**EMPLOYEE ABSENTEEISM**

**Chandrakanth Bachu**

**Oct 2018**

**Contents**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **Introduction** | | | **3** |
|  | 1.1 | Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 3 |
|  | 1.2 | Data . | . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 4 |
| **2** | **Methodology** | | | **5** |
|  | 2.1 | Data Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 5 |
|  |  | 2.1.1 | Univariate Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 5 |
|  |  | 2.1.2 | Bivariate Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 6 |
|  |  | 2.1.3 |  |  |
|  | 2.2 | Preprocessing Techinues. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 10 |
|  |  | 2.2.1 | Missing value analysis | 10 |
|  |  | 2.2.2 | Outlier analysis | 11 |
|  |  | 2.2.3 | Feature Selection | 13 |
|  |  | 2.2.4 | Feature Scaling | 14 |
| **3** | **Modelling** | |  | **15** |
|  | 3.1 | Model Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 15 |
|  |  |  | |  |
|  |  |  | |  |
|  | 3.2 | K Means Clustering . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 15 |
|  |  |  | | 19 |
|  |  |  | |  |
|  |  |  | |  |
| Appendix A – Python code | | | | **20** |
| Reference | | | |  |
|  |  |  | |  |
|  |  |  | |  |
|  |  |  | |  |
|  |  |  | |  |
|  |  |  | |  |
|  |  |  | |  |
|  |  |  | |  |
|  | | |  |  |

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. **Data**

Our task is to build the Clustering model to divide the group of Employees who having more absentees and analyze the reason of absentees. Given below is a sample of the data set that we are using to predict the count:

Table 1.1: Churn Reduction sample data(Columns:1- 8)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| state | account length | area code | phone number | international plan | voice mail plan | number vmail messages | total day minutes |
| KS | 128 | 415 | 382-4657 | no | yes | 25 | 265.1 |
| OH | 107 | 415 | 371-7191 | no | yes | 26 | 161.6 |
| NJ | 137 | 415 | 358-1921 | no | no | 0 | 243.4 |
| OH | 84 | 408 | 375-9999 | yes | no | 0 | 299.4 |
| OK | 75 | 415 | 330-6626 | yes | no | 0 | 166.7 |

Table 1.2: Churn Reduction Sample Data (Columns: 9-18)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| total day calls | total day charge | total eve minutes | total eve calls | total eve charge | total night minutes | total night calls | total night charge | total intl minutes |
| 110 | 45.07 | 197.4 | 99 | 16.78 | 244.7 | 91 | 11.01 | 10 |
| 123 | 27.47 | 195.5 | 103 | 16.62 | 254.4 | 103 | 11.45 | 13.7 |
| 114 | 41.38 | 121.2 | 110 | 10.3 | 162.6 | 104 | 7.32 | 12.2 |
| 71 | 50.9 | 61.9 | 88 | 5.26 | 196.9 | 89 | 8.86 | 6.6 |
| 113 | 28.34 | 148.3 | 122 | 12.61 | 186.9 | 121 | 8.41 | 10.1 |

Churn Reduction Sample data (Columns : 18-21-)

|  |  |  |  |
| --- | --- | --- | --- |
| total intl calls | total intl charge | number customer service calls | Churn |
| 3 | 2.7 | 1 | False. |
| 3 | 3.7 | 1 | False. |
| 5 | 3.29 | 0 | False. |
| 7 | 1.78 | 2 | False. |
| 3 | 2.73 | 3 | False. |

Below are the variables we used to predict the analyze the reason for absentees and absentees in every month

Fig 2.1 Variables of Employee Absenteeism

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | S.no | Column Name |
|  |  | 1 | ID |
|  |  | 2 | Reason for absence |
|  |  | 3 | Month of absence |
|  |  | 4 | Day of the week |
|  |  | 5 | Seasons |
|  |  | 6 | Transportation expense |
|  |  | 7 | Distance from Residence to Work |
|  |  | 8 | Service time |
|  |  | 9 | Age |
|  |  | 10 | Work load Average/day |
|  |  | 11 | Hit target |
|  |  | 12 | Disciplinary failure |
|  |  | 13 | Education |
|  |  | 14 | Son |
|  |  | 15 | Social Smoker |
|  |  | 16 | Social drinker |
|  |  | 17 | Pet |
|  |  | 18 | Weight |
|  |  | 19 | Height |
|  |  | 20 | Body mass index |
|  |  | 21 | Absenteeism time in hours |

**Chapter 2**

**Methodology**

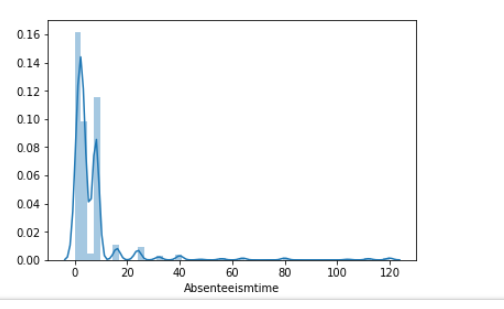
1. **Pre Processing**

Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. To start this process we will first try and look at class imbalance of Target variable in most of the classification class imbalance will create severe problems during the modelling/

**2.1.1 Univariate Analysis**

It seems that 'Absenteeism time in hours' is not normally distributed most of the obsenteeism hours lies between 0-20 hrs and few member having obsentees hours more than 100 also it seems there are outliers are present in the data.

Fig 2.1 Distribution of 'Absenteeism time in hours'

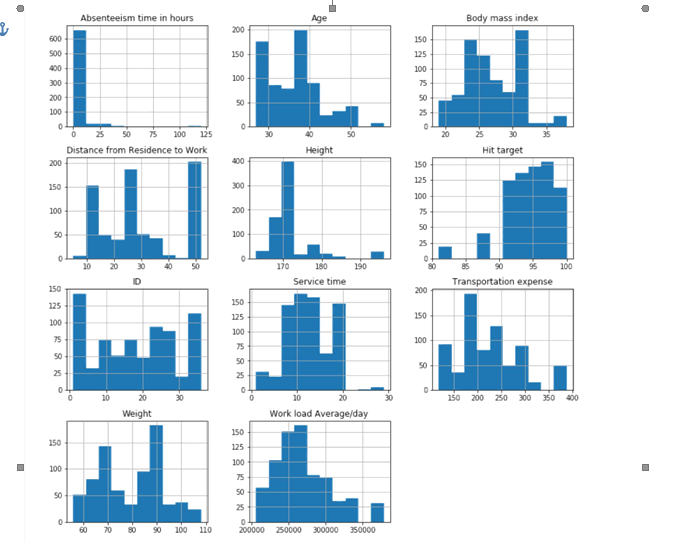


**Distribution of Dependent Numeric Variables :**

In Figure 2.2 it is clearly showing almost all the dependent variables are normally distributed but not perfect normal distributed and outliers are present in the data.

