Data Mining: Supervised Learning Project

Daily Cup Order Projections

Predicting Quantity Demand at Starbucks

Presented by :
Nikhil Panjwani RBA34
Parshav Goel RBA35
Pragya Gupta RBA36

Data Description

Objective: Predict the number of cups of coffee per day based on customer attributes to optimize inventory, staffing, and marketing strategies.

Our dataset comprises of five variables collected to analyze and predict the daily cup order quantities at Starbucks locations. The variables include:

S.no	Variable Name	Data Type	Variable Type	Average Value	Description
1	Amount of Prepaid Card	Continuous	Numerical	31.8	Amount loaded in the Starbucks prepaid card.
2	Age	Discrete	Numerical	32.72	Age of the customer.
3	Days per Month at Starbucks	Discrete	Numerical	10.76	Number of days customer visits Starbucks in a month.
4	Cups of Coffee per Day	Discrete	Numerical	4.6	Number of cups consumed by customer daily.
5	Income	Continuous	Numerical	36.2	Customer's income level.

Based on 25 observations

Correlation

A. Positive Correlations:

Days per month at Starbucks & Cups of Coffee per day: 0.44
↳ Suggests that people who go more frequently to Starbucks tend to drink more coffee and possibly have higher income.

Amount of Prepaid card & Days per month at Starbucks: 0.3
↳ Makes sense; more frequent visitors likely load more on their prepaid cards.

B. Negative Correlations:

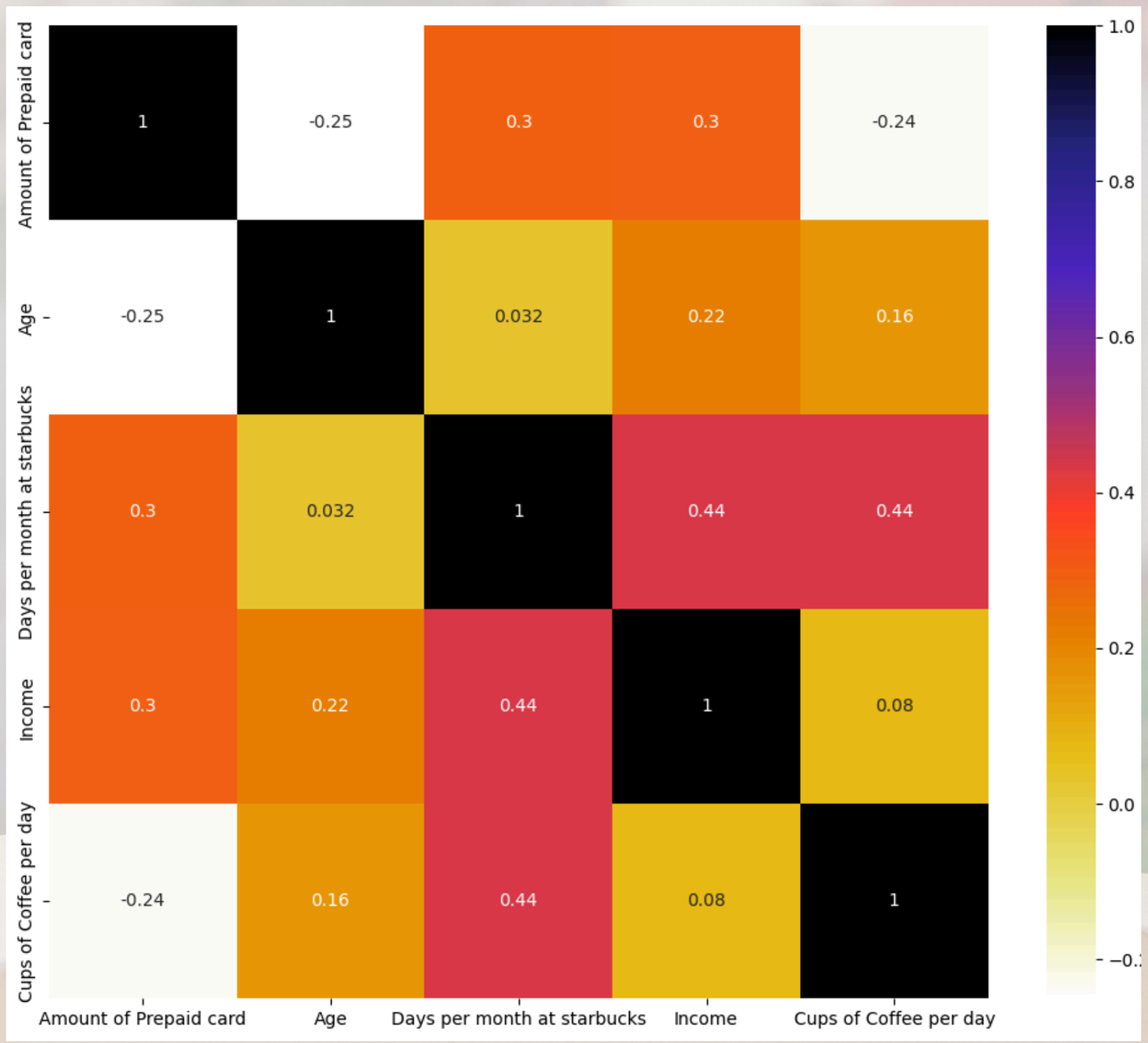
Amount of Prepaid card & Age: -0.25
↳ Older individuals may be less inclined to use prepaid cards.

