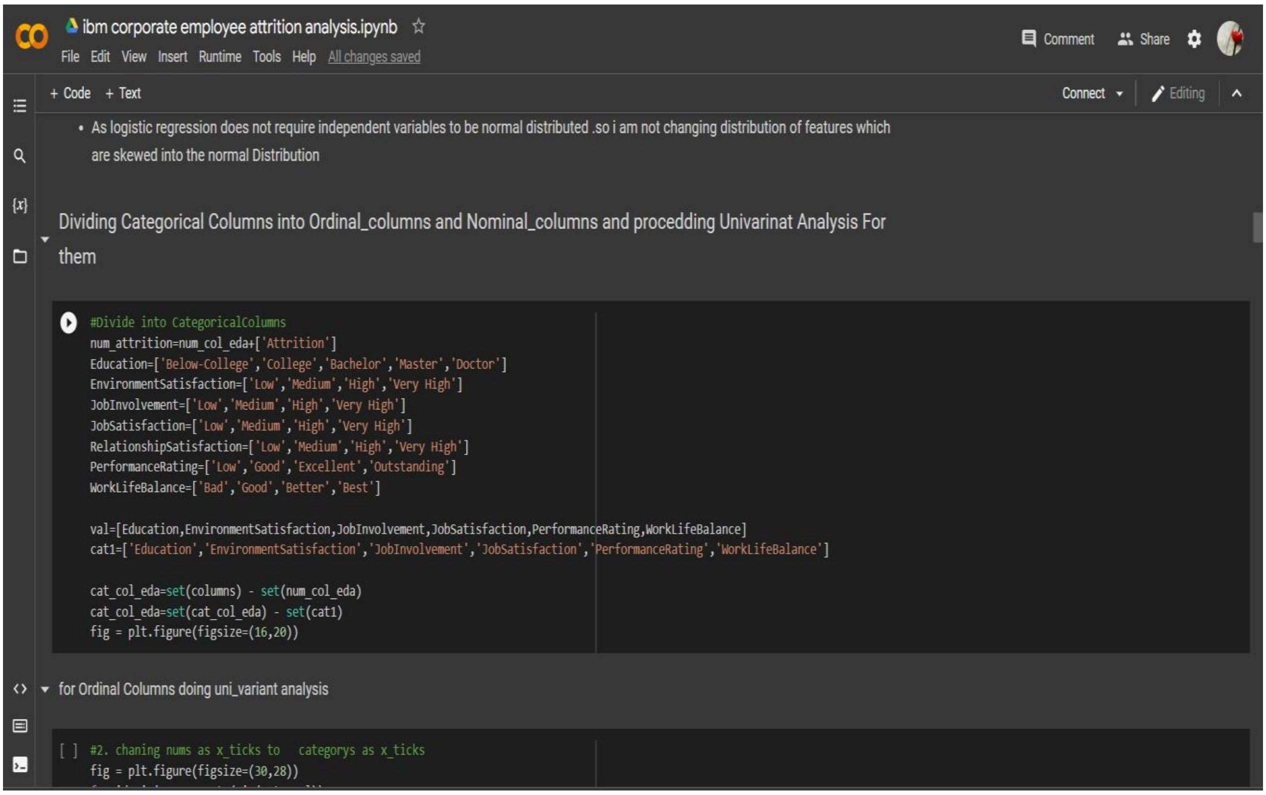
Project Development

Delivery Of Sprint-2

Date	October 2022
Team ID	PNT2022TMID06047
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


ibm corporate employee attrition analysis.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
Connect Editing

- As logistic regression does not require independent variables to be normal distributed .so i am not changing distribution of features which are skewed into the normal Distribution

Dividing Categorical Columns into Ordinal_columns and Nominal_columns and procedding Univarinat Analysis For them

```
#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 doing uni_variant analysis

```
[ ] #2. chaning nums as x_ticks to   categorys as x_ticks
    fig = plt.figure(figsize=(30,28))
```

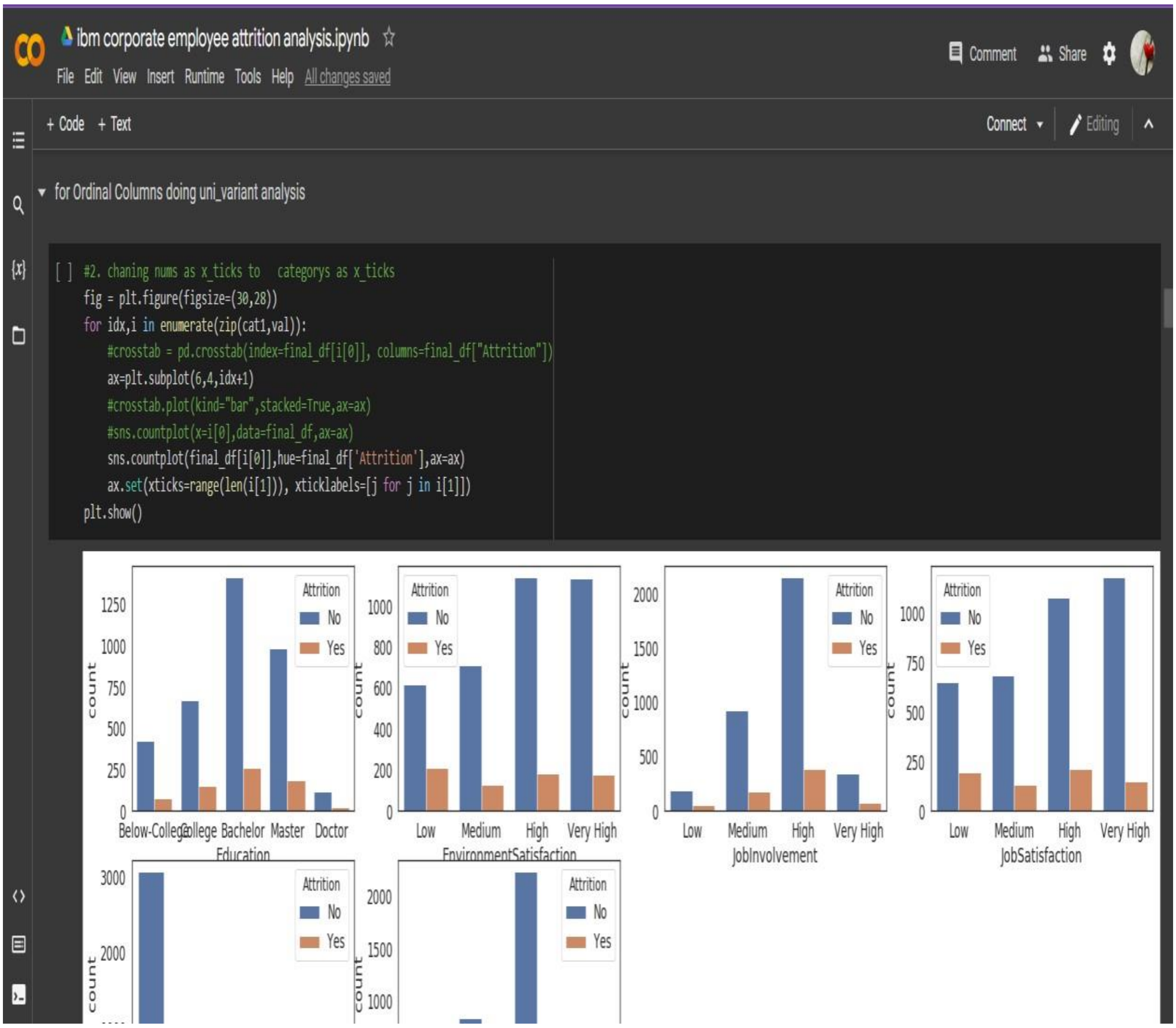
CODING:

```
#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

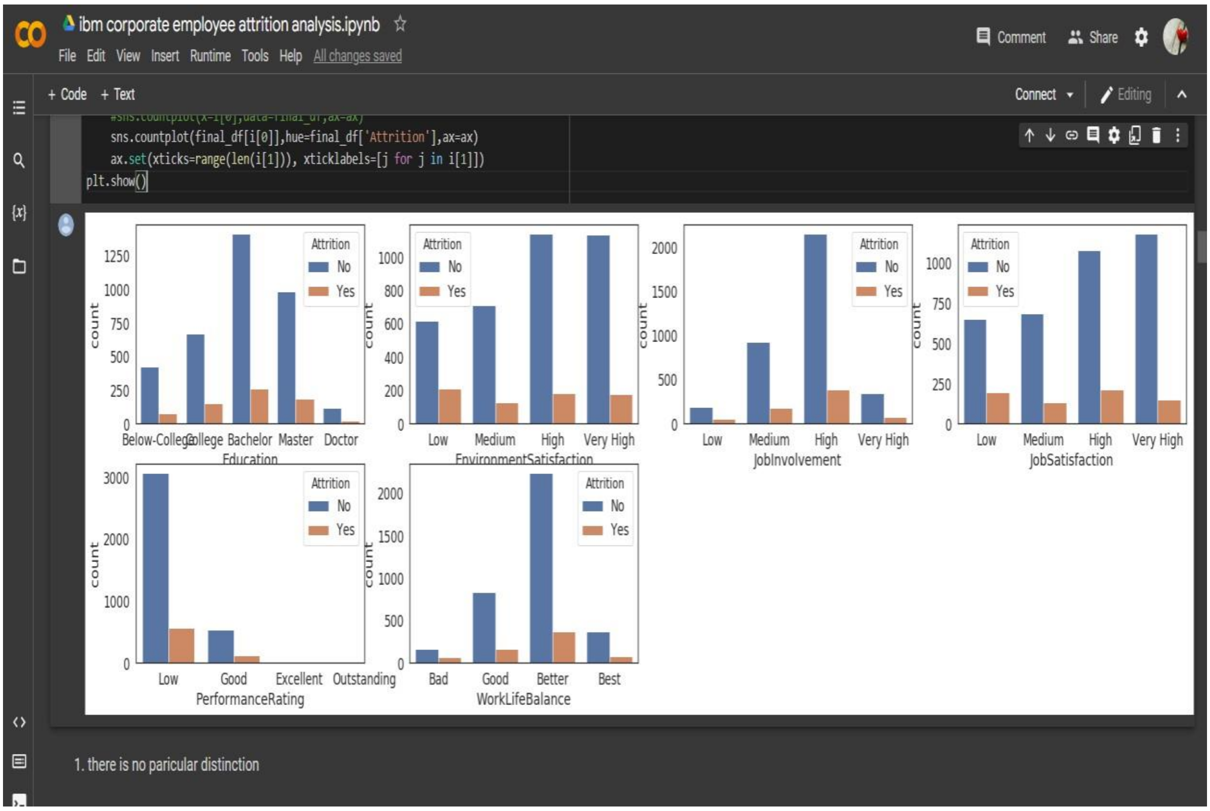
CODING:

```
#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()
```



