Data Cleaning of Earthquake dataset using SQL

Steps to follow while cleaning data:

1. Look the **datatypes** of each column. Check if the datatype of date is in date format and time is in time format, and all numeric columns are either integer or double instead of text datatype. Change datatypes of the columns accordingly.
2. Check the **length of the date and time column**.Check if the length is uniform across all rows of each column
3. Check the **rows with missing values**. Either mean,median or mode can be added to the missing row depending on the context of the column. Certain columns can have zero to be filled in the missing place and check the datatype of such column whether it is numeric or double.
4. **Extract year**, month, dayname, week from the Date column and put them as separate columns
5. Check for **duplicate rows**
6. Check for **outlier**. Remove if there is any.

Dataset:

This dataset includes a record of the date, time, location, depth, magnitude, and source of every earthquake with a reported magnitude 5.5 or higher since 1965.

Schema named earthquake created

```
create schema earthquake;
use earthquake;


-- DATA DISCOVERY--
select * from eqdatabase;
```

Datatypes of the individual column checked

```
-- TO CHECK THE DATA TYPE OF INDIVIDUAL COLUMNS IN TABLE--
select DATA_TYPE from Information_schema.columns where
table_schema='earthquake' and table_name='eqdatabase';
```

Information

Table: eqdatabase

Columns:

Date	text
Time	text
Latitude	double
Longitude	double
Type	text
Depth	double
Depth Error	text
Depth Seismic Stations	text
Magnitude	double
Magnitude Type	text
Magnitude Error	text
Magnitude Seismic Stations	text
Azimuthal Gap	text
Horizontal Distance	text
Horizontal Error	text
Root Mean Square	text
ID	text
Source	text
Location Source	text

DATA_TYPE
text
text
text
double
text
double
double
text
text
text
text
text
text

Date and Time columns are in text format. They need to be changed to date and time format. Before that inconsistency in the date column must be rectified.

Date
1975-02-23T02:58:41.000Z
1985-04-28T02:53:41.530Z

```
-- TO CHECK INCONSISTENCY IN DATE COLUMN--
SELECT length(Date)
from eqdatabase;

SELECT MAX(length(date))
from eqdatabase; -- 24--

SELECT min(length(date))
from eqdatabase; -- 10--

UPDATE eqdatabase
SET date=LEFT(date,10)
where length(date)=24;

SELECT MAX(length(date))
from eqdatabase;-- 10--
```

Date column must be standardised- new column Date2 to be added

```
-- STANDARDIZE DATE FORMAT--
ALTER TABLE eqdatabase
ADD COLUMN Date2 Date after Date;
```

Date	Date2
02-01-1965	NULL
04-01-1965	NULL
05-01-1965	NULL
08-01-1965	NULL
09-01-1965	NULL
10-01-1965	NULL

Text to date conversion of the Date column

```
UPDATE eqdatabase
SET Date2= str_to_date(Date,'%d-%m-%y');
```

Error occurred while updating as there were three rows with different date format.

7 08:57:34 UPDATE eqdatabase SET Date2= str_to_date(Date,'%d-%m-%y') Error Code: 1292. Truncated incorrect date value: '02-01-1965' 0.016 sec

```
SELECT Date,str_to_date(Date,'%d-%m-%y')
from eqdatabase
where str_to_date(Date,'%d-%m-%y') is null;
```

Date	str_to_date(Date,'%d-%m-%y')
1975-02-23	NULL
1985-04-28	NULL
2011-03-13	NULL

Those 3 rows alone replaced with correct date format

```
UPDATE eqdatabase
set Date=replace(Date,'1975-02-23','23-02-1975');
UPDATE eqdatabase
set Date=replace(Date,'1985-04-28','28-04-1985');
UPDATE eqdatabase
set Date=replace(Date,'2011-03-13','13-03-2011');
```

	Date	str_to_date(Date,'%d-%m-%y')

Date2 column in Date format (or datatype) updated successfully

```
UPDATE eqdatabase
SET Date2= str_to_date(`Date`,'%d-%m-%Y');
```

