



CIS5200 Term Project Lab Tutorial



Authors: [Jasmine Diep](#), [William Lam](#), [Nina Roberts](#), [Taron Sarksyan](#)

Instructor: [Dr. Jongwook Woo](#)

Date: 12/09/2020

Analysis of Movie and TV Reviews: A Study of Amazon Data 2005 - 2015

Objectives

In this hands-on lab, you will learn how to:

- Connect to a node of Oracle BDCE (Big Data Compute)
- Get data manually using wget command
- Use other fundamental shell commands
- Use fundamental Hadoop commands
- Use Hive QL commands to perform the analysis
- Visualize the result in Microsoft Power BI

Platform Spec

- Cluster Version: 20.3.3-20
- CPU Speed: MHz: 2195.287
- Cluster # of Nodes: 3
- # of CPUs Cores: 12
- Memory Size: 180GB
- Storage Size: 957GB

Step 1: Get and merge data

SSH to the Hadoop server using the ip address given by the instructor.

NOTE: You have to use your directory name.

```
$ ssh <your directory name>@129.XX.XX.XX
```

This step will download the Amazon data files from AWS.

Get work files from AWS:

The dataset consists of three files of reviews – DVD reviews, digital download reviews, and VHS reviews. These files contain reviews for all Amazon entertainment such as movies, TV, and instructional videos.

```
$ wget -O amazon_data1.tsv.gz https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon\_reviews\_us\_Digital\_Video.Download\_v1\_00.tsv.gz
$ wget -O amazon_data2.tsv.gz https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon\_reviews\_us\_Video.DVD.v1\_00.tsv.gz
$ wget -O amazon_data3.tsv.gz https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon\_reviews\_us\_Video.v1\_00.tsv.gz

$ ls -al --check for 3 files
```

The screenshot shows a terminal window titled 'MINGW64/c/Users/ninag'. It displays the command-line process of downloading three Amazon review datasets from AWS S3 using the wget command. The terminal output includes the URL for each file, the download progress bar, and the final message indicating the file was saved successfully. The terminal window has a dark background with white text and a light gray border.

```
MINGW64:/c/Users/ninag
$ MINGW64:/c/Users/ninag
$ robert@129.150.69.91's password:
-bash-4.1$ wget -O amazon_data1.tsv.gz https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Digital_Video.Download_v1_00.tsv.gz
--2020-11-16 21:14:25-- https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Digital_Video.Download_v1_00.tsv.gz
Resolving s3.amazonaws.com... 52.217.106.166
Connecting to s3.amazonaws.com[52.217.106.166]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 506979922 (483M) [application/x-gzip]
Saving to: "amazon_data1.tsv.gz"

100%[=====] 506,979,922 25.5M/s  in 56s
2020-11-16 21:15:22 (8.61 MB/s) - "amazon_data1.tsv.gz" saved [506979922/506979922]

-bash-4.1$ wget -O amazon_data2.tsv.gz https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video.DVD.v1_00.tsv.gz
--2020-11-16 21:15:22-- https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video.DVD.v1_00.tsv.gz
Resolving s3.amazonaws.com... 52.217.71.78
Connecting to s3.amazonaws.com[52.217.71.78]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1512355451 (1.4G) [application/x-gzip]
Saving to: "amazon_data2.tsv.gz"

100%[=====] 1,512,355,451 20.5M/s  in 87s
2020-11-16 21:16:49 (16.6 MB/s) - "amazon_data2.tsv.gz" saved [1512355451/1512355451]

-bash-4.1$ wget -O amazon_data3.tsv.gz https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video.v1_00.tsv.gz
--2020-11-16 21:16:55-- https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video.v1_00.tsv.gz
Resolving s3.amazonaws.com... 52.216.131.101
Connecting to s3.amazonaws.com[52.216.131.101]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 138929896 (132M) [application/x-gzip]
Saving to: "amazon_data3.tsv.gz"

100%[=====] 138,929,896 15.7M/s  in 9.1s
2020-11-16 21:17:05 (14.6 MB/s) - "amazon_data3.tsv.gz" saved [138929896/138929896]
-bash-4.1$
```

Create work directory:

```
$ hdfs dfs -mkdir amazon_data  
$ hdfs dfs -ls
```

```
[ 100%[=====] 1,512,355,451 27.5M/s  in 42s  
2020-11-09 16:49:07 (34.5 MB/s) - "amazon_data2.tsv.gz" saved [1512355451/1512355451]  
-bash-4.1$ hdfs dfs -mkdir amazon_data  
-bash-4.1$ hdfs dfs -ls  
Found 1 items  
drwxr-xr-x  - nrobert hdfs          0 2020-11-09 16:49 amazon_data  
-bash-4.1$ |
```

Unzip and save .tsv files to HDFS:

```
$ gzip -d amazon_data1.tsv.gz  
$ gzip -d amazon_data2.tsv.gz  
$ gzip -d amazon_data3.tsv.gz
```

List local files:

We see 3 files:

- 1.2GB amazon_data1.tsv
- 3.7GB amazon_data2.tsv
- 337MB amazon_data3.tsv

```
$ ls -al
```

```
-bash-4.1$  
-bash-4.1$  
-bash-4.1$ ls  
amazon_data1.tsv  amazon_data2.tsv  amazon_data3.tsv  
-bash-4.1$ ls -al  
total 5209900  
drwx-----  2 nrobert nrobert      4096 Nov 16 21:26 .  
drwxr-xr-x  40 root    root      4096 Nov 11 22:13 ..  
-rw-rw-r--  1 nrobert nrobert 1288048833 Nov 24 2017 amazon_data1.tsv  
-rw-rw-r--  1 nrobert nrobert 3708889477 Nov 25 2017 amazon_data2.tsv  
-rw-rw-r--  1 nrobert nrobert 337970606 Nov 25 2017 amazon_data3.tsv  
-rw-----  1 nrobert nrobert     786 Nov 16 21:26 .bash_history  
-bash-4.1$
```

Save 3 data files into HDFS:

```
$ hdfs dfs -put amazon_data1.tsv amazon_data  
$ hdfs dfs -put amazon_data2.tsv amazon_data  
$ hdfs dfs -put amazon_data3.tsv amazon_data  
$ hdfs dfs -ls -h amazon_data
```

Review data in amazon_data directory on HDFS:

```
total 5209900
drwx----- 2 nrobert nrobert 4096 Nov 16 21:26 .
drwxr-xr-x 40 root root 4096 Nov 11 22:13 ..
-rw-rw-r-- 1 nrobert nrobert 1288048833 Nov 24 2017 amazon_data1.tsv
-rw-rw-r-- 1 nrobert nrobert 3708889477 Nov 25 2017 amazon_data2.tsv
-rw-rw-r-- 1 nrobert nrobert 337970606 Nov 25 2017 amazon_data3.tsv
-rw----- 1 nrobert nrobert 786 Nov 16 21:26 .bash_history
-bash-4.1$ hdfs dfs -put amazon_data1.tsv amazon_data
-bash-4.1$ hdfs dfs -put amazon_data2.tsv amazon_data
-bash-4.1$ hdfs dfs -put amazon_data3.tsv amazon_data
-bash-4.1$ hdfs dfs -ls -h amazon_data
Found 3 items
-rw-r--r-- 2 nrobert hdfs 1.2 G 2020-11-16 21:34 amazon_data/amazon_data1.tsv
-rw-r--r-- 2 nrobert hdfs 3.5 G 2020-11-16 21:35 amazon_data/amazon_data2.tsv
-rw-r--r-- 2 nrobert hdfs 322.3 M 2020-11-16 21:35 amazon_data/amazon_data3.tsv
-bash-4.1$
```

Remove files on local drive and list files so proof files were deleted:

```
$ rm amazon_data1.tsv
$ rm amazon_data2.tsv
$ rm amazon_data3.tsv
```

```
$ ls -al
```

```
total 5209900
drwx----- 2 nrobert nrobert 4096 Nov 16 21:26 .
drwxr-xr-x 40 root root 4096 Nov 11 22:13 ..
-rw-rw-r-- 1 nrobert nrobert 1288048833 Nov 24 2017 amazon_data1.tsv
-rw-rw-r-- 1 nrobert nrobert 3708889477 Nov 25 2017 amazon_data2.tsv
-rw-rw-r-- 1 nrobert nrobert 337970606 Nov 25 2017 amazon_data3.tsv
-rw----- 1 nrobert nrobert 786 Nov 16 21:26 .bash_history
-bash-4.1$ rm amazon_data1.tsv
-bash-4.1$ rm amazon_data2.tsv
-bash-4.1$ rm amazon_data3.tsv
rm: cannot remove 'amazon_data2.tsv': No such file or directory
-bash-4.1$ ls -al
total 12
drwx----- 2 nrobert nrobert 4096 Nov 16 21:38 .
drwxr-xr-x 40 root root 4096 Nov 11 22:13 ..
-rw----- 1 nrobert nrobert 786 Nov 16 21:26 .bash_history
-bash-4.1$ |
```

Merge 3 Amazon data files together, view size and view file contents:

-Combined file is **5.33GB** – combined.tsv

NOTE: You have to use your directory name.

```
$ hdfs dfs -getmerge /user/<your directory name>/amazon_data/
/home/<your directory name>/combined.tsv
$ ls -al
$ hdfs dfs -put combined.tsv /user/<your directory
name>/amazon_data
$ hdfs dfs -cat combined.tsv | head -20
```

```

MINGW64:/c/Users/ninag
-bash-4.1$ 
-bash-4.1$ 
-bash-4.1$ 
-bash-4.1$ 
-bash-4.1$ hdfs dfs -getmerge /user/nrobert/amazon_data/ /home/nrobert/combined.tsv
-bash-4.1$ ls -al
total 5250592
drwx-----. 2 nrobert nrobert 4096 Nov 16 21:54 .
drwxr-xr-x. 40 root root 4096 Nov 11 22:13 ..
-rw-----. 1 nrobert nrobert 786 Nov 16 21:26 .bash_history
-rw-r--r--. 1 nrobert nrobert 5334908916 Nov 16 21:55 combined.tsv
-rw-r--r--. 1 nrobert nrobert 41678984 Nov 16 21:55 .combined.tsvcrc
-bash-4.1$ cat combined.tsv | head -20
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 12190288 R3FV16928EP5TC B00AYB1482 668895143 Enlightened: season 1 Digital_Video_Download 5 0 0 N
Y I loved it and I wish there was a season 3 I loved it and I wish there was a season 3... I watched season 2 and loved that as well!
2015-08-31
US 30549954 R1IZHHS1MH3AQ4 B00KQD280M 246219280 Vicious Digital_Video_Download 5 0 0 N Y
A s always it seems that the best shows come from England As always it seems that the best shows come from England. best of the best without words
, i cant wait to watch season two. 2015-08-31
US 52895410 R52R85WC6TIAH B0148915LQ 534732318 After Words Digital_Video_Download 4 17 18 N Y
Charming movie This movie isn't perfect, but it gets a lot of things right. Yes, the librarian character played by Marcia Gay Harden is stereotypical and played a bit heavy-handed. But the universal nature of the story, the beautiful setting, and the likability of the characters overcomes this flaw. The quote at the end brought tears to my eyes. If you want to take a break from Hollywood's standard fare of dark, violent, or stupid movies, then give this a try. It is entertaining and thoughtful. 2015-08-31
US 27072354 R7HOOTVIBODS B008LOVTIK 239012694 Masterpiece: Inspector Lewis Season 5 Digital_Video_Download 5 0
0 N Y Five Stars excellant this is what tv should be 2015-08-31
US 26939022 R1XQZN5CD0ZGNX B0094LZMT0 535858974 On The Waterfront Digital_Video_Download 5 0 0 N
Y Brilliant film from beginning to end Brilliant film from beginning to end. All of the performances across the board are flawless. Doesn't get much better than this. 2015-08-31
US 4772040 R1HCST57W334KN B01120SQQE 38517795 Rick and Morty Season 2 Digital_Video_Download 5 5 6 N Y
Best show on TV right now If you don't like this show. Go back to your nickelback cd's you heathen 2015-08-31
US 12910040 R32BUTYQS1ZJBQ B000NPESSA 373323715 Africa Screams Digital_Video_Download 4 1 1 N Y
Very funny. A typical mid 50's comedy very funny. A typical mid 50's comedy. 2015-08-31
US 38805573 RH4SXPL4L9QU B00XWV4QXG 633842417 Entourage: Season 7 Digital_Video_Download 3 0 0 N
Y it was not as good as the series Strange as it is, it was not as good as the series. While it is a good movie, it could have been done better. 2015-08-31

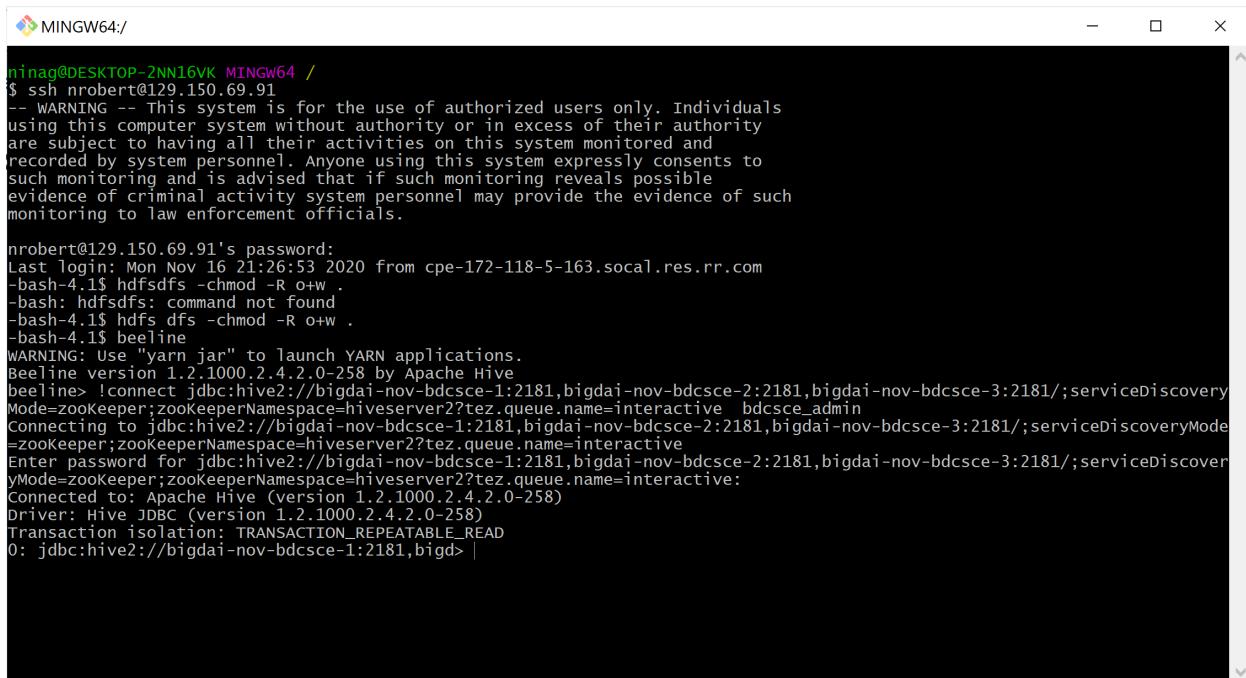
```