for Ordinal Columns doing uni_variant analysis

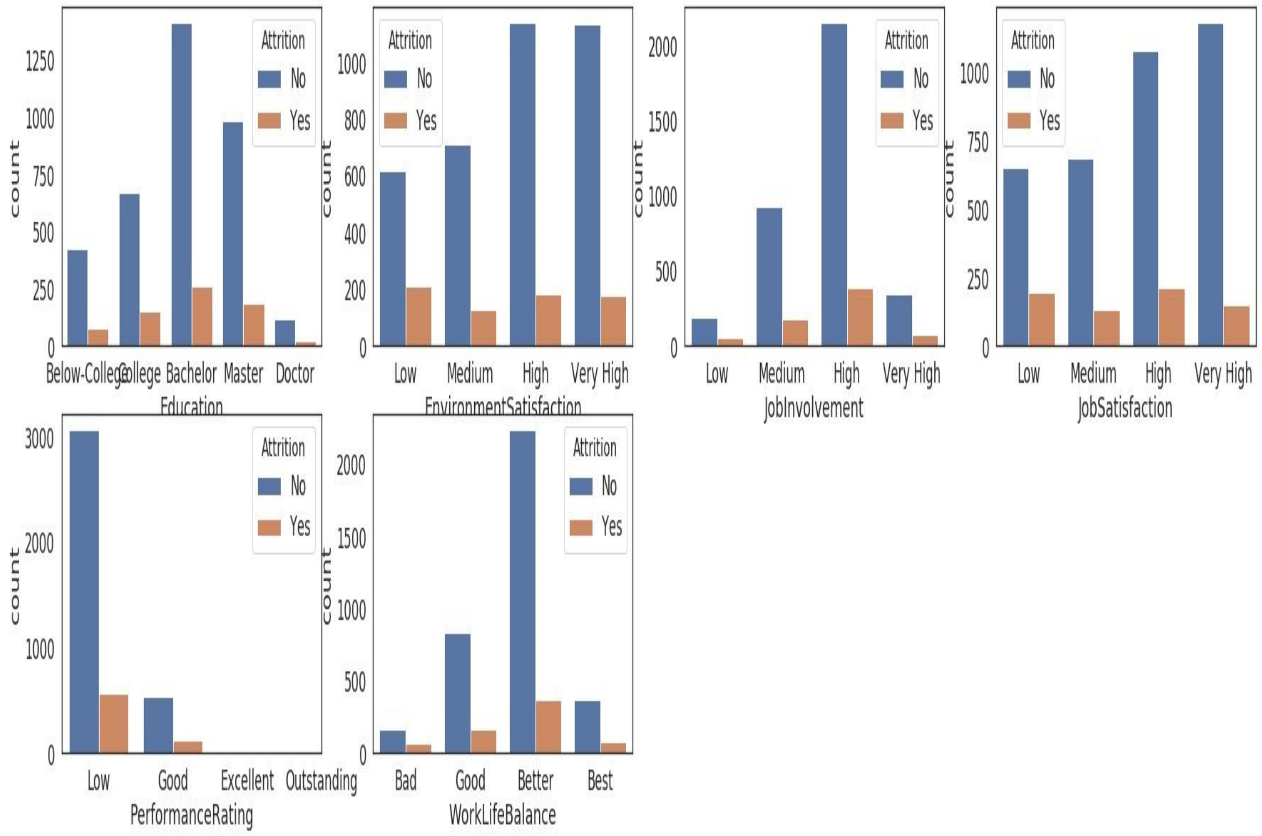
```
[ ] #2. chaning nums as x_ticks to   categorys as x_ticks
    fig = plt.figure(figsize=(30,28))
    for idx,i in enumerate(zip(cat1,val)):
        #crosstab = pd.crosstab(index=final_df[i[0]], columns=final_df["Attrition"])
        ax=plt.subplot(6,4,idx+1)
        #crosstab.plot(kind="bar",stacked=True,ax=ax)
        #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
+ Code + Text
Connect Editing

```
    #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()
```

1. there is no paricular distinction

OUTPUT:

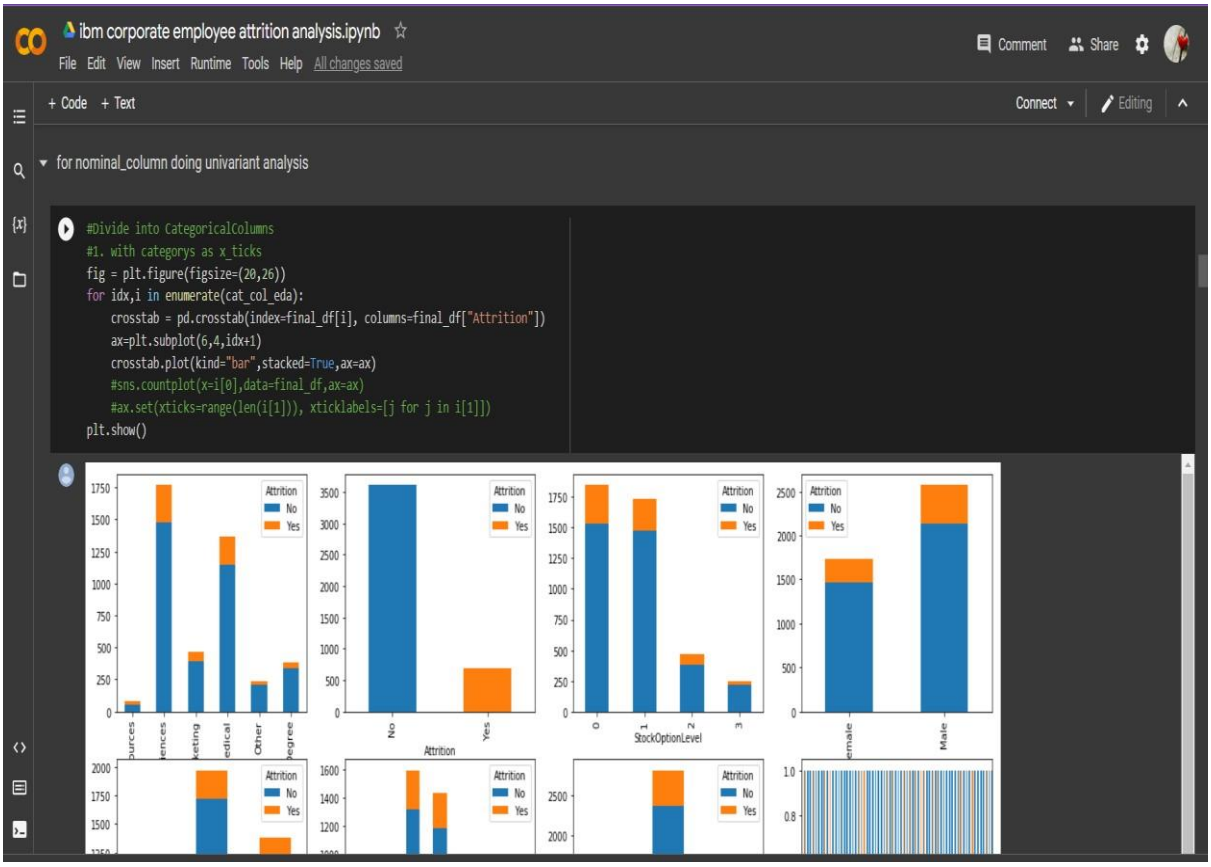


INFERENCES:

- there is no particular distinction

FOR NOMINAL_COLUMN DOING UNIVARIANT ANALYSIS

CODING:

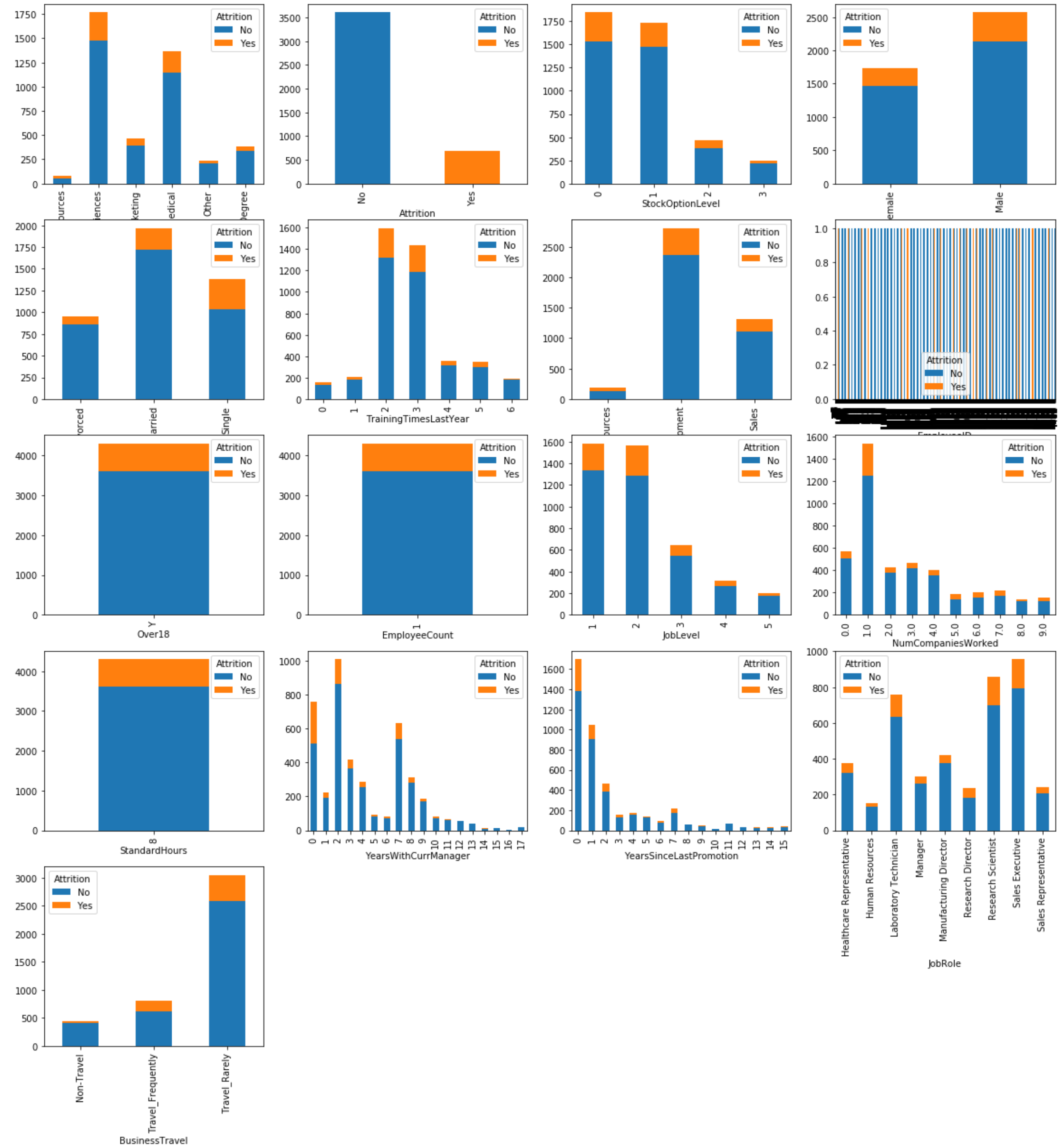

ibm corporate employee attrition analysis.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
Connect
Editing
for nominal_column doing univariant analysis
Attrition
No
Yes
StockOptionLevel