	Date	Date2
	02-01-1965	1965-01-02
	04-01-1965	1965-01-04
	05-01-1965	1965-01-05
	08-01-1965	1965-01-08
	09-01-1965	1965-01-09
	10-01-1965	1965-01-10

Time column must be standardised- new column Time2 to be added

```
-- STANDARDIZE TIME FORMAT--
ALTER TABLE eqdatabase
ADD COLUMN Time2 time after Time;
```

Text to time conversion of the Time column

```
UPDATE eqdatabase
set Time2=cast(Time as time);
```

Inconsistency in the time data


```
select max(length(time))
from eqdatabase;
```

max(length(time))
24

```
select min(length(time))
from eqdatabase;
```

min(length(time))
7

Three rows with inconsistent time data.

```
select time
from eqdatabase
where length(time) = 24;
```

time
1975-02-23T02:58:41.000Z
1985-04-28T02:53:41.530Z
2011-03-13T02:23:34.520Z

Those three rows in the time column must be replaced

```
UPDATE eqdatabase
set time=replace(time,'1975-02-23T02:58:41.000Z',substr(time,12,8))
where time='1975-02-23T02:58:41.000Z';
```

```
UPDATE eqdatabase
set time=replace(time,'1985-04-28T02:53:41.530Z',substr(time,12,8))
where time='1985-04-28T02:53:41.530Z';

 select time
 from eqdatabase
 where length(time) = 24;
```

	time
▶	2011-03-13T02:23:34.520Z

```
UPDATE eqdatabase
set time=replace(time,'2011-03-13T02:23:34.520Z',substr(time,12,8))
where time='2011-03-13T02:23:34.520Z';

  select time
  from eqdatabase
  where length(time) = 24;
```

	time

Time2 column in Time format (or datatype) updated successfully

```
UPDATE eqdatabase
set Time2=cast(Time as time);
```

Time	Time2
13:44:18	13:44:18
11:29:49	11:29:49
18:05:58	18:05:58
18:49:43	18:49:43
13:32:50	13:32:50
13:36:32	13:36:32

Other columns having missing values checked. Text datatype to be modified to double datatype. But before that missing rows should be filled with zeros.

```
-- HANDLING BLANK ROWS IN Depth Error,Depth Seismic,Magnitude Error, Magnitude Seismic Stations,Azimuthal Gap,Horizontal Distance,
-- Horizontal Error, Root Mean Square

SELECT COUNT(`Depth error`) from eqdatabase where `Depth error`='';
```

Count number of rows with no values in each column

```
SELECT COUNT(`Depth error`) from eqdatabase where `Depth error`='';
SELECT COUNT(`Depth Seismic Stations`) from eqdatabase where `Depth Seismic Stations`='';
SELECT COUNT(`Magnitude Error`) from eqdatabase where `Magnitude Error`='';
SELECT COUNT(`Magnitude Seismic Stations`) from eqdatabase where `Magnitude Seismic Stations`='';
SELECT COUNT(`Azimuthal Gap`) from eqdatabase where `Azimuthal Gap`='';
SELECT COUNT(`Horizontal Distance`) from eqdatabase where `Horizontal Distance`='';
SELECT COUNT(`Horizontal Error`) from eqdatabase where `Horizontal Error`='';
SELECT COUNT(`Root Mean Square`) from eqdatabase where `Root Mean Square`='';
```