Step 2: Data Engineering and Analysis - Hadoop Hive

This step uses Hadoop Hive to create a table, load data into the table, view table contents and query data from the table.

New terminal – connect to HIVE:

```
$ hdfs dfs -chmod -R o+w .
$ beeline
$ get connection info from professor
$ use <your database name>;
```



```
MINGW64/
minag@DESKTOP-2NN16VK MINGW64 /
$ ssh nrobert@129.150.69.91
-- WARNING -- This system is for the use of authorized users only. Individuals
using this computer system without authority or in excess of their authority
are subject to having all their activities on this system monitored and
recorded by system personnel. Anyone using this system expressly consents to
such monitoring and is advised that if such monitoring reveals possible
evidence of criminal activity system personnel may provide the evidence of such
monitoring to law enforcement officials.

nrobert@129.150.69.91's password:
Last login: Mon Nov 16 21:26:53 2020 from cpe-172-118-5-163.socal.res.rr.com
-bash-4.1$ hdfsdfs -chmod -R o+w .
-bash: hdfsdfs: command not found
-bash-4.1$ hdfs dfs -chmod -R o+w .
-bash-4.1$ beeline
WARNING: Use "yarn jar" to launch YARN applications.
Beeline version 1.2.1000.2.4.2.0-258 by Apache Hive
beeline> !connect jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigdai-nov-bdcscse-2:2181,bigdai-nov-bdcscse-3:2181;/serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive_bdcscse_admin
Connecting to jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigdai-nov-bdcscse-2:2181,bigdai-nov-bdcscse-3:2181;/serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive
Enter password for jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigdai-nov-bdcscse-2:2181,bigdai-nov-bdcscse-3:2181;/serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive:
Connected to: Apache Hive (version 1.2.1000.2.4.2.0-258)
Driver: Hive JDBC (version 1.2.1000.2.4.2.0-258)
Transaction isolation: TRANSACTION_REPEATABLE_READ
0: jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigd> |
```

```
Create table to store data from "combined.tsv":
```

```
CREATE EXTERNAL TABLE if not exists amazon(
marketplace string,
customer_id string,
review_id string,
product_id string,
product_parent string,
product_title string,
product_category string,
star_rating int,
helpful_votes int,
total_votes int,
vine string,
verified_purchase string,
review_headline string,
review_body string,
review_date date)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE LOCATION '/user/<your database
name>/amazon_data/combined.tsv/'
TBLPROPERTIES ('skip.header.line.count'='1');
```

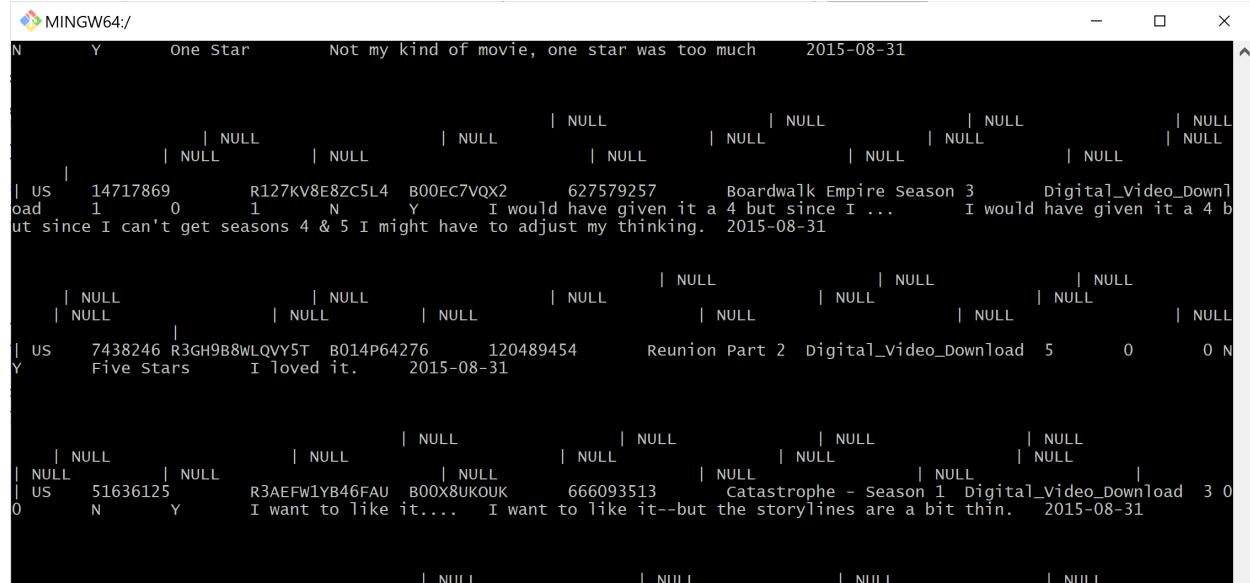
```
--Review table headers:
```

```
DESCRIBE amazon;
```

```
0: jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigd> review_headline string,
0: jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigd> review_body string,
0: jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigd> review_date date
0: jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigd> );
No rows affected (0.226 seconds)
0: jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigd> show tables;
+-----+---+
| tab_name |
+-----+---+
| amazon   |
+-----+---+
1 row selected (0.189 seconds)
0: jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigd> describe amazon;
+-----+-----+-----+
| col_name | data_type | comment |
+-----+-----+-----+
| marketplace | string |           |
| customer_id | string |           |
| review_id | string |           |
| product_id | string |           |
| product_parent | string |           |
| product_title | string |           |
| product_category | string |           |
| star_rating | int |           |
| helpful_votes | int |           |
| total_votes | int |           |
| vine | string |           |
| verified_purchase | string |           |
| review_headline | string |           |
| review_body | string |           |
| review_date | date |           |
+-----+-----+-----+
15 rows selected (0.213 seconds)
0: jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigd>
```

Viewing 20 rows of data:

```
SELECT * from amazon LIMIT 20;
```

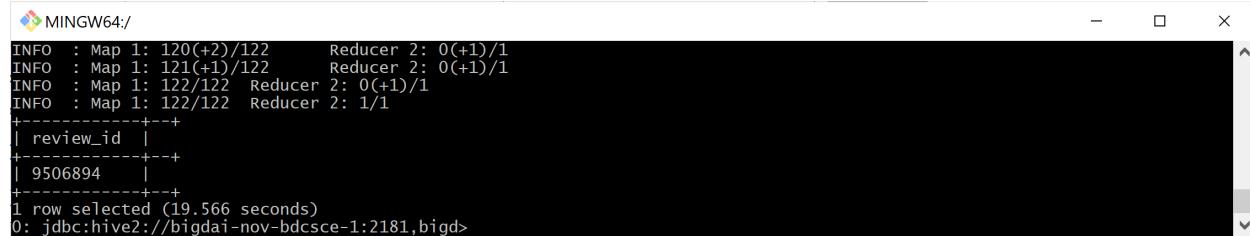


The screenshot shows a terminal window titled "MINGW64:/". It displays 20 rows of data from the "amazon" table. The columns are separated by vertical lines and include review_id, helpfulness, title, content, and various dates. Some fields contain NULL values.

	N	Y	One Star	Not my kind of movie, one star was too much	2015-08-31															
			NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	
1	us	14717869	0	R127KV8E8ZC5L4	B00EC7VQX2	627579257	Boardwalk Empire Season 3	Digital_Video_Download												
2	us	1	1	N	Y	I would have given it a 4 but since I ...	I would have given it a 4 b	ut since I can't get seasons 4 & 5 I might have to adjust my thinking.	2015-08-31											
3	us	7438246	R3GH9B8WLQVY5T	B014P64276	120489454	Reunion Part 2	Digital_Video_Download	5	0	0	N									
4	Y	Five Stars	I loved it.	2015-08-31																
5	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	
6	us	51636125	R3AEFW1YB46FAU	B00X8UKOUK	666093513	Catastrophe - Season 1	Digital_Video_Download	3	0	0	N									
7	0	N	Y	I want to like it....	I want to like it--but the storylines are a bit thin.	2015-08-31														
8	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	

We can query the number of reviews in the data:

```
SELECT COUNT(*) review_id FROM amazon;
```



The screenshot shows a terminal window titled "MINGW64:/". It displays the result of a query to count the number of reviews in the "amazon" table. The output shows 1 row selected in 19.566 seconds, with the result being 9506893.

```
INFO : Map 1: 120(+2)/122 Reducer 2: 0(+1)/1
INFO : Map 1: 121(+1)/122 Reducer 2: 0(+1)/1
INFO : Map 1: 122/122 Reducer 2: 0(+1)/1
INFO : Map 1: 122/122 Reducer 2: 1/1
+---+---+
| review_id |
+---+---+
| 9506894 |
+---+---+
1 row selected (19.566 seconds)
0: jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigd>
```

NOTE: You have to use your directory name.

Use “ANALYZE” function to view statistics on table “amazon”:

```
ANALYZE TABLE amazon COMPUTE STATISTICS;
```

[Table <your database name>.amazon stats: [numFiles=1, numRows=9506893, totalsize=5334908916, rawDataSize=5325401831]

```
INFO : Map 1: 0(+1)/1
INFO : Map 1: 1/1
INFO : Table nrobert.amazon stats: [numFiles=1, numRows=9506893, totalsize=5334908916, rawDataSize=5325401831]
No rows affected (111.466 seconds)
0: jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigd>
```

How many movie/TV titles are in the data?

