LINEAR REGRESSION TRAINING PROJECT: ECOMERCE CUSTOMER

Presented By Group 6

AGENDA

I. Introduction: Dataset & Research question

II. Regression Analysis

III. Conclusion

IV. Discussion & Limitations

I. Introduction

ABOUT THE DATASET

The dataset contains information about **e-commerce customers of a company who sells clothes online**, covering various factors such as:

- Email
- Address
- Avatar
- Avg. Session Length
- Time on App
- Time on Website
- Length of Membership
- Yearly Amount Spent

This dataset is instrumental for analyzing the **influence** of certain factors to the **money spent** on purchasing cloths through online platforms, indicating the potential growth of the **company's revenue**

I. Introduction

KEY DETAILS

- **Total Entries:** The dataset contains 500 entries, each representing a unique customer informations.
- **Columns:** There are 8 columns in the dataset:
  - **Email Address:** A unique identifier for each customer
  - **Address:** The physical address of each customer
  - **Avatar:** A graphical representation or profile picture chosen by each customer
  - **Avg. Session Length**: The average duration of customer sessions on the company's e-commerce platform
  - **Time on App:** The amount of time each customer spends using the company's mobile application
  - **Time on Website:** The amount of time each customer spends on the company's website
  - **Length of Membership:** The duration for which each customer has been a member of the company's ecommerce platform
  - **Yearly Amount Spent:** The total amount of money each customer spends annually on the company's products or services

I. Introduction

RESEARCH QUESTION

Analyzing the customer data to **identify the most significant variables** and help the company decides **whether to focus their efforts on their mobile app experience or their website**

II. Regression Analysis

1. Read file data & preprocessing
2. Descriptive Statistics
3. Outliers
4. Ramsey test
5. Normality of residuals
6. Multicolinearity
7. Heteroskedasicity

1. Read file data & preprocessing

READ FILE DATA

- This step involves loading the dataset from a file into a DataFrame
- The purpose is to **bring the data into a structured format** that can be manipulated and analyzed.

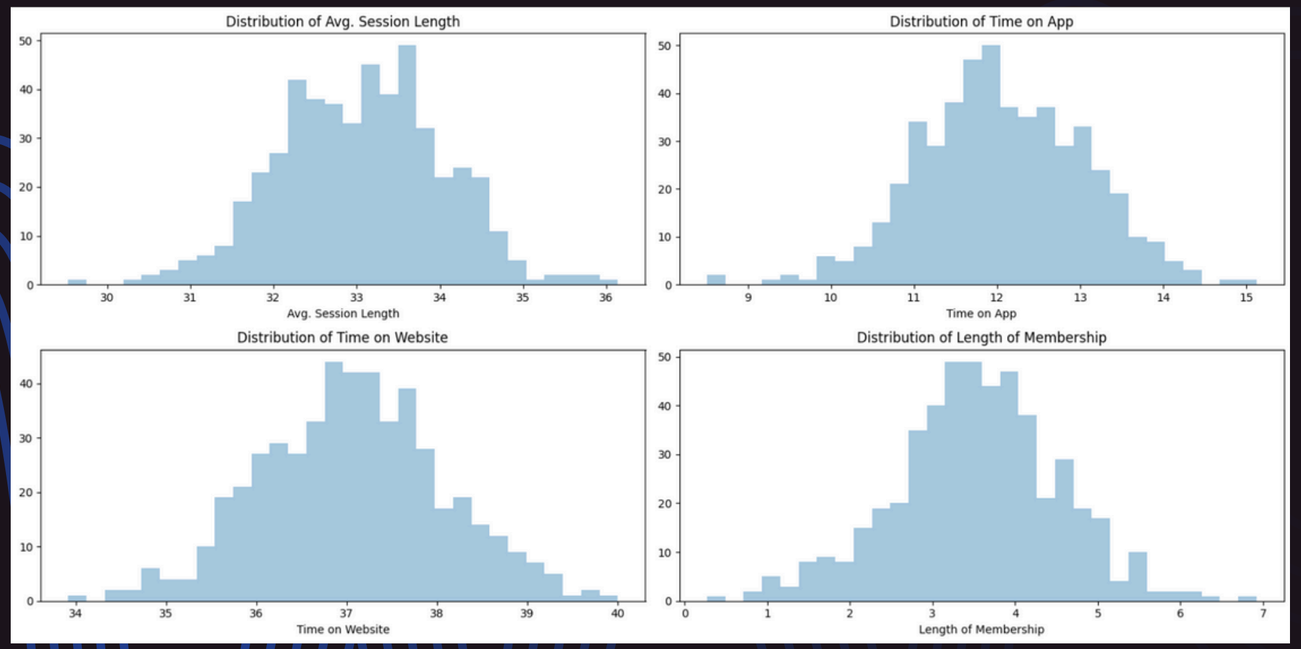
PREPROCESSING

- **df.head(), df.describe(), and df.info()** help in understanding the structure, statistical summary, and gaining insights into the dataset's characteristics
- Dropping irrelevant columns by using **df.drop(columns=[...])**
- **S**egregating the **features (independent variables)** from the **target variable (dependent variable)** in the dataset, organizing data into sets for predictors and the variable to be predicted.
- Specifying a test size of 0.1 (10% of the data) -> ensures a **sufficient amount of data** is reserved
- Using a random state to ensure **reproducibility**

