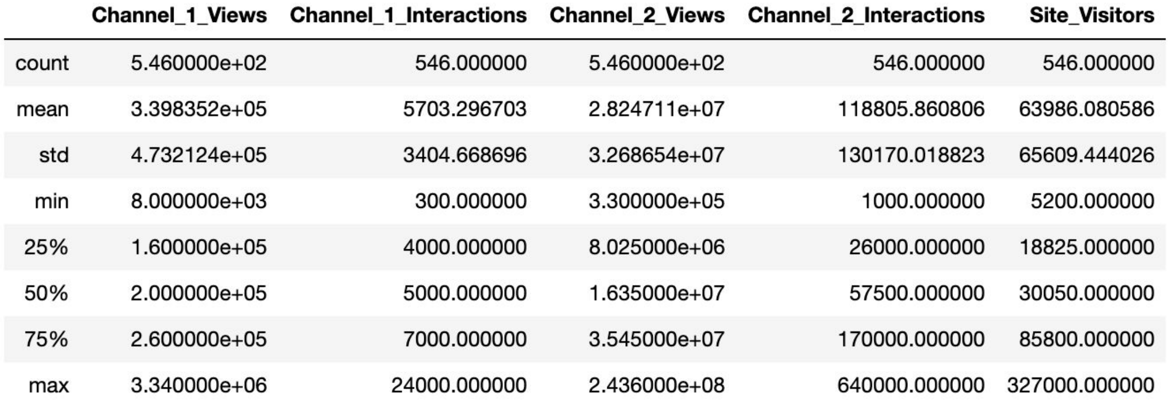
Good Apple Demo Data Analysis

By Anastasia Gorina

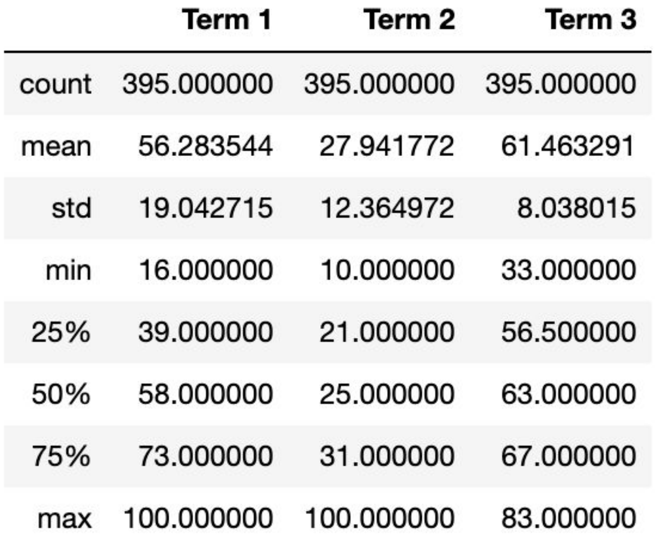
Overview

- Cleaning data and dropping NaN values, converting data to appropriate data types
- Calculating basic statistics (central tendencies and spread)
- Visualizing distributions of the variables (treating them as both discrete and continuous variables)
- Time series to predict search interest by term using Facebook Prophet
- Simple linear regression and multiple regression modeling with OLS to predict the number of site visitors

