TIP INCENTIVE
PHASE -2
DEVELOPING MODEL, FINDINGS

April 16th, 2024

Presenting by:

Sindhuja Pagadala

Practical data science, cs#667

Email: sp71631n@pace.Edu

Seidenberg School Of CSIS

Pace University , NYC

Presented to :

Prof Stephanie Langeland

AGENDA

- EXECUTIVE SUMMARY
- PROJECT PLAN RECAP
- DATA
- EXPLORATORY DATA ANALYSIS(EDA)
- MODELLING METHOD
- FINDINGS
- RECOMMENDATION & NEXT STEPS
- APPENDIX

EXECUTIVE SUMMARY

Business Problem: The business problem is to optimize the revenue and customer satisfaction at Chef's Kitchen restaurant. By analyzing the data on tips and revenue, the restaurant aims to understand which customer segments are more likely to give higher tips and spend more on their dining experience. This information can be used to tailor the restaurant's marketing strategies, customer service training, and menu offerings to maximize revenue and enhance customer satisfaction. Additionally, the analysis can help identify any gaps in service quality or areas where improvements can be made to enhance the overall dining experience and increase customer loyalty.

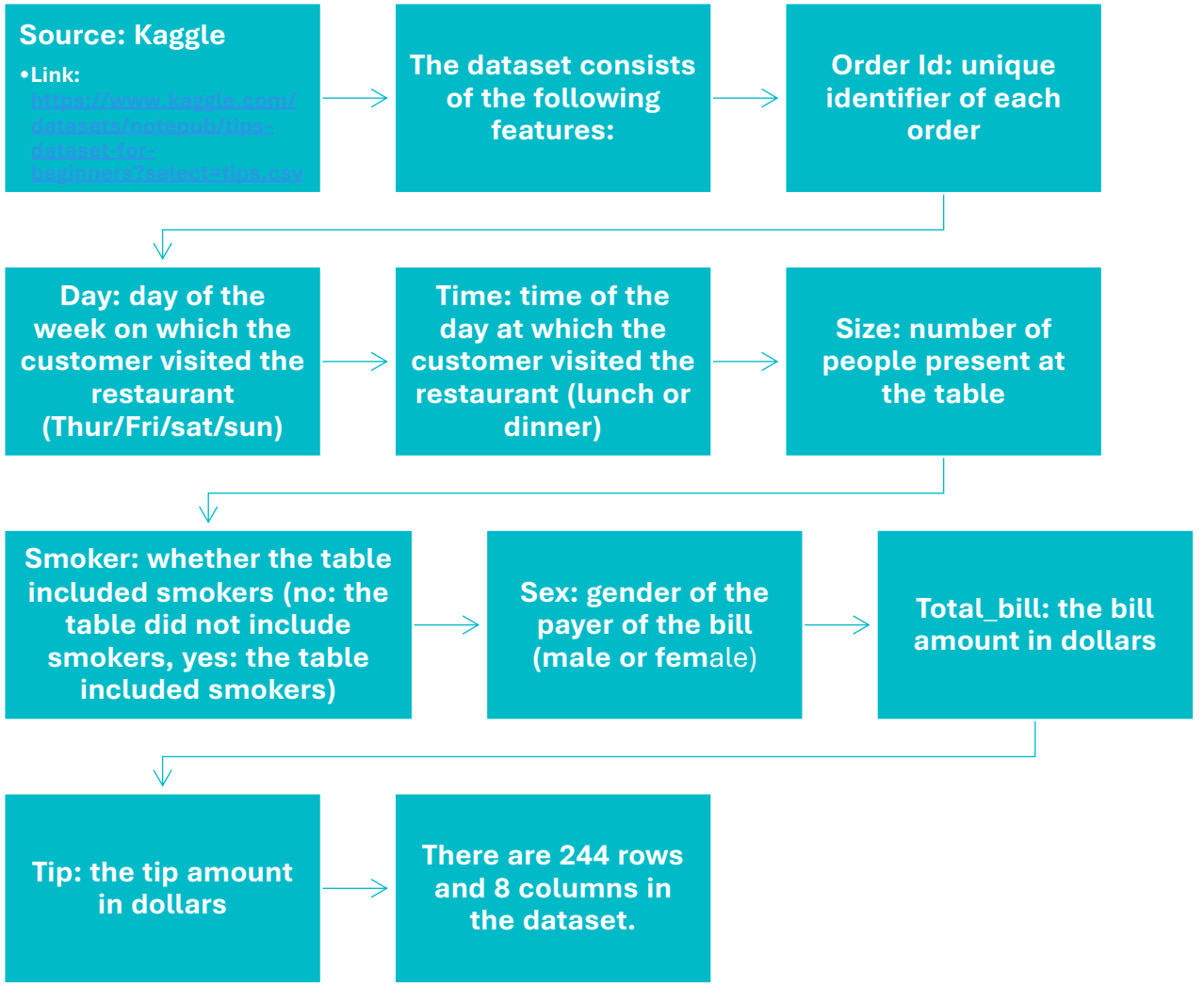
Objective: The objective is to analyze the data on revenue and tips received from customers across different demographics at Chef's Kitchen restaurant in order to identify patterns and trends. The goal is to gain insights into customer behaviour and preferences and use this information to improve the restaurant's operations and customer service.

PLAN OF THE PROJECT

DELIVERABLE	DETAILS	DUE DATES	STATUS OF COMPLETION
DATA & EDA	Identify the patterns and trends in revenue and tips	03-19-2024	Completed
Methods and Findings, Recommendations	Find out the methods and draw recommendations	04-02-2024	Completed
Final Presentation	Final Completion Deck	04-16-2024	Completed

DATA

DATA


Source: Kaggle
•Link: https://www.kaggle.com/datasets/noteasyhttps-dataset-for-beginners?select=tips.csv
The dataset consists of the following features:
Order Id: unique identifier of each order
Day: day of the week on which the customer visited the restaurant (Thur/Fri/sat/sun)
Time: time of the day at which the customer visited the restaurant (lunch or dinner)
Size: number of people present at the table
Smoker: whether the table included smokers (no: the table did not include smokers, yes: the table included smokers)
Sex: gender of the payer of the bill (male or female)
Total_bill: the bill amount in dollars
Tip: the tip amount in dollars
There are 244 rows and 8 columns in the dataset.

