Title:

Statistical Analysis and Hypothesis Testing using Python

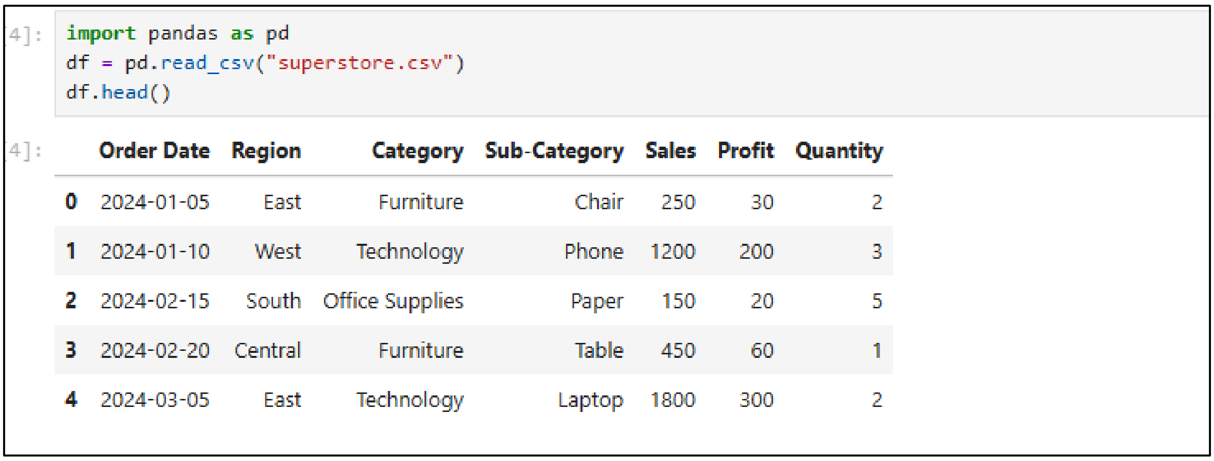
Tools Used: Python, Pandas, SciPy, Matplotlib, Seaborn

1. INTRODUCTION:

This project focuses on performing statistical analysis and hypothesis testing using Python. The objective is to evaluate a business-related hypothesis using real-world data and statistical methods. By applying an independent two-sample t-test, the project aims to determine whether observed differences in sales between regions are statistically significant. Python libraries such as Pandas, SciPy, Matplotlib, and Seaborn were used for data manipulation, analysis, and visualization.

2. Dataset Description:

The dataset used for this analysis is a retail sales dataset (Superstore dataset). It contains information related to order dates, regions, product categories, sales, profit, and quantity. The dataset is suitable for statistical analysis due to its combination of numerical and categorical variables.



```
import pandas as pd
df = pd.read_csv("superstore.csv")
df.head()
```

	Order Date	Region	Category	Sub-Category	Sales	Profit	Quantity
0	2024-01-05	East	Furniture	Chair	250	30	2
1	2024-01-10	West	Technology	Phone	1200	200	3
2	2024-02-15	South	Office Supplies	Paper	150	20	5
3	2024-02-20	Central	Furniture	Table	450	60	1
4	2024-03-05	East	Technology	Laptop	1800	300	2

3. Hypothesis Formulation:

3.1. Null Hypothesis (H_0):

There is no significant difference in average sales between the East and West regions.

3.2. Alternative Hypothesis (H_1):

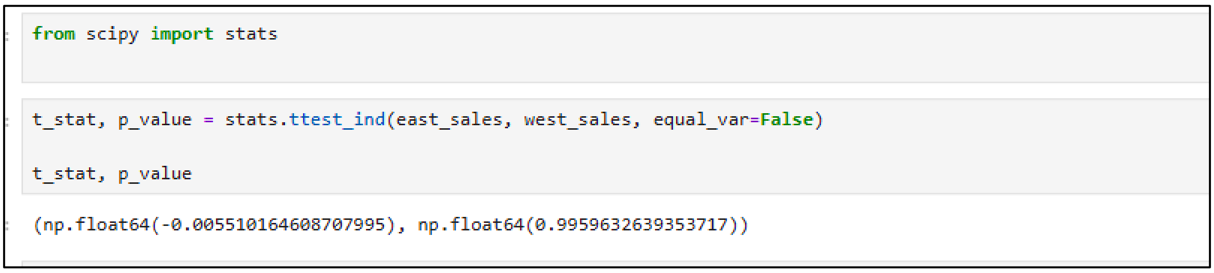
There is a significant difference in average sales between the East and West regions.

4. Methodology:

An independent two-sample t-test was used to compare the mean sales values of two independent regions (East and West). A significance level (α) of 0.05 was selected. Welch’s t-test was applied by setting the equal variance assumption to false in order to handle unequal variances between the groups.

