ETL-Project

Technical Report

github.com/thierzon/etl-project

Team Two-and-a-half-beards

Oscar-Geare
Spyrothebassist
Thierzon

13 March 2021

Introduction

This report will describe the process that Team Two-and-a-half-beards followed to provide the data required to answer the following question:

How did insults in tweets by POTUS Donald Trump influence his approval ratings?

The purpose of the project is to provide the data in a structured format so that an app developer can use this as a basis to create an app to answer the above question.

The following two datasets have been sourced:

- Trump Twitter insults: https://www.kaggle.com/ayushggarg/all-trumps-twitter-insults-20152021
- Trump approval ratings: https://data.world/fivethirtyeight/trump-approval-ratings

The datasets have gone through a data cleanup and analysis process consisting of the following phases:

- **E**xtract
- **T**ransform
- **L**oad

The process for each of the above phases is described in the subsequent sections.

Extracting the data

The datasets selected for this project each consist of a CSV file. The Pandas library was selected to extract the data from the CSV files using Jupyter Notebook. The following Python libraries have been used to complete the project:

- OS
- Numpy
- Pandas
- SQLAlchemy → create_engine

The extraction consisted of the following steps:

- Assign file paths to the file location using ‘os.path.join’ to ensure compatibility with all operating systems.
- Load CSV files into Pandas dataframes using the ‘read_csv’ function. The following dataframes have been created:
    - tweets_df
    - approval_df

This resulted in two dataframes as shown below.

tweets_df.head():

	Unnamed: 0	date	target	insult	tweet
0	1	2014-10-09	thomas-frieden	fool	Can you believe this fool, Dr. Thomas Frieden ...
1	2	2014-10-09	thomas-frieden	DOPE	Can you believe this fool, Dr. Thomas Frieden ...
2	3	2015-06-16	politicians	all talk and no action	Big time in U.S. today - MAKE AMERICA GREAT AG...
3	4	2015-06-24	ben-cardin	It's politicians like Cardin that have destroy...	Politician @SenatorCardin didn't like that I s...
4	5	2015-06-24	neil-young	total hypocrite	For the nonbeliever, here is a photo of @Neily...

approval_df.head():

	president	subgroup	modeldate	startdate	enddate	pollster	grade	samplesize	population	weight	...	disapprove	adjusted_approve
0	Donald Trump	All polls	1/20/2021	1/20/2017	1/22/2017	Morning Consult	B/C	1992.0	rv	0.680029	...	37.0	45.686784
1	Donald Trump	All polls	1/20/2021	1/20/2017	1/22/2017	Gallup	B	1500.0	a	0.262323	...	45.0	45.861441
2	Donald Trump	All polls	1/20/2021	1/20/2017	1/24/2017	Ipsos	B-	1632.0	a	0.153481	...	45.2	43.451563
3	Donald Trump	All polls	1/20/2021	1/21/2017	1/23/2017	Gallup	B	1500.0	a	0.242845	...	46.0	45.861441
4	Donald Trump	All polls	1/20/2021	1/22/2017	1/24/2017	Gallup	B	1500.0	a	0.227380	...	45.0	46.861441

	adjusted_disapprove	multiversions	tracking	url	poll_id	question_id	createddate	timestamp
0	38.055805	NaN	NaN	http://static.politico.com/9b/13/82a3baf542ae9...	49249	77261	1/23/2017	11:47:59 20 Jan 2021
1	43.539189	NaN	T	http://www.gallup.com/poll/201617/gallup-daily...	49253	77265	1/23/2017	11:47:59 20 Jan 2021
2	43.780389	NaN	T	http://polling.reuters.com/#poll/CP3_2/	49426	77599	3/1/2017	11:47:59 20 Jan 2021
3	44.539189	NaN	T	http://www.gallup.com/poll/201617/gallup-daily...	49262	77274	1/24/2017	11:47:59 20 Jan 2021
4	43.539189	NaN	T	http://www.gallup.com/poll/201617/gallup-daily...	49236	77248	1/25/2017	11:47:59 20 Jan 2021

Transforming the data

The Pandas dataframes created during the extraction phase have been further transformed. In the transformation process the datasets are cleaned and reduced to the relevant data.

The date columns in the two dataframes will ultimately be used to link the two datasets and therefore the date format needs to match. The other columns are included as there is potential for insight and further investigation.

Cleaning the data, with the help of the Numpy library, indicated empty cells to be contained in the columns ‘target’ and ‘url’. Further action was not required, however, as these columns are not a priority in this investigation and will in no way hamper the process.

For the ***tweets_df*** the following transformation steps have been performed:

- Removing tweets before the inauguration date of 20 January 2017 using the Pandas ‘drop’ function.
- Extract columns ‘date’, ‘target’ and ‘insult’ to new dataframe ***tweetdates_df*** using the ‘copy’ function (see below).

tweetdates_df.head()

	date	target	insult
2357	2017-01-25	cnn	FAKE NEWS
2358	2017-01-25	chicago	If Chicago doesn't fix the horrible carnage go...
2359	2017-01-26	chelsea-manning	Ungrateful TRAITOR
2360	2017-01-26	chelsea-manning	should never have been released from prison
2361	2017-01-26	chelsea-manning	Terrible!

