DALHOUSIE UNIVERSITY

CSCI 5408 – Data Management, Warehousing, Analytics

Assignment 3

Work done by,

Name: Guturu Rama Mohan Vishnu

Banner ID: B00871849

Email: rm286720@dal.ca

DECLARATION

I, Guturu Rama Mohan Vishnu, declare that in assignment 3 of CSCI 5408 course, data scrapping is not done programmatically or using any online or offline tools. However, the webpages or the domain mentioned in this document are visited manually, and some useful information is gathered for education purpose only. Information, such as email, personal contact numbers, or names of people are not extracted. The course instructor or the Faculty of Computer Science cannot be held responsible for any misuse of the extracted data.

Problem #2: MySQL and Java – Implementation of 2 phase locking

Solution:

The datasets I found in the given link are:

(a) AnnualTicketSales

(b) HighestGrossers

(c) PopularCreativeTypes

(d) TopDistributors

(e) TopGenres

(f) TopGrossingRatings

(g) TopGrossingSources

(h) TopProductionMethods

(i) WideReleasesCount

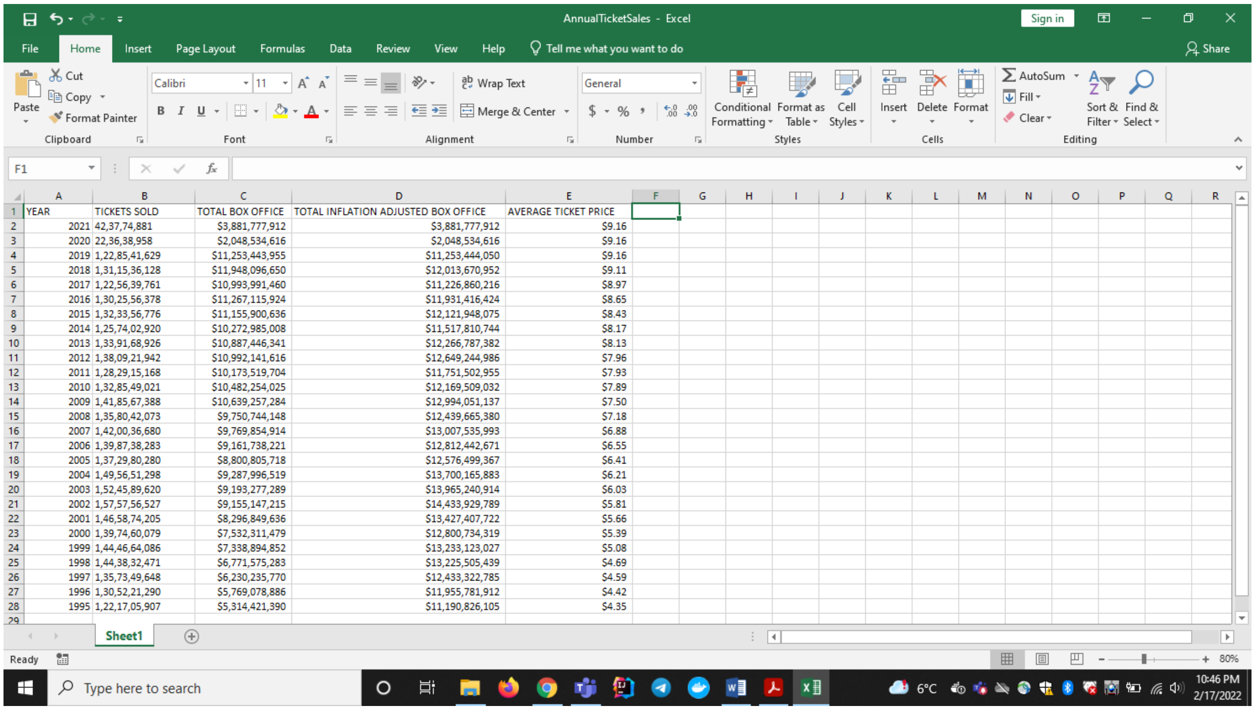
Cleaning and Transformation

AnnualTicketSales

There is no primary key for this table.

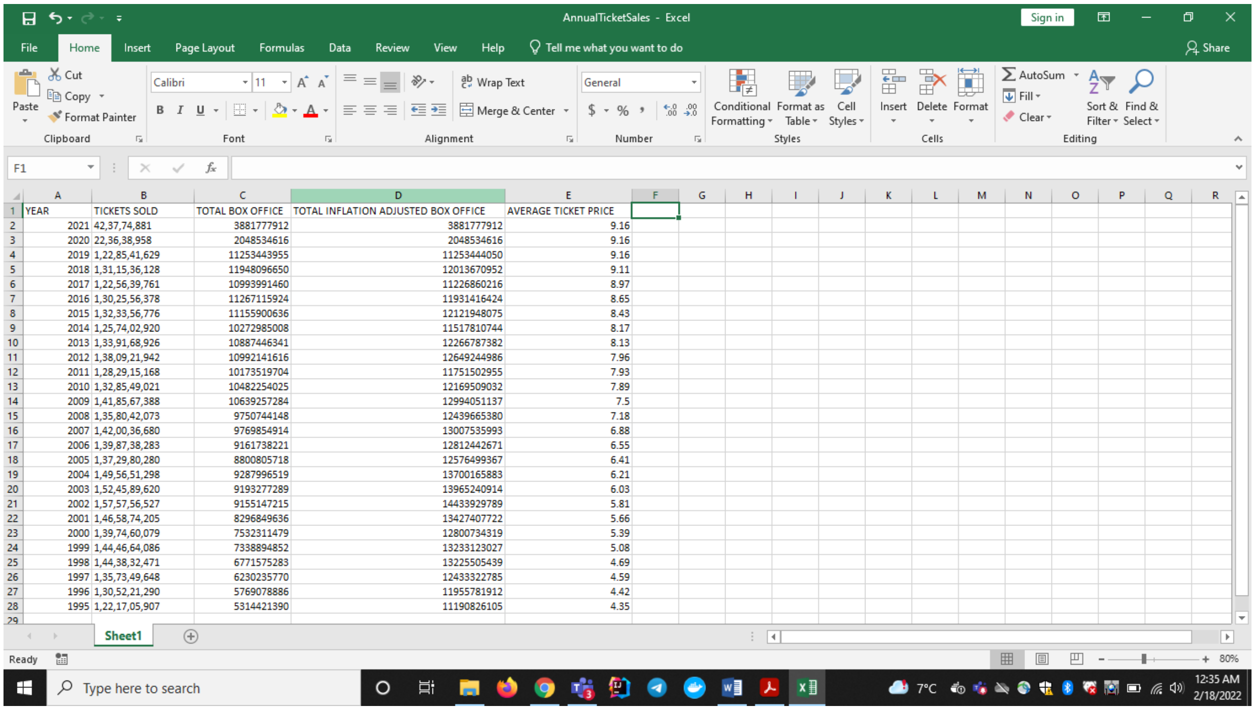
(1) There are also no blank cells or blank columns in this dataset.

(2) But the columns contain values which have “$” sign in them before the actual value. In order to load the data into database and in order to compare the values, we have to remove the $ sign.



	A	B	C	D	E
1	YEAR	TICKETS SOLD	TOTAL BOX OFFICE	TOTAL INFLATION ADJUSTED BOX OFFICE	AVERAGE TICKET PRICE
2	2021	42,37,74,881	$3,881,777,912	$3,881,777,912	$9.16
3	2020	22,36,38,958	$2,048,534,616	$2,048,534,616	$9.16
4	2019	1,22,85,41,629	$11,253,443,955	$11,253,444,050	$9.16
5	2018	1,31,15,36,128	$11,948,096,650	$12,013,670,952	$9.11
6	2017	1,22,56,39,761	$10,993,991,460	$11,226,860,216	$8.97
7	2016	1,30,25,56,378	$11,267,115,924	$11,931,416,424	$8.65
8	2015	1,32,33,56,776	$11,155,900,636	$12,121,948,075	$8.43
9	2014	1,25,74,02,920	$10,272,985,008	$11,517,810,744	$8.17
10	2013	1,33,91,68,926	$10,887,446,341	$12,266,787,382	$8.13
11	2012	1,38,09,21,942	$10,992,141,616	$12,649,244,986	$7.96
12	2011	1,28,29,15,168	$10,173,519,704	$11,751,502,955	$7.93
13	2010	1,32,85,49,021	$10,482,254,025	$12,169,509,032	$7.89
14	2009	1,41,85,67,388	$10,639,257,284	$12,994,051,137	$7.50
15	2008	1,35,80,42,073	$9,750,744,148	$12,439,665,380	$7.18
16	2007	1,42,00,36,680	$9,769,854,914	$13,007,535,993	$6.88
17	2006	1,39,87,38,283	$9,161,738,221	$12,812,442,671	$6.55
18	2005	1,37,29,80,280	$8,800,805,718	$12,576,499,367	$6.41
19	2004	1,49,56,51,298	$9,287,996,519	$13,700,165,883	$6.21
20	2003	1,52,45,89,620	$9,193,277,289	$13,965,240,914	$6.03
21	2002	1,57,57,56,527	$9,155,147,215	$14,433,929,789	$5.81
22	2001	1,46,58,74,205	$8,296,849,636	$13,427,407,722	$5.66
23	2000	1,39,74,60,079	$7,532,311,479	$12,800,734,319	$5.39
24	1999	1,44,46,64,086	$7,338,894,852	$13,233,123,027	$5.08
25	1998	1,44,38,32,471	$6,771,575,283	$13,225,505,439	$4.69
26	1997	1,35,73,49,648	$6,230,235,770	$12,433,322,785	$4.59
27	1996	1,30,52,21,290	$5,769,078,886	$11,955,781,912	$4.42
28	1995	1,22,17,05,907	$5,314,421,390	$11,190,826,105	$4.35

(3) So, I am replacing it with empty string “”, i.e., I am removing that character from the field.

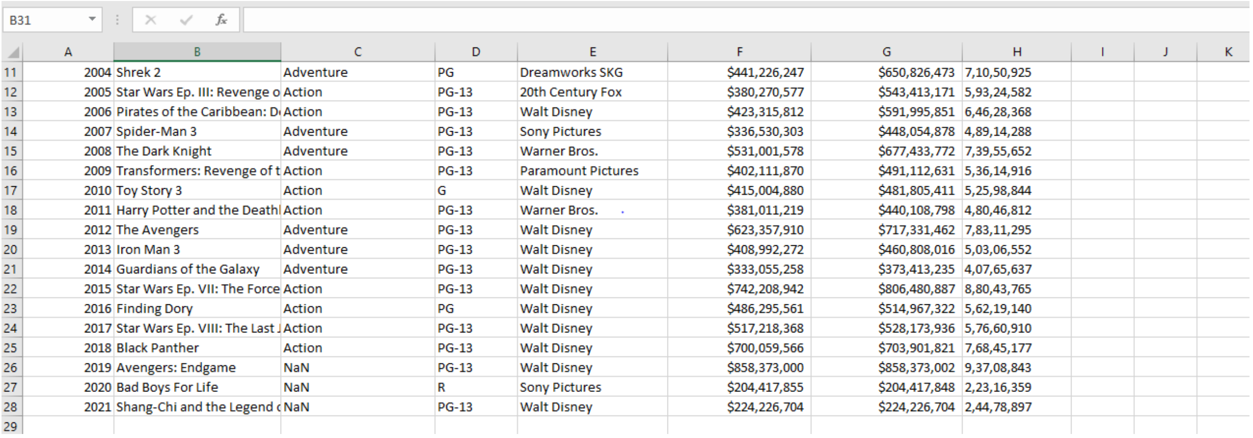


	A	B	C	D	E
1	YEAR	TICKETS SOLD	TOTAL BOX OFFICE	TOTAL INFLATION ADJUSTED BOX OFFICE	AVERAGE TICKET PRICE
2	2021	42,37,74,881	3881777912	3881777912	9.16
3	2020	22,36,38,958	2048534616	2048534616	9.16
4	2019	1,22,85,41,629	11253443955	11253444050	9.16
5	2018	1,31,15,36,128	11948096650	12013670952	9.11
6	2017	1,22,56,39,761	10993991460	11226860216	8.97
7	2016	1,30,25,56,378	11267115924	11931416424	8.65
8	2015	1,32,33,56,776	11155900636	12121948075	8.43
9	2014	1,25,74,02,920	10272985008	11517810744	8.17
10	2013	1,33,91,68,926	10887446341	12266787382	8.13
11	2012	1,38,09,21,942	10992141616	12649244986	7.96
12	2011	1,28,29,15,168	10173519704	11751502955	7.93
13	2010	1,32,85,49,021	10482254025	12169509032	7.89
14	2009	1,41,85,67,388	10639257284	12994051137	7.5
15	2008	1,35,80,42,073	9750744148	12439665380	7.18
16	2007	1,42,00,36,680	9769854914	13007535993	6.88
17	2006	1,39,87,38,283	9161738221	12812442671	6.55
18	2005	1,37,29,80,280	8800805718	12576499367	6.41
19	2004	1,49,56,51,298	9287996519	13700165883	6.21
20	2003	1,52,45,89,620	9193277289	13965240914	6.03
21	2002	1,57,57,56,527	9155147215	14433929789	5.81
22	2001	1,46,58,74,205	8296849636	13427407722	5.66
23	2000	1,39,74,60,079	7532311479	12800734319	5.39
24	1999	1,44,46,64,086	7338894852	13233123027	5.08
25	1998	1,44,38,32,471	6771575283	13225505439	4.69
26	1997	1,35,73,49,648	6230235770	12433322785	4.59
27	1996	1,30,52,21,290	5769078886	11955781912	4.42
28	1995	1,22,17,05,907	5314421390	11190826105	4.35

