Final Report

Using Google Searches to Predict Restaurant Guest Counts

Anthony Lusardi

Fall 2020

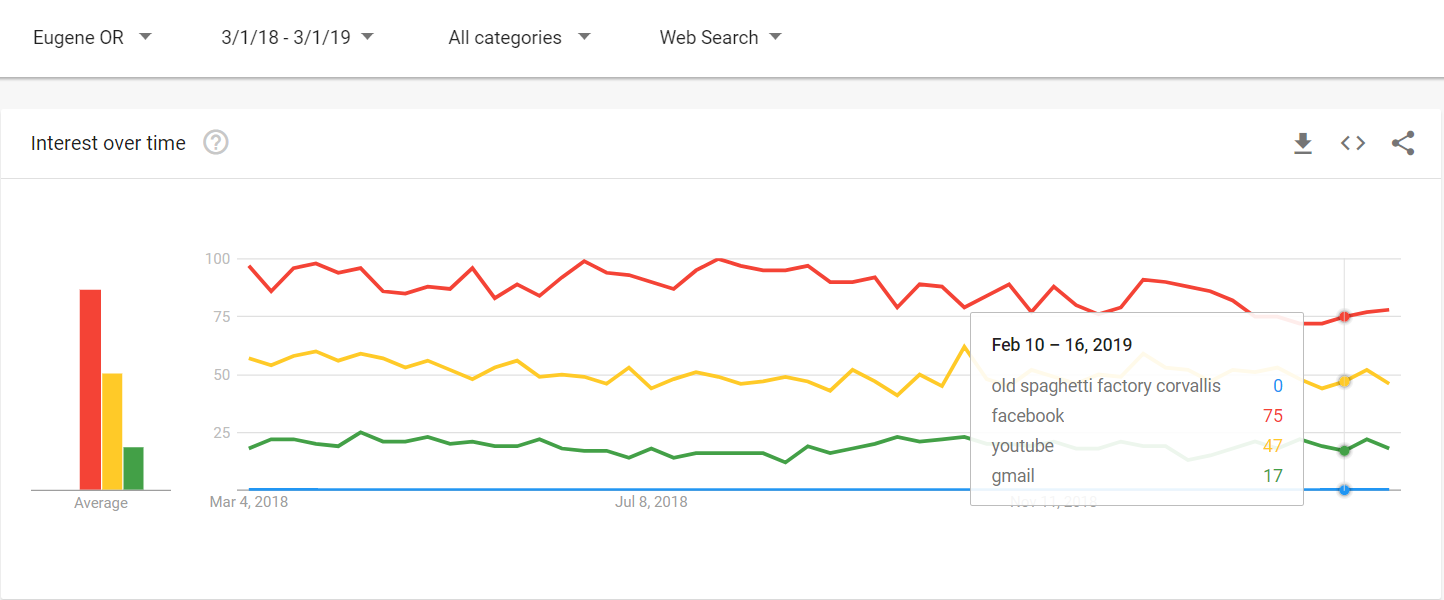
## Overview and Motivation

What I hoped to complete with this project is was use available data collected on google searches for a a restaurant title to answer questions about restaurant business patterns. By using search traffic patterns, a relationship could be drawn from the amount of traffic activity with the amount of clientle the restaurant receives at dinner time. I could then generate a linear regression model to predict the average amount of guests expected to come in on a given day based on the web searches collected from the local area.

Initially it was a far-fetched idea however, I had found In a previous study in 2019, a researcher had been able to predict the forecasted visitors on opening night for a movie theater based on the google searches for the movie title in a geographic area. It was a simple idea that proved to be very successful for consulting a movie theater company in Germany.

