Predicting a Table’s Tips

SAINT GAU

The Problem

- Day-to-day responsibilities: Servers unable to properly prioritize tasks during a rush without knowing how much a table will reward them
- Applicants: Unable to determine their potential salary
- Hiring managers: Unable to properly advertise job openings

Who might care?

Current waitstaff

Job applicants

Hiring managers

What factors might affect a table's tips?

Quantitative Data (Kaggle Dataset)

- Table size
- Smoker present in group?
- Gender of person paying
- Total bill amount
- Day of week and time of day

Qualitative Data (may not be necessary)

- Customer service quality
- Customer satisfaction

Data Information

- Data acquired over a few months
- One server, one restaurant
- Published in 1995
- Number of records: 244
- Number of fields: 7

Engineering Tip Percentage Amounts Feature From Tips Data

Divide tip amount from total bill amount

total_bill	tip	sex	smoker	day	time	size	perc
16.99	1.01	Female	No	Sun	Dinner	2	5.944673

Data Exploration

Correlations

Pairplots

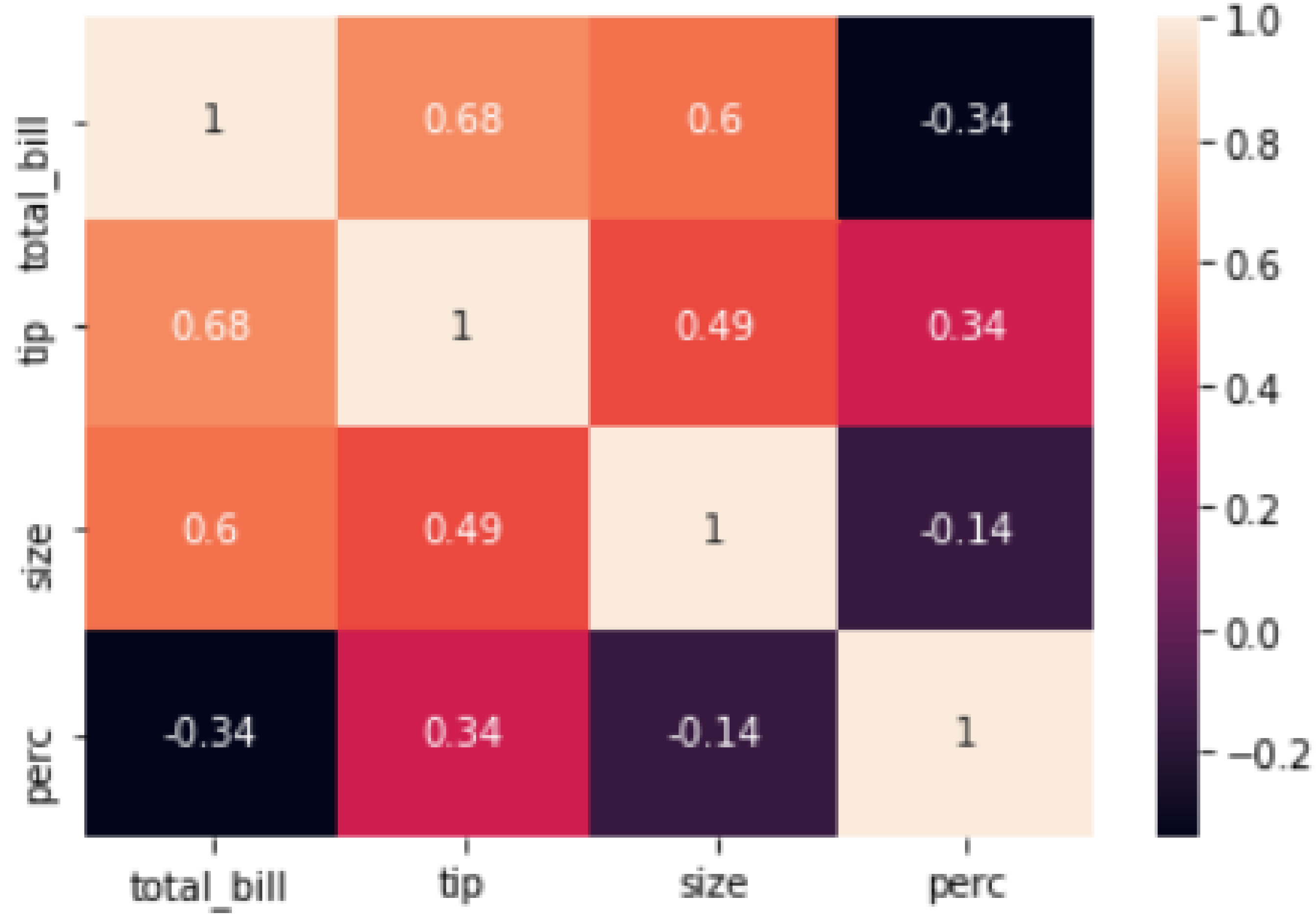
Sex

Smoker

Day

Time

Size



Correlations

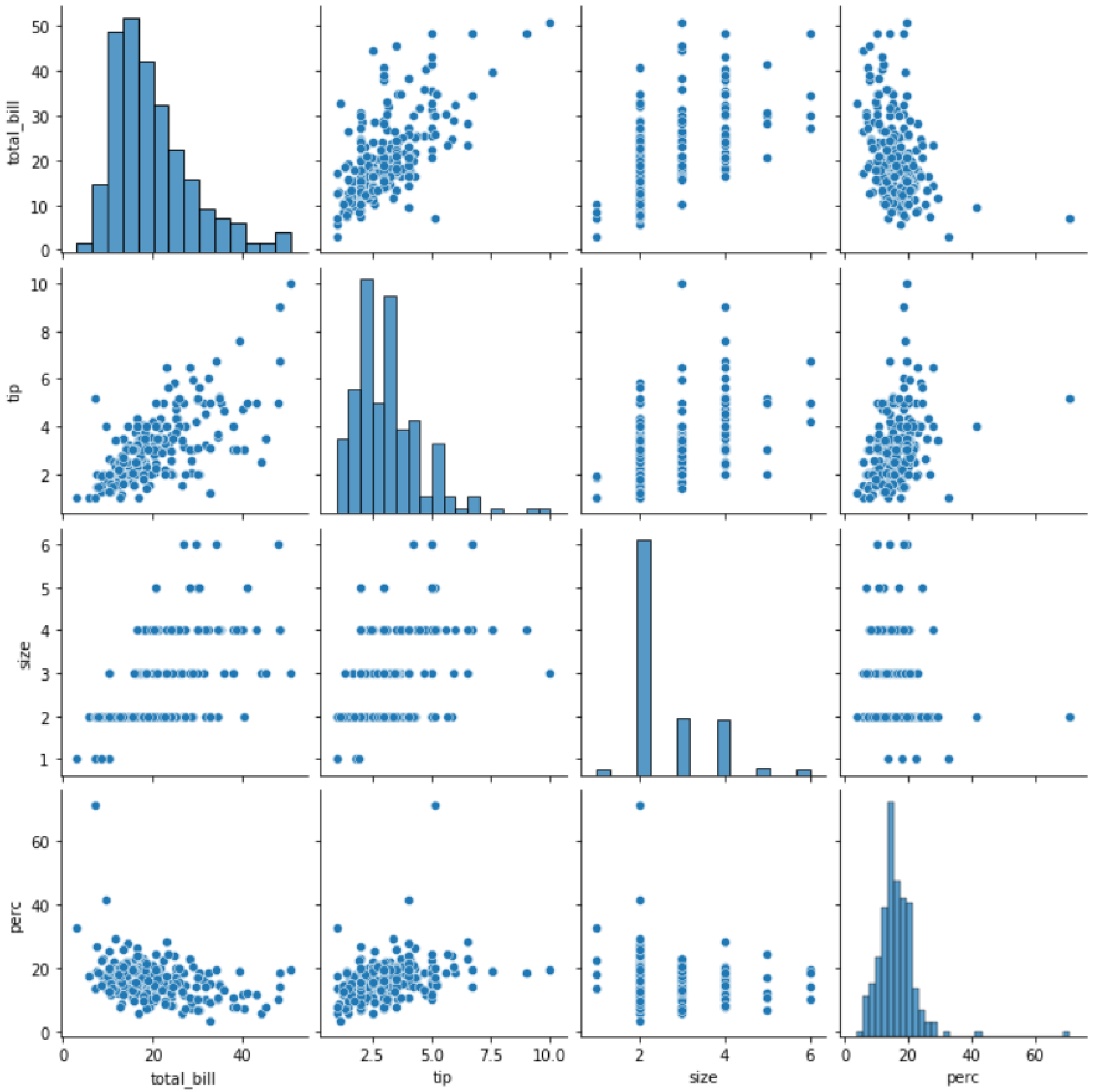
Moderate correlation between tips and total bill