-Count distinct titles

```
SELECT COUNT(DISTINCT product_title) FROM amazon;
```

-407,787 distinct titles exist

```
INFO : Map 1: 1/1    Reducer 2: 79(+1)/80    Reducer 3: 0(+1)/1
INFO : Map 1: 1/1    Reducer 2: 80/80        Reducer 3: 0(+1)/1
INFO : Map 1: 1/1    Reducer 2: 80/80        Reducer 3: 1/1
+-----+-----+
|   c0   |
+-----+-----+
| 407787 |
+-----+-----+
1 row selected (44.712 seconds)
0: jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigd> |
```

--Review years and counts:

```
SELECT DATE_FORMAT(REVIEW_DATE, 'YYYY') AS YEAR,
       COUNT(*) AS YEAR_COUNT
  FROM AMAZON
 GROUP BY DATE_FORMAT(REVIEW_DATE, 'YYYY') AS YEAR,
 ORDER BY YEAR;
```

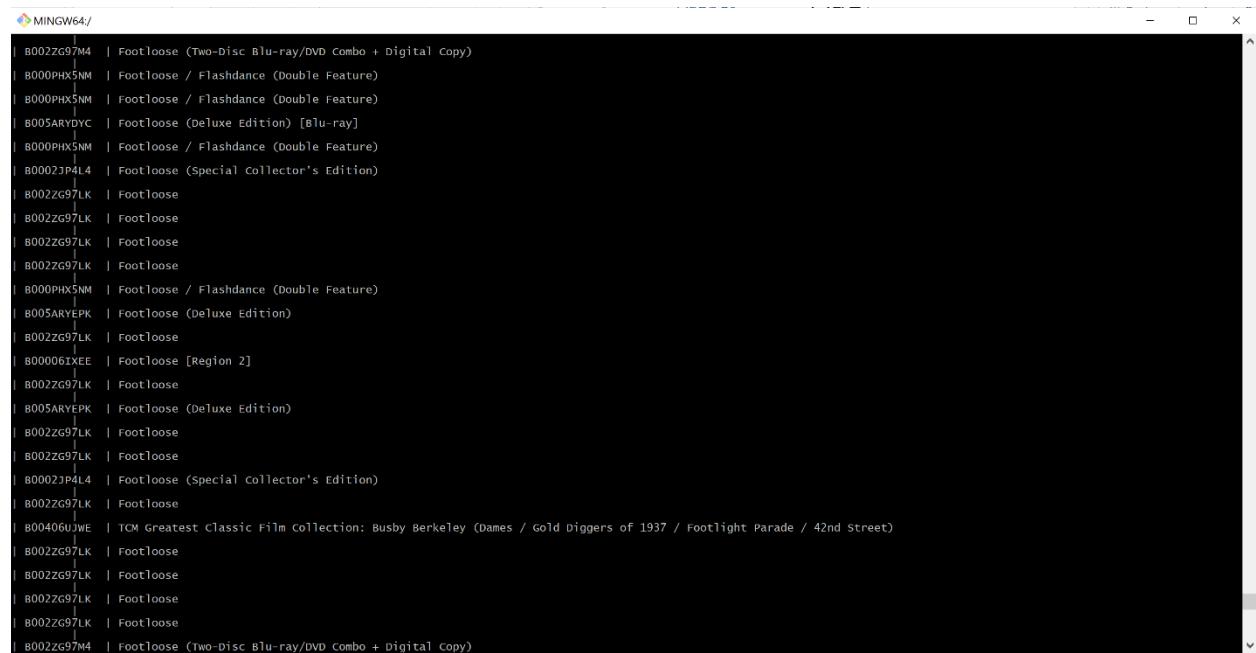
```
INFO : Map 1: 1/1    Reducer 2: 80/80    Reducer 3: 1/1
+-----+-----+
| year | year_count |
+-----+-----+
| NULL | 2           |
| 1995 | 2           |
| 1996 | 14          |
| 1997 | 17          |
| 1998 | 5766         |
| 1999 | 57811        |
| 2000 | 119044       |
| 2001 | 117845       |
| 2002 | 13270         |
| 2003 | 144016         |
| 2004 | 165303         |
| 2005 | 188089         |
| 2006 | 187286         |
| 2007 | 251490         |
| 2008 | 245966         |
| 2009 | 140110         |
| 2010 | 262847         |
| 2011 | 306162         |
| 2012 | 507616         |
| 2013 | 1518498        |
| 2014 | 2548863        |
| 2015 | 2481893        |
+-----+-----+
22 rows selected (58.754 seconds)
0: jdbc:hive2://bigdai-nov-bdcscse-1:2181,bigd> |
```

-Does a title have one Product_ID or many? How should we group products together to determine ratings?

-In the case of the movie “Footloose” there are many product_ids associated with this title and we see that repeating with other product_title(s) in the data.

Example 1 – Title = “Footloose”

```
SELECT * FROM AMAZON WHERE PRODUCT_TITLE = 'Footloose'  
order by review_date desc;
```



The screenshot shows a terminal window titled "MINGW64/" displaying the results of a SQL query. The query selects all columns from the "AMAZON" table where the "PRODUCT_TITLE" is 'Footloose', ordered by "review_date" in descending order. The results are listed in a table format with two columns: the first column contains the product ID (e.g., B002ZG97M4, B000PHX5NM, B000PHX5NM, etc.) and the second column contains the product title (e.g., 'Footloose (Two-Disc Blu-ray/DVD Combo + Digital Copy)', 'Footloose / Flashdance (Double Feature)', 'Footloose / Flashdance (Double Feature)', etc.). The window has a standard Windows-style title bar and scroll bars on the right side.

Product ID	Product Title
B002ZG97M4	'Footloose (Two-Disc Blu-ray/DVD Combo + Digital Copy)
B000PHX5NM	'Footloose / Flashdance (Double Feature)
B000PHX5NM	'Footloose / Flashdance (Double Feature)
B005ARYDYC	'Footloose (Deluxe Edition) [Blu-ray]
B000PHX5NM	'Footloose / Flashdance (Double Feature)
B0002JP4L4	'Footloose (Special Collector's Edition)
B002ZG97LK	'Footloose
B000PHX5NM	'Footloose / Flashdance (Double Feature)
B005ARYEPK	'Footloose (Deluxe Edition)
B002ZG97LK	'Footloose
B000061XEE	'Footloose [Region 2]
B002ZG97LK	'Footloose
B005ARYEPK	'Footloose (Deluxe Edition)
B002ZG97LK	'Footloose
B002ZG97LK	'Footloose
B0002JP4L4	'Footloose (Special Collector's Edition)
B002ZG97LK	'Footloose
B0040GUJWE	'TCM Greatest Classic Film Collection: Busby Berkeley (Dames / Gold Diggers of 1937 / Footlight Parade / 42nd Street)
B002ZG97LK	'Footloose
B002ZG97LK	'Footloose
B002ZG97LK	'Footloose
B002ZG97M4	'Footloose (Two-Disc Blu-ray/DVD Combo + Digital Copy)

Are movie/TV titles distinct? Or are there various spellings for each title?

Likewise, there are various product_title for this one movie. This means we will need to normalize or standardize the data before we can group titles correctly.

Example 2 – Title = “It’s A Wonderful Life”

```
SELECT product_title, COUNT(product_title)
FROM      AMAZON
WHERE
    product_title LIKE '%It's A Wonderful Life%'
GROUP BY
    product_title
ORDER BY
    product_title;
```

product_title	c1
Alvin and the Chipmunks: It's a Wonderful Life Dave [VHS]	1
Archie and the Gang [VHS]	1
Archie and the Gang [VHS] / It's a Wonderful Life (1946) / Mr. Smith Goes to Washington (1939) (Three Disc Set, Asian Import, Plays All-Regions)	2
Collector's Pack (The Bells of St. Mary's / It's a Wonderful Life)	85
Classic Christmas Collection (It's a wonderful Life / White Christmas)	1
Classic Christmas Collection: It's a wonderful Life & White Christmas	1
Frank Capra Presents A Personal Remembrance Plus The Authorized Uncut Version of It's A wonderful Life	2
Frank Capra's "It's a wonderful Life"	2
Frank Capra's It's A wonderful Life (Colorized Edition)	5
Franz Kafka's It's a wonderful Life	2
Franz Kafka's It's a wonderful Life ... and Other Strange Tales [VHS]	13
It Was a wonderful Life	1
It Was a wonderful Life [VHS]	1
It's A wonderful Life	2
It's A wonderful Life (1946) Crown Movie Classics - Black and White	1
It's A wonderful Life (Black & White Version)	1
It's A wonderful Life (Deluxe 50th Anniversary Edition) [VHS]	573
It's A wonderful Life (Two-Disc Collector's Set)	2
It's A wonderful Life [DVD] 1946 - James Stewart (Region Free, PAL)	632
It's A wonderful Life: 45th Anniversary Collector's Edition	1
It's a wonderful Life	250
It's a wonderful Life (60th Anniversary Edition)	725
It's a wonderful Life (Capra's Uncut Classic in Digital Color)	7
It's a wonderful Life (Colorized Version) [VHS]	1
It's a wonderful Life (DVD)	10
It's a wonderful Life (In color) (Hal Roach Film Classics)	1
It's a wonderful Life (Mandarin Chinese Edition)	9
It's a wonderful Life (Original Uncut Version) [VHS]	1
It's a wonderful Life (Two-Disc Collector's Gift Set And Limited Edition Ornament)	49
It's a wonderful Life / Miracle on 34th Street	4
It's a wonderful Life Giftset (Blu-ray + Bell Ornament)	10
It's a wonderful Life [Blu-ray]	309
It's a wonderful Life [VHS]	75
It's a wonderful Life/Miracle on 34th Street [VHS]	1
It's a wonderful Life/Miracle on 34th [VHS]	1
It's a wonderful Life 50th Anniversary [VHS]	4
It's a wonderful Life [VHS]	18
James Stewart's wonderful Life [VHS]	5
Movie Favorite, It's a Wonderful Life, Starring James Stewart/Donna Reed	1
The Best of Soap - Jessica's wonderful Life [VHS]	4
The Little Richard Collection (The Young Ones / Summer Holiday / Wonderful Life)	24
Wonderful Life [VHS]	5
Wonderful Life (Korean Drama) DVD-9	1
Wonderful Life (Korean Drama, English Sub, All Region DVD, 16 Episodes End, 6DVD Set)	1
Wonderful Life - Korean Drama (5 DVD) All Region with English Subtitles	3

We can see that the product_title field in these two examples has these types of notations which will prevent us from grouping successfully on product_title:

- [Blu-Ray]
- [VHS]
- [Theatrical Release]
- [DVD + Digital Copy + UltraViolet]
- [DVD + Digital]
- [Region 2]
- [Ultra HD]
- [PAL]
- [Blu-ray + Digital Copy]
- (Blu-ray/DVD Combo)

Assumptions:

- For the most part, TV listings do not seem problematic for popular titles.
- For movies between 1990 and 2011, Amazon saved the movie year in the product_title field.
- We discovered that if you paste the product_id into amazon.com it will link you to the title.

NOTE: The following will prevent accurate grouping.

- product_title contains both "A" and "a"
- product_title contains both "The" and "the"
- product_title contains both "Its" and "It's"

To delete text inside either brackets or parenthesis in product_title:

```
SELECT product_title =  
regexp_replace(product_title, '\s*\[\[^()]*\]', '')  
FROM amazon;  
  
SELECT product_title =  
regexp_replace(product_title, '\s*\(\[^()]*\)', '')  
FROM amazon;
```

Pitfalls of this method:

There will be some titles that will erroneously be grouped together. These examples were identified and manually fixed:

- Little Women
- Midway

Next query top movie titles:

Run query to select top 100 movie reviews by average star rating and count of ratings.

```
SELECT
product_title,
ROUND(avg(star_rating), 2) as average_rating,
COUNT(review_id) as review_count
FROM amazon
WHERE (review_date >= '01-Jan-05' AND review_date <= '31-Dec-15') and (product_title != 'Pilot')
GROUP by product_title HAVING avg(star_rating) >= 4.5
ORDER by review_count DESC
LIMIT 100;
```

MINGW64:/c/Users/ninag

NFO : Map 1: 1/1 Reducer 2: 9/9 Reducer 3: 1/1

product_title	average_rating	review_count
Bosch Season 1	4.61	47983
Downton Abbey Season 3	4.86	22457
Downton Abbey Season 1	4.87	19497
Justified Season 1	4.68	22155
Downton Abbey Season 2	4.89	13356
Mozart in the Jungle Season 1	4.51	13259
Orphan Black Season 1	4.68	13187
Downton Abbey Season 4	4.84	13000
Vikings Season 1	4.68	11911
The Americans Season 1	4.62	11145
Grimm Season 1	4.65	11119
Alpha House Season 2	4.68	9934
Vikings Season 2	4.82	9761
Suits Season 1	4.77	9748
The Good Wife, Season 1	4.73	9552
The Walking Dead, Season 5	4.76	9507
Game of Thrones - In the High Castle - Season 1	4.73	9081
Band of Brothers Season 1	4.92	9062
Sneaky Pete - Season 1	4.61	8925
Justified Season 4	4.79	8873
Justified Season 2	4.81	8830
Downton Abbey Season 5	4.86	7919
Marvel's The Avengers	4.52	7827
Orphan Black, Season 2	4.79	7810
Sons of Anarchy Season 7	4.72	7525
American Sniper	4.55	7119
The Pacific Season 1	4.62	6994
Masterpiece: Mr. Selfridge Season 1 Original UK Edition	4.54	6830
Breaking Bad The Final Season	4.89	6689
The Wire Season 1	4.52	6676
Justified Season 3	4.79	6611
Grimm Season 2	4.8	6542
The White Queen - Season 1	4.54	6441

Working in Hadoop Hive you can create a table with your data:

Create tmp directory in HDFS:

```
$ hdfs dfs -mkdir tmp;
```

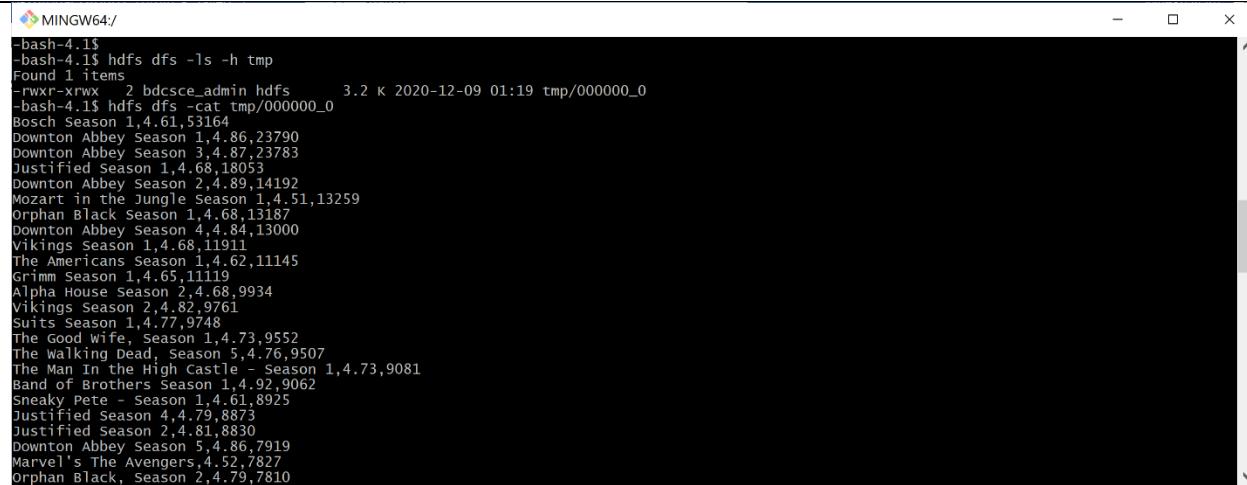
Create Top 100 table and save as comma delimited file in HDFS tmp directory:

NOTE: You have to use your directory name.

```
CREATE TABLE IF NOT EXISTS amazon_top
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ","
STORED AS TEXTFILE LOCATION "/user/<your directory name>/tmp"
AS
SELECT
product_title,
ROUND(avg(star_rating), 2) as average_rating,
COUNT(review_id) as review_count
FROM amazon
WHERE (review_date >= '01-Jan-05' AND review_date <= '31-Dec-15') and (product_title != 'Pilot')
GROUP by product_title HAVING avg(star_rating) >= 4.5
ORDER by review_count DESC
LIMIT 100;
```

View Top 100 file:

```
$ hdfs dfs -cat tmp/000000_0
```



```
MINGW64:
$ hdfs dfs -cat tmp/000000_0
-bash-4.1$ hdfs dfs -ls -h tmp
Found 1 items
-rwxr-xrwx 2 bdcsce.admin hdfs 3.2 K 2020-12-09 01:19 tmp/000000_0
-bash-4.1$ hdfs dfs -cat tmp/000000_0
Bosch Season 1,4.61,53164
Downton Abbey Season 1,4.86,23790
Downton Abbey Season 3,4.87,23783
Justified Season 1,4.68,18053
Downton Abbey Season 2,4.89,14192
Mozart in the Jungle Season 1,4.51,13259
Orphan Black Season 1,4.68,13187
Downton Abbey Season 4,4.84,13000
Vikings Season 1,4.68,11911
The Americans Season 1,4.62,11145
Grimm Season 1,4.65,11119
Alpha House Season 2,4.68,9934
Vikings Season 2,4.82,9761
Suits Season 1,4.77,9748
The Good Wife, Season 1,4.73,9552
The Walking Dead, Season 5,4.76,9507
The Man In the High Castle - Season 1,4.73,9081
Band of Brothers Season 1,4.92,9062
Sneaky Pete - Season 1,4.61,8925
Justified Season 4,4.79,8873
Justified Season 2,4.81,8830
Downton Abbey Season 3,4.86,7919
Marvel's The Avengers,4.52,7827
Orphan Black, Season 2,4.79,7810
```

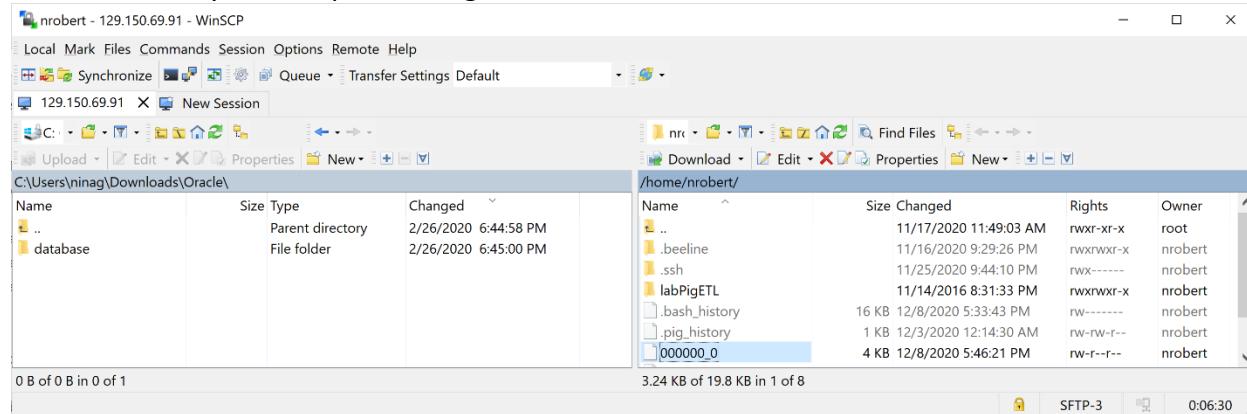
Export File to your local drive and then computer using WinSCP, putty or other method.

Move file to local drive

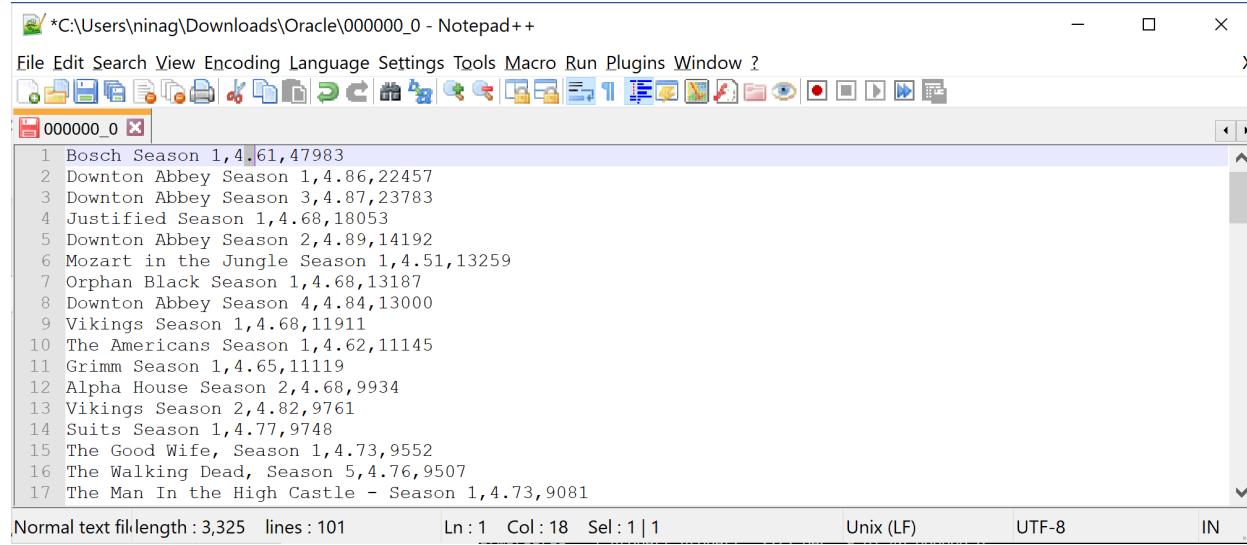
```
hdfs dfs -get tmp/000000_0
```

```
-bash-4.1$ ls -al
total 56
drwx----- 5 nrobert nrobert 4096 Dec  9 01:46 .
drwxr-xr-x  42 root   root   4096 Nov 17 19:49 ..
-rw-r--r--  1 nrobert nrobert 3325 Dec  9 01:46 000000_0
-rw-r--r--  1 nrobert nrobert 15511 Dec  9 01:33 .bash_history
drwxrwxr-x  2 nrobert nrobert 4096 Nov 17 05:29 .beeline
-rw-r--r--  1 nrobert nrobert 1119 Dec  3 06:24 high_cost_sites2
drwxrwxr-x  4 nrobert nrobert 4096 Nov 15  2016 labPigETL
-rw-r--r--  1 nrobert nrobert 1288 Dec  3 07:33 pig_1606980811776.log
```

Transfer file to your computer using WinSCP



View or edit file on computer



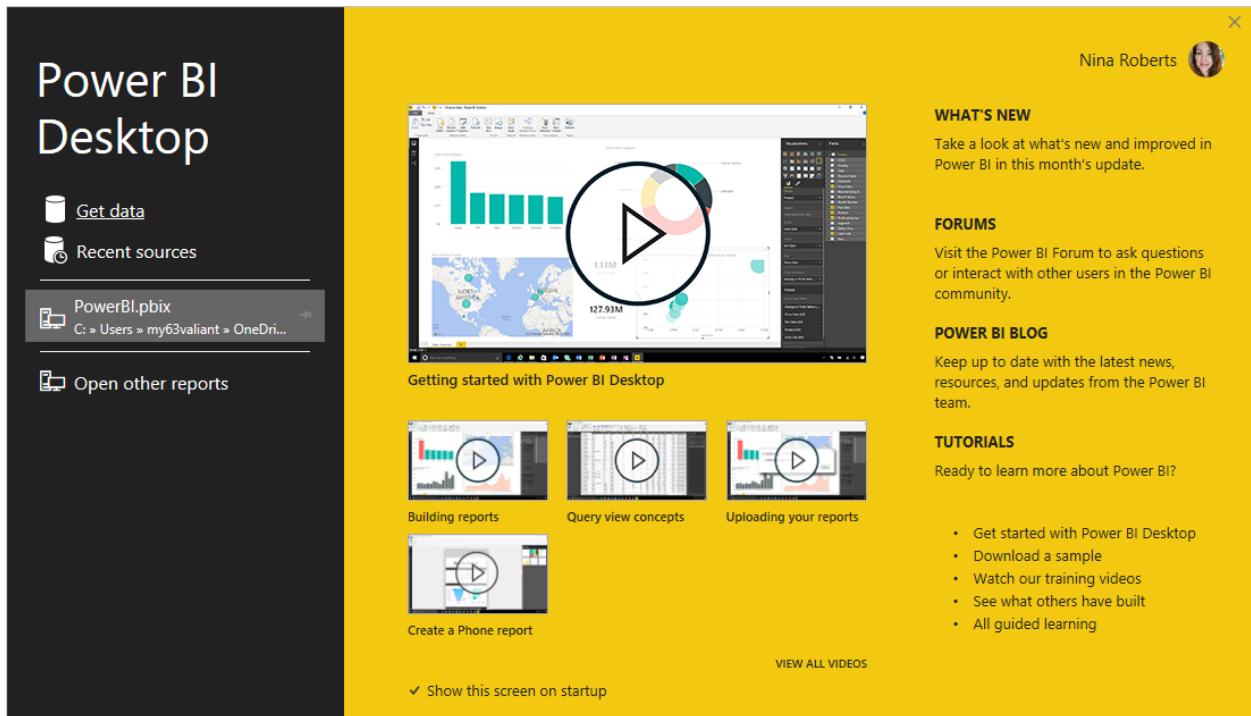
Step 3 : Visualization

Visualization #1 – Power BI

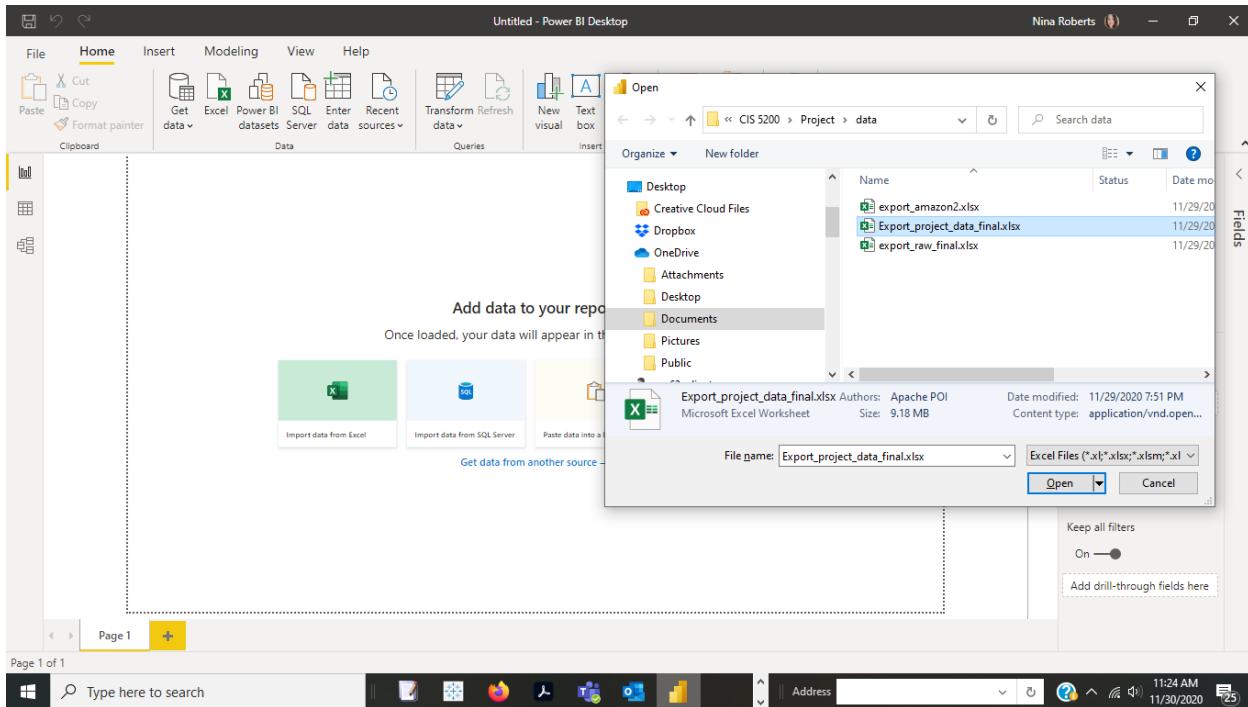
Download and install Power BI Desktop

<https://www.microsoft.com/en-us/download/details.aspx?id=58494>

-Click “Get data” in left hand menu



- Click “import data from Excel” and browse to a data file
- Click “Open”

