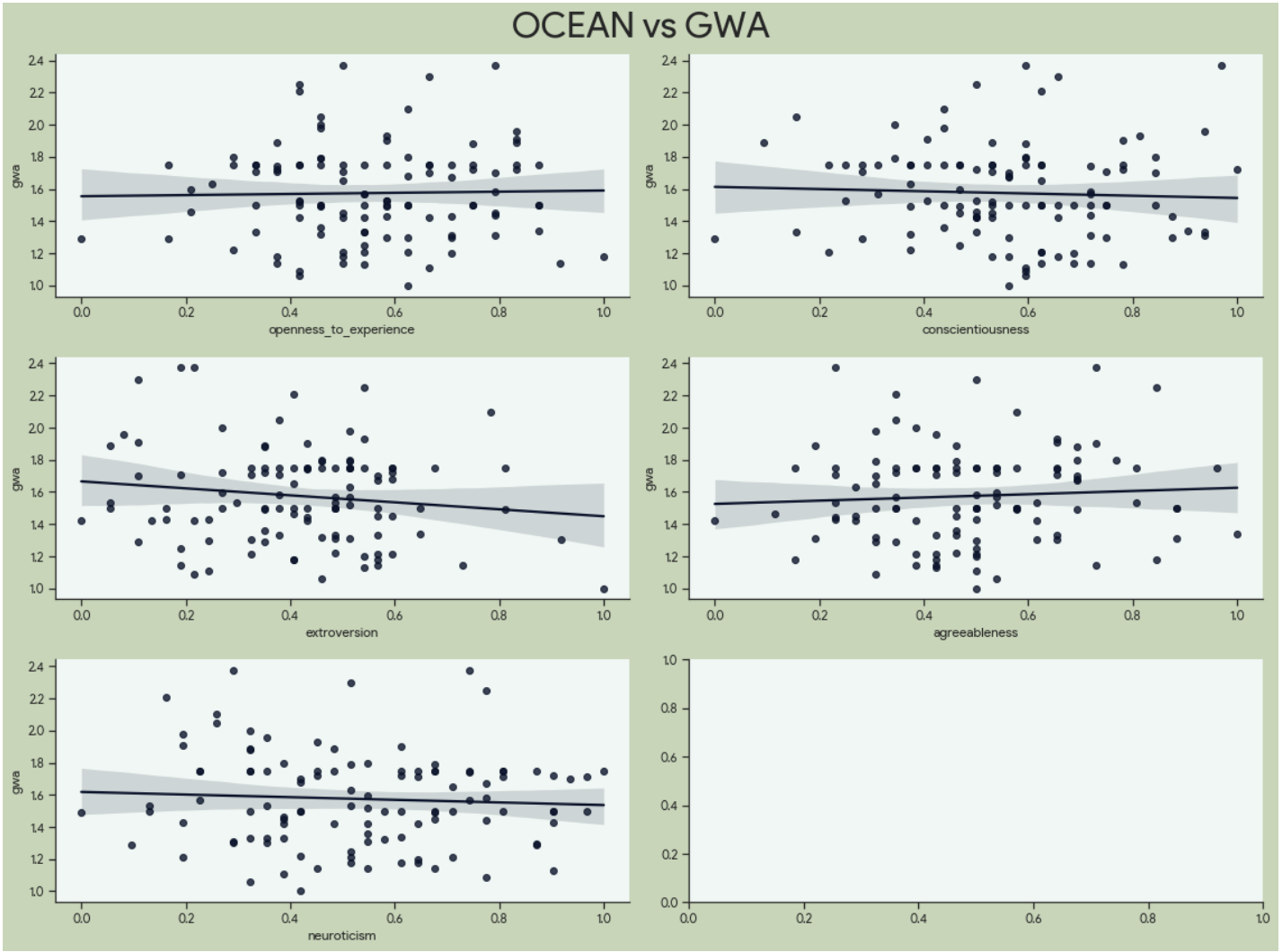
In [11]:
```
fig, ax = plt.subplots(3,2)
ax = ax.flatten()
fig.suptitle('OCEAN vs GWA', fontsize=32)
```

```
i=0

for x in df.columns[5:]:
    sns.regplot(x=df[x], y=df['gwa'], ax=ax[i])
    i += 1

fig.tight_layout()
sns.despine()

plt.savefig('figures/regression_ocean_to_gwa.jpg')
```




# Insights

1. As we can see, the distribution of the OCEAN and GWA seems to be randomly distributed. This signifies no correlation between these two variables and that they are a bad fit.