1. Read file data & preprocessing

	Email	Address	Avatar	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
0	mstephenson@fernandez.com	835 Frank Tunnel\nWrightmouth, MI 82180-9605	Violet	34.497268	12.655651	39.577668	4.082621	587.951054
1	hduke@hotmail.com	4547 Archer Common\nDiazchester, CA 06566-8576	DarkGreen	31.926272	11.109461	37.268959	2.664034	392.204933
2	pallen@yahoo.com	24645 Valerie Unions Suite 582\nCobbborough, D...	Bisque	33.000915	11.330278	37.110597	4.104543	487.547505
3	riverarebecca@gmail.com	1414 David Throughway\nPort Jason, OH 22070-1220	SaddleBrown	34.305557	13.717514	36.721283	3.120179	581.852344
4	mstephens@davidson-herman.com	14023 Rodriguez Passage\nPort Jacobville, PR 3...	MediumAquaMarine	33.330673	12.795189	37.536653	4.446308	599.406092

2. Descriptive Statistics



Distribution plots

Visualize the distributions of four independent variables in the dataset

Avg. Session Length:

-Displays a roughly normal distribution.

-Represents the typical duration users spend during a session on the platform.

Time on App:

-Exhibits a positively skewed distribution.

-Indicates that most users spend shorter durations on the app, but some users spend significantly more time (long tail on the right side).

Time on Website:

-Also shows a positively skewed distribution.

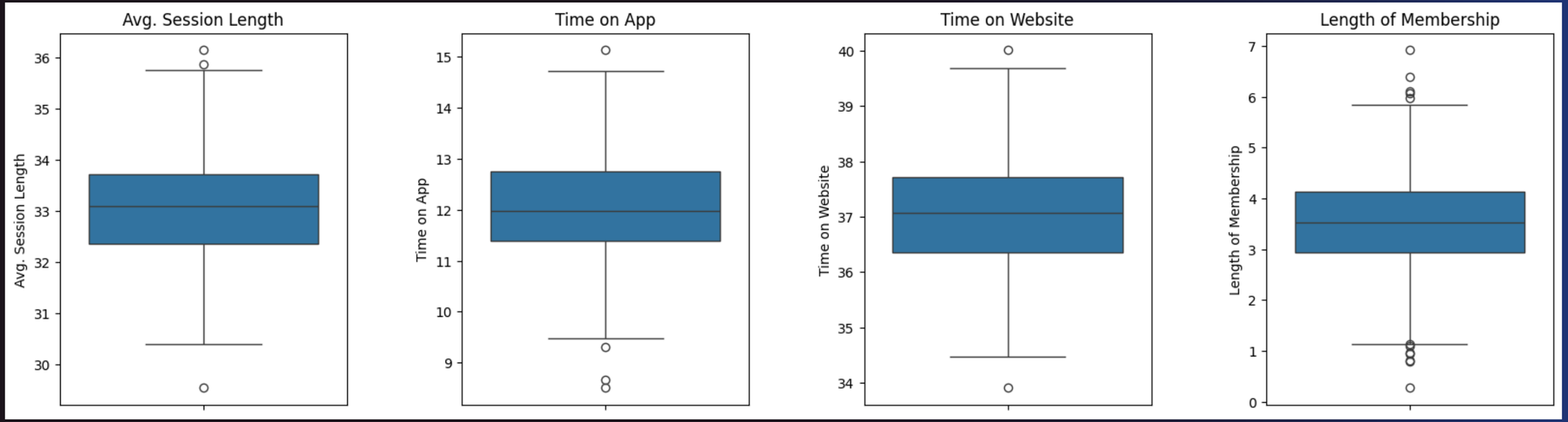
-Users spend relatively less time compared to the app, with shorter durations (shorter tail on the right side).

Length of Membership:

-Displays a roughly normal distribution with a peak around the average length of membership.

-Indicates that most users have memberships of similar durations, with fewer outliers having significantly shorter or longer memberships.

