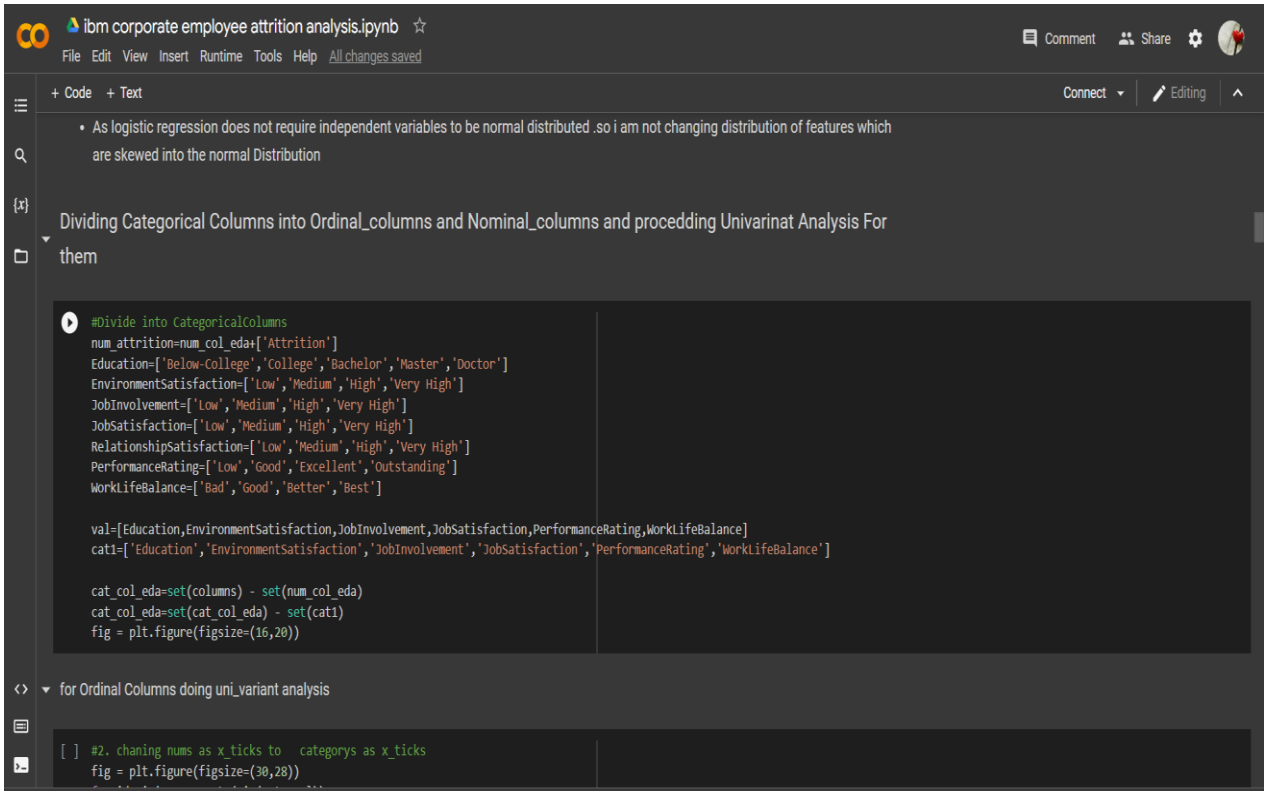
Project Development

Delivery Of Sprint-2

Date	October 2022
Team ID	PNT2022TMID37447
Project Name	Project - Corporate Employee Attrition Analytics

DATA UNDERSTANDING, DATA PREPARATION & EDA

DIVIDING CATEGORICAL COLUMNS INTO ORDINAL_COLUMNS AND NOMINAL_COLUMNS AND PROCEDDING UNIVARIANT ANALYSIS FOR THEM



CODING:

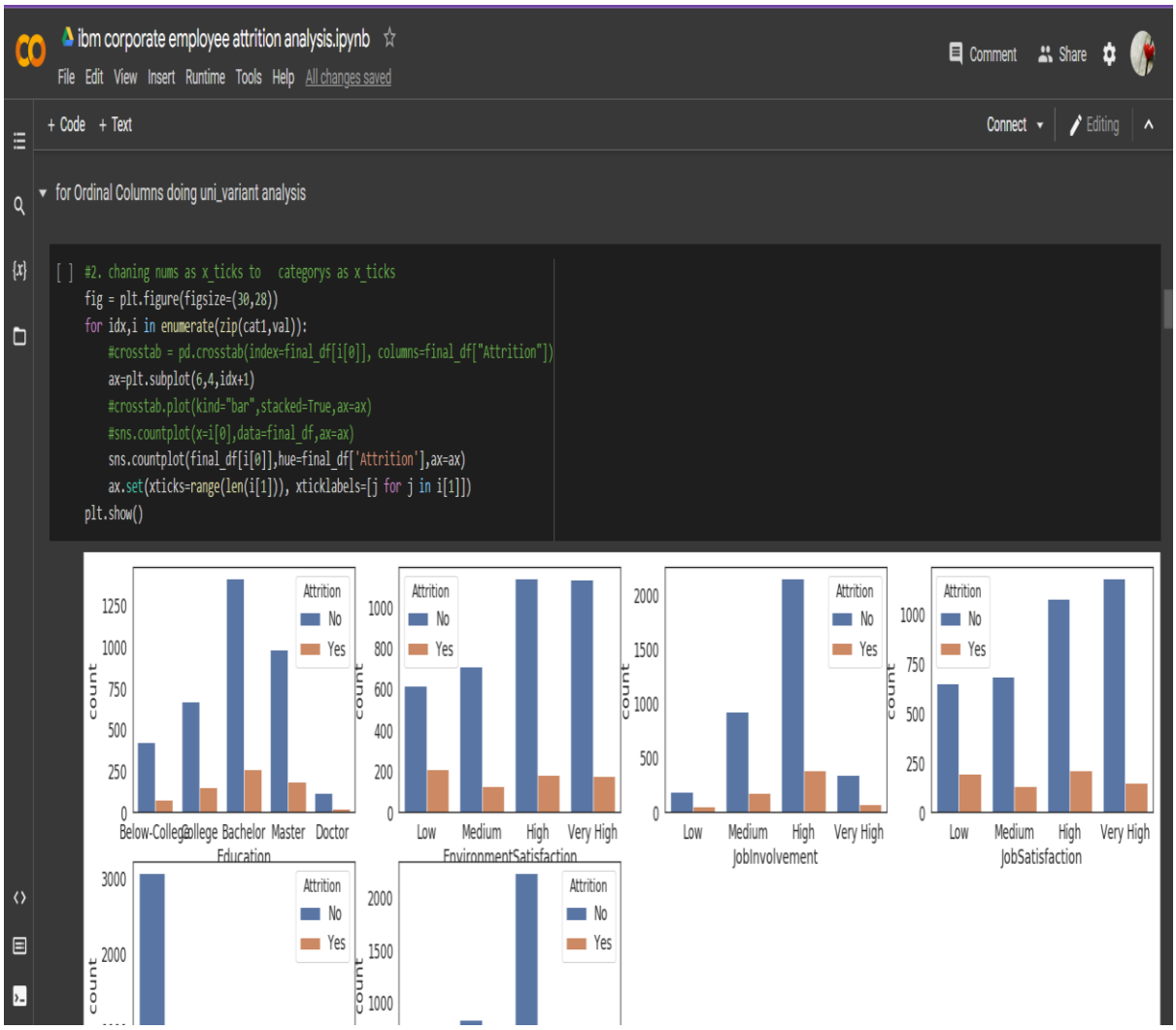
```
#Divide into CategoricalColumns
num_attrition=num_col_eda+['Attrition']
Education=['Below-College','College','Bachelor','Master','Doctor']
EnvironmentSatisfaction=['Low','Medium','High','Very High']
JobInvolvement=['Low','Medium','High','Very High']
JobSatisfaction=['Low','Medium','High','Very High']
RelationshipSatisfaction=['Low','Medium','High','Very High']
PerformanceRating=['Low','Good','Excellent','Outstanding']
WorkLifeBalance=['Bad','Good','Better','Best']

val=[Education,EnvironmentSatisfaction,JobInvolvement,JobSatisfaction,PerformanceRating,WorkLifeBalance]
cat1=['Education','EnvironmentSatisfaction','JobInvolvement','JobSatisfaction','PerformanceRating','WorkLifeBalance']

cat_col_eda=set(columns) - set(num_col_eda)
cat_col_eda=set(cat_col_eda) - set(cat1)
fig = plt.figure(figsize=(16,20))
```

FOR ORDINAL COLUMNS UNIVARIANT ANALYSIS

CODING:

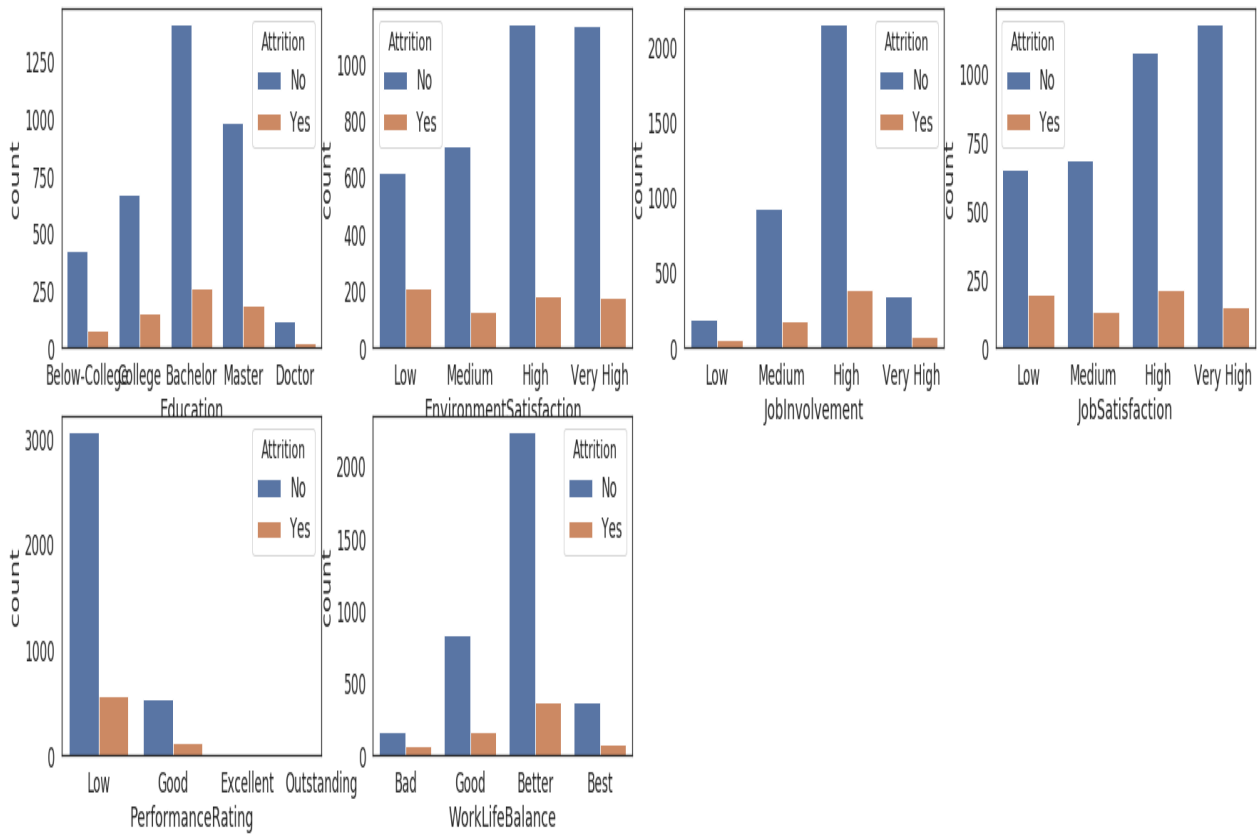
```
#2. chaning nums as x_ticks to   categorys as x_ticks
fig = plt.figure(figsize=(30,28))
for idx,i in enumerate(zip(cat1,val)):
```

```
    #crosstab = pd.crosstab(index=final_df[i[0]], columns=final_df["Attrition"])
    ax=plt.subplot(6,4,idx+1)
    #crosstab.plot(kind="bar",stacked=True,ax=ax)
    #sns.countplot(x=i[0],data=final_df,ax=ax)
    sns.countplot(final_df[i[0]],hue=final_df['Attrition'],ax=ax)
    ax.set(xticks=range(len(i[1])), xticklabels=[j for j in i[1]])
plt.show()
```




ibm corporate employee attrition analysis.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Connect
Editing
+ Code + Text
sns.countplot(final_df[i[0]],hue=final_df['Attrition'],ax=ax)
ax.set(xticks=range(len(i[1])), xticklabels=[j for j in i[1]])
plt.show()
Attrition
No
Yes
count
Below-College College Bachelor Master Doctor
Education
Low Medium High Very High
EnvironmentSatisfaction
Low Medium High Very High
JobInvolvement
Low Medium High Very High
JobSatisfaction
Low Good Excellent Outstanding
PerformanceRating
Bad Good Better Best
WorkLifeBalance
1. there is no paricular distinction

OUTPUT:

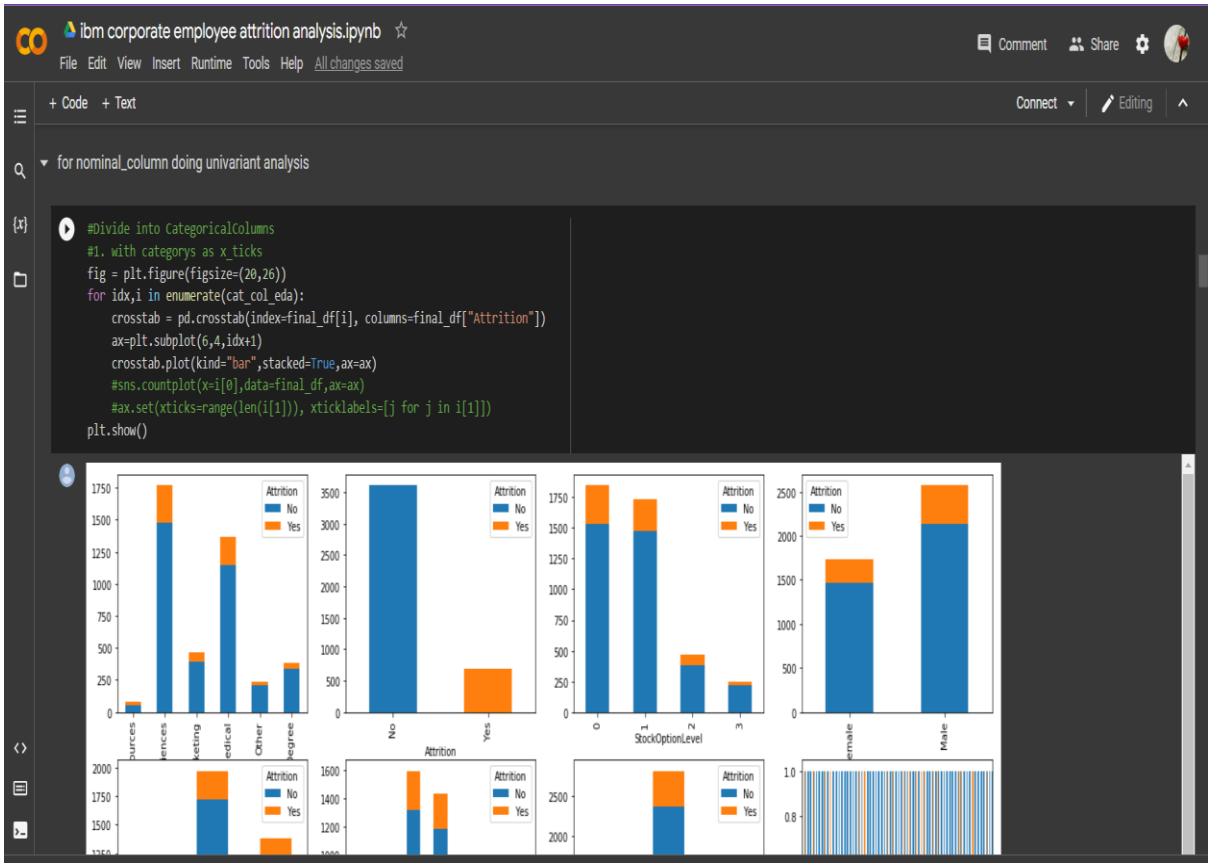

Attrition
No
Yes
count
Below-College College Bachelor Master Doctor
Education
Low Medium High Very High
EnvironmentSatisfaction
Low Medium High Very High
JobInvolvement
Low Medium High Very High
JobSatisfaction
Low Good Excellent Outstanding
PerformanceRating
Bad Good Better Best
WorkLifeBalance

INFERENCES:

- there is no particular distinction

FOR NOMINAL_COLUMN DOING UNIVARIANT ANALYSIS

CODING:



#Divide into CategoricalColumns

#1. with categorys as x_ticks

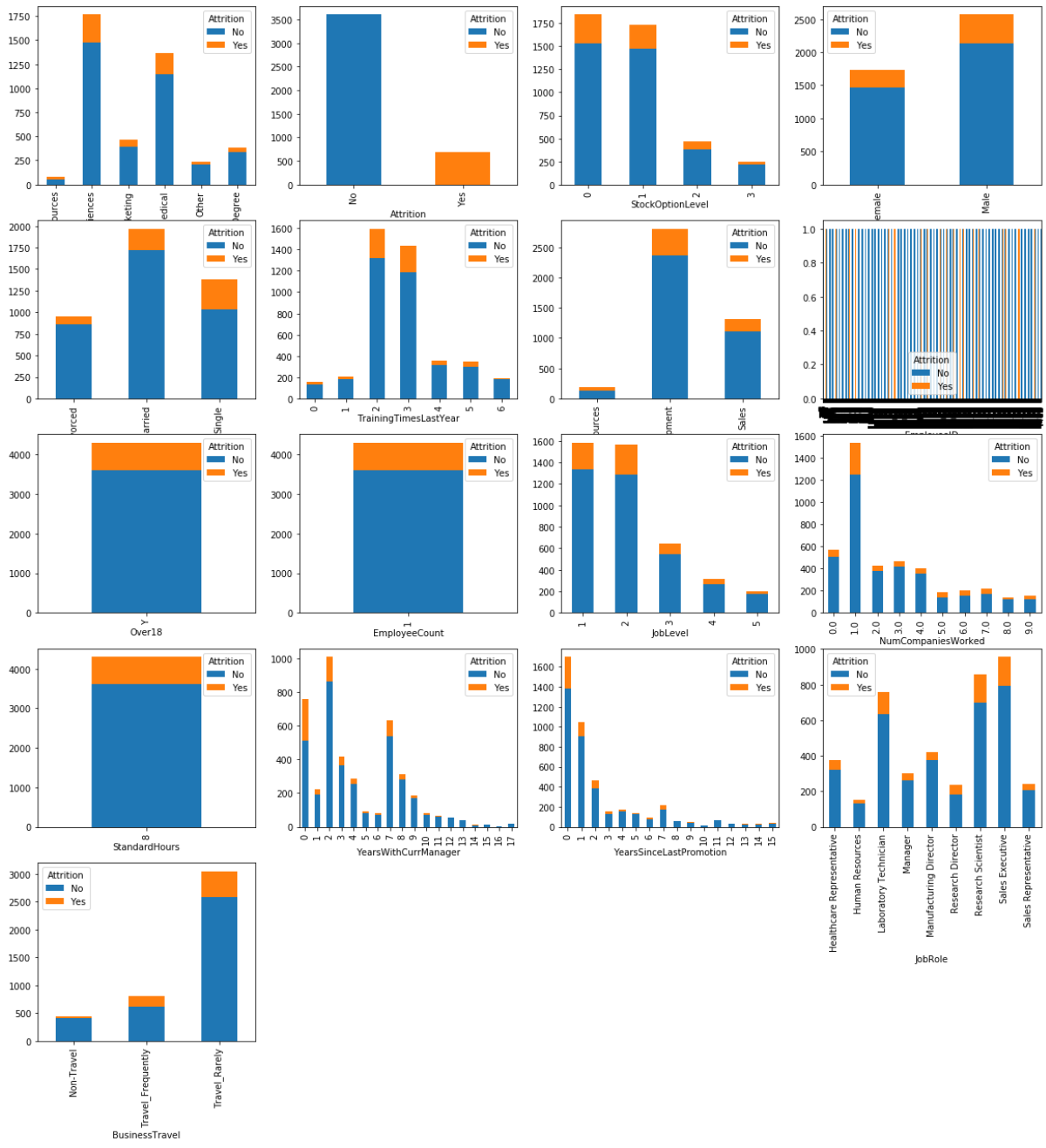
fig = plt.figure(figsize=(20,26))

for idx,i in enumerate(cat_col_eda):

