Team 1: Ecommerce, Consumer Behavior

Project Title: An Investigation of Amazon’s Consumer Behavior

Research Questions:

- Can we predict which products a customer will most likely purchase together within various product segments?
- Can we identify customer segments based on the purchased product categories to better target marketing campaigns?
- Can we extract key topics within product reviews to help companies analyze and interpret customer feedback?

Question 1

- Can we predict which products a customer will most likely purchase together within various product segments?

Goal of Questions 1

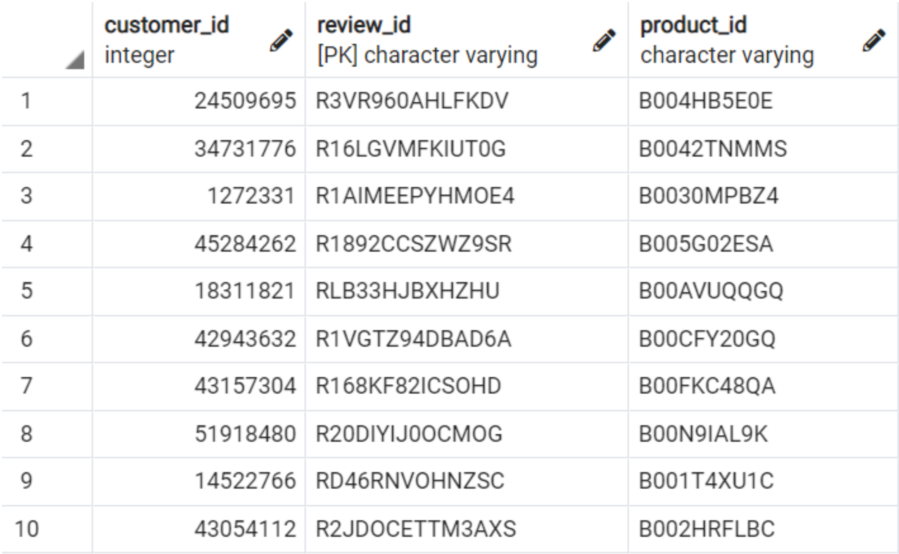
- Help Amazon identify products frequently bought together by customers to increase sales and revenues (cross sell) by analyzing Amazon Marketplace segment data.

Machine Learning Plans

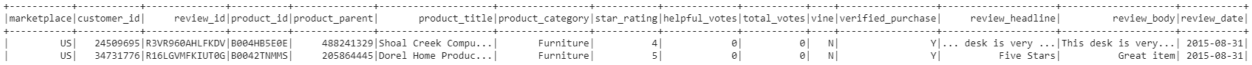
- Association Data Mining
  - Apriori Algorithm:
    - Utilize Apriori Algorithm to populate items that are most frequently bought together within various product segments.

Data Summary

- Team will be using Amazon.com product segment data from S3
- **Data Source:** Amazon S3
- **Datasets:** 8 different product segments
  - Apparel
  - Furniture
  - Music
  - Office Products
  - Personal Care Appliances
  - Video Games
  - Videos
  - Watches
- **Number of Columns:** 15 (raw); 3 (after load in postgres)
- **Type:** Structured



	customer_id integer	review_id [PK] character varying	product_id character varying
1	24509695	R3VR960AHLFKDV	B004HB5E0E
2	34731776	R16LGVMFKIUT0G	B0042TNMMS
3	1272331	R1AIMEEPYHMOE4	B0030MPBZ4
4	45284262	R1892CCSZWZ9SR	B005G02ESA
5	18311821	RLB33HJBXHZHU	B00AVUQQGQ
6	42943632	R1VGTZ94DBAD6A	B00CFY20GQ
7	43157304	R168KF82ICSOHD	B00FKC48QA
8	51918480	R20DIYIJ0OCMOG	B00N9IAL9K
9	14522766	RD46RNVOHNZSC	B001T4XU1C
10	43054112	R2JDOCETTM3AXS	B002HRFLBC



