Data Mining Assignment

GROUP: P1DM6

Business Context & Problem Statement

Context

Important features of Amazing Zone's business model are

- Various products sold online
- Retailer takes responsibility of the shipment of the product
- Products delivered within promised timeline increase happiness among the customer
- Customers are more inclined to purchase the products that has "Retailer fulfilled" tag associated with them

Problem Statement

The company Amazing Zone is facing issues with deliveries due to frequent absenteeism of the delivery personnel.

Due to heavy workload these delivery personnel are not able to perform their optimum limits, which results into their absenteeism from the work, which in turn increases the load on other delivery personnels.

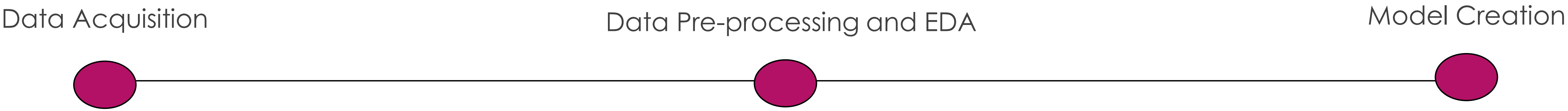
Opportunity Identifications

In order to improve productivity, we need to take a close look at absenteeism record of delivery personnel's, we must

- **Identify the factors causing the absenteeism at work**
- **Suggest a model that can help to determine the absenteeism hours for an employee**
- **Suggest ways by which absenteeism can be reduced**

S-3

Data Understanding & Preparations


Data Acquisition

Data Pre-processing and EDA

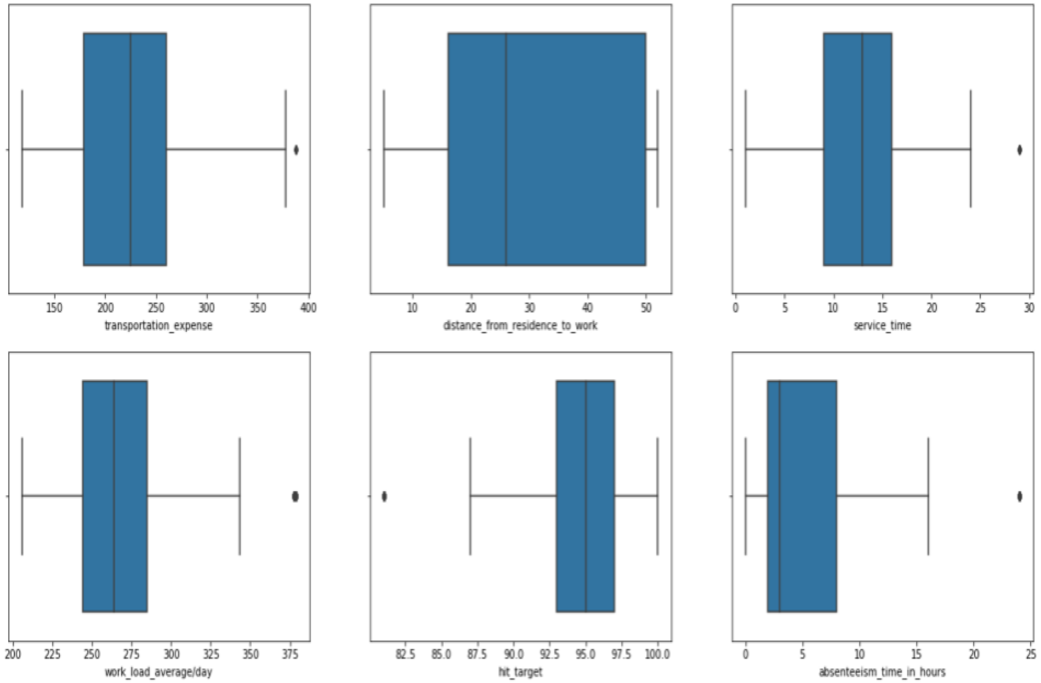
Model Creation

- **Data Acquisition** – Reading the data from CSV file and creating a dataframe
- **Data Pre-Processing and EDA**
  - Check for null values
  - Identify and remove outliers
  - Encoding categorical features
  - Scaling
  - EDA with bar graphs, heatmap and pairplots
- **Model Creation**
  - Splitting the data into training and test data
  - Linear Regression
  - Lasso
  - Model evaluation RMSE and R2

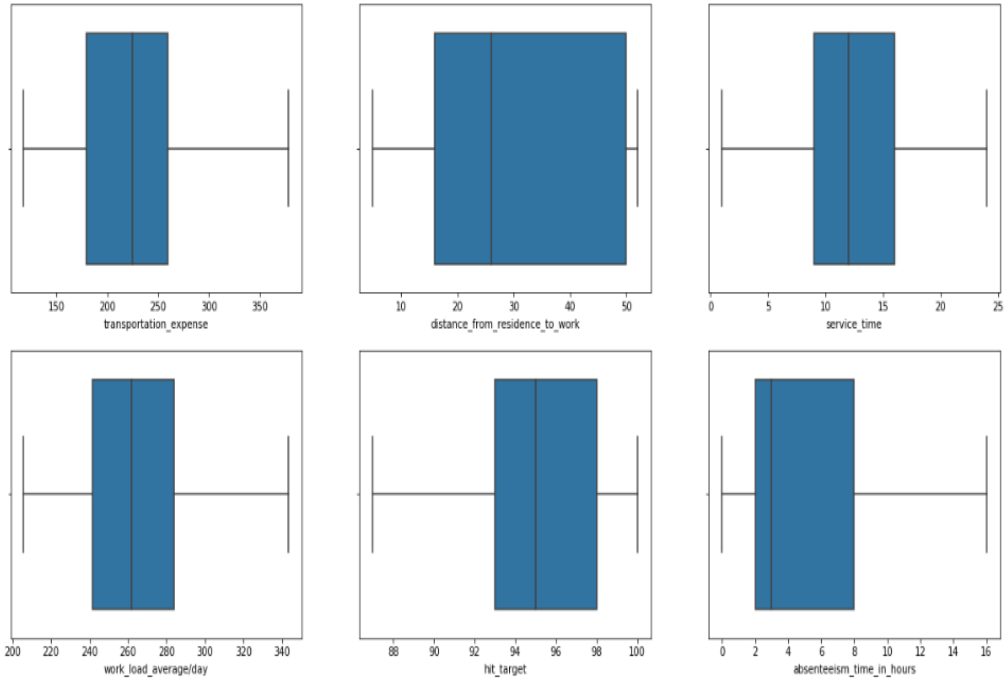
Data Understanding & Preparations

Outlier Analysis

Boxplot on Raw data

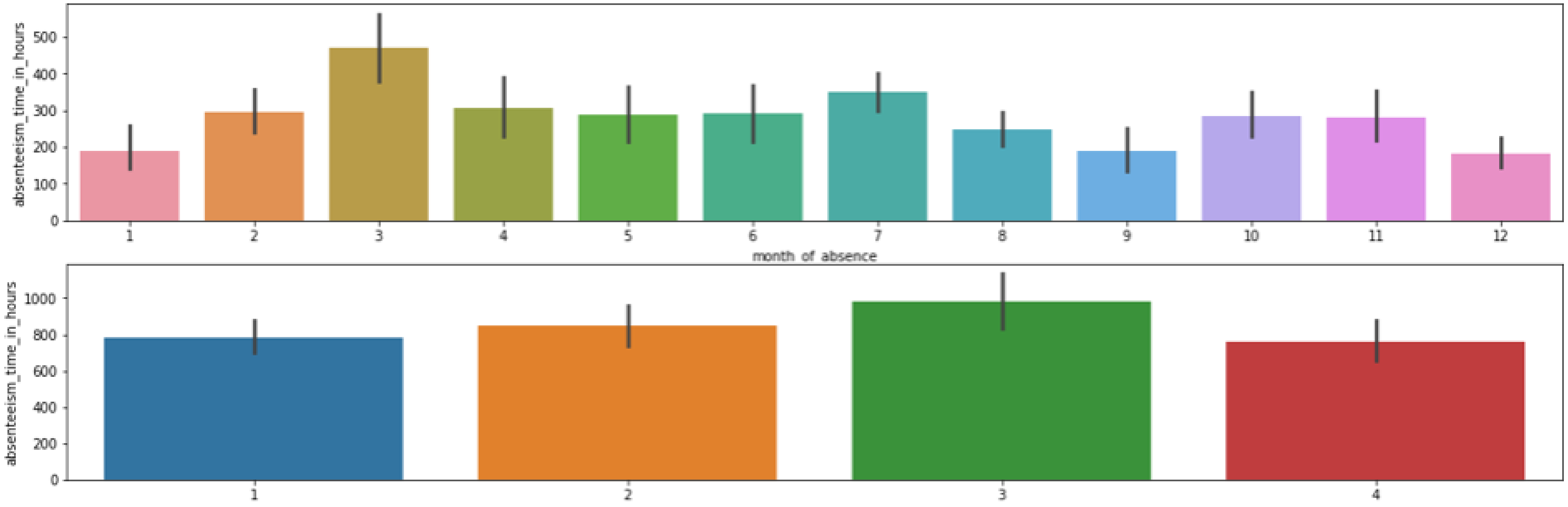


Boxplot post removing outliers



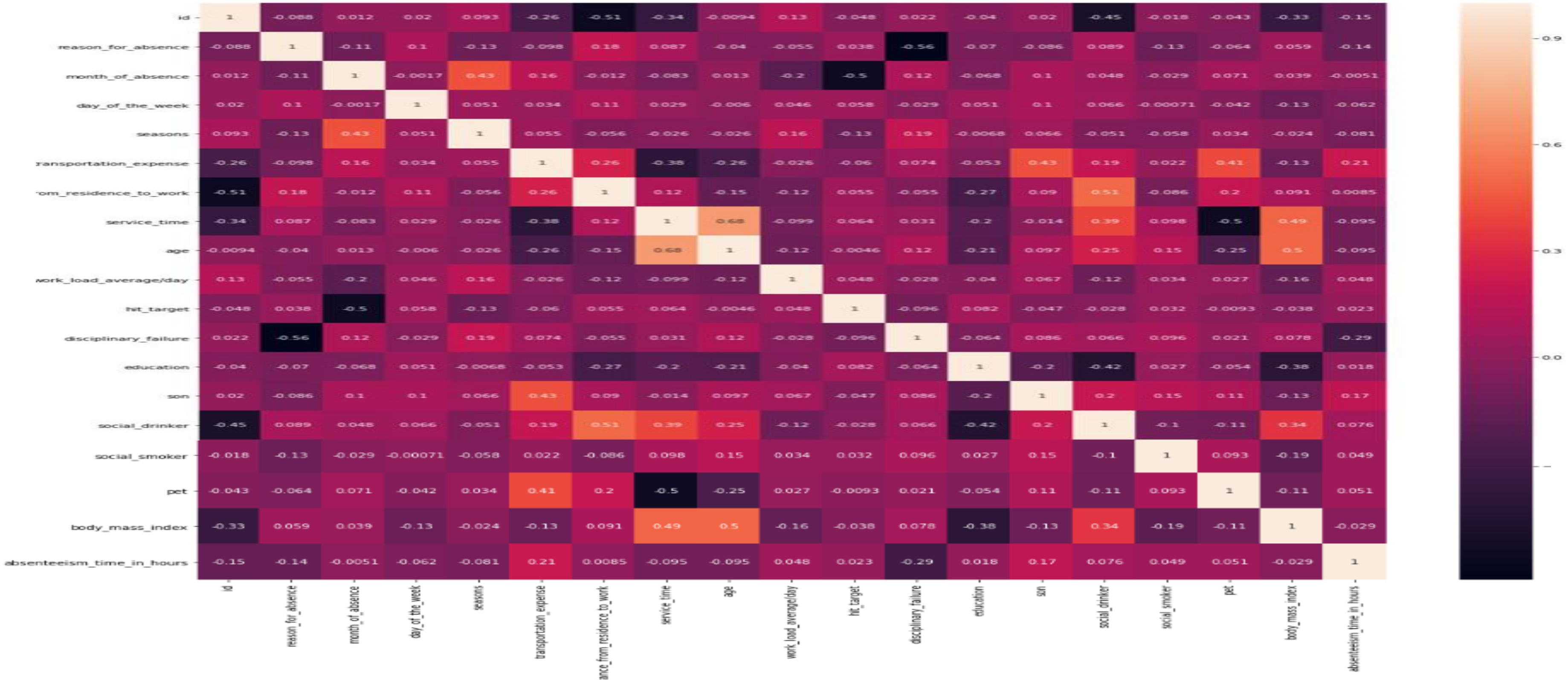
EDA

Data Visualization with Bar Graph



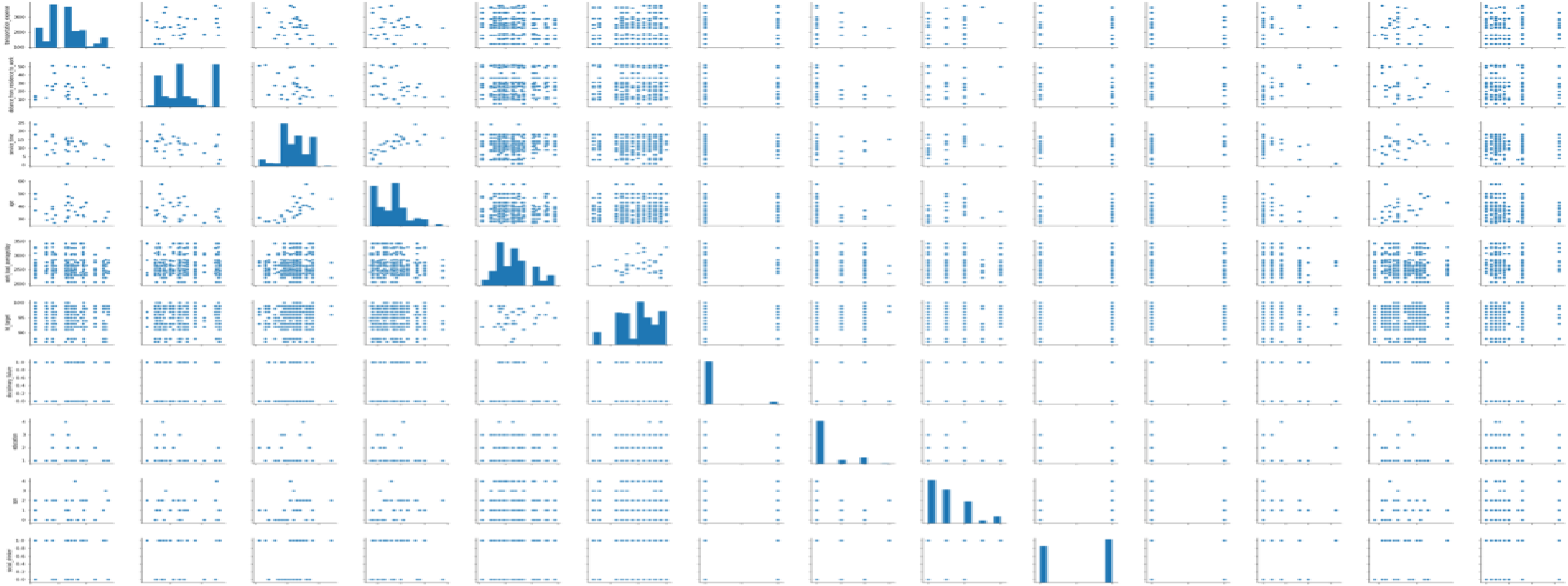
EDA

Heatmap


id 1 -0.088 0.012 0.02 0.093 -0.26 -0.51 -0.34 -0.0094 0.13 -0.048 0.022 -0.04 0.02 -0.45 -0.018 -0.043 -0.33 -0.15
reason_for_absence -0.088 1 -0.11 0.1 -0.13 -0.098 0.18 0.087 -0.04 -0.055 0.038 -0.56 -0.07 -0.086 0.089 -0.13 -0.064 0.059 -0.14
month_of_absence 0.012 -0.11 1 -0.0017 0.43 0.16 -0.012 -0.083 0.013 -0.2 -0.5 0.12 -0.068 0.1 0.048 -0.029 0.071 0.039 -0.0051
day_of_the_week 0.02 0.1 -0.0017 1 0.051 0.034 0.11 0.029 -0.006 0.046 0.058 -0.029 0.051 0.1 0.066 -0.00071 -0.042 -0.13 -0.062
seasons 0.093 -0.13 0.43 0.051 1 0.055 -0.056 -0.026 -0.026 0.16 -0.13 0.19 -0.0068 0.066 -0.051 -0.058 0.034 -0.024 -0.081
ransportation_expense -0.26 -0.098 0.16 0.034 0.055 1 0.26 -0.38 -0.26 -0.026 -0.06 0.074 -0.053 0.43 0.19 0.022 0.41 -0.13 0.21
rom_residence_to_work -0.51 0.18 -0.012 0.11 -0.056 0.26 1 0.12 -0.15 -0.12 0.055 -0.055 -0.27 0.09 0.51 -0.086 0.2 0.091 0.0085
service_time -0.34 0.087 -0.083 0.029 -0.026 -0.38 0.12 1 0.68 -0.099 0.064 0.031 -0.2 -0.014 0.39 0.098 -0.5 0.49 -0.095
age -0.0094 -0.04 0.013 -0.006 -0.026 -0.26 -0.15 0.68 1 -0.12 -0.0046 0.12 -0.21 0.097 0.25 0.15 -0.25 0.5 -0.095
vork_load_average/day 0.13 -0.055 -0.2 0.046 0.16 -0.026 -0.12 -0.099 -0.12 1 0.048 -0.028 -0.04 0.067 -0.12 0.034 0.027 -0.16 0.048
hit_target -0.048 0.038 -0.5 0.058 -0.13 -0.06 0.055 0.064 -0.0046 0.048 1 -0.096 0.082 -0.047 -0.028 0.032 -0.0093 -0.038 0.023
disciplinary_failure 0.022 -0.56 0.12 -0.029 0.19 0.074 -0.055 0.031 0.12 -0.028 -0.096 1 -0.064 0.086 0.066 0.096 0.021 0.078 -0.29
education -0.04 -0.07 -0.068 0.051 -0.0068 -0.053 -0.27 -0.2 -0.21 -0.04 0.082 -0.064 1 -0.2 -0.42 0.027 -0.054 -0.38 0.018
son 0.02 -0.086 0.1 0.1 0.066 0.43 0.09 -0.014 0.097 0.067 -0.047 0.086 -0.2 1 0.2 0.15 0.11 -0.13 0.17
social_drinker -0.45 0.089 0.048 0.066 -0.051 0.19 0.51 0.39 0.25 -0.12 -0.028 0.066 -0.42 0.2 1 -0.1 -0.11 0.34 0.076
social_smoker -0.018 -0.13 -0.029 -0.00071 -0.058 0.022 -0.086 0.098 0.15 0.034 0.032 0.096 0.027 0.15 -0.1 1 0.093 -0.19 0.049
pet -0.043 -0.064 0.071 -0.042 0.034 0.41 0.2 -0.5 -0.25 0.027 -0.0093 0.021 -0.054 0.11 -0.11 0.093 1 -0.11 0.051
body_mass_index -0.33 0.059 0.039 -0.13 -0.024 -0.13 0.091 0.49 0.5 -0.16 -0.038 0.078 -0.38 -0.13 0.34 -0.19 -0.11 1 -0.029
absenteeism_time_in_hours -0.15 -0.14 -0.0051 -0.062 -0.081 0.21 0.0085 -0.095 -0.095 0.048 0.023 -0.29 0.018 0.17 0.076 0.049 0.051 -0.029 1
id
reason_for_absence
month_of_absence
day_of_the_week
seasons
transportation_expense
ance_from_residence_to_work
service_time
age
work_load_average/day
hit_target
disciplinary_failure
education
son
social_drinker
social_smoker
pet
body_mass_index
absenteeism_time_in_hours
0.9
0.6
0.3
0.0

EDA

Pair plots



Data Preparation

Encoding Categorical Data

One Hot encoding has been performed on categorical features of the data to make it viable for analysis.

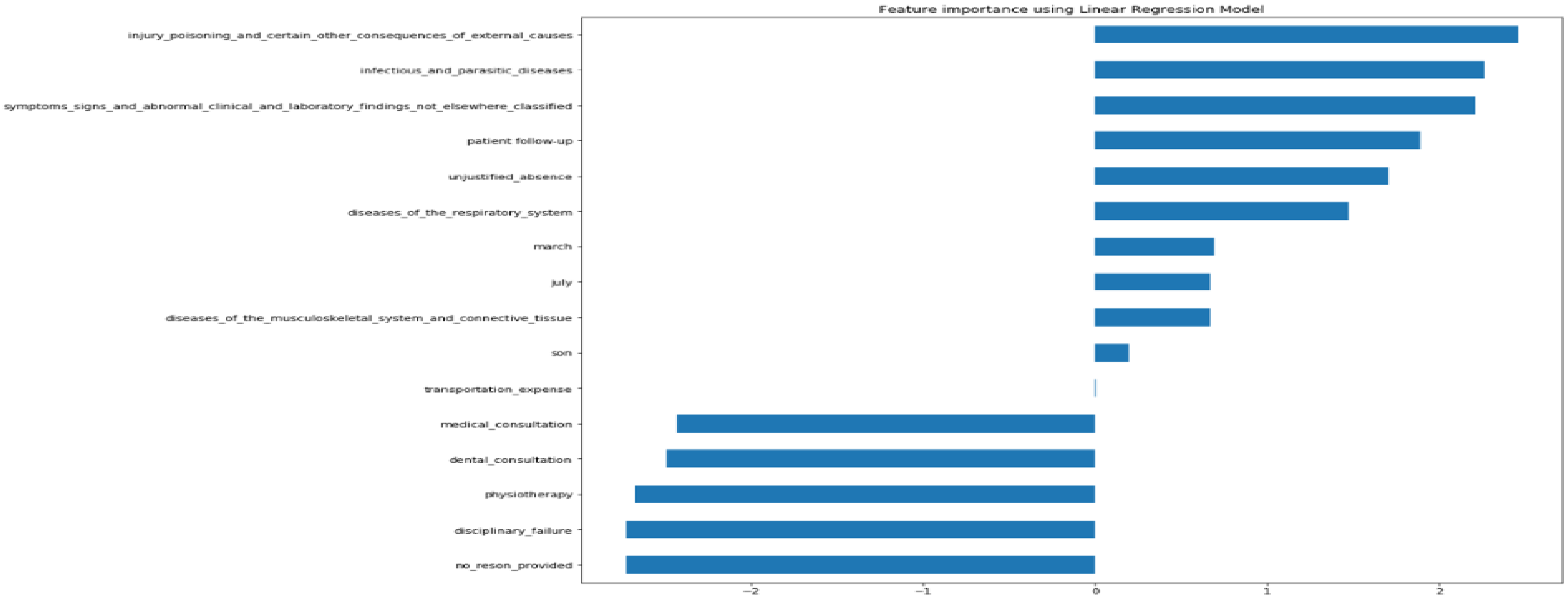
Encoding performed on below columns

- Seasons
- Reason for absence
- Month of absence
- Day of week

Feature selection and model creation

- Calculating correlation coefficient for all the features
- Selecting features with correlation coefficient greater than +/- 0.1
- Splitting the data into dependent and independent features
- Implemented Linear regression model to predict absenteeism
- Implemented LASSO (Least Absolute Shrinkage Selector Operator) Model
- Calculated "Root Mean Square Error" to evaluate the accuracy of the model
- Calculated R2 Score to evaluate strength of correlation between dependent and independent variables
- Displayed linear regression coefficient & intercept
- Plotted feature importance with linear regression model

Feature selection and model creation



Suggestions for Improvement

- Provide periodic medical consultation to the delivery persons which will reduce the occurrences of ailments
- Provide physiotherapy sessions which will not only reduce stress but increase productivity
- Provide transportation or reimburse transportation expense
- Provide periodic dental consultation