Basic Statistics

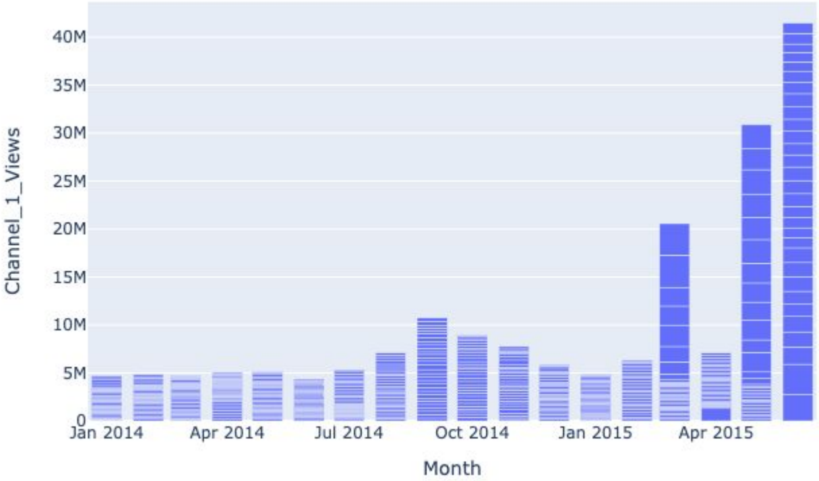


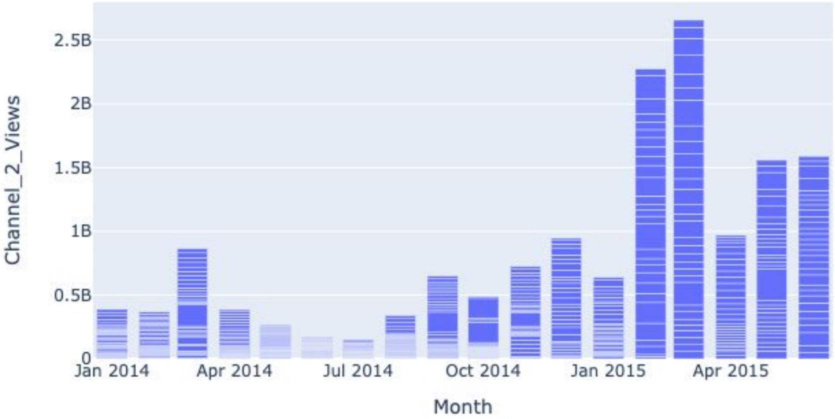
	Channel_1_Views	Channel_1_Interactions	Channel_2_Views	Channel_2_Interactions	Site_Visitors
count	5.460000e+02	546.000000	5.460000e+02	546.000000	546.000000
mean	3.398352e+05	5703.296703	2.824711e+07	118805.860806	63986.080586
std	4.732124e+05	3404.668696	3.268654e+07	130170.018823	65609.444026
min	8.000000e+03	300.000000	3.300000e+05	1000.000000	5200.000000
25%	1.600000e+05	4000.000000	8.025000e+06	26000.000000	18825.000000
50%	2.000000e+05	5000.000000	1.635000e+07	57500.000000	30050.000000
75%	2.600000e+05	7000.000000	3.545000e+07	170000.000000	85800.000000
max	3.340000e+06	24000.000000	2.436000e+08	640000.000000	327000.000000



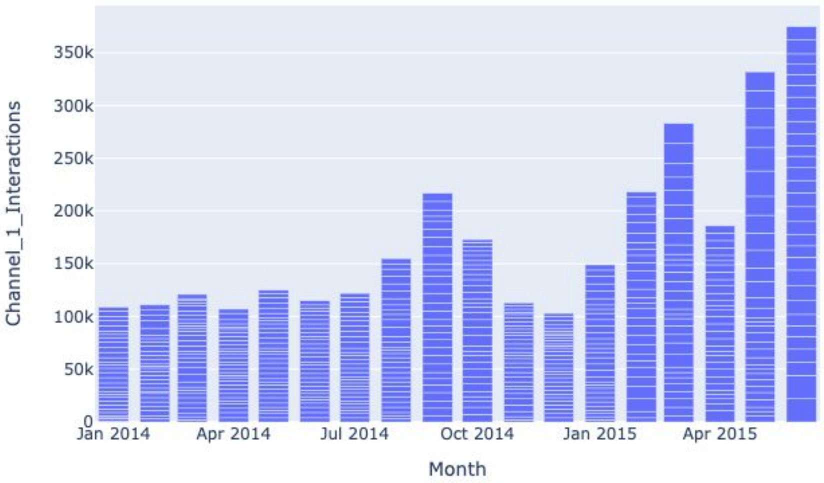
	Term 1	Term 2	Term 3
count	395.000000	395.000000	395.000000
mean	56.283544	27.941772	61.463291
std	19.042715	12.364972	8.038015
min	16.000000	10.000000	33.000000
25%	39.000000	21.000000	56.500000
50%	58.000000	25.000000	63.000000
75%	73.000000	31.000000	67.000000
max	100.000000	100.000000	83.000000