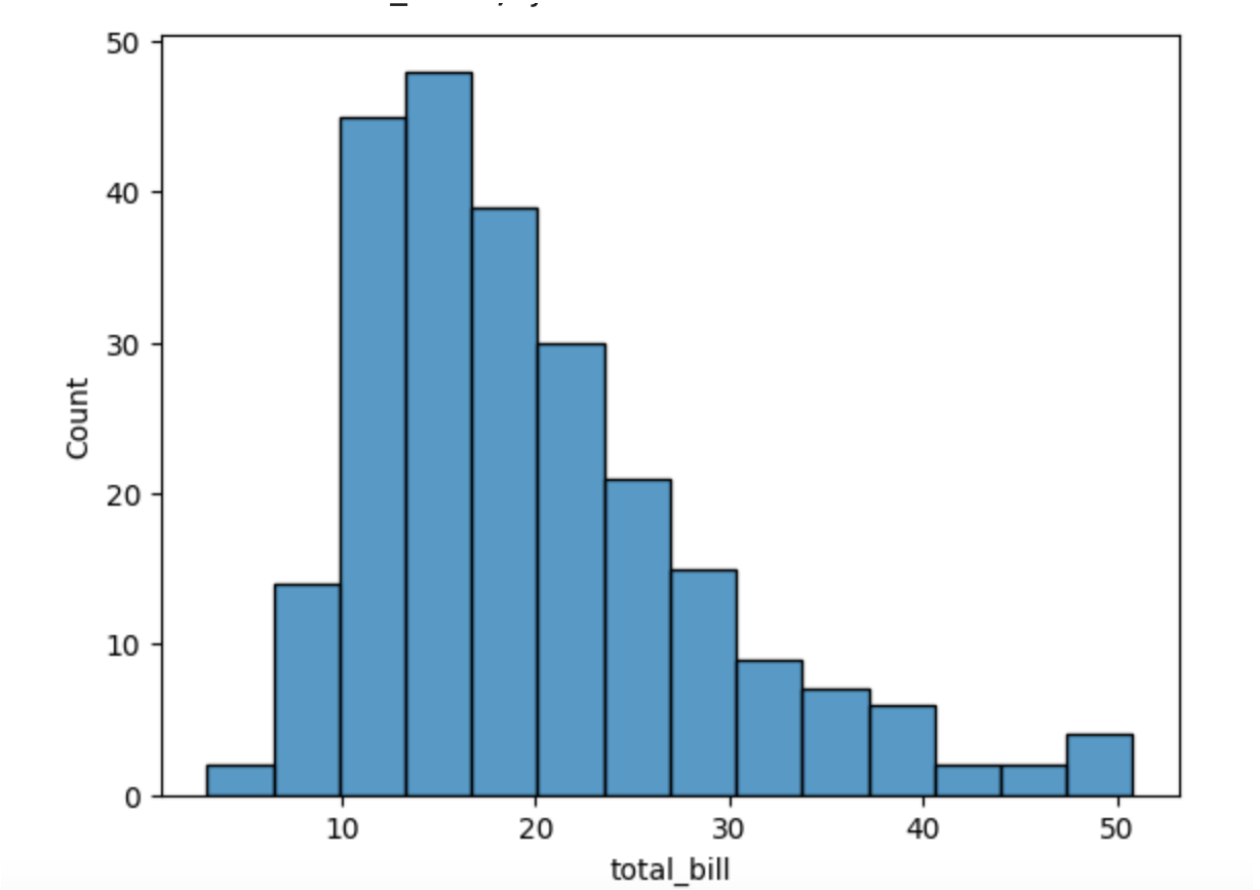
EXPLORATORY DATA ANALYSIS

EXPLORATORY DATA ANALYSIS

Almost half of the customers pay less than 18 dollars for overall bill.

50% of the billing amounts lie between 12 to 25 dollars.

The analysis of the dataset reveals several insights. The bill amount ranges from around 3 to 50 dollars, with an average of 20 dollars. Half of the bill amounts are less than 18 dollars.

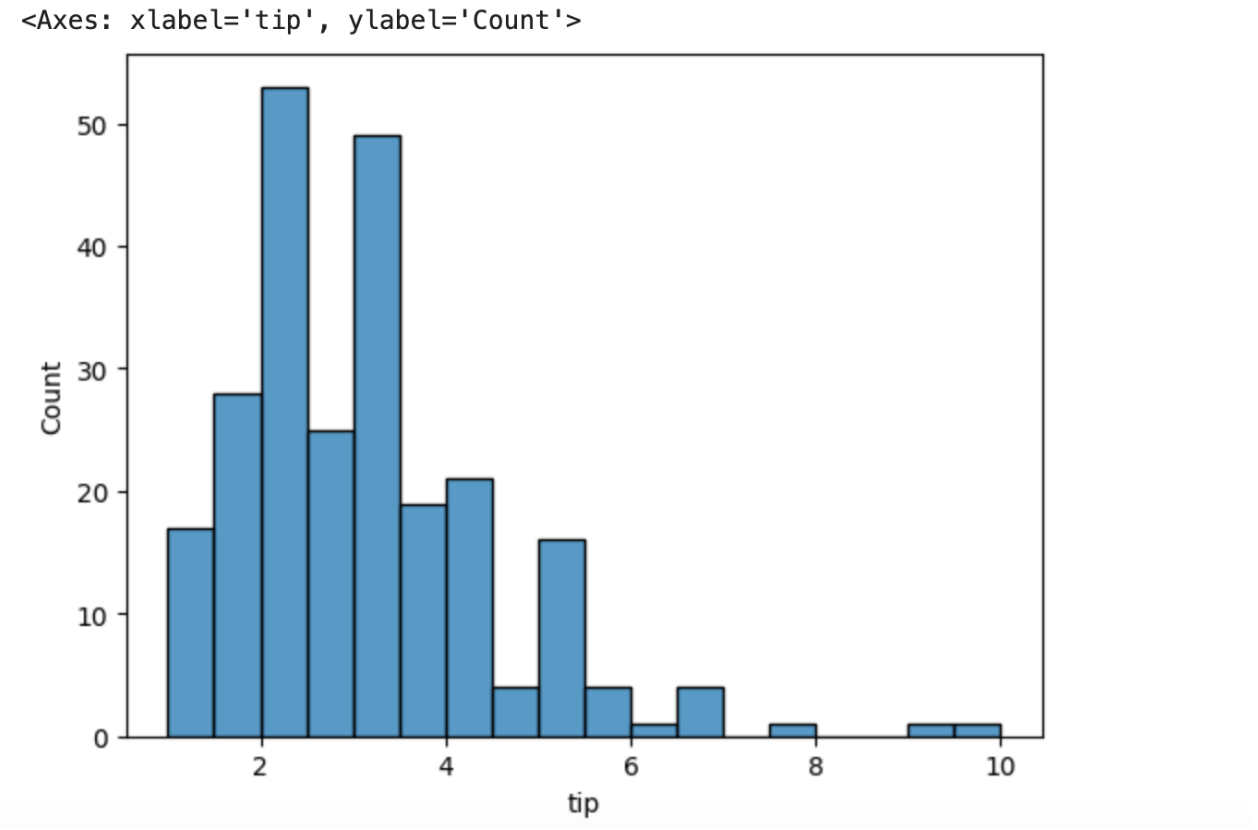


UNDERSTANDING THE TIP

50% of the people pay less than 3 dollars as tip.