HighestGrossers

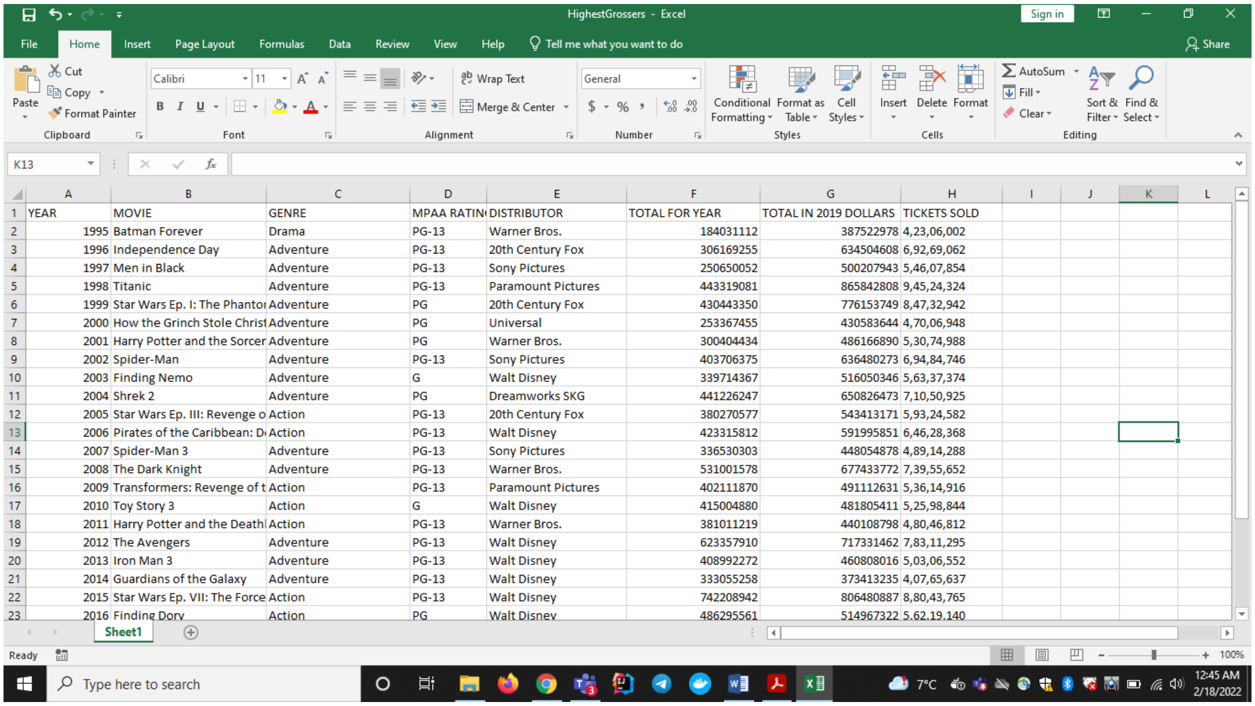
The Primary Key for this table is “MOVIE” column.

(1) We have few blank fields in this data set. We replace them with NaN since they are in text type column.



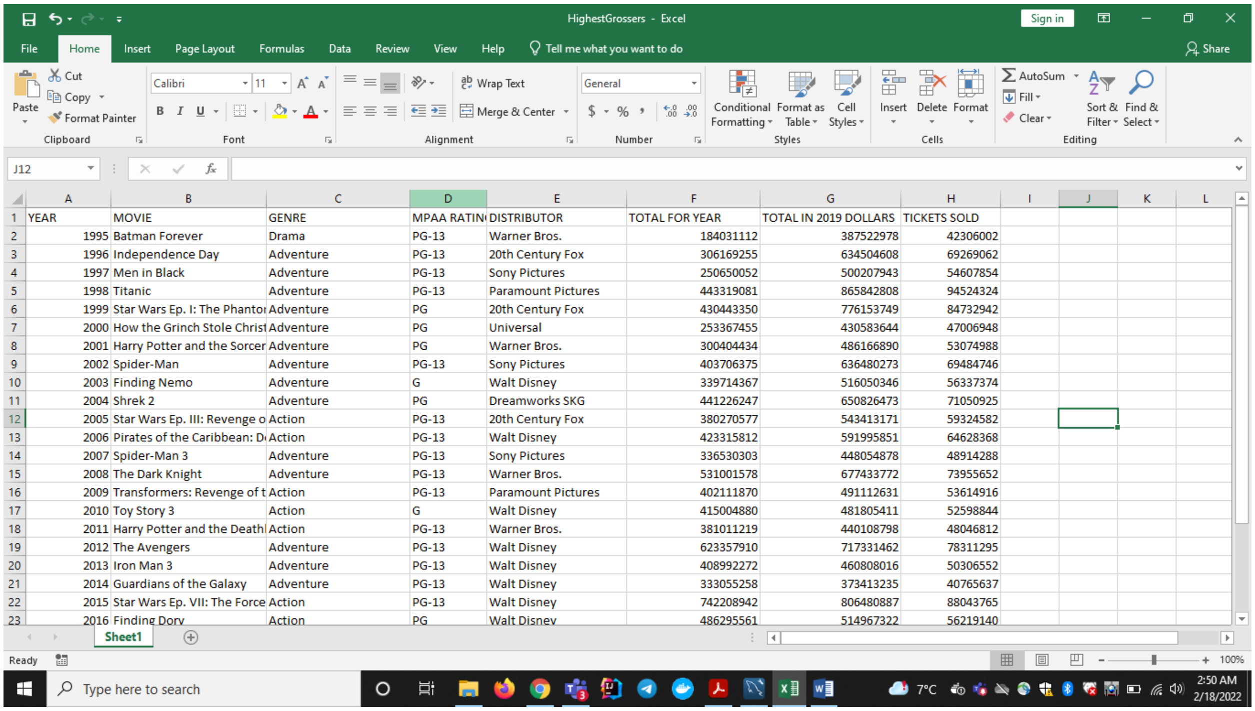
	A	B	C	D	E	F	G	H
11	2004	Shrek 2	Adventure	PG	Dreamworks SKG	$441,226,247	$650,826,473	7,10,50,925
12	2005	Star Wars Ep. III: Revenge o	Action	PG-13	20th Century Fox	$380,270,577	$543,413,171	5,93,24,582
13	2006	Pirates of the Caribbean: D	Action	PG-13	Walt Disney	$423,315,812	$591,995,851	6,46,28,368
14	2007	Spider-Man 3	Adventure	PG-13	Sony Pictures	$336,530,303	$448,054,878	4,89,14,288
15	2008	The Dark Knight	Adventure	PG-13	Warner Bros.	$531,001,578	$677,433,772	7,39,55,652
16	2009	Transformers: Revenge of t	Action	PG-13	Paramount Pictures	$402,111,870	$491,112,631	5,36,14,916
17	2010	Toy Story 3	Action	G	Walt Disney	$415,004,880	$481,805,411	5,25,98,844
18	2011	Harry Potter and the Deathl	Action	PG-13	Warner Bros.	$381,011,219	$440,108,798	4,80,46,812
19	2012	The Avengers	Adventure	PG-13	Walt Disney	$623,357,910	$717,331,462	7,83,11,295
20	2013	Iron Man 3	Adventure	PG-13	Walt Disney	$408,992,272	$460,808,016	5,03,06,552
21	2014	Guardians of the Galaxy	Adventure	PG-13	Walt Disney	$333,055,258	$373,413,235	4,07,65,637
22	2015	Star Wars Ep. VII: The Force	Action	PG-13	Walt Disney	$742,208,942	$806,480,887	8,80,43,765
23	2016	Finding Dory	Action	PG	Walt Disney	$486,295,561	$514,967,322	5,62,19,140
24	2017	Star Wars Ep. VIII: The Last J	Action	PG-13	Walt Disney	$517,218,368	$528,173,936	5,76,60,910
25	2018	Black Panther	Action	PG-13	Walt Disney	$700,059,566	$703,901,821	7,68,45,177
26	2019	Avengers: Endgame	NaN	PG-13	Walt Disney	$858,373,000	$858,373,002	9,37,08,843
27	2020	Bad Boys For Life	NaN	R	Sony Pictures	$204,417,855	$204,417,848	2,23,16,359
28	2021	Shang-Chi and the Legend o	NaN	PG-13	Walt Disney	$224,226,704	$224,226,704	2,44,78,897

(2) As saw in the first table, few columns contain values which have “$” sign in them before the actual value. So, we do the same thing which we did before, in the process of data cleaning.



	A	B	C	D	E	F	G	H
1	YEAR	MOVIE	GENRE	MPAA RATIN	DISTRIBUTOR	TOTAL FOR YEAR	TOTAL IN 2019 DOLLARS	TICKETS SOLD
2	1995	Batman Forever	Drama	PG-13	Warner Bros.	184031112	387522978	4,23,06,002
3	1996	Independence Day	Adventure	PG-13	20th Century Fox	306169255	634504608	6,92,69,062
4	1997	Men in Black	Adventure	PG-13	Sony Pictures	250650052	500207943	5,46,07,854
5	1998	Titanic	Adventure	PG-13	Paramount Pictures	443319081	865842808	9,45,24,324
6	1999	Star Wars Ep. I: The Phantor	Adventure	PG	20th Century Fox	430443350	776153749	8,47,32,942
7	2000	How the Grinch Stole Christ	Adventure	PG	Universal	253367455	430583644	4,70,06,948
8	2001	Harry Potter and the Sorcer	Adventure	PG	Warner Bros.	300404434	486166890	5,30,74,988
9	2002	Spider-Man	Adventure	PG-13	Sony Pictures	403706375	636480273	6,94,84,746
10	2003	Finding Nemo	Adventure	G	Walt Disney	339714367	516050346	5,63,37,374
11	2004	Shrek 2	Adventure	PG	Dreamworks SKG	441226247	650826473	7,10,50,925
12	2005	Star Wars Ep. III: Revenge o	Action	PG-13	20th Century Fox	380270577	543413171	5,93,24,582
13	2006	Pirates of the Caribbean: D	Action	PG-13	Walt Disney	423315812	591995851	6,46,28,368
14	2007	Spider-Man 3	Adventure	PG-13	Sony Pictures	336530303	448054878	4,89,14,288
15	2008	The Dark Knight	Adventure	PG-13	Warner Bros.	531001578	677433772	7,39,55,652
16	2009	Transformers: Revenge of t	Action	PG-13	Paramount Pictures	402111870	491112631	5,36,14,916
17	2010	Toy Story 3	Action	G	Walt Disney	415004880	481805411	5,25,98,844
18	2011	Harry Potter and the Deathl	Action	PG-13	Warner Bros.	381011219	440108798	4,80,46,812
19	2012	The Avengers	Adventure	PG-13	Walt Disney	623357910	717331462	7,83,11,295
20	2013	Iron Man 3	Adventure	PG-13	Walt Disney	408992272	460808016	5,03,06,552
21	2014	Guardians of the Galaxy	Adventure	PG-13	Walt Disney	333055258	373413235	4,07,65,637
22	2015	Star Wars Ep. VII: The Force	Action	PG-13	Walt Disney	742208942	806480887	8,80,43,765
23	2016	Finding Dory	Action	PG	Walt Disney	486295561	514967322	5,62,19,140

(3) The column “TICKETS SOLD” has the integer values but as form of text. So, in order for the database to detect it as a number, I have changed all the fields to number format.



