```
!pip install pyspark
    Collecting pyspark
       Downloading pyspark-3.1.2.tar.gz (212.4 MB)
                             212.4 MB 60 kB/s
    Collecting py4j==0.10.9
       Downloading py4j-0.10.9-py2.py3-none-any.whl (198 kB)
                                      198 kB 62.3 MB/s
    Building wheels for collected packages: pyspark
       Building wheel for pyspark (setup.py) ... done
       Created wheel for pyspark: filename=pyspark-3.1.2-py2.py3-none-any.whl size=212880768
       Stored in directory: /root/.cache/pip/wheels/a5/0a/c1/9561f6fecb759579a7d863dcd846daaa
    Successfully built pyspark
    Installing collected packages: py4j, pyspark
    Successfully installed py4j-0.10.9 pyspark-3.1.2
!ls /
    bin
              datalab home
                              lib64
                                                                    tools
                                    opt
                                           run
     boot
                       lib
                              media
                                                 tensorflow-1.15.2
              dev
                                     proc
                                           sbin
                                                                    usr
     content etc
                       lib32
                             mnt
                                                 tmp
                                     root
                                           srv
                                                                    var
from pyspark.sql import SparkSession
ss=SparkSession.builder.appName("Missingvalues").getOrCreate()
type(ss)
    pyspark.sql.session.SparkSession
                                    me2021-225.csv')
 Saving...
CSHm.show()
          _c0|
                                                  _c2|
                                                                            _c4|
                                               Scheme|revenue20 21|actuals20 21|revenue est21
     |Category|
                        Department|
             1 DOAgriculture, Co...
                                                          81342.71
                                                                       81342.71
                                                                                          116
                                               Total
             1|DOAgriculture, Co...|
                                                          12639.22
                                                                       12639.22
                                                                                           15
                                                  CIS
             1|DOAgriculture, Co...|
                                                                                           21
                                              ISSTCF|
                                                          16218.74
                                                                       16218.74
             1 DOAgriculture, Co...
                                             MIS-PSS
                                                            2004.6
                                                                         2004.6
             1|DOAgriculture, Co...|
                                             PM-AASHA
                                                            313.18
                                                                         313.18
             1|DOAgriculture, Co...|
                                             DPS/UTWS
                                                             733.9
                                                                         733.9
             1|DOAgriculture, Co...|
                                             PM-Kisan|
                                                          48713.84
                                                                       48713.84
                                                                                           7:
             1 DOAgriculture, Co...
                                               PMKMDY |
                                                               125
                                                                            125
             2 DOResearch and Ed...
                                                Total
                                                           2488.61
                                                                        2488.61
```

```
2 DOResearch and Ed...
                                              AE 
                                                        221.15
                                                                     221.15
                                                                                       24
        2 DOResearch and Ed...
                                          AgrEng
                                                         64.02
                                                                      64.02
        2 DOResearch and Ed...
                                            CRAI
                                                            46
                                                                         46
                                                                        635
        2 DOResearch and Ed...
                                    Crop Science
                                                           635
                                                                                       71
        2 DOResearch and Ed...
                                          HorSci
                                                        173.61
                                                                     173.61
        2|DOResearch and Ed...| NationAgrSciFund|
                                                            50
                                                                         50
        2 DOResearch and Ed...
                                  Animal Science
                                                           310
                                                                        310
        2|DOResearch and Ed...|Fisheries Science|
                                                        142.39
                                                                     142.39
        2 DOResearch and Ed...
                                    AgriUnivInst|
                                                           450
                                                                        450
        2|DOResearch and Ed...| EconSatatManage|
                                                         29.76
                                                                      29.76
only showing top 20 rows
```

```
CSHm=CSHm.withColumnRenamed(' c0','Category')\
.withColumnRenamed(' c1', 'Department')\
.withColumnRenamed('_c2','Scheme')\
.withColumnRenamed('_c3','revenue20_21')\
.withColumnRenamed(' c4', 'actuals20 21')\
.withColumnRenamed('_c5','revenue_est21_22')\
.withColumnRenamed('_c6','budget_est21_22')\
.withColumnRenamed('_c7','revised_reven_est21_22')\
.withColumnRenamed(' c8','revised budgt est21 22')\
.withColumnRenamed('_c9','budget_rev_est22_23')\
.withColumnRenamed('_c10','budget_est22_23')
```

