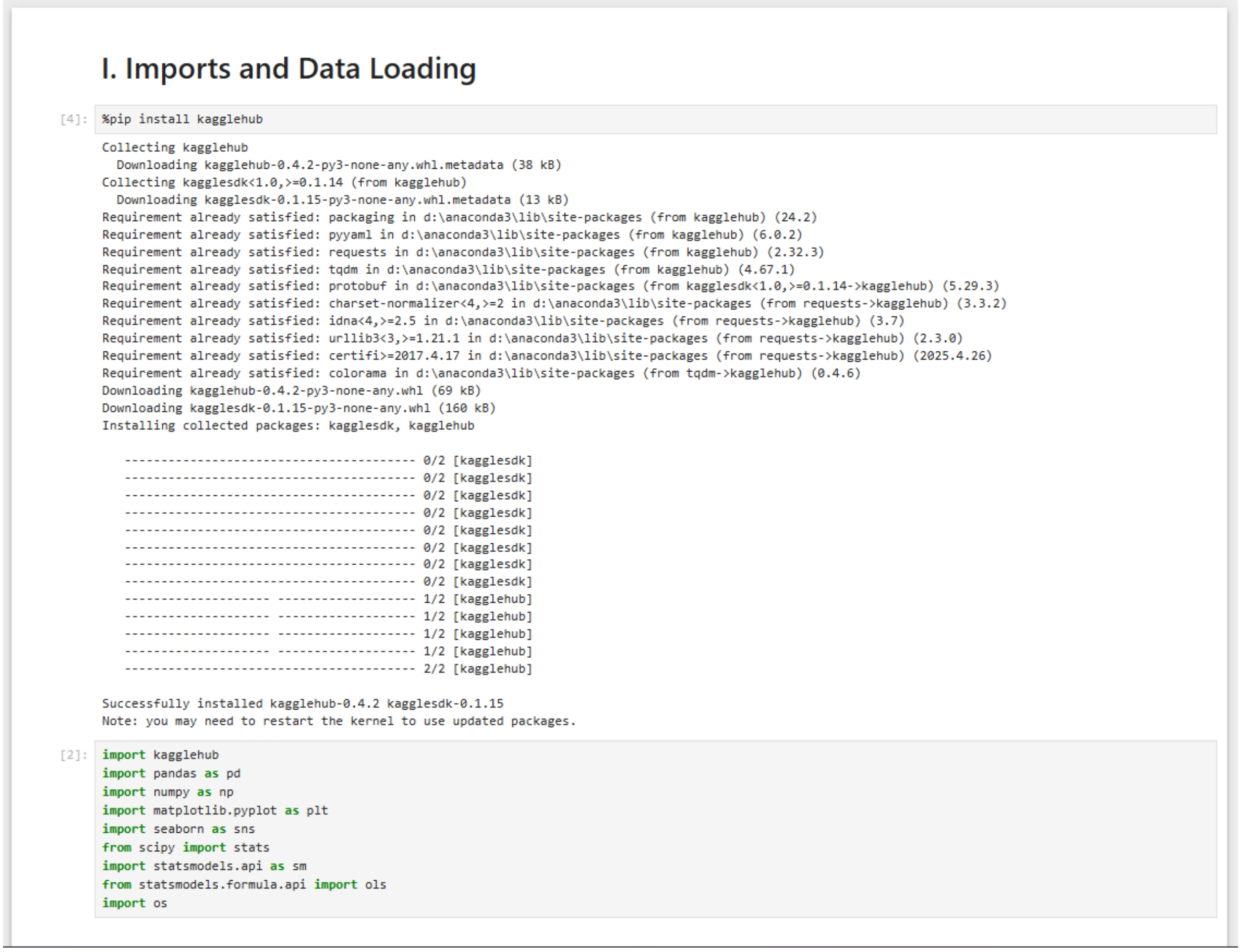
Report: E-commerce Sales and Profitability

Course: Statistics and Data Science 1 (Python)
Instructor: Dilnaz Omarova
Student: Abdanur Ayazbek

Table of contents

I. Dataset Description

- **Data Source:** This study utilizes the "E-commerce Sales and Profit Analysis" dataset obtained from Kaggle via the kagglehub library. [1]
- **Nature of Data:** The dataset represents individual transaction records from a global e-commerce retailer. It includes details such as product categories, shipping modes, geographical regions, and financial metrics (Sales, Profit, Discounts). [1]
- **Observations:** The dataset contains over 51,000 observations, providing a robust foundation for statistical inference. [1]
- **Variables Used:**
    - **Category:** A categorical independent variable (Furniture, Office Supplies, Technology).
    - **Region:** A categorical independent variable representing different global markets.
    - **Profit:** A numerical dependent variable representing the net profit per transaction.
    - **Sales:** A numerical variable used for contextual descriptive analysis.



