Absenteeism at work

This Data Analysis is for the “introduction to data science” course

Objective

- IDENTIFY EMPLOYEES WITH HIGH ABSENTEEISM

  Understand which employees have the highest absenteeism rates.

- ANALYZE REASONS FOR ABSENTEEISM

  Investigate potential reasons for absenteeism, including distance from work, family responsibilities, and workload

- TEMPORAL ANALYSIS

  Identify the months and days with the highest absenteeism.

Objective

- CORRELATION ANALYSIS

  Determine the correlation between absenteeism and other factors.

- REASON DISTRIBUTION

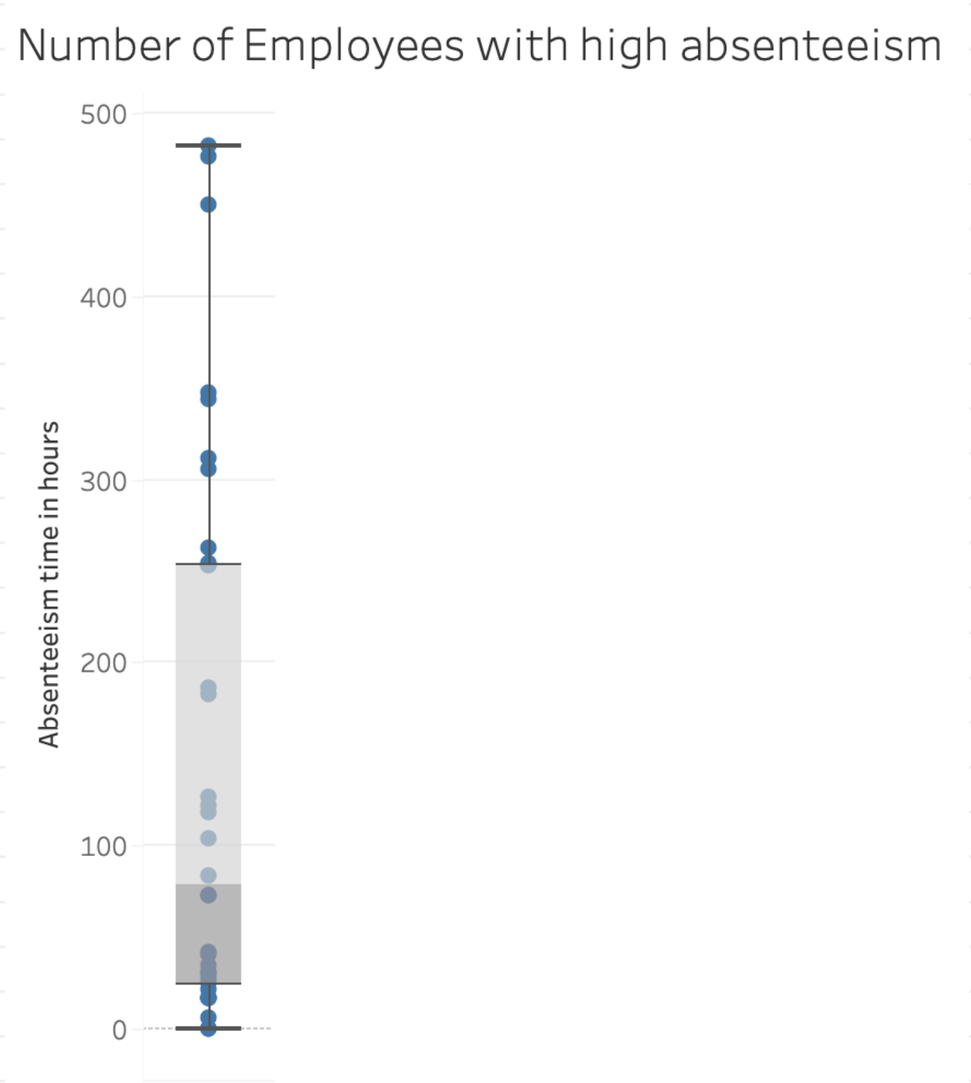
  Analyze the distribution of reasons for absenteeism.

Hypotheses

1. **High Absenteeism Group:** A group of employees consistently exhibit higher absenteeism than others.
2. **Distance and Cost:** Employees living far from the office and incurring high transportation costs have higher absenteeism.
3. **Family Responsibilities:** Employees with more children are more likely to be absent.
4. **Workload Impact:** Employees with high workloads tend to have more absenteeism.
5. **Time Patterns:** Certain months and days exhibit higher absenteeism.
6. **Correlation with Attributes:** Explore correlations between absenteeism and attributes such as age, distance, and health.

1. High Absenteeism Group

- Employees with IDs 3, 14, and 11 show consistently high absenteeism.
- The box plot analysis indicates that most employees fall within a normal range of absenteeism, with only a few outliers.



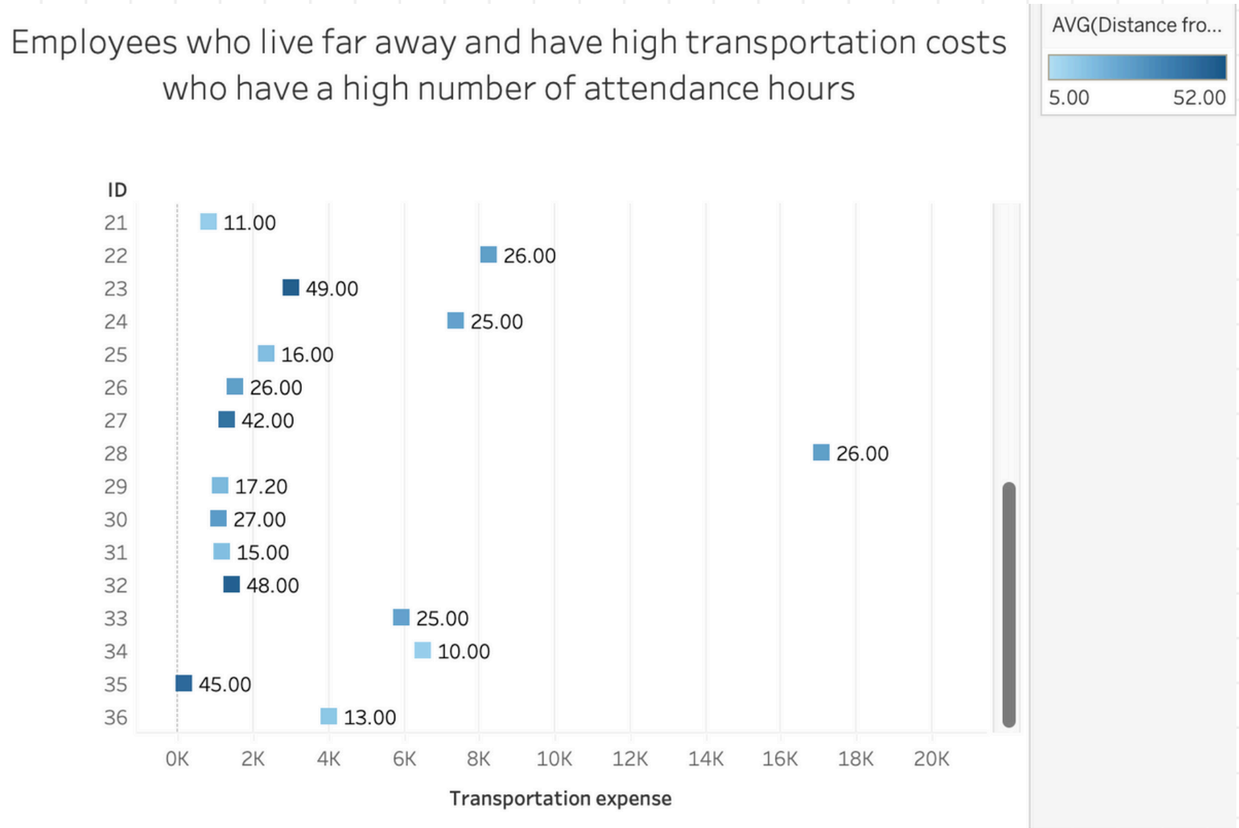
2. Reasons for Absenteeism

a. Distance and Cost

Hypothesis: Employees living farther away have higher absenteeism.

Findings:

- No direct correlation between distance and absenteeism. Employees with high transportation costs also show high absenteeism, but this is not consistent with distance.
- Notable cases: Employees ID 3 (52 km) and ID 28 (26 km) have high costs but different distances.



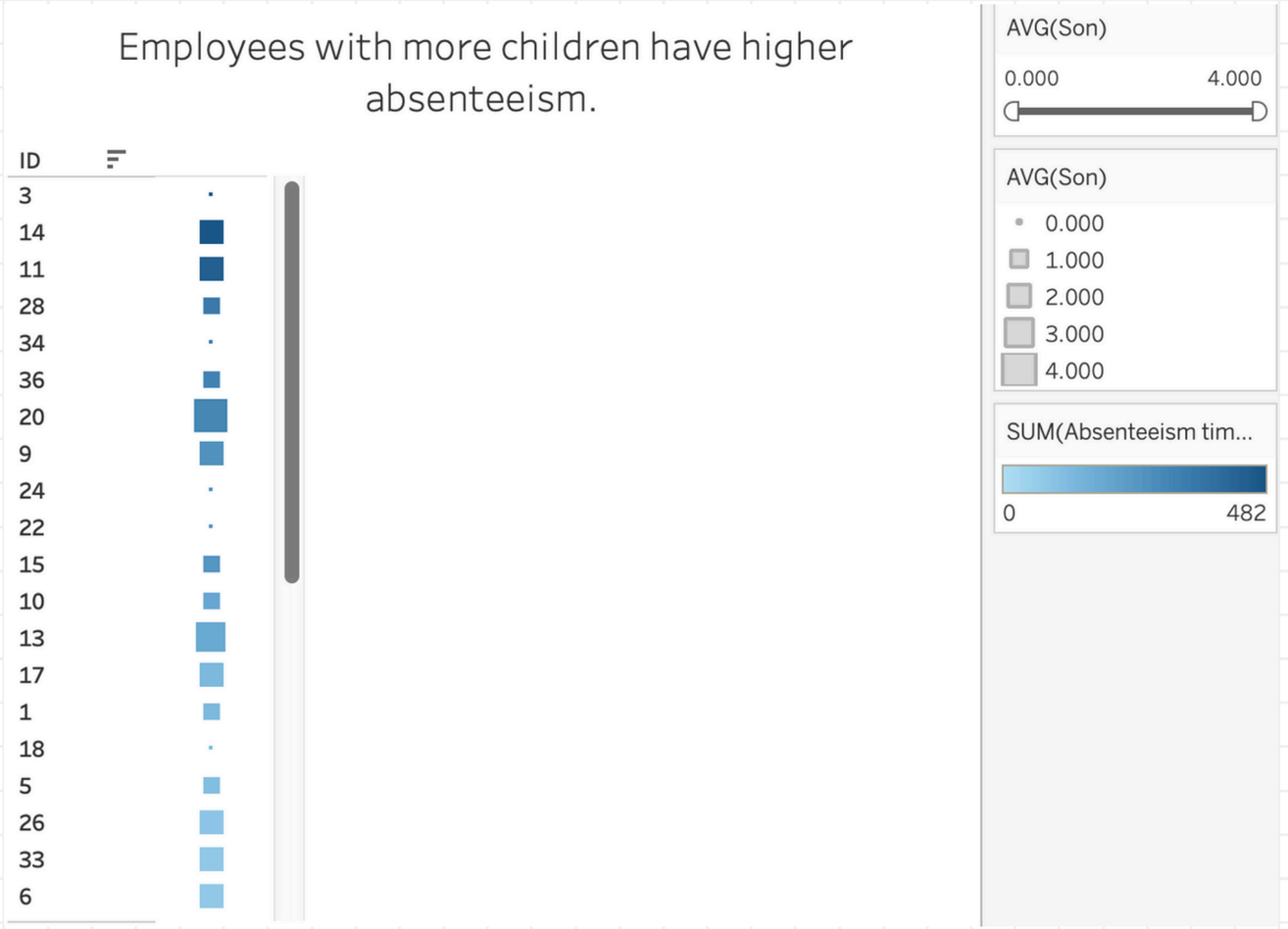
2. Reasons for Absenteeism

b. Family Responsibilities

Hypothesis: More children lead to higher absenteeism.

Findings:

- No significant correlation. Employees with more children do not necessarily have higher absenteeism rates.



2. Reasons for Absenteeism

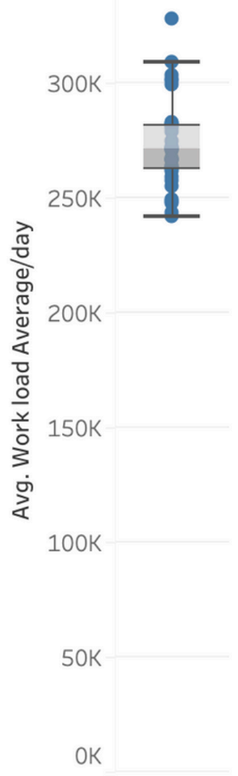
c. Workload Impact

Hypothesis: Higher workloads result in more absenteeism.

Findings:

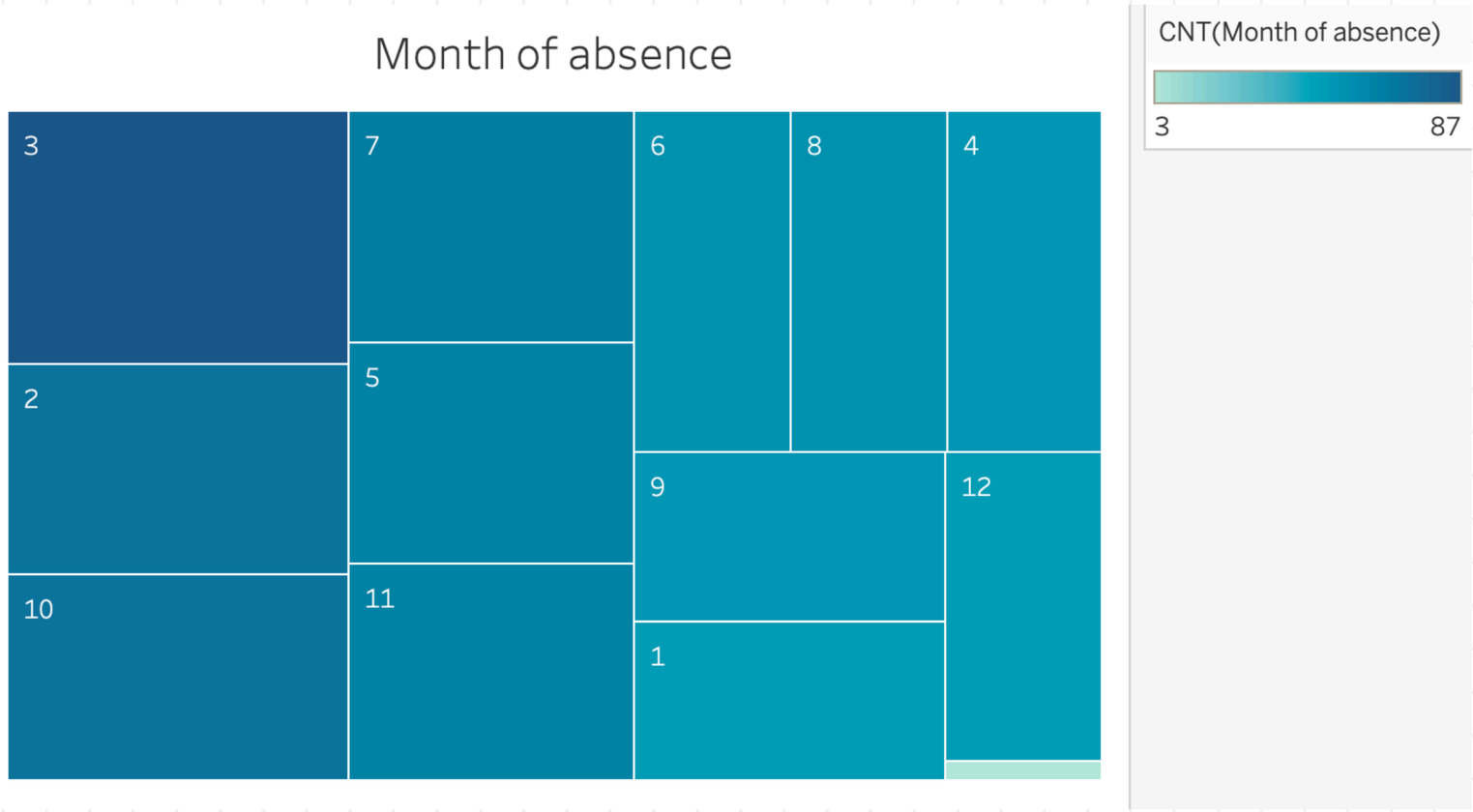
- The analysis shows no strong relationship between workload and absenteeism. Employees with the highest workload do not have high absenteeism rates.

Employees with the highest number of hours of absence have the highest average workload/day



3. Temporal Patterns

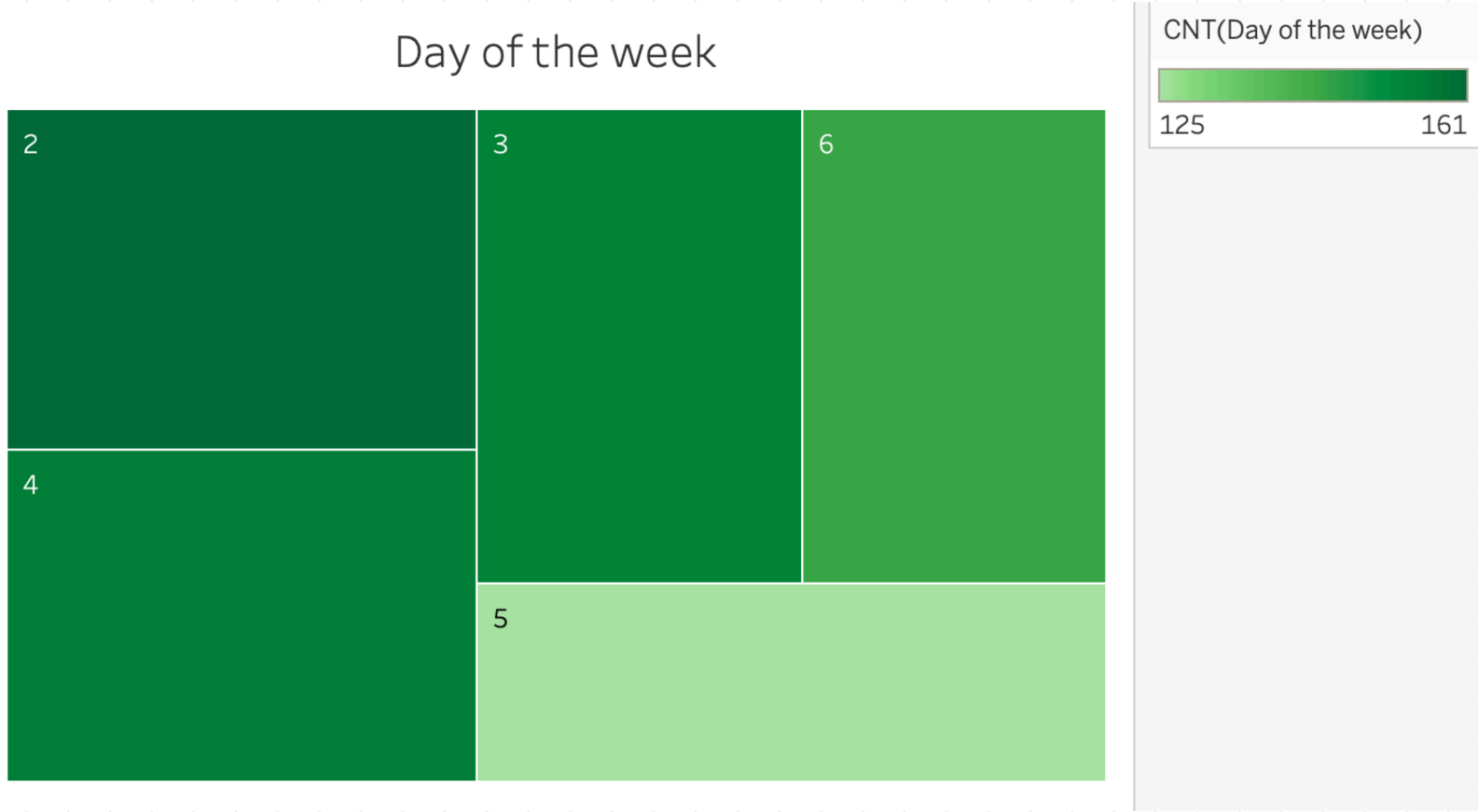
a. Months with Highest Absenteeism



March is the month with the highest absenteeism, followed by **December** with the least absenteeism.

3. Temporal Patterns

a. Days with Highest Absenteeism

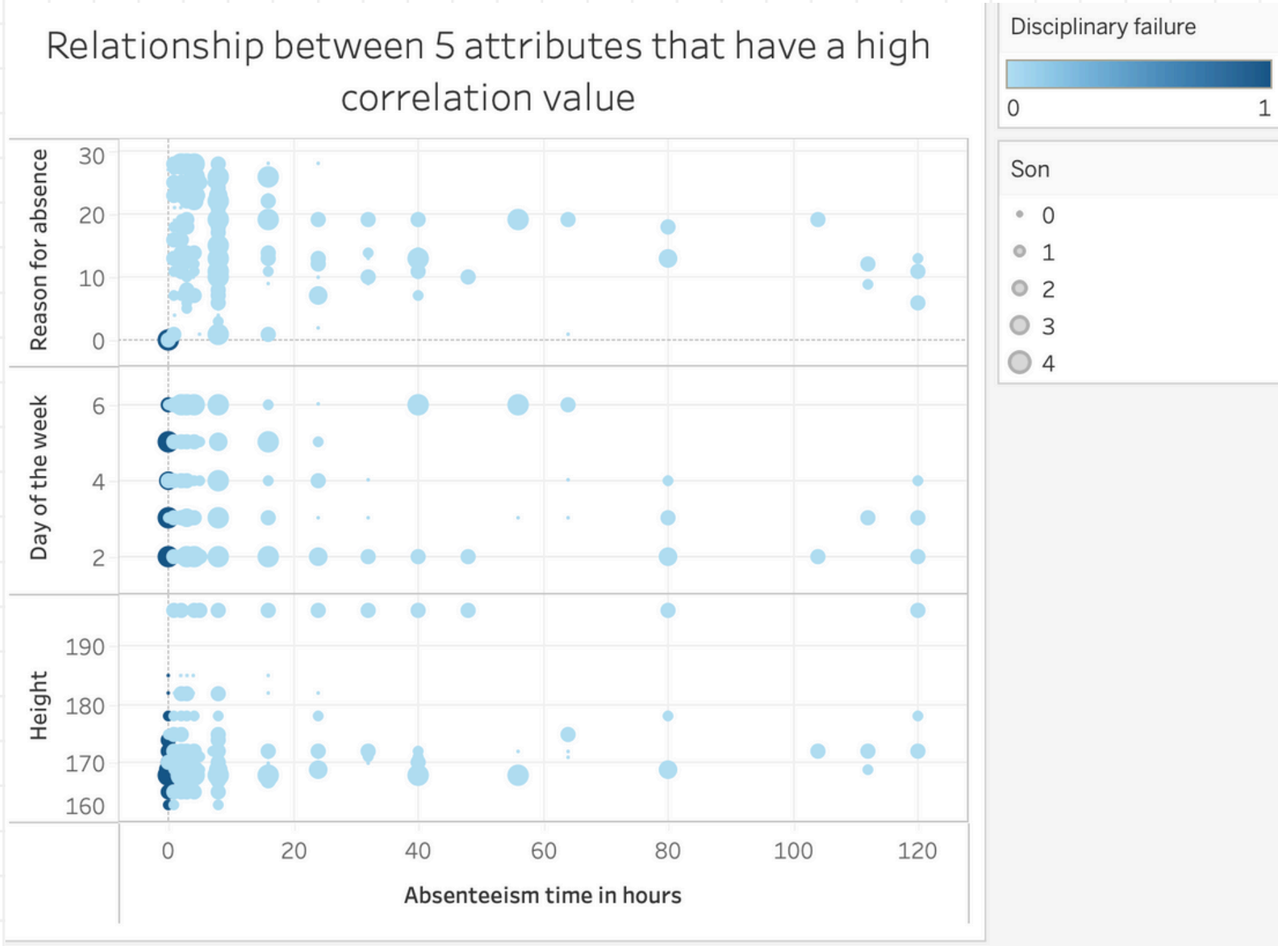


Monday is the day with the highest absenteeism, while **Thursday** has the lowest.

4. Correlation Analysis

Top Correlated Attributes:

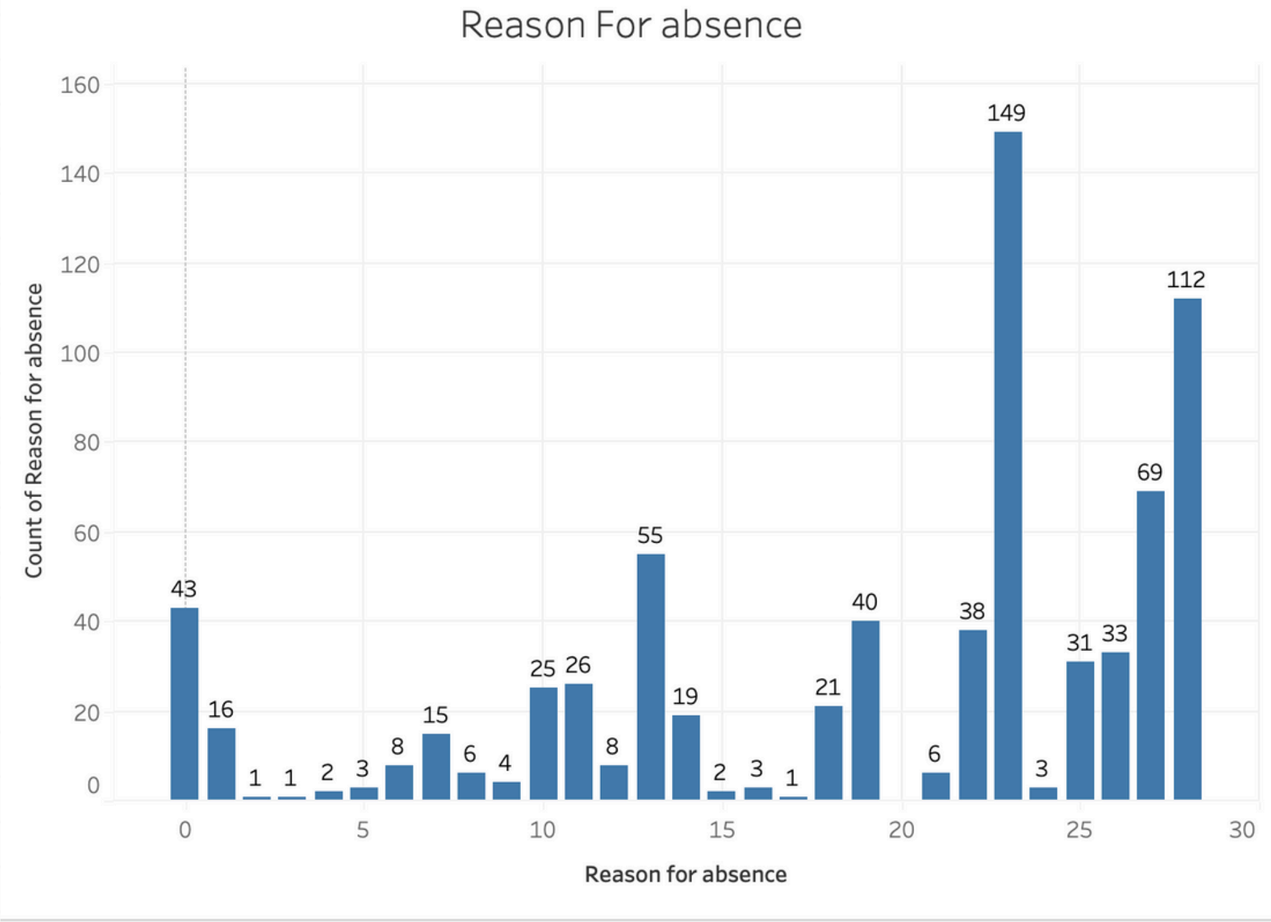
- Attributes with a weak correlation to absenteeism time include:
  - Reason for absence: -0.173
  - Day of the week: -0.124
  - Disciplinary failure: -0.124
  - Son (number of children): 0.114
  - Height: -0.144
- Most attributes show weak correlation, indicating limited impact on absenteeism.



4. Reason Distribution

Top Reasons for Absenteeism:

- **Medical Consultations** (23): 149 instances
- **Dental Consultations** (28): 112 instances
- Least used reasons include **Neoplasms** (2) and **Blood Disorders** (3).



Conclusion

The analysis reveals that while some hypotheses were not supported, key insights were gained:

- Absenteeism is within normal bounds, with a few high absenteeism cases.
- Common reasons for absenteeism are medical consultations, unrelated to distance, family size, or workload.
- Temporal patterns suggest specific months and days have higher absenteeism rates.
- Correlations between absenteeism and other attributes are generally weak, indicating complex interdependencies.