ETL Report

Dean Papadopoulos

Kripanjali Dhungana

Norman Morris

Zheding Zhao

ETL Process Date: 08/08/2022

Introduction:

For our Marketing Analytics project, we are going to determine the factors of a yelp dataset and its impact on the businesses. Our interest specifically generated on looking at the restaurant's reviews. The datasets we are using for this project are the Yelp datasets, Census data and US Labor Statistics and Unemployment. The datasets we will be using consist of the business information, users' information, reviews, ratings, median income, education, unemployment rate based on the zip codes. We brainstormed some hypothetical questions while exploring the datasets that we deemed would be interesting to conduct an analysis on. Below are these questions:

- What are some keywords for good reviews and bad reviews?
- Is there a correlation between review score and length of review text?
- What is the relationship between median income and ratings on expensive restaurants?
- What is the relationship between median income and number of restaurants?
- Is there a correlation between the number of categories and average score?
- Are people who leave more reviews likely to be more critical than people who leave less reviews?
- Does lower unemployment result in more reviews on Yelp?
- Can classification machine learning models be used to group or find patterns in businesses or reviews?

After formulating our hypothesis, we will extract, transform, and load our data to the database before we could perform analysis on them to answer our questions. We will describe this process in detail in the following sections.

Datasets:

We found our datasets in a variety of websites. Our Yelp datasets come from the Yelp website, the unemployment data was generated from U.S. Bureau of Labor Statistics and, the income and education data come from United States Census Bureau.

Yelp. (n.d.). Yelp dataset. Retrieved August 8, 2022, from https://www.yelp.com/dataset/documentation/main

Department of Labor, U. S. (2022, March 14). *Civilian noninstitutional population and associated rate and ratio measures for model-based areas*. U.S. Bureau of Labor Statistics. Retrieved August 8, 2022, from https://www.bls.gov/lau/rdscnp16.htm#data

Bureau, U. S. C. (n.d.). *EDUCATIONAL ATTAINMENT*. Explore census data. Retrieved August 8, 2022, from https://data.census.gov/cedsci/table?q=S1501%3A+EDUCATIONAL+ATTAINMENT&tid=ACSST5Y2020.S1501

Bureau, U. S. C. (n.d.). *INCOME IN THE PAST 12 MONTHS (IN 2020 INFLATION-ADJUSTED DOLLARS)*. Explore census data. Retrieved August 8, 2022, from https://data.census.gov/cedsci/table?q=S1901%3A+INCOME+IN+THE+PAST+12+MONTHS+%28IN+2020+INFLATION-ADJUSTED+DOLLARS%29&g=0100000US%248600000&y=2020&tid=ACSST5Y2020.S1901

List of state abbreviations. List of state abbreviations (download CSV, JSON). (n.d.). Retrieved August 10, 2022, from https://worldpopulationreview.com/states/state-abbreviations

Extraction:

The datasets were extracted from Yelp, U.S. Bureau of Labor Statistics and United States Census Bureau. The Yelp data was extracted in JSON format and the rest of the datasets were extracted in csv format.

1. We went to the above-mentioned data sources to extract all the files.
2. We downloaded the review, business, user, check-in, and tip datasets from Yelp that were in JSON format and then uploaded to our Group 1-Marketing Analytics in the Data Lake Storage container one at a time.
3. We also downloaded the datasets for unemployment rate, education and income that were in csv format and uploaded to our Group 1-Marketing Analytics in the Data Lake Storage container.
4. We downloaded the csv file with state names and its abbreviation and uploaded it to our Group 1-Marketing Analytics in the Data Lake Storage container.

Transformation:

Below are the datasets that we have transformed from their raw form to the usable form in order to use the datasets and perform analysis. We transformed these datasets using Azure Databricks and our detailed ETL process can be found here Datasets ETL-Group 1

Business: Yelp dataset that contains information about businesses on Yelp.com

Transformation Process (Reference: Business_ETL Databrick):

