In [1]:
```
!pip install pyspark
```

```
Requirement already satisfied: pyspark in c:\users\administrator\anaconda3\lib
\site-packages (3.3.0)
Requirement already satisfied: py4j==0.10.9.5 in c:\users\administrator\anacond
a3\lib\site-packages (from pyspark) (0.10.9.5)
```

In [2]:
```
from pyspark.sql import SparkSession
```

In [3]:
```
spark = SparkSession.builder.appName("pysp").getOrCreate()
```

In [4]:
```
df = spark.read.csv("./dataset/DFFjNE.csv", header=True, inferSchema= True)
```

In [5]:
```
df.columns
```

Out[5]:
```
['empno', 'ename', 'job', 'mgr', 'hiredate', 'sal', 'comm', 'deptno']
```

In [6]:
```
df.printSchema()
```

```
root
 |-- empno: integer (nullable = true)
 |-- ename: string (nullable = true)
 |-- job: string (nullable = true)
 |-- mgr: integer (nullable = true)
 |-- hiredate: string (nullable = true)
 |-- sal: integer (nullable = true)
 |-- comm: integer (nullable = true)
 |-- deptno: integer (nullable = true)
```

In [7]:
```
df.show()
```

```
+-----+------+---------+----+----------+----+----+------+
|empno| ename|      job| mgr|  hiredate| sal|comm|deptno|
+-----+------+---------+----+----------+----+----+------+
| 7839|  KING|PRESIDENT|null|11/17/2001|5000|null|    10|
| 7698| BLAKE|  MANAGER|7839|05/01/2001|2850|null|    30|
| 7782| CLARK|  MANAGER|7839|06/09/2001|2450|null|    10|
| 7566| JONES|  MANAGER|7839|04/02/2001|2975|null|    20|
| 7654|MARTIN| SALESMAN|7698|09/28/2001|1250|1400|    30|
| 7499| ALLEN| SALESMAN|7698|02/20/2001|1600| 300|    30|
| 7844|TURNER| SALESMAN|7698|09/08/2001|1500|   0|    30|
| 7900| JAMES|    CLERK|7698|12/03/2001| 950|null|    30|
| 7521|  WARD| SALESMAN|7698|02/22/2001|1250| 500|    30|
| 7902|  FORD|  ANALYST|7566|02/03/2001|3000|null|    20|
| 7369| SMITH|    CLERK|7902|12/17/2000| 800|null|    20|
| 7788| SCOTT|  ANALYST|7566|04/19/2007|3000|null|    20|
| 7876| ADAMS|    CLERK|7788|05/23/2007|1100|null|    20|
| 7934|MILLER|    CLERK|7782|01/23/2002|1300|null|    10|
+-----+------+---------+----+----------+----+----+------+
```

In [8]:
```
df.dtypes
```

Out[8]:
```
[('empno', 'int'),
 ('ename', 'string'),
 ('job', 'string'),
 ('mgr', 'int'),
 ('hiredate', 'string'),
 ('sal', 'int'),
 ('comm', 'int'),
 ('deptno', 'int')]
```

In [9]:
```
from pyspark.sql.types import *
from pyspark.sql.functions import to_date, unix_timestamp, from_unixtime
```

In [10]:
```
df = df.withColumn("empno",df.empno.cast("integer"))\
.withColumn("mgr",df.mgr.cast("integer"))\
.withColumn("sal",df.sal.cast("float"))\
.withColumn("deptno",df.deptno.cast("integer"))
```

In [41]:
```
df = df.withColumn('hiredate',to_date(unix_timestamp(df.hiredate,'MM/dd/yyy').cas
```

In [42]:
```
df.dtypes
```

Out[42]:
```
[('empno', 'int'),
 ('ename', 'string'),
 ('job', 'string'),
 ('mgr', 'int'),
 ('hiredate', 'date'),
 ('sal', 'float'),
 ('comm', 'int'),
 ('deptno', 'int')]
```

