Welcome to Online Engineering at George Washington University

Class will begin shortly

Audio: To eliminate background noise, please be sure your audio is muted. To speak, please click the hand icon at the bottom of your screen (**Raise Hand**). When instructor calls on you, click microphone icon to unmute. When you've finished speaking, ***be sure to mute yourself again***.

Chat: Please type your questions in Chat.

Recordings: Please note the recording of this class meeting will be available to download later today. The class recordings are to be used exclusively by registered students in this particular class. Releasing these recordings is strictly prohibited.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

SEAS 8510
Analytical Methods for Machine Learning

Lecture 9

Dr. Zachary Dennis

Adapted from Rodriguez, Etemadi, Fall 2023, George Washington University

Agenda

9:00 – 9:15		Discussion Group
9:15 – 9:45		Data Analytics Review
9:45 – 10:30		Hypothesis Testing
10:30 – 10:40		*BREAK (10 min)*
10:40 – 11:45		Hypothesis Testing Examples and Applications
11:45 – 12:00		Homework and Discussion Look Ahead

Assignments

Last week: Homework 7 and Discussion 7 due on 5/18 at 9 AM Eastern

This week: Discussion Summary due on 6/1 at 9 AM

Probability Distribution Percentile

- Let X be some continuous r.v., and p be a probability of interest.
- Sometimes we are interested in finding q_p such that

$$F_X(q_p) = P(X \leq q_p) = p$$

  where the smallest value of q_p for which this is true is the p-th quantile (or 100p-th percentile) of the distribution for X. The median of a distribution is its 50th percentile

- **Example**: If exam scores are distributed normally with mean and std. dev. of 80 and 5, what is the 90th percentile score?

```
1 norm.ppf(0.9, loc=80, scale=5)
```

```
86.407757827723
```

Understanding Data

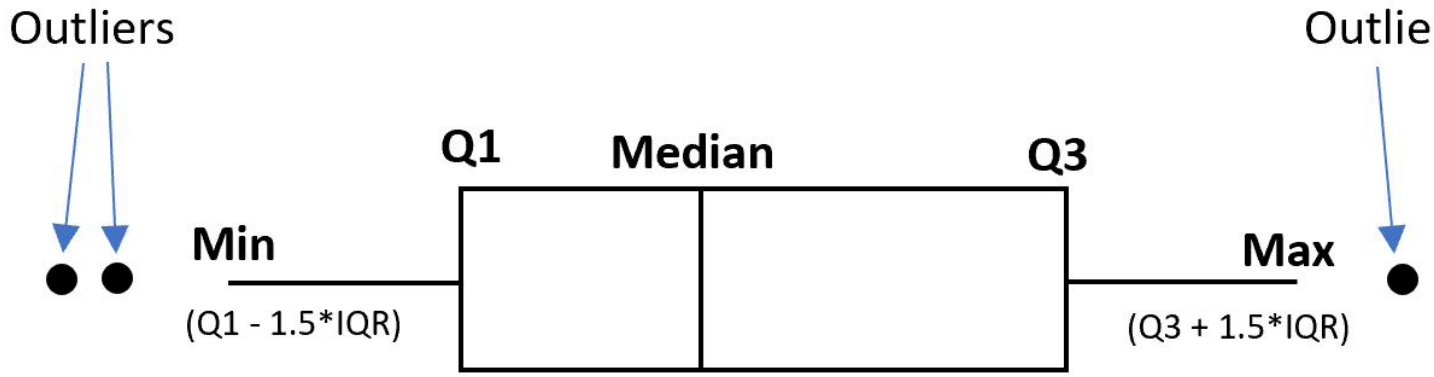
- In machine learning, understanding your data is the first step toward building effective models. Let's explore key data statistics and metrics.
- **Mean**: The mean is a measure of central tendency. It represents the average value of a dataset.
- **Median:** Another measure of central tendency. It represents the middle value when data is sorted (or the average of two middle values). Less affected by outliers compared to the mean
- **Variance**: Variance measures the spread or dispersion of data points. A higher variance indicates greater data variability.

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Standard Deviation**: Standard deviation is the square root of the variance, s_x. It provides a standardized measure of data dispersion.

