Numaira Zaib

Statistical Methods

Z-test, T-test, Chi-Square, Regression analysis

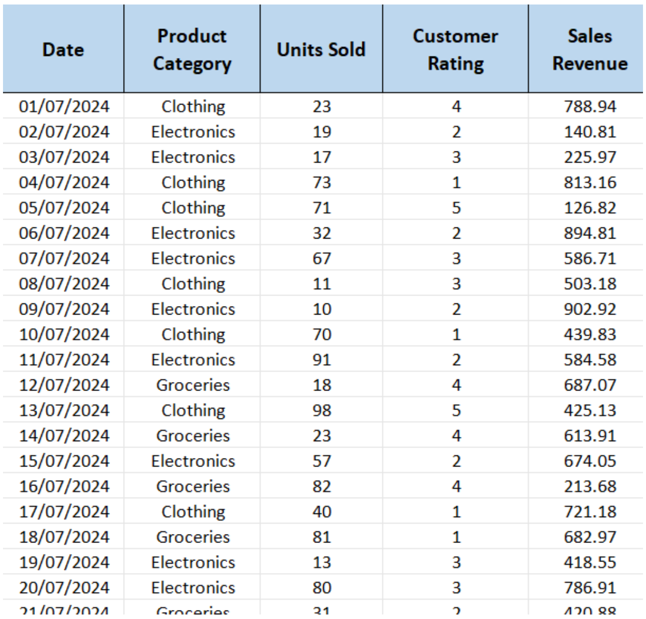
Retail-Sales-Data

This analysis will utilize a comprehensive retail dataset that includes key variables like ‘Product category’, ‘Date’, ‘Sales Revenue’, ‘Units Sold’, and ‘Customer Ratings’. This dataset is ideal for conducting a series of statistical tests Z-tests, T-tests, Chi-square tests, and regression analysis to evaluate the impact of these variables on sales outcomes and inform strategic decision-making.

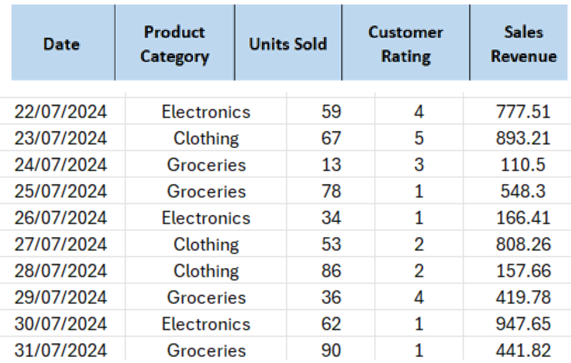
Dataset Overview: The dataset contains sales data with the following variables:

1. **Date**: The date of the sale (daily data).
2. **Product Category**: The category of the product sold (e.g., Clothing, Electronics, Groceries).
3. **Units Sold**: The number of units sold.
4. **Customer Rating**: Customer rating, on a scale from 1 to 5.
5. **Sales Revenue**: The revenue generated from sales.

The dataset has 31 entries, suggesting it might cover a month's data.



Date	Product Category	Units Sold	Customer Rating	Sales Revenue
01/07/2024	Clothing	23	4	788.94
02/07/2024	Electronics	19	2	140.81
03/07/2024	Electronics	17	3	225.97
04/07/2024	Clothing	73	1	813.16
05/07/2024	Clothing	71	5	126.82
06/07/2024	Electronics	32	2	894.81
07/07/2024	Electronics	67	3	586.71
08/07/2024	Clothing	11	3	503.18
09/07/2024	Electronics	10	2	902.92
10/07/2024	Clothing	70	1	439.83
11/07/2024	Electronics	91	2	584.58
12/07/2024	Groceries	18	4	687.07
13/07/2024	Clothing	98	5	425.13
14/07/2024	Groceries	23	4	613.91
15/07/2024	Electronics	57	2	674.05
16/07/2024	Groceries	82	4	213.68
17/07/2024	Clothing	40	1	721.18
18/07/2024	Groceries	81	1	682.97
19/07/2024	Electronics	13	3	418.55
20/07/2024	Electronics	80	3	786.91
21/07/2024	Groceries	31	2	420.88



Date	Product Category	Units Sold	Customer Rating	Sales Revenue
22/07/2024	Electronics	59	4	777.51
23/07/2024	Clothing	67	5	893.21
24/07/2024	Groceries	13	3	110.5
25/07/2024	Groceries	78	1	548.3
26/07/2024	Electronics	34	1	166.41
27/07/2024	Clothing	53	2	808.26
28/07/2024	Clothing	86	2	157.66
29/07/2024	Groceries	36	4	419.78
30/07/2024	Electronics	62	1	947.65
31/07/2024	Groceries	90	1	441.82

Statistical Tests & Regression Analysis: We will perform the following statistical analyses:

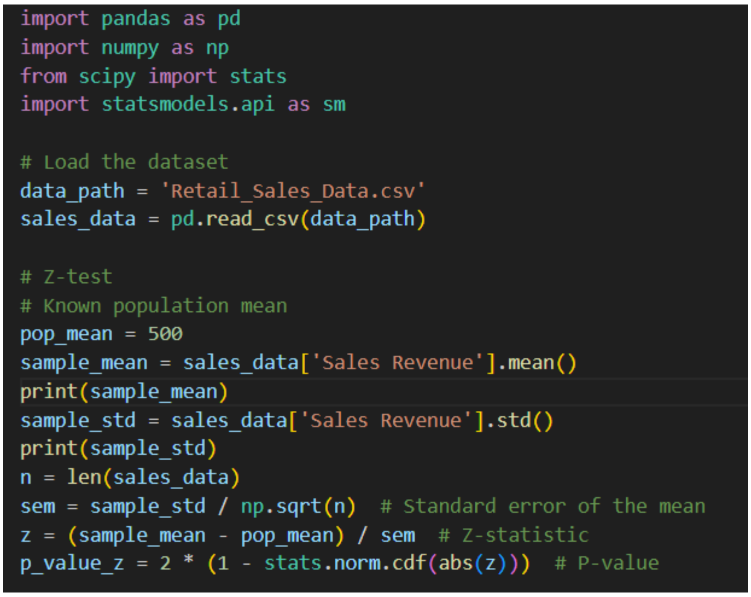
1. **Z-test**: Test if the average Sales Revenue differs from a hypothesized value.
2. **T-test**: Compare the Sales Revenue between two Product Categories.
3. **Chi-square test**: Check if the distribution of Customer Ratings matches an expected distribution.
4. **Regression analysis**: Model Sales Revenue as a function of Units Sold and Customer Rating.

Z - TEST

Hypothesis:

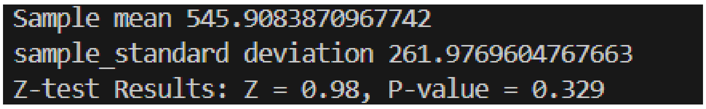
- Null hypothesis (H0): The mean sales revenue is equal to $500.
- Alternative hypothesis (H1): The mean sales revenue is not equal to $500.

Code



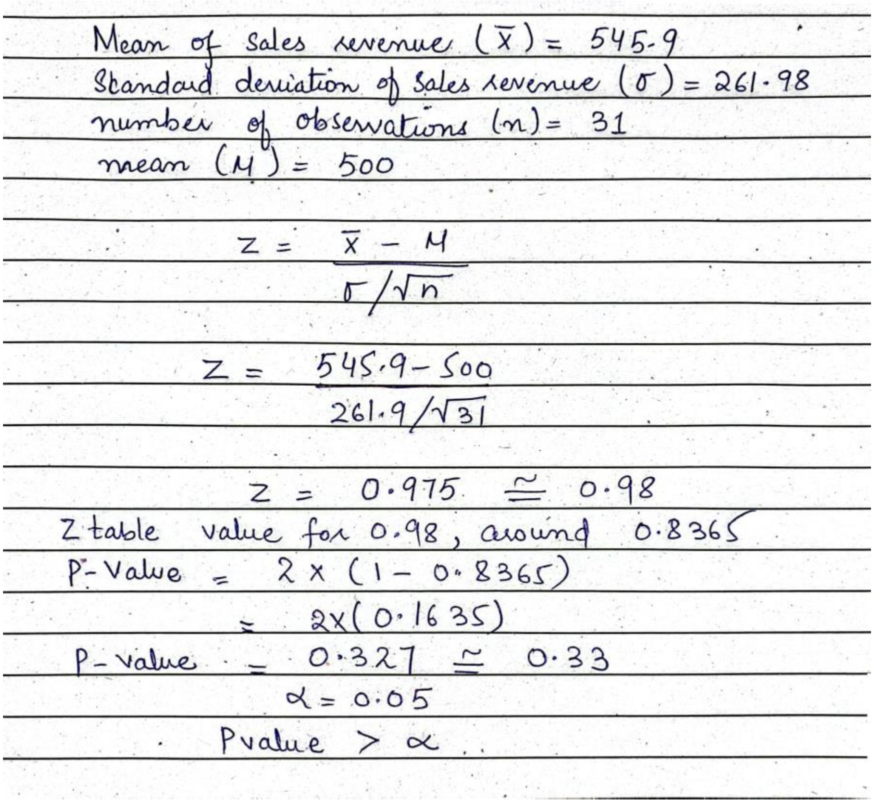
```
import pandas as pd
import numpy as np
from scipy import stats
import statsmodels.api as sm

# Load the dataset
data_path = 'Retail_Sales_Data.csv'
sales_data = pd.read_csv(data_path)

# Z-test
# Known population mean
pop_mean = 500
sample_mean = sales_data['Sales Revenue'].mean()
print(sample_mean)
sample_std = sales_data['Sales Revenue'].std()
print(sample_std)
n = len(sales_data)
sem = sample_std / np.sqrt(n)  # Standard error of the mean
z = (sample_mean - pop_mean) / sem  # Z-statistic
p_value_z = 2 * (1 - stats.norm.cdf(abs(z)))  # P-value
```