Half of the tip amounts lie between 2 to 4 dollars.

The tip amount ranges from around 1 to 10 dollars, with a mean and median of around 3 dollars. The group size varies from 1 to 6 people.

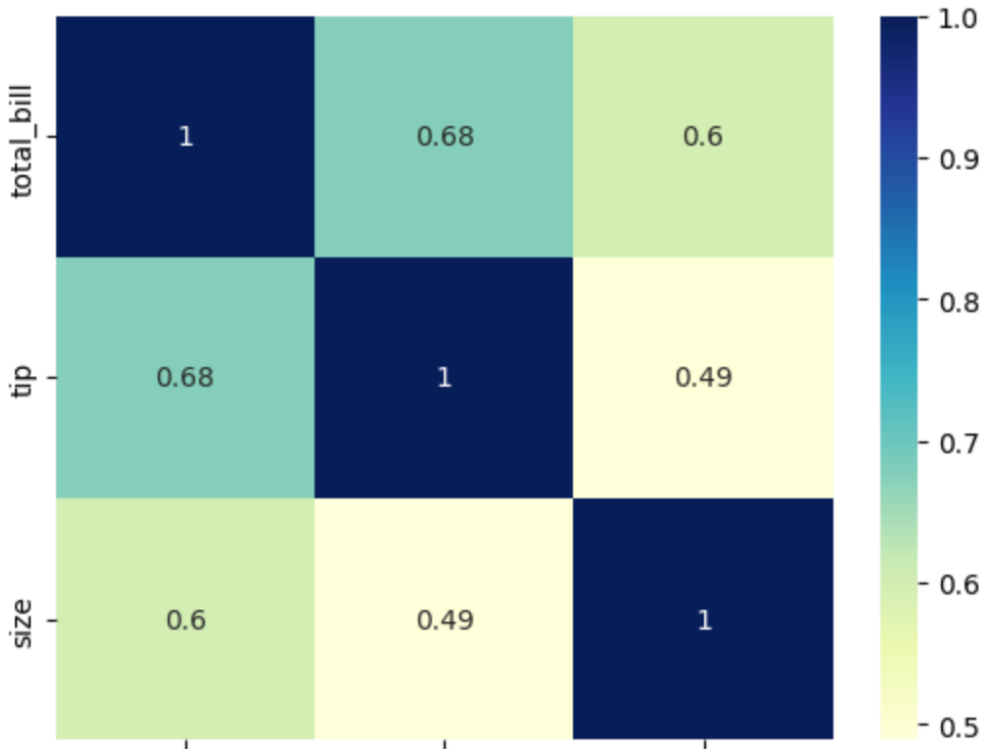


Till now we have seen the distribution of the columns individually. Let us now try to see the relation between various columns so that we can find some meaningful patterns and trends from them. Let us use the heatmap to see the correlation between the numerical columns

we can see that total_bill and tip columns are highly correlated with each other.

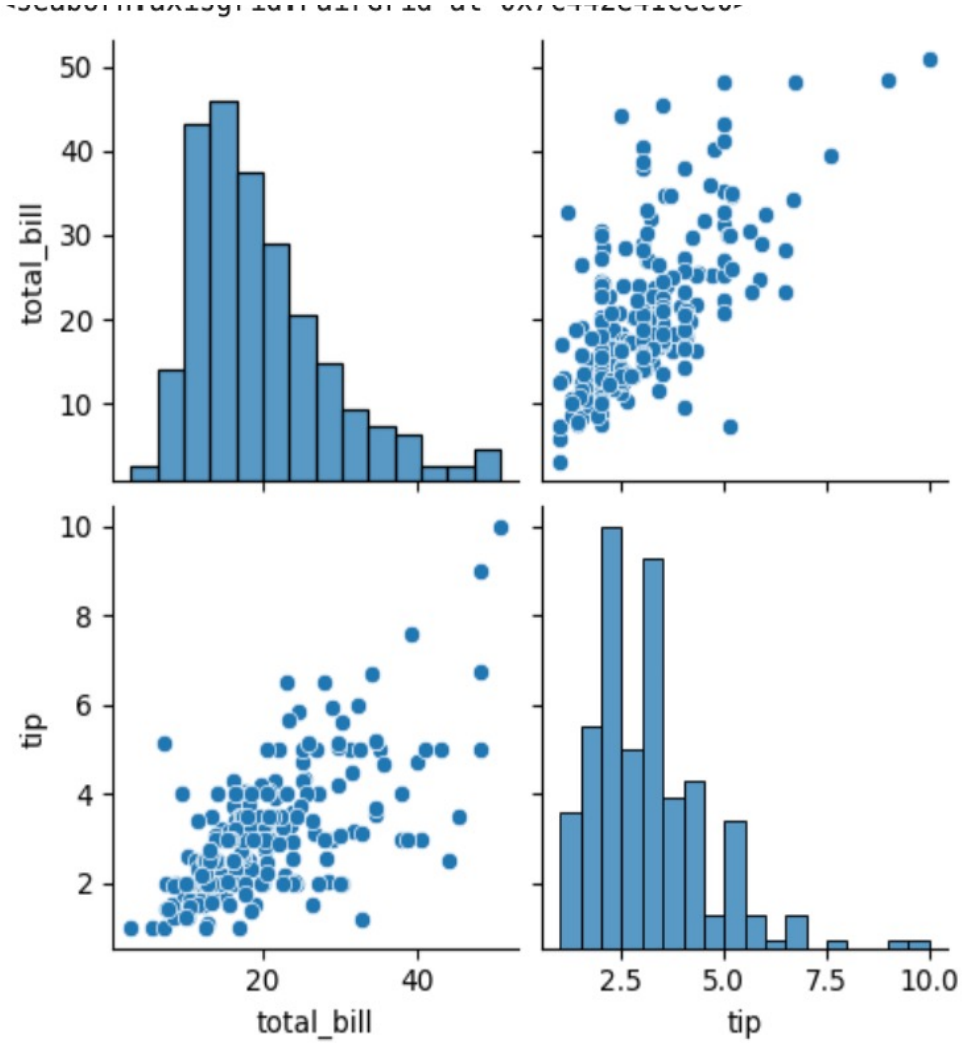
The total_bill and size columns are also correlated with each other which makes sense since higher the group size, higher will be the bill amount and vice-versa.

The tip and size columns are moderately related with eachother.



