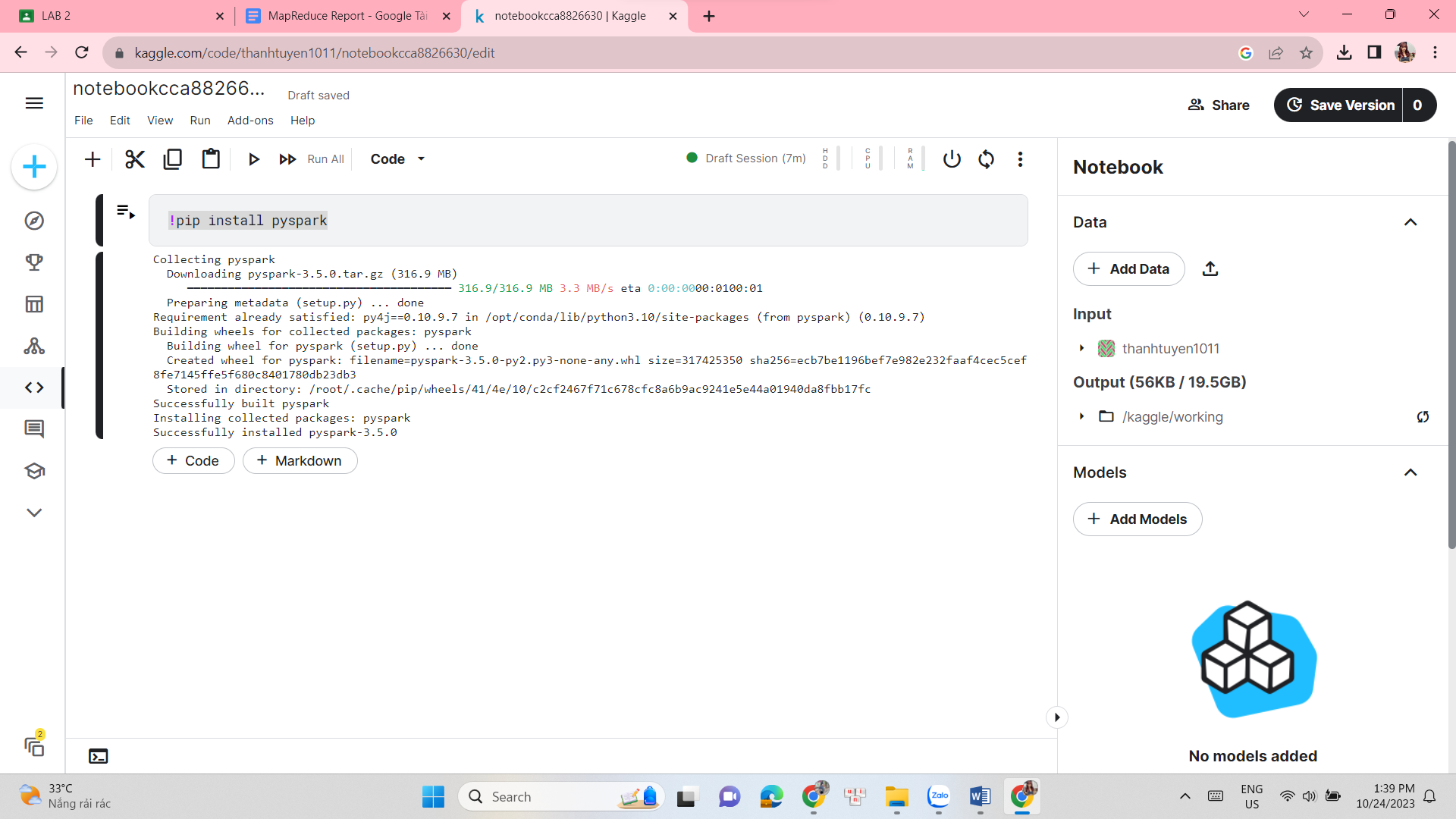
Tên: Bùi Thanh Tuyền

Mssv: 1050080040

!pip install pyspark

#import libraries

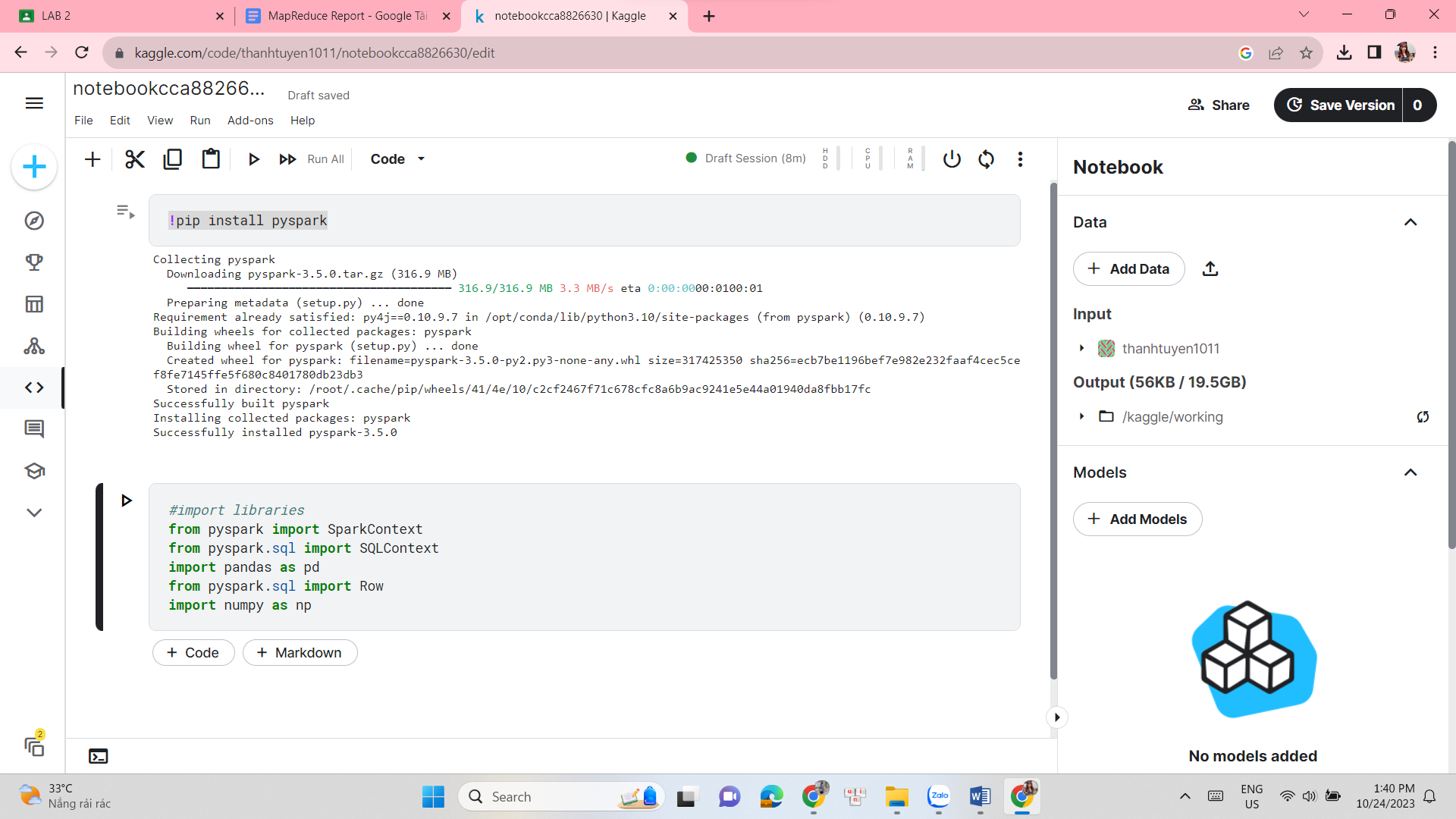
from pyspark import SparkContext

from pyspark.sql import SQLContext

import pandas as pd

from pyspark.sql import Row

import numpy as np