	A	B	C	D	E	F	G	H
1	YEAR	MOVIE	GENRE	MPAA RATIN	DISTRIBUTOR	TOTAL FOR YEAR	TOTAL IN 2019 DOLLARS	TICKETS SOLD
2	1995	Batman Forever	Drama	PG-13	Warner Bros.	184031112	387522978	42306002
3	1996	Independence Day	Adventure	PG-13	20th Century Fox	306169255	634504608	69269062
4	1997	Men in Black	Adventure	PG-13	Sony Pictures	250650052	500207943	54607854
5	1998	Titanic	Adventure	PG-13	Paramount Pictures	443319081	865842808	94524324
6	1999	Star Wars Ep. I: The Phanto	Adventure	PG	20th Century Fox	430443350	776153749	84732942
7	2000	How the Grinch Stole Christ	Adventure	PG	Universal	253367455	430583644	47006948
8	2001	Harry Potter and the Sorcer	Adventure	PG	Warner Bros.	300404434	486166890	53074988
9	2002	Spider-Man	Adventure	PG-13	Sony Pictures	403706375	636480273	69484746
10	2003	Finding Nemo	Adventure	G	Walt Disney	339714367	516050346	56337374
11	2004	Shrek 2	Adventure	PG	Dreamworks SKG	441226247	650826473	71050925
12	2005	Star Wars Ep. III: Revenge o	Action	PG-13	20th Century Fox	380270577	543413171	59324582
13	2006	Pirates of the Caribbean: D	Action	PG-13	Walt Disney	423315812	591995851	64628368
14	2007	Spider-Man 3	Adventure	PG-13	Sony Pictures	336530303	448054878	48914288
15	2008	The Dark Knight	Adventure	PG-13	Warner Bros.	531001578	677433772	73955652
16	2009	Transformers: Revenge of t	Action	PG-13	Paramount Pictures	402111870	491112631	53614916
17	2010	Toy Story 3	Action	G	Walt Disney	415004880	481805411	52598844
18	2011	Harry Potter and the Deathl	Action	PG-13	Warner Bros.	381011219	440108798	48046812
19	2012	The Avengers	Adventure	PG-13	Walt Disney	623357910	717331462	78311295
20	2013	Iron Man 3	Adventure	PG-13	Walt Disney	408992272	460808016	50306552
21	2014	Guardians of the Galaxy	Adventure	PG-13	Walt Disney	333055258	373413235	40765637
22	2015	Star Wars Ep. VII: The Force	Action	PG-13	Walt Disney	742208942	806480887	88043765
23	2016	Finding Dory	Action	PG	Walt Disney	486295561	514967322	56219140

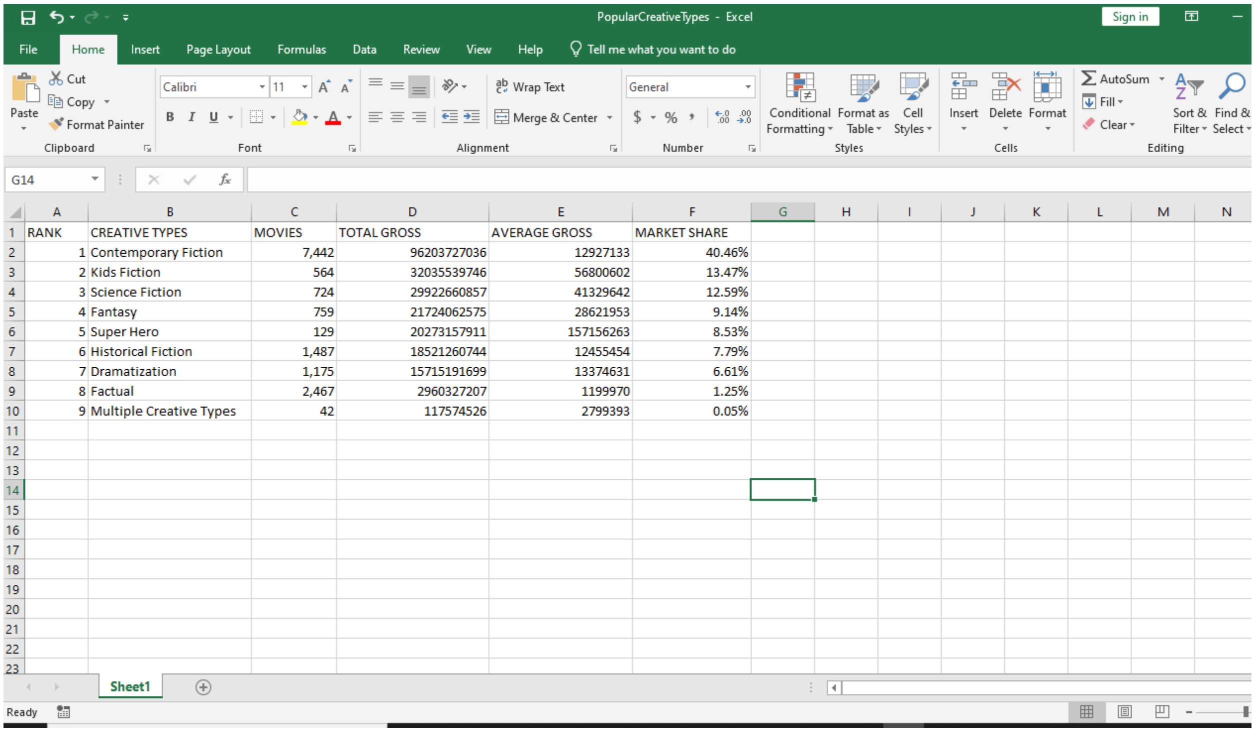
(4) Also, there are no duplicate values in the primary key column.

PopularCreativeTypes

The Primary Key for this table is “CREATIVE TYPES”.

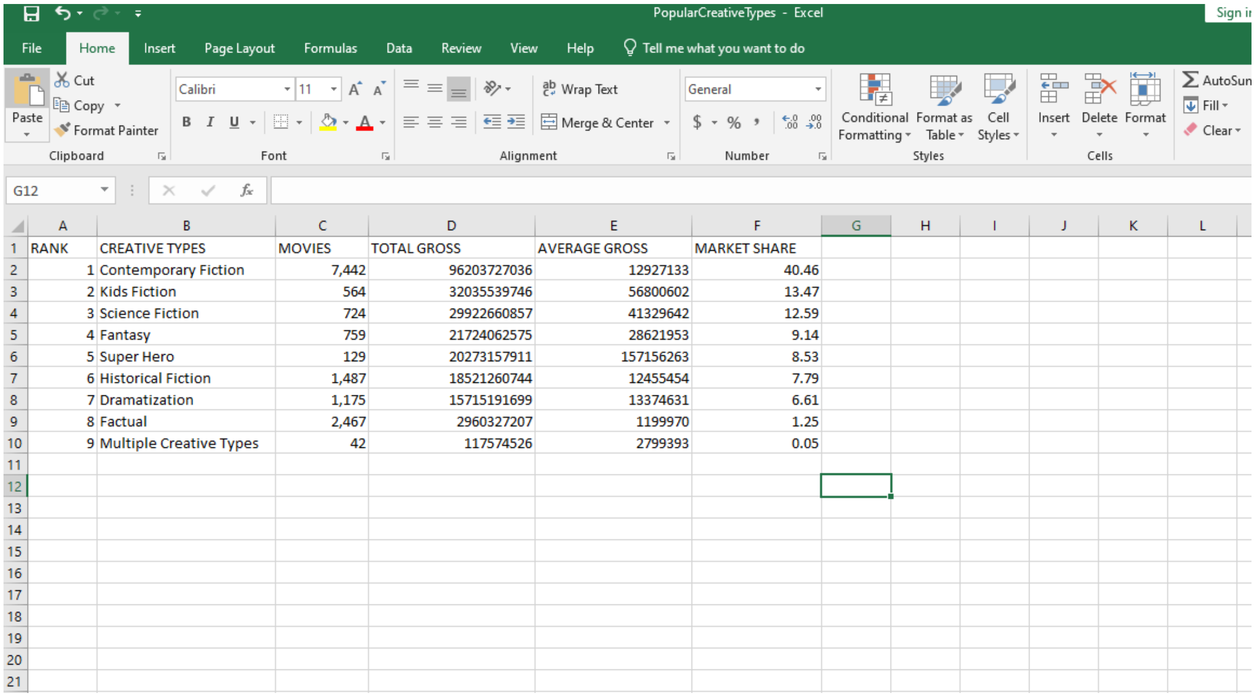
(1) There aren’t any blank cells in this data set.

(2) Few columns contain values which have “$” sign in them before the actual value. So, we do the same thing which we did before, in the process of data cleaning. I removed the signs.



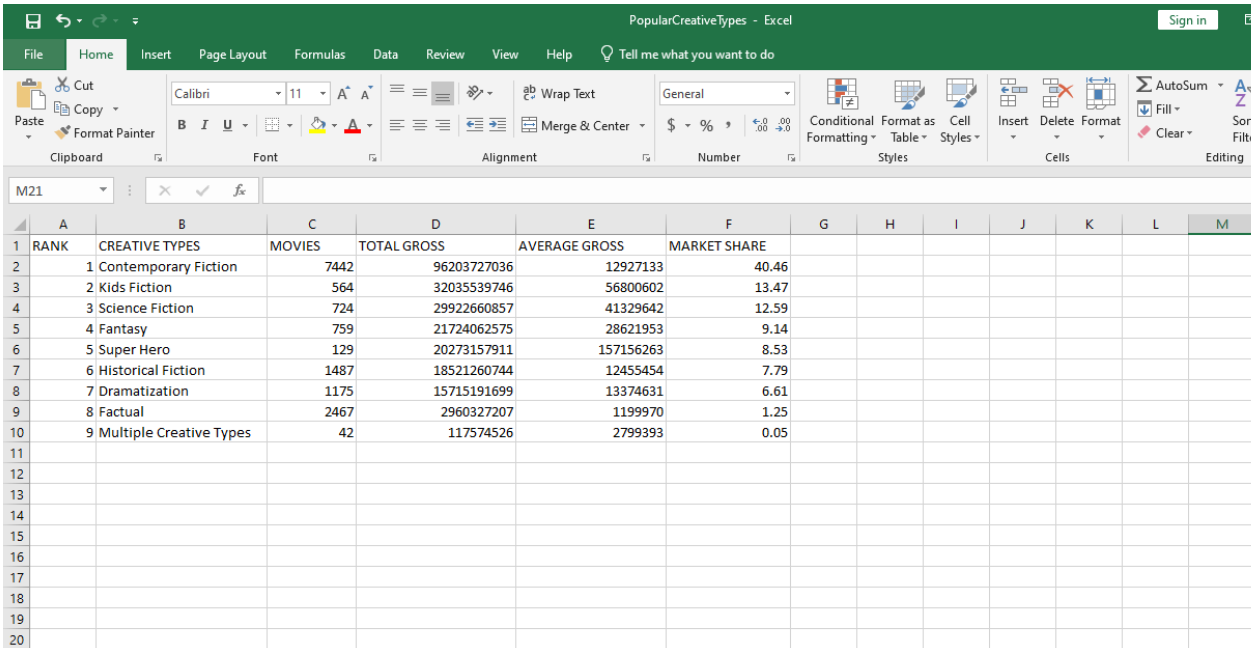
RANK	CREATIVE TYPES	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
1	Contemporary Fiction	7,442	96203727036	12927133	40.46%
2	Kids Fiction	564	32035539746	56800602	13.47%
3	Science Fiction	724	29922660857	41329642	12.59%
4	Fantasy	759	21724062575	28621953	9.14%
5	Super Hero	129	20273157911	157156263	8.53%
6	Historical Fiction	1,487	18521260744	12455454	7.79%
7	Dramatization	1,175	15715191699	13374631	6.61%
8	Factual	2,467	2960327207	1199970	1.25%
9	Multiple Creative Types	42	117574526	2799393	0.05%

(3) Also, there is a column which has a symbol % in it. So, we remove it too for consistency.