Except few ‘Height ‘ and ‘Weight’

Figure 2.2 showing distribution of dependent Numeric variables (python code in Appendix A)



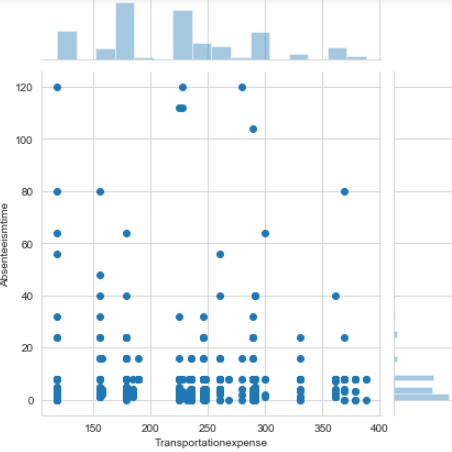
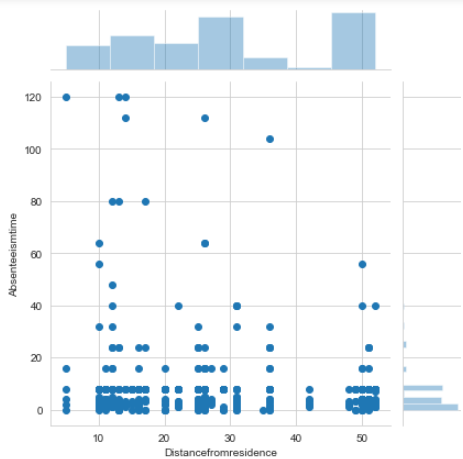
**2.1.2 Bivariate Analysis**

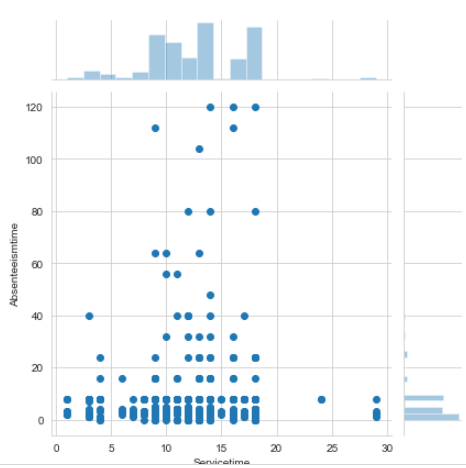
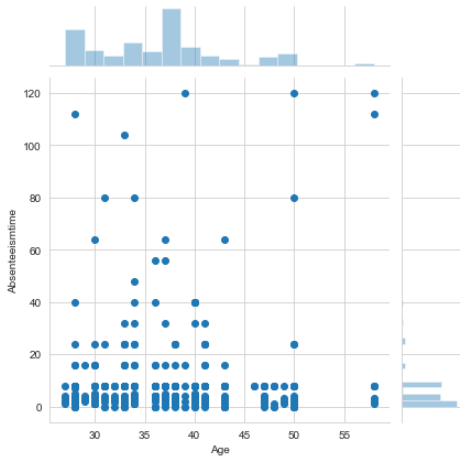
**Relationship between Target Variable “Absenteeism in Hours” and all Numeric variables** :

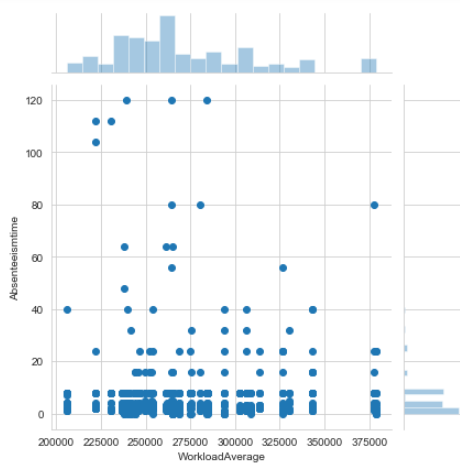
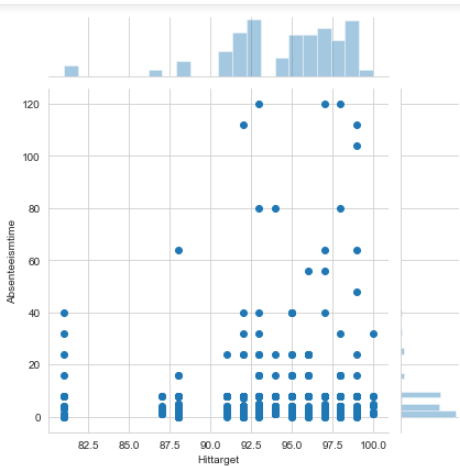
Below Figure 2.3 showing that there are less correlation between Independent Numeric Variables and Target Variable almost all the data points lying between ranges.

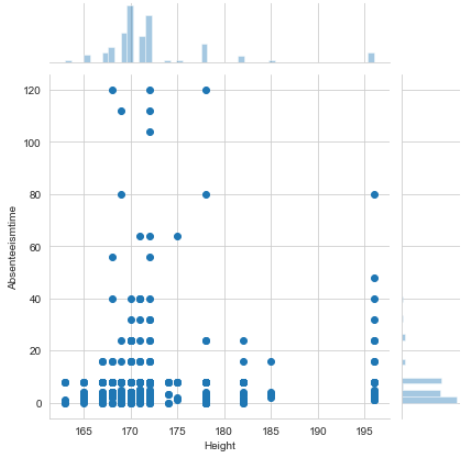
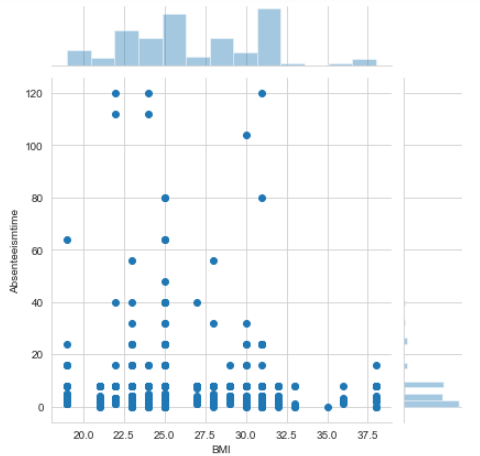
Ex: if we take relation from ‘Absenteeism time’ and ‘Distance from Residence’ than graph is huge data points are lying between obsenteeism time 0 to 20 , it is showing there is less relation between these two variables.And same kind of pattern is following for other variables also.

