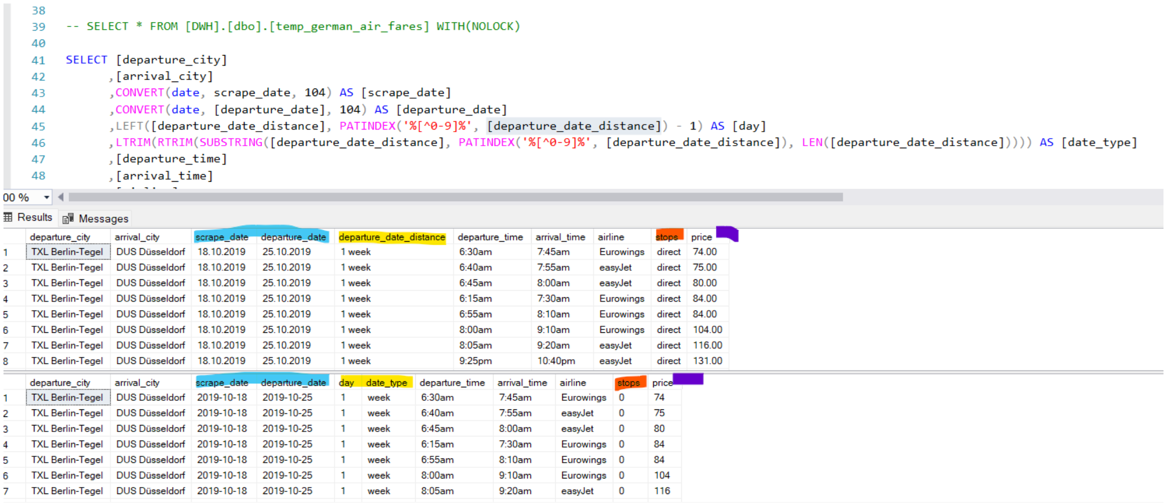
German air fares

Alejandro Meyer Contreras

The first step is to convert to date the fields [scrape_date] and [departure_date], for the field [departure_date_distance] two new fields are created to separate the number and the comments, in the same way that the field [stops] is converted to numbers by removing all the text, finally, we convert [price] to integers.

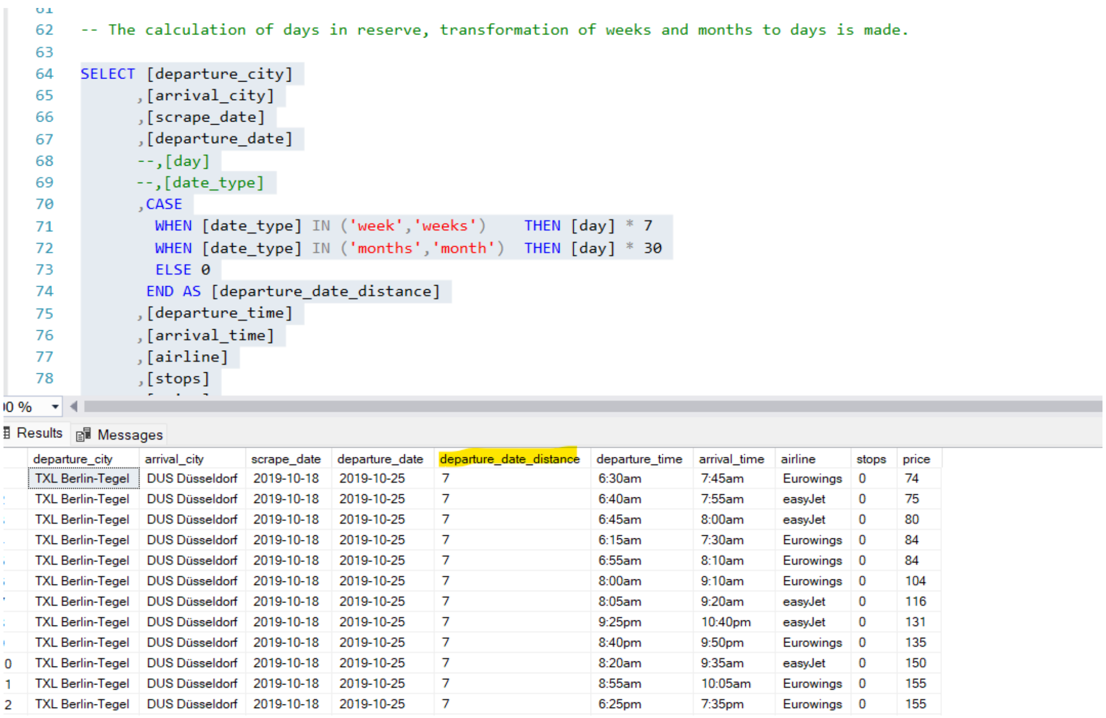


```
-- SELECT * FROM [DWH].[dbo].[temp_german_air_fares] WITH(NOLOCK)

SELECT [departure_city]
      ,[arrival_city]
      ,CONVERT(date, scrape_date, 104) AS [scrape_date]
      ,CONVERT(date, [departure_date], 104) AS [departure_date]
      ,LEFT([departure_date_distance], PATINDEX('%[^0-9]%', [departure_date_distance]) - 1) AS [day]
      ,LTRIM(RTRIM(SUBSTRING([departure_date_distance], PATINDEX('%[^0-9]%', [departure_date_distance]), LEN([departure_date_distance])))) AS [date_type]
      ,[departure_time]
      ,[arrival_time]
```

