Out of range values and inaccurate data

CLEANING DATA IN SQL SERVER DATABASES



Miriam Antona Software Engineer



Dataset - series and episodes

series

```
contact_number | rating | ...
id
                           premiered
                                      | official_site
      name
                          | 2010-04-05 | wwq.cartoonnetwork.com/video/adventuretime/ | 555-906-8845
      Adventure Time
                                                                                                    8,4
                                        ww.sho.com/sho/dexter/home
                                                                                                    8,6
2
      Dexter
                           2006-01-10
                                                                                     555-156-8845
                                        www.cc.com/shows/futurama
3
      Futurama
                           1999-03-28
                                                                                     555-210-9951
                                                                                                    9,2
                                                                                                              . . .
                                                                                                    9,3
      Game of Thrones
                           2011-04-17
                                        www.hbo.com/game-of-thrones
                                                                                     555-abc-6641
                                                                                                             | ... |
5
     Homeland
                           2011-10-02
                                        www.sho.com/sho/homeland/home
                                                                                     555-985-6314
                                                                                                    8,3
     Westworld
                                       www.hbo.com/westworld
                                                                                                    8,6
6
                          | 2016-10-02 |
                                                                                     555-456-1234
     Silicon Valley
                         | 2014-04-06 | www.hbo.com/silicon-valley/
                                                                                     555-604-1234
                                                                                                    | 11,4
                                        www.cbs.com/shows/big_bang_theory/
      The Big Bang Theory | 2007-09-24 |
                                                                                     555-607-1274
8
9
      Paw Patrol
                           2013-08-27
                                        www.nickjr.com/paw-patrol/videos/
                                                                                     555-930-1274
```



Dataset - series and episodes

episodes

episode_id	series_id	name	season	number	airdate	runtime	
					-		-
1	1	Slumber Party Panic	1	1	2010-04-05	15	-
2	1	Trouble in Lumpy Space	1	2	2010-04-05	15	1
3	1	Prisoners of Love	1	3	2010-04-12	15	1
4	1	Tree Trunks	1	4	2010-04-12	15	-
5	1	The Enchiridion!	1	5	2010-04-19	15	-
6	1	The Jiggler	1	6	2010-04-19	15	
7	1	Ricardio the Heart Guy	1	7	2010-04-26	15	
8	1	Business Time	1	8	2010-04-26	15	
9	1	My Two Favorite People	1	9	2010-05-03	15	
			l		1	···	

Out of range values

- Values outside the expected range of valid data
 - e.g. person 400 inches tall
- Can disrupt the results if not detected
- Investigate!



Out of range values - example

```
SELECT * FROM series
```

id	name	premiered	official_site	contact_number	rating	1 1
						-
1	Adventure Time	2010-04-05	wwq.cartoonnetwork.com/video/adventuretime/	555-906-8845	8,4	1
2	Dexter	2006-01-10	ww.sho.com/sho/dexter/home	555-156-8845	8,6	
3	Futurama	1999-03-28	www.cc.com/shows/futurama	555-210-9951	9,2	
4	Game of Thrones	2011-04-17	www.hbo.com/game-of-thrones	555-abc-6641	9,3	1 1
5	Homeland	2011-10-02	www.sho.com/sho/homeland/home	555-985-6314	8,3	1 1
6	Westworld	2016-10-02	www.hbo.com/westworld	555-456-1234	8,6	1 1
7	Silicon Valley	2014-04-06	www.hbo.com/silicon-valley/	555-604-1234	11,4	1 1
8	The Big Bang Theory	2007-09-24	www.cbs.com/shows/big_bang_theory/	555-607-1274	8	1 1
9	Paw Patrol	2013-08-27	www.nickjr.com/paw-patrol/videos/	555-930-1274	11	1 1
10		l I	•••	l	l	



