Halloween Candy Analysis

Prepared by Michael Striffler, Azzy Caceres and Kholiswa Tsotetsi

Rutgers University Coding Bootcamp

October 24th, 2020

Abstract: In this analysis, two datasets were scraped from Kaggle, a crowd-sourced platform which allows users to publish and extract open datasets. The first dataset, published by FiveThirtyEight, ranks common candies offered on Halloween against one another in a series of matchups. The second dataset, uploaded by user Seif Mohmed, is raw survey data containing rankings of Halloween candies. Both of these datasets were extracted from Kaggle, transformed using pandas in Jupyter Notebook and loaded to a PostgreSQL database.

Introduction

What comes to mind when you think of Halloween? Bats? Skeletons? Horror films? During this time of year, stores across the country begin stocking their shelves full of bags of assorted candies in anticipation of hoards of customers who plan to distribute the treats to trick-or-treaters. While part of the fun of Halloween festivities is randomly receiving various kinds of candies, it is only natural to have a predilection for a particular type. This analysis seeks to uncover the most popular candies based on two independent surveys. The value in this work is the discovery of the brand and type of candy people tend to favor. In turn, this can potentially impact brand decision-making when creating new variations of already existing candies.

Data Cleanup & Analysis

Kaggle, a platform which allows users to publish and extract datasets among other uses, was used as the data source. Using specific keywords -- “Halloween” & “candy” -- two fairly robust datasets were located. These two datasets, one from FiveThirtyEight and the other from user Seif Mohmed, are both csv files which were downloaded. The first dataset contains information about the candy name, though the column was labelled competitor name, type of candy (chocolate, fruity, etc), sugar percentage of each candy, the price percentage, and the overall “win” percentage based on the the number of times a particular candy won in a series of over 200,000 matchups. The second dataset contains a series of responses from anonymous survey respondents detailing their thoughts on a given particular candy. Here, survey respondents affiliated each candy to a particular emotion; either “joy”, “meh”, or “despair”.

In Jupyter Notebook, the dependencies imported for this analysis were pandas, sqlalchemy, and Flask. Additionally, sql_keys was created to contain username and password information for PostgreSQL. The first dataset was uploaded in Jupyter Notebook and subsequently transformed into a pandas dataframe. From here, the rename function was used to change the name of the first column from competitor name to candy name. The second dataset required further cleaning prior to being uploaded into Jupyter Notebook. This dataset contained over two thousands responses so a secondary csv sheet was created to tally the number of reactions (joy, meh, or despair) each candy received utilizing the excel function ‘CountIf’. Here, the candies that were present on both datasets were identified and cleaned to ensure that each candy name was in the exact same format on both sheets. In addition to the raw counts, the percentage of each emotion was calculated for each candy. From here, the second dataset was uploaded into Jupyter Notebook for further analysis. In Jupyter Notebook, the second dataset was also transformed into a pandas dataframe. At this point, using the pandas merge function, specific columns from the two datasets were merged. Specifically, the win percent, price percent ,and sugar percent from the first dataset and the calculated percent joy, meh, and despair from the second dataset.

Once the datasets were transformed, identical tables were created in the relational database, PostgreSQL. Because information is linked between two tables using a primary key for the analysis, a relational database was chosen over a non-relational database such as MongoDB. ERD, entity relational diagrams, was utilized in order to create a schematic of the tables. The output was loaded into PostgreSQL. Back in Jupyter Notebook, an engine was created to connect to PostgreSQL database. The pandas to sql function was used to load each table into the corresponding table in the PostgreSQL database. In order to check if each table loaded successfully into PostgreSQL, the read sql query function was used to verify.

Next, Flask was utilized to visualize the top candies from both datasets. Here, similar to the functions used in Jupyter Notebook, an engine was created to connect to the PostgreSQL database. An app route was created detailing two additional routes containing lists of the top candies. The first app route (/top_candy) looked at the win percent and percent joy from the merged table. Here, the candies were filtered so that only the candies that had a win percent of over 50% were selected. Using this route, it is evident that Reese's Peanut Butter Cups were the most popular candy. The next app route (/candy_hierarchy), was filtered in the same manner and ordered by the percent joy in descending order. Again, the top candy was Reese's Peanut Butter Cups.

These findings were highly interesting as both datasets were independent of one another. While one dataset is comprised of over 200,000 matchups where Reese's Peanut Butter Cups won at the highest rate, in the other dataset, which is comprised of over two thousand survey replies, Reese's Peanut Butter Cups also received the highest percent joy. This potentially gives some insight as to why there are so many Reese's peanut butter cup varieties --- Reese's Pieces, Reese's peanut butter cup minis, Reese's Stuffed with Pieces, etc-- as it is evident people like a peanut butter-chocolate combination. In addition to Reese's Peanut Butter Cups, the other candies that were also highly ranked amongst both datasets were Twix, Kit Kats, and Snickers. Therefore, it can be concluded that people tend to favor chocolate over fruity candies.