**Day-7 Assignment**

1. Load data to Pandas Dataframe

|  |
| --- |
| import pandas as pd  import matplotlib.pyplot as plt  data=pd.read\_csv('general\_data.csv')  masterData=pd.DataFrame(data) |

1. Check for nulls and removed null rows.

*Found null values in 2 columns ‘NumCompaniesWorked’(19 Rows) and ‘TotalWorkingYears’(9 Rows)*

|  |
| --- |
| #Checking for NULLs  masterData.isnull().any()  TotalWorkingYears\_nulls = masterData['TotalWorkingYears'].isnull().value\_counts()  NumCompaniesWorked\_nulls = masterData['NumCompaniesWorked'].isnull().value\_counts()  print(TotalWorkingYears\_nulls)  print(NumCompaniesWorked\_nulls) |

1. Removing null Rows.

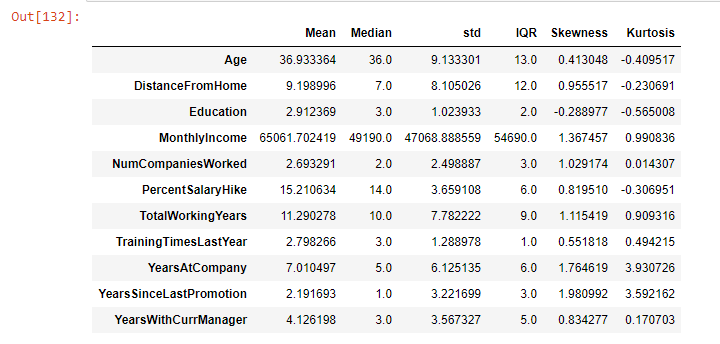
|  |
| --- |
| nullExclued= masterData.dropna() |

1. Taking the required columns into consideration

|  |
| --- |
| columns\_to\_consider=['Age','DistanceFromHome','Education','MonthlyIncome',  'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',  'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager'] |

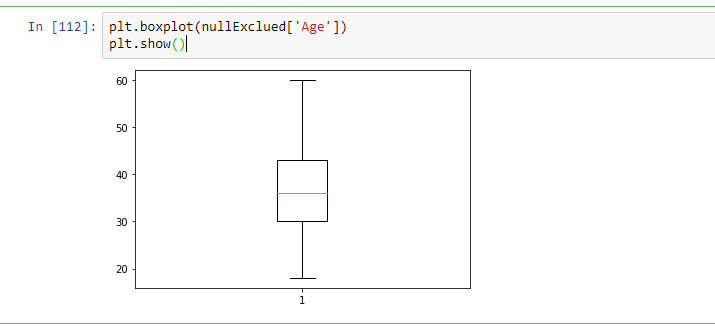
1. Calculating Mean,Median, Skewness, Peakness of the columns in consideration

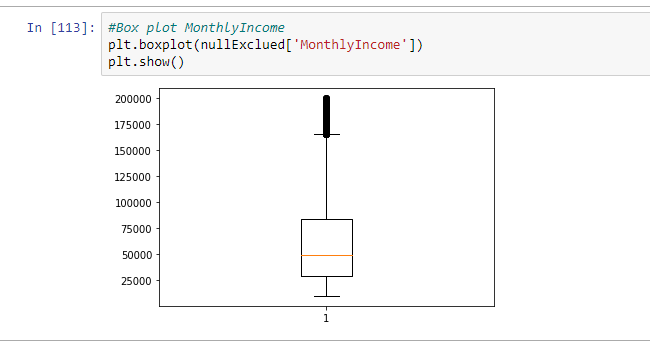
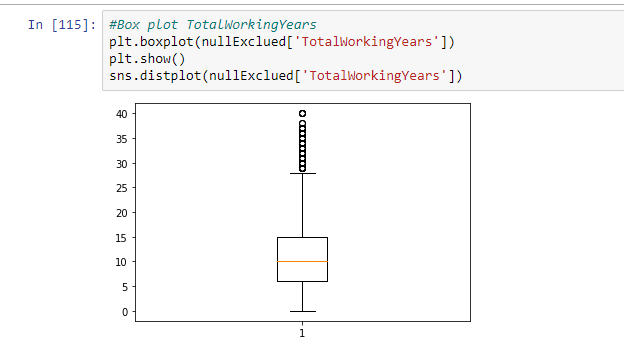
|  |
| --- |
| desc= masterData[columns\_to\_consider].describe().T  df\_mean=pd.Series(nullExclued[columns\_to\_consider].mean())  df\_median=pd.Series(nullExclued[columns\_to\_consider].median())  #df\_mode1=pd.Series(nullExclued[columns\_to\_consider].mode())  df\_skew=pd.Series(nullExclued[columns\_to\_consider].skew())  df\_kurt=pd.Series(nullExclued[columns\_to\_consider].kurt())  df\_IQR =pd.Series(desc['75%']-desc['25%'])  df\_std=pd.Series(desc['std'])  df\_stats=pd.DataFrame([df\_mean,df\_median,df\_std,df\_IQR,df\_skew,df\_kurt]).T  df\_stats.columns=['Mean','Median','std','IQR','Skewness','Kurtosis']  df\_stats |

