In [3]:

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, mean, stddev, count, when, isnan, approx_count_distinct
import matplotlib.pyplot as plt
import pandas as pd
# Initialize Spark
spark = SparkSession.builder.appName("AcademicStressAnalysis").getOrCreate()
# Load dataset
file_path = "academicstresslevel.csv"
df = spark.read.option("header", True).csv(file_path, inferSchema=True)
```

In [4]:

```
numeric_cols = [f.name for f in df.schema.fields if "int" in f.dataType.simpleString().lower() or "double" in f
string_cols = [f.name for f in df.schema.fields if f.name not in numeric_cols]
```

In [6]:

```
missing_summary = (
df.select([
        count(when(col(c).isNull(), c)).alias(c)
        if c in string_cols else
        count(when(isnan(col(c)) | col(c).isNull(), c)).alias(c)
        for c in df.columns
    ])
    .toPandas()
    .T.reset_index()
 )
missing_summary.columns = ["Column", "Missing_Count"]
print("\n Missing Values Summary:")
print(missing_summary)
```

```
 Missing Values Summary:
                                              Column  Missing_Count
0                                          Timestamp              0
1                                Your Academic Stage              0
2                                      Peer pressure              0
3                   Academic pressure from your home              0
4                                  Study Environment              1
5         What coping strategy you use as a student?              0
6  Do you have any bad habits like smoking, drink...              0
7  What would you rate the academic  competition ...              0
8                   Rate your academic stress index               0
```

In [9]:

```
 unique_counts = [
    (col_name, df.select(approx_count_distinct(col(col_name))).collect()[0][0])
    for col_name in string_cols
 ]
 unique_counts_df = pd.DataFrame(unique_counts, columns=["Column", "Unique_Count"])
 print("\n Unique Categorical Value Counts:")
 print(unique_counts_df)
```

```
 Unique Categorical Value Counts:
                                              Column  Unique_Count
0                                          Timestamp           149
1                                Your Academic Stage             3
2                                  Study Environment             3
3         What coping strategy you use as a student?             3
4  Do you have any bad habits like smoking, drink...             3
```

In [10]:

```
 # Numerical summary (mean, std, median, IQR)
 summary_stats = df.select(
    *[mean(c).alias(f"{c}_mean") for c in numeric_cols],
    *[stddev(c).alias(f"{c}_stddev") for c in numeric_cols],
 ).toPandas()
 print("\n Mean & Stddev Summary:")
 print(summary_stats.T)
```

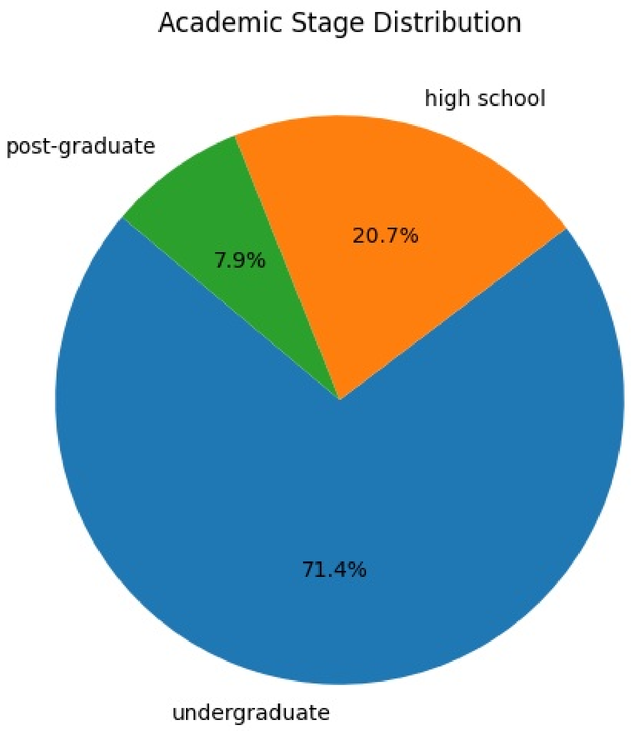
```
 Mean & Stddev Summary:
                                                           0
Peer pressure_mean                                  3.071429
Academic pressure from your home_mean               3.178571
What would you rate the academic  competition i...  3.492857
Rate your academic stress index _mean               3.721429
Peer pressure_stddev                                1.083844
Academic pressure from your home_stddev             1.276618
What would you rate the academic  competition i...  1.028349
Rate your academic stress index _stddev             1.032339
```

In [11]:

```
 pdf = df.select(numeric_cols).toPandas()
 num_summary = pdf.describe(percentiles=[0.25, 0.5, 0.75]).T
 num_summary["IQR"] = num_summary["75%"] - num_summary["25%"]
 print("\n Detailed Numeric Summary:")
 print(num_summary)
 Detailed Numeric Summary:
                                                    count      mean       std  \
Peer pressure                                       140.0  3.071429  1.083844
Academic pressure from your home                    140.0  3.178571  1.276618
What would you rate the academic  competition i...  140.0  3.492857  1.028349
Rate your academic stress index                     140.0  3.721429  1.032339

                                                    min  25%  50%  75%  max  \
Peer pressure                                       1.0  2.0  3.0  4.0  5.0
Academic pressure from your home                    1.0  2.0  3.0  4.0  5.0
What would you rate the academic  competition i...  1.0  3.0  4.0  4.0  5.0
Rate your academic stress index                     1.0  3.0  4.0  4.0  5.0

                                                    IQR
Peer pressure                                       2.0
Academic pressure from your home                    2.0
What would you rate the academic  competition i...  1.0
Rate your academic stress index                     1.0
```

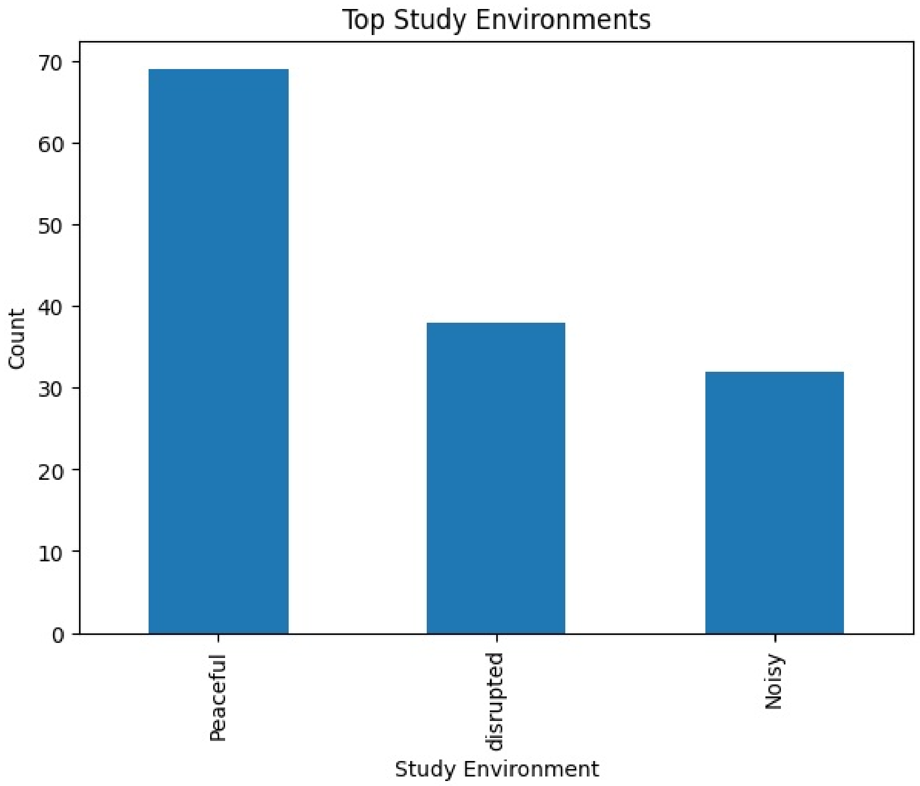
In [15]:

```
# Convert PySpark DataFrame to Pandas
pdf_full = df.toPandas()