2. Descriptive Statistics



#Boxplot

2. Descriptive Statistics

#Scatterplots

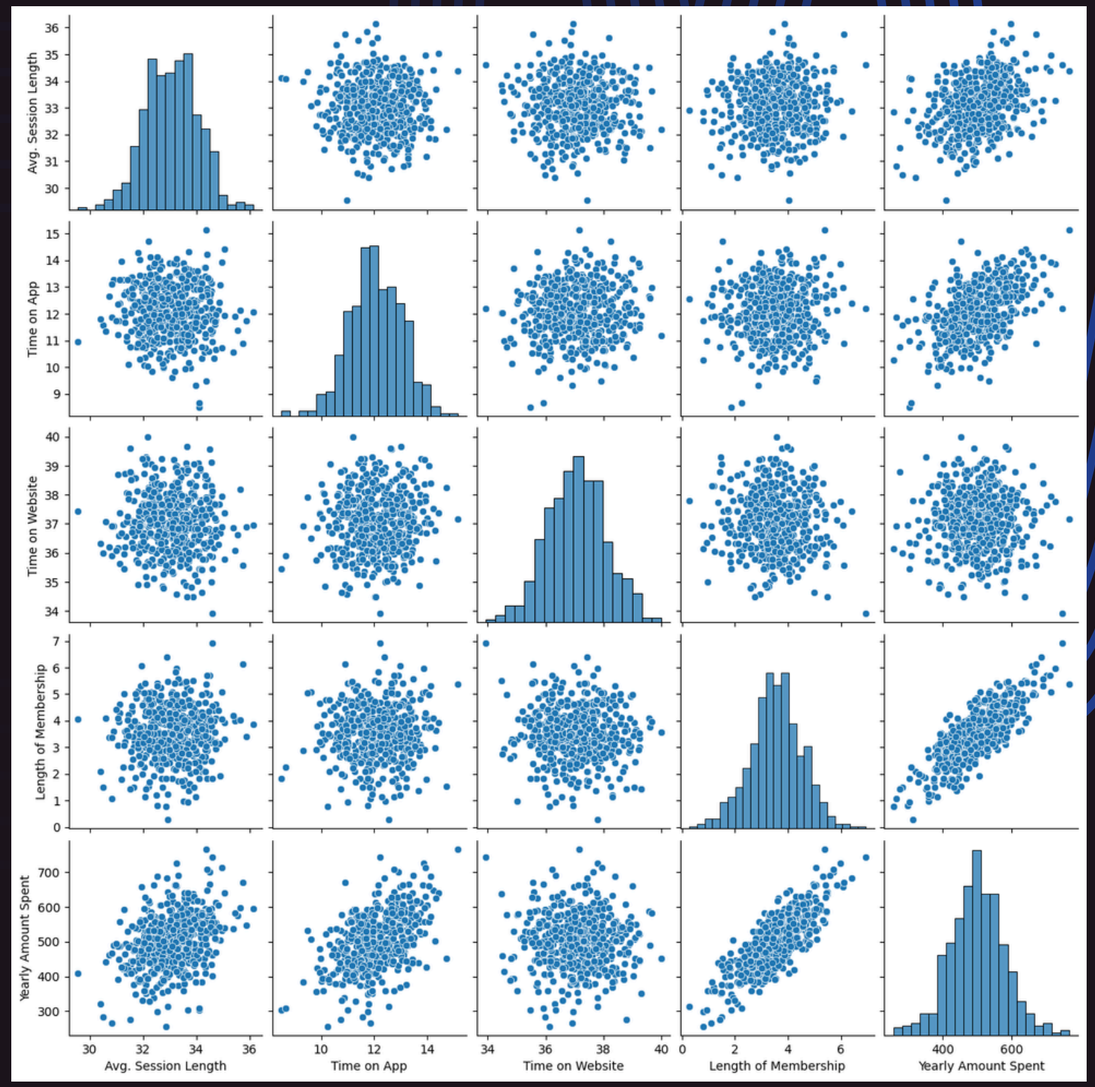
Using a **pairplot** is like a grid of **scatterplots** and **histograms**.

- **Histograms (Along the Diagonal):** Shows the distribution of each variable by itself.

- **Scatterplots (Off-diagonal):** Show how two variables relate to each other.

+ If points go up together, it's a positive relationship.
+ f points go down together, it's a negative relationship.
+ If points are all over the place, there's no clear relationship.
+ Points close to a line mean a strong relationship, while scattered points mean a weak one.
+ Outliers, which are unusual points, can be spotted.

->It's symmetric, so comparing variables is easy.



2. Descriptive Statistics

#Correlation matrix

Helps understand **relationships** between variables. Identifies **influential variables** and their **strength** of relationship.

- **Correlation Values**

+ Range: -0.048 to 1.
+ Closer to 1/-1: stronger correlation, closer to 0: weaker correlation.