```
#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()
```

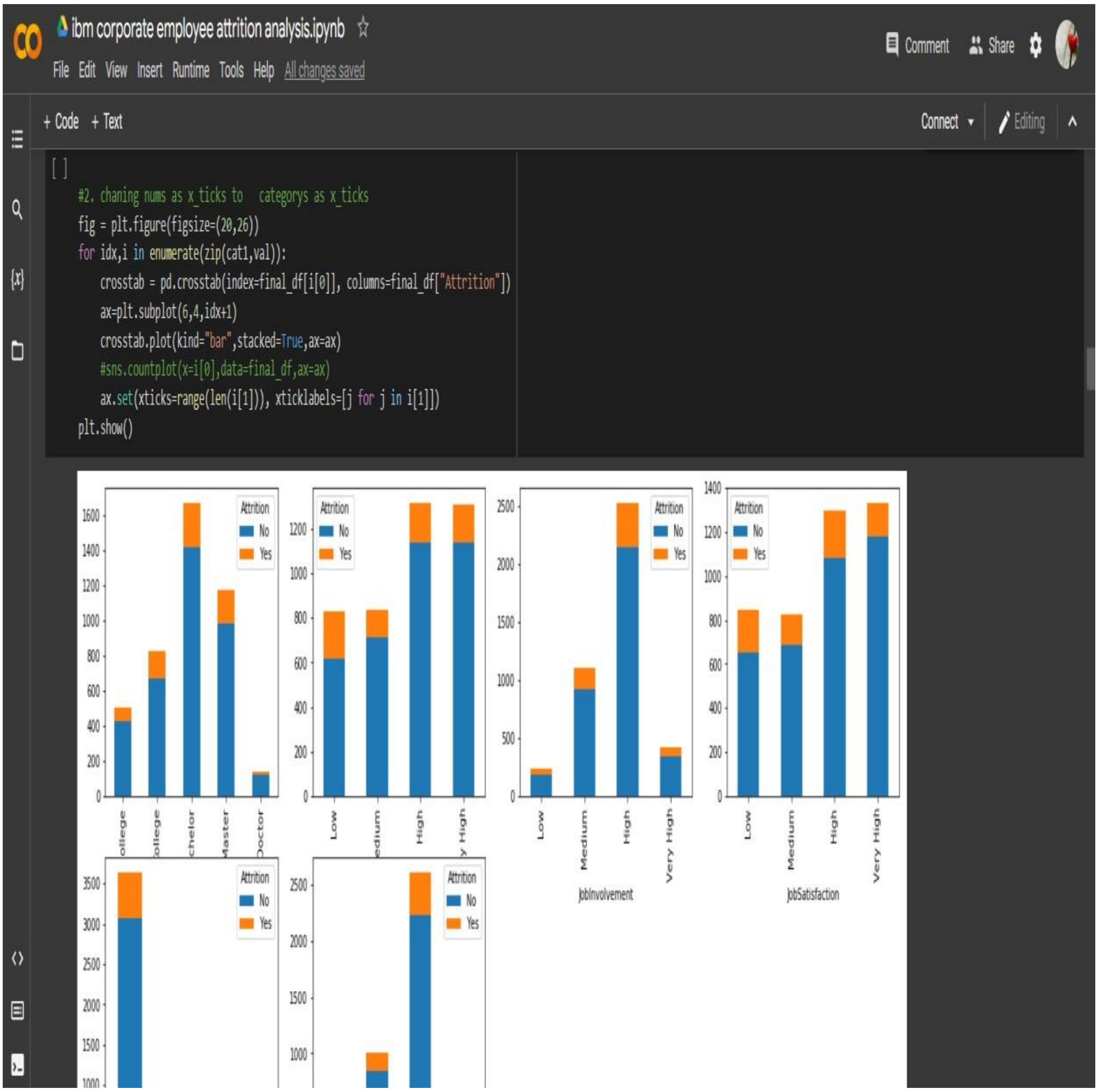
```
#Divide into CategoricalColumns
#1. with categorys as x_ticks fig =
plt.figure(figsize=(20,26)) for idx,i
in enumerate(cat_col_eda):
   crosstab = pd.crosstab(index=final_df[i], columns=final_df["Attrition"])
ax=plt.subplot(6,4,idx+1)
   crosstab.plot(kind="bar",stacked=True,ax=ax)
```

```
    #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()
```

#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"])

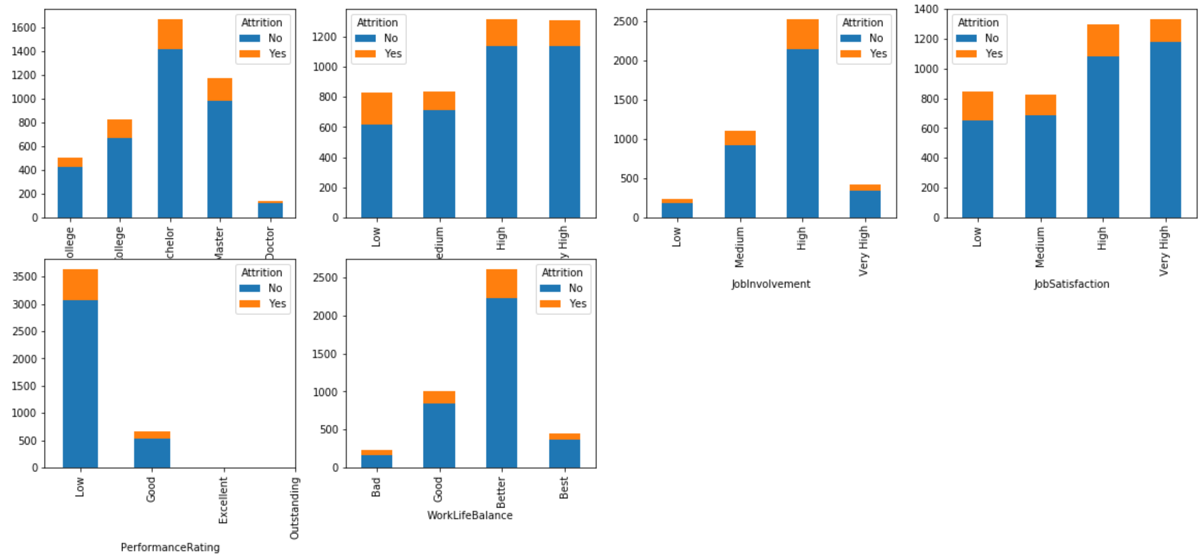
ax=plt.subplot(6,4,idx+1)

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

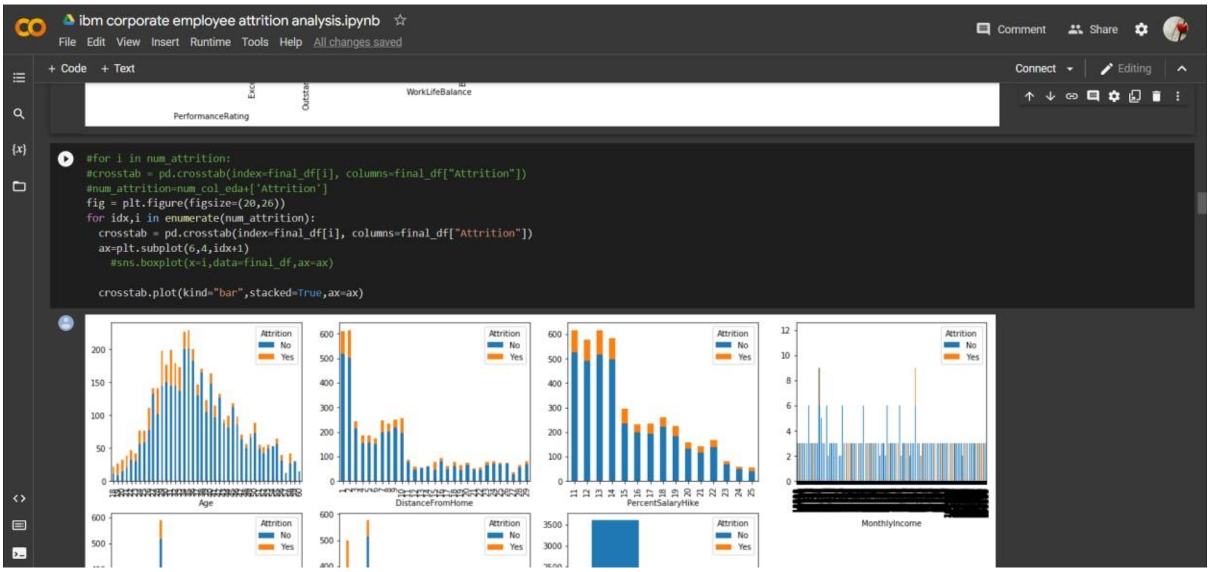
#sns.countplot(x=i[0],data=final_df,ax=ax)   ax.set(xticks=range(len(i[1])),

xticklabels=[j for j in i[1]]) plt.show()

OUTPUT:



CODING:



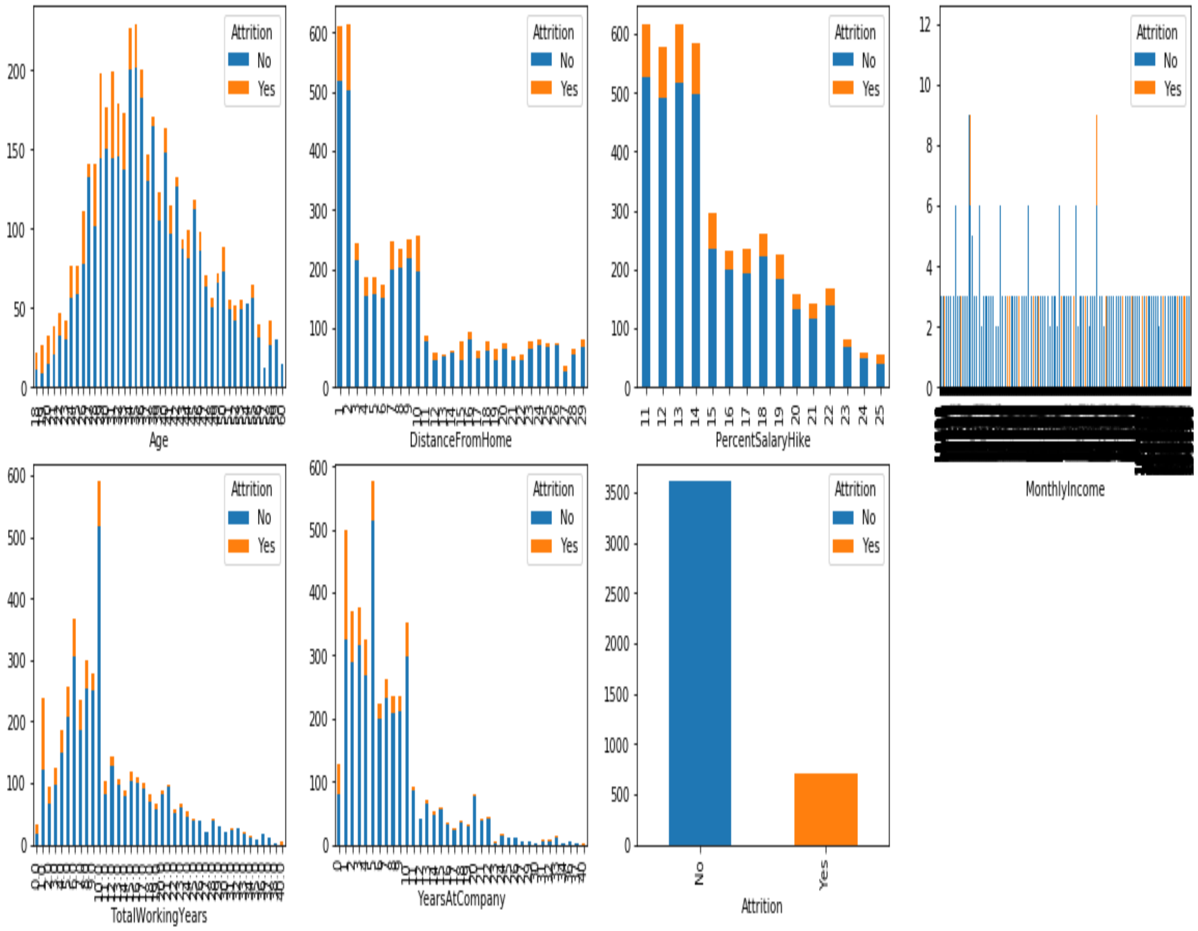
```
#for i in num_attrition:
#crosstab = pd.crosstab(index=final_df[i], columns=final_df["Attrition"])
#num_attrition=num_col_eda+['Attrition']
fig = plt.figure(figsize=(20,26))
for idx,i in enumerate(num_attrition):
  crosstab = pd.crosstab(index=final_df[i], columns=final_df["Attrition"])
  ax=plt.subplot(6,4,idx+1)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