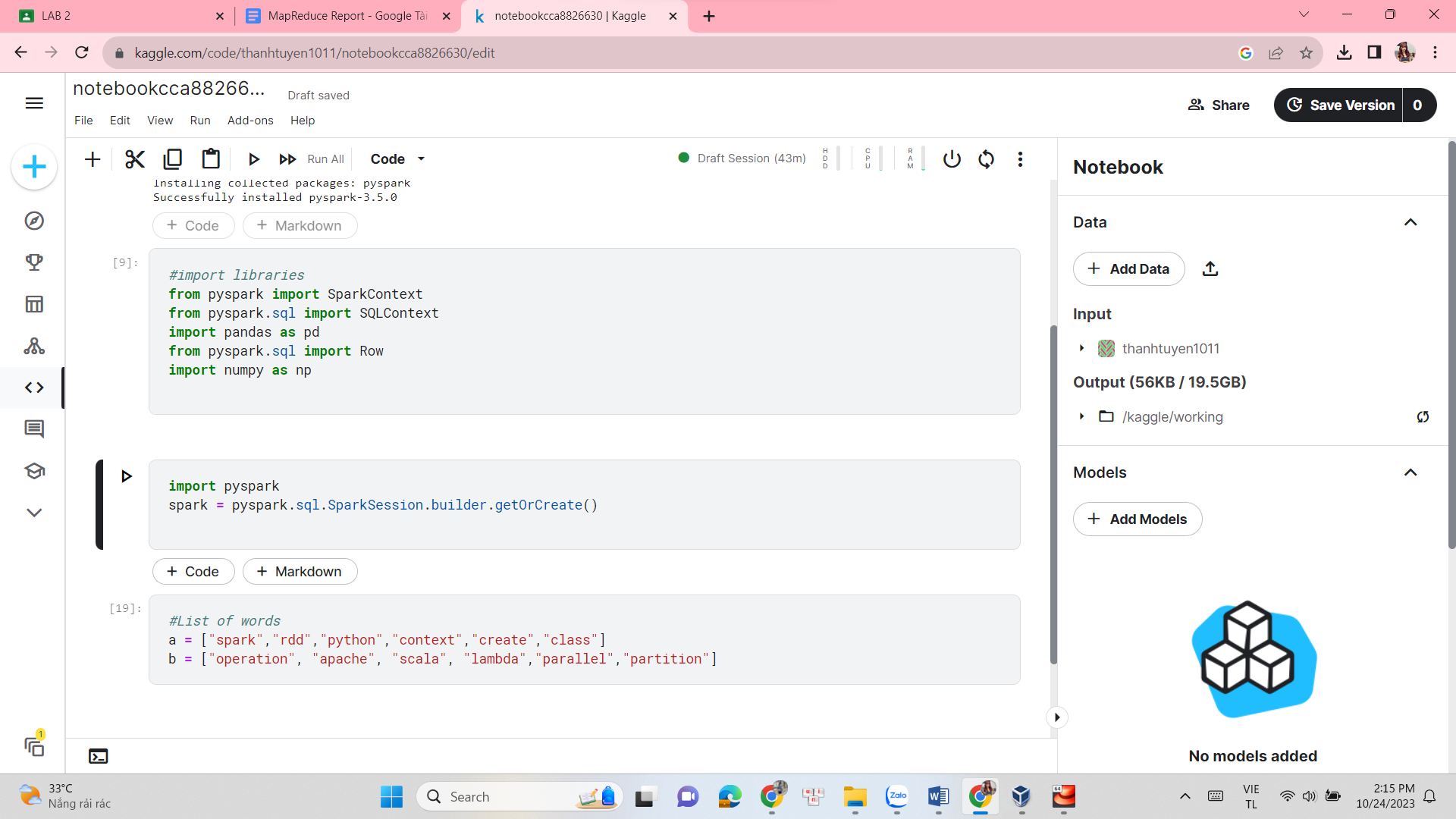
import pyspark

spark = pyspark.sql.SparkSession.builder.getOrCreate()

#List of words

a = ["spark","rdd","python","context","create","class"]

b = ["operation", "apache", "scala", "lambda","parallel","partition"]



#Making RDDs

rdd\_A = sc.parallelize(a)

rdd\_B= sc.parallelize(b)

#Mapping the names of RDD

A = rdd\_A.map(lambda x: Row(name = x))

B = rdd\_B.map(lambda x: Row(name = x))

#Creating dataframe from RDD

dfA = sqlContext.createDataFrame(A)

dfB = sqlContext.createDataFrame(B)

#Creating Alias

df1 = dfA.alias("df1")

df2 = dfB.alias("df2")