Understanding Data

- **Quantiles:** Divide data into equal-sized subsets. Common quantiles include quartiles (25th, 50th, 75th percentiles). Useful for understanding data distribution and identifying outliers.
- **Interquartile Range (IQR):** IQR measures the spread of data around the median. It is the difference between the 75th percentile (Q3) and the 25th percentile (Q1). Useful for identifying outliers and assessing data variability.
- **Skewness:** Quantifies the asymmetry of the data distribution.
  - Positive skew: Data is skewed to the right (tail on the right).
  - Negative skew: Data is skewed to the left (tail on the left).



https://www.statology.org/how-to-read-box-plot-with-outliers/

Understanding Data

- **Covariance:** Covariance measures the degree to which two variables change together.
  - Positive covariance: Variables move in the same direction.
  - Negative covariance: Variables move in opposite directions.

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- **Correlation:** Correlation is a standardized measure of covariance. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation). Helps assess the linear relationship between two variables.

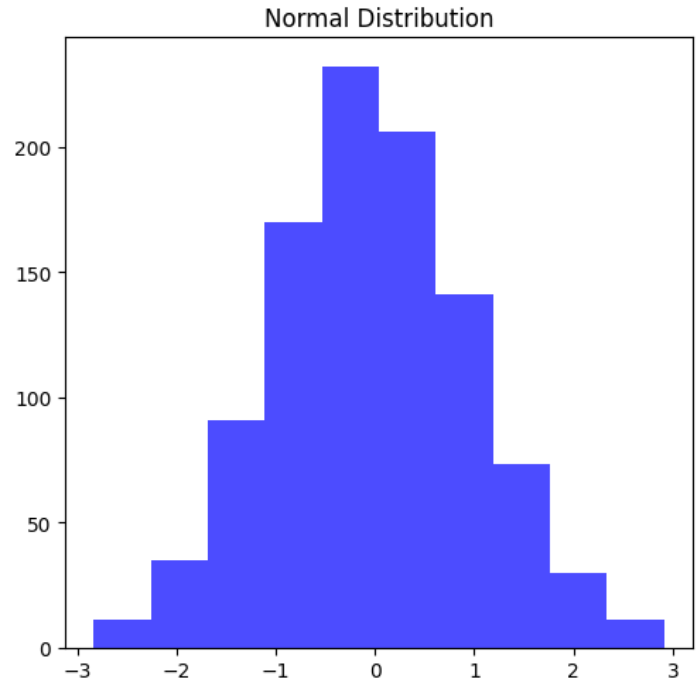
$$r = \frac{\mathrm{Cov}(X,Y)}{s_X s_Y}$$