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

#for i in num_attrition:

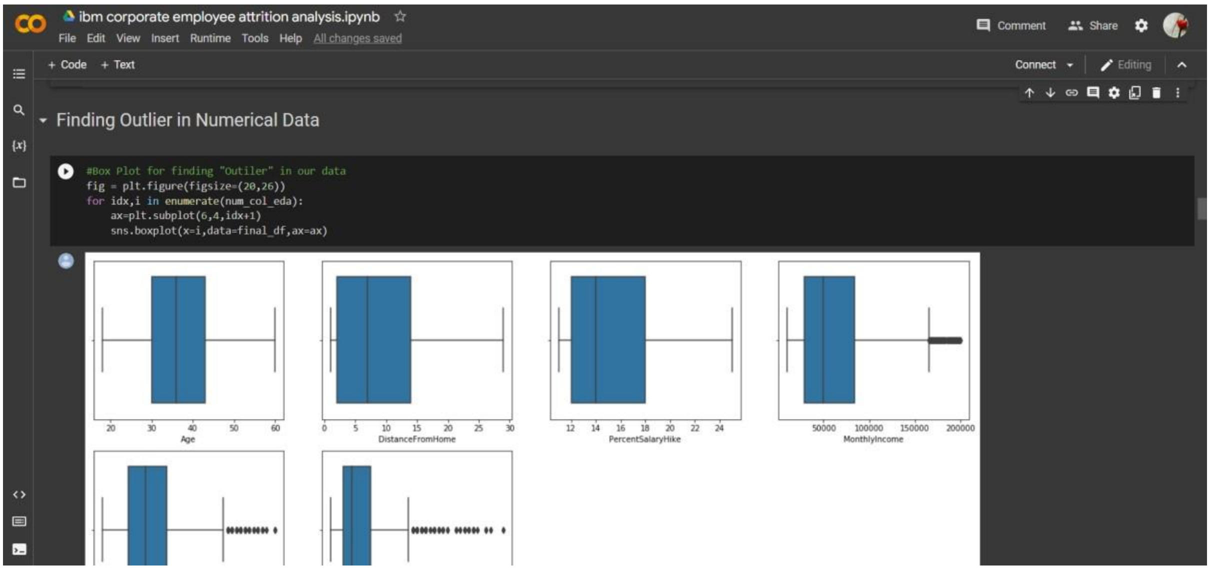
#crosstab = pd.crosstab(index=final_df[i], columns=final_df["Attrition"])

```
#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


ibm corporate employee attrition analysis.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
Connect
Editing

Finding Outlier in Numerical Data

```
#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)
```

Age
DistanceFromHome
PercentSalaryHike
MonthlyIncome

CODING:

#Box Plot for finding "Outiler" in our

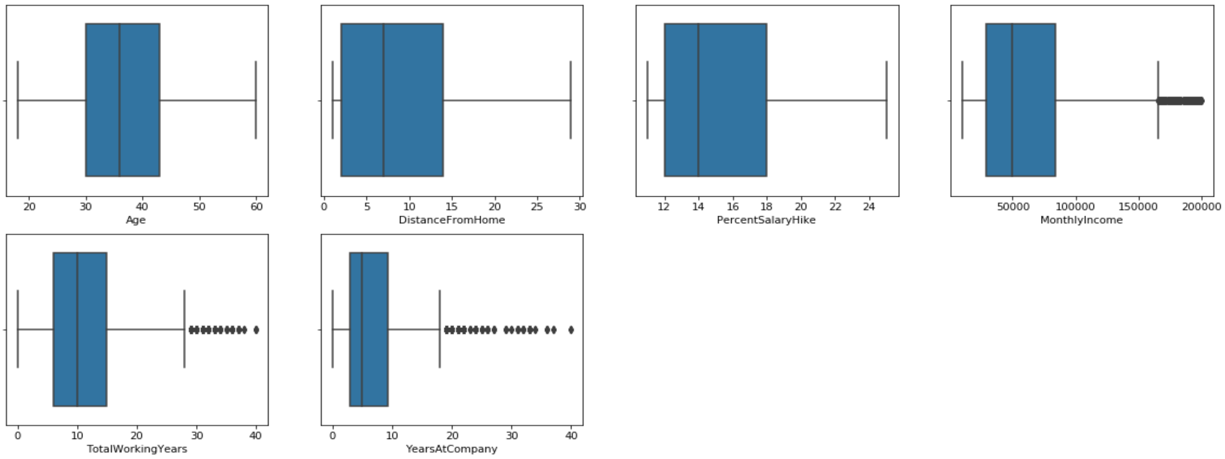
data fig = plt.figure(figsize=(20,26)) for

idx,i in enumerate(num_col_eda):

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

sns.boxplot(x=i,data=final_df,ax=ax)

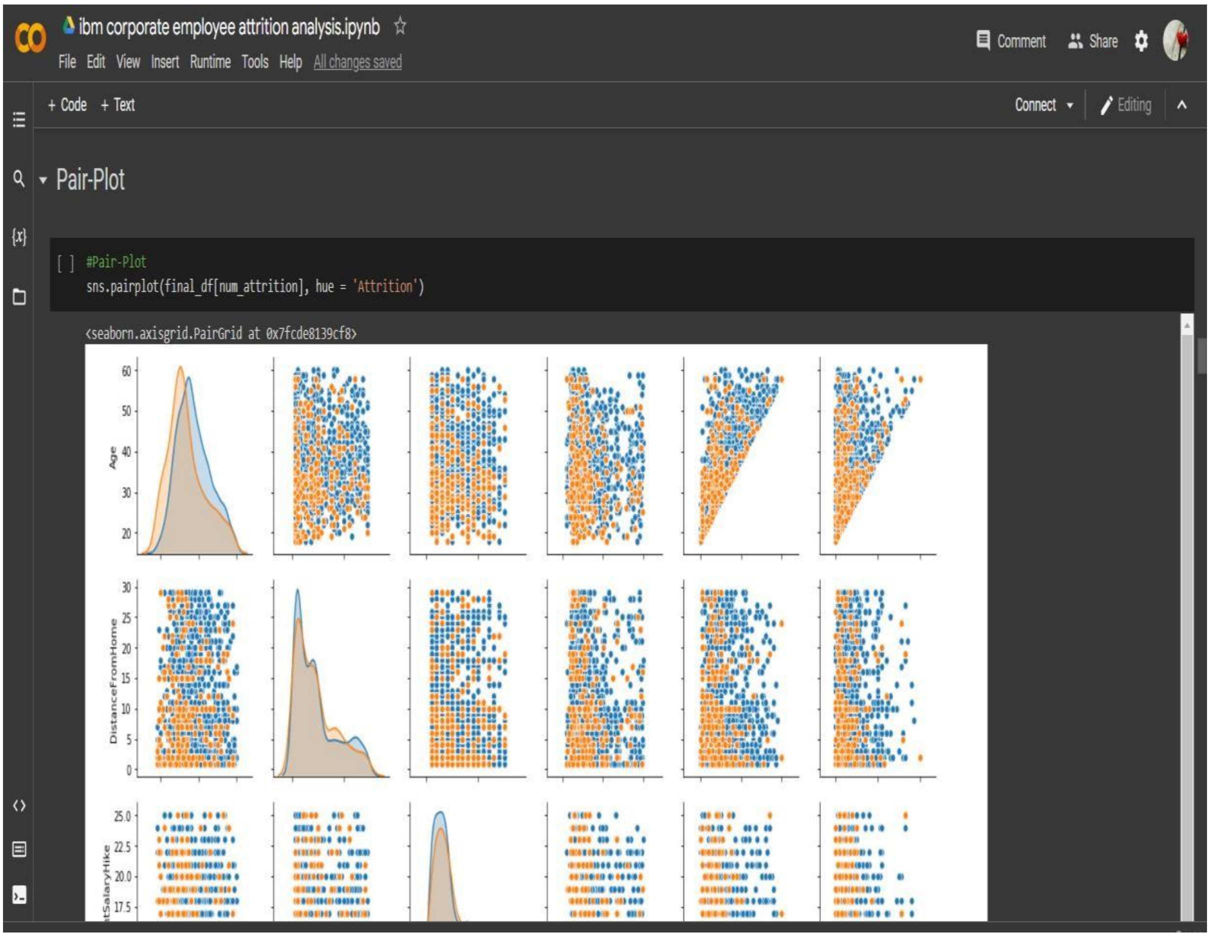
OUTPUT:



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



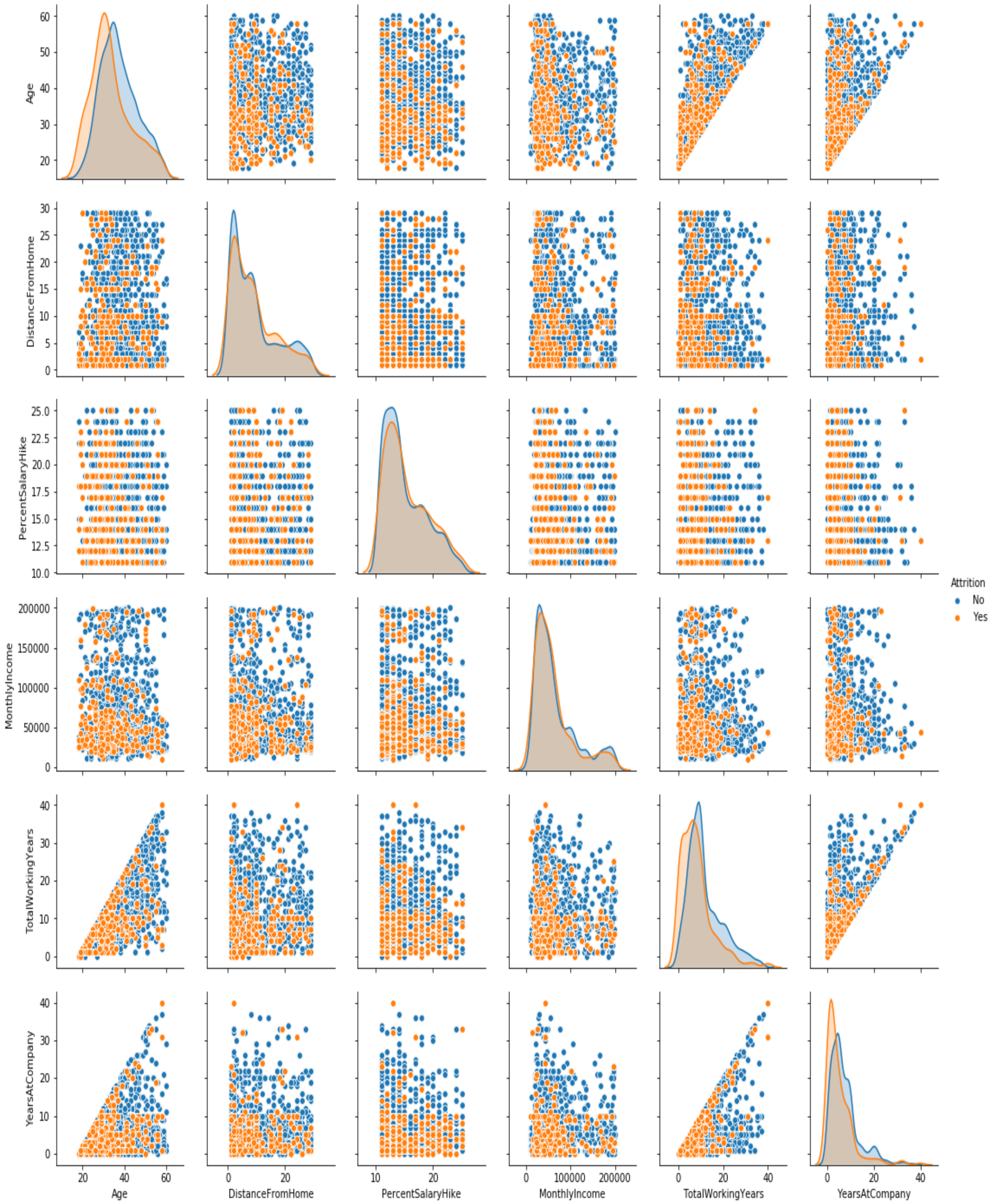
```
[ ] #Pair-Plot
sns.pairplot(final_df[num_attrition], hue = 'Attrition')
```

