IST 782 Portfolio Milestone

M.S. Applied Data Science Syracuse University

Table of Contents

I. Introduction

Introduce Myself

Educational background

- B.B.A. in Operations and Supply Chain Management
- University of Wisconsin Whitewater, December 2017

Work experience

Position	Company	Location	Time
Solution Consultant	Smith & Nephew	Remote	Aug 2021 - Present
Global Demand Planner	Smith & Nephew	Remote	Nov 2020 – Aug 2021
Supply Chain Planner	Schreiber Foods	Green Bay, WI	Nov 2019 – Nov 2020
Global Demand Planner	Smith & Nephew	Austin, TX	Apr 2018 – Nov 2019
Supply Chain Intern	Mahindra USA	Houston, TX	May 2017 – Aug 2017

Professional accomplishments

- 4x dean’s list academic excellence award
- Co-treasurer and board member of campus APICS club
- 2nd place CSCMP supply chain case study competition
- APICS certified in production and inventory management (CPIM)
- Microsoft Excel Expert certification
- Amazon Warehouse Services (AWS) certified cloud practitioner
- Rapid Response certified contributor level 1
- Rapid Response certified author level 1

Skills

- R
- SQL
- Python
- Data Visualization
- Machine Learning
- Data Science
- Microsoft Excel
- Tableau
- Critical Thinking
- Team Leadership
- Public Speaking
- Analytic Problem Solving
- Alteryx
- Statistical Forecasting
- Supply Chain Analytics
- Amazon Web Services (AWS)
- Web Scraping

Why Data Science

The supply chain industry needs data scientists

- Statistical forecasting
- Inventory optimization
- Complex supply networks
- Manufacturing automation
- Risk management
- Real time data processing
- Predictive analytics

Data science is a growing field

- Per LinkedIn, there has been a 650% increase in data science jobs since 2012
- IBM says the demand for data scientists will continue to be strong for years
- The U.S. Bureau of Labor Statistics expects 11.5 Mil new data science jobs through 2026
- In 2020, data scientist was listed as the third best job in America according to Glassdoor

Flextronics Pulse Center, Silicon Valley

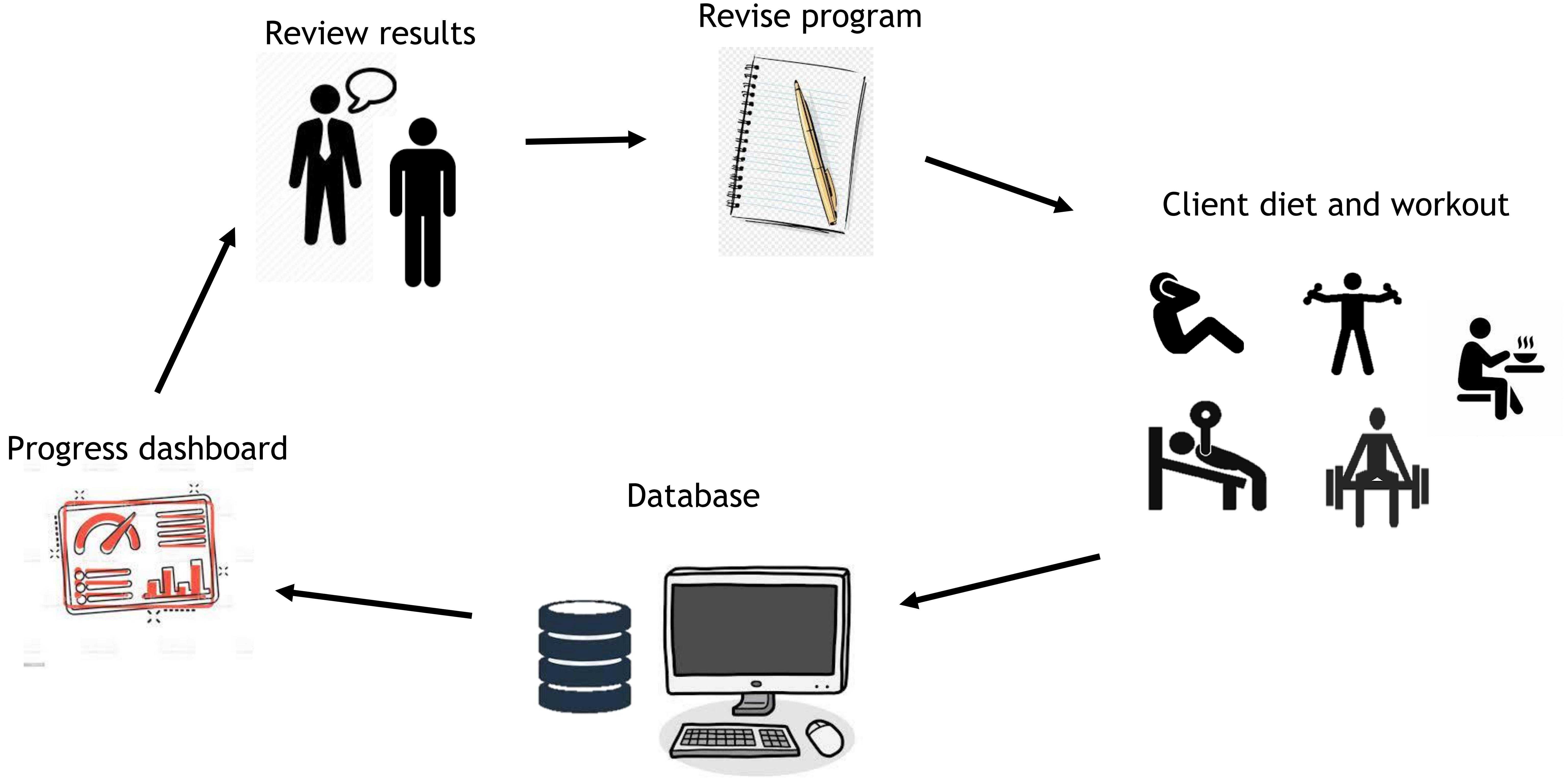
Statistics cited from https://towardsdatascience.com/is-data-science-still-a-rising-career-in-2021-722281f7074c

II. Project 1

Health and fitness database

Project Introduction

Goal: create a database for storing fitness related data such as exercise history and nutrition logs and use it to enhance the client experience



Review results

Revise program

Client diet and workout

Database

Progress dashboard

Database Development

CLIENT:A personal training client who is taking part in a weightlifting program and is using or did use the database for keeping track of their progression.

MEASUREMENT: a measurement taken by a client at a particular point in time

MUSCLE GROUP: a body part that is affected by lifting and can be measured

LIFT: a movement performed in a gym during a workout, that typically includes resistance, and impacts one or more muscle groups.

SET: a collection of repetitions

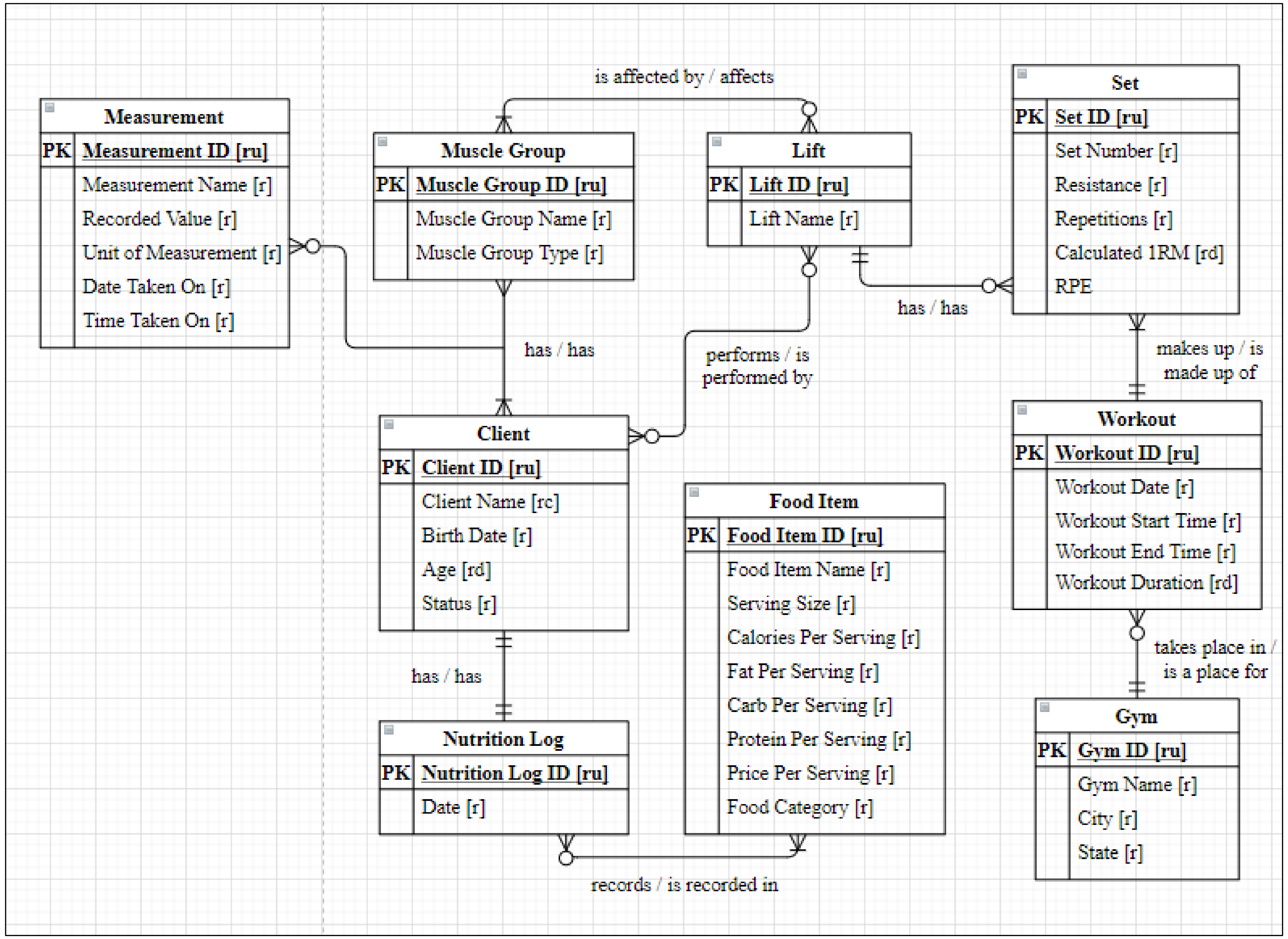
WORKOUT: a collection of sets

GYM: a facility that has the necessary equipment that can be used by a client to perform lifts

NUTRITION LOG: a diary that contains records of food or beverage items that are consumed

FOOD ITEM: an article of food or beverage that has calories and that is entered into the nutrition log

Entity Relation Diagram



Measurement
PK **Measurement ID [ru]**
Measurement Name [r]
Recorded Value [r]
Unit of Measurement [r]
Date Taken On [r]
Time Taken On [r]

Muscle Group
PK **Muscle Group ID [ru]**
Muscle Group Name [r]
Muscle Group Type [r]

Lift
PK **Lift ID [ru]**
Lift Name [r]

Set
PK **Set ID [ru]**
Set Number [r]
Resistance [r]
Repetitions [r]
Calculated 1RM [rd]
RPE

Client
PK **Client ID [ru]**
Client Name [rc]
Birth Date [r]
Age [rd]
Status [r]

Workout
PK **Workout ID [ru]**
Workout Date [r]
Workout Start Time [r]
Workout End Time [r]
Workout Duration [rd]