UNDERSTANDING THE RELATIONSHIP BETWEEN TOTAL_BILL AND TIP USING SUITABLE PLOTS

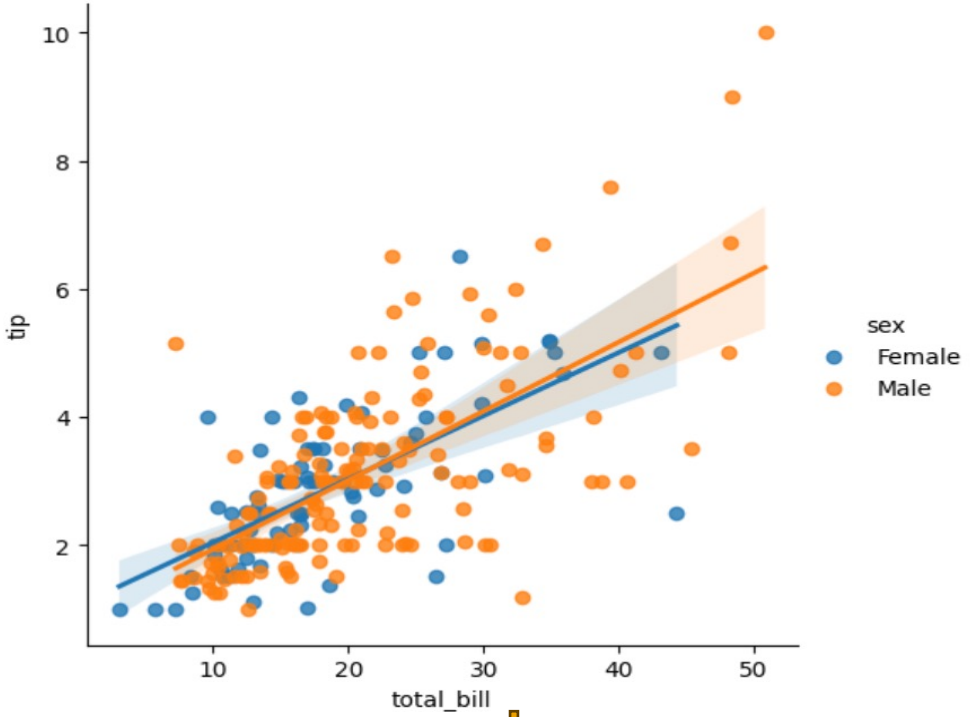
We have seen that total_bill and tip columns are highly correlated with each other. Let us verify this relationship with the help of various plots



Based on the analysis of the pair plots, it can be observed that there is a clear linear relationship between the variables "total_bill" and "tip". This means that as the total bill amount increases, the tip amount also tends to increase, and vice versa.

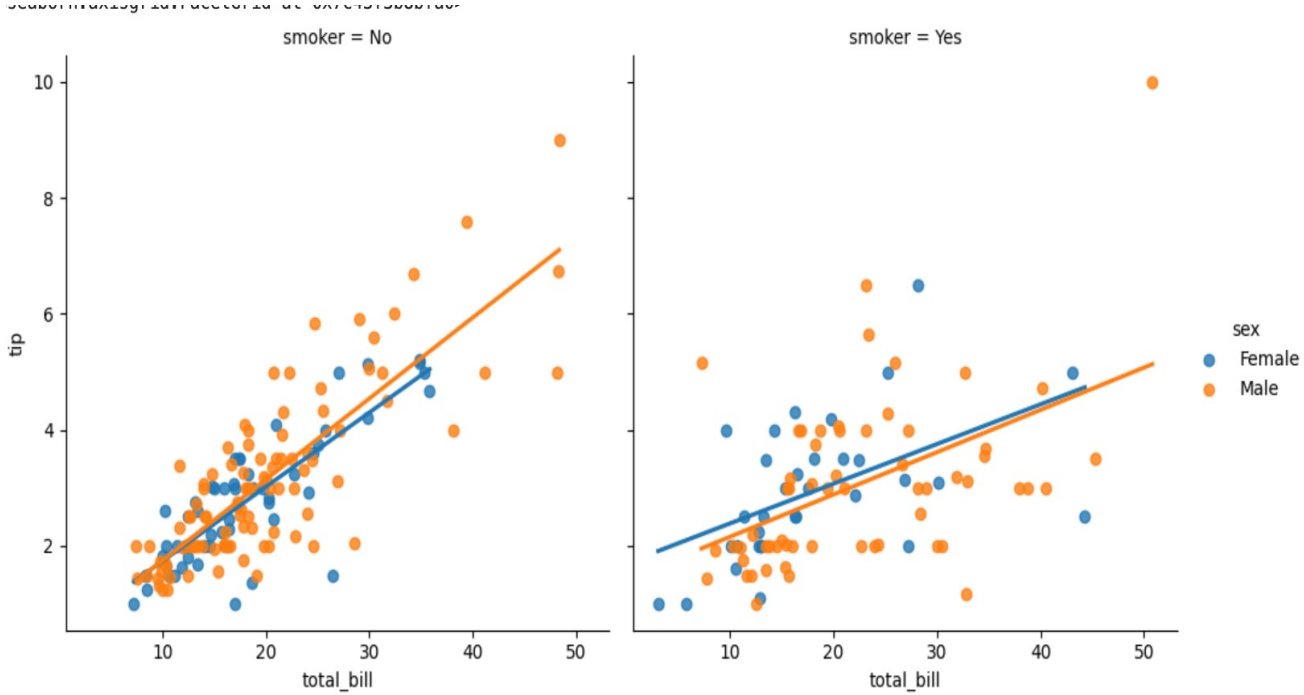
The scatter plot in the graphs shows a positive correlation between the total bill and tip. The points on the plot are mostly clustered around a diagonal line that slopes upwards from left to right, indicating the positive relationship between the two variables. This suggests that customers tend to tip more when they have a higher bill amount.

Is the relationship between total_bill and tip valid with respect to the gender of the bill payer, smoking status, days and time at which the customer has visited the restaurant and also the group size?



