Employee Absenteeism

Muddassir Mohammed

24-Jul-2018

CONTENTS

Chapter 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 Used

Abstract:

This dataset contains records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil.

Number of Instances: 740

Number of Attributes: 21

1.3 Attribute 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. Work load 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)

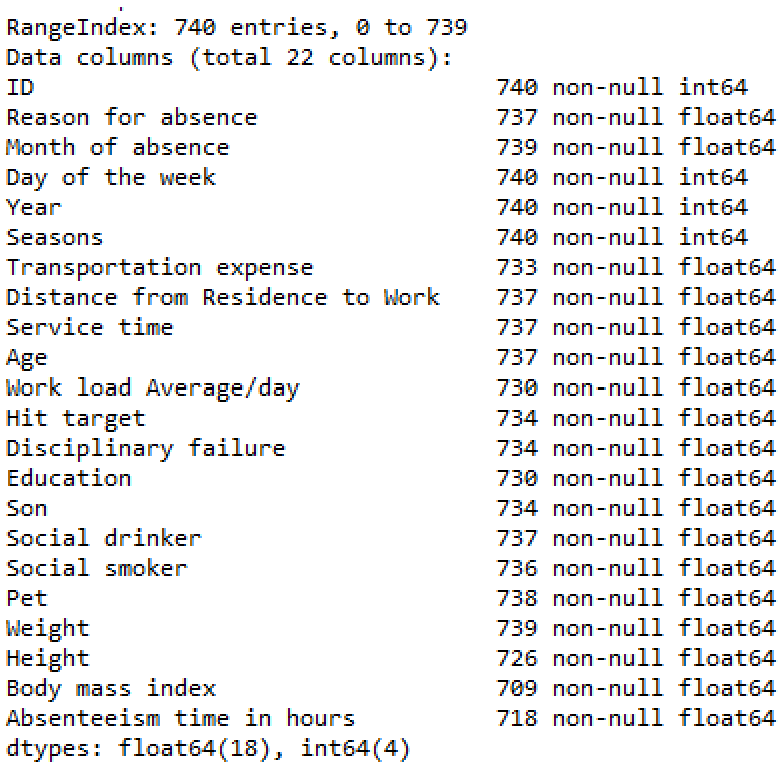
1.3 Hypothesis:

- Persons who have more BMI are most absentees.
- Persons who are having large family are most absentees
- Absenteeism increases with age
- Smoking Habits are more prone to diseases
- More Absentees due more transportation cost
- High Workload/More Work pressure leads to more absentees
- More Absentees on festival seasons
- More Absentees in winter and Summer Seasons
- .More Absentees on start and end Week

Let us test these hypothesis with the help of exploratory data analysis

Chapter 2

Exploratory Data Analysis:

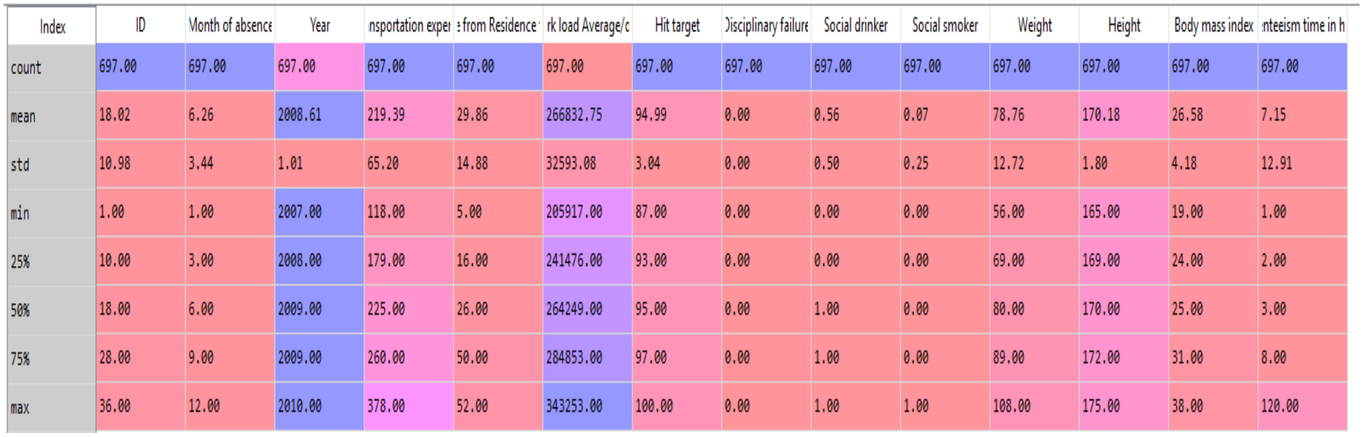


```
RangeIndex: 740 entries, 0 to 739
Data columns (total 22 columns):
ID                                 740 non-null int64
Reason for absence                 737 non-null float64
Month of absence                   739 non-null float64
Day of the week                    740 non-null int64
Year                               740 non-null int64
Seasons                            740 non-null int64
Transportation expense             733 non-null float64
Distance from Residence to Work    737 non-null float64
Service time                       737 non-null float64
Age                                737 non-null float64
Work load Average/day              730 non-null float64
Hit target                         734 non-null float64
Disciplinary failure               734 non-null float64
Education                          730 non-null float64
Son                                734 non-null float64
Social drinker                     737 non-null float64
Social smoker                      736 non-null float64
Pet                                738 non-null float64
Weight                             739 non-null float64
Height                             726 non-null float64
Body mass index                    709 non-null float64
Absenteeism time in hours          718 non-null float64
dtypes: float64(18), int64(4)
```

We can see that there are total 740 rows and 22 columns with datatypes integer and float

Here we are exploring numerical variables after imputation with missing values.

2.1 Exploration of Numerical Variables:



Index	ID	Month of absence	Year	nsportation exper	: from Residence	rk load Average/c	Hit target	Disciplinary failure	Social drinker	Social smoker	Weight	Height	Body mass index	nteeism time in h
count	697.00	697.00	697.00	697.00	697.00	697.00	697.00	697.00	697.00	697.00	697.00	697.00	697.00	697.00
mean	18.02	6.26	2008.61	219.39	29.86	266832.75	94.99	0.00	0.56	0.07	78.76	170.18	26.58	7.15
std	10.98	3.44	1.01	65.20	14.88	32593.08	3.04	0.00	0.50	0.25	12.72	1.80	4.18	12.91
min	1.00	1.00	2007.00	118.00	5.00	205917.00	87.00	0.00	0.00	0.00	56.00	165.00	19.00	1.00
25%	10.00	3.00	2008.00	179.00	16.00	241476.00	93.00	0.00	0.00	0.00	69.00	169.00	24.00	2.00
50%	18.00	6.00	2009.00	225.00	26.00	264249.00	95.00	0.00	1.00	0.00	80.00	170.00	25.00	3.00
75%	28.00	9.00	2009.00	260.00	50.00	284853.00	97.00	0.00	1.00	0.00	89.00	172.00	31.00	8.00
max	36.00	12.00	2010.00	378.00	52.00	343253.00	100.00	0.00	1.00	1.00	108.00	175.00	38.00	120.00

There are total 697 values

Mean and Median are almost for most of the variables which are uniformly distributed.

But Weight, Height, Body Mass Index and absenteeism time are non-uniformly distributed

Disciplinary failure has only one value which can be eliminated.

Transportation expense ranges from 118 to 378 BRL(Brazil Currency)

Distance from residences ranges from 5 to 52 km

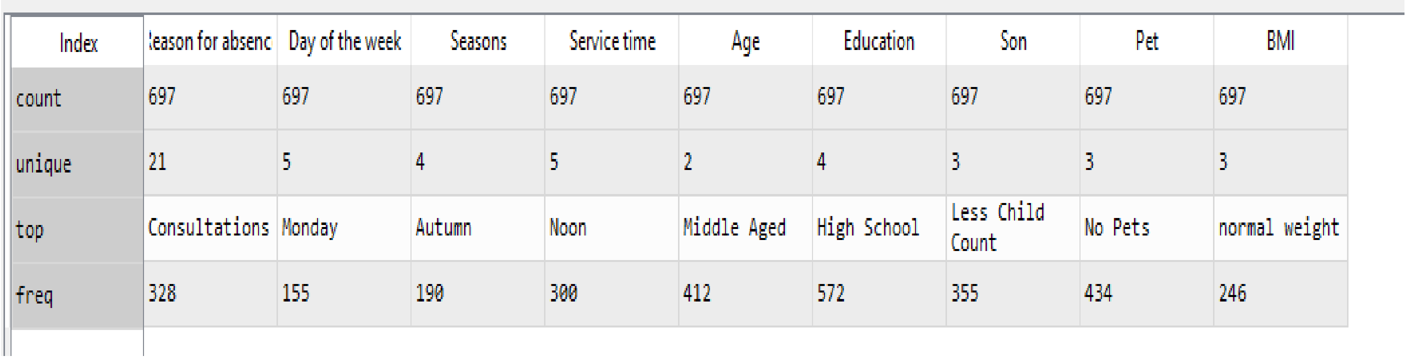
Absenteeism time ranges from 1 to 120 hours

There are total 36 employees who suffered from several diseases.

Year ranges from 2007 to 2010

Months ranges from 1 to 12

2.2 Exploration of Categorical Variables:



Index	eason for absenc	Day of the week	Seasons	Service time	Age	Education	Son	Pet	BMI
count	697	697	697	697	697	697	697	697	697
unique	21	5	4	5	2	4	3	3	3
top	Consultations	Monday	Autumn	Noon	Middle Aged	High School	Less Child Count	No Pets	normal weight
freq	328	155	190	300	412	572	355	434	246

We can see that

There are more employees who were absent due to medical consultations.

Frequency of Mondays are more

More absenteeism in autumn season

More absenteeism at Noon

More Middle Aged group in the dataset

More High School Children

Frequency of employees having children 1 or 2 are more

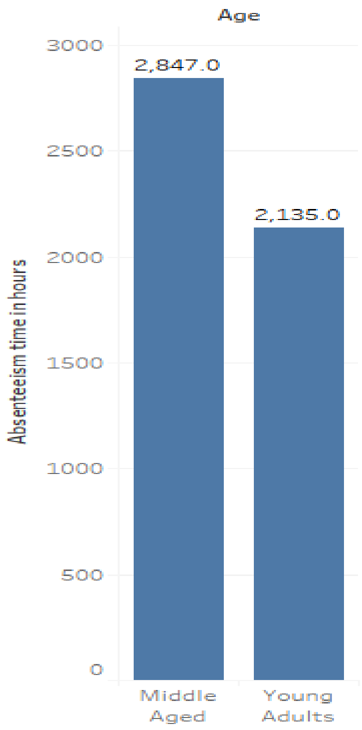
Most of them have no pets

There are more persons who have good Body Mass Index

Let us explore the data by answering some basic questions.

2.3 Testing Hypothesis

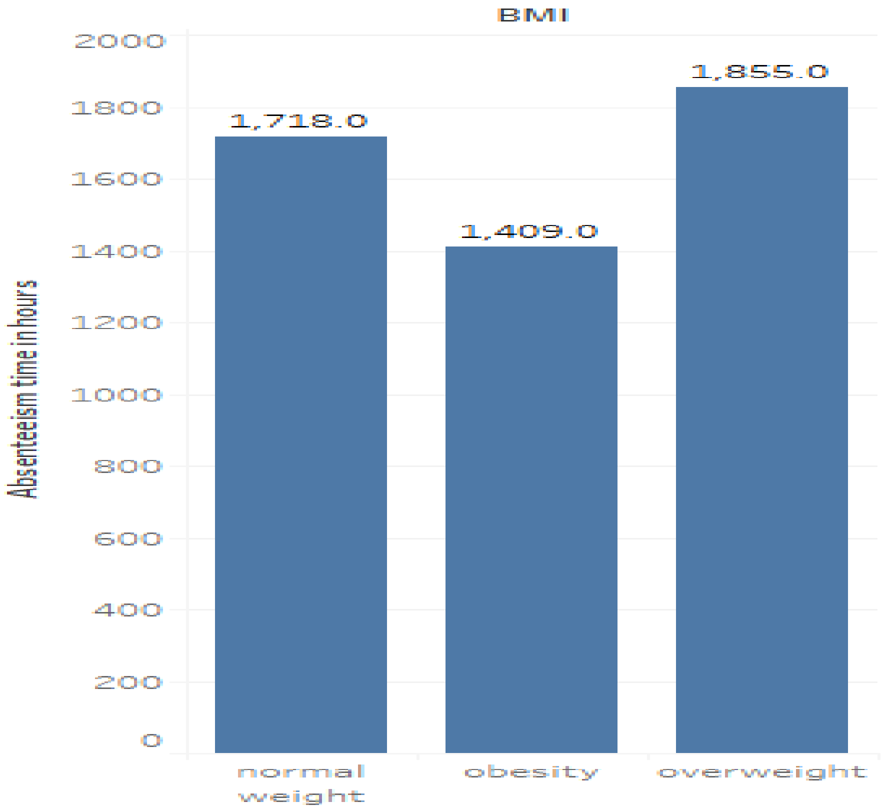
Does Age impact absenteeism?



We can see that Middle Aged (34-55) have more absenteeism.

But there is nothing much drastic change of absenteeism between Middle Aged persons and Young Adults.

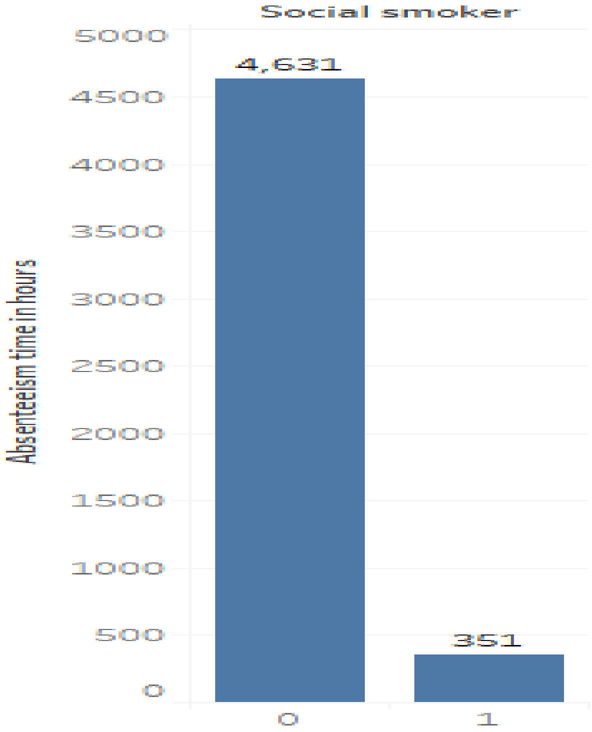
Does BMI of person impact Absenteeism?

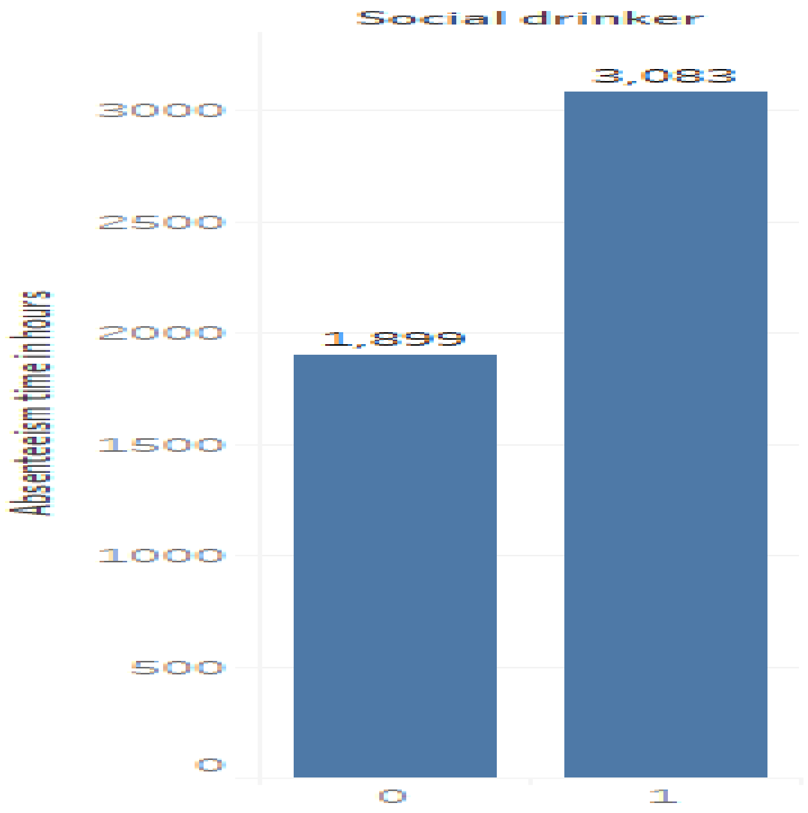


There are more persons who are unhealthy and they are having more absenteeism.

There are more unhealthy people, may be overweight results in several other diseases which in turn results in more absenteeism

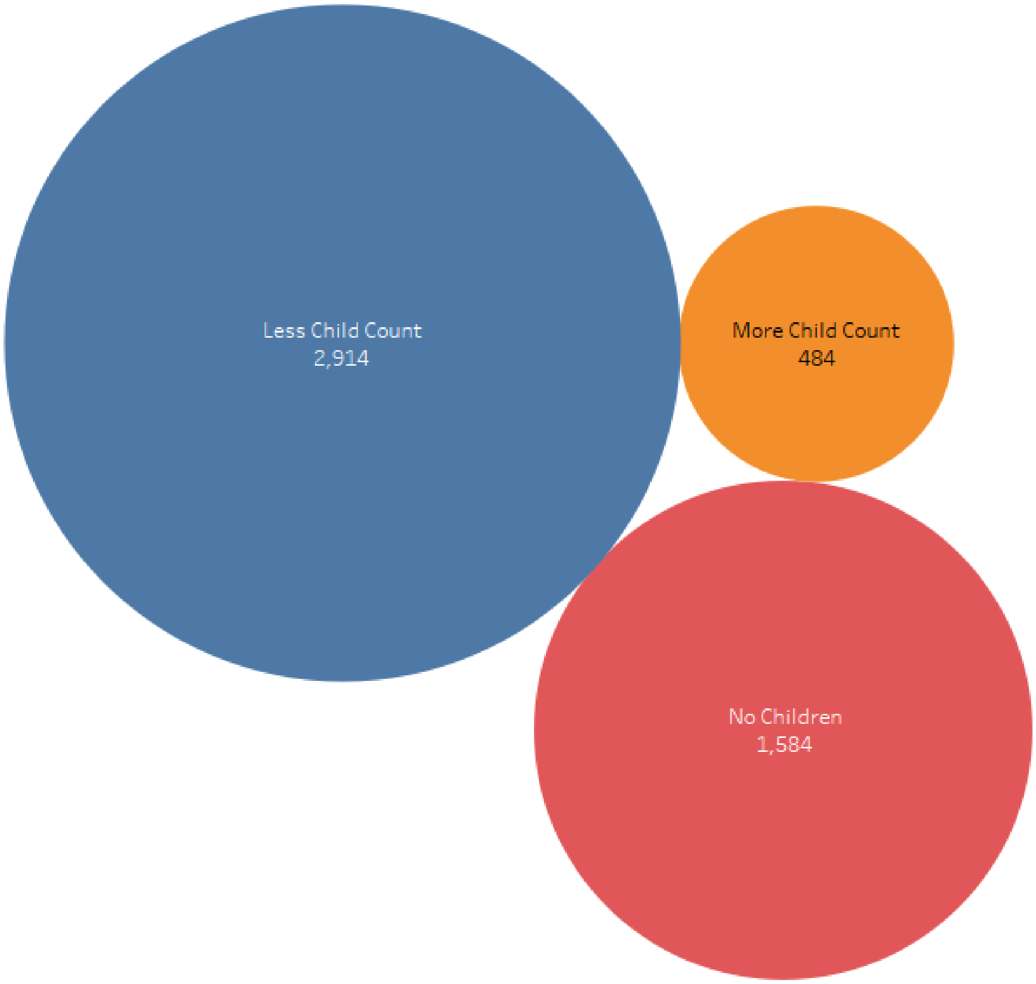
Does Smoking and Drinking habits impact Absenteeism?





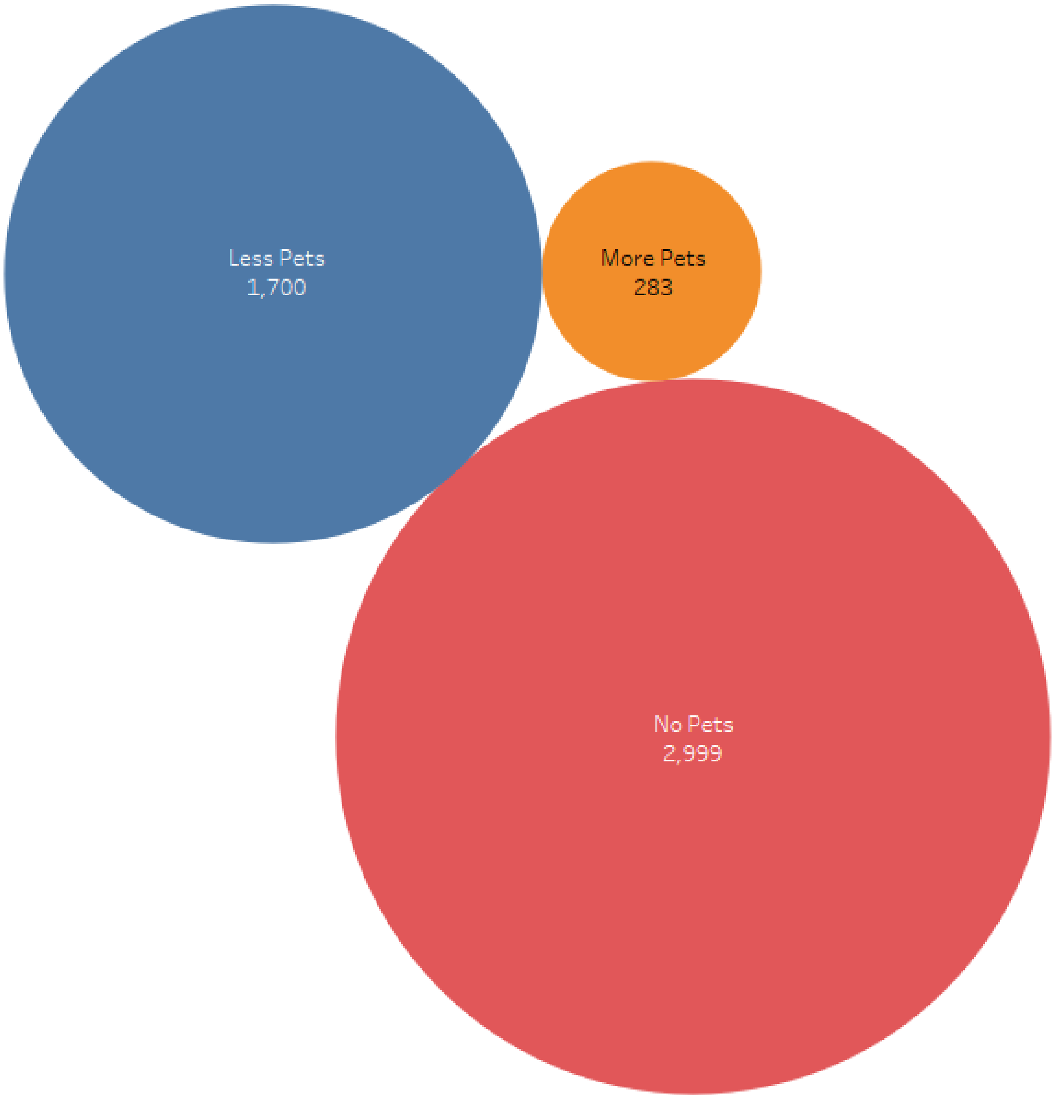
We can see that Social Smokers are more prone to absenteeism whereas social drinker doesn't impact much.