	COUNT(`Depth error`)
▶	18951

	COUNT(`Depth Seismic Stations`)
▶	16315

	COUNT(`Magnitude Error`)
▶	23085

	COUNT(`Magnitude Seismic Stations`)
▶	20848

	COUNT(`Azimuthal Gap`)
	16113

	COUNT(`Horizontal Distance`)
▶	21808

COUNT(`Horizontal Error`)
22256

COUNT(`Root Mean Square`)
6060

Update zeros to the missing rows of specific columns

```
-- Update zeros in the missing rows of the text columns so that those columns could be changed to double-
UPDATE eqdatabase
set `Depth error`= CASE WHEN `Depth error` ='' THEN 0.0
                        ELSE `Depth error`
                        END;

UPDATE eqdatabase
set `Depth Seismic Stations`= CASE WHEN `Depth Seismic Stations` ='' THEN 0.0
                        ELSE `Depth Seismic Stations`
                        END;

UPDATE eqdatabase
set `Magnitude Error`= CASE WHEN `Magnitude Error` ='' THEN 0.0
                        ELSE `Magnitude Error`
                        END;

UPDATE eqdatabase
set `Magnitude Error`= CASE WHEN `Magnitude Error` ='' THEN 0.0
                          ELSE `Magnitude Error`
                          END;

UPDATE eqdatabase
set `Magnitude Seismic Stations`= CASE WHEN `Magnitude Seismic Stations` ='' THEN 0.0
                          ELSE `Magnitude Seismic Stations`
                          END;

UPDATE eqdatabase
set `Azimuthal Gap`= CASE WHEN `Azimuthal Gap` ='' THEN 0.0
                          ELSE `Azimuthal Gap`
                          END;

UPDATE eqdatabase
set `Horizontal Distance`= CASE WHEN `Horizontal Distance` ='' THEN 0.0
                          ELSE `Horizontal Distance`
                          END;
```

Modified the datatype to double

```
-- MODIFY THE DATA TYPE OF THE ABOVE TEXT COLUMNS TO DOUBLE--
ALTER TABLE eqdatabase MODIFY COLUMN `Depth error` double;
ALTER TABLE eqdatabase MODIFY COLUMN `Depth Seismic Stations` double;
ALTER TABLE eqdatabase MODIFY COLUMN `Magnitude Error` double;
ALTER TABLE eqdatabase MODIFY COLUMN `Magnitude Seismic Stations` double;
ALTER TABLE eqdatabase MODIFY COLUMN `Azimuthal Gap` double;
ALTER TABLE eqdatabase MODIFY COLUMN `Horizontal Distance` double;
ALTER TABLE eqdatabase MODIFY COLUMN `Horizontal Error` double;
ALTER TABLE eqdatabase MODIFY COLUMN `Root Mean Square` double;
```

Checked for duplicate rows

```
-- TO CHECK FOR DUPLICATES
with t1 as
(SELECT *, row_number() over(partition by Date2,Time2,Latitude,Longitude order by id) as rownum
FROM eqdatabase)
select count(*) from t1 where rownum>1
```

	count(*)
▶	0

Extracted Year, month, day name, week from Date2 column and added in separate columns

```
-- TO EXTRACT YEAR , MONTH AND DAY FROM Date column
SELECT EXTRACT(Year FROM Date2) from eqdatabase;

ALTER TABLE eqdatabase
ADD COLUMN Year int after Time2;

UPDATE eqdatabase
set Year= EXTRACT(Year FROM Date2);


SELECT EXTRACT(Month FROM Date2) from eqdatabase;

ALTER TABLE eqdatabase
ADD COLUMN Month int after Year;

UPDATE eqdatabase
set Month= EXTRACT(Month FROM Date2);
```

```sql
-- EXTRACT WEEK--
SELECT WEEK(DATE2,0) from eqdatabase;

ALTER TABLE eqdatabase
ADD COLUMN WEEK int after MONTH;

UPDATE eqdatabase
set Week= WEEK(DATE2,0);

ALTER TABLE eqdatabase
RENAME COLUMN WEEK TO Week;
```

```sql
-- Extract dayname
SELECT dayname(DATE2) from eqdatabase;

ALTER TABLE eqdatabase
ADD COLUMN `Day name` character after Week;

UPDATE eqdatabase
set `Day name`= dayname(DATE2);

ALTER TABLE eqdatabase
modify COLUMN `Day name` character(15) after Week;

UPDATE eqdatabase
set `Day name`= dayname(DATE2);
```

Checked for outliers

```sql
-- check for outliers--
select * from eqdatabase where Magnitude < 5.5;
select * from eqdatabase where Year > 2016 or Year < 1965;
```