Correlation

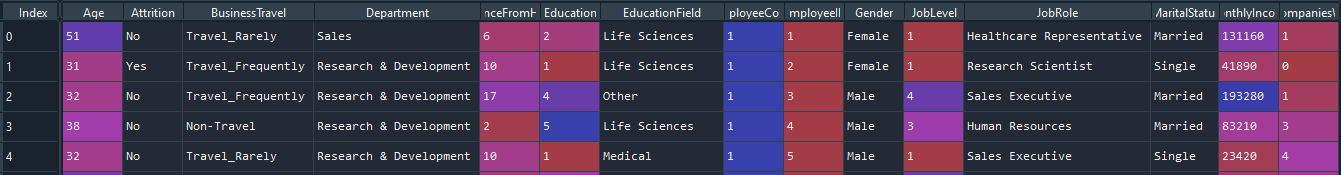
import pandas as pd

import matplotlib.pyplot as plt

**# Read dataset and display**

dataset = pd.read\_csv("general\_data.csv")

dataset.head()



**# check null dataset**

dataset.isnull().sum()

**# drop null values in dataset**

df = dataset.dropna()

data1 = df.isnull().sum()



**# Replace Attrition values yes = 0 and no = 1**

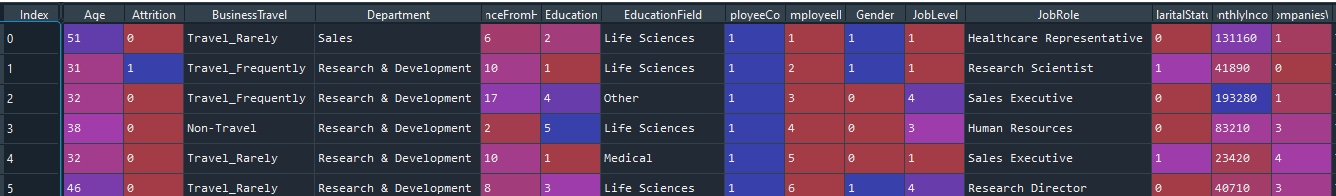
df['Attrition'].replace({"Yes" : 1,"No" : 0},inplace=True)

**# Replace Gender values male = 0 and female = 1**

df['Gender'].replace({"Male":0,"Female":1},inplace=True)

**# Replace MaritalStatus values us Married = 0, single= 1, divorced= 2**

df['MaritalStatus'].replace({"Married": 0, "Single" : 1,"Divorced": 2}, inplace= True)



**Correlation**

from scipy.stats import pearsonr

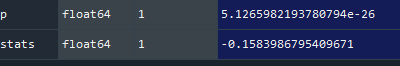
**# Correlation of Attrition and Age**

H0 -> There is no significant different between Attrition and Age

H1 -> There is significant different between Attrition and Age

stats,p = pearsonr(df.Attrition,df.Age)

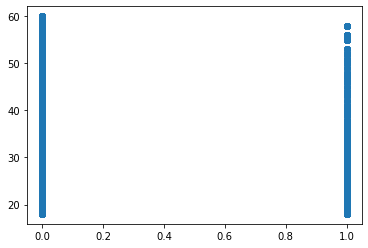
print(stats,p)



IF the **P <0.05,**  so H0 is rejected, we accept alternative H1

plt.scatter(df.Attrition,df.Age)

plt.show()



**# Correlation of Attrition and Distancefromhome**

H0 -> There is no significant different between Attrition and Distancefromhome

H1 -> There is significant different between Attrition and Distancefromhome

stats,p = pearsonr(df.Attrition,df.DistanceFromHome)

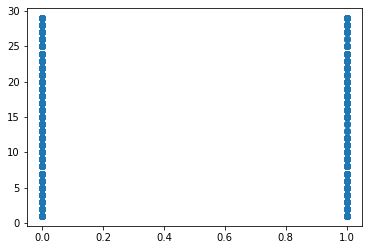
print(stats,p)



If **P>0.05**, we accept H0

plt.scatter(df.Attrition,df.DistanceFromHome)

plt.show()



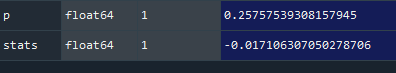
**# Correlation of Attrition and Education**

H0 -> There is no significant different between Attrition and Education

H1 -> There is significant different between Attrition and Education

stats,p = pearsonr(df.Attrition,df.Education)

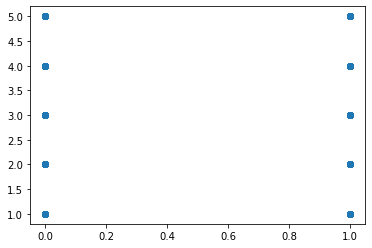
print(stats,p)



If **P>0.05** , we accept Ho

plt.scatter(df.Attrition,df.Education)

plt.show()



**# Correlation of Attrition and job level**

H0 -> There is no significant different between Attrition and Joblevel

H1 -> There is significant different between Attrition and Joblevel

stats,p = pearsonr(df.Attrition,df.JobLevel)

print(stats,p)