Does Large Family and More number of Pets Impact Absenteeism?



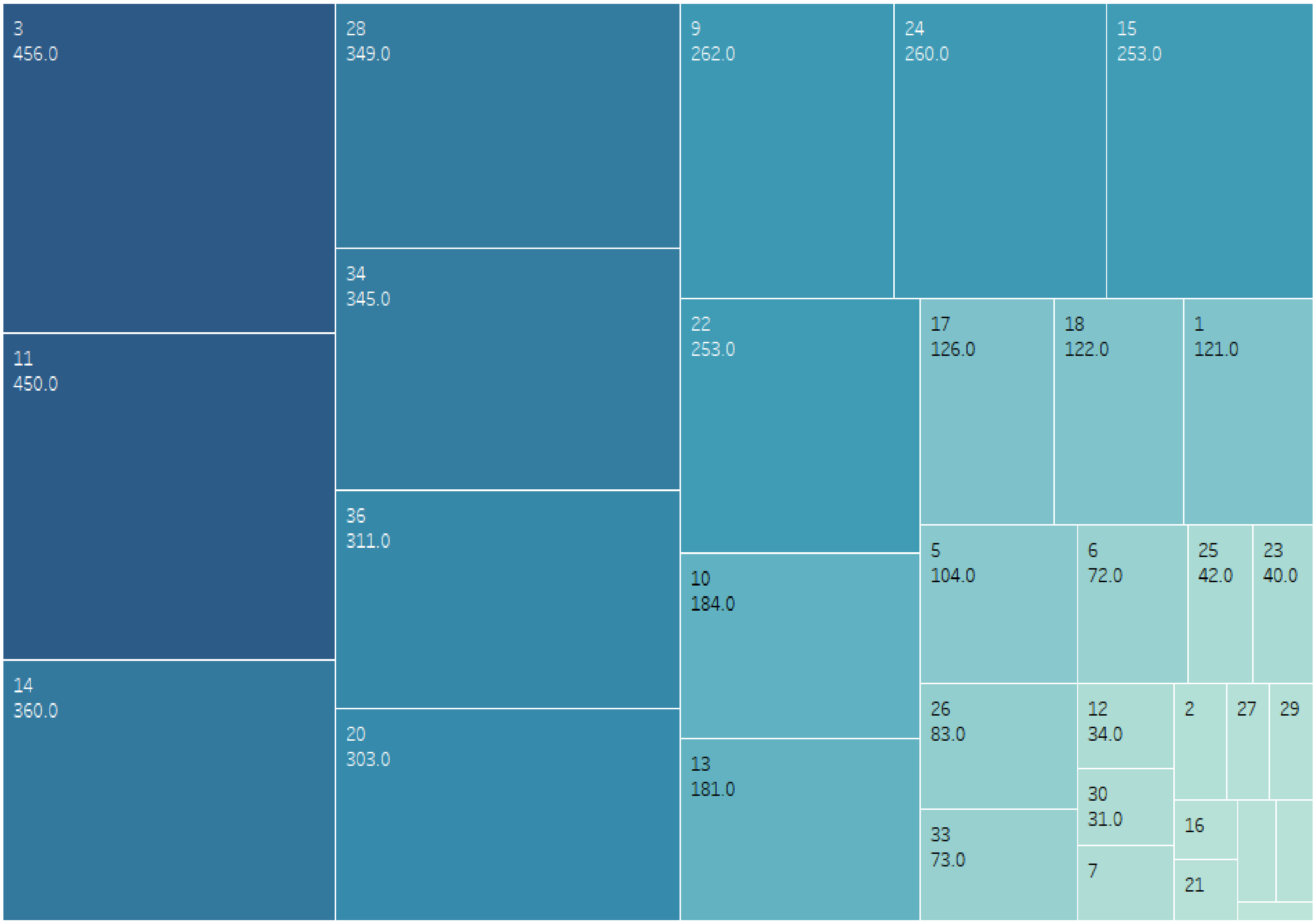
Usually, Person having more children are more prone to absenteeism as they have to look after their children.

But here, employees having 1 or 2 children have more absenteeism which is contrary to basic assumption.



Persons having more pets are less prone to absenteeism which is contrary to our basic assumptions.

Top employees who has more absenteeism?

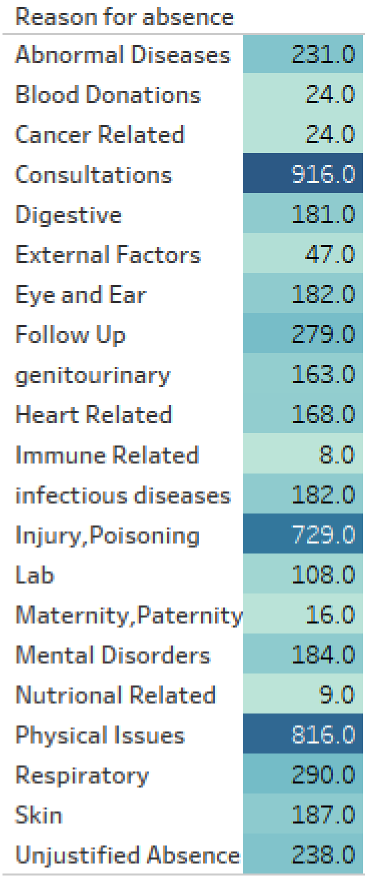


Employee Id's:

3,11,14,28,34,36,20,9,24,15,22

We need to concentrate more on these employees and check whether their absenteeism is genuine or not.

Diseases which Impact Absenteeism:



Reason for absence	
Abnormal Diseases	231.0
Blood Donations	24.0
Cancer Related	24.0
Consultations	916.0
Digestive	181.0
External Factors	47.0
Eye and Ear	182.0
Follow Up	279.0
genitourinary	163.0
Heart Related	168.0
Immune Related	8.0
infectious diseases	182.0
Injury,Poisoning	729.0
Lab	108.0
Maternity,Paternity	16.0
Mental Disorders	184.0
Nutrional Related	9.0
Physical Issues	816.0
Respiratory	290.0
Skin	187.0
Unjustified Absence	238.0

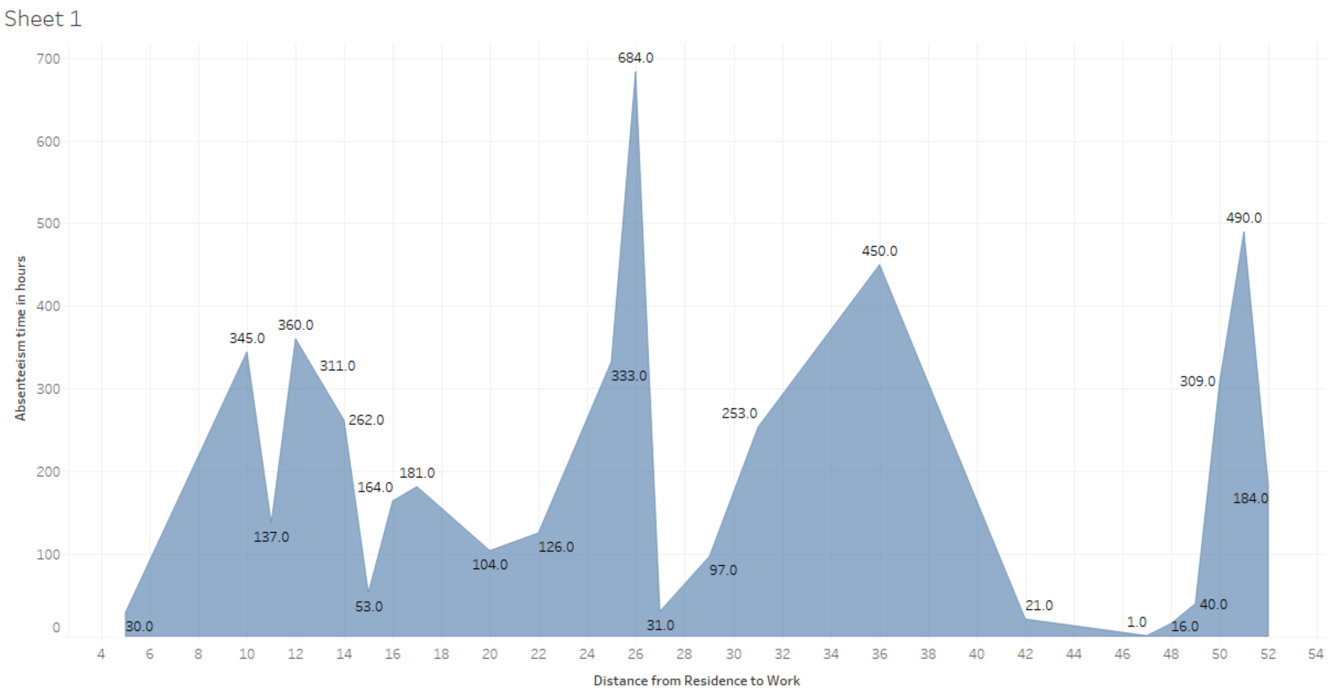
We can see that more absenteeism is due to

1).Medical Consultations.

2).Injury, Poisoning

3) Physical Issues such as musculoskeletal system and connective tissue

Relation between Distance from Residence to Work and Absenteeism in hours:



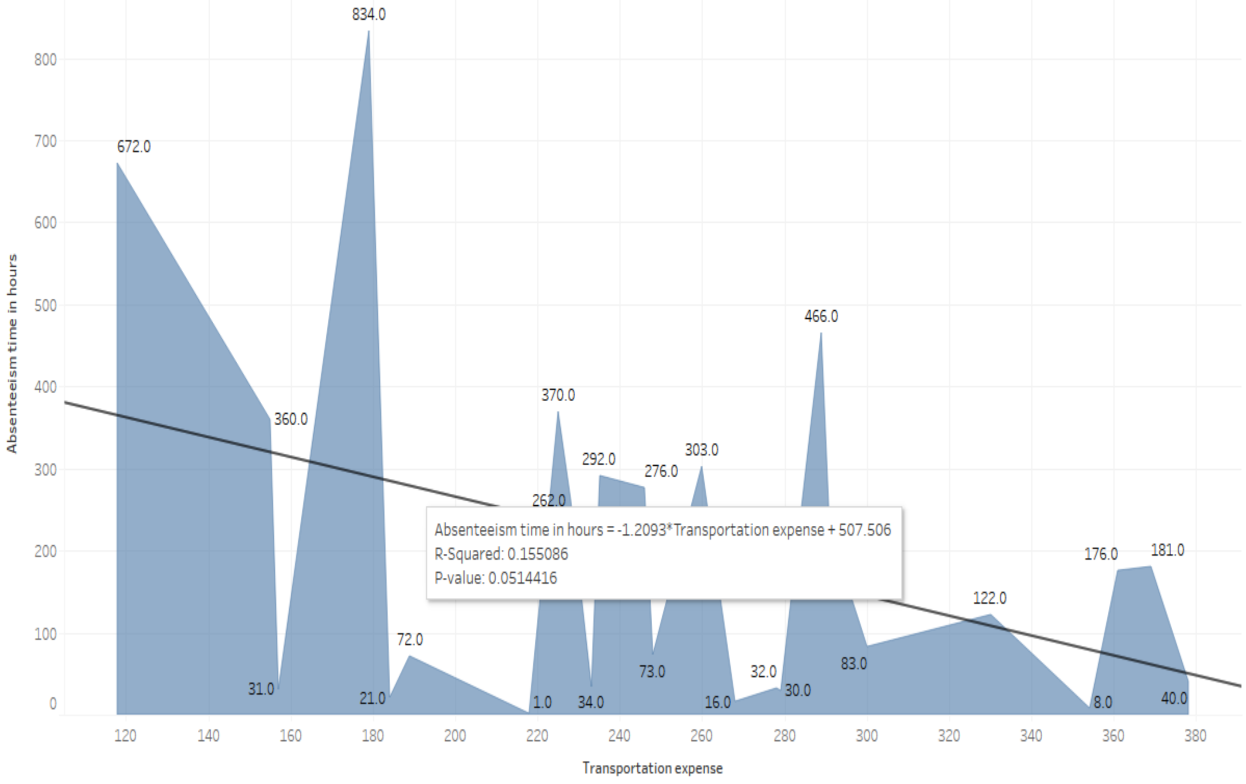
We can see that there is no linear relation between distance from residence and Absenteeism.

But we can see that there are some peaks at 10, 12,26,36,51

This peaks might be due to employees or reason of absence

But we can say that distance from residence has no significant impact on absenteeism.

Does More Transportation cost impacts Absenteeism?

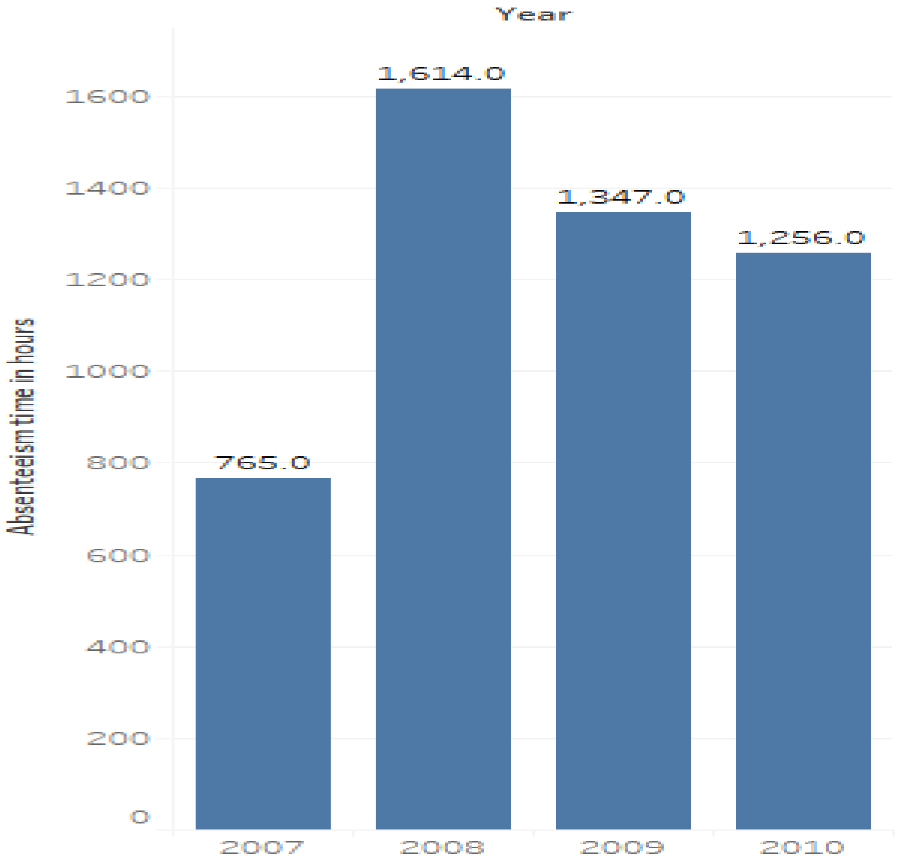


Absenteeism time in hours = -1.2093*Transportation expense + 507.506
R-Squared: 0.155086
P-value: 0.0514416

We can see that there is slightly decreasing trend with R square value of 15%

From this we can say Transportation too has slight impact on absenteeism.

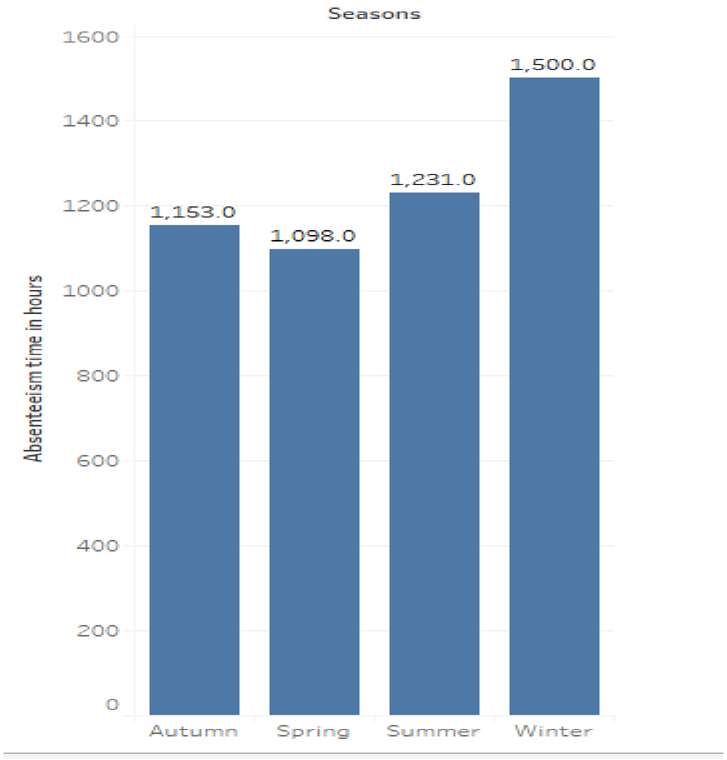
Which Year has more Absenteeism?



We can see that 2008 has more Absenteeism.

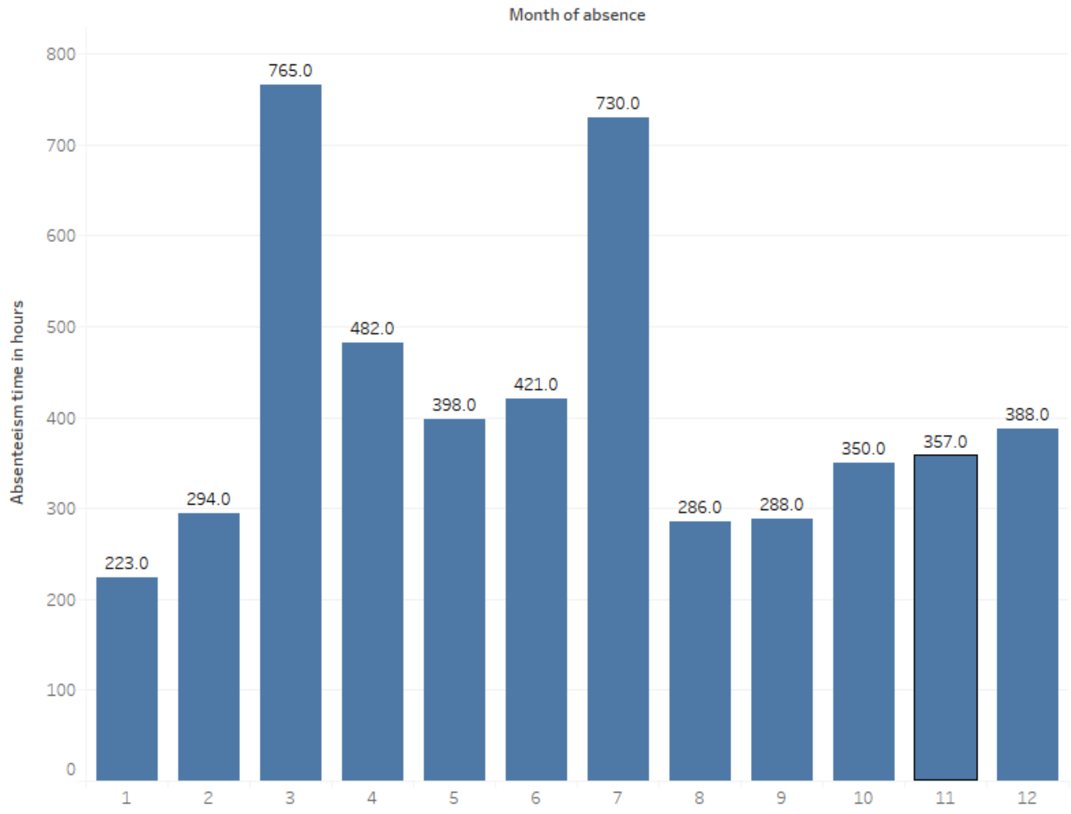
2010 with 6 months of data has more absenteeism. So we can expect more absenteeism for 2010 and 2011

Which Season has more absenteeism?



We can see that winter has more absenteeism.

Which Month has more Absenteeism?

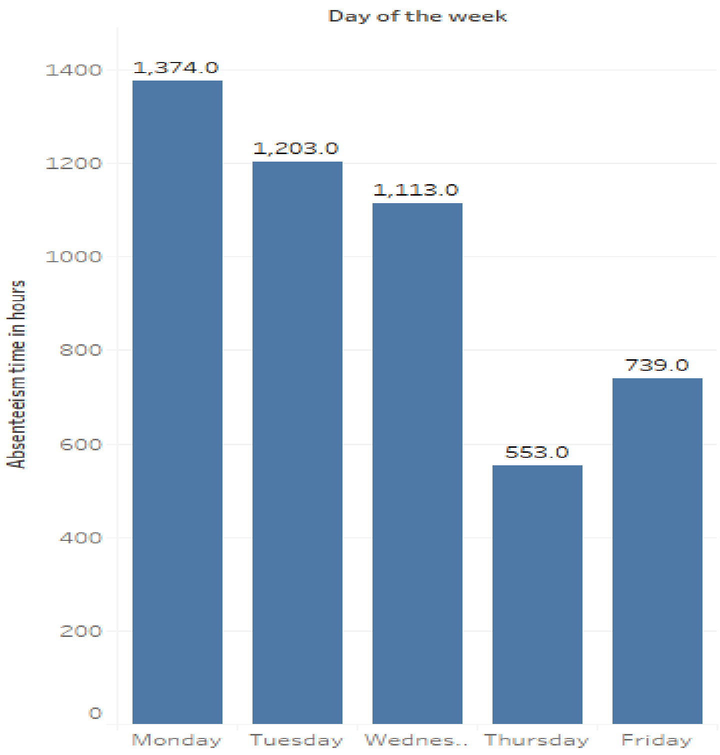


We can see that March and July has more absenteeism

April, May and June has fair absenteeism.