- **Heatmap Visualization**

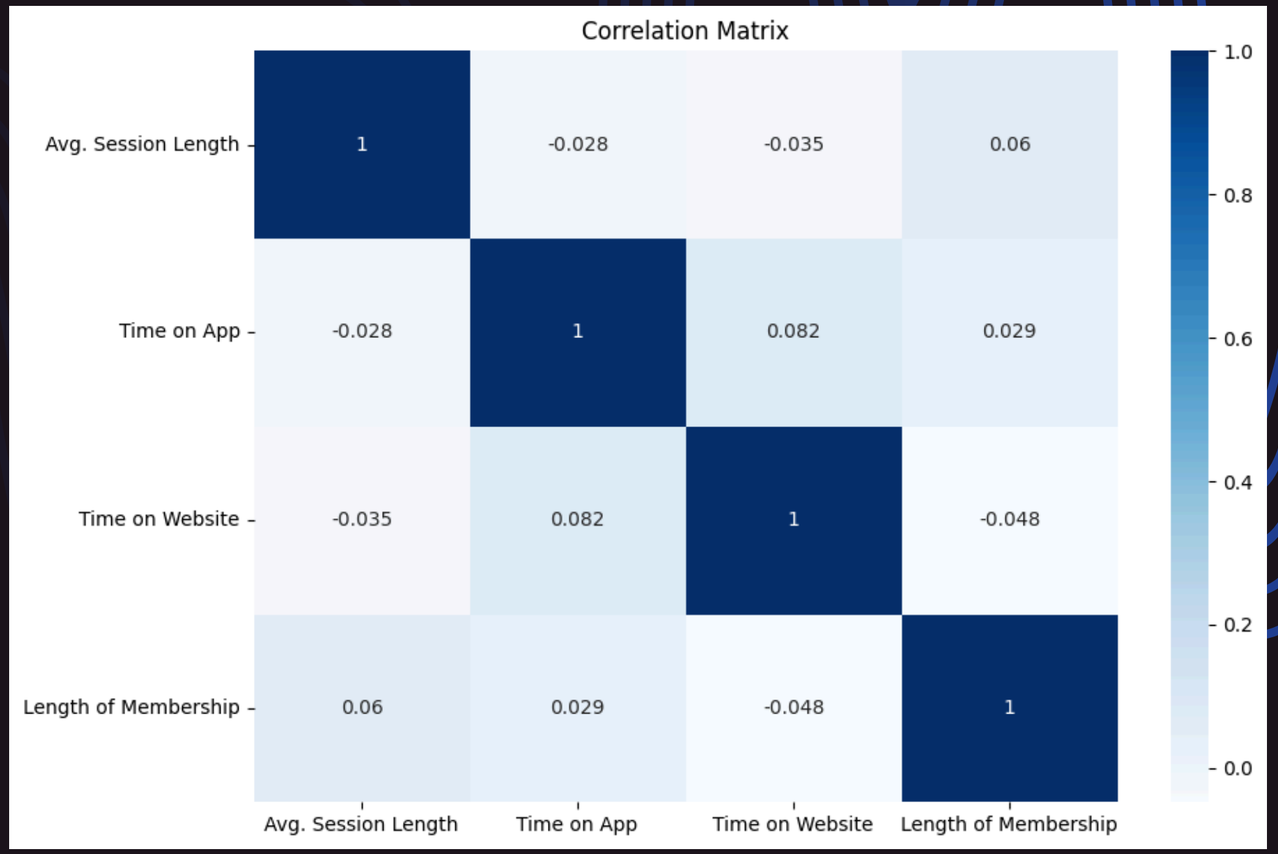
+ Darker blue: higher positive correlation, lighter: lower or no correlation.

- **Identifying Relationships**

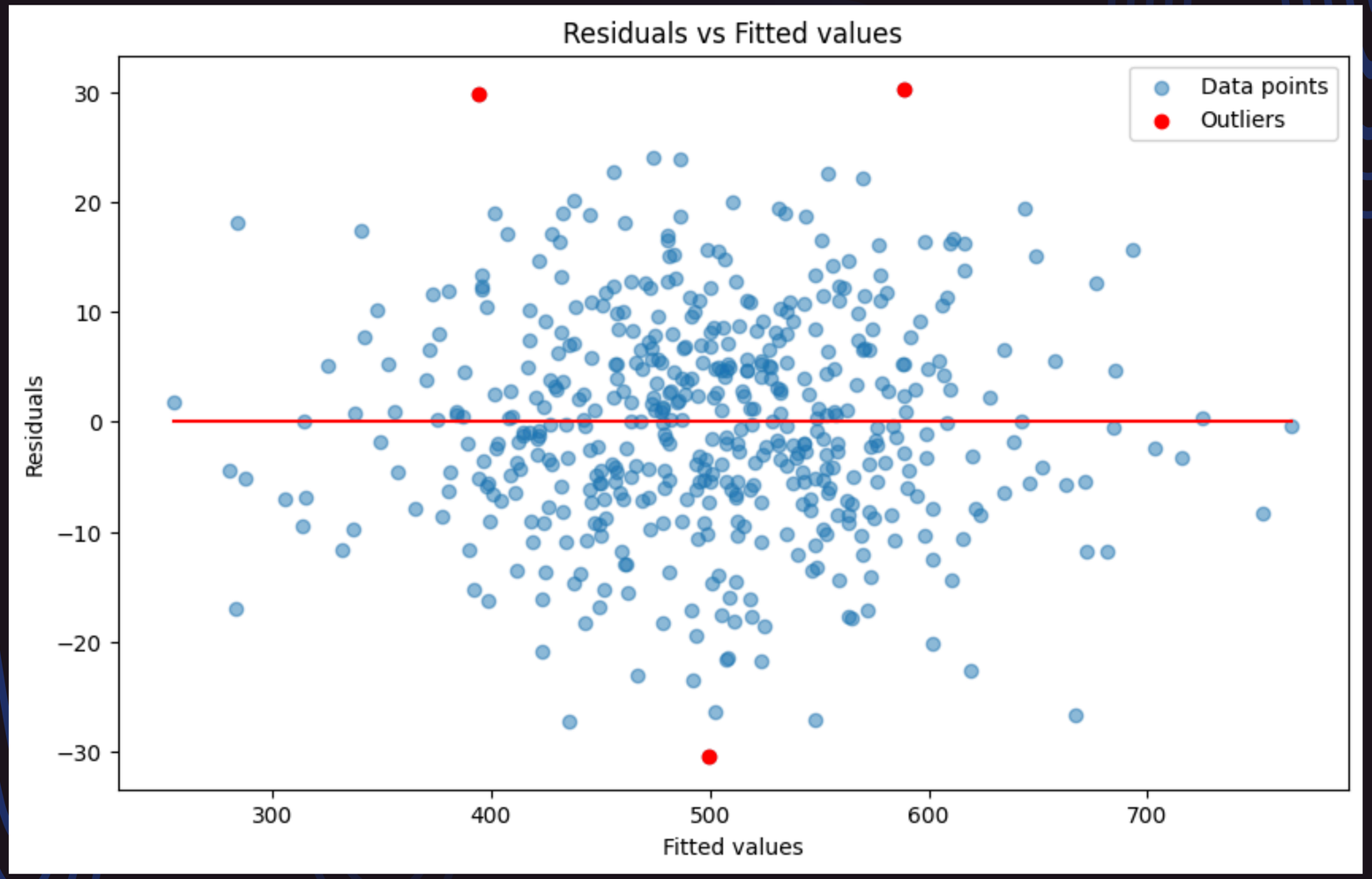
+ Dark squares: strong correlations.
+ Positive: both increase together, negative: one increases while the other decreases.