Amount of Prepaid card & Cups of Coffee per day: -0.24
↳ Slightly unexpected; could imply heavy coffee drinkers don't rely on prepaid cards as much.

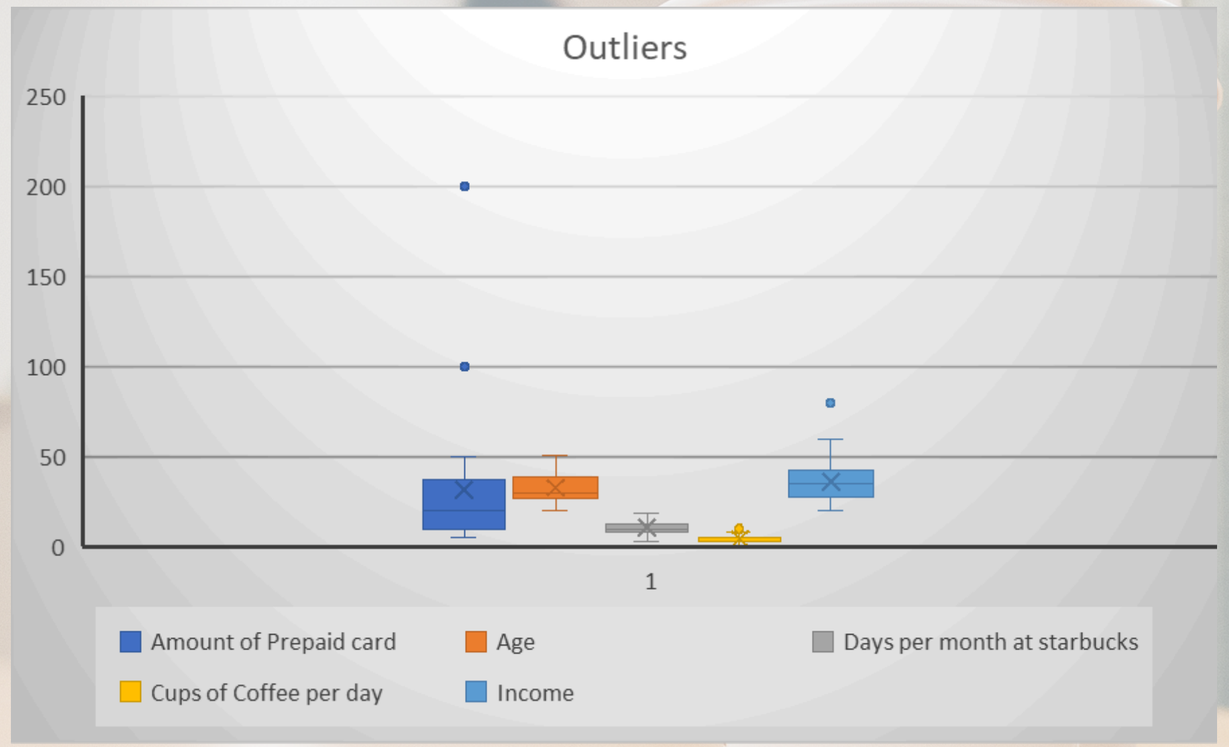
C. Weak or No Correlation:

Age & Days per month at Starbucks: 0.032
↳ Almost no relationship.