# Pie chart — Academic Stage
plt.figure(figsize=(6, 6))
pdf_full["Your Academic Stage"].value_counts().plot.pie(autopct="%1.1f%%", startangle=140)
plt.title("Academic Stage Distribution")
plt.ylabel("")
plt.show()
```



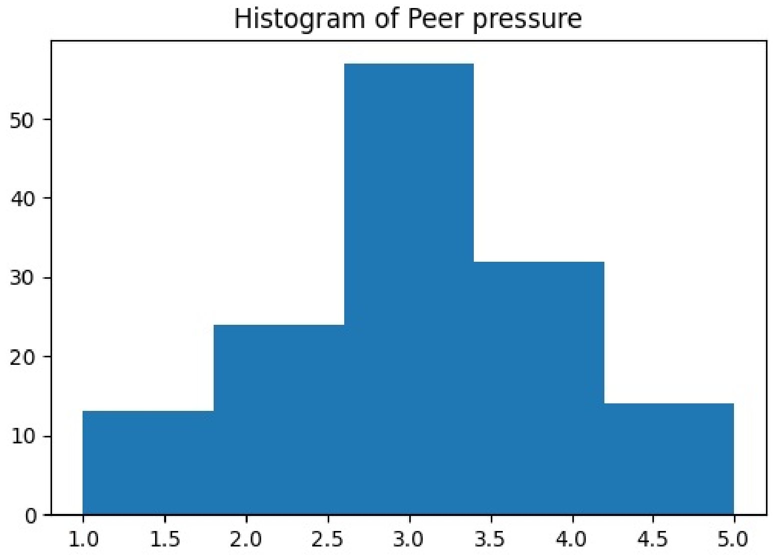
In [17]:

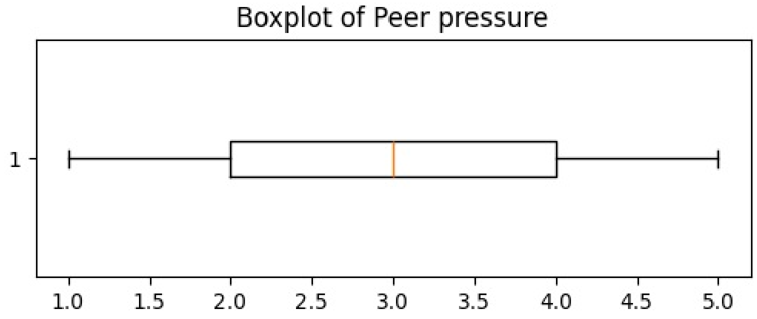
```
plt.figure(figsize=(7, 5))
pdf_full["Study Environment"].value_counts().head(10).plot(kind="bar")
plt.title("Top Study Environments")
plt.ylabel("Count")
plt.show()
```



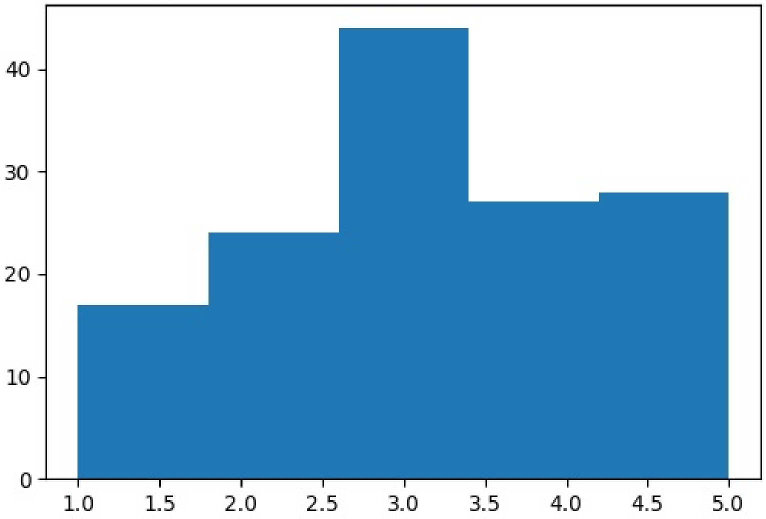
In [18]:

```python
for col_name in numeric_cols:
    plt.figure(figsize=(6, 4))
    plt.hist(pdf_full[col_name].dropna(), bins=5)
    plt.title(f"Histogram of {col_name}")
    plt.show()
    plt.figure(figsize=(6, 2))
    plt.boxplot(pdf_full[col_name].dropna(), vert=False)
    plt.title(f"Boxplot of {col_name}")
    plt.show()
```

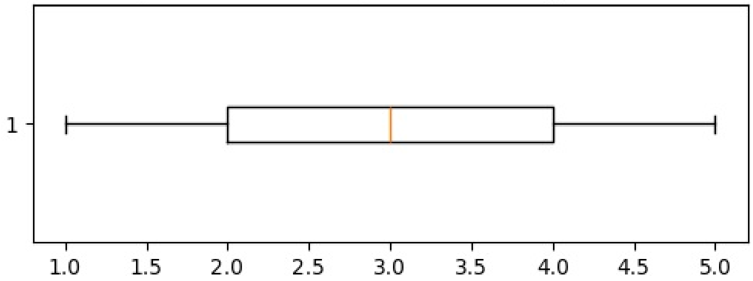




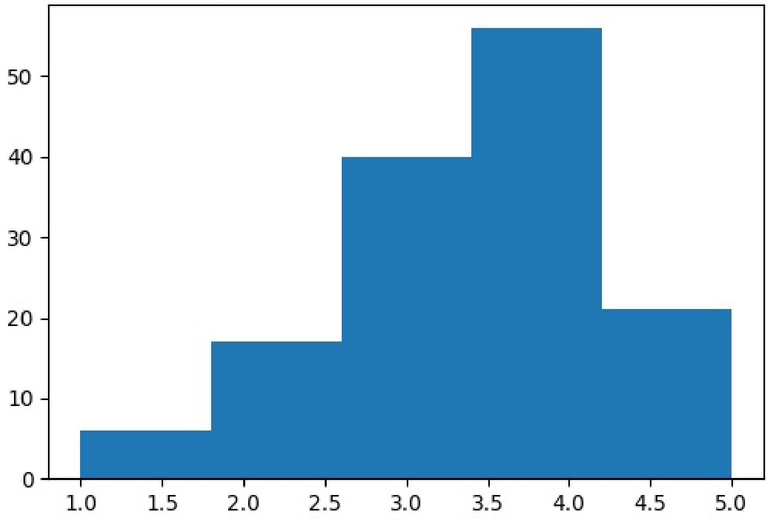
Histogram of Academic pressure from your home



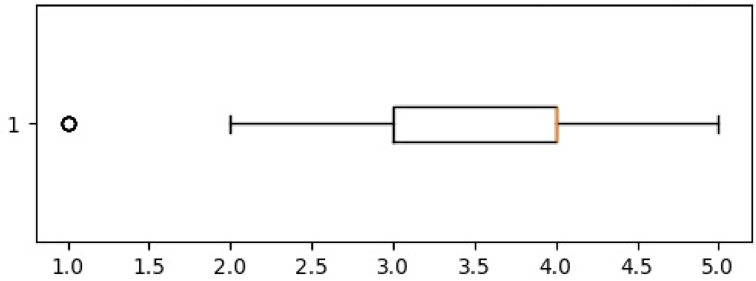
Boxplot of Academic pressure from your home

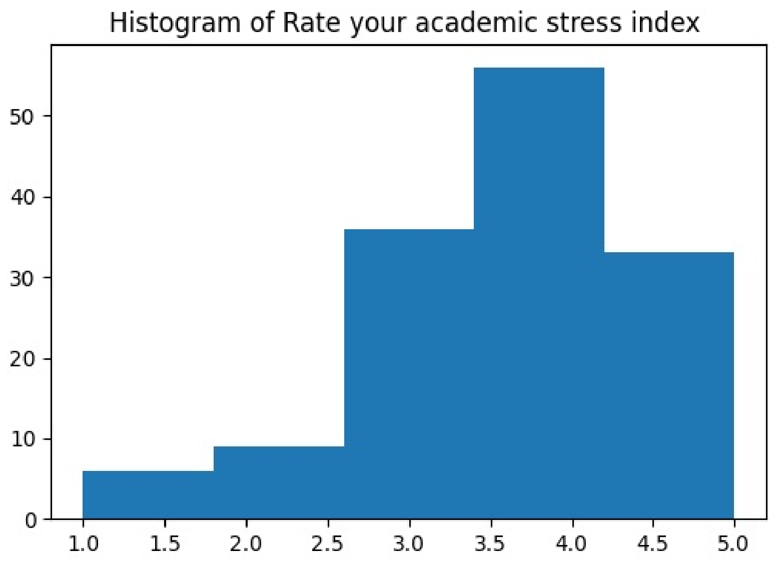