- **Identifying Redundancy**

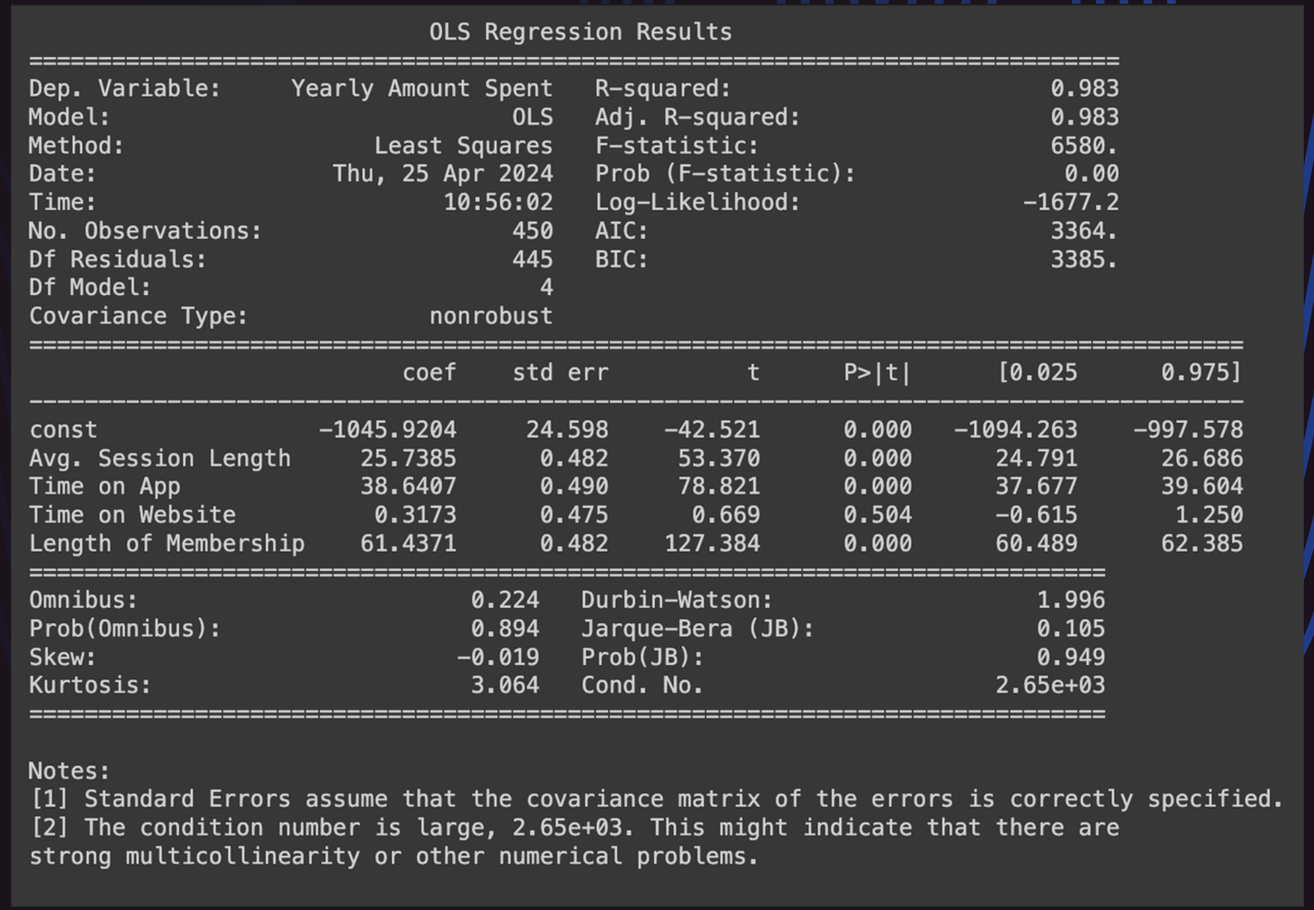
High correlations suggest redundancy. Redundant variables can be removed to simplify models.