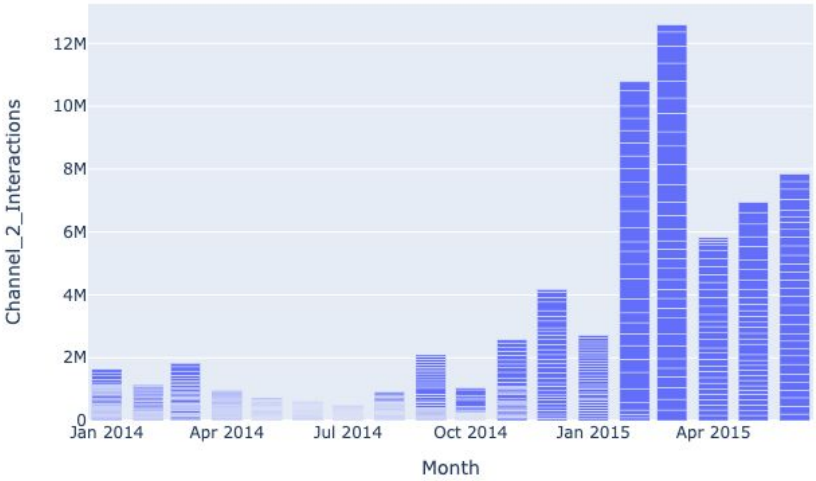
Channel Views by Month (discrete)



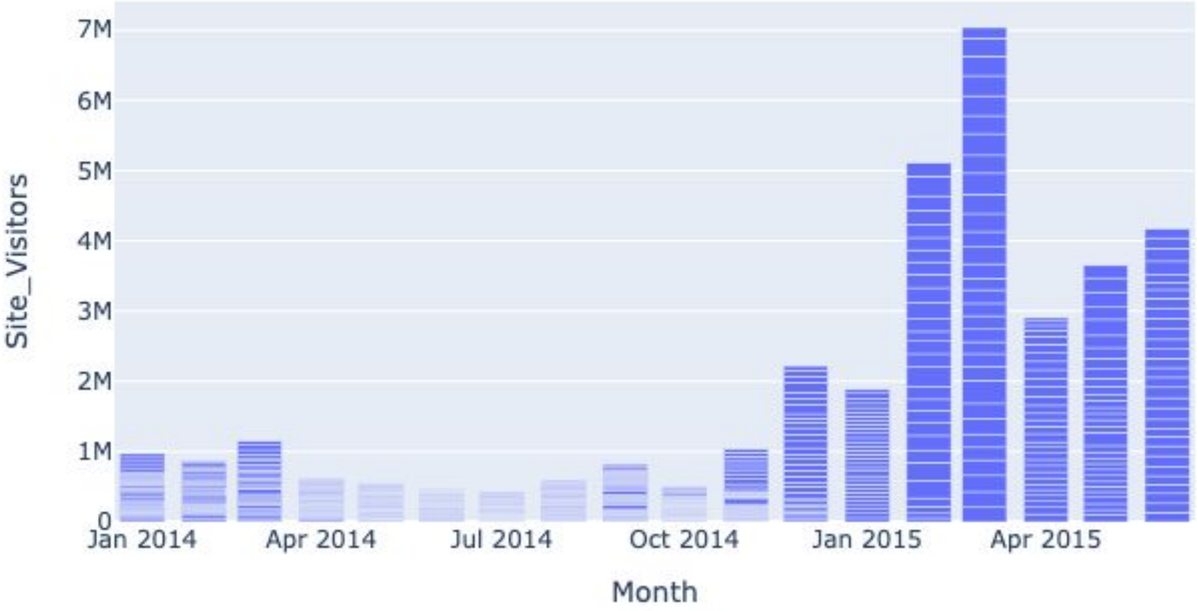


Channel Interactions by Month (discrete)