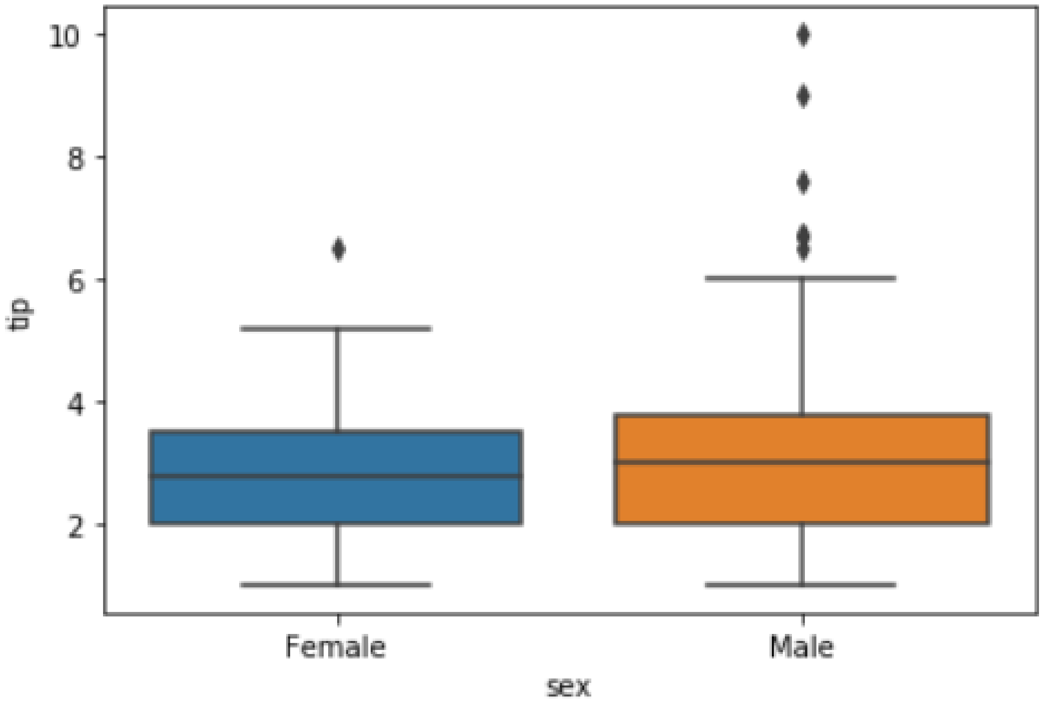
Weak correlation between group size and tip percentage