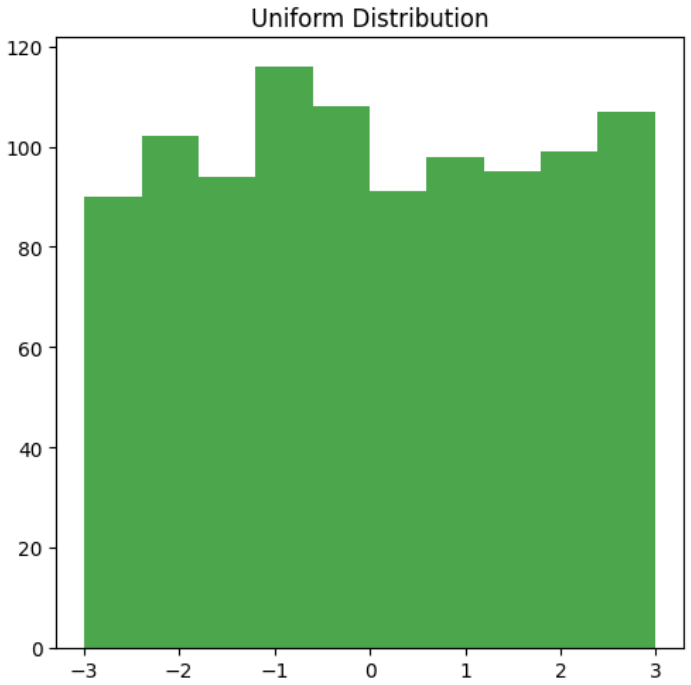
Understanding Data

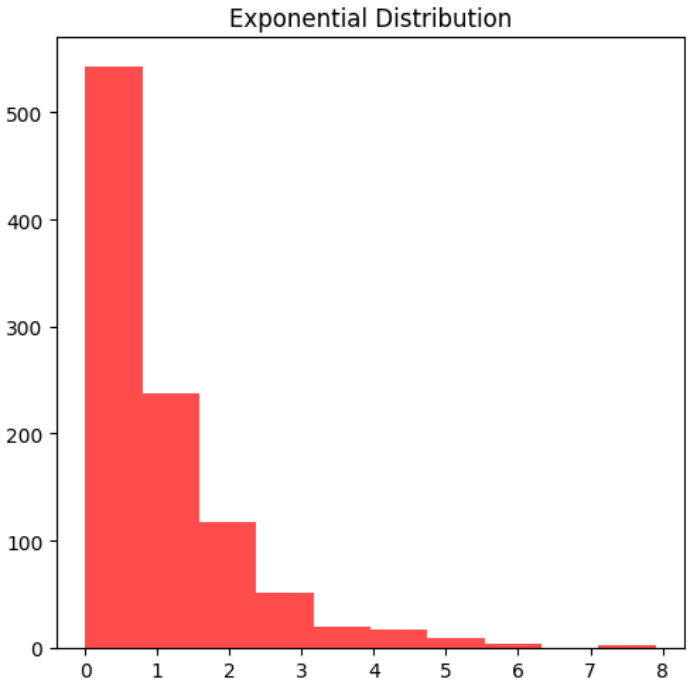
```
mean = np.mean(data)
variance = np.var(data, ddof = 1) #Denominator of N-1, sample var
std_dev = np.std(data)
median = np.median(data)
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
cov_matrix = np.cov(data_X, data_Y)
corr_matrix = np.corrcoef(data_X, data_Y)

from scipy.stats import iqr, skew
iqr(data)
skew(data)
```

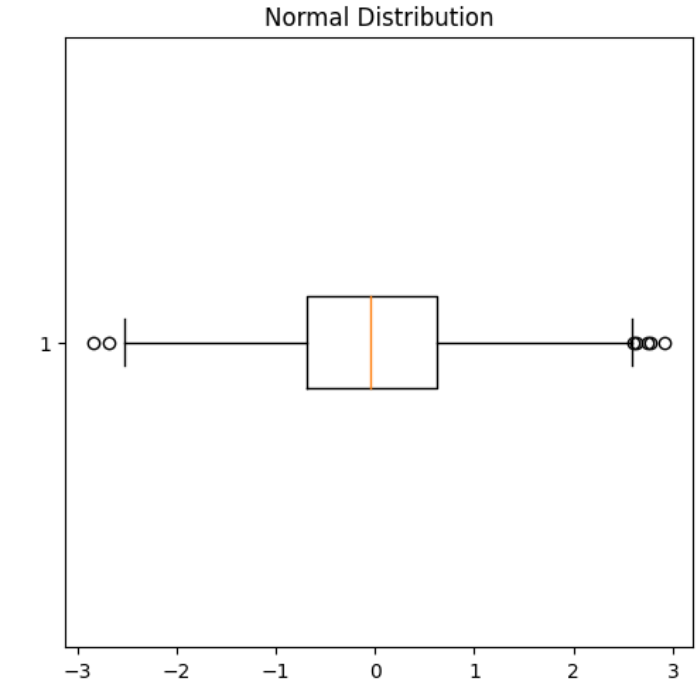
- `import matplotlib.pyplot as plt`
- `plt.hist(data)`
- `plt.boxplot(data)`