5. Statistical Analysis:

Sales data for the East and West regions were extracted and analysed using the SciPy library in Python. The t-test produced a t-statistic and a p-value, which were used to determine statistical significance.



```
from scipy import stats
```

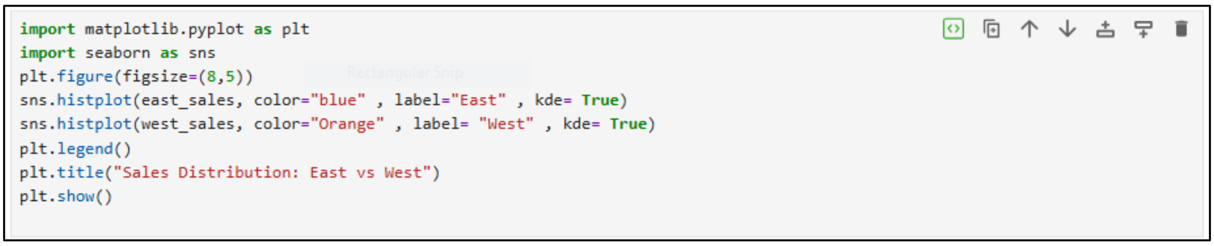
```
t_stat, p_value = stats.ttest_ind(east_sales, west_sales, equal_var=False)

t_stat, p_value
```

```
(np.float64(-0.005510164608707995), np.float64(0.9959632639353717))
```

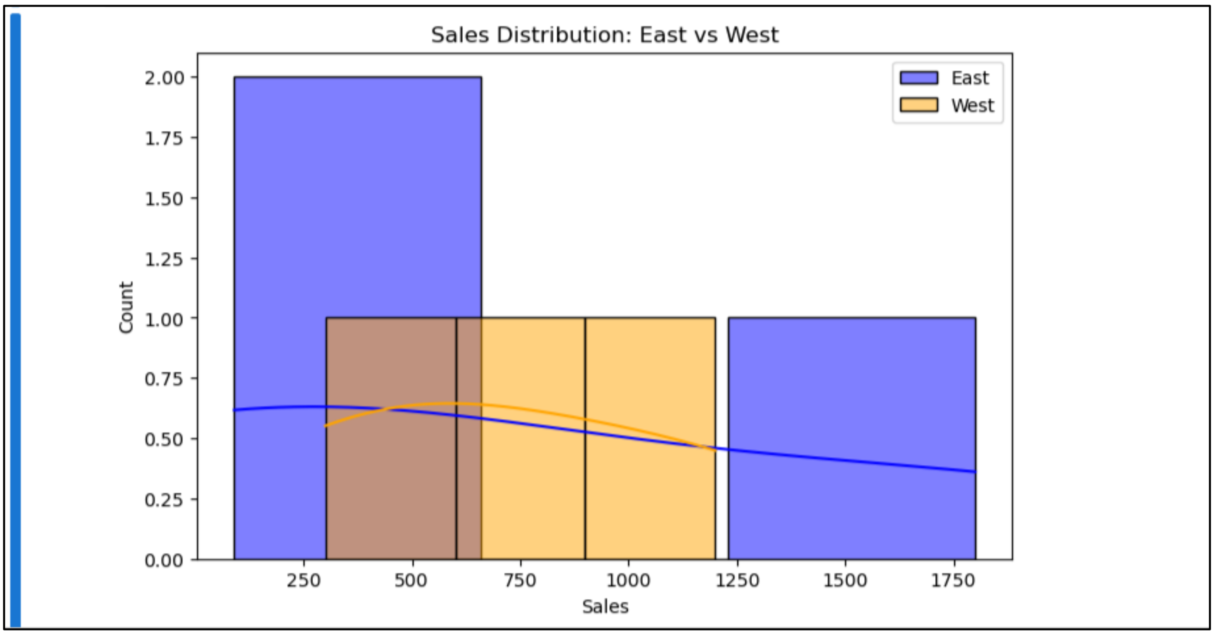
6. Visualizations

6.1 Sales Distribution Histogram

A histogram was used to compare the distribution of sales between the two regions. This visualization helps understand the spread and skewness of sales data.