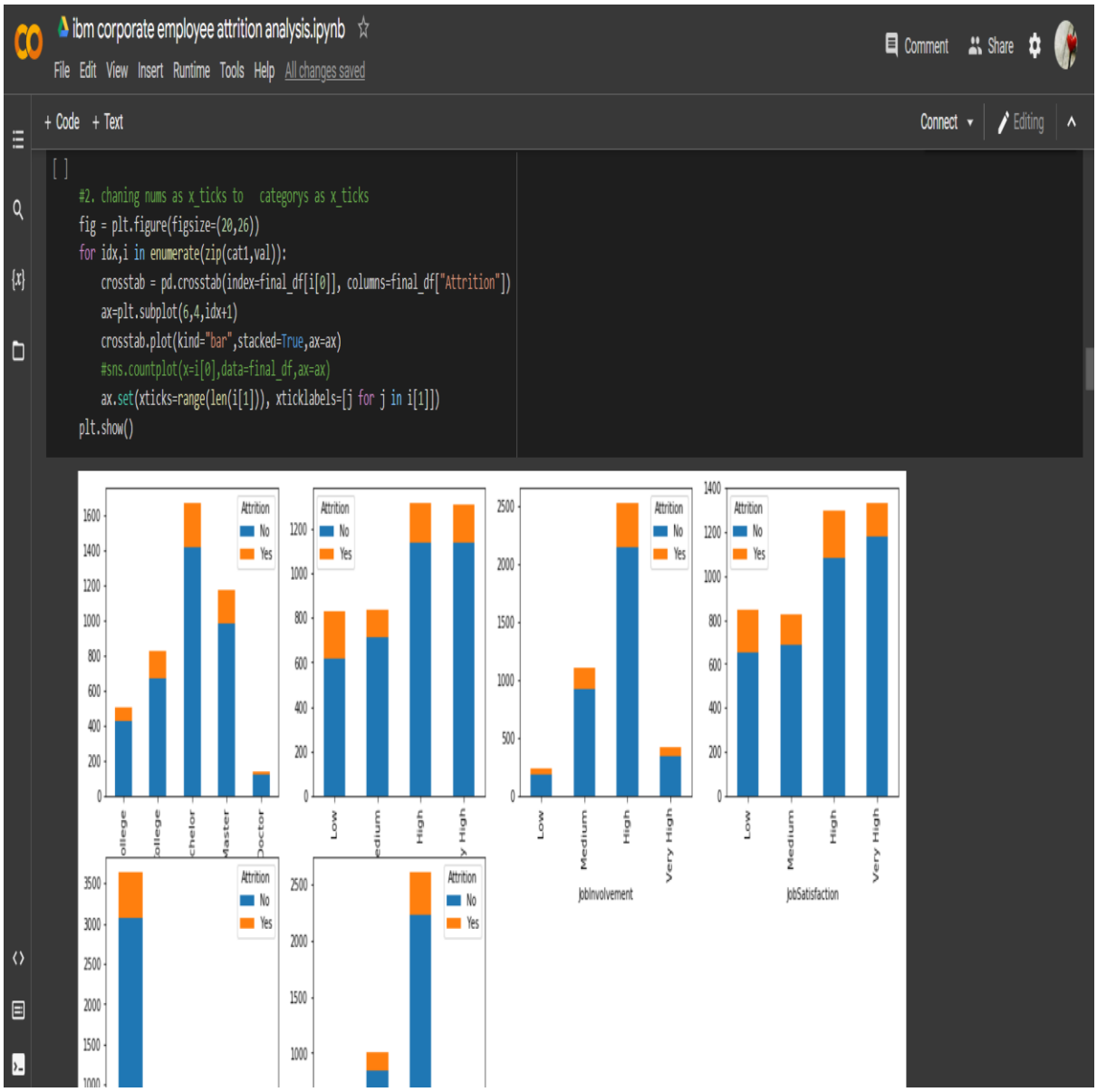
crosstab = pd.crosstab(index=final_df[i], columns=final_df["Attrition"])

```
    ax=plt.subplot(6,4,idx+1)
    crosstab.plot(kind="bar",stacked=True,ax=ax)
    #sns.countplot(x=i[0],data=final_df,ax=ax)
    #ax.set(xticks=range(len(i[1])), xticklabels=[j for j in i[1]])
plt.show()
```

OUTPUT:



CODING:



#2. chaning nums as x_ticks to categorys as x_ticks

fig = plt.figure(figsize=(20,26))

for idx,i in enumerate(zip(cat1,val)):

crosstab = pd.crosstab(index=final_df[i[0]], columns=final_df["Attrition"])

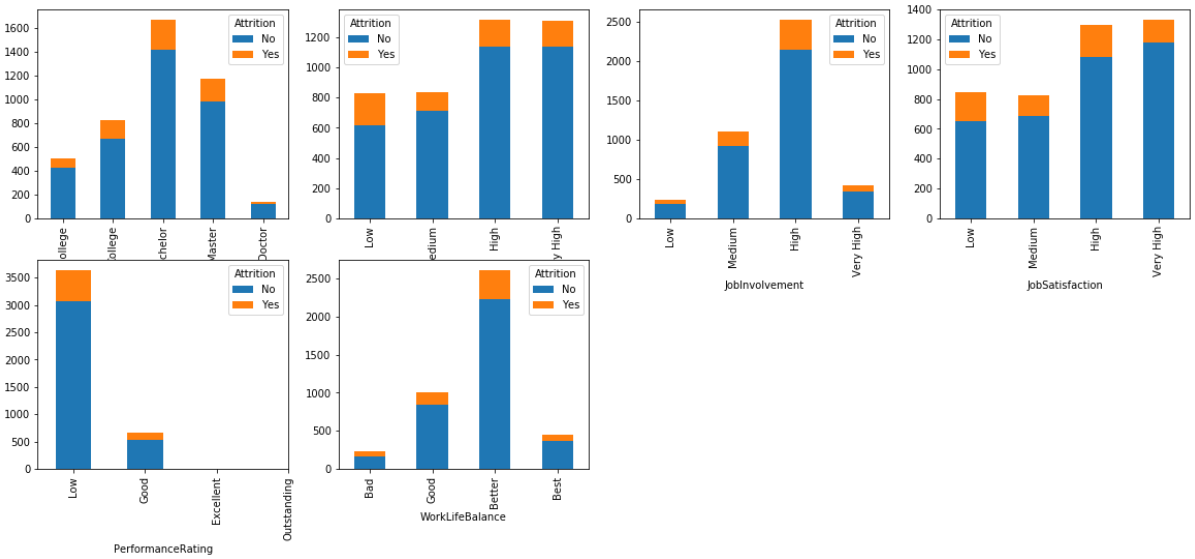
ax=plt.subplot(6,4,idx+1)

crosstab.plot(kind="bar",stacked=True,ax=ax)

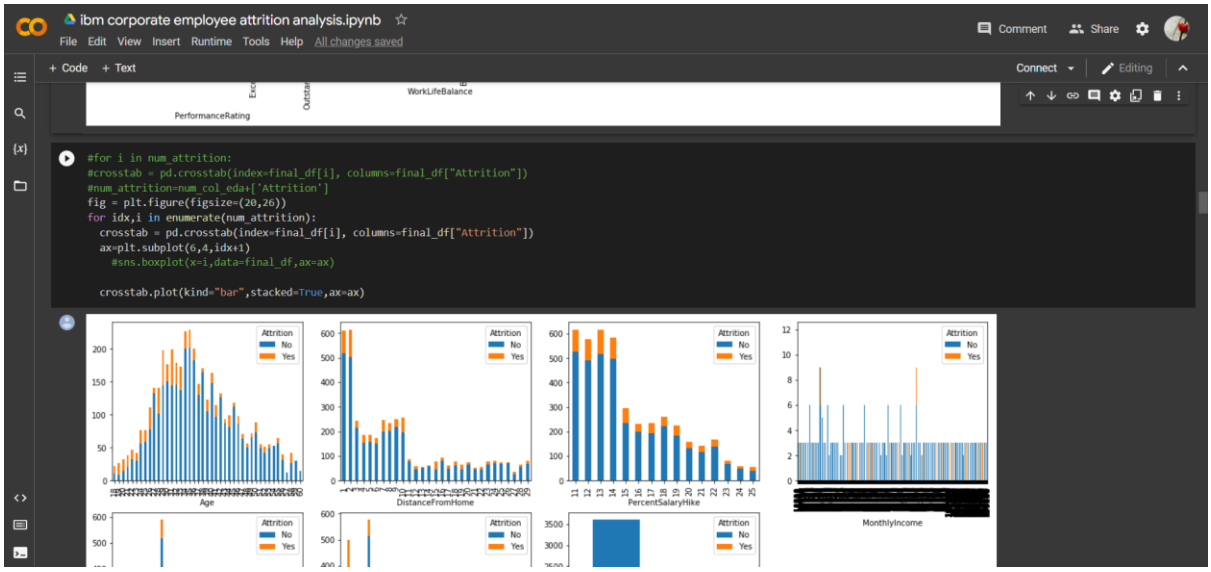
#sns.countplot(x=i[0],data=final_df,ax=ax)

```
    ax.set(xticks=range(len(i[1])), xticklabels=[j for j in i[1]])
plt.show()
```

OUTPUT:

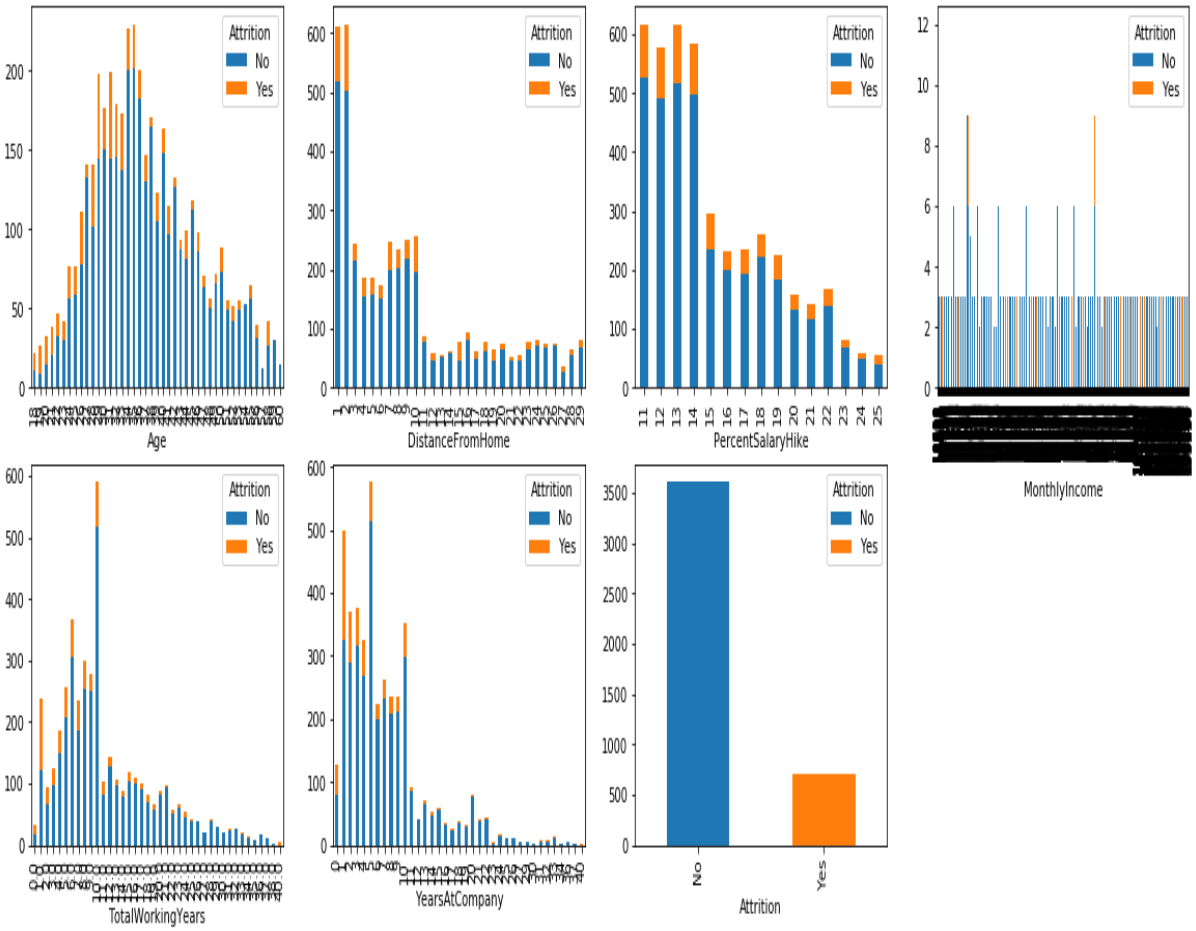


CODING:

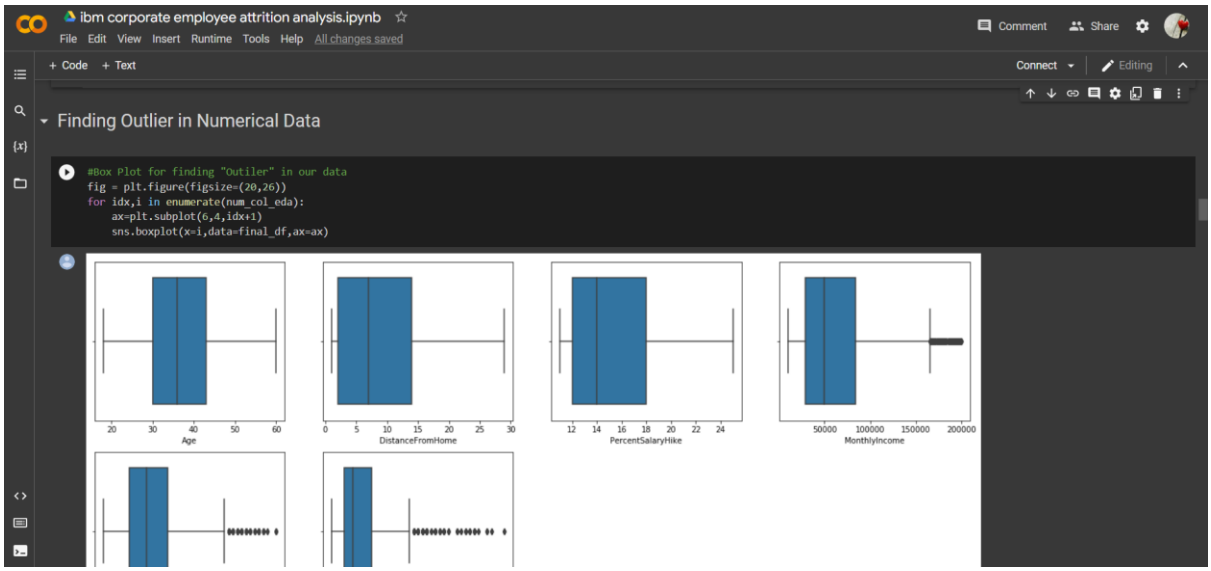


```
#for i in num_attrition:
#crosstab = pd.crosstab(index=final_df[i], columns=final_df["Attrition"])
#num_attrition=num_col_eda+['Attrition']
fig = plt.figure(figsize=(20,26))
for idx,i in enumerate(num_attrition):
  crosstab = pd.crosstab(index=final_df[i], columns=final_df["Attrition"])
  ax=plt.subplot(6,4,idx+1)
    #sns.boxplot(x=i,data=final_df,ax=ax)
  crosstab.plot(kind="bar",stacked=True,ax=ax)
```

OUTPUT:



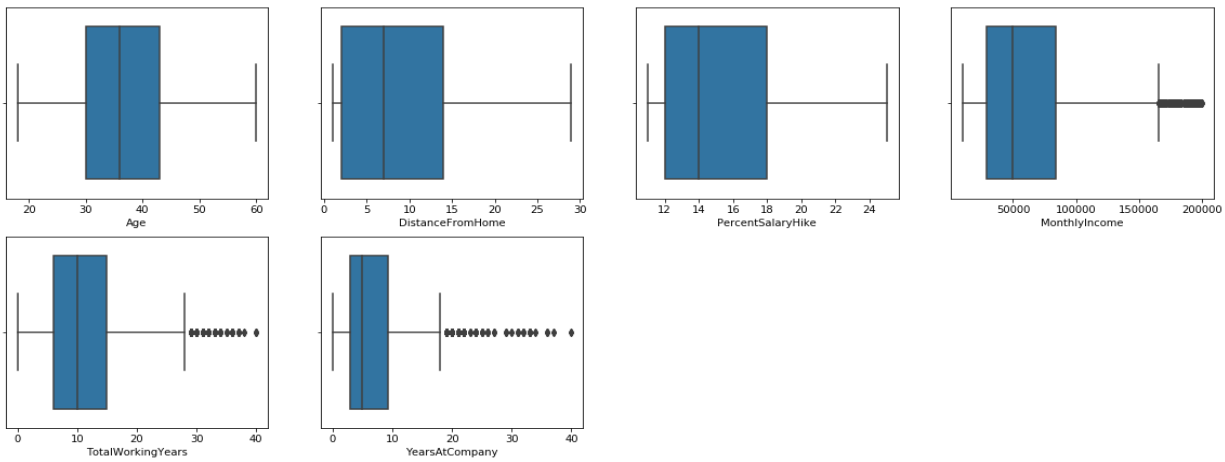
FINDING OUTLIER IN NUMERICAL DATA



CODING:

```
#Box Plot for finding "Outiler" in our data
fig = plt.figure(figsize=(20,26))
for idx,i in enumerate(num_col_eda):
    ax=plt.subplot(6,4,idx+1)
    sns.boxplot(x=i,data=final_df,ax=ax)
```