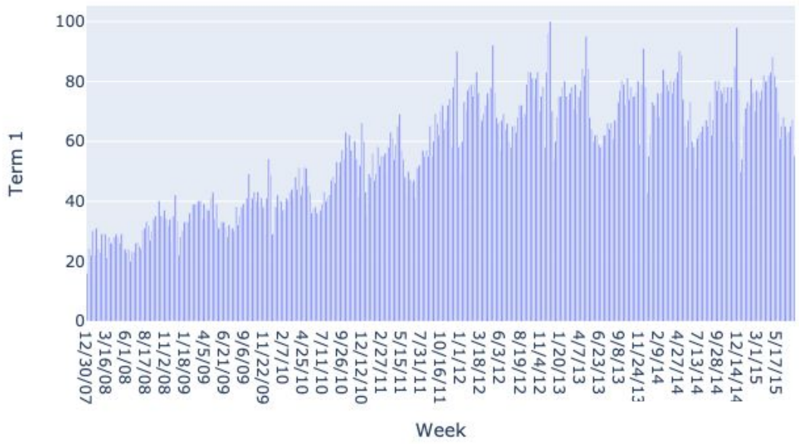


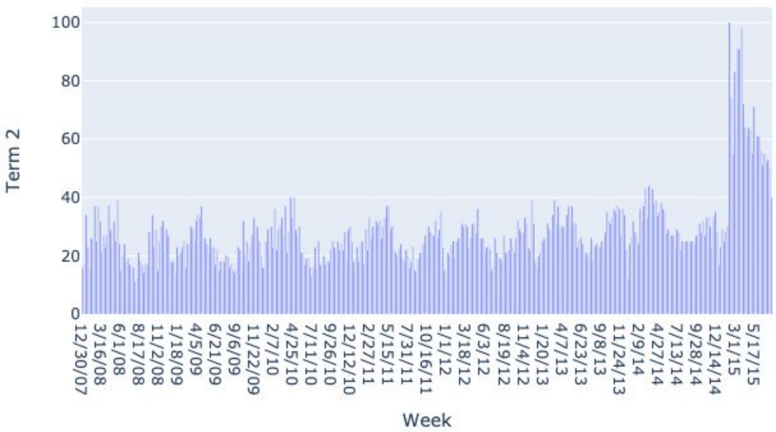


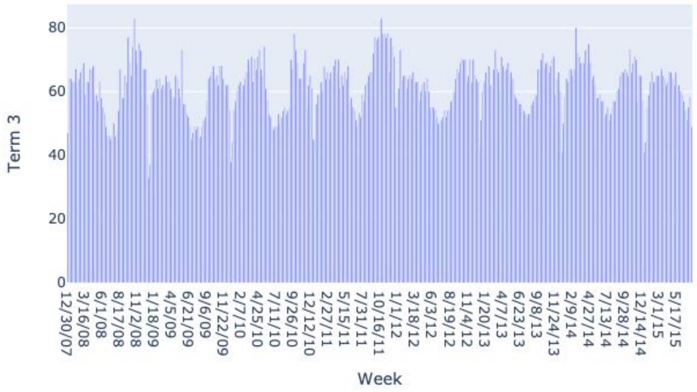
Site Visitors by Month (discrete)