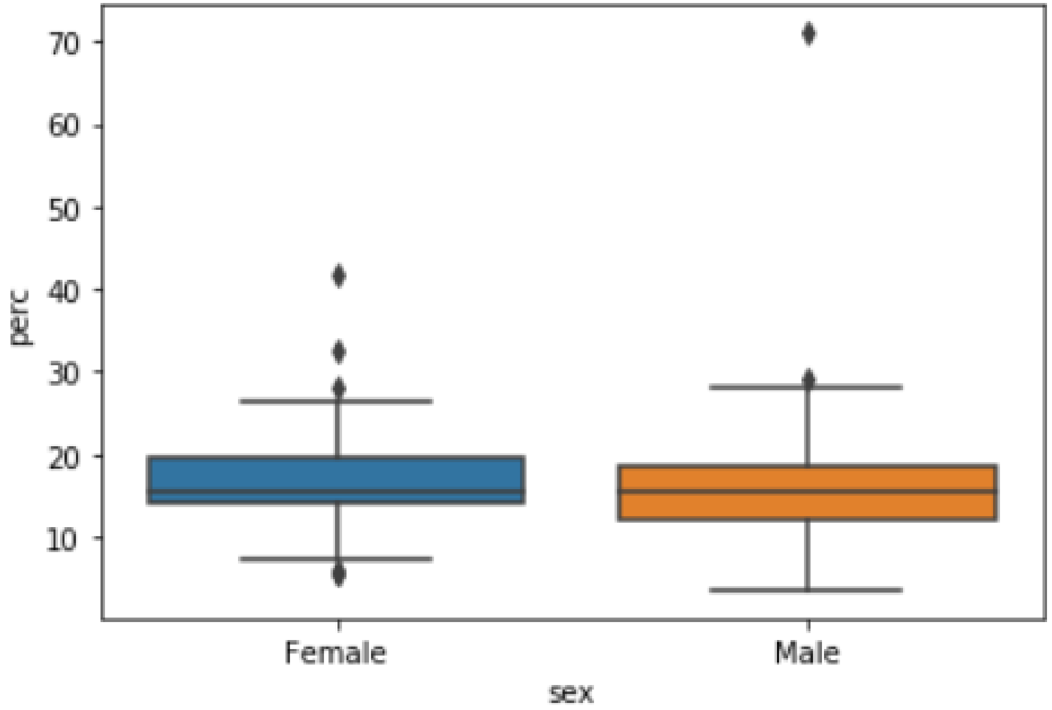
Pairplots

All numerical features skew right

Tips increase as total bill increases

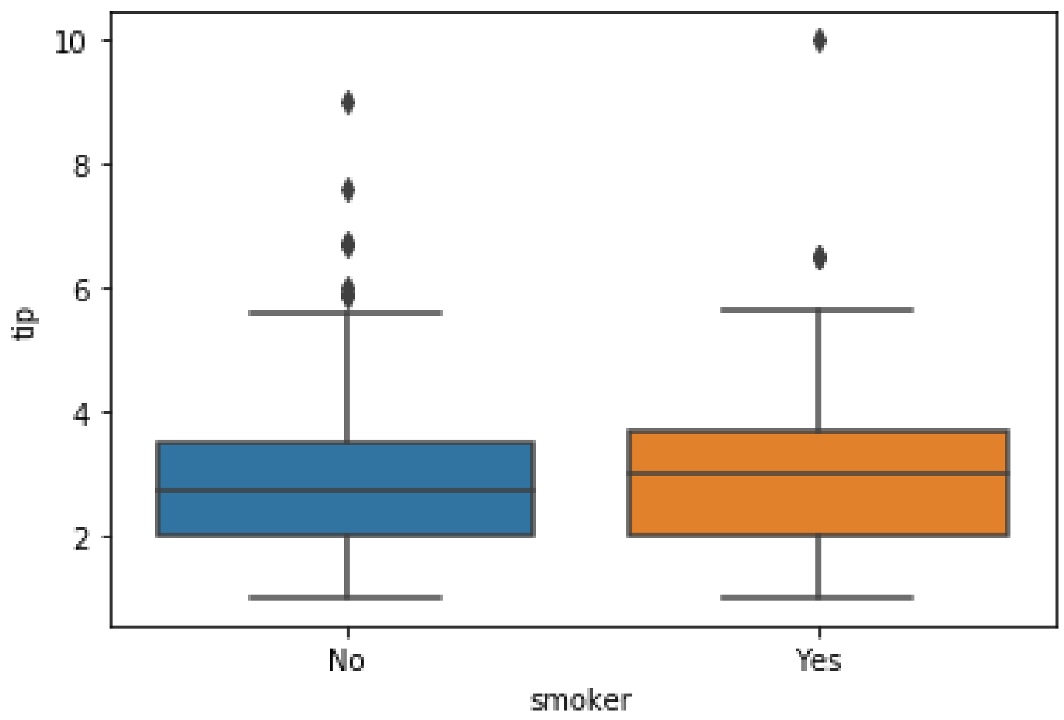


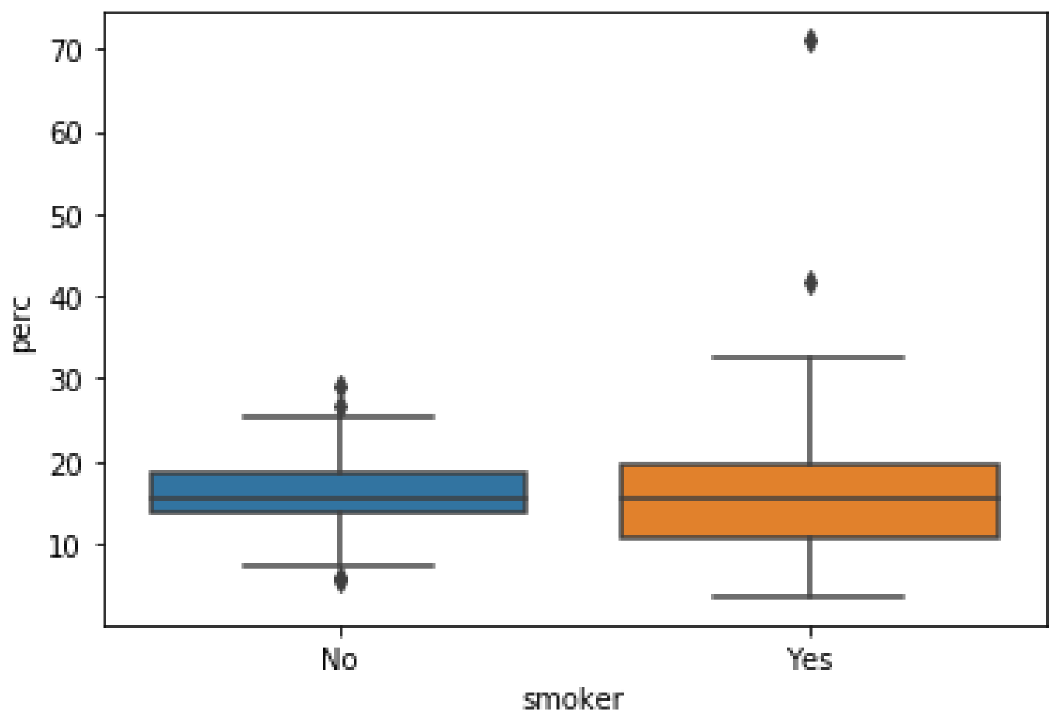




Sex

NO SIGNIFICANT DIFFERENCES IN EITHER TIP AMOUNT OF PERCENTAGES

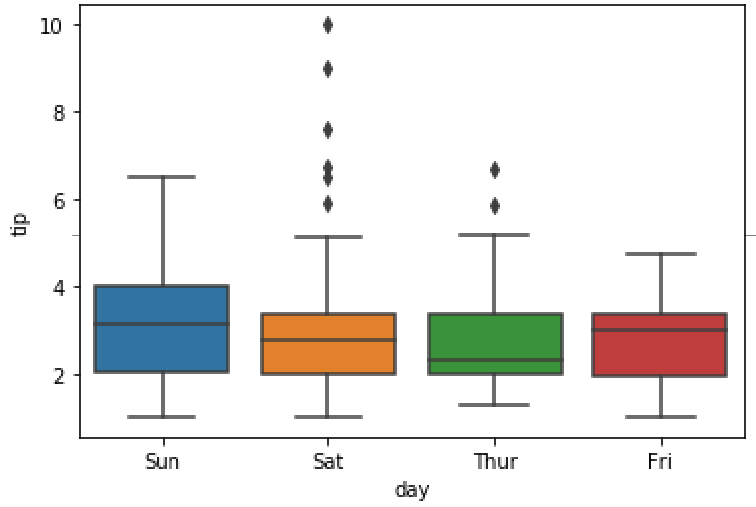


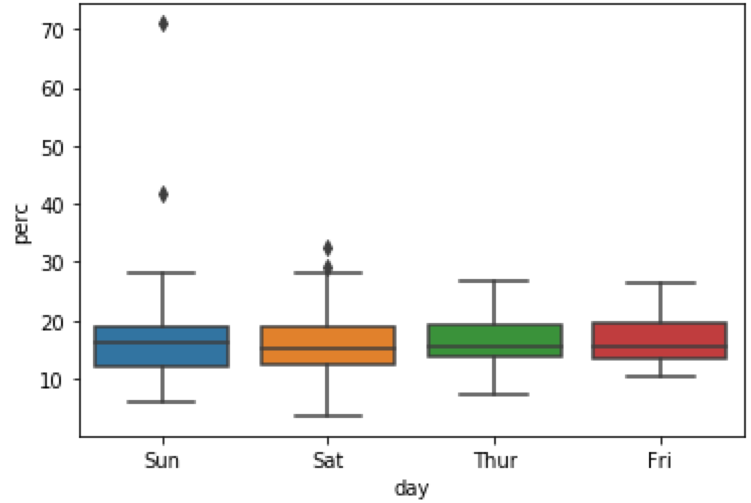


Smoker