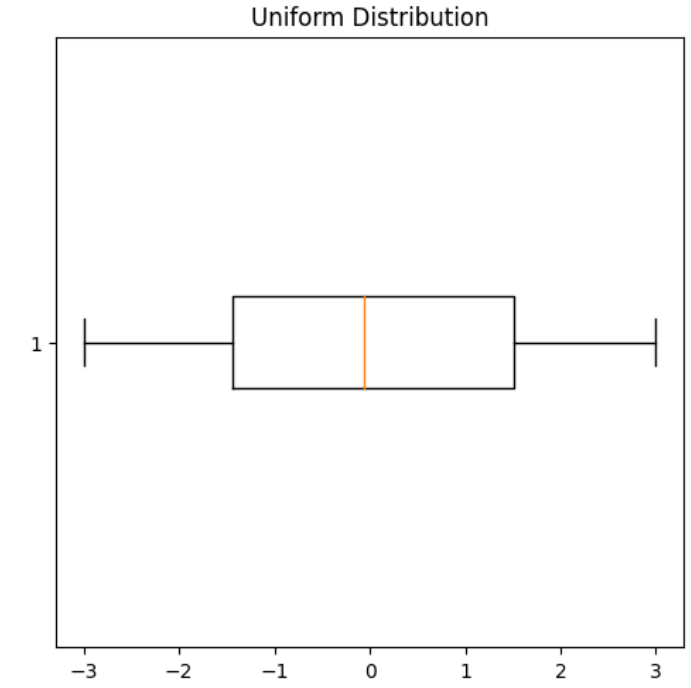
Understanding Data

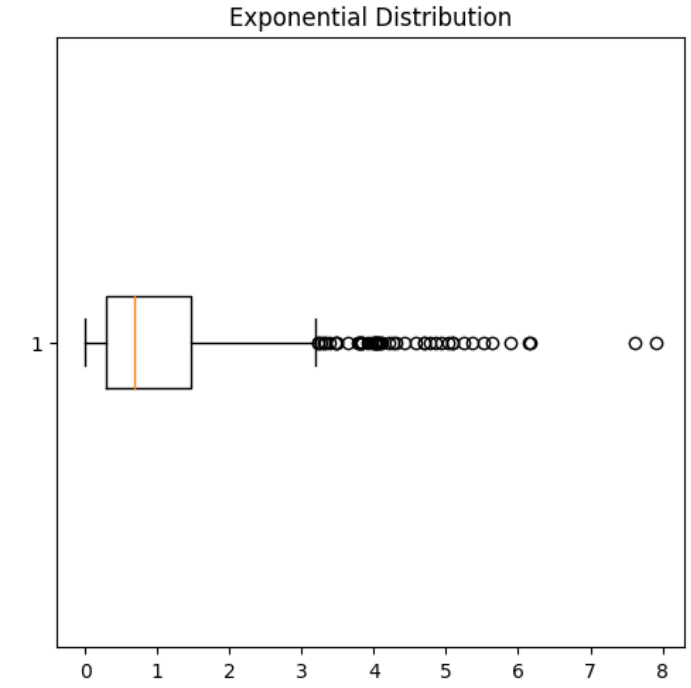












Hypothesis Testing

- A fundamental tool in data analysis. Used to make informed decisions based on data.
- To determine if there is enough evidence to reject a hypothesis about a population parameter.
- Involves formulating hypotheses, collecting data, and drawing conclusions
- Two types of hypotheses:
  - Null Hypothesis (H_0): Represents the status quo or no effect.
  - Alternative Hypothesis (H_a): Represents the proposed effect or change.
- Example:
  - H_0: There is no difference in test scores before and after a training program.
  - H_a: There is a significant improvement in test scores after the training program.

Hypothesis Testing Error

- Type 1 Error (False Positive): Occurs when a null hypothesis that is actually true is rejected.
  - It represents a situation where the test incorrectly detects an effect or difference that doesn't exist.
  - The probability of Type 1 error is denoted as "alpha" (α) and is typically set as the significance level (e.g., 0.05) in hypothesis testing.
  - Minimizing Type 1 error is important when the cost or consequences of making a false positive decision are high.
- Type 2 Error (False Negative): Occurs when a null hypothesis that is actually false is not rejected.
  - It represents a situation where the test fails to detect a real effect or difference that exists.
  - The probability of Type 2 error is denoted as "beta" (β).
  - Minimizing Type 2 error is crucial when failing to detect a true effect can have significant implications, such as in medical testing or quality control.
- Example: There is a glass of water on a table. Null: It is water. Alternative: It is H2SO4. What type of error is more dangerous?

Hypothesis Testing Steps

- Data Collection: Gather relevant data through experiments, surveys, or observations.
- Formulate Hypotheses: Define null and alternative hypotheses based on research questions.
- Select Significance Level (α): Determine the acceptable level of Type I error (false positive).
- Perform Statistical Test: Choose an appropriate statistical test (e.g., t-test, chi-square, ANOVA) based on data type and research question.
- Determine P-value: Calculate the p-value, which represents the probability of obtaining results as extreme as those observed, assuming the null hypothesis is true.
  - Smaller p-values indicate stronger evidence against the null hypothesis.
- Make a Decision: Compare the p-value to the chosen significance level (α).
  - If p-value $< \alpha$, reject the null hypothesis.
  - If p-value $\geq \alpha$, fail to reject the null hypothesis.