Results:



```
Sample mean 545.9083870967742
sample_standard deviation 261.9769604767663
Z-test Results: Z = 0.98, P-value = 0.329
```

Manual Calculations:



Mean of Sales revenue ($\bar{x}$) = 545.9

Standard deviation of Sales revenue (σ) = 261.98

number of observations (n) = 31

mean (μ) = 500

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$Z = \frac{545.9 - 500}{261.9/\sqrt{31}}$$

$$Z = 0.975 \simeq 0.98$$

Z table value for 0.98, around 0.8365

P-Value = 2 x (1 − 0.8365)

= 2x(0.1635)

P-value = 0.327 ≃ 0.33

$\alpha = 0.05$

Pvalue > α

Interpretation: With a p-value of 0.33, we fail to reject the null hypothesis at the 5% significance level. This suggests that there is not enough evidence to conclude that the mean sales revenue significantly differs from $500.

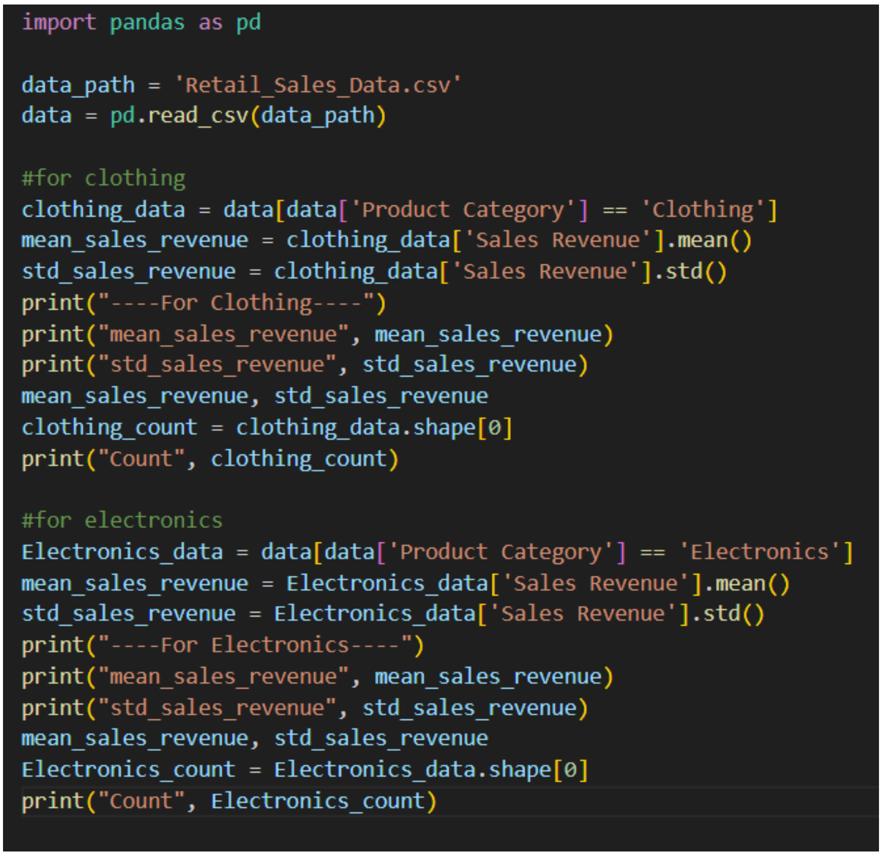
T-TEST

Hypothesis:

- Null hypothesis (H0): The mean sales revenue for 'Clothing' and 'Electronics' are equal.
- Alternative hypothesis (H1): The mean sales revenue for 'Clothing' and 'Electronics' are not equal.