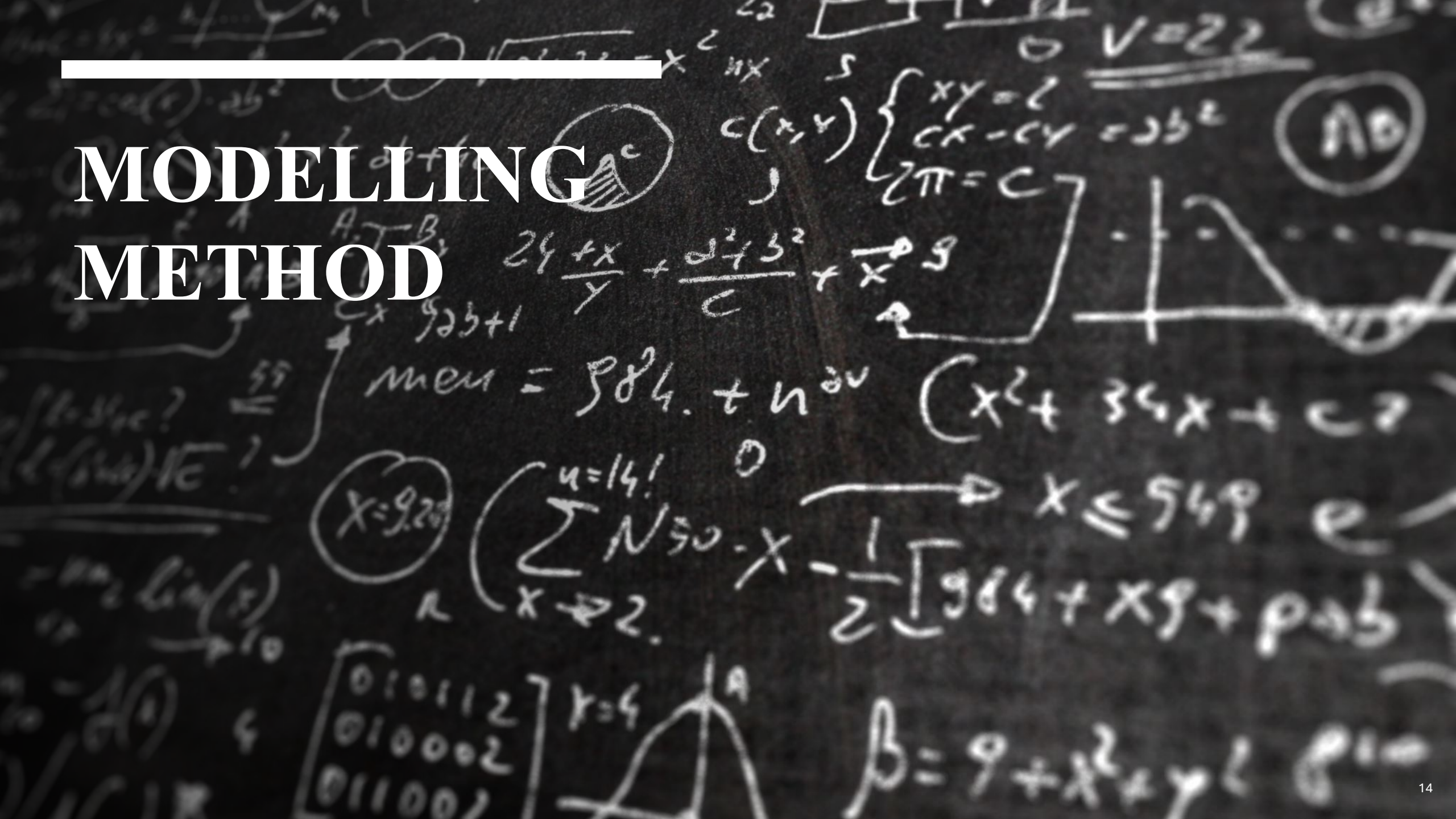
Based on the data in the provided scatter plot, there is a linear relationship between the total_bill and tip amounts for both genders, with some outliers where high billing amounts correspond to high tip amounts, particularly among male bill payers.

The analysis indicates that the highest tip was given by a table that included smokers, and non-smokers exhibit a stronger linear relationship between total_bill and tip compared to smokers.



NOTE: Refer to Appendix for more Data Visualizations

MODELLING METHOD



MODEL SELECTION

OUTCOME VARIABLE:

Neural network model is choosen to analyse and predict tip amount based on bill amount and size of party

FEATURES:

Day, Time, Size of the group , Gender of the bill Payer, Total Bill, Tip

MODEL TYPE: NEURAL NETWORK

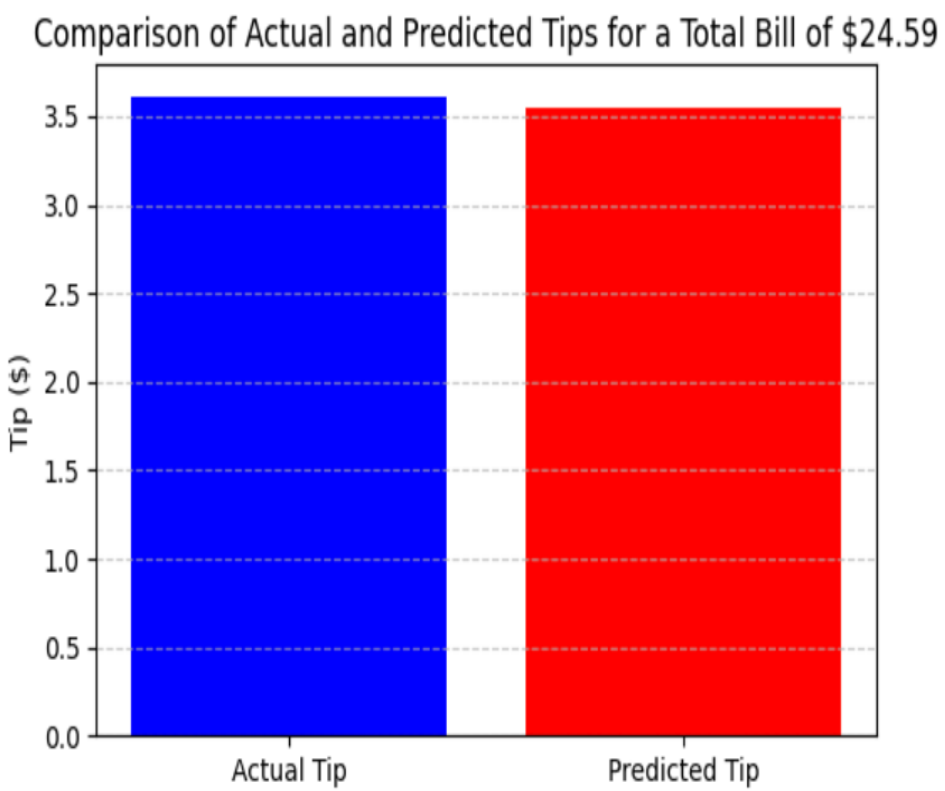
- Neural networks are well-suited for analyzing complex relationships and patterns in data.
- They can capture non-linear relationships between input and output variables, making them suitable for predicting tip amounts based on factors like total_bill and group size.
- Neural networks have the capacity to handle large amounts of data, enabling them to learn intricate patterns and make more accurate predictions.
- Compared to simpler models, neural networks are capable of capturing and leveraging the subtleties of the data, resulting in more sophisticated and precise predictions.

NOTE: Refer to appendix slide for Data Transformation for Modelling

MODEL OUTPUT AND ANALYSIS

```
2/2 [==============================] - 0s 9ms/step - loss: 0.6819
Test Loss: 0.6818665862083435
```

```
1/1 [==============================] - 0s 47ms/step
Predicted Tip: $3.54
```



- We are using neural network to predict how much tip we might get at a restaurant. It learns from past data about the total bill and the size of your group to make accurate predictions.
- The program tests the neural network by checking how close its predictions are to the actual tip amounts. The test loss, which is a measure of prediction accuracy, is impressively low at just 0.6819. This means the neural network is often very accurate in its predictions.