Figure 2.3 relationship between Numeric variables (python code in Appendix A)



**Relationship between Target Variable Churn and Categorical Variables.**

**Reason For Absence and Absenteeism Time :** there are ICD code 13,19,23,28,22 are the top five reason for absentees.

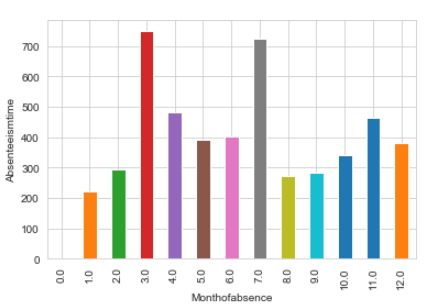
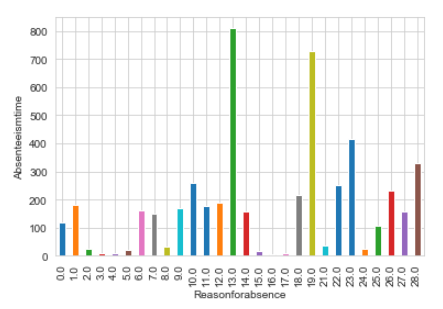
**Month of Absence and Absenteeism Time :**  In March ,April ,July and November months having more Absentees.

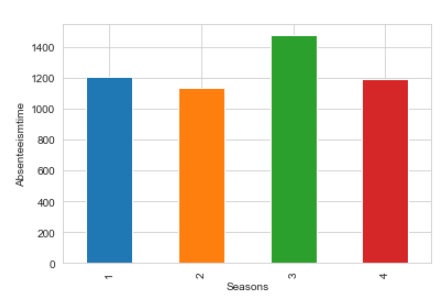
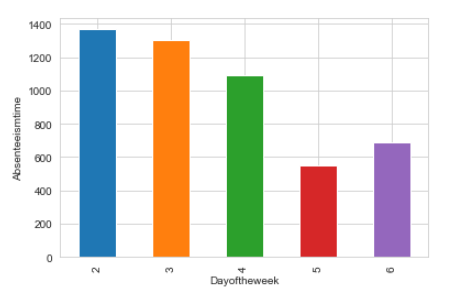
**Days Of Week and Absenteeism Time :**  In Monday , Tuesday and Wednesday having more Absentees.

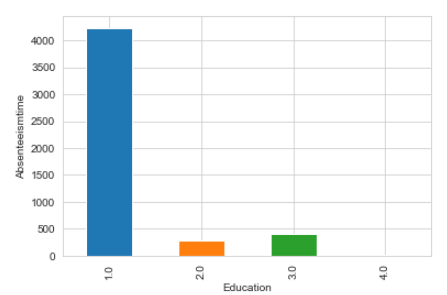
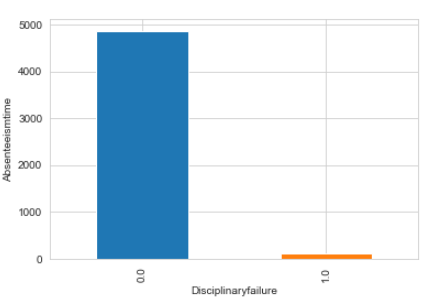
**Son and Absenteeism Time :**  Those who having 0,1 and 2 sons are having more absenteeism rate..

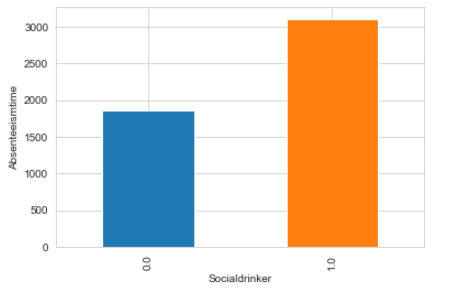
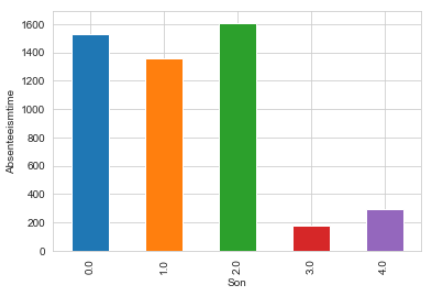
**Pet and Absenteeism Time :**  Those who having 0,1 Pets are having more absenteeism rate..

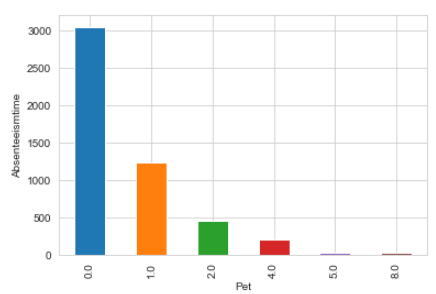
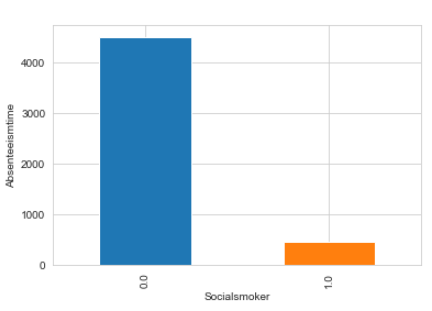
And other Variables like Descliplinary failure , Social Drinker , Social Smoker having class imbalance











**2.2.1 Missing Value Analysis**

Missing values in data is a common phenomenon in real world problems. Knowing how to handle missing values effectively is a required step to reduce bias and to produce powerful models.

Below table illustrate no missing value present in the data.