	departure_city	arrival_city	scrape_date	departure_date	departure_date_distance	departure_time	arrival_time	airline	stops	price
1	TXL Berlin-Tegel	DUS Düsseldorf	18.10.2019	25.10.2019	1 week	6:30am	7:45am	Eurowings	direct	74.00
2	TXL Berlin-Tegel	DUS Düsseldorf	18.10.2019	25.10.2019	1 week	6:40am	7:55am	easyJet	direct	75.00
3	TXL Berlin-Tegel	DUS Düsseldorf	18.10.2019	25.10.2019	1 week	6:45am	8:00am	easyJet	direct	80.00
4	TXL Berlin-Tegel	DUS Düsseldorf	18.10.2019	25.10.2019	1 week	6:15am	7:30am	Eurowings	direct	84.00
5	TXL Berlin-Tegel	DUS Düsseldorf	18.10.2019	25.10.2019	1 week	6:55am	8:10am	Eurowings	direct	84.00
6	TXL Berlin-Tegel	DUS Düsseldorf	18.10.2019	25.10.2019	1 week	8:00am	9:10am	Eurowings	direct	104.00
7	TXL Berlin-Tegel	DUS Düsseldorf	18.10.2019	25.10.2019	1 week	8:05am	9:20am	easyJet	direct	116.00
8	TXL Berlin-Tegel	DUS Düsseldorf	18.10.2019	25.10.2019	1 week	9:25pm	10:40pm	easyJet	direct	131.00

	departure_city	arrival_city	scrape_date	departure_date	day	date_type	departure_time	arrival_time	airline	stops	price
1	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	1	week	6:30am	7:45am	Eurowings	0	74
2	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	1	week	6:40am	7:55am	easyJet	0	75
3	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	1	week	6:45am	8:00am	easyJet	0	80
4	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	1	week	6:15am	7:30am	Eurowings	0	84
5	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	1	week	6:55am	8:10am	Eurowings	0	84
6	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	1	week	8:00am	9:10am	Eurowings	0	104
7	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	1	week	8:05am	9:20am	easyJet	0	116

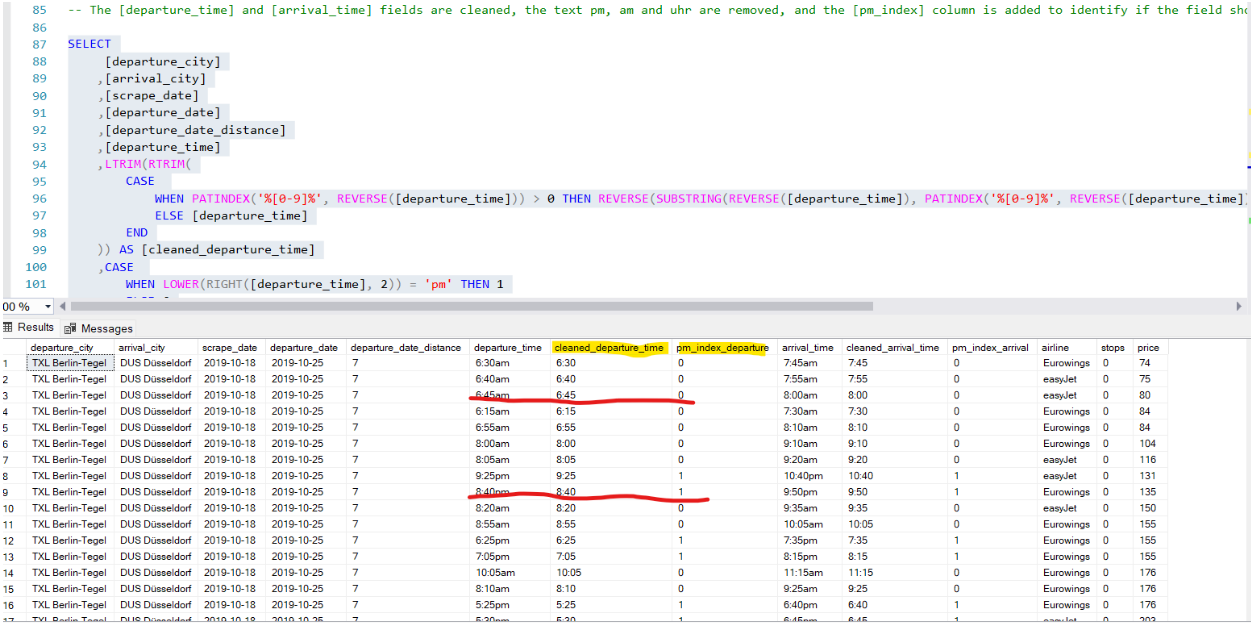
Now we use the new fields [day] and [date_type] to calculate the days in advance of the ticket purchase.



```
-- The calculation of days in reserve, transformation of weeks and months to days is made.

SELECT [departure_city]
      ,[arrival_city]
      ,[scrape_date]
      ,[departure_date]
      --,[day]
      --,[date_type]
      ,CASE
        WHEN [date_type] IN ('week','weeks')    THEN [day] * 7
        WHEN [date_type] IN ('months','month')  THEN [day] * 30
        ELSE 0
       END AS [departure_date_distance]
      ,[departure_time]
      ,[arrival_time]
      ,[airline]
      ,[stops]
```

	departure_city	arrival_city	scrape_date	departure_date	departure_date_distance	departure_time	arrival_time	airline	stops	price
	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	6:30am	7:45am	Eurowings	0	74
	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	6:40am	7:55am	easyJet	0	75
	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	6:45am	8:00am	easyJet	0	80
	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	6:15am	7:30am	Eurowings	0	84
	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	6:55am	8:10am	Eurowings	0	84
	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	8:00am	9:10am	Eurowings	0	104
	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	8:05am	9:20am	easyJet	0	116
	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	9:25pm	10:40pm	easyJet	0	131
	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	8:40pm	9:50pm	Eurowings	0	135
0	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	8:20am	9:35am	easyJet	0	150
1	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	8:55am	10:05am	Eurowings	0	155
2	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	6:25pm	7:35pm	Eurowings	0	155

With the following code we create the new field [cleaned_departure_time] where we delete “am, pm, uhr” and also create the field [pm_index_departure] to identify the fields where we must add 12 hours to have everything in 24 hours format.



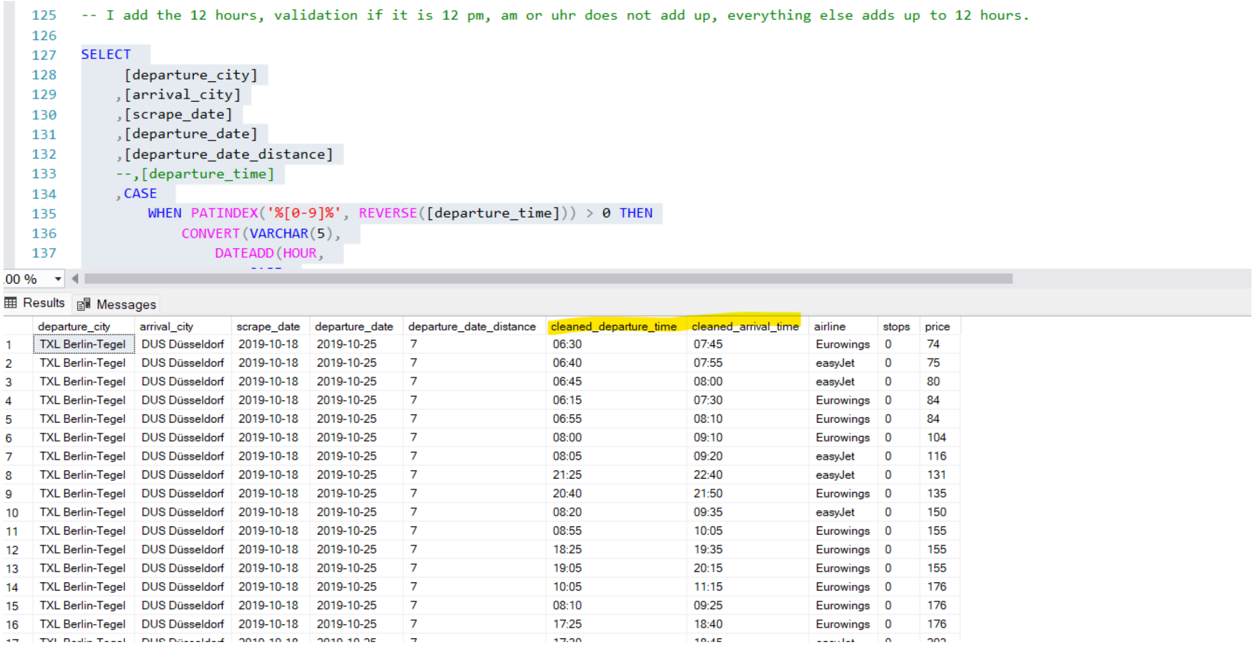
```
-- The [departure_time] and [arrival_time] fields are cleaned, the text pm, am and uhr are removed, and the [pm_index] column is added to identify if the field sh

SELECT
     [departure_city]
    ,[arrival_city]
    ,[scrape_date]
    ,[departure_date]
    ,[departure_date_distance]
    ,[departure_time]
    ,LTRIM(RTRIM(
        CASE
            WHEN PATINDEX('%[0-9]%', REVERSE([departure_time])) > 0 THEN REVERSE(SUBSTRING(REVERSE([departure_time]), PATINDEX('%[0-9]%', REVERSE([departure_time])
            ELSE [departure_time]
        END
    )) AS [cleaned_departure_time]
    ,CASE
        WHEN LOWER(RIGHT([departure_time], 2)) = 'pm' THEN 1
```

00 %

Results | Messages

	departure_city	arrival_city	scrape_date	departure_date	departure_date_distance	departure_time	cleaned_departure_time	pm_index_departure	arrival_time	cleaned_arrival_time	pm_index_arrival	airline	stops	price
1	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	6:30am	6:30	0	7:45am	7:45	0	Eurowings	0	74
2	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	6:40am	6:40	0	7:55am	7:55	0	easyJet	0	75
3	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	6:45am	6:45	0	8:00am	8:00	0	easyJet	0	80
4	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	6:15am	6:15	0	7:30am	7:30	0	Eurowings	0	84
5	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	6:55am	6:55	0	8:10am	8:10	0	Eurowings	0	84
6	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	8:00am	8:00	0	9:10am	9:10	0	Eurowings	0	104
7	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	8:05am	8:05	0	9:20am	9:20	0	easyJet	0	116
8	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	9:25pm	9:25	1	10:40pm	10:40	1	easyJet	0	131
9	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	8:40pm	8:40	1	9:50pm	9:50	1	Eurowings	0	135
10	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	8:20am	8:20	0	9:35am	9:35	0	easyJet	0	150
11	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	8:55am	8:55	0	10:05am	10:05	0	Eurowings	0	155
12	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	6:25pm	6:25	1	7:35pm	7:35	1	Eurowings	0	155
13	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	7:05pm	7:05	1	8:15pm	8:15	1	Eurowings	0	155
14	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	10:05am	10:05	0	11:15am	11:15	0	Eurowings	0	176
15	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	8:10am	8:10	0	9:25am	9:25	0	Eurowings	0	176
16	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	5:25pm	5:25	1	6:40pm	6:40	1	Eurowings	0	176

Then we do the sum of 12 hours using our new fields and we will have the standardized fields.