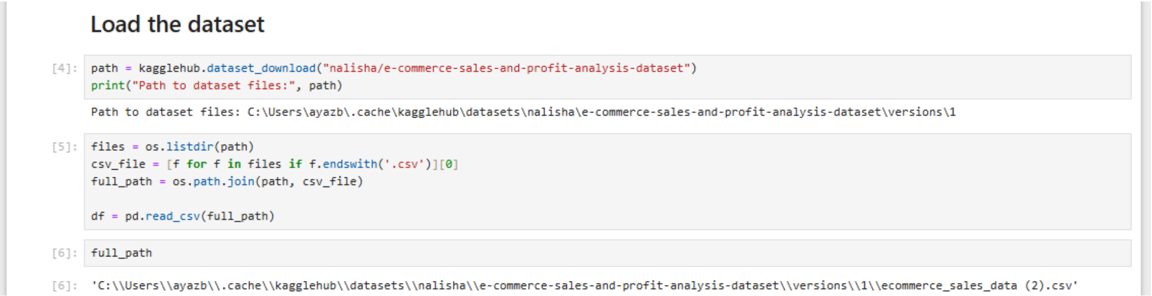
I. Imports and Data Loading

```
[4]: %pip install kagglehub
```

```
Collecting kagglehub
  Downloading kagglehub-0.4.2-py3-none-any.whl.metadata (38 kB)
Collecting kagglesdk<1.0,>=0.1.14 (from kagglehub)
  Downloading kagglesdk-0.1.15-py3-none-any.whl.metadata (13 kB)
Requirement already satisfied: packaging in d:\anaconda3\lib\site-packages (from kagglehub) (24.2)
Requirement already satisfied: pyyaml in d:\anaconda3\lib\site-packages (from kagglehub) (6.0.2)
Requirement already satisfied: requests in d:\anaconda3\lib\site-packages (from kagglehub) (2.32.3)
Requirement already satisfied: tqdm in d:\anaconda3\lib\site-packages (from kagglehub) (4.67.1)
Requirement already satisfied: protobuf in d:\anaconda3\lib\site-packages (from kagglesdk<1.0,>=0.1.14->kagglehub) (5.29.3)
Requirement already satisfied: charset-normalizer<4,>=2 in d:\anaconda3\lib\site-packages (from requests->kagglehub) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in d:\anaconda3\lib\site-packages (from requests->kagglehub) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in d:\anaconda3\lib\site-packages (from requests->kagglehub) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in d:\anaconda3\lib\site-packages (from requests->kagglehub) (2025.4.26)
Requirement already satisfied: colorama in d:\anaconda3\lib\site-packages (from tqdm->kagglehub) (0.4.6)
Downloading kagglehub-0.4.2-py3-none-any.whl (69 kB)
Downloading kagglesdk-0.1.15-py3-none-any.whl (160 kB)
Installing collected packages: kagglesdk, kagglehub

   ---------------------------------------- 0/2 [kagglesdk]
   ---------------------------------------- 0/2 [kagglesdk]
   ---------------------------------------- 0/2 [kagglesdk]
   ---------------------------------------- 0/2 [kagglesdk]
   ---------------------------------------- 0/2 [kagglesdk]
   ---------------------------------------- 0/2 [kagglesdk]
   ---------------------------------------- 0/2 [kagglesdk]
   ---------------------------------------- 0/2 [kagglesdk]
   -------------------- ------------------- 1/2 [kagglehub]
   -------------------- ------------------- 1/2 [kagglehub]
   -------------------- ------------------- 1/2 [kagglehub]
   -------------------- ------------------- 1/2 [kagglehub]
   ---------------------------------------- 2/2 [kagglehub]

Successfully installed kagglehub-0.4.2 kagglesdk-0.1.15
Note: you may need to restart the kernel to use updated packages.
```

```
[2]: import kagglehub
     import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
     from scipy import stats
     import statsmodels.api as sm
     from statsmodels.formula.api import ols
     import os
```

