GitHub repository for files: https://github.com/knolin804/ETL_project
Project Team: Derek Hagman, Kristine Nolin, Carol Wittig
ETL Steps:

EXTRACT

FINDING DATA

From Kaggle we downloaded the following datasets because they provided related book information and could provide data for analysis based on user ratings across three different rating groups—one a reader advisory (GoodRead) and two e-commerce sites: Amazon and Flipkart.

- "GoodReads database" (csv file) which includes over 13,000 books, title, author, isbn, avg reader rating, numbers of reviews, numbers of ratings, categories.

- Amazon () vs Flipkart (1400 entries) book prices data set(s), which contained the same set of variables: author, title, isbn, ratings count, stars (rating), and price.

TRANSFORM

DATA CLEANUP & ANALYSIS

Sources of Data:

1) We reviewed the original csv files for their structure and columns but didn't do any do manipulation before bringing into pandas [CSV FILES IN FOLDER: amazon.csv, books.csv, flipkart.csv]

Types of Transformation Needed:

1) Each of us created a Jupyter notebook and imported our csv file into Panadas, creating a dataframe [JUPYTER NOTEBOOKS IN FOLDER: ETL_FINAL.ipynb (GoodReads), Final_Amazon,ipynb, and flipkart.ipynb]

2) Cleaning of the dataframes included:
   a. Removing duplicate entries (Flipkart) and also removed empty rows with missing data

   b. Formatted the text in the title column to be consistent across the three notebooks

   c. Converted rupee prices to American dollars (Flipkart and Amazon)

   d. Dropped some of the original columns that wouldn't be needed for further analysis

   e. Split the author name into first and last names. Jupyter notebooks show the details, but this was the most intensive work step, although the output doesn't reflect how long it took or how unnecessary it ended up being based on what was actually needed in the database.
      i. Additional step with dataframes needed for Flipkart because author order ws inconsistent and in Amazon, author formatting was inconsistent. GoodReads data was the cleanest, but still had challenges in that first name, middle initial, last name was not easily separated.

ii. In some instances the authors' names were split over multiple columns and these needed to be appropriately reordered. For Flipkart, Derek used the isbn value to confirm authors of the books.

iii. Saved cleaned csvs to files (amazon_df2.csv, books_cleaned.csv, and final_flipkar.csv)

LOAD

Final Production Database – we selected PostGres and chose a relational database for our structure because each of our datasets were csv files in a table format with structured fields. Each contained related items in ISBN, title, and author. This type of structured and similar data, with multi-rows of similar data, and a predefined schema (column names in a table structure) lends itself to a relational database.

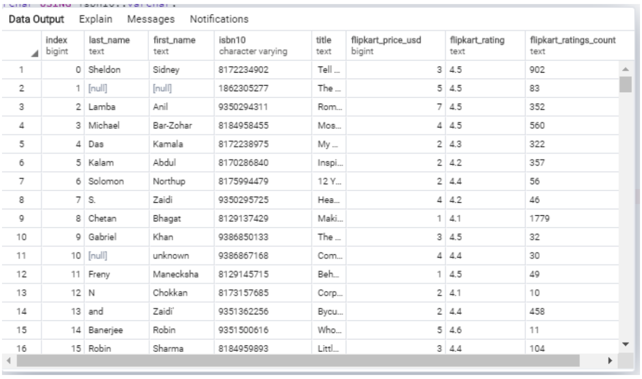
Setup the PostGres Database [SQL DATABASE IN FOLDER: project_queries.sql]

a) Established the PostGres database GoodReads.

FINAL TABLES USED IN PRODUCTION DATABASE [SQL TABLES IN FOLDER – project_queries.sql]

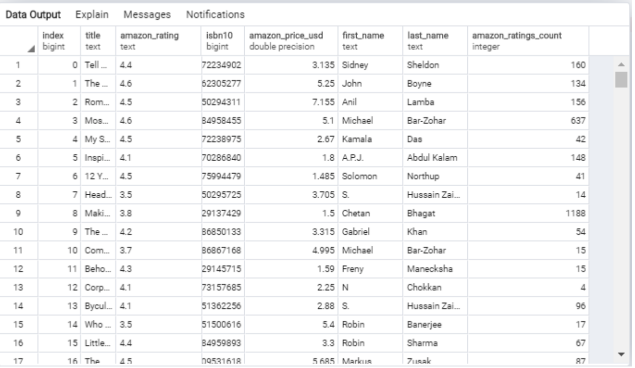
In Jupyter, created the engine to make the connection with PostGres

a) Uploaded the dataframes to create the tables in PostGres
b) In PostGres, changed the datatypes because import assigned datatypes to the columns that had to be changed before you could run queries across the tables.
   i. Tables show query for each dataset - fields included author, title, ISBN, rating.