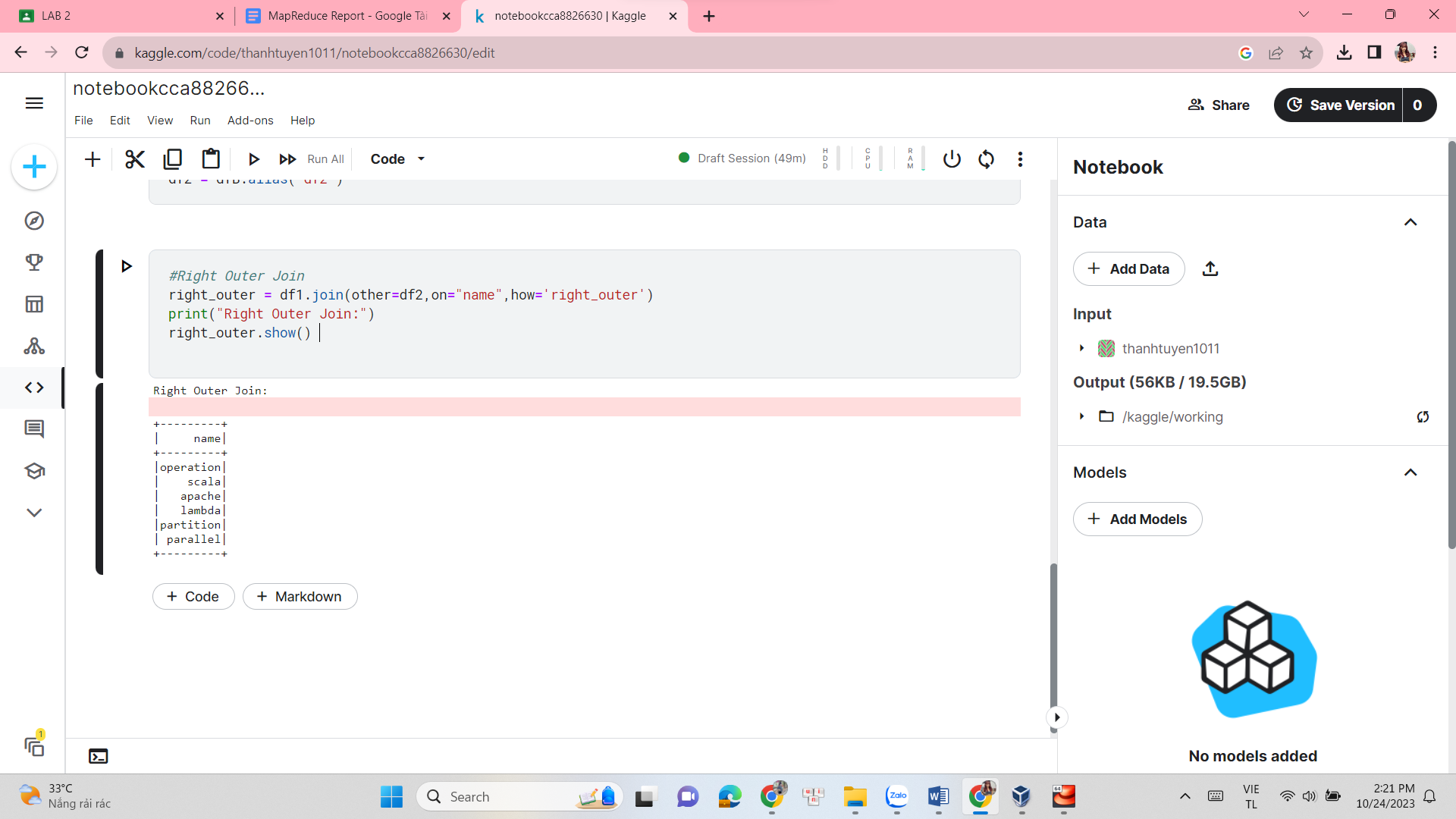


#Right Outer Join

right\_outer = df1.join(other=df2,on="name",how='right\_outer')

print("Right Outer Join:")

right\_outer.show()