Income & Cups of Coffee per day: 0.08
↳ Weak connection between income and daily coffee intake.



Outliers



	count	mean	std	min	25%	50%	75%
Amount of Prepaid card	25.0	25.60	19.635745	5.0	10.0	20.0	35.0
Age	25.0	32.72	8.403967	20.0	27.0	30.0	38.0
Days per month at starbucks	25.0	10.72	3.931921	3.0	8.0	10.0	12.0
Income	25.0	35.00	10.801234	20.0	30.0	35.0	40.0
Cups of Coffee per day	25.0	4.52	2.181742	1.0	3.0	5.0	5.0

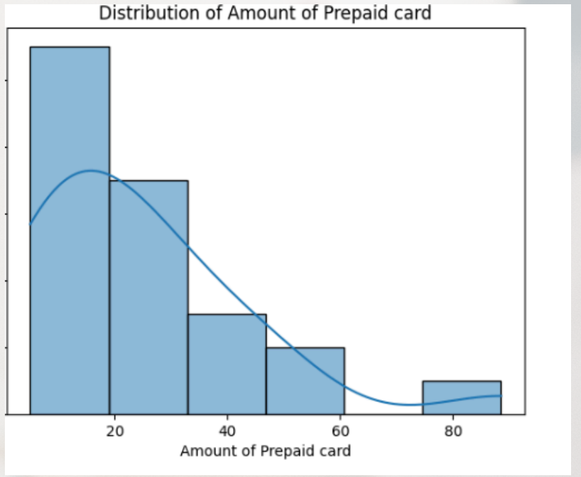
	max
Amount of Prepaid card	72.5
Age	51.0
Days per month at starbucks	18.0
Income	55.0
Cups of Coffee per day	8.0

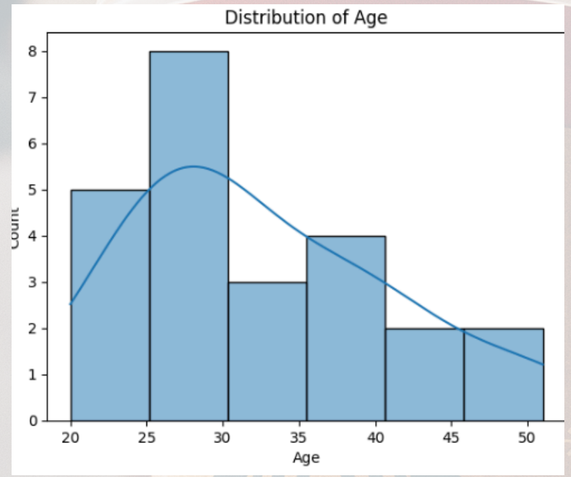
One customer with $200 prepaid only consumes 2 cups/day (likely a corporate buyer), while another customer wit $30 prepaid drinks 10 cups/day (possibly a barista or heavy caffeine user).

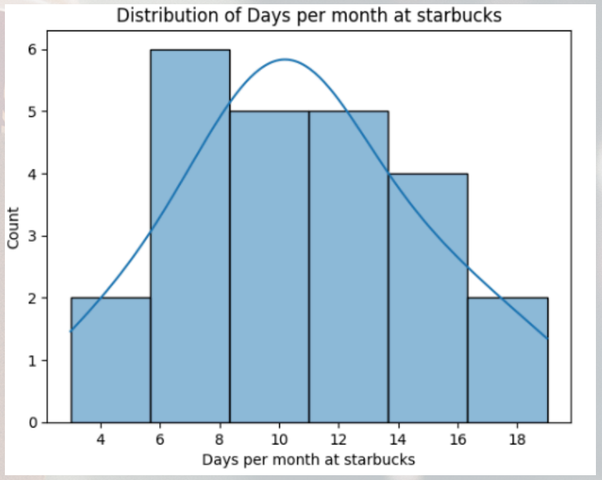
Treatment:

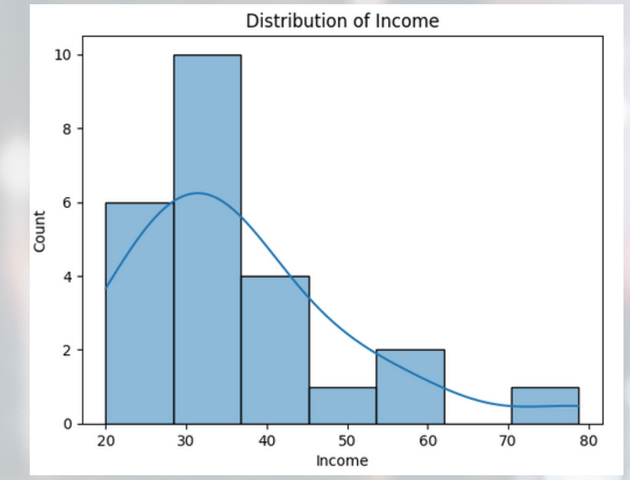
- When we ran the box whisker plot on the raw data, we identified two outlier values in the price column and one outlier in the income column.
- We have treated outliers through IQR, and replaced it with maximum acceptable value.

Univariate Analysis









Spending Habits (Prepaid Card Usage)

Most customers load 5–50.
High prepaid amounts (>100) correlate with frequent visits but fewer cups/day.

Customer Demographic (Age Distribution)

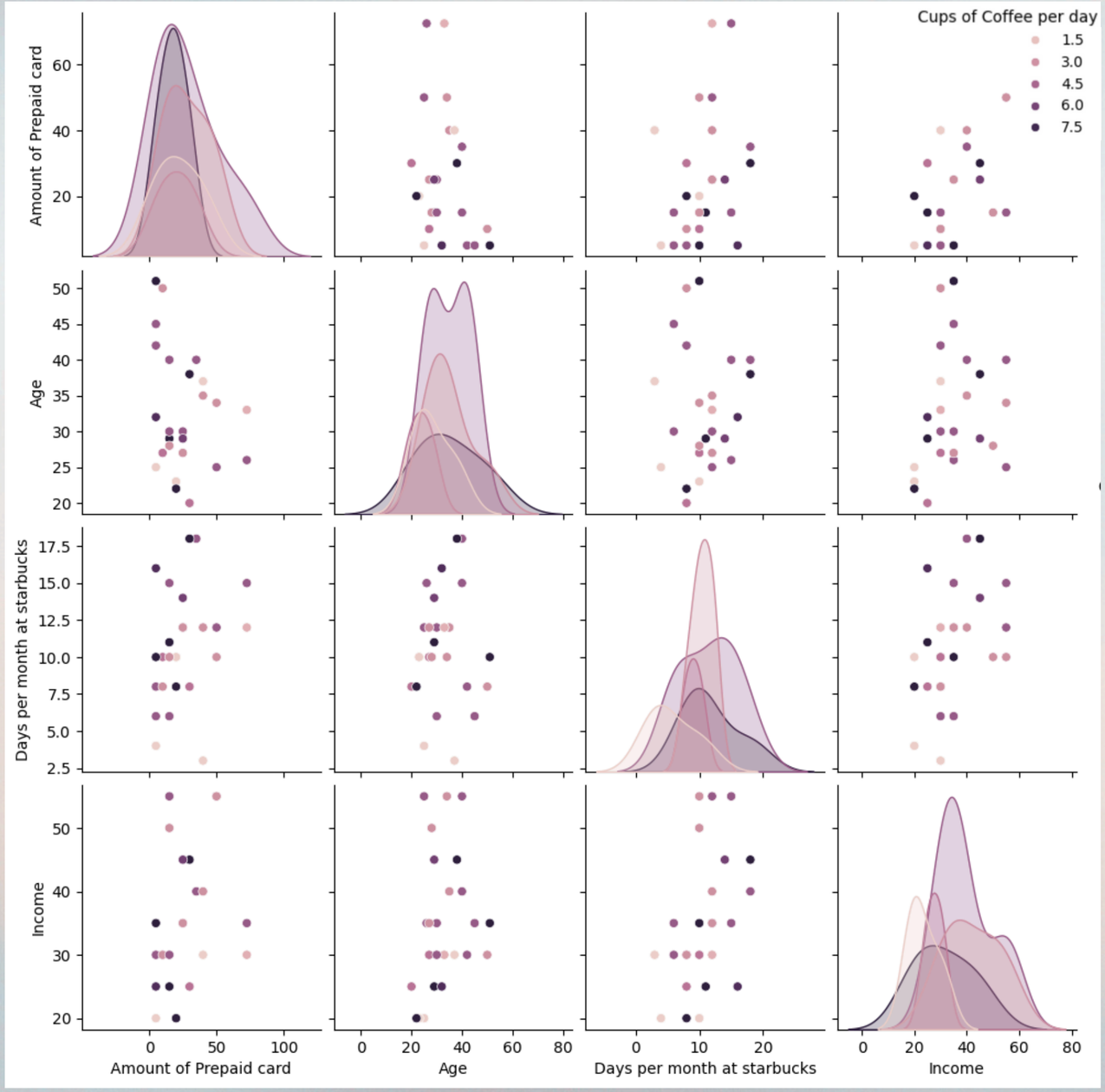
Majority aged 25-40 (young professionals).
Older customers (50+) visit less frequently but order more cups per visit.