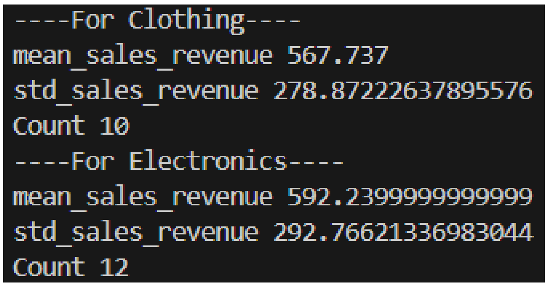
Finding Mean, Standard deviation and count for further calculations

Code

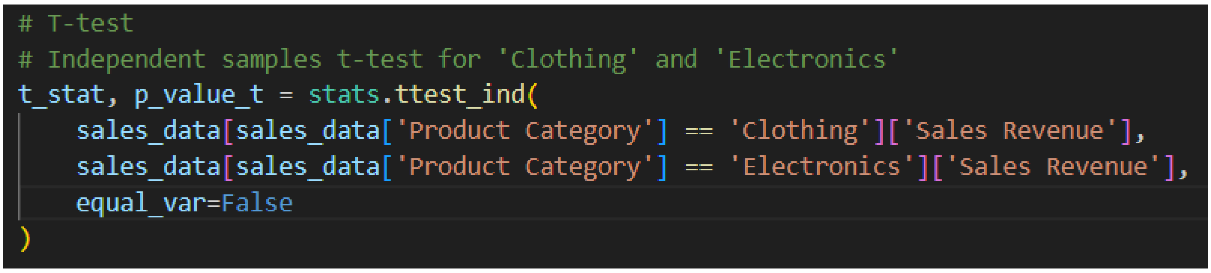


```
import pandas as pd

data_path = 'Retail_Sales_Data.csv'
data = pd.read_csv(data_path)

#for clothing
clothing_data = data[data['Product Category'] == 'Clothing']
mean_sales_revenue = clothing_data['Sales Revenue'].mean()
std_sales_revenue = clothing_data['Sales Revenue'].std()
print("----For Clothing----")
print("mean_sales_revenue", mean_sales_revenue)
print("std_sales_revenue", std_sales_revenue)
mean_sales_revenue, std_sales_revenue
clothing_count = clothing_data.shape[0]
print("Count", clothing_count)

#for electronics
Electronics_data = data[data['Product Category'] == 'Electronics']
mean_sales_revenue = Electronics_data['Sales Revenue'].mean()
std_sales_revenue = Electronics_data['Sales Revenue'].std()
print("----For Electronics----")
print("mean_sales_revenue", mean_sales_revenue)
print("std_sales_revenue", std_sales_revenue)
mean_sales_revenue, std_sales_revenue
Electronics_count = Electronics_data.shape[0]
print("Count", Electronics_count)
```

Results

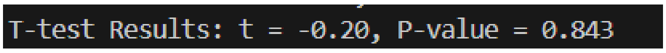


```
----For Clothing----
mean_sales_revenue 567.737
std_sales_revenue 278.87222637895576
Count 10
----For Electronics----
mean_sales_revenue 592.2399999999999
std_sales_revenue 292.76621336983044
Count 12
```



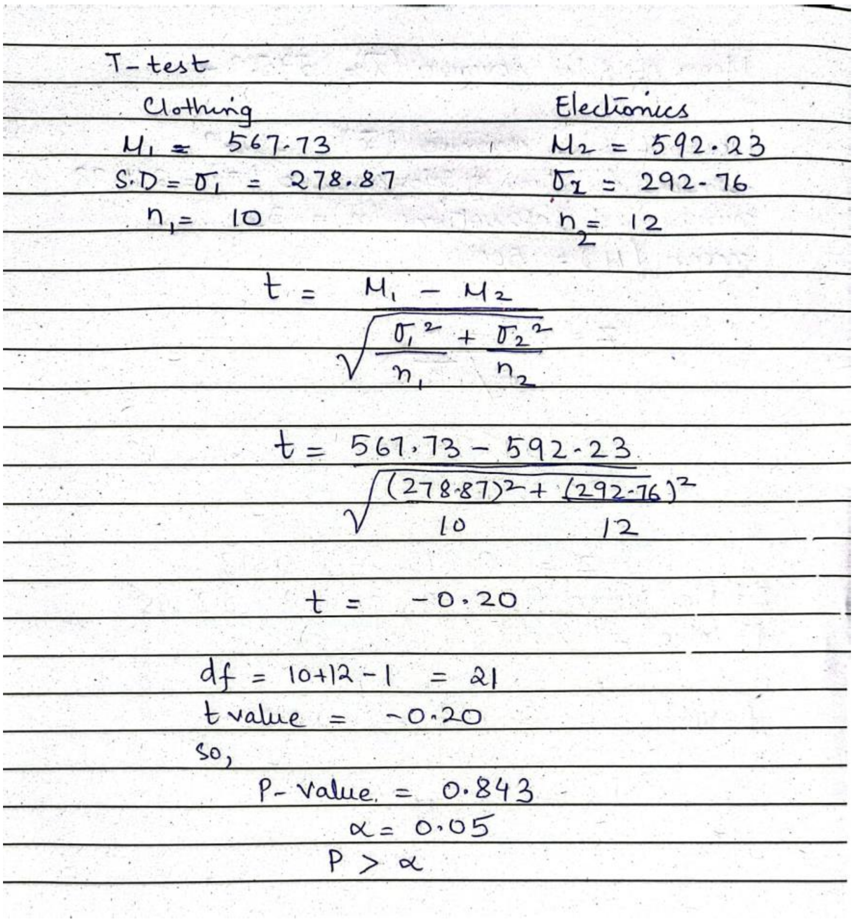
```
# T-test
# Independent samples t-test for 'Clothing' and 'Electronics'
t_stat, p_value_t = stats.ttest_ind(
    sales_data[sales_data['Product Category'] == 'Clothing']['Sales Revenue'],
    sales_data[sales_data['Product Category'] == 'Electronics']['Sales Revenue'],
    equal_var=False
)
```