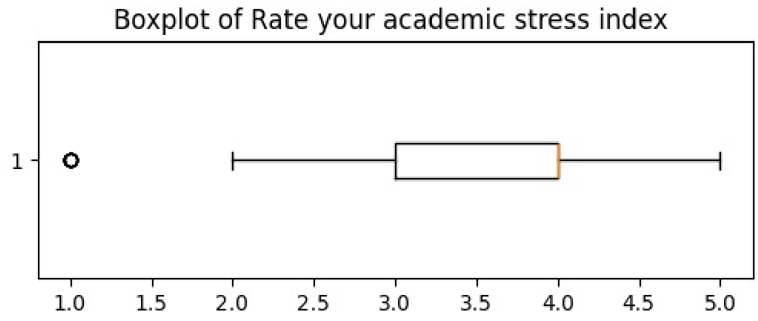
Histogram of What would you rate the academic competition in your student life



Boxplot of What would you rate the academic competition in your student life

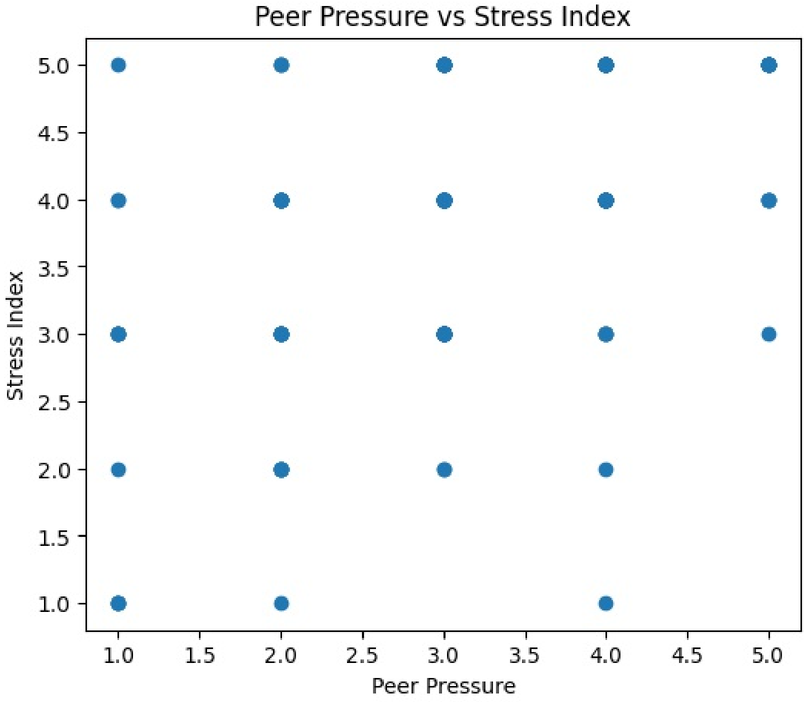


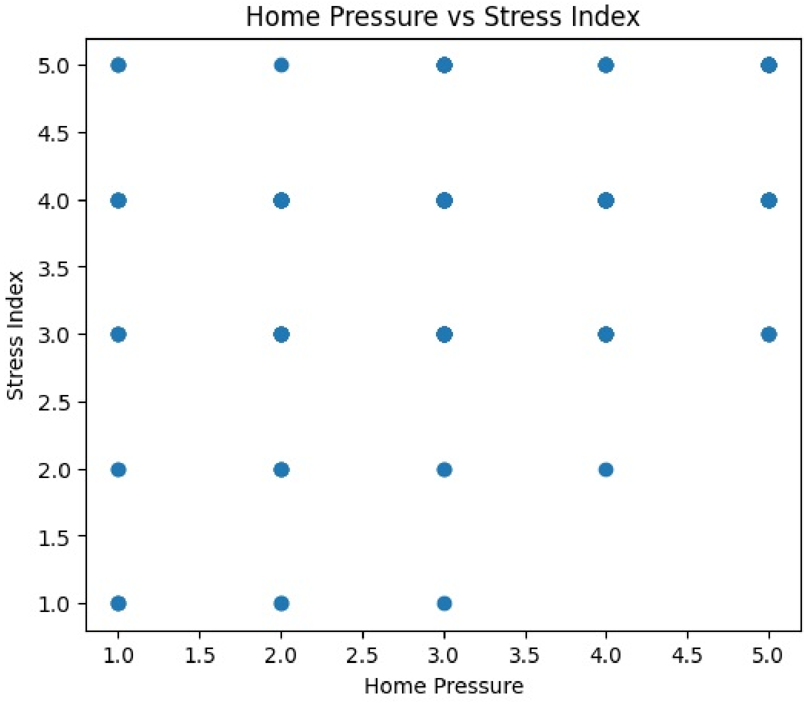




In [19]:

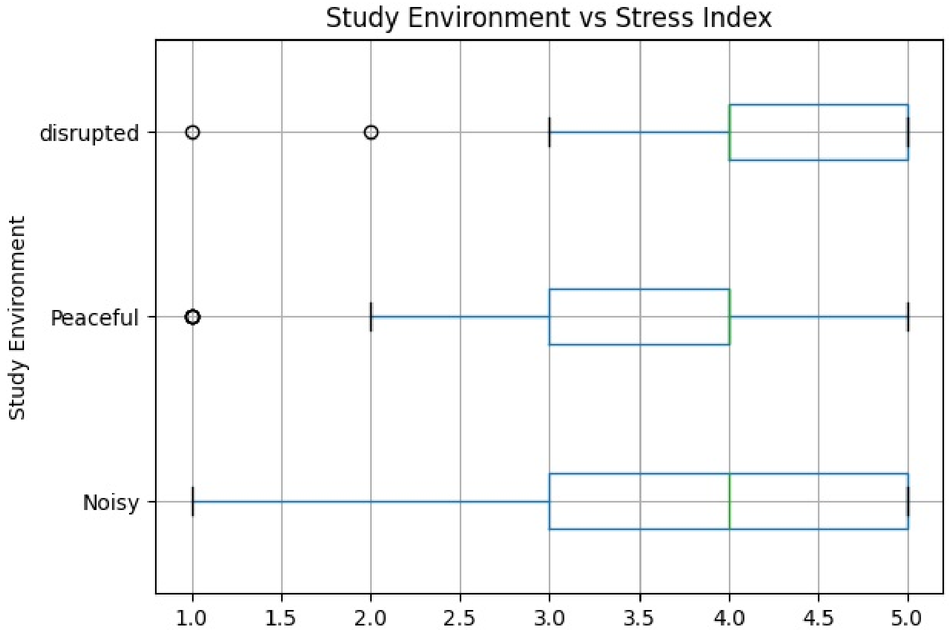
```
# 2.3 Relationship Analysis
 # ----------------------------------------------------------
# Scatter plots
plt.figure(figsize=(6, 5))
plt.scatter(pdf_full["Peer pressure"], pdf_full["Rate your academic stress index "])
plt.title("Peer Pressure vs Stress Index")