Average Customer Visits: 8 to 12 Times per Month

15+ visits/month-
Highly loyal segment
Light & Heavy Users Exist in Smaller Numbers

Income:

Income peaks at $30k-$40k.
Rightward skewness indicates predominance of moderate to upper-middle income levels.



Multi-Variate Analysis

- **Days per month at Starbucks:** people drinking more coffee tend to visit more frequently.
- **Age vs. Coffee Consumption:** No strong visible pattern. Coffee drinkers are fairly spread out across ages.
- **Prepaid Card vs. Cups of Coffee:** Moderate prepaid card usage correlates with more coffee, but not strongly.
- **Higher income ($50K+)** → More prepaid spending, but not necessarily more cups/day.
- **Lower income (<$30K)** → Fewer visits, but higher per-visit consumption (e.g., 8–10 cups/day).
- **High-frequency visitors** (12+ days/month): 3–5 cups/day.
- **Low-frequency visitors** (<8 days/month): 1–2 cups/day.
- There is **no strong linear trend** visible in most plots.
- Days at Starbucks may be **weakly positively** associated with Cups per Day, but still noisy.

Multiple Linear Regression Analysis

OLS Regression Results

Dep. Variable:	Cups of Coffee per day	R-squared:	0.344
Model:	OLS	Adj. R-squared:	0.212
Method:	Least Squares	F-statistic:	2.617
Date:	Wed, 09 Apr 2025	Prob (F-statistic):	0.0659
Time:	21:31:33	Log-Likelihood:	-49.204
No. Observations:	25	AIC:	108.4
Df Residuals:	20	BIC:	114.5
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.0994	2.037	1.031	0.315	-2.150	6.348
Amount of Prepaid card	-0.0408	0.023	-1.792	0.088	-0.088	0.007
Age	0.0182	0.051	0.354	0.727	-0.089	0.125
Days per month at starbucks	0.3195	0.114	2.806	0.011	0.082	0.557
Income	-0.0158	0.044	-0.363	0.720	-0.107	0.075

Omnibus:	0.604	Durbin-Watson:	1.216
Prob(Omnibus):	0.739	Jarque-Bera (JB):	0.411
Skew:	0.299	Prob(JB):	0.814
Kurtosis:	2.806	Cond. No.	300.

As per anova, the F-Stat value > 0.05, which means the model is insignificant.

Hypotheses

H_0: $\beta_1 = \beta 2 = \beta 3 = \beta 4 = 0$ (no effect on Cups of Coffee per Day).