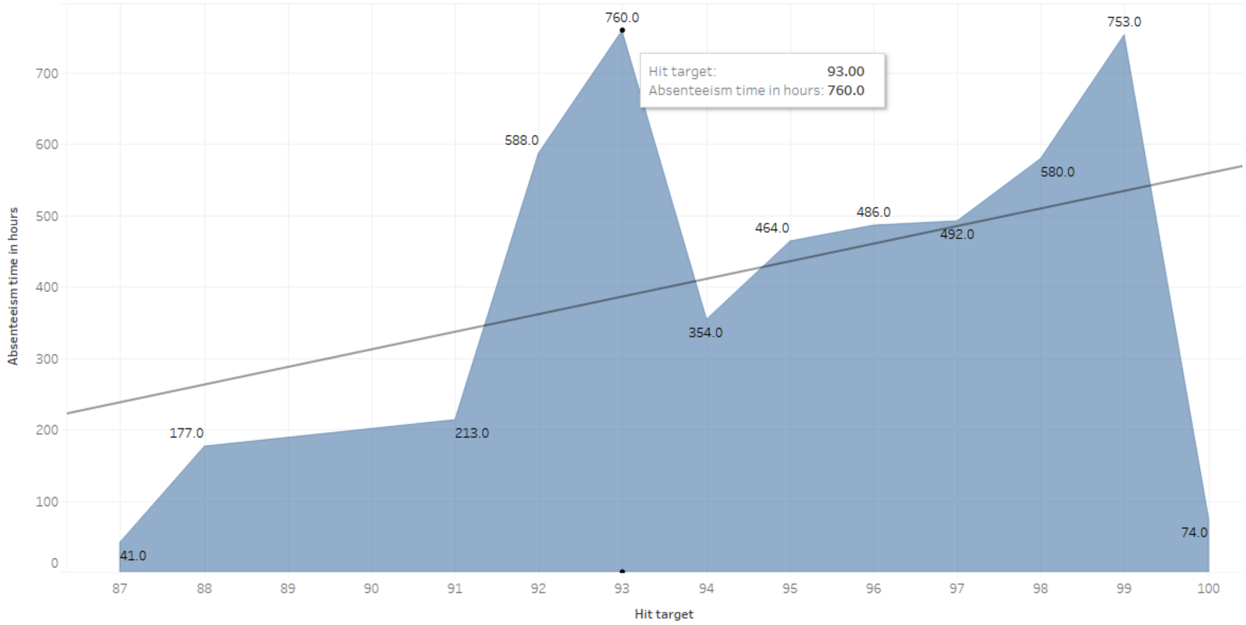
Hence we can conclude that more absenteeism can be expected in winter season (3, 4, 5, 6) and at the start of summer and fair absenteeism for last three months.

Which Weekday has more absenteeism?



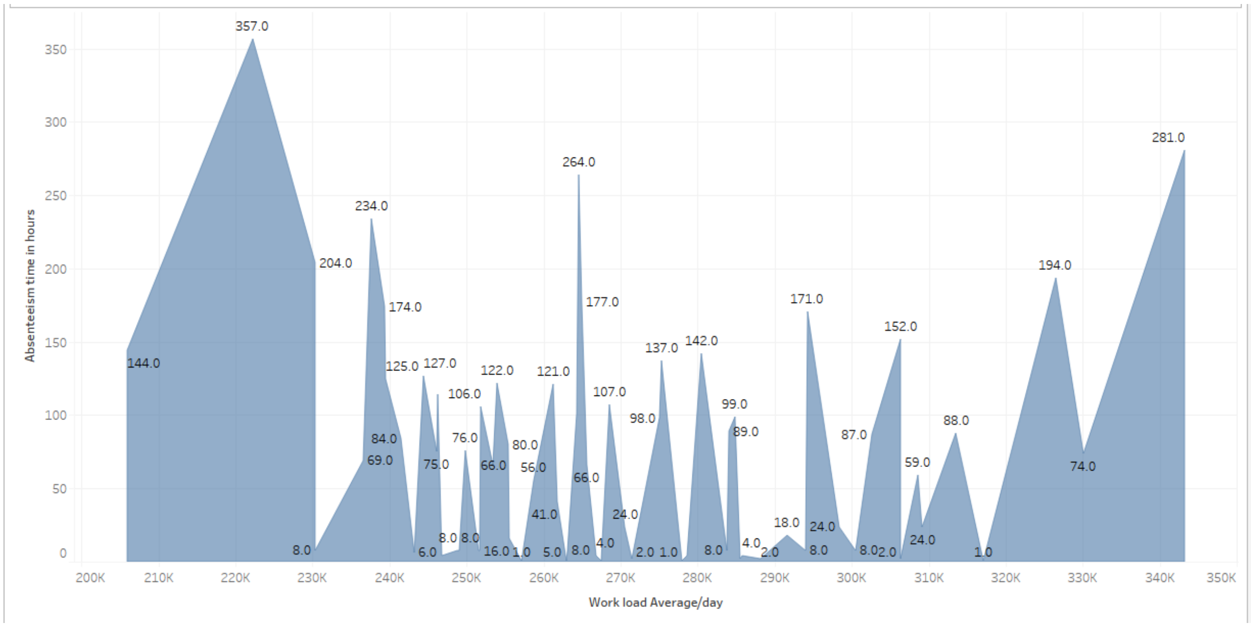
We can see that clearly there is high absenteeism on Monday.

Relation between Hit Target and Absenteeism:



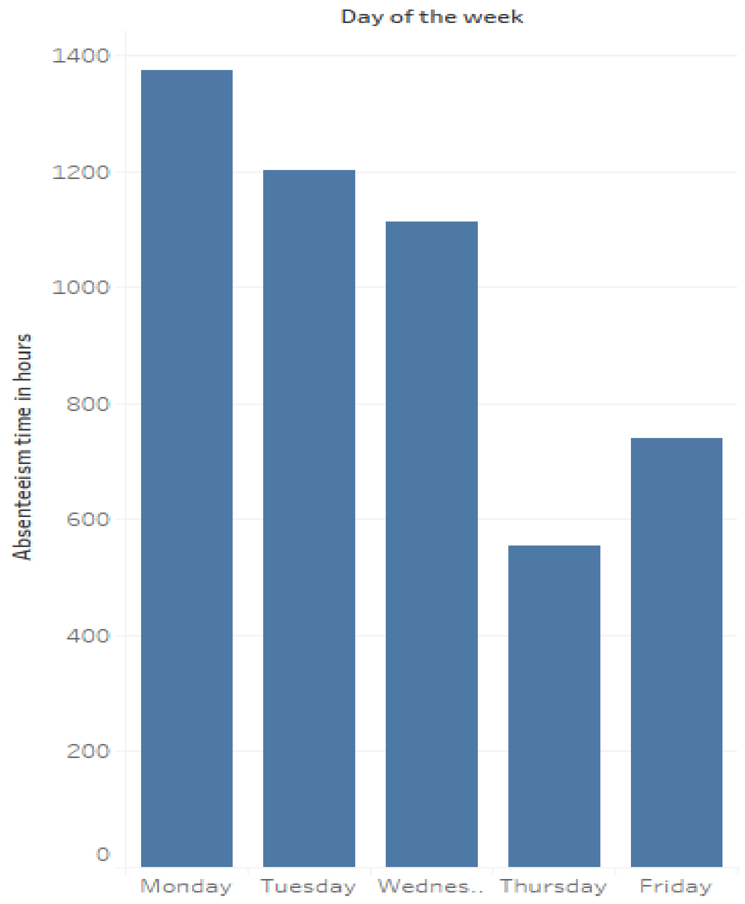
We can see some peaks at hit target 93 and 99.But there is some linear relationship with some noise.

Relation between Workload and Absenteeism:



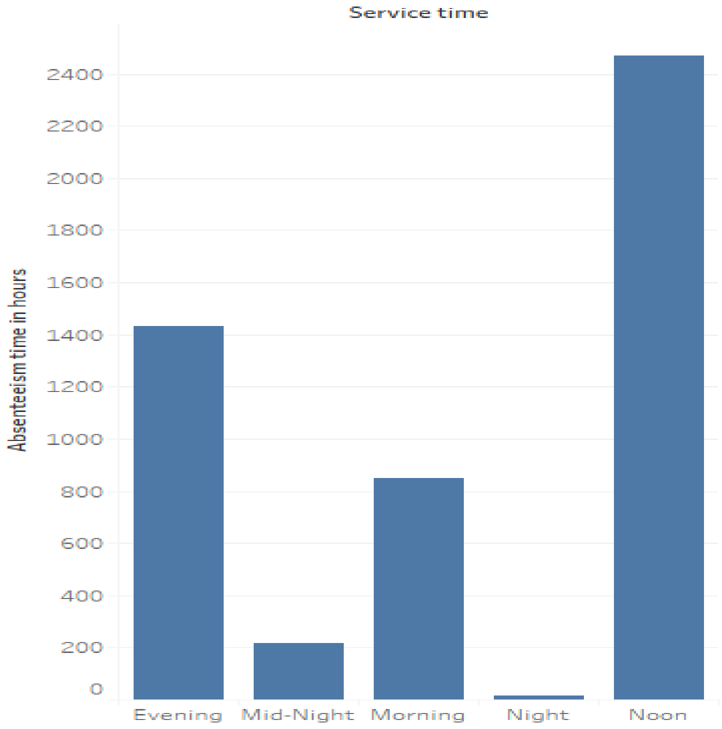
We can see more random peaks which doesn’t explain much about absenteeism.

Does Absenteeism have impact depends on Weekdays?

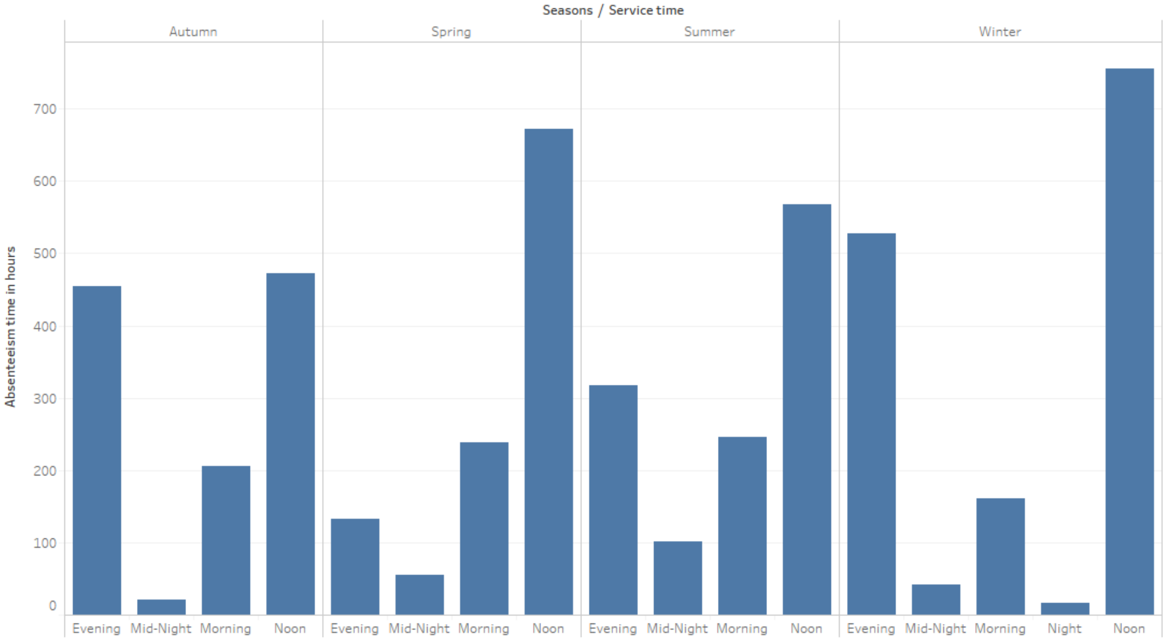


We can see that absenteeism tends to decrease from Monday to Friday

Relation between Service time and Absenteeism:



We ca see that there is more absenteeism at Noon



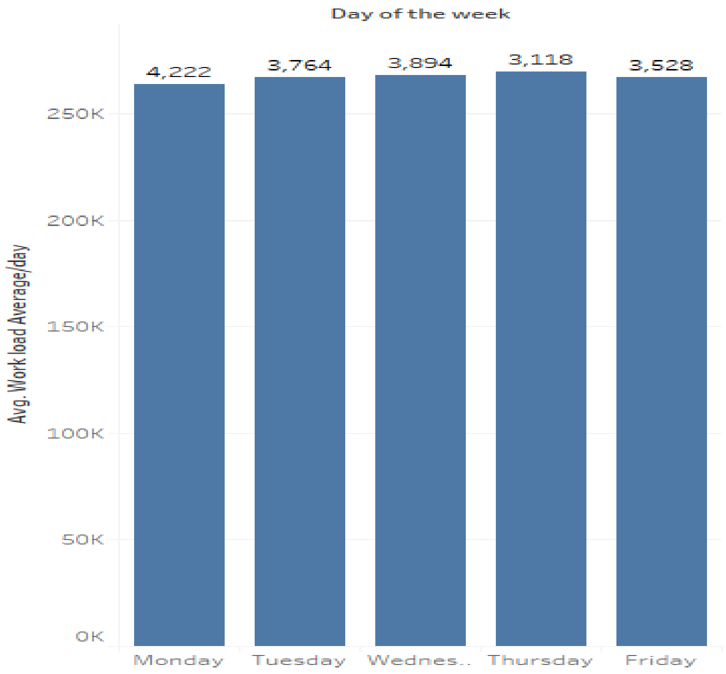
We can see that in summer and spring seasons, there is relatively more absenteeism at noon because of high temperatures in summer and end of spring season.

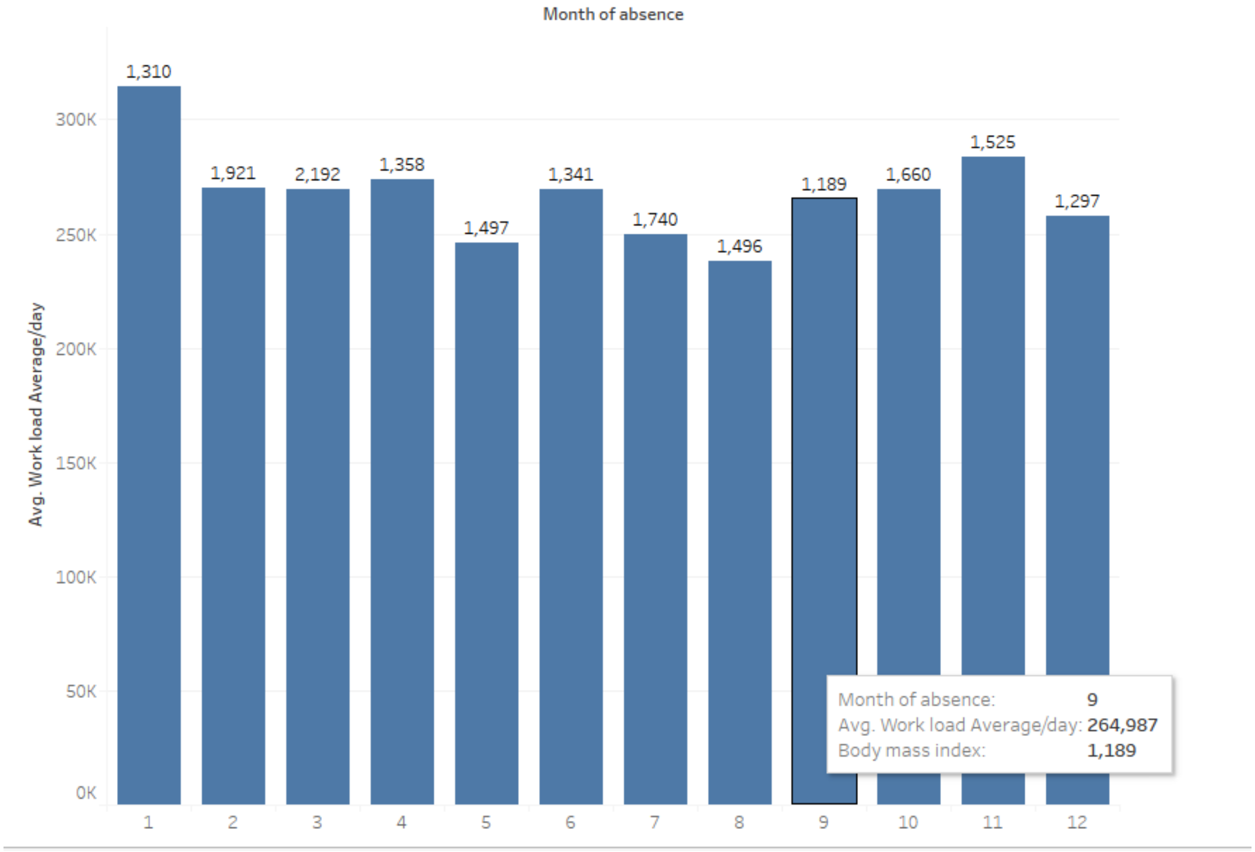
End of spring is December which is festival season.

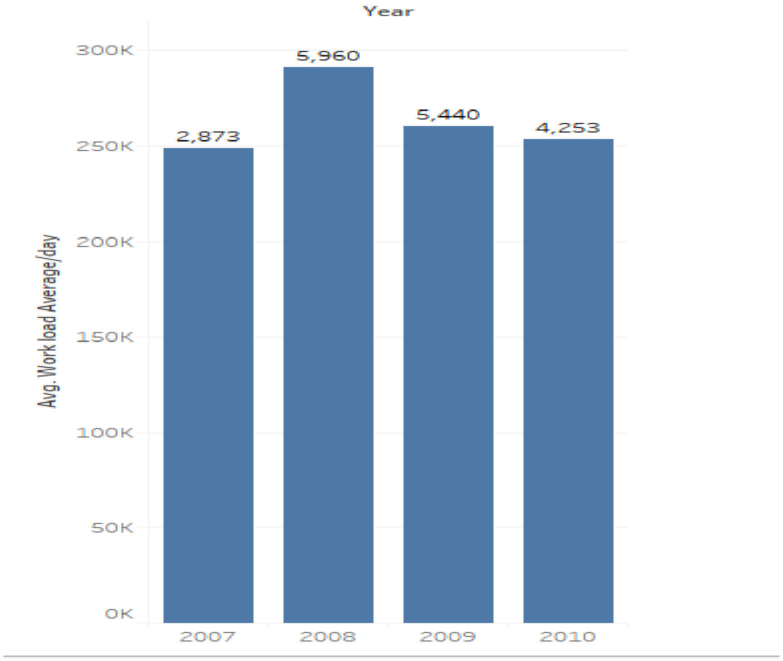
So we can expect more absentees.

Due to high temperatures in summer, there is more absenteeism because of skin related diseases.

Average Work load on weekdays, Months and years:



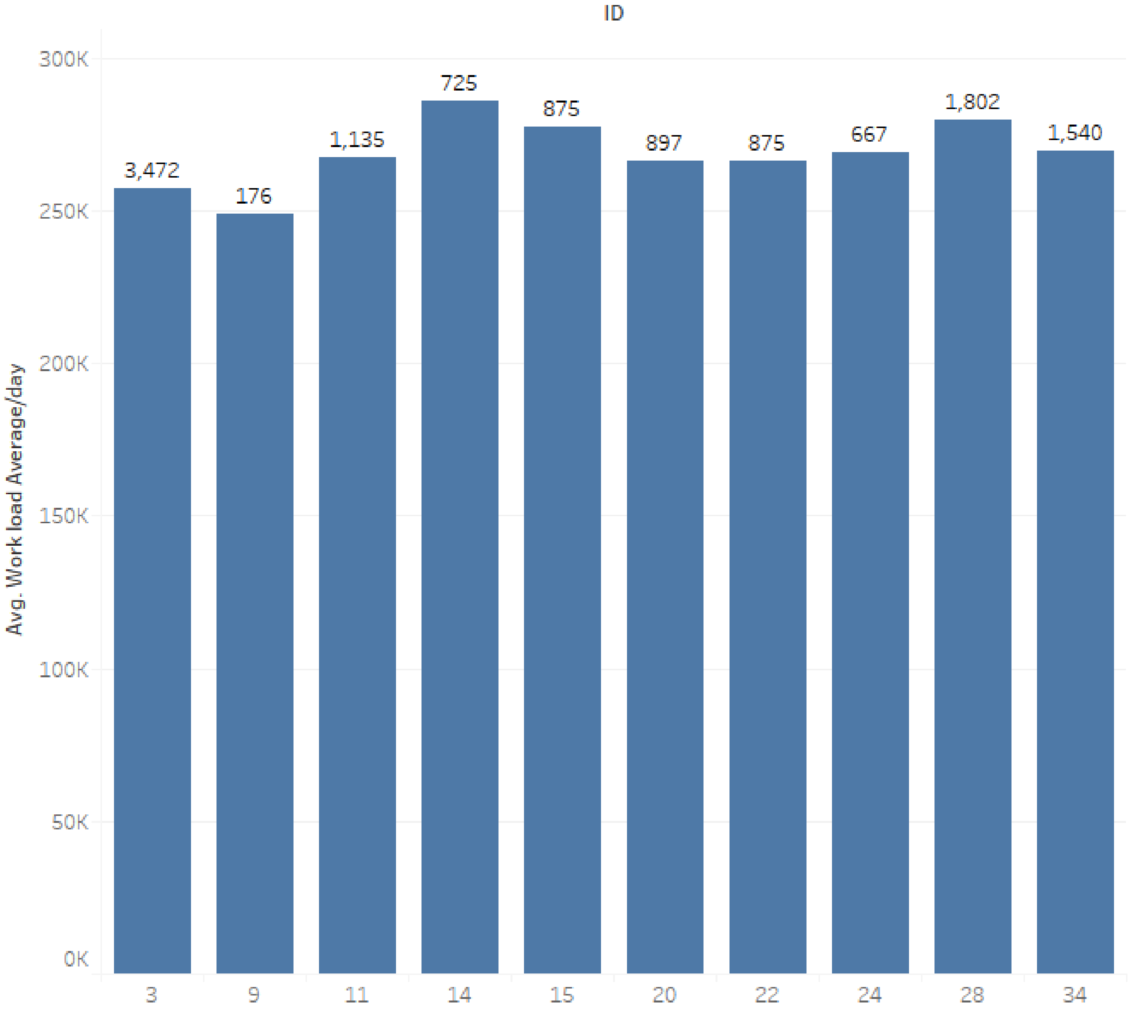

Month of absence
Avg. Work load Average/day
300K
250K
200K
150K
100K
50K
0K
1,310
1,921
2,192
1,358
1,497
1,341
1,740
1,496
1,189
1,660
1,525
1,297
1
2
3
4
5
6
7
8
9
10
11
12
Month of absence: 9
Avg. Work load Average/day: 264,987
Body mass index: 1,189


Year
Avg. Work load Average/day
300K
250K
200K
150K
100K
50K
0K
2,873
5,960
5,440
4,253
2007
2008
2009
2010

We can see that workload is same throughout the week, month and years.