For this data set Imputed missing values using KNN Imputation Method.

|  |  |  |
| --- | --- | --- |
| S.no | Column Name | Missing Values |
| 1 | ID |  |
| 2 | Reason for absence | 3 |
| 3 | Month of absence | 1 |
| 4 | Day of the week | 0 |
| 5 | Seasons | 0 |
| 6 | Transportation expense | 7 |
| 7 | Distance from Residence to Work | 3 |
| 8 | Service time | 3 |
| 9 | Age | 3 |
| 10 | Work load Average/day | 10 |
| 11 | Hit target | 6 |
| 12 | Disciplinary failure | 6 |
| 13 | Education | 10 |
| 14 | Son | 6 |
| 15 | Social Smoker | 4 |
| 16 | Social drinker | 3 |
| 17 | Pet | 2 |
| 18 | Weight | 1 |
| 19 | Height | 14 |
| 20 | Body mass index | 31 |
| 21 | Absenteeism time in hours | 22 |

**2.2.2 Outlier Analysis**

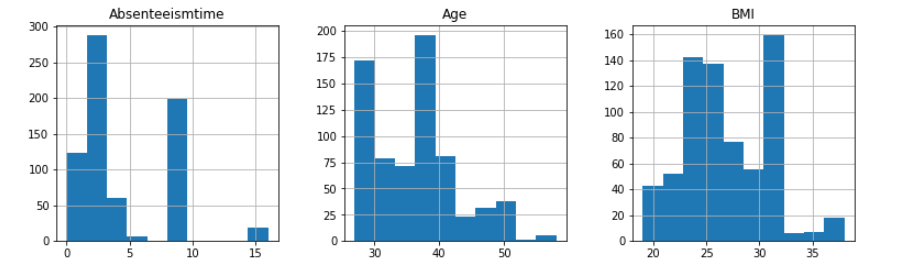
The Other steps of Preprocessing Technique is Outliers analysis , an outlier is an observation point that is distant from other observations. Outliers in data can distort predictions and affect the accuracy, if you don’t detect and handle them appropriately especially in regression models..

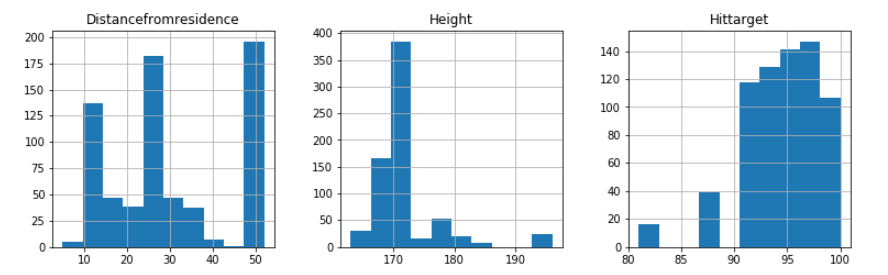
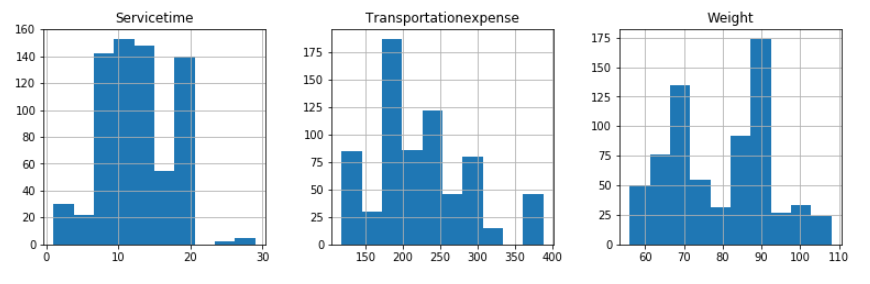
As we are observed in fig 2.2 the data is skewed so, there is chance of outlier in independent variable ‘Total\_Customer\_service\_calls’ ,”number\_Vmail\_messages” and “total\_intl\_calls”

one of the best method to detect outliers is Boxplot

Fig 2.4 shows presence of Outliers in variable ‘casual’ and relationship between ‘casual’ and ‘cnt’ before removing Outliers.

Figure 2.4 Distribution of numeric variables after removing Outliers(Python code in Appendix B)



We are losing almost 5% of data after treating outliers , but still we are going by removing those data since outliers are distributed in almost all the variables it may cause bias to the model.

Boxplot :-  boxplot is a method for graphically depicting groups of numerical data through their [quartiles](https://en.wikipedia.org/wiki/Quartile). Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles

**2.2.3 Features Selections**

Machine learning works on a simple rule – if you put garbage in, you will only get garbage to come out. By garbage here, I mean noise in data.

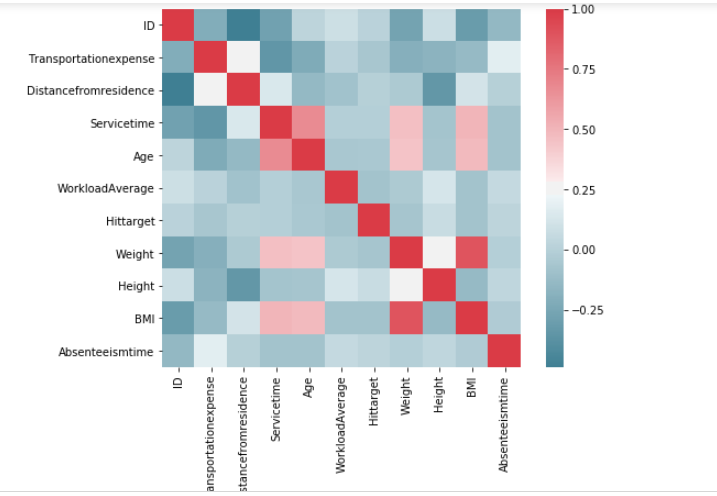
This becomes even more important when the number of features are very large. You need not use every feature at your disposal for creating an algorithm. You can assist your algorithm by feeding in only those features that are really important. I have myself witnessed feature subsets giving better results than complete set of feature for the same algorithm or – “Sometimes, less is better!”.

We should consider the selection of feature for model based on below criteria

1. The relationship between two independent variable should be less and
2. The relationship between Independent and Target variables should be high.

Below fig 2.5 illustrates that relationship between all numeric variables using Corrgram plot .

Figure 2.5 correlation plot of numeric variables (Python code in Appendix A)



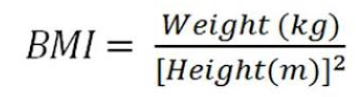
Corrgram : it help us visualize the data in correlation matrices. correlograms are implimented through the **corrgram(x, order = , panel=, lower.panel=, upper.panel=, text.panel=, diag.panel=)**

**2.4.1 Dimensionality Reduction for numeric variables**

Above Fig 2.5 is showing

This plot is showing clearly that relation ship between ‘Height’ , ‘Weight’ and ’BMI’

Hence BMI is depends on Height and Weight than there is no need of Height and Weight variables



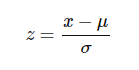
If we observe the figure 2.5 there is almost less relationship between Independent numerical variables and target Variable (Absenteeism in Hours)

**2.2.4 Features Scaling Using Standardization**

Most of the Machine Learning algorithms performance depends on data we are passing through it ,

If two variable are in different ranges than there is chance that Model will bias towards that higher range variable so it is important to Scale Numeric variables in same range.