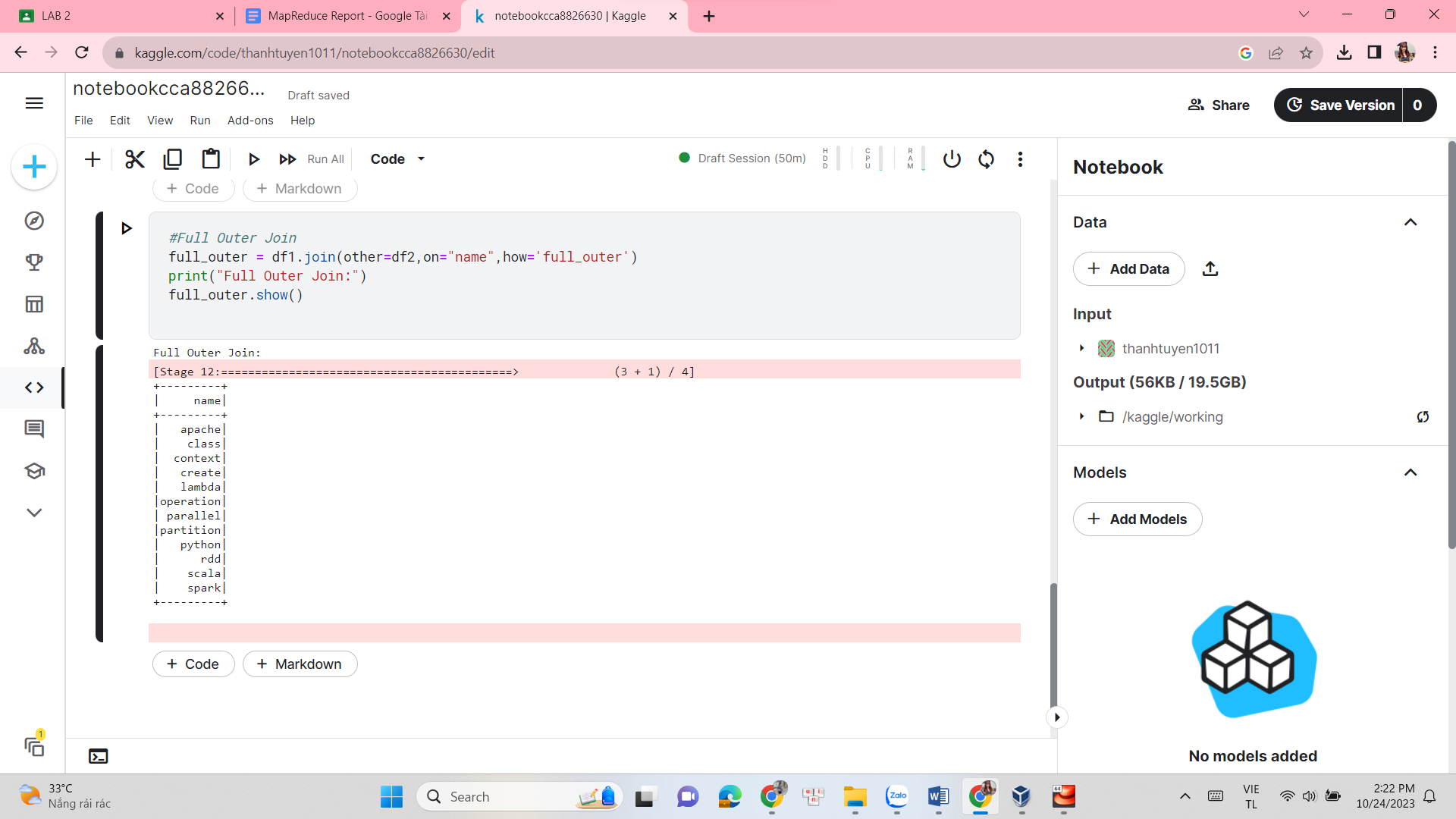


#Full Outer Join

full\_outer = df1.join(other=df2,on="name",how='full\_outer')

print("Full Outer Join:")

full\_outer.show()