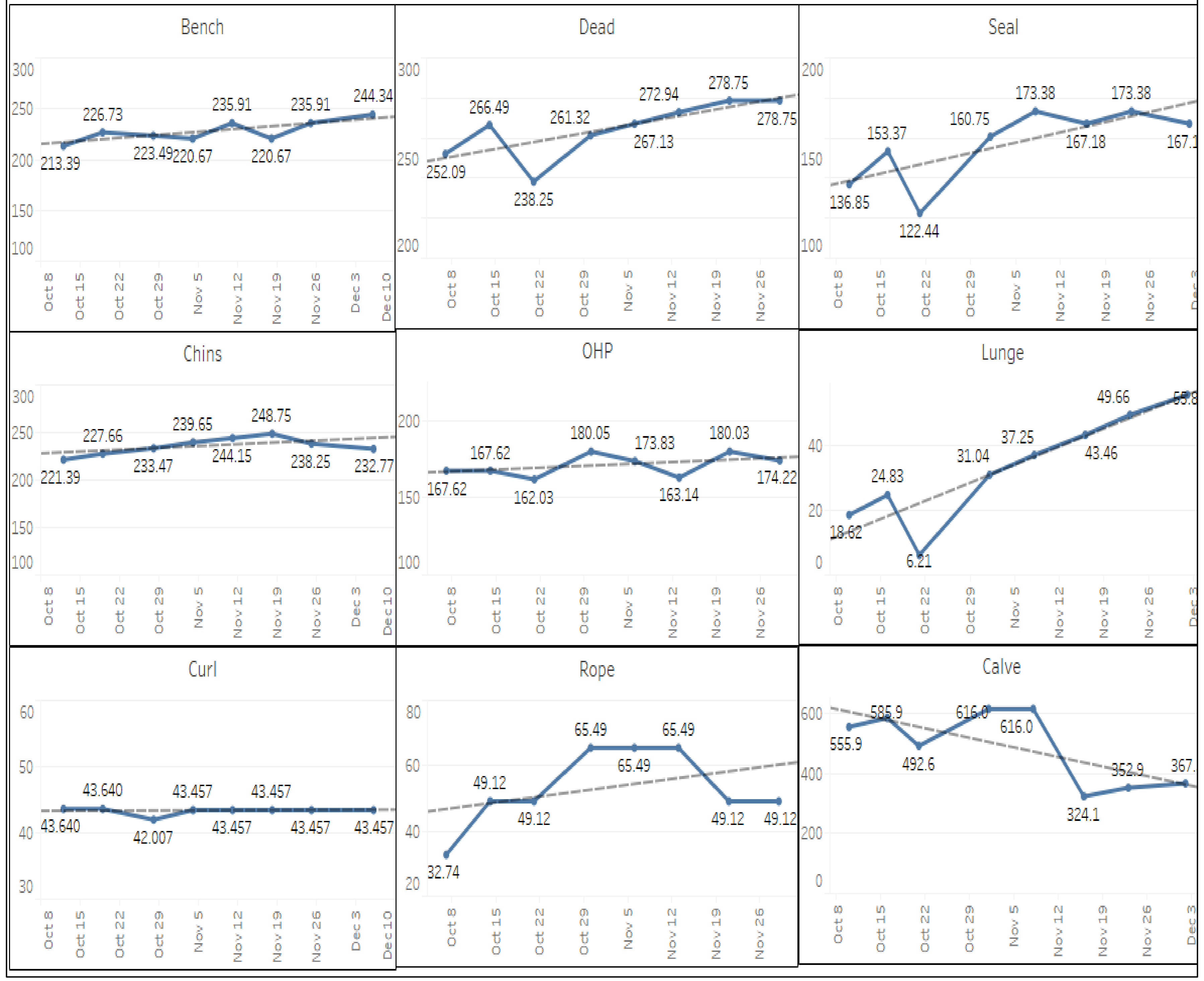
Food Item
PK **Food Item ID [ru]**
Food Item Name [r]
Serving Size [r]
Calories Per Serving [r]
Fat Per Serving [r]
Carb Per Serving [r]
Protein Per Serving [r]
Price Per Serving [r]
Food Category [r]

Nutrition Log
PK **Nutrition Log ID [ru]**
Date [r]

Gym
PK **Gym ID [ru]**
Gym Name [r]
City [r]
State [r]

Progress Dashboard

Estimated 1 RM Over Time Per Lift



Starting Weight
as of: 10/5/2020

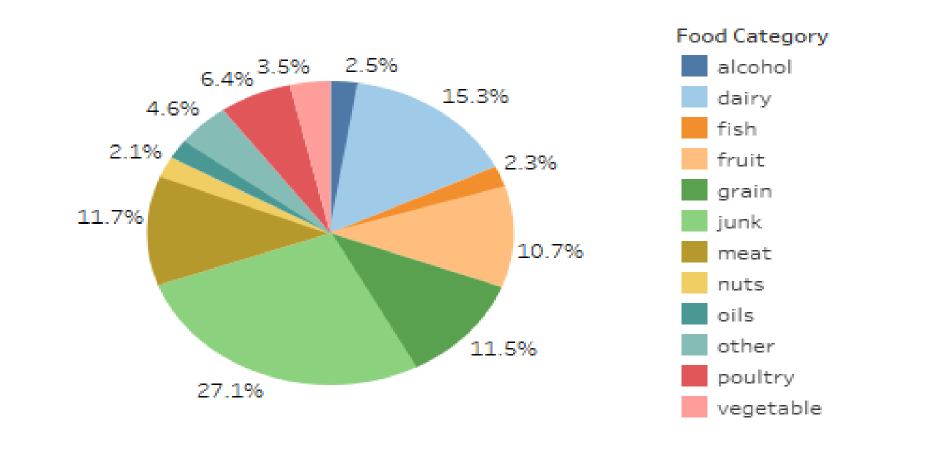
167.9

Current Weight
as of: 12/7/2020

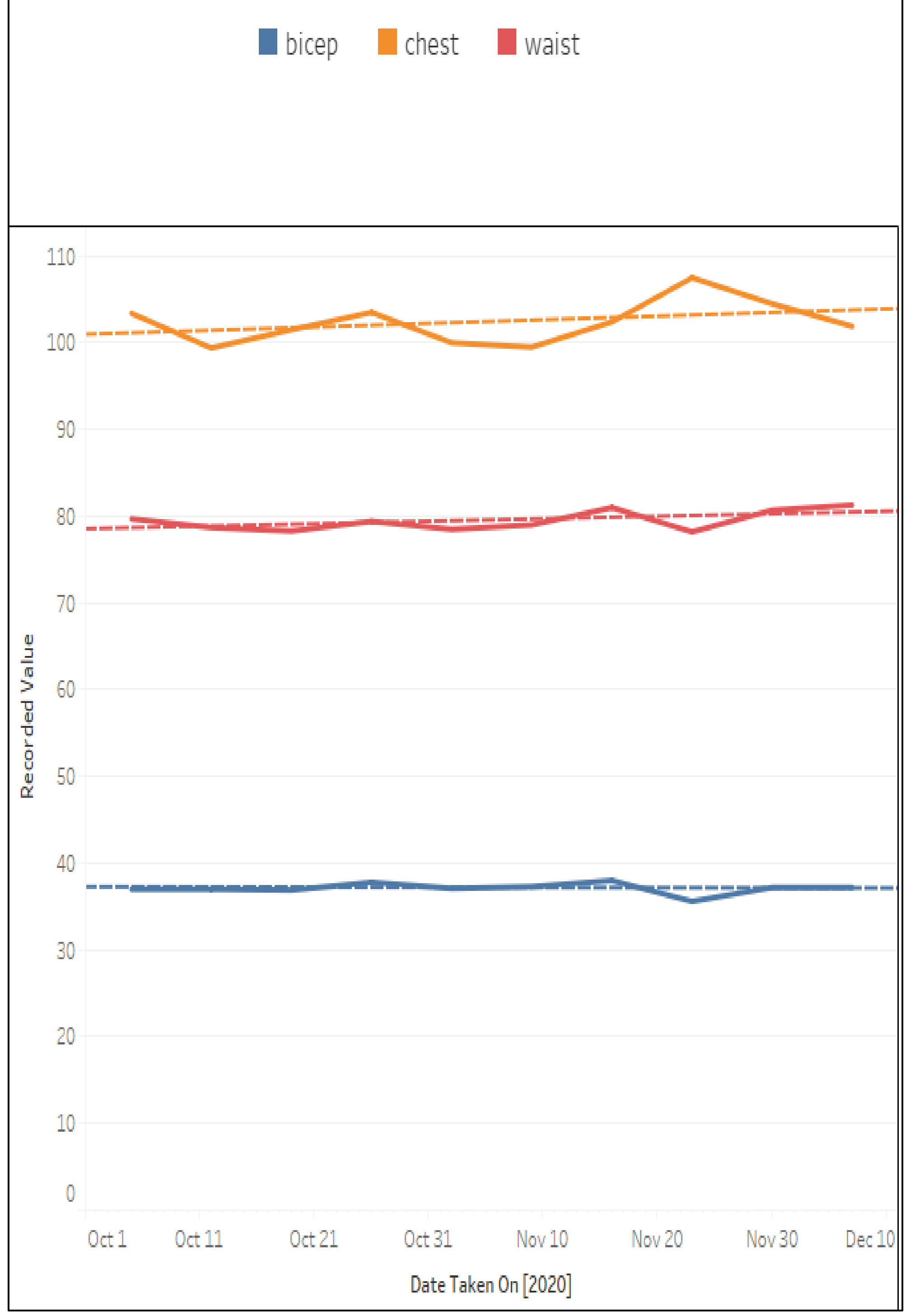
177.0

Week over Week Change

Date Taken On	Recorded Value	% Difference
10/5/2020	167.9	
10/12/2020	166.4	-0.9%
10/19/2020	170.0	2.2%
10/26/2020	169.0	-0.6%
11/2/2020	167.7	-0.8%
11/9/2020	172.1	2.6%
11/16/2020	178.5	3.7%
11/23/2020	173.2	-3.0%
11/30/2020	180.7	4.3%
12/7/2020	177.0	-2.0%



Measurement Recordings



III. Project 2

Machine Learning with Song Data

Part 1: Predicting Song Popularity

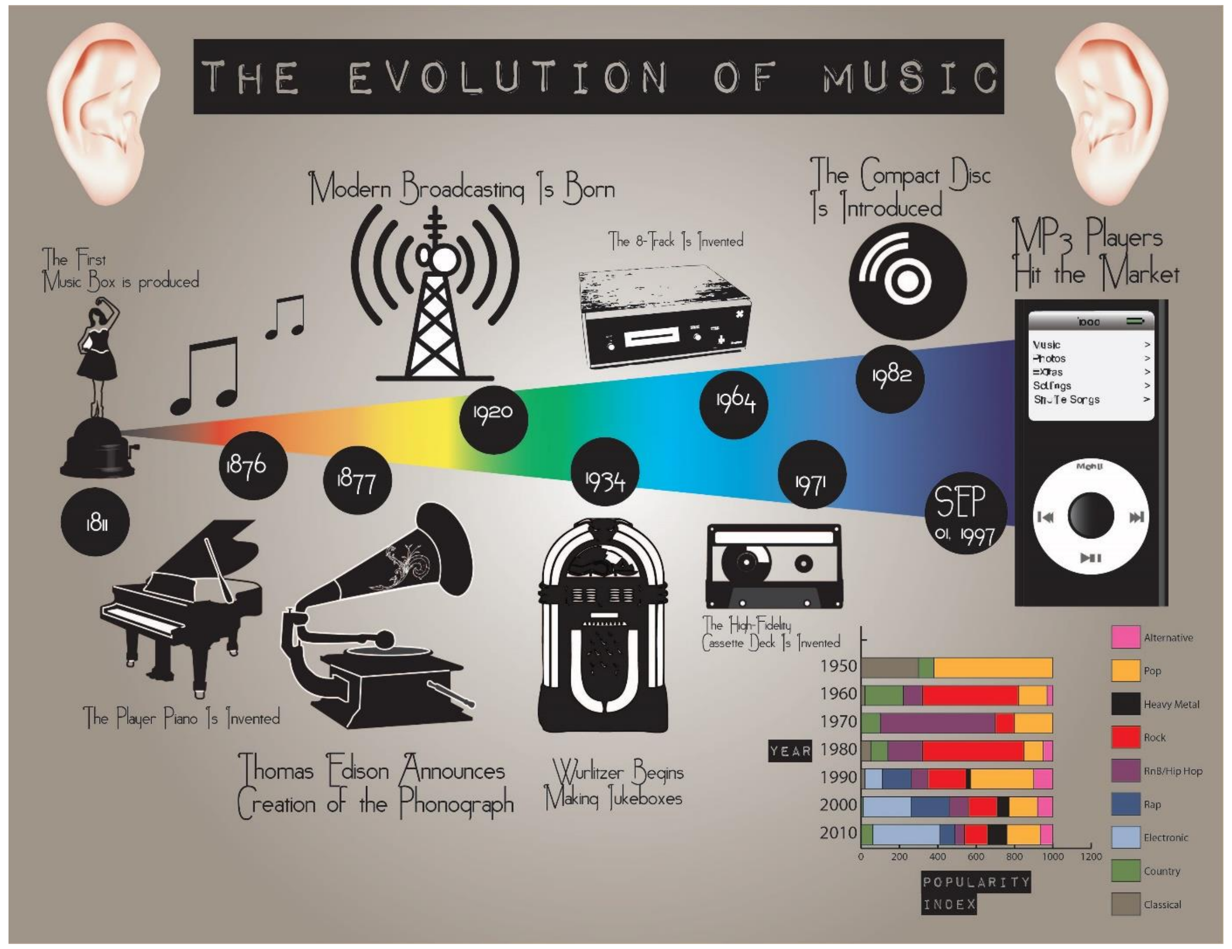
Goal: predict the popularity of songs based on audio features and other meta data