plt.xlabel("Peer Pressure")
plt.ylabel("Stress Index")
plt.show()
plt.figure(figsize=(6, 5))
plt.scatter(pdf_full["Academic pressure from your home"], pdf_full["Rate your academic stress index "])
plt.title("Home Pressure vs Stress Index")
plt.xlabel("Home Pressure")
plt.ylabel("Stress Index")
plt.show()
```





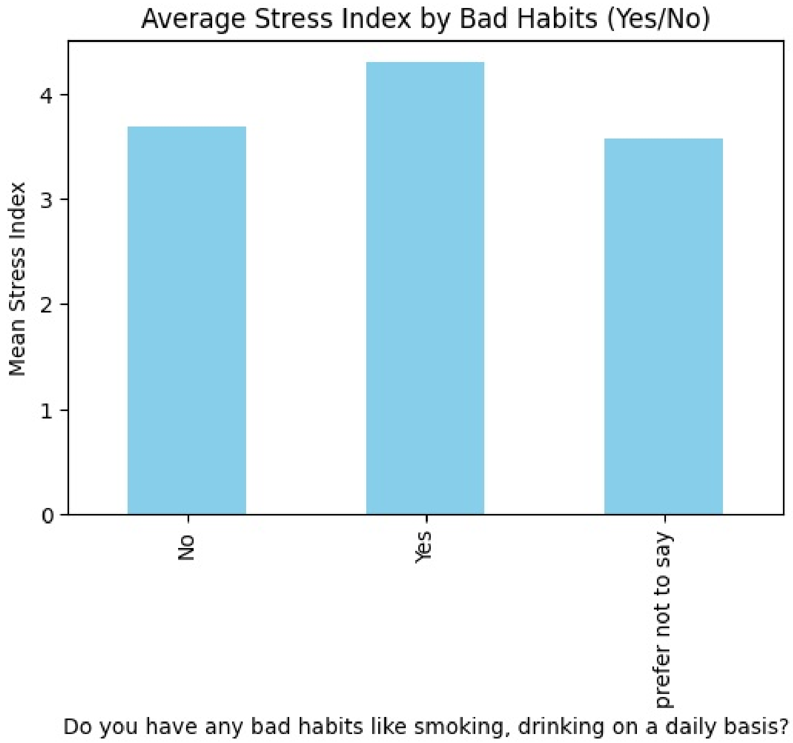
```
In [20]: plt.figure(figsize=(10, 5))
         pdf_full.boxplot(column="Rate your academic stress index ", by="Study Environment", vert=False)
         plt.title("Study Environment vs Stress Index")
         plt.suptitle("")
         plt.show()
```