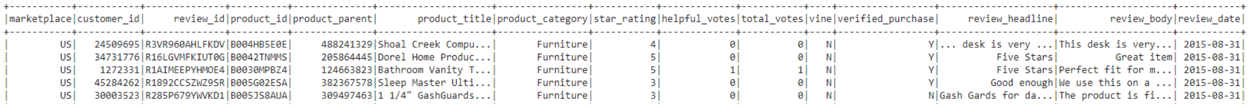
```
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+--------------------+-----------+
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_purchase|     review_headline|         review_body|review_date|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+--------------------+-----------+
|         US|   24509695|R3VR960AHLFKDV|B004HB5E0E|     488241329|Shoal Creek Compu...|       Furniture|          4|            0|          0|   N|                Y|... desk is very ...|This desk is very...| 2015-08-31|
|         US|   34731776|R16LGVMFKIUT0G|B0042TNMMS|     205864445|Dorel Home Produc...|       Furniture|          5|            0|          0|   N|                Y|          Five Stars|          Great item| 2015-08-31|
```

Data Processing Plan

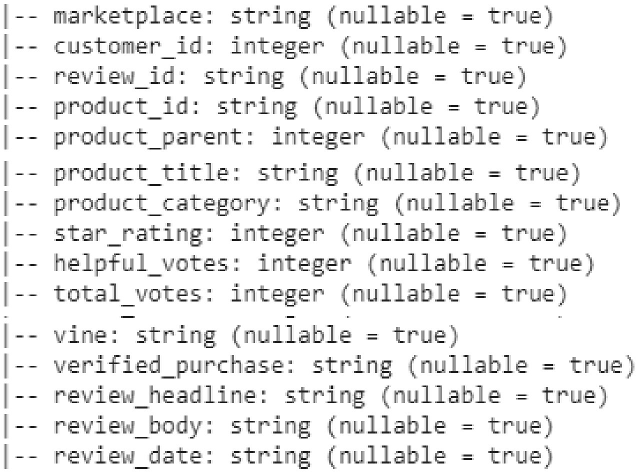
Extract

The team selected 8 different **product segments** from Amazon data:

- Music
- Video Games
- Videos
- Watches
- Furniture
- Office Products
- Personal Care Appliances
- Apparel



```
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+--------------------+-----------+
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_purchase|     review_headline|         review_body|review_date|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+--------------------+-----------+
|         US|   24509695|R3VR960AHLFKDV|B004HB5E0E|     488241329|Shoal Creek Compu...|       Furniture|          4|            0|          0|   N|                Y|... desk is very ...|This desk is very...| 2015-08-31|
|         US|   34731776|R16LGVMFKIUT0G|B0042TNMMS|     205864445|Dorel Home Produc...|       Furniture|          5|            0|          0|   N|                Y|          Five Stars|          Great item| 2015-08-31|
|         US|    1272331|R1AIMEEPYHMOE4|B0030MPBZ4|     124663823|Bathroom Vanity T...|       Furniture|          5|            1|          1|   N|                Y|          Five Stars|Perfect fit for m...| 2015-08-31|
|         US|   45284262|R1892CCSZWZ9SR|B005G02ESA|     382367578|Sleep Master Ulti...|       Furniture|          3|            0|          0|   N|                Y|         Good enough|We use this on a ...| 2015-08-31|
|         US|   30003523|R285P679YWVKD1|B005JS8AUA|     309497463|1 1/4" GashGuards...|       Furniture|          3|            0|          0|   N|                N|Gash Gards for da...|The product is fi...| 2015-08-31|
```



```
 |-- marketplace: string (nullable = true)
 |-- customer_id: integer (nullable = true)
 |-- review_id: string (nullable = true)
 |-- product_id: string (nullable = true)
 |-- product_parent: integer (nullable = true)
 |-- product_title: string (nullable = true)
 |-- product_category: string (nullable = true)
 |-- star_rating: integer (nullable = true)
 |-- helpful_votes: integer (nullable = true)
 |-- total_votes: integer (nullable = true)
 |-- vine: string (nullable = true)
 |-- verified_purchase: string (nullable = true)
 |-- review_headline: string (nullable = true)
 |-- review_body: string (nullable = true)
 |-- review_date: string (nullable = true)
```