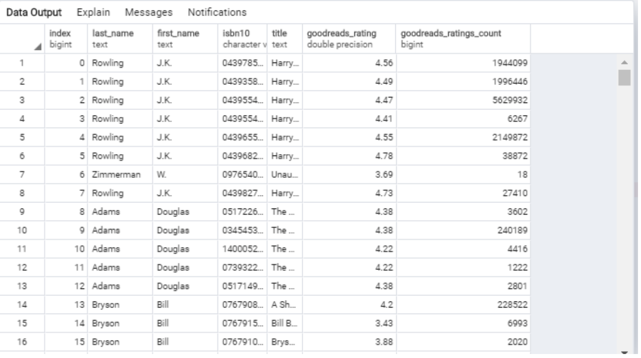
	index bigint	last_name text	first_name text	isbn10 character varying	title text	flipkart_price_usd bigint	flipkart_rating text	flipkart_ratings_count text
1	0	Sheldon	Sidney	8172234902	Tell ...	3	4.5	902
2	1	[null]	[null]	1862305277	The ...	5	4.5	83
3	2	Lamba	Anil	9350294311	Rom...	7	4.5	352
4	3	Michael	Bar-Zohar	8184958455	Mos...	4	4.5	560
5	4	Das	Kamala	8172238975	My ...	2	4.3	322
6	5	Kalam	Abdul	8170286840	Inspi...	2	4.2	357
7	6	Solomon	Northup	8175994479	12 Y...	2	4.4	56
8	7	S.	Zaidi	9350295725	Hea...	4	4.2	46
9	8	Chetan	Bhagat	8129137429	Maki...	1	4.1	1779
10	9	Gabriel	Khan	9386850133	The ...	3	4.5	32
11	10	[null]	unknown	9386867168	Com...	4	4.4	30
12	11	Freny	Manecksha	8129145715	Beh...	1	4.5	49
13	12	N	Chokkan	8173157685	Corp...	2	4.1	10
14	13	and	Zaidi'	9351362256	Bycu...	2	4.4	458
15	14	Banerjee	Robin	9351500616	Who...	5	4.6	11
16	15	Robin	Sharma	8184959893	Littl...	3	4.4	104

Figure 1: Flipkart Table



	index bigint	title text	amazon_rating text	isbn10 bigint	amazon_price_usd double precision	first_name text	last_name text	amazon_ratings_count integer
1	0	Tell ...	4.4	72234902	3.135	Sidney	Sheldon	160
2	1	The ...	4.6	62305277	5.25	John	Boyne	134
3	2	Rom...	4.5	50294311	7.155	Anil	Lamba	156
4	3	Mos...	4.6	84958455	5.1	Michael	Bar-Zohar	637
5	4	My S...	4.5	72238975	2.67	Kamala	Das	42
6	5	Inspi...	4.1	70286840	1.8	A.P.J.	Abdul Kalam	148
7	6	12 Y...	4.5	75994479	1.485	Solomon	Northup	41
8	7	Head...	3.5	50295725	3.705	S.	Hussain Zai...	14
9	8	Maki...	3.8	29137429	1.5	Chetan	Bhagat	1188
10	9	The ...	4.2	86850133	3.315	Gabriel	Khan	54
11	10	Com...	3.7	86867168	4.995	Michael	Bar-Zohar	15
12	11	Beho...	4.3	29145715	1.59	Freny	Manecksha	15
13	12	Corp...	4.1	73157685	2.25	N	Chokkan	4
14	13	Bycul...	4.1	51362256	2.88	S.	Hussain Zai...	96
15	14	Who ...	3.5	51500616	5.4	Robin	Banerjee	17
16	15	Little...	4.4	84959893	3.3	Robin	Sharma	67
17	16	The	4.5	09531618	5.685	Markus	Zusak	87

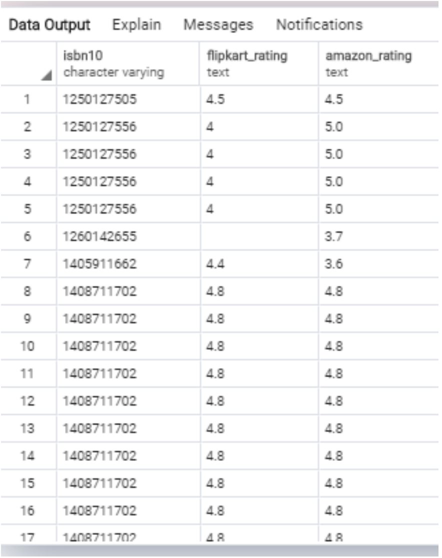
Figure 2: Amazon Table



	index bigint	last_name text	first_name text	isbn10 character v	title text	goodreads_rating double precision	goodreads_ratings_count bigint
1	0	Rowling	J.K.	0439785...	Harry...	4.56	1944099
2	1	Rowling	J.K.	0439358...	Harry...	4.49	1996446
3	2	Rowling	J.K.	0439554...	Harry...	4.47	5629932
4	3	Rowling	J.K.	0439554...	Harry...	4.41	6267
5	4	Rowling	J.K.	0439655...	Harry...	4.55	2149872
6	5	Rowling	J.K.	0439682...	Harry...	4.78	38872
7	6	Zimmerman	W.	0976540...	Unau...	3.69	18
8	7	Rowling	J.K.	0439827...	Harry...	4.73	27410
9	8	Adams	Douglas	0517226...	The ...	4.38	3602
10	9	Adams	Douglas	0345453...	The ...	4.38	240189
11	10	Adams	Douglas	1400052...	The ...	4.22	4416
12	11	Adams	Douglas	0739322...	The ...	4.22	1222
13	12	Adams	Douglas	0517149...	The ...	4.38	2801
14	13	Bryson	Bill	0767908...	A Sh...	4.2	228522
15	14	Bryson	Bill	0767915...	Bill B...	3.43	6993
16	15	Bryson	Bill	0767910...	Brys...	3.88	2020

Figure 3: GoodReads Table

ii. A sample query included joining tables to compare ratings across Flipkart and Amazon.



	isbn10 character varying	flipkart_rating text	amazon_rating text
1	1250127505	4.5	4.5
2	1250127556	4	5.0
3	1250127556	4	5.0
4	1250127556	4	5.0
5	1250127556	4	5.0
6	1260142655		3.7
7	1405911662	4.4	3.6
8	1408711702	4.8	4.8
9	1408711702	4.8	4.8
10	1408711702	4.8	4.8
11	1408711702	4.8	4.8
12	1408711702	4.8	4.8
13	1408711702	4.8	4.8
14	1408711702	4.8	4.8
15	1408711702	4.8	4.8
16	1408711702	4.8	4.8
17	1408711702	4.8	4.8