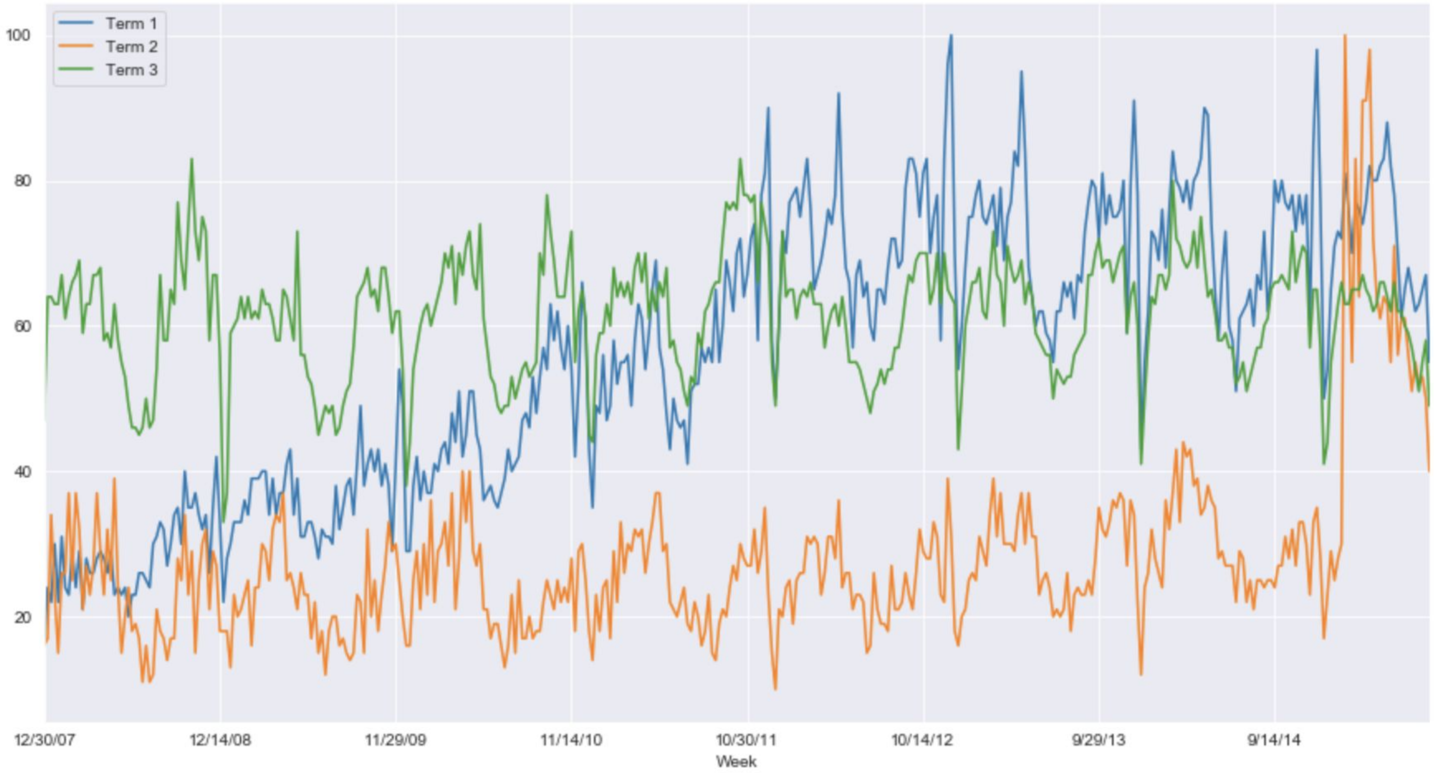
Site activity by Week



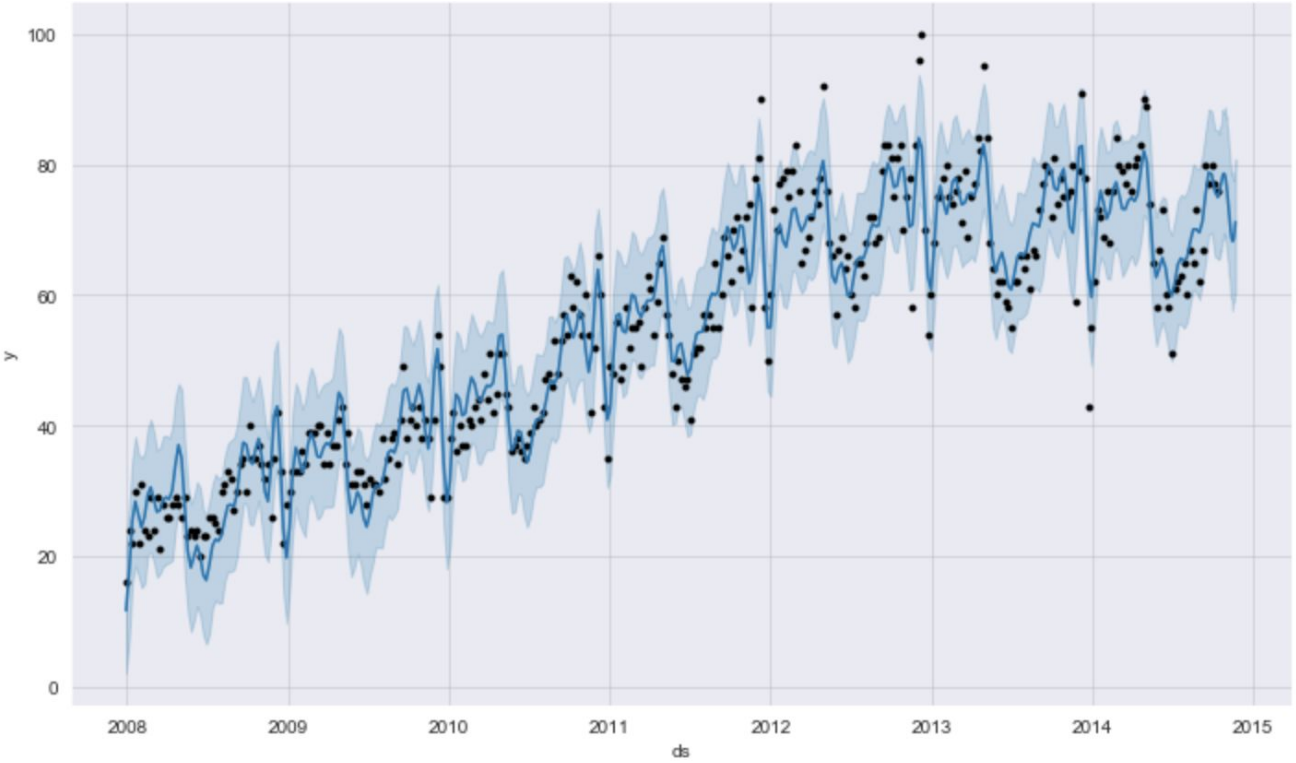