3. Outliers


Residuals vs Fitted values
Data points
Outliers
Residuals
Fitted values
30
20
10
0
−10
−20
−30
300
400
500
600
700

- **R-squared and Adjusted R-squared:**
  - Very high R-squared and Adjusted R-squared
  -> The independent variables explain about 98.3% of the variance in the dependent variable
- **F-statistic and Prob (F-statistic):**
  - The F-statistic is extremely high (6580.0), and the Prob (F-statistic) is 0.00
  ->The model is statistically significant and at least one of the independent variables is useful for predicting the dependednt variable.
- **Coefficient Significance:**
  - Avg. Session Length, Time on App, and Length of Membership are statistically significant (p-values < 0.05)
  - Time on Website is not statistically significant (p-value = 0.504).
- **Multicollinearity:**
  - The condition number (2.65e+03) is extremely large
  -> the presence of strong multicollinearity among the independent variables.
- **Residual Analysis:**
  -The residual analysis (Omnibus, Jarque-Bera, Skewness, and Kurtosis) suggests that the residuals are normally distributed, which is a desirable property for OLS regressionn.



OLS Regression Results

Dep. Variable:	Yearly Amount Spent	R-squared:	0.983
Model:	OLS	Adj. R-squared:	0.983
Method:	Least Squares	F-statistic:	6580.
Date:	Thu, 25 Apr 2024	Prob (F-statistic):	0.00
Time:	10:56:02	Log-Likelihood:	-1677.2
No. Observations:	450	AIC:	3364.
Df Residuals:	445	BIC:	3385.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1045.9204	24.598	-42.521	0.000	-1094.263	-997.578
Avg. Session Length	25.7385	0.482	53.370	0.000	24.791	26.686
Time on App	38.6407	0.490	78.821	0.000	37.677	39.604
Time on Website	0.3173	0.475	0.669	0.504	-0.615	1.250
Length of Membership	61.4371	0.482	127.384	0.000	60.489	62.385

Omnibus:	0.224	Durbin-Watson:	1.996
Prob(Omnibus):	0.894	Jarque-Bera (JB):	0.105
Skew:	-0.019	Prob(JB):	0.949
Kurtosis:	3.064	Cond. No.	2.65e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.65e+03. This might indicate that there are strong multicollinearity or other numerical problems.