H_1: Atleast one β is not equal to 0

Since p value >0.05 in our three Independent variables,

- Amount of Prepaid Card
- Age
- Income

we will perform step-wise regression

Insights

- Each extra Starbucks visit per month increases daily coffee consumption by 0.32 cups
- The model explains only 34.4% of the variation in coffee consumption can be explained by the variables that we have taken into consideration.
- Prepaid card balance and income have no significant effect on coffee consumption, despite slight negative trends.

Multiple Linear Regression Analysis

```
Removing column: Age
                            OLS Regression Results
==============================================================================
Dep. Variable:     Cups of Coffee per day   R-squared:                       0.339
Model:                               OLS   Adj. R-squared:                  0.245
Method:                    Least Squares   F-statistic:                     3.598
Date:                   Wed, 09 Apr 2025   Prob (F-statistic):             0.0306
Time:                           21:31:37   Log-Likelihood:                -49.282
No. Observations:                     25   AIC:                             106.6
Df Residuals:                         21   BIC:                             111.4
Df Model:                              3
Covariance Type:               nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                            2.6033      1.427      1.825      0.082      -0.363       5.570
Amount of Prepaid card          -0.0435      0.021     -2.070      0.051      -0.087       0.000
Days per month at starbucks      0.3191      0.111      2.863      0.009       0.087       0.551
Income                          -0.0111      0.041     -0.274      0.787      -0.096       0.073
==============================================================================
Omnibus:                        0.396   Durbin-Watson:                   1.256
Prob(Omnibus):                  0.820   Jarque-Bera (JB):                0.261
Skew:                           0.231   Prob(JB):                        0.878
Kurtosis:                       2.806   Cond. No.                         179.
==============================================================================
```

- **Since age was the most insignificant variable, we dropped it first and then income.**
- The model here now has become significant, since Prob (F-statistic) < 0.05
- About 33.7% of the variation in daily coffee consumption can be explained by the model. This is moderate, suggesting other unmodeled factors (e.g., taste preferences, location, or marketing) also play a role.
- For every $1 increase in the prepaid card amount, coffee consumption decreases by approximately 0.0447 cups per day.

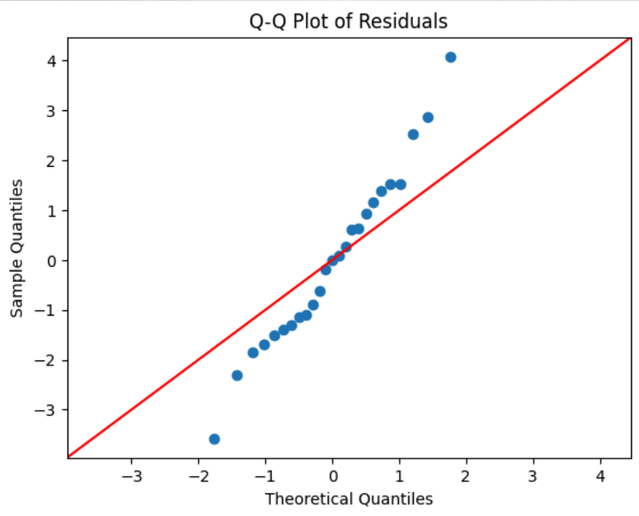
- After dropping the variable income, the independent variable, Amount of Prepaid Card has also become significant.
- This indicates Multi-collinearity