For the ease of analysis outside of the scope of this report, the number of insults per date have been extracted. The ***tweetsdates_df*** was transformed using the ‘groupby’ function grouping on the column ‘date’ and using the ‘count’ function. This resulted in ***grouped_tweets*** as shown below.

grouped_tweets.head()

	count
date	
2017-01-25	2
2017-01-26	4
2017-01-28	8
2017-01-29	5
2017-01-31	6

For the ***approval_df*** the following transformation steps have been performed:

- Extract columns ‘startdate’, ‘enddate’, ‘pollster’, ‘samplesize’, ’adjusted_approve’, ‘adjusted_disapprove’ and ‘url’ to new dataframe ***approvedfinal_df*** using the ‘copy’ function (see below).

approvedfinal_df.head()

	startdate	enddate	pollster	samplesize	adjusted_approve	adjusted_disapprove	url
0	1/20/2017	1/22/2017	Morning Consult	1992.0	45.686784	38.055805	http://static.politico.com/9b/13/82a3baf542ae9...
1	1/20/2017	1/22/2017	Gallup	1500.0	45.861441	43.539189	http://www.gallup.com/poll/201617/gallup-daily...
2	1/20/2017	1/24/2017	Ipsos	1632.0	43.451563	43.780389	http://polling.reuters.com/#poll/CP3_2/
3	1/21/2017	1/23/2017	Gallup	1500.0	45.861441	44.539189	http://www.gallup.com/poll/201617/gallup-daily...
4	1/22/2017	1/24/2017	Gallup	1500.0	46.861441	43.539189	http://www.gallup.com/poll/201617/gallup-daily...

- The ‘startdate’ and ‘enddate’ columns in the ***approvedfinal_df*** are in US format and have been converted using the Pandas ‘to_datetime’ function to match the date format in ***tweetdates_df*** (see below).

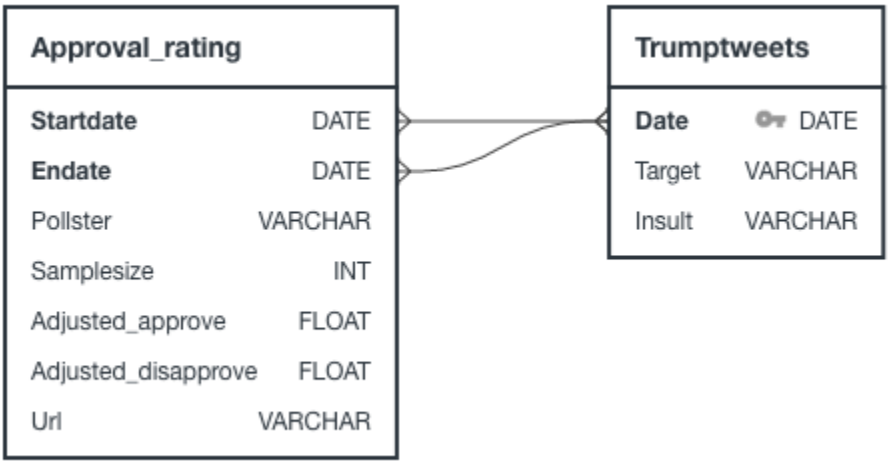
approvedfinal_df.head()

	startdate	enddate	pollster	samplesize	adjusted_approve	adjusted_disapprove	url
0	2017-01-20	2017-01-22	Morning Consult	1992.0	45.686784	38.055805	http://static.politico.com/9b/13/82a3baf542ae9...
1	2017-01-20	2017-01-22	Gallup	1500.0	45.861441	43.539189	http://www.gallup.com/poll/201617/gallup-daily...
2	2017-01-20	2017-01-24	Ipsos	1632.0	43.451563	43.780389	http://polling.reuters.com/#poll/CP3_2/
3	2017-01-21	2017-01-23	Gallup	1500.0	45.861441	44.539189	http://www.gallup.com/poll/201617/gallup-daily...
4	2017-01-22	2017-01-24	Gallup	1500.0	46.861441	43.539189	http://www.gallup.com/poll/201617/gallup-daily...

Ultimately, the app developer needs to take the next step in amending the ‘startdate’ and ‘enddate’ columns in the approval dataframe as per their requirements.

Loading the data

In the loading phase the two dataframes are loaded into a relational database. To support the app developer a ERD was prepared to provide the structure of the relational database that will be used (see below).



The relational database has been set up in PostgreSQL. The following steps have been performed to load the ***tweetsdates_df*** and ***approvedfinal_df*** dataframes into the database:

- Setup the tables in the SQL database as per ERD prepared using the SQL query function (refer to queries.sql).
- Connect to the database using SQLALchemy create_engine.
- Confirm tables have been created in the database.
- Load dataframes into the database using the Pandas function ‘to_sql’.
- Verify that the data has been successfully loaded into the dataframe using the function ‘read_sql_query’.

Upon completing the last step, the dataframes have been successfully loaded into the SQL database and ready to be used by the app developer.