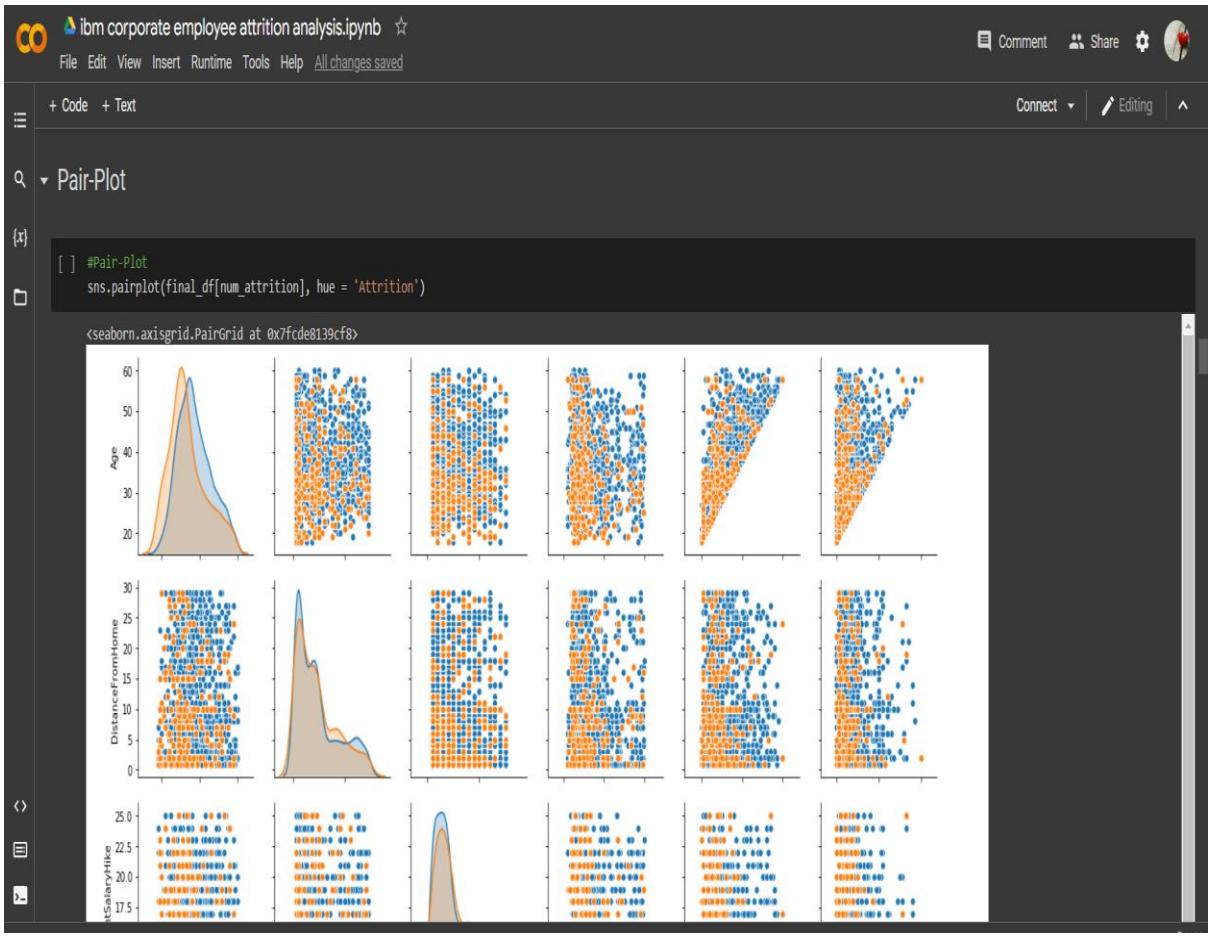
OUTPUT:



RESULTS FROM ABOVE GRAPH:

- from above Boxplots, we are trying to find is there any outliers in Numerical columns
- We can Observe outliers on Monthly Income, Total Working Years and Years at Company Columns that those columns don't outliers Because there is highly possibilities on occurring

PAIR-PLOT



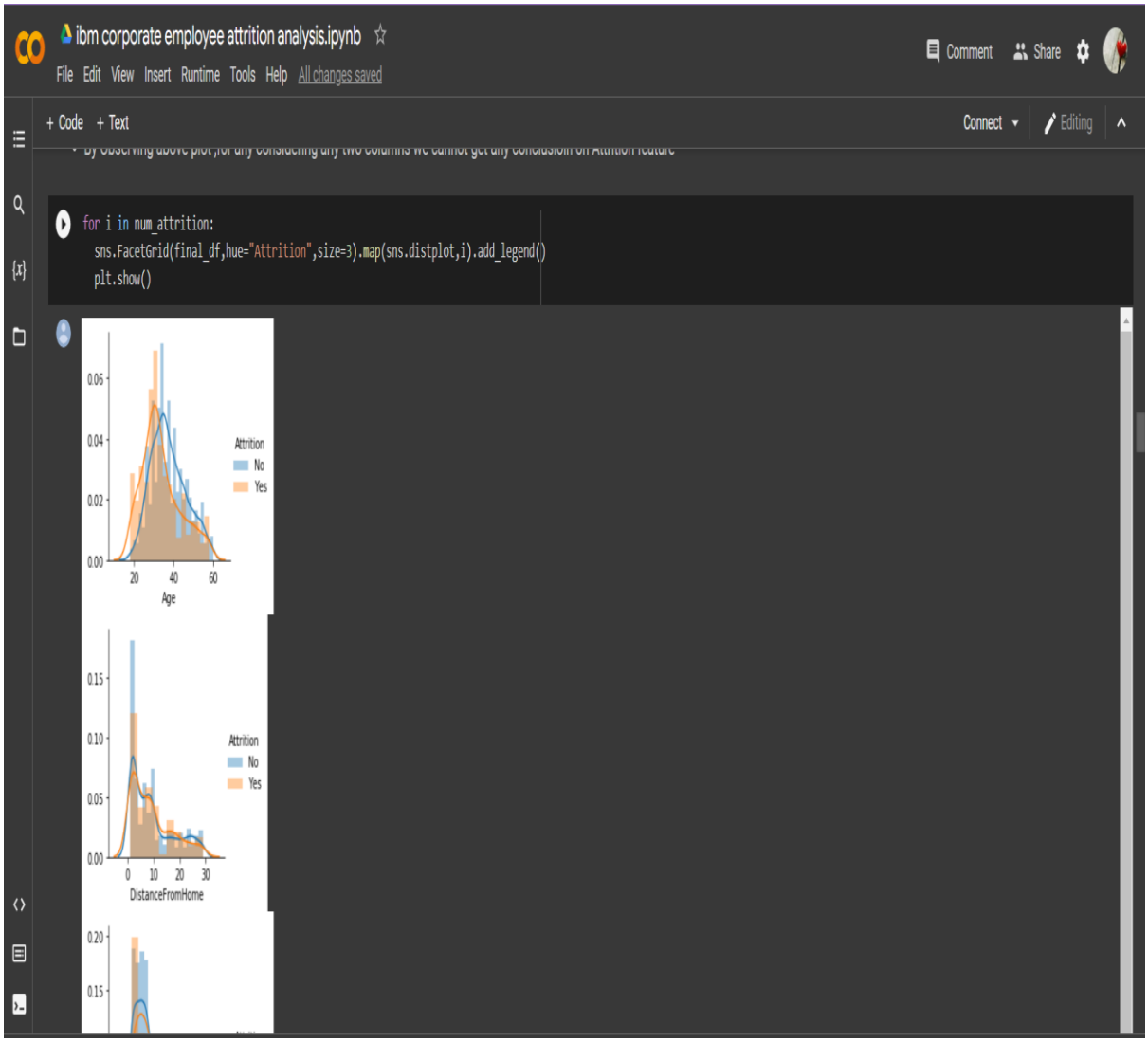
CODING:

```
#Pair-Plot

sns.pairplot(final_df[num_attrition], hue = 'Attrition')
```

OUTPUT:

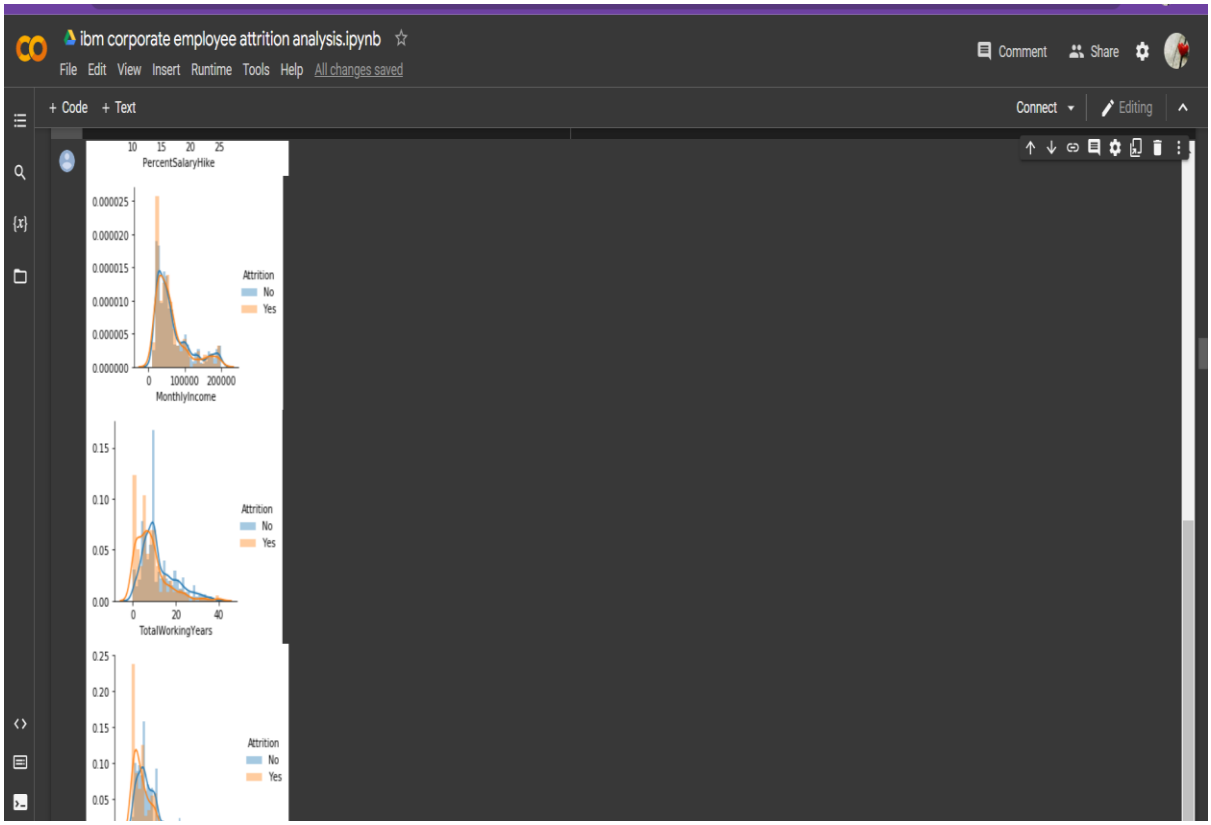
- By Observing above plot, for any considering any two columns we cannot get any conclusion on Attrition feature