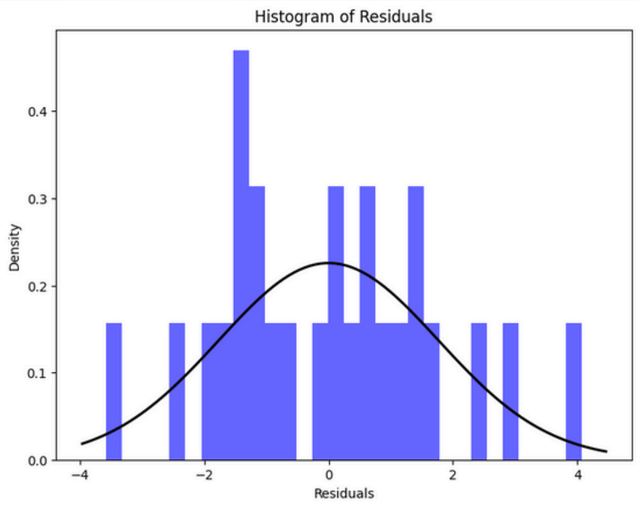
```
Removing column: Income
                            OLS Regression Results
==============================================================================
Dep. Variable:     Cups of Coffee per day   R-squared:                       0.337
Model:                               OLS   Adj. R-squared:                  0.277
Method:                    Least Squares   F-statistic:                     5.595
Date:                   Wed, 09 Apr 2025   Prob (F-statistic):             0.0109
Time:                           21:31:37   Log-Likelihood:                -49.327
No. Observations:                     25   AIC:                             104.7
Df Residuals:                         22   BIC:                             108.3
Df Model:                              2
Covariance Type:               nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                            2.3681      1.115      2.124      0.045       0.056       4.680
Amount of Prepaid card          -0.0447      0.020     -2.213      0.038      -0.087      -0.003
Days per month at starbucks      0.3074      0.101      3.050      0.006       0.098       0.516
==============================================================================
Omnibus:                        0.713   Durbin-Watson:                   1.248
Prob(Omnibus):                  0.700   Jarque-Bera (JB):                0.469
Skew:                           0.324   Prob(JB):                        0.791
Kurtosis:                       2.828   Cond. No.                         101.
==============================================================================
```

MLR Equation

$$y = 2.3681 - 0.0447X1 + 0.3074X2$$

Model Diagnostics





Normality Test:

A. Q-Q Plot:

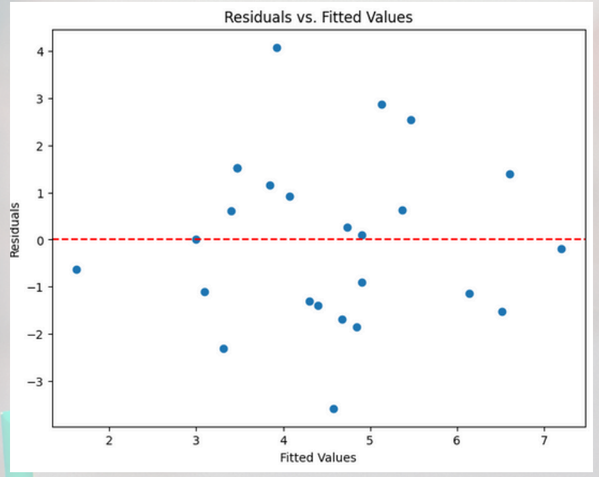
- The Q-Q (quantile-quantile) plot compares the distribution of residuals with a theoretical normal distribution.
- Most points fall fairly close to the red line, suggesting approximate normality.

B. Histogram of Residuals:

- The histogram is fairly symmetrical and roughly bell-shaped.
- The overlaid density curve matches a normal distribution fairly well, though there is some variability.

C. Shapiro-Wilk Test:

- **Null Hypothesis (H_0):** The residuals are normally distributed.
- **Alternative Hypothesis (H_1):** The residuals are not normally distributed.
- Statistic: 0.9844 & P-value: 0.9563
- Since the p-value is much greater than 0.05, we fail to reject the null hypothesis that the residuals are normally distributed.



Homoscedasticity Test:

A. Residuals vs. Fitted Values Scatter Plot:

- This plot checks for homoscedasticity (constant variance of residuals across fitted values).
- No clear pattern (like a funnel shape or curve) is evident.