NOTE: Refer to the appendix slide technical reasoning on neural network

FINDINGS

FINDINGS FROM THE MODEL

- The neural network model is utilized to predict restaurant tips based on the total bill and party size.
- The test loss of the model, a measure of prediction accuracy, is impressively low at 0.6819, indicating that the predictions are often very accurate.
- The model demonstrates its prowess by accurately forecasting tip amounts with a small deviation from the expected values. For example, it predicts a $3.54 tip on a $24.59 bill, which is only 3% different from the actual tipping behavior of $3.61.
- This level of precision, within a 3% margin, highlights the model's reliability in estimating tips and its potential to aid in understanding and anticipating customer tipping practices.

RECOMMENDATIONS

&

NEXT STEPS

RECOMMENDATIONS & NEXT STEPS

1.Target weekends and dinner time: Since the number of orders is higher on weekends, particularly during dinner time, businesses can focus on offering promotions or incentives during these periods to encourage higher tips.

2.Focus on non-smokers: Non-smokers show a more prominent linear relationship between the total bill and tip amount compared to smokers. Businesses can target non-smokers for promotions, as they are more likely to have a higher tip amount based on their total bill.

3.Encourage higher spending on overall bill: Businesses could consider offering incentives to customers who spend more on their overall bill. As there is a positive linear relationship between the total bill amount and tip amount, encouraging higher spending can potentially lead to higher tips.

4.Consider day and time-specific strategies: The median billing amount is higher on Saturdays and Sundays compared to other days, and the median tip amount is higher on Fridays, Saturdays, and Sundays compared to Thursdays. Businesses can plan day and time-specific promotions or offerings to capitalize on these higher spending and tipping trends.

NEXT STEPS

1.RETRAIN THE MODEL:

Combine the existing dataset with the newly collected data.

Use the combined dataset to retrain the chosen advanced model.

2.EVALUATE MODEL PERFORMANCE:

Employ appropriate evaluation metrics, such as mean squared error (MSE) or R-squared, to assess the performance of the refined model.

Utilize validation techniques like cross-validation or train-test splits to evaluate the model's generalization ability.

Compare the performance of the refined model with the previous version to measure improvement.

3.ASSESS IMPACT OF EXPANDED DATA:

Analyze how the inclusion of additional data has affected the model's ability to address the business problem or answer open questions.

Consider the overall predictive accuracy, robustness, and generalization of the refined model.

Assess whether the expanded data landscape has provided new insights, addressed previous limitations, or improved the model's performance.

4.ITERATE AND FINE-TUNE:

Fine-tune the model by adjusting hyperparameters, exploring different architectures, or employing regularization techniques.

Repeat the retraining and evaluation process until the model achieves the desired performance and effectively addresses the business problem.

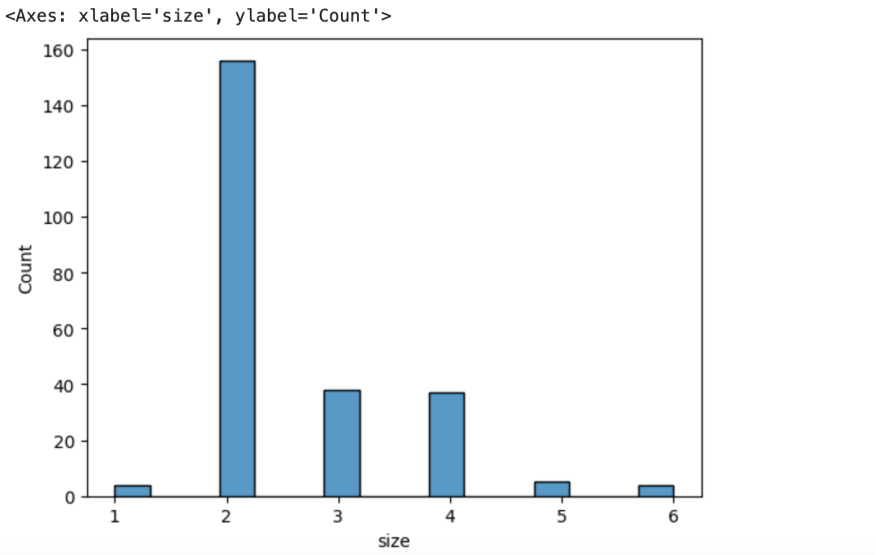
APPENDIX

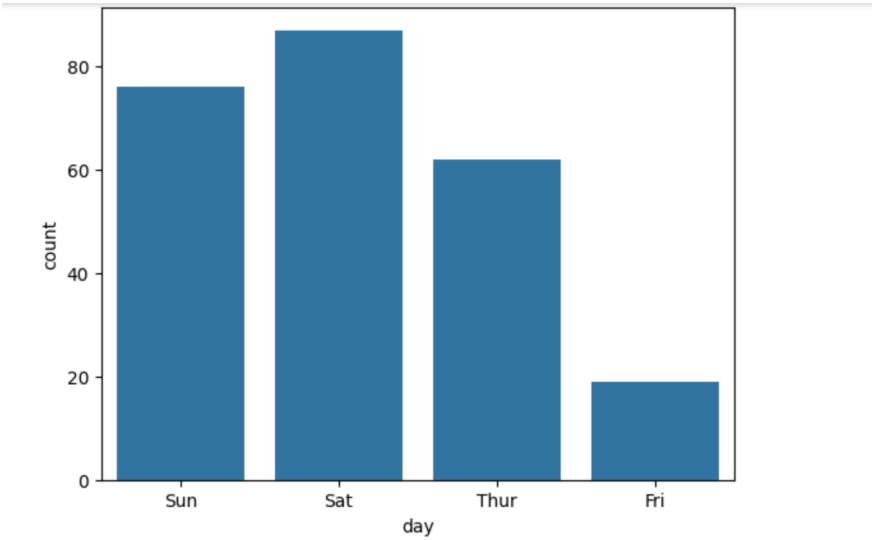
DATA VISUALIZATIONS

UNDERSTANDING THE SIZE

Majority of the customers come in groups of 2 people (~160).

50% of the groups have 2 to 3 people in them.