Hypothesis Testing: Example

- It is claimed that a machine learning classification model is accurate 90% of the time. Out of 1000 observations, 876 are classified correctly. Do we have sufficient evidence against the claim?

Hypothesis test for proportion

One sample Z-Test because we are comparing sample statistic/proportional to a hypothesized population parameter.

Two sample Z-Test would be used to compare to proportions from two samples to each other.

Steps:

1. State null and alt. hypotheses
2. Calculate sample proportion $\hat{p} = \frac{number\ of\ correct\ classifications}{total\ number\ of\ observations} = \frac{876}{1000}$
3. Calculate Standard Error of the sampling distribution $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$
4. Calculate test statistic z:

$$z = \frac{\hat{p} - p_0}{SE}$$

5. Find p-value associated with the test statistic
6. Compare to significance level (0.05)

T-tests

- A T-test is a statistical test used to compare the means of two groups.
- It helps determine if there are significant differences between the groups.
- **Types of T-Tests**:
  - Independent samples t-test: Compares means between two different groups.
  - Paired sample t-test: Compares means from the same group at different times.
  - One-sample t-test: Tests the mean of a single group against a known mean.
- **Criteria for Use**:
  - Normally distributed data.
  - Scale (interval or ratio) data.
  - Random sampling from the population.
- **Common Applications**:
  - Comparing test scores of two different groups of students.
  - Assessing the effect of a treatment in a before-and-after study.
  - Testing hypotheses in experimental research.

T-tests: Example 1

- **Example**: A group of students are given a pre-test before a study session and a post-test after the session to measure the effectiveness of the study method.

Pre-Test Scores: [70, 75, 65, 80, 85, 68, 90, 85, 88, 70]

Post-Test Scores: [75, 80, 68, 83, 88, 73, 93, 90, 90, 75]

Use t-test to see if the study session had a statistically significant effect on the students' test scores.

```
from scipy.stats import ttest_rel
pre_test_scores = [70, 75, 65, 80, 85, 68, 90, 85, 88, 70]
post_test_scores = [75, 80, 68, 83, 88, 73, 93, 90, 90, 75]
t_statistic, p_value = ttest_rel(post_test_scores, pre_test_scores) t_statistic, p_value
```

Steps to calculate the t-statistic:

1. **Calculate the differences** between each pair of observations (post-test score - pre-test score).
2. **Compute the mean of these differences** $\bar{d}$:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

3. **Compute the standard deviation of these differences** s_d:

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}$$

4. **Compute the t-statistic** using the formula:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

