Project 2:

Data :

Context

Understanding what will be the Burn Rate for the employee working in an organization based on the current pandemic situation where work from home is a boon and a bane. How are employees' Burn Rate affected based on various conditions provided?

Content

Globally, World Mental Health Day is celebrated on October 10 each year. The objective of this day is to raise awareness about mental health issues around the world and mobilize efforts in support of mental health. According to an anonymous survey, about 450 million people live with mental disorders that can be one of the primary causes of poor health and disability worldwide. These days when the world is suffering from a pandemic situation, it becomes really hard to maintain mental fitness.

* Employee ID: The unique ID allocated for each employee (example: fffe390032003000)
* Date of Joining: The date-time when the employee has joined the organization (example: 2008-12-30)
* Gender: The gender of the employee (Male/Female)
* Company Type: The type of company where the employee is working (Service/Product)
* WFH Setup Available: Is the work from home facility available for the employee (Yes/No)
* Designation: The designation of the employee of work in the organization.
  + In the range of [0.0, 5.0] bigger is higher designation.
* Resource Allocation: The amount of resource allocated to the employee to work, ie. number of working hours.
  + In the range of [1.0, 10.0] (higher means more resource)
* Mental Fatigue Score: The level of fatigue mentally the employee is facing.
  + In the range of [0.0, 10.0] where 0.0 means no fatigue and 10.0 means completely fatigue.
* Burn Rate: The value we need to predict for each employee telling the rate of Bur out while working.
  + In the range of [0.0, 1.0] where the higher the value is more is the burn out.

1. Loading file on HDFS
2. Reading file from HDFS to Sparks

>>> df = spark.read.csv('/project\_2/project\_2\_data.csv',inferSchema=True,header=True).toDF('Sr No','Employee ID','Date of Joining','Gender','Company Type','WFH Setup Av

ailable','Designation','Resource Allocation','Mental Fatigue Score','Burn Rate')

Check records in file

>>> df.show(10)

1. Import required modules

>>> from pyspark.sql.types import StructType,StructField,StringType, IntegerType, DateType

>>> from pyspark.sql import Row

1. Create RDD using file

# loading csv file into spark context

>>> dl = sc.textFile('/project\_2/employeeburndataset.csv')

1. Create schema

>>> Columns=StructType([StructField("Employee id",StringType(),False),StructField("Date of joining",DateType(),False), StructField("Gender",StringType(),True), StructFi

eld("Company Type",StringType(),True),StructField("WFH Setup Available",IntegerType(),True), StructField("Designation", IntegerType(),True), StructField("Resource Alloc

ation",IntegerType(),True), StructField("Mental Fatigue Score", IntegerType(),True),StructField("Burn Rate", IntegerType(),True)])

1. Create dataframe

>>> dataframe = Spark.createDataFrame(rdd,Columns)

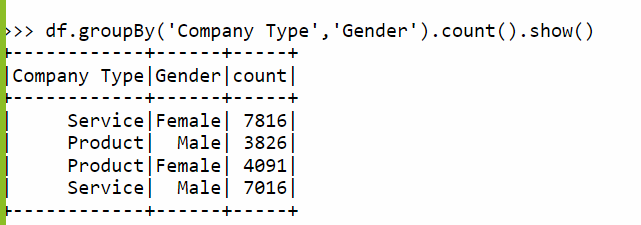
1. Change datatype of columns

>>> from pyspark.sql.functions import col

>>> df.withColumn('df.DOJ',col('DOJ').cast("Date")).show(10)

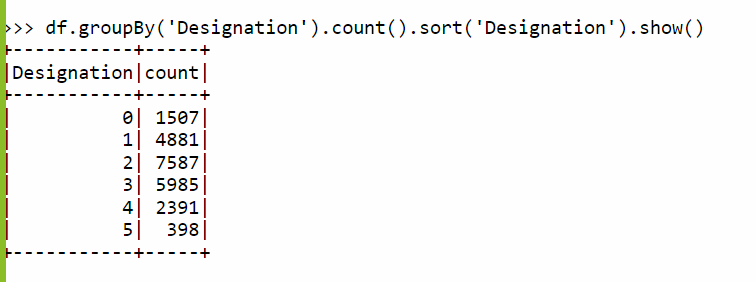
Q1) Find the count of employee based on gender in all company

type.