Out of range values - example

```
SELECT * FROM series
```

id	name	premiered	official_site	contact_number	rating
1	Adventure Time	2010-04-05	wwq.cartoonnetwork.com/video/adventuretime/	555-906-8845	8,4
2	Dexter	2006-01-10	ww.sho.com/sho/dexter/home	555-156-8845	8,6
3	Futurama	1999-03-28	www.cc.com/shows/futurama	555-210-9951	9,2
4	Game of Thrones	2011-04-17	www.hbo.com/game-of-thrones	555-abc-6641	9,3
5	Homeland	2011-10-02	www.sho.com/sho/homeland/home	555-985-6314	8,3
6	Westworld	2016-10-02	www.hbo.com/westworld	555-456-1234	8,6
7	Silicon Valley	2014-04-06	www.hbo.com/silicon-valley/	555-604-1234	11,4 **
8	The Big Bang Theory	2007-09-24	www.cbs.com/shows/big_bang_theory/	555-607-1274	8
9	Paw Patrol	2013-08-27	www.nickjr.com/paw-patrol/videos/	555-930-1274	11 **
10		l I	•••	l	l l l



Out of range values - detecting the values

```
SELECT * FROM series
WHERE rating NOT BETWEEN 0 AND 10
```

```
SELECT * FROM series
WHERE rating < 0 OR rating > 10
```

Out of range values - excluding the values

```
SELECT * FROM series
WHERE rating BETWEEN 0 AND 10
```

id	name	premiered	official_site	contact_number	rating	summary
1	Adventure Time	2010-05-04	wwq.cartoonnetwork.com/video/adventuretime/	555-906-8845	8,4	l
2	Dexter	2006-01-10	ww.sho.com/sho/dexter/home	555-156-8845	8,6	l
3	Futurama	1999-03-03	www.cc.com/shows/futurama	555-210-9951	9,2	
4	Game of Thrones	2011-04-04	www.hbo.com/game-of-thrones	555-abc-6641	9,3	
5	Homeland	2011-02-10	www.sho.com/sho/homeland/home	555-985-6314	8,3	
6	Westworld	2016-02-10	www.hbo.com/westworld	555-456-1234	8,6	
8	The Big Bang Theory	2007-09-24	www.cbs.com/shows/big_bang_theory/	555-607-1274	8	
	<u> </u>	l	l	l		·



Out of range values - excluding the values

```
SELECT * FROM series
WHERE rating >= 0 AND rating <=10</pre>
```

id	name	premiered	official_site	contact_number	rating	summary	
	- Adventure Time			 555-906-8845	 0 /	 !	
1 2	Dexter		<pre> wwq.cartoonnetwork.com/video/adventuretime/ ww.sho.com/sho/dexter/home</pre>		8,4 8,6	 	
3	Futurama		www.cc.com/shows/futurama		9,2		
4	Game of Thrones	2011-04-04	www.hbo.com/game-of-thrones	555-abc-6641	9,3	l	I
5	Homeland	2011-02-10	www.sho.com/sho/homeland/home	555-985-6314	8,3	l	
6	Westworld	2016-02-10	www.hbo.com/westworld	555-456-1234	8,6	l	
8	The Big Bang Theory	2007-09-24	www.cbs.com/shows/big_bang_theory/	555-607-1274	8		
· · · ·	1	l	l	l	l	· · · ·	



Inaccurate data

- Two or more values are contradictory
- Can disrupt the results if not detected
- Investigate!



Inaccurate data - example

series

```
| id | name | premiered | official_site | contact_number | rating | summary |
|----|-----|-----|-----|-----|
```

episodes

```
| episode_id | series_id | name | season | number | airdate | runtime |
|-----|----|----|
```

Valid: episodes.airdate >= series.premiered

Inaccurate data - example

Inaccurate data - detecting the values

```
SELECT
    series.id,
    series.name,
    series.premiered,
    episodes.episode_id,
    episodes.name,
    episodes.airdate
FROM series
INNER JOIN episodes ON series.id = episodes.series_id
WHERE episodes.airdate < series.premiered</pre>
```

```
premiered
                             episode_id |
                                                      l airdate
id
    name
                                         name
                                                       _____
                                                      2010-10-02
               2011-10-02 | 58
    Homeland
                                         Pilot
    Westworld | 2016-10-02 | 70
                                         The Original | 2015-10-02 |
                                                      | 2015-10-30 |
    Westworld | 2016-10-02 | 74
                                         Contrapasso
```

Inaccurate data - excluding the values

```
SELECT
    series.id,
    series.name,
    series.premiered,
    episodes.episode_id,
    episodes.name,
    episodes.airdate
FROM series
INNER JOIN episodes ON series.id = episodes.series_id
WHERE episodes.airdate >= series.premiered
```

1	episode_id	series_id	name	season	number	airdate	runtime
-1							
- [1	1	Slumber Party Panic	1	1	2010-04-05	15
-1	2	1	Trouble in Lumpy Space	1	2	2010-04-05	15
-1	3	1	Prisoners of Love	1	3	2010-04-12	15
-1	4	1	Tree Trunks	1	4	2010-04-12	15
1		1	l	l	l	l	l



Let's practice!