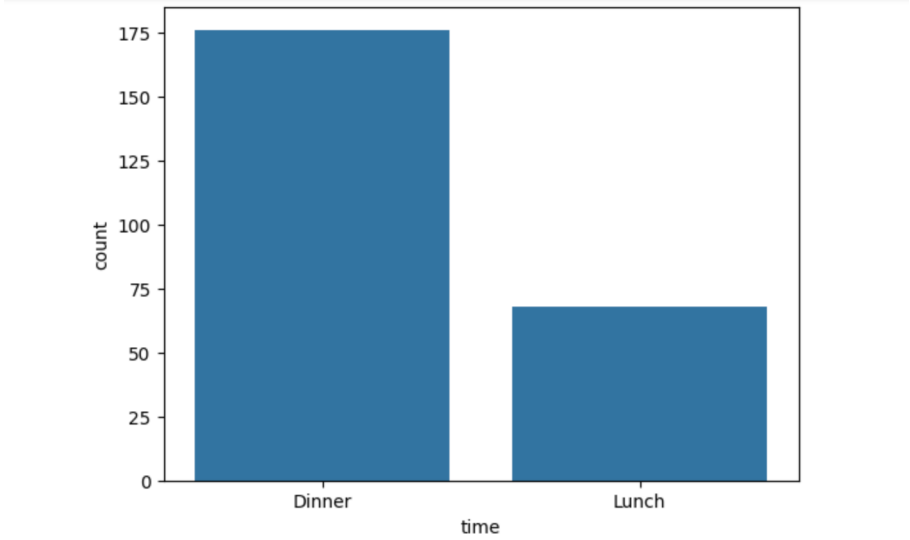
UNDERSTANDING THE DISTRIBUTION OF DAY

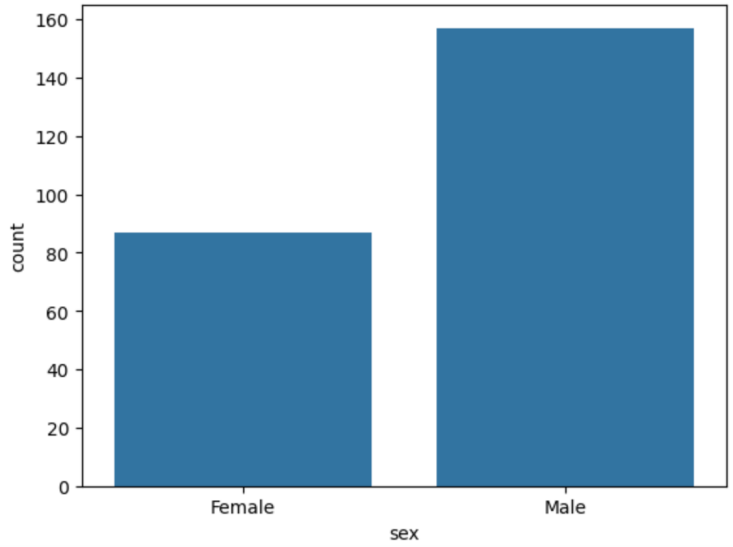
The number of orders is more during the weekends than the weekdays.
The highest number of orders is received on saturdays

DATA VISUALIZATIONS

UNDERSTANDING THE DISTRIBUTION OF TIME

The restaurant receives more orders during dinner time as compared to lunch.



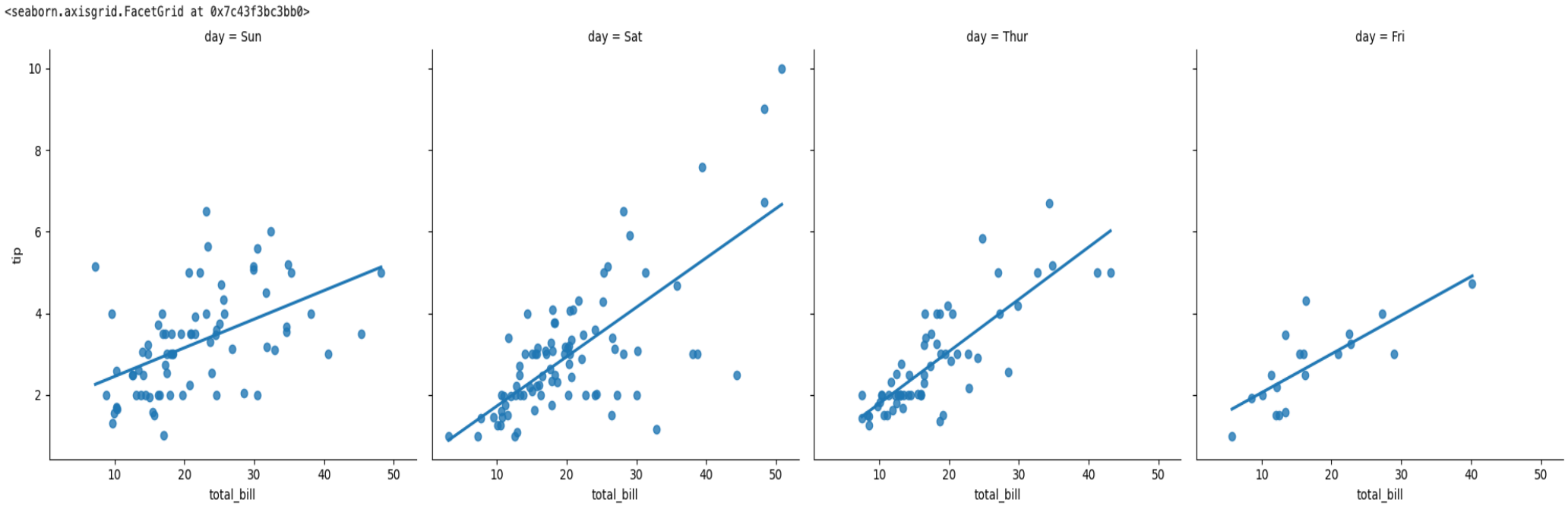


UNDERSTANDING THE DISTRIBUTION OF SEX

The number of male bill payers is around 155 while the number of female bill payers is around 85.