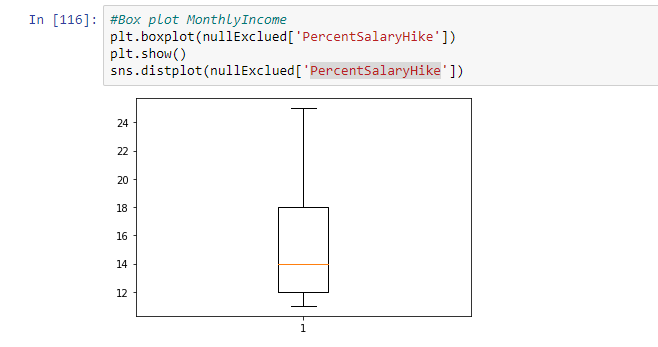


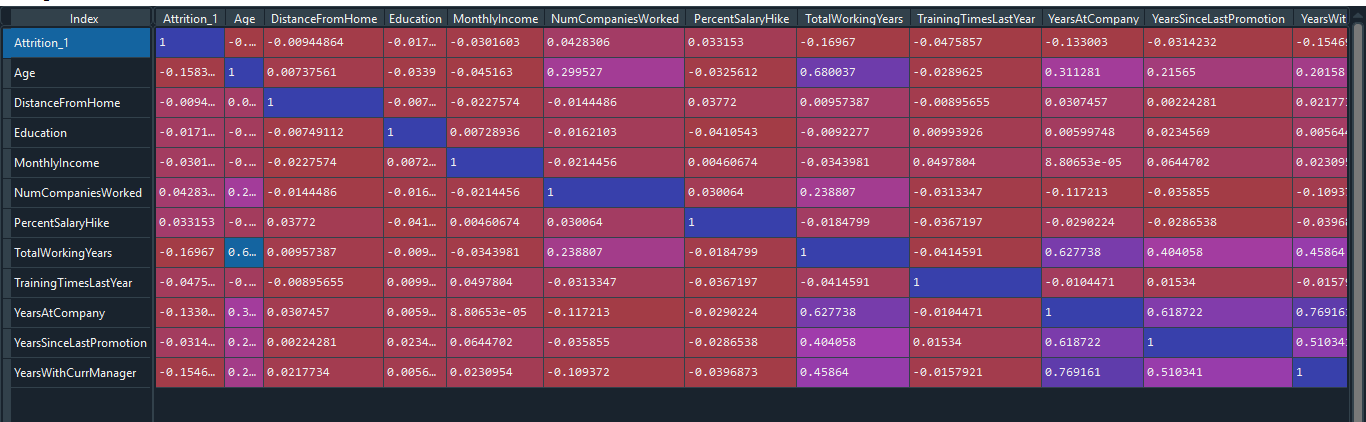
1. Box plot for Age.

Age is normally distributed in the company. Its is symmetrical and no outlayers.



1. Box plot MonthlySalary  
   Monthly salary has outlayers. It is positively skewed  
     
   
2. Box plot of TotalWorkingYears  
   TotalWorkingYears has outliners and is positively skewed. TotalWorkingYears (1.11) and MonthlySalary(1.37) has similar Skewness.  
   
3. Box plot for **PercentSalaryHike.**PercentSalaryHike has no outliners and is positively skewed.



1. **Correlation matrix**   
     
   

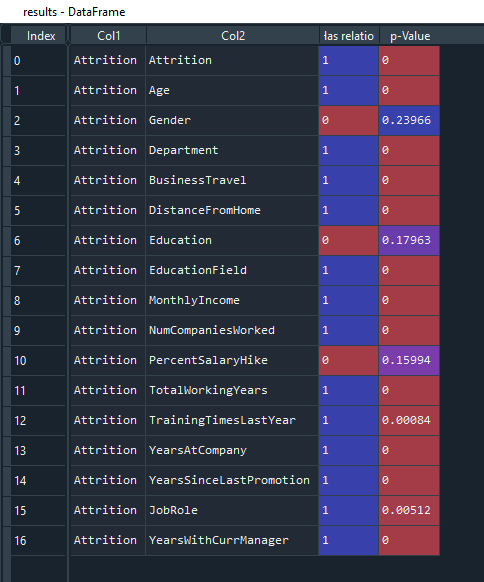
**Observations:**

* There is almost zero correlation between Attrition and other columns.
* There is slight correlation between ‘TotalWorkingYears’(-0.17),YearsAtCompany(-0.13), YearsWithCurrentManager(-0.15)

1. Performed Chi2 test on the categorical columns with Attrition column.

**Analysis:**

* There is no significant relation between Attrition and Gender,Eduction,PercentSalaryHike.



1. Performed MannWhiteny test, indendent ttest, PearsonR test with the columns affect on Attrition.  
   Following are the inferences:  
   DistanceFromHome, Education has no impact on Attrition.  
     
   