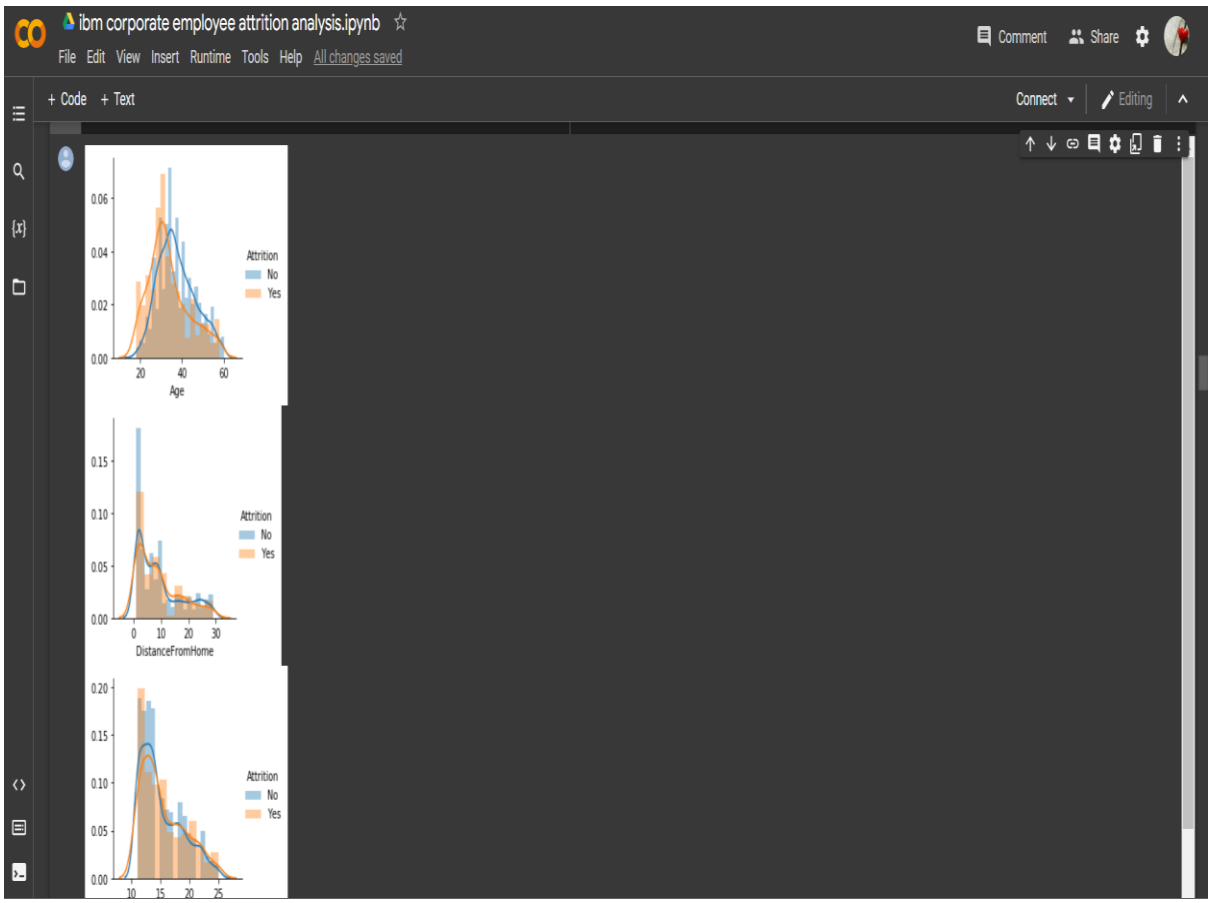
CODING:




```
for i in num_attrition:
  sns.FacetGrid(final_df,hue="Attrition",size=3).map(sns.distplot,i).add_legend()
  plt.show()
```

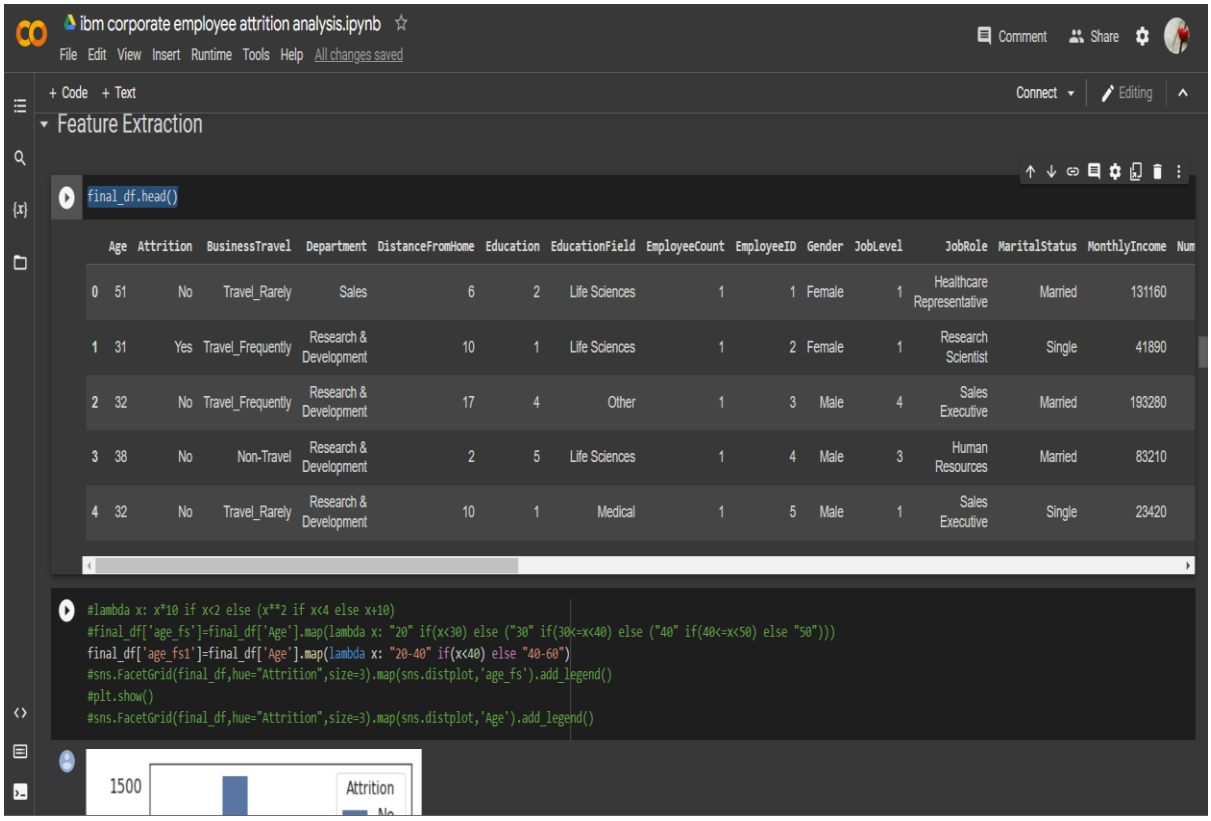
OUTPUT:

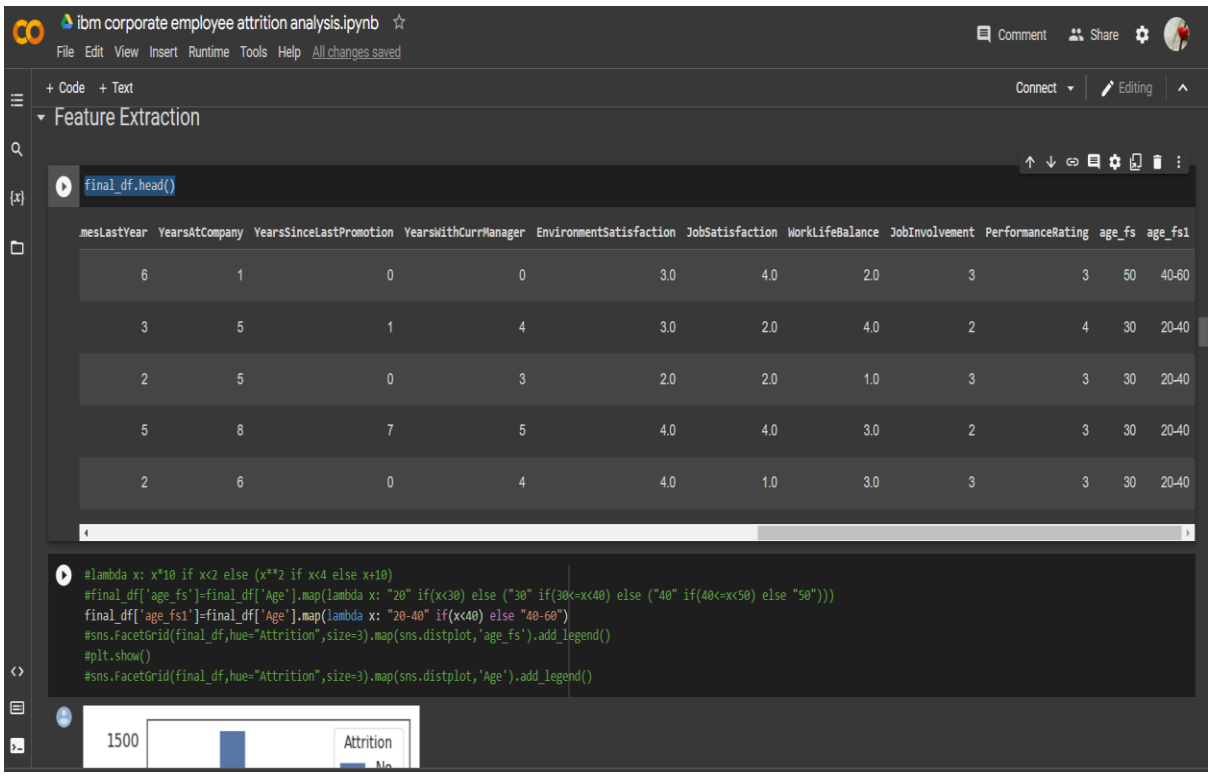

ibm corporate employee attrition analysis.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
Connect
Editing
PercentSalaryHike
MonthlyIncome
TotalWorkingYears
Attrition
No
Yes


ibm corporate employee attrition analysis.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
Connect
Editing
Age
DistanceFromHome
Attrition
No
Yes

FEATURE EXTRACTION

CODING:



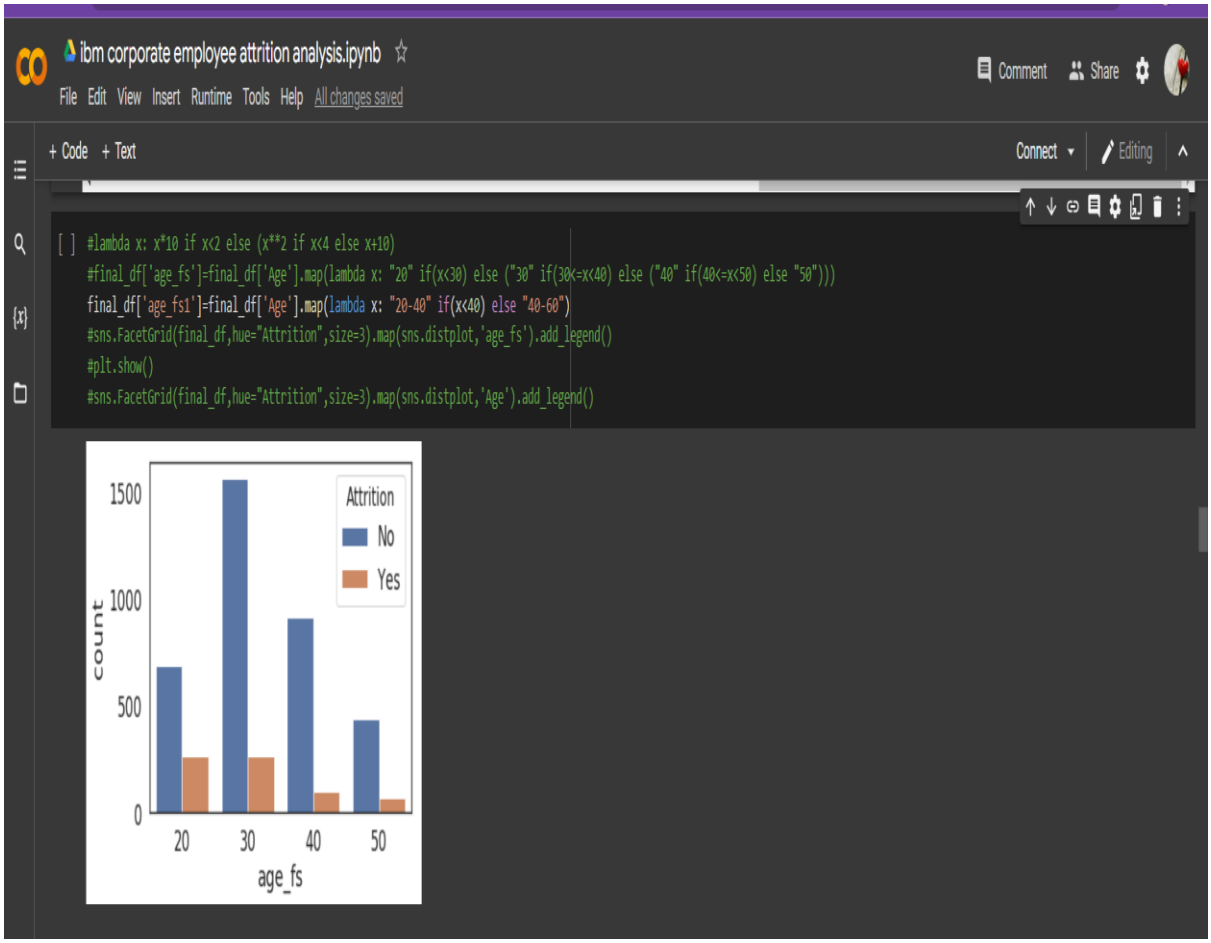


final_df.head()

OUTPUT:

Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	JobLevel	JobRole	MaritalStatus	MonthlyIncome	NumCompaniesWorked	Over18	PercentSalaryHike	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager	EnvironmentSatisfaction	JobSatisfaction	WorkLifeBalance	JobInvolvement	PerformanceRating	age_fs	age_fs1
0	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	1	Healthcare Representative	Married	131160	1.0	Y	11	8	0	1.0	6	1	0	0	3.0	4.0	2.0	3	3	50	40-60
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	2	Female	1	Research Scientist	Single	41890	0.0	Y	23	8	1	6.0	3	5	1	4	3.0	2.0	4.0	2	4	30	20-40
2	32	No	Travel_Frequently	Research & Development	17	4	Other	1	3	Male	4	Sales Executive	Married	193280	1.0	Y	15	8	3	5.0	2	5	0	3	2.0	2.0	1.0	3	3	30	20-40
3	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	4	Male	3	Human Resources	Married	83210	3.0	Y	11	8	3	13.0	5	8	7	5	4.0	4.0	3.0	2	3	30	20-40
4	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	5	Male	1	Sales Executive	Single	23420	4.0	Y	12	8	2	9.0	2	6	0	4	4.0	1.0	3.0	3	3	30	20-40

CODING:



#lambda x: x*10 if x<2 else (x**2 if x<4 else x+10)

#final_df['age_fs']=final_df['Age'].map(lambda x: "20" if(x<30) else ("30" if(30<=x<40) else ("40" if(40<=x<50) else "50")))

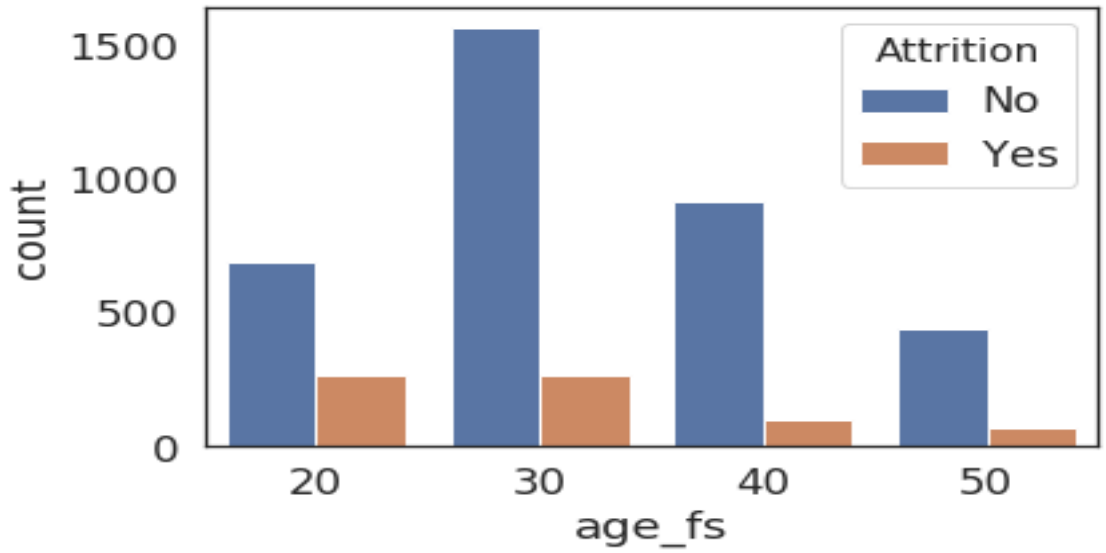
final_df['age_fs1']=final_df['Age'].map(lambda x: "20-40" if(x<40) else "40-60")

#sns.FacetGrid(final_df,hue="Attrition",size=3).map(sns.distplot,'age_fs').add_legend()

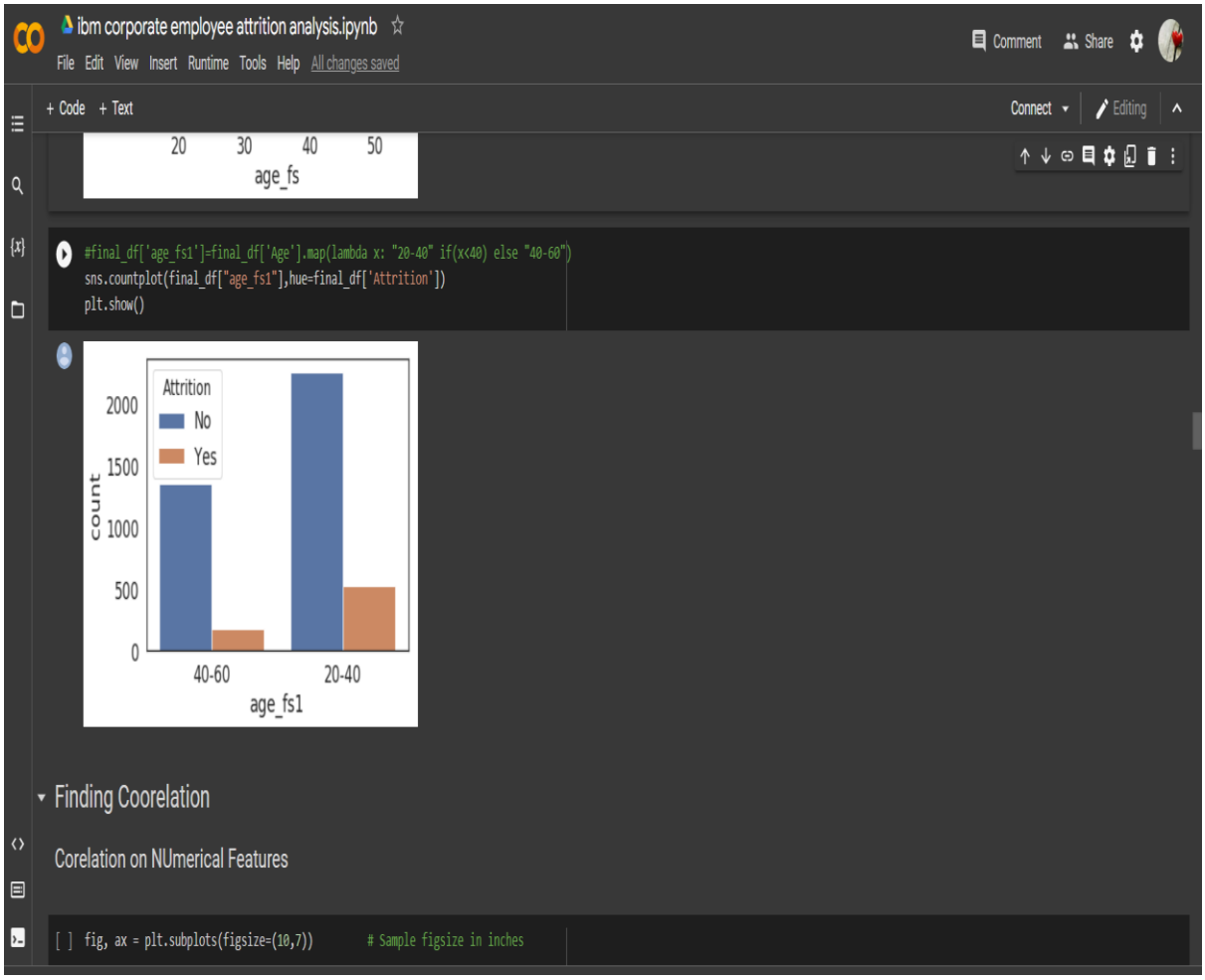
#plt.show()