NO SIGNIFICANT DIFFERENCES IN EITHER TIP AMOUNT OF PERCENTAGES

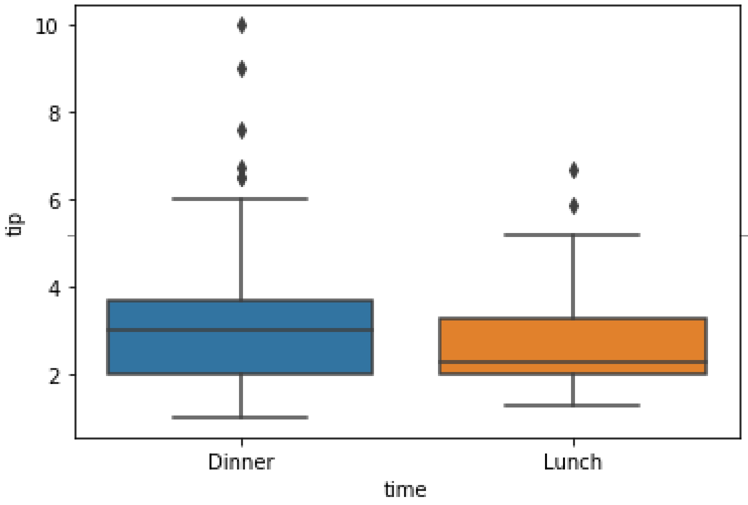
Day

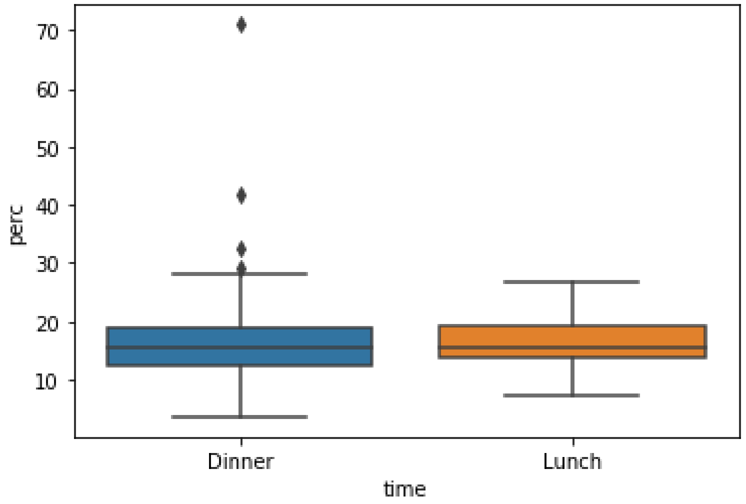




Significant difference found in tip amounts between different days

No significant differences found in tip percentages between different days

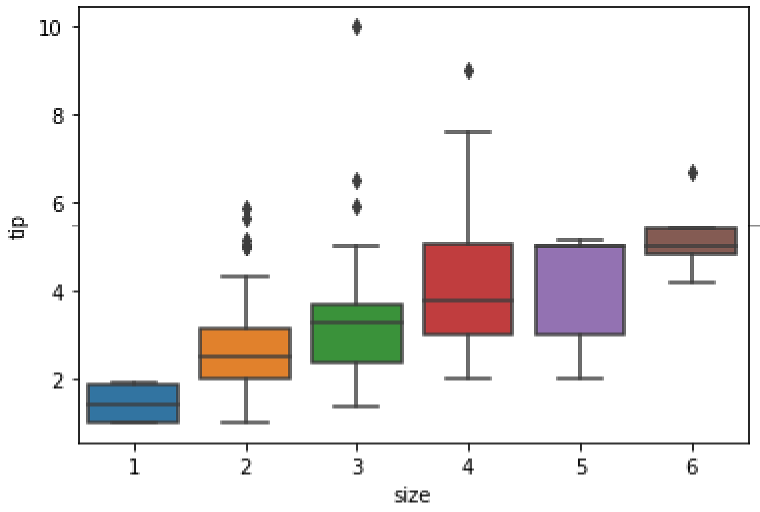


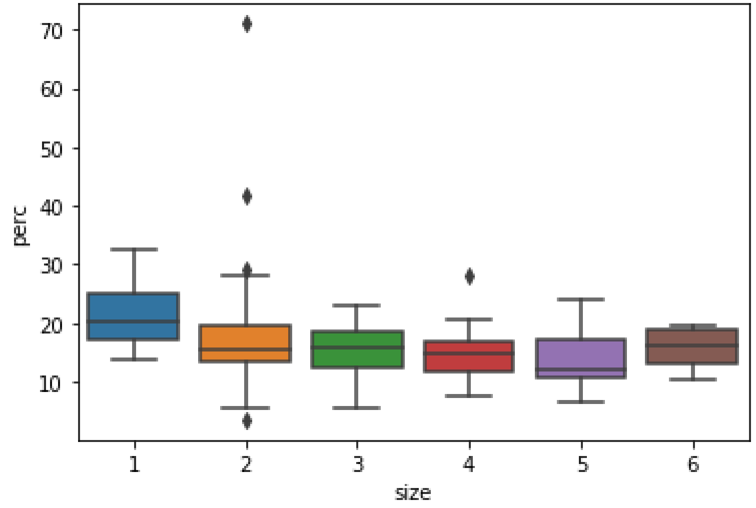


Time