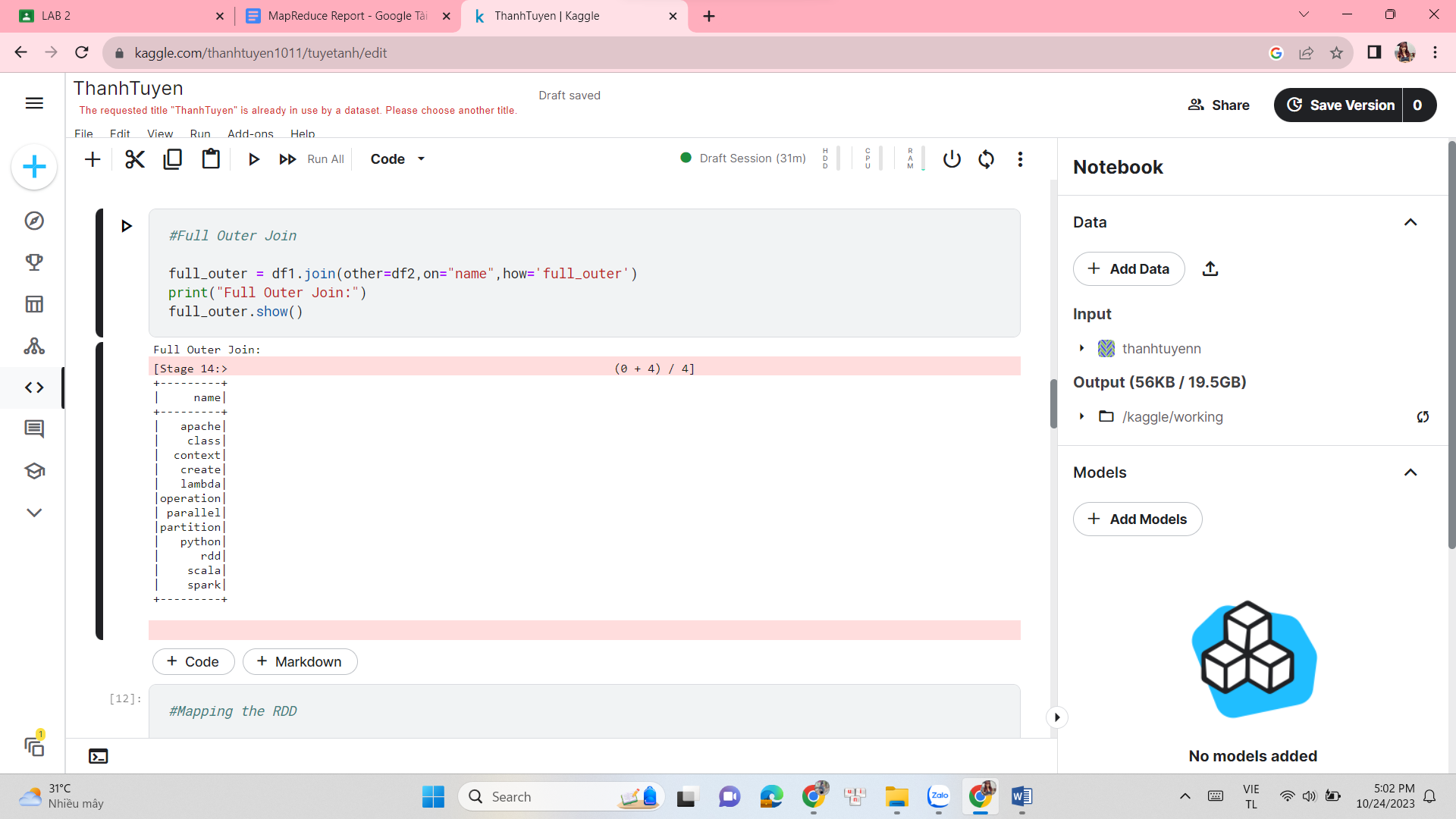


#Full Outer Join

full\_outer = df1.join(other=df2,on="name",how='full\_outer')

print("Full Outer Join:")

full\_outer.show()



