```
In [35]:
           import pandas as pd
           import numpy as np
           import matplotlib.pyplot as plt
           import seaborn as sns
           import plotly.express as px
           import warnings
           warnings.filterwarnings('ignore')
           df = pd.read csv("adult.csv")
 In [2]:
           df.head(5)
 Out[2]:
                  workclass
                              fnlwgt education education.num marital.status occupation relationship
              age
                                                                                                 ra
                                                                                         Not-in-
            0
               90
                           ?
                              77053
                                       HS-grad
                                                          9
                                                                 Widowed
                                                                                                 Wh
                                                                                          family
                                                                               Exec-
                                                                                          Not-in-
            1
                82
                      Private 132870
                                       HS-grad
                                                          9
                                                                 Widowed
                                                                                                 Wh
                                                                           managerial
                                                                                          family
                                        Some-
            2
                66
                             186061
                                                          10
                                                                 Widowed
                                                                                   ?
                                                                                       Unmarried
                                                                                                 Bla
                                        college
                                                                            Machine-
            3
               54
                      Private 140359
                                        7th-8th
                                                          4
                                                                  Divorced
                                                                                       Unmarried Wh
                                                                            op-inspct
                                        Some-
                                                                                Prof-
               41
                      Private 264663
                                                          10
                                                                Separated
                                                                                       Own-child Wh
                                                                             specialty
                                        college
 In [3]: df.shape
 Out[3]: (32561, 15)
```

In [4]: | df.dtypes

Out[4]: age int64 workclass object fnlwgt int64 education object education.num int64 marital.status object occupation object relationship object race object sex object capital.gain int64 capital.loss int64 hours.per.week int64 native.country object income object dtype: object

In [5]: df.describe

Out[5]:	<body> d ation</body>		d NDFrame tion.num	.des		e of ital.st		_	orkcl	.ass	fn ⁻	lwgt e	duc
	0 idowed	90	?	770	053		HS−gr				9		W
	1 idowed	82	Private	1328	870	ŀ	HS−gr	ad			9		W
	2 idowed	66	?	1860	061	Some-o	colle	ge			10		W
	3 vorced	54	Private	1403	359	7	7th-8	th			4		Di
	4 arated	41	Private	2640	663	Some-o	colle	ge			10		Sep
	• • •												
	32556 arried	22	Private	310	152	Some-d	colle	ge			10	Neve	r-m
	32557 spouse	27	Private	257	302	Asso	oc–ac	dm			12	Married-c	iv-
	32558 spouse	40	Private	1543	374	H	HS−gr	ad			9	Married-c	iv-
	32559 idowed	58	Private	1519	910	H	HS−gr	ad			9		W
	32560 arried	22	Private	2014	490	ŀ	∃S-gr	ad			9	Neve	r-m
			occupat	ion	re	lations	shin	race		sex	car	pital.gain	\
	0		occupac	?		-in-fan		White		ale	cup	0 () Dica c	
	1	Exe	c-manager	ial		-in-fan	nily	White		ale		0	
	2			?		Unmarı		Black		ale		0	
	3		ne-op-ins	•		Unmarı		White		ale		0	
	4	Pr	of-specia	lty		0wn-ch		White		ale		0	
	32556	Dro	tective-s	erv	No+	-in-fan	nilv	White		lale		0	
	32557		Tech-supp		NOL		√ife	White		ale		0	
	32558		ne-op-ins			Hust		White		lale		0	
	32559		Adm-cleri					White		ale		0	
	32560		Adm-cleri			0wn-ch		White		lale		0	
		capit	al.loss	hour	s.pe	r.week	nati	ve.cou	intry	inco	me		
	0		4356			40	Uni	ted-St	ates	<=5	0K		
	1		4356			18		ted-St		<=5			
	2		4356			40		ted-St		<=5			
	3		3900			40		ted-St		<=5			
	4		3900			40	Unı	ted-St	ates	<=5			
	32556		0			40	Uni	ted-St	246	- <=5	ok Ok		
	32557		0			38		ted-St		<=5			
	32558		0			40		ted-St		>5			
	32559		0			40		ted-St		<=5			
	32560		0			20		ted-St		<=5			