```
<seaborn.axisgrid.PairGrid at 0x7fcde8139cf8>
```

CODING:

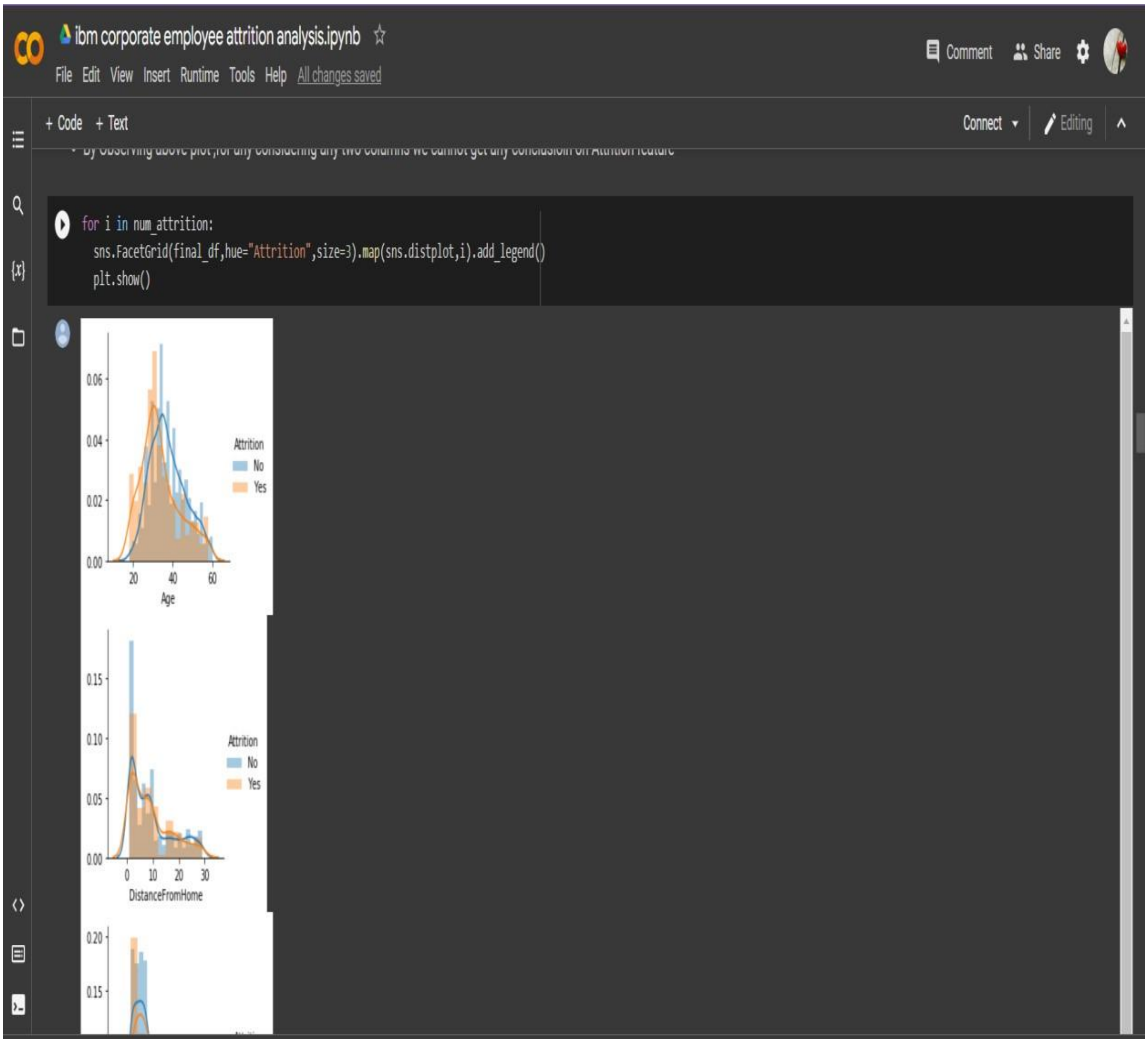
```
#Pair-Plot
sns.pairplot(final_df[num_attrition], hue = 'Attrition')
```

OUTPUT:



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

CODING:



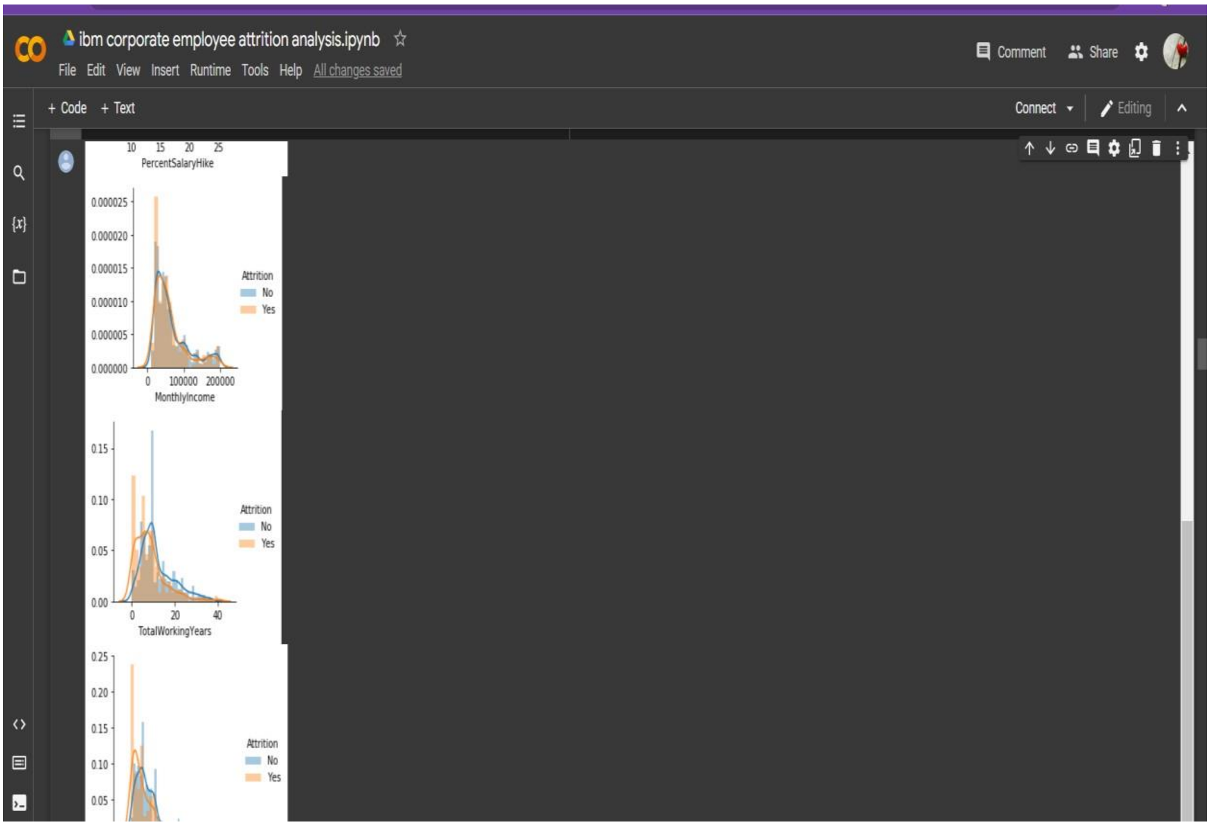
- By observing above plot ,for any considering any two columns we cannot get any conclusion on Attrition feature

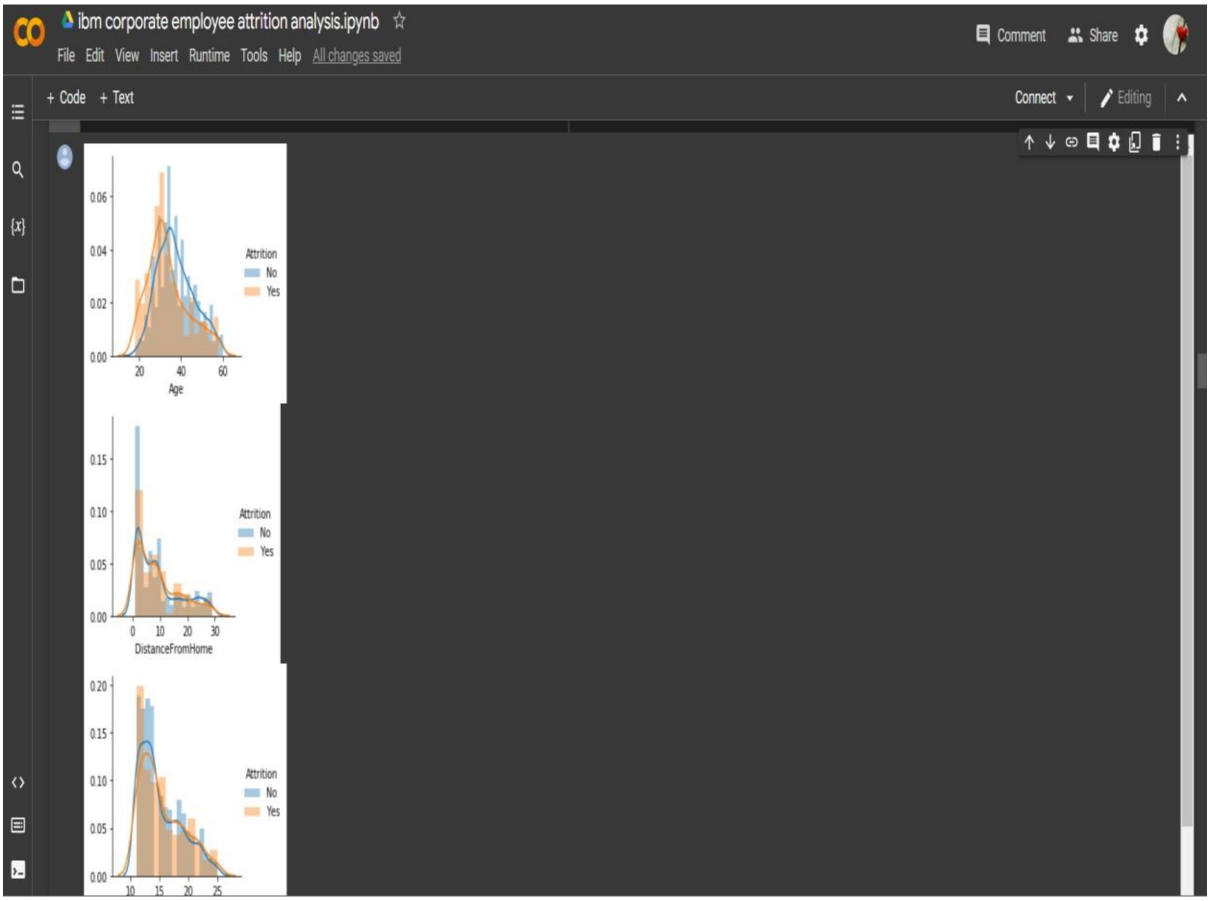
```
for i in num_attrition:
  sns.FacetGrid(final_df,hue="Attrition",size=3).map(sns.distplot,i).add_legend()
  plt.show()
```

for i in num_attrition:

sns.FacetGrid(final_df,hue="Attrition",size=3).map(sns.distplot,i).add_l e gend()   plt.show()

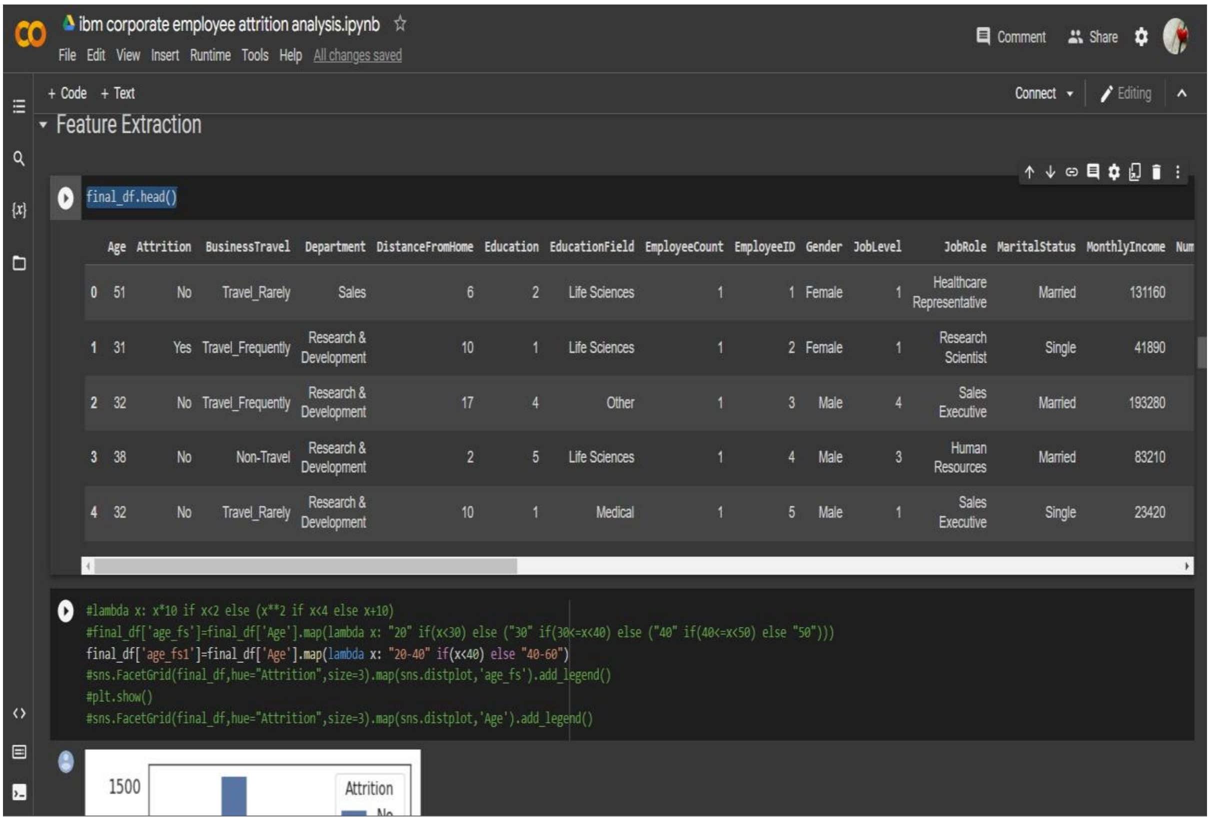
OUTPUT:


ibm corporate employee attrition analysis.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
Connect
Editing
10 15 20 25
PercentSalaryHike
0.000025
0.000020
0.000015
0.000010
0.000005
0.000000
0 100000 200000
MonthlyIncome
Attrition
No
Yes
0.15
0.10
0.05
0.00
0 20 40
TotalWorkingYears
Attrition
No
Yes
0.25
0.20
0.15
0.10
0.05
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
0.06
0.04
0.02
0.00
20 40 60
Age
Attrition
No
Yes
0.15
0.10
0.05
0.00
0 10 20 30
DistanceFromHome
Attrition
No
Yes
0.20
0.15
0.10
0.05
0.00
10 15 20 25
Attrition
No
Yes

FEATURE EXTRACTION

CODING:

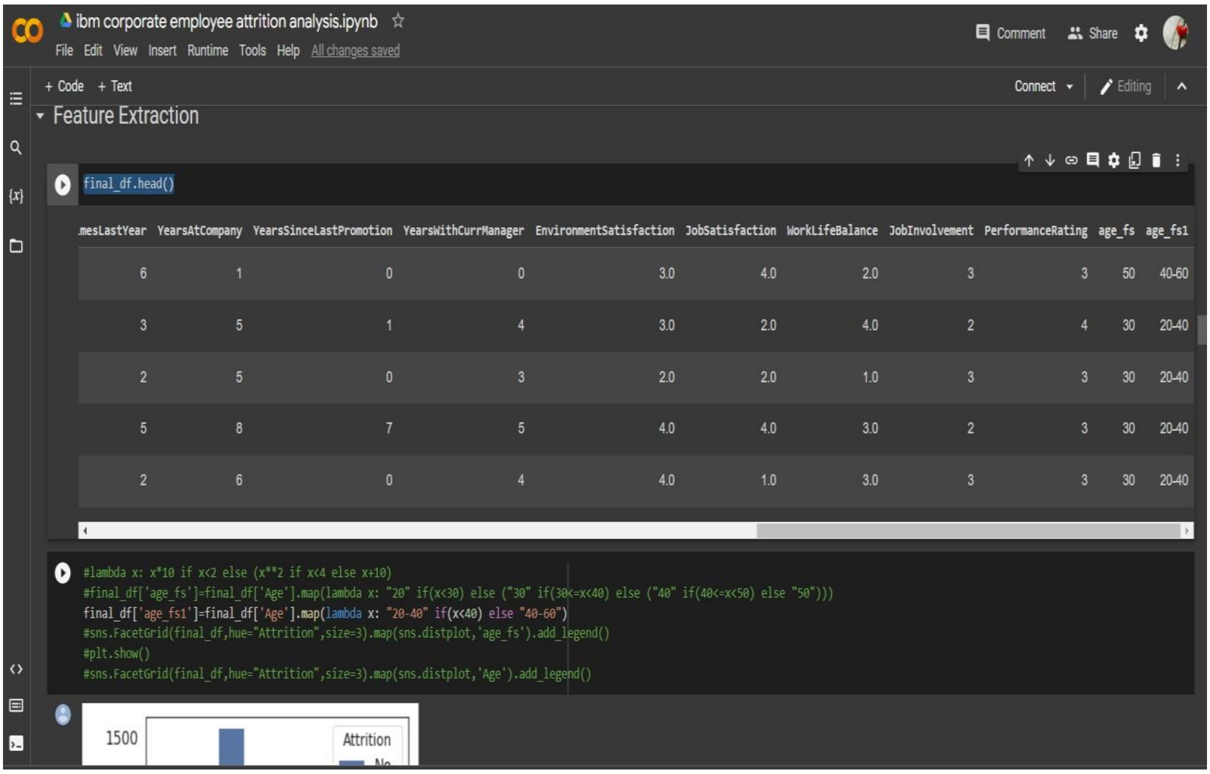

ibm corporate employee attrition analysis.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code
+ Text
Connect
Editing
1500
Attrition
No

Feature Extraction