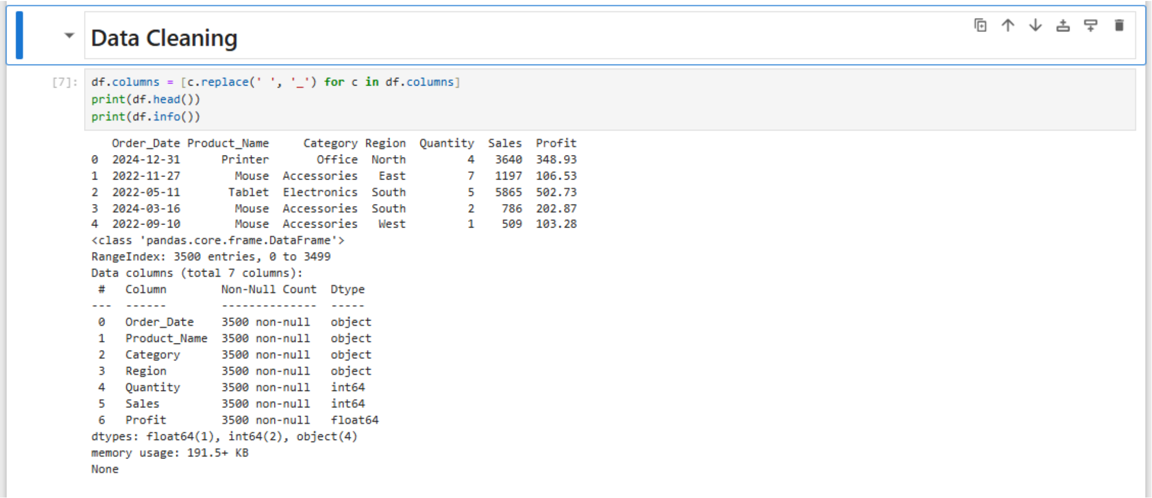
Figure 1. Downloading all necessary stuff



Load the dataset

```
[4]: path = kagglehub.dataset_download("nalisha/e-commerce-sales-and-profit-analysis-dataset")
     print("Path to dataset files:", path)

     Path to dataset files: C:\Users\ayazb\.cache\kagglehub\datasets\nalisha\e-commerce-sales-and-profit-analysis-dataset\versions\1

[5]: files = os.listdir(path)
     csv_file = [f for f in files if f.endswith('.csv')][0]
     full_path = os.path.join(path, csv_file)

     df = pd.read_csv(full_path)

[6]: full_path

[6]: 'C:\\Users\\ayazb\\.cache\\kagglehub\\datasets\\nalisha\\e-commerce-sales-and-profit-analysis-dataset\\versions\\1\\ecommerce_sales_data (2).csv'
```