RANK	CREATIVE TYPES	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
1	Contemporary Fiction	7,442	96203727036	12927133	40.46
2	Kids Fiction	564	32035539746	56800602	13.47
3	Science Fiction	724	29922660857	41329642	12.59
4	Fantasy	759	21724062575	28621953	9.14
5	Super Hero	129	20273157911	157156263	8.53
6	Historical Fiction	1,487	18521260744	12455454	7.79
7	Dramatization	1,175	15715191699	13374631	6.61
8	Factual	2,467	2960327207	1199970	1.25
9	Multiple Creative Types	42	117574526	2799393	0.05

(4) Movies column has comma in the values. To maintain consistency, I removed them too.



	A	B	C	D	E	F
1	RANK	CREATIVE TYPES	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
2	1	Contemporary Fiction	7442	96203727036	12927133	40.46
3	2	Kids Fiction	564	32035539746	56800602	13.47
4	3	Science Fiction	724	29922660857	41329642	12.59
5	4	Fantasy	759	21724062575	28621953	9.14
6	5	Super Hero	129	20273157911	157156263	8.53
7	6	Historical Fiction	1487	18521260744	12455454	7.79
8	7	Dramatization	1175	15715191699	13374631	6.61
9	8	Factual	2467	2960327207	1199970	1.25
10	9	Multiple Creative Types	42	117574526	2799393	0.05

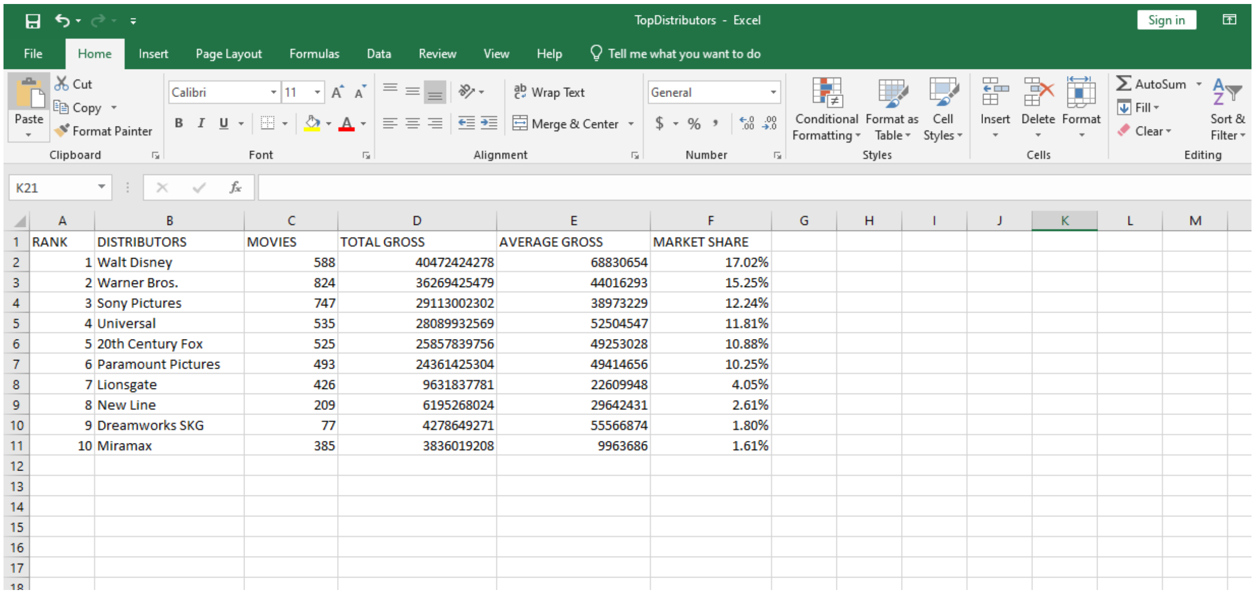
(5) There are no duplicate values in the primary key column.

TopDistributors

The Primary Key for this table is “DISTRIBUTORS”.

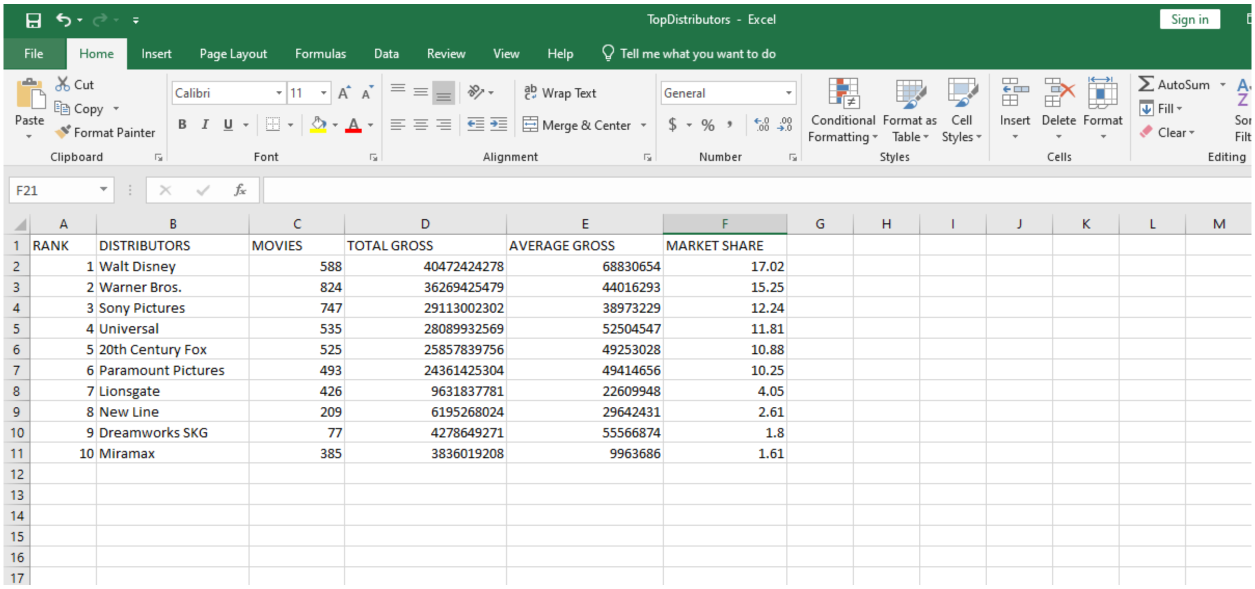
(1) There aren’t any blank cells in this data set.

(2) Few columns contain values which have “$” sign in them before the actual value. So, we do the same thing which we did before, in the process of data cleaning. I removed the signs.



RANK	DISTRIBUTORS	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
1	Walt Disney	588	40472424278	68830654	17.02%
2	Warner Bros.	824	36269425479	44016293	15.25%
3	Sony Pictures	747	29113002302	38973229	12.24%
4	Universal	535	28089932569	52504547	11.81%
5	20th Century Fox	525	25857839756	49253028	10.88%
6	Paramount Pictures	493	24361425304	49414656	10.25%
7	Lionsgate	426	9631837781	22609948	4.05%
8	New Line	209	6195268024	29642431	2.61%
9	Dreamworks SKG	77	4278649271	55566874	1.80%
10	Miramax	385	3836019208	9963686	1.61%

(3) Also, there is a column which has a symbol % in it. So, we remove it too for consistency.



RANK	DISTRIBUTORS	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
1	Walt Disney	588	40472424278	68830654	17.02
2	Warner Bros.	824	36269425479	44016293	15.25
3	Sony Pictures	747	29113002302	38973229	12.24
4	Universal	535	28089932569	52504547	11.81
5	20th Century Fox	525	25857839756	49253028	10.88
6	Paramount Pictures	493	24361425304	49414656	10.25
7	Lionsgate	426	9631837781	22609948	4.05
8	New Line	209	6195268024	29642431	2.61
9	Dreamworks SKG	77	4278649271	55566874	1.8
10	Miramax	385	3836019208	9963686	1.61

(4) There are no duplicate values in the primary key column.

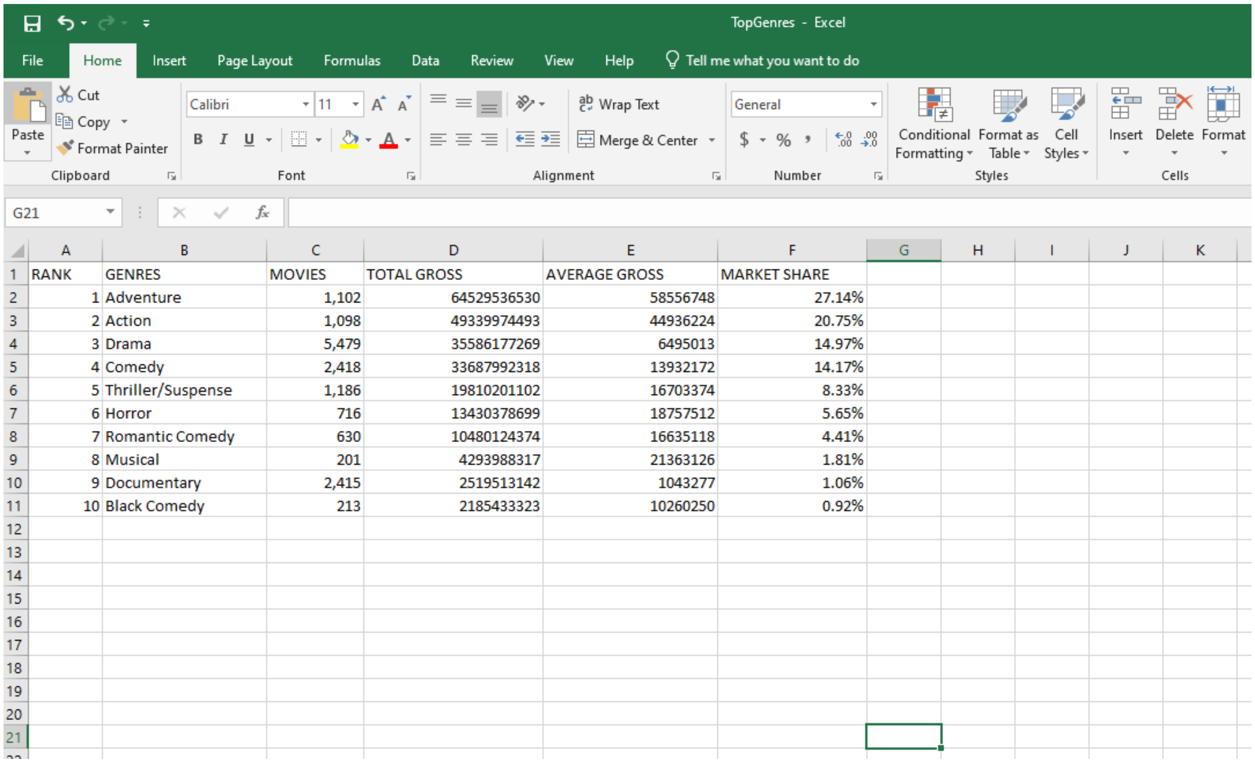
(5) There are a few empty columns in this dataset which can be seen while importing data in to the database. I deleted those columns too.

TopGenres

The Primary Key for this table is “GENRES”.