CSHm.columns

```
['Category',
     'Department',
     'Scheme',
     'revenue20 21',
     'actuals20_21',
     'revenue est21 22',
     | h...d==+24 22 |
Saving...
     budget rev est22 23,
     'budget est22 23',
     'revenue est21 23',
     'revised est21 22']
```

CSHm.printSchema()

```
root
```

```
|-- Category: string (nullable = true)
|-- Department: string (nullable = true)
|-- Scheme: string (nullable = true)
|-- revenue20 21: string (nullable = true)
|-- actuals20 21: string (nullable = true)
|-- revenue est21 22: float (nullable = true)
|-- budget est21 22: string (nullable = true)
```

```
|-- revised_reven_est21_22: float (nullable = true)
|-- revised_budgt_est21_22: string (nullable = true)
|-- budget_rev_est22_23: float (nullable = true)
|-- budget_est22_23: string (nullable = true)
|-- revenue_est21_23: float (nullable = true)
|-- revised_est21_22: float (nullable = true)
```

CSHm.na.drop().show()

Category	Depa	rtment	Scheme	revenue20_21	actuals20_21	revenue_es
1	DOAgriculture,	Co	Total	81342.71	81342.71	11
1	DOAgriculture,	Co	CIS	12639.22	12639.22] 1
1	DOAgriculture,	Co	ISSTCF	16218.74	16218.74	1
1	DOAgriculture,	Co	MIS-PSS	2004.6	2004.6	
1	DOAgriculture,	Co	PM-AASHA	313.18	313.18	
1	DOAgriculture,	Co	DPS/UTWS	733.9	733.9	
1	DOAgriculture,	Co	PM-Kisan	48713.84	48713.84] 7
1	DOAgriculture,	Co	PMKMDY	125	125	
2	DOResearch and	Ed	Total	2488.61	2488.61	
2	DOResearch and	Ed	AE	221.15	221.15	
2	DOResearch and	Ed	AgrEng	64.02	64.02	
2	DOResearch and	Ed	CRAI	46	46	
2	DOResearch and	Ed	Crop Science	635	635	
2	DOResearch and	Ed	HorSci	173.61	173.61	
2	DOResearch and	Ed	NationAgrSciFund	50	50	
2	DOResearch and	Ed	Animal Science	310	310	
2	DOResearch and	Ed	Fisheries Science	142.39	142.39	
2	DOResearch and	Ed	AgriUnivInst	450	450	
2	DOResearch and	Ed	EconSatatManage	29.76	29.76	
2	DOResearch and	Ed	NAgriHighEduPro(EAP)	207.85	207.85	

only showing top 20 rows

◀		→
Saving	×	

evenue_es
11
1
2
7
•

	2 DOResearch and Ed	Crop Science	635	635	
	2 DOResearch and Ed	HorSci	173.61	173.61	
	2 DOResearch and Ed	NationAgrSciFund	50	50	
	2 DOResearch and Ed	Animal Science	310	310	
	2 DOResearch and Ed	Fisheries Science	142.39	142.39	
	2 DOResearch and Ed	AgriUnivInst	450	450	
	2 DOResearch and Ed	EconSatatManage	29.76	29.76	
	2 DOResearch and Ed	NAgriHighEduPro(EAP)	207.85	207.85	
_		L			

only showing top 20 rows

CSHm.na.drop(how='all').show()

_est21	revenue	actuals20_21	revenue20_21	Scheme	Department	Category
r	 	actuals20_21	revenue20_21	Scheme	Department	Category
11649		81342.71	81342.71	Total	DOAgriculture, Co	1
1569		12639.22	12639.22	CIS	DOAgriculture, Co	1
2117		16218.74	16218.74	ISSTCF	DOAgriculture, Co	1
206		2004.6	2004.6	MIS-PSS	DOAgriculture, Co	1
5(313.18	313.18	PM-AASHA	DOAgriculture, Co	1
86		733.9	733.9	DPS/UTWS	DOAgriculture, Co	1
7506		48713.84	48713.84	PM-Kisan	DOAgriculture, Co	1
22		125	125	PMKMDY	DOAgriculture, Co	1
272		2488.61	2488.61	Total	DOResearch and Ed	2
24		221.15	221.15	AE	DOResearch and Ed	2
7		64.02	64.02	AgrEng	DOResearch and Ed	2
5		46	46	CRAI	DOResearch and Ed	2
71		635	635	Crop Science	DOResearch and Ed	2
19		173.61	173.61	HorSci	DOResearch and Ed	2
		50	50	NationAgrSciFund	DOResearch and Ed	2
33		310	310	Animal Science	DOResearch and Ed	2
15		142.39	142.39	Fisheries Science	DOResearch and Ed	2
48		450	450	AgriUnivInst	DOResearch and Ed	2
3		29.76	29.76	EconSatatManage	NOResearch and Ed	2