Transform

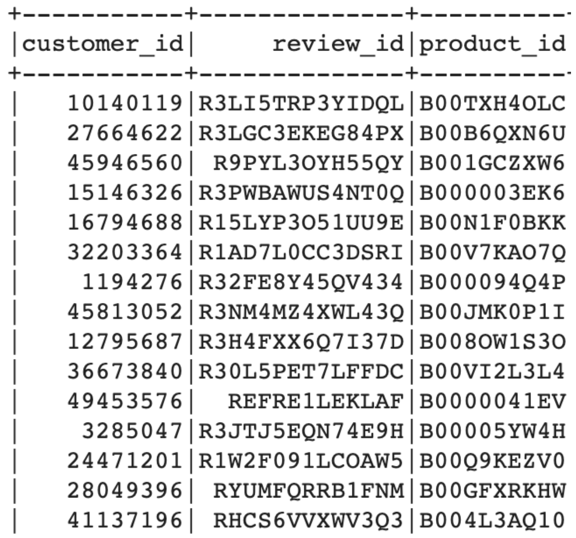
- Load Amazon product segment into PySpark DataFrame
- Perform preliminary cleaning
  - Drop unnecessary columns
    - Columns: ‘marketplace’, ‘product_parent’, ‘vine’, ‘review_headline’, ‘review_headline’, ‘review_body’, ‘review_date’
  - Filter data to present only verified purchases
    - verified_purchase = ‘Y’
  - Drop the verified purchased column after filtering
    - Column: verified_purchase
- Create Apriori Analysis dataframe
  - Drop additional unnecessary columns in preparation for Apriori Analysis
    - Columns: ‘review_id’, ‘product_id’, ‘product_title’, ‘star_rating’, ‘helpful_votes’, ‘total_votes’
- Repeat this process with various product segments.

Load

- Download Postgres driver that will allow Spark to interact with PostgresSQL

- Configure settings for PostgresSQL

- Write the cleaned table into PostgresSQL.
  - Write cleaned product segment table that is prepped for Apriori Analysis into PostgresSQL

- 

```
+-----------+--------------+----------
|customer_id|     review_id|product_id
+-----------+--------------+----------
|   10140119|R3LI5TRP3YIDQL|B00TXH4OLC
|   27664622|R3LGC3EKEG84PX|B00B6QXN6U
|   45946560| R9PYL3OYH55QY|B001GCZXW6
|   15146326|R3PWBAWUS4NT0Q|B000003EK6
|   16794688|R15LYP3O51UU9E|B00N1F0BKK
|   32203364|R1AD7L0CC3DSRI|B00V7KAO7Q
|    1194276|R32FE8Y45QV434|B000094Q4P
|   45813052|R3NM4MZ4XWL43Q|B00JMK0P1I
|   12795687|R3H4FXX6Q7I37D|B008OW1S3O
|   36673840|R30L5PET7LFFDC|B00VI2L3L4
|   49453576|  REFRE1LEKLAF|B000041EV
|    3285047|R3JTJ5EQN74E9H|B00005YW4H
|   24471201|R1W2F091LCOAW5|B00Q9KEZV0
|   28049396| RYUMFQRRB1FNM|B00GFXRKHW
|   41137196| RHCS6VVXWV3Q3|B004L3AQ10
```

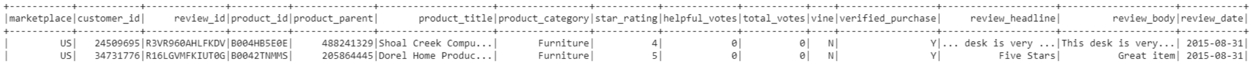
Question 2: Can we identify customer segments based on the purchased product categories to better target marketing campaigns?

Goal of Question 2: Help Amazon learn and predict which customers are more likely to purchase products within product segments. Identifying this trend can help Amazon target advertisements to specific customers within certain product segments in efforts to increase sales and revenues.

