BEER CONSUMPTION IN BRAZIL

GENERAL ASSEMBLY DATA SCIENCE FINAL PROJECT

PROBLEM STATEMENT

- Beer is the most consumed alcoholic beverage in Brazil.
- Brazil is the third largest beer producer in the world.
- Brazil is also the third largest consumer of beer in the world, after China and the United States.
- Brazil entered a recession in 2014. The country started to recover in 2017, but corruption, the lack of public policies, and the economic crisis have led to an increase in poverty.
- Beer sales may help to drive the economy. On the other end, the lack of sales may hurt the economy.
- The main goal of this exercise is to predict the consumption of beer in one of the country's most populous cities, Sao Paulo.

HYPOTHESIS

- Beer consumption can be affected by many factors:
  - Temperature
  - Precipitation
  - Day of the week
  - Month
  - Season

THE DATA

- The data contains conditions for each day in a calendar year.

	Date	Average_Temp	Min_Temp	Max_Temp	Precipitation	Weekend	Beer_Consumption
0	1/1/15	27.30	23.9	32.5	0.0	0	25.461
1	1/2/15	27.02	24.5	33.5	0.0	0	28.972
2	1/3/15	24.82	22.4	29.9	0.0	1	30.814
3	1/4/15	23.98	21.5	28.6	1.2	1	29.799
4	1/5/15	23.82	21.0	28.3	0.0	0	28.9

THE DATA

- More columns can be created by parsing the "Date" column.

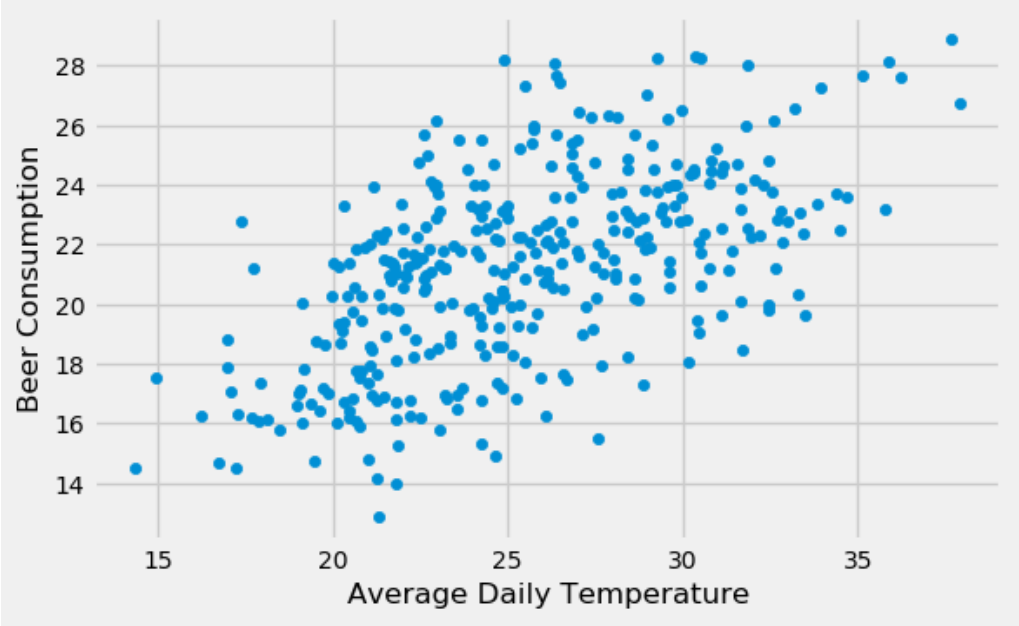
	Date	Average_Temp	Min_Temp	Max_Temp	Precipitation	Weekend	Beer_Consumption	Month	Season
0	1/1/15	27.30	23.9	32.5	0.0	0	25.461	1	Summer
1	1/2/15	27.02	24.5	33.5	0.0	0	28.972	1	Summer
2	1/3/15	24.82	22.4	29.9	0.0	1	30.814	1	Summer
3	1/4/15	23.98	21.5	28.6	1.2	1	29.799	1	Summer
4	1/5/15	23.82	21.0	28.3	0.0	0	28.900	1	Summer

METRICS AND LIMITATIONS

- The success metric is limited to the ability to predict the amount of beer consumed in Sao Paulo and whether there is a trend in the data based on the feature columns.
- Look at R-squared test set values and RMSE values.
- The data is only based on one year (2015).
- The dataset focuses on an area in the city, rather than the whole city.