```
<Figure size 1000x500 with 0 Axes>
```

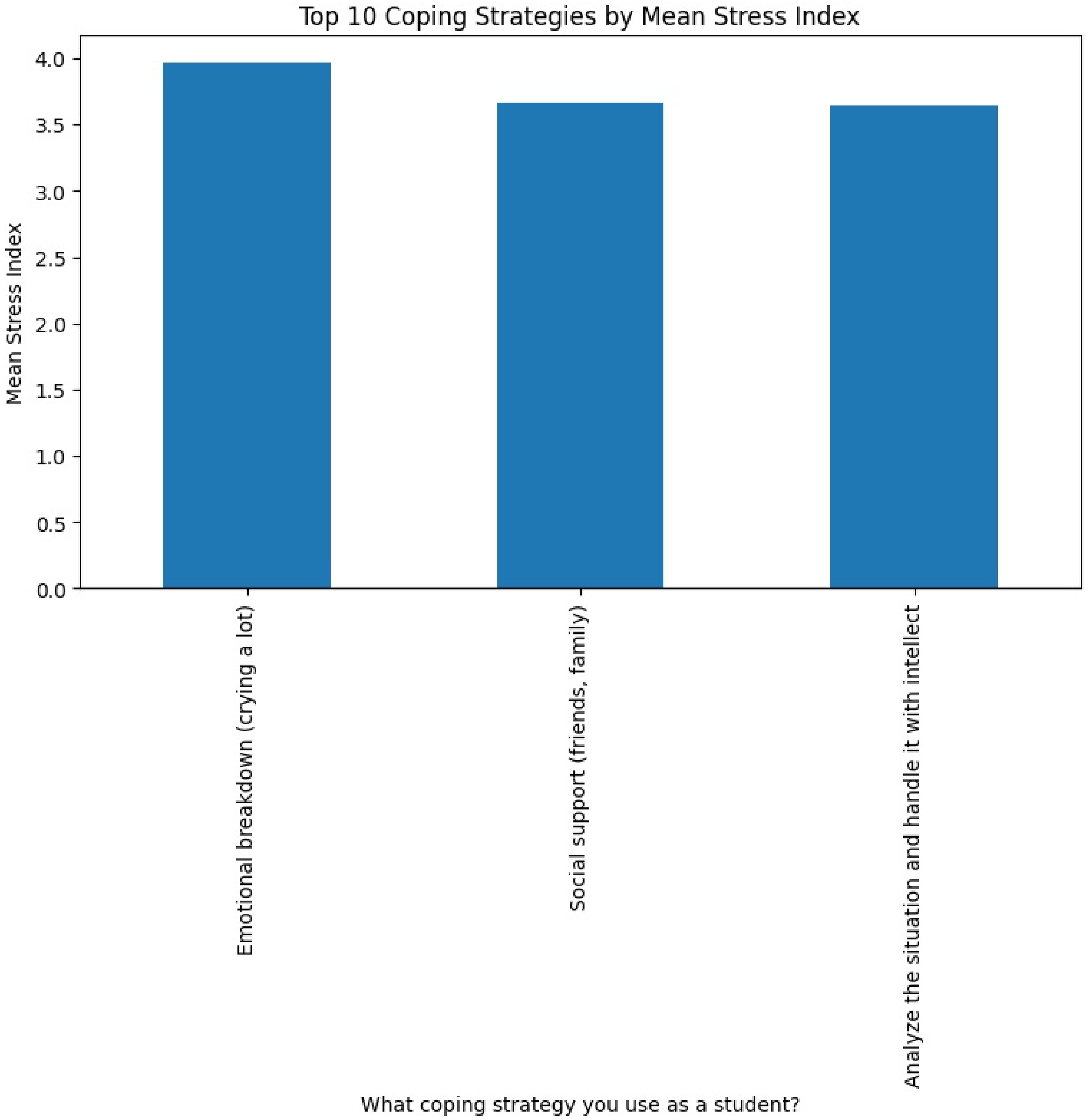


In [21]:

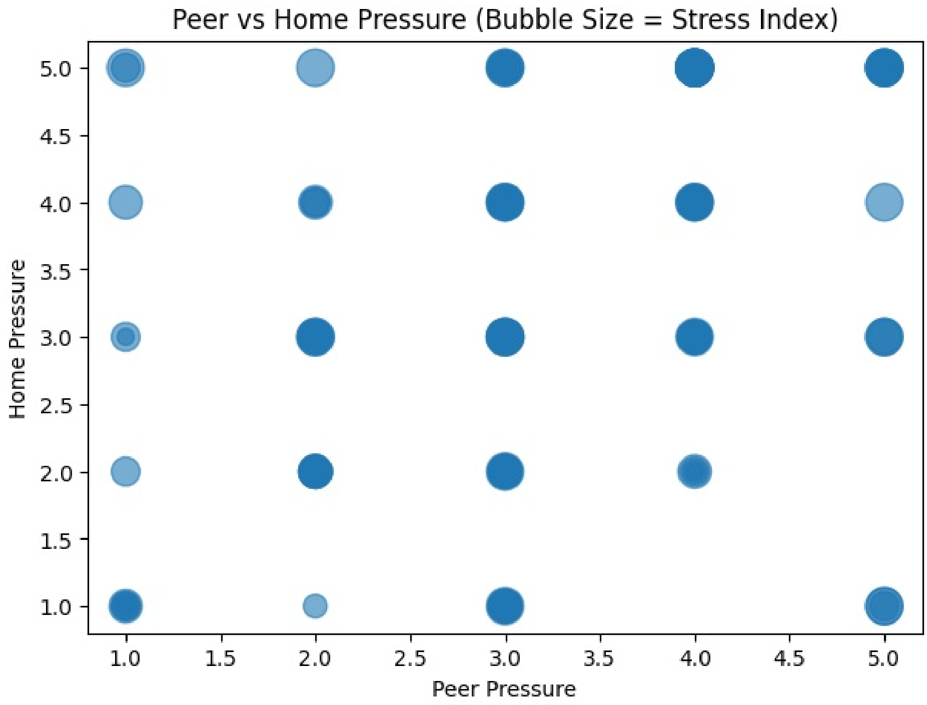
```
pdf_full.groupby("Do you have any bad habits like smoking, drinking on a daily basis?")[
"Rate your academic stress index "
].mean().plot(kind="bar", color="skyblue", figsize=(6, 4))
plt.title("Average Stress Index by Bad Habits (Yes/No)")
plt.ylabel("Mean Stress Index")
plt.show()
```



In [22]:
```
pdf_full.groupby("What coping strategy you use as a student?")[
"Rate your academic stress index "
].mean().sort_values(ascending=False).head(10).plot(kind="bar", figsize=(9, 5))
plt.title("Top 10 Coping Strategies by Mean Stress Index")
plt.ylabel("Mean Stress Index")
plt.show()
```



In [24]:
```
plt.figure(figsize=(7, 5))
plt.scatter(
pdf_full["Peer pressure"],
pdf_full["Academic pressure from your home"],
s=pdf_full["Rate your academic stress index "] * 60,
alpha=0.6,
)
plt.title("Peer vs Home Pressure (Bubble Size = Stress Index)")
plt.xlabel("Peer Pressure")
plt.ylabel("Home Pressure")
plt.show()
```



```
In [ ]:
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js