**Using Machine Learning for Business Operations**

**Focus: Employee Absenteeism**

Adriana Rubalcaba

Laura Sun

USC Data Analytics Boot Camp

University of Southern California, 2019

[arubalcaba001@gmail.com](mailto:arubalcaba001@gmail.com)

[sun.laura@outlook.com](mailto:lauar@gmail.com)

**Abstract**

This study uses a machine learning approach to conduct exploratory analysis on employee absenteeism. We explore different ML models through Scikit-learn, a Python machine learning library, and data modification techniques through Pandas, a Python data manipulation library, to predict and evaluate hours of absenteeism. Dataset is derived from a Brazilian courier and dated between July 2007 to July 2010.

**Introduction**

Lost productivity due to absenteeism in the U.S. cost employers $225.8 billion annually, or $1,685 per employee (U.S. Bureau of Labor Statistics, 2019). That's a big dent — and all due to a combination of direct and indirect costs. This study explores a courier company’s absentee records with the aim of identifying unique trends and providing insight on absenteeism through machine learning.

Because the variable we’d like to predict (hours in absenteeism) is continuous by nature, we assume this study is a regression problem.

**Data**

This dataset and its related information are available from the UCI Machine Learning Repository. With 740 records and 20 distinctive features collected for 36 different employees during a four-year period (July 2007 to July 2010). These features range from reason for absence, age, Body Mass Index (BMI) to service time, number of children and number of pets among others.

***Variables Information:***

**1.** Individual identification (ID)

**2.** Reason for absence (ICD) -

Absences attested by the **International Code of Diseases** (ICD) stratified into 21 categories (I to XXI) as follows:

**I**. Certain infectious and parasitic diseases

**II**. Neoplasms

**III.** Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

**IV**. Endocrine, nutritional and metabolic diseases

**V**. Mental and behavioral disorders

**VI**. Diseases of the nervous system

**VII**. Diseases of the eye and adnexa

**VIII**. Diseases of the ear and mastoid process

**IX**. Diseases of the circulatory system

**X**. Diseases of the respiratory system

**XI**. Diseases of the digestive system

**XII**. Diseases of the skin and subcutaneous tissue

**XIII**. Diseases of the musculoskeletal system and connective tissue

**XIV**. Diseases of the genitourinary system

**XV**. Pregnancy, childbirth and the puerperium

**XVI**. Certain conditions originating in the perinatal period

**XVII**. Congenital malformations, deformations and chromosomal abnormalities

**XVIII**. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

**XIX**. Injury, poisoning and certain other consequences of external causes

**XX.** External causes of morbidity and mortality

**XXI**. Factors influencing health status and contact with health services

And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).

**3.** Month of absence

**4.** Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))

**5.** Seasons (summer (1), autumn (2), winter (3), spring (4))

**6.** Transportation expense

**7.** Distance from Residence to Work (kilometers)

**8.** Service time

**9.** Age

**10.** Workload Average/day

**11.** Hit target

**12.** Disciplinary failure (yes=1; no=0)

**13.** Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))

**14.** Son (number of children)

**15.** Social drinker (yes=1; no=0)

**16.** Social smoker (yes=1; no=0)

**17.** Pet (number of pet)

**18.** Weight

**19.** Height

**20.** Body mass index

**21**. Absenteeism time in hours (target)

**Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics. In the given dataset, there are 21 variables and data types of all variables are either float64 or int64. There are 740 observations and 21 columns in our data set. Missing value is also present in our data.

***List of columns and their number of unique values***

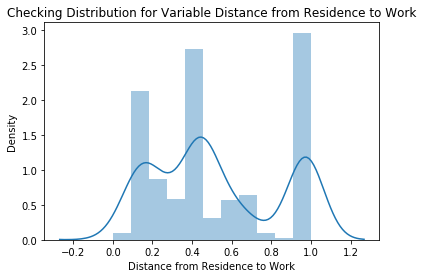
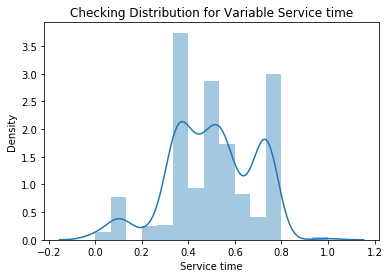
|  |  |
| --- | --- |
| Columns | Unique Values |
| ID | 36 |
| Reason for absence | 28 |
| Month of absence | 13 |
| Day of the week | 5 |
| Seasons | 4 |
| Transportation expense | 24 |
| Distance from Residence to Work | 25 |
| Service time | 18 |
| Age  Work load Average/day  Hit target  Disciplinary failure  Education  Son  Social drinker  Social smoker  Pet  Weight  Height  Body mass index  Absenteeism time in hours | 22  38  13  2  4  5  2  2  6  26  14  17  19 |

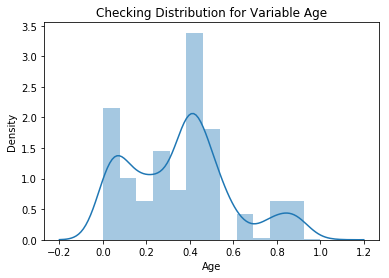
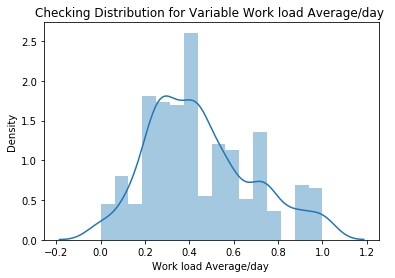
**From EDA we have concluded that there are 10 continuous variable and 11 categorical variables in nature.**

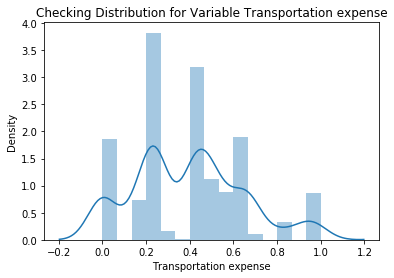
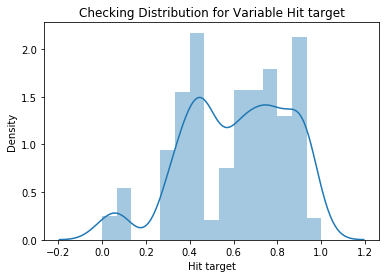
**Methodology**

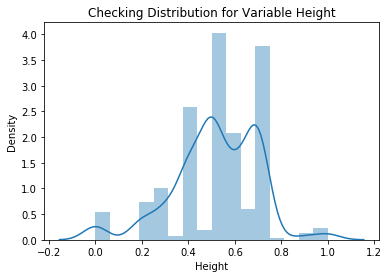
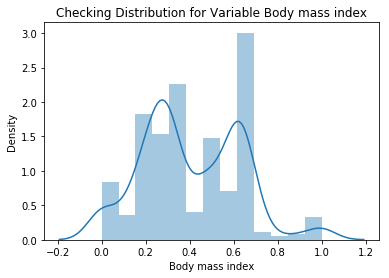
***Preprocessing***

A substantial portion of the project revolved around understanding the data, standardizing the features and taking care of extreme outliers to better feed our machine learning models. Any predictive modeling requires a thorough examination of the data before modeling. We begin by assessing the probability distributions of all variables. Most analysis like regression, requires the data to be normally distributed. We can visualize that in a glance by looking at the probability distributions or probability density functions of the variables.

** **

** **

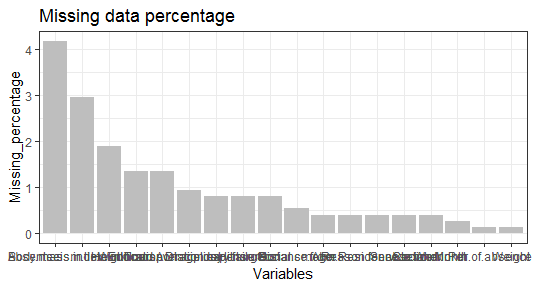
** **

** **

***Missing Value Analysis***

In statistics, missing data, or missing values, occur when there aren’t any data for the variable in an observation. Missing values are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. If a column has more than 30% of data as missing value, we can either ignore the entire column or we can ignore those observations. In the given data the maximum percentage of missing value is 4.189% for **body mass index** column. So, we will compute missing values for all the columns.

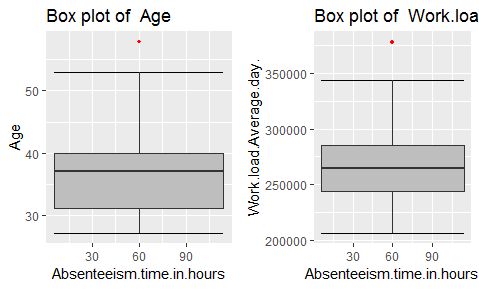
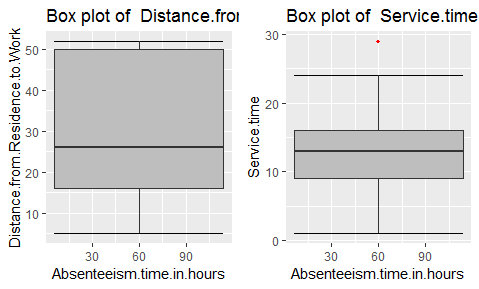
**In this project we have used KNN imputation method to impute missing value**.

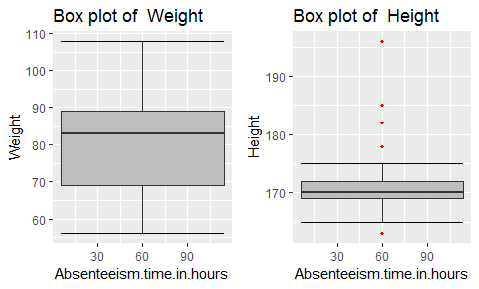
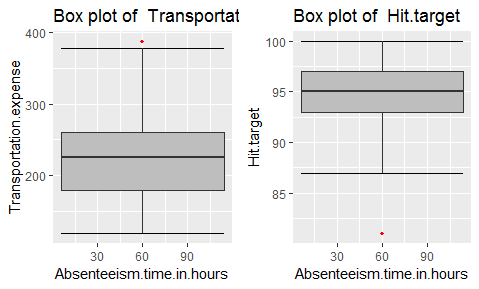


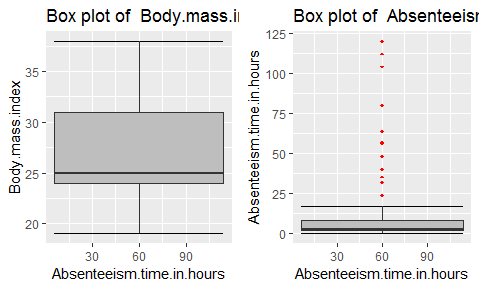
***Outlier Analysis***

We can clearly observe from these probability distributions that most of the variables are skewed. The skew in these distributions can be explained by the presence of outliers and extreme values in the data. One of the other steps of pre-processing apart from checking for normality is the presence of outliers. In this case we use a classic approach of removing outliers. See next page for boxplot visualizations of outliers.

In our figures, we plot the 11 predictor variables with respect to **Absenteeism time in hour**. A lot of useful inferences can be made from these plots. First as you can see, we have a lot of outliers and extreme values in each of the data set.

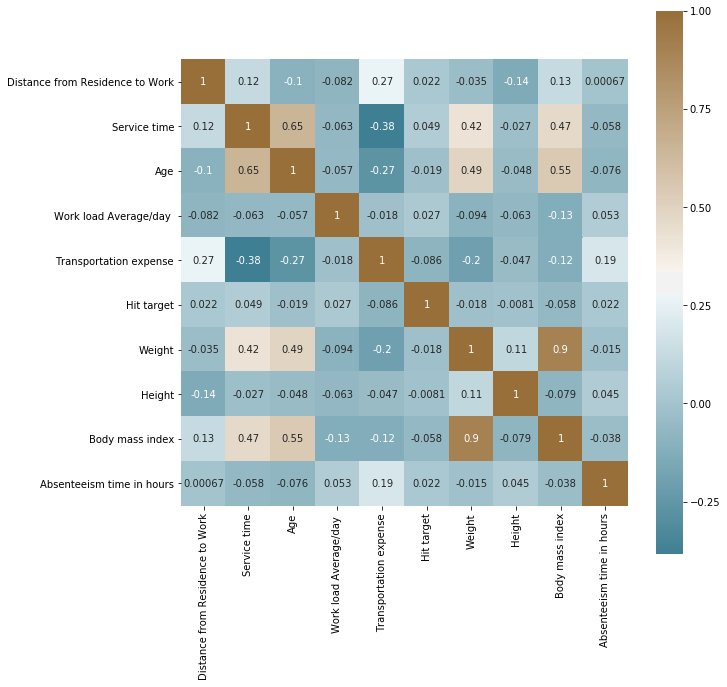




From the boxplot nearly all variables, with the **exception of “Distance from residence to work”, “Weight” and “Body mass index”**, have outliers. We convert the outliers (data beyond minimum and maximum values) as NA i.e. missing values and fill them by **KNN** imputation method.

***Feature Selection***

To improve modeling performance, we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction. Selecting a subset of relevant columns for the model construction is known as Feature Selection. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase overhead. To reduce overhead, we adopt feature selection techniques to extract meaningful features out of data. This in turn helps us to avoid the problem of multi-collinearity. In this project we have selected **Correlation Analysis** for numerical variable and **ANOVA** (Analysis of variance) for categorical variable.



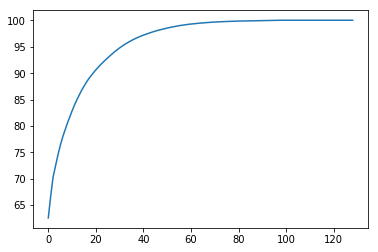
From correlation analysis we have found that **Weight** and **Body mass index** has high correlation (>0.7), so we have excluded the **Weight** column.

***Feature Scaling***

**Feature scaling** is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and generally performed during the data preprocessing step. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, most classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, that particular feature will govern the distance. Therefore, the range of all features should be normalized so that each feature contributes proportionately to the final distance. Since our data is not uniformly distributed we will use **normalization** as a Feature Scaling Method.

***Principal Component Analysis***

Principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set. It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. With fewer variables, visualization also becomes much more meaningful. PCA is more useful when dealing with 3 or higher dimensional data. After creating dummy variable of categorical variables, the shape of our data became 107 columns and 714 observations. This high number of columns leads to bad accuracy.



We apply the PCA algorithm on our data and from the above graph we have conclude that 45 variables out of 107 explains more than 95% of data. Thus, we can select only those 45 variables to feed our models.

**Modeling**

After a thorough pre-processing, we will be using some regression models on our processed data to predict the target variable. The following are the models which we have built –

***Decision Tree***

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Each branch connects nodes with “and” and multiple branches are connected by “or”. It can be used for classification and regression. It is a supervised machine learning algorithm that accepts continuous and categorical variables as independent variables. The appeal of this model derives from the extremely easy comprehensibility by the business users. The RMSE value and R^2 value for our project in R and Python are –

|  |  |  |
| --- | --- | --- |
| **Decision Tree** | **R** | **PYTHON** |
| **RMSE Train** | 0.410 | 0.573 |
| **RMSE Test** | 0.363 | 0.568 |
| **R^2 Test** | 0.98 | 0.96 |

***Random Forest***

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree. It means to build each decision tree on random forest we are not going to use the same data. The RMSE value and R^2 value for our project in R and Python are –

|  |  |  |
| --- | --- | --- |
| **Random Forest** | **R** | **PYTHON** |
| **RMSE Train** | 0.369 | 0.033 |
| **RMSE Test** | 0.614 | 0.029 |
| **R^2 Test** | 0.96 | 0.99 |

***Linear Regression***

Linear Regression is one of the statistical methods of prediction. It is applicable only on continuous data. To build any model we have some assumptions to put on data and model. Here are the assumptions to the linear regression model.

|  |  |  |
| --- | --- | --- |
| **Linear Regression** | **R** | **PYTHON** |
| **RMSE Train** | 0.001 | 4.34e-15 |
| **RMSE Test** | 0.001 | 4.14e-15 |
| **R^2 Test** | 0.99 | 1 |

***Gradient boosting***

**Gradient boosting** is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

|  |  |  |
| --- | --- | --- |
| **Gradient boosting** | **R** | **PYTHON** |
| **RMSE Train** | 2.276 | 0.001 |
| **RMSE Test** | 1.818 | 0.001 |
| **R^2** | 0.91 | 0.99 |

**Conclusion**

***Model Evaluation***

We have seen the **Root Mean Square Error** (RMSE) and **R-Squared** Value of different models. **Root Mean Square Error** (RMSE) is the standard deviation of the residuals (prediction **errors**). Residuals are a measure of how far from the regression line data points are, RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Whereas **R**-**squared** is a relative measure of fit, **RMSE** is an absolute measure of fit. As the square root of a variance, **RMSE** can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of **RMSE** and higher value of **R-Squared Value** indicate better fit.

***Model Selection***

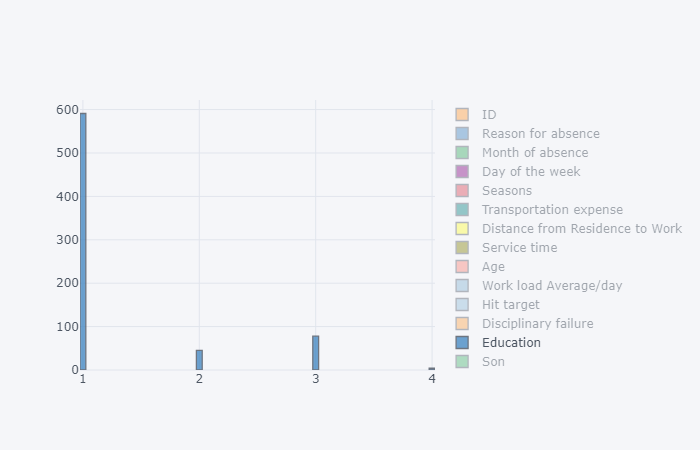
From the observation of all **RMSE Value** and **R-Squared** Value we have concluded that **Linear Regression Model**  has minimum value of RMSE and it’s **R-Squared** Value is also maximum (i.e. 1).

The RMSE value of Testing data and Training does not differs a lot this implies that it is not the case of overfitting.

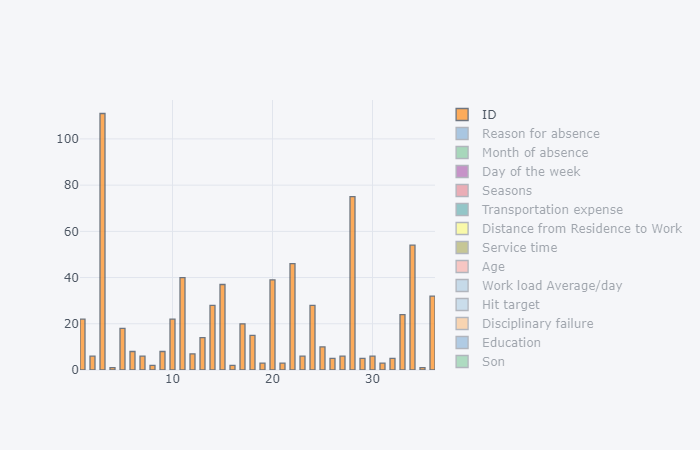
***Key Insights***

**The Changes which company should bring to reduce the number of absenteeism –**

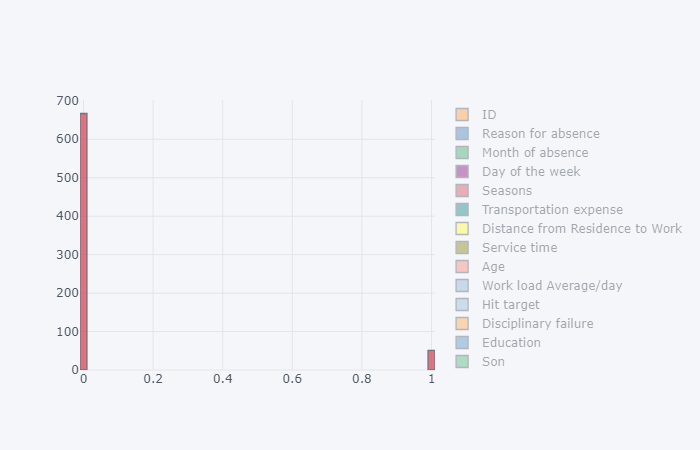
1. We find that employees with low education have maximum absentee time.

****

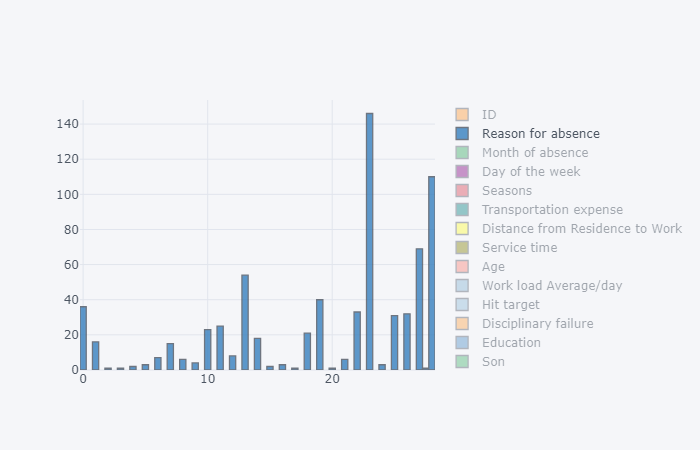
1. Employees with **ID 3, 28, 34** have the highest numbers of absences.

****

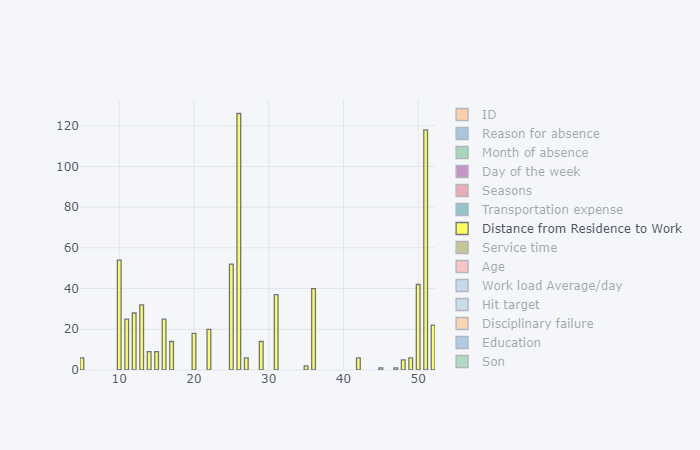
1. Employees who are social smoker have more absentee hours than non-smokers.

****

1. Medical consultation and dental consultation are the top reasons for absence.

****

1. Employees who have a greater Distance from Residence to Work tend to have a higher number of absences.

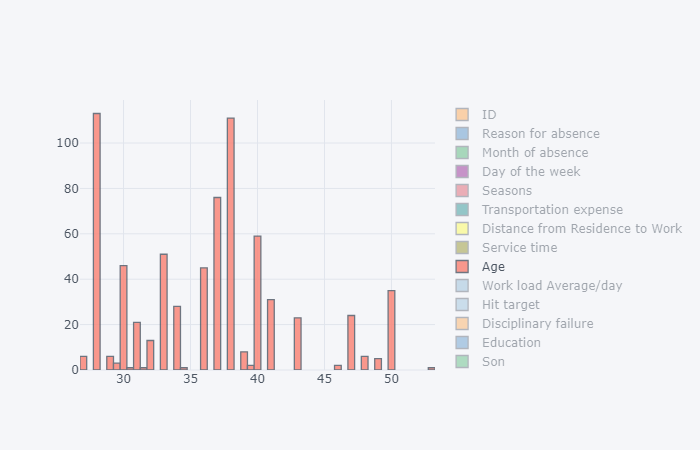
****

**Appendix**

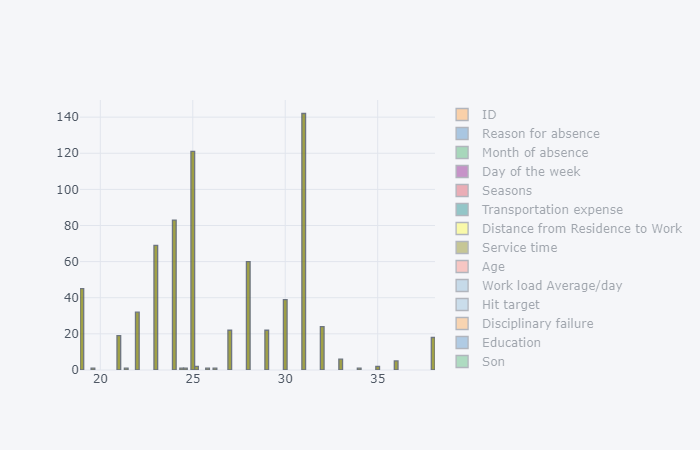
**Extra Figures**

**Relationship of our target variable (Absentee time in hour) with other variables.**

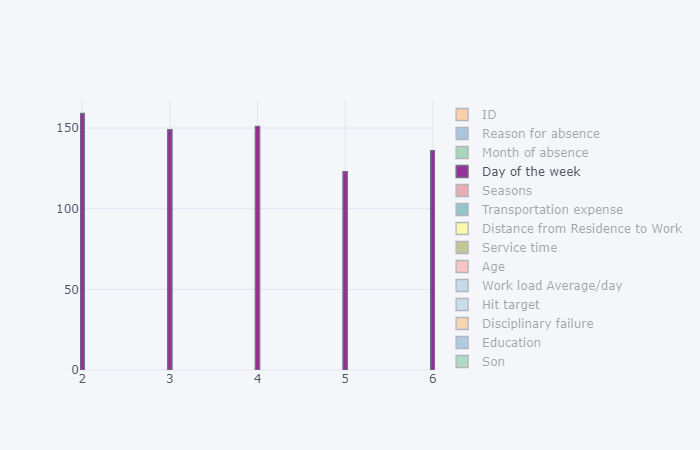
1. **With “Age”**

****

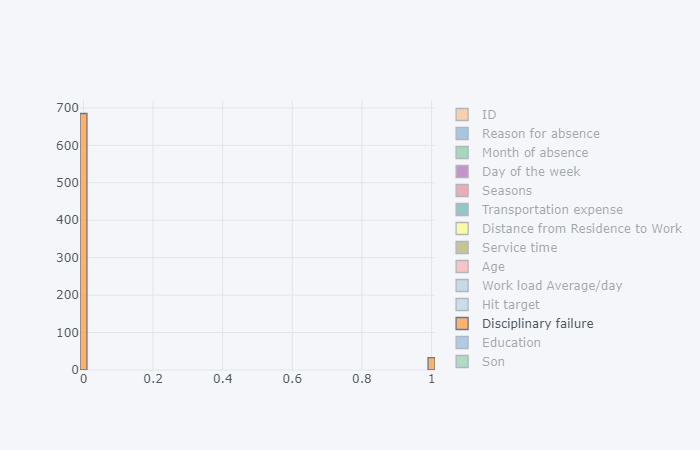
1. **With “Body mass index”**

****

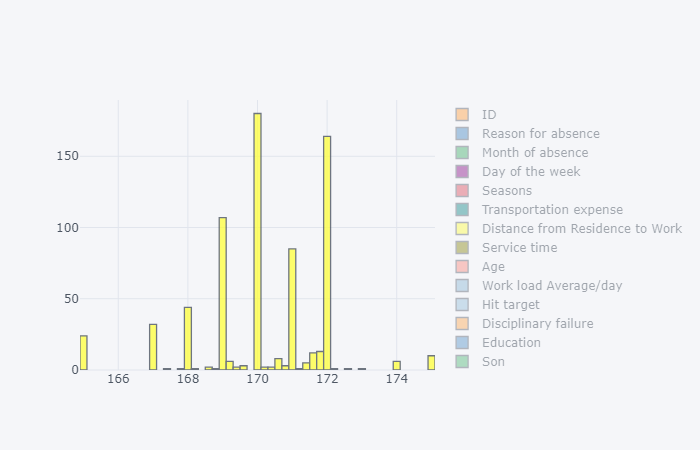
1. **With “Day of week”**

****

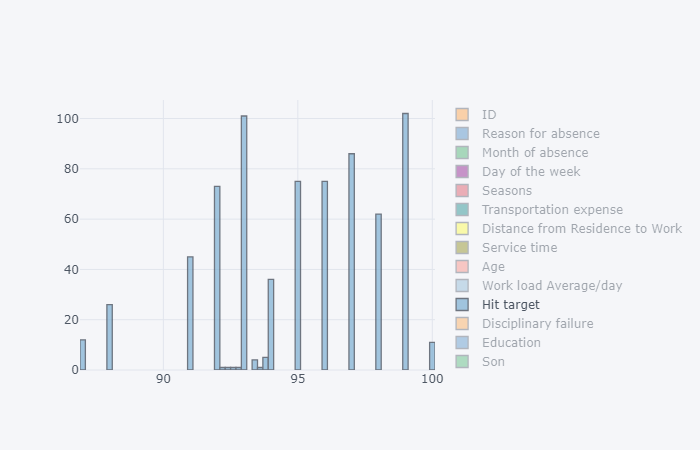
1. **With “Disciplinary failure”**

****

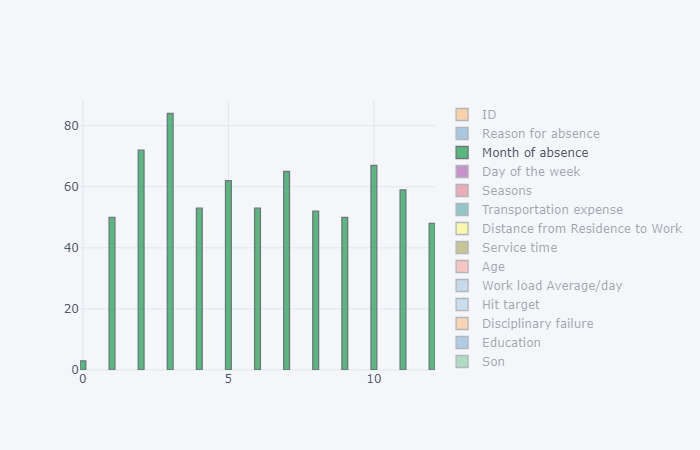
1. **With “Height”**

****

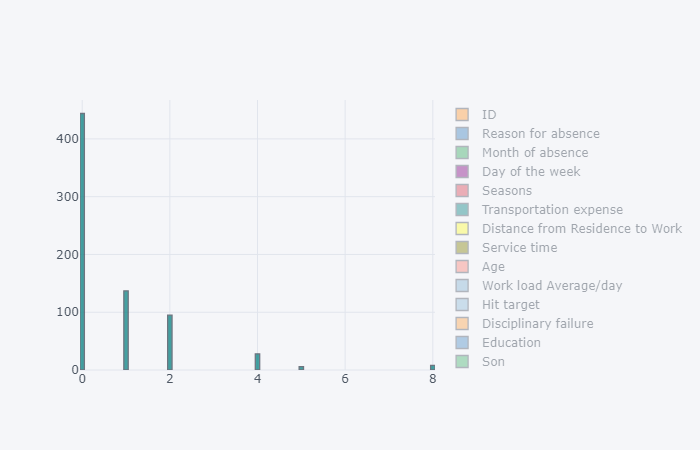
1. **With “Hit target”**

****

1. **With “Month of absent”**

****

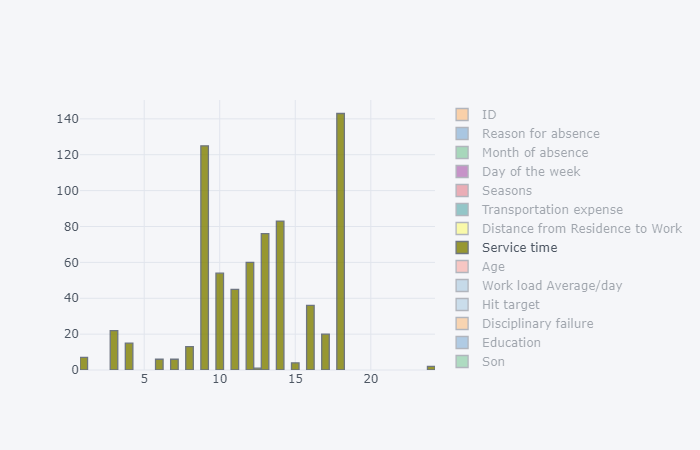
1. **With “Pet”**

****

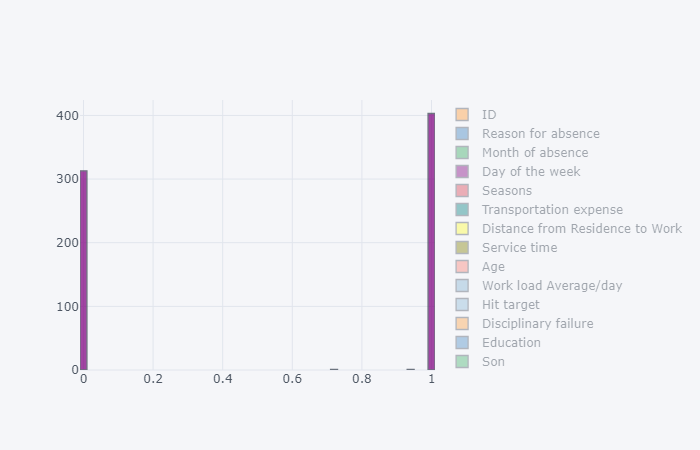
1. **With “Seasons”**

****

1. **With “Service time”**

****

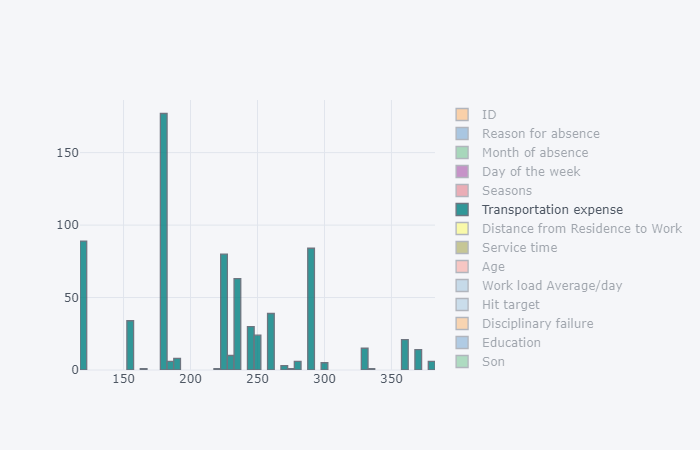
1. **With “Social drinker”**

****

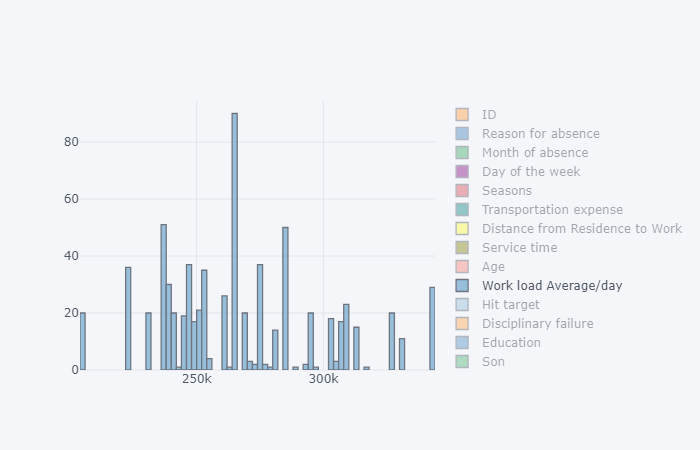
1. **With “Son”**

****

1. **With “Transportation Expense”**

****

1. **With “Work load/day”**

****

**Resources**

“Absences from Work of Employed Full-Time Wage and Salary Workers by Occupation and Industry.” *U.S. Bureau of Labor Statistics*, U.S. Bureau of Labor Statistics, 18 Jan. 2019, www.bls.gov/cps/cpsaat47.htm.

“Absenteeism at Work.” *UCI Machine Learning Repository: Absenteeism at Work Data Set*, 5 Apr. 2018, archive.ics.uci.edu/ml/datasets/Absenteeism+at+work.

Analytics Vidhya Content. “Practical Guide to Principal Component Analysis (PCA) in R & Python.” *Analytics Vidhya*, 24 June 2019, www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/.

“EdWisor Home.” *Get Skilled.Get Hired*, edwisor.com/career-data-scientist.