[32561 rows x 15 columns]>

In [6]: df = df.replace('?', np.NaN)
df.head

	ar incac	•										
Out[6]:	<pre><bound method="" ndframe.head="" of<="" td=""><td>educa</td><td>tio</td></bound></pre>										educa	tio
	0 idowed	90	NaN	770	53	HS	-gra	ad		9		W
	1 idowed	82	Private	1328	70	HS	-gra	ad		9		W
	2 idowed	66	NaN	1860	61	Some-co	lleg	je		10		W
	3 vorced	54	Private	1403	59	7t	h–8t	:h		4		Di
	4 arated	41	Private	2646	63	Some-co	lleg	je		10		Sep
		•••			• •			•				
	32556 arried	22	Private	3101	.52	Some-co	lleg	je		10	Neve	r-m
	32557 spouse	27	Private	2573	02	Assoc	-acc	dm		12	Married-c	iv-
	32558 spouse	40	Private	1543	74	HS	-gra	ad		9	Married-c	iv-
	32559 idowed	58	Private	1519	10	HS	-gra	ad		9		W
	32560 arried	22	Private	2014	90	HS	-gra	ad		9	Neve	r-m
			occupat	ion	ro	lationsh	in	race		SAV C	nital dain	
	0		•			-in-fami		White	Fema		pital.gain 0	
	1	Exe	ec-manager			-in-fami	•	White	Fema		0	
	2			NaN		Unmarri	-	Black	Fema	ale	0	1
	3		ine-op-ins	•		Unmarri			Fema		0	
	4	Pı	of-specia	ılty		Own-chi	ld	White	Fema	ale	0	ļ
	22556	D		• • •	NI - I		1					
	32556 32557	Pro	tective-s		NOT.	-in-fami Wi	•	White White	Fema	ale	0	
	32558	Mach	Tech-supp ine-op-ins			WI Husba		White		ale	0	
	32559	Hach	Adm-cleri			Unmarri		White	Fema		0	
	32560		Adm-cleri			Own-chi		White		ale	0	
		capit	tal.loss	hours	.pe	r.week n	ativ	e.coun	try :	income		
	0		4356					ed-Sta		<=50K		
	1		4356					ed–Sta		<=50K		
	2		4356					ed–Sta		<=50K		
	3		3900					ed-Sta		<=50K		
	4		3900				υniτ	ed–Sta		<=50K		
	32556		0			40	Unit	ed-Sta	tes	<=50K		
	32557		0					ed-Sta		<=50K		
	32558		0					ed–Sta		>50K		
	32559		0			40	Unit	ed-Sta	tes	<=50K		
	32560		0			20	Unit	ed–Sta	tes	<=50K		

[32561 rows x 15 columns]>

```
In [7]: | df.isna().sum()
Out[7]: age
                               0
         workclass
                            1836
         fnlwgt
                               0
         education
                               0
         education.num
                               0
         marital.status
                               0
         occupation
                            1843
         relationship
                               0
                               0
         race
                               0
         sex
         capital.gain
                               0
         capital.loss
                               0
         hours.per.week
                               0
         native.country
                             583
         income
                               0
         dtype: int64
In [8]: df["workclass"].mode()[0]
Out[8]: 'Private'
In [9]: df["occupation"].mode()[0]
Out[9]: 'Prof-specialty'
In [10]: df["native.country"].mode()[0]
Out[10]: 'United-States'
In [11]: df["workclass"].fillna(df["workclass"].mode()[0], inplace=True)
         df["occupation"].fillna(df["occupation"].mode()[0], inplace=True)
         df["native.country"].fillna(df["native.country"].mode()[0], inplace=True
         df["workclass"].describe()
Out[11]: count
                      32561
         unique
                          8
         top
                    Private
                      24532
         freq
         Name: workclass, dtype: object
```