So we can see that with respect to time, there is no change in workload.

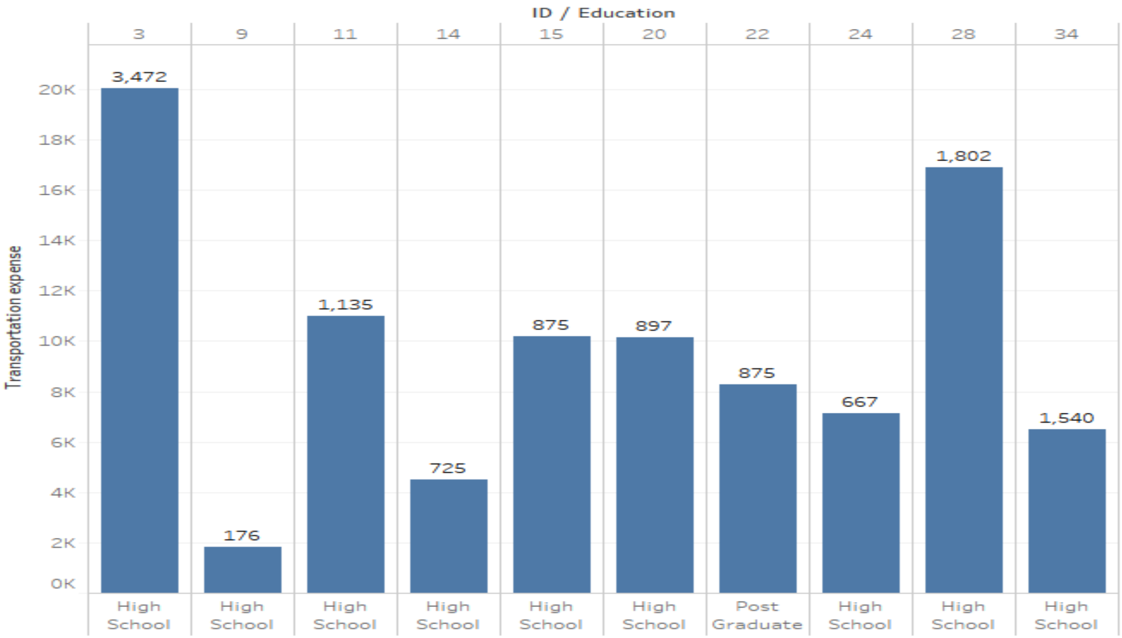
Let us check the workload on absentees who are taking more leaves.



Workload seems to be same for them when compared to other employees.

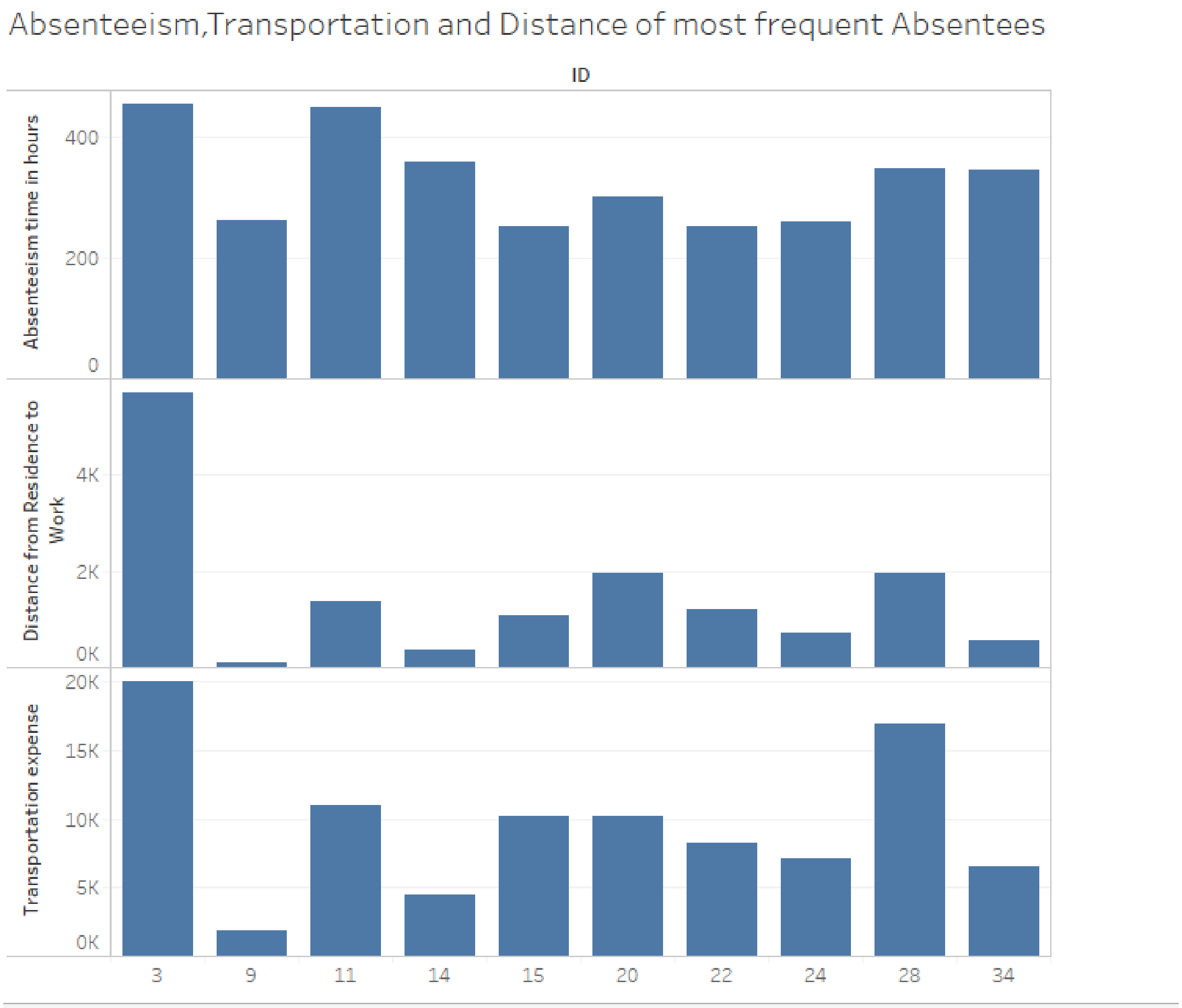
So there is no work pressure on them.

Transportation cost of Frequent absentees:



We can see that almost everyone’s education is high school and only for employees 3 and 28 transportation cost is high.

Let us look at the distance from residence for these employees

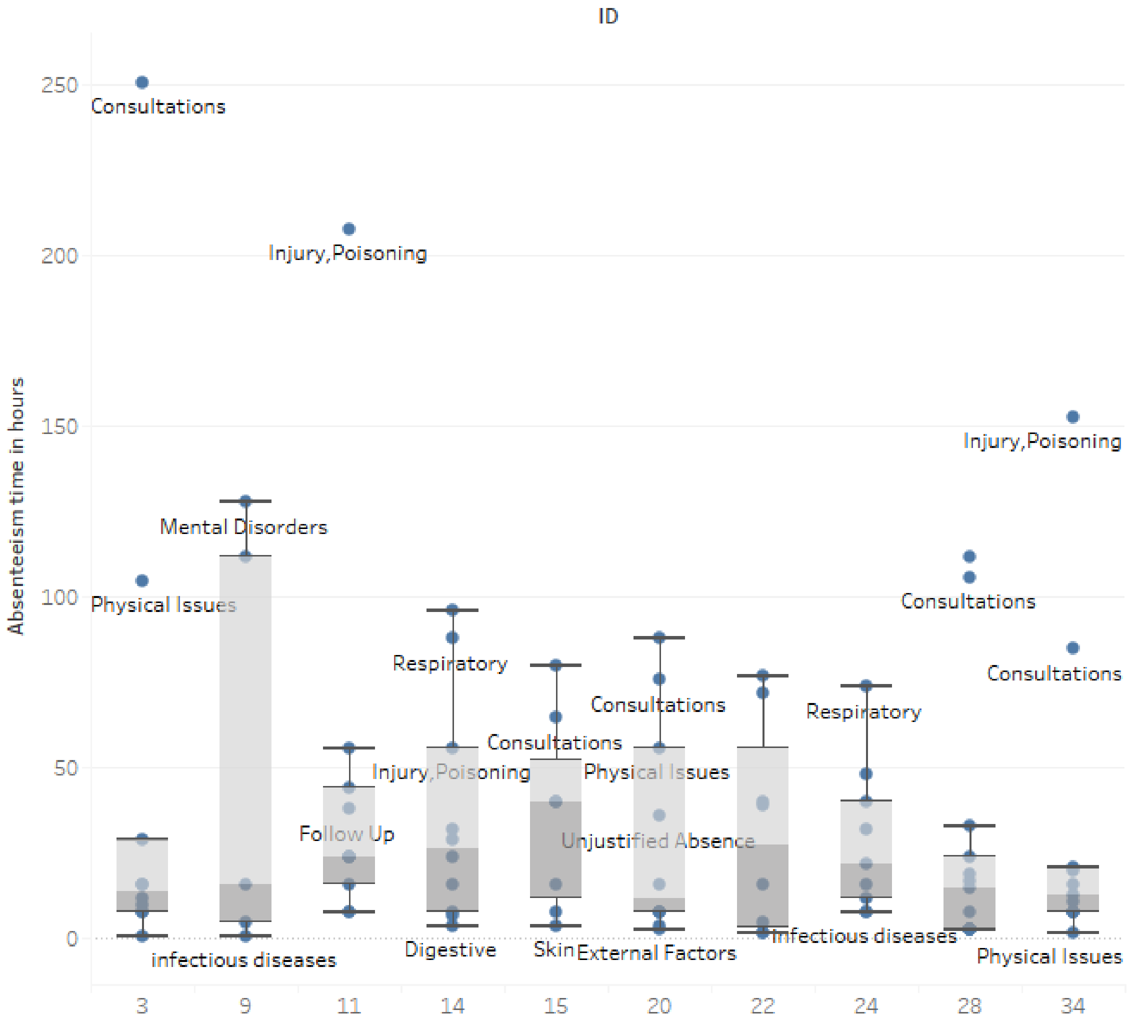


Absenteeism,Transportation and Distance of most frequent Absentees

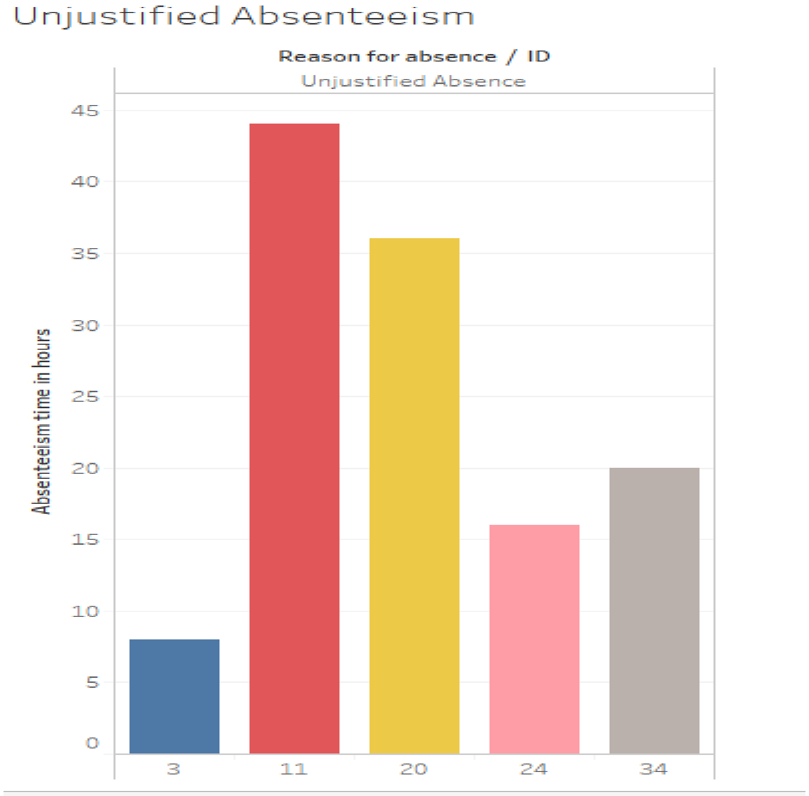
We can see that for employees 3 and 28 transportation cost is high. This might be the cause for their absenteeism. For others Transportation and distance doesn’t affect their absenteeism. Let’s look at other factors

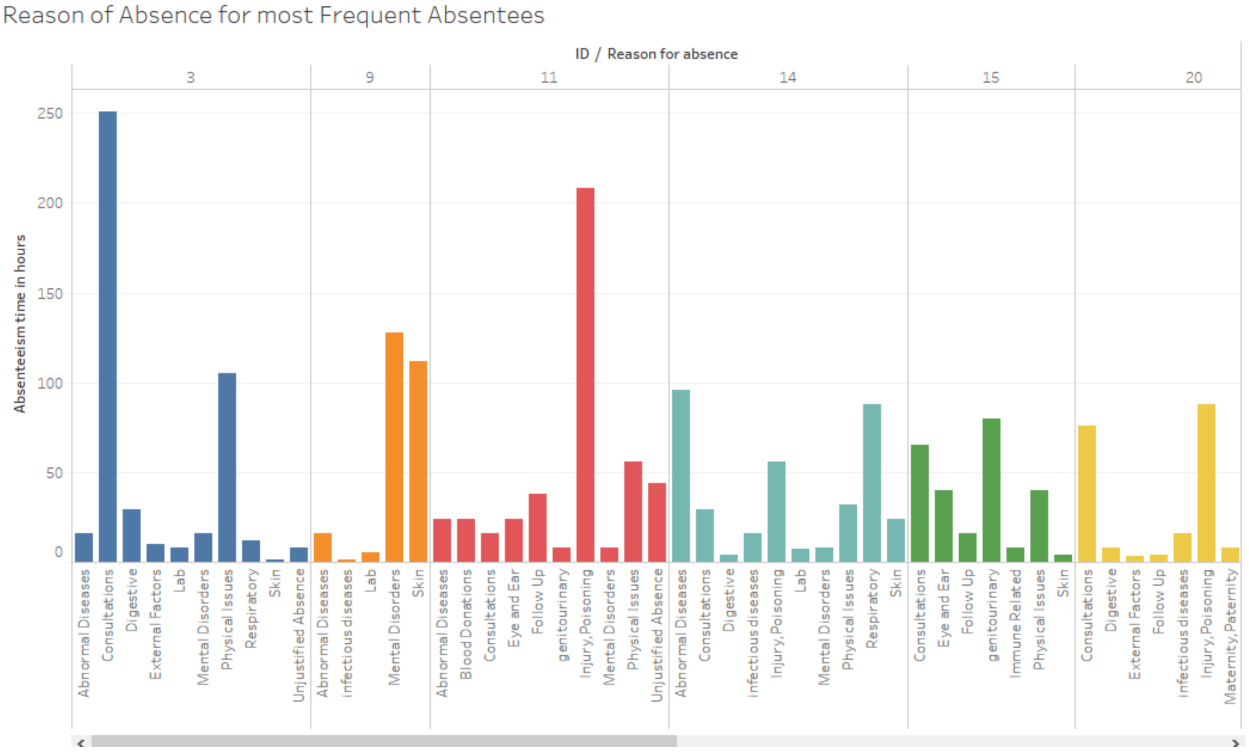
Let’s look at the reason for their absenteeism

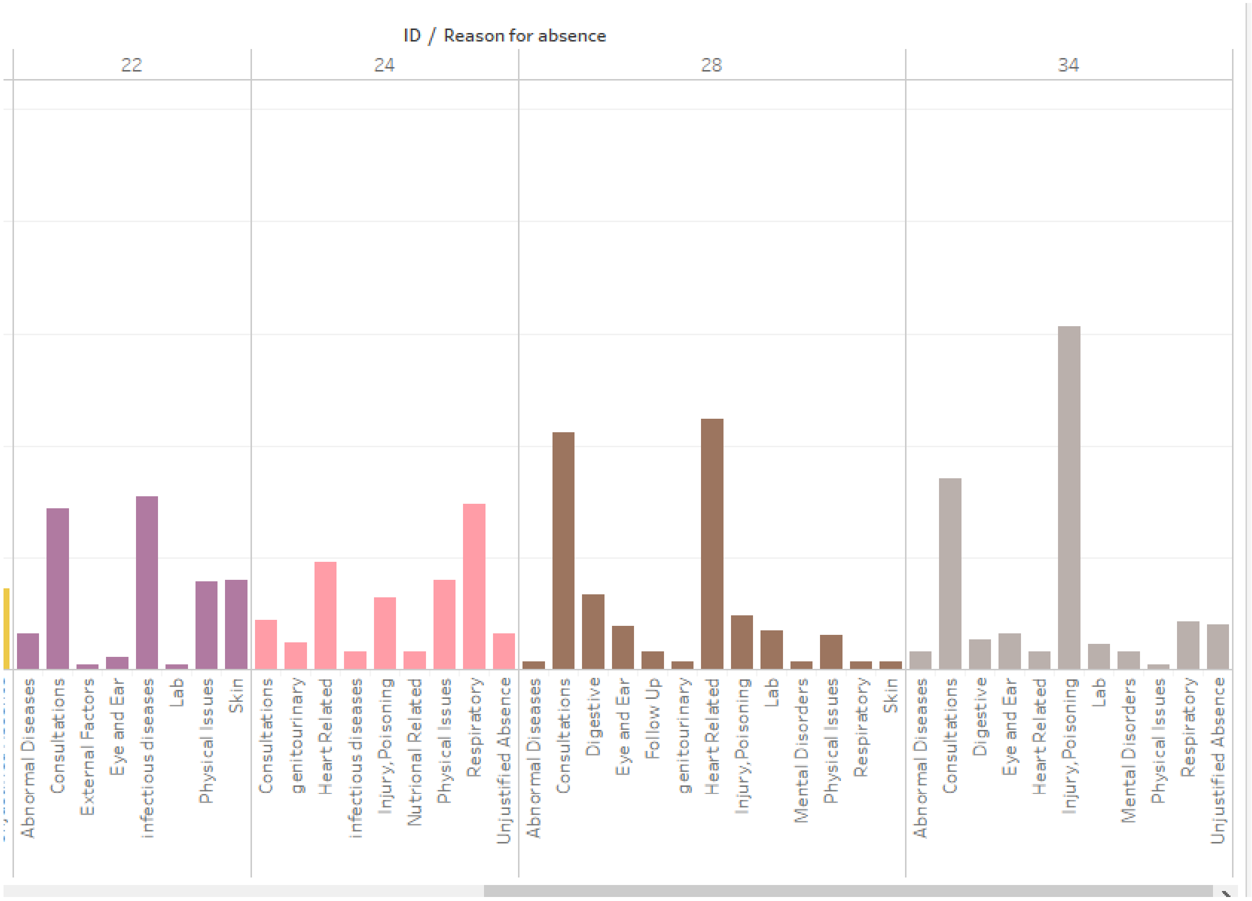
Box plots for most frequent absentees



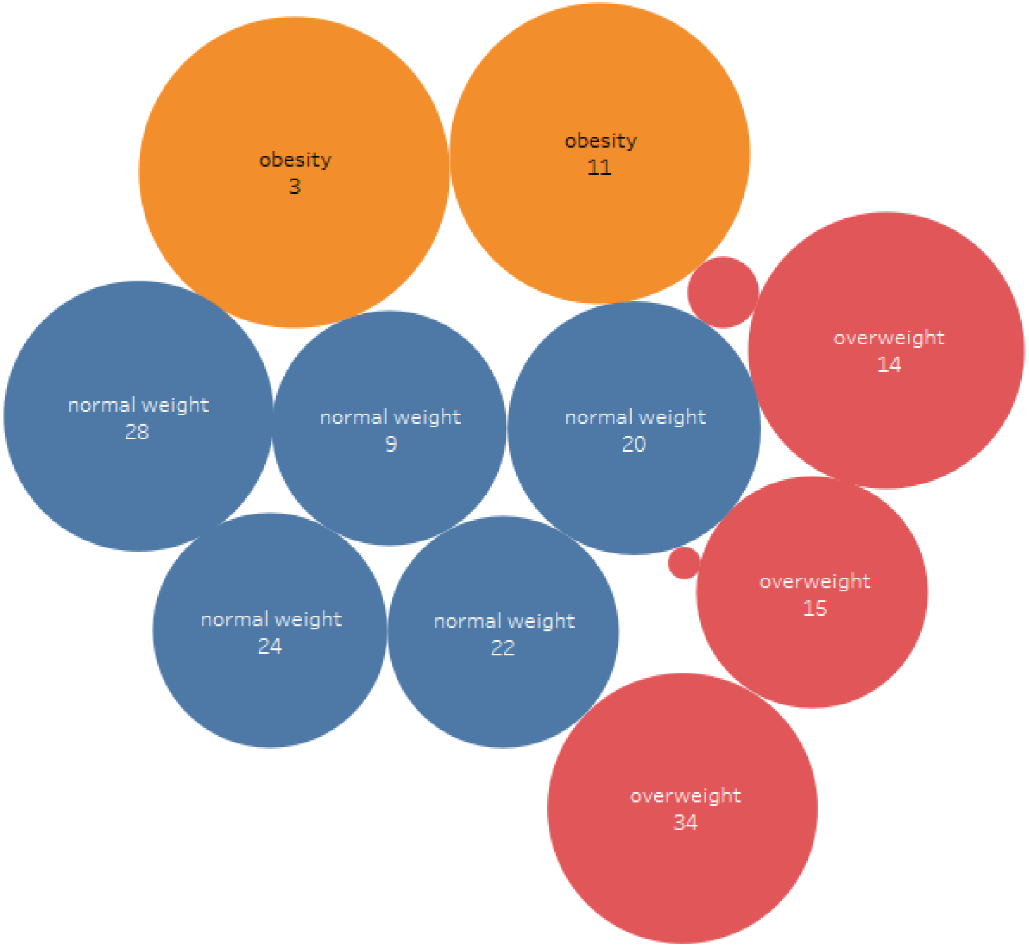
Let’s have a look at Unjustified Absenteesim