<u>Real world application</u>

- Data science is used in developing songs
- How do record labels produce hit song after hit song?
- They have figured out the "formula"

<u>Questions to answer</u>

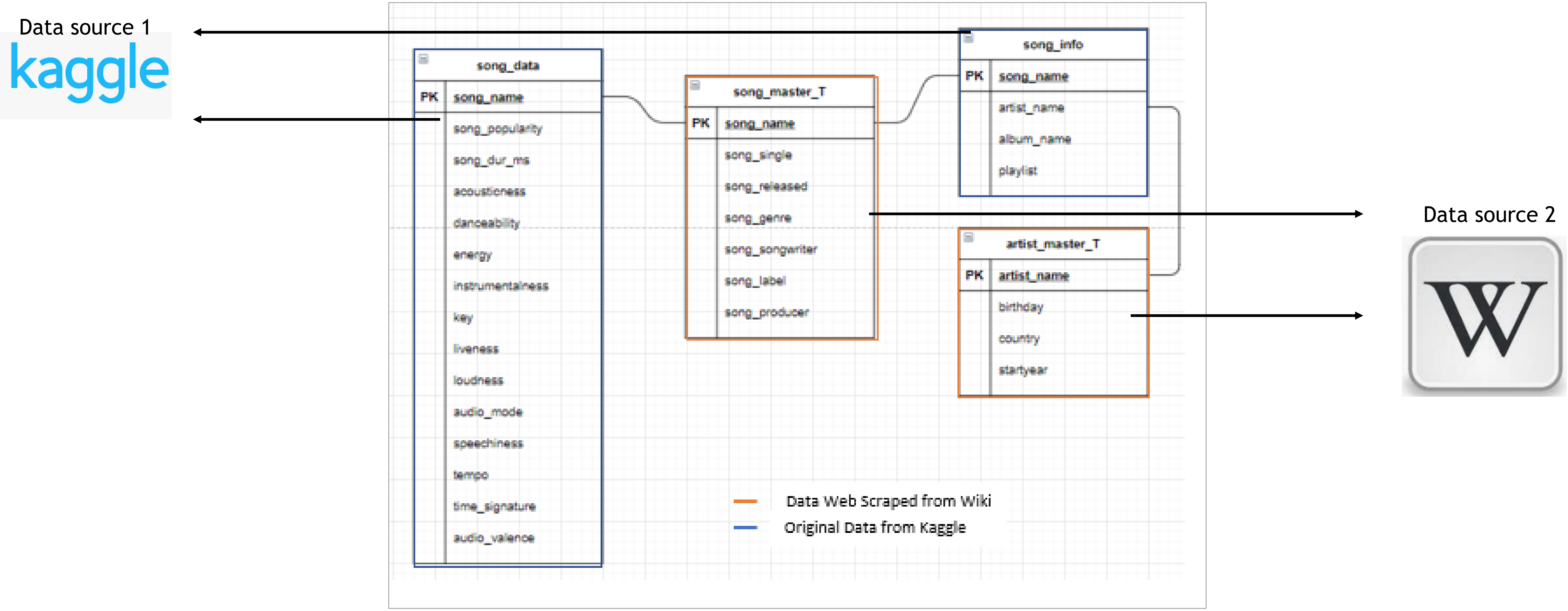
- How have popular songs change overtime?
- Are there certain attributes that correlate with popular songs?
- Can the popularity of a song be predicted based on its attributes?


THE EVOLUTION OF MUSIC
The First Music Box is produced
1811
1876
1877
The Player Piano Is Invented
Thomas Edison Announces Creation of the Phonograph
Modern Broadcasting Is Born
1920
1934
Wurlitzer Begins Making Jukeboxes
The 8-Track Is Invented
1964
1971
The High-Fidelity Cassette Deck Is Invented
The Compact Disc Is Introduced
1982
SEP 01, 1997
MP3 Players Hit the Market
YEAR
1950
1960
1970
1980
1990
2000
2010
0
200
400
600
800
1000
1200
POPULARITY INDEX
Alternative
Pop
Heavy Metal
Rock
RnB/Hip Hop
Rap
Electronic
Country
Classical

Data Collection

Original dataset from Kaggle and then collected additional metadata through web scraping Wikipedia

Database diagram



	song_data
PK	song_name
	song_popularity
	song_dur_ms
	acousticness
	danceability
	energy
	instrumentalness
	key
	liveness
	loudness
	audio_mode
	speechiness
	tempo
	time_signature
	audio_valence

	song_master_T
PK	song_name
	song_single
	song_released
	song_genre
	song_songwriter
	song_label
	song_producer

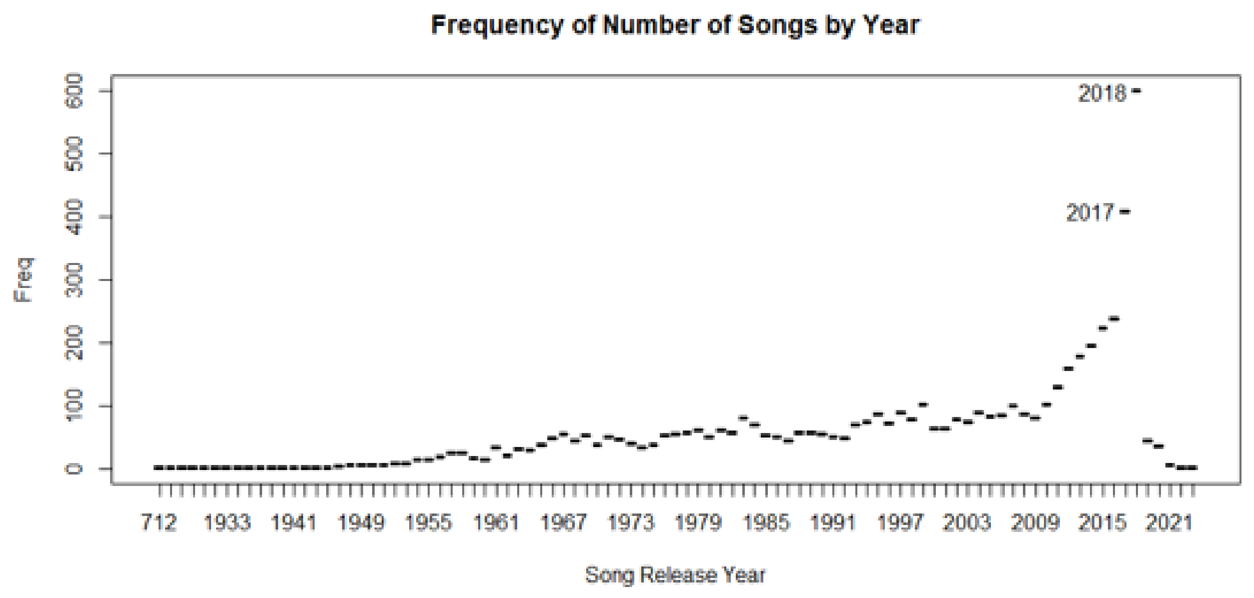
	song_info
PK	song_name
	artist_name
	album_name
	playlist

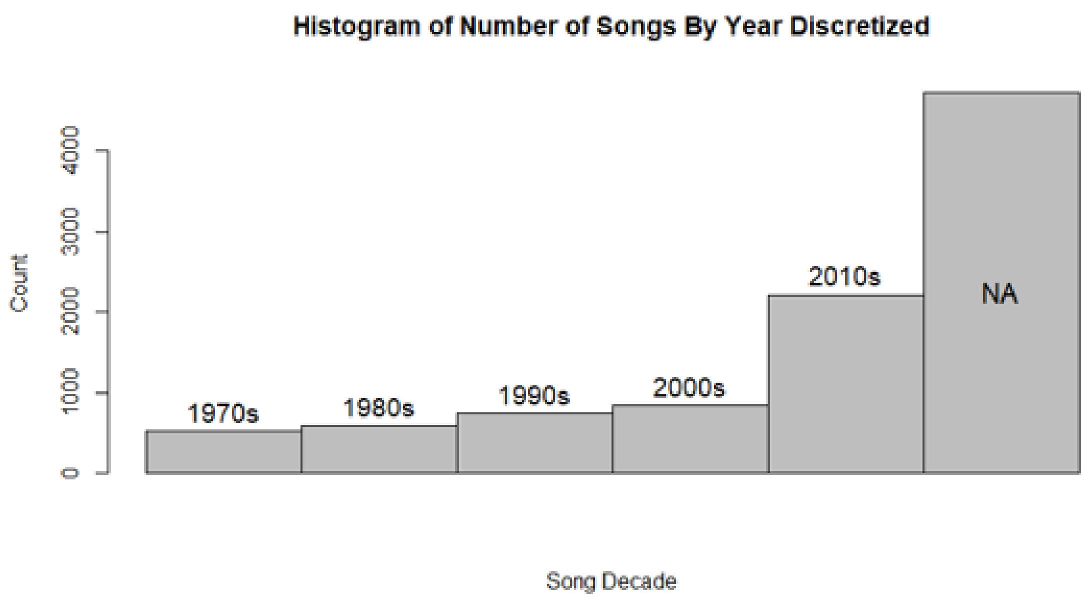
	artist_master_T
PK	artist_name
	birthday
	country
	startyear