4. Ramsey test

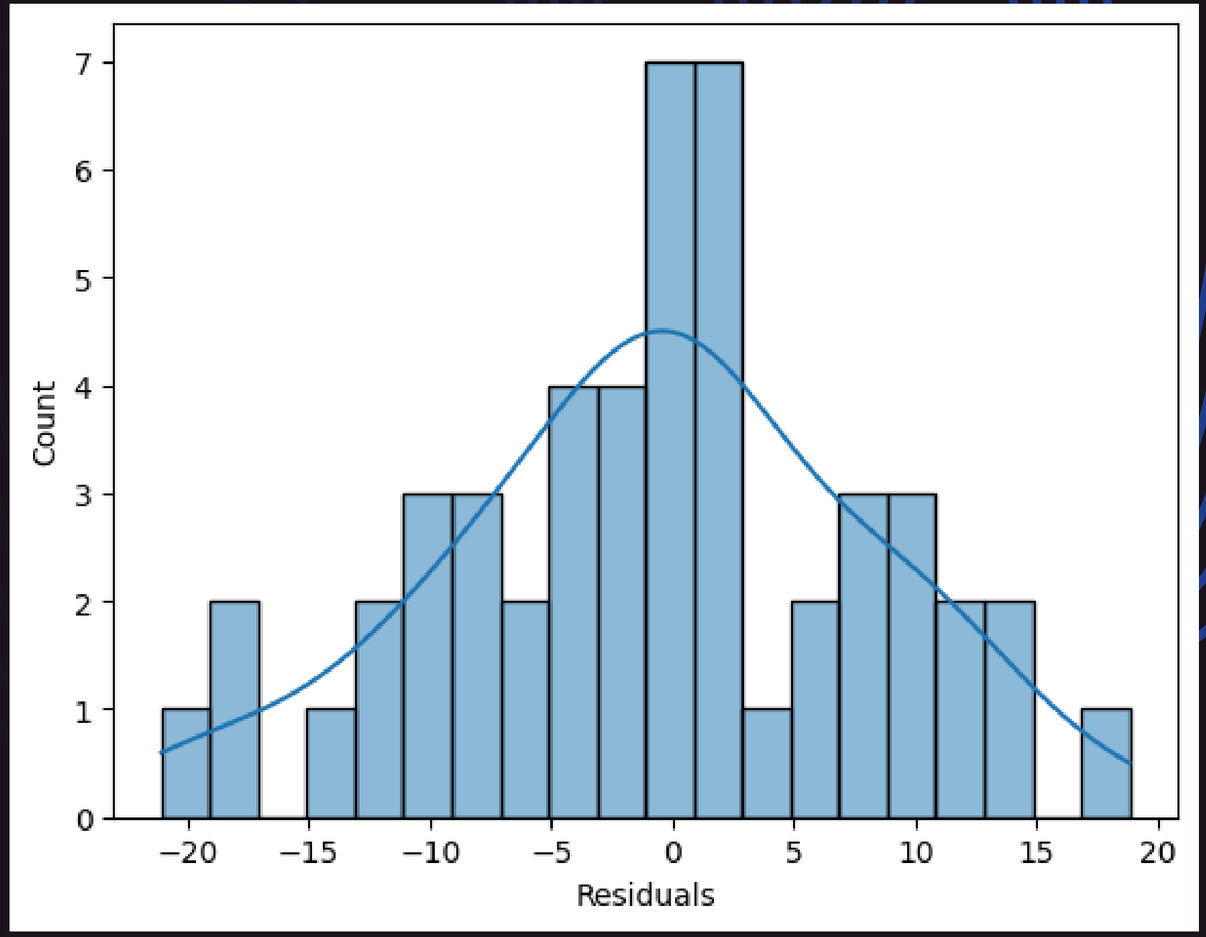
```
reset = linear_reset(model_fit, power=2, use_f = True)
print("Ramsey's RESET test F-statistic: ", reset.fvalue)
print("P-value: ", reset.pvalue)
threshold = 0.05
if reset.pvalue > threshold:
  print("No error, no omitted variable")
else:
  print("There's an error and one or many omitted variables")
```

```
Ramsey's RESET test F-statistic:  0.035594105206125 65
P-value:  0.8504423706969509
No error, no omitted variable
```

The **Ramsey RESET (Regression Specification Error Test)** is a diagnostic test used to **examine** whether the **linear regression model** suffers from **specification errors**, particularly omitted variables or functional form misspecification.

5. Normality of residuals

We use JB_pvalue to test for the normality of residuals, which is a **diagnostic test** used to **assess** whether the residuals (errors) of a regression model **follow a normal distribution**. Deviations from normality can indicate **potential issues** with the model's assumptions, such as **misspecification or outliers**.



6. Multicolinearity

	Features	VIF
3	Length of Membership	3.913564
1	Time on App	8.810480
0	Avg. Session Length	11.318800
2	Time on Website	68.414661

- **Multicollinearity**

+ Refers to **high correlation** among **independent variables** in a regression model.

- **Analysis Objective**

Aim to assess **multicollinearity** in the linear regression model and its **impact** on model validity.

- **Variance Inflation Factor (VIF)**

+ Measures the extent to which the **variance** of regression coefficients is **inflated** due to multicollinearity.
+ **High VIF values** (typically above 10) indicate moderate **multicollinearity**, implying **high correlation** among predictors in the model

6. Multicolinearity

```
from sklearn.linear_model import Lasso
model_2 = Lasso(alpha=1.0)
model_2.fit(X_train, Y_train)
```

```
Lasso
Lasso()
```

From the following table, we can see **a significant multicolinearity problem**. **Lasso Regression** introduces an **regularization term** into the loss function. Lasso Regression can perform **automatic variable selection** by shrinking the coefficients of unimportant variables to **near zero**. This helps reduce the effect of multicolinearity.

7. Heteroskedasicity

- The code **detects heteroskedasticity** by performing the **Breusch-Pagan test**, and if the **p-value** is **less than the significance level**, it concludes that homoskedasticity is **violated**
-

-> Indicating the presence of heteroskedasticity in the regression model.

-> Using fitted values directly as weights does not help reduce heteroskedasticity.

```
X_test_with_const = sm.add_constant(X_test)
test_data_prediction_lasso = model_2.predict(X_test)
residuals_lasso = Y_test - test_data_prediction_lasso
lm, lm_p_value, fvalue, f_p_value = het_breuschpagan(residuals_lasso, X_test_with_const)

print("Breusch-Pagan test LM statistic:", lm)
print("Breusch-Pagan test LM p-value:", lm_p_value)
print("Breusch-Pagan test F-statistic:", fvalue)
print("Breusch-Pagan test F p-value:", f_p_value)
```

```
Breusch-Pagan test LM statistic: 10.97120718330053
Breusch-Pagan test LM p-value: 0.02688951882202681
Breusch-Pagan test F-statistic: 3.1624365475972382
Breusch-Pagan test F p-value: 0.022537735629463616
```

```
if lm_p_value <0.05:
    print("Homoskedasticity is violated.")
else:
    print("Homoskedasticity assumption is not violated.")
```

```
Homoskedasticity is violated.
```