If **P>0.05 ,** we accept H0

plt.scatter(df.Attrition,df.JobLevel)

plt.show()

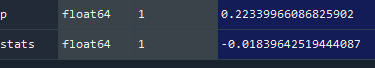
**# Correlation of Attrition and Gender**

H0 -> There is no significant different between Attrition and Gender

H1 -> There is significant different between Attrition and Gender

stats,p = pearsonr(df.Attrition,df.Gender)

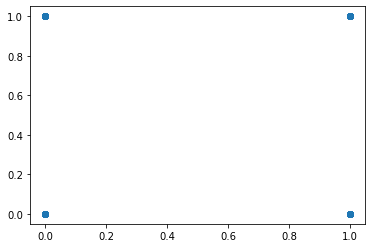
print(stats,p)



If **P>0.05** we accept H0

plt.scatter(df.Attrition,df.Gender)

plt.show()



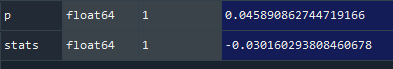
**# Correlation of Attrition and MonthlyIncome**

H0 -> There is no significant different between Attrition and MonthlyIncome

H1 -> There is significant different between Attrition and Monthly Income

stats,p = pearsonr(df.Attrition,df.MonthlyIncome)

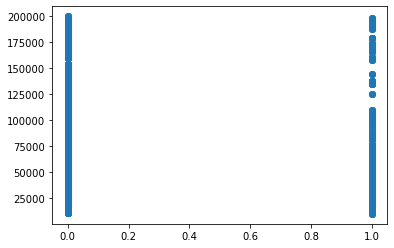
print(stats,p)



If **P<0.05** we reject H0 and Accept H1

plt.scatter(df.Attrition,df.MonthlyIncome)

plt.show()



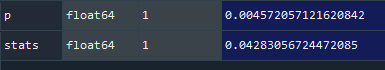
**# Correlation of Attrition and NumCompaniesWorked**

H0 -> There is no significant different between Attrition and NumCompaniesWorked

H1 -> There is significant different between Attrition and NumCompaniesWorked

stats,p = pearsonr(df.Attrition,df.NumCompaniesWorked)

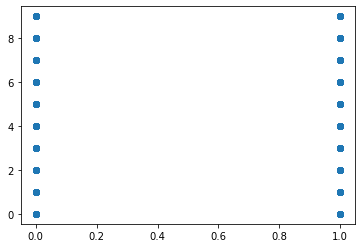
print(stats,p)



If **P<0.05**  we reject H0

plt.scatter(df.Attrition,df.NumCompaniesWorked)

plt.show()



**# Correlation of Attrition and PercentSalaryHike**

H0 -> There is no significant different between Attrition and PercentSalaryHike

H1 -> There is significant different between Attrition and PercentSalaryHike

stats,p = pearsonr(df.Attrition,df.PercentSalaryHike)

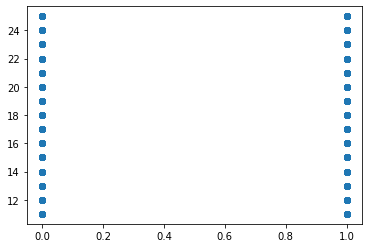
print(stats,p)



If **P<0.05** we reject H0

plt.scatter(df.Attrition,df.PercentSalaryHike)

plt.show()



**# Correlation of Attrition and TotalWorkingYears**

H0 -> There is no significant different between Attrition and TotalWorkingYears

H1 -> There is significant different between Attrition and TotalWorkingYears

stats,p = pearsonr(df.Attrition,df.TotalWorkingYears)

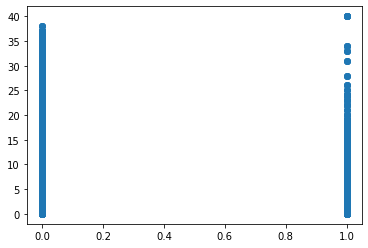
print(stats,p)



If **P<0.05 we reject H0**

plt.scatter(df.Attrition,df.TotalWorkingYears)

plt.show()



**# Correlation of Attrition and YearsAtCompany**

H0 -> There is no significant different between Attrition and YearsAtCompany

H1 -> There is significant different between Attrition and YearsAtCompany

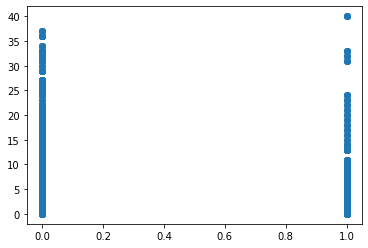
stats,p = pearsonr(df.Attrition,df.YearsAtCompany)

print(stats,p)