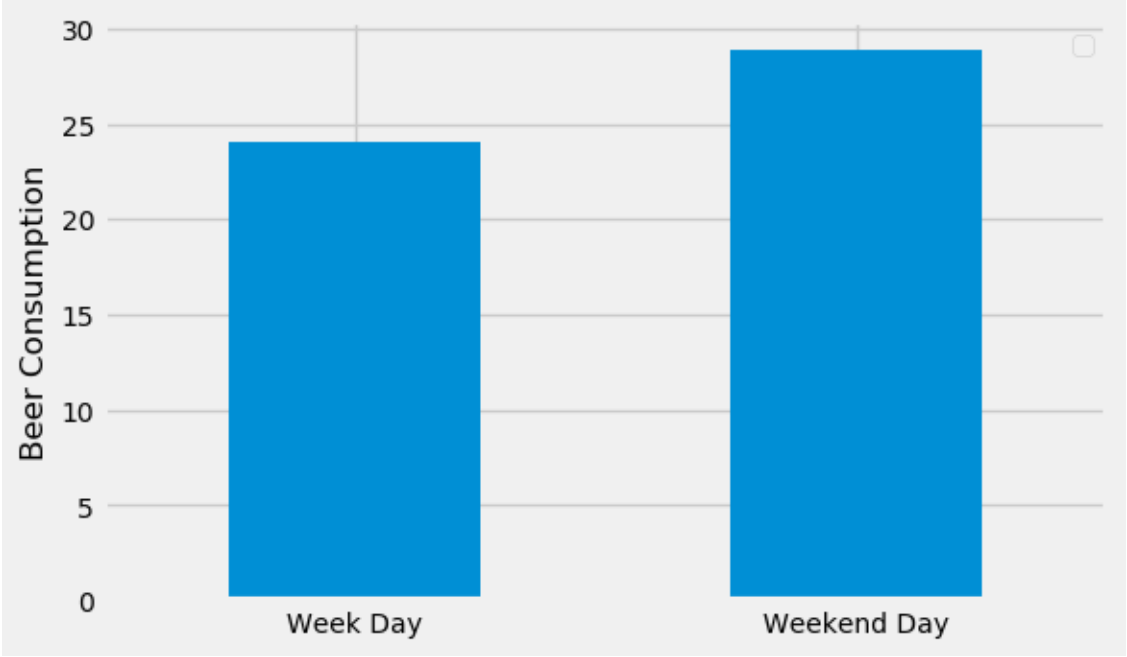
DATA EXPLORATION

- Investigate relationship between beer consumption and feature columns.
- Beer consumption increases with temperature.



DATA EXPLORATION

- Average beer consumption is higher on weekend days than on week days.



APPROACH

- I tried out three different models to see which would best work with my dataset:
  - Linear Regression
  - Decision Tree Regressor
  - Random Forest Regressor

MODEL EVALUATION

- Linear regression:
  - R-squared test value: 0.724
  - RMSE test value: 2.28
- Decision tree:
  - R-squared test value: 0.553
  - RMSE test value: 2.90
- Random forest:
  - R-squared test value: 0.700
  - RMSE test value: 2.38

MODEL EVALUATION

- The model performs the best with linear regression.
- Feature engineering helped a little but not much.
- More data is needed for a better score.

IMPACT

- This model can help the Brazilian government predict beer consumption with an R-squared value of 0.724.
- This model can be used to forecast the economic trends based on the time of year and the weather.

CONCLUSION

- Recommendations
  - This model can be incorporated but with warning.
  - More data should be researched and added to the dataset for better results.
- Next Steps
  - Incorporate time of day into the dataset.
  - Expand to more areas in the metropolitan area and eventually to other cities and rural areas in the country.
  - Look at data from other years- how does weather change and how does this affect beer consumption?