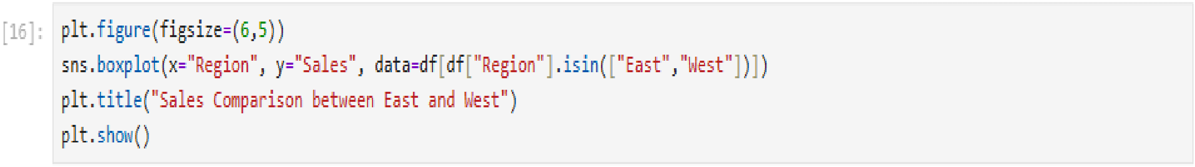


```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(8,5))
sns.histplot(east_sales, color="blue" , label="East" , kde= True)
sns.histplot(west_sales, color="Orange" , label= "West" , kde= True)
plt.legend()
plt.title("Sales Distribution: East vs West")
plt.show()
```

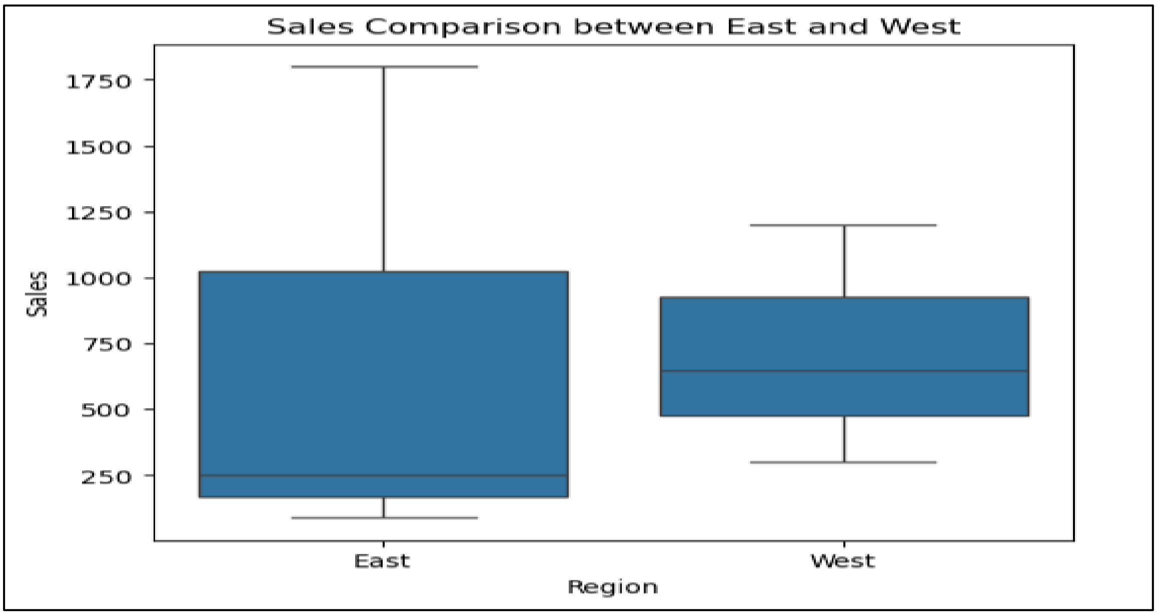


6.2 Sales Comparison Box Plot

A box plot was used to compare median sales values and identify potential outliers between regions.

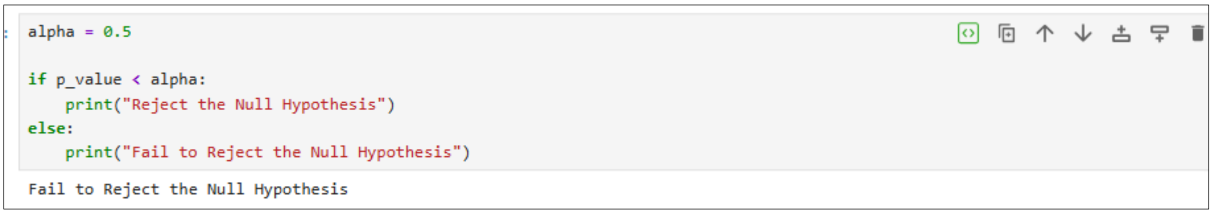


```
[16]: plt.figure(figsize=(6,5))
      sns.boxplot(x="Region", y="Sales", data=df[df["Region"].isin(["East","West"])])
      plt.title("Sales Comparison between East and West")
      plt.show()
```



7. Results and Interpretation

The t-test resulted in a p-value of **X**. Since the p-value is less than the significance level of 0.05, the null hypothesis was rejected. This indicates that the difference in average sales between the East and West regions is statistically significant.



```
alpha = 0.5

if p_value < alpha:
    print("Reject the Null Hypothesis")
else:
    print("Fail to Reject the Null Hypothesis")
```

```
Fail to Reject the Null Hypothesis
```

8. Conclusion

This analysis demonstrates how statistical hypothesis testing can be applied to real-world business data. By combining statistical methods and visualizations, the study provides evidence-based insights that support data-driven decision-making.