## Challenges

#### Collecting data from a Data Monopoly

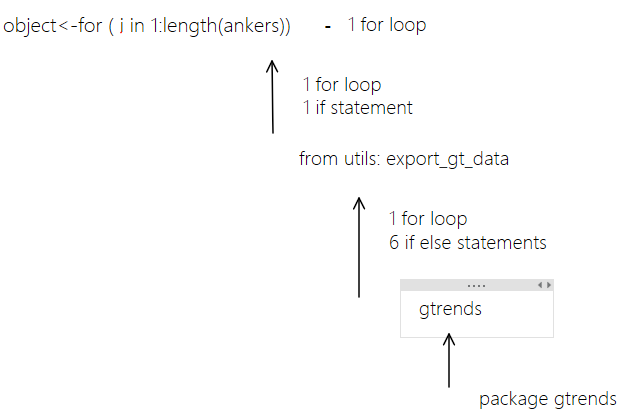
The single most important element of this project is scaling search activity. Google only shares data that is scaled from 0 to 100 and relative to search activity within the area with no indication as to how many searches there actually are.(Fig 1) Therefore, I needed to create a dynamic scaling system that creates a relationship between the search term and the constant search traffic terms call "anchor terms" eg:("facebook", "gmail").Part of this process included working with code that will use weighting function to assign a relative value of a Google search in relation with all searches in the area.[]

#### Working with another person's work

At first look, I was very impressed by the amount of documentation available for this repository. In addition, it was all written in R which was a great advantage when working with expediting the code-writing process. However, after taking a look under the hood especially on the utils folder, I found that although it was well-documented, there were quite a few inefficiencies and some unneeded functions such as adjusting for a premiere movie night schedules in Germany and multiple nested for loops .

#### Coding inefficiencies

Here is an example of a for loop that gathers data on anchor terms but depends on 2 other functions that are also for loops to do so (Fig 2)



## Data gathering -preprocessing and data analysis

The dataset used I had complied with conjunction with restaurant manager. It contains key data for recorded guest activity at the Old Spaghetti Factory (OSF) in Corvallis OR, between 01/01/2018 and 12/31/2019. It contains 17 features of key data associated with predicted guest activity such as guest counts from the previous years, managers predictions, and six week trends. The guest counts for dinner will be the target variable that will be predicted.

After preprocessing of the raw OSF data is done in this script(insert link) the Google Trends data will be collected and preprocessed.

The order how to run the scripts to gather and preprocess the Google Trends data is

* Collecting the Google Trends data for the anchor terms running this script(insert link)
* Collecting and scaling the Google Trends data for the defined search terms running this script(insert link)

For data analysis and testing following a rmd document here (insert link) Using a program called exploratory note

## Results - Searches and Guest Counts

After preprocessing and scaling the two highest search terms "old spaghetti factory" and "spaghetti factory" from the 52 weeks from March 2018 to March 2019, we can see that the relationship of guest counts and search activity in the area is not a significant relationship.



Unfortunately, the visualization as well as the correlation and linear models show that scaled google searches do not seem to help predict weekly guest activity.

#### Linear Regression with Keyword behavior - Coef. Table

After fitting the model with the scaled keywords, we created a linear model to predict the dinner guest counts.

Unfortunately, the summary for this model is not looking promising, even with the most direct keywords to predict guest counts, with a high -pvalue and low R squared value there is no significant evidence to say that keywords predict guest behavior.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Term | Coefficient | Std Error | t Ratio | P Value | Conf High | Conf Low | Base Level |
| (Intercept) | 1715.66747 | 34.26999 | 50.0632667 | 0.0000000 | 1782.83664 | 1648.49829 |  |
| old spaghetti factory | 24.21650 | 43.52042 | 0.5564401 | 0.5804951 | 109.51652 | -61.08351 |  |
| spaghetti factory | -20.55753 | 43.51396 | -0.4724353 | 0.6387571 | 64.72984 | -105.84490 |  |

## Learning Opportunities

*There are no mistakes just learning opportunities - Dad*

#### Code Efficiency

In order to make the code more efficient, I would need more time to really pick apart the he nested for loops. Much of this could be improved with a total reconstruction of the repository. At the end of it I found that the only necessary functions could be summarized into 25% of the volumes written in the repository. Fortunately, I had spent most of my time understanding and mapping the repository with notes, which led to a greater understanding of the process.

#### Geographical Issues

In the future I would focus more time on finding better geographic areas to isolate search terms. The anchor terms that are used were taken from the entire state of OR as designated with the gtrends function. I could not find a way to isolate searches within the Eugene Area as indicated by google trends interface in (Fig1)

An alternative could be to download the scaled title with every search term in .csv's to use the scale function to then isolate a more accurate geographic area.

#### Prediction

Consistent with the results, I found that seasonality and time dependency was an enormous factor in predicting guest counts.A GAM time series model was also used in the box office prediction. Initially, I had the idea to do a linear regression model because there was a high correlation of variables and that it was the most simple version of predicting data. For future studies, I will use a time series.

#### References