**Assignment-14**

**Topics Covered Today**

* Read text file into Pyspark DataFrame
* GroupBy and Aggregate function
* Using Pivot/ UnPivot
* Handling Missing Values Pyspark
* Sorting
* Joins

**Read text file into Pyspark DataFrame**

**Using csv()**

**A screenshot of a computer

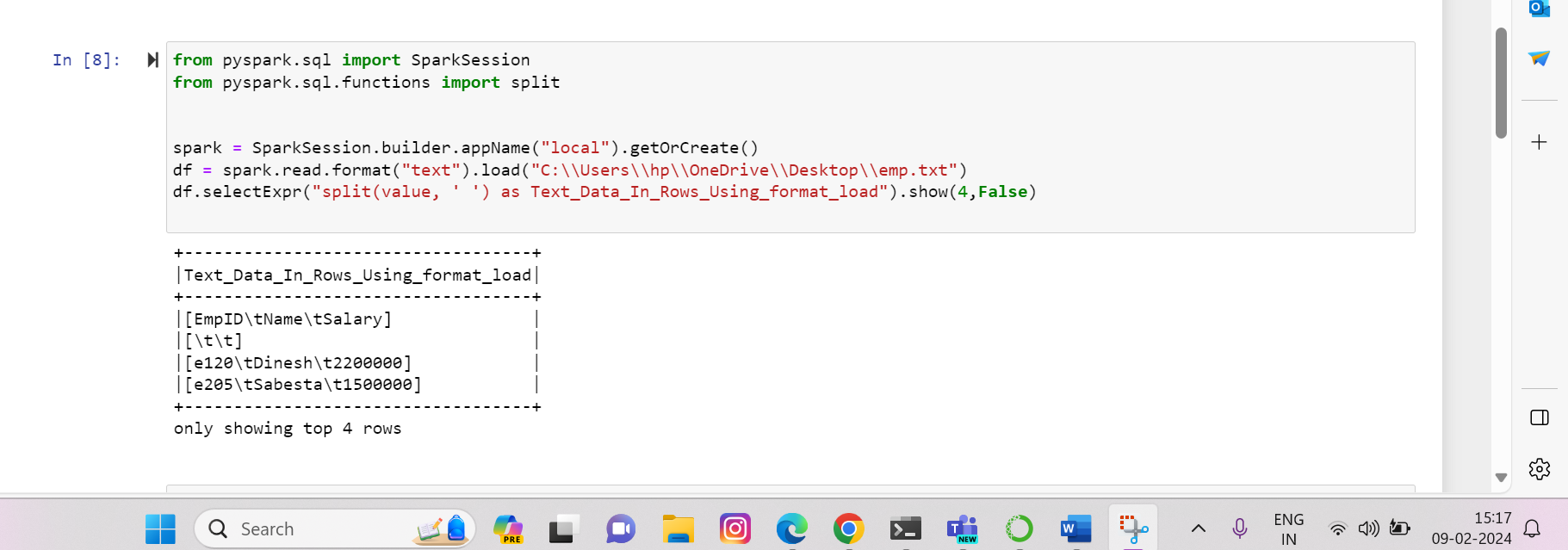
Description automatically generated**

**Using text()**

**A screenshot of a computer

Description automatically generated**

**Using format()**

****

**Adding a new column**

**Using lit()**

**A screenshot of a computer

Description automatically generated**

**From an existing column**

**A screenshot of a computer

Description automatically generated**

**Using concat\_ws()**

**A screenshot of a computer

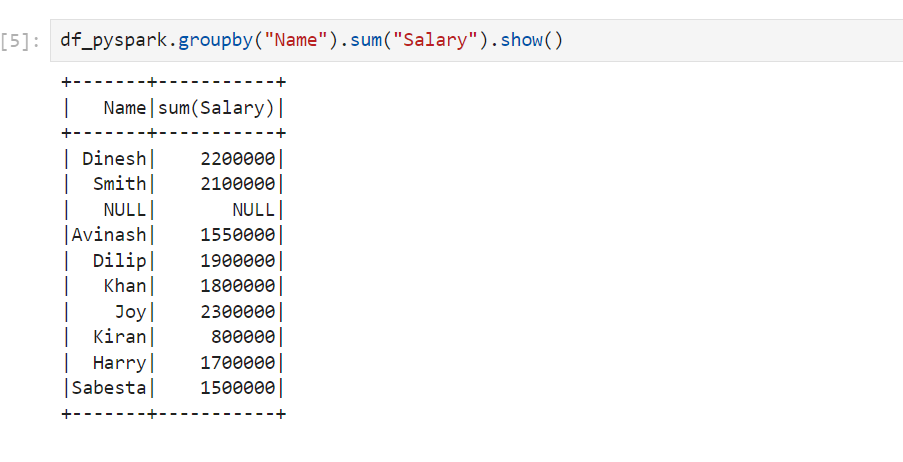
Description automatically generated**

**Using if not in()**

**A screenshot of a computer

Description automatically generated**

**GroupBy and Aggregate functions**

**With sum()**

**With min():A screenshot of a computer

Description automatically generated**

**With max():A screenshot of a computer

Description automatically generated**

**With avg():A screen shot of a computer

Description automatically generated**

**With mean():A screenshot of a computer

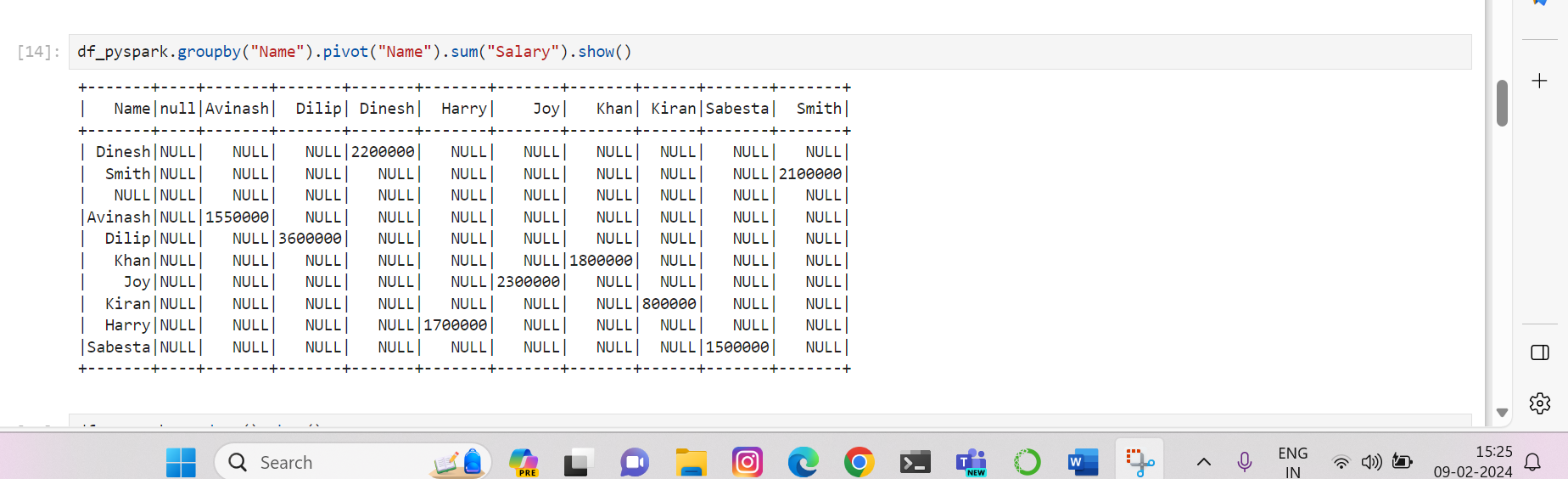
Description automatically generated**

**With count():A screenshot of a computer program

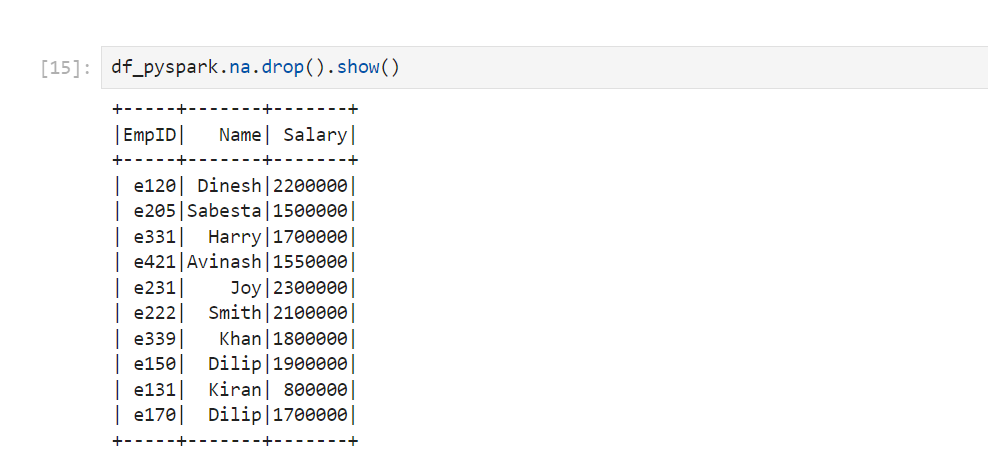
Description automatically generated**

**Using Pivot/ UnPivot**

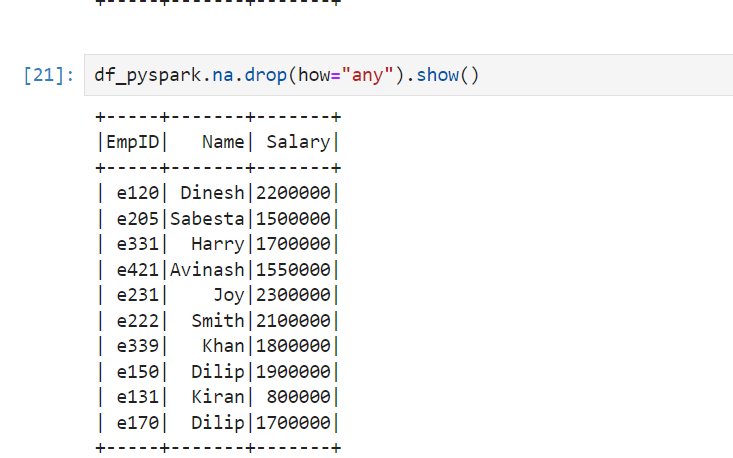
**Spark SQL provides pivot() function to rotate the data from one column into multiple columns (transpose row to column).**

****

**Handling Missing Values Pyspark**

**Dropping rows based on null values:**

**Using how=”any”:**

****

**Using how =”all”: A screenshot of a computer

Description automatically generated**

**Using threshold()**

**A screenshot of a computer

Description automatically generated**

**Using subset():A screenshot of a computer

Description automatically generated**

**Sort() Function:**

**By default the sort function gives us ascending order:A screenshot of a computer

Description automatically generated**

**……..Continued below……..**

**For descending Order:**

**A screenshot of a computer

Description automatically generated**

**Sorting with two Columns:**

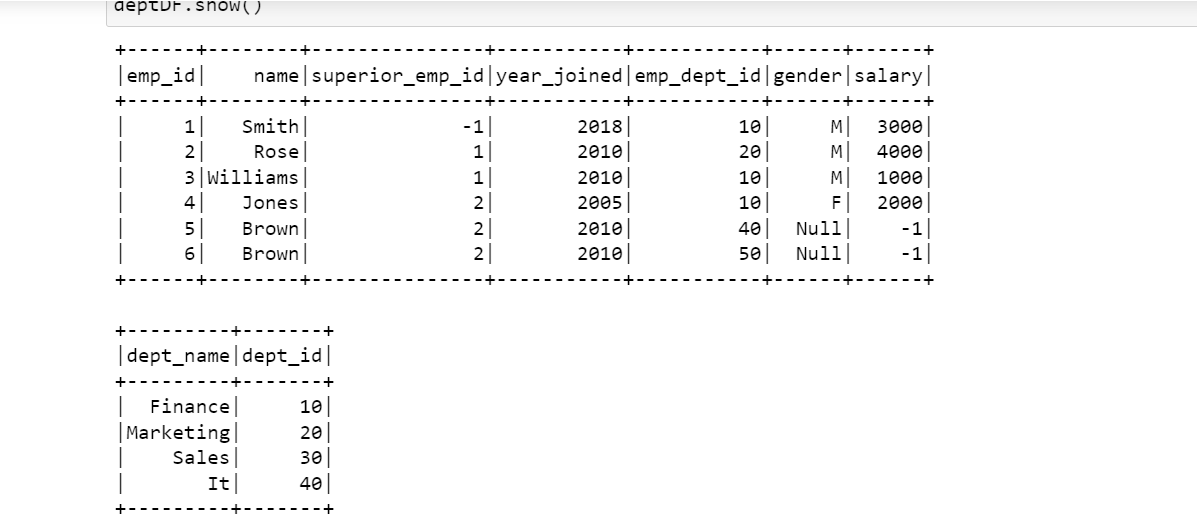
**A screenshot of a computer

Description automatically generated**

**----------Continued Below----------**

**Joins**

**Two datafrmaes on which we will perform joins:**

****

**Inner Join:**

**A screenshot of a computer

Description automatically generated**

**Outer Join:**

**A screenshot of a computer

Description automatically generated**

**Left Join:**

**A screenshot of a computer

Description automatically generated**

**Right Join:**

**A screenshot of a computer code

Description automatically generated**

**LeftSemi Join:**

**A screenshot of a computer

Description automatically generated**

**Leftanti Join:**

**A screenshot of a computer

Description automatically generated**