Significant difference found in tip amounts between different shifts

No significant differences found in tip percentages between different shifts





Size

Significant difference found in tip amounts between different sizes

No significant differences found in tip percentages between different sizes

Machine Learning Modeling

Type: Supervised learning

Tools: Python’s scikit-learn

Low amount of data: Bootstrapping required

Modeling Steps

Pipeline

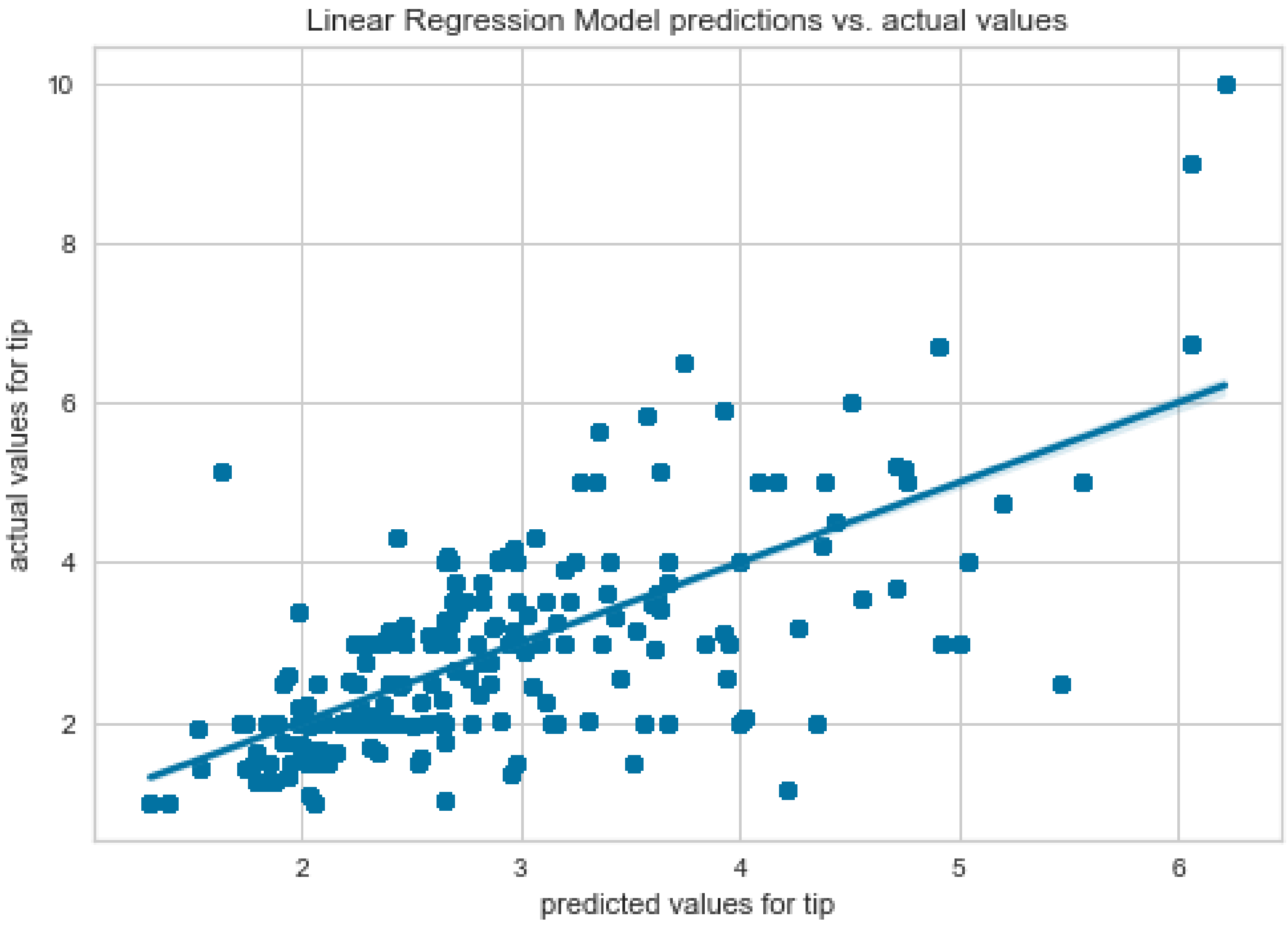
- Data Pre-Processing
  - 1. One-hot encoding
  - 2. Data splitting into training and test sets (80%-20%)
  - 3. Scaling
- Cross-Validation (CV) for Hyperparameter Tuning
  - 1. 5 fold CV
  - 2. Using scikit-learn's grid search method
  - 3. Evaluation metric: Mean absolute error

