Notebook to analyze the internal air temperature values regarding the ambient temperature

In [6]:

```python
# Python standard library imports
import time

# Third-party imports for database connection and data manipulation
from sqlalchemy import create_engine
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import pearsonr
# Third-party imports for mapping
import folium
```

In [7]:

```python
# Database connection parameters
dbname = 'ar41'
user = 'postgres'
password = '1234'
host = 'localhost'  # localhost or the server address
port = '5432'  # default PostgreSQL port is 5432

# Establish a connection to the database
connection_str = f"postgresql://{user}:{password}@{host}:{port}/{dbname}"
engine = create_engine(connection_str)
```

1. Choosing sensors to analyze

- Out of all the internal sensors, both the Internal Air Temperature PC1 and PC2 are the most correlated to the ambient temperature obtained during the enrichment phase of the project.

1. Logical thinking and procedure:

A. Null values or values above the acceptable boundaries (65°C) are not considered for this analysis
B. We first perform a frequency analysis
C. Using cumulative frequency analysis of the absolute difference between the sensors and the temperature outside of the train we choose the outlier

1. Bucket analysis

In [8]:

```python
query_temperature_differences = f"""
WITH TemperatureDifferences AS (
    SELECT
        mapped_veh_id,
        "timestamps_UTC",
        "Temperature",
        "RS_E_InAirTemp_PC1",
        ABS("Temperature" - "RS_E_InAirTemp_PC1") AS temp_difference
    FROM
        vehicle_data_enriched
    WHERE
```

```
        "Temperature" IS NOT NULL
        AND "RS_E_InAirTemp_PC1" < 70
)
SELECT
    bucket_range,
    COUNT(*) AS occurrences
FROM (
    SELECT
        mapped_veh_id,
        "timestamps_UTC",
        "Temperature",
        "RS_E_InAirTemp_PC1",
        temp_difference,
        floor(temp_difference / 5) * 5 AS bucket_range
    FROM
        TemperatureDifferences
) AS temp_diff_buckets
GROUP BY
    bucket_range
ORDER BY
    bucket_range;
"""
```

In [9]:

```
df_temperature_differences = pd.read_sql_query(query_temperature_differences
print(df_temperature_differences)
```

```
    bucket_range  occurrences
0            0.0      1240241
1            5.0      2584300
2           10.0      3144885
3           15.0      3789691
4           20.0      3177038
5           25.0      1862560
6           30.0      1024313
7           35.0       492889
8           40.0       203760
9           45.0        87887
10          50.0        40473
11          55.0        13404
12          60.0         3127
13          65.0          409
```

After here, there are two paths:

- Arbitrarily choosing a threshold or value to consider anomaly (field knowledge)
- Using cumulative frequency analysis to consider only values above a cumulative percentage (e.g. 99%)

In [24]:

```
# Calculate the cumulative sum of frequencies
df_temperature_differences['Cumulative_Frequency'] = df_temperature_differen

# Calculate the total number of occurrences
total_occurrences = df_temperature_differences['occurrences'].sum()

# Set a threshold percentage
threshold_percentage = 99.95

# Find the bucket where the cumulative frequency exceeds the threshold perce
outlier_bucket = df_temperature_differences[df_temperature_differences['Cumu

print(f"The outlier bucket is {outlier_bucket}")
```

```
The outlier bucket is 55.0
```

With this analysis, we can say that any timestamp with a temperature difference above 55 shouldn't be normal.

And we can proceed to create our final queries to detect anomalies. For PC1 sensor:

In [25]:

```
# Analysis of PC1 in air temperature vs outside temperature
query_temperature_differences_PC1 = f"""
SELECT mapped_veh_id, "timestamps_UTC", "Temperature", "RS_E_InAirTemp_PC1",
FROM vehicle_data_enriched
WHERE "Temperature" is not null AND ABS("Temperature"-"RS_E_InAirTemp_PC1")
GROUP BY mapped_veh_id, "timestamps_UTC", "Temperature", "RS_E_InAirTemp_PC
ORDER BY ABS("Temperature"-"RS_E_InAirTemp_PC1") DESC;
"""
```