As we observed in Univariate analysis that there are almost all the variable are normal form so, we are using Standardization(Z - Score) technique to scale the Numeric Variable.



**Chapter 3**

**Modelling**

**3.1 Model Development**

**K Means Clustering**

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

The centroids of the K clusters, which can be used to label new data

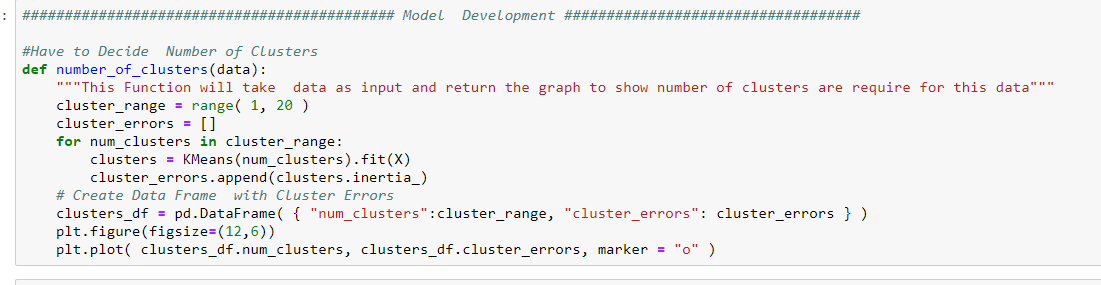
Labels for the training data (each data point is assigned to a single cluster)

**Selection of Number of Clusters**

Selection of number of clusters is a biggest challenge in K- Means clusters

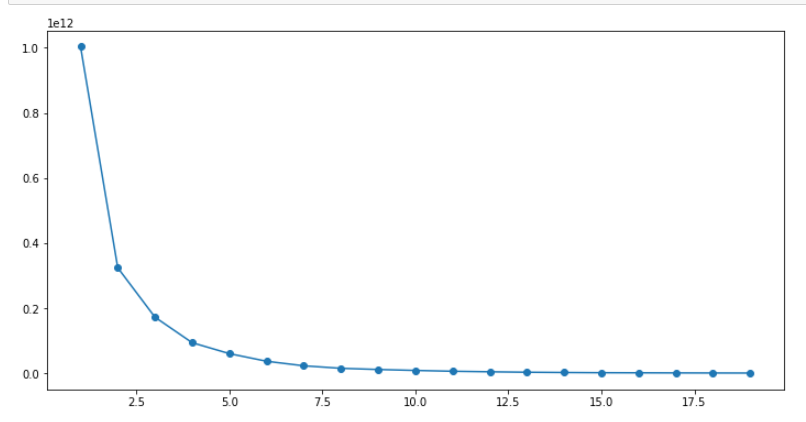
This is an iterative method here we are choosing elbow method to choose the number of clusters (K)

Below code will give the graph of elbow shaped to choose the number of clusters

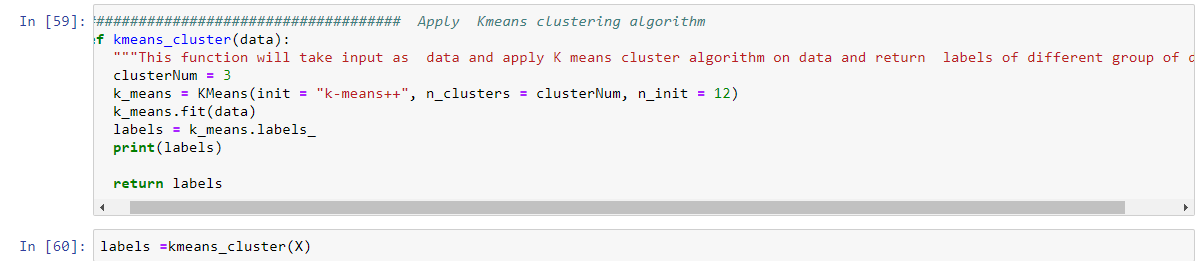


The Below Graph is showing that graph is decrease drastically till more than 2.5 pint and it started decreasing slowly , so number of clusters for this data set are 3

Graph illustrate the Elbow method of Kmeans clustering

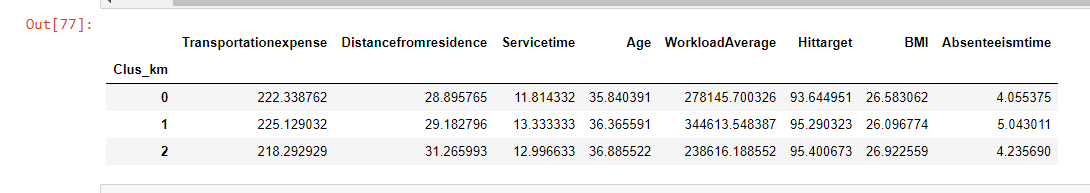


Below figure shows developing of clustering algorithm by selecting number of clusters (K) = 3



**Reason For Absentees based on Numerical Variable**

After assigning this labels to the data and analyze which numeric variable is more average



By above figure it is clearly showing that cluster which is having more average AbseentismTime i.e. 5.04 are having more ‘TransportationExpenses’ ‘serviceTime’ and ‘Workload/day’

As per above figure Company should change below things

1. Should provider Transportation facilities to the Employees
2. Should give less Workload on employees (especially young employess)

**For Categorical Variables**

**Reason for absenteeism and absenteeism time for label =1**

Below figure Illustrate Cluster Label=1 which is having more obsenteeism hours having below reason

**ICD - 1** :- Certain infectious and parasitic diseases

**ICD - 3** :- Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

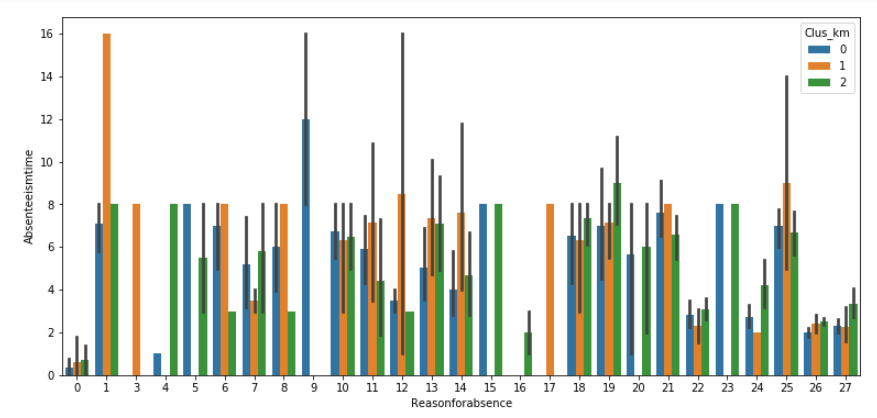
**ICD – 6** :- Diseases of the nervous system