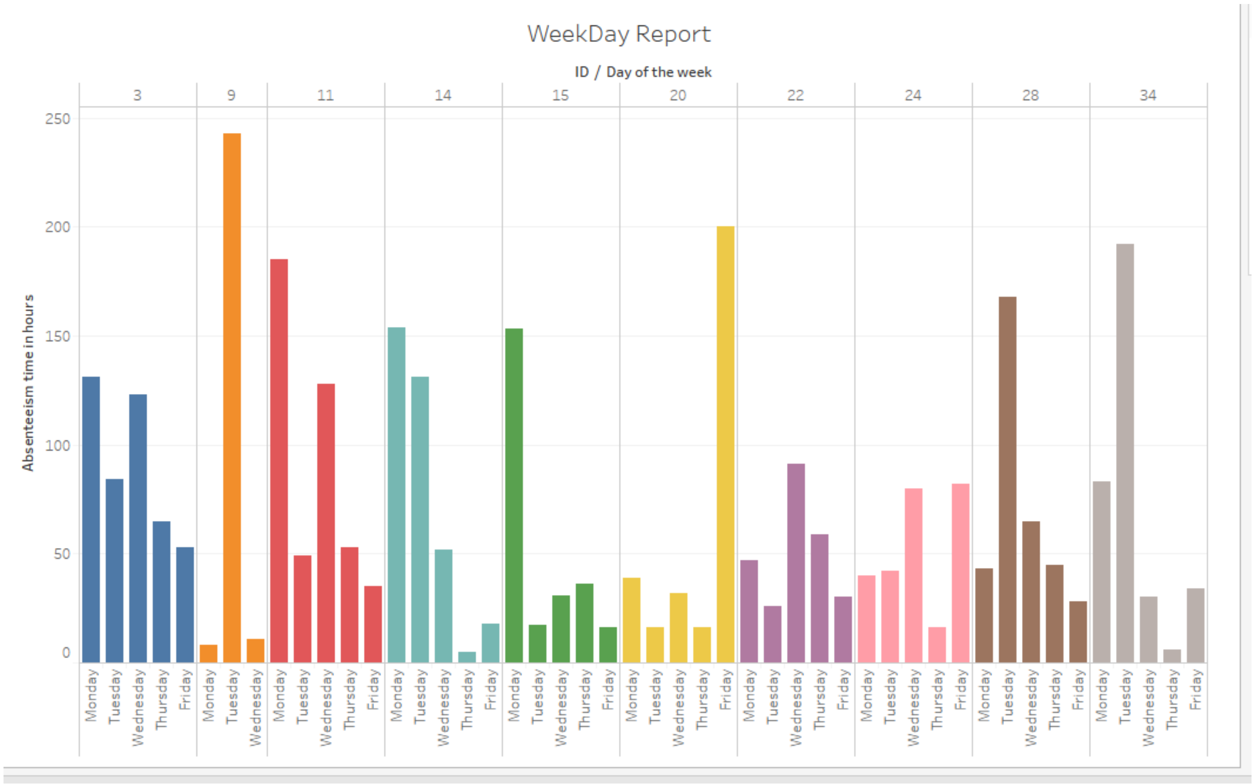





ID / Reason for absence
22
24
28
34
Abnormal Diseases
Consultations
External Factors
Eye and Ear
infectious diseases
Lab
Physical Issues
Skin
Consultations
genitourinary
Heart Related
infectious diseases
Injury,Poisoning
Nutrional Related
Physical Issues
Respiratory
Unjustified Absence
Abnormal Diseases
Consultations
Digestive
Eye and Ear
Follow Up
genitourinary
Heart Related
Injury,Poisoning
Lab
Mental Disorders
Physical Issues
Respiratory
Skin
Abnormal Diseases
Consultations
Digestive
Eye and Ear
Heart Related
Injury,Poisoning
Lab
Mental Disorders
Physical Issues
Respiratory
Unjustified Absence

BMI of Most Frequent Absentees





Inferences:

3:

- Consultations
- Physical Issues
- Obese
- Middle Aged
- Transportation High
- Distance from Residence High
- More Absenteeism on Monday

9:

- Mental Disorders
- Skin
- Middle Aged
- More Absenteeism on Tuesday

11:

- Injury, Poisoning
- More Unjustified Absence
- Obese

- More Absenteeism on Monday

20:

- Consultations
- Injury Poisoning
- Middle Aged
- More Absenteeism on Friday

28:

- More Unjustified Absence
- Consultations
- Heart Related
- Transportation High
- More Absenteeism on Tuesday

34:

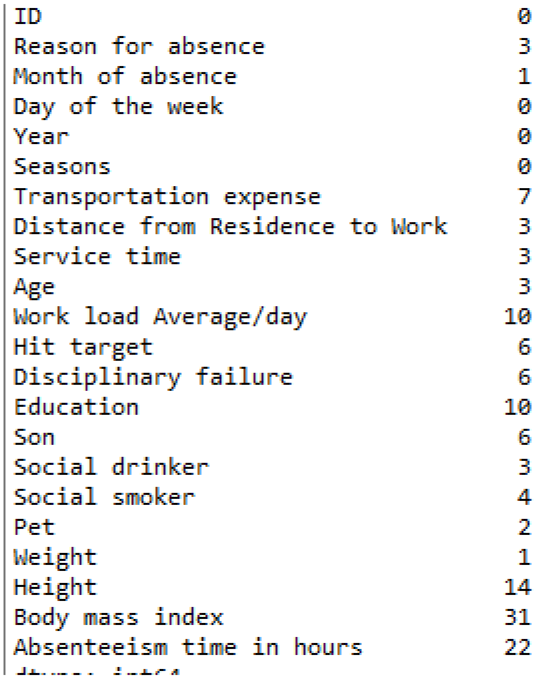
- Consultations
- Injury, poisoning
- Overweight
- Middle Aged
- More Absenteeism on Tuesday

Thus we have explored the data by answering some basic questions, testing hypothesis, univariate analysis, bivariate analysis and looking into factors affecting most frequent employees

Chapter 3

Data Pre Processing:

3.1 Missing Value Analysis:



```
ID                                0
Reason for absence                3
Month of absence                  1
Day of the week                   0
Year                              0
Seasons                           0
Transportation expense            7
Distance from Residence to Work   3
Service time                      3
Age                               3
Work load Average/day            10
Hit target                        6
Disciplinary failure              6
Education                        10
Son                               6
Social drinker                    3
Social smoker                     4
Pet                               2
Weight                            1
Height                           14
Body mass index                  31
Absenteeism time in hours        22
dtype: int64
```

Year data is added in excel as year data is not given

1) Imputing Missing value with 20

data["Reason for absence"]=data["Reason for absence"].fillna(20)

2) Deleting the rows for which for which reason of absence is not recorded

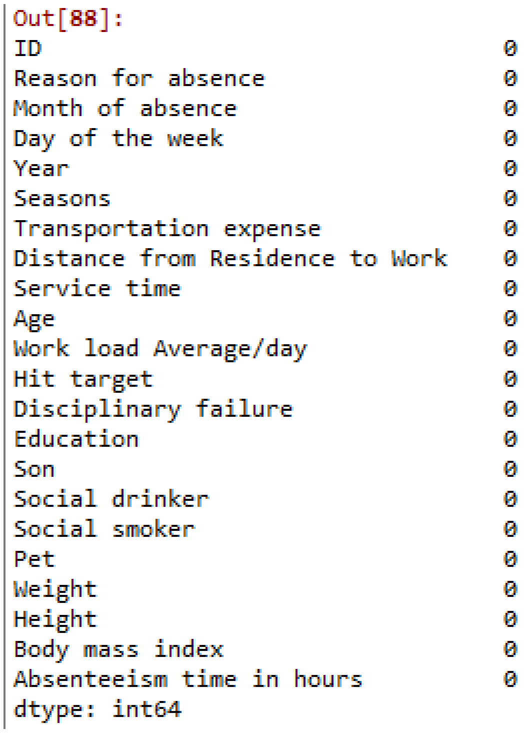
data=data[data["Reason for absence"]!=0]

3) Imputing Absenteeism time in hours column with 1 as zero absenteeism is does not make any sense

data[data["Absenteeism time in hours"]==0]["Absenteeism time in hours"]=1

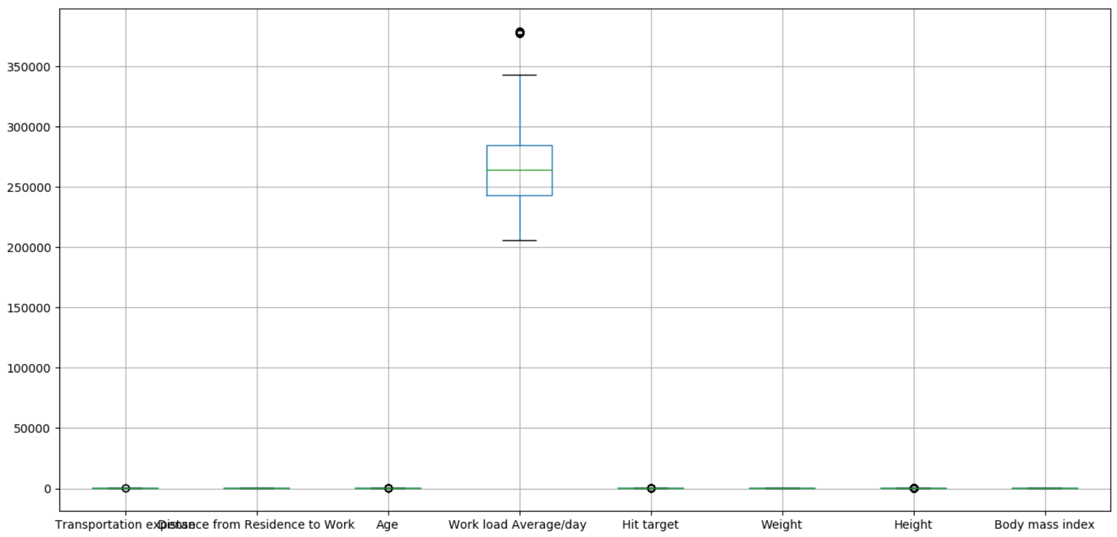
KNN Imputation:

Let's check the missing values after KNN Imputation



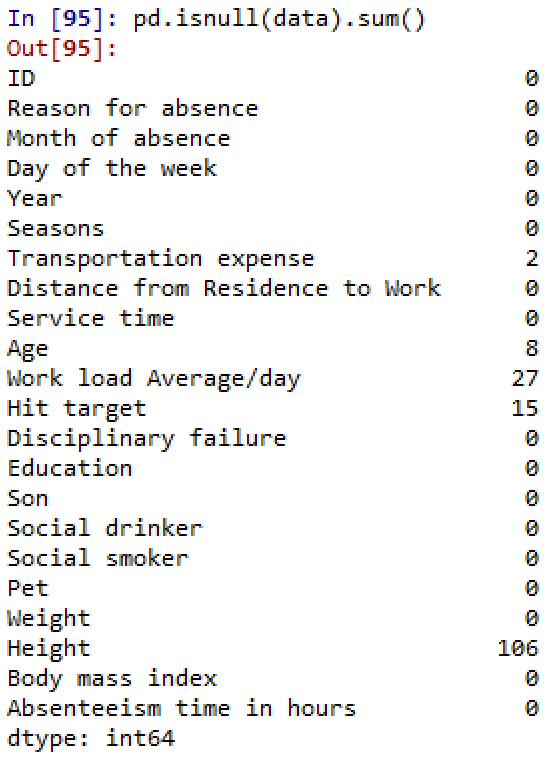
```
Out[88]:
ID                                 0
Reason for absence                 0
Month of absence                   0
Day of the week                    0
Year                               0
Seasons                            0
Transportation expense             0
Distance from Residence to Work    0
Service time                       0
Age                                0
Work load Average/day              0
Hit target                         0
Disciplinary failure               0
Education                          0
Son                                0
Social drinker                     0
Social smoker                      0
Pet                                0
Weight                             0
Height                             0
Body mass index                    0
Absenteeism time in hours          0
dtype: int64
```

3.2 Outlier Analysis:



We can see that there are some outliers.

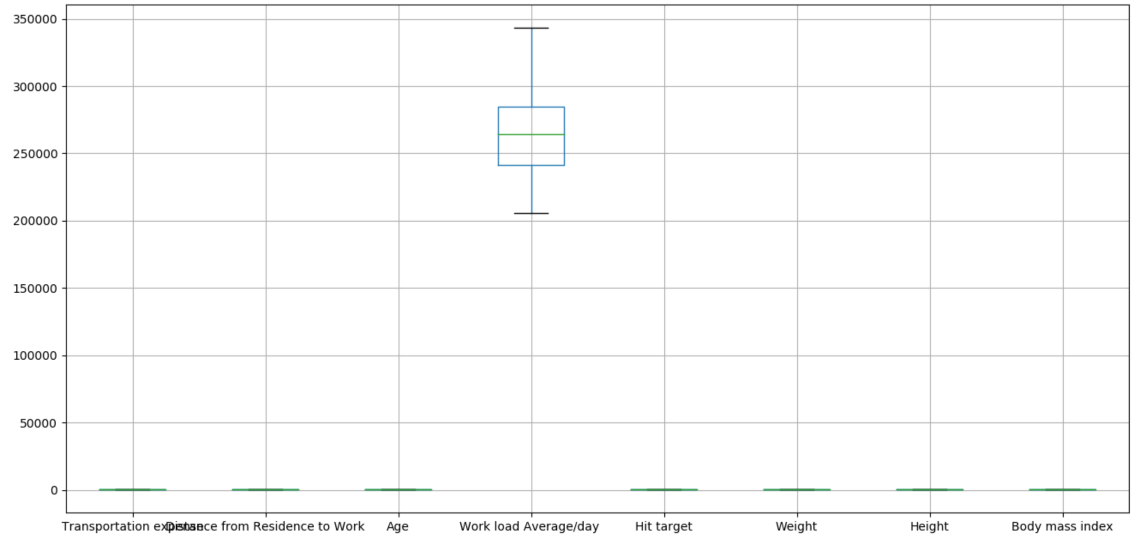
Let us check the number of outliers of each column by imputing outliers with unknowns



```
In [95]: pd.isnull(data).sum()
Out[95]:
ID                                  0
Reason for absence                  0
Month of absence                    0
Day of the week                     0
Year                                0
Seasons                             0
Transportation expense              2
Distance from Residence to Work     0
Service time                        0
Age                                 8
Work load Average/day              27
Hit target                         15
Disciplinary failure                0
Education                           0
Son                                 0
Social drinker                      0
Social smoker                       0
Pet                                 0
Weight                              0
Height                            106
Body mass index                     0
Absenteeism time in hours           0
dtype: int64
```

Let us impute these unknowns using KNN.

Let us check the outliers after KNN.

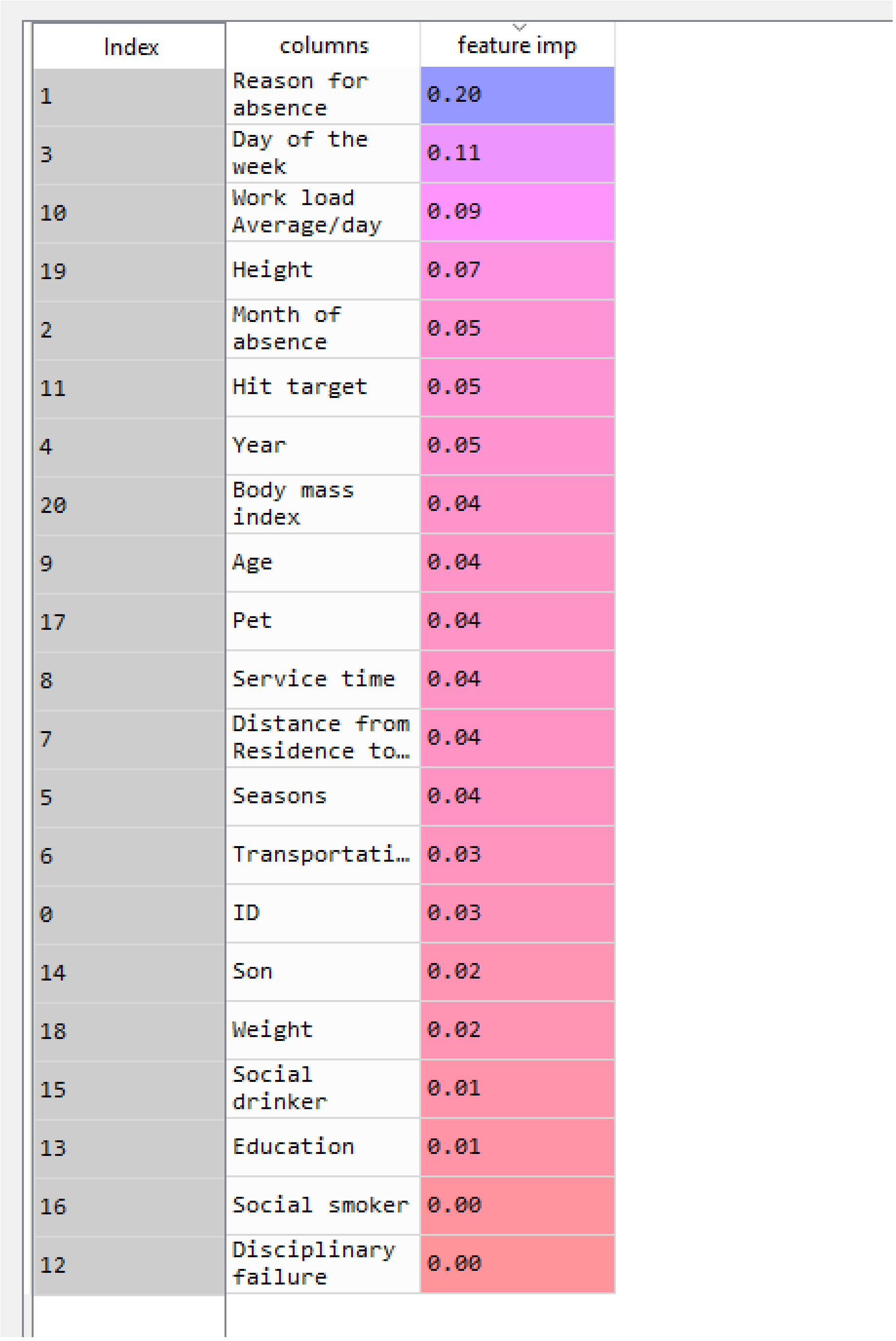


3.3 Feature Selection:

Categorical:

Let us check feature importance of categorical variables using ExtraTreesRegressor

```
dependent=data.drop(['Absenteeism time in hours'],axis=1)

target=data['Absenteeism time in hours']


from sklearn.ensemble import ExtraTreesRegressor

model=ExtraTreesRegressor(random_state=0)

model.fit(dependent,target)

print(model.feature_importances_)

f_s_df=pd.DataFrame()

f_s_df["columns"]=list(dependent.columns)

f_s_df["feature imp"]=model.feature_importances_
```



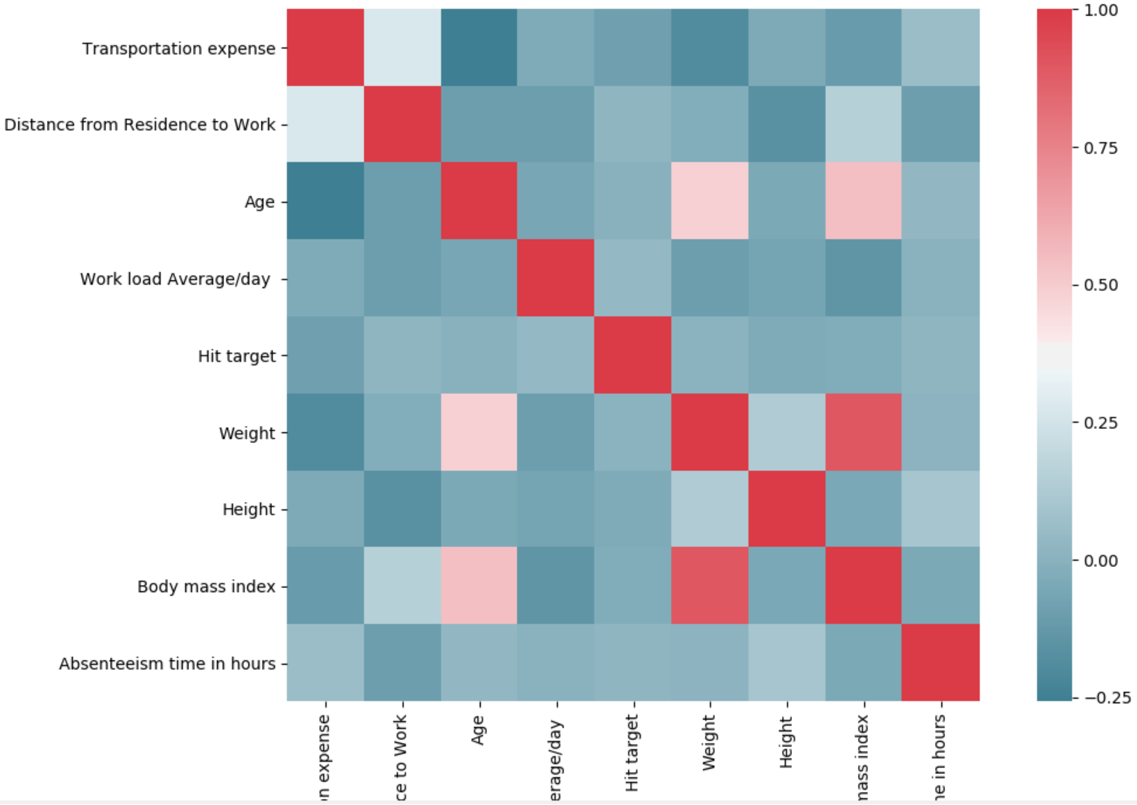
Index	columns	feature imp
1	Reason for absence	0.20
3	Day of the week	0.11
10	Work load Average/day	0.09
19	Height	0.07
2	Month of absence	0.05
11	Hit target	0.05
4	Year	0.05
20	Body mass index	0.04
9	Age	0.04
17	Pet	0.04
8	Service time	0.04
7	Distance from Residence to…	0.04
5	Seasons	0.04
6	Transportati…	0.03
0	ID	0.03
14	Son	0.02
18	Weight	0.02
15	Social drinker	0.01
13	Education	0.01
16	Social smoker	0.00
12	Disciplinary failure	0.00

Feature Importance with 0 and 0.1 are dropped.