In [26]:

```
df_temperature_differences_PC1 = pd.read_sql_query(query_temperature_differe
print(df_temperature_differences_PC1)
df_temperature_differences_PC1.to_csv('OutsideDiff1.csv', sep=',', index=Fal
```

```
       mapped_veh_id       timestamps_UTC  Temperature  RS_E_InAirTemp_PC1
\
0              154.0  2023-02-07 22:50:37         -2.7                65.0
1              154.0  2023-02-07 22:51:37         -2.7                65.0
2              154.0  2023-02-07 22:51:57         -2.7                65.0
3              154.0  2023-02-07 22:52:37         -2.7                65.0
4              154.0  2023-02-07 22:53:37         -2.7                65.0
...              ...                  ...          ...                 ...
13536          197.0  2023-02-08 12:40:26          5.0                60.0
13537          197.0  2023-02-08 12:41:18          5.0                60.0
13538          197.0  2023-02-08 12:41:25          5.0                60.0
13539          197.0  2023-03-01 21:52:23         -1.0                54.0
13540          197.0  2023-03-01 21:52:27         -1.0                54.0

        abs
0      67.7
1      67.7
2      67.7
3      67.7
4      67.7
...     ...
13536  55.0
13537  55.0
13538  55.0
13539  55.0
13540  55.0

[13541 rows x 5 columns]
```

In [28]:

```
# Analysis of PC1 in air temperature vs outside temperature (dashboard outpu
query_temperature_differences_PC1_2 = f"""
SELECT *, ABS("Temperature"-"RS_E_InAirTemp_PC1")
FROM vehicle_data_enriched
WHERE "Temperature" is not null AND ABS("Temperature"-"RS_E_InAirTemp_PC1")
ORDER BY ABS("Temperature"-"RS_E_InAirTemp_PC1") DESC;
"""

df_temperature_differences_PC1_2 = pd.read_sql_query(query_temperature_diffe
df_temperature_differences_PC1_2['outlier_type'] = 'Outisde temperature'
print(df_temperature_differences_PC1_2)
df_temperature_differences_PC1_2.to_csv('R9-1.csv', sep=',', index=False, en
       mapped_veh_id      timestamps_UTC        lat       lon  \
0              154.0  2023-02-07 22:53:37  51.015433  3.775966   
1              154.0  2023-02-07 22:52:37  51.015444  3.775948   
2              154.0  2023-02-07 22:50:37  51.015397  3.775909   
3              154.0  2023-02-07 22:54:40  51.015426  3.776021   
4              154.0  2023-02-07 22:51:37  51.015409  3.775935   
...              ...                  ...        ...       ...   
13536          157.0  2023-04-03 09:58:28  51.015223  3.776540   
13537          171.0  2023-02-01 00:21:38  50.400397  4.459595   
13538          153.0  2023-02-13 07:29:37  51.014134  3.778928   
13539          155.0  2023-02-06 22:31:02  51.016140  3.774164   
13540          168.0  2023-04-14 21:42:33  51.015991  3.774759   

       RS_E_InAirTemp_PC1  RS_E_InAirTemp_PC2  RS_E_OilPress_PC1  \
0                    65.0                52.0                0.0   
1                    65.0                49.0                0.0   
2                    65.0                26.0                0.0   
3                    65.0                55.0                0.0   
4                    65.0                 0.0                0.0   
...                   ...                 ...                ...   
13536                61.0                55.0                3.0   
13537                61.0                30.0                0.0   
13538                56.0                55.0                3.0   
13539                55.0                12.0                0.0   
13540                63.0                56.0                0.0   

       RS_E_OilPress_PC2  rs_e_rpm_pc1  rs_e_rpm_pc2  ...  \
0                    3.0           0.0           0.0  ...   
1                    3.0           0.0           0.0  ...   
2                  210.0           0.0         596.0  ...   
3                    3.0           0.0           0.0  ...   
4                    0.0           0.0           0.0  ...   
...                  ...           ...           ...  ...   
13536                6.0           0.0           0.0  ...   
13537              672.0           0.0         803.0  ...   
13538                3.0           0.0           0.0  ...   
13539              286.0           0.0         839.0  ...   
13540               10.0           0.0           0.0  ...   

          timestamps_floor  nearest_point_id        Lat       Lon  \
0      2023-02-07 22:00:00               208  51.015433  3.775966   
1      2023-02-07 22:00:00               208  51.015444  3.775948   
2      2023-02-07 22:00:00               208  51.015397  3.775909   
3      2023-02-07 22:00:00               208  51.015426  3.776021   
4      2023-02-07 22:00:00               208  51.015409  3.775935   
...                    ...               ...        ...       ...   
13536  2023-04-03 09:00:00               208  51.015223  3.776540   
13537  2023-02-01 00:00:00               449  50.400397  4.459595   
13538  2023-02-13 07:00:00               208  51.014134  3.778928   
13539  2023-02-06 22:00:00               208  51.016140  3.774164   
13540  2023-04-14 21:00:00               208  51.015991  3.774759   

                      Time  Temperature  Humidity  Rain   abs  \
0      2023-02-07 22:00:00         -2.7      93.0   0.0  67.7   
1      2023-02-07 22:00:00         -2.7      93.0   0.0  67.7   
2      2023-02-07 22:00:00         -2.7      93.0   0.0  67.7   
3      2023-02-07 22:00:00         -2.7      93.0   0.0  67.7   
4      2023-02-07 22:00:00         -2.7      93.0   0.0  67.7   
...                    ...          ...       ...   ...   ...   
13536  2023-04-03 09:00:00          6.0      74.0   0.0  55.0   
13537  2023-02-01 00:00:00          6.0      97.0   0.0  55.0   
13538  2023-02-13 07:00:00          1.0      97.0   0.0  55.0   
13539  2023-02-06 22:00:00          0.0      93.0   0.0  55.0   
13540  2023-04-14 21:00:00          8.0      79.0   0.0  55.0   
```

```
              outlier_type
0      Outisde temperature
1      Outisde temperature
2      Outisde temperature
3      Outisde temperature
4      Outisde temperature
...                    ...
13536  Outisde temperature
13537  Outisde temperature
13538  Outisde temperature
13539  Outisde temperature
13540  Outisde temperature

[13541 rows x 24 columns]
```