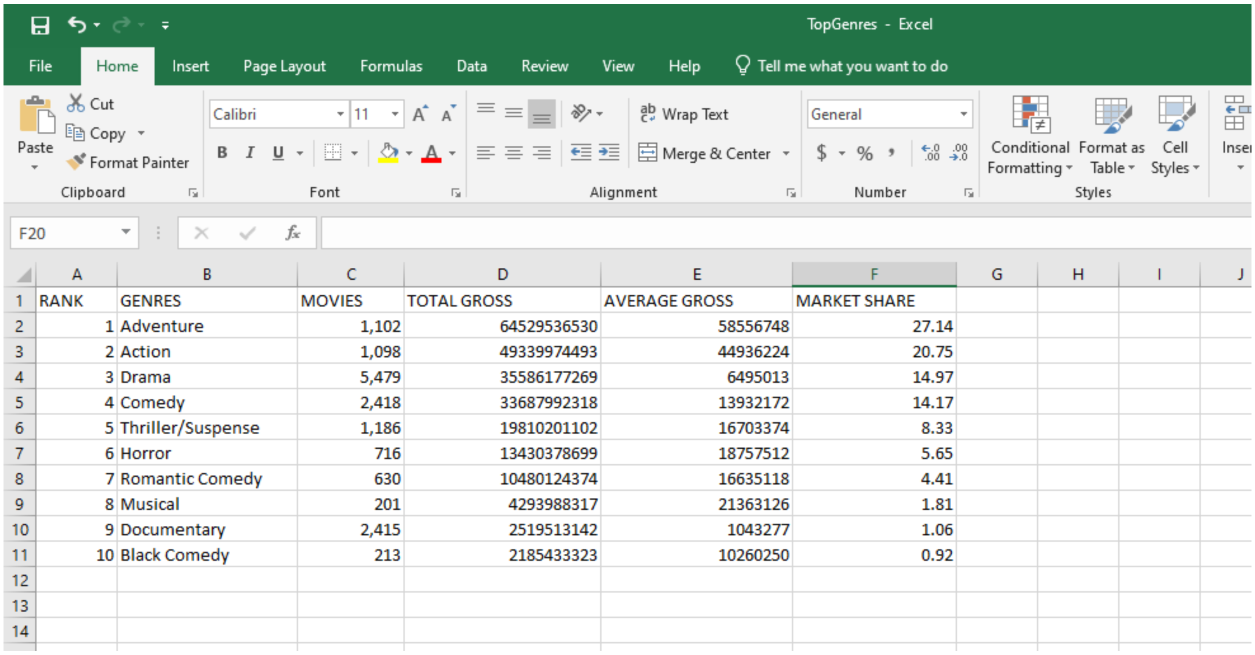
(1) There aren’t any blank cells in this data set.

(2) Few columns contain values which have “$” sign in them before the actual value. So, we do the same thing which we did before, in the process of data cleaning. I removed the signs.



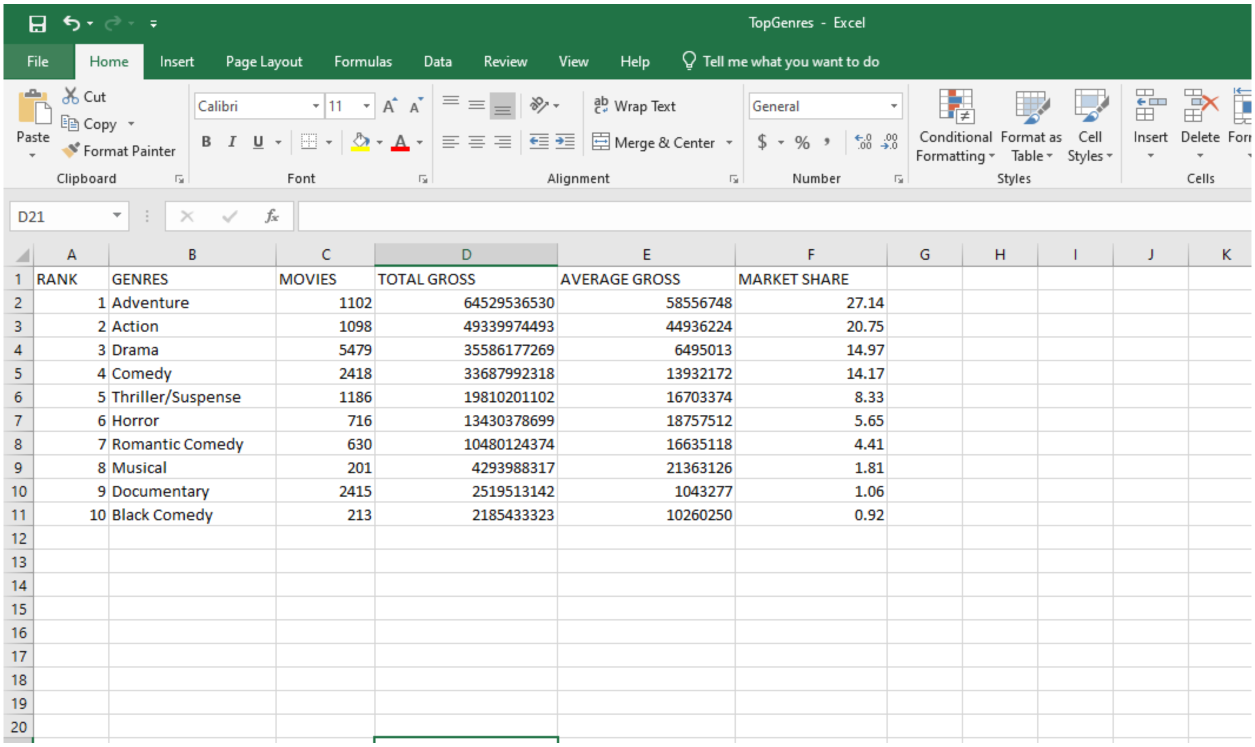
RANK	GENRES	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
1	Adventure	1,102	64529536530	58556748	27.14%
2	Action	1,098	49339974493	44936224	20.75%
3	Drama	5,479	35586177269	6495013	14.97%
4	Comedy	2,418	33687992318	13932172	14.17%
5	Thriller/Suspense	1,186	19810201102	16703374	8.33%
6	Horror	716	13430378699	18757512	5.65%
7	Romantic Comedy	630	10480124374	16635118	4.41%
8	Musical	201	4293988317	21363126	1.81%
9	Documentary	2,415	2519513142	1043277	1.06%
10	Black Comedy	213	2185433323	10260250	0.92%

(3) Also, there is a column which has a symbol % in it. So, we remove it too for consistency.



	A	B	C	D	E	F
1	RANK	GENRES	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
2	1	Adventure	1,102	64529536530	58556748	27.14
3	2	Action	1,098	49339974493	44936224	20.75
4	3	Drama	5,479	35586177269	6495013	14.97
5	4	Comedy	2,418	33687992318	13932172	14.17
6	5	Thriller/Suspense	1,186	19810201102	16703374	8.33
7	6	Horror	716	13430378699	18757512	5.65
8	7	Romantic Comedy	630	10480124374	16635118	4.41
9	8	Musical	201	4293988317	21363126	1.81
10	9	Documentary	2,415	2519513142	1043277	1.06
11	10	Black Comedy	213	2185433323	10260250	0.92

(4) Movies column has comma in the values. To maintain consistency, I removed them too.



	A	B	C	D	E	F
1	RANK	GENRES	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
2	1	Adventure	1102	64529536530	58556748	27.14
3	2	Action	1098	49339974493	44936224	20.75
4	3	Drama	5479	35586177269	6495013	14.97
5	4	Comedy	2418	33687992318	13932172	14.17
6	5	Thriller/Suspense	1186	19810201102	16703374	8.33
7	6	Horror	716	13430378699	18757512	5.65
8	7	Romantic Comedy	630	10480124374	16635118	4.41
9	8	Musical	201	4293988317	21363126	1.81
10	9	Documentary	2415	2519513142	1043277	1.06
11	10	Black Comedy	213	2185433323	10260250	0.92

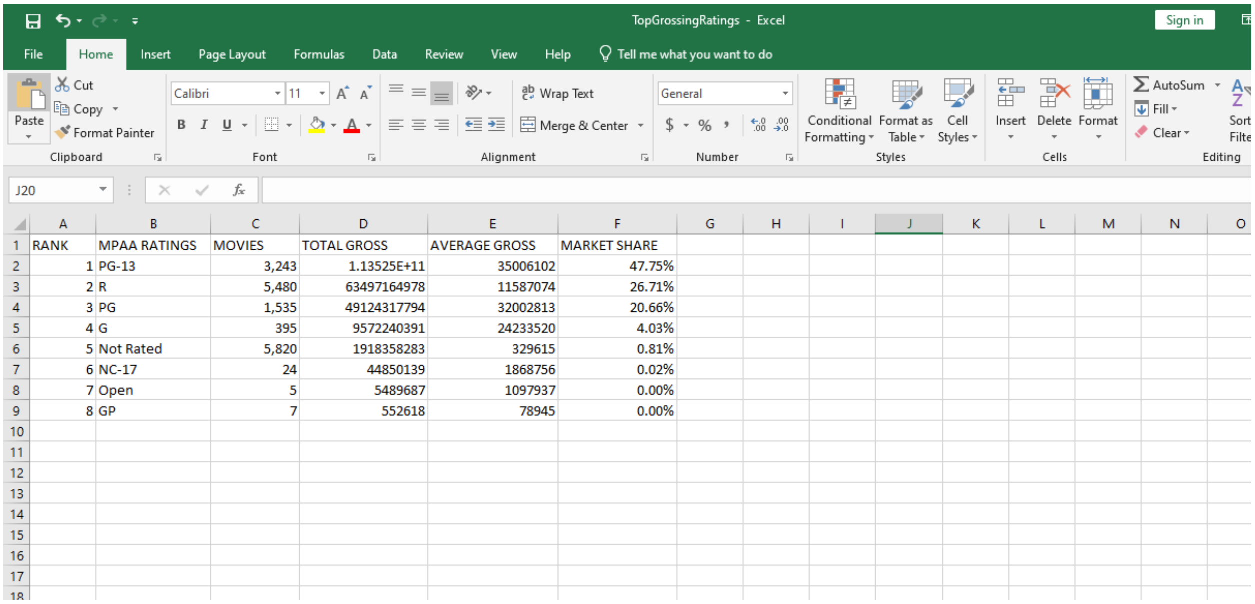
(5) There are no duplicate values in the primary key column.

TopGrossingRatings

The Primary Key for this table is “MPAA RATINGS”.

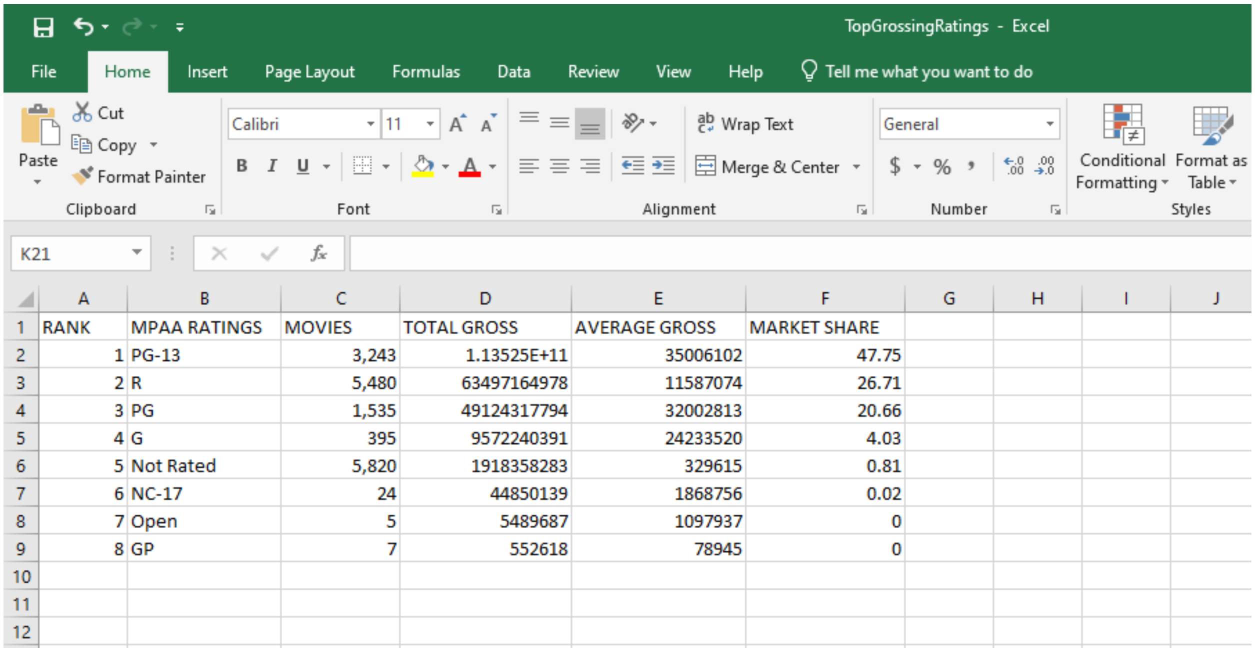
(1) There aren’t any blank cells in this data set.

(2) Few columns contain values which have “$” sign in them before the actual value. So, we do the same thing which we did before, in the process of data cleaning. I removed the signs.