CLEANING DATA IN SQL SERVER DATABASES



Converting data with different types

CLEANING DATA IN SQL SERVER DATABASES



Miriam Antona Software Engineer



Introduction to undesirable data types

```
SELECT * FROM series
WHERE rating BETWEEN 0 AND 10

SELECT * FROM series
WHERE rating >= 0 AND rating <= 10</pre>
```



Introduction to undesirable data types

```
SELECT * FROM series
WHERE rating BETWEEN 0 AND 10
```

```
SELECT * FROM series
WHERE rating >= 0 AND rating <= 10</pre>
```

id	name	rating
	-	-
1	Adventure Time	8.4
2	Dexter	8.6
3	Futurama	9.2
4	Game of Thrones	9.3
<u> </u>	1	1

Column_name	Type
id	int
name	varchar
premiered	date
official_site	varchar
contact_number	varchar
rating	float
summary	varchar

Introduction to undesirable data types

Column_name	Type
id	int
name	varchar
premiered	date
official_site	varchar
contact_number	varchar
rating	varchar
summary	varchar

Undesirable data types - examples

```
SELECT * FROM series
WHERE rating BETWEEN 0 AND 10
```

```
SELECT * FROM series
WHERE rating >= 0 AND rating <= 10</pre>
```

Conversion failed when converting the varchar value '8.4' to data type int.

Undesirable data types - examples

SELECT AVG(rating)

FROM series

Operand data type varchar is invalid for avg operator.



Solving type conversion problems

```
WHERE CAST(rating AS FLOAT) BETWEEN 0 AND 10

SELECT * FROM series
WHERE CONVERT(FLOAT, rating) BETWEEN 0 AND 10
```

SELECT * **FROM** series

Solving type conversion problems

```
SELECT AVG(CAST(rating AS FLOAT)) AS rating_casted
FROM series
WHERE CAST(rating AS FLOAT) BETWEEN 0 AND 10

SELECT AVG(CONVERT(FLOAT, rating)) AS rating_casted
FROM series
WHERE CONVERT(FLOAT, rating) BETWEEN 0 AND 10
```



Let's practice!

CLEANING DATA IN SQL SERVER DATABASES



Pattern matching

CLEANING DATA IN SQL SERVER DATABASES



Miriam Antona Software Engineer



Pattern matching - Introduction

```
SELECT * FROM series
```

- - first and fourth numbers between 2 and 9
 - the rest between 0 and 9
- Invalid number: 555-156-8845



Pattern matching - Introduction

- SQL Server doesn't provide full blown set of regular expressions.
- SQL Server can match patters using LIKE.
- To have the full blown set of regular expressions -> create and install extensions.

Pattern matching - LIKE

Determines if a string matches a specified pattern



Pattern matching - LIKE

• Determines if a string matches a specified pattern

Wildcard character	Description	Example
%	Any string of zero or more characters	WHERE contact_number LIKE '555-%'
- (underscore)	Any single character	WHERE contact_number LIKE ''
	Any single character within the specified range or set	WHERE contact_number LIKE '[2-9][0-9][0-9]-[2-9][0-9][0-9][0-9][0-9][0-9]
[^]	Any single character not within the specified range or set	WHERE contact_number LIKE '[^2-9]'

Pattern matching - example with %

```
SELECT name, contact_number
FROM series
WHERE contact_number LIKE '555%'
```

Pattern matching - example with %

```
SELECT
  name,
  contact_number
FROM series
WHERE contact_number NOT LIKE '555%'
```

Pattern matching - example with [] (brackets)

```
SELECT
    name,
    contact_number
FROM series
WHERE contact_number LIKE '[2-9][0-9][0-9]-[2-9][0-9][0-9]-[0-9][0-9][0-9]'
```

Pattern matching - example with [] (brackets)

```
SELECT
    name,
    contact_number
FROM series
WHERE contact_number NOT LIKE '[2-9][0-9][0-9]-[2-9][0-9][0-9]-[0-9][0-9][0-9]'
```

Let's practice!

CLEANING DATA IN SQL SERVER DATABASES

