Urban Data Analysis

Week 11: Introduction to ANOVA

- WHAT IS ANOVA?

- WHY AND WHEN TO USE IT?

- RUNNING ANOVA WITH PYTHON (STATSMODELS)

- NYC EMISSIONS CASE STUDY

- INTERPRETING RESULTS

What is ANOVA?

Analysis of Variance (ANOVA) tests whether the means of three or more groups are significantly different.

- One-Way ANOVA: One categorical independent variable

- Tests group mean differences

- Extension of t-test

Use ANOVA when:

- One numeric dependent variable

- One or more categorical independent variables

- You want to test group mean differences

Example: Does average NYC NOx emissions vary by borough?

ANOVA Hypotheses

Null Hypothesis (H0): All group means are equal

Alternative Hypothesis (H1): At least one group mean differs

We evaluate using the F-statistic and p-value.

Case Study: NYC Emissions

Dataset: NYC Community Air Survey (NYCCAS)

Goal: Examine if average NOx emissions differ by borough

Step 1: Load and Explore Data

```
import pandas as pd
```

```
import statsmodels.api as sm
```

```
from statsmodels.formula.api import ols
```

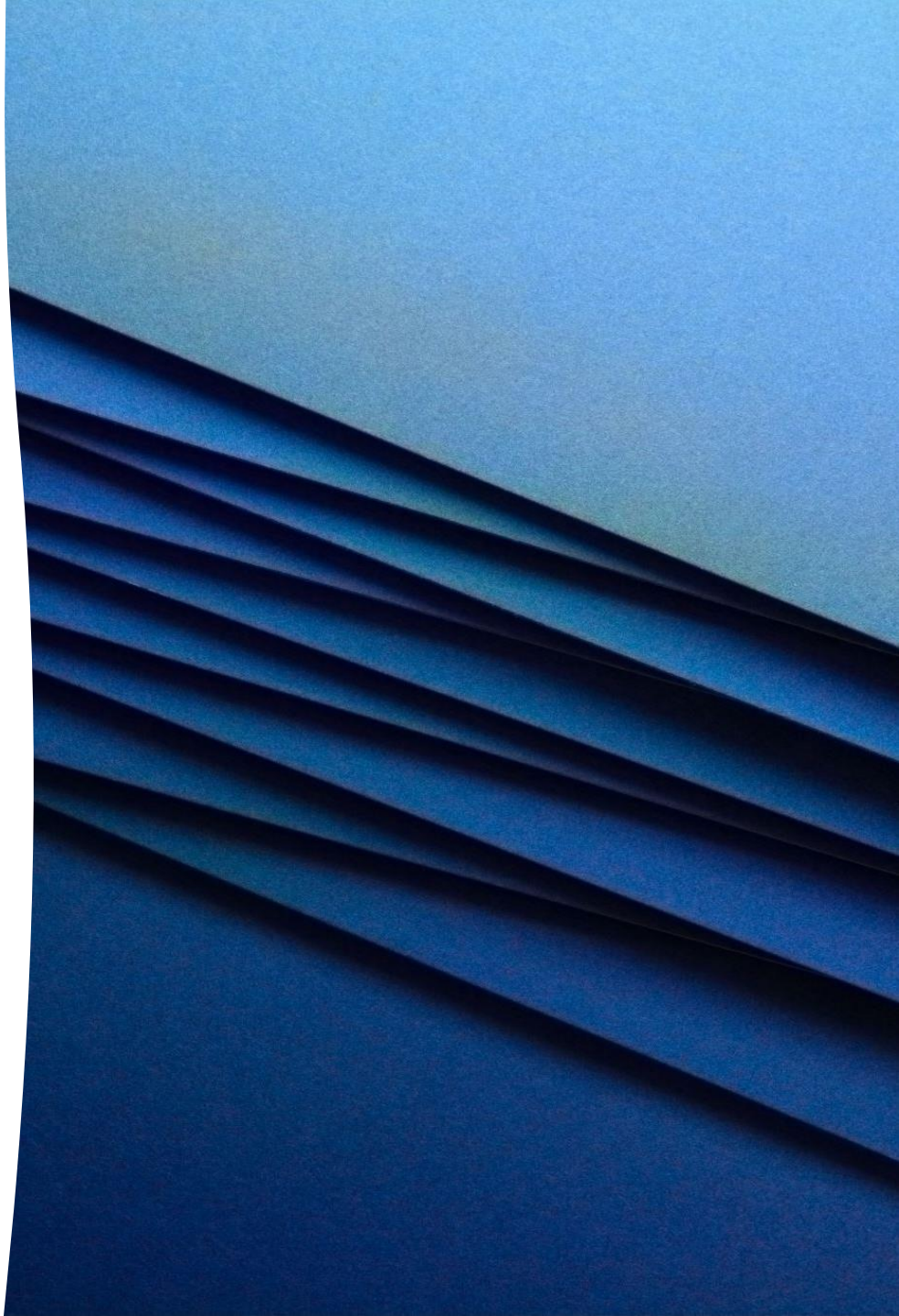
```
emissions =
pd.read_csv('nyc_emissions_by_borough.csv')
```

```
emissions.head()
```