In [12]: df.describe

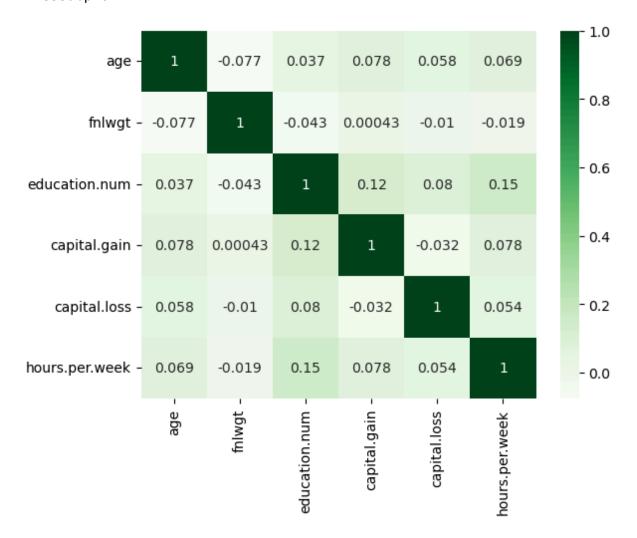
Out[12]:	<body> d ation</body>		d NDFrame tion.num	.des		e of ital.st	2+110	_	workcl	ass	fnl	wgt	educ
	0 idowed	90	Private	77	053		S-gra				9		W
	1 idowed	82	Private	132	870	Н	S-gra	ad			9		W
	2 idowed	66	Private	186	061	Some-c	olle	ge		1	LØ		W
	3 vorced	54	Private	140	359	7	th-8	th			4		Di
	4 arated	41	Private	264	663	Some-c	olle	ge		1	LØ		Sep
	• • •	• • •	• • • •		• • •		•	• •		• •	•		
	32556 arried	22	Private	310	152	Some-c	olle	ge		1	LØ	Nev	er-m
	32557 spouse	27	Private	257	302	Asso	c–ac	dm		1	L2 I	Married-	-civ-
	32558 spouse	40	Private	1543	374	Н	S-gra	ad			9 1	Married-	civ-
	32559	58	Private	1519	910	Н	S-gra	ad			9		W
	idowed 32560 arried	22	Private	201	490	Н	S–gra	ad			9	Nev	er-m
			occupat	ion	ro	lations	hin	rac	•	cov	can	ital gai	n \
	0	Pr	occupat of-specia			lations -in-fam		rac Whit		sex ale	сар	ital.gai	.n \ 0
	1		c-manager	-		-in-fam		White		ale			0
	2		of-specia			Unmarr		Blac	k Fem	ale			0
	3	Machi	ne-op-ins	pct		Unmarr	ied	Whit	e Fem	ale			0
	4	Pr	of-specia	lty		0wn-ch		Whit		ale			0
	22556	Dro	tostivo s		No+	in fom	 . 1	 		ale		• • •	
	32556 32557		tective-s Tech-supp		NOL	−in−fam W	ife	White White		ale			0 0
	32558		ne-op-ins			w Husb		Whit		ale			0
	32559		Adm-cleri	•		Unmarr		White		ale			0
	32560		Adm-cleri			0wn-ch		Whit		ale			0
		canit	al.loss	hour	s.ne	r.week	nativ	ve.co	untrv	incon	ne		
	0	Сартс	4356		J. PC	40			tates	<=50			
	1		4356			18			tates	<=50			
	2		4356			40			tates	<=50			
	2 3		3900			40			tates	<=50			
	4		3900			40			tates	<=50			
	32556		0			40	lln i	+_d_S	••• tates	<=50			
	32557		0			38			tates	<=50			
	32558		0			40			tates	>50			
	32559		0			40			tates	<=50			
	32560		0			20			tates	<=50			
								_					

[32561 rows x 15 columns]>

In [13]: | df.isna().sum() Out[13]: age 0 workclass 0 fnlwgt 0 education 0 education.num 0 marital.status 0 occupation 0 relationship 0 0 race sex 0 capital.gain 0 capital.loss 0 hours.per.week native.country 0 0 income dtype: int64

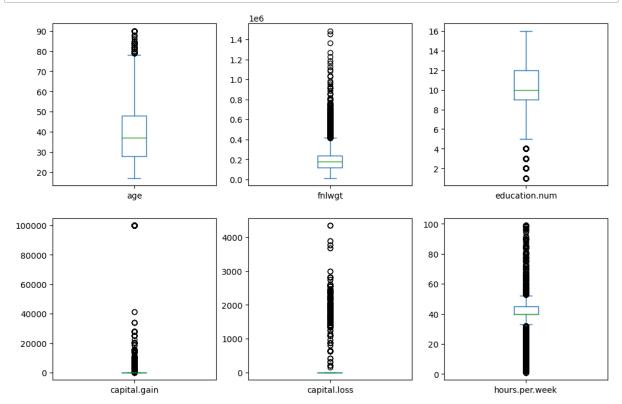
In [14]: | sns.heatmap(df.corr(),cmap = 'Greens',annot = True)

Out[14]: <AxesSubplot:>



```
In [15]: df.hist(figsize = (10,10), color = "Green")
Out[15]: array([[<AxesSubplot:title={'center':'age'}>,
                     <AxesSubplot:title={'center':'fnlwgt'}>],
                    [<AxesSubplot:title={'center':'education.num'}>,
                     <AxesSubplot:title={'center':'capital.gain'}>],
                    [<AxesSubplot:title={'center':'capital.loss'}>,
                     <AxesSubplot:title={'center':'hours.per.week'}>]], dtype=objec
           t)
                                                                             fnlwgt
                                 age
             6000
                                                         15000
             5000
                                                         12500
             4000
                                                         10000
             3000
                                                          7500
             2000
                                                          5000
             1000
                                                          2500
                0
                                                             0
                                     60
                                                                              0.75
                                                                                   1.00
                   20
                            40
                                             80
                                                              0.00
                                                                   0.25
                                                                         0.50
                                                                                         1.25
                                                                                              1.50
                                                                                              1e6
                             education.num
                                                                           capital.gain
            10000
                                                         30000
                                                         25000
             8000
                                                         20000
             6000
                                                         15000
             4000
                                                         10000
             2000
                                                          5000
                0
                                                             0
                          5.0
                                7.5
                                    10.0
                                          12.5
                                               15.0
                                                                    20000
                                                                          40000
                                                                                60000
                                                                                       80000 100000
                     2.5
                              capital.loss
                                                                         hours.per.week
            30000
                                                         15000
            25000
            20000
                                                         10000
            15000
            10000
                                                          5000
             5000
                0
                                                             0
```

In [16]: df.plot(kind = 'box', figsize = (12,12), layout = (3,3), sharex = False,

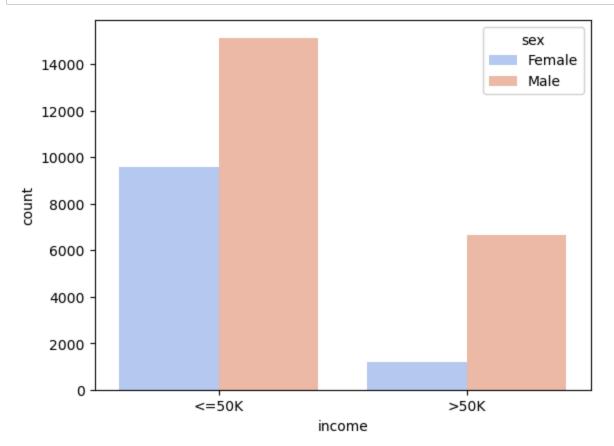


In [18]: df['income'].value_counts()

Out[18]: <=50K 24720 >50K 7841

Name: income, dtype: int64

In [19]: sns.countplot(df['income'], palette='coolwarm', hue='sex', data=df);



In [20]:
 from sklearn.pipeline import Pipeline
 from sklearn.preprocessing import StandardScaler,OneHotEncoder

num_col = ['age', 'fnlwgt','education.num','capital.gain','capital.loss'
 df.num = df[num_col]
 df.num

