Data Cleaning in SQL

Data cleaning in SQL refers to the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in a database to ensure the data is accurate, reliable, and suitable for analysis or other purposes. It involves various techniques and operations to handle missing values, handle outliers, standardize data formats, remove duplicate records, and resolve inconsistencies.

For example we have this data about Product in some company using Sql Developer and we want to clean them. So what we can do is :

	ID_PROD...	NAME	PRICE	YEAR_PROD	MATERIAL	REVIEW
1	L001	ARKHA	12000	08-JUL-07	Wood	3.58
2	L002	BIAN	12000	07-JUL-07	Vinyl	4.12
3	L003	MAHA	18000	08-JUL-07	Glass	3.18
4	L004	BRIX	18500	09-JUL-07	Wood	3.75
5	L005	ANEKA	19000	07-JUL-07	Wood	4.28
6	L006	RAHYA	10000	09-JUL-07	Not Ava...	3.78
7	L007	KEFAS	12500	08-JUL-08	N/A	4.18
8	L008	(null)	15000	08-JUL-08	Wood	3.58

1. Remove Irrelevant Data

   It can figure out what data is relevant to my analyses and the question I'm asking. Let's say I'm only interested in product that using the material wood. Data from Product that using outside of the material wood will skew my results, and I should remove them from the dataset. I can filter them out with the following statement :

```
select * from PRODUCT where MATERIAL = 'Wood';
```

   With this statement, we will only display the data that has the material wood.

	ID_PRODUCT	NAME	PRICE	YEAR_PROD	MATERIAL	REVIEW
1	L001	ARKHA	12000	08-JUL-07	Wood	3.58
2	L004	BRIX	18500	09-JUL-07	Wood	3.75
3	L005	ANEKA	19000	07-JUL-07	Wood	4.28
4	L008	(null)	15000	08-JUL-08	Wood	3.58

2. Using mathematic function

   Sometimes the data is too complicated and messy to look at. So we can use the mathematics function to make it cleaner and more tidy. Let's say I want to make the review column don't have any decimal value. I can use a statement to make it happen :

```
select ID_PRODUCT, NAME, PRICE,
YEAR_PROD, MATERIAL, ROUND(REVIEW)
from PRODUCT;
```

Rounding numbers after the decimal point will be adjusted automatically according to the general rules of mathematics. If the number is above 5, it will be rounded up. If the number is below 5, it will be rounded down.

	ID_PRODUCT	NAME	PRICE	YEAR_PROD	MATERIAL	ROUND(REVIEW)
1	L001	ARKHA	12000	08-JUL-07	Wood	4
2	L002	BIAN	12000	07-JUL-07	Vinyl	4
3	L003	MAHA	18000	08-JUL-07	Glass	3
4	L004	BRIX	18500	09-JUL-07	Wood	4
5	L005	ANEKA	19000	07-JUL-07	Wood	4
6	L006	RAHYA	10000	09-JUL-07	Not Available	4
7	L007	KEFAS	12500	08-JUL-08	N/A	4
8	L008	(null)	15000	08-JUL-08	Wood	4

3. Find Missing Data

   In data above we have a missing value in NAME column. If I'm trying to understand which product is having the best review from customer there is no way else than removing the missing value. We can use this statement to remove the missing value :

```
select * from PRODUCT where NAME IS NOT NULL;
```

	ID_PRODUCT	NAME	PRICE	YEAR_PROD	MATERIAL	REVIEW
1	L001	ARKHA	12000	08-JUL-07	Wood	3.58
2	L002	BIAN	12000	07-JUL-07	Vinyl	4.12
3	L003	MAHA	18000	08-JUL-07	Glass	3.18
4	L004	BRIX	18500	09-JUL-07	Wood	3.75
5	L005	ANEKA	19000	07-JUL-07	Wood	4.28
6	L006	RAHYA	10000	09-JUL-07	Not Available	3.78
7	L007	KEFAS	12500	08-JUL-08	N/A	4.18

   Actually, the missing data can be replaced by other observations and category but in this case we don't really have any idea or option to replace on other observation.

4. Fix Structural Errors

In data above we have Material column that have the values N/A and Not Available. To make it clear and not ambiguous we can update the data by using this statement :

```
UPDATE PRODUCT SET MATERIAL = NULL WHERE ID_PRODUCT IN('L006', 'L007');
```

	ID_PROD...	NAME	PRICE	YEAR_PROD	MATERIAL	REVIEW
1	L001	ARKHA	12000	08-JUL-07	Wood	3.58
2	L002	BIAN	12000	07-JUL-07	Vinyl	4.12
3	L003	MAHA	18000	08-JUL-07	Glass	3.18
4	L004	BRIX	18500	09-JUL-07	Wood	3.75
5	L005	ANEKA	19000	07-JUL-07	Wood	4.28
6	L006	RAHYA	10000	09-JUL-07	(null)	3.78
7	L007	KEFAS	12500	08-JUL-08	(null)	4.18
8	L008	(null)	15000	08-JUL-08	Wood	3.58

So it can help us to see the data more clearly and also can thrives us to a better conclusion or analyses.