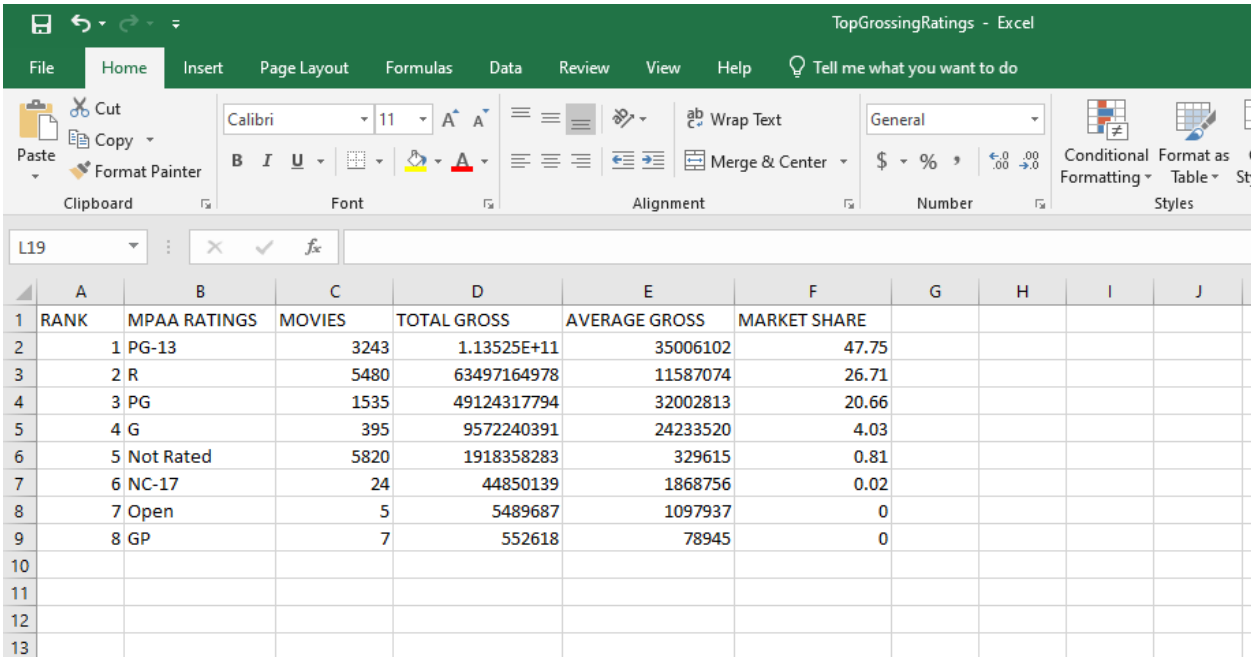
RANK	MPAA RATINGS	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
1	PG-13	3,243	1.13525E+11	35006102	47.75%
2	R	5,480	63497164978	11587074	26.71%
3	PG	1,535	49124317794	32002813	20.66%
4	G	395	9572240391	24233520	4.03%
5	Not Rated	5,820	1918358283	329615	0.81%
6	NC-17	24	44850139	1868756	0.02%
7	Open	5	5489687	1097937	0.00%
8	GP	7	552618	78945	0.00%

(3) Also, there is a column which has a symbol % in it. So, we remove it too for consistency.



	A	B	C	D	E	F
1	RANK	MPAA RATINGS	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
2	1	PG-13	3,243	1.13525E+11	35006102	47.75
3	2	R	5,480	63497164978	11587074	26.71
4	3	PG	1,535	49124317794	32002813	20.66
5	4	G	395	9572240391	24233520	4.03
6	5	Not Rated	5,820	1918358283	329615	0.81
7	6	NC-17	24	44850139	1868756	0.02
8	7	Open	5	5489687	1097937	0
9	8	GP	7	552618	78945	0

(4) Movies column has comma in the values. To maintain consistency, I removed them too.



	A	B	C	D	E	F
1	RANK	MPAA RATINGS	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
2	1	PG-13	3243	1.13525E+11	35006102	47.75
3	2	R	5480	63497164978	11587074	26.71
4	3	PG	1535	49124317794	32002813	20.66
5	4	G	395	9572240391	24233520	4.03
6	5	Not Rated	5820	1918358283	329615	0.81
7	6	NC-17	24	44850139	1868756	0.02
8	7	Open	5	5489687	1097937	0
9	8	GP	7	552618	78945	0

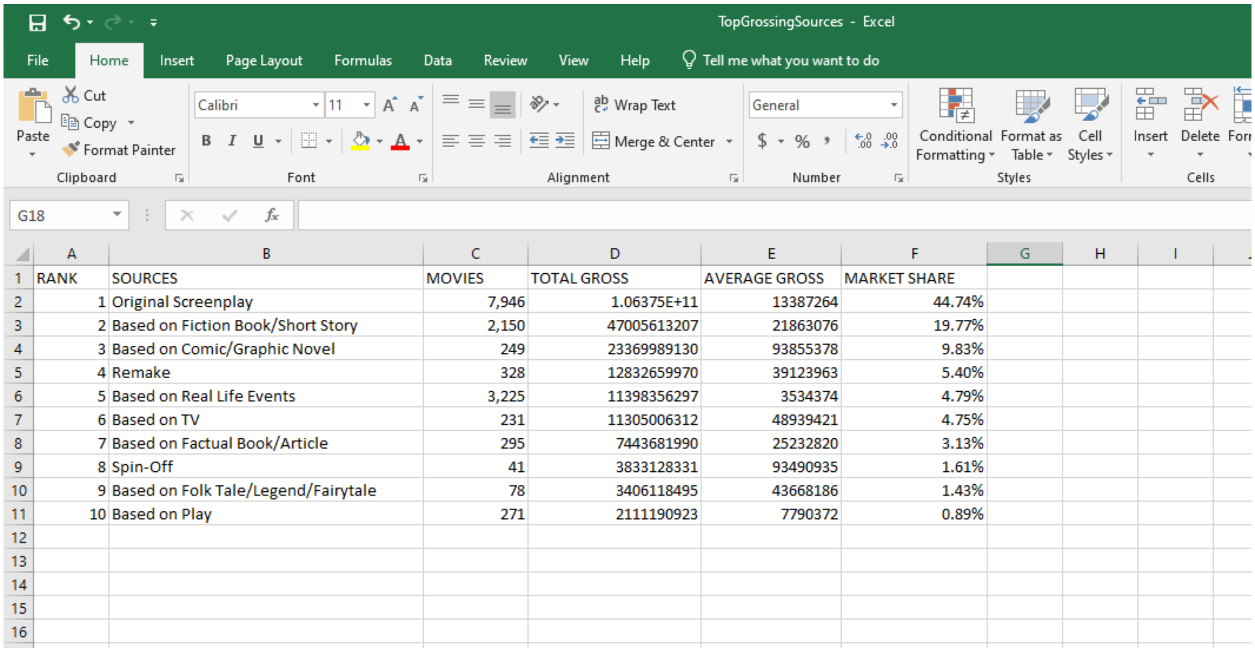
(5) There are no duplicate values in the primary key column.

TopGrossingSources

The Primary Key for this table is “SOURCES”.

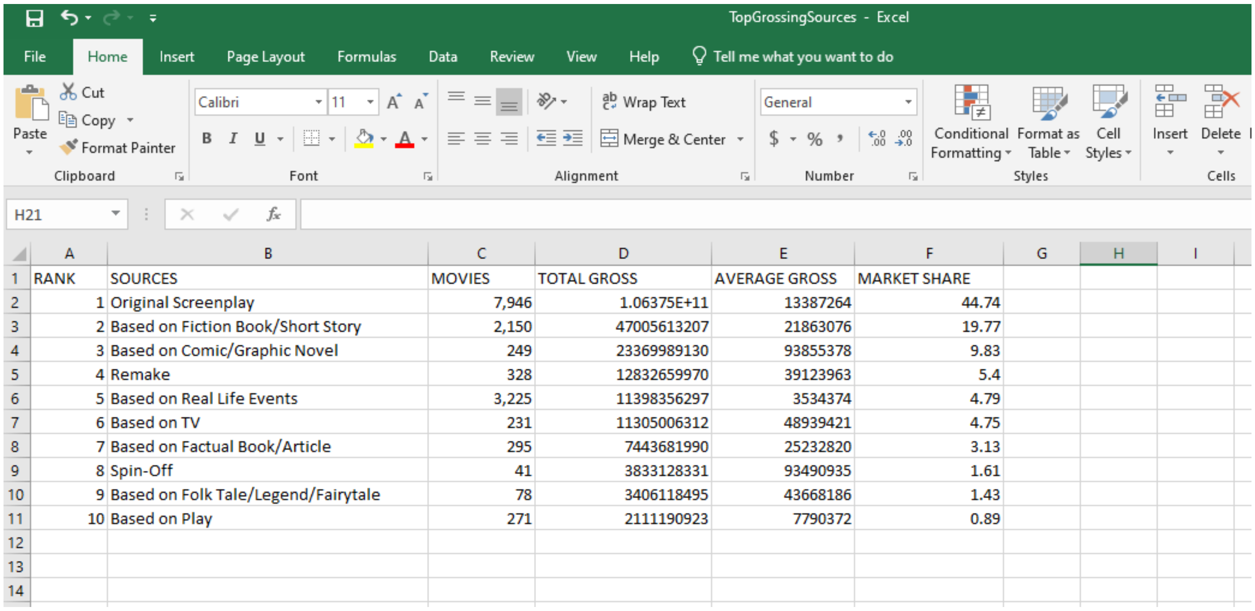
(1) There aren’t any blank cells in this data set.

(2) Few columns contain values which have “$” sign in them before the actual value. So, we do the same thing which we did before, in the process of data cleaning. I removed the signs.



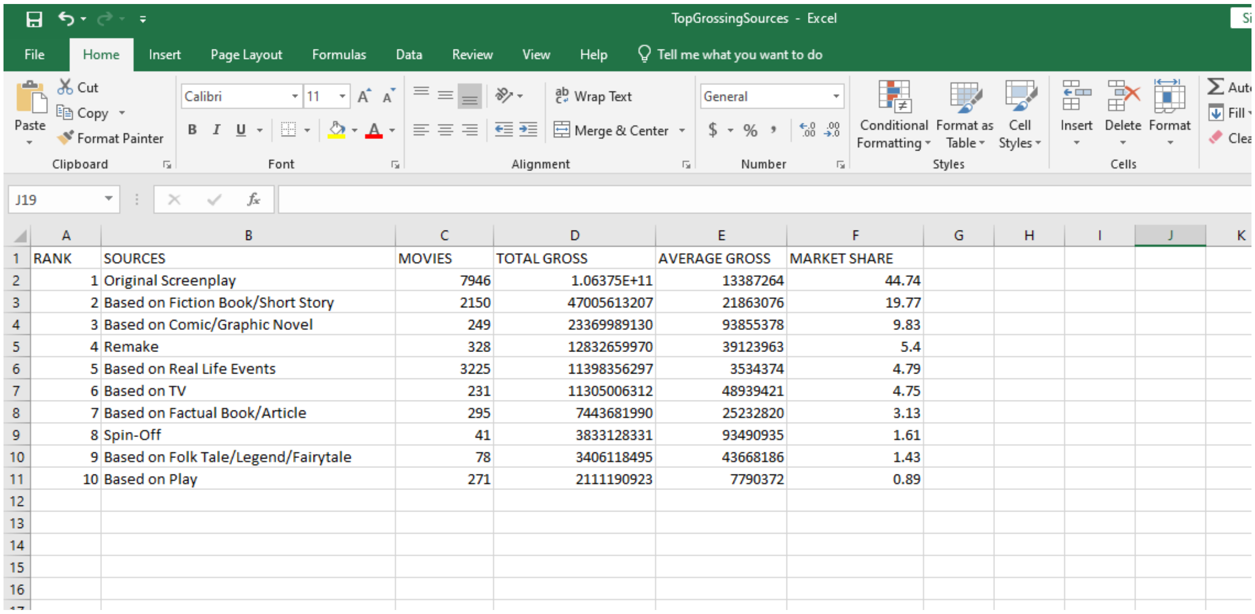
	A	B	C	D	E	F
1	RANK	SOURCES	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
2	1	Original Screenplay	7,946	1.06375E+11	13387264	44.74%
3	2	Based on Fiction Book/Short Story	2,150	47005613207	21863076	19.77%
4	3	Based on Comic/Graphic Novel	249	23369989130	93855378	9.83%
5	4	Remake	328	12832659970	39123963	5.40%
6	5	Based on Real Life Events	3,225	11398356297	3534374	4.79%
7	6	Based on TV	231	11305006312	48939421	4.75%
8	7	Based on Factual Book/Article	295	7443681990	25232820	3.13%
9	8	Spin-Off	41	3833128331	93490935	1.61%
10	9	Based on Folk Tale/Legend/Fairytale	78	3406118495	43668186	1.43%
11	10	Based on Play	271	2111190923	7790372	0.89%

(3) Also, there is a column which has a symbol % in it. So, we remove it too for consistency.