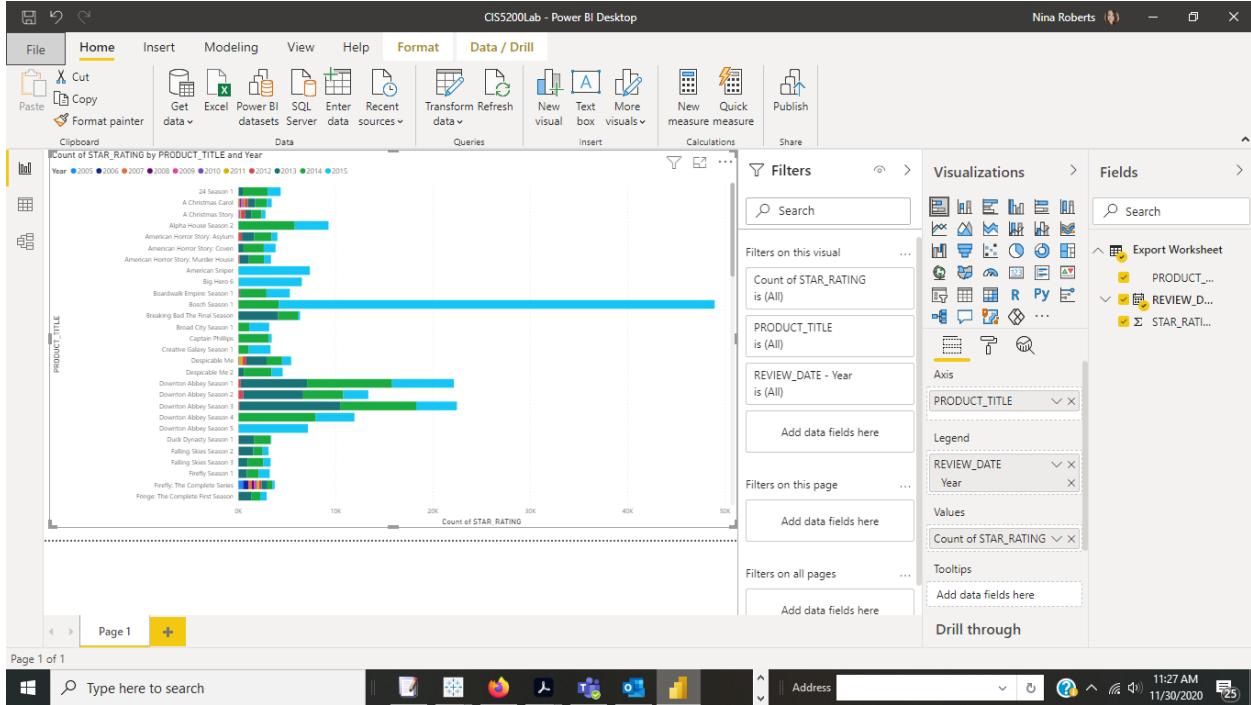


- Click sheet in Excel workbook that contains the data
- Click “Load”

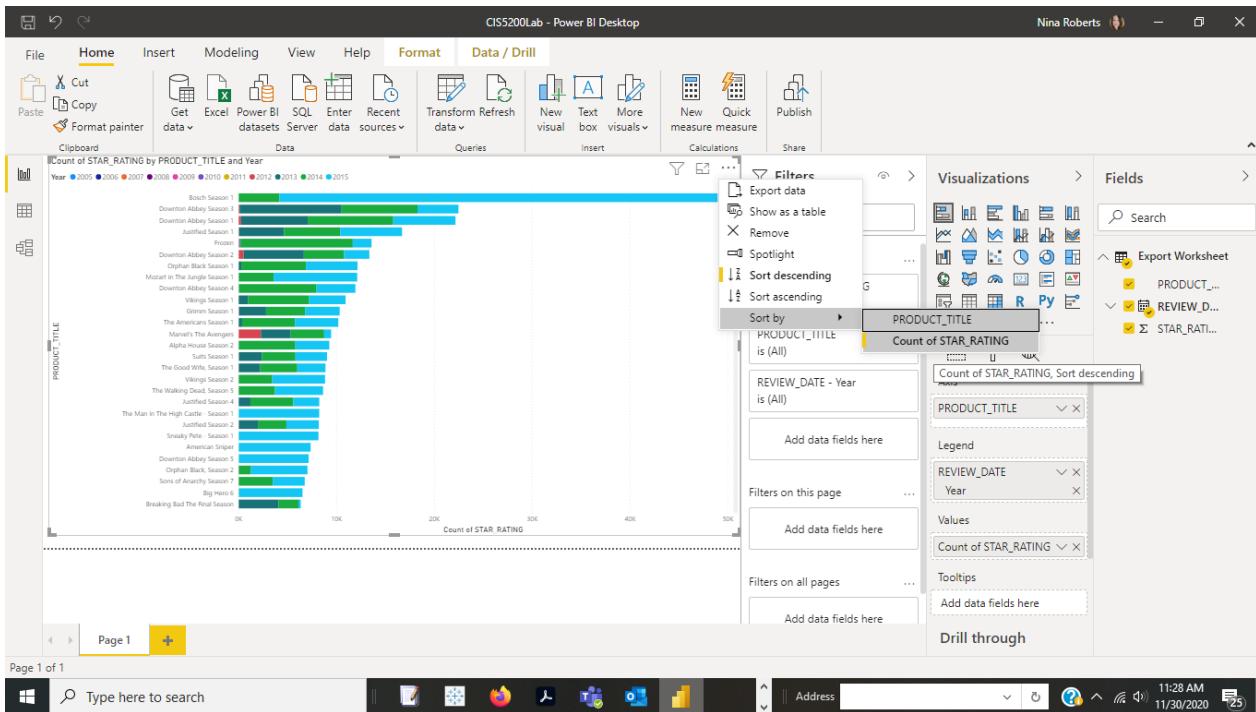
The screenshot shows the Power BI Desktop application window. The 'Navigator' pane is open, showing the 'Export Worksheet' section. A table of data is displayed with columns: PRODUCT_TITLE, STAR_RATING, and REVIEW_DATE. The data includes various Christmas-related titles like 'A Christmas Carol', 'A Christmas Story', and 'National Lampoon's Christmas Vacation'. The main Power BI interface shows a blank report canvas with a dotted grid, a ribbon menu at the top, and a search bar at the bottom.

PRODUCT_TITLE	STAR_RATING	REVIEW_DATE
A Christmas Carol	5	8/11/2015
A Christmas Story	5	8/8/2015
A Christmas Story	5	7/28/2015
A Christmas Story	5	7/18/2015
A Christmas Carol	5	6/28/2015
A Christmas Carol	4	4/19/2015
A Christmas Carol	5	3/18/2015
A Christmas Story	5	3/8/2015
A Christmas Carol	3	3/8/2015
A Christmas Story	5	3/5/2015
National Lampoon's Christmas Vacation	5	2/25/2015
National Lampoon's Christmas Vacation	5	2/14/2015
A Christmas Carol	5	2/7/2015
A Christmas Carol	2	1/20/2015
A Christmas Carol	5	1/3/2015
National Lampoon's Christmas Vacation	1	12/31/2014
A Christmas Carol	5	12/30/2014
A Christmas Story	5	9/13/2014
National Lampoon's Christmas Vacation	5	8/15/2014
A Christmas Carol	5	7/14/2014
A Christmas Carol	5	7/3/2014
A Christmas Carol	3	5/18/2014
A Christmas Carol	5	2/22/2014
A Christmas Carol	5	1/30/2014

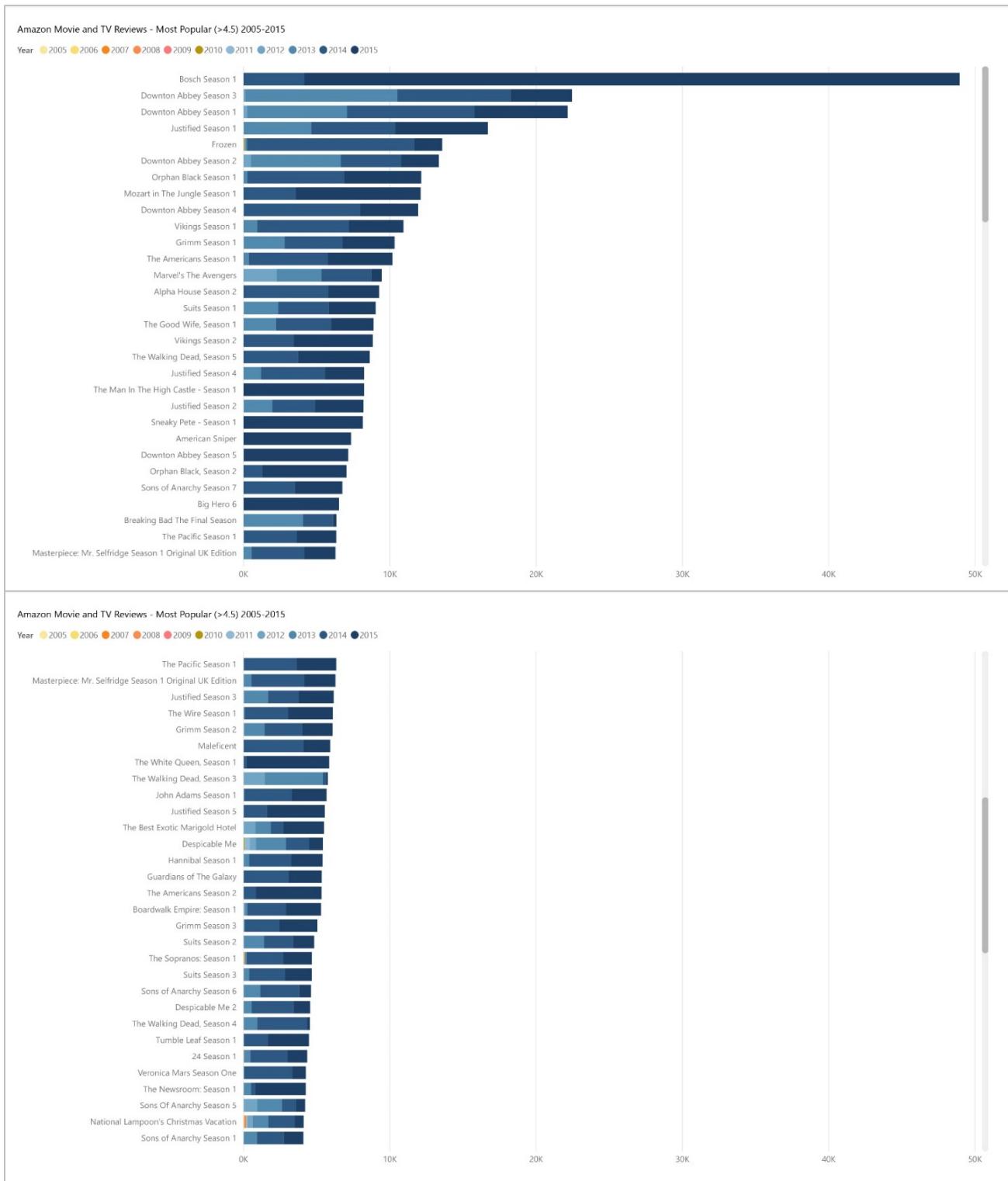
- Select “stacked bar chart” option in “Visualizations” menu
- Assign PRODUCT_TITLE, REVIEW_DATE(YEAR), AND COUNT OF STAR_RATING as graph values