B. White's Test:

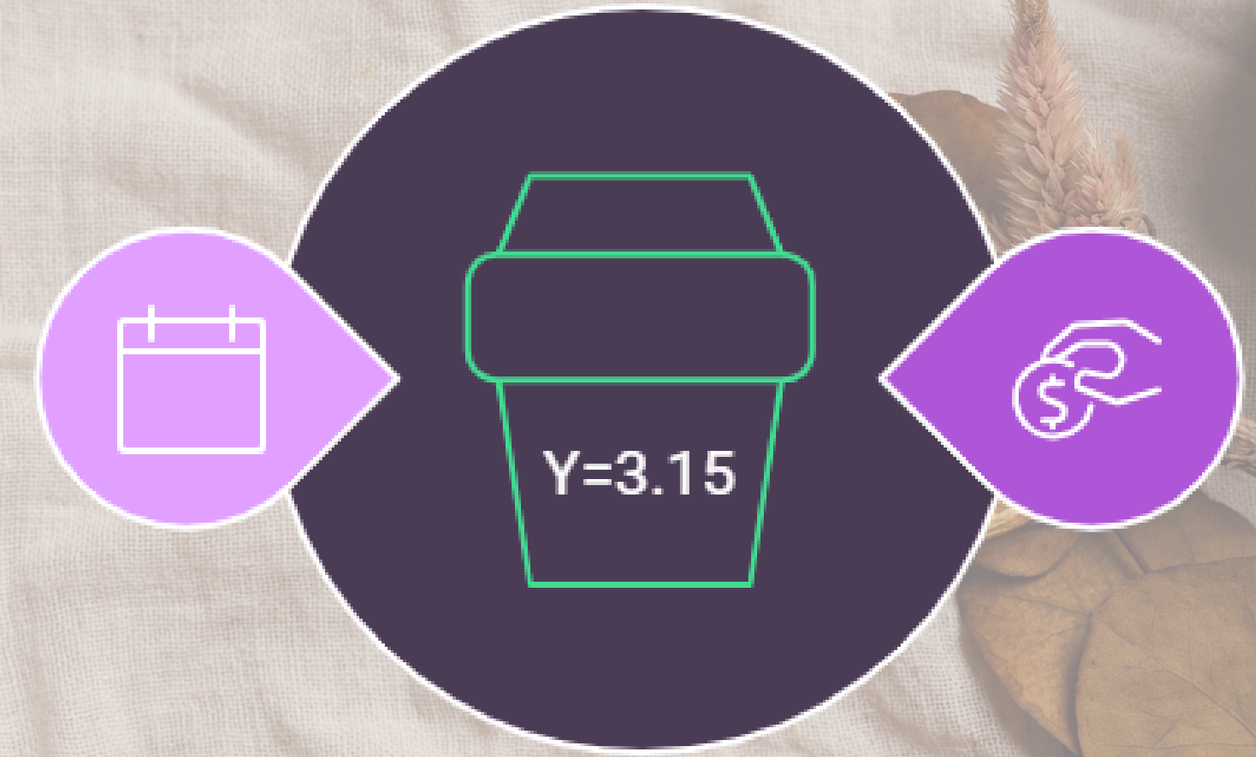
- **Null Hypothesis (H_0):** The residuals exhibit homoscedasticity .
- **Alternative Hypothesis (H_1):** The residuals do not exhibit homoscedasticity.
- Test Statistic p-value: 0.2401 & F-Test p-value: 0.2275
- Both p-values are greater than 0.05, so we fail to reject the null hypothesis of homoscedasticity.

Predicting Daily Cup Consumption

X2=4

Days per Month

The number of days considered for the calculation



X1=10

Prepaid Amount

The amount paid in advance for services

Recommendations

Incorporate Additional Behavioral & Contextual Data

- Include variables like time of day, weather, promotional offers, mobile app usage, and customer location to enhance model accuracy.

Revise Prepaid Strategy

- Since more money on prepaid cards does not equate to higher cup sales, redesign prepaid incentives to encourage higher frequency of use, not just top-ups.

Target High-Frequency Visitors

- Promote personalized deals or exclusive perks for customers who visit often—this group has the highest correlation to cup orders.

Segment and Personalize Marketing

- While age and income aren't significant across the whole dataset, further segmentation might reveal sub-group patterns. Use clustering techniques to uncover niche behaviors.

Deploy Predictive Analytics in Store Operations

- Use the demand prediction model to optimize staffing and inventory—especially for peak days based on visit frequency trends.

App-Based Behavioral Data Collection

- Encourage app check-ins or mobile orders to capture more granular data per visit, enabling more accurate demand forecasting.