In [43]:
```
df.show()
```

```
+-----+------+---------+----+----------+------+----+------+
|empno| ename|      job| mgr|  hiredate|   sal|comm|deptno|
+-----+------+---------+----+----------+------+----+------+
| 7839|  KING|PRESIDENT|null|2001-11-17|5000.0|null|    10|
| 7698| BLAKE|  MANAGER|7839|2001-05-01|2850.0|null|    30|
| 7782| CLARK|  MANAGER|7839|2001-06-09|2450.0|null|    10|
| 7566| JONES|  MANAGER|7839|2001-04-02|2975.0|null|    20|
| 7654|MARTIN| SALESMAN|7698|2001-09-28|1250.0|1400|    30|
| 7499| ALLEN| SALESMAN|7698|2001-02-20|1600.0| 300|    30|
| 7844|TURNER| SALESMAN|7698|2001-09-08|1500.0|   0|    30|
| 7900| JAMES|    CLERK|7698|2001-12-03| 950.0|null|    30|
| 7521|  WARD| SALESMAN|7698|2001-02-22|1250.0| 500|    30|
| 7902|  FORD|  ANALYST|7566|2001-02-03|3000.0|null|    20|
| 7369| SMITH|    CLERK|7902|2000-12-17| 800.0|null|    20|
| 7788| SCOTT|  ANALYST|7566|2007-04-19|3000.0|null|    20|
| 7876| ADAMS|    CLERK|7788|2007-05-23|1100.0|null|    20|
| 7934|MILLER|    CLERK|7782|2002-01-23|1300.0|null|    10|
+-----+------+---------+----+----------+------+----+------+
```

In [44]:
```
df.select("ename","sal").show(5)
```

```
+------+------+
| ename|   sal|
+------+------+
|  KING|5000.0|
| BLAKE|2850.0|
| CLARK|2450.0|
| JONES|2975.0|
|MARTIN|1250.0|
+------+------+
only showing top 5 rows
```

In [45]:
```
df1 = df.select("ename","sal")


df1.show(5)
```

```
+------+------+
| ename|   sal|
+------+------+
|  KING|5000.0|
| BLAKE|2850.0|
| CLARK|2450.0|
| JONES|2975.0|
|MARTIN|1250.0|
+------+------+
only showing top 5 rows
```

In [46]:
```
df.select("ename","sal").describe().show()
```

```
+-------+-----+-----------------+
|summary|ename|              sal|
+-------+-----+-----------------+
|  count|   14|               14|
|   mean| null|2073.214285714286|
| stddev| null|1182.5032235162716|
|    min|ADAMS|            800.0|
|    max| WARD|           5000.0|
+-------+-----+-----------------+
```

In [47]:
```
df.select("job").distinct().show()
```

```
+---------+
|      job|
+---------+
|  ANALYST|
| SALESMAN|
|    CLERK|
|  MANAGER|
|PRESIDENT|
+---------+
```

In [48]:
```
df.select("job").dropDuplicates().show()
```

```
+---------+
|      job|
+---------+
|  ANALYST|
| SALESMAN|
|    CLERK|
|  MANAGER|
|PRESIDENT|
+---------+
```

In [49]:
```
df2 = df.dropna(how="any", subset = ["comm","mgr"])
print(df2.count())
print(df.count())
```

```
4
14
```

In [50]:
```
df2 = df.dropna(how="all", subset = ["comm","mgr"])
print(df2.count())
print(df.count())
```

```
13
14
```

In [51]:
```
df.show()
```

```
+-----+------+---------+----+----------+------+----+------+
|empno| ename|      job| mgr|  hiredate|   sal|comm|deptno|
+-----+------+---------+----+----------+------+----+------+
| 7839|  KING|PRESIDENT|null|2001-11-17|5000.0|null|    10|
| 7698| BLAKE|  MANAGER|7839|2001-05-01|2850.0|null|    30|
| 7782| CLARK|  MANAGER|7839|2001-06-09|2450.0|null|    10|
| 7566| JONES|  MANAGER|7839|2001-04-02|2975.0|null|    20|
| 7654|MARTIN| SALESMAN|7698|2001-09-28|1250.0|1400|    30|
| 7499| ALLEN| SALESMAN|7698|2001-02-20|1600.0| 300|    30|
| 7844|TURNER| SALESMAN|7698|2001-09-08|1500.0|   0|    30|
| 7900| JAMES|    CLERK|7698|2001-12-03| 950.0|null|    30|
| 7521|  WARD| SALESMAN|7698|2001-02-22|1250.0| 500|    30|
| 7902|  FORD|  ANALYST|7566|2001-02-03|3000.0|null|    20|
| 7369| SMITH|    CLERK|7902|2000-12-17| 800.0|null|    20|
| 7788| SCOTT|  ANALYST|7566|2007-04-19|3000.0|null|    20|
| 7876| ADAMS|    CLERK|7788|2007-05-23|1100.0|null|    20|
| 7934|MILLER|    CLERK|7782|2002-01-23|1300.0|null|    10|
+-----+------+---------+----+----------+------+----+------+
```

In [29]:
```
df2.show()
```

```
+-----+------+--------+----+----------+------+----+------+
|empno| ename|     job| mgr|  hiredate|   sal|comm|deptno|
+-----+------+--------+----+----------+------+----+------+
| 7698| BLAKE| MANAGER|7839|05/01/2001|2850.0|null|    30|
| 7782| CLARK| MANAGER|7839|06/09/2001|2450.0|null|    10|
| 7566| JONES| MANAGER|7839|04/02/2001|2975.0|null|    20|
| 7654|MARTIN|SALESMAN|7698|09/28/2001|1250.0|1400|    30|
| 7499| ALLEN|SALESMAN|7698|02/20/2001|1600.0| 300|    30|
| 7844|TURNER|SALESMAN|7698|09/08/2001|1500.0|   0|    30|
| 7900| JAMES|   CLERK|7698|12/03/2001| 950.0|null|    30|
| 7521|  WARD|SALESMAN|7698|02/22/2001|1250.0| 500|    30|
| 7902|  FORD| ANALYST|7566|02/03/2001|3000.0|null|    20|
| 7369| SMITH|   CLERK|7902|12/17/2000| 800.0|null|    20|
| 7788| SCOTT| ANALYST|7566|04/19/2007|3000.0|null|    20|
| 7876| ADAMS|   CLERK|7788|05/23/2007|1100.0|null|    20|
| 7934|MILLER|   CLERK|7782|01/23/2002|1300.0|null|    10|
+-----+------+--------+----+----------+------+----+------+
```

In [30]:
```
df2 = df.fillna({"comm":0, "mgr":0})
```

In [31]:
```
df2.show()
```

```
+-----+------+---------+----+----------+------+----+------+
|empno| ename|      job| mgr|  hiredate|   sal|comm|deptno|
+-----+------+---------+----+----------+------+----+------+
| 7839|  KING|PRESIDENT|   0|11/17/2001|5000.0|   0|    10|
| 7698| BLAKE|  MANAGER|7839|05/01/2001|2850.0|   0|    30|
| 7782| CLARK|  MANAGER|7839|06/09/2001|2450.0|   0|    10|
| 7566| JONES|  MANAGER|7839|04/02/2001|2975.0|   0|    20|
| 7654|MARTIN| SALESMAN|7698|09/28/2001|1250.0|1400|    30|
| 7499| ALLEN| SALESMAN|7698|02/20/2001|1600.0| 300|    30|
| 7844|TURNER| SALESMAN|7698|09/08/2001|1500.0|   0|    30|
| 7900| JAMES|    CLERK|7698|12/03/2001| 950.0|   0|    30|
| 7521|  WARD| SALESMAN|7698|02/22/2001|1250.0| 500|    30|
| 7902|  FORD|  ANALYST|7566|02/03/2001|3000.0|   0|    20|
| 7369| SMITH|    CLERK|7902|12/17/2000| 800.0|   0|    20|
| 7788| SCOTT|  ANALYST|7566|04/19/2007|3000.0|   0|    20|
| 7876| ADAMS|    CLERK|7788|05/23/2007|1100.0|   0|    20|
| 7934|MILLER|    CLERK|7782|01/23/2002|1300.0|   0|    10|
+-----+------+---------+----+----------+------+----+------+
```