Machine Learning Plans

- Unsupervised Machine Learning
    - K-Means Cluster Analysis
        - A K-Means Cluster Analysis model will be performed to cluster customers into various product types based on purchasing behavior within various product segments.

Data Summary

- Team will be using Amazon.com product segment data from S3
- **Data Source:** Amazon S3
- **Datasets:** 8 different product segments
    - Music
    - Video Games
    - Videos
    - Watches
    - Furniture
    - Office Products
    - Personal Care Appliances
    - Apparel
- **Number of Columns:** 15
- **Type:** Structured
- 

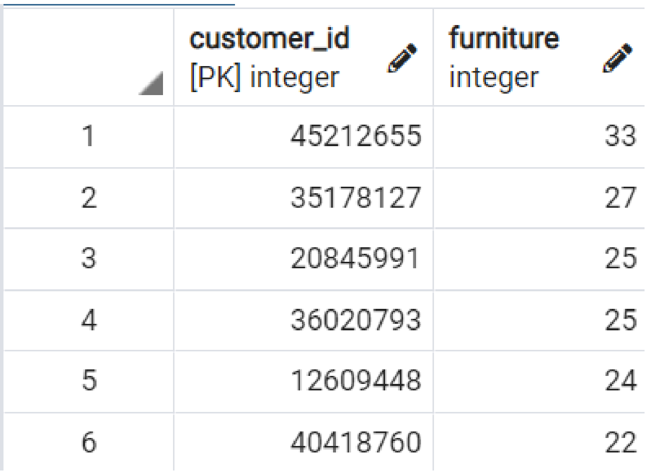
```
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+--------------------+-----------+
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_purchase|     review_headline|         review_body|review_date|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+--------------------+-----------+
|         US|   24509695|R3VR960AHLFKDV|B004HB5E0E|     488241329|Shoal Creek Compu...|       Furniture|          4|            0|          0|   N|                Y|... desk is very ...|This desk is very...| 2015-08-31|
|         US|   34731776|R16LGVMFKIUT0G|B0042TNMMS|     205864445|Dorel Home Produc...|       Furniture|          5|            0|          0|   N|                Y|          Five Stars|          Great item| 2015-08-31|
```

ETL Pipeline

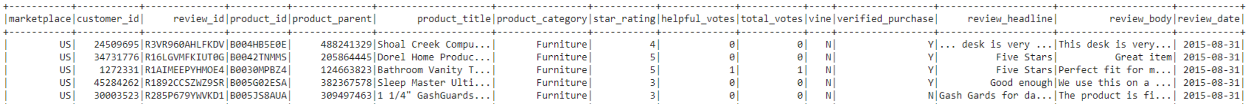
Extract

The team selected 8 different **product segments** from Amazon data:

- Music
- Video Games
- Videos
- Watches
- Furniture
- Office Products
- Personal Care Appliances
- Apparel



	customer_id [PK] integer	furniture integer
1	45212655	33
2	35178127	27
3	20845991	25
4	36020793	25
5	12609448	24
6	40418760	22



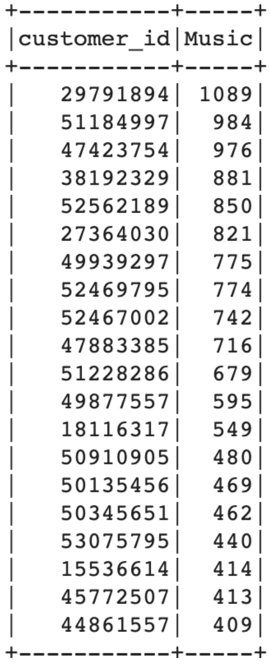
```
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+--------------------+-----------+
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_purchase|     review_headline|         review_body|review_date|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+--------------------+-----------+
|         US|   24509695|R3VR960AHLFKDV|B004HB5E0E|     488241329|Shoal Creek Compu...|       Furniture|          4|            0|          0|   N|                Y|... desk is very ...|This desk is very...| 2015-08-31|
|         US|   34731776|R16LGVMFKIUT0G|B0042TNMMS|     205864445|Dorel Home Produc...|       Furniture|          5|            0|          0|   N|                Y|          Five Stars|          Great item| 2015-08-31|
|         US|    1272331|R1AIMEEPYHMOE4|B0030MPBZ4|     124663823|Bathroom Vanity T...|       Furniture|          5|            1|          1|   N|                Y|          Five Stars|Perfect fit for m...| 2015-08-31|
|         US|   45284262|R1892CCSZWZ9SR|B005G02ESA|     382367578|Sleep Master Ulti...|       Furniture|          3|            0|          0|   N|                Y|         Good enough|We use this on a ...| 2015-08-31|
|         US|   30003523|R285P679YWVKD1|B005JS8AUA|     309497463|1 1/4" GashGuards...|       Furniture|          3|            0|          0|   N|                N|Gash Gards for da...|The product is fi...| 2015-08-31|
```