1. Import Config Databrick for SQL Server and Mounting Credential: Cmd 1
2. Mount the Container for importing and exporting the data: Cmd 2
3. Import relevant packages and functions: Cmd 3
4. Read the Business dataset JSON file: Cmd 4
5. Filter out null rows: Cmd 4
6. Drop irrelevant columns: Cmd 4
7. Rename columns to fit SQL Database schema: Cmd 4
8. Filter Business table by Zip Codes in zipCodeTable: Cmd 5
9. List all fields in attributes column: Cmd 6
10. Remove BusinessParking and Ambience from attributes: Cmd 6
11. Create a new DataFrame from Business that only contains BusinessID and Attributes: Cmd 7
12. Replace ‘None’ in attributes to None (Null values): Cmd 8
13. Replace ‘True’ and ‘False’ into ‘1’ and ‘0’ respectively: Cmd 8
14. Resolve u-string issues in attribute values: Cmd 9
15. Drop attributes that contain more than 50% Null values: Cmd 10
16. Create Attributes DataFrame: Cmd 11
17. Filter unique attribute values: Cmd 12
18. Create AttributesValue DataFrame: Cmd 13
19. Create BusinessAttributes DataFrame: Cmd 14
20. Export the business table to the container for other datasets to use: Cmd 15
21. Write to the Business table in SQL Database from Corresponding DataFrame: Cmd 16
22. Write to the Attributes Table in SQL Database from Corresponding DataFrame: Cmd 17
23. Write to the AttributesValues Table in SQL Database from Corresponding DataFrame: Cmd 18
24. Write to the BusinessAttributes Table in SQL Database from Corresponding DataFrame: Cmd 19

Reviews: Yelp dataset that contains information about reviews on Yelp.com

Transformation Process: (Reference: Reviews_ETL Databrick):

1. Import Config Databrick for SQL Server and Mounting Credential: Cmd 2
2. Mount the Container for importing and exporting the data: Cmd 3
3. Import and read the yelp_academic_dataset_review.json into a pyspark dataframe: Cmd 5
4. Covert the date column from a string to a sate type: Cmd 6
5. Check columns of file for nulls in any of the rows: Cmd 7
6. Check types of columns to ensure they are all in proper formats: Cmd 8

7. Rename the columns in the data frame to match naming conventions to the SQL databse: Cmd 9
8. Import and read the cleaned business csv: Cmd 10
9. Create a list of all the BusinessIDs used in the Business Table: Cmd 11
10. Filter the main Dataframe to only include the list of BusinessIDs created in the previous command: Cmd 12
11. Import and read the cleaned user csv file: Cmd 13
12. Create a dataframe of only the UserID column: Cmd 14
13. Join the main dataframe with the UserID data frame on the UserID column: Cmd 15
14. Export the main cleaned dataframe to the container in JSON format: Cmd 16
15. Write to the Review Table in the SQL Database from the main dataframe: Cmd 17

Users Yelp dataset that contains information about users on Yelp.com

Transformation Process (Reference: Users_ETL Databrick):

1. Import and read JSON file into a Spark dataframe: Cmd 6
2. Drop ‘friends’ and ‘elite’ columns: Cmd 7
3. Drop nulls and duplicates: Cmd 8
4. Rename columns to match SQL schema: Cmd 9
5. Write to JSON called user_cleaned.json: Cmd 10

Tips Yelp dataset that contains information about tips on Yelp.com

Transformation Process (Reference: Tip_ETL Databrick):

1. Import and read yelp_academic_dataset_tip.json file into a pyspark dataframe: Cmd 2
2. Drop nulls and duplicates: Cmd 3
3. Rename columns to match SQL schema: Cmd 4
4. Write to JSON file named tip_cleaned.json in the data lake

Merge_And_Populate Databrick (Loads Users and Tips into Database)

1. Import and load user_cleaned.json file into dataframe: Cmd 7
2. Write cleaned user dataframe into SQL Server: Cmd 8
3. Import and load tip_cleaned.json and business.csv into dataframes: Cmd 10, Cmd 11
4. Perform an inner merge to get rid of entries that don’t match up: Cmd 12
5. Drop all columns but BusinessID, Date, ComplimentCount, Text, UserID: Cmd 12
6. Export dataframes to SQL Server: Cmd 13

Unemployment: Data about Unemployment Rate

Transformation Process (Reference: Unemployment_ETL Databrick):

1. Read in the excel file: Cmd 3
2. Rename the columns to their proper headings: Cmd 4
3. Check for nulls in each row: Cmd 5
4. Filter out the two codes that are not states: Cmd 6
5. Filter out unnecessary rows: Cmd 7
6. Drop any rows with nulls and only include rows from the year 2000 or greater: Cmd 8
7. Cast the Year & Month columns to integer and cast the UnemploymentRate Column to a float: Cmd 9
8. Import the cleaned state csv from the storage container and create proper headings for each row: Cmd 10
9. Join the state and unemployment data frames on the 'State' Column: Cmd 11
10. Create a list of the 13 states that our datasets contain and filter out the states that will not be used: Cmd 12
11. Write the final cleaned dataframe to the container in JSON format: Cmd 13
12. Write to the Unemployment Table in SQL Database from the dataframe: Cmd 15

