Project 1 Report

Eric Sclafani

1 Data description

For this project, the data set I used is called “Absenteeism at work” (click here to view the UCI link). The idea of the data set is to gain insight into the most common reasons for people to be absent from work. This data was taken from a courier company in Brazil from July 2007 to July 2010, as reported on the UCI website. Each record is an instance of someone being absent from this company, with 21 different attributes recorded for each.

2 Attributes

For some of the attributes, there is no description for what they mean or what units they’re measured in. I have contacted the data set author for further clarification, but have not received a response thus far. I have marked the ambiguous ones with a (?).

- **ID** - unique identifier
- **Reason for absence** - the reason for the absence according to the International code of Diseases (ICD). Numbers (1-21) represent a different reason detailed by the ICD and numbers (22-28) are non-ICD reasons for absence (all reasons are explained in the UCI repository).
- **Month of absence** - the month the absence occurred
- **Day of the week** - the day the absence occurred
- **Season** - the season the absence occurred
- **Transportation expense** (?) - money spent on transportation

- **Distance from residence to Work** - distance to work in kilometers
- **Service time** (?) - I think this is how many years soent at the company
- **Age** - age
- **Work load Average/day** (?) - not sure what the numbers in this column mean
- **Hit target** (?) - not sure what this means
- **Disciplinary failure** (?) - indicator for whether they failed school
- **Education** - highest level of education
- **Son** - number of children
- **Social drinker** - whether they're a social drinker
- **Social smoker** - whether they're a social smoker
- **Pet** - number of pets
- **Weight** - weight in kilograms
- **Height** - height in centimeters
- **Body mass index** - the body mass index
- **Absenteeism time in hours** - amount of hours absent from work

I think this data set is interesting because this type of data can be used by a company to find correlations in their employee absences. For example, if numerous employees have a large workload and also have lots of absence time, they may be getting overworked. Or maybe people who live farther from work tend to be absent more. There are plenty of features that let you ask probing questions like that.

3 Preprocessing

The first step in preprocessing was converting the numerical categorical columns back into their string representations. Had I not done this, the categorical names in my charts would all be meaningless numbers. The only exception to this is the “Reason for absence” column, in which the strings are way too long to show in the application.

After that, I removed the following columns because they were either meaningless or I didn’t feel like they contributed anything useful: ID, Hit target, Social drinker, Social smoker, Disciplinary failure.

I then renamed most of my other columns and also added a new one called “Disease”, which indicates whether the person was absent due to a disease or not. This column was attained by using the “Reason for absence” column. This was indeed meaningful because it shows that the majority of absences were not due to a disease.

4 Implementation notes

I developed this application using Plotly and Dash because I am a computational linguistics master’s student who has an extensive python background. Since the amount of code is relatively small for this project, I kept all the functioning code in one file (*app.py*). I sectioned my code using tildes like so: ∼∼∼ section_name ∼∼∼

In future projects, I will likely divide my different components and such into multiple files for organization and readability sake.