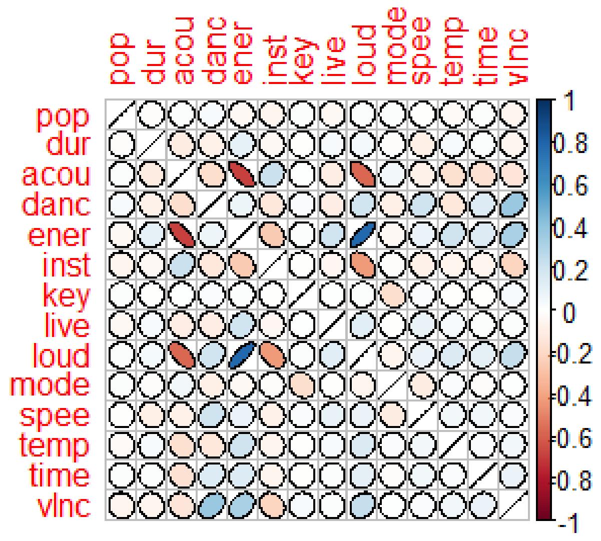
— Data Web Scraped from Wiki

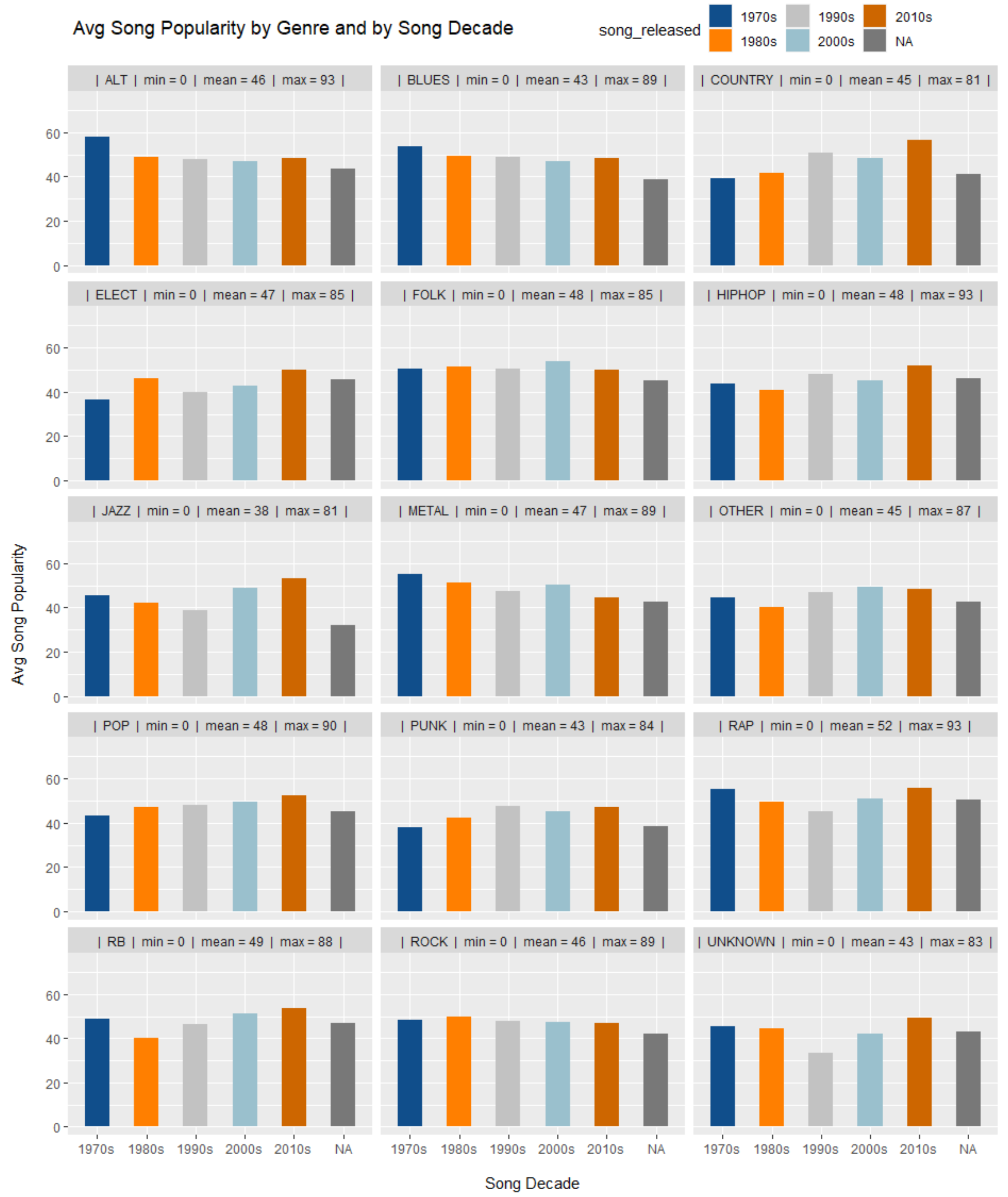
— Original Data from Kaggle

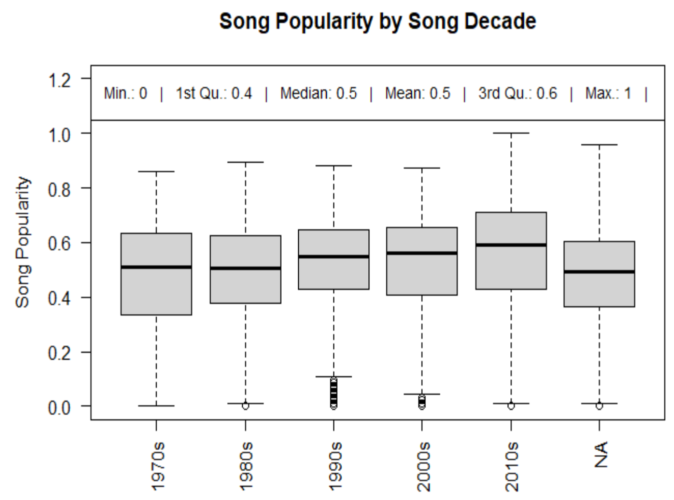
Exploratory Data Analysis

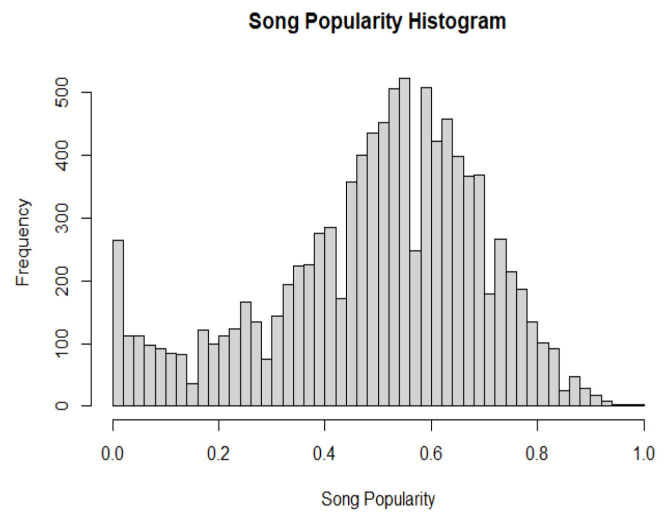


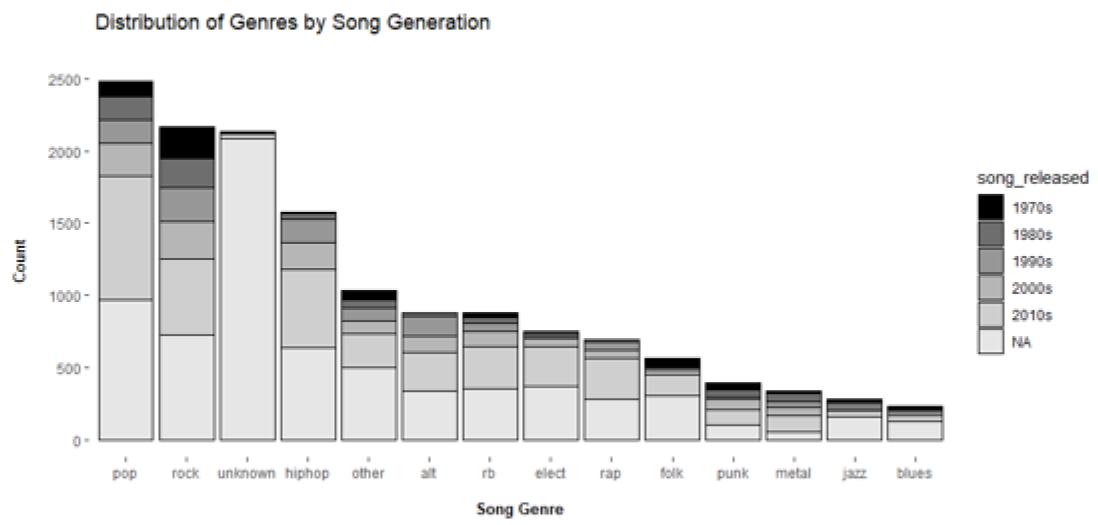


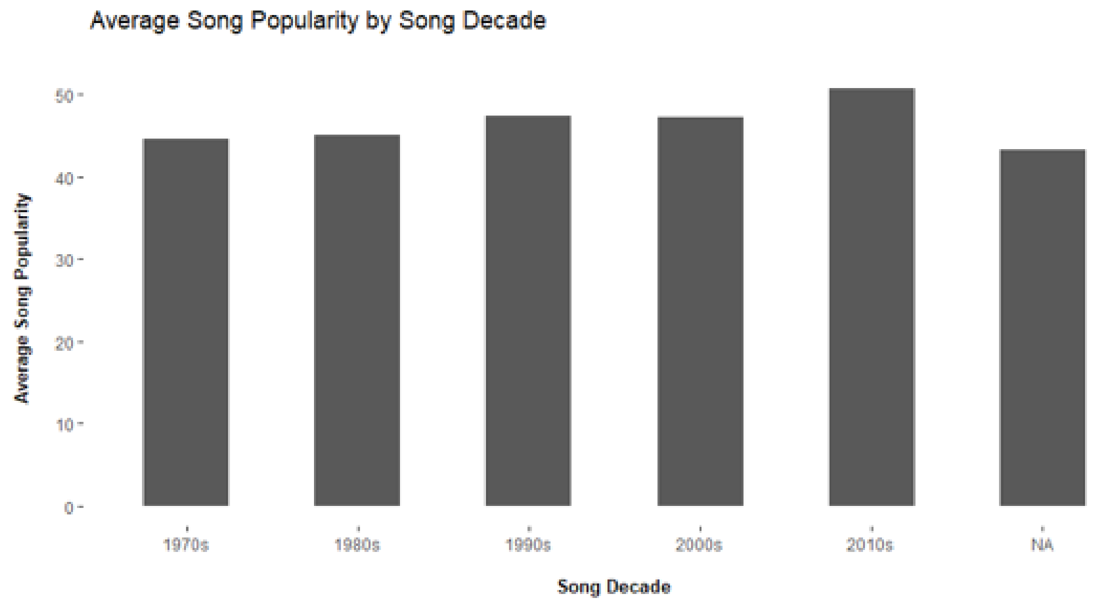


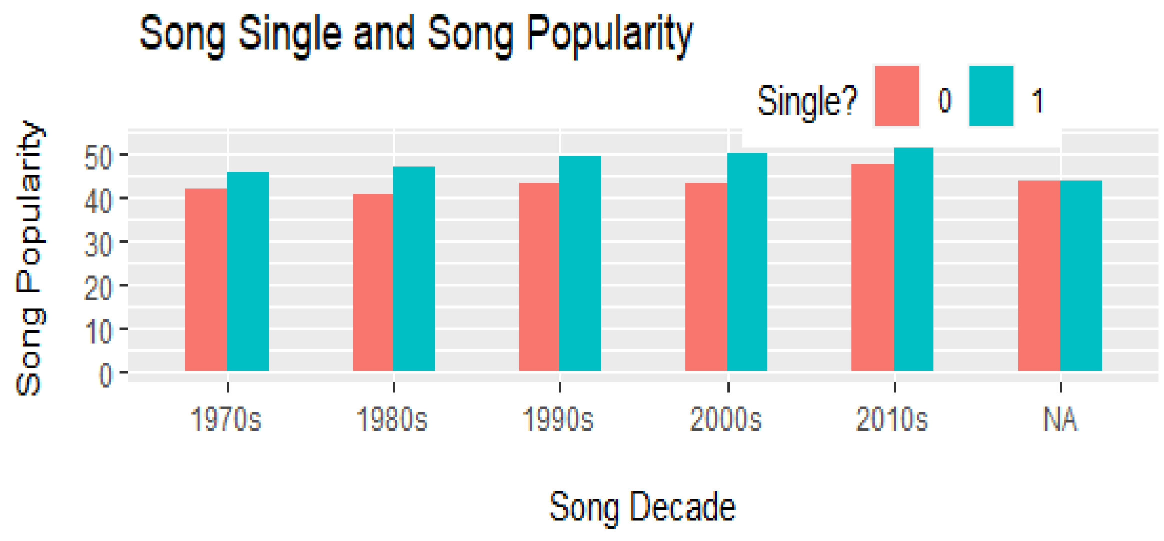












Data Modeling

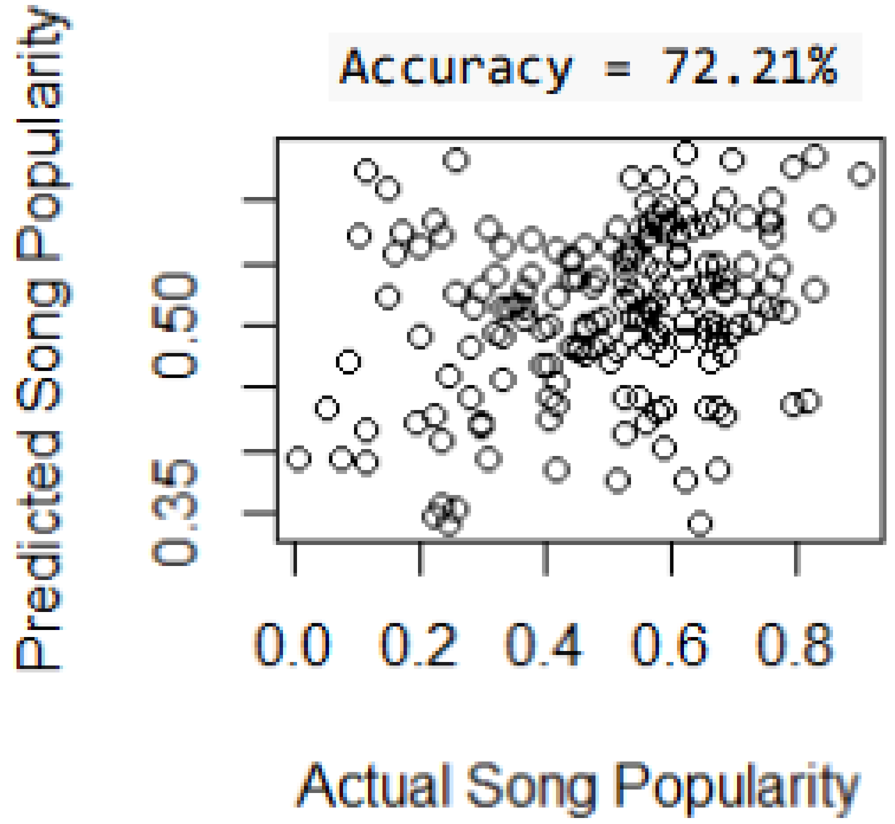
Type of problem: supervised regression

Song Audio features

Other Song Metadata

Random forest model

Results



Part 2: Creating Song Clusters

Goal: create clusters of songs based on their attributes

Real world application

- Music streaming services use machine learning to make recommendations
- Auto generating playlists
- Discovering new music

Questions to answer

- Is it possible to create song clusters based on their attributes?
- How do the clusters compare to genres?