Education: Census Data on Education

Transformation Process (Reference: Census_Education_ETL Databrick):

1. Import and read ACSST5Y2020.S1501_data_with_overlays_2022-04-22T163802.csv file into Spark dataframe: Cmd 4
2. Drop first row: Cmd 4
3. Select only 'GEO_ID' and 'S1501_C02_012E' columns: Cmd 5
4. Rename columns to 'ZipCode' and 'CensusEducationPct': Cmd 5
5. Extract zip code by doing a substring on ZipCode column for the 10^{th} to 14^{th} characters: Cmd 5
6. Write dataframe to JSON file named education_cleaned.json in the Data Lake: Cmd 6

Income: Census Data on Income

Transformation Process (Reference: Census_Income_ETL Databrick):

1. Import Config Databrick for SQL Server and Mounting Credential: Cmd 2
2. Mount the Container for importing and exporting the data: Cmd 3
3. Import the dataset and filter out the first row: Cmd 5

4. Next read in this filtered dataset via pyspark and set the first column as the header: Cmd 6
5. Create a ‘ZipCode’ column by taking a substring of the ‘id’ column using characters 10-14: Cmd 7
6. All columns besides ‘ZipCode’ and 'Estimate!!Households!!Median income (dollars)' income will be filtered out: Cmd 9
7. Next any rows with the value of ‘-’ need to be filtered out since these are used to depict nulls in the data: Cmd 10
8. Finally cast the median income to IntegerType: Cmd 12
9. Write cleaned dataframe to container in JSON format: Cmd 16

ZipCode: Table that contains Zip code and respective states

Transformation Process (Reference: ZipCode:_ETL):

1. Import Config Databrick for SQL Server and Mounting Credential: Cmd 2
2. Mount the Container for importing and exporting the data: Cmd 3
3. Import cleaned income JSON file: Cmd 5
4. Import cleaned education JSON file: Cmd 6
5. Import ZipCode csv file: Cmd 7
6. Create a new dataframe by joining both income and education dataframes on their ZipCode columns: Cmd 8
7. Filter zipcode dataframe to only keep zipcode and state abbreviations, rename these columns properly: Cmd 9
8. Join the zipcode dataframe to the main dataframe: Cmd 10
9. Filter the dataframe to only include the 13 states which are needed: Cmd 11
10. Rename the columns in the main data frame to match with the those set up in the database: Cmd 12
11. Write the cleaned ZipCode file to the storage container in JSON format: Cmd 12
12. Write to the ZipCodeTable in the SQL Database: Cmd 13

Kafka Producer: A producer that produces messages about state names and their abbreviations to the pre-made topic.

Transformation Process (Reference: Creating_Producer Databrick):

1. Mount the data lake container to import data: Cmd 1
2. Define functions that handle error messages: Cmd 2
3. Set up Kafka connections: Cmd 3
4. Initialize a Kafka Producer: Cmd 4
5. Import relevant packages: Cmd 5
6. Read the business table for filtering: Cmd 6
7. Filter the business table: Cmd 7
8. Read the state and abbreviation table: Cmd 8
9. Filter the state and abbreviation table: Cmd 8

10. Send the message to the topic via Kafka Producer: Cmd 9

Kafka Consumer: A consumer that consumes messages from the pre-made topic and sends it to SQL Server and Data Lake.

Transformation Process (Reference: Creating_Consumer Databrick):

1. Mount the data lake container to import data: Cmd 1
2. Define functions that handle error messages: Cmd 2
3. Set up Kafka connections: Cmd 3
4. Initialize a Kafka Consumer: Cmd 3
5. Consumer the message from the Kafka Topic: Cmd 4
6. SQL Connection String: Cmd 5
7. Make a Pyspark Dataframe based on consumer messages: Cmd 6
8. Write the dataframe to the data lake container as a csv: Cmd 7
9. Send the dataframe to the SQL Server Database: Cmd 8

Load:

After uploading the datasets in the Databricks we performed extraction and transformation on them and loaded the data to MS SQL Server.

Conclusion:

To conclude our ETL report, we initially started by exploring various datasets, but we decided to use Yelp's dataset to determine the factors that affect the businesses when a user leaves a review or how users choose restaurants based on the reviews they have. Then we chose the income unemployment data from U.S. Bureau of Labor Statistics and, the income and education data come from United States Census Bureau. After the process of extraction, transformation, and loading, we prepared visualizations to answer our questions and queries described in the introduction section of this report. The answers will be discussed more in our detailed project report.