Transform

1. Load Amazon product segment into PySpark DataFrame
2. Perform preliminary cleaning
   - Drop unnecessary columns
     - Columns: ‘marketplace’, ‘product_parent’, ‘vine’, ‘review_headline’, ‘review_headline’, ‘review_body’, ‘review_date’
   - Filter data to present only verified purchases
     - verified_purchase = ‘Y’
   - Drop the verified purchased column after filtering
     - Column: verified_purchase
3. Create Segmentation Analysis dataframe
   - Drop additional unnecessary columns in preparation for K-Means Cluster Analysis
     - Columns: ‘review_id’, ‘product_id’, ‘product_title’, ‘star_rating’, ‘helpful_votes’, ‘total votes’
   - Group the data by customer id and product category
     - Group by ‘customer_id’ and count ‘product_category’ (# reviews = # transactions)
   - Filter the top results
     - Filter the data to show 100,000 rows displaying the top customer purchases within the chosen product category
4. Repeat this process with various product segments.

Load

- Download Postgres driver that will allow PySpark to interact with PostgresSQL
- Configure settings for PostgresSQL
- Write the cleaned table into PostgresSQL.
    - Write cleaned product segment table that is prepped for K-Means Cluster Analysis into PostgresSQL



```
+-----------+-----+
|customer_id|Music|
+-----------+-----+
|   29791894| 1089|
|   51184997|  984|
|   47423754|  976|
|   38192329|  881|
|   52562189|  850|
|   27364030|  821|
|   49939297|  775|
|   52469795|  774|
|   52467002|  742|
|   47883385|  716|
|   51228286|  679|
|   49877557|  595|
|   18116317|  549|
|   50910905|  480|
|   50135456|  469|
|   50345651|  462|
|   53075795|  440|
|   15536614|  414|
|   45772507|  413|
|   44861557|  409|
+-----------+-----+
```

Question 3: Can we extract key topics within product reviews to help companies analyze customer feedback?

Goal of Question 3: Help companies easily and readily extract key topics within product reviews to understand the customer feedback of their products. This will help companies identify positive or negative trends with their products and allow them to improve their products and customer service without having to read review by review.

Product Reviews for a specific product (B000M0MJU2; air mattress)

Machine Learning Plans

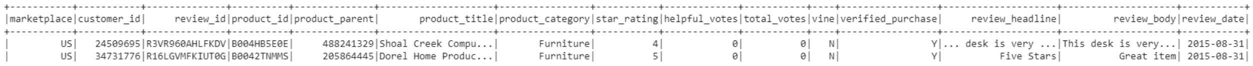
- Natural Language Processing
    - Topic Analysis
        - Use NLP to remove words that are not aggregating to analysis
        - Utilize Topic Analysis to enable companies to easily and readily view key topics from their product reviews in efforts to improve customer and product services.
        - Use Latent Dirichlet Allocation (LDA) machine learning model for topic