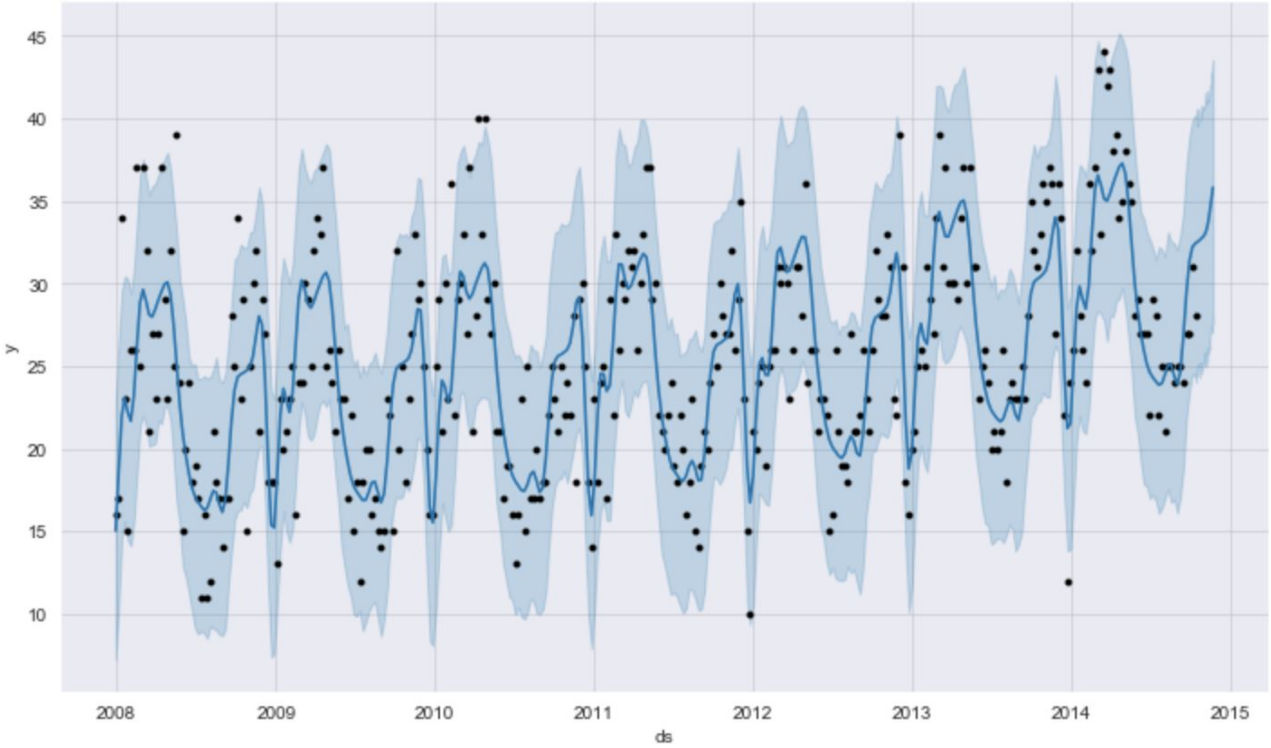
Search interest by Term (as continuous)