**ICD – 8** :- Diseases of the ear and mastoid process

**ICD – 13** :- Diseases of the musculoskeletal system and connective tissue

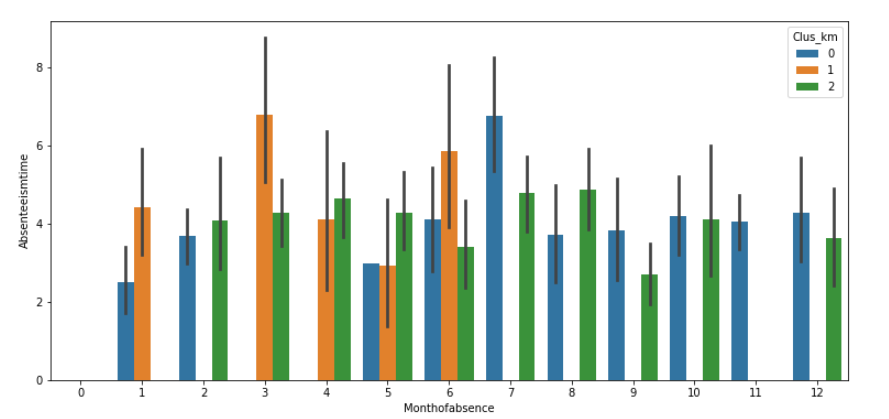
**ICD - 19** : - Injury, poisoning and certain other consequences of external causes

**ICD – 25**:- laboratory examination

* 

**Month of absenteeism and absenteeism time for label =1**

Below figure illustrate most of the absentees are happening in the month of January,March,April,May,June.



**1 .Question**

**What changes company should bring to reduce the number of absenteeism**

Ans :

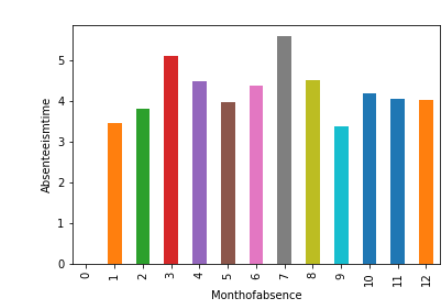
1. Company should **decrease** the **Service Time** and **Work Load** of Employees(**Especially the young Employees**)
2. Company should Provide **Doctor Consultant in Office** Especially in **Spring and Summer Season** since most of the employees suffering with **infectious diseases** in **winter and Spring Season** and might be because of **high work pressure** most of the employees are suffering with **ICD – 6** ( Diseases of the nervous system) mostly during summer season.

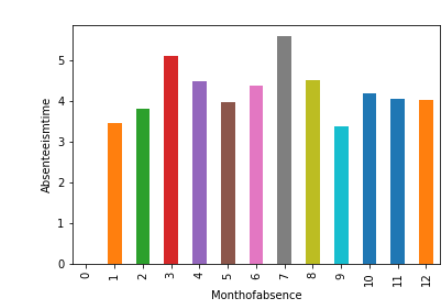
**2. Question**

How much losses every month can we project in 2011 if same trend of absenteeism continues?

Ans :

|  |  |  |
| --- | --- | --- |
|  | Month of absence | average Absenteeism Time in hours per day |
|  | 1 | 4.44 |
|  | 2 | 4.083333 |
|  | 3 | 8.647727 |
|  | 4 | 9.09434 |
|  | 5 | 6.21875 |
|  | 6 | 7.518519 |
|  | 7 | 10.895522 |
|  | 8 | 5.148148 |
|  | 9 | 5.528302 |
|  | 10 | 5.028571 |
|  | 11 | 7.539683 |
|  | 12 | 7.857143 |





**Appendix A**

**import** **pandas** **as** **pd** *# to do summary operations on data*

**import** **numpy** **as** **np** *# to do mathematical calculations on data*

**import** **os** *# to interact with local system directories*

**import** **matplotlib.pyplot** **as** **plt** *# for plotting*

**import** **seaborn** **as** **sns** *# for better plotting*

**import** **sys** *# To Interact with System folders/libraries*

**from** **scipy.stats** **import** chi2\_contingency *# FOr Chi square Test*

os.path.dirname(sys.executable) *# To Interact with System folders/libraries*

*#from fancyimpute import KNN # Impute Missing values*

**from** **sklearn.cluster** **import** KMeans *## To Implement K Means Clustering*

In [102]:

**def** change\_data\_type(data,col\_names,convert\_type):

*"""" This function will take the data frame and columns and conversion type as input*

*will give the converted columns from one datatype to another datatype*

*"""*

**for** col **in** col\_names:

**print**(col , "before convert" , data[col].dtype)

data[col] = data[col].astype(convert\_type)

**print**(col,"after convert",data[col].dtype)

**return** data

*# """ This function will take the data frame and Numerical columns as input and*

*# will give scatter plot with regression line and correlationship as out[put to show the relationship between two numerical variables*

*# """*

**def** joint\_plots(data,numeric\_columns,target\_column):

**for** col **in** numeric\_columns:

fig = plt.figure(figsize=(3,3)) *# define plot area*

ax=sns.jointplot(x=col,y=target\_column,data=df\_obsent)

*#ax = sns.regplot(x="Age", y='Absenteeismtime',data=df\_obsent)*

plt.show()

**def** reg\_plots(data,numeric\_columns,target\_column):

*""" This function will take the data frame and Numerical columns as input and*

*will give scatter plot with regression line and correlationship as out[put to show the relationship between two numerical variables*

*"""*

**for** col **in** numeric\_columns:

core=data[[col,target\_column]].corr()

**print**(core)

fig = plt.figure(figsize=(6,6)) *# define plot area*

ax = sns.regplot(x=col, y=target\_column,data=df\_obsent)

plt.show()

**def** plot\_bar(data,cat\_columns,col\_y):

*""" This plot will take Categorical variables and numerica target variable as input and give*

*out put as barplot which categorical variable on x axis and numerical variable on Y axis"""*