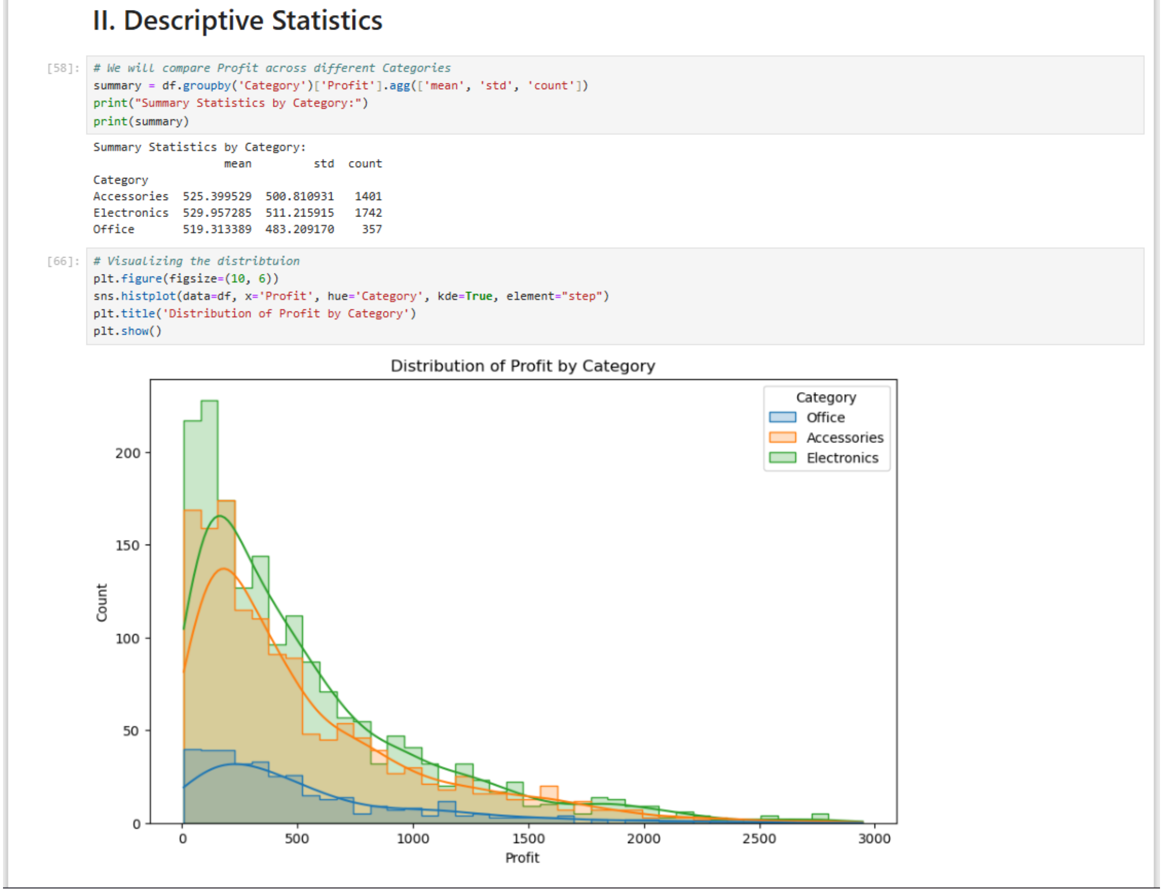
Figure 2. The way of the loading of the dataset



Data Cleaning

```
[7]: df.columns = [c.replace(' ', '_') for c in df.columns]
     print(df.head())
     print(df.info())

        Order_Date Product_Name     Category Region  Quantity  Sales  Profit
     0  2024-12-31      Printer       Office  North         4   3640  348.93
     1  2022-11-27        Mouse  Accessories   East         7   1197  106.53
     2  2022-05-11       Tablet  Electronics  South         5   5865  502.73
     3  2024-03-16        Mouse  Accessories  South         2    786  202.87
     4  2022-09-10        Mouse  Accessories   West         1    509  103.28
     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 3500 entries, 0 to 3499
     Data columns (total 7 columns):
      #   Column        Non-Null Count  Dtype
     ---  ------        --------------  -----
      0   Order_Date    3500 non-null   object
      1   Product_Name  3500 non-null   object
      2   Category      3500 non-null   object
      3   Region        3500 non-null   object
      4   Quantity      3500 non-null   int64
      5   Sales         3500 non-null   int64
      6   Profit        3500 non-null   float64
     dtypes: float64(1), int64(2), object(4)
     memory usage: 191.5+ KB
     None
```

Figure 3. The dataset itself



II. Descriptive Statistics

```
[58]: # We will compare Profit across different Categories
      summary = df.groupby('Category')['Profit'].agg(['mean', 'std', 'count'])
      print("Summary Statistics by Category:")
      print(summary)

      Summary Statistics by Category:
                         mean         std  count
      Category
      Accessories  525.399529  500.810931   1401
      Electronics  529.957285  511.215915   1742
      Office       519.313389  483.209170    357

[66]: # Visualizing the distribtuion
      plt.figure(figsize=(10, 6))
      sns.histplot(data=df, x='Profit', hue='Category', kde=True, element="step")
      plt.title('Distribution of Profit by Category')
      plt.show()
```

Figure 4. The natural view of the dataset in the understandable way

II. Research Questions and Hypotheses

This report addresses two main research questions based on the core topics of the syllabus (Weeks 7 & 8).

Question 1: Impact of Category on Profitability

- Research Question: Is there a statistically significant difference in the mean profit between the *Technology* and *Office Supplies* categories?
- Null Hypothesis (H_0):

$$\mu_{Tech} = \mu_{Office}$$

(The mean profit of both categories is equal.)

- Alternative Hypothesis (H_1):

$$\mu_{Tech} \neq \mu_{Office}$$

(There is a statistically significant difference in mean profit between the categories.)