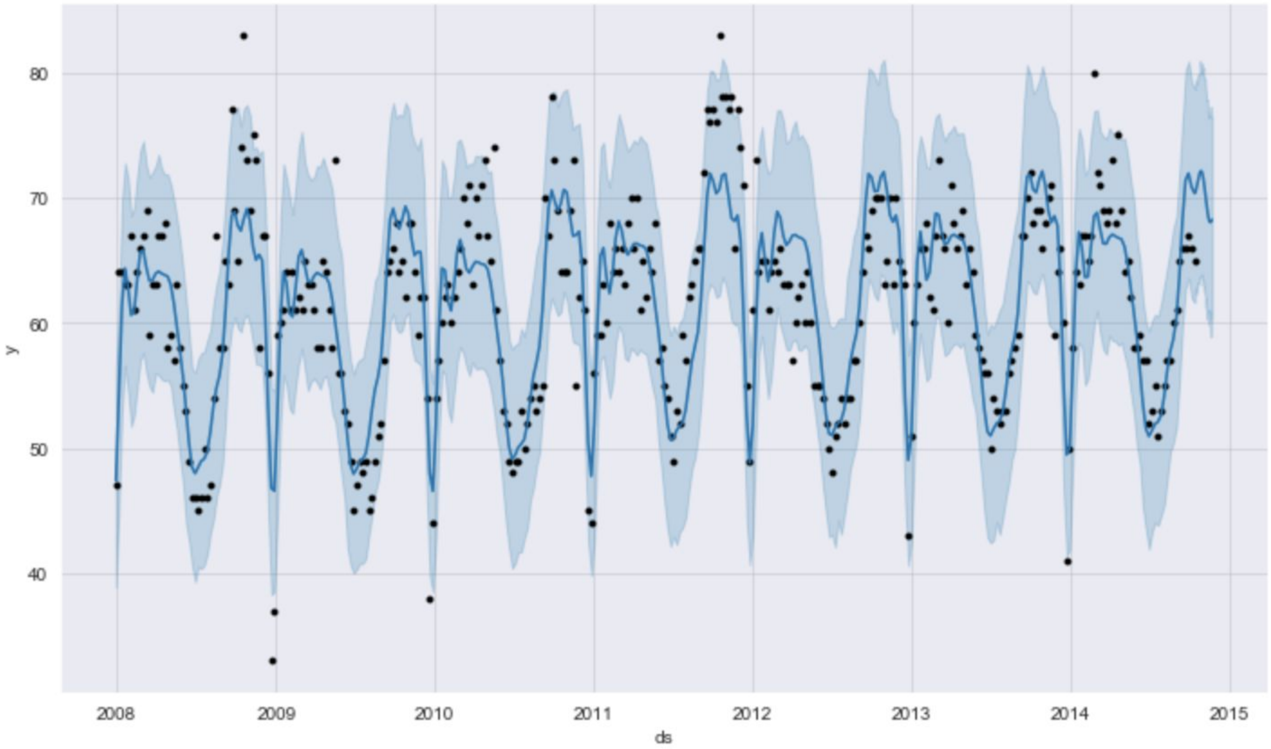
Time Series (Facebook Prophet) predictions for Term 1