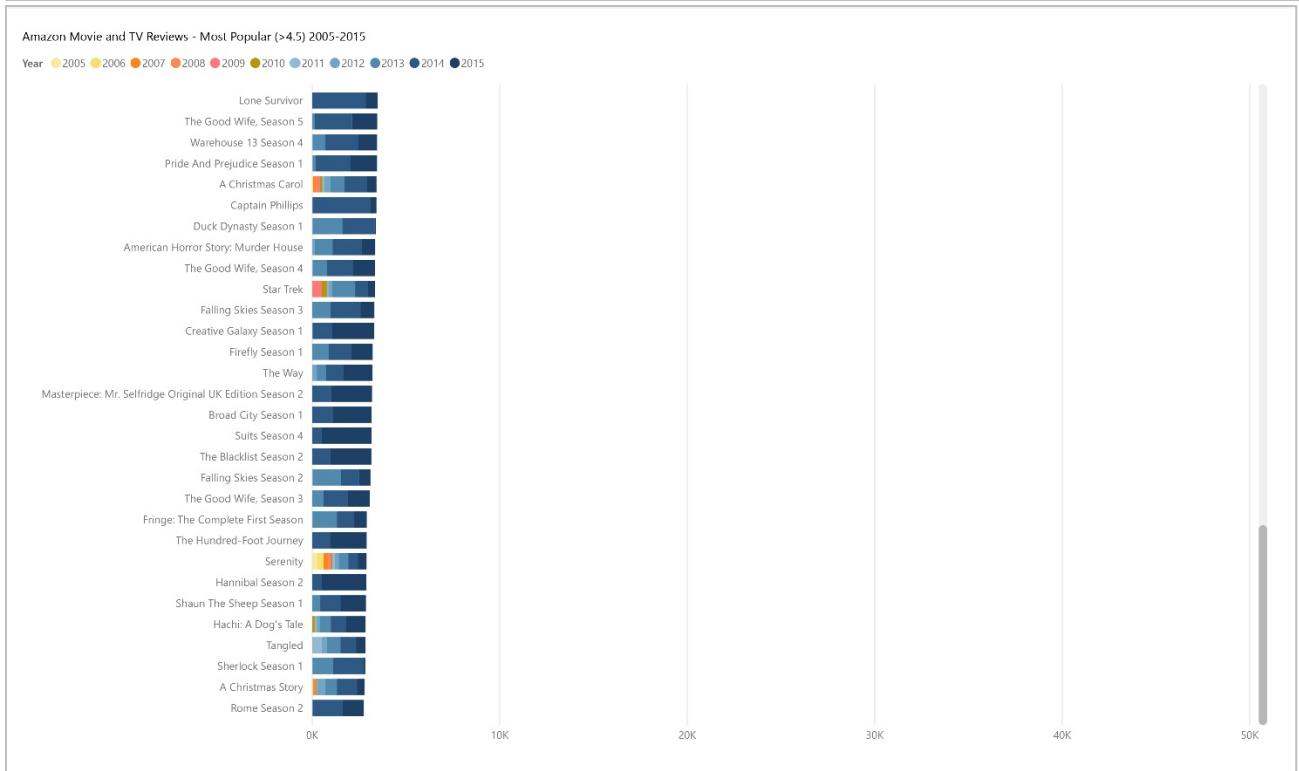
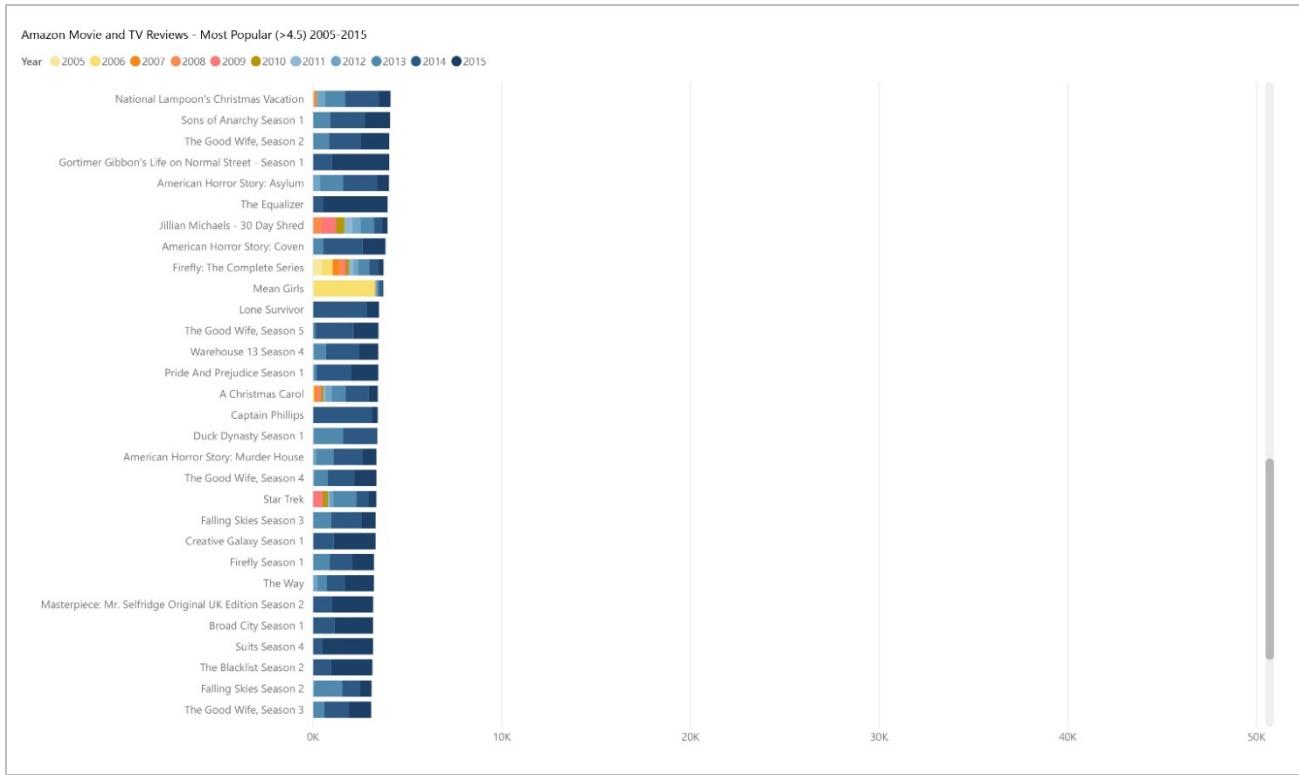


- Select menu under three dots in upper right hand corner to sort by “Count of STAR_RATING”



-Format colors, title and legend and export or print



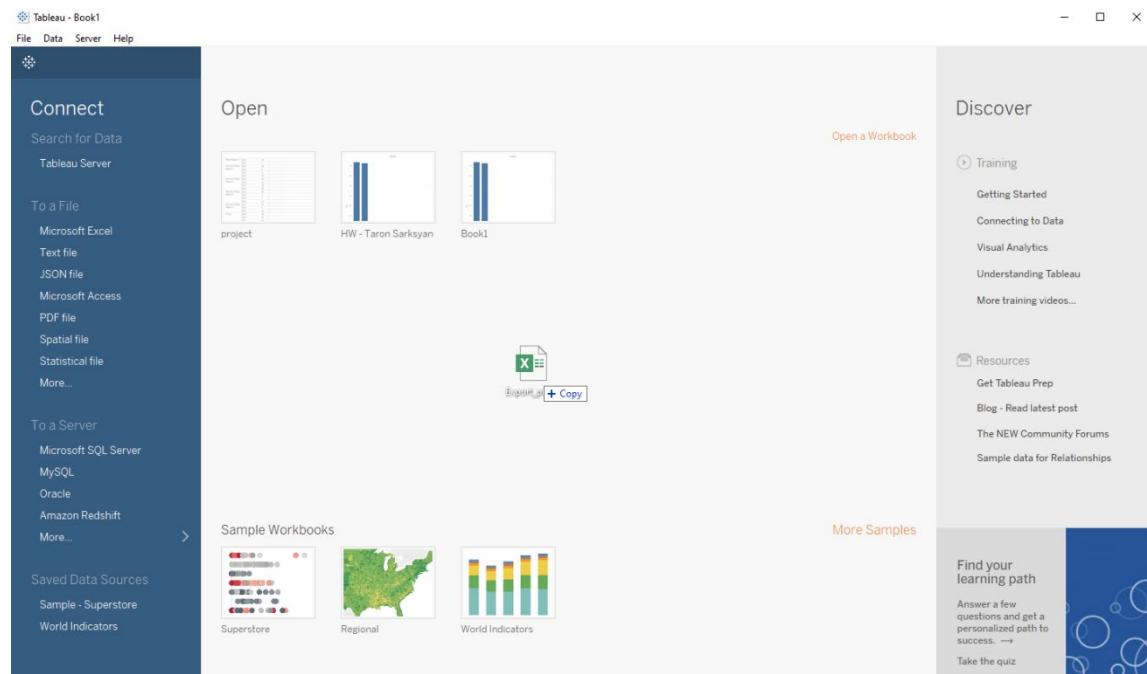


Visualization #2 – Tableau (Animated Data Visualization)

(Total Count of Star Ratings vs Product Title/Year/Rank)

Step 1:

By starting on the animated visualization graph, first you open Tableau and then drag your Excel file to Tableau.



Step 2:

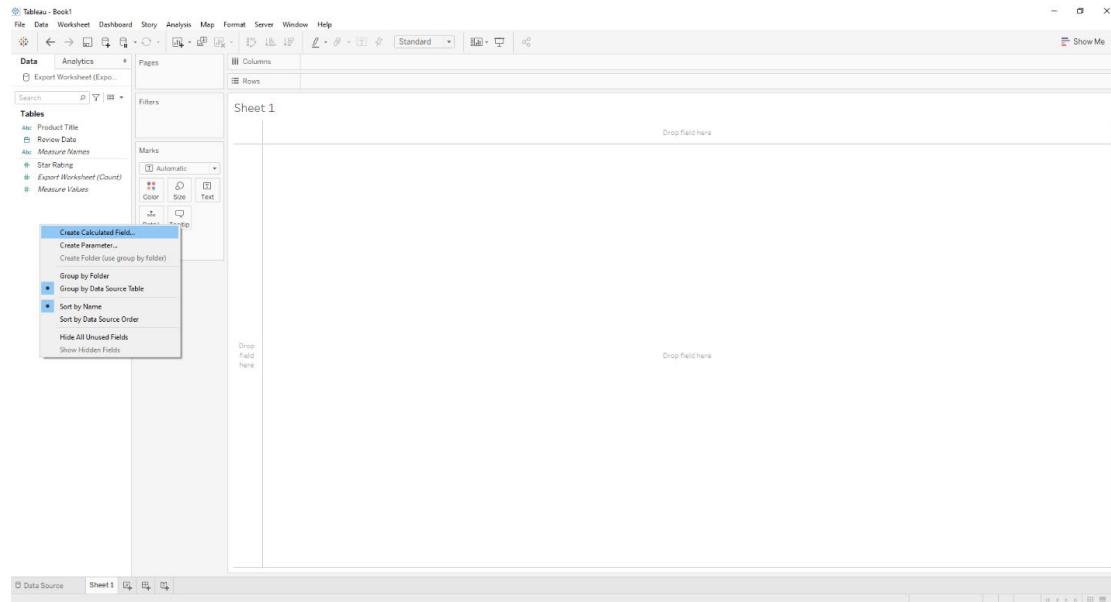
After making sure the data displayed is correct, create a new sheet by clicking on Sheet1.

This screenshot shows a Tableau worksheet titled 'New Union'. It displays a data table with columns: 'Product Title', 'Star Rating', and 'Review Date'. The data includes various TV show titles and their ratings. At the bottom, there are navigation buttons: 'Data Source', 'Sheet1' (which is highlighted in orange), and other sheet icons. A small tooltip 'Go to Worksheet' appears over the sheet icons.

