-- Data Cleaning

-- Get an overview of the data

SELECT \*

FROM world\_life\_expectancy;

-- Find Duplicates

SELECT Country, Year, CONCAT(Country, Year), COUNT(CONCAT(Country, Year))

FROM world\_life\_expectancy

GROUP BY Country, Year, CONCAT(Country, Year)

HAVING COUNT(CONCAT(Country, Year)) > 1;

-- Remove Duplicates

SELECT \*

FROM (

SELECT Row\_ID, CONCAT(Country, Year),

ROW\_NUMBER() OVER (PARTITION BY CONCAT(Country, Year) ORDER BY CONCAT(Country, Year)) AS Row\_Num

FROM world\_life\_expectancy

) AS Row\_table

WHERE Row\_Num > 1;

DELETE FROM world\_life\_expectancy

WHERE

Row\_ID IN (

SELECT Row\_ID

FROM (

SELECT Row\_ID,

CONCAT(Country, Year),

ROW\_NUMBER() OVER (PARTITION BY CONCAT(Country, Year) ORDER BY CONCAT(Country, Year)) AS Row\_Num

FROM world\_life\_expectancy

) AS Row\_table

WHERE Row\_Num > 1

);

-- Check for blanks

SELECT \*

FROM world\_life\_expectancy

WHERE Status = '';

SELECT DISTINCT (Status)

FROM world\_life\_expectancy

WHERE Status <> '';

SELECT DISTINCT (Country)

FROM world\_life\_expectancy

WHERE Status = 'Developing';

-- populate missing data

UPDATE world\_life\_expectancy

SET Status = 'Developing'

WHERE Country IN (

SELECT DISTINCT (Country)

FROM world\_life\_expectancy

WHERE Status = 'Developing');

UPDATE world\_life\_expectancy t1

JOIN world\_life\_expectancy t2

ON t1.Country = t2.Country

SET t1.Status = 'Developing'

WHERE t1.Status = ''

AND t2.Status <> ''

AND t2.Status = 'Developing';

-- Do it for developed too

UPDATE world\_life\_expectancy t1

JOIN world\_life\_expectancy t2

ON t1.Country = t2.Country

SET t1.Status = 'Developed'

WHERE t1.Status = ''

AND t2.Status <> ''

AND t2.Status = 'Developed';

-- Check for NULLS

SELECT \*

FROM world\_life\_expectancy

WHERE Status IS NULL;

-- Check for blanks in life expectancy

SELECT \*

FROM world\_life\_expectancy

WHERE `Life expectancy` = '';

-- populate missing life expectancy data by calculating the average of the previous and next years' life expectancy values for the same country.

SELECT t1.Country, t1.Year, t1.`Life expectancy`,

t2.Country, t2.Year, t2.`Life expectancy`,

t3.Country, t3.Year, t3.`Life expectancy`,

ROUND((t2.`Life expectancy` + t3.`Life expectancy`)/2, 1)

FROM world\_life\_expectancy t1

JOIN world\_life\_expectancy t2

ON t1.Country = t2.Country

AND t1.Year = t2.Year - 1

JOIN world\_life\_expectancy t3

ON t1.Country = t3.Country

AND t1.Year = t3.Year + 1

WHERE t1.`Life expectancy` = '';

UPDATE world\_life\_expectancy t1

JOIN world\_life\_expectancy t2

ON t1.Country = t2.Country

AND t1.Year = t2.Year - 1

JOIN world\_life\_expectancy t3

ON t1.Country = t3.Country

AND t1.Year = t3.Year + 1

SET t1.`Life expectancy` = ROUND((t2.`Life expectancy` + t3.`Life expectancy`)/2, 1)

WHERE t1.`Life expectancy` = '';

SELECT \*

FROM world\_life\_expectancy;