Data Collection

Sample random song from Spotify API

- Over 12,000 “pseudo” random songs via Spotify API
- There is no true random method for doing this
- Songs dating back from 1970 to present

Get audio features from Spotify API

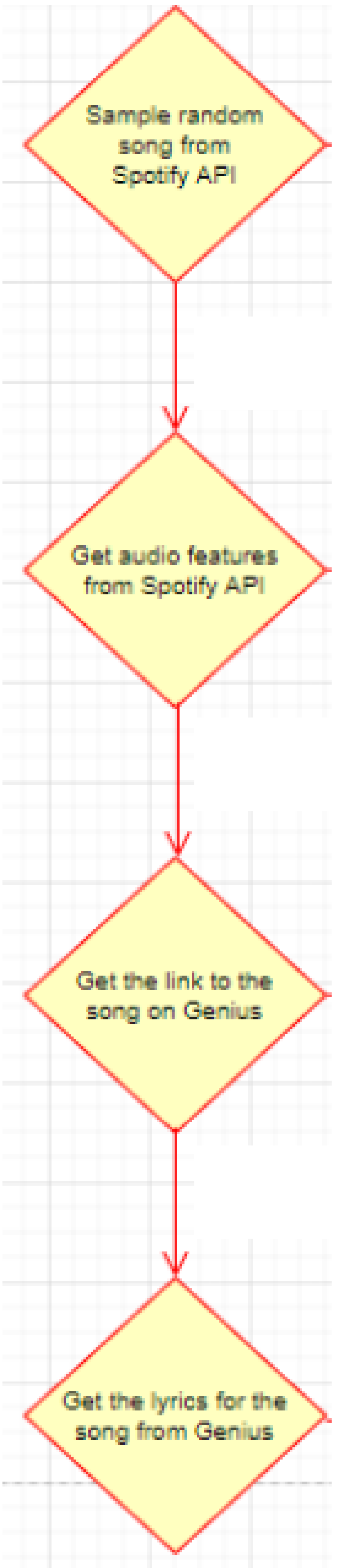
- Same audio features from part 1

Get the link to the song on Genius

- Find the correct link by web scraping a google search
- Conduct search for [Song Name] + “Genius.com”
- If there is one it appears in the top results

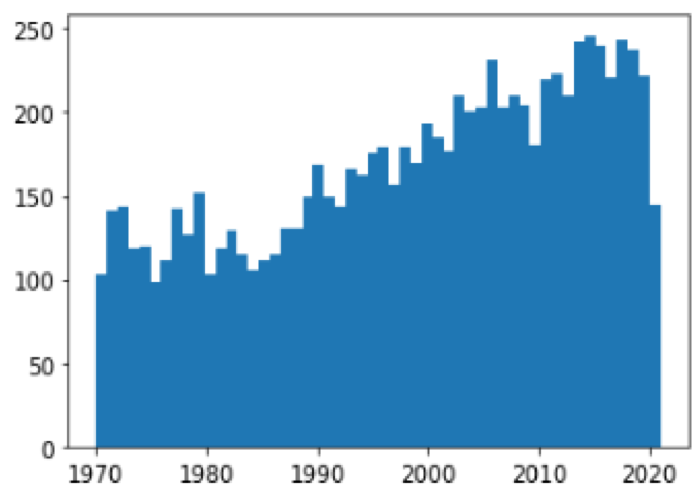
Get the lyrics for the song from Genius

- Web scrape the HTML code for the genius page
- Parse it to extract the text of the song lyrics

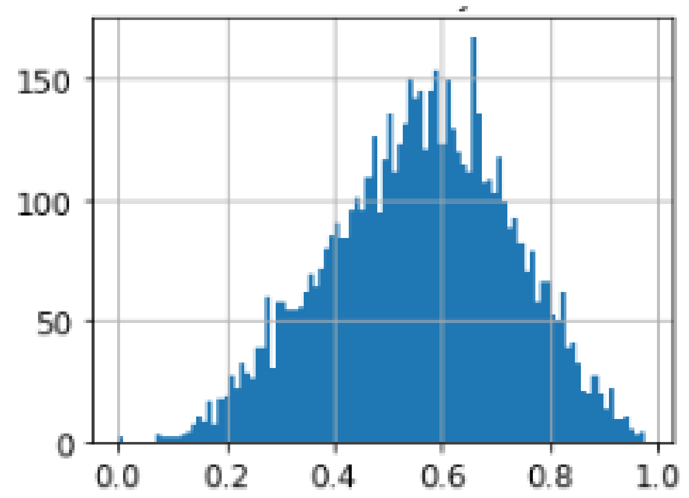

Sample random song from Spotify API
Get audio features from Spotify API
Get the link to the song on Genius
Get the lyrics for the song from Genius