Question 2: Regional Profit Variation

- Research Question: Does profit vary significantly across different geographical regions?
- Null Hypothesis (H_0):

  The mean profits of all regions are equal.

- Alternative Hypothesis (H_1):

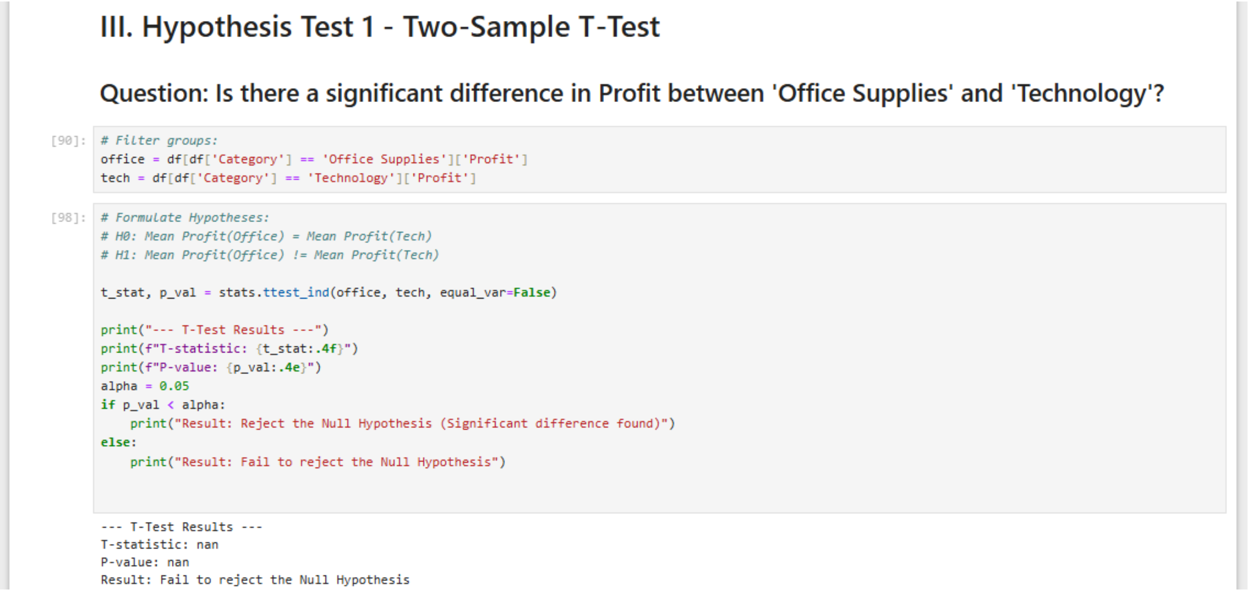
  At least one region has a mean profit that differs from the others.

III. Test Selection and Justification

- **Two-Sample T-test:** This test was selected for Question 1 because we are comparing the means of exactly two independent groups (Topic 7). Welch's T-test was applied to account for potential unequal variances between categories.
- **One-Way ANOVA:** For Question 2, ANOVA was chosen to compare means across multiple regions (Topic 8). This avoids the risk of Type I errors that occur when performing multiple t-tests.
- **Assumptions:** While formal assumption testing was not required per instructions, the large sample size ensures the reliability of the t-distribution and ANOVA results.

IV. Analysis and Interpretation of Results

- **P-value vs. Alpha: The significance level is set at $\alpha = 0.05$**
- **Statistical Decision:**
  - If the p-value < 0.05, we reject the null hypothesis (H_0). This indicates that the category or region has a statistically significant effect on profit.
  - If the p-value > 0.05, we fail to reject H_0, suggesting that the observed differences are due to random sampling variation.