DATA VISUALIZATIONS

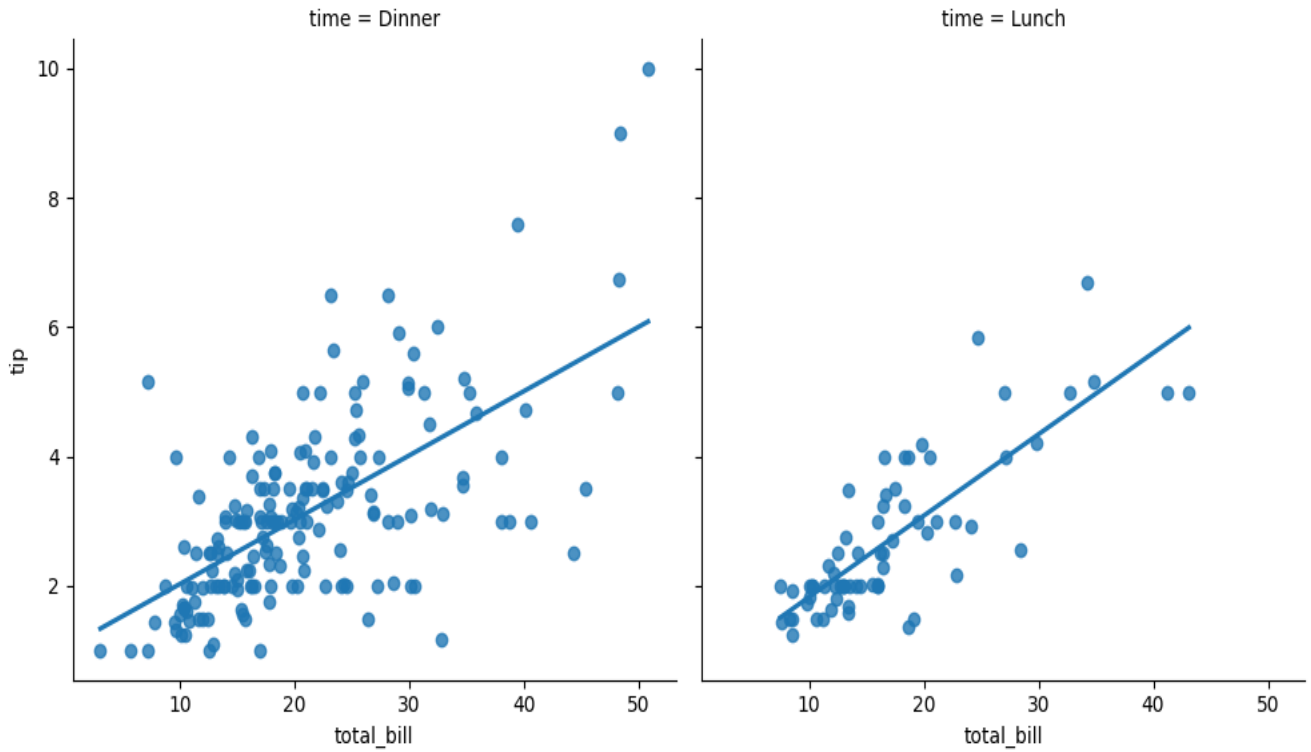
With respect to day



The analysis reveals a generally linear relationship between total_bill and tip across different days of the week, indicating that as the total_bill amount increases, the tip amount also tends to increase.

DATA VISUALIZATIONS

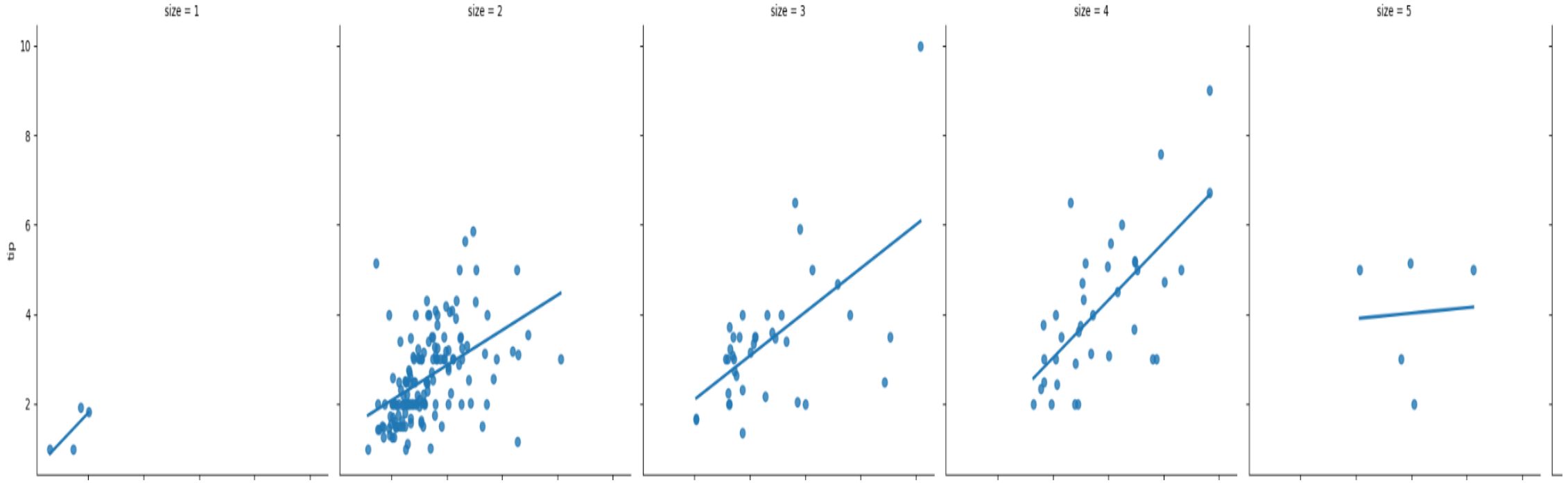
With respect to time



The analysis of the relationship between total_bill and tip with respect to the time reveals that there is a generally consistent and linear trend.

DATA VISUALIZATIONS

With respect to size



The analysis reveals that as the size of the group increases, the relationship between total_bill and tip becomes more constant, indicating that the tip amount does not vary significantly with larger group sizes. This suggests that customers tend to tip a consistent percentage or amount regardless of the size of their dining party..

DATA TRANSFORMATION

Transforming categorical variables (sex, smoker, day, and time) in the "tips" DataFrame into numerical values for analysis.

The "sex" column is mapped to 0 for "Female" and 1 for "Male".

The "smoker" column is mapped to 0 for "No" and 1 for "Yes".

The "day" column is mapped to numerical values: "Thur" to 1, "Fri" to 2, "Sat" to 3, and "Sun" to 4.

The "time" column is mapped to 1 for "Lunch" and 0 for "Dinner".

MODELLING METHOD: TECHNICAL REASONING ON NEURAL NETWORK

- The model has two layers: the first layer has 64 units and uses the ReLU activation function, while the second layer has just one unit (output layer) without an activation function.
- ReLU (Rectified Linear Unit) is an activation function commonly used in neural networks that returns zero for negative input values and the input value itself for positive input values.
- The input shape of the model depends on the number of features in the input data, which in this case is determined by the columns in the xtrain array. In this case, there are two features: 'total_bill' and 'size'.
- The ReLU activation function helps the model understand complex patterns in the data by introducing non-linearity.
- The model uses the mean squared error (MSE) loss function, which calculates the average squared difference between predicted and actual values. The goal is to minimize this loss during training.
- The Adam optimizer is used to update the model's weights during training. It dynamically adjusts the learning rate and is known for its good performance in various scenarios.
- During training, the model is trained on the xtrain and ytrain data for 500 epochs, with 20% of the training data reserved for validation.
- After training, the model's performance is evaluated on the unseen test data (xtest and ytest) using the model.evaluate() function. The test loss is calculated, representing the model's prediction error on the test data. A lower test loss indicates better performance, as it means the model's predictions are closer to the actual values.

ADDITIONAL INFORMATION

Git repository:

https://github.com/Sindhu-2498/TIPS-INCENTIVE


thank
you