Results:



```
T-test Results: t = -0.20, P-value = 0.843
```

Manual calculations:



T-test

Clothing

$M_1 = 567.73$

$S.D = \sigma_1 = 278.87$

$n_1 = 10$

Electronics

$M_2 = 592.23$

$\sigma_2 = 292.76$

$n_2 = 12$

$$t = \frac{M_1 - M_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$t = \frac{567.73 - 592.23}{\sqrt{\frac{(278.87)^2}{10} + \frac{(292.76)^2}{12}}}$$

$$t = -0.20$$

$df = 10+12-1 = 21$

t value $= -0.20$

so,

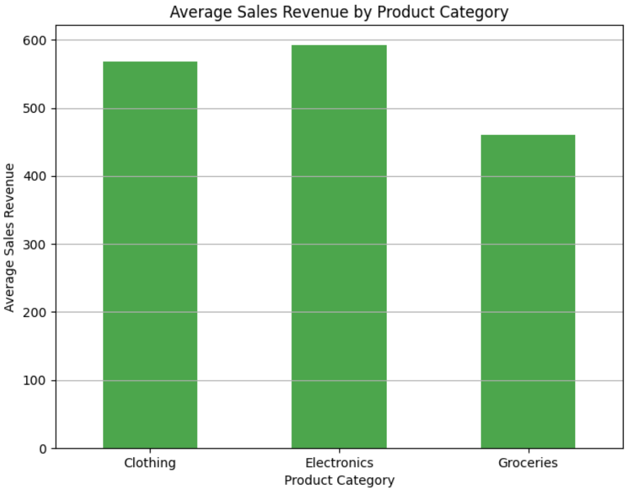
P-value $= 0.843$

$\alpha = 0.05$

$P > \alpha$

Interpretation: The p-value of 0.84 indicates that there is no significant difference in the mean sales revenue between the 'Clothing' and 'Electronics' categories. We fail to reject the null hypothesis, suggesting the means of the two categories are statistically similar.