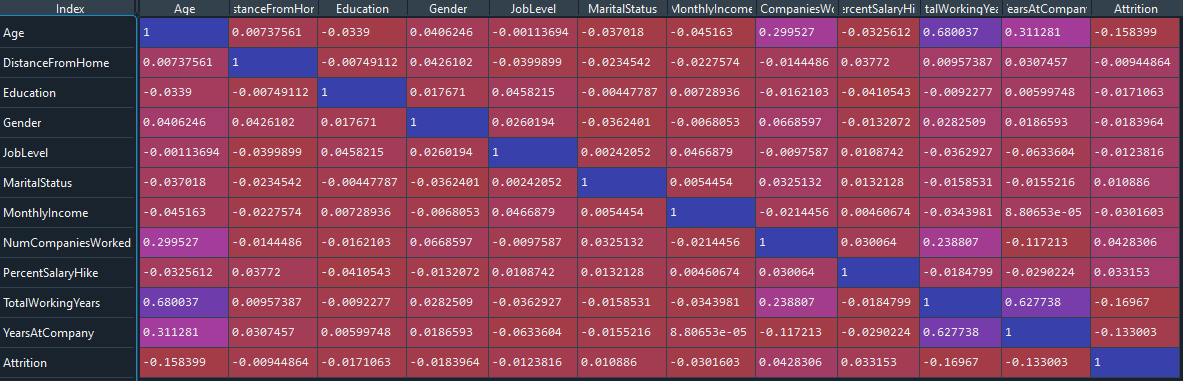
plt.scatter(df.Attrition,df.YearsAtCompany)

plt.show()



**# Correlation Matrix and Visualization**

corr\_mat = df[['Age','DistanceFromHome','Education','Gender','JobLevel','MaritalStatus', 'MonthlyIncome','NumCompaniesWorked','PercentSalaryHike','TotalWorkingYears', 'YearsAtCompany','Attrition']].copy()

corr = corr\_mat.corr() 

import seaborn as sns

plt.figure(figsize=(12,10),dpi=100)

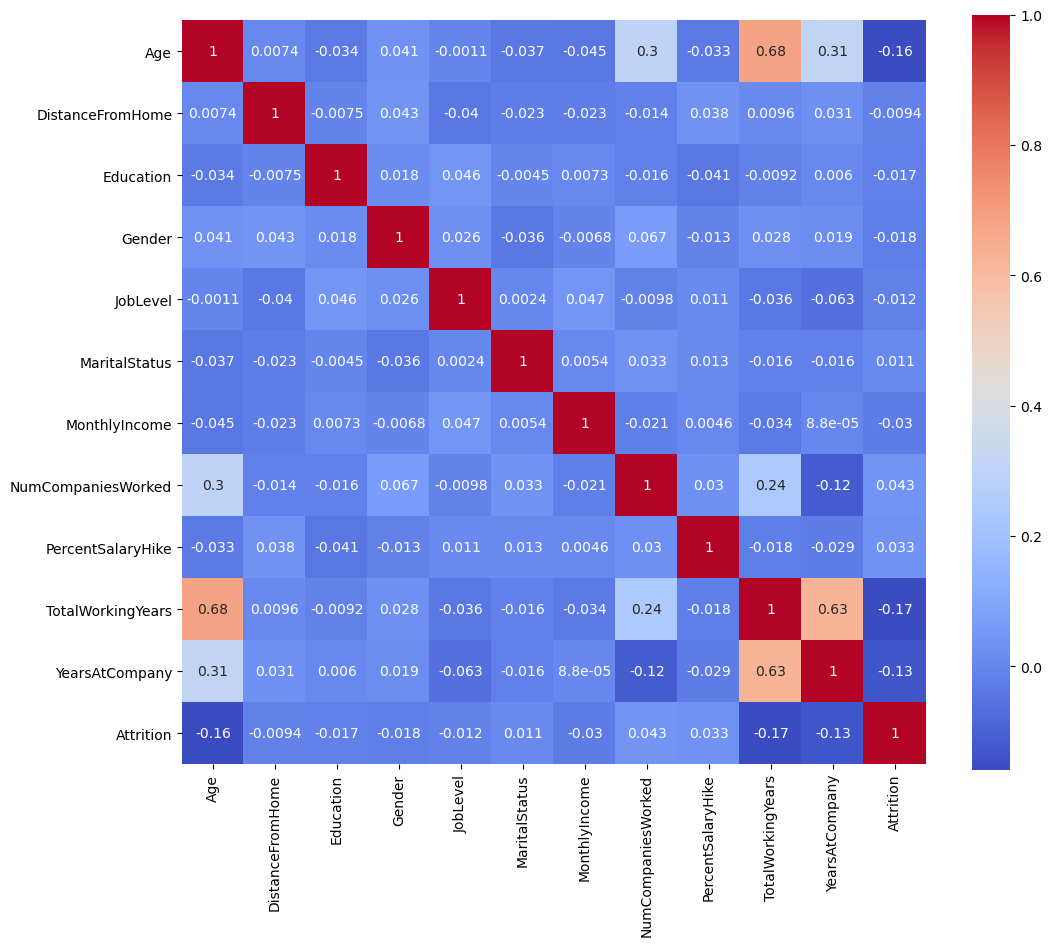
sns.heatmap(corr\_mat.corr(),

cmap='coolwarm',

annot=True,

square=True,

robust=True);



top\_corr = corr\_mat.nlargest(5,"Attrition")

“top 5 correlated values “