```
final_df.head()
```

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	JobLevel	JobRole	MaritalStatus	MonthlyIncome	Num
0	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	1	Healthcare Representative	Married	131160	
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	2	Female	1	Research Scientist	Single	41890	
2	32	No	Travel_Frequently	Research & Development	17	4	Other	1	3	Male	4	Sales Executive	Married	193280	
3	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	4	Male	3	Human Resources	Married	83210	
4	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	5	Male	1	Sales Executive	Single	23420	

```
#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()
```


ibm corporate employee attrition analysis.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code
+ Text
Connect
Editing
1500
Attrition
No

Feature Extraction

```
final_df.head()
```

mesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager	EnvironmentSatisfaction	JobSatisfaction	WorkLifeBalance	JobInvolvement	PerformanceRating	age_fs	age_fs1
6	1	0	0	3.0	4.0	2.0	3	3	50	40-60
3	5	1	4	3.0	2.0	4.0	2	4	30	20-40
2	5	0	3	2.0	2.0	1.0	3	3	30	20-40
5	8	7	5	4.0	4.0	3.0	2	3	30	20-40
2	6	0	4	4.0	1.0	3.0	3	3	30	20-40

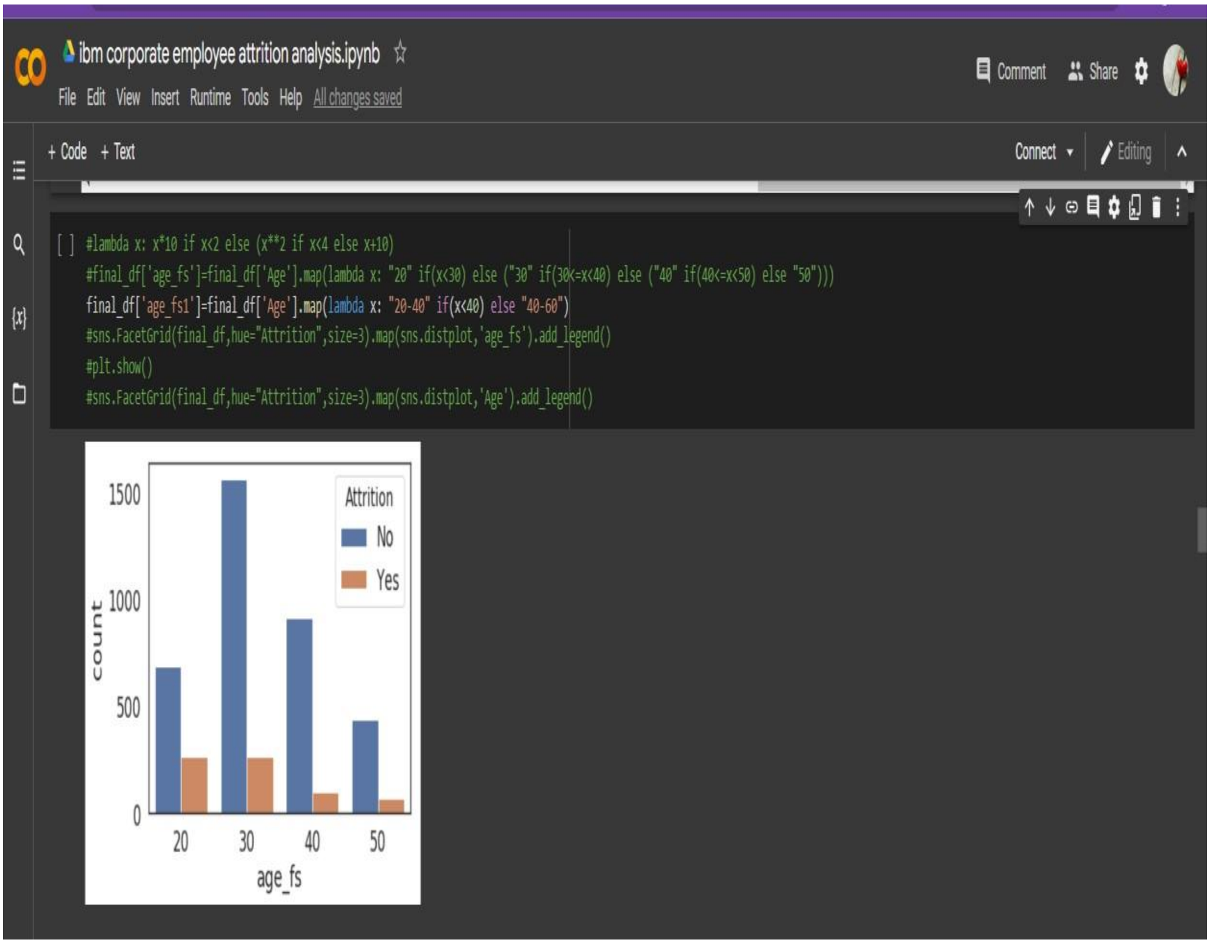
```
#lambda x: x*10 if x<2 else (x**2 if x<4 else x+10)
#final_df['age_fs']=final_df['Age'].map(lambda x: "20" if(x<30) else ("30" if(30<=x<40) else ("40" if(40<=x<50) else "50")))
final_df['age_fs1']=final_df['Age'].map(lambda x: "20-40" if(x<40) else "40-60")
#sns.FacetGrid(final_df,hue="Attrition",size=3).map(sns.distplot,'age_fs').add_legend()
#plt.show()
#sns.FacetGrid(final_df,hue="Attrition",size=3).map(sns.distplot,'Age').add_legend()
```

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
```

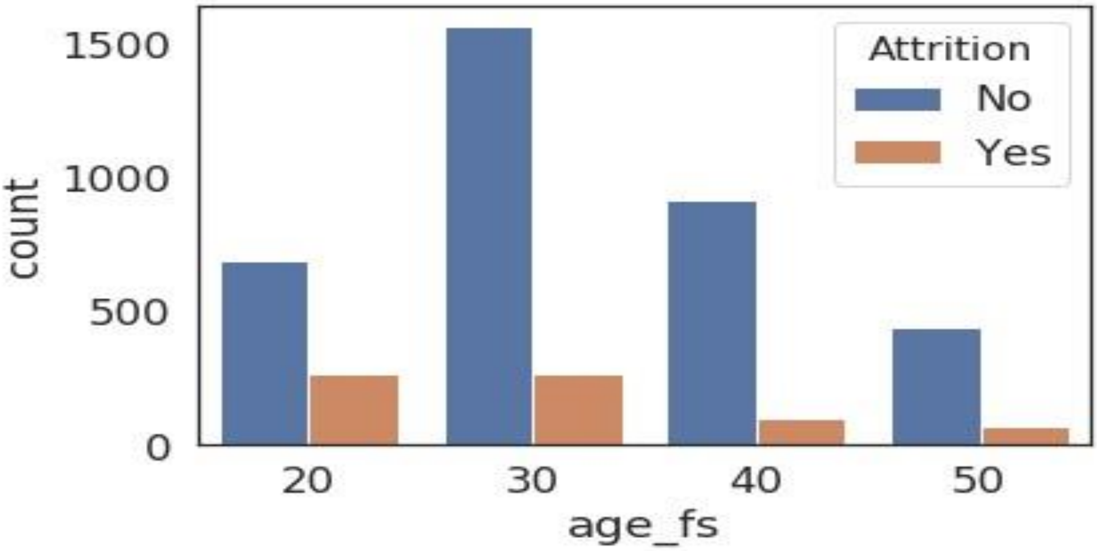
CODING:


ibm corporate employee attrition analysis.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code
+ Text
Connect
Editing
[ ]
1500
1000
500
0
count
Attrition
No
Yes
20
30
40
50
age_fs

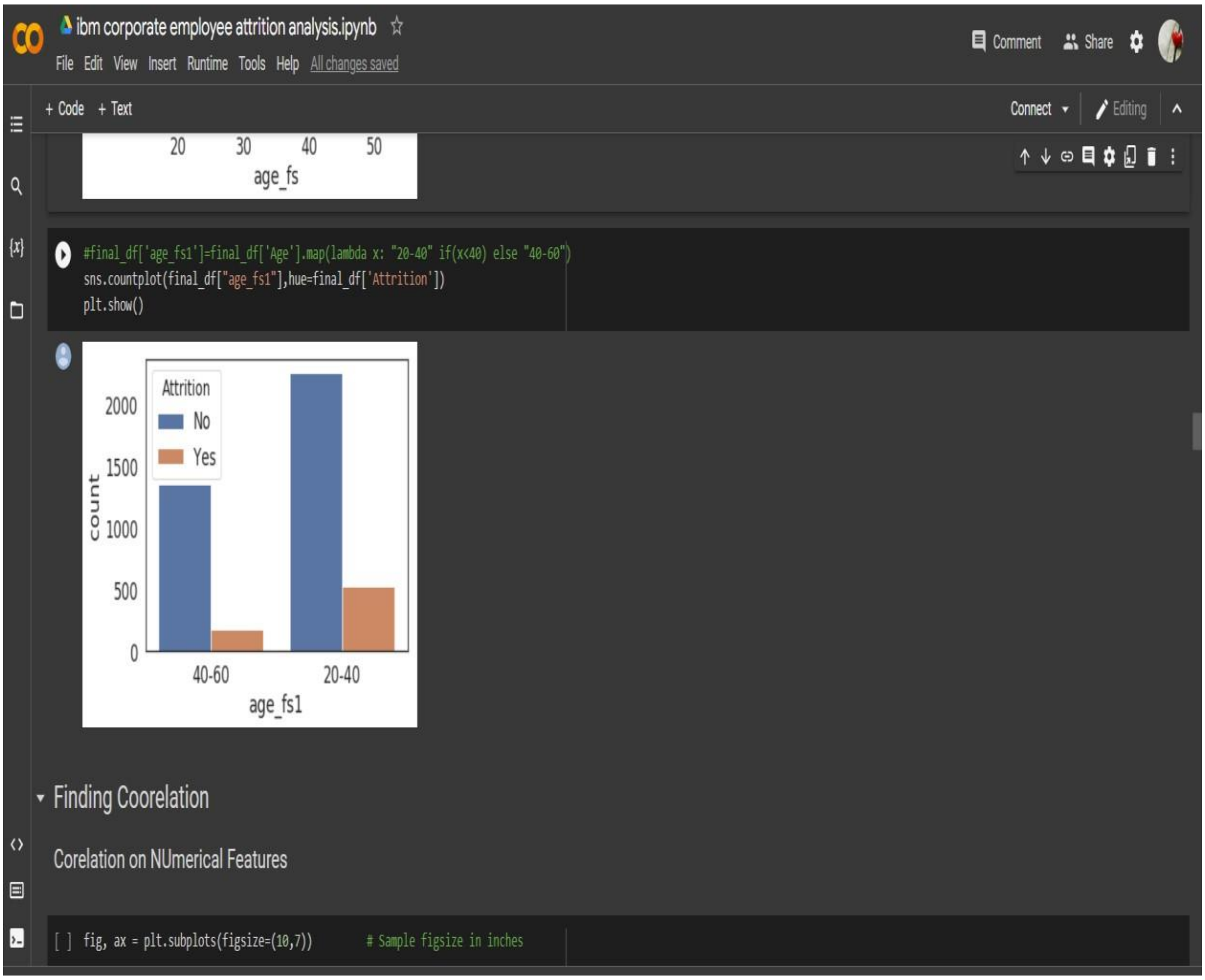
```
#lambda x: x*10 if x<2 else (x**2 if x<4 else x+10)
#final_df['age_fs']=final_df['Age'].map(lambda x: "20" if(x<30) else ("30" if(30<=x<40) else ("40" if(40<=x<50) else "50")))
final_df['age_fs1']=final_df['Age'].map(lambda x: "20-40" if(x<40) else "40-60")
#sns.FacetGrid(final_df,hue="Attrition",size=3).map(sns.distplot,'age_fs').add_legend()
#plt.show()
#sns.FacetGrid(final_df,hue="Attrition",size=3).map(sns.distplot,'Age').add_legend()
```

```
#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:



CODING:



```
#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()
```

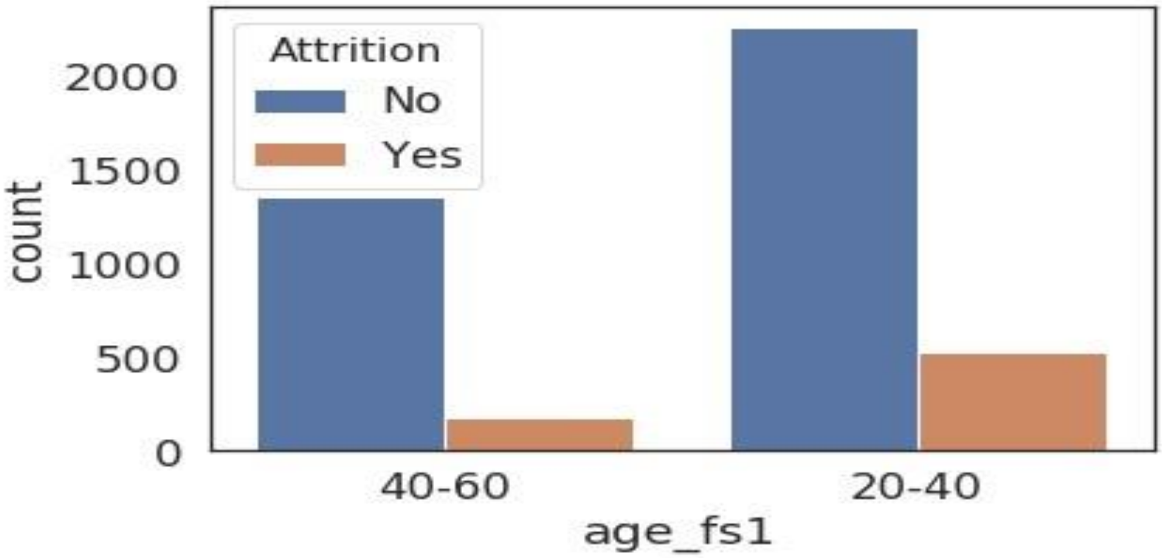
Finding Coorelation

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:

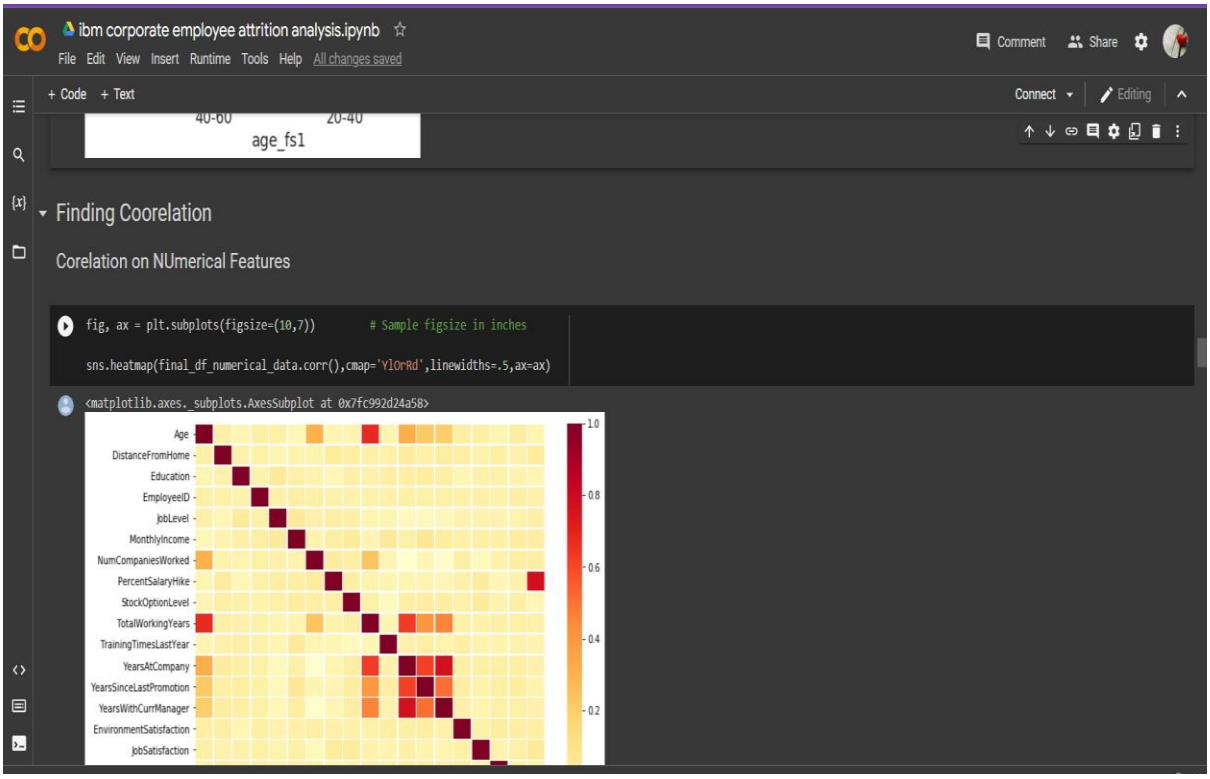


FINDING CORRELATION

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)
```

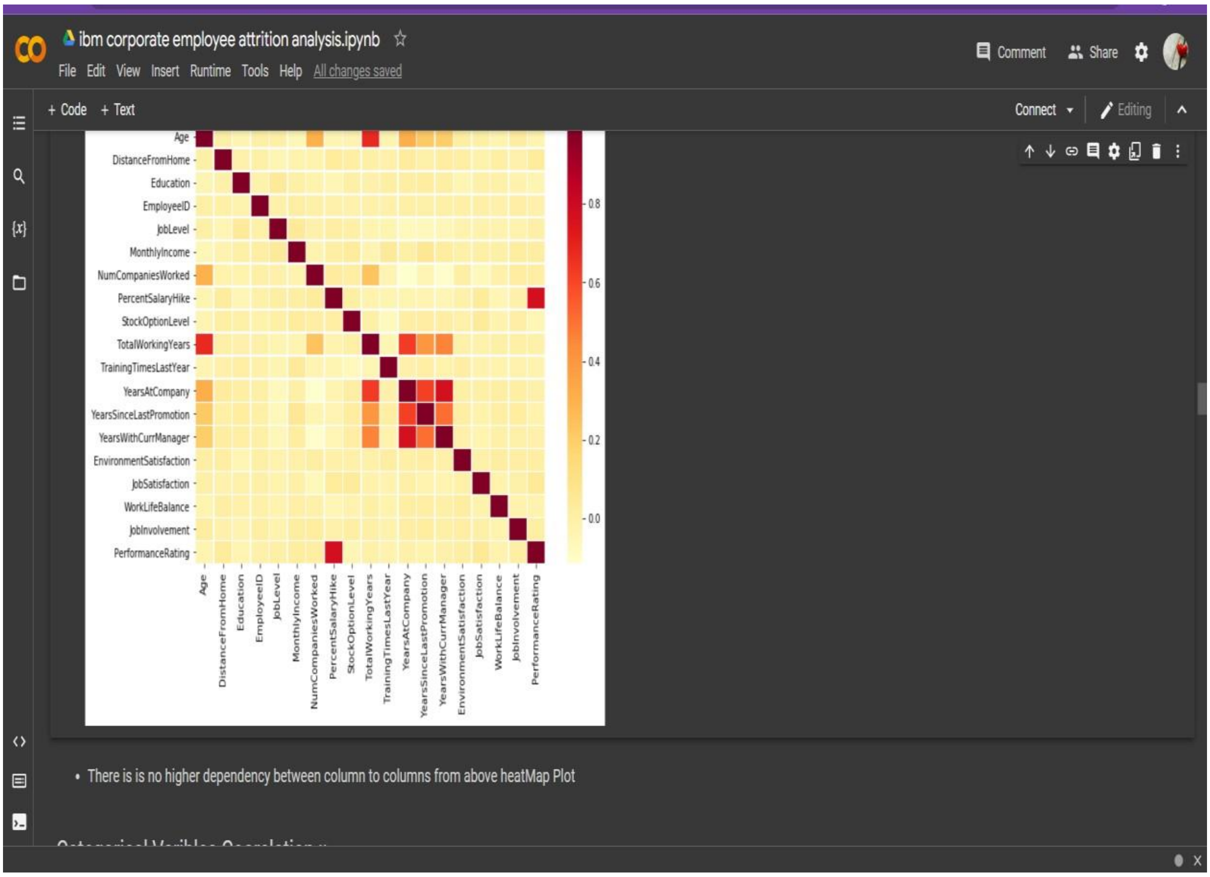


Finding Coorelation

Corelation on NUmerical Features

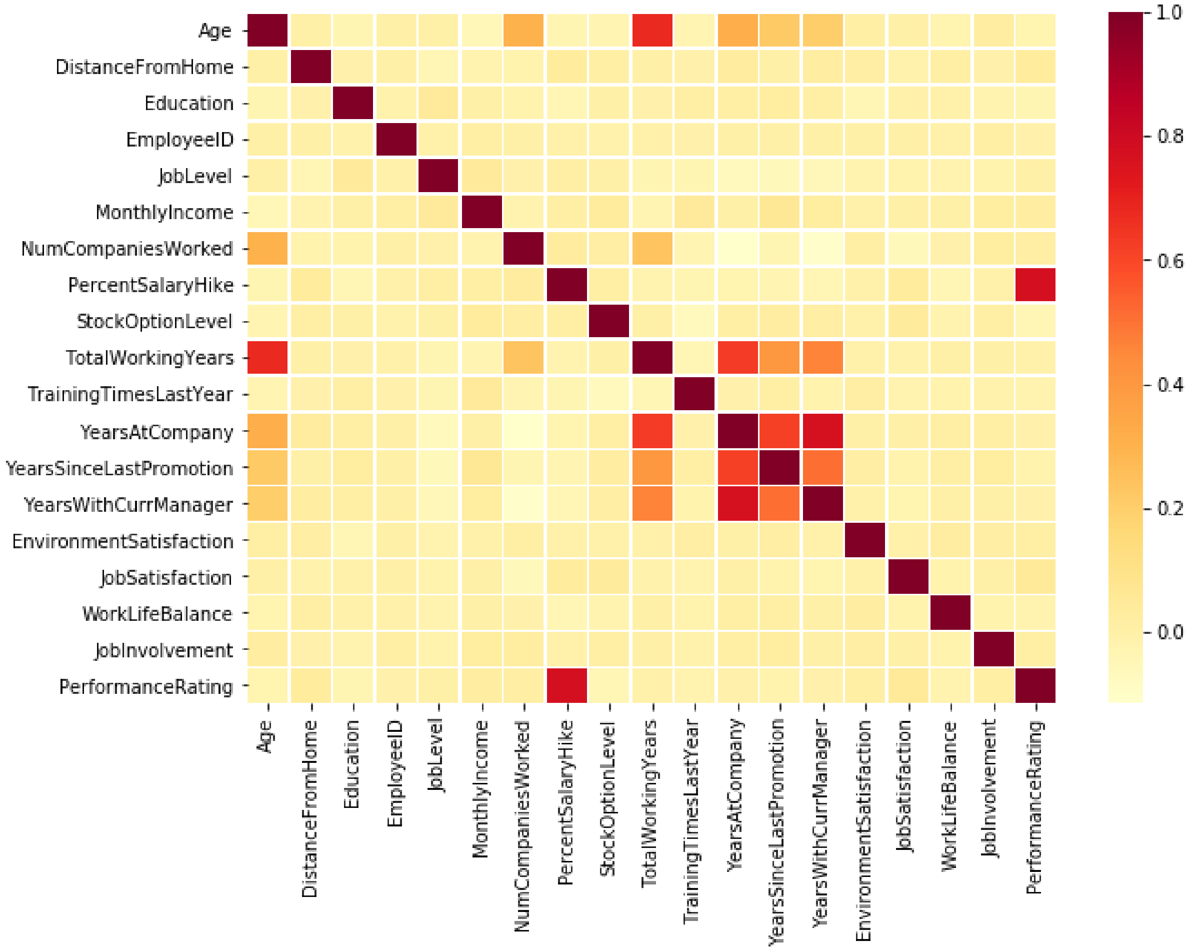
```
fig, ax = plt.subplots(figsize=(10,7))        # Sample figsize in inches

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

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



- 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