Out [20]:

	age	fnlwgt	education.num	capital.gain	capital.loss	hours.per.week
0	90	77053	9	0	4356	40
1	82	132870	9	0	4356	18
2	66	186061	10	0	4356	40
3	54	140359	4	0	3900	40
4	41	264663	10	0	3900	40
•••						
32556	22	310152	10	0	0	40
32557	27	257302	12	0	0	38
32558	40	154374	9	0	0	40
32559	58	151910	9	0	0	40
32560	22	201490	9	0	0	20

32561 rows × 6 columns

PCA

```
In [21]: from sklearn.decomposition import PCA
         pca = PCA(n components=3)
         pca.fit transform(df.num)
Out[21]: array([[-1.12725567e+05, -1.08159367e+03,
                                                     4.26250405e+03],
                 [-5.69085666e+04, -1.08329643e+03,
                                                     4.26463275e+03],
                 [-3.71756658e+03, -1.08491460e+03,
                                                     4.26671954e+03],
                 [-3.54043958e+04, -1.07641859e+03, -9.05501713e+01],
                 [-3.78683960e+04, -1.07634102e+03, -9.06102320e+01],
                 [ 1.17116044e+04, -1.07785772e+03, -8.87784724e+01]])
In [22]: print(pca.components )
         print(sum(pca.explained_variance_ratio_))
         [-9.90507714e-06 \quad 9.99999999e-01 \quad -1.05283840e-06 \quad 3.03678811e-05
           -3.91389877e-05 -2.19555372e-061
          [ 1.43514070e-04 -3.04337871e-05 4.27228452e-05 9.99998483e-01
           -1.72989546e-03 1.31095844e-04]
          [ 2.01710563e-03 3.91106215e-05 5.32933664e-04 1.72935042e-03
            9.99994820e-01 1.73646184e-03]]
```

0.999999969574739

```
In [23]: pca.explained_variance_ratio_
Out[23]: array([9.95113642e-01, 4.87183949e-03, 1.44878108e-05])
In [24]: X= df.drop(['income'], axis=1)
          y = df['income']
In [25]: from sklearn.preprocessing import StandardScaler, LabelEncoder
In [26]: | df2= df.copy()
         df2= df2.apply(LabelEncoder().fit_transform)
          df2.head()
Out[26]:
             age workclass fnlwgt education education.num marital.status occupation relationship race
          0
             72
                        3
                           2649
                                      11
                           6514
              65
                                      11
                                                   8
                                                              6
                                                                        3
                                                                                  1
          1
              49
                       3 11175
                                      15
                                                   9
                                                              6
                                                                        9
          3
             37
                           7009
                                      5
                                                                        6
                       3 16850
                                      15
                                                              5
                                                                        9
                                                                                  3
             24
In [27]: | ss= StandardScaler().fit(df2.drop('income', axis=1))
In [28]: X= ss.transform(df2.drop('income', axis=1))
          y= df['income']
In [29]: from sklearn.model selection import train test split
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
```

Logistic Regression

```
In [30]: from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import accuracy_score

    lr = LogisticRegression()

model = lr.fit(X_train, y_train)
    prediction = model.predict(X_test)

print("Accuracy on training data: {:,.3f}".format(lr.score(X_train, y_train))
    print("Accuracy on test data: {:,.3f}".format(lr.score(X_test, y_test)))
```

Random Forest Classifier

Accuracy on training data: 0.824 Accuracy on test data: 0.823

```
In [34]: from sklearn.ensemble import RandomForestClassifier

rfc = RandomForestClassifier()

model1 = rfc.fit(X_train, y_train)
prediction1 = model1.predict(X_test)

print("Accuracy on training data: {:,.3f}".format(rfc.score(X_train, y_t print("Accuracy on test data: {:,.3f}".format(rfc.score(X_test, y_test))
```

Accuracy on training data: 1.000 Accuracy on test data: 0.854