Bar chart sales by category

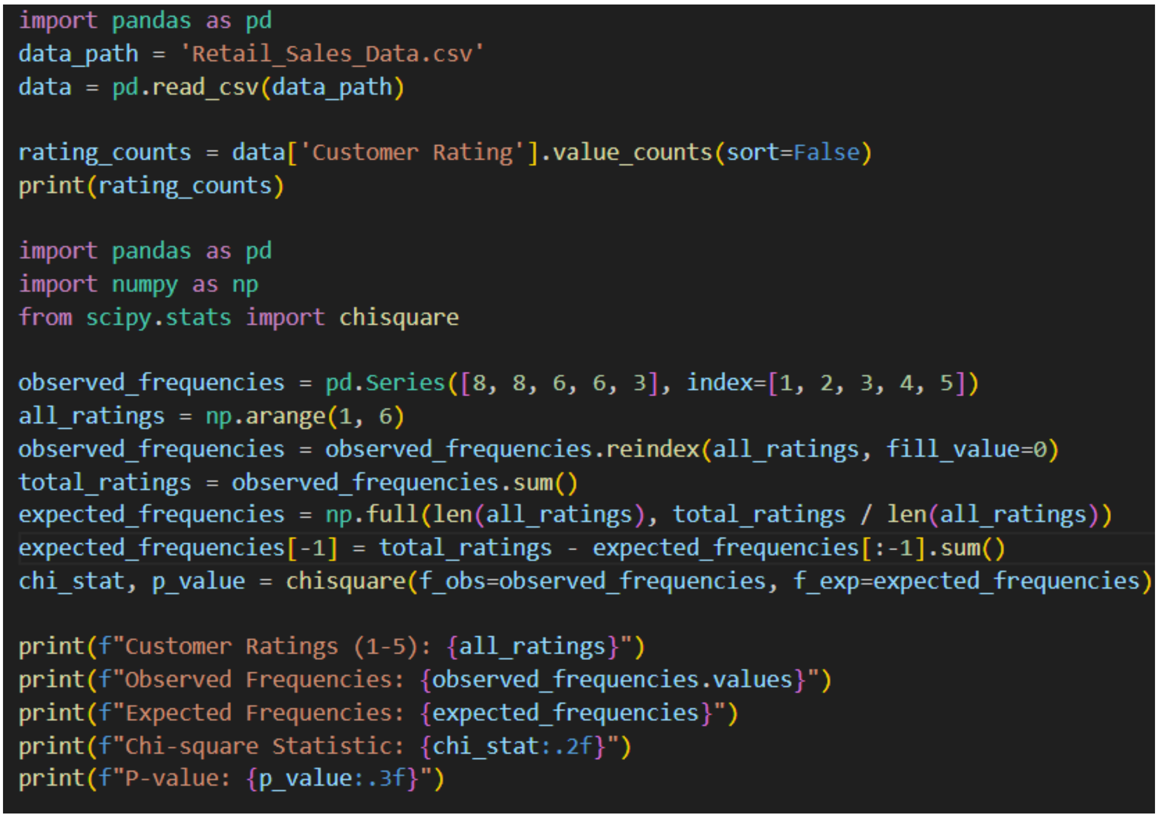


Chi-square test

Hypothesis:

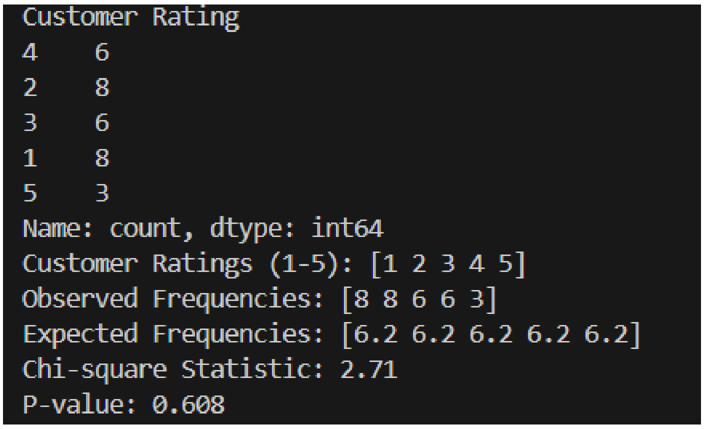
- Null hypothesis (H0): The distribution of customer ratings follows a uniform distribution.
- Alternative hypothesis (H1): The distribution of customer ratings does not follow a uniform distribution.

Code



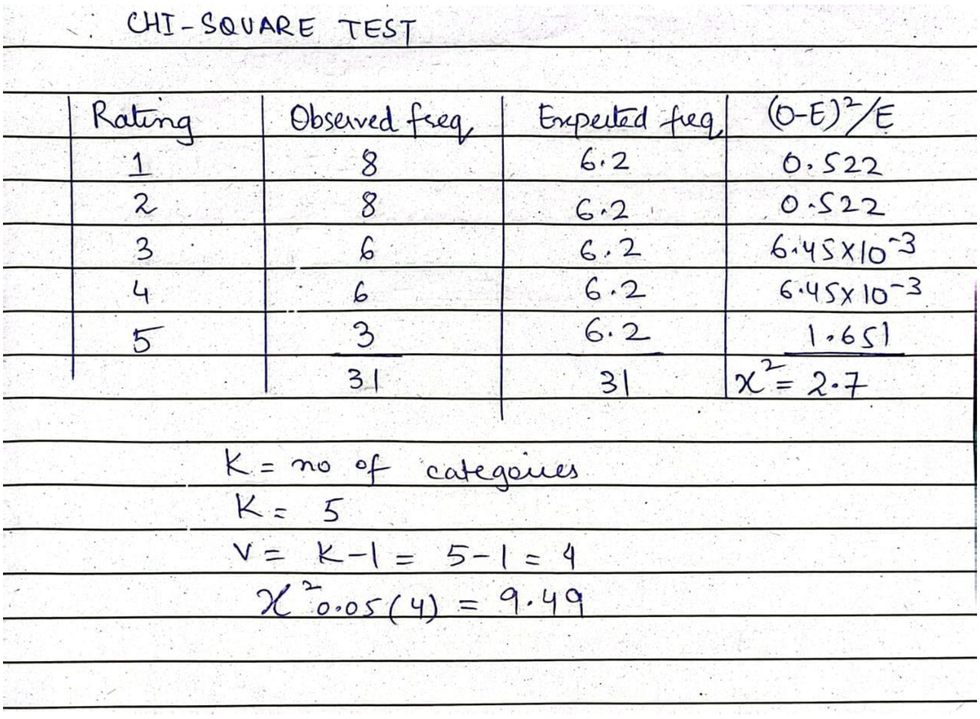
```python
import pandas as pd
data_path = 'Retail_Sales_Data.csv'
data = pd.read_csv(data_path)

rating_counts = data['Customer Rating'].value_counts(sort=False)
print(rating_counts)

import pandas as pd
import numpy as np
from scipy.stats import chisquare

observed_frequencies = pd.Series([8, 8, 6, 6, 3], index=[1, 2, 3, 4, 5])
all_ratings = np.arange(1, 6)
observed_frequencies = observed_frequencies.reindex(all_ratings, fill_value=0)
total_ratings = observed_frequencies.sum()
expected_frequencies = np.full(len(all_ratings), total_ratings / len(all_ratings))
expected_frequencies[-1] = total_ratings - expected_frequencies[:-1].sum()
chi_stat, p_value = chisquare(f_obs=observed_frequencies, f_exp=expected_frequencies)

print(f"Customer Ratings (1-5): {all_ratings}")
print(f"Observed Frequencies: {observed_frequencies.values}")
print(f"Expected Frequencies: {expected_frequencies}")
print(f"Chi-square Statistic: {chi_stat:.2f}")
print(f"P-value: {p_value:.3f}")
```

