Chocolate Bar Ratings

An open-sourced data to help analyse the relationship between the percentage of cocoa bean in chocolates produced and the rating given to these chocolates.

Objective To uncover the relationship between the chocolate rating and the percentage of cocoa bean in the chocolate.

Data

- This data is publicly available (open source) and can be downloaded from Chocolate Bar Ratings.

Limitations

- Some of the values in column 'Broad Bean Origin' and 'Company Location' has more than one country, I had to select one of the countries for the analysis.
- In terms of ethics, the data does not contain any personal information.

Skills Data Cleaning: Wrangling and subsetting. Data consistency checks. Combining and exporting data. Deriving new variables combination. Grouping data and aggregating variables. Used Python and Tableau for visualization and excel for reporting.

Tools MS Excel, Python, Anaconda, Pandas, Numpy, Jupyter, Seaborn, Matplotlib. Data wrangling and visualization In Python and Tableau for visualization.

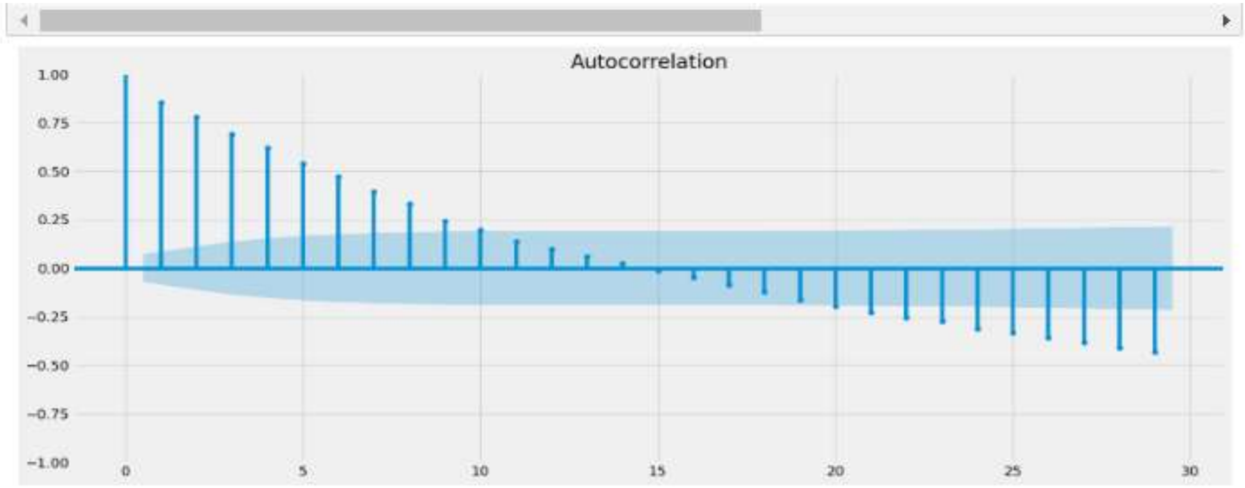
- Check the Tableau Storyboard here
- Check the Python scripts here

Initial Analysis

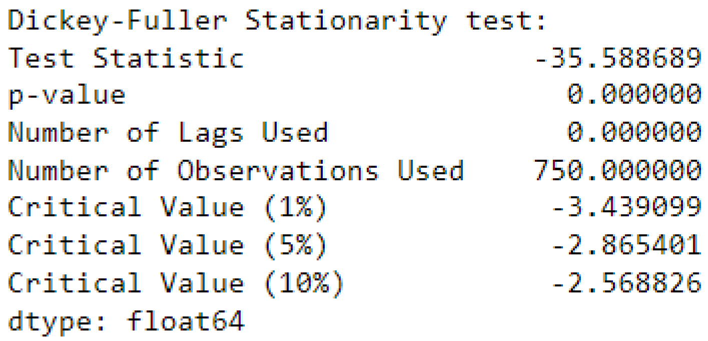
•Tested the stationarity of the relationship between Cocoa Percent and Rating using Dickey-Fuller Stationarity test.

We are not able to affirm that Cocoa Percent affects chocolate Ratings. It's possible that we need to have more data points or more source data to make better assumptions.