Product Title	Star Rating	Review Date
Justified Season 2	5	8/16/2015
Sneaky Pete - Season 1	5	8/16/2015
The Newsroom: Seaso...	5	8/16/2015
Downton Abbey Seas...	5	8/16/2015
The Newsroom: Seaso...	5	8/16/2015
Sneaky Pete - Season 1	5	8/16/2015
The Newsroom: Seaso...	4	8/16/2015
Veronica Mars Season...	5	8/16/2015
Sneaky Pete - Season 1	5	8/16/2015
Justified Season 3	5	8/16/2015
Sneaky Pete - Season 1	5	8/16/2015
Bosch Season 1	5	8/16/2015
Downton Abbey Seas...	3	8/16/2015
The Newsroom: Seaso...	3	8/16/2015
Tumble Leaf Season 1	5	8/16/2015

Step 3:

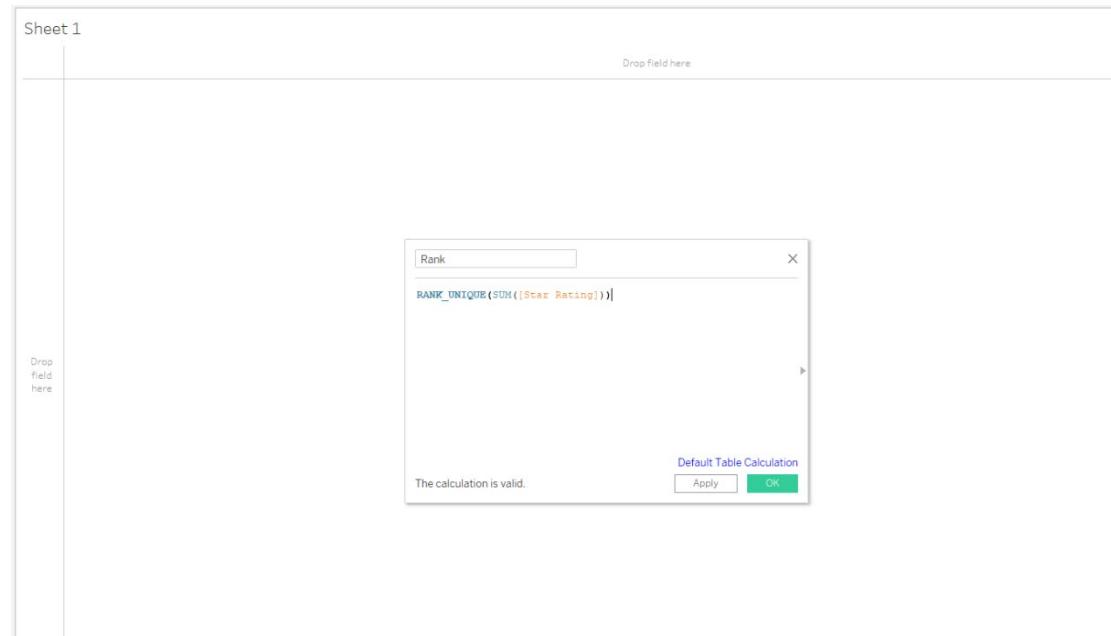
Right click on the empty area under tables to create a calculated field for “Rank.”



Step 4:

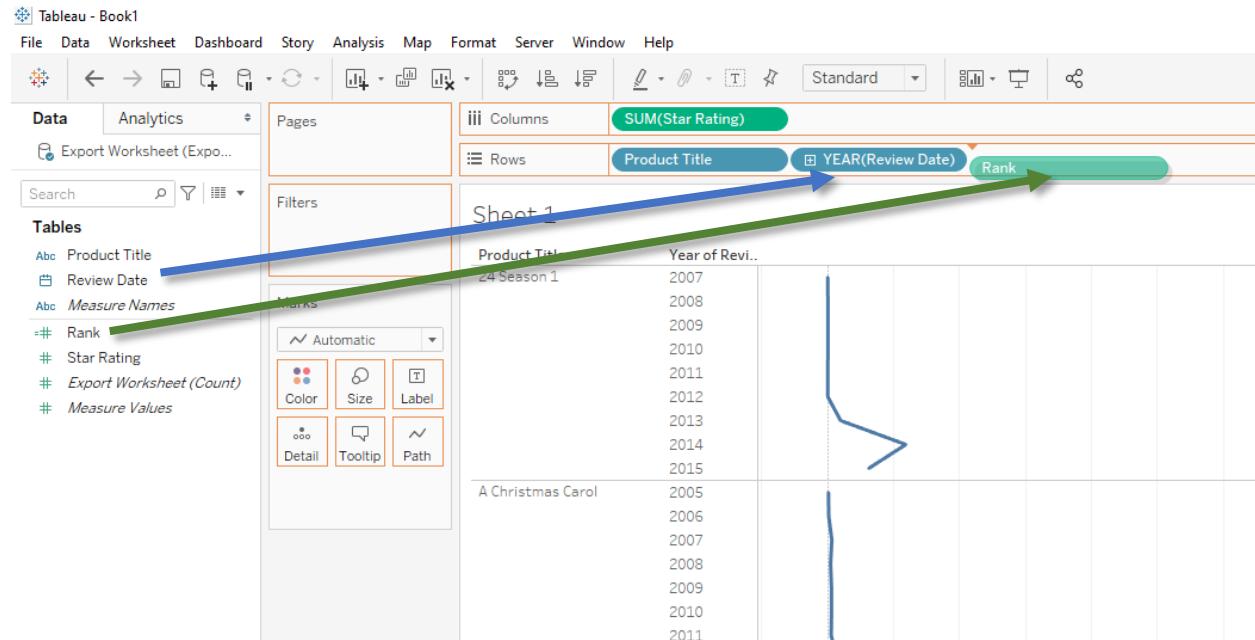
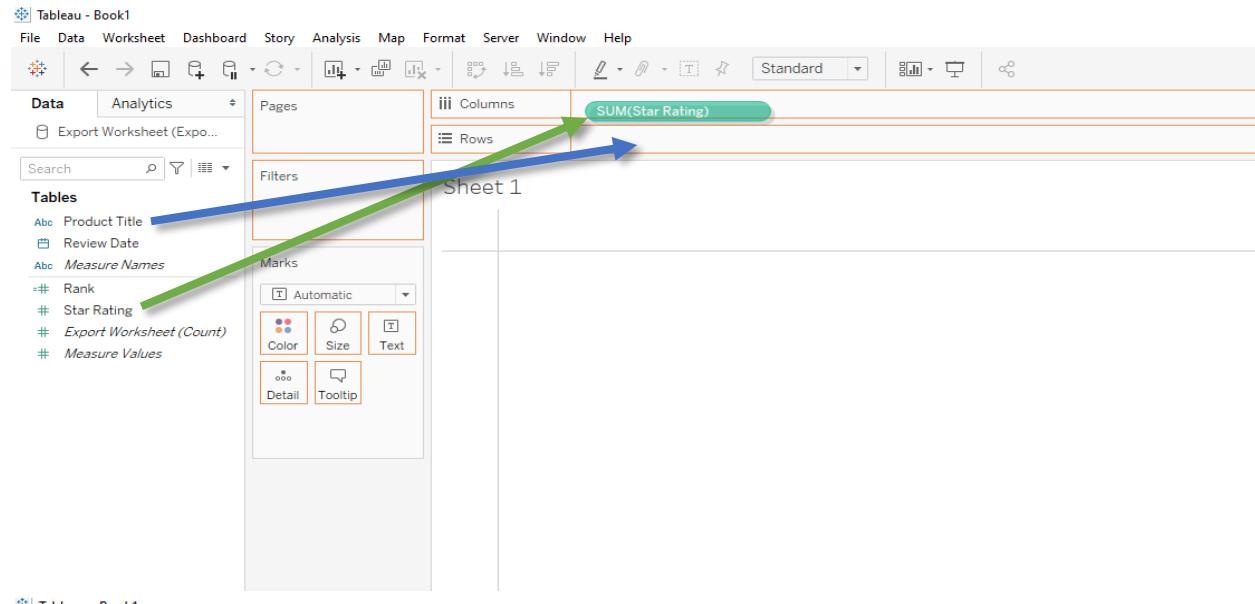
You must make sure to label your field as “Rank” to differentiate your fields. The code

`RANK_UNIQUE(SUM([Insert_Your_Field_Here]))`, allows you to have ranking for your rows or columns.



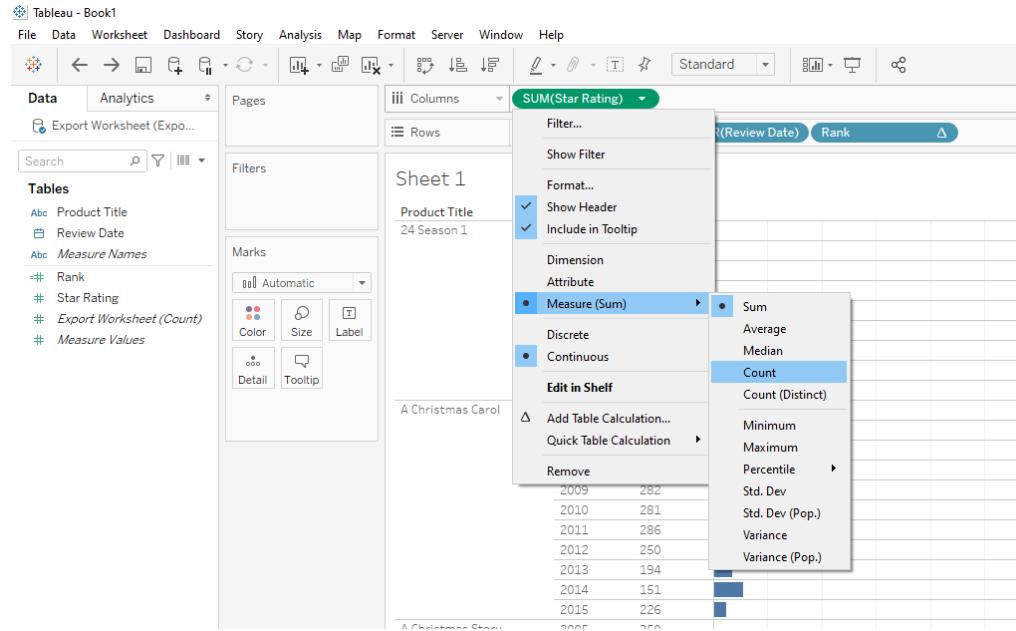
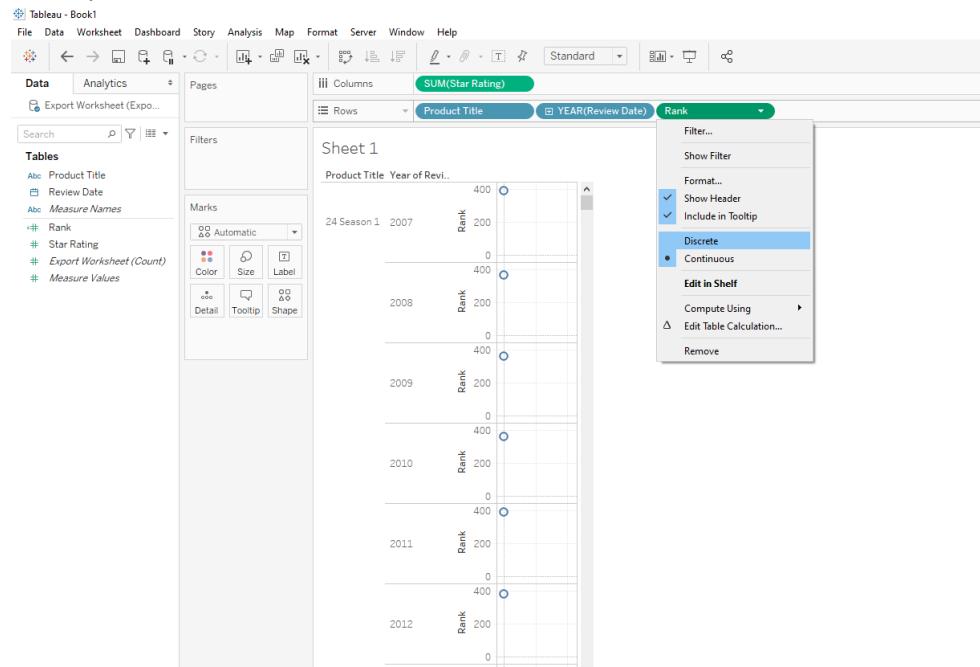
Step 5:

Drag the fields (Under Tables), “Star Rating” to Columns, “Product Title,” “Review Date,” and “Rank” to Rows.



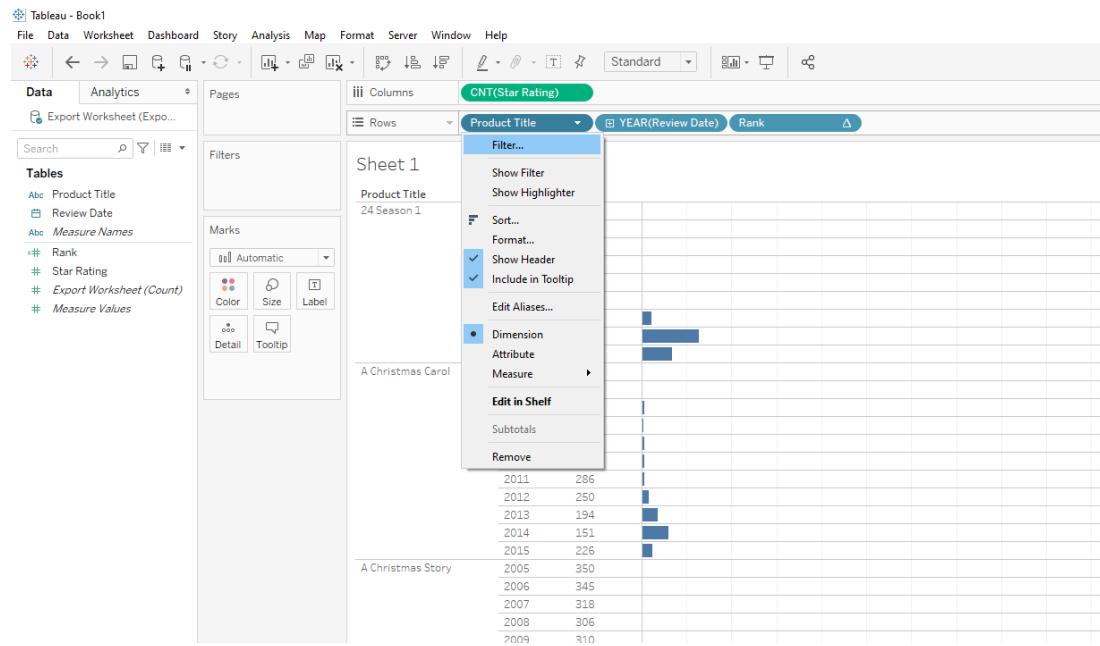
Step 6:

To change the layout of the chart to a Horizontal Bar Chart, move your mouse to rank and click on the arrow. Then click on “Discrete.” After that, move your mouse to Sum(Star Rating) and click on the arrow. Move your mouse down to Measure (Sum), then click on count.



