ASSIGNMENT

**NAME:**M.Gnana sai likhitha

Code:

from google.colab import files

uploaded = files.upload()

# Import libraries

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from scipy.stats import ttest\_ind, chi2\_contingency, f\_oneway

# Load dataset

df = pd.read\_csv("Superstore.csv", encoding='latin1')

# Select relevant numeric columns

numeric\_cols = ['Sales', 'Quantity', 'Discount', 'Profit']

# Summary statistics

summary\_stats = df[numeric\_cols].describe()

print("Summary Statistics:\n", summary\_stats)

# Save summary stats

summary\_stats.to\_csv("summary\_statistics.csv")

# Mode

mode\_vals = df[numeric\_cols].mode().iloc[0]

print("\nMode:\n", mode\_vals)

# Save mode

mode\_vals.to\_frame().T.to\_csv("mode\_values.csv", index=False)

# T-test: Profit comparison between East and West

east\_profit = df[df['Region'] == 'East']['Profit']

west\_profit = df[df['Region'] == 'West']['Profit']

t\_stat, p\_val = ttest\_ind(east\_profit, west\_profit)

print(f"\nT-Test (East vs West Profit): t-stat = {t\_stat:.4f}, p-value = {p\_val:.4f}")

print("Result:", "Reject H₀ → Profits are significantly different." if p\_val < 0.05 else "Fail to reject H₀ → No significant difference.")

# Chi-Square Test: Region vs Ship Mode

contingency = pd.crosstab(df['Region'], df['Ship Mode'])

chi2, chi\_p, dof, expected = chi2\_contingency(contingency)

print(f"\nChi-Square Test (Region vs Ship Mode): chi2 = {chi2:.4f}, p-value = {chi\_p:.4f}")

print("Result:", "Reject H₀ → Region and Ship Mode are dependent." if chi\_p < 0.05 else "Fail to reject H₀ → They are independent.")

# ANOVA: Profit across all 4 regions

south\_profit = df[df['Region'] == 'South']['Profit']

central\_profit = df[df['Region'] == 'Central']['Profit']

anova\_stat, anova\_p = f\_oneway(east\_profit, west\_profit, south\_profit, central\_profit)

print(f"\nANOVA (Profit by Region): F-stat = {anova\_stat:.4f}, p-value = {anova\_p:.4f}")

print("Result:", "Reject H₀ → At least one region has different profit." if anova\_p < 0.05 else "Fail to reject H₀ → No significant difference.")

# Pivot table: Total sales by region

region\_sales = pd.pivot\_table(df, index='Region', values='Sales', aggfunc='sum')

print("\nTotal Sales by Region:\n", region\_sales)

# Save pivot to CSV

region\_sales.to\_csv("region\_sales.csv")

# Visualizations

plt.figure(figsize=(14, 5))

# Sales Distribution

plt.subplot(1, 3, 1)

sns.histplot(df['Sales'], bins=30, kde=True)

plt.title("Sales Distribution")

# Profit by Region

plt.subplot(1, 3, 2)

sns.boxplot(x='Region', y='Profit', data=df)

plt.title("Profit by Region")

# Discount Distribution

plt.subplot(1, 3, 3)

sns.histplot(df['Discount'], bins=20, kde=True)

plt.title("Discount Distribution")

plt.tight\_layout()

plt.savefig("visualizations.png", dpi=300)

plt.show()

# Correlation heatmap

plt.figure(figsize=(6, 4))

sns.heatmap(df[numeric\_cols].corr(), annot=True, cmap='coolwarm')

plt.title("Correlation Heatmap")

plt.savefig("correlation\_heatmap.png", dpi=300)

plt.show()

output:



