Analysis of Employee Absenteeism

by
Kalaimani Muthu

Introduction

The dataset contains records of absenteeism at work from July 2007 to July 2010 .

The data includes employee information, reasons for absence, and other factors that might affect absenteeism.

Data Dictionary Overview

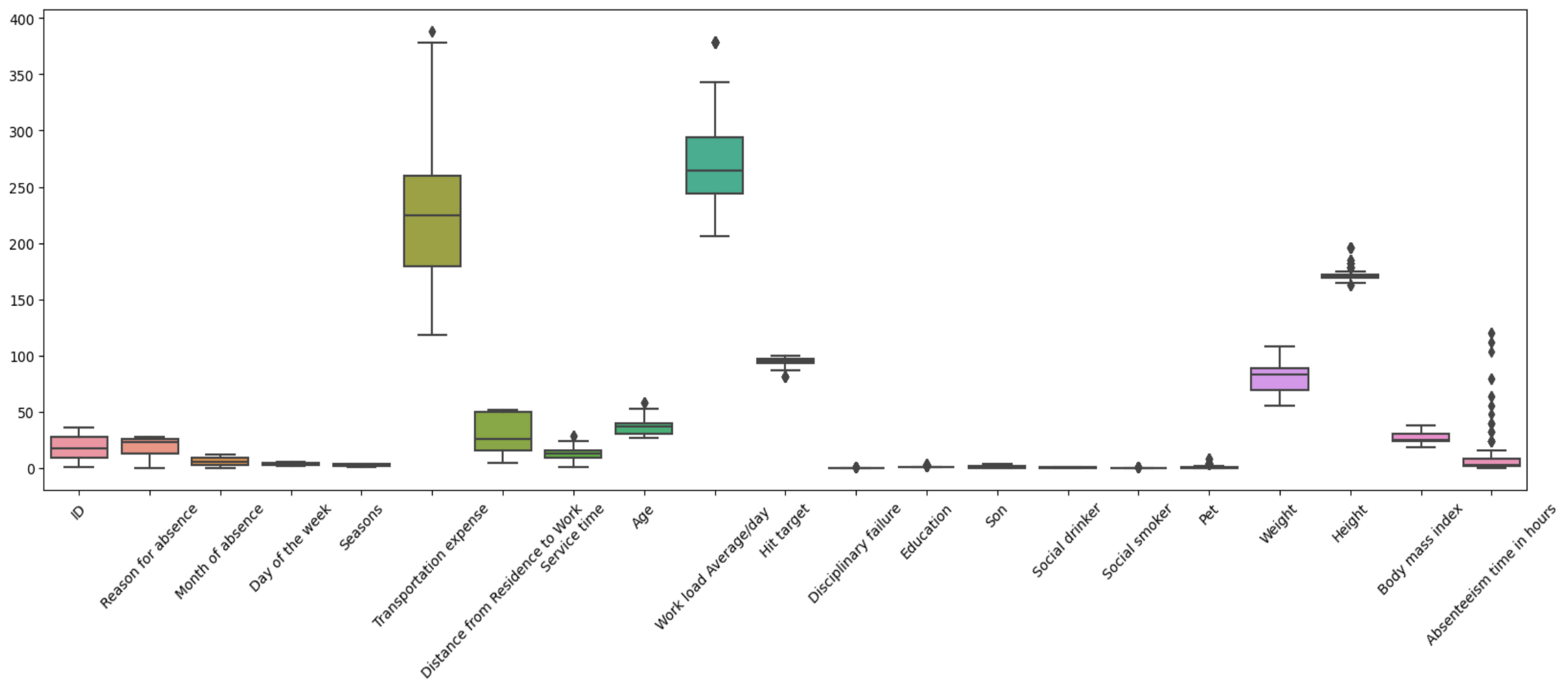
- Key columns:
- - ID: Unique identifier for each employee
- - Reason for absence: 21 categories of absences
- - Month of absence: Month during which absence occurred
- - Age, Service time, Education level, Social habits, etc.

Data Cleaning Steps

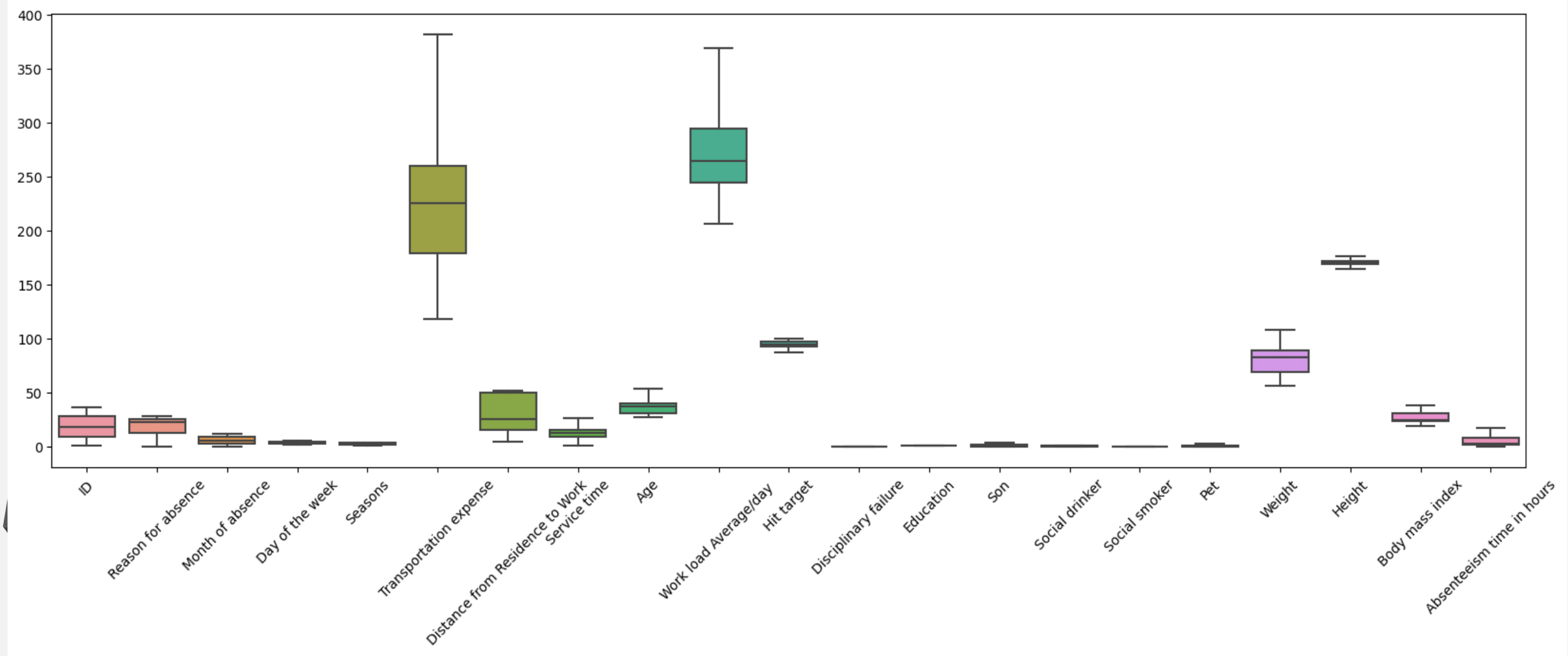
```
ID                                0
Reason for absence                0
Month of absence                  0
Day of the week                   0
Seasons                           0
Transportation expense            0
Distance from Residence to Work   0
Service time                      0
Age                               0
Work load Average/day             0
Hit target                        0
Disciplinary failure              0
Education                         0
Son                               0
Social drinker                    0
Social smoker                     0
Pet                               0
Weight                            0
Height                            0
Body mass index                   0
Absenteeism time in hours         0
Absenteeism time in hours_        0
dtype: int64
There is no missing values in the dataset
```

- The data does not have any missing values

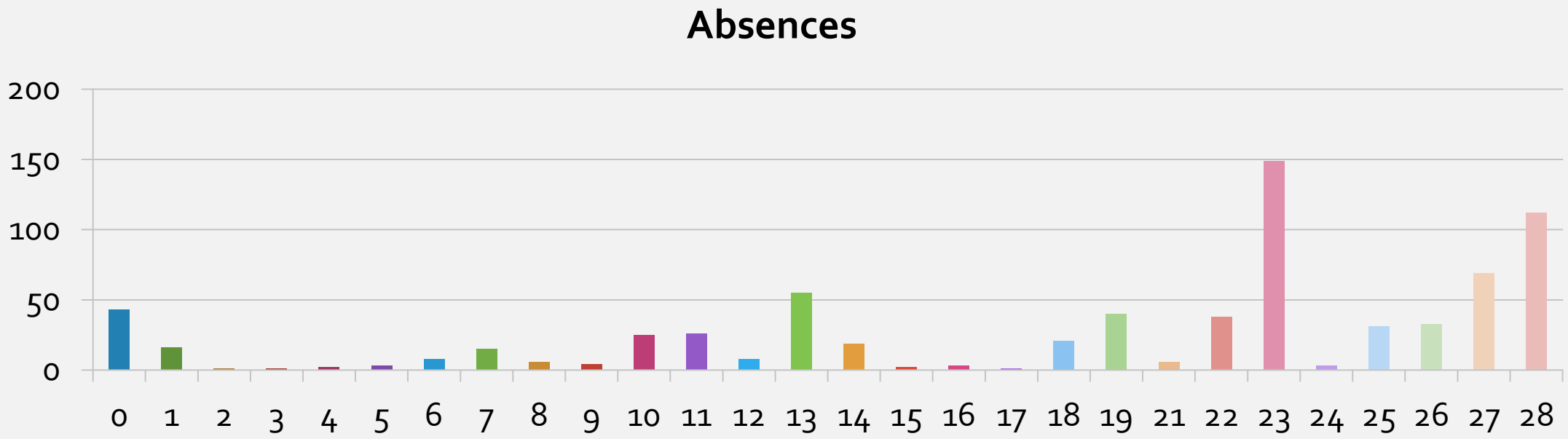
Before Outlier Treatment



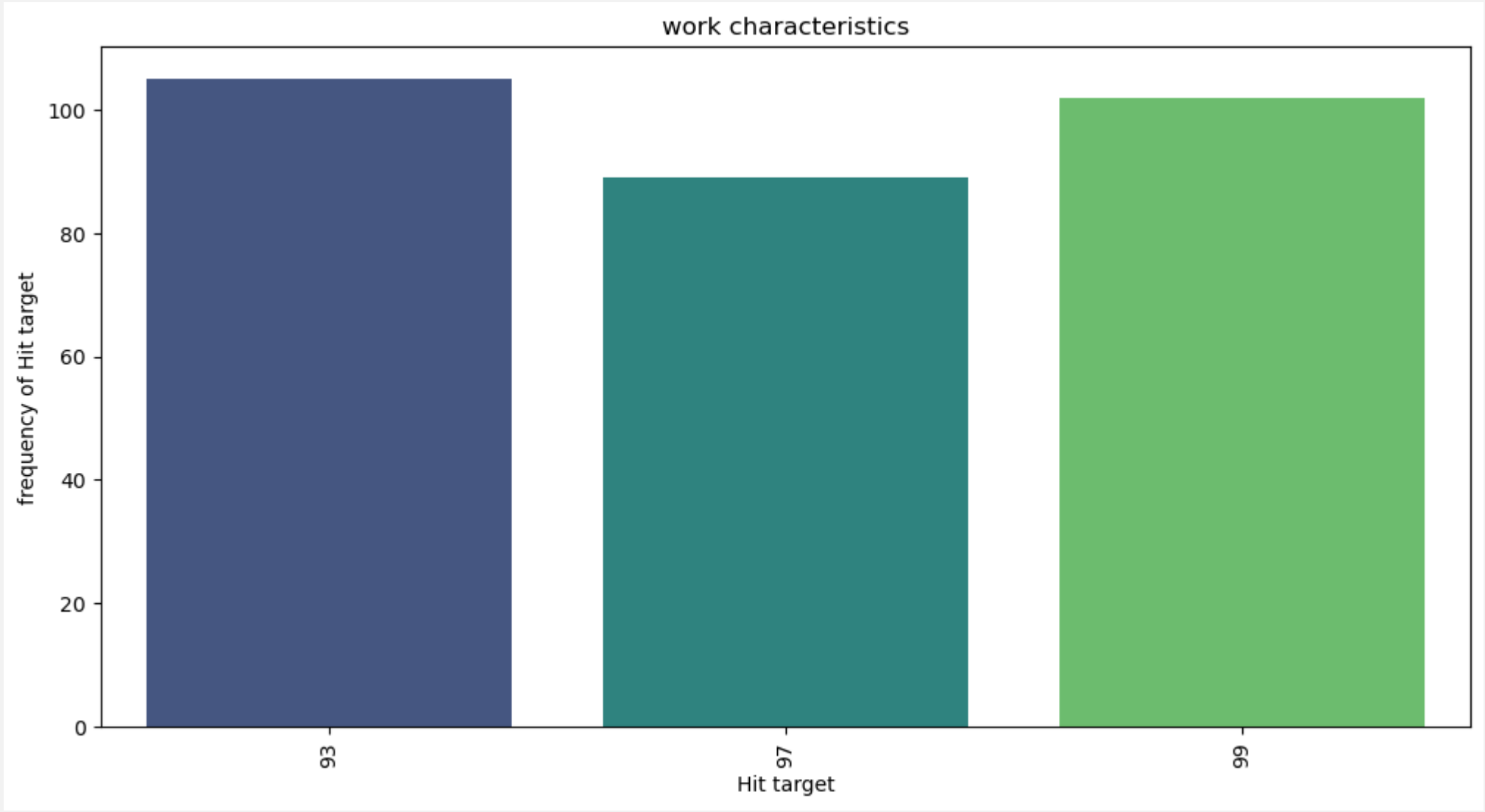
After Outlier Treatment



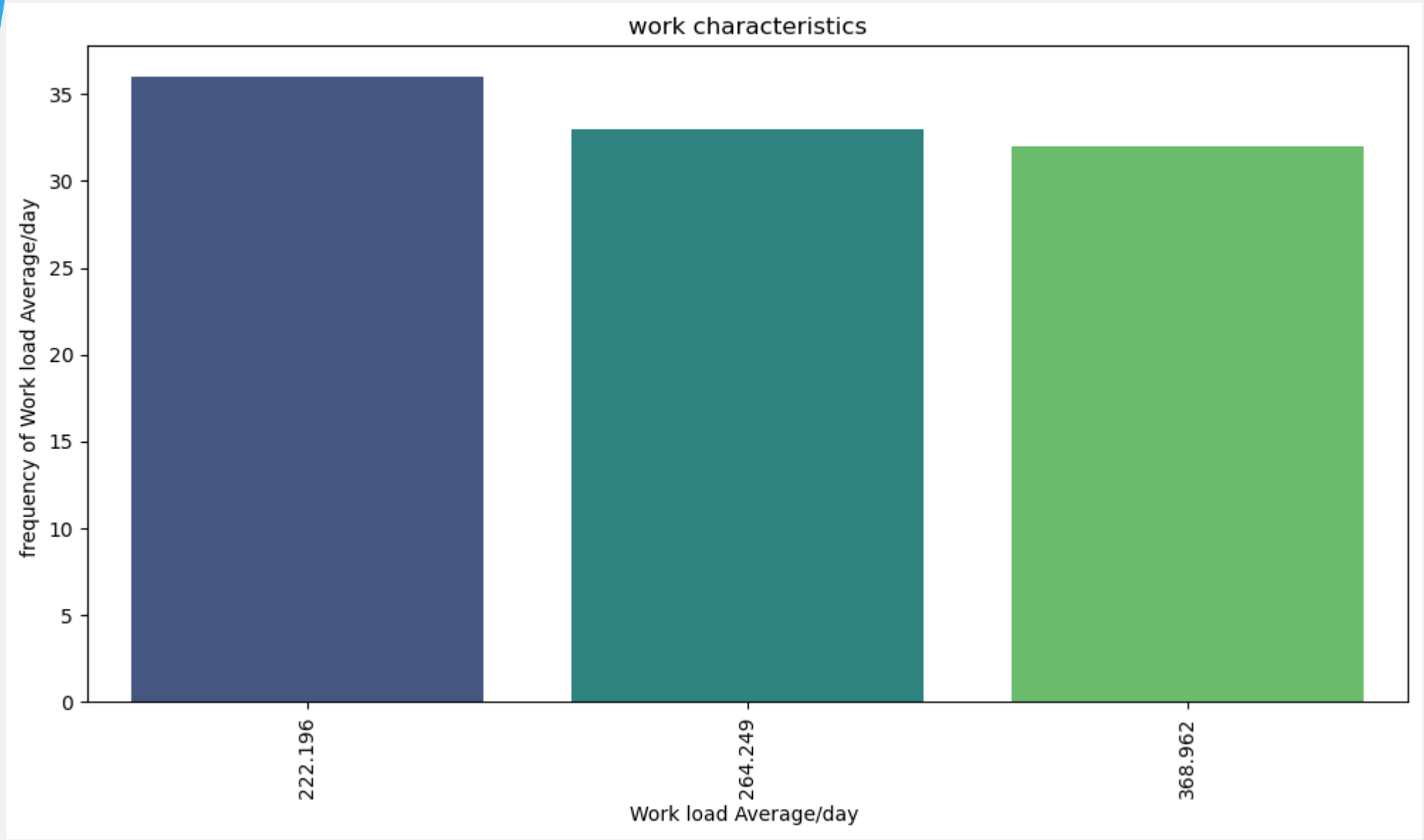
Distribution of Absenteeism by Reason



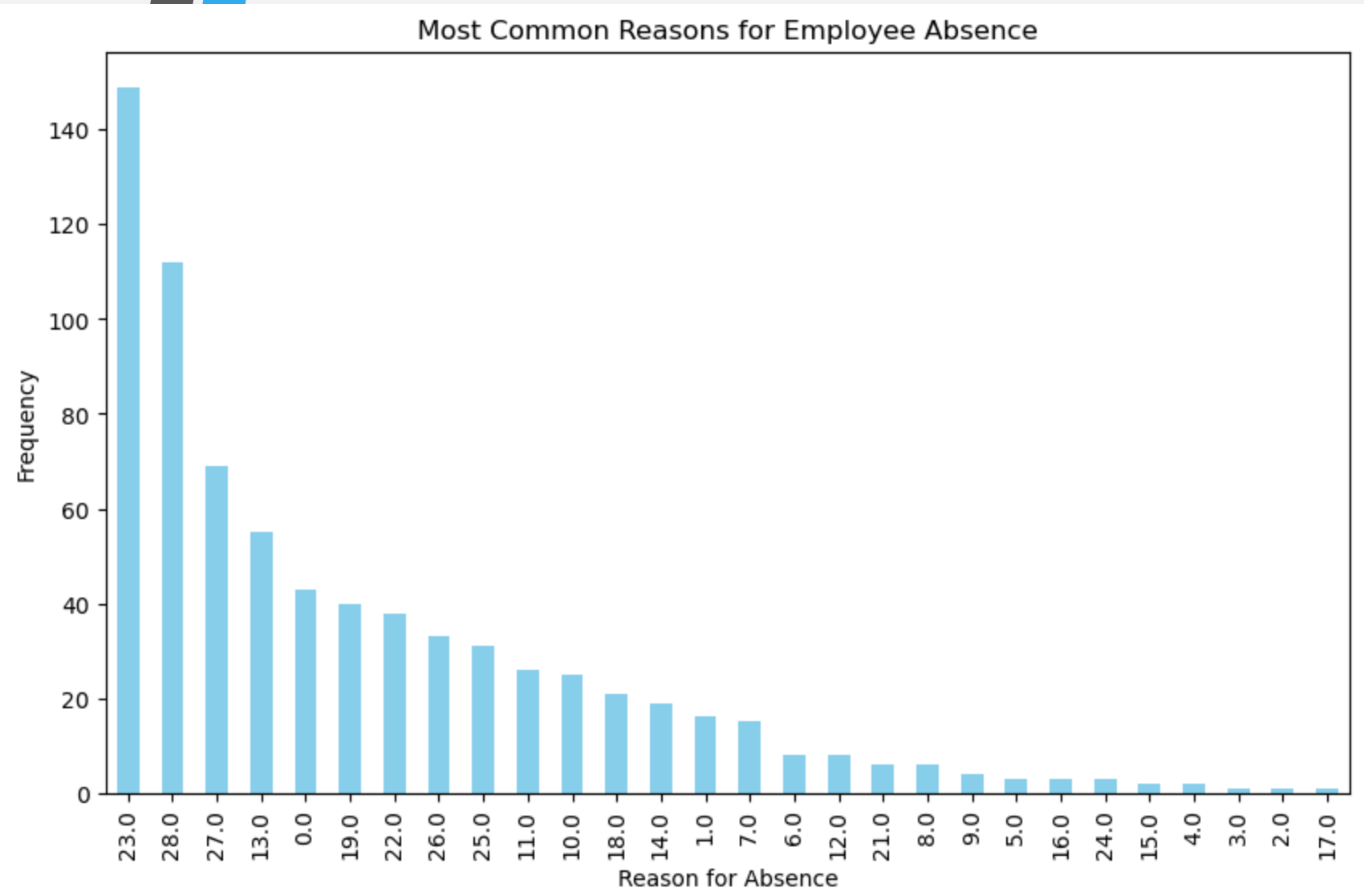
The chart shows the frequency of different reasons for absenteeism. Most absences are related to medical consultations and respiratory diseases, indicating these are major factors affecting employee absenteeism.



This charts shows how work characteristics vary across Hit target

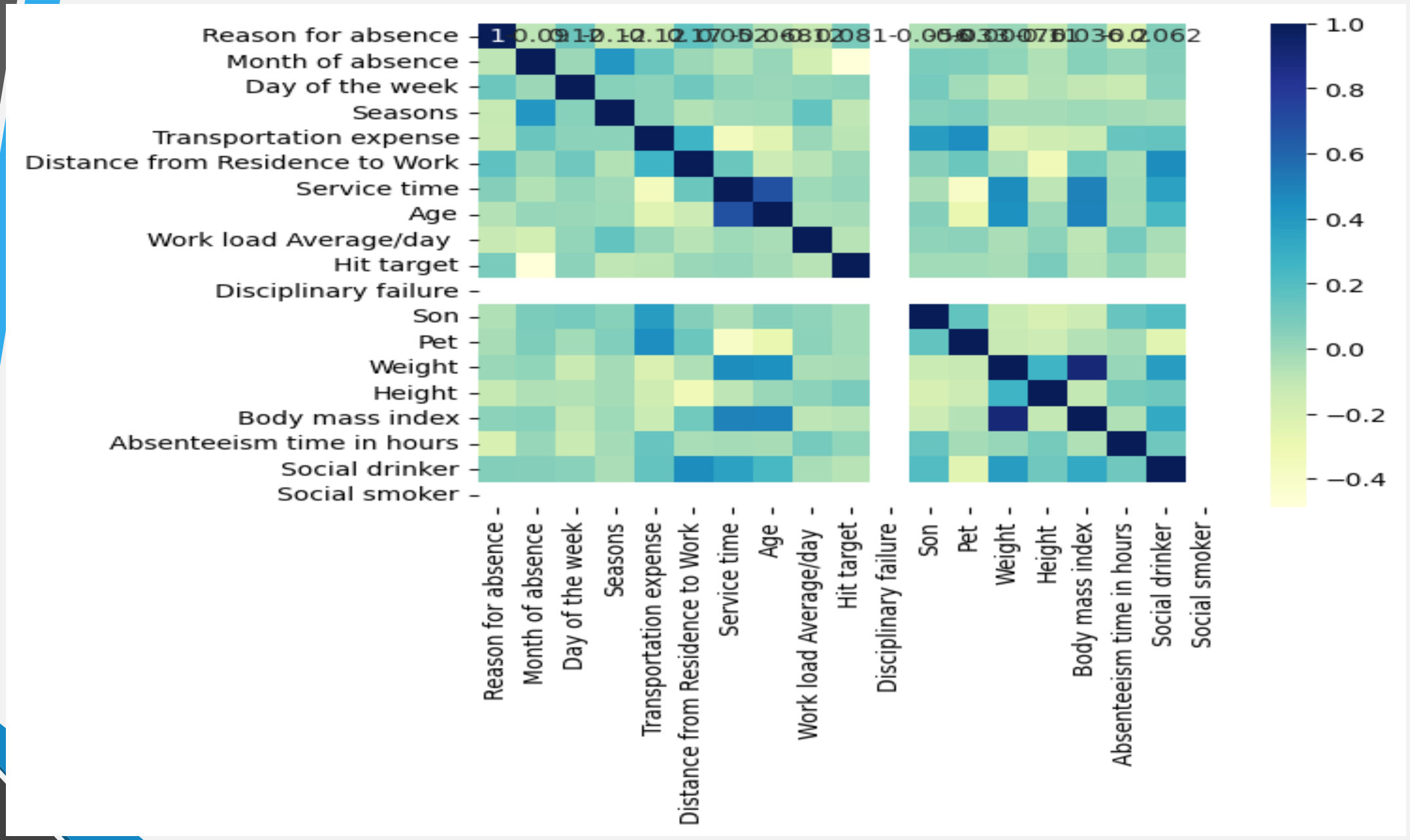


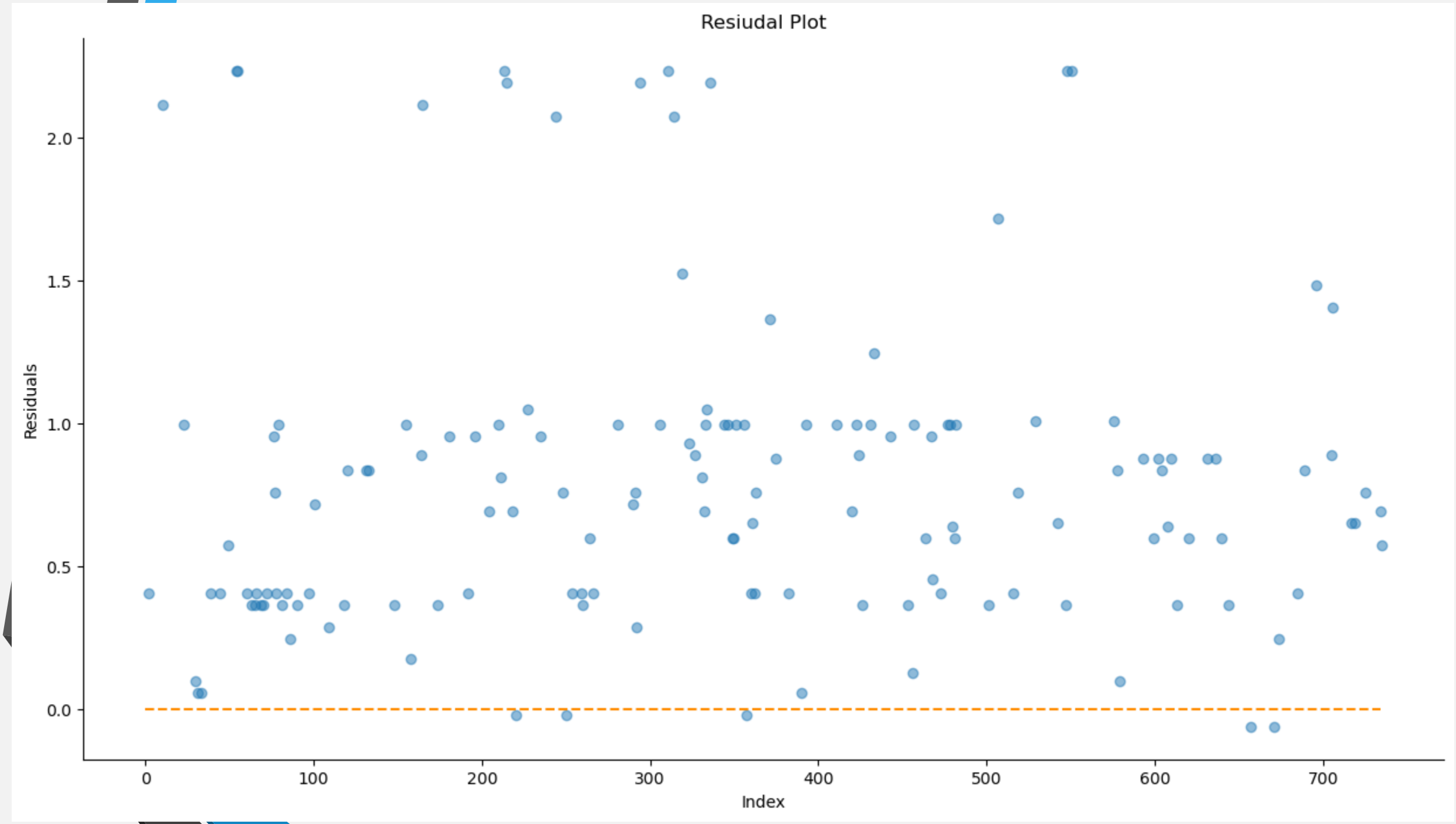
This charts shows how work characteristics vary across Work load Average/day

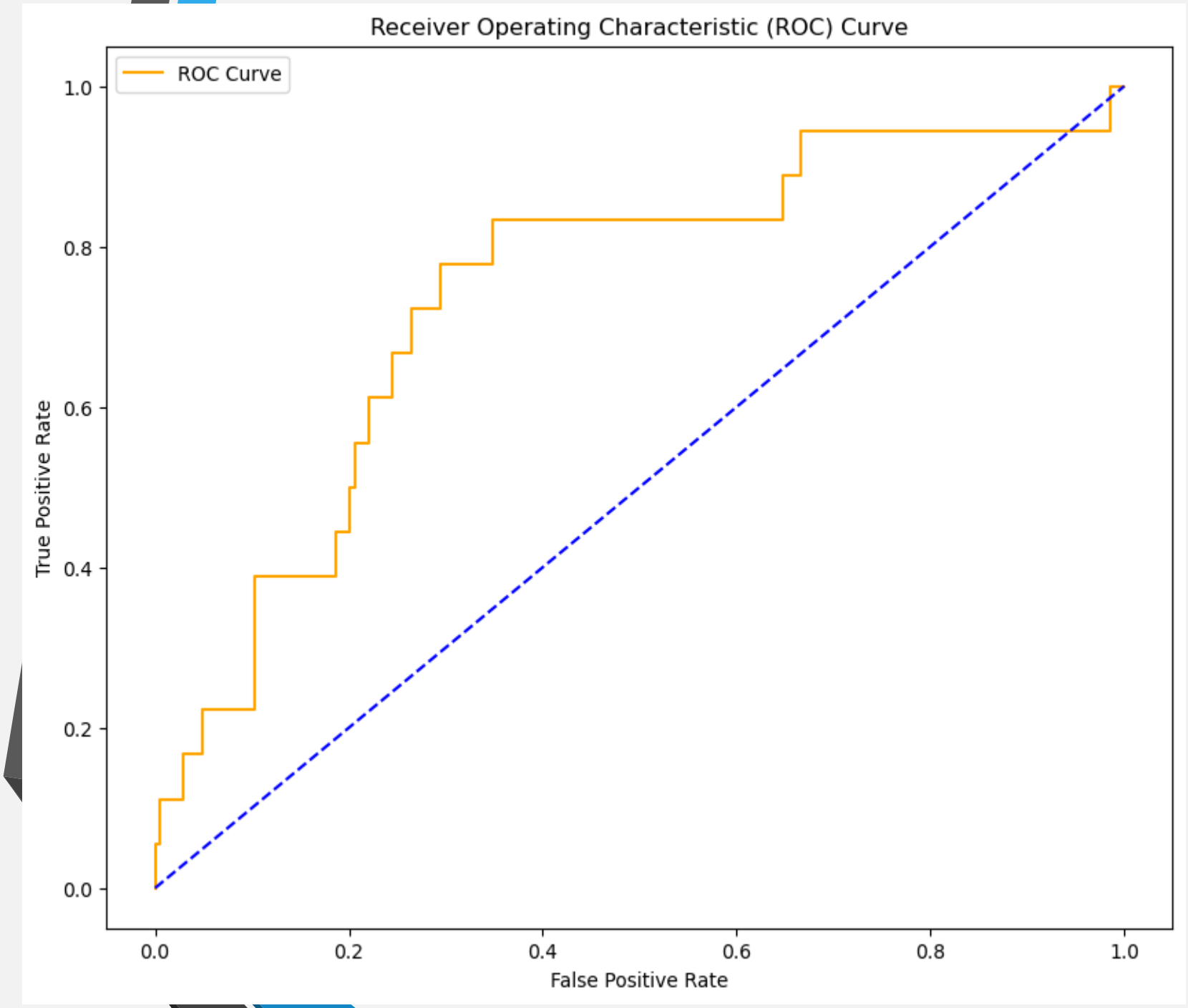


The most common reasons for employee absence are 23, 28, and 27.

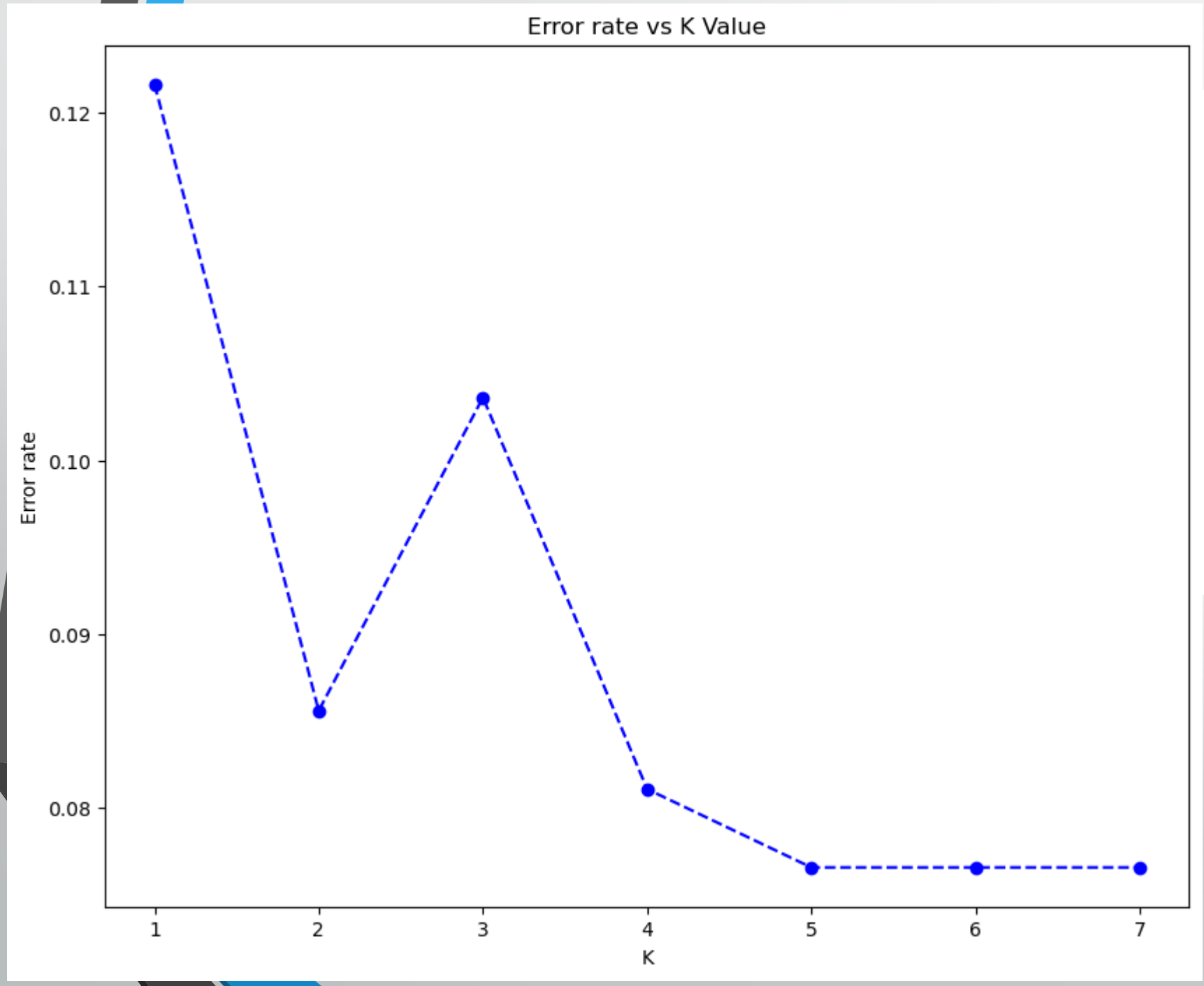
- 23 stands for Medical Consultation,
- 28 stands for Dental Consultation
- And 27 stands for Physiotherapy.


Reason for absence
Month of absence
Day of the week
Seasons
Transportation expense
Distance from Residence to Work
Service time
Age
Work load Average/day
Hit target
Disciplinary failure
Son
Pet
Weight
Height
Body mass index
Absenteeism time in hours
Social drinker
Social smoker
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Resiudal Plot
Residuals
Index
2.0
1.5
1.0
0.5
0.0
0
100
200
300
400
500
600
700



The model excels at identifying employees not at risk of absenteeism with high precision (0.93) and perfect recall (1.00), ensuring efficient resource allocation. With an overall accuracy of 92%, it provides a reliable basis for initial assessments in managing absenteeism. The ROC score of 0.741 indicates a fair ability to distinguish between different risk levels.

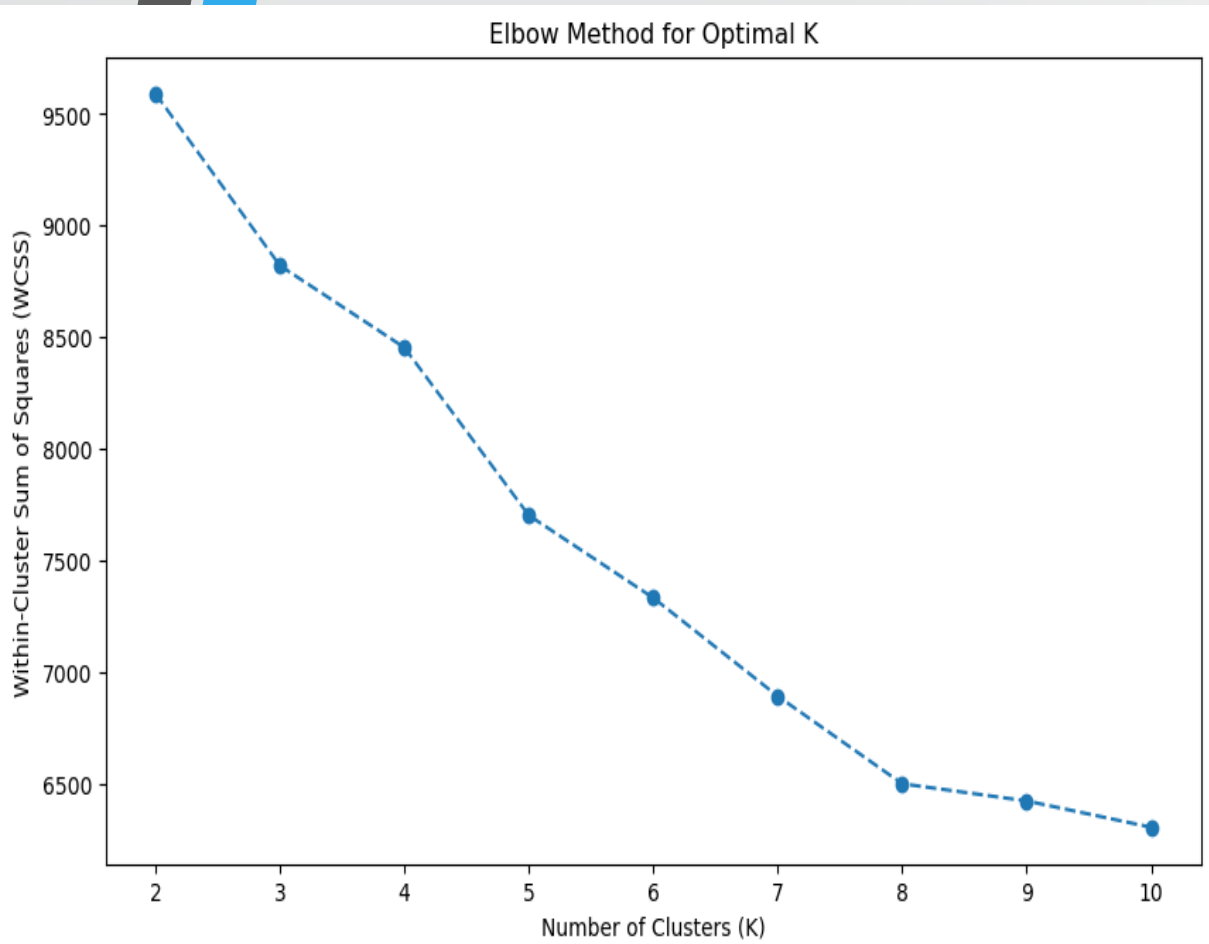


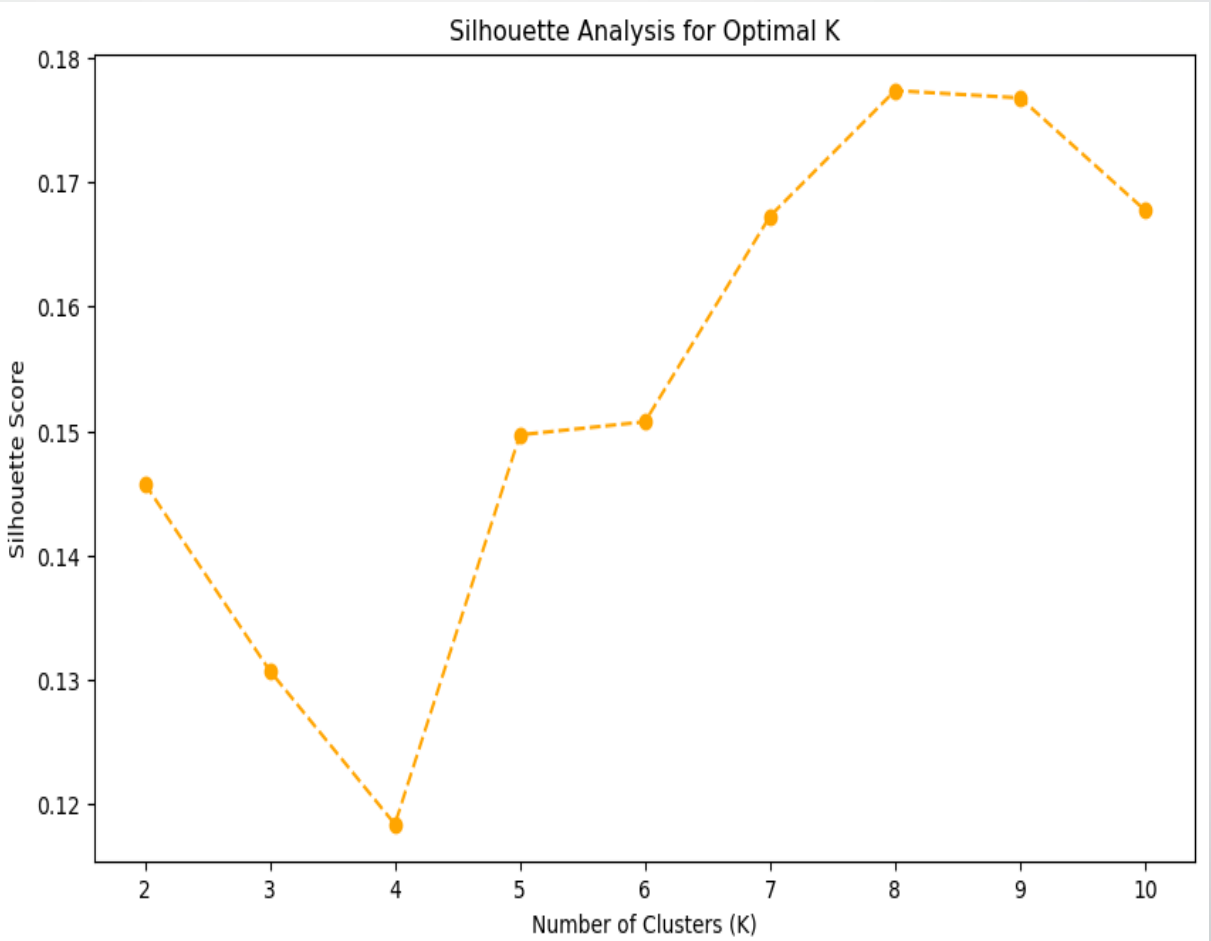

```
Accuracy: 0.918918918918919
precision_score: 0.918918918918919
recall_score: 0.918918918918919
f1_score: 0.918918918918919
confusion_matrix: [[203   1]
 [ 17   1]]
classification_report:              precision    recall  f1-score   support

           0       0.92      1.00      0.96       204
           1       0.50      0.06      0.10        18

    accuracy                           0.92       222
   macro avg       0.71      0.53      0.53       222
weighted avg       0.89      0.92      0.89       222
```

The KNN model achieves an overall high performance with an accuracy of 91.9%. The optimal K value is around 5, where the error rate minimizes and stabilizes.





The plot suggests a bend (elbow) around K=6, indicating that six clusters might be the optimal choice for minimizing within-cluster variance without having too many clusters. Silhouette Analysis:
The silhouette scores peak at K=6 with a score of approximately 0.178, suggesting that six clusters provide the best separation and cohesion of the data points.