Step 2: Summary Statistics

```
emissions.groupby('borough')['nox'].describe()
```

Check group means and standard deviations

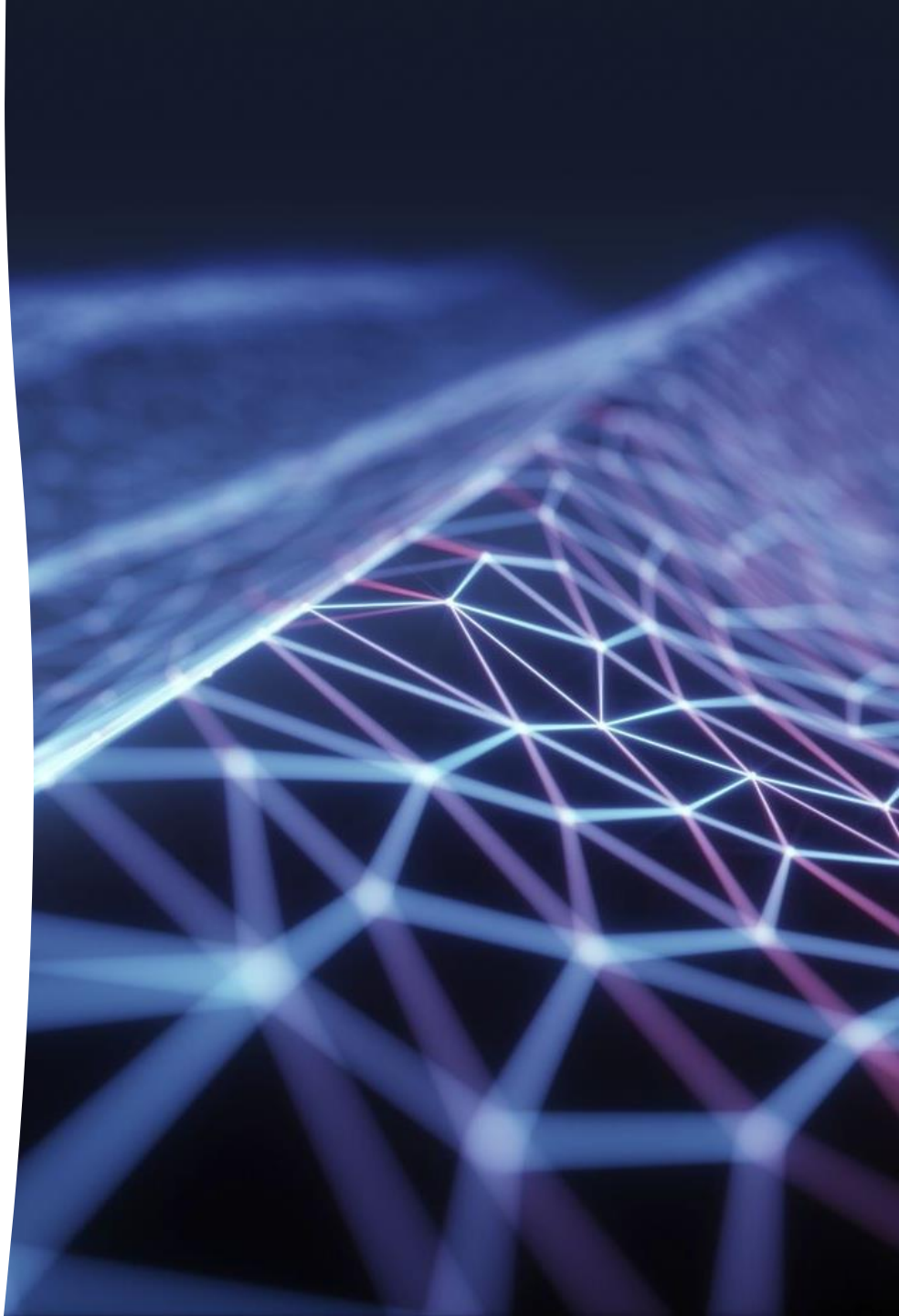


Step 3: Run One-Way ANOVA

```
model = ols('nox ~ C(borough)', data=emissions).fit()
```

```
anova_table = sm.stats.anova_lm(model, typ=2)
```

```
print(anova_table)
```