	A	B	C	D	E	F	G	H	I
1	RANK	SOURCES	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE			
2	1	Original Screenplay	7,946	1.06375E+11	13387264	44.74			
3	2	Based on Fiction Book/Short Story	2,150	47005613207	21863076	19.77			
4	3	Based on Comic/Graphic Novel	249	23369989130	93855378	9.83			
5	4	Remake	328	12832659970	39123963	5.4			
6	5	Based on Real Life Events	3,225	11398356297	3534374	4.79			
7	6	Based on TV	231	11305006312	48939421	4.75			
8	7	Based on Factual Book/Article	295	7443681990	25232820	3.13			
9	8	Spin-Off	41	3833128331	93490935	1.61			
10	9	Based on Folk Tale/Legend/Fairytale	78	3406118495	43668186	1.43			
11	10	Based on Play	271	2111190923	7790372	0.89			
12									
13									
14									

(4) Movies column has comma in the values. To maintain consistency, I removed them too.



	A	B	C	D	E	F	G	H	I	J	K
1	RANK	SOURCES	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE					
2	1	Original Screenplay	7946	1.06375E+11	13387264	44.74					
3	2	Based on Fiction Book/Short Story	2150	47005613207	21863076	19.77					
4	3	Based on Comic/Graphic Novel	249	23369989130	93855378	9.83					
5	4	Remake	328	12832659970	39123963	5.4					
6	5	Based on Real Life Events	3225	11398356297	3534374	4.79					
7	6	Based on TV	231	11305006312	48939421	4.75					
8	7	Based on Factual Book/Article	295	7443681990	25232820	3.13					
9	8	Spin-Off	41	3833128331	93490935	1.61					
10	9	Based on Folk Tale/Legend/Fairytale	78	3406118495	43668186	1.43					
11	10	Based on Play	271	2111190923	7790372	0.89					
12											
13											
14											
15											
16											

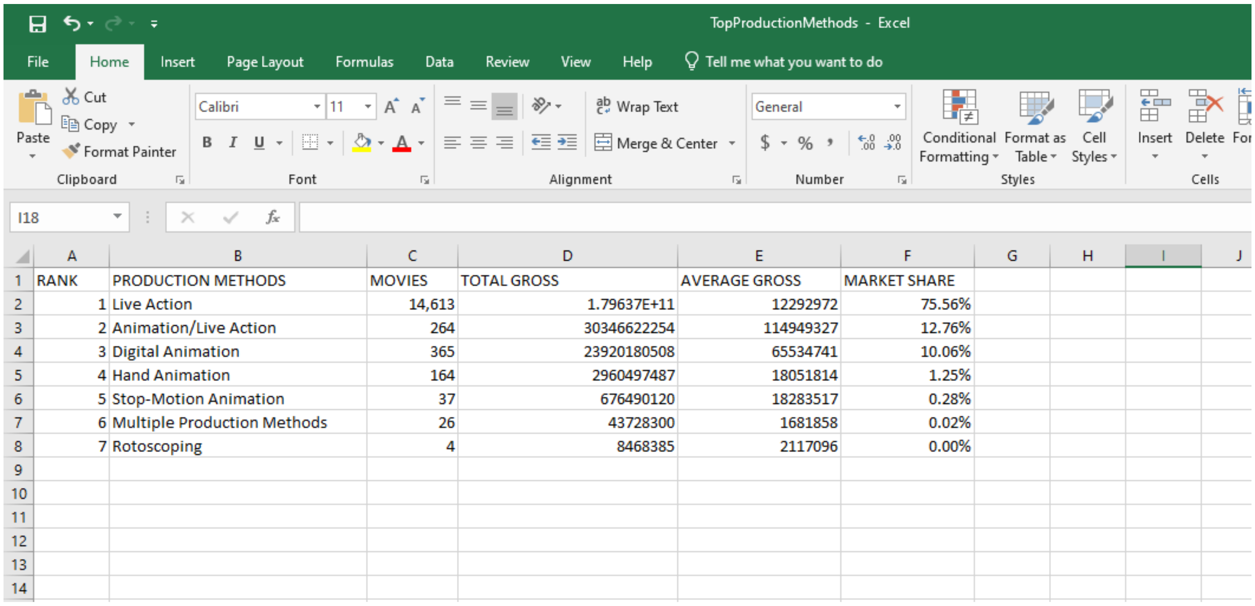
(5) There are no duplicate values in the primary key column.

TopProductionMethods

The Primary key for this table is “PRODUCTION METHODS”.

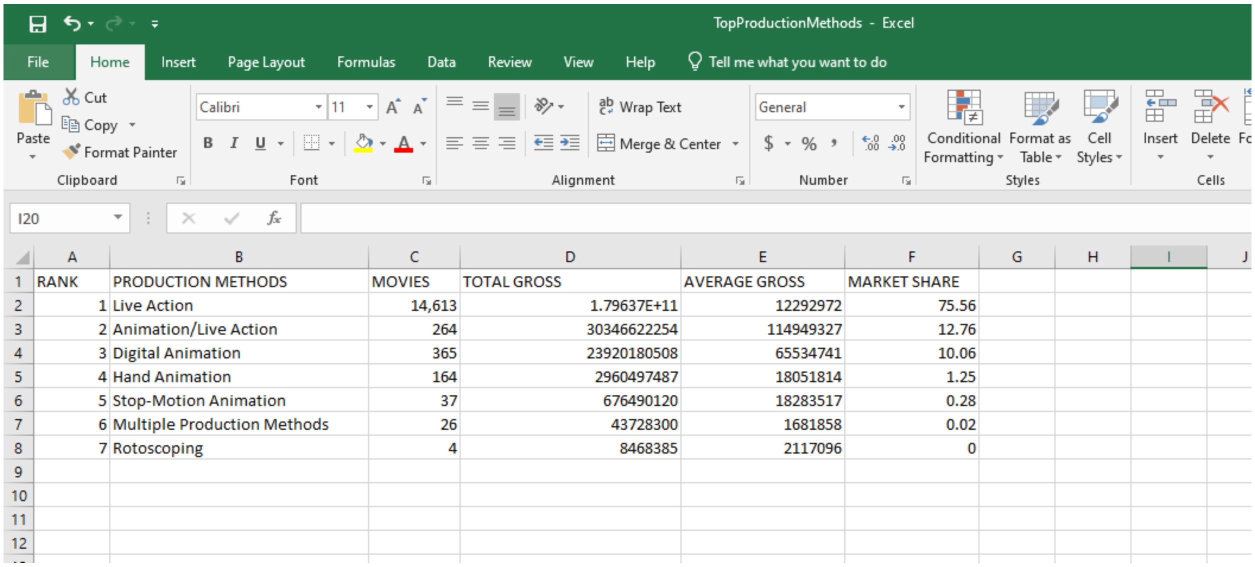
(1) There aren’t any blank cells in this data set.

(2) Few columns contain values which have “$” sign in them before the actual value. So, we do the same thing which we did before, in the process of data cleaning. I removed the signs.



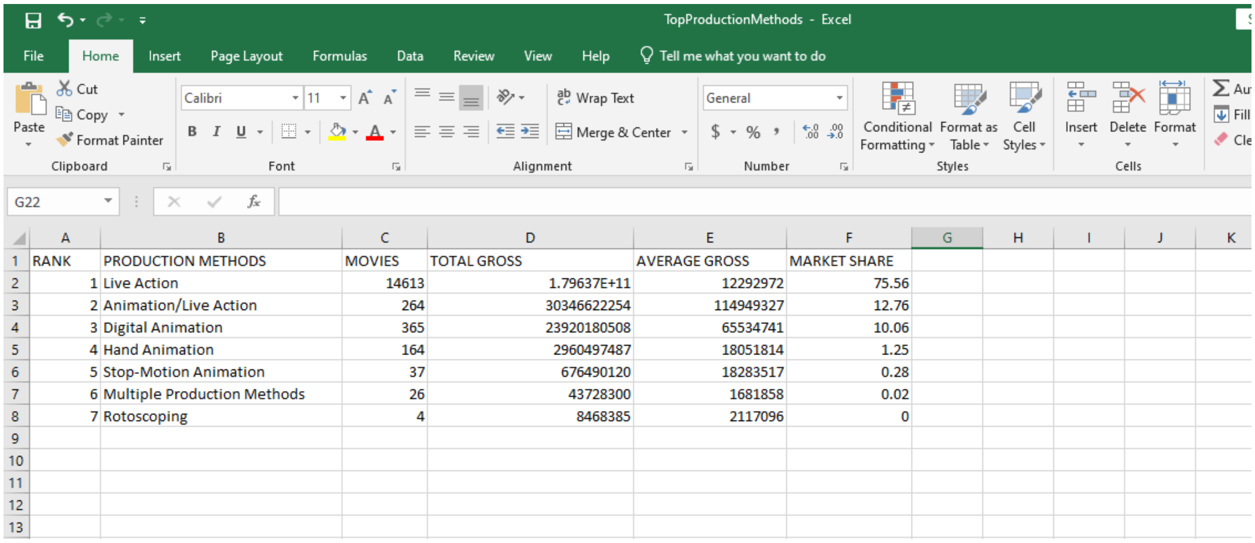
	A	B	C	D	E	F
1	RANK	PRODUCTION METHODS	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
2	1	Live Action	14,613	1.79637E+11	12292972	75.56%
3	2	Animation/Live Action	264	30346622254	114949327	12.76%
4	3	Digital Animation	365	23920180508	65534741	10.06%
5	4	Hand Animation	164	2960497487	18051814	1.25%
6	5	Stop-Motion Animation	37	676490120	18283517	0.28%
7	6	Multiple Production Methods	26	43728300	1681858	0.02%
8	7	Rotoscoping	4	8468385	2117096	0.00%

(3) Also, there is a column which has a symbol % in it. So, we remove it too for consistency.



	A	B	C	D	E	F
1	RANK	PRODUCTION METHODS	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
2	1	Live Action	14,613	1.79637E+11	12292972	75.56
3	2	Animation/Live Action	264	30346622254	114949327	12.76
4	3	Digital Animation	365	23920180508	65534741	10.06
5	4	Hand Animation	164	2960497487	18051814	1.25
6	5	Stop-Motion Animation	37	676490120	18283517	0.28
7	6	Multiple Production Methods	26	43728300	1681858	0.02
8	7	Rotoscoping	4	8468385	2117096	0

(4) Movies column has comma in the values. To maintain consistency, I removed them too.



	A	B	C	D	E	F
1	RANK	PRODUCTION METHODS	MOVIES	TOTAL GROSS	AVERAGE GROSS	MARKET SHARE
2	1	Live Action	14613	1.79637E+11	12292972	75.56
3	2	Animation/Live Action	264	30346622254	114949327	12.76
4	3	Digital Animation	365	23920180508	65534741	10.06
5	4	Hand Animation	164	2960497487	18051814	1.25
6	5	Stop-Motion Animation	37	676490120	18283517	0.28
7	6	Multiple Production Methods	26	43728300	1681858	0.02
8	7	Rotoscoping	4	8468385	2117096	0

(5) There are no duplicate values in the primary key column.