Exploratory Data Analysis

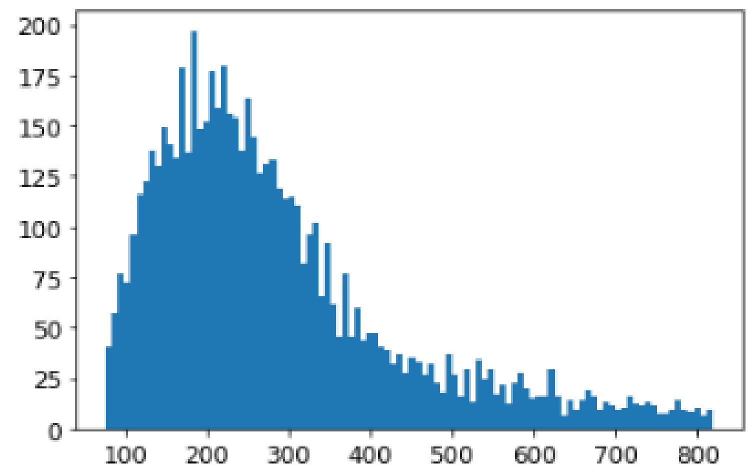
Histogram of song years



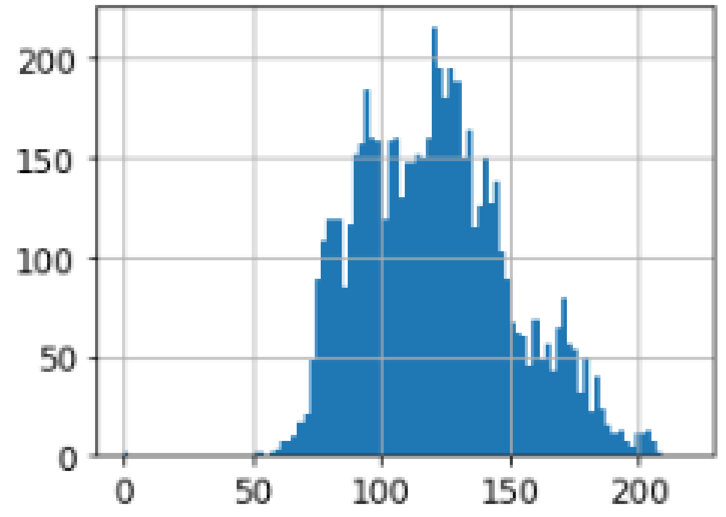
Histogram of song danceability



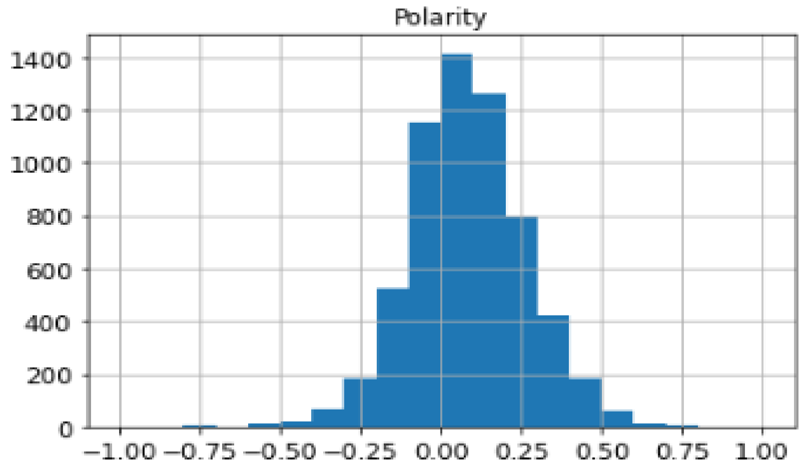
Histogram of song length



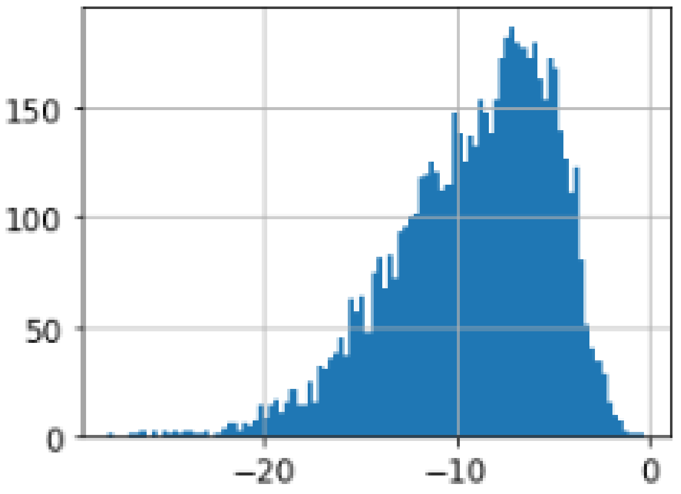
Histogram of song tempo



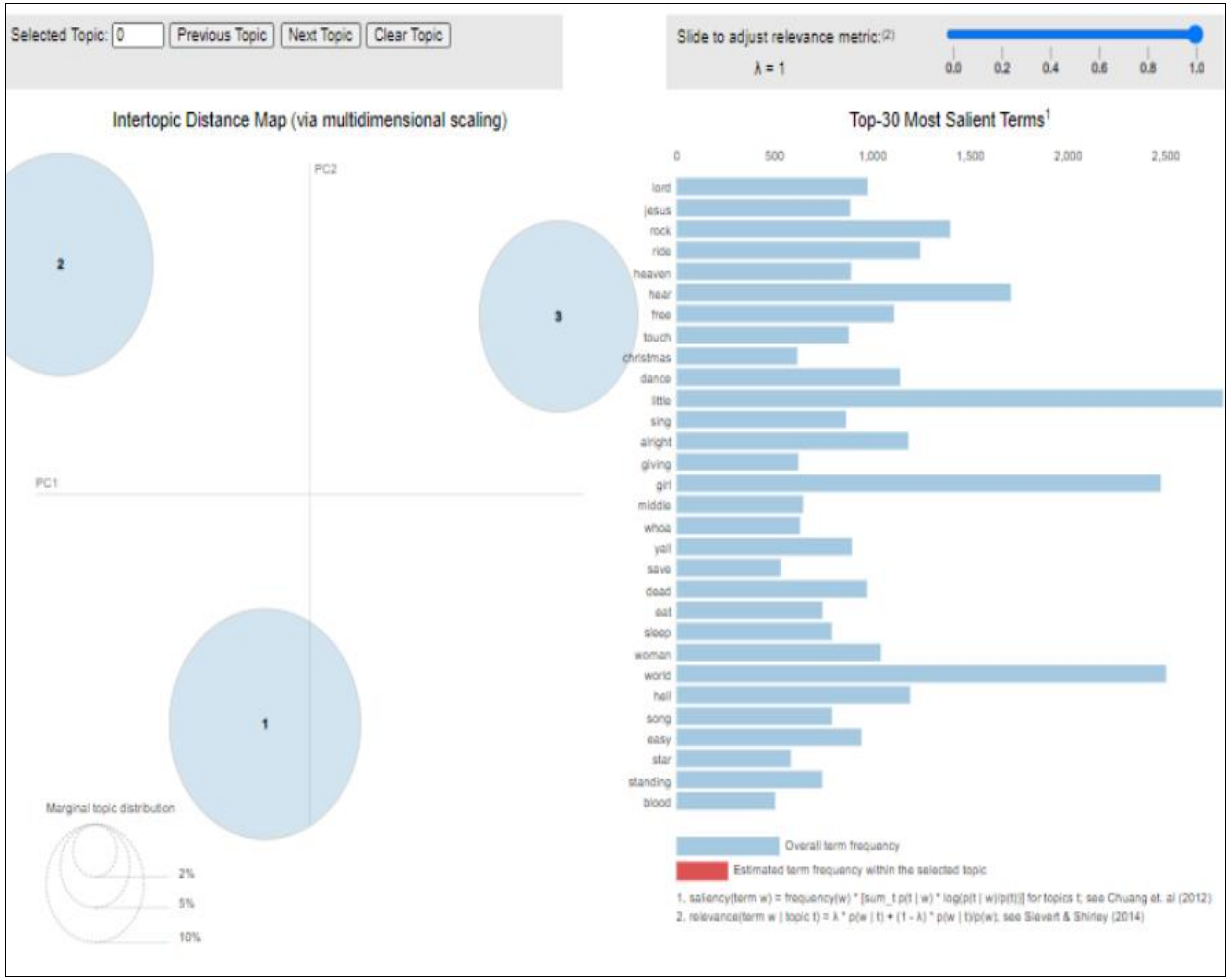
Histogram of polarity scores



Histogram of song loudness



LDA Topic Modeling Results



▨ Overall term frequency

🟥 Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Data Modeling

Type of problem: unsupervised clustering

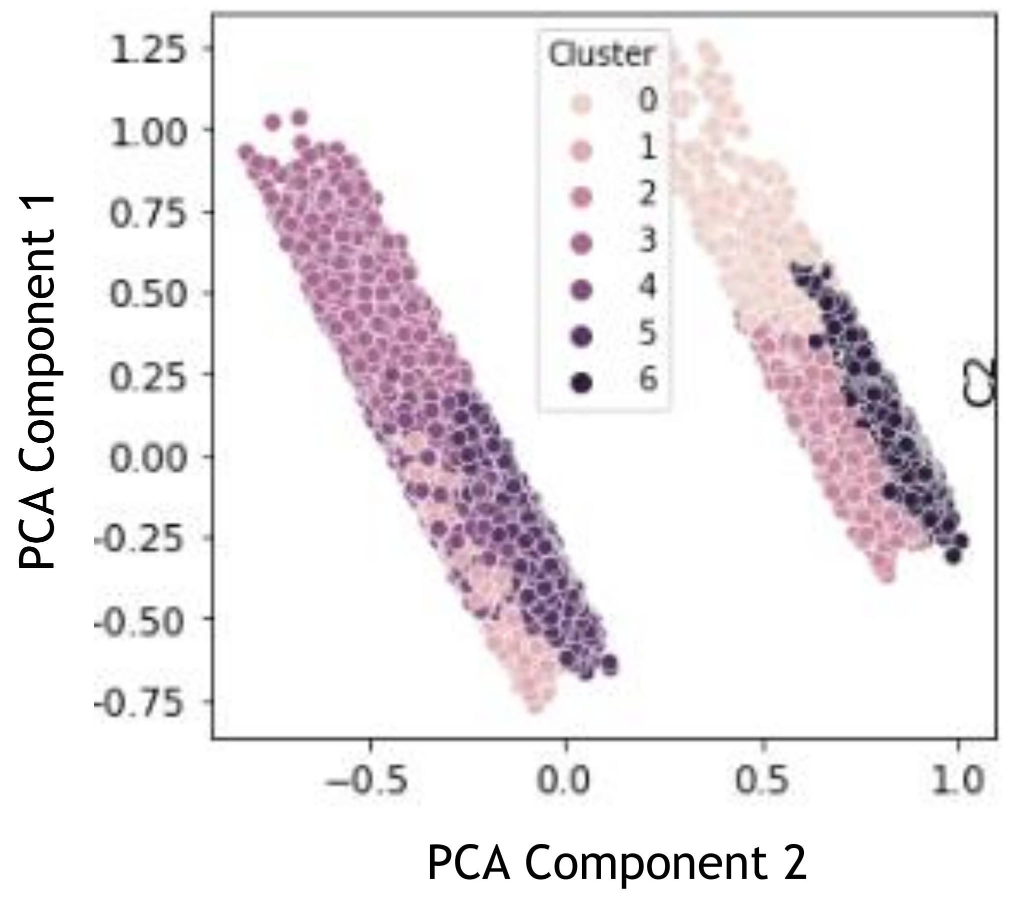
How the song sounds (audio features)

What the song is about (LDA topic modeling)

Are the song lyrics positive or negative?

K means clustering algorithm

Song clusters k = 7



IV. Project 3

Cryptocurrency Price Prediction

Data Collection

Asset table

- Collected by web scraping Coinbase asset directory page
- This page has a list of all tradeable cryptocurrencies on Coinbase
- Currently about 165 cryptocurrencies that can be traded

Price table

- Collected through the yfinance Python package
- Historical price and volume data by day for all data available on Yahoo Finance

CoinMarket table

- Collected through the CoinMarketCap API
- Daily snapshots of circulating supply, max supply, and coin market cap rank
- Cannot get historical data can only take the daily snapshots

Youtube table

- Collected through Google Cloud / Youtube API
- Retrieve a count of published Youtube videos for crypto slugs by day
- Capped at 1000 per day (100 for 10 API keys), but can get historical data
- Starting from the top down and building the history up overtime

Twitter table

- Collected through Twitter API
- Retrieve a count of Tweets that contain hashtag of the ticker by day
- Cannot get historical data without elevated access. Snapshot taken every day

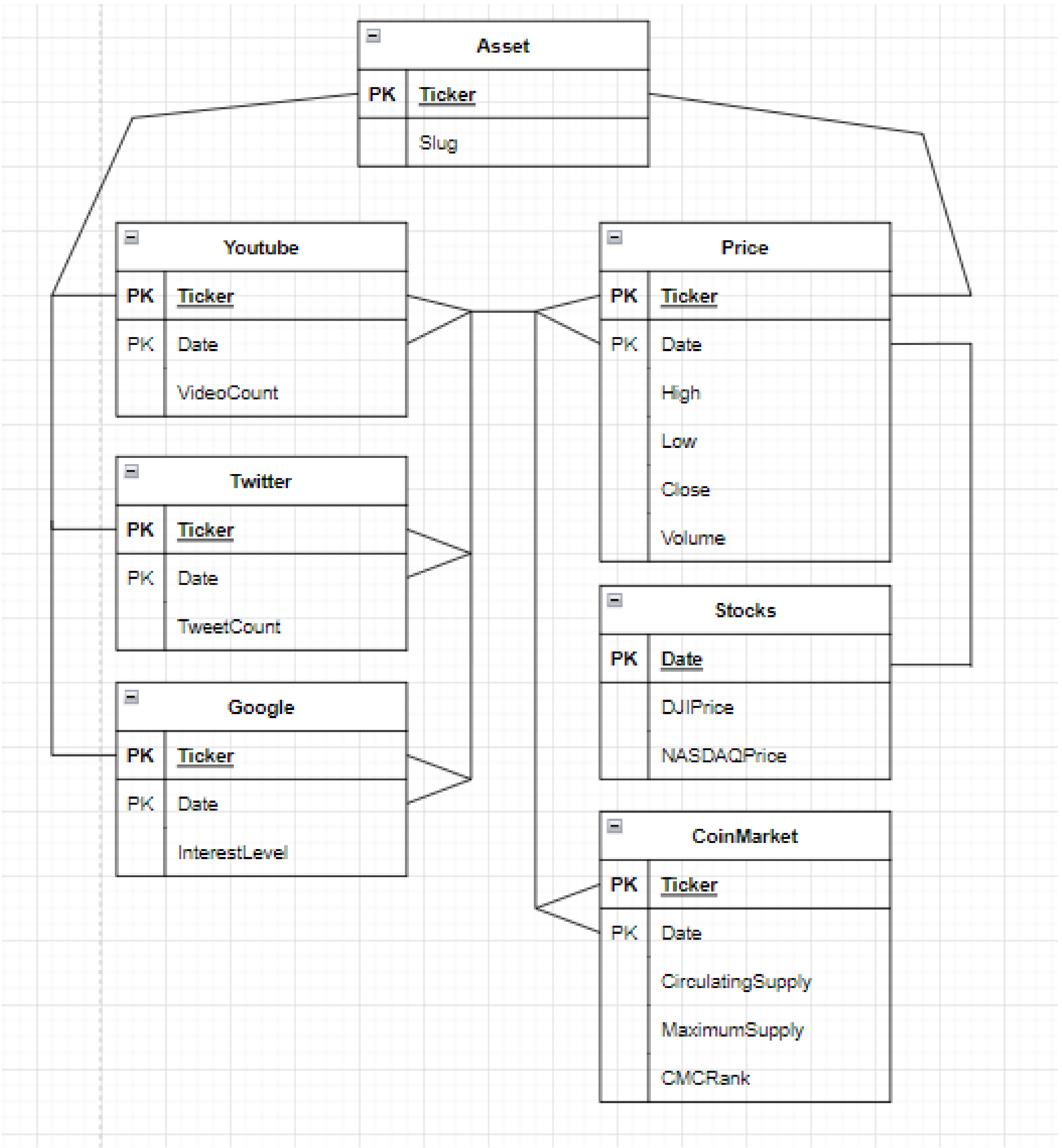
Stocks table

- Collected through yfinance Python package
- Historical closing prices for all data available on Yahoo Finance

Google table

- Collected through pytrends python package
- Retrieve data from Google Trends about cryptocurrency searches by day
- Might need to use some extrapolation methods due to data availability

Database diagram



Asset	
PK	Ticker
	Slug

Youtube	
PK	Ticker
PK	Date
	VideoCount

Twitter	
PK	Ticker
PK	Date
	TweetCount

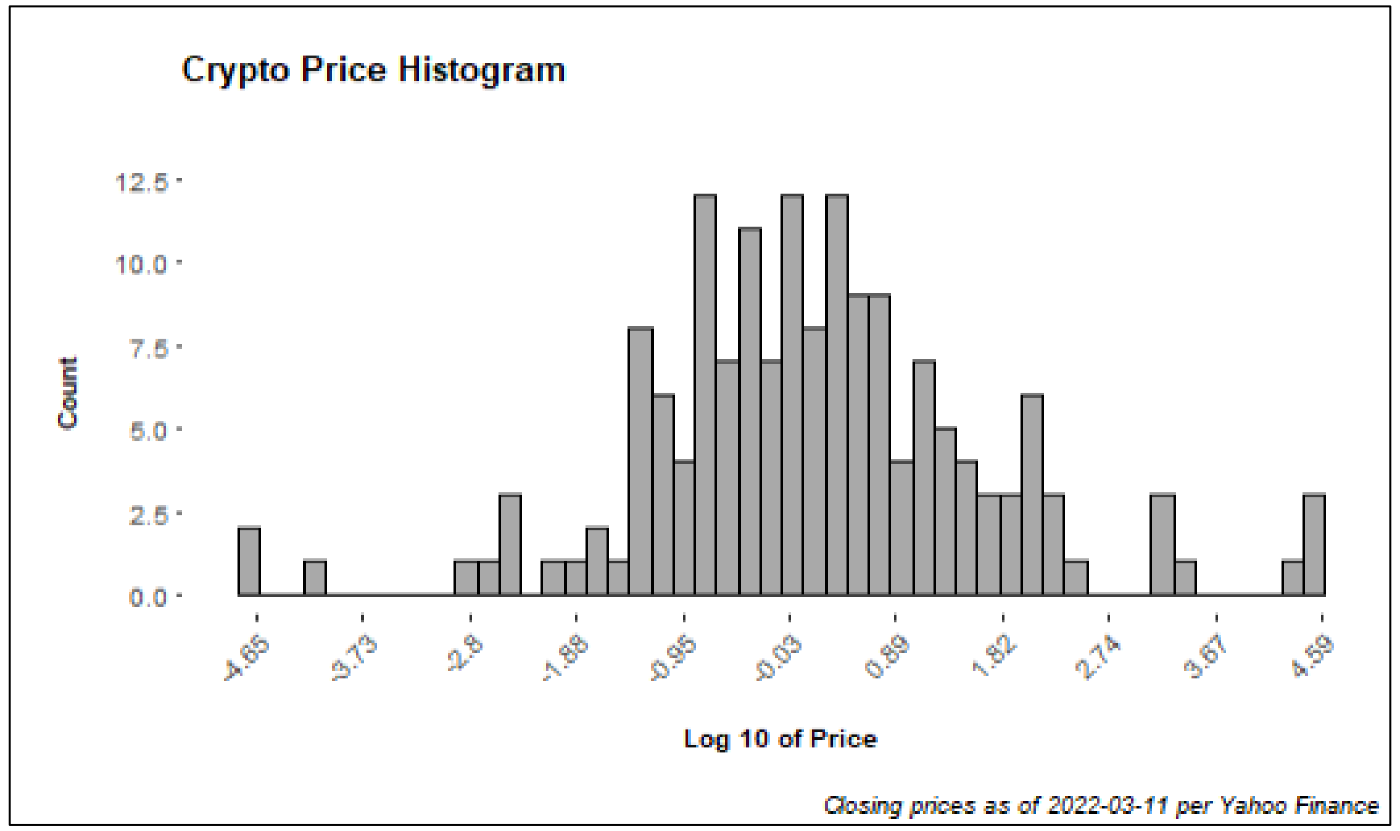
Google	
PK	Ticker
PK	Date
	InterestLevel