In [34]:
```
from pyspark.sql.functions import mean
```

In [35]:
```
avg = df.select(mean(df.comm)).collect()[0][0]
print(avg)
```

```
550.0
```

In [36]:
```
df2 = df.fillna({"comm":avg})
df2.show()
```

```
+-----+------+---------+----+----------+------+----+------+
|empno| ename|      job| mgr|  hiredate|   sal|comm|deptno|
+-----+------+---------+----+----------+------+----+------+
| 7839|  KING|PRESIDENT|null|11/17/2001|5000.0| 550|    10|
| 7698| BLAKE|  MANAGER|7839|05/01/2001|2850.0| 550|    30|
| 7782| CLARK|  MANAGER|7839|06/09/2001|2450.0| 550|    10|
| 7566| JONES|  MANAGER|7839|04/02/2001|2975.0| 550|    20|
| 7654|MARTIN| SALESMAN|7698|09/28/2001|1250.0|1400|    30|
| 7499| ALLEN| SALESMAN|7698|02/20/2001|1600.0| 300|    30|
| 7844|TURNER| SALESMAN|7698|09/08/2001|1500.0|   0|    30|
| 7900| JAMES|    CLERK|7698|12/03/2001| 950.0| 550|    30|
| 7521|  WARD| SALESMAN|7698|02/22/2001|1250.0| 500|    30|
| 7902|  FORD|  ANALYST|7566|02/03/2001|3000.0| 550|    20|
| 7369| SMITH|    CLERK|7902|12/17/2000| 800.0| 550|    20|
| 7788| SCOTT|  ANALYST|7566|04/19/2007|3000.0| 550|    20|
| 7876| ADAMS|    CLERK|7788|05/23/2007|1100.0| 550|    20|
| 7934|MILLER|    CLERK|7782|01/23/2002|1300.0| 550|    10|
+-----+------+---------+----+----------+------+----+------+
```

In [39]:
```
from pyspark.sql.functions import year, month, dayofyear, weekofyear, hour, minut
```

In [52]:
```
df.select(df.hiredate,year(df.hiredate).alias("dt_year"),month(df.hiredate).alias
```

```
+----------+-------+--------+-------------+-----------+-------------+
|  hiredate|dt_year|dt_month|dt_dayofmonth|dt_dayofyear|dt_weekofyear|
+----------+-------+--------+-------------+-----------+-------------+
|2001-11-17|   2001|      11|           17|        321|           46|
|2001-05-01|   2001|       5|            1|        121|           18|
|2001-06-09|   2001|       6|            9|        160|           23|
|2001-04-02|   2001|       4|            2|         92|           14|
|2001-09-28|   2001|       9|           28|        271|           39|
|2001-02-20|   2001|       2|           20|         51|            8|
|2001-09-08|   2001|       9|            8|        251|           36|
|2001-12-03|   2001|      12|            3|        337|           49|
|2001-02-22|   2001|       2|           22|         53|            8|
|2001-02-03|   2001|       2|            3|         34|            5|
|2000-12-17|   2000|      12|           17|        352|           50|
|2007-04-19|   2007|       4|           19|        109|           16|
|2007-05-23|   2007|       5|           23|        143|           21|
|2002-01-23|   2002|       1|           23|         23|            4|
+----------+-------+--------+-------------+-----------+-------------+
```

In [ ]:
```
```