WideReleasesCount

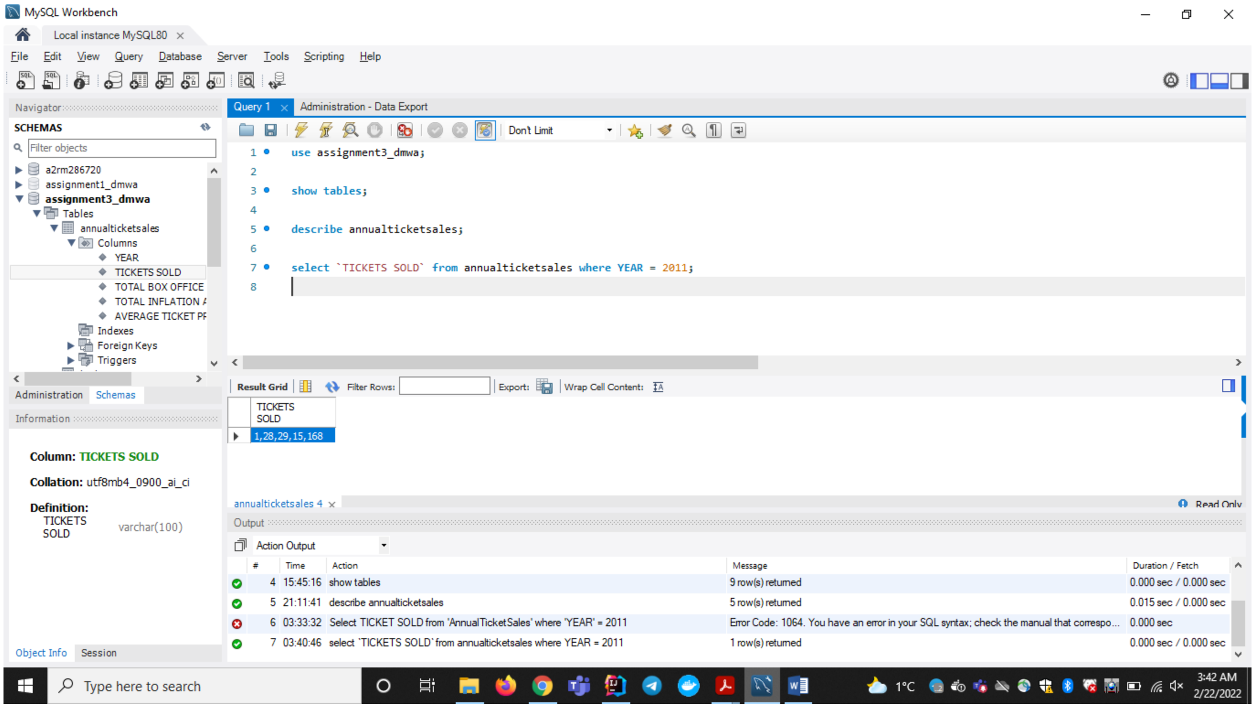
There is no primary key for this table.

(1) There are no blank values in this table.

(2) There are no special characters like **$** or **,** or **%**. Hence there is no cleaning to do in this dataset.

Two Phase Locking Protocol

Before Transaction:



```
use assignment3_dmwa;

show tables;

describe annualticketsales;

select `TICKETS SOLD` from annualticketsales where YEAR = 2011;
```

TICKETS SOLD
1,28,29,15,168

Column: TICKETS SOLD

Collation: utf8mb4_0900_ai_ci

Definition:
TICKETS SOLD varchar(100)

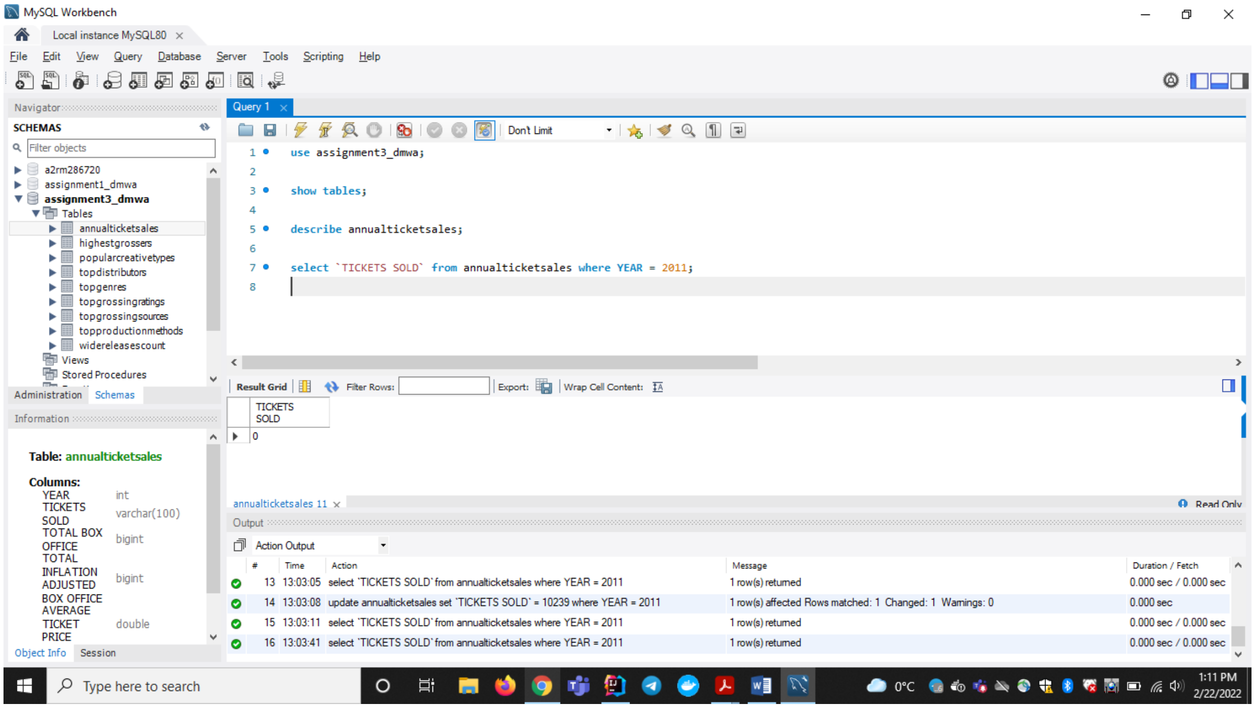
annualticketsales 4

Output

Action Output

#	Time	Action	Message	Duration / Fetch
4	15:45:16	show tables	9 row(s) returned	0.000 sec / 0.000 sec
5	21:11:41	describe annualticketsales	5 row(s) returned	0.015 sec / 0.000 sec
6	03:33:32	Select TICKET SOLD from 'AnnualTicketSales' where 'YEAR' = 2011	Error Code: 1064. You have an error in your SQL syntax; check the manual that correspo...	0.000 sec
7	03:40:46	select `TICKETS SOLD` from annualticketsales where YEAR = 2011	1 row(s) returned	0.000 sec / 0.000 sec

After Transaction:



```
use assignment3_dmwa;

show tables;

describe annualticketsales;

select `TICKETS SOLD` from annualticketsales where YEAR = 2011;
```

TICKETS SOLD
0

Table: annualticketsales

Columns:

YEAR	int
TICKETS SOLD	varchar(100)
TOTAL BOX OFFICE	bigint
TOTAL INFLATION ADJUSTED BOX OFFICE	bigint
AVERAGE TICKET PRICE	double

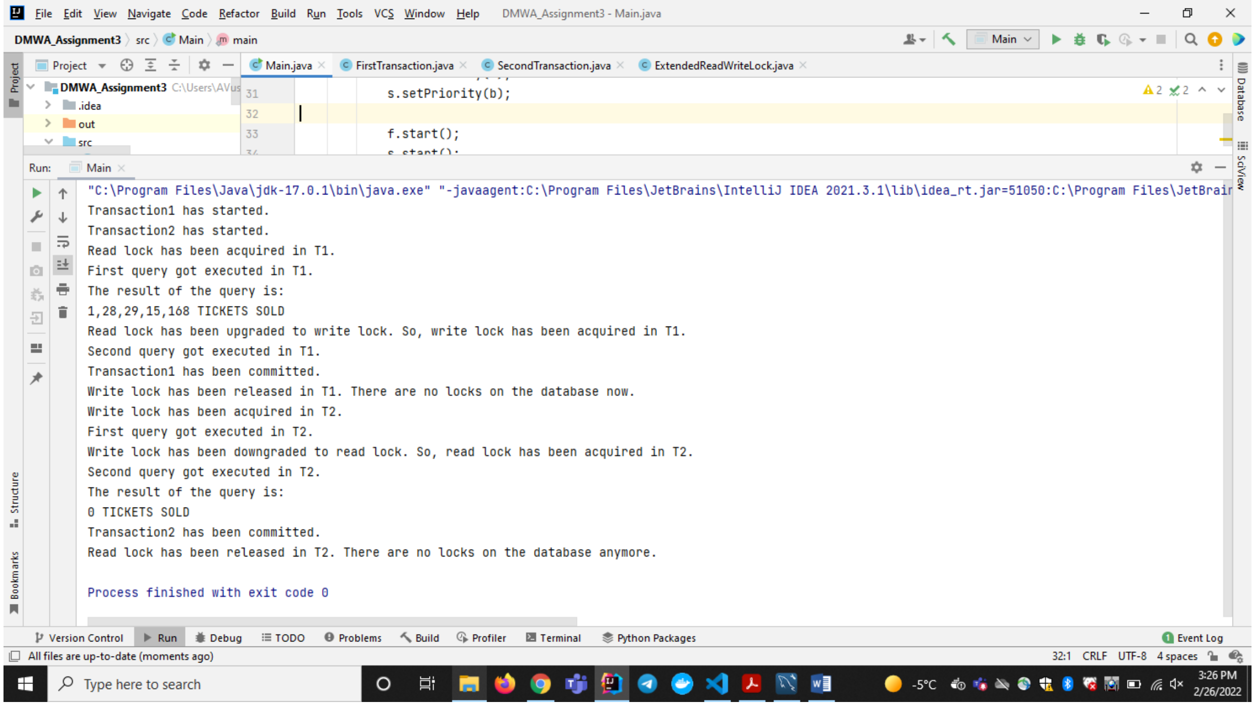
annualticketsales 11

Output

Action Output

#	Time	Action	Message	Duration / Fetch
13	13:03:05	select `TICKETS SOLD` from annualticketsales where YEAR = 2011	1 row(s) returned	0.000 sec / 0.000 sec
14	13:03:08	update annualticketsales set `TICKETS SOLD` = 10239 where YEAR = 2011	1 row(s) affected Rows matched: 1 Changed: 1 Warnings: 0	0.000 sec
15	13:03:11	select `TICKETS SOLD` from annualticketsales where YEAR = 2011	1 row(s) returned	0.000 sec / 0.000 sec
16	13:03:41	select `TICKETS SOLD` from annualticketsales where YEAR = 2011	1 row(s) returned	0.000 sec / 0.000 sec

Proof that code got executed:



```
        s.setPriority(b);

        f.start();
```

```
"C:\Program Files\Java\jdk-17.0.1\bin\java.exe" "-javaagent:C:\Program Files\JetBrains\IntelliJ IDEA 2021.3.1\lib\idea_rt.jar=51050:C:\Program Files\JetBrair
Transaction1 has started.
Transaction2 has started.
Read lock has been acquired in T1.
First query got executed in T1.
The result of the query is:
1,28,29,15,168 TICKETS SOLD
Read lock has been upgraded to write lock. So, write lock has been acquired in T1.
Second query got executed in T1.
Transaction1 has been committed.
Write lock has been released in T1. There are no locks on the database now.
Write lock has been acquired in T2.
First query got executed in T2.
Write lock has been downgraded to read lock. So, read lock has been acquired in T2.
Second query got executed in T2.
The result of the query is:
0 TICKETS SOLD
Transaction2 has been committed.
Read lock has been released in T2. There are no locks on the database anymore.

Process finished with exit code 0
```

Git URL for code: https://git.cs.dal.ca/rguturu/csci-5408-w2022-b00871849-gutururamamohanvishnu/-/tree/main/assignment-3

References:

[1] Tanwar, V., "*Two Phase Locking Protocol – GeeksforGeeks*," GeeksforGeeks [Online]. Available at: https://www.geeksforgeeks.org/two-phase-locking-protocol/ [Accessed: February 22, 2022].

[2] Purswani, A., "*Difference between Shared Lock and Exclusive Lock – GeeksforGeeks*," GeeksforGeeks [Online]. Available at: https://www.geeksforgeeks.org/difference-between-shared-lock-and-exclusive-lock/ [Accessed: February 22, 2022].

[3] "*ReentrantReadWriteLock - why can't reader acquire writer's lock?*," Stack Overflow [Online]. Available at: https://stackoverflow.com/questions/43099144/reentrantreadwritelock-why-cant-reader-acquire-writers-lock [Accessed: February 22, 2022].