Exploratory Data Analysis on Employee Absenteeism

Using pandas, matplotlib and seaborn.

Dataset is available on **Kaggle**

Tolulope Okerayi

BUSINESS TASK

This dataset is the absenteeism records from a courier company in Brazil from July 2007 to July 2010.
The goal is to determine the likely cause of **‘Employee Absenteeism’** at the company during working hours.
I will analyse patterns of Absenteeism, by exploring the following questions:

- What are the most common reasons for employee absenteeism?
- How does the season, day of the week or month affect absenteeism?
- Are there patterns in absenteeism based on certain employee demographics or characteristics?

KEY VARIABLES

- Reasons for Absence
- Month of Absence
- Day of the Week
- Seasons
- Absenteeism Time in Hours
- Distance from Residence to Work
- Transportation Expense
- Age
- Body Mass Index
- Social Drinker
- Social Smoker
- Pet
- Children

What is the total employee absenteeism time?

Total Absenteeism Hours

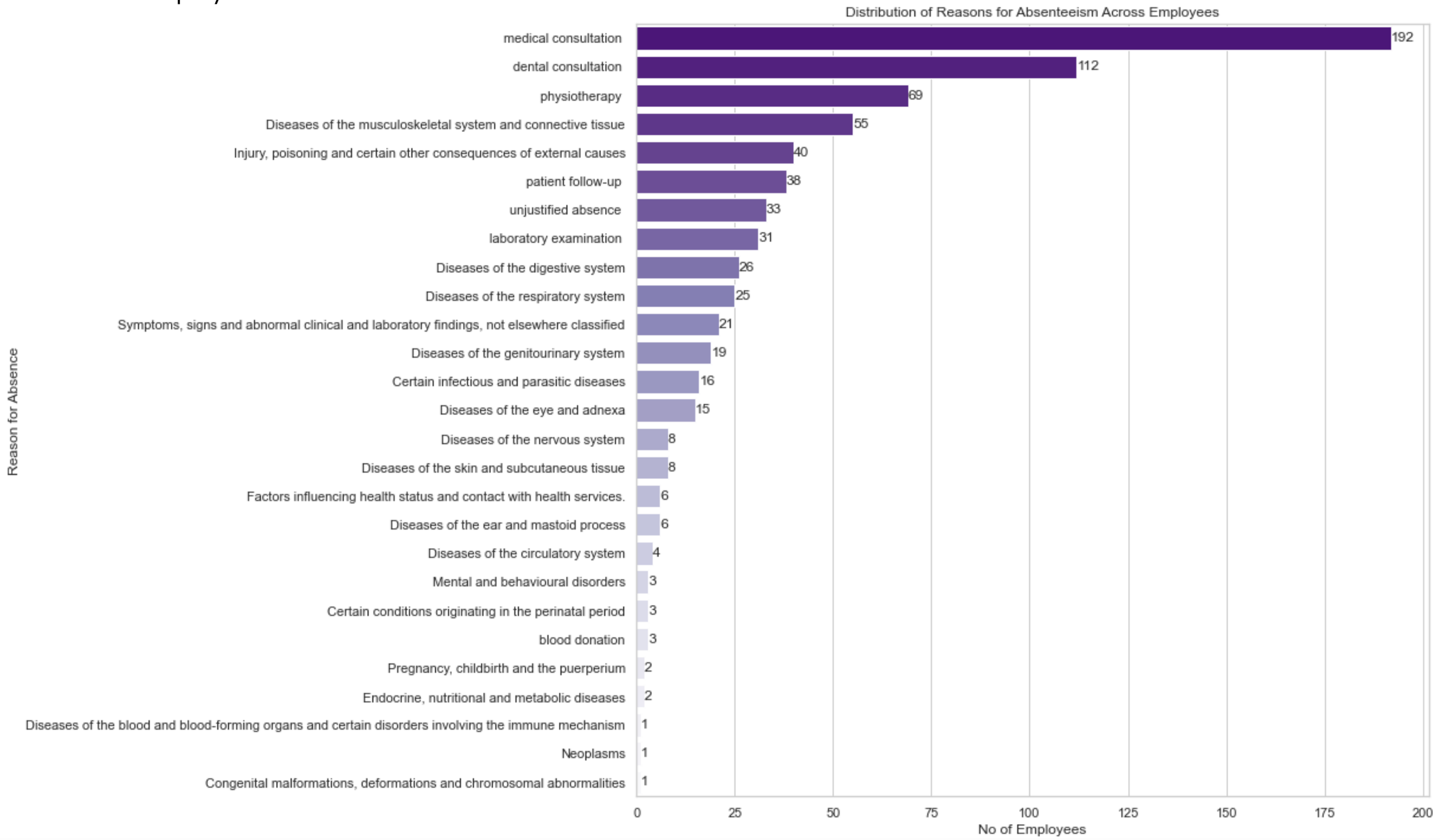
```
#Summing the absenteeism hours

total_absenteeism_hours = merged_df['absenteeism_time_in_hours'].sum()
print(f'Total Absenteeism Hours: {total_absenteeism_hours} hours')
```

The total number of hours that employees were absent at the company is **‘5432 hours’**

What is the most common reason for employee absenteeism?

```python
# How many times did each reason for employee absenteeism occur?

reason_counts = merged_df['reason'].value_counts()

sns.set(style="whitegrid")

# Creating a bar plot
plt.figure(figsize=(12, 12))
colors = sns.color_palette("Purples", n_colors=len(reason_counts))[::-1]
sns.barplot(x=reason_counts.values, y=reason_counts.index, palette=colors)

# Setting the text labels
for i, value in enumerate(reason_counts.values):
    plt.text(value, i, f'{value}', ha='left', va='center', fontsize=12)

# Customizing the plot
plt.xlabel('No of Employees')
plt.ylabel('Reason for Absence')
plt.title('Distribution of Reasons for Absenteeism Across Employees')

plt.show()
```

Reasons for Employee Absenteeism

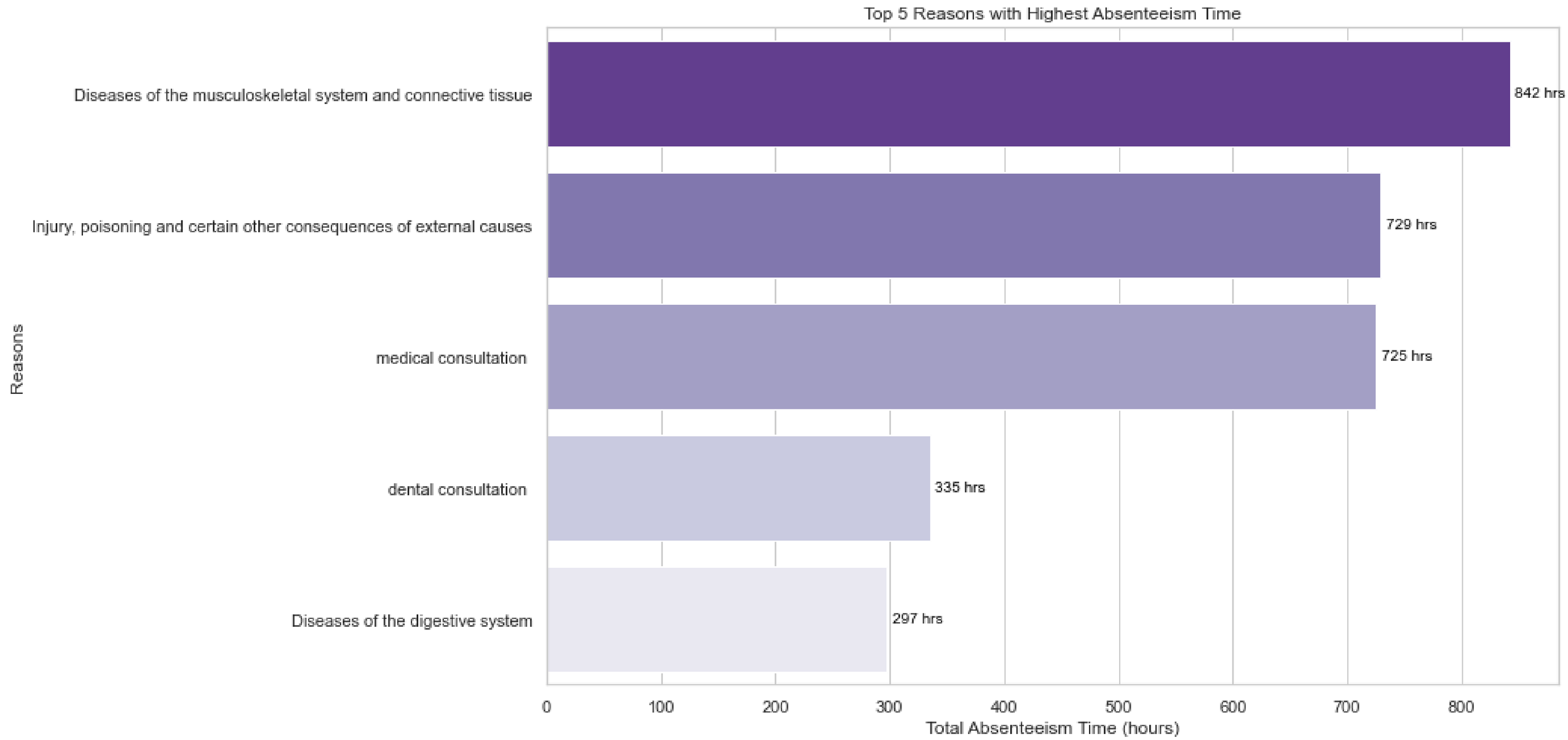


What is the most common reason for employee absenteeism?

The most frequently cited reason for employee absenteeism is **'Medical Consultation'**, with a total of **`192`** occurrences, indicating that this reason has the highest prevalence among employees who were absent.

What are the top 5 reasons with the highest absenteeism time?

```
# Grouping by reason and calculating the total absenteeism time
reason_absenteeism = merged_df.groupby('reason')['absenteeism_time_in_hours'].sum()

# Selecting the top 5 reasons with the highest absenteeism time
top5_reasons = reason_absenteeism.nlargest(5)

sns.set(style="whitegrid")

# Creating a bar plot
plt.figure(figsize=(12, 8))
colors = sns.color_palette("Purples", n_colors=len(top5_reasons))[::-1]
barplot = sns.barplot(x=top5_reasons.values, y=top5_reasons.index, palette=colors)

# Adding the text labels
for i, value in enumerate(top5_reasons.values):
    plt.text(value + 5, i, f'{value:.0f} hrs', va='center', fontsize=10, color='black')

# Customizing the plot
plt.title('Top 5 Reasons with Highest Absenteeism Time')
plt.xlabel('Total Absenteeism Time (hours)')
plt.ylabel('Reasons')


plt.show()
```


Top 5 Reasons with Highest Absenteeism Time
Diseases of the musculoskeletal system and connective tissue
Injury, poisoning and certain other consequences of external causes
medical consultation
dental consultation
Diseases of the digestive system
842 hrs
729 hrs
725 hrs
335 hrs
297 hrs
Reasons
0
100
200
300
400
500
600
700
800
Total Absenteeism Time (hours)

What are the top 5 reasons with the highest absenteeism time?

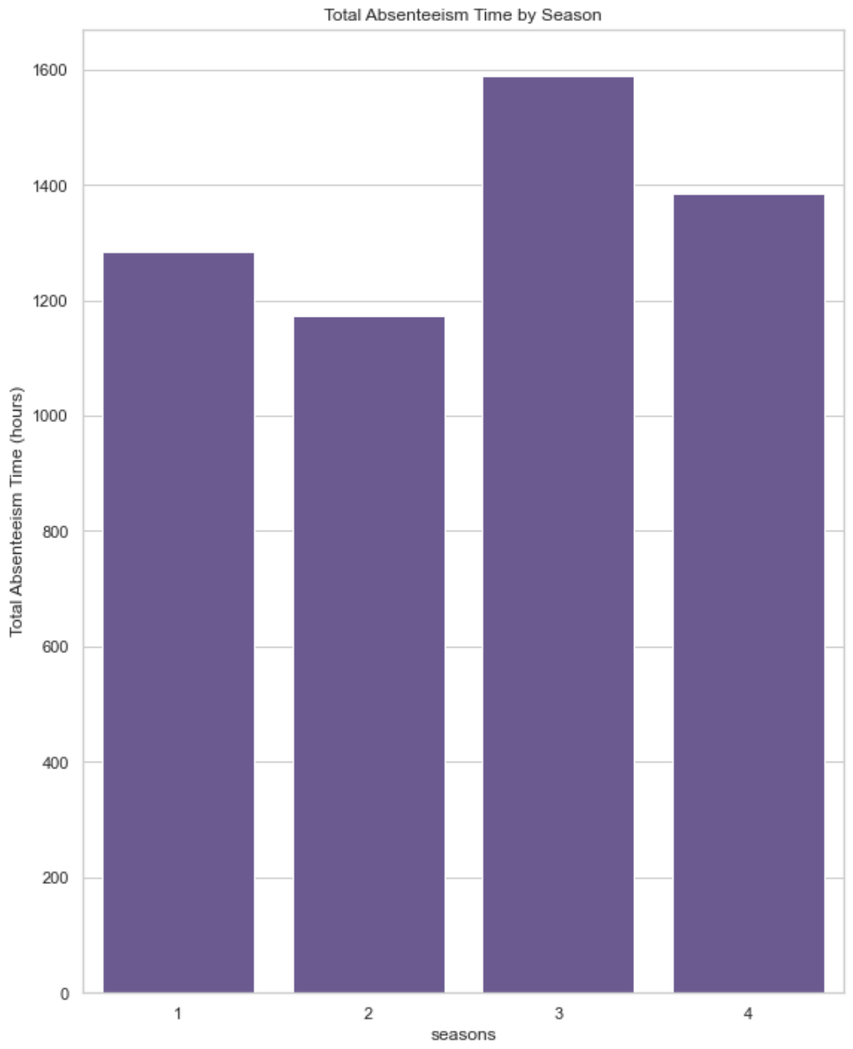
The primary reason is attributed to **'Diseases of the musculoskeletal system and connective tissue',** which leads with the highest absenteeism time of **842 hours**, followed by **'Injury, poisoning, and certain other consequences of external causes'** with **729 hours** and **'Medical Consultation'** ranking third at **725 hours.**

Dispatch riders often face challenges such as **prolonged periods of riding, exposure to vibrations, navigating through traffic, and handling packages**, which collectively increase the risk of musculoskeletal strains and injuries.

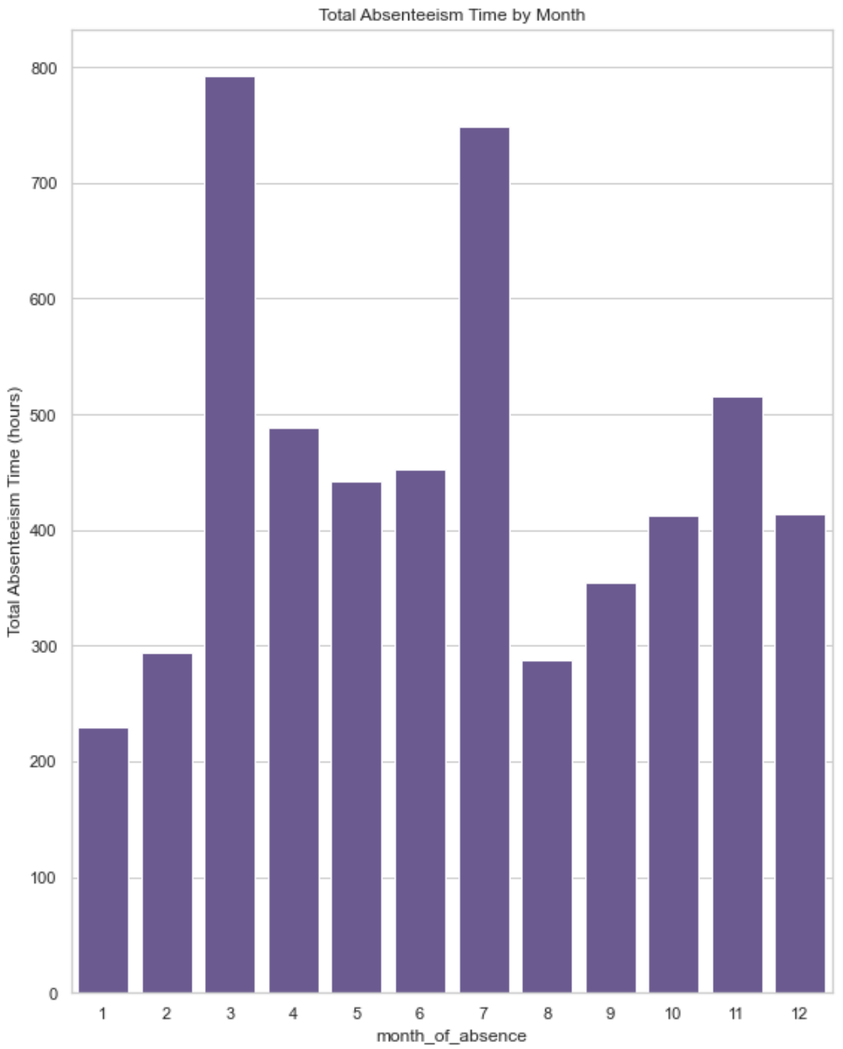
What effects does season, month and day of the week have on employee absenteeism?

```
# Grouping by season, month, and day of the week and calculating the total absenteeism time
grouped_season = merged_df.groupby('seasons')['absenteeism_time_in_hours'].sum().reset_index()
grouped_month = merged_df.groupby('month_of_absence')['absenteeism_time_in_hours'].sum().reset_index()
grouped_day = merged_df.groupby('day_of_the_week')['absenteeism_time_in_hours'].sum().reset_index()

#Setting the plot style
sns.set(style="whitegrid")
fig, axes = plt.subplots(nrows=1, ncols=3, figsize=(18, 10))
plot_color = '#645394'

# Plotting seasons
sns.barplot(x='seasons', y='absenteeism_time_in_hours', data=grouped_season, ax=axes[0], color=plot_color)
axes[0].set_title('Total Absenteeism Time by Season')

# Plotting month_of absence
sns.barplot(x='month_of_absence', y='absenteeism_time_in_hours', data=grouped_month, ax=axes[1], color=plot_color)
axes[1].set_title('Total Absenteeism Time by Month')

# Plotting day_of_the_week
sns.barplot(x='day_of_the_week', y='absenteeism_time_in_hours', data=grouped_day, ax=axes[2], color=plot_color)
axes[2].set_title('Total Absenteeism Time by Day of the Week')

# Setting a common y-axis label
for ax in axes:
    ax.set_ylabel('Total Absenteeism Time (hours)')

plt.tight_layout()
plt.show()l Absenteeism Time (hours)')
plt.ylabel('Reasons')


plt.show()
```

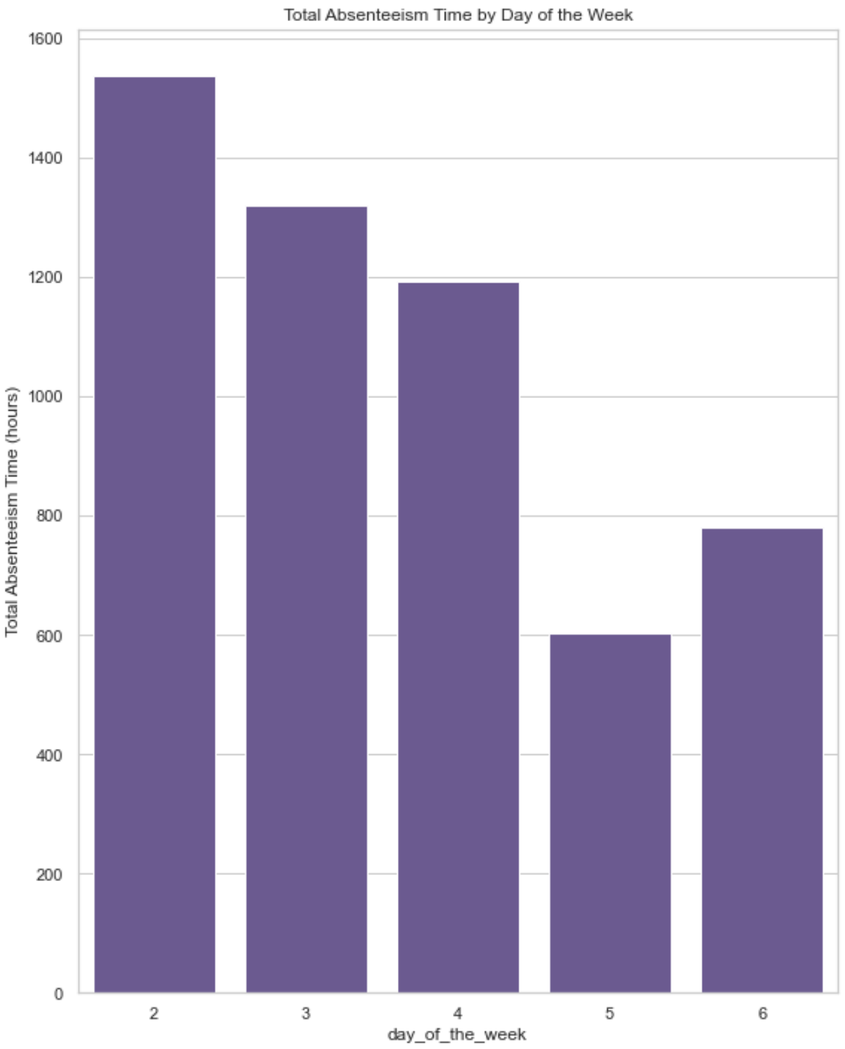
Seasons



Month of Absence



Day of the week



What effects does season, month and day of the week have on employee absenteeism?

The Seasons represents; Summer (1), Autumn (2), Winter (3), and Spring (4)**.** It is evident that **‘Winter’** records the highest absenteeism time among employees. This could be attributed to various factors, including **weather conditions, potential health issues associated with winter**, or **increased workload during the winter season.**

What effects does season, month and day of the week have on employee absenteeism?

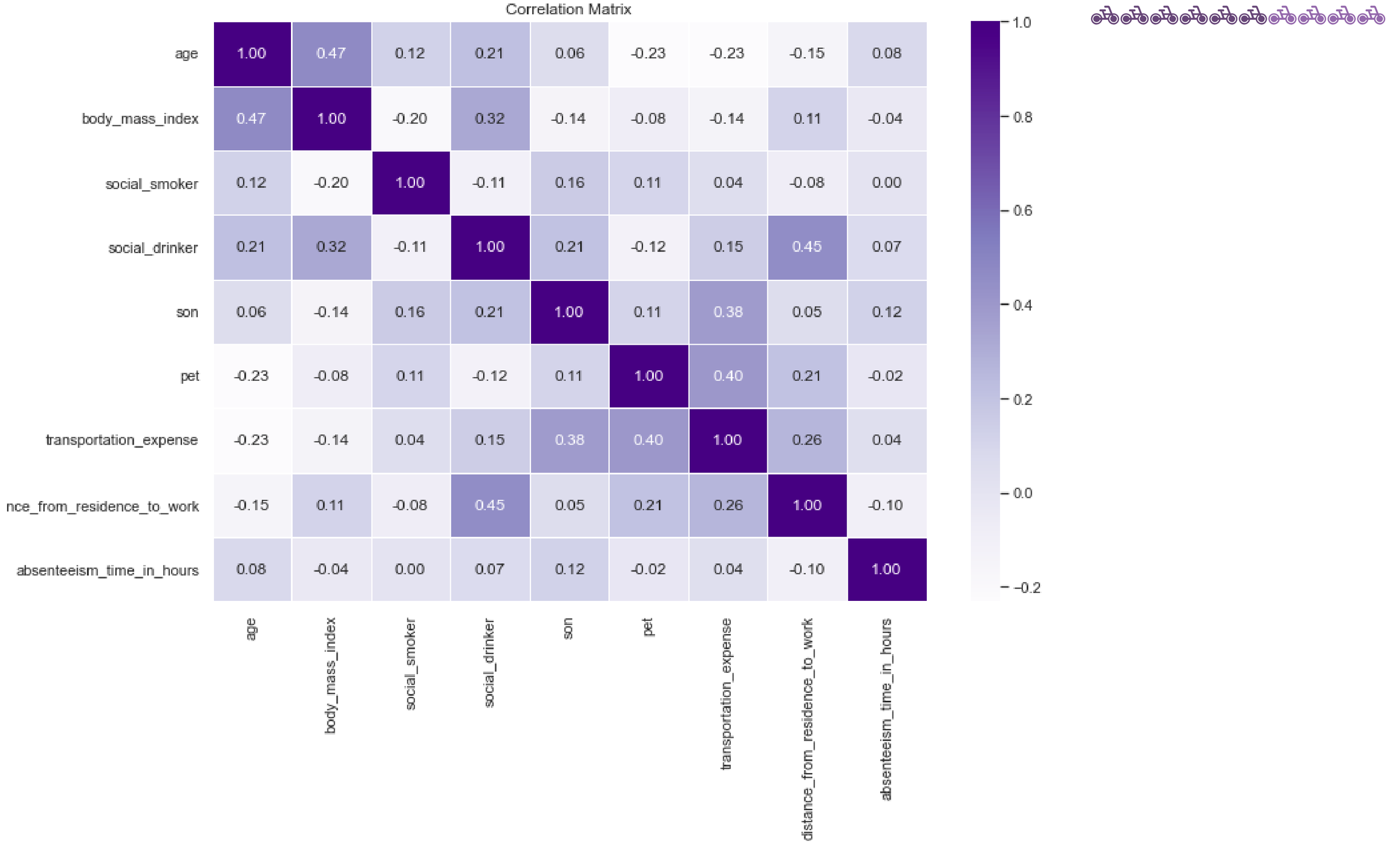
Month of absence is from January (1) to December (12). The month with the highest absenteeism time is **‘March’**, closely followed by **‘July’**. Conversely, ‘January’ has the lowest absenteeism time. This could be related to a fresh start in the new year, with employees generally being more present and engaged.

What effects does season, month and day of the week have on employee absenteeism?

Days are represented as Monday (2) to Friday (6). **‘Monday’** stands out as the day with the highest absenteeism time. This could be attributed to the "Monday blues" phenomenon, where employees might experience higher stress levels or challenges returning to work after the weekend.

What is the correlation between various employee demographics and absenteeism time?

```python
# Selecting the relevant columns for the correlation analysis
correlation_columns = ['age', 'body_mass_index','social_smoker','social_drinker','workload_average_per_day', 'son', 'pet',
'transportation_expense', 'distance_from_residence_to_work', 'absenteeism_time_in_hours']

# Creating a correlation matrix
correlation_matrix = merged_df[correlation_columns].corr()

# Displaying the correlation matrix
print("Correlation Matrix:")
print(correlation_matrix)

# Visualize the correlation matrix using a heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='Purples', fmt='.2f', linewidths=.5)
plt.title('Correlation Matrix')
plt.show()
```


Correlation Matrix
age 1.00 0.47 0.12 0.21 0.06 -0.23 -0.23 -0.15 0.08
body_mass_index 0.47 1.00 -0.20 0.32 -0.14 -0.08 -0.14 0.11 -0.04
social_smoker 0.12 -0.20 1.00 -0.11 0.16 0.11 0.04 -0.08 0.00
social_drinker 0.21 0.32 -0.11 1.00 0.21 -0.12 0.15 0.45 0.07
son 0.06 -0.14 0.16 0.21 1.00 0.11 0.38 0.05 0.12
pet -0.23 -0.08 0.11 -0.12 0.11 1.00 0.40 0.21 -0.02
transportation_expense -0.23 -0.14 0.04 0.15 0.38 0.40 1.00 0.26 0.04
nce_from_residence_to_work -0.15 0.11 -0.08 0.45 0.05 0.21 0.26 1.00 -0.10
absenteeism_time_in_hours 0.08 -0.04 0.00 0.07 0.12 -0.02 0.04 -0.10 1.00
age
body_mass_index
social_smoker
social_drinker
son
pet
transportation_expense
distance_from_residence_to_work
absenteeism_time_in_hours
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

What is the correlation between various employee demographics and absenteeism time?

Here's an interpretation of some key employee demographics and their correlation with the absenteeism time:

- There is a very **weak positive correlation** between ‘Age and Absenteeism time’ indicating almost no relationship.
- There is a very **weak negative correlation** between ‘Body mass index and Absenteeism time’, indicating almost no relationship.
- There is almost **no correlation** between ‘social smoker’ and Absenteeism time’.

What is the correlation between various employee demographics and absenteeism time?

- There is a very **weak positive correlation** between being a ‘Social drinker and Absenteeism time’, indicating almost no relationship.
- There is a **weak positive correlation** between the employee's number of children ‘(Son) and Absenteeism time’, suggesting a slight tendency for individuals with more children to have higher absenteeism time.
- There is almost **no correlation** between the employee's number of ‘Pets and Absenteeism time’.

What is the correlation between various employee demographics and absenteeism time?

- There is almost **no correlation** between ‘Transportation expenses of the employees and Absenteeism time’.
- There is a **‘weak negative correlation’** between the ‘Distance from residence to work and Absenteeism time’, indicating a slight tendency for individuals who live farther from work to have lower absenteeism time.

CONCLUSION

Diseases of the musculoskeletal system and connective tissue contributes significantly to the highest absenteeism hours. The nature of their work also involves **repetitive motions, exposure to external elements, and the potential for accidents,** contributing to health issues related to the musculoskeletal system and injuries.

The seasonal analysis identifies **winter** as the peak season for absenteeism, potentially influenced by the physically demanding nature of a courier job.

The **weak correlations** suggests that age, body mass index, and other work and lifestyle factors have **minimal impact on absenteeism**

RECOMMENDATION

To address this, the company might consider implementing measures to enhance employee well-being, such as **ergonomic improvements**, **health and safety programs**, or **periodic health check-ups**.

Additionally, providing adequate support and resources for employees in such physically demanding roles could contribute to a healthier and more productive work environment with less absenteeism time.