7. Heteroskedasicity

- We calculated the **squared residuals** and took their **absolute values** to **mitigate the influence** of outliers.
- Then, we predicted the **y values** and regressed the **absolute squared residuals** on the **predicted** y values.
- By doing this, we created a **new model** to address heteroskedasticity. The result of the Breusch-Pagan test after using the new model showed **no signs of heteroskedasticity**

```
temp = abs(residuals_lasso)
temp_squared = temp ** 2

y_pred = model_2.predict(X_test)

model_abs_resid = sm.OLS(temp_squared, sm.add_constant(y_pred)).fit()

weights_lasso_abs = model_abs_resid.fittedvalues

model_wls_lasso_abs = sm.WLS(Y_test, X_test, weights=abs(weights_lasso_abs))
result_wls_lasso_abs = model_wls_lasso_abs.fit()

residuals_wls_lasso_abs = result_wls_lasso_abs.resid
lm_abs, lm_p_value_abs, fvalue_abs, f_p_value_abs = het_breuschpagan(residuals_wls_lasso_abs, X_test_with_const)

print("Breusch-Pagan test LM statistic:", lm_abs)
print("Breusch-Pagan test LM p-value:", lm_p_value_abs)
print("Breusch-Pagan test F-statistic:", fvalue_abs)
print("Breusch-Pagan test F p-value:", f_p_value_abs)

if lm_p_value_abs < 0.05:
    print("Homoskedasticity is violated.")
else:
    print("Homoskedasticity assumption is not violated.")
```

```
Breusch-Pagan test LM statistic: 4.36921232319663
Breusch-Pagan test LM p-value: 0.35833838326714296
Breusch-Pagan test F-statistic: 1.0772033782127675
Breusch-Pagan test F p-value: 0.3790282551766583
Homoskedasticity assumption is not violated.
```

7. Heteroskedasicity

- **Model Fit**

High R-squared value (0.997) indicates strong explanatory power.
+ **Significant F-statistic (p < 0.05)** suggests overall model significance.

- **Coefficients**

+ All coefficients are **statistically significant** (p < 0.05).
+ **Time on App** has the **largest positive impact** on Yearly Amount Spent.
+ Time on Website has a negative impact on Yearly Amount Spent.
+ **Length of Membership** has the **most substantial** positive impact.

- **Assumptions and Notes**

+ **No** significant autocorrelation detected.
+ Residuals are **normally** distributed.

WLS Regression Results

Dep. Variable:	Yearly Amount Spent	R-squared (uncentered):	0.997
Model:	WLS	Adj. R-squared (uncentered):	0.997
Method:	Least Squares	F-statistic:	4375.
Date:	Thu, 25 Apr 2024	Prob (F-statistic):	1.02e-58
Time:	15:36:21	Log-Likelihood:	-231.33
No. Observations:	50	AIC:	470.7
Df Residuals:	46	BIC:	478.3
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Avg. Session Length	12.3502	2.576	4.794	0.000	7.165	17.536
Time on App	35.2842	3.154	11.187	0.000	28.935	41.633
Time on Website	-15.0318	2.349	-6.398	0.000	-19.761	-10.303
Length of Membership	61.1742	3.479	17.584	0.000	54.171	68.177

Omnibus:	1.835	Durbin-Watson:	2.062
Prob(Omnibus):	0.400	Jarque-Bera (JB):	1.003
Skew:	-0.263	Prob(JB):	0.606
Kurtosis:	3.453	Cond. No.	53.9

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

III. Conclusion

Identify the most significant variables

- **Time on App** and **Length of Membership** are the **most significant variables** influencing **Yearly Amount Spent** by customers.

Whether to focus their efforts on their mobile app experience or their website

- The coefficient for **Time on App (35.2842)** is higher than that of **Time on Website (-15.0318)**, indicating that increasing app usage correlates with greater yearly spending compared to website usage.

  -> The company should put efforts on enhancing the mobile app experience could yield higher returns for the company.
- **Length of Membership** also plays **a substantial role (coefficient: 61.1742)**, suggesting that longer memberships lead to increased spending.

  -> Strategies aimed at retaining customers and encouraging longer memberships could further boost revenue.
- **Prioritizing improvements to the mobile app interface** and **fostering customer loyalty through membership** retention initiatives are key areas for maximizing profitability.

IV. Discussion & Limitations

The dataset contains a **limited** number of independent variables causing difficult to predict the **denpendent variable**, and there's a constant implication of **multicollinearity** among the independent variables

The data sample is **not large**, suggests that their **range** might be **limited** and the **R^2 score** is also **big**
-> Leading to **overfitting**.
-> May suffer from some **errors when predict** with other data.

THANK YOU FOR LISTENING!

Presented By Group 6