Audible dataset cleaning

1, Created the table

```
CREATE TABLE audible

(

        NAME VARCHAR(300),

        author VARCHAR(200),

        narrator VARCHAR(200),

        time VARCHAR(80),

        releasedate DATE,

        language VARCHAR(150),

        stars VARCHAR(100),

        price VARCHAR(40)

)
```

2, updated all **not rated yet** columns to **0** stars

```
UPDATE audible

SET stars = 0.0

WHERE stars = 'Not rated yet'
```

3, added (number of ratings) column into dataset

```
ALTER TABLE audible

ADD COLUMN ratings VARCHAR(10);
```

```
UPDATE audible

SET ratings = CASE

                WHEN stars ILIKE '%rating' THEN SUBSTRING(stars, POSITION('stars' IN stars) + 5, (POSITION('rating' IN stars)-1) - (POSITION('stars' IN stars) + 5))
```

```
                                        WHEN stars ILIKE '%ratings' THEN SUBSTRING(stars,
POSITION('stars' IN stars) + 5, (POSITION('ratings' IN  stars)-1) - (POSITION('stars' IN stars) + 5))

                                        ELSE '0.0'

                                        END
```

4, before converting it into integer first replace the ‘,’(comma) with empty string eg 2,343 to 2343

```
UPDATE audible

SET ratings = REPLACE( ratings, ',', '')
```

5, CHANGE DATA TYPE using ALTER TABLE ,ALTER COLUMN

```
ALTER TABLE audible

ALTER COLUMN ratings SET DATA TYPE INTEGER USING ratings::INTEGER
```

6, Changed price field where it was free in price column

```
UPDATE audible

SET price = 0

WHERE price = 'Free'
```

7, again replaced commas in varchar for price column

```
UPDATE audible

SET price = REPLACE(price, ',' , '')
```

8, changed data type from character to Decimal in price column using following because it has some values with cents.

```
ALTER TABLE audible

ALTER COLUMN price SET DATA TYPE DECIMAL(10,4) USING price::DECIMAL(10,4)
```

9, updated stars info to only give actual stars removing all unnecessary info

```
UPDATE audible
SET stars = LEFT(stars , POSITION('out of' IN stars) - 2)
```

10, changed data type to decimal for stars

```
ALTER TABLE audible
ALTER COLUMN stars SET DATA TYPE DECIMAL(2,1) USING stars::DECIMAL(2,1)
```

11, Updated author column to remove Written by text

```
update audible
SET author = SUBSTRING(author, POSITION('by:' IN author)+3, LENGTH(author))
```

IMPORTANT !! , if you run update command twice you might lose some words from your data and that can create the mess, also it is advised to update data on copy of your dataset and don't run command twice even if your system lags and it didn't do it on the first go, just wait for some time before hitting the run button and other important advice BEFORE UPDATING CHECK WHAT IS THE OUTPUT OF YOUR CODE BY USING IT IN SELECT STATEMENT !!

ALSO creating dummy columns for reference can help avoiding errors!

12, SIMILARLY for narrator removed narrated by

```
update audible
SET narrator = SUBSTRING(narrator, POSITION('by:' IN narrator)+3, LENGTH(narrator))
```

U can always check results by using logic in SELECT statement eg,

```
SELECT SUBSTRING(narrator, POSITION('by:' IN narrator)+3, LENGTH(narrator)) FROM audible
```

13, added a column to extract hours only

```
ALTER TABLE audible
```

```
ADD COLUMN hours VARCHAR(10)
```

14, extracted the hours in separate column

```
UPDATE audible
SET hours = CASE WHEN time ILIKE '%hr%' THEN LEFT(time , POSITION(' ' IN time)) ELSE '0' END
```

15, added mins column and then altered it coz format was not fitting in varchar(10) , **(later while cleaning found there are some values like ‘Less than 1 minute’)**

```
ALTER TABLE audible
SET mins DATA TYPE VARCHAR(30) USING mins::VARCHAR(30)
```

16, updated mins table

```
UPDATE audible
SET mins = CASE
                        WHEN time ILIKE '%and%' THEN SUBSTRING(time , POSITION('and ' IN time) + 4 , LENGTH(time) )
                        WHEN time ILIKE '%hr' THEN '0'
                  ELSE LEFT(time , POSITION(' min' IN time)) END
```

17, removed empty cells using

```
update audible
SET mins = 0
WHERE mins = ''
```

18, removed min or mins text from min column

```
update audible
SET mins = CASE WHEN mins ILIKE '%min%' THEN LEFT(mins, POSITION(' min' IN mins)) ELSE mins END
```

19, mins column nees to be trimmed

```
update audible
SET mins = TRIM(mins)
```

20, Updated the mins column to handle ‘less than 1 minute’ text

```
UPDATE audible
SET mins = '1'
WHERE mins ILIKE '%Less%
```

21, Altered data type of mins to smallint

```
ALTER TABLE audible
ALTER COLUMN mins SET DATA TYPE SMALLINT USING mins::SMALLINT
```

22, changed time column to show only mins instead of mix of hour and min data

```
UPDATE audible
SET time = hours*60 + mins
```

23, Changed data type to integer again for time column for comparison purposes

```
ALTER TABLE audible
ALTER COLUMN time SET DATA TYPE SMALLINT USING time::SMALLINT
```

24, DELETED the dummy columns that were created for reference like author_copy , hours , mins

```
ALTER TABLE audible

DROP COLUMN author_copy;


ALTER TABLE audible

DROP COLUMN hours;


ALTER TABLE audible

DROP COLUMN mins;
```

WE OBTAINED THE CLEAN DATASET THAT CAN BE USED FOR FURTHER ANALYSIS.

THANK YOU!