#sns.FacetGrid(final_df,hue="Attrition",size=3).map(sns.distplot,'Age').add_legend()

OUTPUT:


1500
1000
500
0
count
20
30
40
50
age_fs
Attrition
No
Yes

CODING:

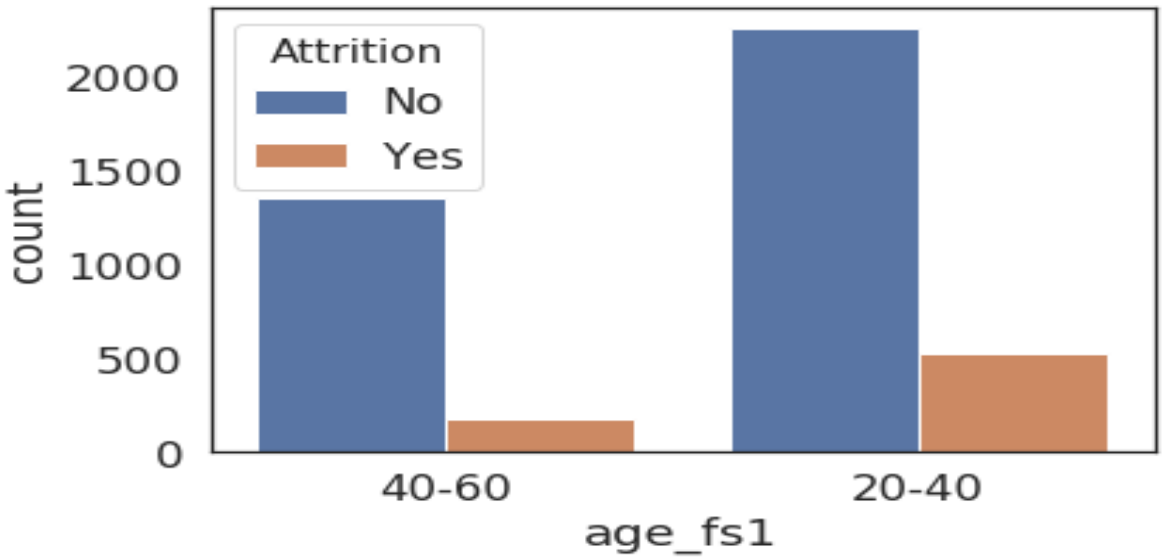

ibm corporate employee attrition analysis.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
Connect
Editing
20 30 40 50
age_fs
#final_df['age_fs1']=final_df['Age'].map(lambda x: "20-40" if(x<40) else "40-60")
sns.countplot(final_df["age_fs1"],hue=final_df['Attrition'])
plt.show()
Attrition
No
Yes
2000
1500
1000
500
0
count
40-60
20-40
age_fs1
Finding Coorelation
Corelation on NUmerical Features
[ ] fig, ax = plt.subplots(figsize=(10,7)) # Sample figsize in inches

#final_df['age_fs1']=final_df['Age'].map(lambda x: "20-40" if(x<40) else "40-60")

sns.countplot(final_df["age_fs1"],hue=final_df['Attrition'])

plt.show()

OUTPUT:



FINDING CORRELATION

CORRELATION ON NUMERICAL FEATURES

CODING:

fig, ax = plt.subplots(figsize=(10,7)) # Sample figsize in inches

sns.heatmap(final_df_numerical_data.corr(),cmap='YlOrRd',linewidths=.5,ax=ax)

ibm corporate employee attrition analysis.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

Connect Editing

40-60 20-40

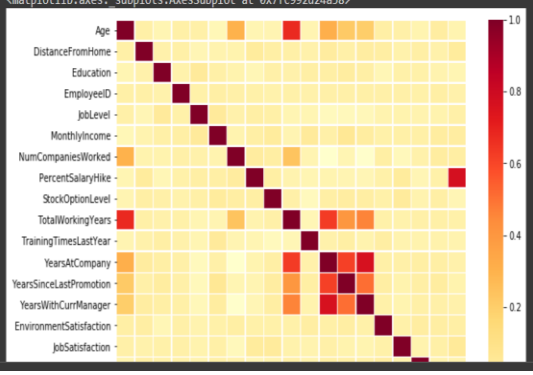
age_fs1

Finding Coorelation

Corelation on NUmerical Features

```
fig, ax = plt.subplots(figsize=(10,7))        # Sample figsize in inches

sns.heatmap(final_df_numerical_data.corr(),cmap='YlOrRd',linewidths=.5,ax=ax)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc992d24a58>
```



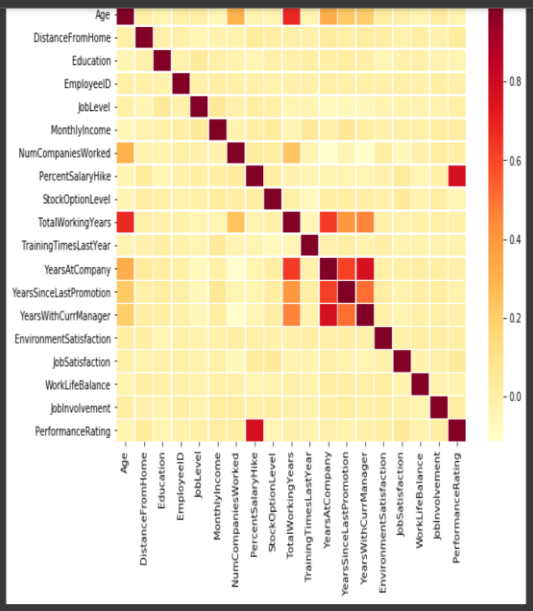
ibm corporate employee attrition analysis.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

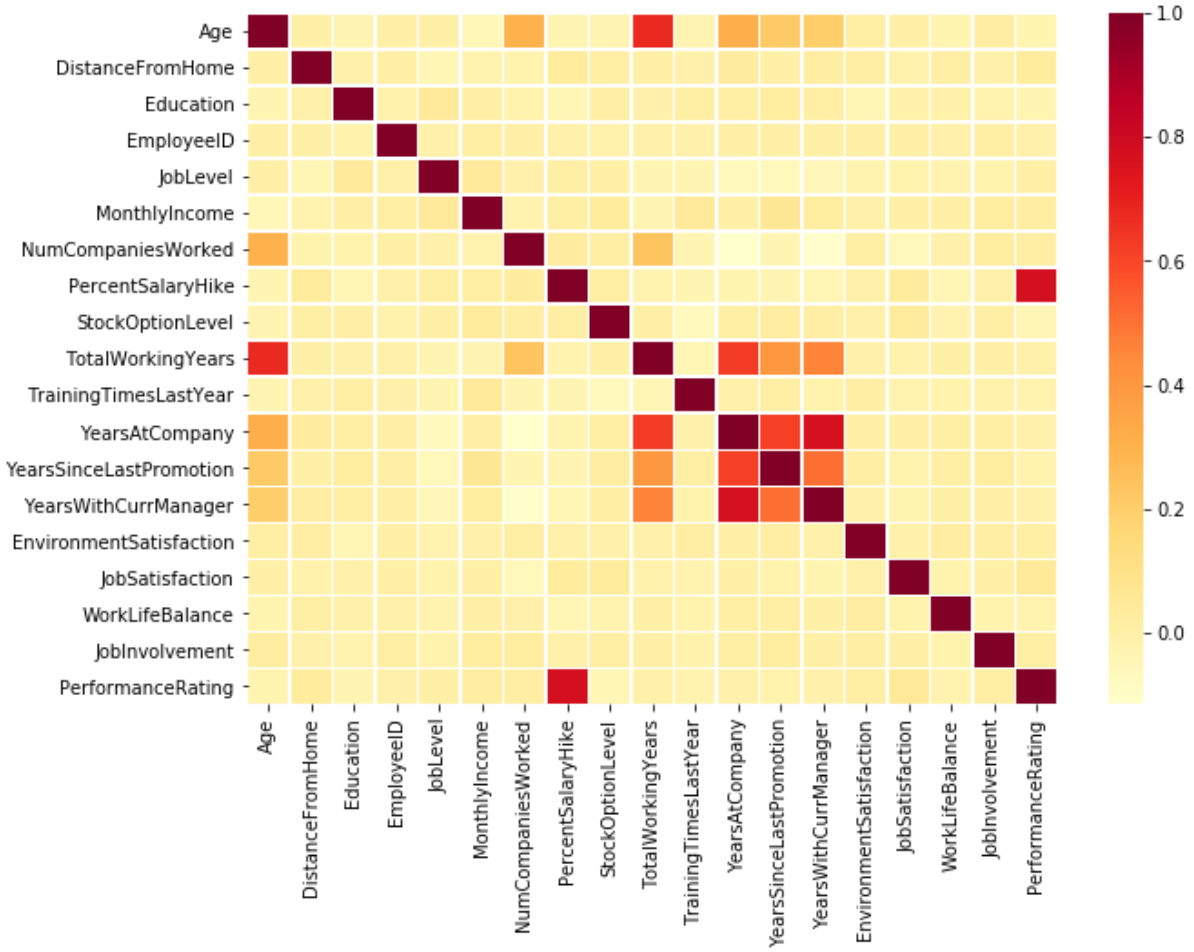
+ Code + Text

Connect Editing



- There is is no higher dependency between column to columns from above heatMap Plot

OUTPUT:



INFERENCES:

- There is no higher dependency between column to columns from above heat Map Plot