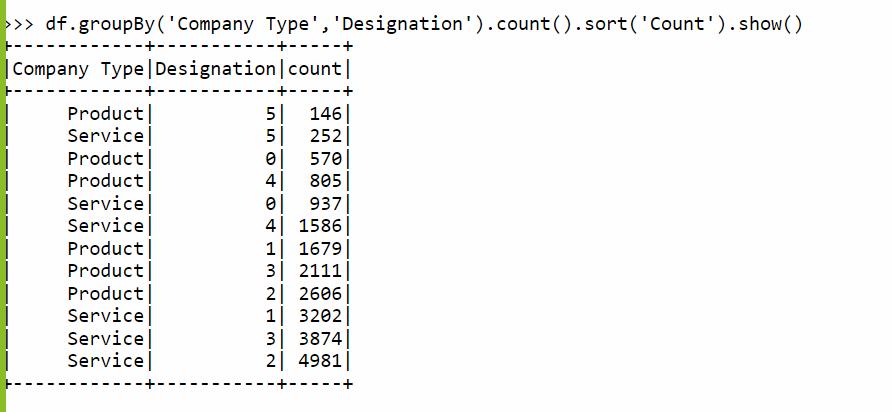


From this we can see that most of employee like to work on Service based company but few people like to work on Product based too.

Q2) Find number of employees working in company based on Designation

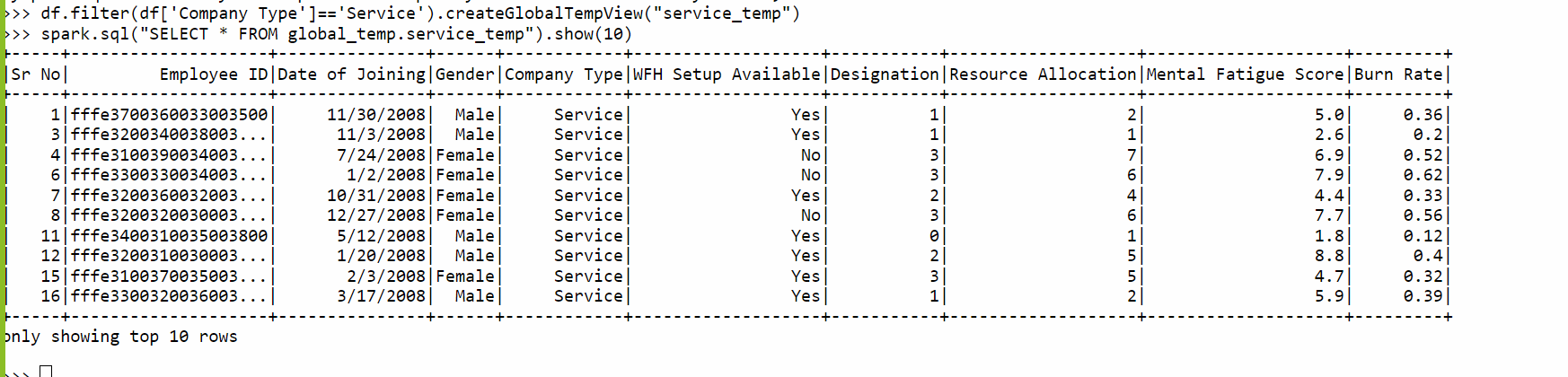


Q3) Find the count of employee based on Designation and Company Type

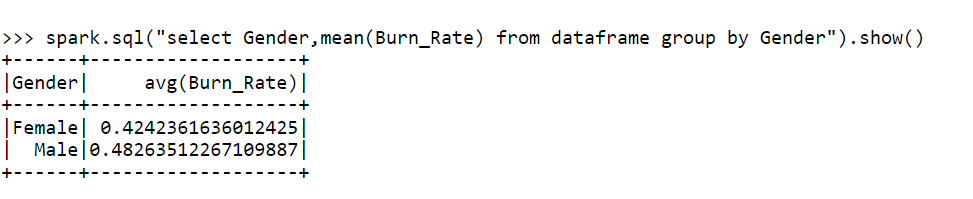


From this we can say that most of employee working to service company in Designation 1 / 2 or 3.

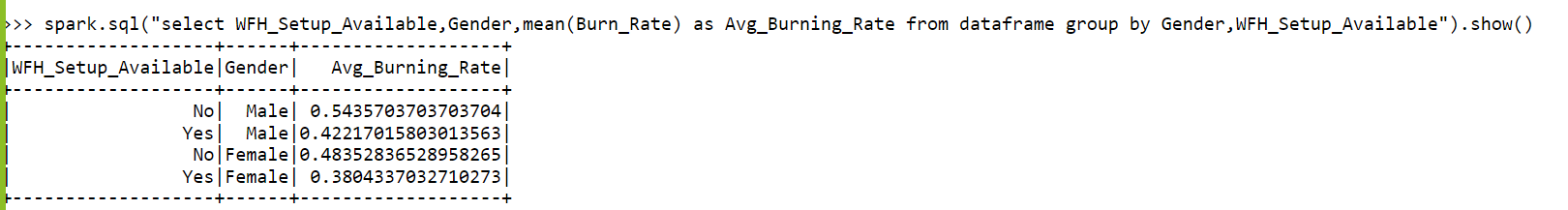
Q4) Creating global view name as ‘service\_temp’ which content only data of service type company



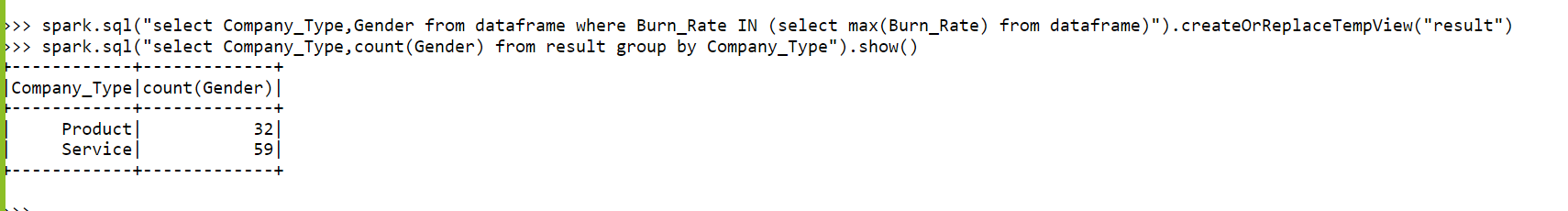
Q5) Find the average of burning rate based on gender



Q6) Find the average of burning rate of employee who didn’t have a WFH setup based on gender

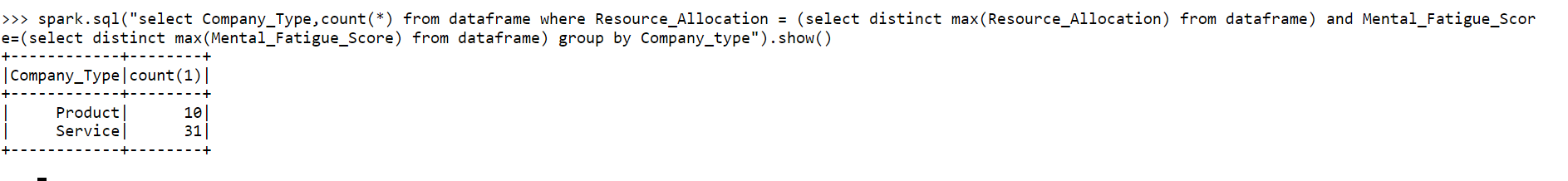


Q7) Find the count of employee who having max burning rate based on company type.



Q8) Total number of employee who having high Resource available but having High mental Fatigue Score

spark.sql("select Company\_Type,count(\*) from dataframe where Resource\_Allocation = (select distinct max(Resource\_Allocation) from dataframe) and Mental\_Fatigue\_Score = (select distinct max(Mental\_Fatigue\_Score) from dataframe) group by Company\_type").show()



Q9) Partition by month stored as parquet file

Q10) Collect random 5000 data and create 2 different temps them perform Joins (python,submit,HDFS)

Q11) Partition by Designation and stored data in to HDFS

Q12) describe().show()

Q13)

Q14)

Q15)

Q16)

Q17)

Q18)

Q19)

Q20)