Step 4: Interpret ANOVA Output

ANOVA Table Output:

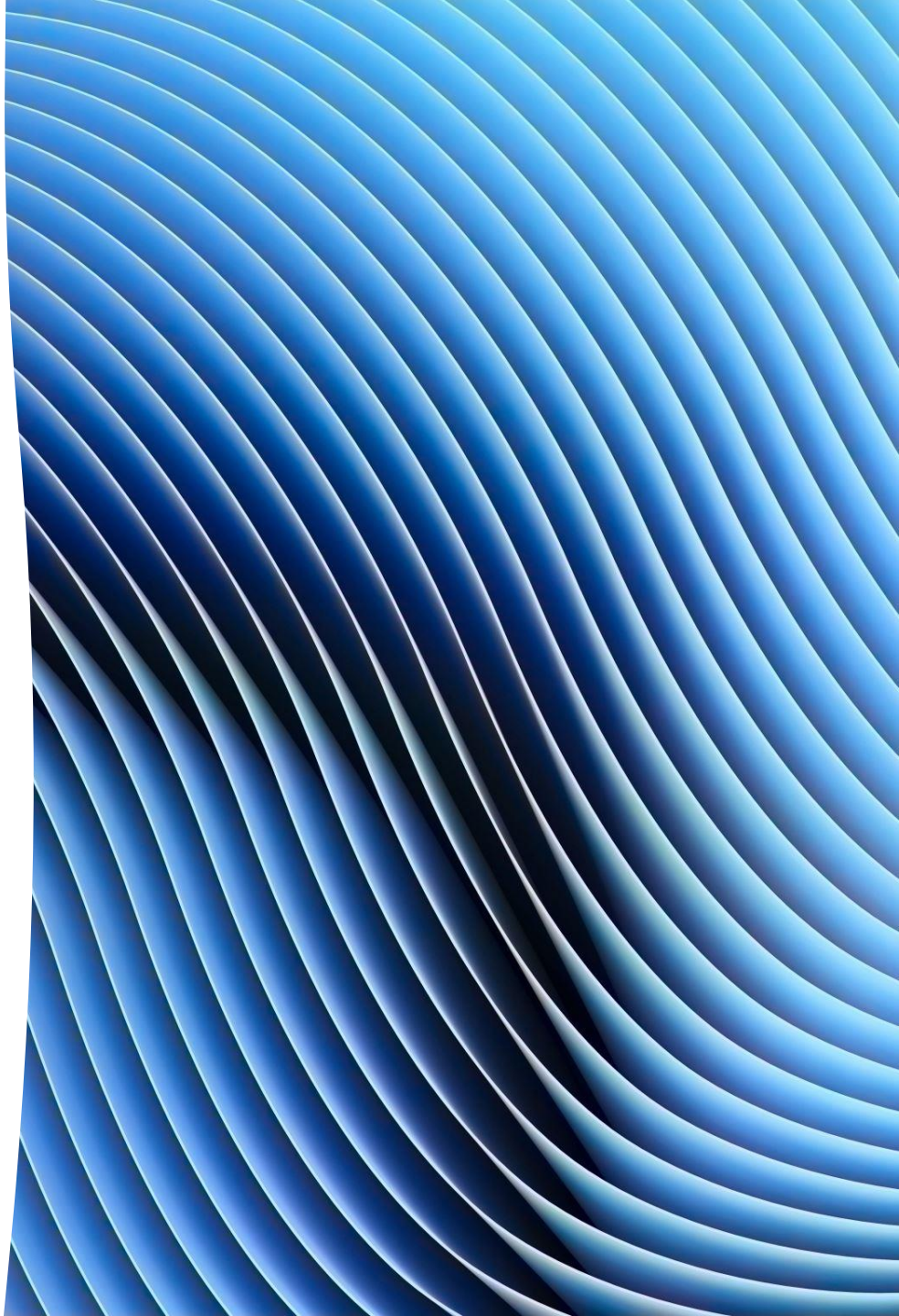
- Sum of Squares

- Degrees of Freedom

- F-Statistic

- p-value

Key: If $p < 0.05$, reject H0 (means differ)