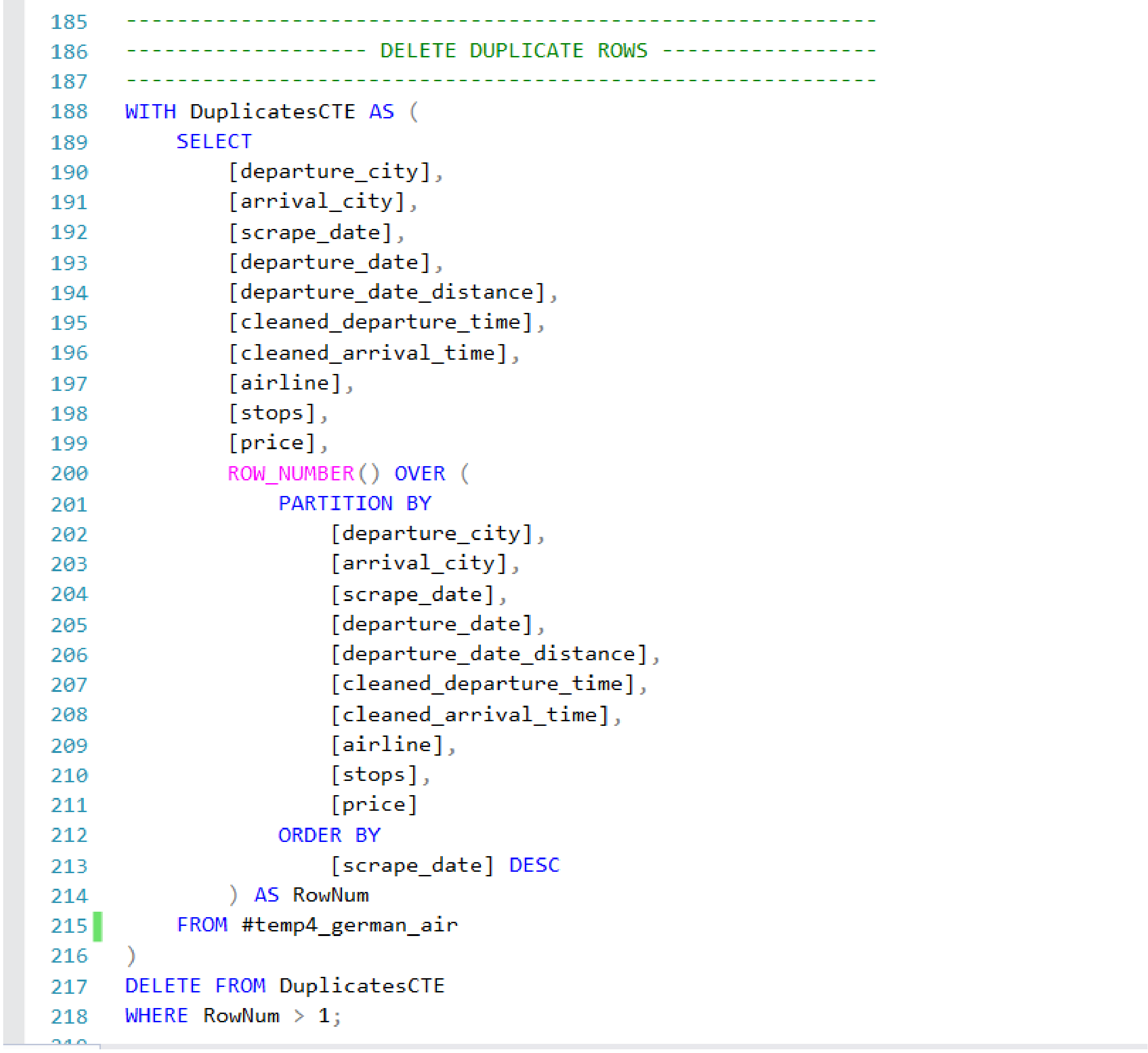
```
-- I add the 12 hours, validation if it is 12 pm, am or uhr does not add up, everything else adds up to 12 hours.

SELECT
     [departure_city]
    ,[arrival_city]
    ,[scrape_date]
    ,[departure_date]
    ,[departure_date_distance]
    --,[departure_time]
    ,CASE
        WHEN PATINDEX('%[0-9]%', REVERSE([departure_time])) > 0 THEN
            CONVERT(VARCHAR(5),
                DATEADD(HOUR,
```