Numerical Variables:

Let us check the correlation between numerical variable.

```
sns.heatmap(corr_mat,mask=np.zeros_like(corr_mat,dtype=np.bool),cmap=sns.diverging_palette(220,10,as_cmap=True),square=True)
```



Weight and BMI are highly positively correlated

Height and BMI are highly negatively correlated

Hence Weight and Height columns can be eliminated

Disciplinary failure can be eliminated as it has only one unique data.

3.4 Feature Engineering:

Feature Engineering is done for data to do EDA.

Refer to the code in EDA section

To make the model perform better binning is done and new features are added and dummy columns are extracted for newly added features.

Let us check what the new features are:

BMI:

["underweight","normal weight","overweight","obesity"]

Age:

["Young Adults","Middle Aged","Old"]

Education:

1:"High School",2:"Graduate",3:"Post Graduate",4:"Master and Doctor"

Son:

["No Children","Less Child Count","More Child Count"]

Pets:

["No Pets","Less Pets","More Pets"]

Service Time:

["Mid-Night","Morning","Noon","Evening","Night"]

ID:

["Less Absentees","Medium Absentees","High Absentees"]

Seasons:

{1:"Summer",2:"Autumn",3:"Winter",4:"Spring"}

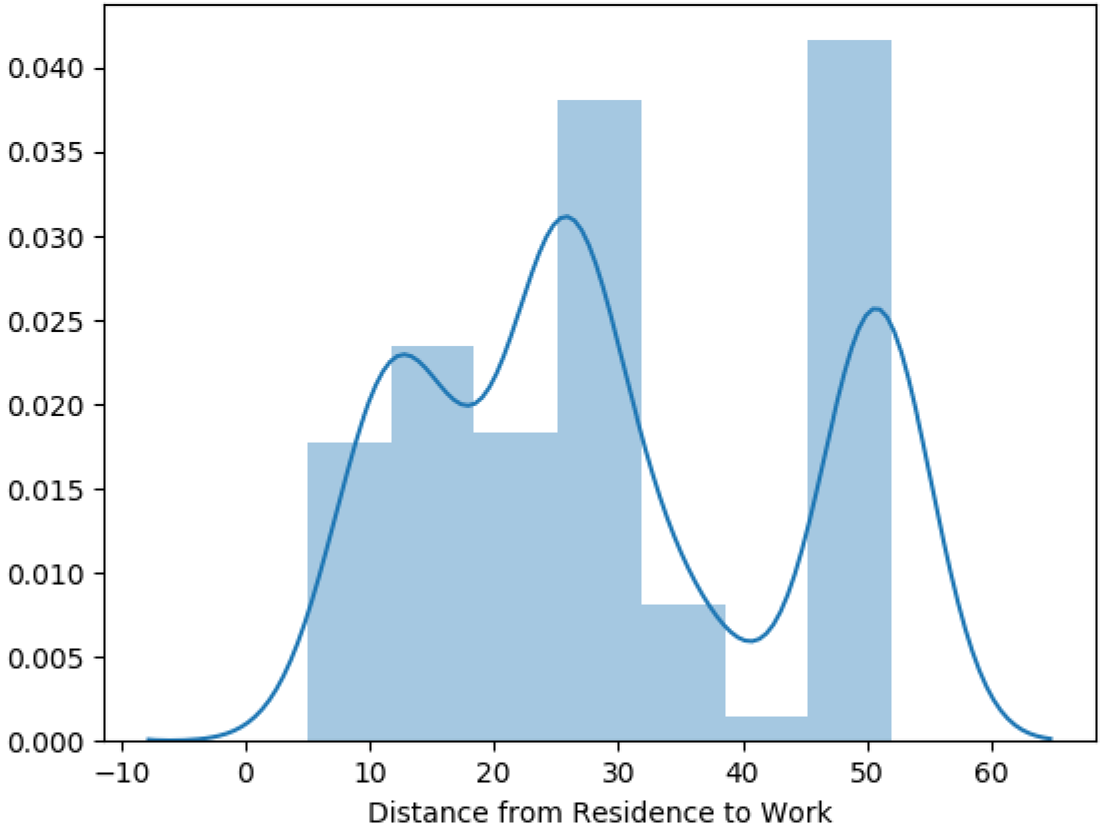
Dummy Encoding and Grouping the data:

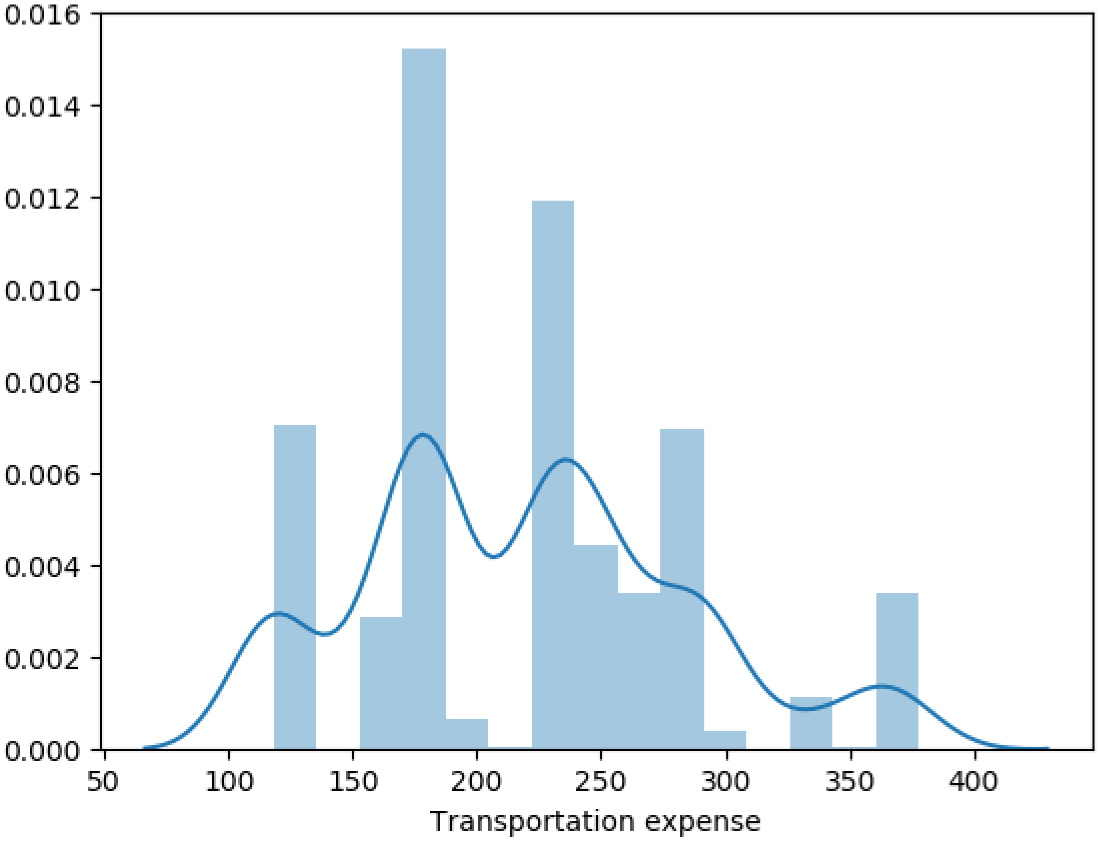
Dummy encoding is done on newly added features.

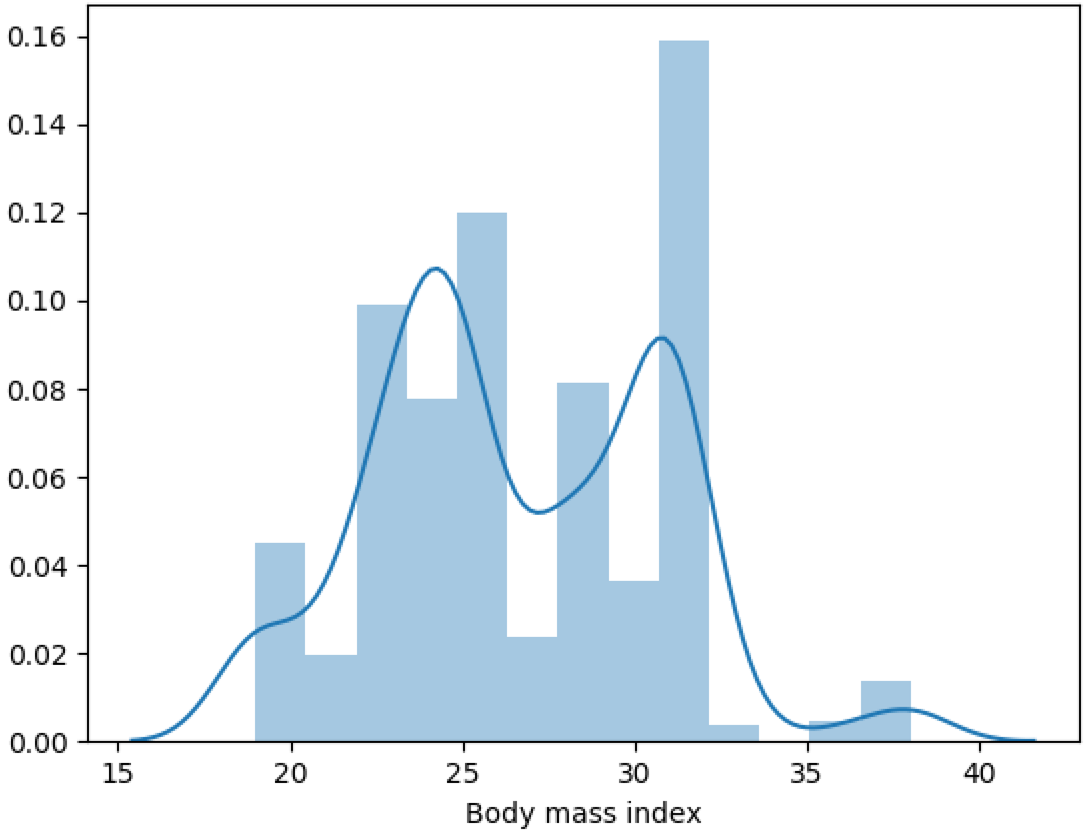
Our aim is to predict losses per month in 2011.Hence we group the entire data by year and month.

3.5 Feature Scaling:

Let us check the distribution of each variable




0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
50
100
150
200
250
300
350
400
Transportation expense



We can see that above variables are not normally distributed.

Hence, it is better using Normalization process for feature scaling

Sampling:

Here we are going to train data for the year 2007, 2008,2009

And test data for 2010 data.

Chapter 4

Modelling

4.1 ML Techniques:

Let us try with different models and compare their performances.

I have created a function which will fit respective models and fine tune the parameters using GridSearchCV and prints MSE and RMSE for values for the model which is used to compare the models for evaluation

```
def fitting_model(model_p,param,xtrain_scaled,ytrain,xtest_scaled,ytest):

    gs_model=GridSearchCV(model_p,cv=5,param_grid=param,scoring='mean_squared_error')

    gs_model.fit(xtrain_scaled,ytrain)

    best_est=gs_model.best_estimator_

    print(best_est)

    print("*********CV Score(mean square)**********************")

    print(gs_model.best_score_)

    y_pred=best_est.predict(xtest_scaled)

    print("*********mean_squared_error**************")

    print(mean_squared_error(ytest,y_pred))

    print("*********RMSE*****************************")

    print(np.sqrt(mean_squared_error(ytest,y_pred)))
```

Random Forest:

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
           max_features='auto', max_leaf_nodes=None,
           min_impurity_decrease=0.0, min_impurity_split=None,
           min_samples_leaf=5, min_samples_split=5,
           min_weight_fraction_leaf=0, n_estimators=10, n_jobs=1,
           oob_score=False, random_state=None, verbose=0, warm_start=False)
*********CV Score(mean_square)************
7.79221316195
*********mean_squared_error**************
30.2045847692
*********RMSE*****************************
5.49586979187
```

Gradient Boosting:

```
GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse', init=None,
             learning_rate=0.1, loss='ls', max_depth=None,
             max_features=None, max_leaf_nodes=None,
             min_impurity_decrease=0.0, min_impurity_split=None,
             min_samples_leaf=20, min_samples_split=3,
             min_weight_fraction_leaf=0.0, n_estimators=10, presort='auto',
             random_state=0, subsample=1.0, verbose=0, warm_start=False)
```

```
********* CV Score(mean_square)**********************
8.65853406441
*********mean_squared_error**************
36.0250007721
*********RMSE*****************************
6.00208303609
```

KNN:

```
KNeighborsRegressor(algorithm='auto', leaf_size=30, metric='minkowski',
          metric_params=None, n_jobs=1, n_neighbors=7, p=2,
          weights='uniform')
********* CV Score(mean_square)**********************
6.81413751155
*********mean_squared_error**************
36.4788953573
*********RMSE*****************************
6.03977610159
```

Decision Tree:

```
DecisionTreeRegressor(criterion='mse', max_depth=None, max_features='log2',
           max_leaf_nodes=None, min_impurity_decrease=0.0,
           min_impurity_split=None, min_samples_leaf=1,
           min_samples_split=2, min_weight_fraction_leaf=0.5,
           presort=False, random_state=0, splitter='best')
```

```
********* CV Score(mean_square)**********************

8.74166794586

*********mean_squared_error**************

35.6422901335

*********RMSE******************************

5.97011642546
```

SVR:

```
SVR(C=7.0, cache_size=200, coef0=0.0, degree=3, epsilon=0.1, gamma=0.5,
  kernel='rbf', max_iter=-1, shrinking=True, tol=0.001, verbose=False)

********* CV Score(mean_square)**********************

-8.07814621653

*********mean_squared_error**************

35.297627233

*********RMSE******************************

5.94118062619
```

LSTM:

LSTM will work with multivariate time series data.

Here we are adding previous lags to our data by "Shift" operator

```
data_pre["T_"+str(i)]=data_pre["Absenteeism time in hours"].shift(i)
```

this will add previous lags to every observation.

This way we have to convert series to supervised data so that our LSTM will work on previous lags and present lags and analyze the output of present observation

Code:

```
model_k=Sequential()

model_k.add(LSTM(1,input_shape=(xtrain_reshape.shape[1], xtrain_reshape.shape[2])))

model_k.add(Dense(1))

model_k.compile(loss='mae', optimizer='adam')

history = model_k.fit(xtrain_reshape, train_data["Absenteeism time in hours"], epochs=50, batch_size=72, validation_data=(xtest_reshape, test["Absenteeism time in hours"]), verbose=2, shuffle=False)
```

Results:

MSE: 56.65

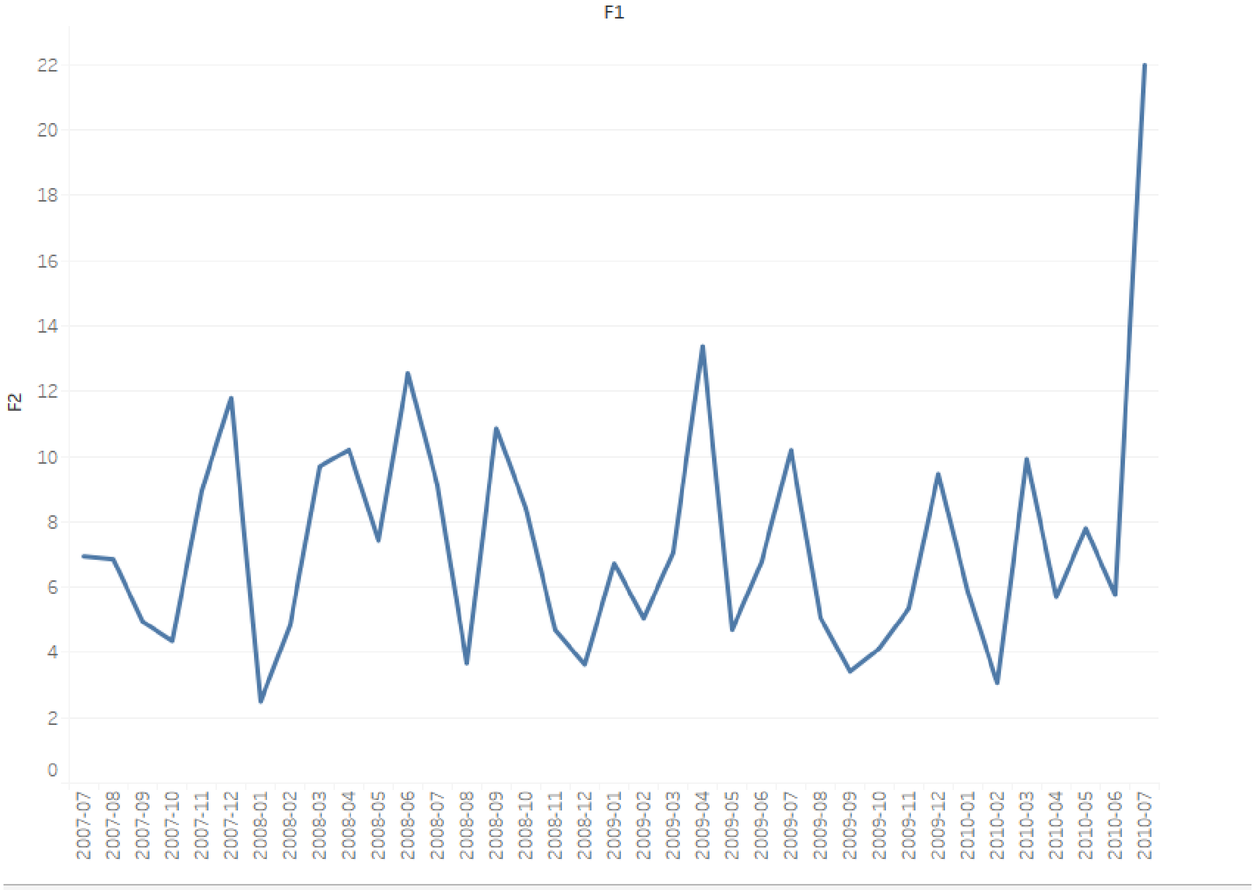
RMSE:7.52

Now let us work with time series and check how it performs

4.2 Time series modelling:

This can be interpreted as time series model too

Let us plot the time series

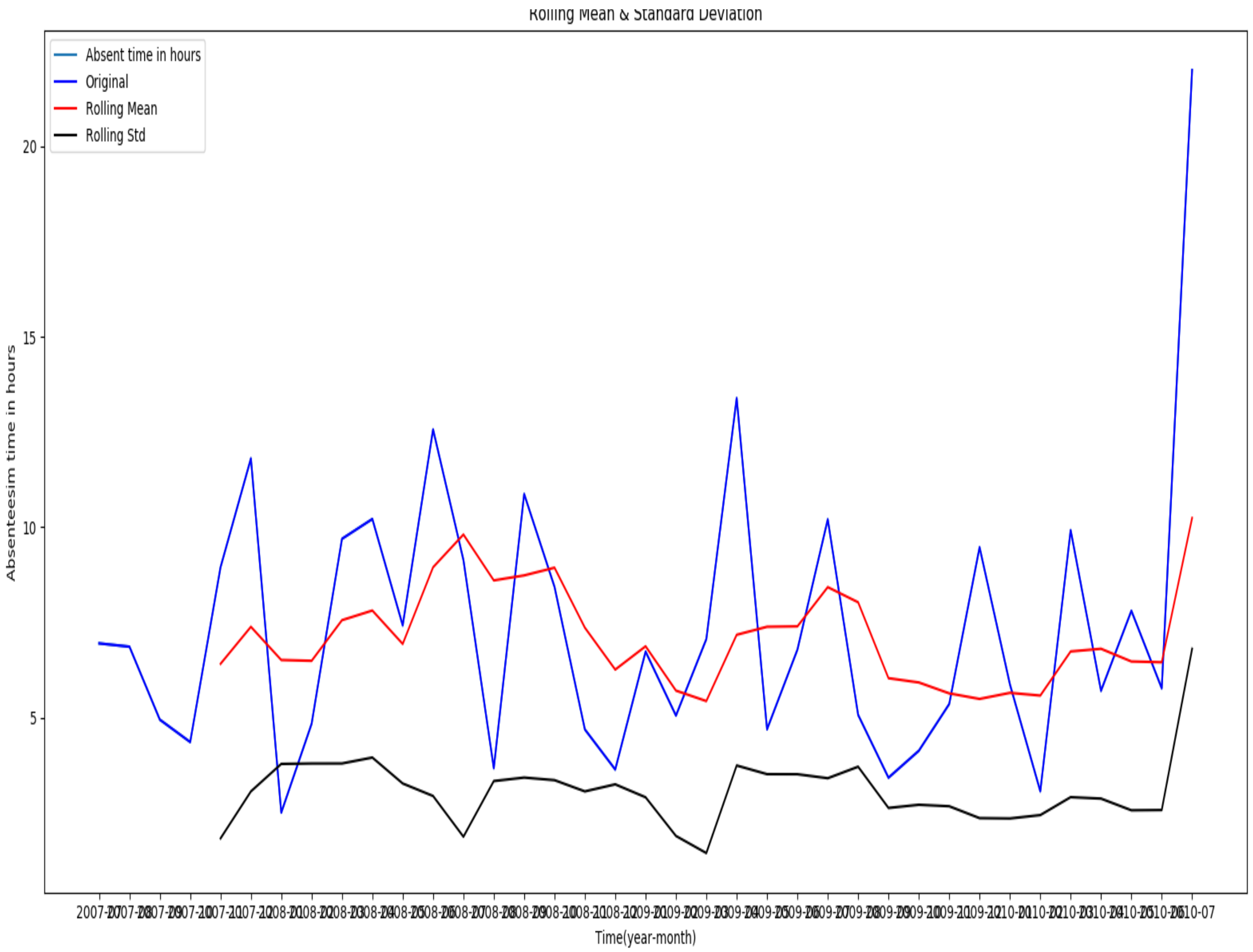


Dickey Fuller Test for Stationarity:

We have constructed a function which will perform dicker fuller test for given time series

Let us run this function and check results.