CSHm.na.drop(thresh=3).show()

+		.+		+	
Category	Departmen	Scheme	revenue20_21	actuals20_21	revenue_est21
Category	Departmen	: Scheme	revenue20_21	actuals20_21	r
1	DOAgriculture, Co	Total	81342.71	81342.71	11649
1	DOAgriculture, Co	CIS	12639.22	12639.22	1569
1	DOAgriculture, Co	ISSTCF	16218.74	16218.74	2117
1	DOAgriculture, Co	MIS-PSS	2004.6	2004.6	206
1	DOAgriculture, Co	PM-AASHA	313.18	313.18	56
1	DOAgriculture, Co	DPS/UTWS	733.9	733.9	86

	1 DOAgriculture, Co	PM-Kisan	48713.84	48713.84	7506
ĺ	1 DOAgriculture, Co	PMKMDY	125	125	22
ĺ	2 DOResearch and Ed	Total	2488.61	2488.61	272
ĺ	2 DOResearch and Ed	AE	221.15	221.15	24
ĺ	2 DOResearch and Ed	AgrEng	64.02	64.02	7
ĺ	2 DOResearch and Ed	CRAI	46	46	ī
ĺ	2 DOResearch and Ed	Crop Science	635	635	71
ĺ	2 DOResearch and Ed	HorSci	173.61	173.61	19
ĺ	2 DOResearch and Ed	NationAgrSciFund	50	50	ī
	2 DOResearch and Ed	Animal Science	310	310	33
	2 DOResearch and Ed	Fisheries Science	142.39	142.39	1:
	2 DOResearch and Ed	AgriUnivInst	450	450	48
	2 DOResearch and Ed	EconSatatManage	29.76	29.76	3
4		+			

only showing top 20 rows

←

CSHm.na.drop(subset='Department').show()

				+		
revenue_est21	actuals20_21	20_21	revenue20	Scheme	Department	Category
r	actuals20_21	1 20_21	revenue20	Scheme	Department	Category
11649	81342.71	42.71	81342	Total	DOAgriculture, Co	1
1569	12639.22	39.22	12639	CIS	DOAgriculture, Co	1
2117	16218.74	18.74	16218	ISSTCF	DOAgriculture, Co	1
200	2004.6	004.6	2004	MIS-PSS	DOAgriculture, Co	1
56	313.18	13.18	313	PM-AASHA	DOAgriculture, Co	1
86	733.9	733.9	73:	DPS/UTWS	DOAgriculture, Co	1
7500	48713.84	13.84	48713	PM-Kisan	DOAgriculture, Co	1
22	125	125] :	PMKMDY	DOAgriculture, Co	1
272	2488.61	88.61	2488	Total	DOResearch and Ed	2
24	221.15	21.15	221	AE	DOResearch and Ed	2
7	64.02	64.02	64	AgrEng	DOResearch and Ed	2
-	46	46		CRAI	DOResearch and Ed	2
71	635	635		Crop Science	DOResearch and Ed	2
19	173.61	73.61	173	HorSci	NOResearch and Ed)
-	50	50		NationAgrSciFund	<u> </u>	ng
33	310	310]	Animal Science		iig
15	142.39	42.39	142	Fisheries Science	DOMESEAL CIT ALIA EA	4
48	450	450	4	AgriUnivInst	DOResearch and Ed	2
3	29.76	29.76	29	EconSatatManage	DOResearch and Ed	2

only showing top 20 rows

CSHm=CSHm.na.drop(subset='Department')

CSHm.na.fill('4.1', 'actuals20_21').show()

+	+				+	
Category	Depart	tment	Scheme	revenue20_21	actuals20_21	r
1	DOAgriculture, (Co	Total	81342.71	81342.71	11649
1	DOAgriculture, (Co	CIS	12639.22	12639.22	1569
1	DOAgriculture, (Co	ISSTCF	16218.74	16218.74	2117
1	DOAgriculture, (Co	MIS-PSS	2004.6	2004.6	206
1	DOAgriculture, (Co	PM-AASHA	313.18	313.18	56
1	DOAgriculture, (Co	DPS/UTWS	733.9	733.9	86
1	DOAgriculture, (Co	PM-Kisan	48713.84	48713.84	7506
1	DOAgriculture, (Co	PMKMDY	125	125	22
2	DOResearch and E	Ed	Total	2488.61	2488.61	272
2	DOResearch and E	Ed	AE	221.15	221.15	24
2	DOResearch and E	Ed	AgrEng	64.02	64.02	7
2	DOResearch and E	Ed	CRAI	46	46	Ē
2	DOResearch and E	Ed	Crop Science	635	635	71
2	DOResearch and E	Ed	HorSci	173.61	173.61	19
2	DOResearch and E	Ed	NationAgrSciFund	50	50	Ē
2	DOResearch and E	Ed	Animal Science	310	310	33
2	DOResearch and E	Ed	Fisheries Science	142.39	142.39	15
2	DOResearch and E	Ed	AgriUnivInst	450	450	48
2	DOResearch and E	Ed	EconSatatManage	29.76	29.76	3
+	+				+	

only showing top 20 rows

4

CSHm.na.fill('4.1',['revenue20_21','actuals20_21']).show()

]	Category	De	epartment	Scheme	revenue20_21	actuals20_21	revenue_est21
	Category	De	epartment	Scheme	revenue20_21	actuals20_21	r
	1	DOAgricultu	re, Co	Total	81342.71	81342.71	11649
	1	DOAgricultu	re, Co	CIS	12639.22	12639.22	1569
	1	DOAgricultu	re, Co	ISSTCF	16218.74	16218.74	2117
	1	DOAgricultu	re, Co	MIS-PSS	2004.6	2004.6	206
	1	DOAgricultu	re, Co	PM-AASHA	313.18	313.18	5(
	1	ln∩∆oricultuu	re (n	DPS/UTWS	733.9	733.9	86
Covir	20			PM-Kisan	48713.84	48713.84	750€
Savir	ıg			PMKMDY	125	125	22
		DOVESEALCH (anu cu	Total	2488.61	2488.61	272
	2	DOResearch a	and Ed	AE	221.15	221.15	24
	2	DOResearch a	and Ed	AgrEng	64.02	64.02	j
	2	DOResearch a	and Ed	CRAI	46	46	į į
	2	DOResearch a	and Ed	Crop Science	635	635	71
	2	DOResearch a	and Ed	HorSci	173.61	173.61	19
	2	DOResearch a	and Ed	NationAgrSciFund	50	50	ī
	2	DOResearch a	and Ed	Animal Science	310	310	33
	2	DOResearch a	and Ed	Fisheries Science	142.39	142.39	1!
	2	DOResearch a	and Ed	AgriUnivInst	450	450	48
	2	DOResearch a	and Ed	EconSatatManage	29.76	29.76	3
+		+					

only showing top 20 rows

→

```
CSHm=CSHm.withColumn('revenue est21 22',CSHm['revenue est21 22'].cast('float'))
CSHm=CSHm.withColumn('revenue est21 23',CSHm['budget est21 22'].cast('float'))
CSHm=CSHm.withColumn('revised_reven_est21_22',CSHm['revised_reven_est21_22'].cast('float'))
CSHm=CSHm.withColumn('revised est21 22',CSHm['revised reven est21 22'].cast('float'))
CSHm=CSHm.withColumn('budget_rev_est22_23',CSHm['budget_rev_est22_23'].cast('float'))
CSHm=CSHm.withColumn('budget rev est22 23',CSHm['budget rev est22 23'].cast('float'))
CSHm.columns
     ['Category',
      'Department',
      'Scheme',
      'revenue20 21',
      'actuals20 21',
      'revenue est21 22',
      'budget est21 22',
      'revised reven est21 22',
      'revised budgt est21 22',
      'budget rev est22 23',
      'budget est22 23',
      'revenue est21 23',
      'revised est21 22']
from pyspark.ml.feature import Imputer
CSHm.head()
     Row(Category='Category', Department='Department', Scheme='Scheme', revenue20 21='revenue
CSHm.tail(5)
     [Row(Category='25', Department='DOHigher Education ', Scheme='IScGF', revenue20_21='1674
                                     'DOHigher Education ', Scheme='SCUs', revenue20_21='368.7
                                     'DOHigher Education ', Scheme='PMRF', revenue20_21='26.26
 Saving...
                                     'DOHigher Education ', Scheme='NMETICT', revenue20 21='10
      Row(Category='25', Department='DOHigher Education', Scheme='Svc(MOOCs)', revenue20 21=
CSHm
     DataFrame[Category: string, Department: string, Scheme: string, revenue20_21: string, ac
CSHm.groupby('Department')
     <pyspark.sql.group.GroupedData at 0x7f56052ff290>
```

type(CSHm)

pyspark.sql.dataframe.DataFrame

CSHm.groupby('Department').avg().show()

Department	avg(revenue_est21_22)	avg(revised_reven_est21_22)	avg(budget_rev_@
Ministry of Culture	192.04333353042603	130.92666721343994	151.27166
DOHigher Education	997.6888780328962	695.3955697748396	1156.33896
DOAgriculture, Co	28985.0	25658.324584960938	25967.2028
DOFood Distribution	18573.5262770286	67209.59737946437	38464.762
DOSchool Education	251.0	176.625	
Department	null	null	
DO Petrochemicals	38.1966667175293	17.09999984105427	37.819999
Ministry of Defen	500.0	756.6666666666666	566.66666
Ministry of Corpo	8.0	6.0	
Ministry of Coal	409.9899959564209	276.8649950027466	209.990003
DOTelecom	2319.6971345629013	2075.9971765790665	2599.2913
DOConsumer Affairs	389.0181907307018	2164.992723768408	513.53635
Department for pr	1289.601953125	1833.201953125	1728.8119
Ministry of Devel	512.9022284613716	337.984448923005	446.55555
DOPosts	317.17999903361004	245.24666341145834	239.25333
Ministry of Civil	1740.2024765014648	1862.729995727539	1434.49999
Ministry of Earth	388.3333333333333	191.83333333333334	339.389999
DOResearch and Ed	440.3333333333333	369.71833260854083	431.41666
DOFertilisers	47539.333333333336	89298.19921875	53019.7864
DOPharma	112.39999923706054	163.3319969177246	164.30406

CSHm.describe()

DataFrame[summarv: string Category: string, Department: string, Scheme: string, revenue Saving...

CSHm.describe().show()

	summary	Category	Department	Scheme	revenue20_21
İ	count			135	
	mean	12.791044776119403	null	null	4533.800298507461
	stddev	7.159907827845493	null	null	16631.58525419928
	min	1	DO Petrochemicals	ACROSS	0.01
ĺ	max	Category	Ministry of Earth	WirelessPlanCoord	revenue20_21
į			· · · · · · · · · · · · · · · · · · ·	· ·	

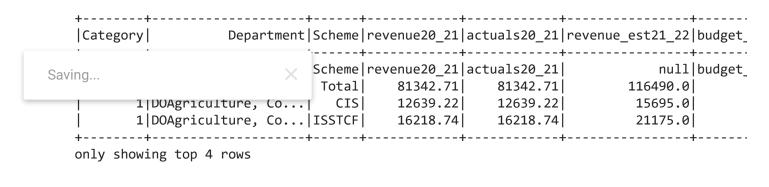
CSHm.show()

+	+	+	+		
Category	Department	Scheme	revenue20_21	actuals20_21	revenue_est21
+	+	+	+		·
Category	Department	Scheme	revenue20_21	actuals20_21	r
1	DOAgriculture, Co	Total	81342.71	81342.71	11649
1	DOAgriculture, Co	CIS	12639.22	12639.22	1569
1	DOAgriculture, Co	ISSTCF	16218.74	16218.74	2117
1	DOAgriculture, Co	MIS-PSS	2004.6	2004.6	206
1	DOAgriculture, Co	PM-AASHA	313.18	313.18	56
1	DOAgriculture, Co	DPS/UTWS	733.9	733.9	86
1	DOAgriculture, Co	PM-Kisan	48713.84	48713.84	7506
1	DOAgriculture, Co	PMKMDY	125	125	22
2	DOResearch and Ed	Total	2488.61	2488.61	272
2	DOResearch and Ed	AE	221.15	221.15	24
2	DOResearch and Ed	AgrEng	64.02	64.02	<u> </u>
2	DOResearch and Ed	CRAI	46	46	
2	DOResearch and Ed	Crop Science	635	635	71
2	DOResearch and Ed	HorSci	173.61	173.61	19
2	DOResearch and Ed	NationAgrSciFund	50	50	
2	DOResearch and Ed	Animal Science	310	310	3:
2	DOResearch and Ed	Fisheries Science	142.39	142.39	1!
2	DOResearch and Ed	AgriUnivInst	450	450	48
2	DOResearch and Ed	EconSatatManage	29.76	29.76	[
+	+	<u> </u>	+	 	·
only show	ing ton 20 nows				

only showing top 20 rows

from pyspark.ml.feature import VectorAssembler, StringIndexer

CSHm.show(4)



CSHm.select('Department').distinct().show()

```
|DOAgriculture, Co...|
|DOFood Distribution |
   DOSchool Education
           Department |
   DO Petrochemicals |
|Ministry of Defen...|
|Ministry of Corpo...|
    Ministry of Coal |
            DOTelecom |
 DOConsumer Affairs |
Department for pr...
|Ministry of Devel...|
             DOPosts |
|Ministry of Civil...|
|Ministry of Earth...|
DOResearch and Ed...
        DOFertilisers|
             DOPharma|
only showing top 20 rows
```

DepartmentEncoder=StringIndexer(inputCol="Department", outputCol="Dpt").fit(CSHm)
CSHm=DepartmentEncoder.transform(CSHm)
CSHm.show()

+	+			+	+	+	
Catego	ry	Depa	rtment	Scheme	revenue20_21	actuals20_21	revenue_est21
Catego	 ry	Depa	rtment	Scheme	revenue20_21	actuals20_21	r r
	1	DOAgriculture,	Co	Total	81342.71	81342.71	11649
	1	DOAgriculture,	Co	CIS	12639.22	12639.22	1569
	1	DOAgriculture,	Co	ISSTCF	16218.74	16218.74	2117
	1	DOAgriculture,	Co	MIS-PSS	2004.6	2004.6	206
	1	DOAgriculture,	Co	PM-AASHA	313.18	313.18	56
	1	DOAgriculture,	Co	DPS/UTWS	733.9	733.9	86
		DOAgriculture,		•	48713.84	48713.84	7500
	1 l	N∩∆oriculture	$C \circ$	I PMKMDY	125	125	22
			×	Total	2488.61	2488.61	272
				AE	221.15	221.15	24
	۷	DOVEZEULCII GIIC	ı Eu	I AgrEng	64.02	64.02]
	2	DOResearch and	l Ed	CRAI	46	46	[
	2	DOResearch and	l Ed	Crop Science	635	635	71
	2	DOResearch and	l Ed	HorSci	173.61	173.61	19
	2	DOResearch and	l Ed	NationAgrSciFund	50	50	
•	2	DOResearch and	l Ed	Animal Science	310	310	3:
	2	DOResearch and	l Ed	Fisheries Science	142.39	142.39	1!
	2	DOResearch and	l Ed	AgriUnivInst	450	450	48
	2	DOResearch and	l Ed	EconSatatManage	29.76	29.76	[

DepartmentEncoder.labels

```
['DOCommerce',
С→
     'DOFood Distribution ',
     'DOResearch and Education',
     'DOConsumer Affairs ',
     'DOHigher Education ',
     'Ministry of Develope of North Region',
     'DOAgriculture, Cooperation',
     'DOTelecom',
     'Ministry of Culture',
     'Ministry of Earth Sciences ',
     'DOPharma',
     'Department for promotion of Industry',
     'DOSchool Education',
     'Ministry of Civil Aviation ',
     'Ministry of Coal ',
     'DO Petrochemicals ',
     'DOFertilisers',
     'DOPosts ',
     'Ministry of Defence (Civil) ',
     'Ministry of Corporate Affairs ',
     'Department']
```

Saving... ×

0s completed at 9:30 PM

X