discovery

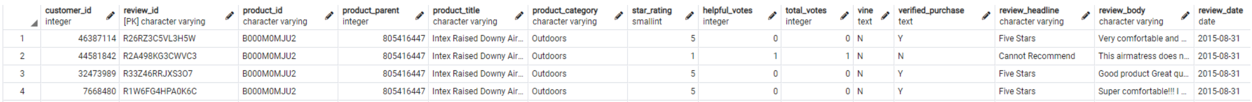
Data Summary

Team will be using Amazon.com product segment data from S3

- **Data Source:** Amazon S3
- **Datasets:** 1 product segments
  - Outdoors
- **Number of Columns: 15**
- **Type: Structured**



```
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+--------------------+-----------+
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_purchase|     review_headline|         review_body|review_date|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+--------------------+-----------+
|         US|   24509695|R3VR960AHLFKDV|B004HB5E0E|     488241329|Shoal Creek Compu...|       Furniture|          4|            0|          0|   N|                Y|... desk is very ...|This desk is very...| 2015-08-31|
|         US|   34731776|R16LGVMFKIUT0G|B0042TNMMS|     205864445|Dorel Home Produc...|       Furniture|          5|            0|          0|   N|                Y|          Five Stars|          Great item| 2015-08-31|
```



	customer_id integer	review_id [PK] character varying	product_id character varying	product_parent integer	product_title character varying	product_category character varying	star_rating smallint	helpful_votes integer	total_votes integer	vine text	verified_purchase text	review_headline character varying	review_body character varying	review_date date
1	46387114	R26RZ3C5VL3H5W	B000M0MJU2	805416447	Intex Raised Downy Air...	Outdoors	5	0	0	N	Y	Five Stars	Very comfortable and ...	2015-08-31
2	44581842	R2A498KG3CWVC3	B000M0MJU2	805416447	Intex Raised Downy Air...	Outdoors	1	1	1	N	N	Cannot Recommend	This airmatress does n...	2015-08-31
3	32473989	R33Z46RRJXS3O7	B000M0MJU2	805416447	Intex Raised Downy Air...	Outdoors	5	0	0	N	Y	Five Stars	Good product Great qu...	2015-08-31
4	7668480	R1W6FG4HPA0K6C	B000M0MJU2	805416447	Intex Raised Downy Air...	Outdoors	5	0	0	N	Y	Five Stars	Super comfortable!!! I ...	2015-08-31

ETL Process

Extract

The team selected 1 specific **product segments** from Amazon data and selected 1 product:

- Product segment: outdoors
- Product_id: B000M0MJU2

Transform

- Load Amazon product segment into PySpark outdoors dataFrame
- Identify products with the larger number of reviews
- Select the product with the highest volume of reviews
- Filter data by specific product id (B000M0MJU2) the air mattress
- Drop not needed columns: marketplace
- Transform date_review in datetime
- Convert to pandas dataframe to clean up data
- Use NTLKto remove punctuation, make it lower case and handle strange characters for review_body and review_headline

Load

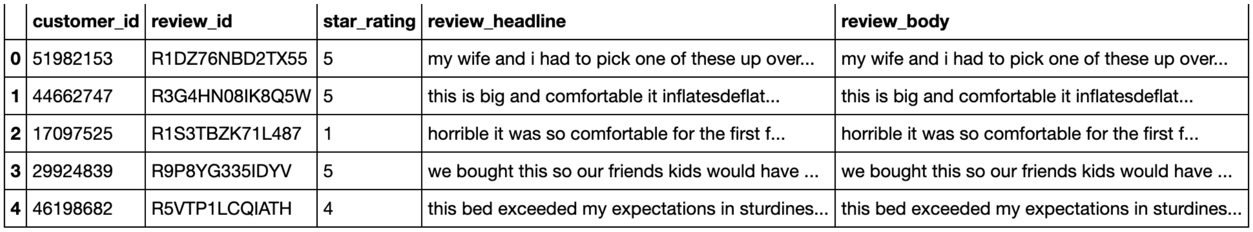
- Download Postgres driver that will allow PySpark to interact with PostgresSQL

- Configure settings for PostgresSQL

- Write the cleaned table into PostgresSQL.