Results:



```
Customer Rating
4    6
2    8
3    6
1    8
5    3
Name: count, dtype: int64
Customer Ratings (1-5): [1 2 3 4 5]
Observed Frequencies: [8 8 6 6 3]
Expected Frequencies: [6.2 6.2 6.2 6.2 6.2]
Chi-square Statistic: 2.71
P-value: 0.608
```

Manual Calculations:



CHI-SQUARE TEST

Rating	Observed freq,	Expected freq,	$(O-E)^2/E$
1	8	6.2	0.522
2	8	6.2	0.522
3	6	6.2	6.45×10^{-3}
4	6	6.2	6.45×10^{-3}
5	3	6.2	1.651
	31	31	$\chi^2 = 2.7$

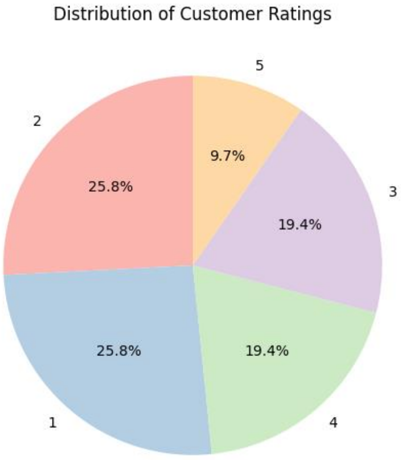
K = no of categories

K = 5

$V = K-1 = 5-1 = 4$

$\chi^2_{0.05(4)} = 9.49$

Pie Chart of customer rating



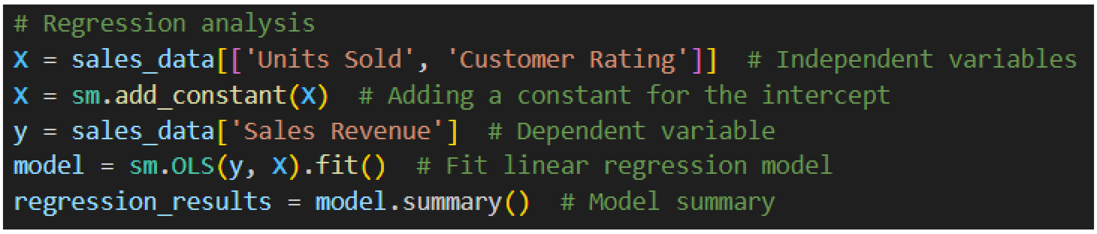
Interpretation: The p-value of 0.61 indicates that we fail to reject the null hypothesis. This suggests that the observed distribution of customer ratings does not significantly differ from a uniform distribution.

Regression analysis

Hypothesis:

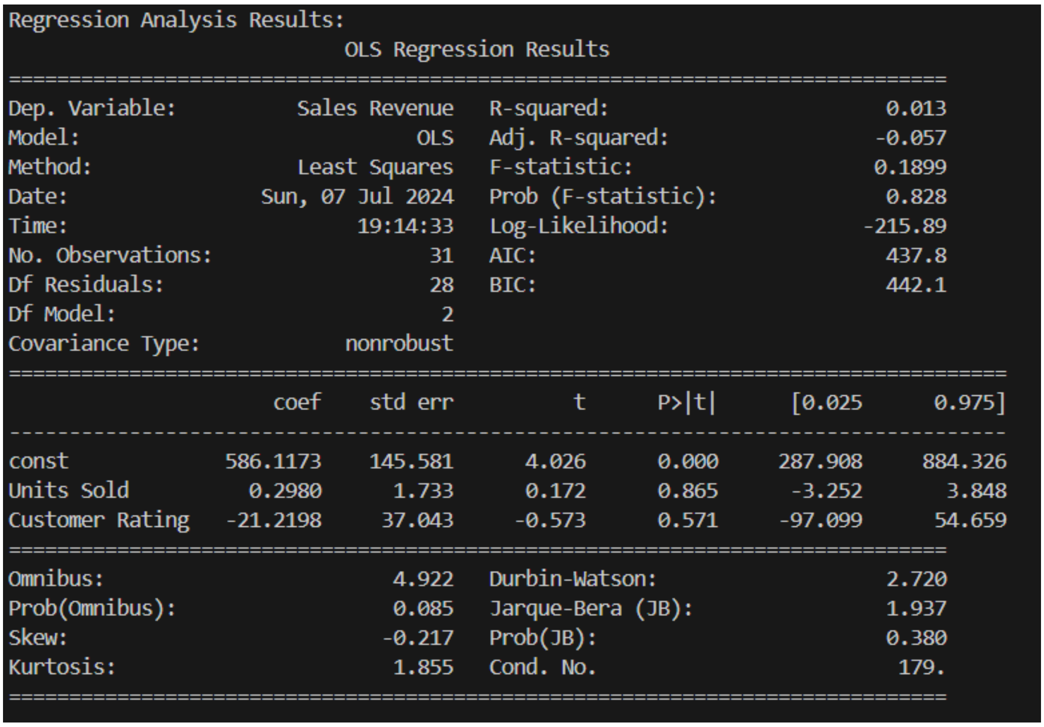
- Null hypothesis (H0): There is no relationship between the independent variables (Units Sold, Customer Rating) and the dependent variable (Sales Revenue).
- Alternative hypothesis (H1): There is a relationship between the independent variables and the dependent variable.

Code



```
# Regression analysis
X = sales_data[['Units Sold', 'Customer Rating']]  # Independent variables
X = sm.add_constant(X)  # Adding a constant for the intercept
y = sales_data['Sales Revenue']  # Dependent variable
model = sm.OLS(y, X).fit()  # Fit linear regression model
regression_results = model.summary()  # Model summary
```

Results:



```
Regression Analysis Results:
                            OLS Regression Results                            
==============================================================================
Dep. Variable:          Sales Revenue   R-squared:                       0.013
Model:                            OLS   Adj. R-squared:                 -0.057
Method:                 Least Squares   F-statistic:                    0.1899
Date:                Sun, 07 Jul 2024   Prob (F-statistic):              0.828
Time:                        19:14:33   Log-Likelihood:                -215.89
No. Observations:                  31   AIC:                             437.8
Df Residuals:                      28   BIC:                             442.1
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const             586.1173    145.581      4.026      0.000     287.908     884.326
Units Sold          0.2980      1.733      0.172      0.865      -3.252       3.848
Customer Rating   -21.2198     37.043     -0.573      0.571     -97.099      54.659
==============================================================================
Omnibus:                        4.922   Durbin-Watson:                   2.720
Prob(Omnibus):                  0.085   Jarque-Bera (JB):                1.937
Skew:                          -0.217   Prob(JB):                        0.380
Kurtosis:                       1.855   Cond. No.                         179.
==============================================================================
```

Interpretation: The model has a very low R-squared value of 0.013, indicating that only 1.3% of the variability in 'Sales Revenue' is explained by 'Units Sold' and 'Customer Rating'. The high p-values for both coefficients suggest that neither 'Units Sold' nor 'Customer Rating' significantly predicts 'Sales Revenue'.

Despite the intuitive expectation that 'Units Sold' and 'Customer Ratings' would significantly impact 'Sales Revenue,' the current statistical analysis did not yield significant results. This discrepancy may stem from several factors, including potential inadequacies in data quality or the simplicity of the model used, which might not adequately capture the complex dynamics between the variables. To address these issues, it is recommended to enhance the regression model by incorporating interaction effects or nonlinear relationships, which could provide a more accurate depiction of how these factors influence sales revenue. Additionally, expanding the dataset to include a broader range of variables and a larger sample size will likely improve the robustness of the findings. Implementing advanced analytical techniques, such as machine learning, could also uncover deeper insights that traditional methods might miss. Further, conducting segmented analyses based on different product types or customer demographics could reveal specific patterns or effects that are not apparent in a more generalized analysis. These steps are expected to refine the understanding of the data and align the findings more closely with typical business expectations.

Conclusion:

This project used a variety of statistical tests, like Z-tests, T-tests, Chi-square tests, and regression analysis, to explore how ‘product category’, ‘Units Sold’, and ‘Customer Ratings’ affect ‘Sales Revenue’ in a retail setting. This challenges the usual belief that selling more units and high customer satisfaction directly boost revenue. I also created detailed charts and processed data carefully to make the findings clearer and more useful. This work sets the stage for further research to look into other factors or to use more advanced methods to better understand what drives sales.