T-tests: Example 1

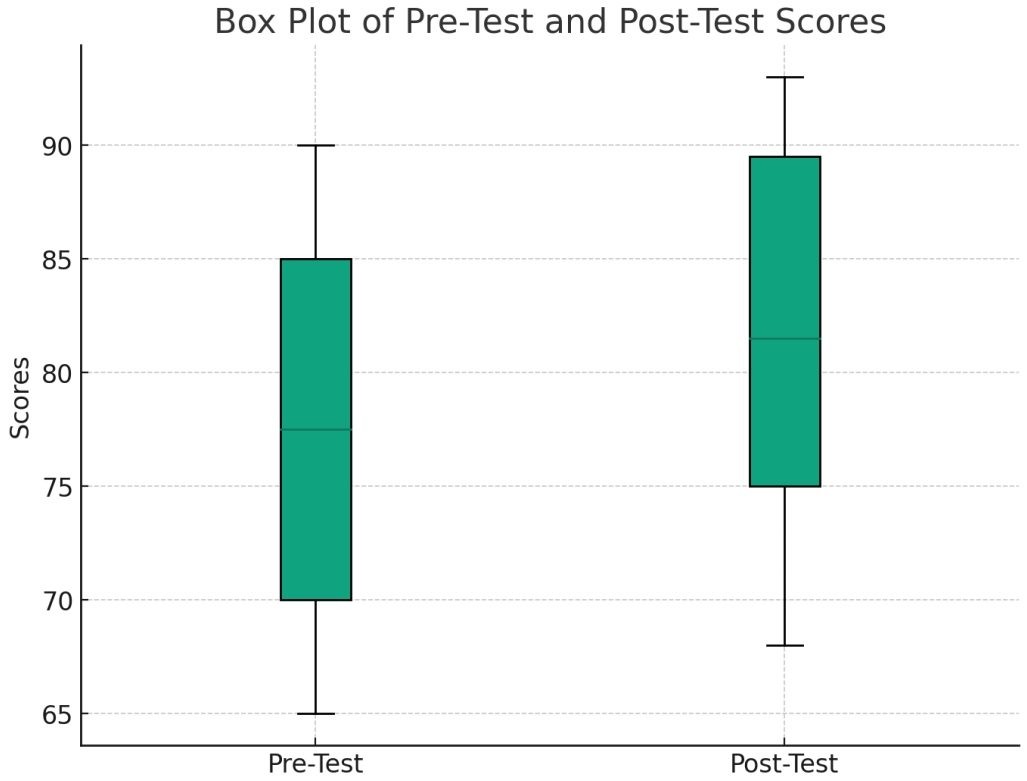
- **Example**: A group of students are given a pre-test before a study session and a post-test after the session to measure the effectiveness of the study method.

Pre-Test Scores: [70, 75, 65, 80, 85, 68, 90, 85, 88, 70]

Post-Test Scores: [75, 80, 68, 83, 88, 73, 93, 90, 90, 75]

Use t-test to see if the study session had a statistically significant effect on the students' test scores.

```
(10.301275604009799, 2.7934158896131296e-06)
```



T-tests: Example 2

- **Example**: A company wants to predict customer churn (whether or not a customer will stop using the company's product or service). The company collects various customer metrics and wants to determine which features are significantly different between customers who churn and those who do not.

We have a feature called "Monthly Usage" (hours per month a customer uses the company's service). Monthly Usage for a random sample of customers:

Group A (Churned): [10, 12, 9, 11, 8, 15, 7, 6, 9, 10]

Group B (Not Churned): [20, 22, 19, 25, 24, 21, 23, 26, 22, 20]

Is "Monthly Usage" a valuable feature?

```
from scipy.stats import ttest_ind
churned = [10, 12, 9, 11, 8, 15, 7, 6, 9, 10]
not_churned = [20, 22, 19, 25, 24, 21, 23, 26, 22, 20]
t_statistic, p_value = ttest_ind(churned, not_churned)
t_statistic, p_value
```

```
(-11.426768162550097, 1.1051347301945896e-09)
```

ANalysis Of VAriance (ANOVA)

- ANOVA is a statistical method used to test the differences between two or more means.
- Commonly used when comparing three or more groups.
- Types of ANOVA include:
  - One-way ANOVA: Tests the effect of a single factor.
  - Two-way ANOVA: Tests the effect of two independent variables.
  - MANOVA (Multivariate ANOVA): Tests multiple dependent variables.
- Assumptions:
  - Normal distribution of the dependent variable.
  - Homogeneity of variances (equal variances across groups).
  - Independent observations
- **Post-hoc Analysis**: To determine which specific groups differ after an ANOVA indicates a significant difference exists.
  - Helps in interpreting the results of an ANOVA by providing detailed pairwise comparisons.
- Most Common: Tukey's Honestly Significant Difference (HSD) Test

ANOVA Example

- We are working on a dataset related to predicting house prices, and we have several categorical features along with the target variable (house price). Features:
- Neighborhood: Categorical (e.g., 'Downtown', 'Suburb', 'Countryside')
- House Type: Categorical (e.g., Single-family', 'Town House', 'Apartment')

Objective: To determine which categorical features significantly affect house prices, thus are important for our predictive model.

Neighborhood	Type	Price
Downtown	Single_family	500000
Downtown	Town_House	520000
Downtown	Apartment	490000
Downtown	Single_family	410000
Downtown	Town_House	480000
Downtown	Apartment	560000
Downtown	Single_family	470000
Suburb	Town_House	510000
Suburb	Apartment	400000
Suburb	Single_family	420000
Suburb	Town_House	380000
Suburb	Apartment	430000
Suburb	Single_family	450000
Suburb	Town_House	390000
Countryside	Apartment	410000
Countryside	Single_family	407000
Countryside	Town_House	399000
Countryside	Apartment	370000
Countryside	Single_family	320000
Countryside	Town_House	290000
Countryside	Apartment	340000

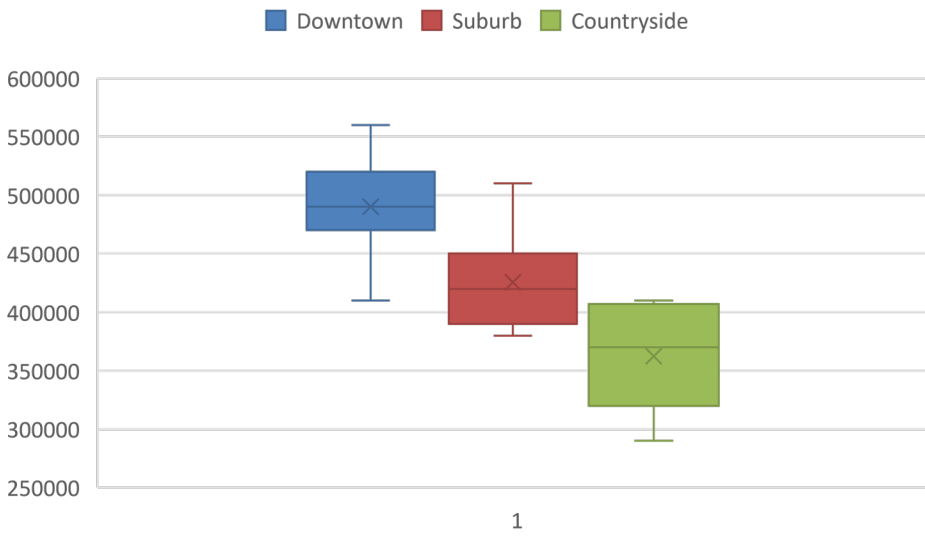
```
import pandas as pd
data = {'Neighborhood':
['Downtown','Downtown','Downtown','Downtown','Downtown','Downtown','Downtown','S
uburb','Suburb','Suburb','Suburb','Suburb','Suburb','Suburb','Countryside','Coun
tryside','Countryside','Countryside','Countryside','Countryside','Countryside',]
,
        'Type': ['Single-Family', 'Townhouse', 'Apartment', 'Single-Family',
'Townhouse', 'Apartment', 'Single-Family', 'Townhouse', 'Apartment', 'Single-
Family', 'Townhouse', 'Apartment','Single-Family', 'Townhouse',
'Apartment','Single-Family', 'Townhouse', 'Apartment', 'Single-Family',
'Townhouse', 'Apartment'],
        'Price':[500000,520000,490000,410000,480000,560000,470000,510000,400000,
420000,380000,430000,450000,390000,410000,407000,399000,370000,320000,290000,
340000]}
df = pd.DataFrame(data)
```