Performance evaluation using holdout dataset (20% of whole data)

Regression Algorithms Used

1. Linear Regression

2. Random Forest Regression



Model Comparisons

Model	Mean Absolute Error
Linear Regression Model	~.60
Random Forest Model	almost zero

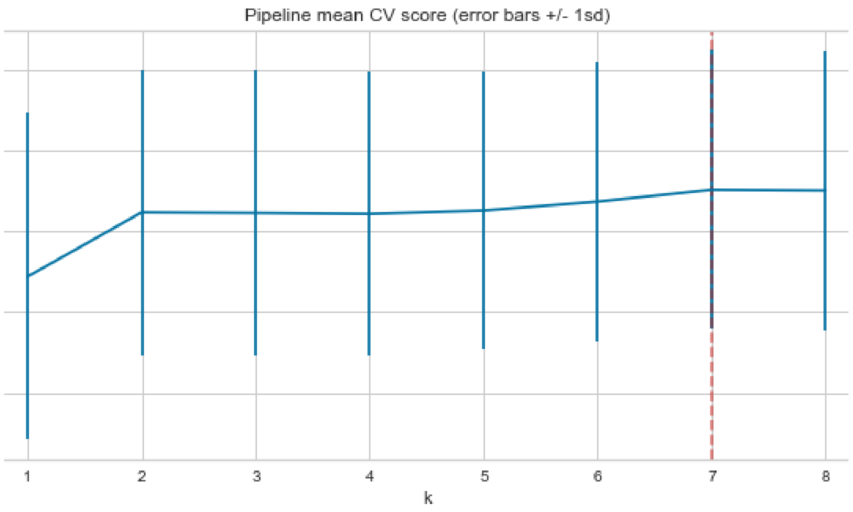
Model Comparisons

RANDOM FOREST DOES BETTER ON TRAINING DATA

Some Details on the Best Model

Features used: 7

Standard Scaling



total_bill	0.837978
size	0.169831
time_Lunch	0.122636
smoker_Yes	0.060106
day_Sun	0.047272
sex_Male	0.002088
day_Thur	-0.092271

Some Details on the Best Model

Total bill is most important feature for predicting tip amounts

Total bill and Table Size are most important features for predicting tip percentages

Testing on Under-Sampled Test Data

Model	Mean Absolute Error
Linear Regression	~0.60
Random Forest	~0.79

Testing on Under-Sampled Test Data

LINEAR REGRESSION DOES BETTER ON TEST SET

Input features used in Linear Regression model

Use model pipeline on new data and predict tip amounts

Using the Model

Using the Model

```
new_data = pd.DataFrame.from_dict(
    {'total_bill':[30.00],
              'size':[2],
              'sex_Male':[1],
              'smoker_Yes':[1],
              'day_Sat':[1],
              'day_Sun':[0],
              'day_Thur':[0],
              'time_Lunch':[0]})
tip_lr_grid.predict(new_data)
```

```
array([3.91518522])
```

Tip Percentage	
< 10%	Bottom priority
10% – 15%	Low Priority
15% – 18%	Medium Priority
18% - 20%	High Priority
> 20%	Top Priority

An Example of Model Usage: Possible Recommendations

Conclusion

Model might not be applicable to other restaurants

Success of this model will be encouraging for other restaurants

Assumptions, Limitations, and Disclaimers

WE ASSUME THAT ALL TABLES ARE INDEPENDENT, THOUGH THAT WOULD NOT BE THE CASE FOR REGULARS

USED ONLY ONE SERVER'S DATA FROM ONE RESTAURANT OVER THE COURSE OF A FEW MONTHS

THE MODEL MAY BEHAVE POORLY IF WE TRY TO PREDICT TIPS AND TIP PERCENTAGES OF OTHER RESTAURANTS

Diversify	Diversify information with more features
Extract	Extract information from more servers/bartenders and other restaurants
Include	Include dates so monthly/annual salaries can be calculated from tips predictions

More Ideas to Improve the Model in the Future

Thank you!

Saint Gau

Email: transaintgau@gmail.com

https://www.linkedin.com/in/saintgau/

https://github.com/transaint/Professional-Portfolio

Final project report: https://github.com/transaint/Springboard-Projects/blob/master/Springboard%20Projects/Predicting%20a%20Table's%20Tips/Final%20Project%20Report.ipynb