Absenteeism at Work Data Modeling, Tranformation, Cleaning and Creation

Tracking employee absences helps businesses understand how much time is lost and why. Since absences can cost millions each year, it's crucial to monitor them and find ways to bring them down. This reduces disruption and saves the company money.

All information used in this notebook is based on source data:

> Martiniano, Andrea and Ferreira, Ricardo. (2018). Absenteeism at work. UCI Machine Learning repository. https://doi.org/10.24432/C5X882.

In [108…

```
-- Note: This database was created manually also the csv file
-- was imported to the database in the same way.

-- Using the database
USE Absenteeism_at_work
```

Commands completed successfully.
Total execution time: 00:00:00.001

In [109…

```
-- Check all the columns from the new imported csv table
SELECT
    TABLE_NAME,
    COLUMN_NAME,
    ORDINAL_POSITION
FROM INFORMATION_SCHEMA.COLUMNS WHERE TABLE_NAME = 'Absenteeism'
ORDER BY ORDINAL_POSITION ASC
```

(21 rows affected)
Total execution time: 00:00:00.097

Out[109]:

TABLE_NAME	COLUMN_NAME	ORDINAL_POSITION
Absenteeism	ID	1
Absenteeism	Reason_for_absence	2
Absenteeism	Month_of_absence	3
Absenteeism	Day_of_the_week	4
Absenteeism	Seasons	5
Absenteeism	Transportation_expense	6
Absenteeism	Distance_from_Residence_to_Work	7
Absenteeism	Service_time	8
Absenteeism	Age	9
Absenteeism	Work_load_Average_day	10
Absenteeism	Hit_target	11
Absenteeism	Disciplinary_failure	12
Absenteeism	Education	13
Absenteeism	Son	14
Absenteeism	Social_drinker	15

Absenteeism	Social_smoker	16
Absenteeism	Pet	17
Absenteeism	Weight	18
Absenteeism	Height	19
Absenteeism	Body_mass_index	20
Absenteeism	Absenteeism_time_in_hours	21

In [110…

```
-- Verifying data imported from csv file
SELECT TOP 10 * FROM Absenteeism
```

(10 rows affected)

Total execution time: 00:00:00.013

Out[110]:

ID	Reason_for_absence	Month_of_absence	Day_of_the_week	Seasons	Transportation_expense	Distance_from_Resi
11	26	7	3	1	289	
36	0	7	3	1	118	
3	23	7	4	1	179	
7	7	7	5	1	279	
11	23	7	5	1	289	
3	23	7	6	1	179	
10	22	7	6	1	361	
20	23	7	6	1	260	
14	19	7	2	1	155	
1	22	7	2	1	235	

Based on data modeling design, transforming, cleaning and creating the necessary tables.

In [111…

```
--> Creating each new table based on the original data: Absenteeism

-- Creating: Absenteeism_AWK
IF OBJECT_ID(N'dbo.Absenteeism_AWK', N'U') IS NOT NULL
    DROP TABLE dbo.Absenteeism_AWK;
SELECT
    id AS IdEmployee,
    Reason_for_absence AS IdAbsence,
    Transportation_expense,
    Service_time,
    Work_load_Average_day AS Work_load_avg_day,
    Hit_target,
    Absenteeism_time_in_hours,
    Disciplinary_failure,
        Day_of_the_week,
        CASE WHEN Day_of_the_week = 1 THEN 'Monday'
             WHEN Day_of_the_week = 2 THEN 'Tuesday'
                 WHEN Day_of_the_week = 3 THEN 'Wednesday'
                 WHEN Day_of_the_week = 4 THEN 'Thursday'
                 WHEN Day_of_the_week = 5 THEN 'Friday'
                 WHEN Day_of_the_week = 6 THEN 'Saturday'
                 WHEN Day_of_the_week = 7 THEN 'Sunday'
    END AS Day_name,
        Month_of_absence,
```

```
        CASE WHEN Month_of_absence = 1 THEN 'January'
             WHEN Month_of_absence = 2 THEN 'February'
                 WHEN Month_of_absence = 3 THEN 'March'
                 WHEN Month_of_absence = 4 THEN 'April'
                 WHEN Month_of_absence = 5 THEN 'May'
                 WHEN Month_of_absence = 6 THEN 'June'
                 WHEN Month_of_absence = 7 THEN 'July'
                 WHEN Month_of_absence = 8 THEN 'August'
                 WHEN Month_of_absence = 9 THEN 'September'
                 WHEN Month_of_absence = 10 THEN 'October'
                 WHEN Month_of_absence = 11 THEN 'November'
          -- month with 0 also setting as december
                 WHEN Month_of_absence IN(12,0) THEN 'December'
  END AS Month_name,
  Seasons,
  CASE WHEN Seasons = 1 THEN 'Summer'
       WHEN Seasons = 2 THEN 'Fall'
           WHEN Seasons = 3 THEN 'Winter'
           WHEN Seasons = 4 THEN 'Spring'
   END AS Season_name
INTO Absenteeism_AWK
FROM Absenteeism;

-- UPDATE Month_of_absence WITH 0 BY 12 (December)
UPDATE Absenteeism_AWK SET Month_of_absence = 12 WHERE Month_of_absence = 0


-- Because the dataset is based on a Brazilian company, the year stations are different
-- from North America, so I will update this to have the stations based on
-- Canada where I live.

UPDATE [Absenteeism_AWK]
SET SEASONS = 1, SEASON_NAME = 'Summer'
WHERE Month_of_absence IN (6,7,8)

UPDATE [Absenteeism_AWK]
SET SEASONS = 2, SEASON_NAME = 'Fall'
WHERE Month_of_absence IN (9,10,11)

UPDATE [Absenteeism_AWK]
SET SEASONS = 3, SEASON_NAME = 'Winter'
WHERE Month_of_absence IN (12,1,2)

UPDATE [Absenteeism_AWK]
SET SEASONS = 4, SEASON_NAME = 'Spring'
WHERE Month_of_absence IN (3,4,5)


-- Checking some sample data from Fact_Absenteeism
SELECT TOP 5 * FROM Absenteeism_AWK
```

(740 rows affected)

(3 rows affected)

(175 rows affected)

(187 rows affected)

(174 rows affected)

(204 rows affected)

(5 rows affected)

Total execution time: 00:00:00.060

Out[111]:

IdEmployee	IdAbsence	Transportation_expense	Service_time	Work_load_avg_day	Hit_target	Absenteeism_time_i
11	26	289	13	239.5540008544922	97	

36	0	118	18	239.5540008544922	97
3	23	179	18	239.5540008544922	97
7	7	279	14	239.5540008544922	97
11	23	289	13	239.5540008544922	97

In [112…

```
-- Checking the distinct Absence reason id
SELECT DISTINCT Reason_for_absence
FROM Absenteeism
ORDER BY 1
```

(28 rows affected)

Total execution time: 00:00:00.021

Out[112]:

Reason_for_absence
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
21
22
23
24
25
26
27
28

In [113…
```sql
-- We found an absence with ID zero so I will replace it with "unexcused absence" to avo
-- those employee records with absence ID = 0.
UPDATE Absenteeism SET Reason_for_absence = 26 WHERE Reason_for_absence = 0
```

(43 rows affected)

Total execution time: 00:00:00.008

In [114…
```sql
-- Creating: Reason_Absense
IF OBJECT_ID(N'dbo.Reason_Absense', N'U') IS NOT NULL
    DROP TABLE dbo.Reason_Absense;
SELECT DISTINCT
   Reason_for_absence AS IdAbsence,
    CASE WHEN Reason_for_absence = 1 THEN 'Certain infectious and parasitic diseases'
        WHEN Reason_for_absence = 2 THEN 'Neoplasms'
        WHEN Reason_for_absence = 3 THEN 'Diseases of the blood and blood-forming organs
        WHEN Reason_for_absence = 4 THEN 'Endocrine, nutritional and metabolic diseases'
        WHEN Reason_for_absence = 5 THEN 'Mental and behavioural disorders'
        WHEN Reason_for_absence = 6 THEN 'Diseases of the nervous system'
        WHEN Reason_for_absence = 7 THEN 'Diseases of the eye and adnexa'
        WHEN Reason_for_absence = 8 THEN 'Diseases of the ear and mastoid process'
        WHEN Reason_for_absence = 9 THEN 'Diseases of the circulatory system'
        WHEN Reason_for_absence = 10 THEN 'Diseases of the respiratory system'
        WHEN Reason_for_absence = 11 THEN 'Diseases of the digestive system'
        WHEN Reason_for_absence = 12 THEN 'Diseases of the skin and subcutaneous tissue'
        WHEN Reason_for_absence = 13 THEN 'Diseases of the musculoskeletal system and co
        WHEN Reason_for_absence = 14 THEN 'Diseases of the genitourinary system'
        WHEN Reason_for_absence = 15 THEN 'Pregnancy, childbirth and the puerperium'
        WHEN Reason_for_absence = 16 THEN 'Certain conditions originating in the perinat
        WHEN Reason_for_absence = 17 THEN 'Congenital malformations, deformations and ch
        WHEN Reason_for_absence = 18 THEN 'Symptoms, signs and abnormal clinical and lab
        WHEN Reason_for_absence = 19 THEN 'Injury, poisoning and certain other consequen
        WHEN Reason_for_absence = 20 THEN 'External causes of morbidity and mortality'
        WHEN Reason_for_absence = 21 THEN 'Factors influencing health status and contact
        WHEN Reason_for_absence = 22 THEN 'Patient follow-up'
        WHEN Reason_for_absence = 23 THEN 'Medical consultation'
        WHEN Reason_for_absence = 24 THEN 'Blood donation'
        WHEN Reason_for_absence = 25 THEN 'Laboratory examination'
        WHEN Reason_for_absence = 26 THEN 'Unjustified absence'
        WHEN Reason_for_absence = 27 THEN 'Physiotherapy'
        WHEN Reason_for_absence = 28 THEN 'Dental consultation'
    END AS Absence_description
INTO Reason_Absense
FROM Absenteeism;

-- Checking some sample data from Reason_Absense
SELECT TOP 5 * FROM Reason_Absense
```

(27 rows affected)

(5 rows affected)

Total execution time: 00:00:00.025

Out[114]:

IdAbsence	Absence_description
1	Certain infectious and parasitic diseases
2	Neoplasms
3	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
4	Endocrine, nutritional and metabolic diseases
5	Mental and behavioural disorders

In [115…
```sql
-- > I have checked the original Absenteeism table and it does not contain any absence
```

```
-- records for category # 20 "External causes of morbidity and mortality".
-- I will proceed to insert inside the previously created table: "Dim_Reason_Absense"
-- for future use only.

IF EXISTS(SELECT IdAbsence FROM Reason_Absense WHERE IdAbsence = 20)
BEGIN
 PRINT 'Id reason #20 already exists'
END ELSE
PRINT 'Inserting the id reason #20'
 DELETE FROM Reason_Absense WHERE IdAbsence = 20
 INSERT INTO Reason_Absense VALUES (20, 'External causes of morbidity and mortality')

-- Checking the new inserted record
SELECT * FROM Reason_Absense WHERE IdAbsence = 20
```

Inserting the id reason #20

(0 rows affected)

(1 row affected)

(1 row affected)

Total execution time: 00:00:00.026

Out[115]:

IdAbsence	Absence_description
20	External causes of morbidity and mortality

In [116…

```
-- Creating: Employee_Demographic
IF OBJECT_ID(N'dbo.Employee_Demographic', N'U') IS NOT NULL
    DROP TABLE dbo.Employee_Demographic;
SELECT DISTINCT
        ID AS IdEmployee,
        Age,
        [Weight],
        Height,
        Body_mass_index,
        CASE WHEN Education = 1 THEN 'High school'
             WHEN Education = 2 THEN 'Graduate'
                 WHEN Education = 3 THEN 'Postgraduate'
                 WHEN Education = 4 THEN 'Master and Doctor'
        END AS Education_level,
        Son AS Children_number,
        CASE WHEN Social_drinker = 1 THEN 'Yes' ELSE 'No' END AS Social_drinker,
        CASE WHEN Social_smoker = 1 THEN 'Yes' ELSE 'No' END AS Social_smoker,
        Pet AS Pets_number,
        Distance_from_Residence_to_Work AS Home_work_distance_km
INTO Employee_Demographic
FROM Absenteeism;

-- Updating the id of employee #29 because by mistake it is repeated
-- with different level of education.
UPDATE Employee_Demographic SET IdEmployee = 37 WHERE IdEmployee = 29 AND Education_leve

-- Checking some sample data from Employee_Demographic
SELECT top 5 * FROM Employee_Demographic ORDER BY 1
```

(37 rows affected)

(1 row affected)

(5 rows affected)

Total execution time: 00:00:00.071

Out[116]:

IdEmployee	Age	Weight	Height	Body_mass_index	Education_level	Children_number	Social_drinker	Social_smo
1	37	88	172	29	Postgraduate	1	No	
2	48	88	163	33	High school	1	No	

3	38	89	170	31	High school	0	Yes
4	40	98	170	34	High school	1	Yes
5	43	106	167	38	High school	1	Yes

Final step, export each table created as a CSV file to be used in Power BI to build the necessary data modeling and graphics.

Absenteeism_AWK.csv

Reason_Absense.csv

Employee_Demographic.csv