**for** col **in** cat\_columns:

data.groupby(col)[col\_y].sum().plot.bar()

*#df\_obsent.groupby(col)[col\_y].reset\_index().sort\_values(col\_y).plot.bar()*

plt.xlabel(col) *# Set text for the x axis*

plt.ylabel(col\_y)*# Set text for y axis*

plt.show()

**def** plot\_box(data, cols, col\_x):

*"""" This function will display the box plot, to show relationship between numeric*

*variables(Cols) and and target categorical variable (Col\_x)*

*"""*

**for** col **in** cols:

sns.set\_style("whitegrid")

sns.boxplot(col\_x, col, data=data)

plt.xlabel(col\_x) *# Set text for the x axis*

plt.ylabel(col)*# Set text for y axis*

plt.show()

**def** fun\_numeric\_relation(data):

*"""" This function will give output of plot of relationship between numeric variables in data frame """*

f, ax = plt.subplots(figsize=(8, 6))

corr = data.corr()

sns.heatmap(corr, mask=np.zeros\_like(corr, dtype=np.bool), cmap=sns.diverging\_palette(220, 10, as\_cmap=True), square=True, ax=ax)

**def** encoding\_categorical(data , cat\_columns):

*""" This function will take take dataframe and cat\_columns as input and turn those categorical values into encoding form*

*0's and 1' if columns is not category it will convert this into category and encode the values*

*"""*

**for** col **in** cat\_columns:

data[col]=data[col].astype("category")

data[col]=data[col].cat.codes

**return** data

**def** standardform\_convert(data ,numeric\_columns):

*""" This functin will take input as data frame and numerical columns and convert those numerical data into standardization*

*form and gives output and converted data frame"""*

**for** i **in** numeric\_columns:

**print**(i)

data[i] = (data[i] - data[i].mean())/data[i].std()

**return** data

**def** group\_plot(data , cat\_columns,target\_col):

*""" This function will show the relationship between two categorical variables in grouped bar chart"""*

**for** col **in** cat\_columns:

**print**(col ,"and",target\_col," target Variable")

carat\_table = pd.crosstab(index=data[col],columns=data[target\_col])

*#print(carat\_table)*

carat\_table.plot(kind="bar", figsize=(10,10),stacked=False)

**def** encoding\_categorical(data , cat\_columns):

*""" This function will take take dataframe and cat\_columns as input and turn those categorical values into encoding form*

*0's and 1' if columns is not category it will convert this into category and encode the values*

*"""*

**for** col **in** cat\_columns:

data[col]=data[col].astype("category")

data[col]=data[col].cat.codes

**return** data

**def** treat\_outlier(data,numeric\_columns):

*""" This function will take the input as data frame and numeric values and return output dataframe*

*after treating the outliers"""*

**for** i **in** numeric\_columns:

**print**(i)

q75, q25 = np.percentile(data.loc[:,i], [75 ,25])

iqr = q75 - q25

mini = q25 - (iqr\*1.5)

maxi = q75 + (iqr\*1.5)

**print**(mini)

**print**(maxi)

data\_outlier = data.drop(data[data.loc[:,i] < mini].index)

data\_outlier = data.drop(data[data.loc[:,i] > maxi].index)

**return** data\_outlier

**def** number\_of\_clusters(data):

*"""This Function will take data as input and return the graph to show number of clusters are require for this data"""*

cluster\_range = range( 1, 20 )

cluster\_errors = []

**for** num\_clusters **in** cluster\_range:

clusters = KMeans(num\_clusters).fit(X)

cluster\_errors.append(clusters.inertia\_)

*# Create Data Frame with Cluster Errors*

clusters\_df = pd.DataFrame( { "num\_clusters":cluster\_range, "cluster\_errors": cluster\_errors } )

plt.figure(figsize=(12,6))

plt.plot( clusters\_df.num\_clusters, clusters\_df.cluster\_errors, marker = "o" )

**def** kmeans\_cluster(data):

*"""This function will take input as data and apply K means cluster algorithm on data and return labels of different group of data"""*

clusterNum = 3

k\_means = KMeans(init = "k-means++", n\_clusters = clusterNum, n\_init = 12)

k\_means.fit(data)

labels = k\_means.labels\_

**print**(labels)

**return** labels

In [ ]:

*# train data from .csv file*

*# getting and setting current working directories*

os.getcwd()

os.chdir("D:/Edwisor assignments/obsenteeism/")

os.getcwd()

*#get the list of files in the directy*

**print**(os.listdir(os.getcwd()))

In [ ]:

*############### Load the Data ##############################*

df\_obsent = pd.read\_excel('Absenteeism\_at\_work\_Project.xls', sheet\_name="Absenteeism\_at\_work")

In [ ]:

df\_obsent.columns = ["ID","Reasonforabsence" ,"Monthofabsence","Dayoftheweek","Seasons","Transportationexpense" ,"Distancefromresidence"

,"Servicetime","Age","WorkloadAverage","Hittarget","Disciplinaryfailure","Education","Son", "Socialdrinker"

,"Socialsmoker","Pet","Weight","Height","BMI","Absenteeismtime"]

df\_obsent.columns

In [ ]:

*# Under standing data*

df\_obsent.head()

*# Summary Of Data*

df\_obsent.info()

*# this data set contains 740 rows and 21 columns out of this 21 columns nine columns are categorical and remaining*

*#columns are Numeric*

In [ ]:

*# Creating the copy of Data Frame*

df\_obsentees = df\_obsent.copy()

cat\_variables = ['Reasonforabsence','Monthofabsence','Dayoftheweek','Seasons','Disciplinaryfailure','Education','Son','Socialdrinker','Socialsmoker','Pet']

*# Turn above Variables to categorical*

df\_obsent = change\_data\_type(df\_obsent,cat\_variables,'category')

numeric\_variables = df\_obsent.select\_dtypes(exclude=['object','category']).columns

*# Check the variable data types after conversion*

df\_obsent.info()

In [ ]:

*#descriptive statistics summary*

df\_obsent['Absenteeismtime'].describe()

A = df\_obsent['Absenteeismtime']

Anan=A[~np.isnan(A)] *# Remove the NaNs*

*#Check whether target variable is normallly Distributed or not*

sns.distplot(Anan)

In [ ]:

pd.DataFrame.hist(df\_obsent.loc[:,numeric\_variables], figsize = [13,13]);

*#In Below figure it is showing that Valriables 'Age','BMI','distanceFromResidence','Hittarget','Eervicetime',*