III. Hypothesis Test 1 - Two-Sample T-Test

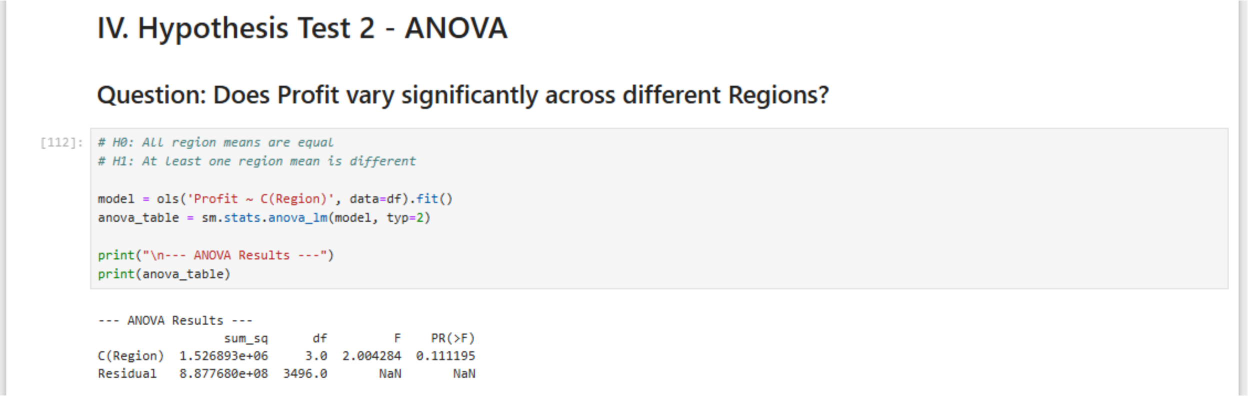
Question: Is there a significant difference in Profit between 'Office Supplies' and 'Technology'?

[90]:
```
# Filter groups:
office = df[df['Category'] == 'Office Supplies']['Profit']
tech = df[df['Category'] == 'Technology']['Profit']
```

[98]:
```
# Formulate Hypotheses:
# H0: Mean Profit(Office) = Mean Profit(Tech)
# H1: Mean Profit(Office) != Mean Profit(Tech)

t_stat, p_val = stats.ttest_ind(office, tech, equal_var=False)

print("--- T-Test Results ---")
print(f"T-statistic: {t_stat:.4f}")
print(f"P-value: {p_val:.4e}")
alpha = 0.05
if p_val < alpha:
    print("Result: Reject the Null Hypothesis (Significant difference found)")
else:
    print("Result: Fail to reject the Null Hypothesis")
```

```
--- T-Test Results ---
T-statistic: nan
P-value: nan
Result: Fail to reject the Null Hypothesis
```