00 %

Results | Messages

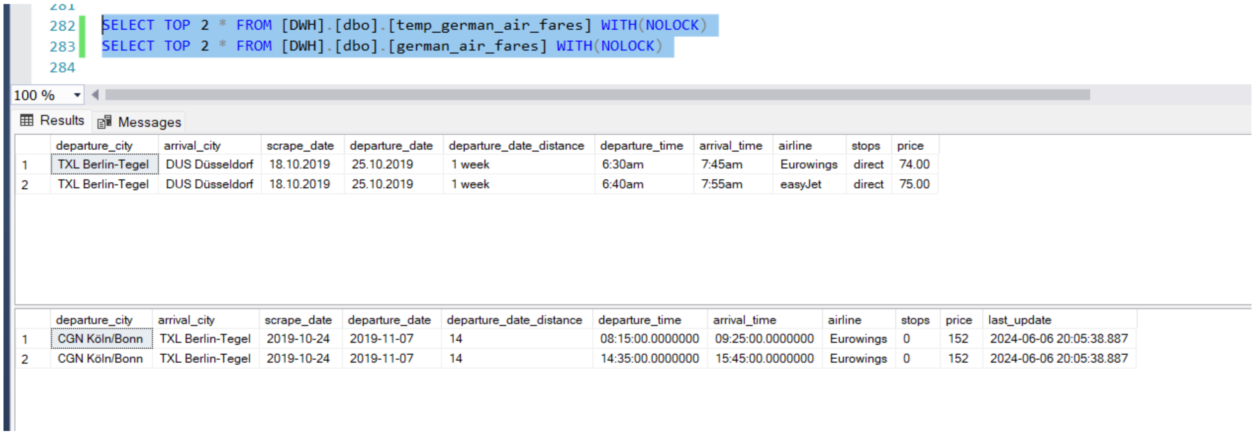
	departure_city	arrival_city	scrape_date	departure_date	departure_date_distance	cleaned_departure_time	cleaned_arrival_time	airline	stops	price
1	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	06:30	07:45	Eurowings	0	74
2	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	06:40	07:55	easyJet	0	75
3	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	06:45	08:00	easyJet	0	80
4	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	06:15	07:30	Eurowings	0	84
5	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	06:55	08:10	Eurowings	0	84
6	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	08:00	09:10	Eurowings	0	104
7	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	08:05	09:20	easyJet	0	116
8	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	21:25	22:40	easyJet	0	131
9	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	20:40	21:50	Eurowings	0	135
10	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	08:20	09:35	easyJet	0	150
11	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	08:55	10:05	Eurowings	0	155
12	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	18:25	19:35	Eurowings	0	155
13	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	19:05	20:15	Eurowings	0	155
14	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	10:05	11:15	Eurowings	0	176
15	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	08:10	09:25	Eurowings	0	176
16	TXL Berlin-Tegel	DUS Düsseldorf	2019-10-18	2019-10-25	7	17:25	18:40	Eurowings	0	176

We eliminate duplicate data since we do not have an ID for each flight.



```sql
-----------------------------------------------------------
------------------- DELETE DUPLICATE ROWS -----------------
-----------------------------------------------------------
WITH DuplicatesCTE AS (
    SELECT
        [departure_city],
        [arrival_city],
        [scrape_date],
        [departure_date],
        [departure_date_distance],
        [cleaned_departure_time],
        [cleaned_arrival_time],
        [airline],
        [stops],
        [price],
        ROW_NUMBER() OVER (
            PARTITION BY
                [departure_city],
                [arrival_city],
                [scrape_date],
                [departure_date],
                [departure_date_distance],
                [cleaned_departure_time],
                [cleaned_arrival_time],
                [airline],
                [stops],
                [price]
            ORDER BY
                [scrape_date] DESC
        ) AS RowNum
    FROM #temp4_german_air
)
DELETE FROM DuplicatesCTE
WHERE RowNum > 1;
```

Finally, we merge the production table and compare the original data with the final data.



```sql
SELECT TOP 2 * FROM [DWH].[dbo].[temp_german_air_fares] WITH(NOLOCK)
SELECT TOP 2 * FROM [DWH].[dbo].[german_air_fares] WITH(NOLOCK)
```

	departure_city	arrival_city	scrape_date	departure_date	departure_date_distance	departure_time	arrival_time	airline	stops	price
1	TXL Berlin-Tegel	DUS Düsseldorf	18.10.2019	25.10.2019	1 week	6:30am	7:45am	Eurowings	direct	74.00
2	TXL Berlin-Tegel	DUS Düsseldorf	18.10.2019	25.10.2019	1 week	6:40am	7:55am	easyJet	direct	75.00

	departure_city	arrival_city	scrape_date	departure_date	departure_date_distance	departure_time	arrival_time	airline	stops	price	last_update
1	CGN Köln/Bonn	TXL Berlin-Tegel	2019-10-24	2019-11-07	14	08:15:00.0000000	09:25:00.0000000	Eurowings	0	152	2024-06-06 20:05:38.887
2	CGN Köln/Bonn	TXL Berlin-Tegel	2019-10-24	2019-11-07	14	14:35:00.0000000	15:45:00.0000000	Eurowings	0	152	2024-06-06 20:05:38.887