*#'Transportationexpenses' , 'Weight' and 'WorkloadAverage' are looking bit normally distributed but not perfect normal distribution still few outliers are present in each variable*

In [ ]:

*# ################ Univariate Analysis for Categorival Variables*

create\_frequncy\_tables\_plot(df\_obsent,cat\_variables)

In [ ]:

*############################################ Bibvariate Relationship between Numerical Variable ################*

*# Analyse realationship using Joint Plot*

joint\_plots(df\_obsent,numeric\_variables,'Absenteeismtime')

*# Analyse Relationship using Reg Plot*

reg\_plots(df\_obsent,numeric\_variables,'Absenteeismtime')

In [ ]:

*######################## Bivariate Relationship btween categorical Variable and Target Variable ##########*

plot\_bar(df\_obsent,cat\_variables,'Absenteeismtime')

In [ ]:

*############################### Impute Missing Values ###########################*

*#Impute Missing values for Numerical COlumns*

**for** col **in** numeric\_variables:

df\_obsent[col] = df\_obsent[col].fillna(df\_obsent[col].median())

**for** col **in** cat\_variables:

df\_obsent[col] = df\_obsent[col].fillna(df\_obsent[col].mode()[0])

In [ ]:

df\_obsent.isnull().sum()

In [ ]:

*####################################### Outlier Analysis ####################################*

numeric\_variables

num\_col= ['Transportationexpense','Distancefromresidence',

'Servicetime','Age','Hittarget','Weight',

'Height','BMI','Absenteeismtime']

df\_outlier =treat\_outlier(df\_obsent,num\_col)

In [ ]:

*# shape of data after treating Outliers*

df\_outlier.shape

*# We lost 5% of data*

*# Verify the distribution of numeric variable after treating utliers*

pd.DataFrame.hist(df\_outlier.loc[:,num\_col], figsize = [13,13]);

In [ ]:

numeric\_variables

num\_col= ['Transportationexpense','Distancefromresidence',

'Servicetime','Age','Hittarget','Weight',

'Height','BMI','Absenteeismtime']

In [ ]:

*########################################### Feature Engineering ###############################*

fun\_numeric\_relation(df\_outlier)

*# If we observe the figure 2.5 there is almost less relationship between Independent numerical variables and target Variable (Absenteeism in Hours)*

In [ ]:

*#This plot is showing clearly that relation ship between ‘Height’ , ‘Weight’ and ’BMI’*

*#Hence BMI is depends on Height and Weight than there is no need of Height and Weight variables*

df\_outlier = df\_outlier.drop(["Height","Weight"],axis=1)

*# Verify the shape of data frame*

df\_outlier.shape

In [ ]:

*################################ Encode Categorical variables*

encoding\_categorical(df\_outlier,cat\_variables)

In [ ]:

*################################# Standarddised Numerical Variables ###############################*

*#SInce Most of the data is normally distributes Except few outliers so going for the Standardization*

**from** **sklearn.preprocessing** **import** StandardScaler

X = df\_outlier.values[:,1:]

X = np.nan\_to\_num(X)

Clus\_dataSet = StandardScaler().fit\_transform(X)

*#Clus\_dataSet*

X

In [ ]:

*############################################ Model Development ###################################*

*#Have to Decide Number of Clusters*

In [ ]:

*### choose number of clusters*

number\_of\_clusters(X)

*#below graph showing that graph is decrease drastically when it is greater than 2.5 and it slow down from than*

*# an elbow shape is there at nearly 3*

*# Number of clusters for this data is 3*

In [ ]:

*###################################### Apply Kmeans clustering algorithm*

**def** kmeans\_cluster(data):

*"""This function will take input as data and apply K means cluster algorithm on data and return labels of different group of data"""*

clusterNum = 3

k\_means = KMeans(init = "k-means++", n\_clusters = clusterNum, n\_init = 12)

k\_means.fit(data)

labels = k\_means.labels\_

**print**(labels)

**return** labels

In [ ]:

labels =kmeans\_cluster(X)

In [ ]:

*####### Analyse the labels by appending to the data*

*#Asign These Labels to the original DataFrame*

df\_outlier["Clus\_km"] = labels

df\_obsent.head(5)

df\_outlier["Clus\_km"].value\_counts()

In [ ]:

*# Verify the average of each columns values by group with clusters labels*

numeric\_col= ['Transportationexpense','Distancefromresidence','Servicetime','Age','WorkloadAverage','Hittarget','BMI','Absenteeismtime']

df\_outlier.groupby('Clus\_km')[numeric\_col].mean()

In [ ]:

*# Reasoni for Obsentees analyse for categorical variables*

*#y\_c0l= df\_obsent['Absenteeismtime'].mean()*

plt.figure(figsize=(13,6))

sns.barplot(data=df\_outlier, x="Reasonforabsence", y="Absenteeismtime", hue="Clus\_km")

In [ ]:

*# Reasoni for Obsentees analyse for categorical variables*

*#y\_c0l= df\_obsent['Absenteeismtime'].mean()*

plt.figure(figsize=(13,6))

sns.barplot(data=df\_outlier, x="Monthofabsence", y="Absenteeismtime", hue="Clus\_km")

*#Monthofabsence Dayoftheweek*

In [ ]:

*#y\_c0l= df\_obsent['Absenteeismtime'].mean()*

plt.figure(figsize=(13,6))

sns.barplot(data=df\_outlier, x="Socialdrinker", y="Absenteeismtime", hue="Clus\_km")

*#Monthofabsence Dayoftheweek*

In [ ]:

*################*

*#What changes company should bring to reduce the number of absenteeism*

*#Ans :*

*#i. Company should decrease the Service Time and Work Load of Employees(Especially the young Employees)*

*#ii. Company should Provide Doctor Consultant in Office Especially in Spring and Summer Season since most of the employees suffering with infectious diseases in winter and Spring Season and might be because of high work pressure most of the employees are suffering with ICD – 6 ( Diseases of the nervous system) mostly during summer season.*

In [ ]:

*###2. Question*

*##How much losses every month can we project in 2011 if same trend of absenteeism continues?*

df\_outlier.groupby('Monthofabsence')['Absenteeismtime'].mean().plot.bar()

*#df\_obsent.groupby(col)[col\_y].reset\_index().sort\_values(col\_y).plot.bar()*

plt.xlabel('Monthofabsence') *# Set text for the x axis*

plt.ylabel('Absenteeismtime')*# Set text for y axis*

plt.show()

In [ ]:

df\_obsent.groupby('Monthofabsence')['Absenteeismtime'].mean()