First autocorrelation test



Second Dickey-Fuller Stationarity test

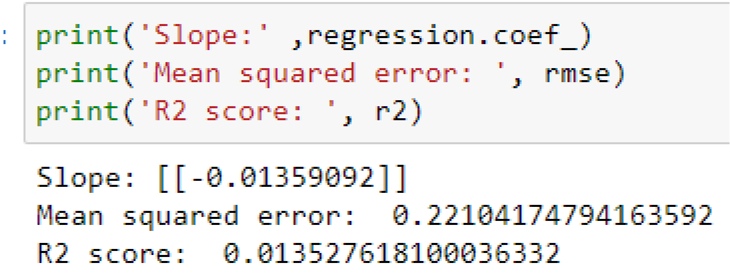


```
Dickey-Fuller Stationarity test:
Test Statistic                -35.588689
p-value                         0.000000
Number of Lags Used             0.000000
Number of Observations Used   750.000000
Critical Value (1%)            -3.439099
Critical Value (5%)            -2.865401
Critical Value (10%)           -2.568826
dtype: float64
```

For both first and second Dickey-Fuller Stationarity tests.

Analysis of Relationship

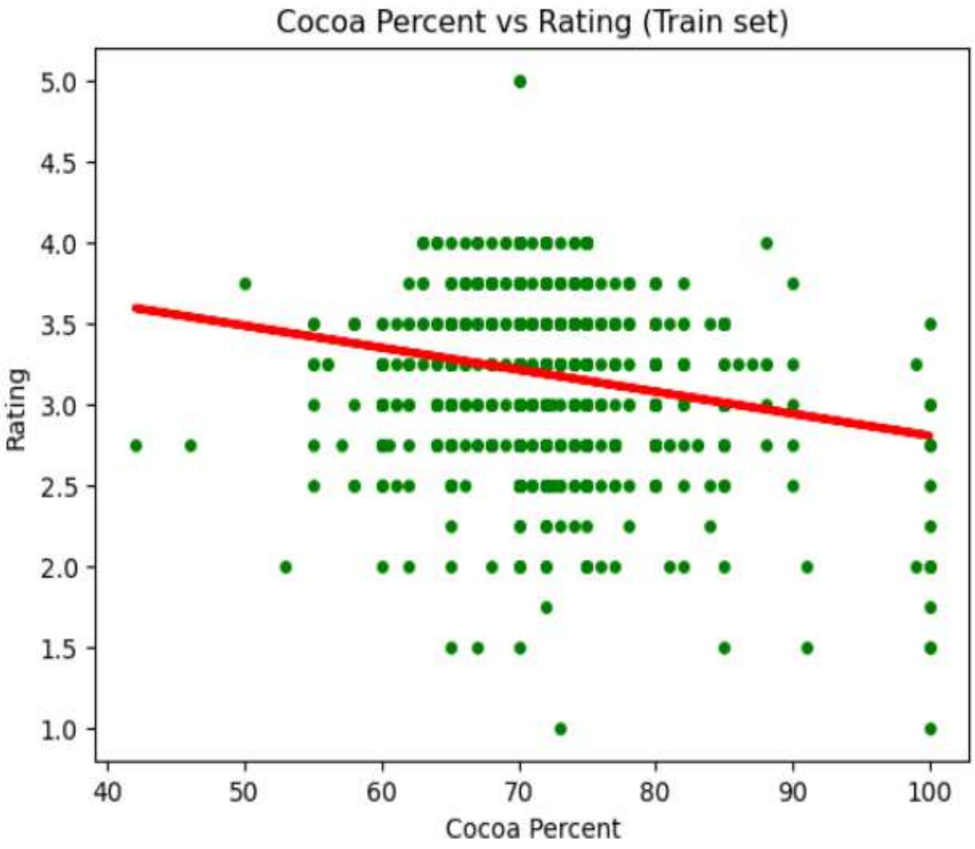
After conducting a Regression Analysis, it was found out that there is no relationship between Cocoa Percent and Rating, using Train set.



```
print('Slope:' ,regression.coef_)
print('Mean squared error: ', rmse)
print('R2 score: ', r2)
```

```
Slope: [[-0.01359092]]
Mean squared error:  0.22104174794163592
R2 score:  0.013527618100036332
```

summary statistics of Test Set



The model's outcome on the training set is very similar to that on the test set. With the MSE (mean_squared_error) being even larger on the train set. This proves that the Cocoa Percent are not the driving factor of chocolate Ratings. There might be other variables contributing to the chocolate Ratings apart from the Cocoa percent in each chocolate.

RECOMMENDATIONS

There is no relationship between the Percentage of cocoa in the chocolates and the Ratings. A high or low rating for each of the chocolate does not depend on the quantity of cocoa in them.

Therefore, there is a need to explore more information outside of the data available.