Price	
PK	Ticker
PK	Date
	High
	Low
	Close
	Volume

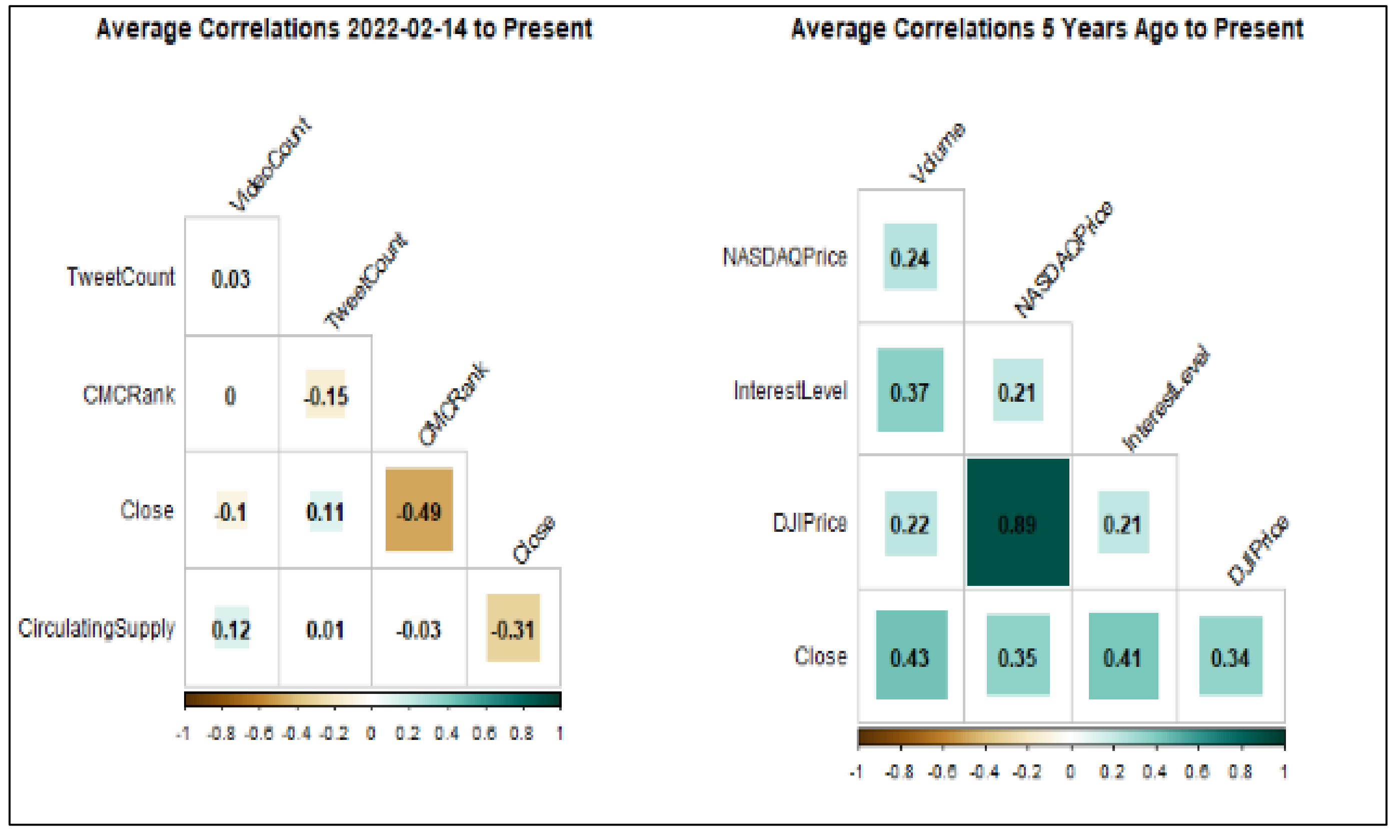
Stocks	
PK	Date
	DJIPrice
	NASDAQPrice

CoinMarket	
PK	Ticker
PK	Date
	CirculatingSupply
	MaximumSupply
	CMCRank

Exploratory Data Analysis



Closing prices as of 2022-03-11 per Yahoo Finance

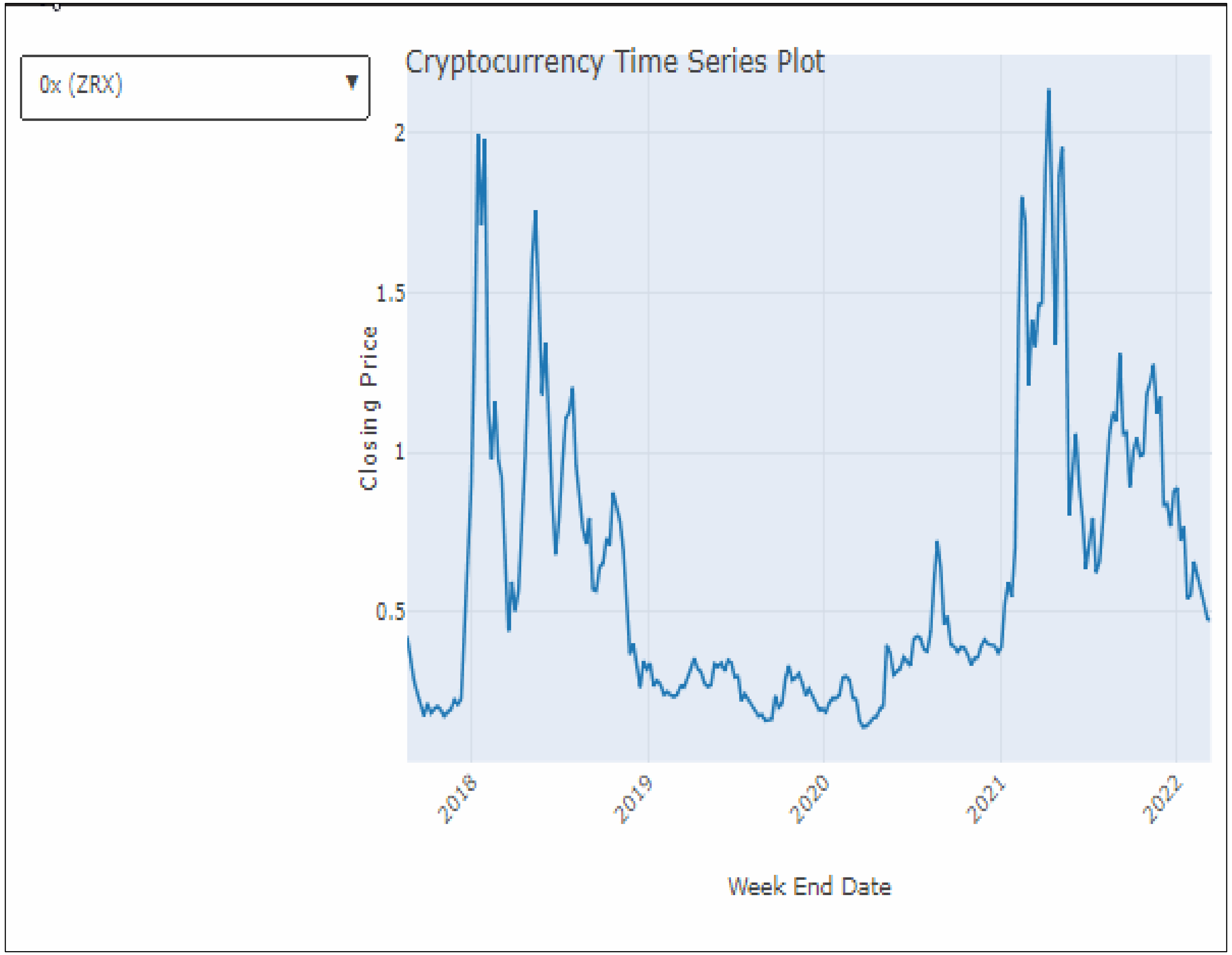


Average Correlations 2022-02-14 to Present

	VideoCount	TweetCount	CMCRank	Close
TweetCount	0.03			
CMCRank	0	-0.15		
Close	-0.1	0.11	-0.49	
CirculatingSupply	0.12	0.01	-0.03	-0.31

Average Correlations 5 Years Ago to Present

	Volume	NASDAQPrice	InterestLevel	DJIPrice
NASDAQPrice	0.24			
InterestLevel	0.37	0.21		
DJIPrice	0.22	0.89	0.21	
Close	0.43	0.35	0.41	0.34



V. Conclusion

Reflection

Project I

Flexibility as a point of emphasis

- Once the database is developed it can be difficult to make changes
- But changes are bound to happen, so flexibility is important
- Discuss with stakeholders, put yourself in their shoes

Project II Part 1

Importance of data quality

- There is only so much that can be done to remedy bad data
- If the data that goes into a model is bad the data that comes out is also likely bad
- If the data scientist is involved in data collection, make sure its done right!