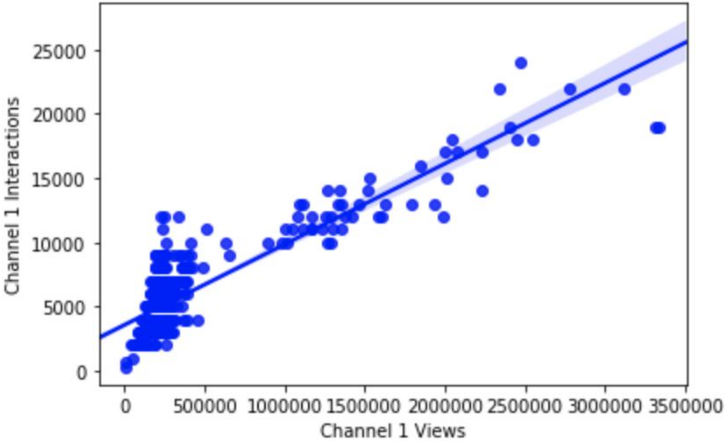
Time Series (Facebook Prophet) predictions for Term 2

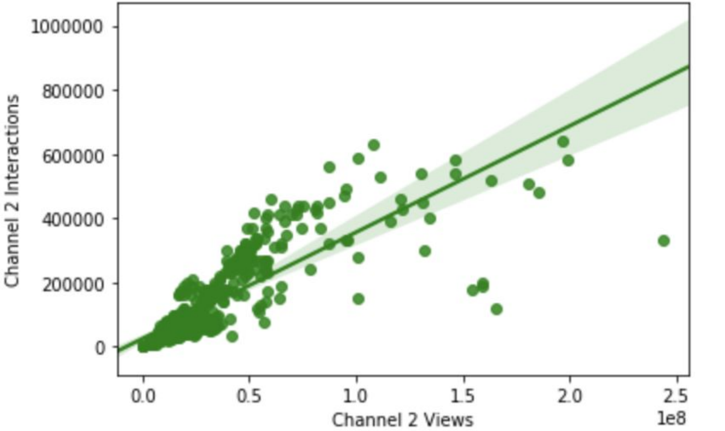


Time Series (Facebook Prophet) predictions for Term 3



Modeling results (Simple Linear Regression)





- Channel 1 interactions correlate with channel 1 views with Pearson R coefficient of 0.87

- Channel 2 interactions correlate with channel 2 views with Pearson R coefficient of 0.83

Modeling Results (OLS)

- The proportion of the variance for a dependent variable (site visitors) is explained by independent variables (both channels views and interactions) by adjusted R-squared of 0.96
- And predicting the same variable (site visitors) using just the views from both channels results in adjusted R-squared of 0.70
- However, only 15% of the variance for that variable is explained by the variance in search interest (taking X=term1, term2, term3), even in a log-transformed data

Notes

- Jupyter notebook with all the code, cleaned data, and visualizations can be found in my GitHub repo -- https://github.com/agorina91/Good_Apple

- Distributions of the variables were plotted using Plotly Express, making them interactive