HYPOTHESIS TESTING PROJECT

THANH PHONG, HO (PARKER)

INSTITUTE OF DATA - 26 OCTOBER 2021

BUSINESS PROBLEM:

The problem is that the work environment of today is more competitive, businesses want to:

Reduce the cost of employee's absenteeism

Achieve business goals.

Provide better healthcare for employees

BUSINESS CONTEXT:

To understand whether employees with certain characteristics are expected to be away from work at some point in time or not.

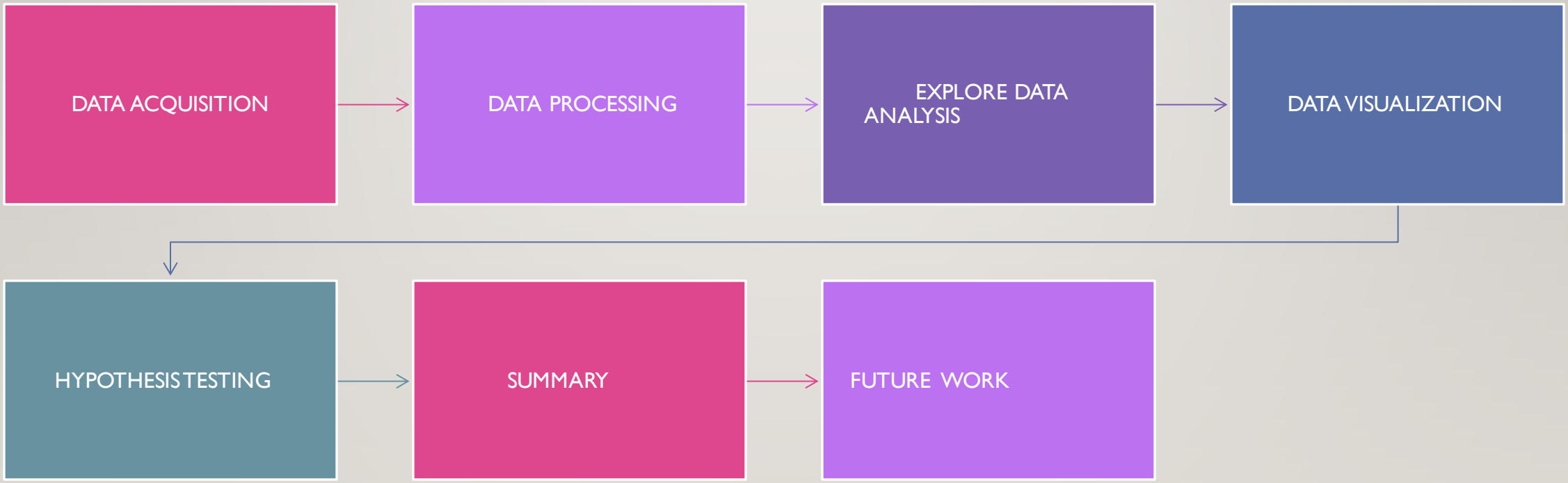
We want to know how many working hours any employee could be away from work based on information like:

How far they live from their workplace.

Will the expense of transportation affect their absenteeism?

What are the reasons for absenteeism?

DATA PIPELINE



DATA OVERVIEW

Data was collected from 2015 July 06 to 2018 May 31

700 observations

12 features

No missing values

1 object feature

10 integer features

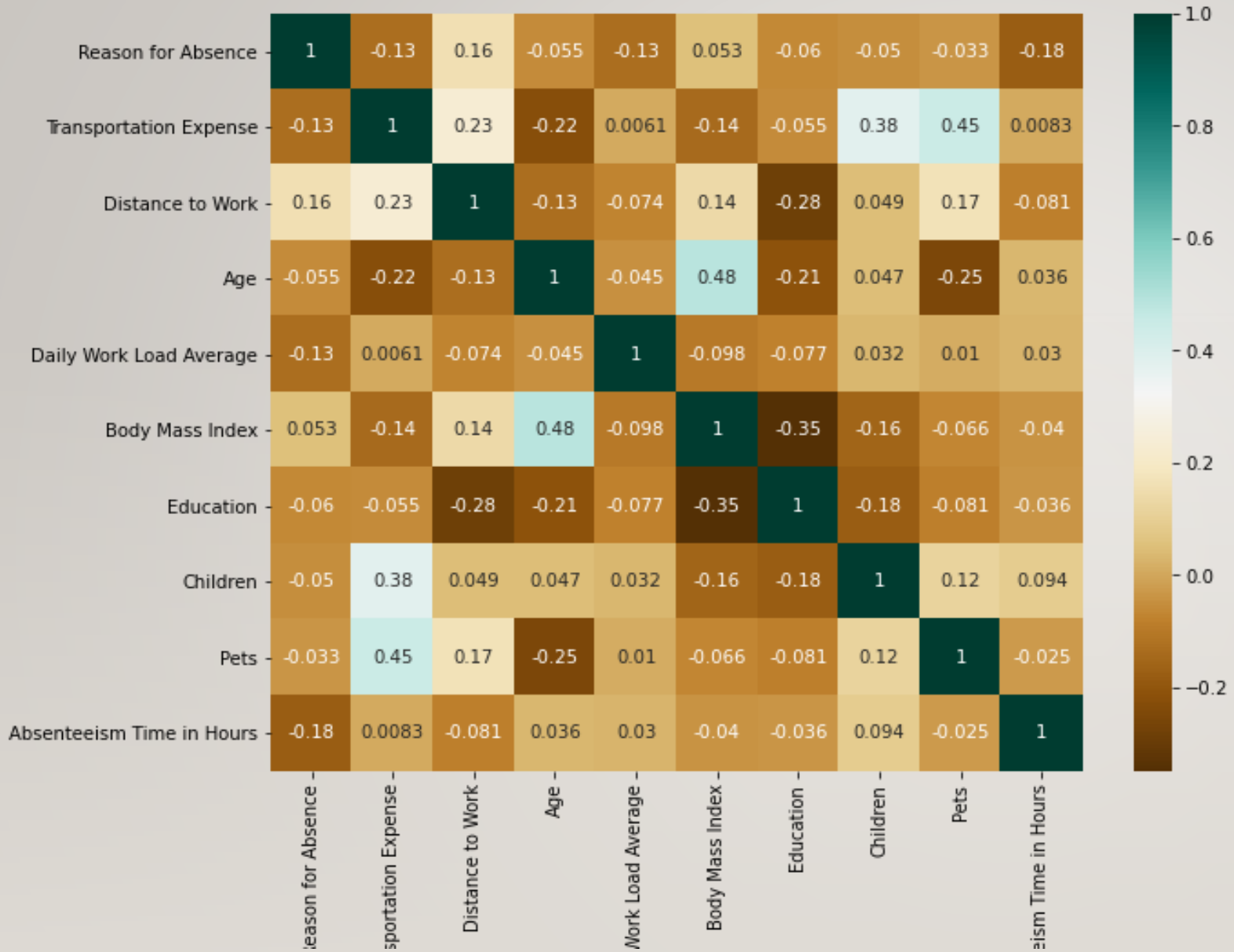
1 float feature

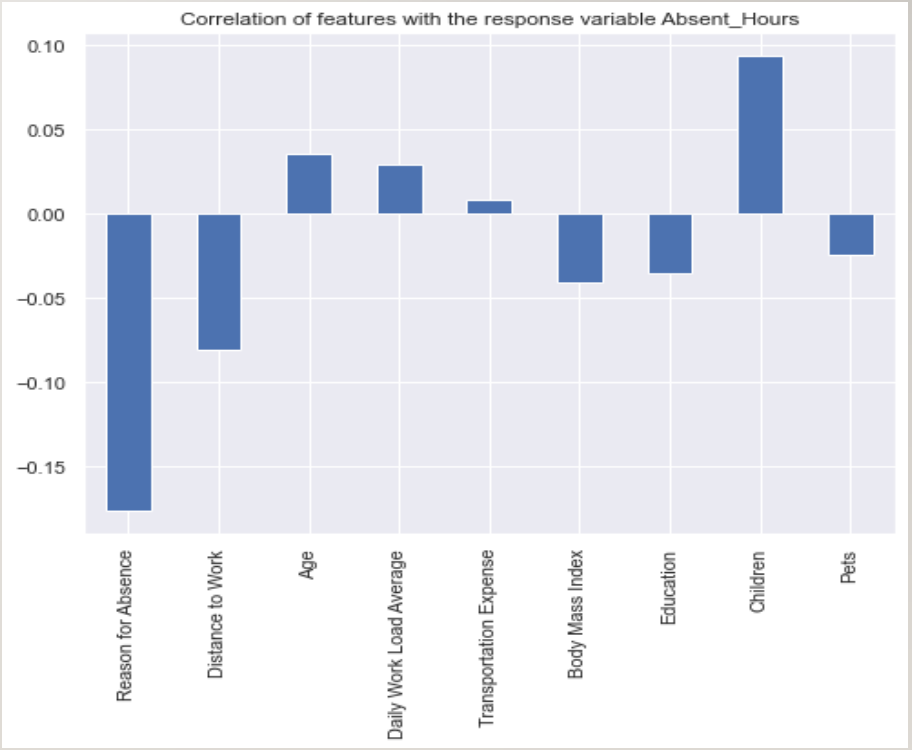
12 FEATURES

- ID
- Reason for Absence
- Date
- Transportation Expense
- Distance to Work
- Age
- Daily Workload Average
- Body Mass Index
- Children
- Education
- Pets
- Absenteeism Time in Hours

Feature	Description
1	ID: Individual Identification
2	**Reason for Absence:** (Reasons 1-21 are registered in the International Classification of Diseases (ICU), reasons 22-28 are not) 1: Certain infectious and parasitic diseases 2: Neoplasms 3: Disease of the blood and blooding-forming organs and certain disorders involving in the immune mechanism 4: Endocrine, nutritional and metabolic diseases 5: Mental and behavioral disorders 6: Diseases of the nervous system 7: Diseases of the eye and adnexa 8: Diseases of the ear and mastoid process 9: Diseases of the circulatory system 10: Diseases of the respiratory system 11: Diseases of the digestive system 12: Diseases of the skin and subcutaneous tissue 13: Diseases of the musculoskeletal and connective tissue 14: Diseases of the genitourinary system 15: Pregnancy, childbirth and puerperium 16: Certain conditions originating in the perinatal period 17: Congenital malformations, deformations and chromosomal abnormalities 18: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified. 19: Injury, poisoning and certain other consequences of external causes 20: External causes of morbidity and mortality 21: Factors influencing health status and contact with health services 22: Patient follow-up 23: Medical consultation 24: Blood donation 25: Laboratory examination 26: Unjustified absence 27: Physiotherapy 28: Dental consultation
3	**Date**: date of absence
4	**Transportation Expense**: costs related to business travel such as fuel, parking and meals
5	**Distance to Work**: measured in kilometers
6	**Age**: years of age
7	**Daily Workload Average**: measured in minutes
8	**Body Mass Index**
9	**Education:** a categorical variable and presenting different levels of education
10	**Children:** number of children in the family
11	**Pets**: number of pets in the family
12	**Absenteeism Time in Hours:**

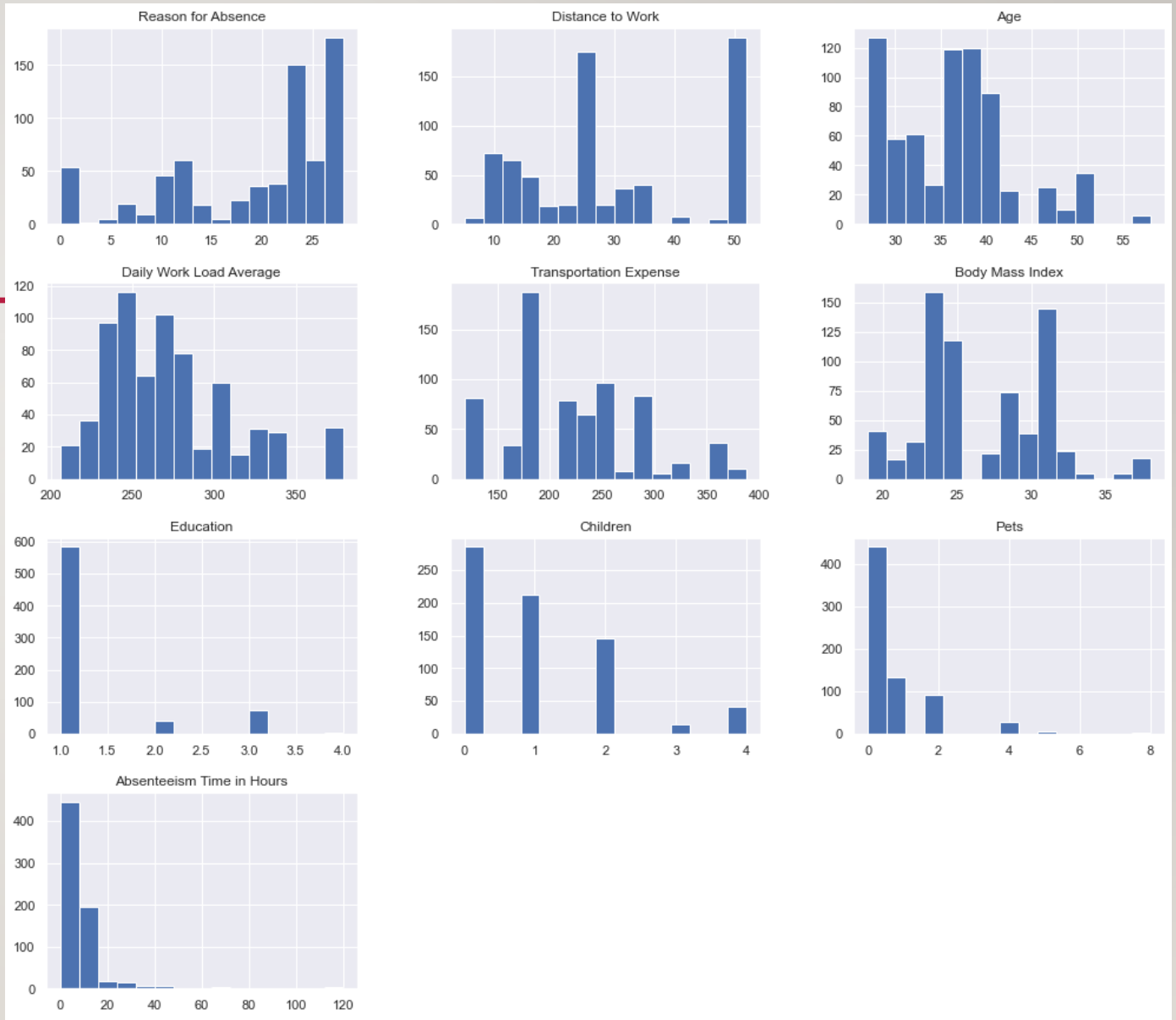
DATA CORRELATION

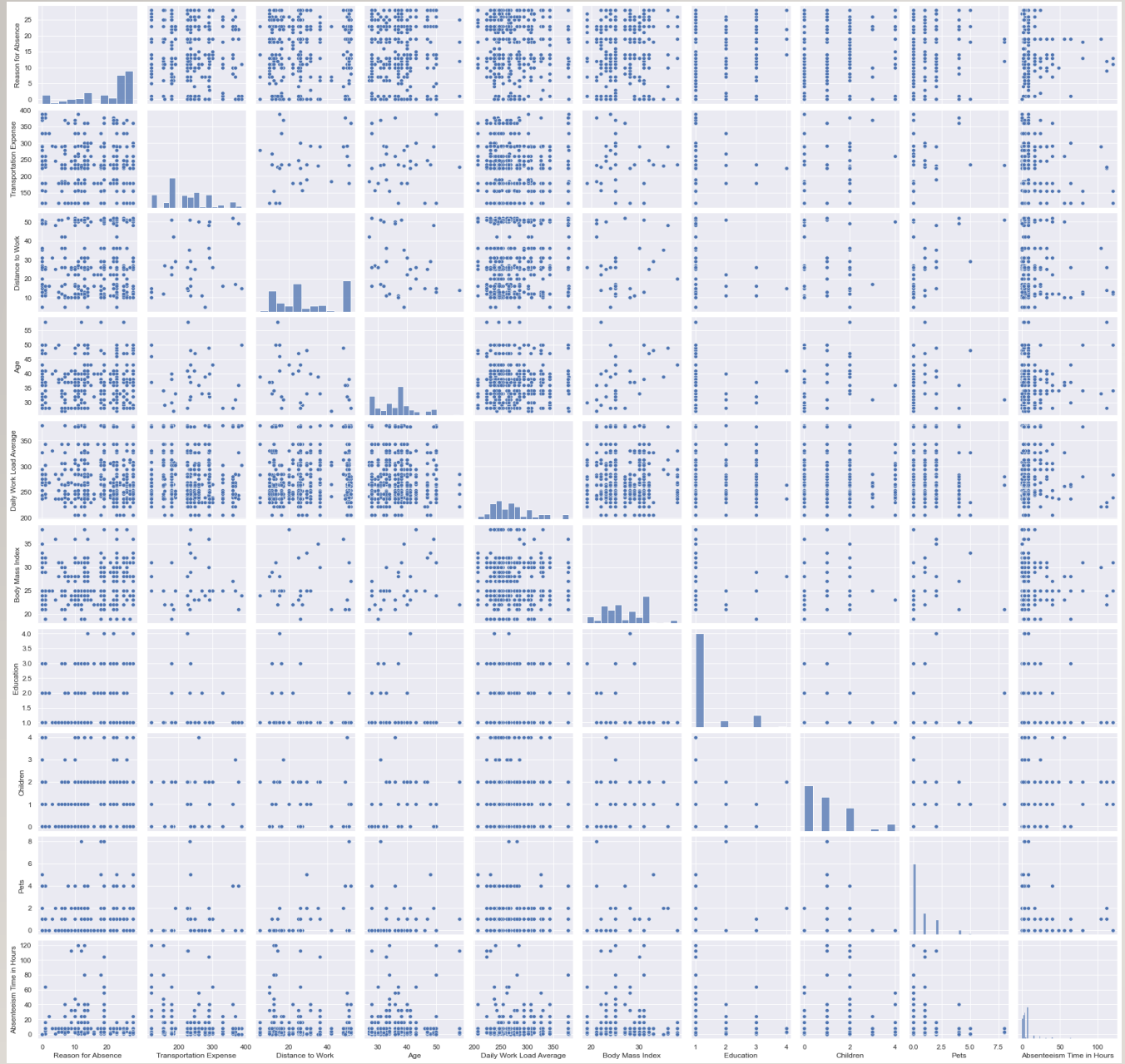




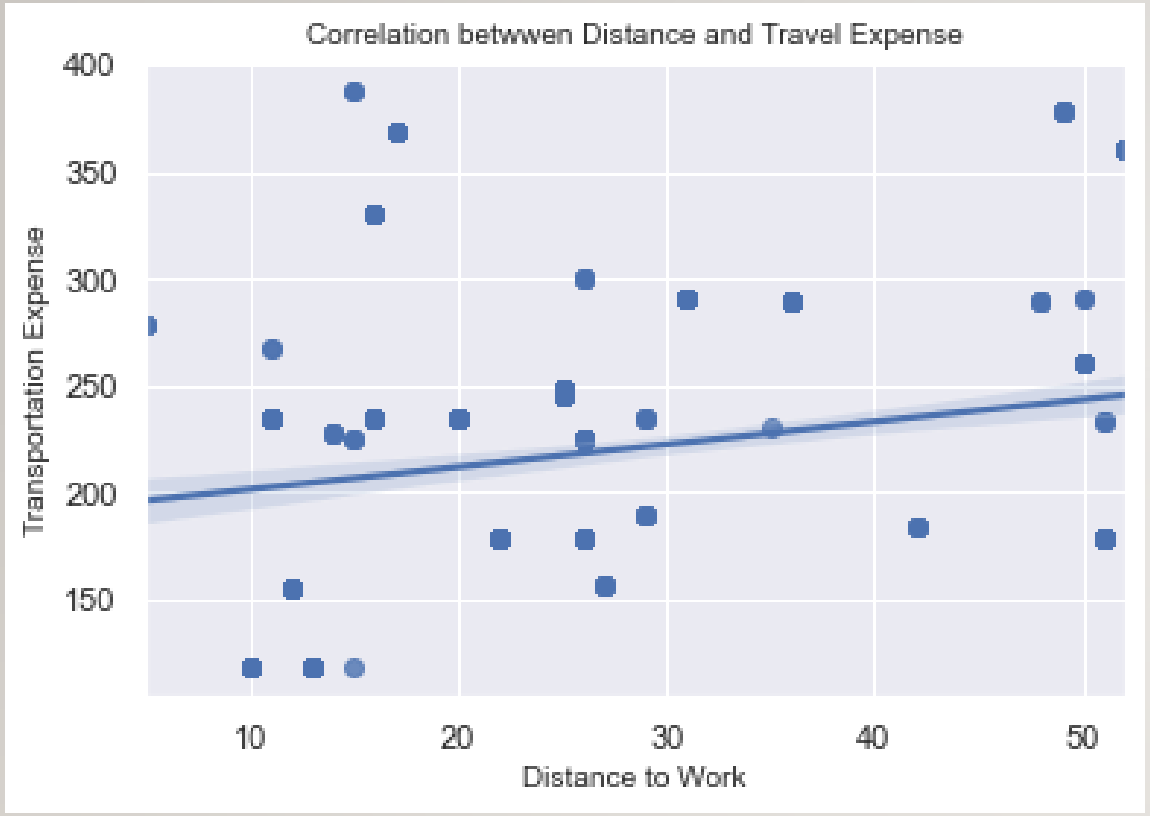
DISTRIBUTION OF NUMERIC FEATURES

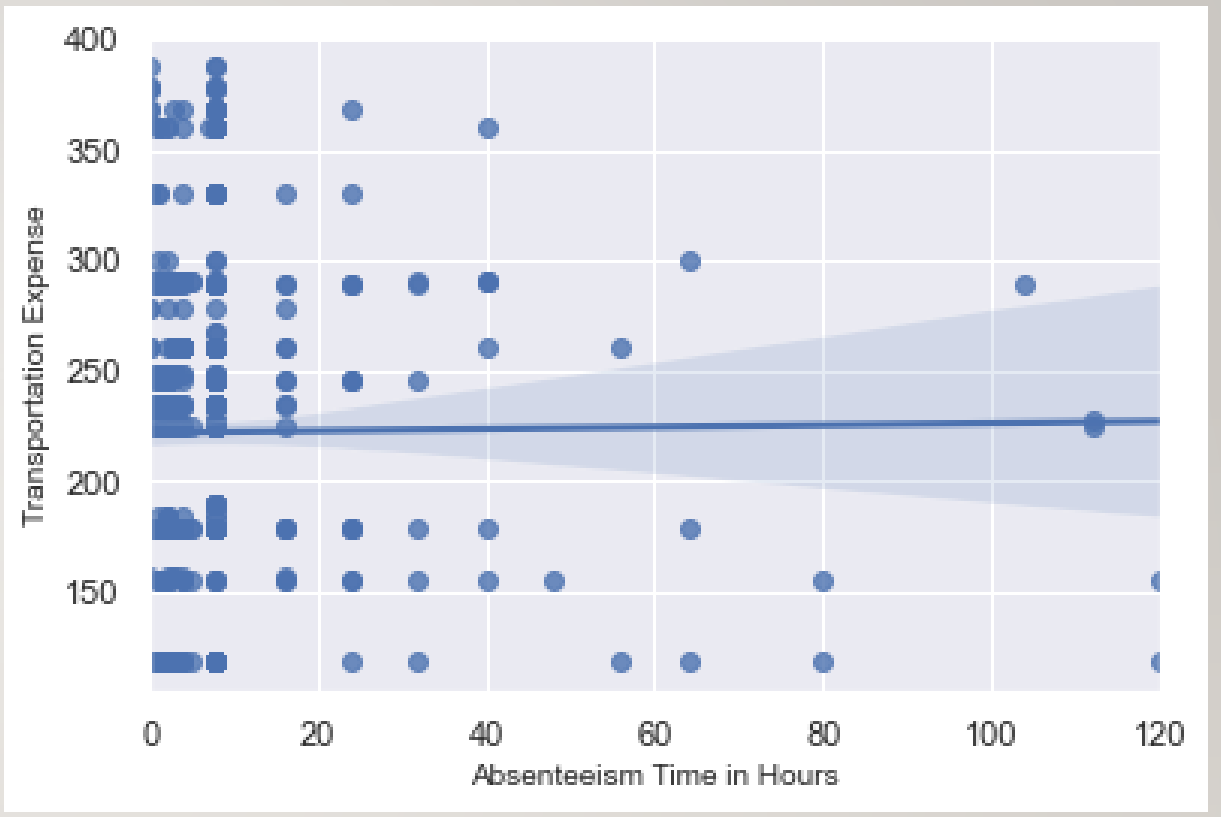
- Click to add text





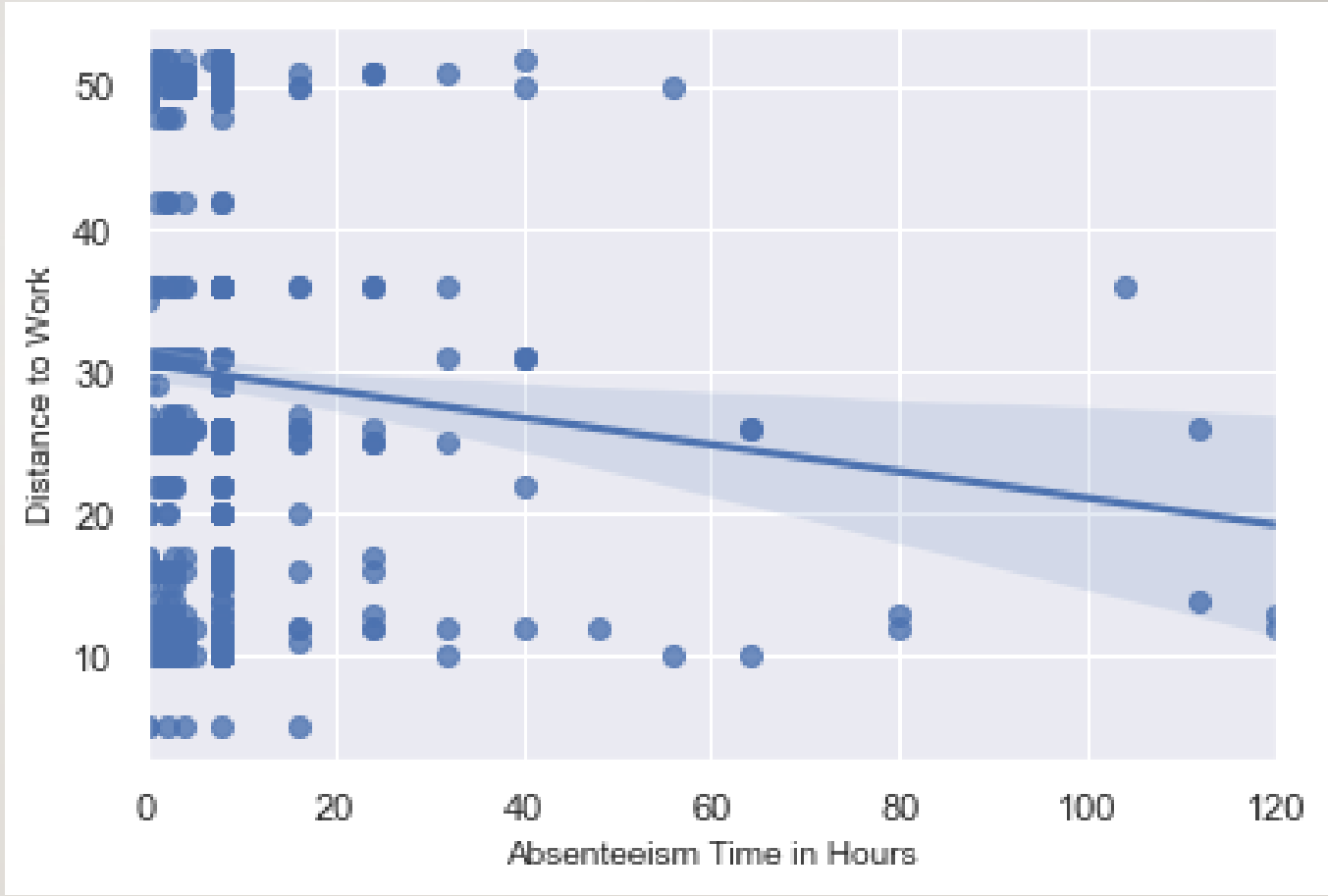
PAIRPLOT SHOWING CORRELATION ALL NUMERIC FEATURES

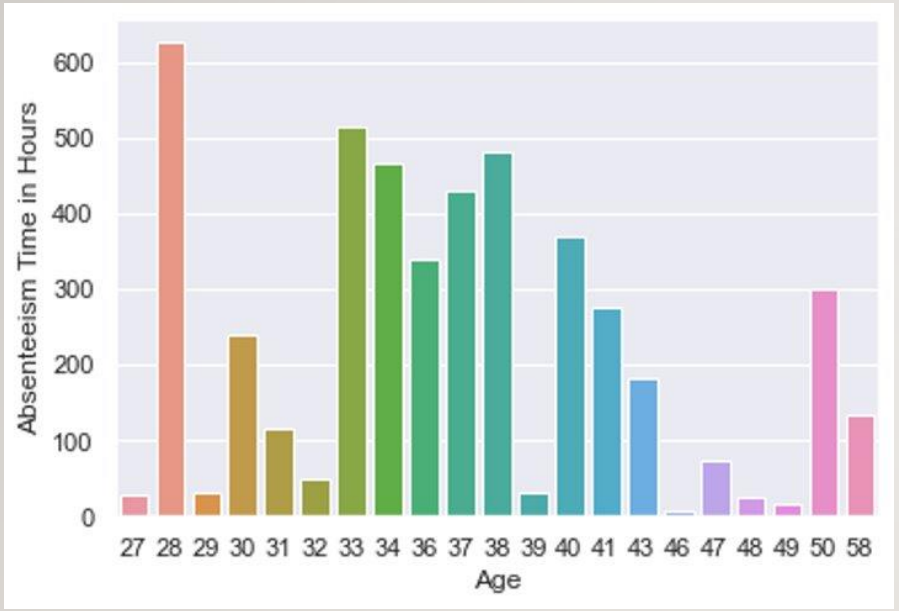

Correlation betwwen Distance and Travel Expense
Transportation Expense
400
350
300
250
200
150
10
20
30
40
50
Distance to Work

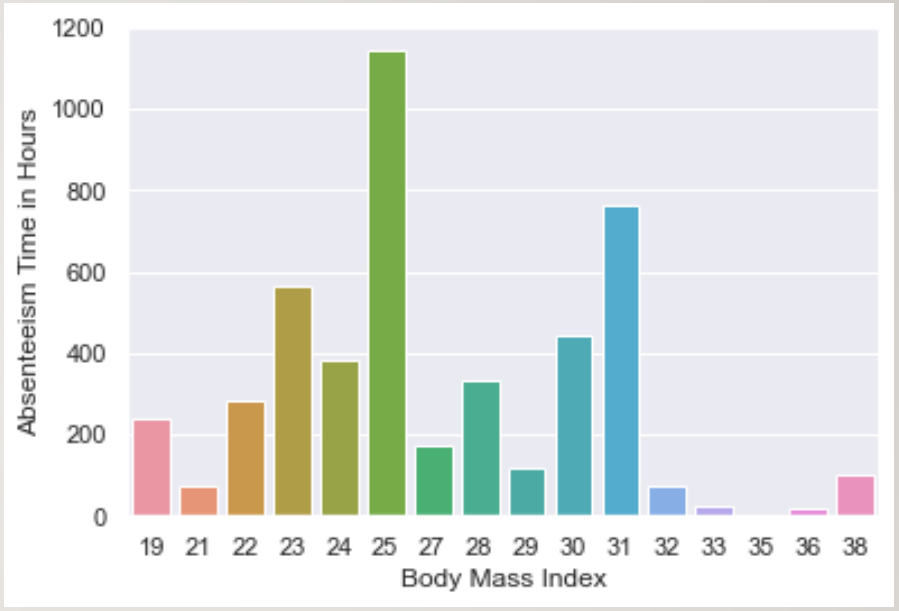

Transportation Expense
400
350
300
250
200
150
0
20
40
60
80
100
120
Absenteeism Time in Hours

EDA

EDA

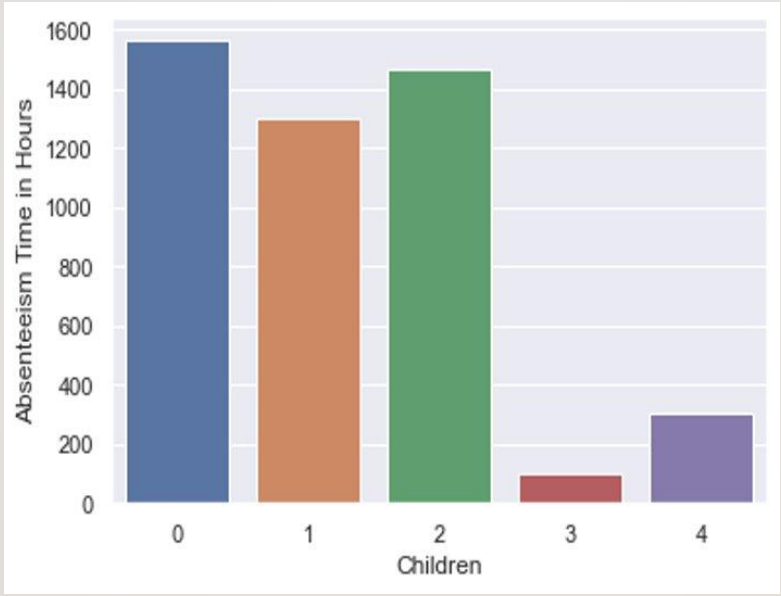

Distance to Work
50
40
30
20
10
0
20
40
60
80
100
120
Absenteeism Time in Hours

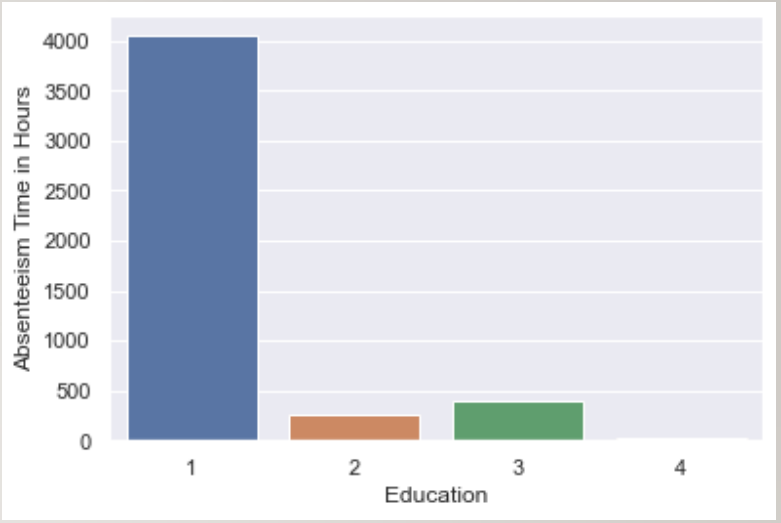

Absenteeism Time in Hours
0
100
200
300
400
500
600
27 28 29 30 31 32 33 34 36 37 38 39 40 41 43 46 47 48 49 50 58
Age

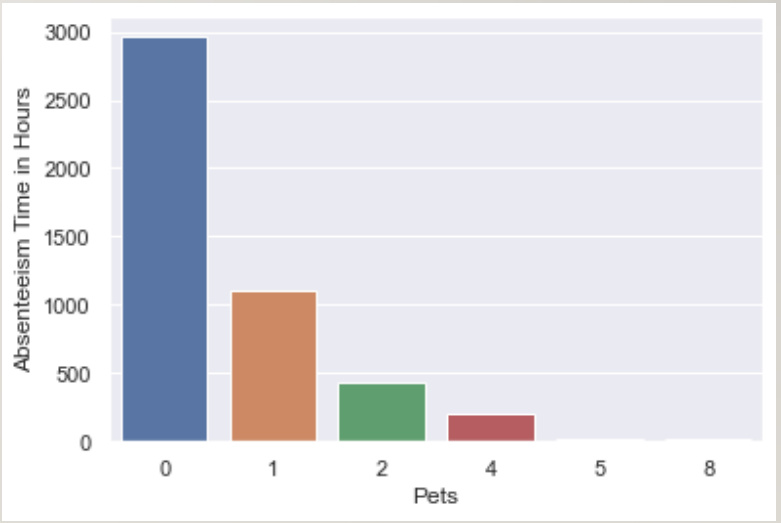

Absenteeism Time in Hours
0
200
400
600
800
1000
1200
19 21 22 23 24 25 27 28 29 30 31 32 33 35 36 38
Body Mass Index

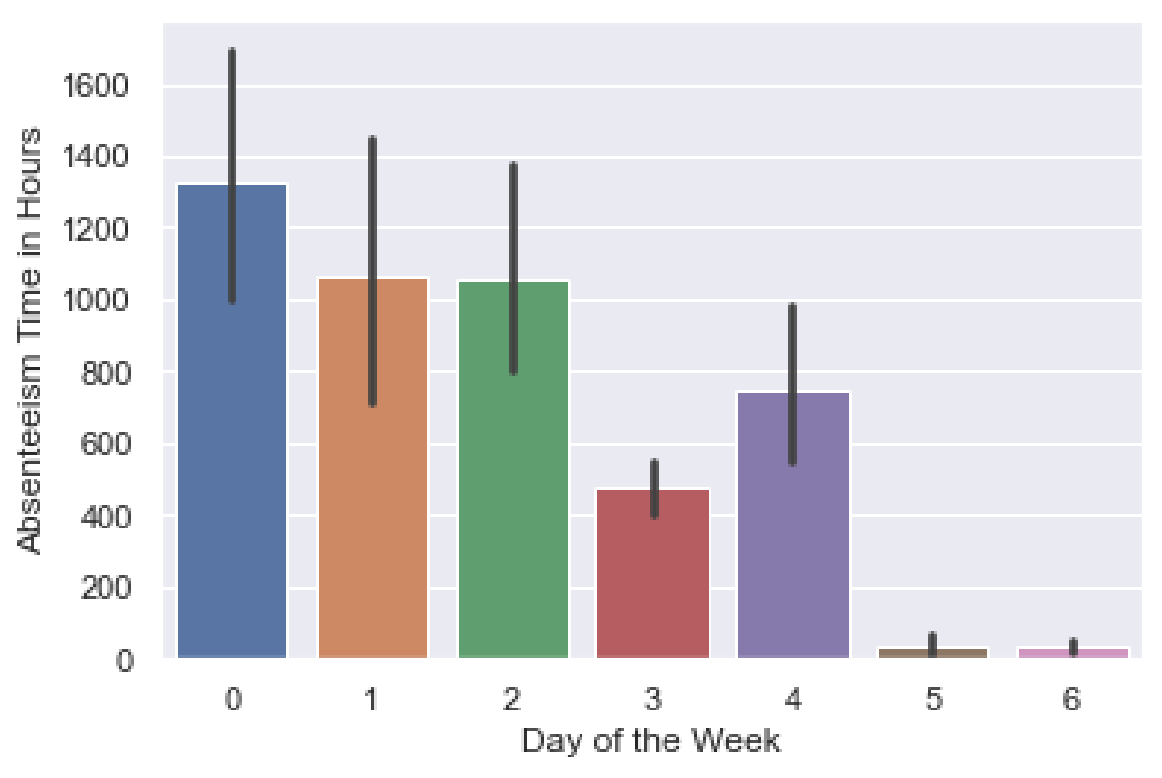
EDA

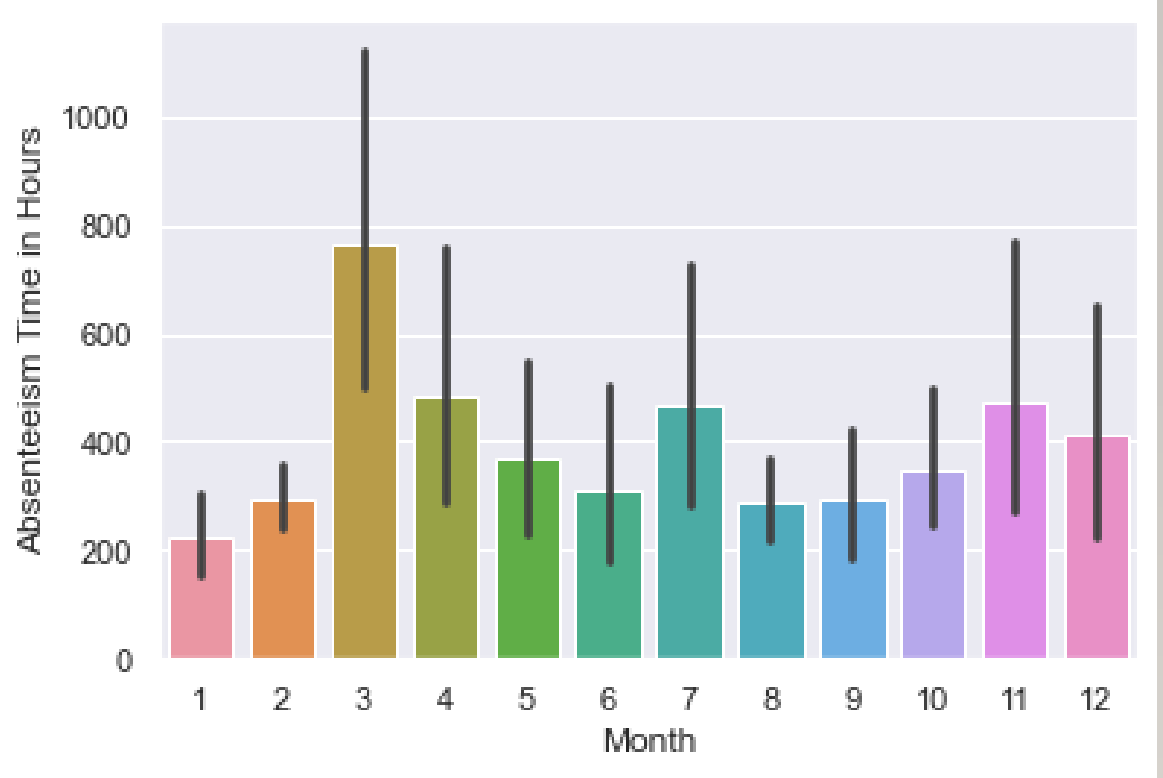
EDA


Absenteeism Time in Hours
1600
1400
1200
1000
800
600
400
200
0
0
1
2
3
4
Children

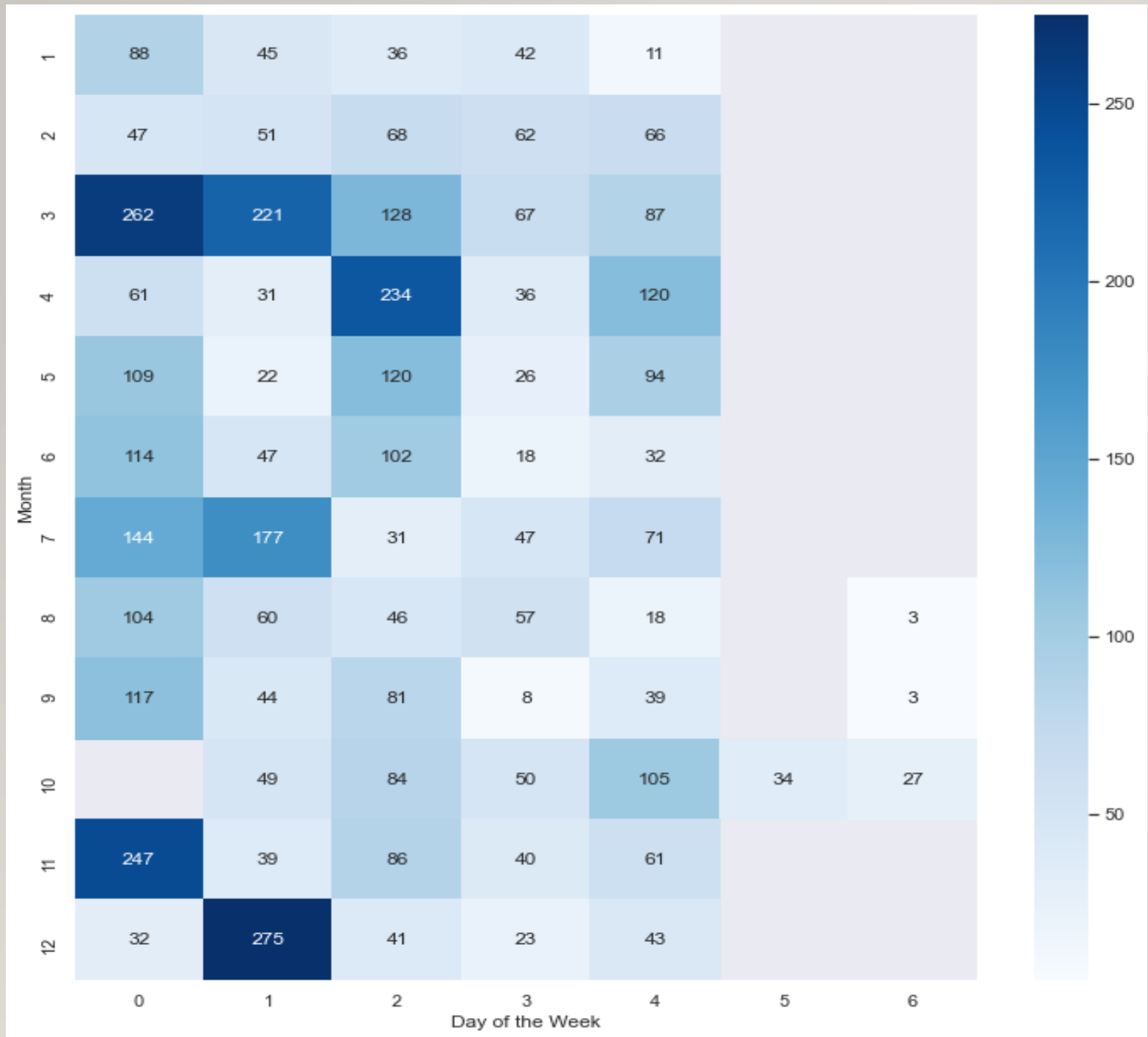

Absenteeism Time in Hours
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
Education


Absenteeism Time in Hours
3000
2500
2000
1500
1000
500
0
0
1
2
4
5
8
Pets


Absenteeism Time in Hours
1600
1400
1200
1000
800
600
400
200
0
0
1
2
3
4
5
6
Day of the Week

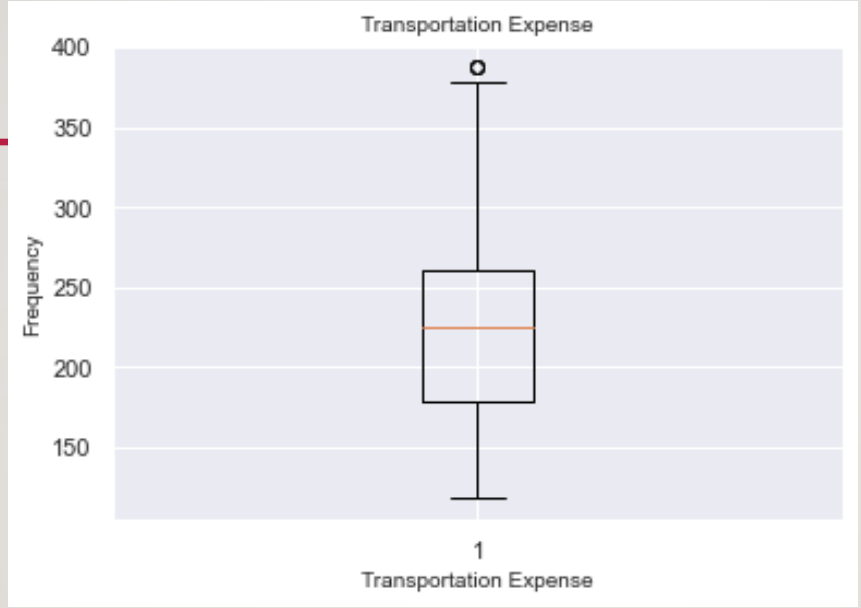

Absenteeism Time in Hours
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
11
12
Month

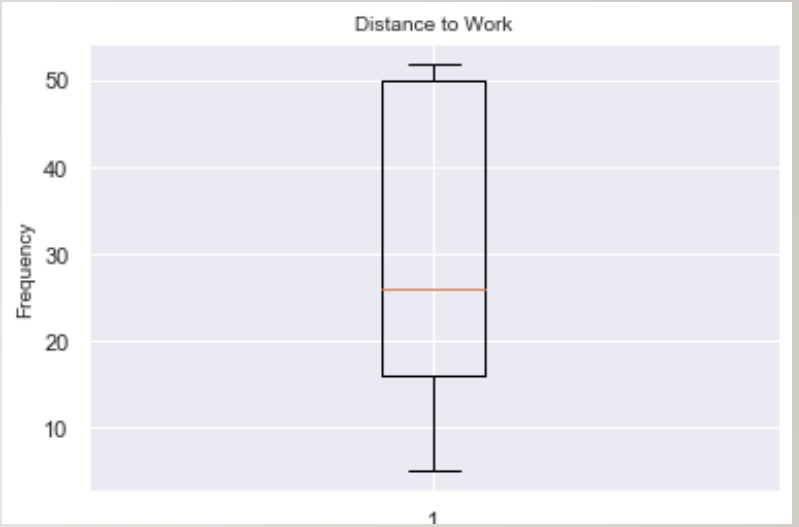
EDA



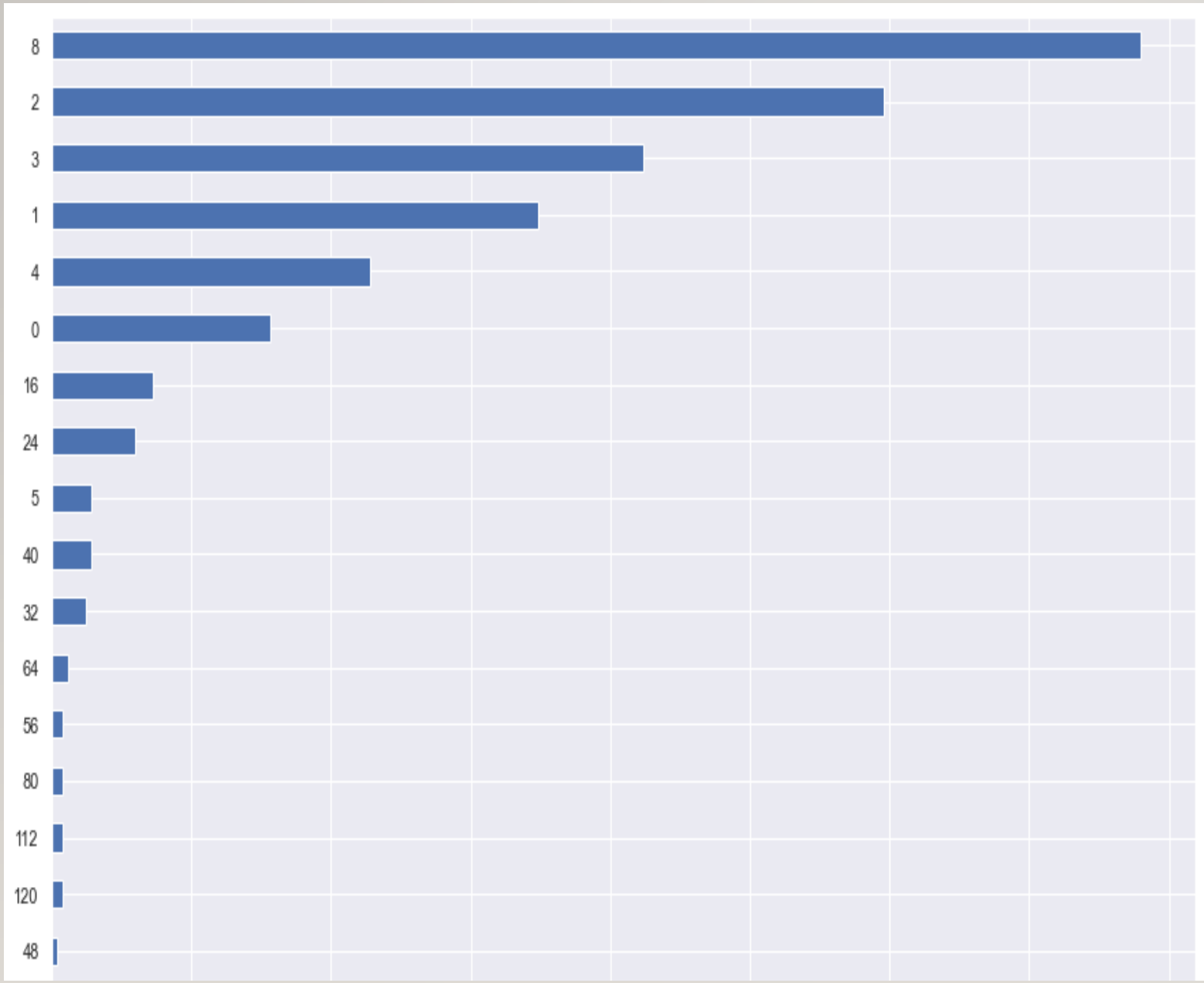
Heatmap showing absent hours throughout weekdays and months

EDA

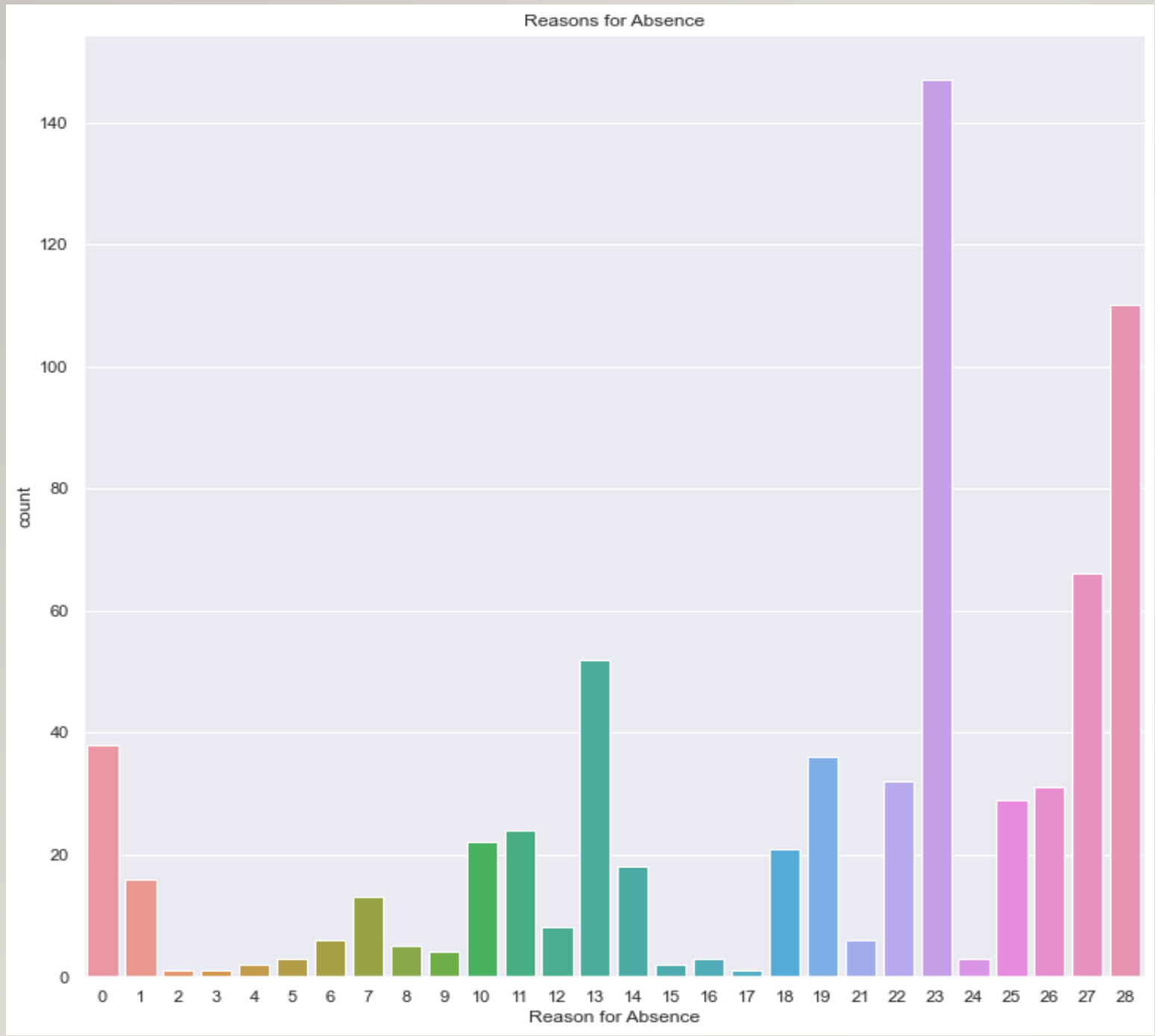

Transportation Expense
400
350
300
250
200
150
Frequency
1
Transportation Expense


Distance to Work
50
40
30
20
10
Frequency
1


120
100
80
60
40
20
0
Frequency
1
Absenteeism Time in Hours



Absenteeism Time in Hours

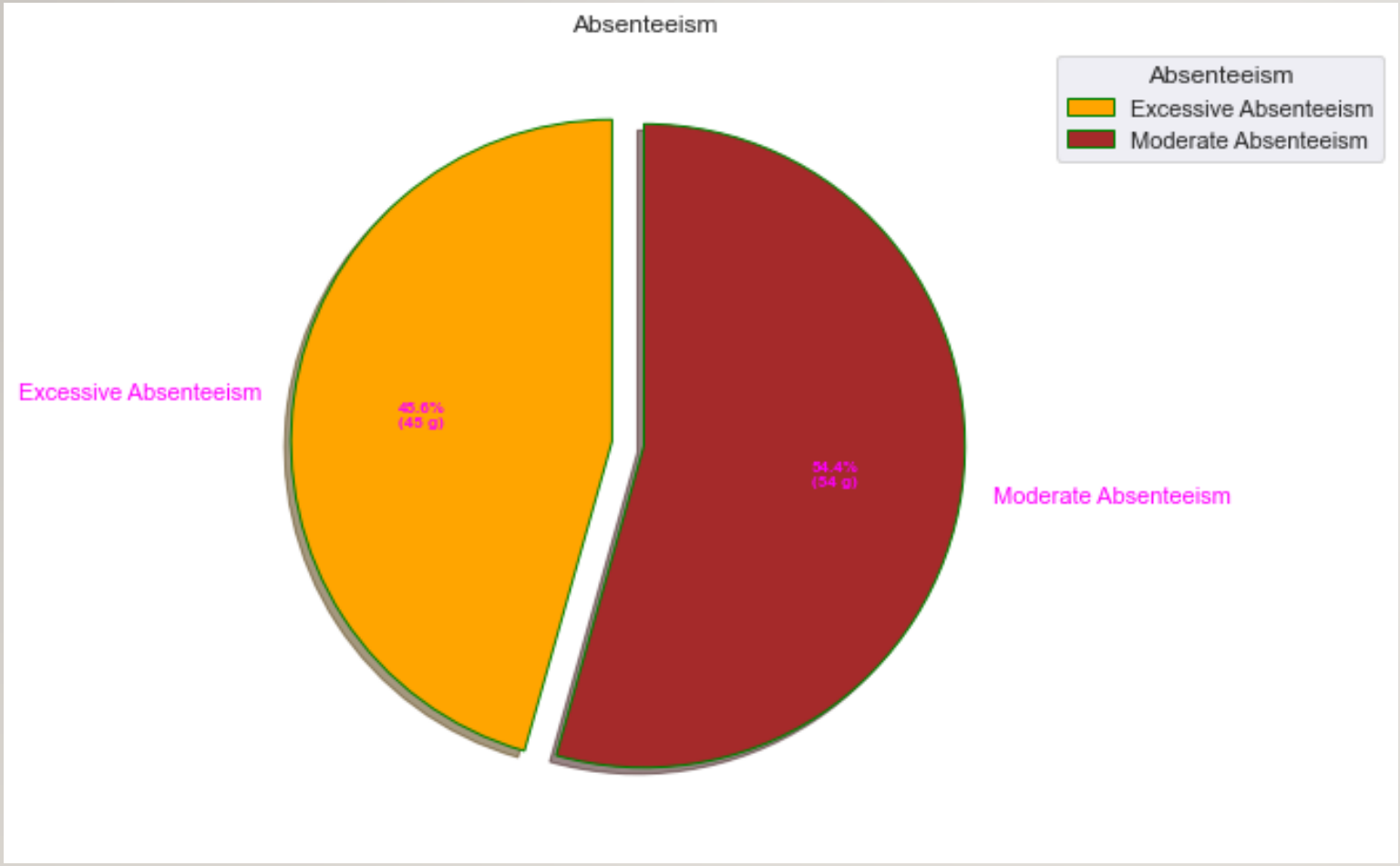


Reasons for Absent

23: Medical consultation

27: Physiotherapy
28: Dental consultation

PIE CHART SHOWING PERCENTAGE OF ABSENTEEISM HOUR



HYPOTHESIS TESTING

ALPHA = 0.05

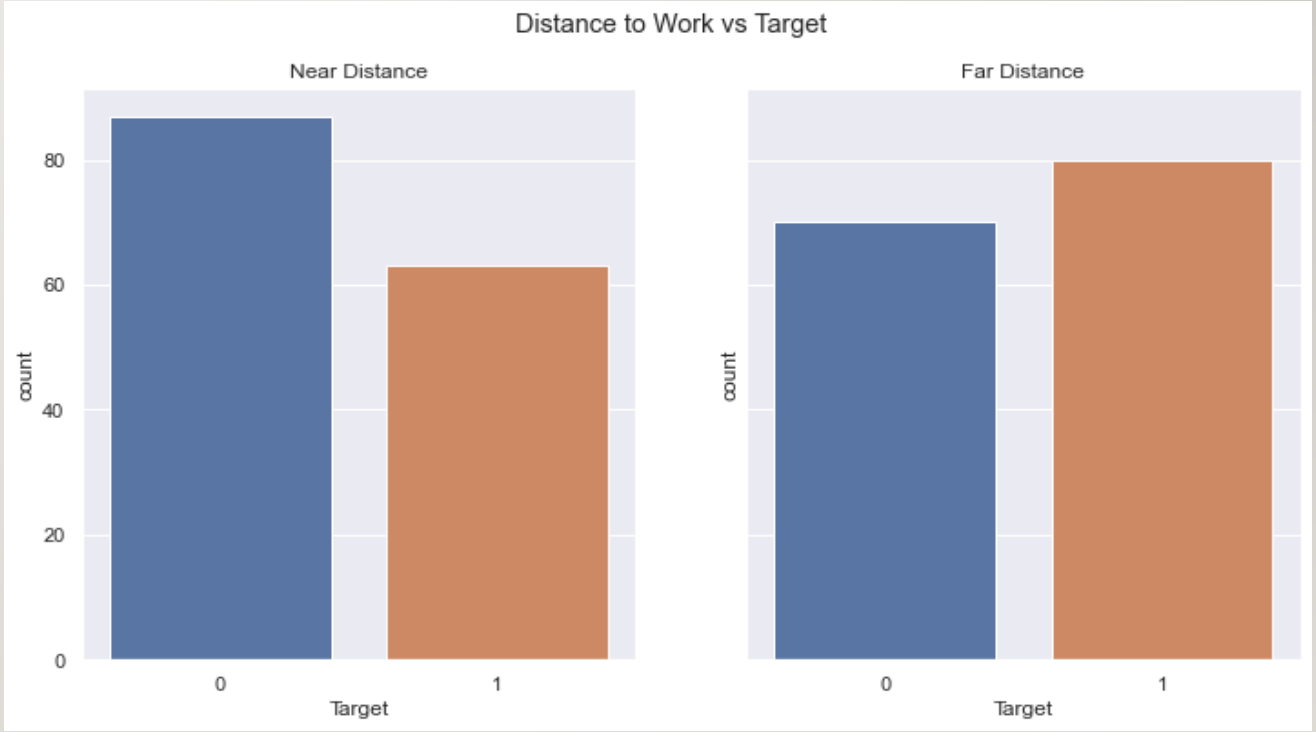
T = -1.971299347442617

P = 0.04961460049096922

WE REJECT OUR NULL HYPOTHESIS

Null Hypothesis (H0): There is no difference in the Absenteeism time between the near and far distance to workplace.

Alternative Hypothesis (HA): There is a difference in the Absenteeism time between the near and far distance to workplace.



HYPOTHESIS TESTING

alpha = 0.05
t = -6.2864120267018455
p = 8.547685104806746e-10
We reject our null hypothesis.

Null Hypothesis (H0): There is no difference in the Absenteeism time between the less and more expense to travel to workplace.

Alternative Hypothesis (HA): There is a difference in the Absenteeism time between the less and more expense to travel to workplace.

HYPOTHESIS TESTING

ALPHA = 0.05

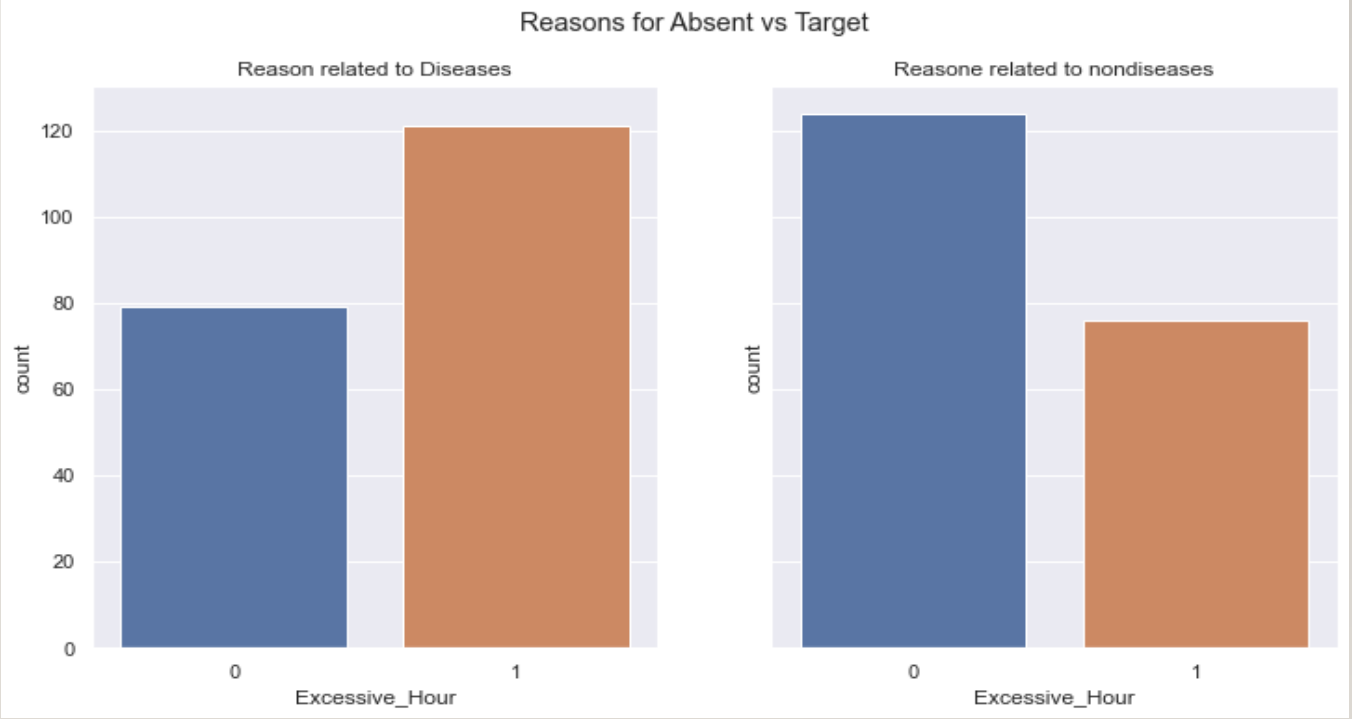
T = 4.6074075338187495

P = 5.497664640181139E-06

WE REJECT OUR NULL HYPOTHESIS.

Null Hypothesis (H0): There is no difference in the Absenteeism time between the disease reasons and non- disease reasons.

Alternative Hypothesis (HA): There is a difference in the Absenteeism time between the disease reasons and non- disease reasons.



SUMMARY

Employee is living far from workplace is more likely to absent from work

Employee who spend more expense for travel to work is also likely to absent from work

The reasons for absence related to diseases would spend more absent hour than reasons related to non-diseases.

FUTURE WORK

Analysis other features to be more accuracy for predict absenteeism hours purposes.

Use logistic regression model to classify the target.

Collect more data to be able to get better analysis