5. Gathering Columns in SQL

   To make the data more easy and more less columns we can combine those columns that can be added together. In example we have the player table.

	PLAYERNO	NAME	INITIALS	BIRTH_DATE	SEX	JOINED	STREET	HOUSENO	POSTCODE	TOWN	PHONENO	LEAGUENO
1	2	Everett	R	01-SEP-48	M	1975	Stoney Road	43	3575NH	Stratford	070-237893	2411
2	6	Parmenter	R	25-JUN-64	M	1977	Haseltine Lane	80	1234KK	Stratford	070-476537	8467
3	7	Wise	GWS	11-MAY-63	M	1991	Edgecombe Way	39	9758VB	Stratford	070-347689	(null)
4	8	Newcastle	B	08-JUL-62	F	1990	Station Road	4	6584WO	Inglewood	070-458458	2983
5	27	Collins	DD	28-DEC-64	F	1993	Long Drive	804	8457DK	Eltham	079-234857	2513
6	28	Collins	C	22-JUN-63	F	1993	Old Main Road	10	1294QK	Midhurst	010-659599	(null)
7	39	Bishop	D	29-OCT-56	M	1990	Eaton Square	78	9629CD	Stratford	070-393435	(null)
8	44	Baker	E	09-JAN-63	M	1990	Lewis Street	23	4444LJ	Inglewood	070-368753	1124
9	57	Brown	M	17-AUG-71	M	1995	Edgecombe Way	16	4377CB	Stratford	070-473458	6409
10	83	Hope	PK	11-NOV-56	M	1992	Magdalene Road	16A	1812UP	Stratford	070-353548	1608
11	95	Miller	P	14-MAY-63	M	1972	High Street	33A	5746OP	Douglas	070-867564	(null)
12	100	Parmenter	P	28-FEB-63	M	1979	Haseltine Lane	80	6494SG	Stratford	070-494593	6524
13	104	Moorman	D	10-MAY-70	F	1994	Stout Street	65	9437AO	Eltham	079-987571	7060
14	112	Bailey	IP	01-OCT-63	F	1994	Vixen Road	8	6392LK	Plymouth	010-548745	1319

In the Street and Houseno column we can gather those columns to 1 columns with the statement below. The | | function can be used when we want to gather some column and the ‘ ‘ function can be used to gives us a space between the gathered columns.

```
select PLAYERNO, NAME, INITIALS, BIRTH_DATE, SEX, JOINED, STREET || ' ' || HOUSENO AS ADDRESS, PHONENO, LEAGUENO
from PLAYERS;
```

	PLAYERNO	NAME	INITIALS	BIRTH_DATE	SEX	JOINED	ADDRESS	PHONENO	LEAGUENO
1	2	Everett	R	01-SEP-48	M	1975	Stoney Road 43	070-237893	2411
2	6	Parmenter	R	25-JUN-64	M	1977	Haseltine Lane 80	070-476537	8467
3	7	Wise	GWS	11-MAY-63	M	1991	Edgecombe Way 39	070-347689	(null)
4	8	Newcastle	B	08-JUL-62	F	1990	Station Road 4	070-458458	2983
5	27	Collins	DD	28-DEC-64	F	1993	Long Drive 804	079-234857	2513
6	28	Collins	C	22-JUN-63	F	1993	Old Main Road 10	010-659599	(null)
7	39	Bishop	D	29-OCT-56	M	1990	Eaton Square 78	070-393435	(null)
8	44	Baker	E	09-JAN-63	M	1990	Lewis Street 23	070-368753	1124
9	57	Brown	M	17-AUG-71	M	1995	Edgecombe Way 16	070-473458	6409
10	83	Hope	PK	11-NOV-56	M	1992	Magdalene Road 16A	070-353548	1608
11	95	Miller	P	14-MAY-63	M	1972	High Street 33A	070-867564	(null)
12	100	Parmenter	P	28-FEB-63	M	1979	Haseltine Lane 80	070-494593	6524
13	104	Moorman	D	10-MAY-70	F	1994	Stout Street 65	079-987571	7060
14	112	Bailey	IP	01-OCT-63	F	1994	Vixen Road 8	010-548745	1319