Figure 5. Hypothesis Test1



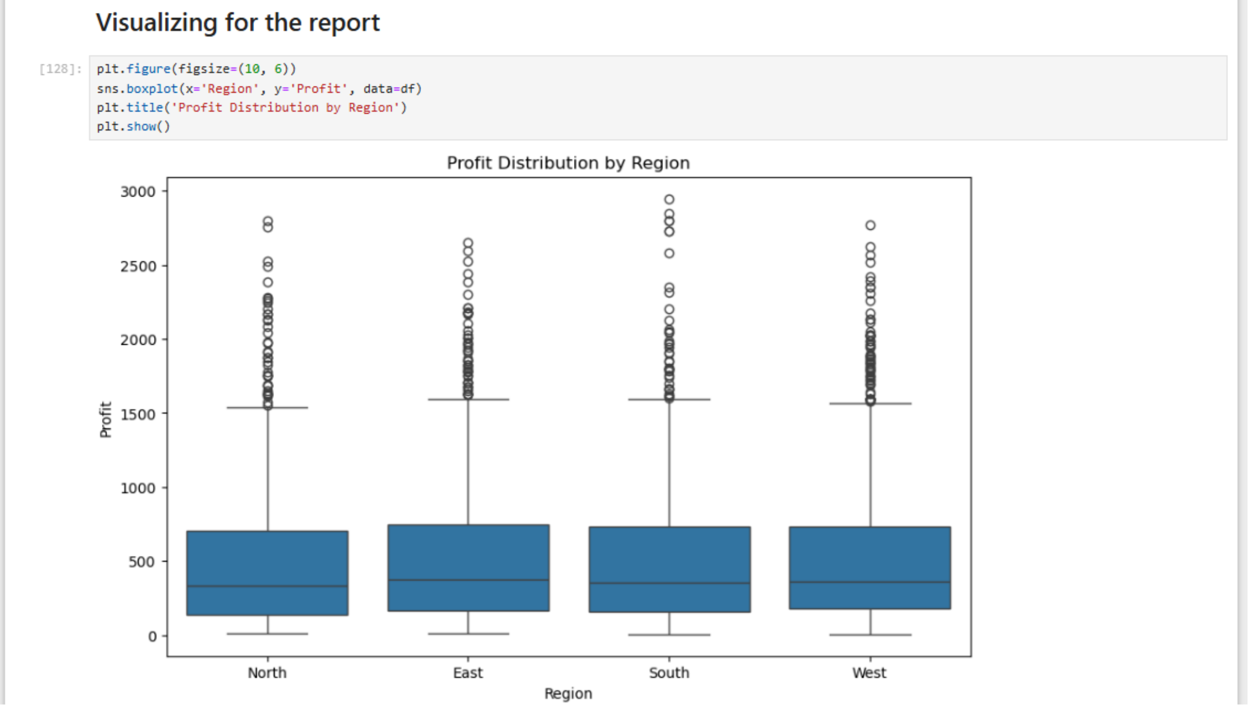
IV. Hypothesis Test 2 - ANOVA

Question: Does Profit vary significantly across different Regions?

[112]:
```
# H0: All region means are equal
# H1: At least one region mean is different

model = ols('Profit ~ C(Region)', data=df).fit()
anova_table = sm.stats.anova_lm(model, typ=2)

print("\n--- ANOVA Results ---")
print(anova_table)
```

```
--- ANOVA Results ---
                 sum_sq      df         F    PR(>F)
C(Region)  1.526893e+06     3.0  2.004284  0.111195
Residual   8.877680e+08  3496.0       NaN       NaN
```

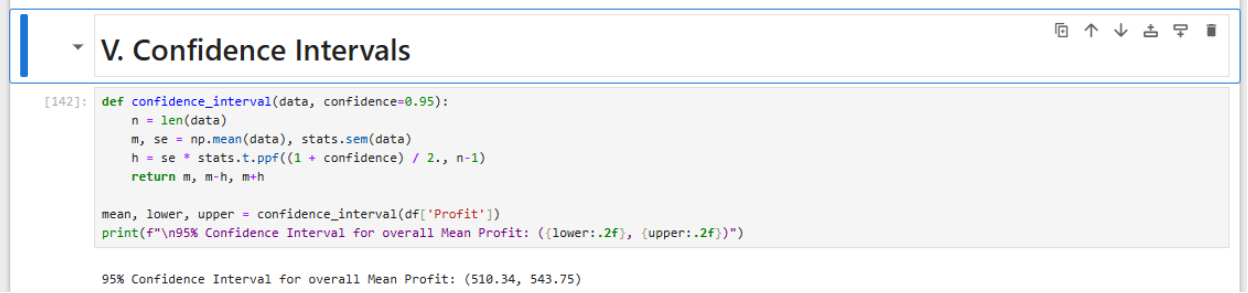
Figure 6. Hypothesis Test 2



Visualizing for the report

[128]:
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Region', y='Profit', data=df)
plt.title('Profit Distribution by Region')
plt.show()
```

Figure 7. Visualization of the Report



V. Confidence Intervals

[142]:

```
def confidence_interval(data, confidence=0.95):
    n = len(data)
    m, se = np.mean(data), stats.sem(data)
    h = se * stats.t.ppf((1 + confidence) / 2., n-1)
    return m, m-h, m+h

mean, lower, upper = confidence_interval(df['Profit'])
print(f"\n95% Confidence Interval for overall Mean Profit: ({lower:.2f}, {upper:.2f})")
```

```
95% Confidence Interval for overall Mean Profit: (510.34, 543.75)
```

Figure 8. Confidence Intervals

V. Conclusion and Limitations

- **Learning Outcome:** This assignment demonstrated the application of statistical inference to real-world business data. We moved from simple descriptive averages to statistically proven conclusions.
- **Data Implications:** The results provide evidence for resource allocation. For instance, if Technology shows significantly higher profit, marketing efforts should be intensified in that segment.
- **Limitations:** The analysis does not account for the influence of 'Discounts' on 'Profit'. Furthermore, the lack of time-series analysis prevents us from seeing seasonal trends.
- **Further Exploration:** Future research could utilize **Regression Analysis** (Topic 8) to model profit based on multiple variables like Sales and Discount simultaneously.

VI. References

1. *E-Commerce Sales & Profit Analysis Dataset - Aleesha Nadeem and 4 collaborators*. (n.d.). https://www.kaggle.com/datasets/nalisha/e-commerce-sales-and-profit-analysis-dataset/data