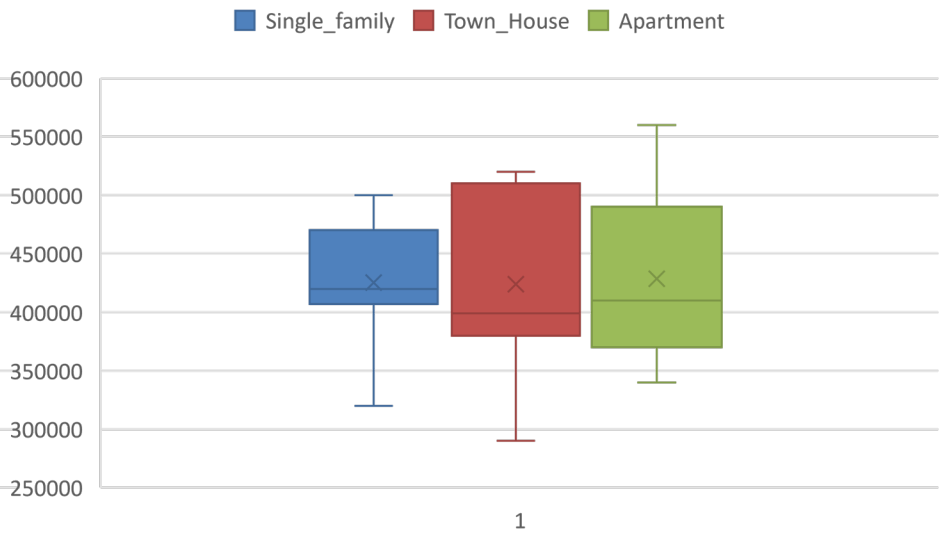
ANOVA Example

```
from scipy.stats import f_oneway

anova_neighborhoods = f_oneway(df[df['Neighborhood'] == 'Downtown']['Price'],
                               df[df['Neighborhood'] == 'Suburb']['Price'],
                               df[df['Neighborhood'] == 'Countryside']['Price'])

print("ANOVA results for Neighborhoods are ", anova_neighborhoods)

anova_Types = f_oneway(df[df['Type'] == 'Single-Family']['Price'],
                               df[df['Type'] == 'Townhouse']['Price'],
                               df[df['Type'] == 'Apartment']['Price'])

print("ANOVA results for house Type are ", anova_Types)

from statsmodels.stats.multicomp import pairwise_tukeyhsd

tukey = pairwise_tukeyhsd(endog=df['Price'],      # Data
                          groups=df['Neighborhood'],   # Groups
                          alpha=0.05)          # Significance level

tukey_results = tukey.summary()  # Summary of test results

print("Tukey HSD results for Neighborhoods are ", tukey_results)
```

ANOVA Example

```
ANOVA results for Neighborhoods are  F_onewayResult(statistic=13.605301981433989, pvalue=0.0002513562955021209)
ANOVA results for house Type are  F_onewayResult(statistic=0.007026798902722093, pvalue=0.9930005538091575)
Tukey HSD results for Neighborhoods are           Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================================================
   group1    group2    meandiff  p-adj     lower        upper    reject
---------------------------------------------------------------------------------
Countryside Downtown 127714.2857 0.0002   65228.3766 190200.1948   True
Countryside   Suburb  63428.5714 0.0463     942.6623 125914.4805   True
   Downtown   Suburb -64285.7143 0.0432 -126771.6234  -1799.8052   True
---------------------------------------------------------------------------------
```





Chi-Square Test of Independence

- The Chi-Square Test of Independence is a statistical method used to determine if there is a significant association between two categorical variables.
- Purpose:
  - To test the independence of two variables.
  - Commonly used in feature selection, hypothesis testing, and market research.
- Formulate Hypotheses:
  - Null Hypothesis (H0): Variables are independent.
  - Alternative Hypothesis (H1): Variables are not independent.
- If the p-value is less than α, reject H0.

Chi-Square Test: Example

- Suppose we have a dataset related to a marketing campaign for a bank. The dataset contains various customer features and a target variable indicating whether the customer subscribed to a term deposit (Yes or No). Our goal is to identify which customer features are significantly related to the subscription outcome.
- Features (Categorical):
  - Job (e.g., admin, technician, entrepreneur, ...)
  - Marital Status (e.g., single, married, divorced)
  - Education (e.g., primary, secondary, tertiary, unknown)
  - Default: has credit in default? (yes, no)
  - Housing: has housing loan? (yes, no)
  - Loan: has personal loan? (yes, no)

	Job	Marital	Education	Default	Housing	Loan	Subscription
0	admin	married	tertiary	no	yes	no	yes
1	technician	single	secondary	no	yes	yes	no
2	entrepreneur	married	tertiary	yes	no	no	yes
3	admin	divorced	primary	no	yes	no	no
4	technician	married	secondary	no	no	yes	yes

Chi-Square Test: Example

- Let's see if Marital Status is a significant feature for classification. First create a contingency table:
-
```
contingency_table =
    pd.crosstab(df['Marital'],  df['Subscription'])
```

Subscription	no	yes
Marital		
divorced	1	0
married	0	3
single	1	0

- Then perform a chi-square test (**Results may not be valid due to the size of data set**)

```
chi2, p, dof, expected = chi2_contingency(contingency_table)
print(f"The p-value is {p}.")
```

```
The p-value is 0.08208499862389880.
```

- Results indicate that the two variables are NOT independent, hence Marital Status is a valuable feature for predicting subscription.