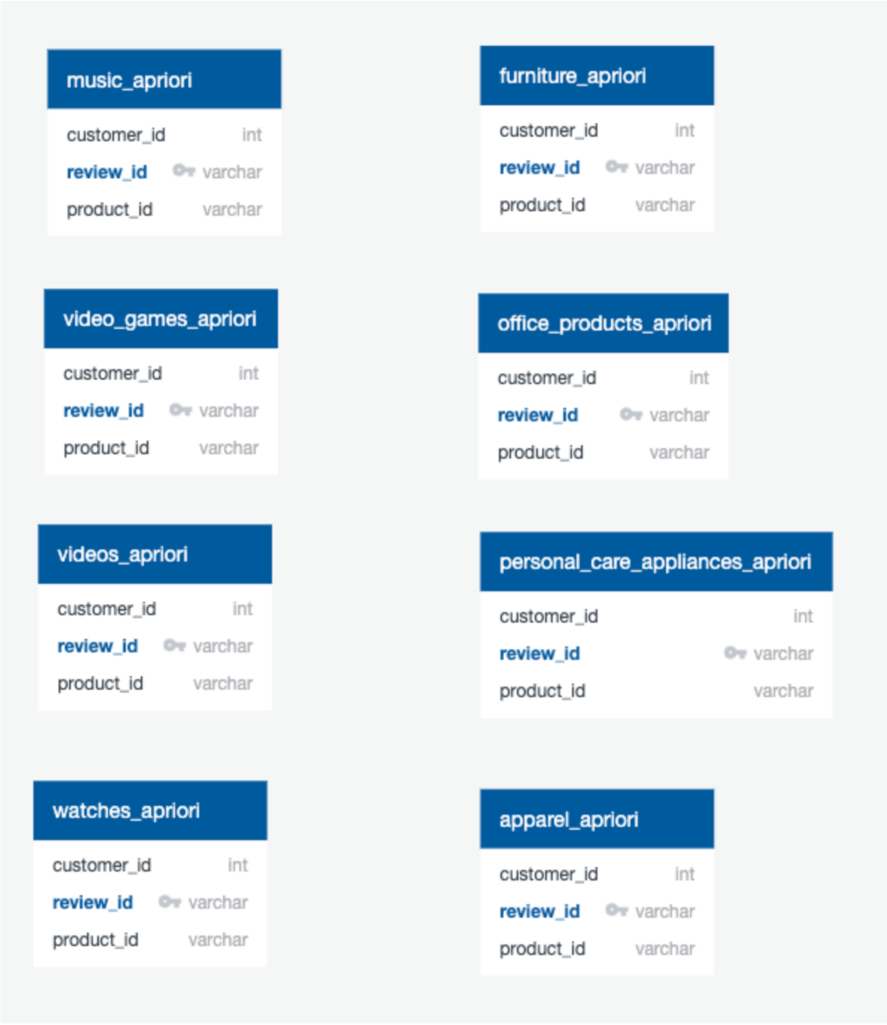
- Write the cleaned air mattress table containing the review data into PostgresSQL in preparation for Topic Analysis.