```
def test_stationarity(timeseries):

    #Determing rolling statistics
    rolmean = pd.rolling_mean(timeseries,window=5)
    rolstd = pd.rolling_std(timeseries,window=5)

    #Plot rolling statistics:
    orig = plt.plot(timeseries, color='blue',label='Original')
    mean = plt.plot(rolmean, color='red', label='Rolling Mean')
    std = plt.plot(rolstd, color='black', label = 'Rolling Std')
    plt.legend(loc='best')
    plt.title('Rolling Mean & Standard Deviation')
    plt.show(block=False)

    #Perform Dickey-Fuller test:
    print ('Results of Dickey-Fuller Test:')
    dftest = adfuller(timeseries, autolag='AIC')
    dfoutput = pd.Series(dftest[0:4], index=['Test Statistic','p-value','#Lags Used','Number of Observations Used'])
    for key,value in dftest[4].items():
        dfoutput['Critical Value (%s)'%key] = value
    print (dfoutput)
```



Results of Dickey-Fuller Test:

```
Test Statistic                 -4.837780
p-value                      0.000046
#Lags Used                      0.000000
Number of Observations Used    36.000000
Critical Value (1%)           -3.626652
Critical Value (5%)           -2.945951
Critical Value (10%)          -2.611671
dtype: float64
```

Null Hypothesis:

Time series is not stationary

Alternative Hypothesis:

Time series is stationary

If Test statistic is less than critical value then reject Null Hypothesis

Here

Test statistic is less than critical value, hence we reject Null Hypothesis concluding that time series is stationary.

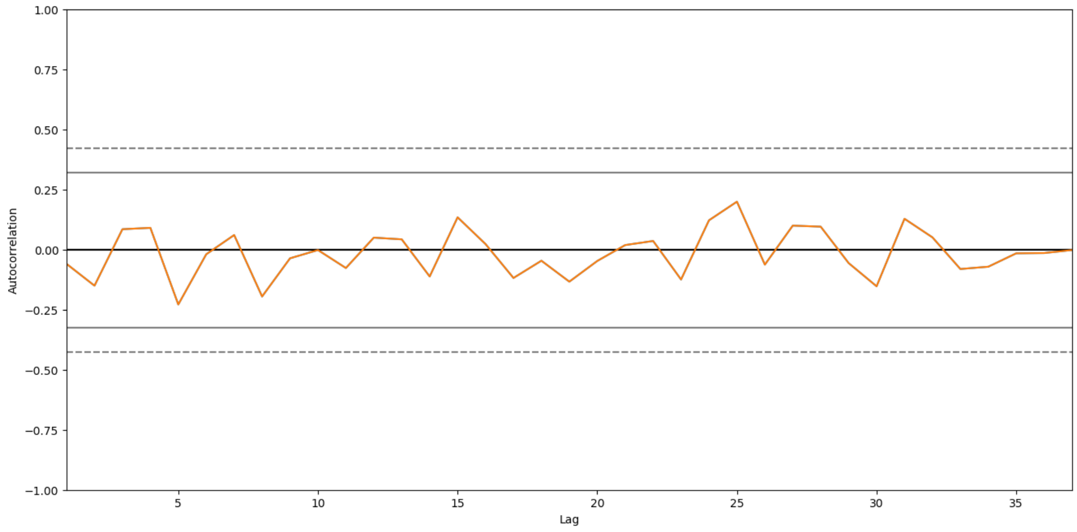
ARIMA Model:

In order to determine p,d,q parameters of ARIMA model, let’s plot ACF plot

ACF Plot:

from pandas.tools.plotting import autocorrelation_plot

autocorrelation_plot(ts)



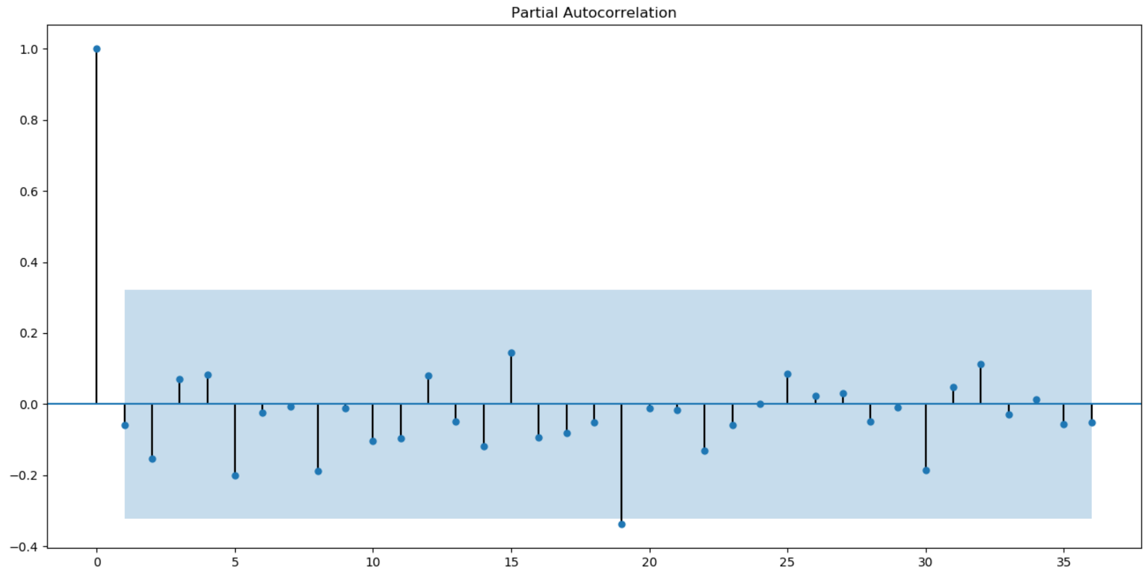
We can see that lags doesn’t even touch boundary lines.

Hence we can try p value with 0

PACF plot(Partial AutoCorrelation plot):

from statsmodels.graphics.tsaplots import plot_pacf

plot_pacf(ts)



We can see that there are two lags out of the boundary. Hence we can try with

q=1 OR 2

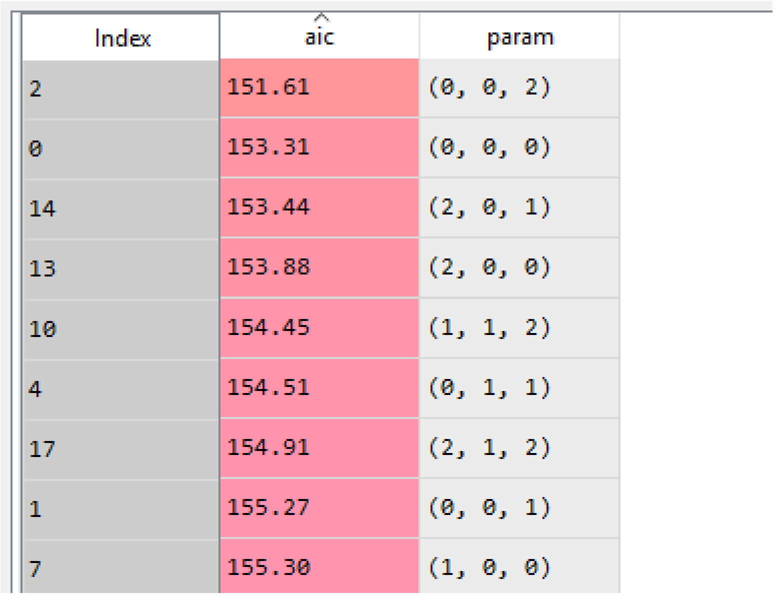
Let’s apply ARIMA model with p,d,q parameters.

We can also tune p,d,q parameters with the help of below code:

```
import statsmodels.api as sm

p=range(3)

d=range(1)

q=range(3)

pdq=list(itertools.product(p,d,q))


train_ts_df=pd.DataFrame(train_ts)

test_ts_df=pd.DataFrame(test_ts)

train_list=[x for x in train_ts_df["Absenteeism time in hours"]]

test_list=[x for x in test_ts_df["Absenteeism time in hours"]]
```

```
min_aic_list=[]

for param in pdq:

    try:

        model=ARIMA(train_list,order=param)

        results = model.fit()

        min_aic_list.append([results.aic,param])

    except:

        continue


min_aic_df=pd.DataFrame(min_aic_list,columns=["aic","param"])
```

Here p,d,q are iterated for number of values and we can consider p,d,q parameters for which aic value is less



Index	aic	param
2	151.61	(0, 0, 2)
0	153.31	(0, 0, 0)
14	153.44	(2, 0, 1)
13	153.88	(2, 0, 0)
10	154.45	(1, 1, 2)
4	154.51	(0, 1, 1)
17	154.91	(2, 1, 2)
1	155.27	(0, 0, 1)
7	155.30	(1, 0, 0)

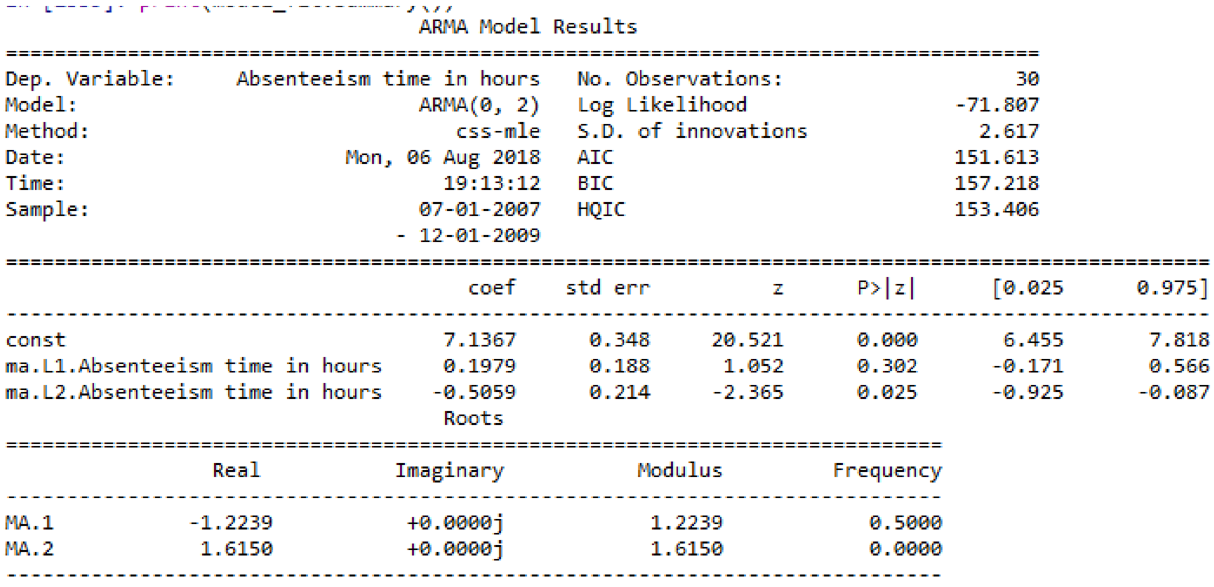
We can see that (0,0,2) parameter has less value of 151.

Hence order (0,0,2) is considered for the model.

Let us fit the model with (p,d,q) as (0,0,2) and check the results:

Train Results:

Model Summary:



```
                              ARMA Model Results
==============================================================================
Dep. Variable:     Absenteeism time in hours   No. Observations:                   30
Model:                            ARMA(0, 2)   Log Likelihood                 -71.807
Method:                              css-mle   S.D. of innovations              2.617
Date:                       Mon, 06 Aug 2018   AIC                            151.613
Time:                               19:13:12   BIC                            157.218
Sample:                           07-01-2007   HQIC                           153.406
                                - 12-01-2009
=================================================================================================
                                     coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                              7.1367      0.348     20.521      0.000       6.455       7.818
ma.L1.Absenteeism time in hours    0.1979      0.188      1.052      0.302      -0.171       0.566
ma.L2.Absenteeism time in hours   -0.5059      0.214     -2.365      0.025      -0.925      -0.087
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
MA.1           -1.2239           +0.0000j            1.2239            0.5000
MA.2            1.6150           +0.0000j            1.6150            0.0000
-----------------------------------------------------------------------------
```

Residual Plot:

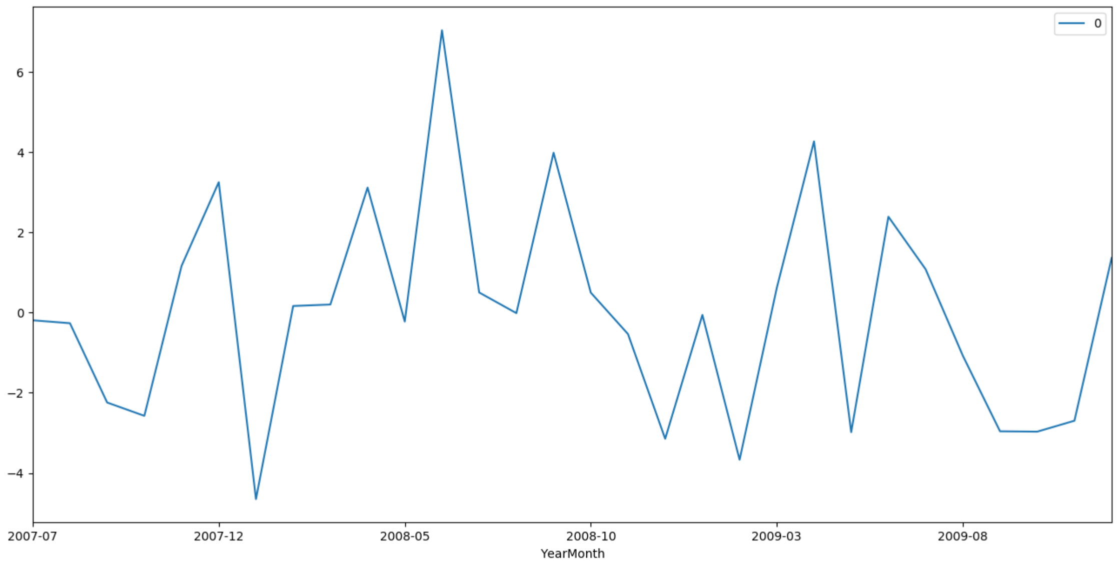
residuals_df=pd.DataFrame(model_fit.resid)

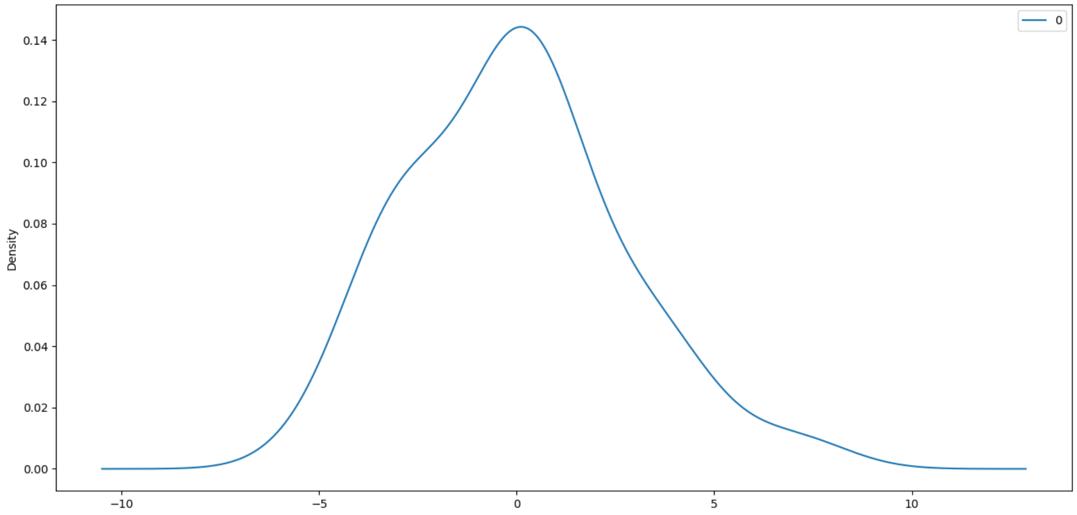
residuals_df.plot()

Residual Distribution:

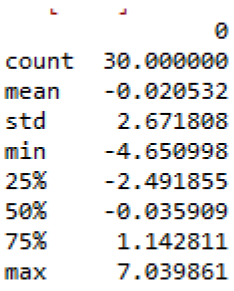
residuals_df.plot(kind='kde')

residuals_df.describe()





Residual Statistics:



```
               0
count  30.000000
mean   -0.020532
std     2.671808
min    -4.650998
25%    -2.491855
50%    -0.035909
75%     1.142811
max     7.039861
```

Test Results:

Train data: 2007, 2008, 2009

Test data: 2010

MSE: 44.22636137027308

RMSE: 6.650290322254592

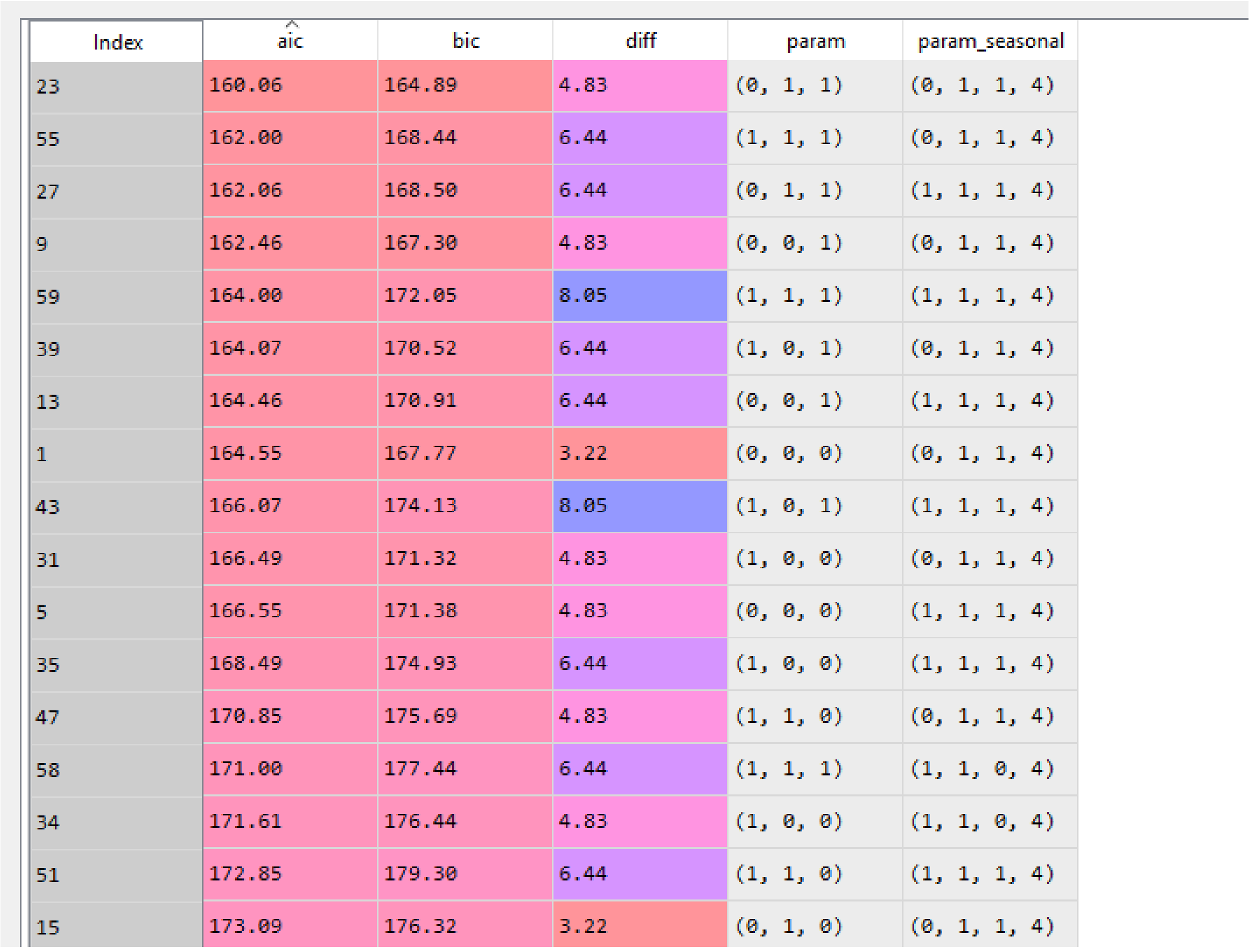
AIC score: 161

Let us apply SARIMAX model and check the results:

SARIMAX:

Tuning parameters for order and seasonal order:

```
p=range(2)
d=range(2)
q=range(2)
pdq=list(itertools.product(p,d,q))
seasonal_pdq = [(x[0], x[1], x[2], 4) for x in list(itertools.product(p, d, q))]
min_aic_list=[]
for param in pdq:
    for param_seasonal in seasonal_pdq:
        try:
            mod = sm.tsa.statespace.SARIMAX(ts,
                order=param,
                seasonal_order=param_seasonal,
                enforce_stationarity=False,
                enforce_invertibility=False)
            results = mod.fit()
            min_aic_list.append([results.aic,results.bic,(results.bic-results.aic),param,param_seasonal])
        except:
          continue

min_aic_df1=pd.DataFrame(min_aic_list,columns=["aic","bic","diff","param","param_seasonal"])
```



Index	aic	bic	diff	param	param_seasonal
23	160.06	164.89	4.83	(0, 1, 1)	(0, 1, 1, 4)
55	162.00	168.44	6.44	(1, 1, 1)	(0, 1, 1, 4)
27	162.06	168.50	6.44	(0, 1, 1)	(1, 1, 1, 4)
9	162.46	167.30	4.83	(0, 0, 1)	(0, 1, 1, 4)
59	164.00	172.05	8.05	(1, 1, 1)	(1, 1, 1, 4)
39	164.07	170.52	6.44	(1, 0, 1)	(0, 1, 1, 4)
13	164.46	170.91	6.44	(0, 0, 1)	(1, 1, 1, 4)
1	164.55	167.77	3.22	(0, 0, 0)	(0, 1, 1, 4)
43	166.07	174.13	8.05	(1, 0, 1)	(1, 1, 1, 4)
31	166.49	171.32	4.83	(1, 0, 0)	(0, 1, 1, 4)
5	166.55	171.38	4.83	(0, 0, 0)	(1, 1, 1, 4)
35	168.49	174.93	6.44	(1, 0, 0)	(1, 1, 1, 4)
47	170.85	175.69	4.83	(1, 1, 0)	(0, 1, 1, 4)
58	171.00	177.44	6.44	(1, 1, 1)	(1, 1, 0, 4)
34	171.61	176.44	4.83	(1, 0, 0)	(1, 1, 0, 4)
51	172.85	179.30	6.44	(1, 1, 0)	(1, 1, 1, 4)
15	173.09	176.32	3.22	(0, 1, 0)	(0, 1, 1, 4)

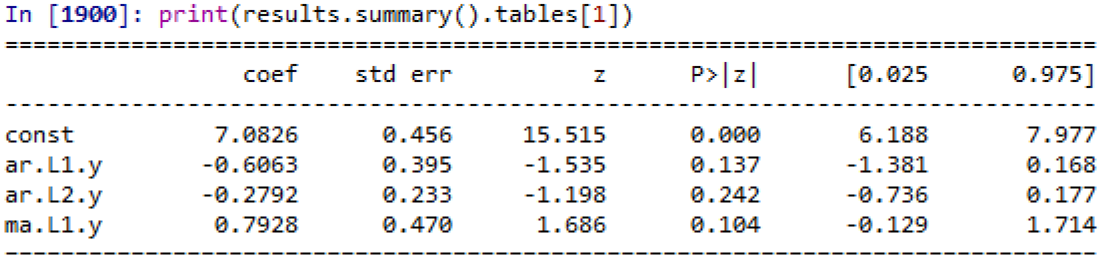
Order : (0,1,1)

Seasonal Order: (0,1,1,4)

AIC: 160

Let us fit the model for training data and check the resultsprint(results.summary().tables[1])

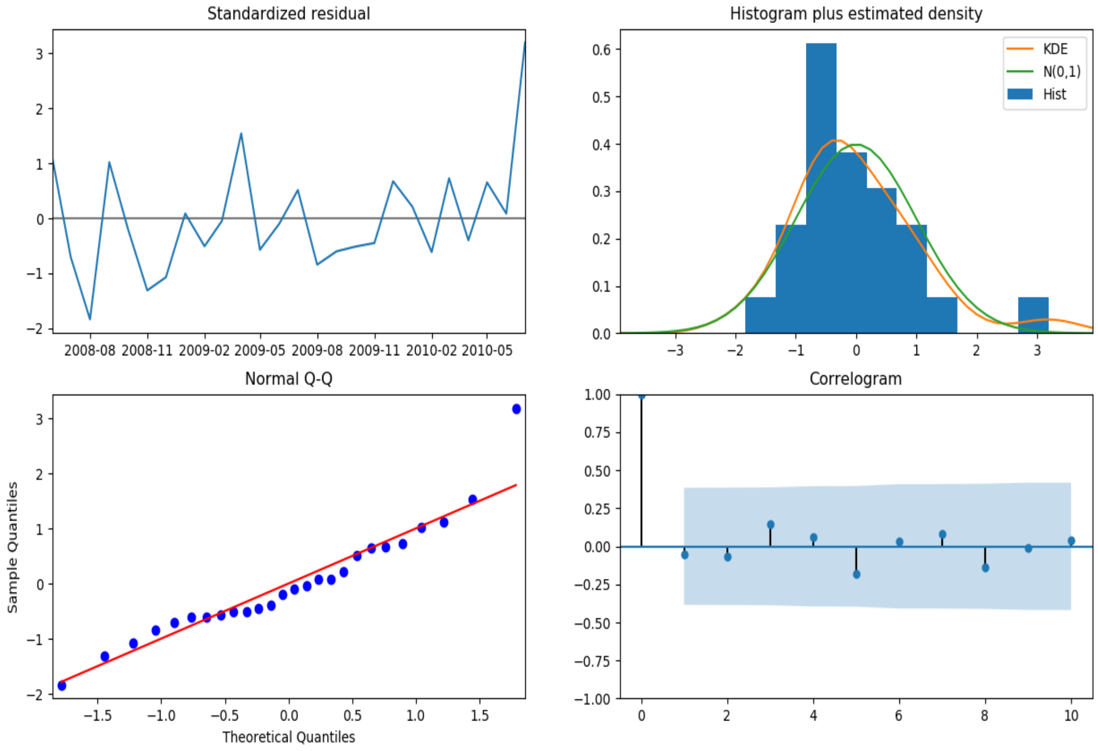
Model Summary:



```
In [1900]: print(results.summary().tables[1])
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          7.0826      0.456     15.515      0.000       6.188       7.977
ar.L1.y       -0.6063      0.395     -1.535      0.137      -1.381       0.168
ar.L2.y       -0.2792      0.233     -1.198      0.242      -0.736       0.177
ma.L1.y        0.7928      0.470      1.686      0.104      -0.129       1.714
------------------------------------------------------------------------------
```

Diagnostics:

results.plot_diagnostics(figsize=(15, 12))



Residual distribution is slightly left skewed.

We can see there is one lag out of region.

We can see that standardized residuals follows training data.

From Normal Q-Q, we can see that fit is good.

Test Results:

MSE: 33.20086064163015

RMSE: 5.762018799138905

AIC: 160

We got less AIC score than ARIMA model.

Chapter 5

Model Selection for Forecasting:

Let us check the MSE of basic models:

Random Forest: 30.20

Gradient Boost: 36.02

KNN: 36.47

Decision Tree: 35.64

SVR: 35.29

Clearly we can see that Random Forest performs better than all other models

Let us compare MSE for Time Series Models:

SARIMAX: 33.20

ARIMA: MSE: 44.22

Clearly we can see that SARIMAX has less MSE and also we can see that residuals of SARIMAX follows normal distribution and fitting of SARIMAX is better than ARIMA model.

AIC of both time series models are almost same.

AIC and BIC difference is not much high for SARIMAX

Random Forest and SARIMAX MSE's are approximately same.

Our aim is to predict losses for 2011, hence SARIMAX model is known for better forecasting.

Also predictor variables for 2011 are not given. Hence SARIMAX model is chosen for solving the problem.

SARIMAX models has following advantages:

- Season variations and peaks are better captured.
- Good AIC and BIC score
- Less MSE and RMSE
- Better Forecasting than other models
- Good Fit
- Residuals follow normal distribution

Chapter 6

Forecasting 2011 Absentees

Our aim is to predict losses for 2011 year.

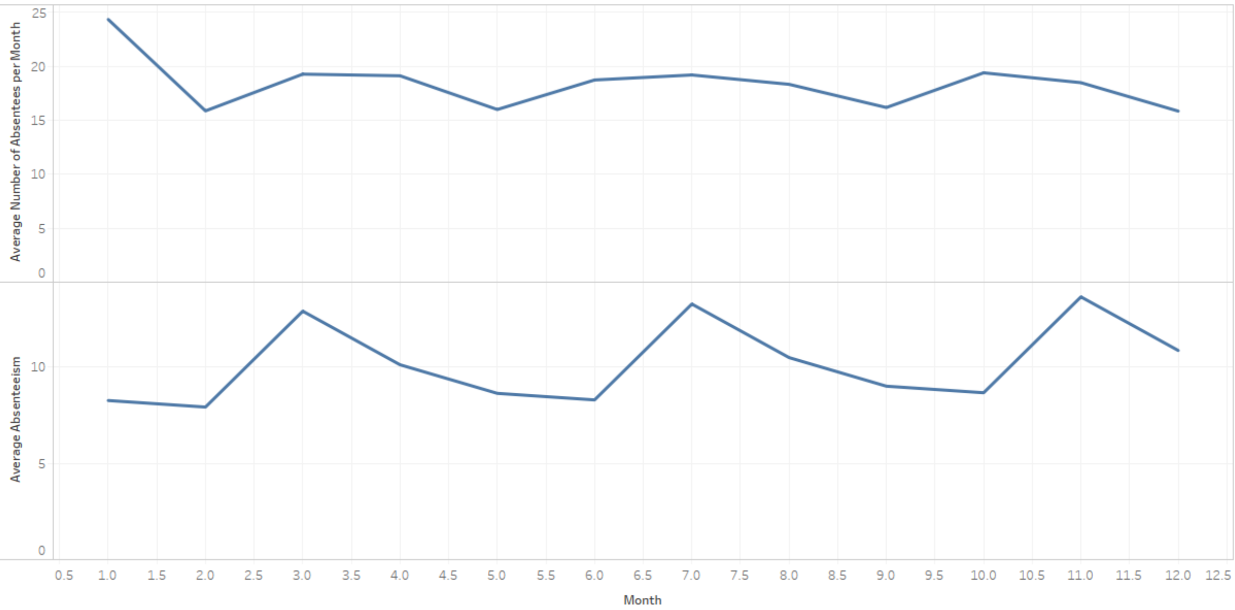
So let’s forecast Absenteeism for the year 2011.

It’s been said that predict losses for 2011 if same trend of absenteeism is continued.

Total Absenteeism time in hours/month= (Number of Absentees per month * Average Absenteeism per month)

Let us check the variation of

Number of Absentees per month
Average Absenteeism per month



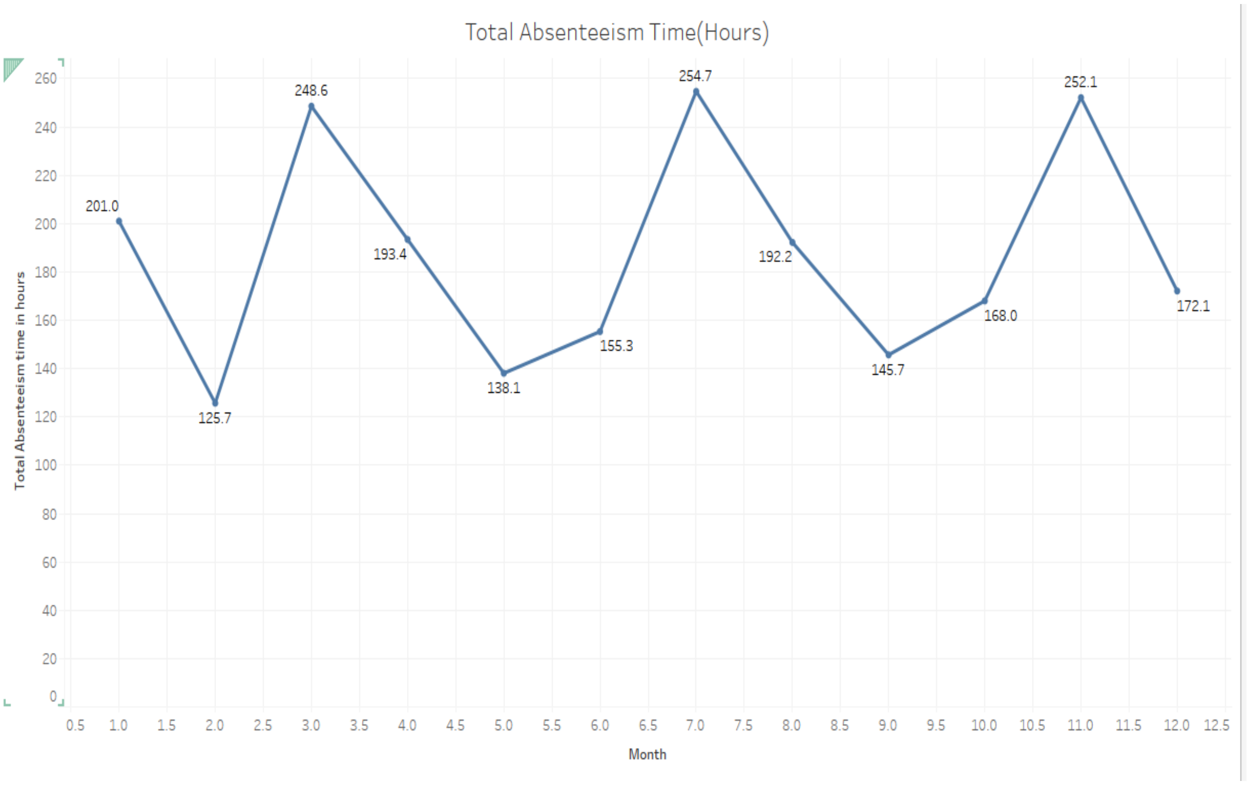
We have plotted **Number of Absentees per month**

We can see the trend of Absenteeism for 2008, 2009, 2010.

Let us assume this trend continues for 2011.

From SARIMAX output, we got **Average Absenteeism per month.**

From these two results we calculate Total Absenteeism/month (in hours) for 2011 year



We can see some peaks at

March:

- May be because of winter
- Vacation time

July(May be because of summer)

Summer and winters are extreme in Brazil which may leads to diseased absenteeism

November:

May be because of

- year end
- long leaves
- vacations
- Famous Brazil carnival
- Christmas eve

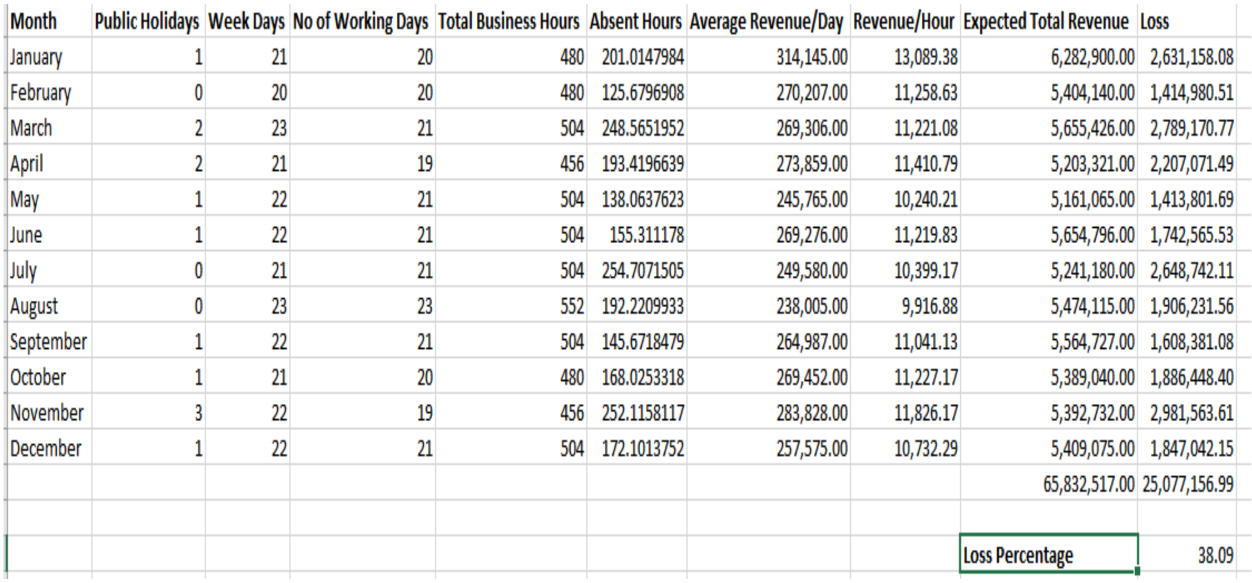
Carnival (approximately a month before Easter)

CC-BY Prefeitura de Olinda on Flickr

This festival has become synonymous with Brazil and is celebrated all over the country. Rio is the best place to be during Carnival, as it is celebrated on a much more grandiose scale than anywhere else. This festival takes place over 5 days, of which all days are not national holidays but rather days of observance. However, Shrove Tuesday (one of the last days of Carnival) is considered a national holiday. During this period, the whole of Brazil goes into festival mode.

Chapter 7

Predicting losses based on Total Absentees per Month



Month	Public Holidays	Week Days	No of Working Days	Total Business Hours	Absent Hours	Average Revenue/Day	Revenue/Hour	Expected Total Revenue	Loss
January	1	21	20	480	201.0147984	314,145.00	13,089.38	6,282,900.00	2,631,158.08
February	0	20	20	480	125.6796908	270,207.00	11,258.63	5,404,140.00	1,414,980.51
March	2	23	21	504	248.5651952	269,306.00	11,221.08	5,655,426.00	2,789,170.77
April	2	21	19	456	193.4196639	273,859.00	11,410.79	5,203,321.00	2,207,071.49
May	1	22	21	504	138.0637623	245,765.00	10,240.21	5,161,065.00	1,413,801.69
June	1	22	21	504	155.311178	269,276.00	11,219.83	5,654,796.00	1,742,565.53
July	0	21	21	504	254.7071505	249,580.00	10,399.17	5,241,180.00	2,648,742.11
August	0	23	23	552	192.2209933	238,005.00	9,916.88	5,474,115.00	1,906,231.56
September	1	22	21	504	145.6718479	264,987.00	11,041.13	5,564,727.00	1,608,381.08
October	1	21	20	480	168.0253318	269,452.00	11,227.17	5,389,040.00	1,886,448.40
November	3	22	19	456	252.1158117	283,828.00	11,826.17	5,392,732.00	2,981,563.61
December	1	22	21	504	172.1013752	257,575.00	10,732.29	5,409,075.00	1,847,042.15
								65,832,517.00	25,077,156.99
								Loss Percentage	38.09

Formulae:

No of Working days=No of Weekdays-Public holidays

We got absent hours from model

Total Business hours=No of Working Days*24

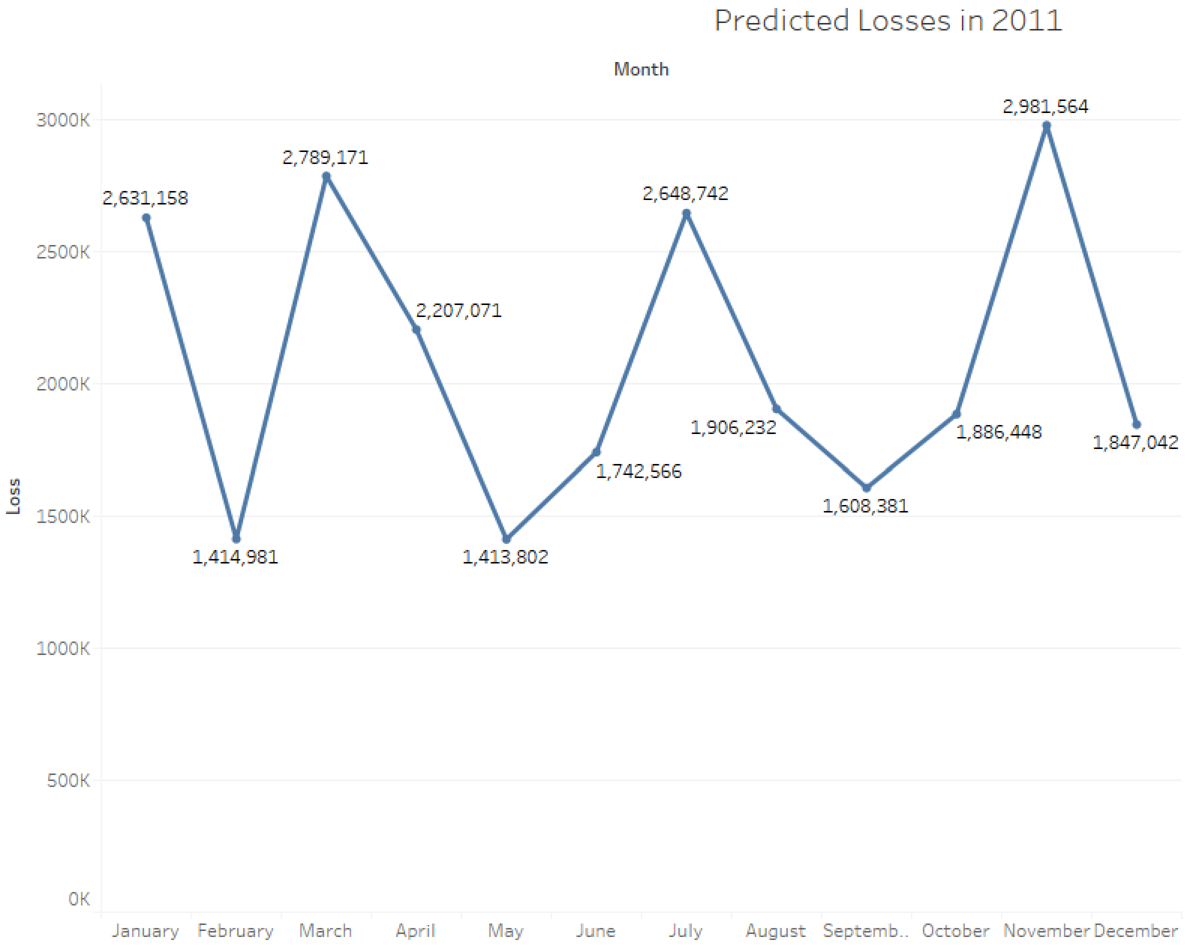
We got Average Revenue/Day for 2008 and 2009 years per month

Revenue per Hour=Average Revenue/24

Expected Total Revenue=Total Hours of Business*Revenue/Hour

Loss=Absent Hours*Revenue/Hour

Let us check loss per month in the year 2011



Total losses for the year=25,077,156.99

Total Expected Revenue= 65,832,517.00

Loss Percentage: 38%

Chapter 8:

Strategies to Reduce Losses

Company is losing 38% of earnings due to absenteeism.

We can decrease these loss percentage by following some strategies.

1. **Hiring of Temporary Labor in March, July and November:**

   For March, July, November, losses are more which are average loss is 25 million (approximately) which is too high.

   Hiring Temporary labor costs us about, let's say 2-5 million.

   In this way, we can save 15-20 million to the company by hiring temporary labor

2. **Call for Overtime or Work for Weekend in Febraury,May,June,September,October:**

   In these months, less absenteeism is recorded which is about 100 hours.So we can encourage employees to extend their time and also allocating shifts on weekends. We can offer incentives to customers which is much less than the losses suffered by the company.

3. **Unjustified Absence:**
   Unjustified Absence should not encouraged.Strict policies must be imposed on employees

4. **Absenteeism on Week Start:**
   We can see that there is more absentees on Monday which is not recommended. Employees are properly monitored and instructed to avoid this issue. Leaves have to be approved only if their reason is genuine

5. **Noon Time Absenteeism:**
   Persons who are near to office are tend to be absent at noon and employee absenteeism is recorded more due to summer season because of high temperatures.

Partial shifts: Arriving late, leaving early and taking longer breaks than allowed are considered forms of absenteeism and can affect productivity and workplace morale.

6. **Fighting Seasonal Effect:**

   In Brazil, summer and winter seasons are extreme. It will results to so many skin related issues. High Temperatures in summer results in skin related and physical issues.
   And winter season results in so many contagious diseases.Precautionary measures are taken to fight seasonality

7. **Continuous Monitoring of Employee Behavior:**

   These are some of employee ids

   3,11,14,28,34,36,20,9,24,15,22 whose absenteeism is more. These employees are properly checked whether their absenteeism is genuine or not. If their absenteeism is not genuine.

   Employees may call in sick to attend a job interview.

8. **Dealing with the reason "Consultations":**

   Injuries, illness and medical appointments are the most commonly reported reasons for missing work (though not always the actual reason).
   For consultations alone we lose 900 hours which are huge ones.
   Employees are strictly advised to take medical consultations on weekends.
   If they are dental, ENT consultations they can be postponed to weekends.
   If they are emergency issues, they are properly checked for genuinity
   by checking medical slips and appointment letters.

9. **Blood Donations:**

   Blood donations cannot be allowed on weekdays. If it is emergency, it can be allowed maximum up to 1 hour

10. **Injuries, Poisoning and Physical Issues:**

    More absenteeism is due to follow issues.
    Injuries are due employee's negligence, hurry, no proper time maintenance.
    These issues are properly addressed .They must be properly guided to avoid injuries due to accidents.
    Awareness has to be created about the food intake to avoid food poisoning

11. **Handling Long Leaves before Christmas:**

    More absenteeism occurs in November due to Christmas season.
    People tend to take long leaves on the eve of Christmas. Employees have to be given extra bonus who work in November and December.

Conclusion:

Thus our model helps to assist in decision-making, since production can be maintained with planned measures such as distribution holidays and measures of maintenance of production as

Call for overtime and / or work over the weekend or holiday, hiring of temporary labor, etc. With the advantage of the planning time for the summoning of employees in advance or contracting temporary.