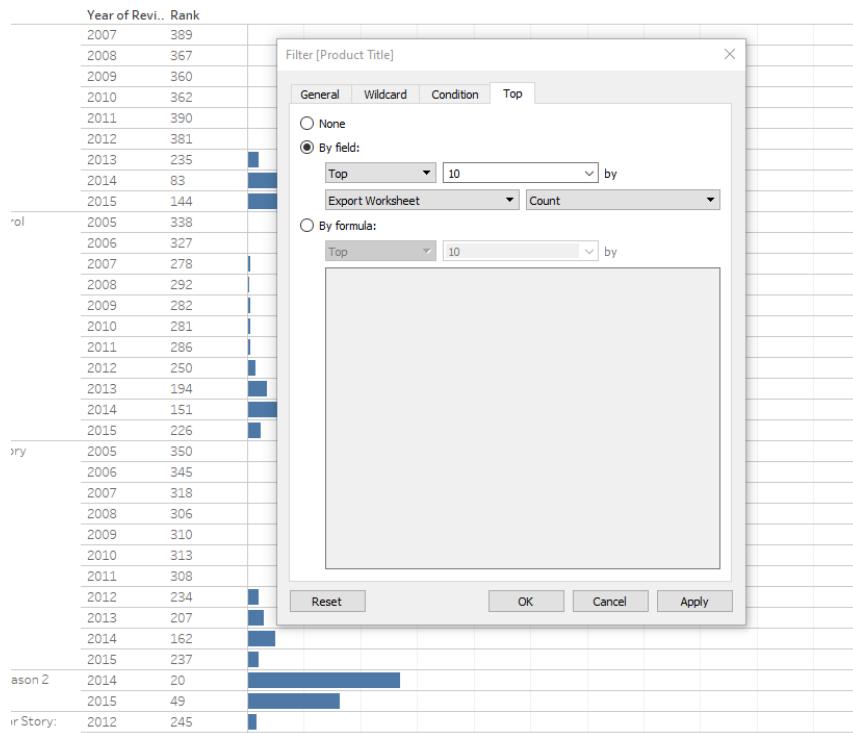
Step 7:

To show only the Top 10 Product Titles, move the mouse to Product Title and click on the arrow. Then click on Filter.



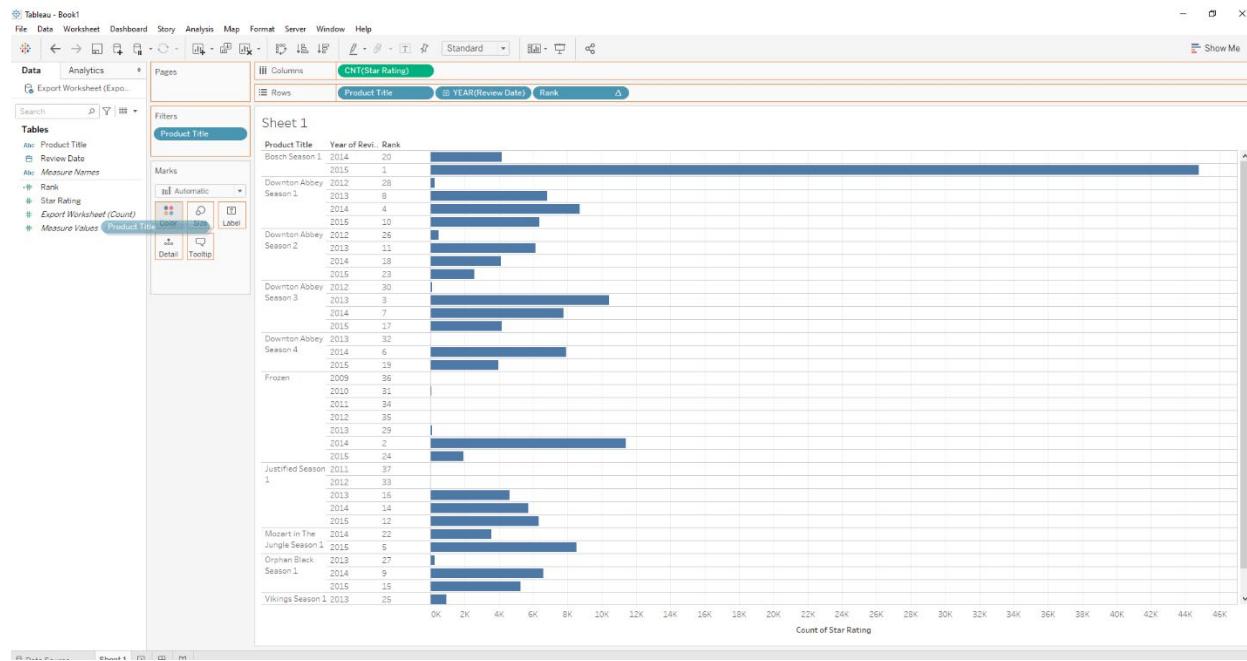
Step 8:

Click on the Top tab and then click on “By Field.” Since it will automatically show “Top 10” like the diagram below, click on Apply and OK.



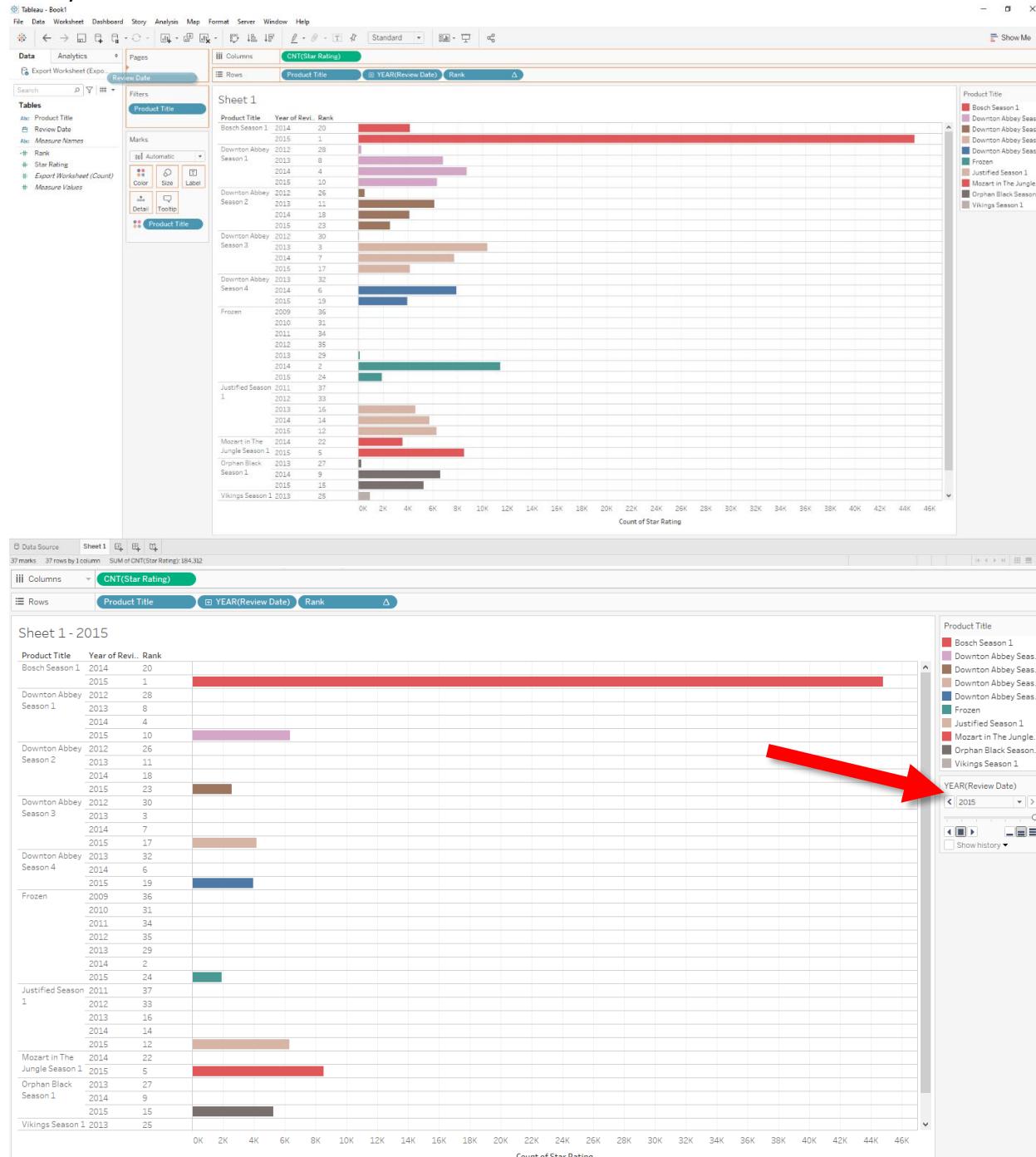
Step 9:

To assign different bar colors for each product title, drag the Product Title field to the Color mini box in the Marks box.

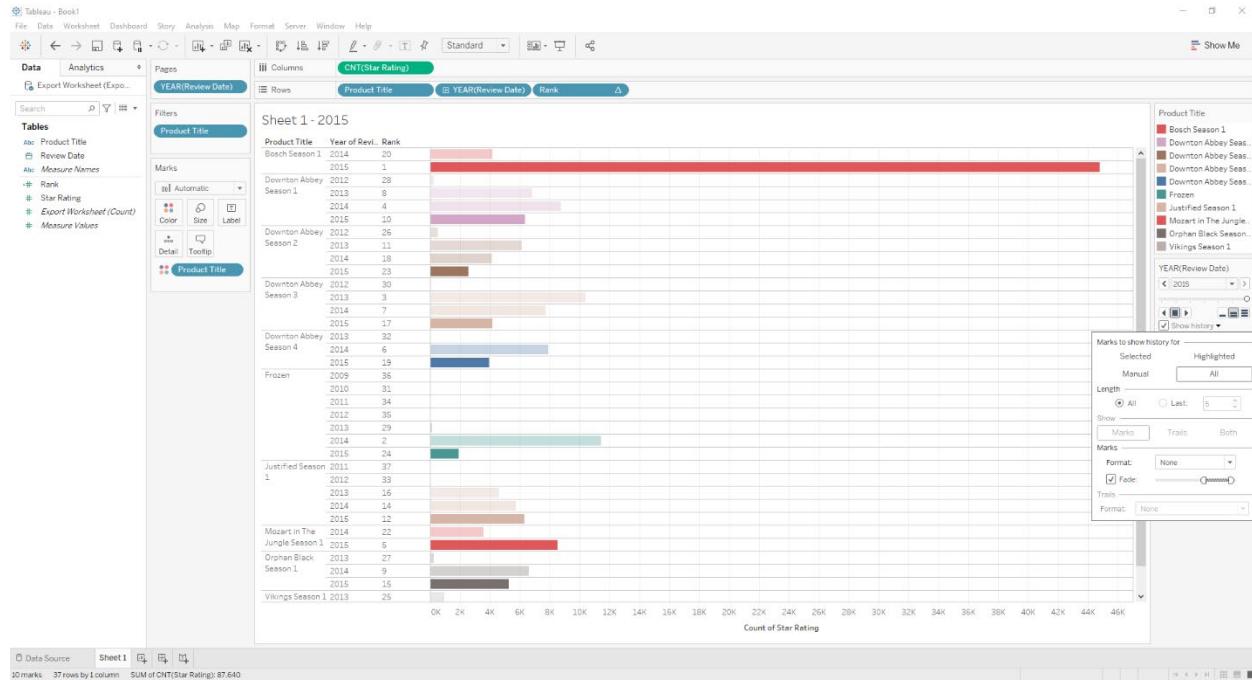


Step 10:

To begin with the animation, drag the Review Date field to the Pages box. Then the Animation box will appear on the right side of Tableau. To see the chart animation, click on the right arrow (Play Button) to see progress year after year. Left arrow goes backward (Reverse Button), and the middle button stops the animation. You can also drag the circle button both ways to see each year.



- Also, when clicking on the Show History Button, click on All to see bars faded from previous years as the animation is being played. You can drag the fade to your preference.



Final Graphs from Tableau:



References

1. URL of Data Source, URL of Data Source, <https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt>
2. URL of Project Github: https://github.com/NinaRo2/CIS_5200
3. Other reference:
<https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF>