Project Report

Employee Absenteeism

Diksha Bhalerao

15-Aug-2019

Table of Contents

Chapter 1

1. Introduction

1.1 Problem statement

XYZ is a courier company. As we appreciate that human capital plays an important role in collection, transportation and delivery. The company is passing through genuine issue of Absenteeism. The company has shared it dataset and requested to have an answer on the following areas:

1. What changes company should bring to reduce the number of absenteeism?
2. How much losses every month can we project in 2011 if same trend of absenteeism continues?

1.2 Data exploration

Data exploration is a task involves outlining of characteristics of a data set, i.e. its size, accuracy, initial patterns and other attributes in the data. That's why it is important to do a general hypothesis of data.

1.2.1 Data

Our primary task is to find which factors are affecting absenteeism in company. So here's a sample of data we've been provided.

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense
11	26	7	3	1	289
36	0	7	3	1	118
3	23	7	4	1	179

7	7	7	5	1	279
11	23	7	5	1	289

Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Disciplinary failure
36	13	33	239,554	97	0
13	18	50	239,554	97	1
51	18	38	239,554	97	0
5	14	39	239,554	97	0
36	13	33	239,554	97	0

Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
1	2	1	0	1	90	172	30	4
1	1	1	0	0	98	178	31	0
1	0	1	0	0	89	170	31	2
1	2	1	1	0	68	168	24	4
1	2	1	0	1	90	172	30	2

1.2.2 Structure of data

Size: 740 obs. of 21 variables

Variable	Data Type
ID	int64
Reason for absence	float64
Month of absence	float64
Day of the week	int64
Seasons	int64
Transportation issue	float64
Distance from residence to work	float64
Service time	float64
Age	float64
Work load average/day	float64
Hit target	float64
Disciplinary failure	float64
Education	float64
Son	float64
Social drinker	float64
Social smoker	float64
Pet	float64
Weight	float64
Height	float64
Body mass index	float64
Absenteeism time in hours	float64

1.2.3 Uniqueness of data

Variable	No. of unique values
ID	36
Reason for absence	28
Month of absence	13
Day of the week	5
Seasons	4
Transportation issue	24
Distance from residence to work	25
Service time	18
Age	22
Work load average/day	38
Hit target	13
Disciplinary failure	2
Education	4
Son	5
Social drinker	2
Social smoker	2
Pet	6
Weight	26
Height	14
Body mass index	17
Absenteeism time in hours	19

1.2.4 Completeness of data

Missing values in a data often cause inconsistency in output. So lets first check number of missing values in each variable.

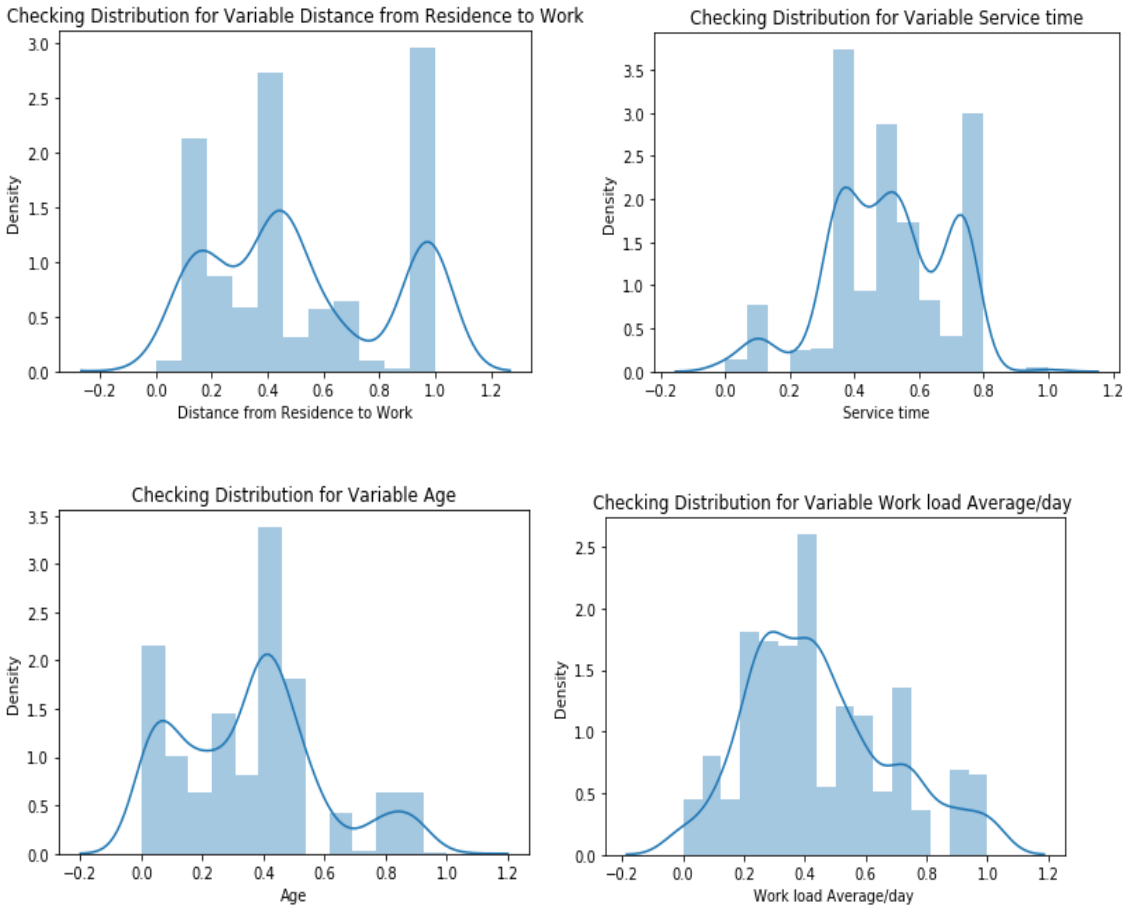
Variable	Missing count
ID	0
Reason for absence	3
Month of absence	1
Day of the week	0
Seasons	0
Transportation issue	6
Distance from residence to work	3
Service time	3
Age	2
Work load average/day	8
Hit target	6
Disciplinary failure	5
Education	10
Son	6
Social drinker	3
Social smoker	4
Pet	2
Weight	1
Height	14
Body mass index	29
Absenteeism time in hours	0

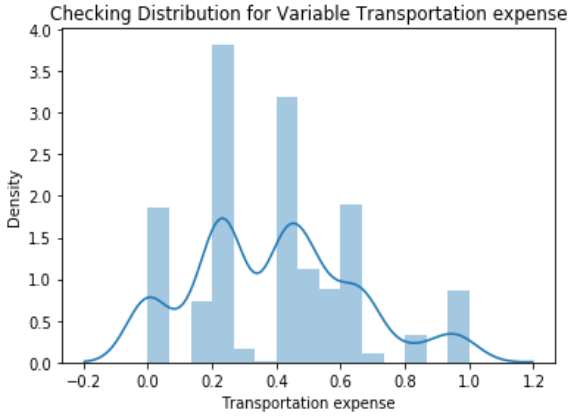
from EDA we have concluded that there are 10 continuous variable and 11 categorical variable in nature.

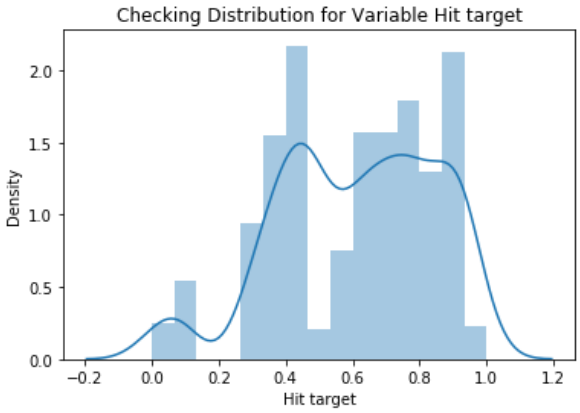
Chapter 2

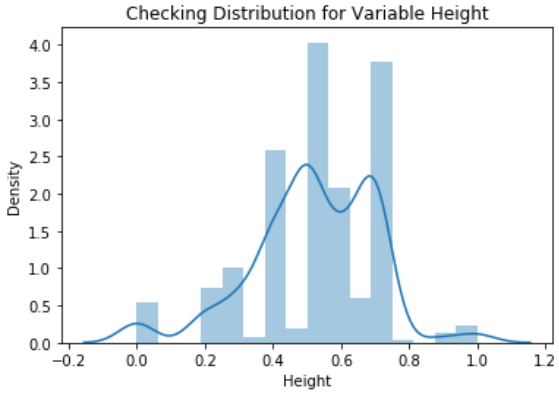
2.1 Data Preprocessing

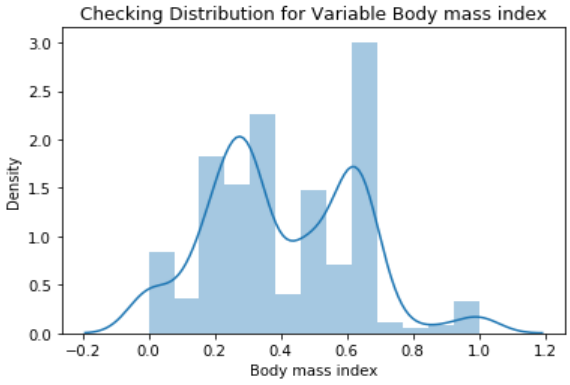
For predictive modeling before we start modeling our data. We need to clean given data. This maintains consistency and accuracy in our modeling. In data cleaning we process missing values to avoid any discrepancy in our data. We start with visualizing our data through Plots ,graphs and histograms. We will look into probability distributions of each variable So that we can check if data is normally distributed or not.





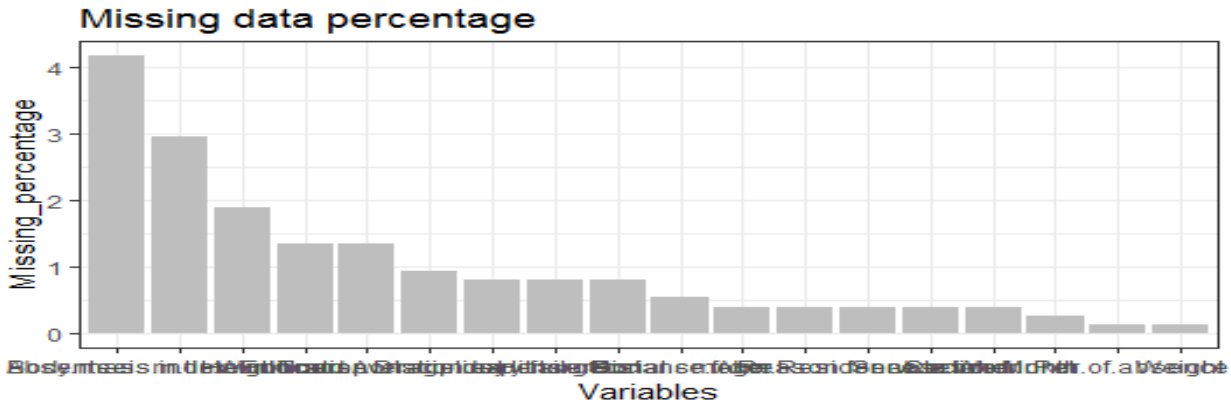






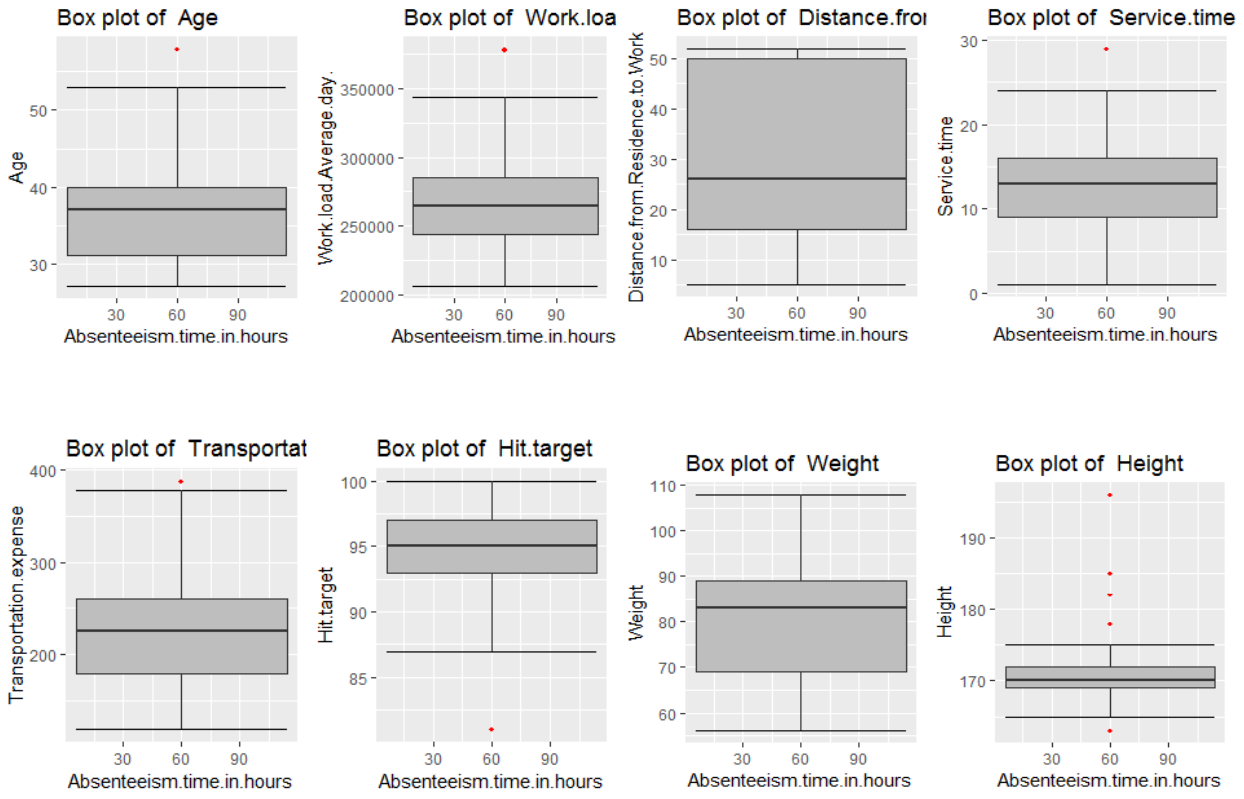
2.1.1 Missing value analysis

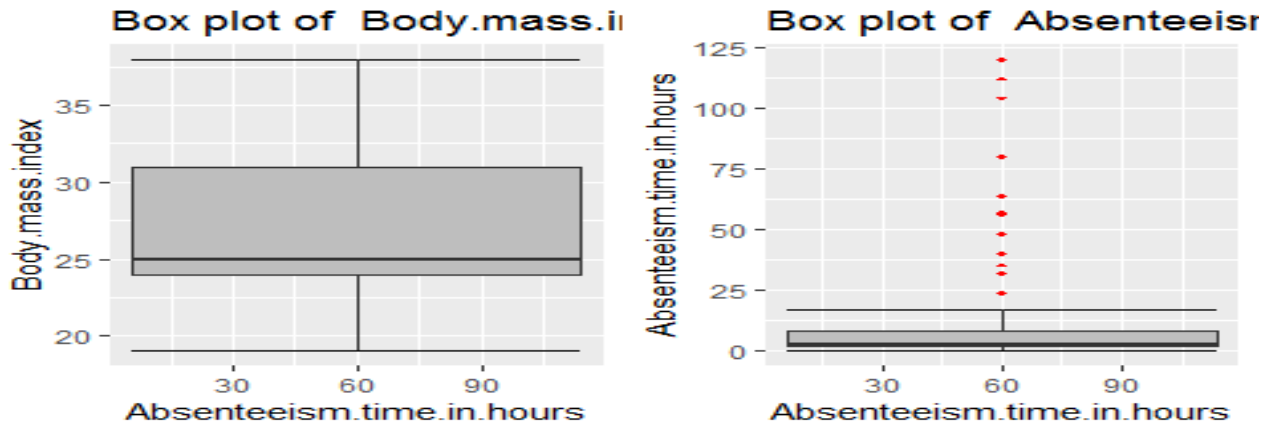
Missing value analysis is the most important task in data analysis. Also it is a really crucial task. After we check for missing value, if a column has more than 30% observations missing then we either ignore whole column or neglect those observations. In this project we have used KNN imputation method to impute missing value.



2.1.2 Outlier analysis

From the probability distributions given here. We can clearly say that variables are mostly skewed. Outliers and extreme values are main causes of skewness. For removing these outliers we use boxplot method. Following are boxplots of predictor variables w.r.t. Absenteeism time in hour.



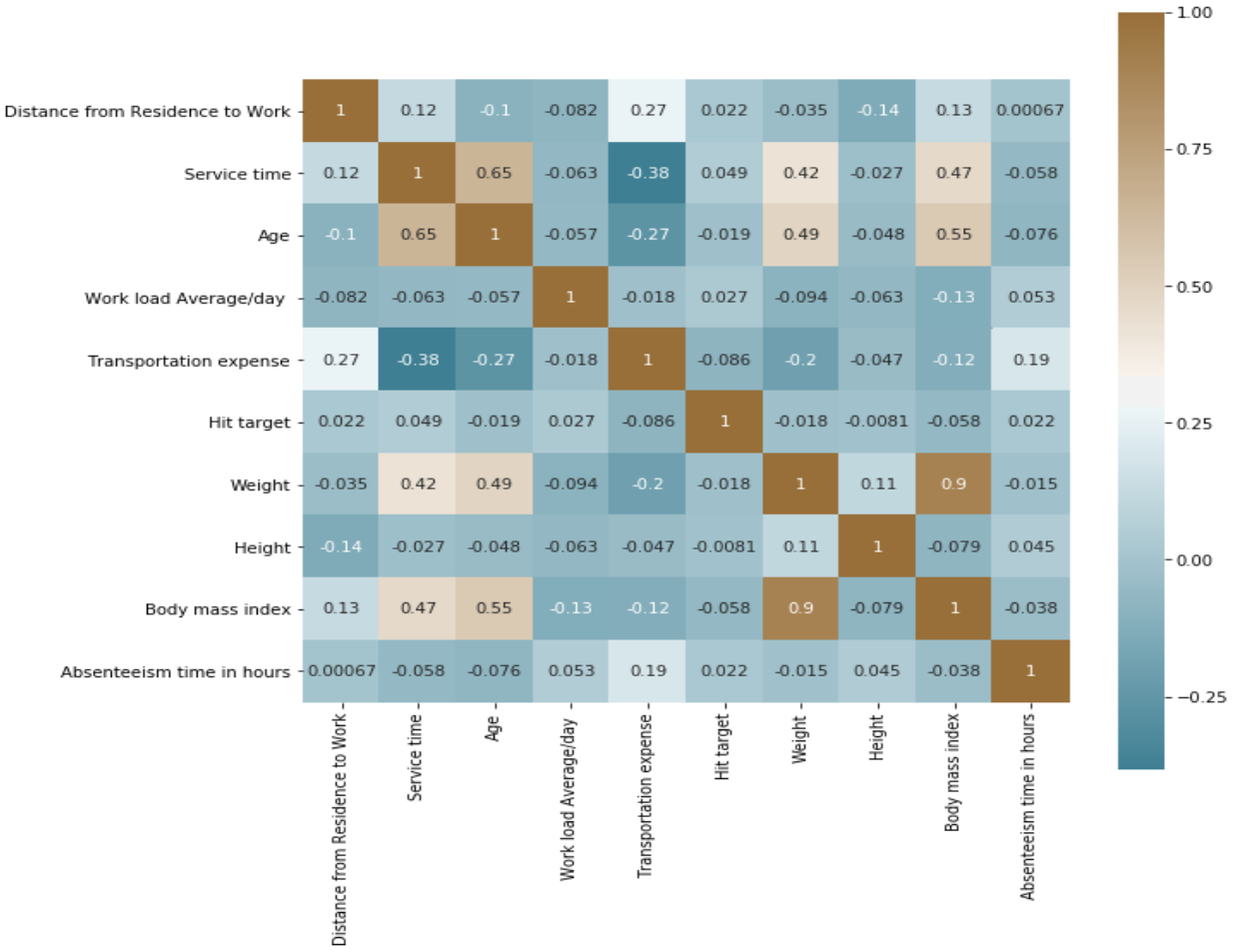


According to these boxplots almost all the variables except “Distance from residence to work”, “Weight” and “Body mass index” consists of outliers. We have converted the outliers as NA i.e. missing values and fill them by KNN imputation method.

2.1.3. Feature selection

Feature selection process of selection of a subset of relevant features (predictors, variables) to use in model construction.

Before performing any type of modeling 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 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 technique to extract meaningful features out of data. This in turn helps us to avoid the problem of multicollinearity. In this project we have selected Correlation Analysis for numerical variable and ANOVA (Analysis of variance) for categorical variable.



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

2.1.4 Feature Scaling

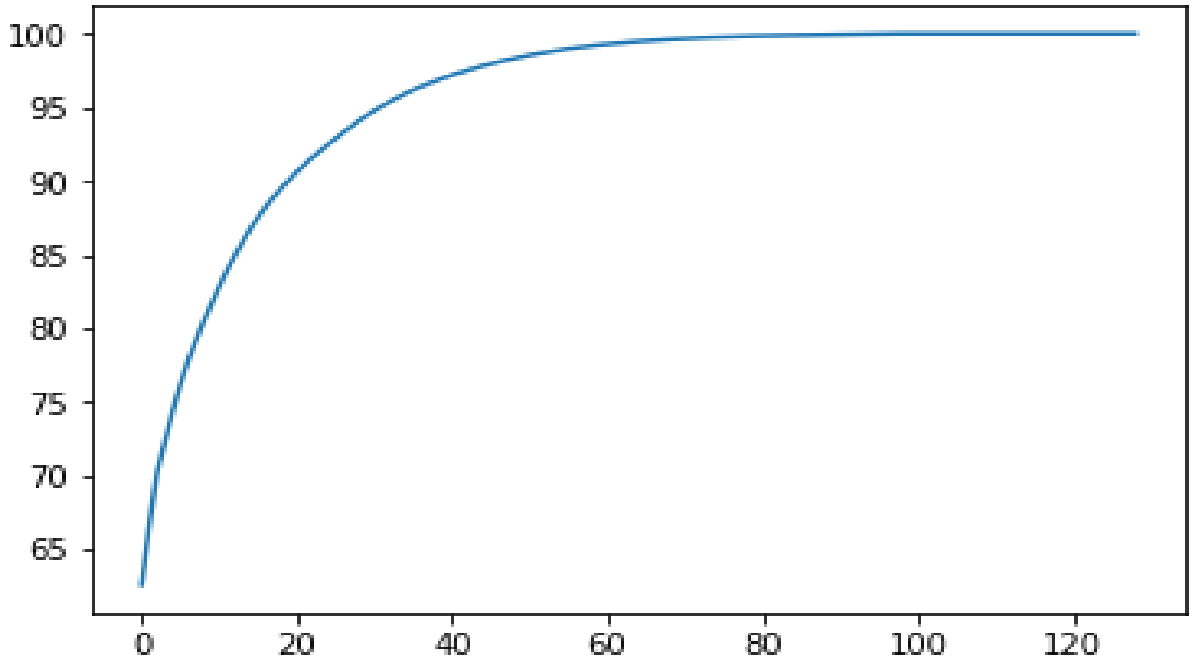
Feature scaling could be a technique wont to standardize the vary of freelance variables or options of information. Inprocessing, it's conjointly called information standardization and is usually performed throughout the info preprocessing step.

Since the vary of values of data varies wide, in some machine learning algorithms, objective functions won't work properly while not standardization. as an example, the bulk of classifiers calculate the gap between 2 points by the gap. If one among the options incorporates a broad vary of values, the gap are going to be ruled by this explicit feature. Therefore,

the vary of all options ought to be normalized so every feature contributes some proportionately to the ultimate distance.. As our data is not equally distributed we'll use Normalization as Feature Scaling Method.

2.1.5 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 information as much 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 have applied PCA algorithm on our data and from the above graph we have concluded that 45 variables out of 107 explains more than 95% of data. So we have selected only those 45 variables to feed our models.

Chapter 3

3. Modeling

After a thorough preprocessing we will be using some regression models on our processed data to predict the target variable. Following are the models which we have built –

3.1 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. Every branch is connected to nodes with “and” and multiple branch nodes are connected with “or”. It can be used for classification and regression. It is a supervised machine learning algorithm. Accept continuous and categorical variables as independent variables. Split of decision tree is seen in the below tree. The RMSE value and R^2 value for our project in R and Python are –

Decision Tree	R	PYTHON
RMSE Test	0.12428932	0.5755419840270428
R^2 Test	0.69948786	0.9702022454372837

3.2 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 Test	0.09219915	0.006151198268317667
R^2 Test	0.89601668	0.9999965963180563

3.3 Liner 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 Test	1.156764e-15	0.00658891418853747
R^2 Test	1.000000e+00	0.999996094674632

Chapter 4

4. Conclusion

In this chapter we are going to evaluate our models, select the best model for our dataset and try to get answers of the asked questions.

4.1 Model Evaluation

Previously 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.

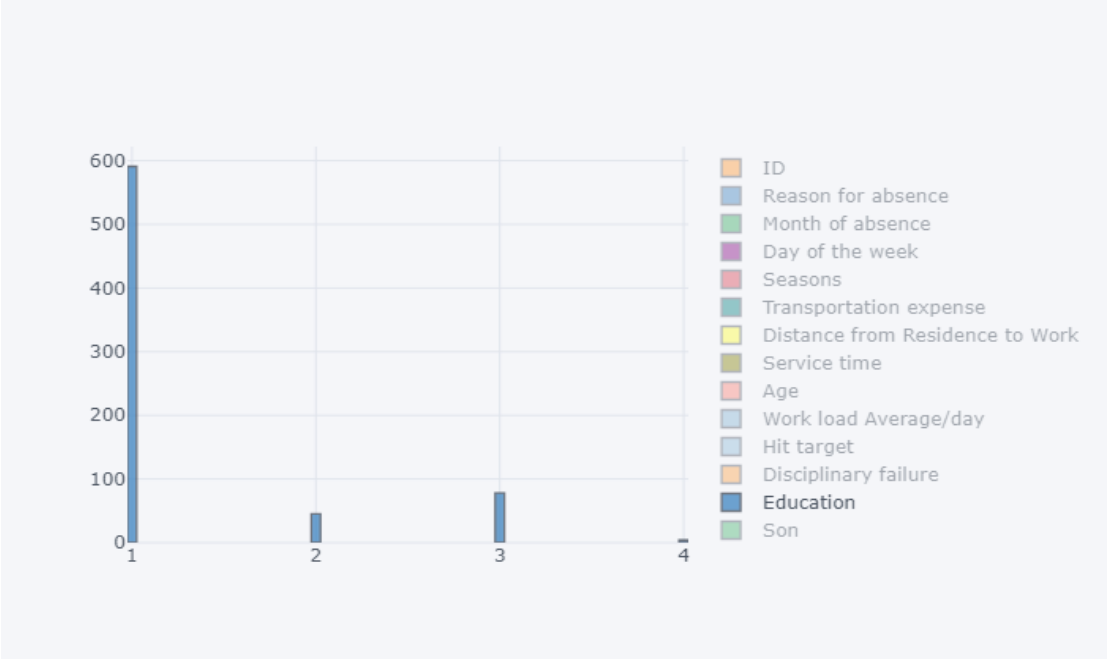
4.2 Model Selection

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).

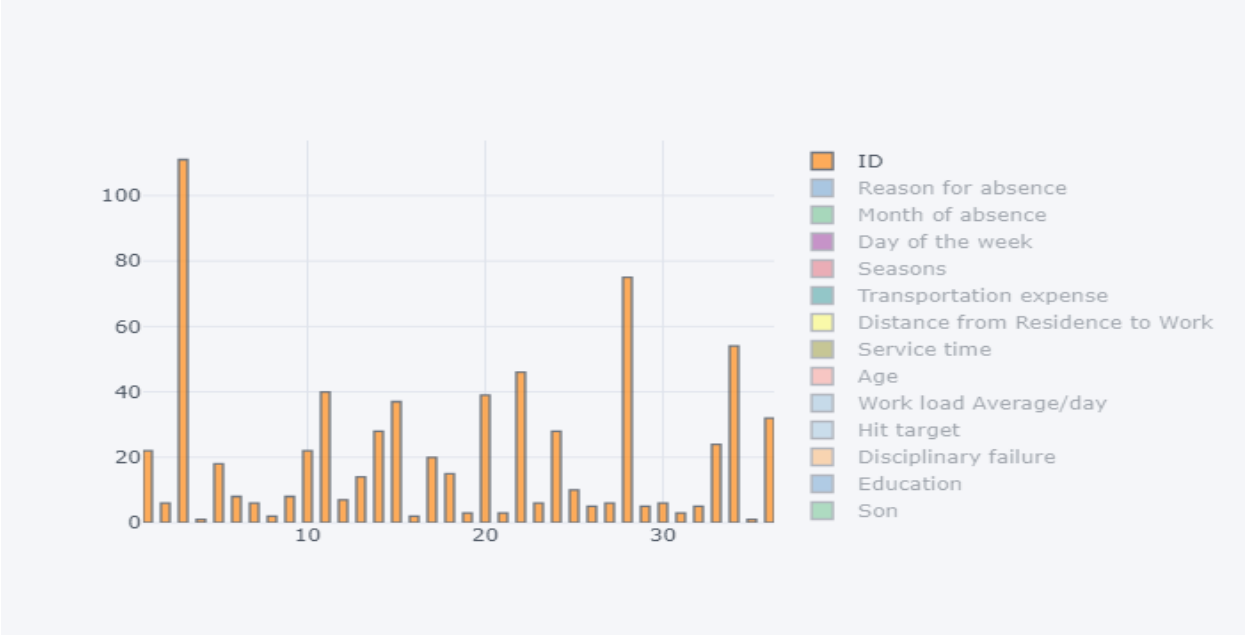
4.3 Solutions to asked questions

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

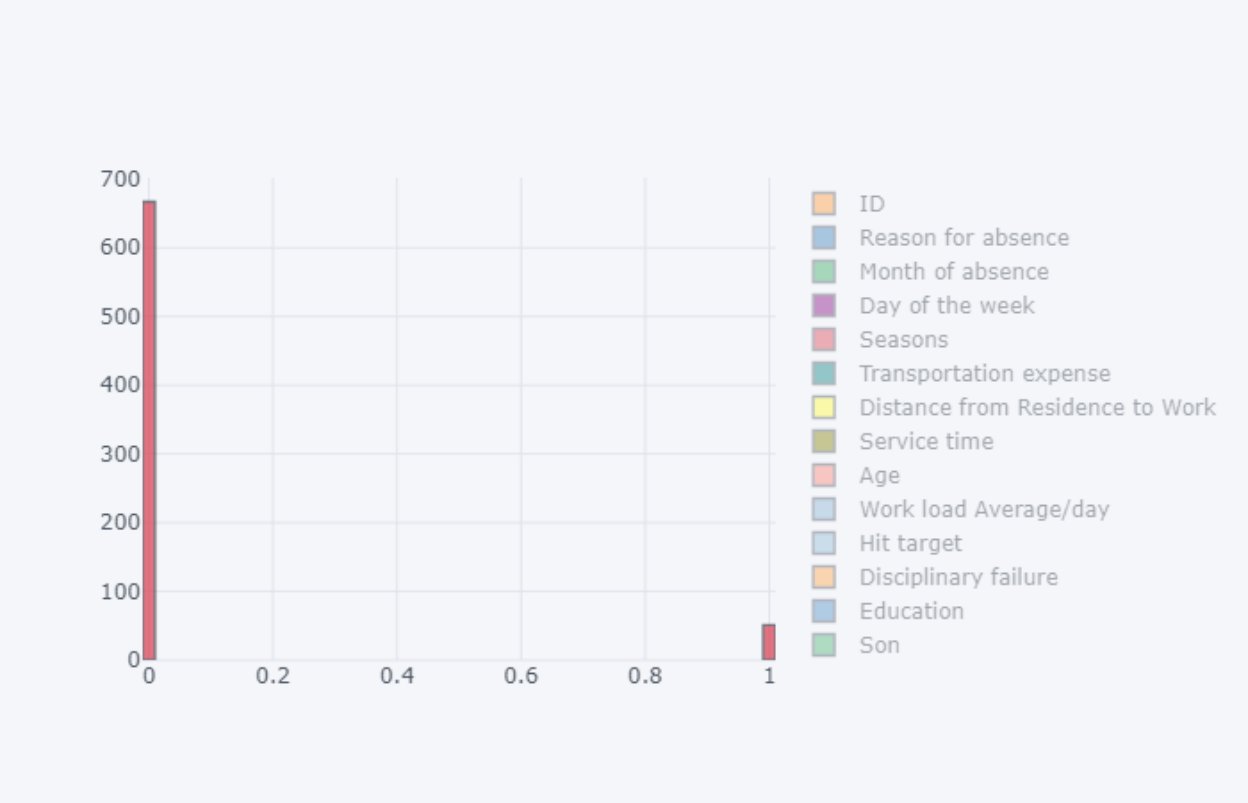
1. It is observed that employee with low education have maximum absentee time.



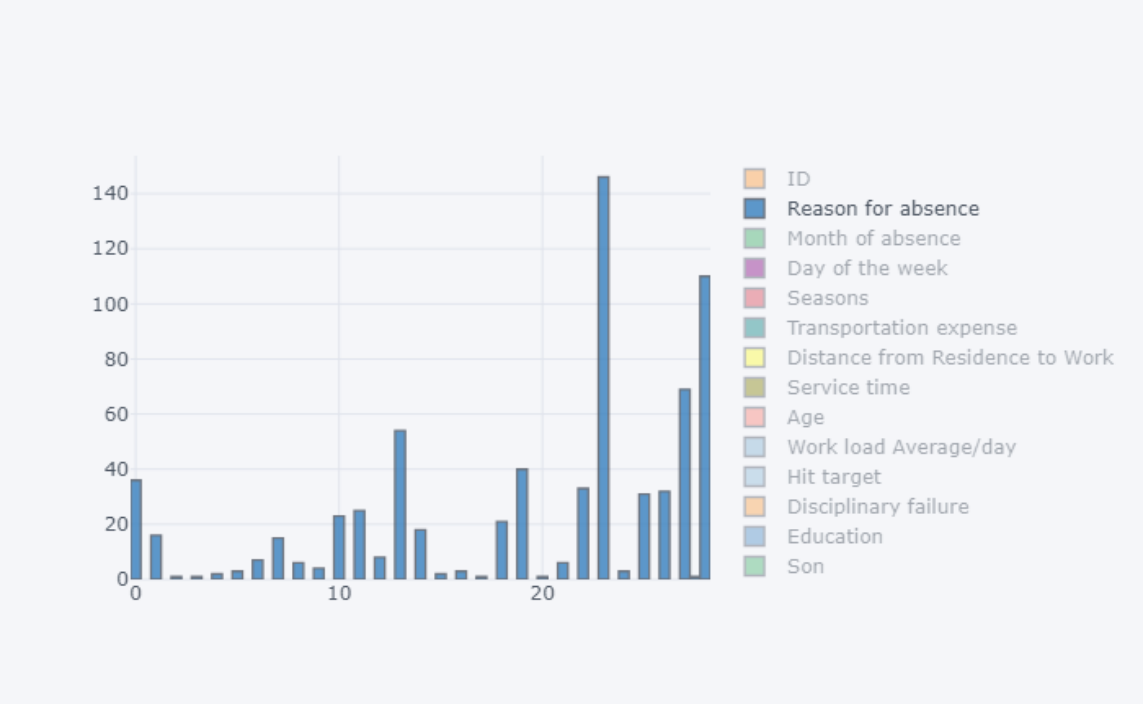
2. Some employee with ID 3, 28, 34 are often absent from work, company should take action against them.



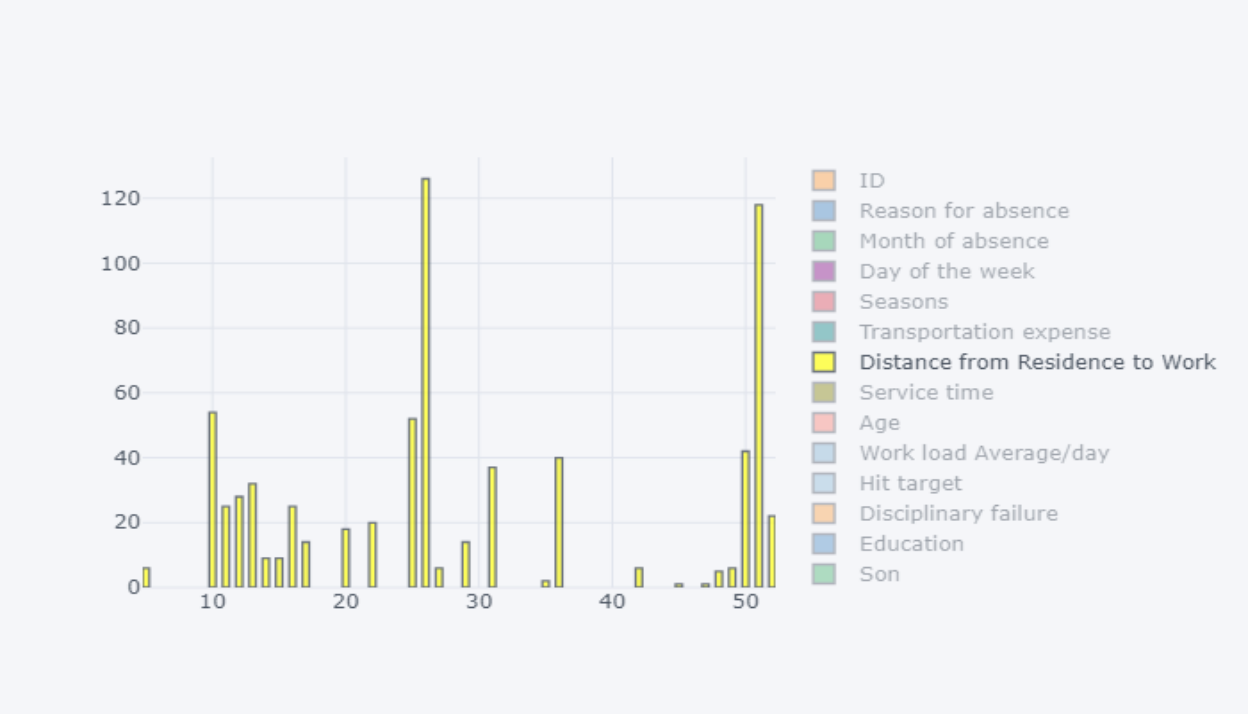
3. Employees who are social smoker have more absentee hour than who are not social smoker.



4. Most often Reason for absence are medical consultation and dental consultation, company should take care of it.



5. Employees who has Distance from Residence to Work high more tends to absent more.

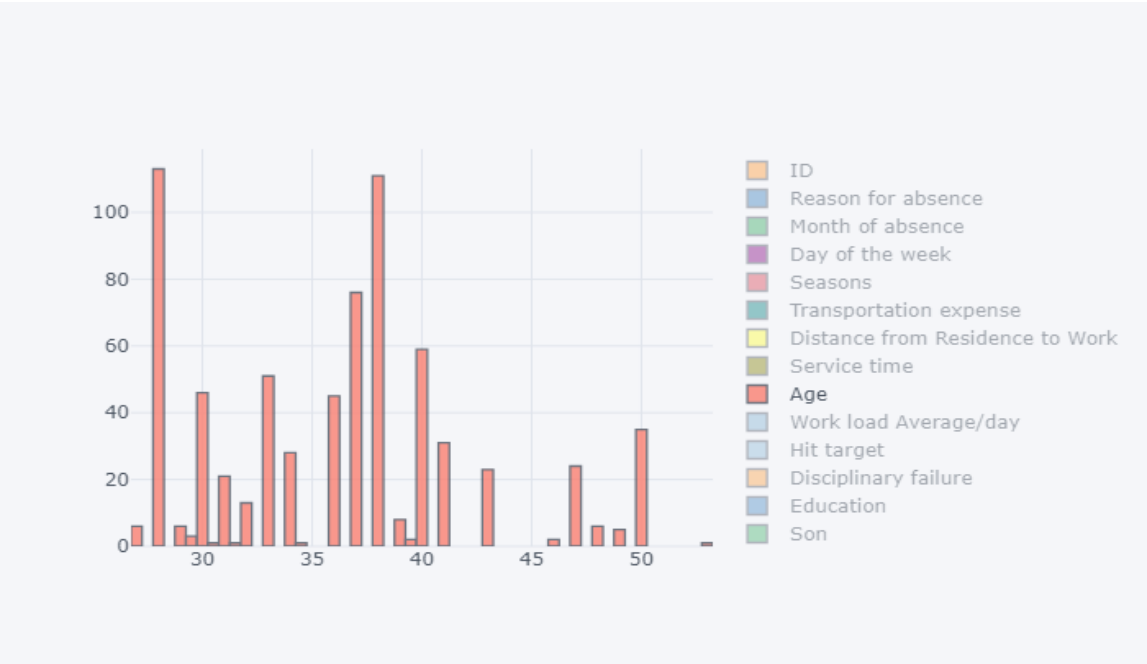


Appendix

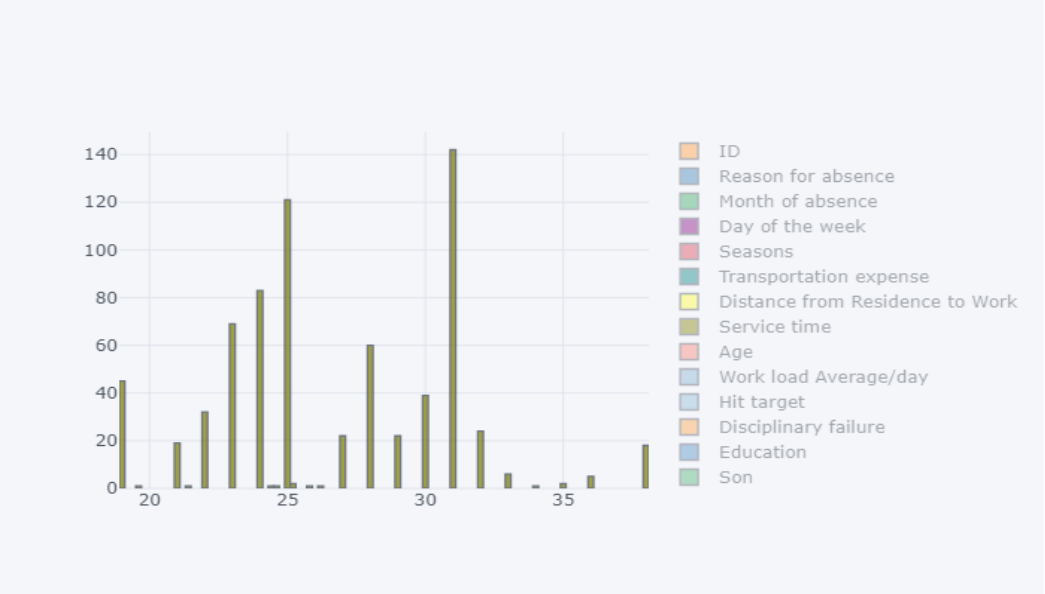
Extra Figures

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

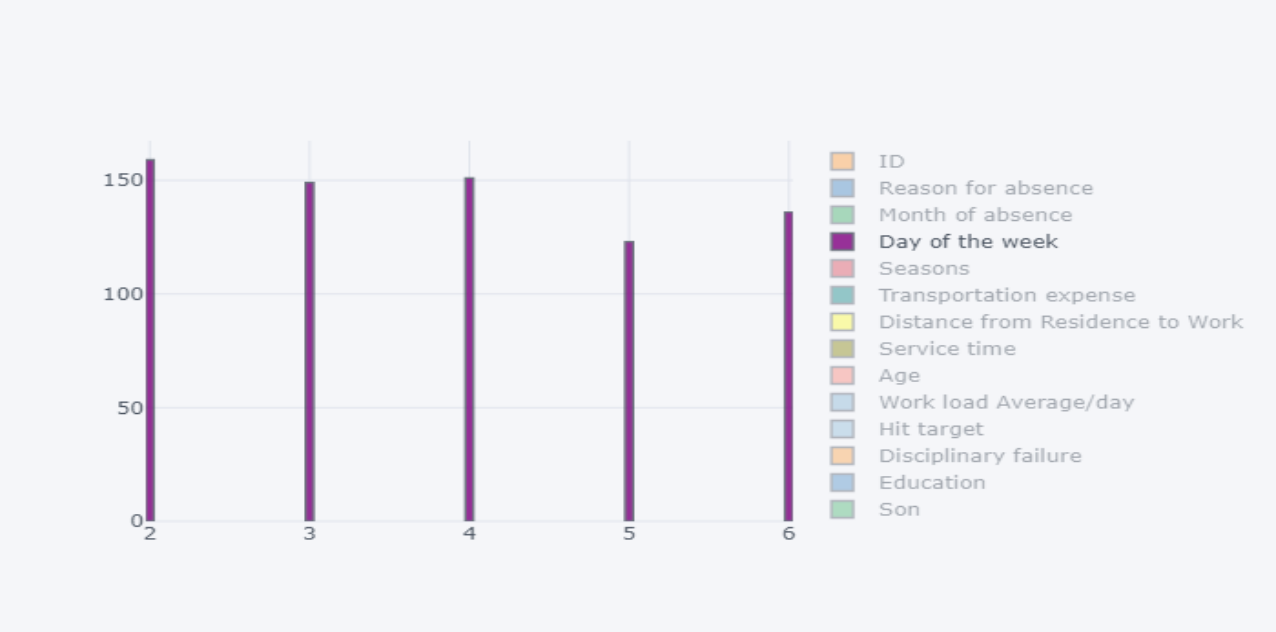
1. With “Age”



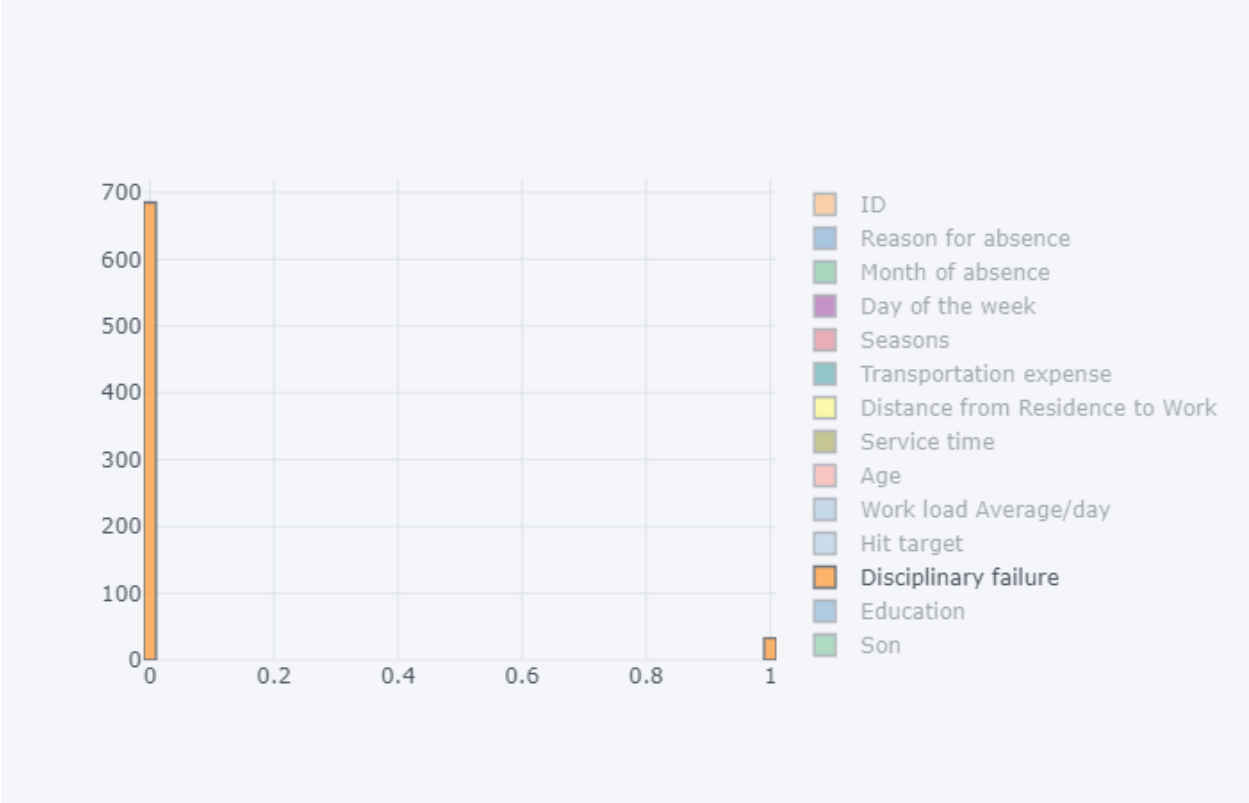
2. With “Body mass index”



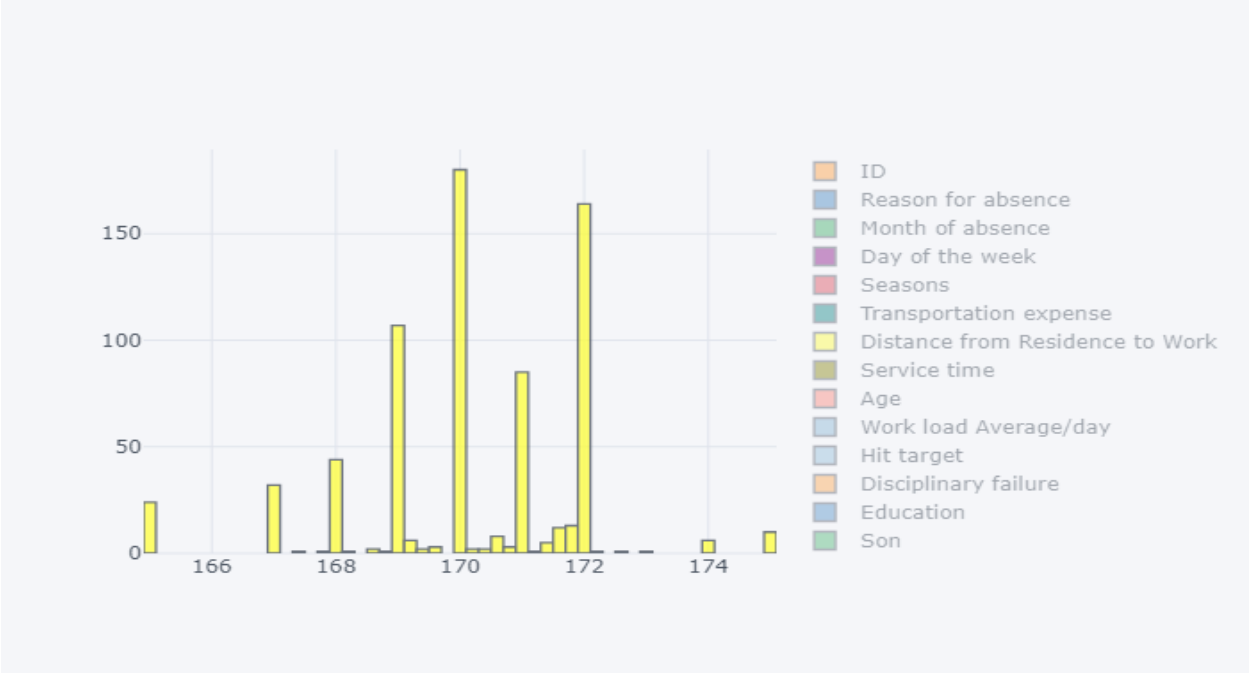
3. With “Day of week”



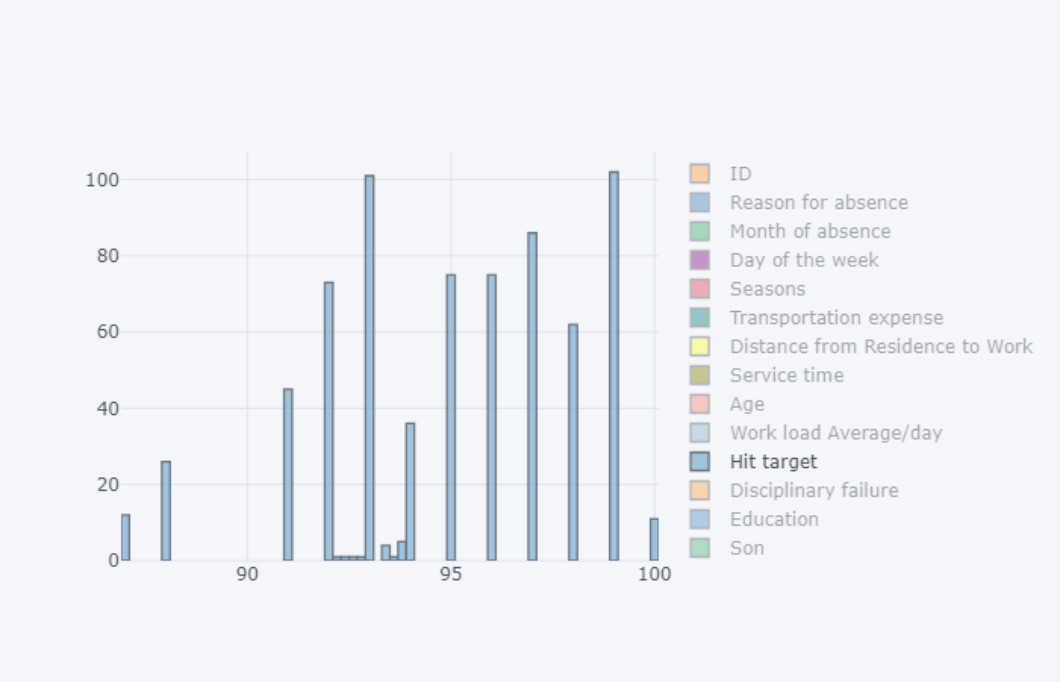
4. With “Disciplinary failure”



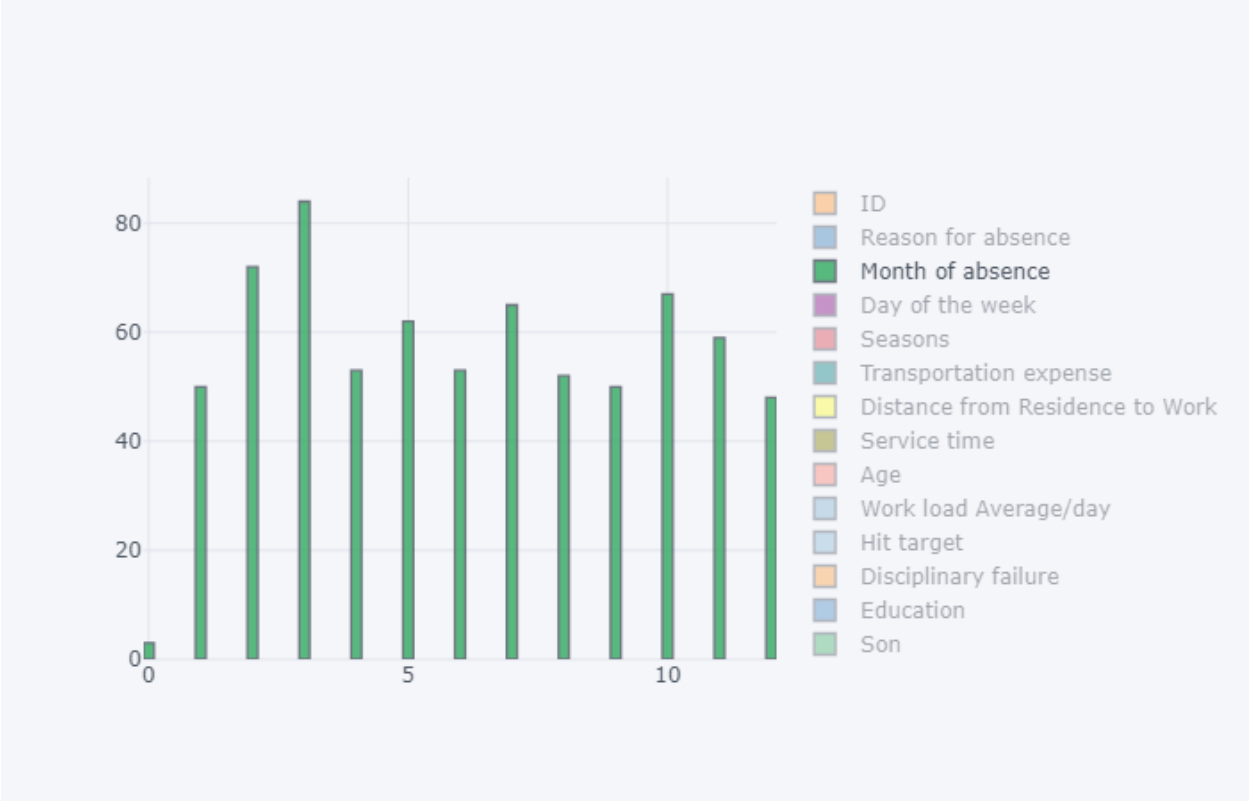
5. With “Height”



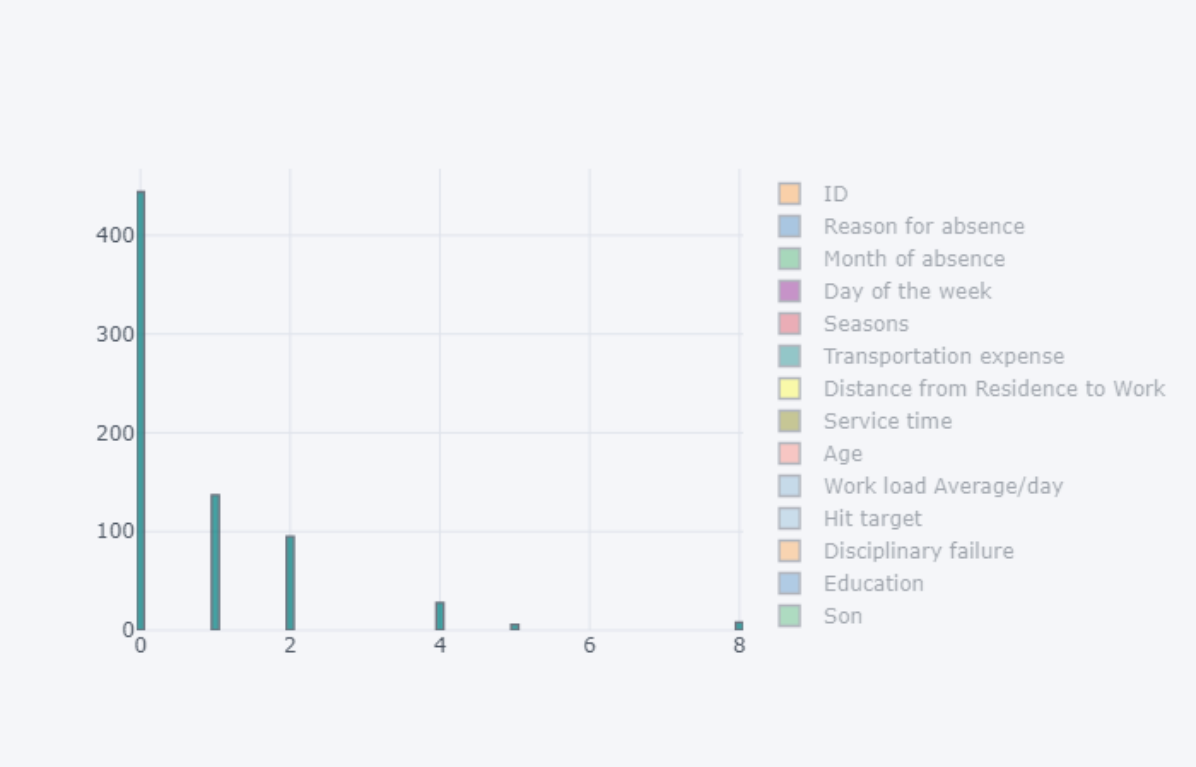
6. With “Hit target”



7. With “Month of absent”

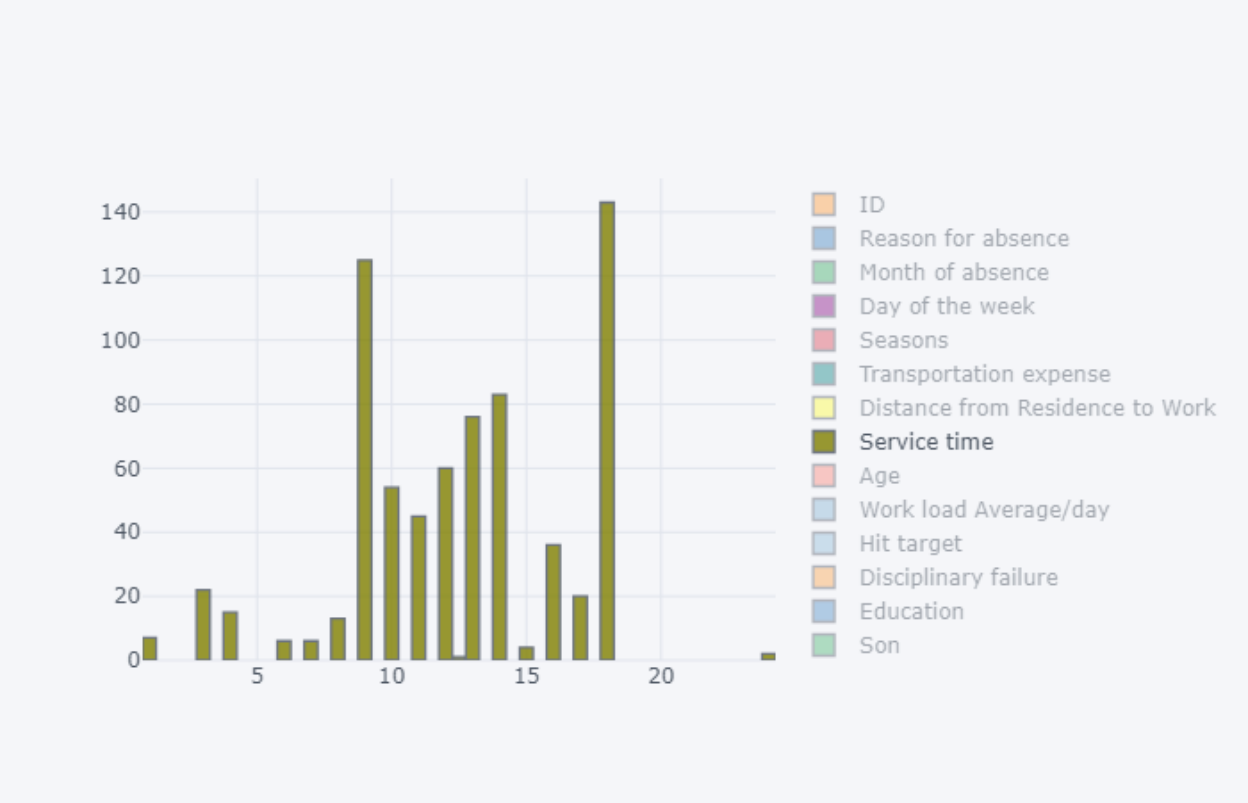


8. With “Pet”

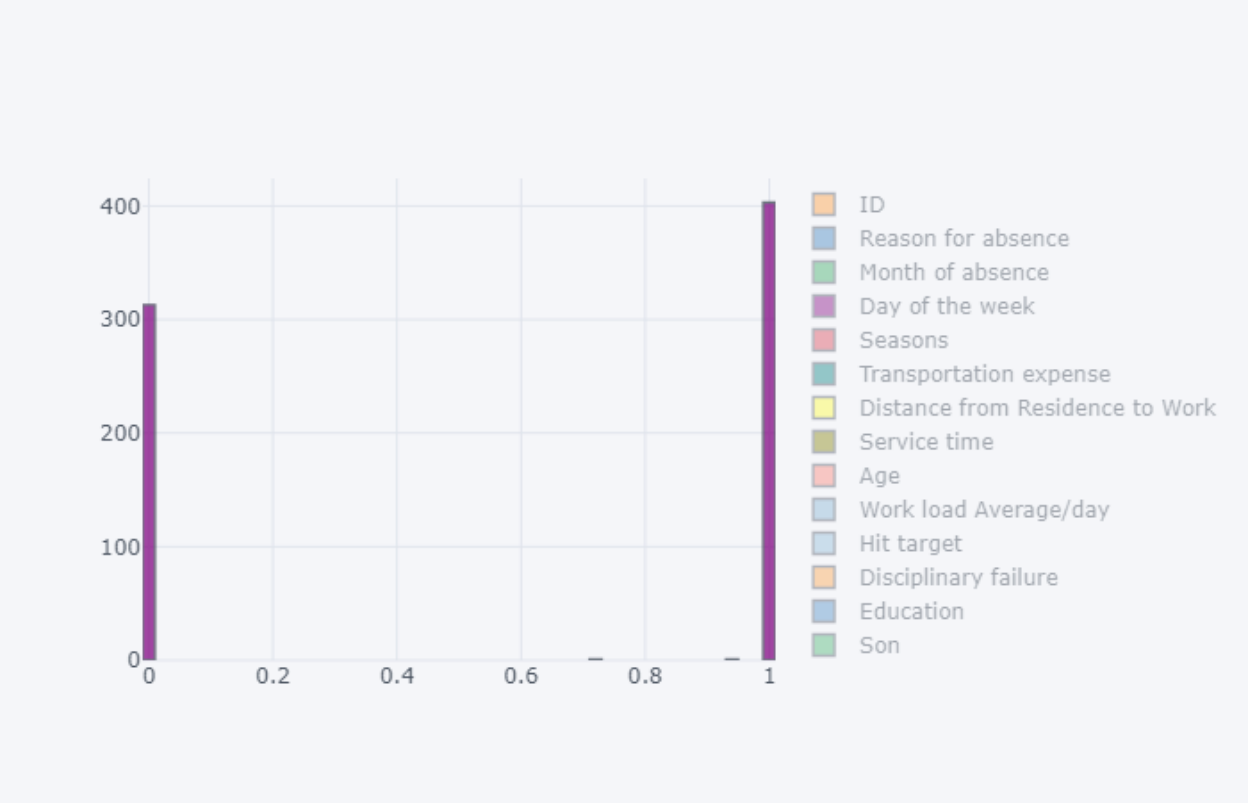


9. With “Seasons”

10. With “Service time”



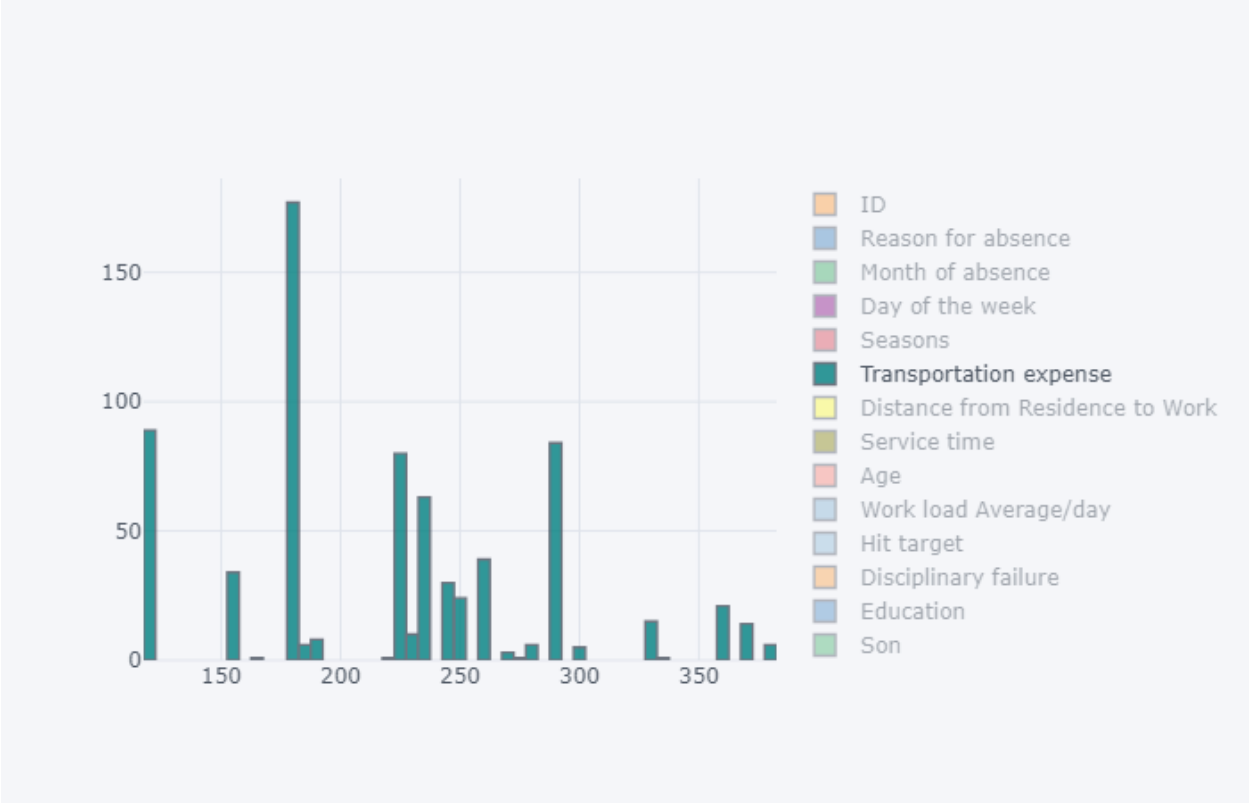
11. With “Social drinker”



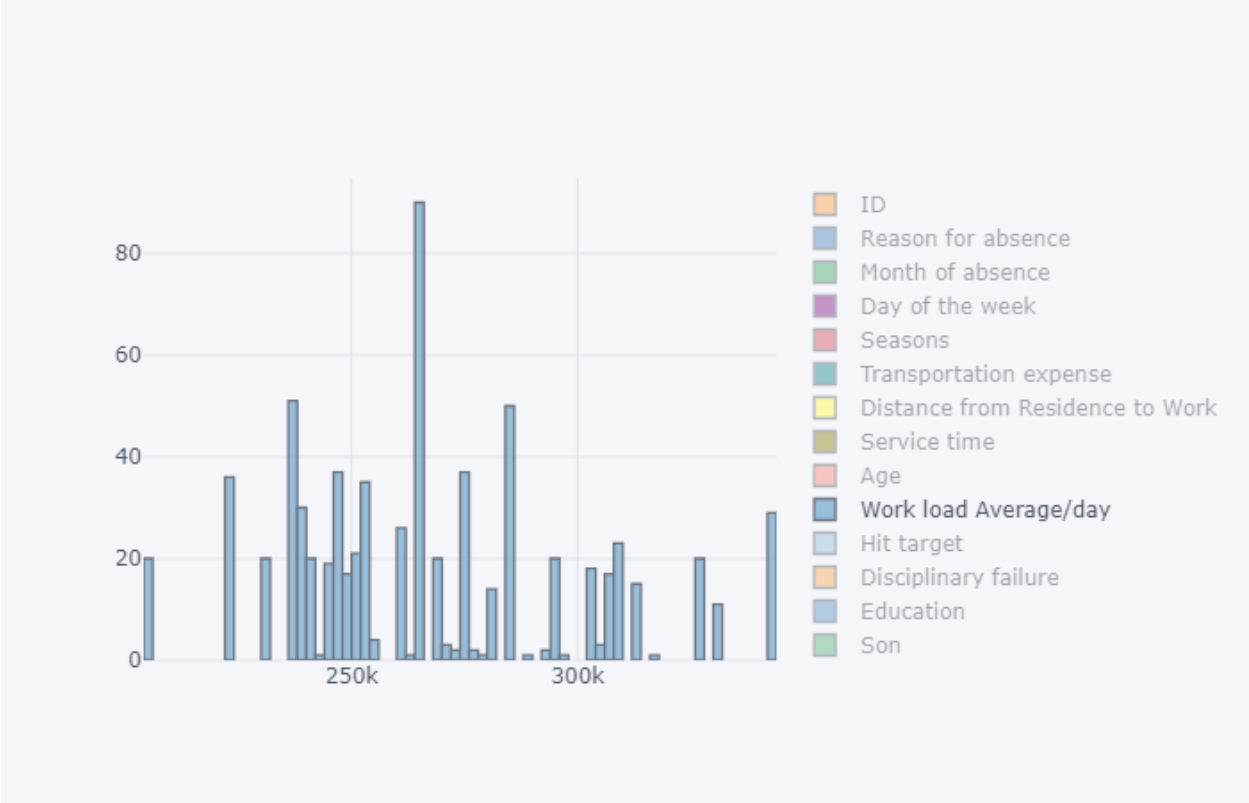
12. With “Son”


300
250
200
150
100
50
0
0
1
2
3
4
ID
Reason for absence
Month of absence
Day of the week
Seasons
Transportation expense
Distance from Residence to Work
Service time
Age
Work load Average/day
Hit target
Disciplinary failure
Education
Son

13. With “Transportation Expense”



14. With “Work load/day”



References

1. For Data Cleaning and Model Development - https://edwisor.com/career-data-scientist
2. For PCA - https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/
3. For Visualization – https://www.udemy.com/python-for-data-science-and-machine-learning-bootcamp/