And for PC2 sensor:

```
In [29]: # Analysis of PC2 in air temperature vs outside temperature
query_temperature_differences_PC2 = f"""
SELECT mapped_veh_id, "timestamps_UTC", "Temperature", "RS_E_InAirTemp_PC2",
FROM vehicle_data_enriched
WHERE "Temperature" is not null AND ABS("Temperature"-"RS_E_InAirTemp_PC2")
GROUP BY mapped_veh_id, "timestamps_UTC", "Temperature", "RS_E_InAirTemp_PC
ORDER BY ABS("Temperature"-"RS_E_InAirTemp_PC2") DESC;
"""
```

```
In [30]: df_temperature_differences_PC2 = pd.read_sql_query(query_temperature_differe
print(df_temperature_differences_PC2)
df_temperature_differences_PC2.to_csv('OutsideDiff2.csv', sep=',', index=Fal
```

```
       mapped_veh_id       timestamps_UTC  Temperature  RS_E_InAirTemp_PC2  
\
0              126.0  2023-03-02 06:57:38         -1.6                65.0   
1              126.0  2023-03-02 07:12:18         -1.5                65.0   
2              126.0  2023-03-02 07:13:21         -1.5                65.0   
3              126.0  2023-03-02 07:14:21         -1.5                65.0   
4              126.0  2023-03-02 07:14:32         -1.5                65.0   
...              ...                  ...          ...                 ...   
17563          192.0  2023-03-20 19:21:58         10.0                65.0   
17564          194.0  2023-03-09 18:01:43         10.0                65.0   
17565          194.0  2023-03-09 18:02:46         10.0                65.0   
17566          194.0  2023-03-27 07:39:38          4.0                59.0   
17567          194.0  2023-04-21 22:02:43          5.0                60.0   

        abs  
0      66.6  
1      66.5  
2      66.5  
3      66.5  
4      66.5  
...     ...  
17563  55.0  
17564  55.0  
17565  55.0  
17566  55.0  
17567  55.0  

[17568 rows x 5 columns]
```

With this we have a total of ~100K outliers:

```
- Values with absolute difference from outside temperature
PC1: 13541
- Values with absolute difference from outside temperature
PC2: 17568
```