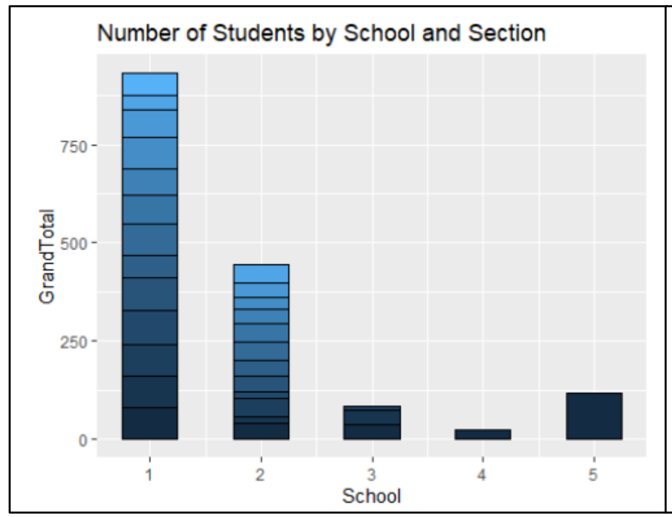
Project II Part 2

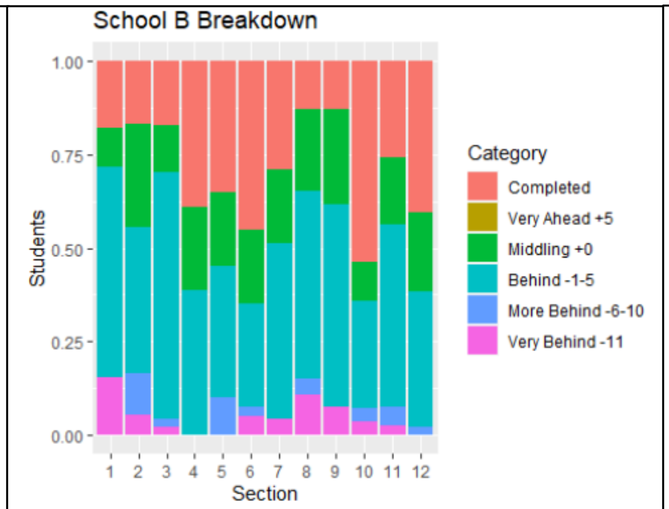
Dealing with ambiguity

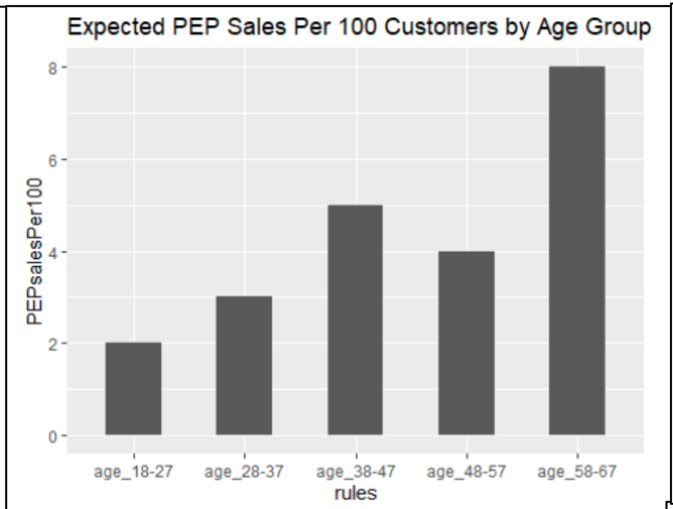
- Ambiguity is a reoccurring theme in data science, especially with text data
- Crowdsourcing is a good way to approach ambiguity
- Do not expect out of the box solutions to be enough, configure for the task at hand

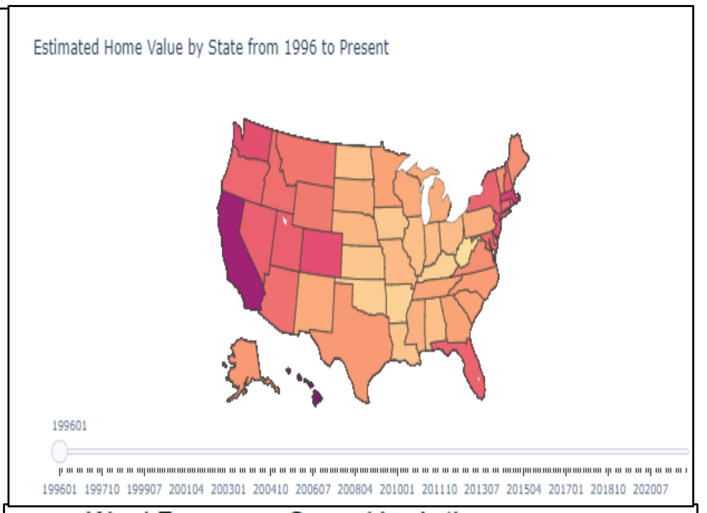
Gallery

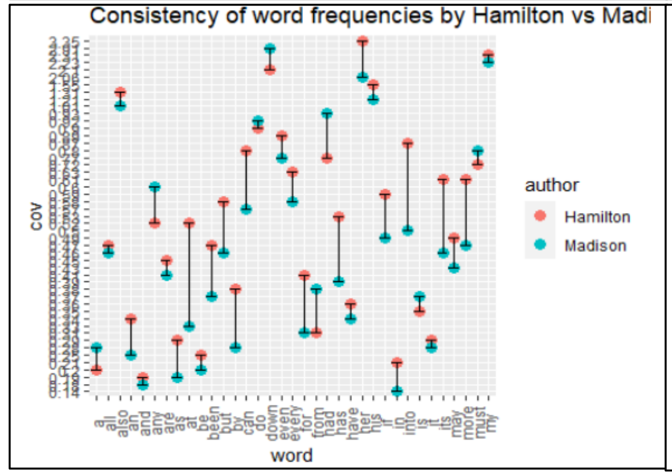
A collection of other data visualizations I have created for assignments, labs, or projects outside of the 3 projects discussed in this presentation

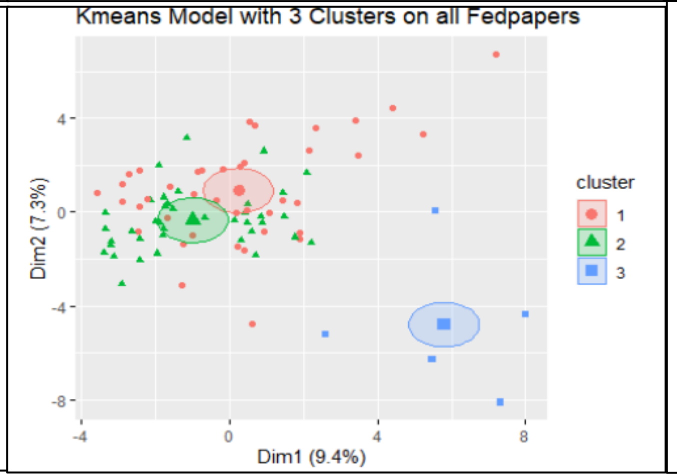

Number of Students by School and Section

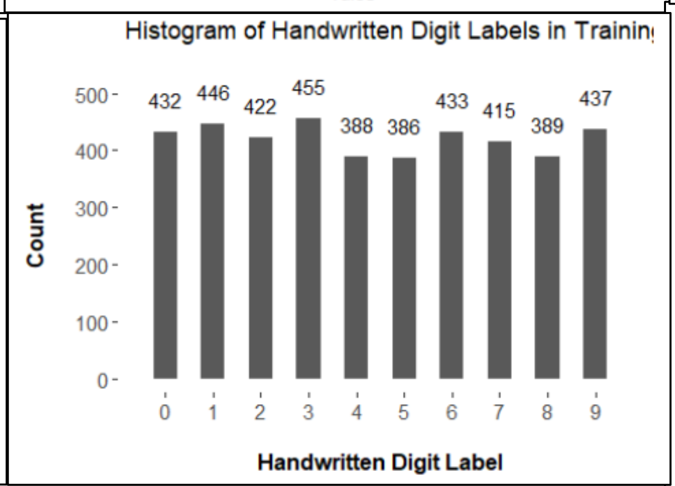

School B Breakdown

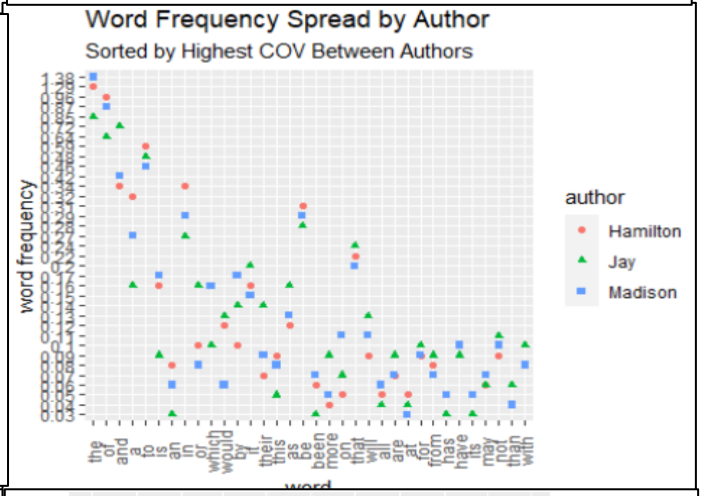

Expected PEP Sales Per 100 Customers by Age Group

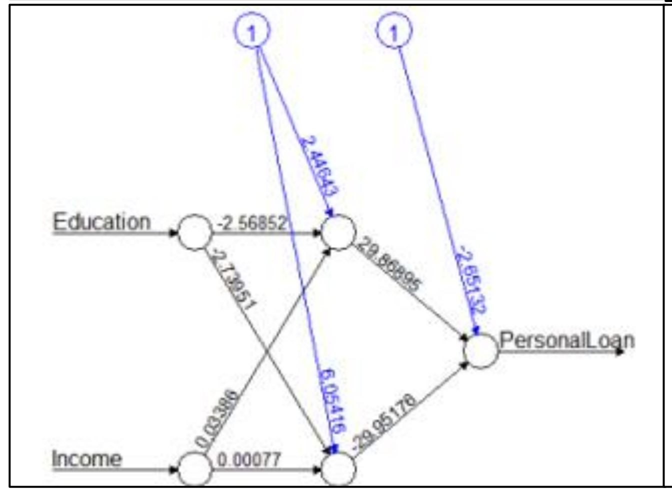

Estimated Home Value by State from 1996 to Present

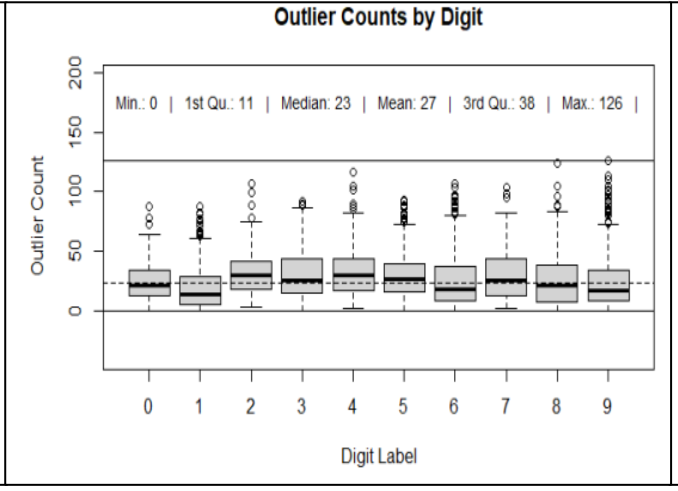

Consistency of word frequencies by Hamilton vs Madi

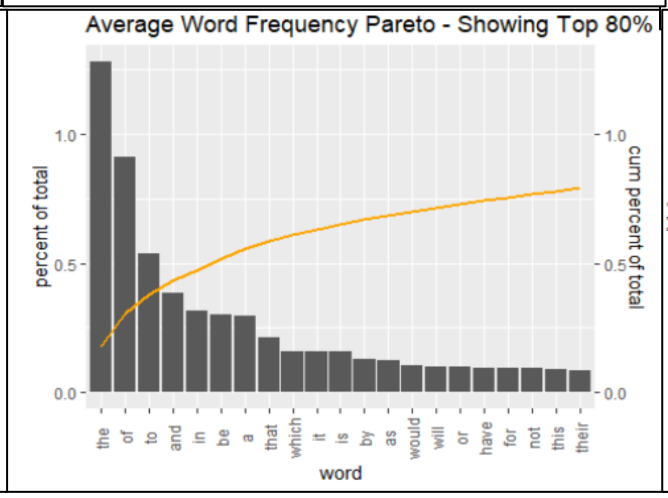

Kmeans Model with 3 Clusters on all Fedpapers

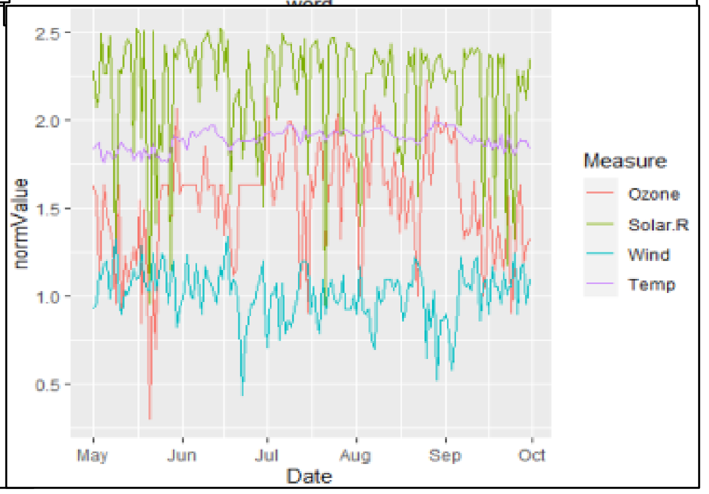

Histogram of Handwritten Digit Labels in Trainin


Word Frequency Spread by Author
Sorted by Highest COV Between Authors




Outlier Counts by Digit


Average Word Frequency Pareto - Showing Top 80%



End Presentation