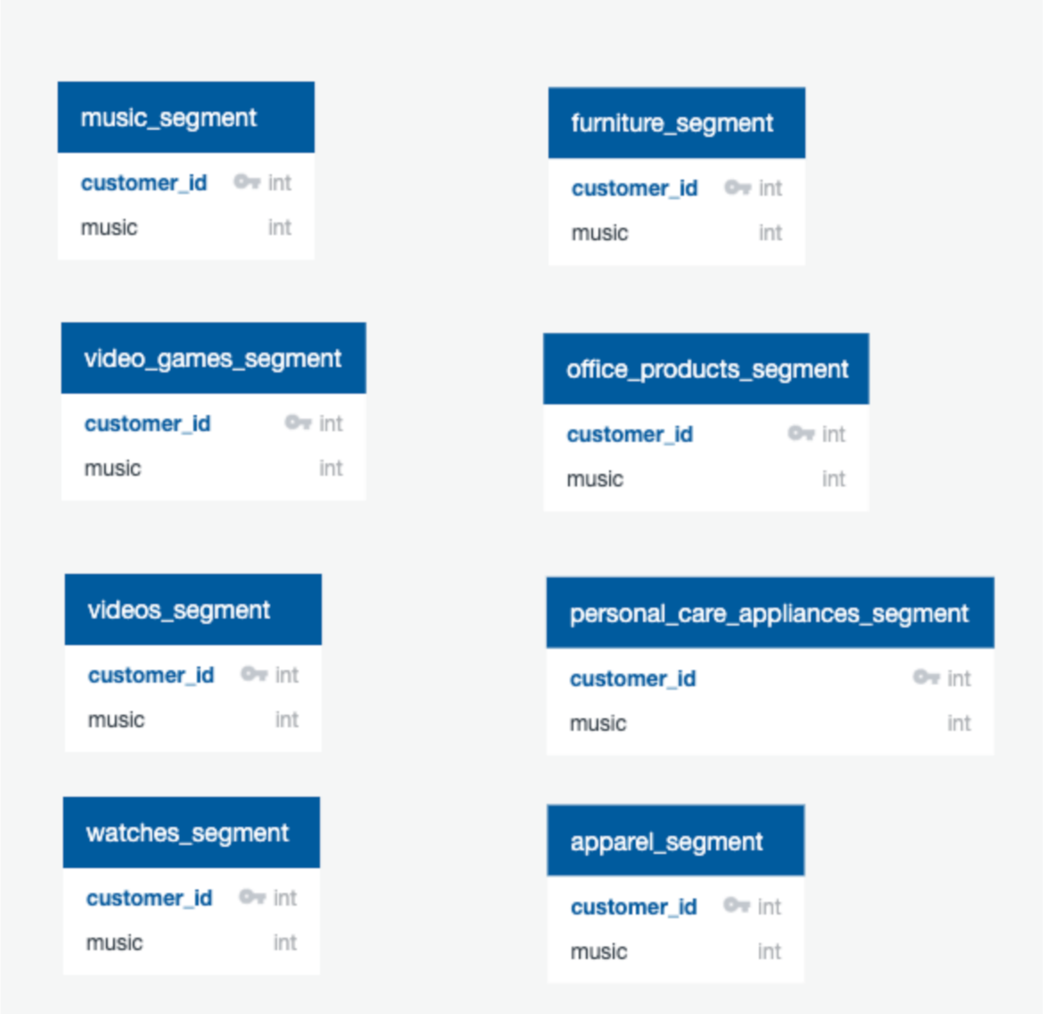
- Use Amazon RDS to connect to database and load dataframe into jupyter notebook for machine learning model



	customer_id	review_id	star_rating	review_headline	review_body
0	51982153	R1DZ76NBD2TX55	5	my wife and i had to pick one of these up over...	my wife and i had to pick one of these up over...
1	44662747	R3G4HN08IK8Q5W	5	this is big and comfortable it inflatesdeflat...	this is big and comfortable it inflatesdeflat...
2	17097525	R1S3TBZK71L487	1	horrible it was so comfortable for the first f...	horrible it was so comfortable for the first f...
3	29924839	R9P8YG335IDYV	5	we bought this so our friends kids would have ...	we bought this so our friends kids would have ...
4	46198682	R5VTP1LCQIATH	4	this bed exceeded my expectations in sturdines...	this bed exceeded my expectations in sturdines...

ERD’s For Apriori Analysis Tables and K-Means Tables Before Joins





music_segment		
customer_id	🔑	int
music		int

furniture_segment		
customer_id	🔑	int
music		int

video_games_segment		
customer_id	🔑	int
music		int

office_products_segment		
customer_id	🔑	int
music		int

videos_segment		
customer_id	🔑	int
music		int

personal_care_appliances_segment		
customer_id	🔑	int
music		int

watches_segment		
customer_id	🔑	int
music		int

apparel_segment		
customer_id	🔑	int
music		int

Why These Topics?

- Data analysis is key for strategic and well-informed decision making
- Big data allows e-commerce businesses to understand customers better through customer behavior analysis
- Helps target specific customers segments to upsell products, increase conversion rates and grow sales
- Better customer segmentation to improve targeted marketing campaigns and increase sales
- Product reviews is a great source of customer feedback and one of the main drivers for conversion rates, developing an automated way to process them can help drive product enhancements and accelerate decision making