In [31]:

```
# Analysis of PC2 in air temperature vs outside temperature (dashboard outpu
query_temperature_differences_PC2_2 = f"""
SELECT *, ABS("Temperature"-"RS_E_InAirTemp_PC2")
FROM vehicle_data_enriched
WHERE "Temperature" is not null AND ABS("Temperature"-"RS_E_InAirTemp_PC2")
ORDER BY ABS("Temperature"-"RS_E_InAirTemp_PC2") DESC;
"""

df_temperature_differences_PC2_2 = pd.read_sql_query(query_temperature_diffe
df_temperature_differences_PC2_2['outlier_type'] = 'Outisde temperature'
print(df_temperature_differences_PC2_2)
df_temperature_differences_PC2_2.to_csv('R9-2.csv', sep=',', index=False, en
       mapped_veh_id       timestamps_UTC        lat       lon  \
0              126.0  2023-03-02 06:57:38  51.138971  3.641456   
1              126.0  2023-03-02 07:13:21  51.143728  3.631155   
2              126.0  2023-03-02 07:16:22  51.145184  3.628014   
3              126.0  2023-03-02 07:14:32  51.144350  3.629814   
4              126.0  2023-03-02 07:19:15  51.148530  3.623480   
...              ...                  ...        ...       ...   
17563          181.0  2023-05-16 06:19:55  51.014812  3.777485   
17564          172.0  2023-04-08 22:19:41  51.189699  5.110142   
17565          120.0  2023-01-25 18:15:28  50.403879  4.438693   
17566          123.0  2023-04-18 22:11:30  51.190693  5.113100   
17567          173.0  2023-03-15 00:21:53  50.094601  4.526841   

       RS_E_InAirTemp_PC1  RS_E_InAirTemp_PC2  RS_E_OilPress_PC1  \
0                    63.0                65.0                6.0   
1                    13.0                65.0              217.0   
2                    15.0                65.0              220.0   
3                    13.0                65.0              224.0   
4                    16.0                65.0              213.0   
...                   ...                 ...                ...   
17563                30.0                64.0              265.0   
17564                51.0                61.0                3.0   
17565                22.0                53.0              189.0   
17566                33.0                64.0              238.0   
17567                25.0                56.0              203.0   

       RS_E_OilPress_PC2  rs_e_rpm_pc1  rs_e_rpm_pc2  ...  \
0                    3.0           0.0           0.0  ...   
1                    3.0         799.0           0.0  ...   
2                    3.0         806.0           0.0  ...   
3                    3.0         803.0           0.0  ...   
4                    3.0         801.0           0.0  ...   
...                  ...           ...           ...  ...   
17563               20.0         800.0           0.0  ...   
17564                3.0           0.0           0.0  ...   
17565                3.0         799.0           0.0  ...   
17566                3.0         803.0           0.0  ...   
17567                0.0         776.0           0.0  ...   

          timestamps_floor  nearest_point_id        Lat       Lon  \
0      2023-03-02 06:00:00               168  51.138971  3.641456   
1      2023-03-02 07:00:00               168  51.143728  3.631155   
2      2023-03-02 07:00:00               168  51.145184  3.628014   
3      2023-03-02 07:00:00               168  51.144350  3.629814   
4      2023-03-02 07:00:00               168  51.148530  3.623480   
...                    ...               ...        ...       ...   
17563  2023-05-16 06:00:00               208  51.014812  3.777485   
17564  2023-04-08 22:00:00               143  51.189699  5.110142   
17565  2023-01-25 18:00:00               449  50.403879  4.438693   
17566  2023-04-18 22:00:00               144  51.190693  5.113100   
17567  2023-03-15 00:00:00               567  50.094601  4.526841   

                      Time  Temperature  Humidity  Rain   abs  \
0      2023-03-02 06:00:00         -1.6      81.0   0.0  66.6   
1      2023-03-02 07:00:00         -1.5      80.0   0.0  66.5   
2      2023-03-02 07:00:00         -1.5      80.0   0.0  66.5   
3      2023-03-02 07:00:00         -1.5      80.0   0.0  66.5   
4      2023-03-02 07:00:00         -1.5      80.0   0.0  66.5   
...                    ...          ...       ...   ...   ...   
17563  2023-05-16 06:00:00          9.0      81.0   0.0  55.0   
17564  2023-04-08 22:00:00          6.0      91.0   0.0  55.0   
17565  2023-01-25 18:00:00         -2.0      89.0   0.0  55.0   
17566  2023-04-18 22:00:00          9.0      76.0   0.0  55.0   
17567  2023-03-15 00:00:00          1.0      93.0   0.0  55.0   
```

```
              outlier_type
0      Outisde temperature
1      Outisde temperature
2      Outisde temperature
3      Outisde temperature
4      Outisde temperature
...                    ...
17563  Outisde temperature
17564  Outisde temperature
17565  Outisde temperature
17566  Outisde temperature
17567  Outisde temperature

[17568 rows x 24 columns]
```