Step 5: Visualize Differences

```
import seaborn as sns
```

```
import matplotlib.pyplot as plt
```

```
sns.boxplot(x='borough', y='nox', data=emissions)
```

```
plt.title('NOx Levels by Borough')
```

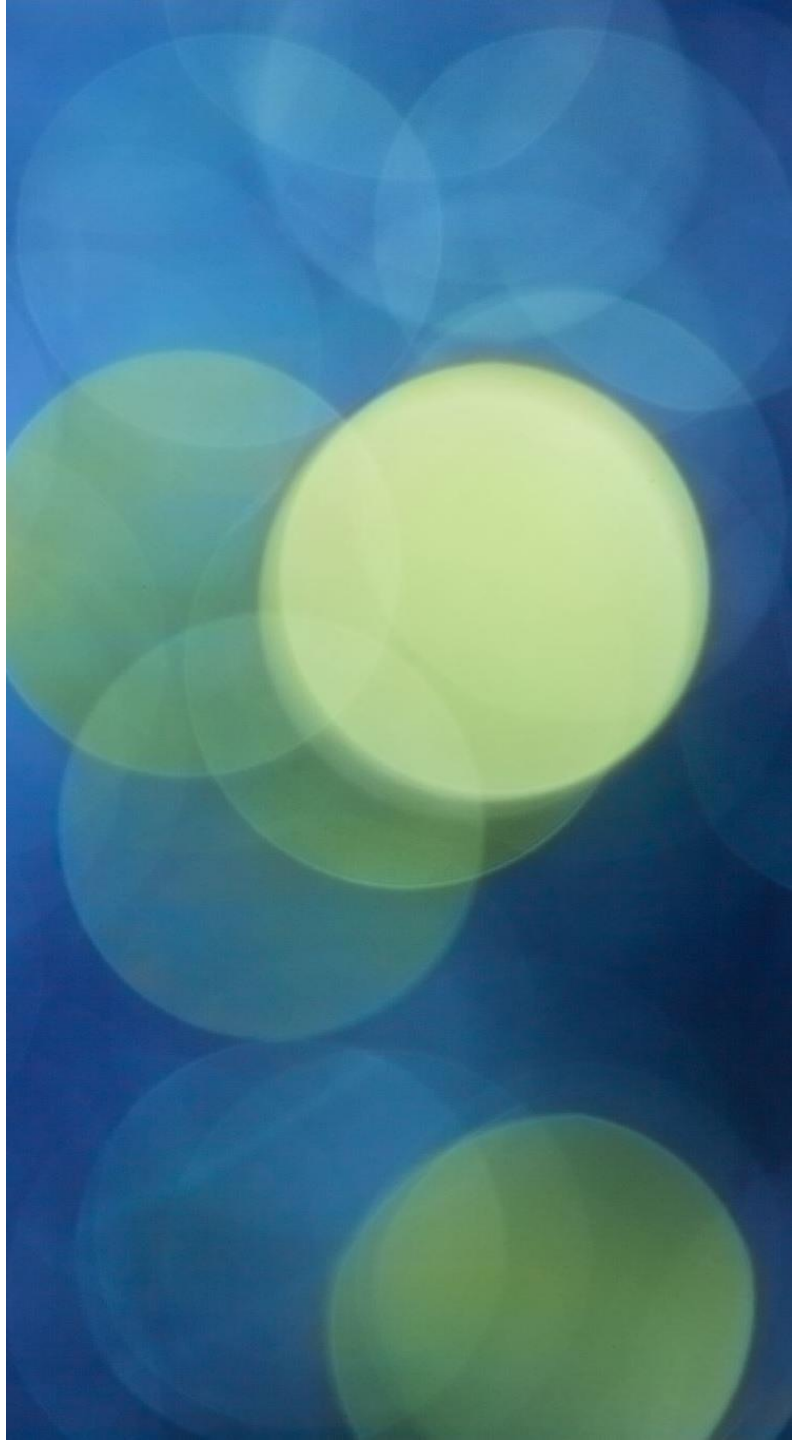
```
plt.show()
```

Wrap-Up

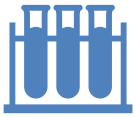
- ANOVA compares multiple group means

- Statsmodels simplifies analysis

- Interpretation relies on p-values and F-stat



Discussion Questions



- When is a t-test more appropriate than ANOVA?

- What assumptions must ANOVA meet?

- How can visualizations confirm statistical results?