#Mapping the RDD

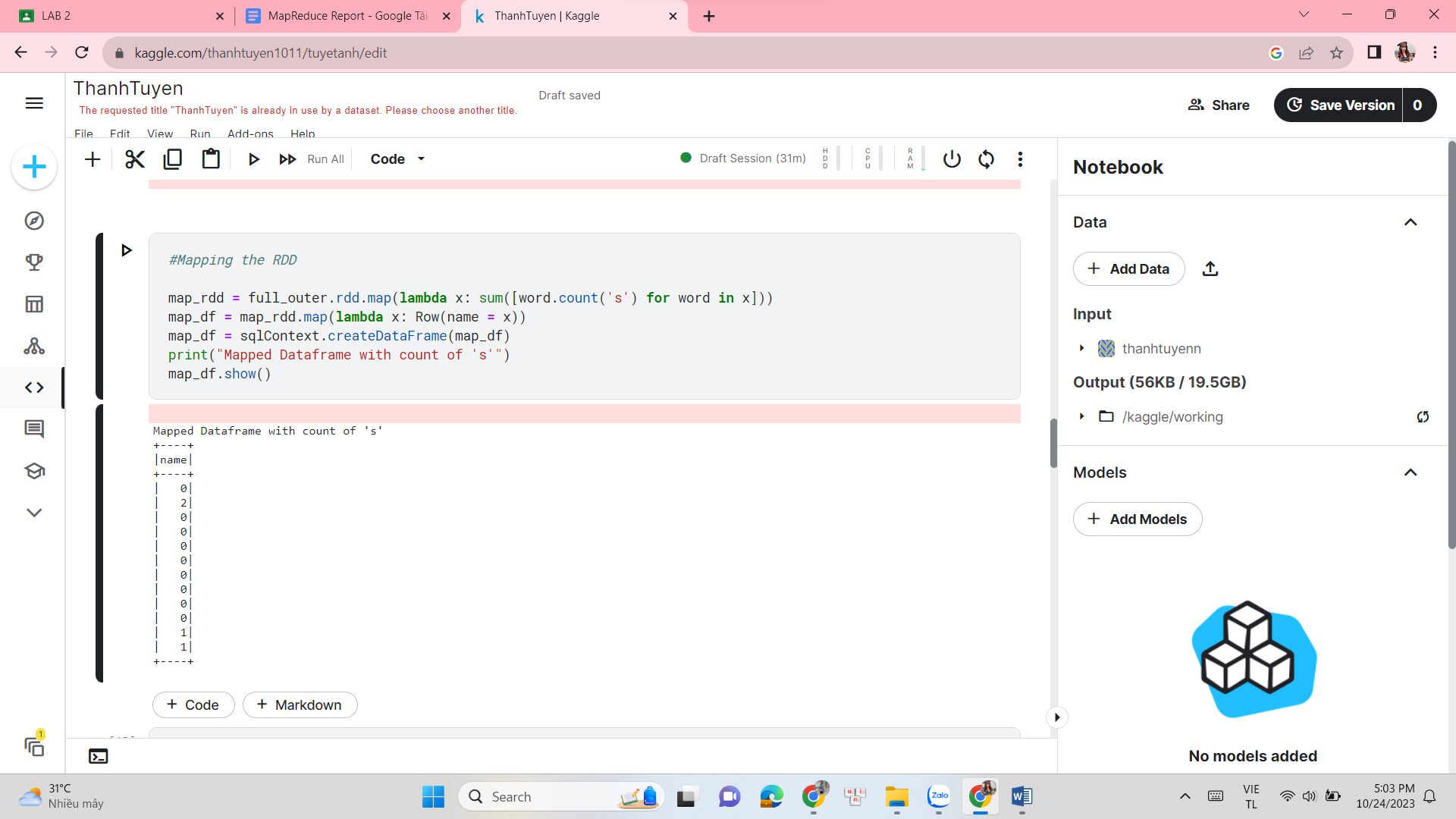
map\_rdd = full\_outer.rdd.map(lambda x: sum([word.count('s') for word in x]))

map\_df = map\_rdd.map(lambda x: Row(name = x))

map\_df = sqlContext.createDataFrame(map\_df)

print("Mapped Dataframe with count of 's'")

map\_df.show()



reduce\_rdd=map\_rdd.reduce(lambda x,y: x+y)

print("Using Map-Reduce, the character \"s\" appears",reduce\_rdd,"times in all a and b.\n" )

#Aggregate function

count = full\_outer.rdd.aggregate(0, lambda i, x: i + x[0].count('s'), lambda i, j: i+j)

print("Using aggregate function, the character \"s\" appears",count, "times in all a and b." )

from pyspark.sql.functions import col

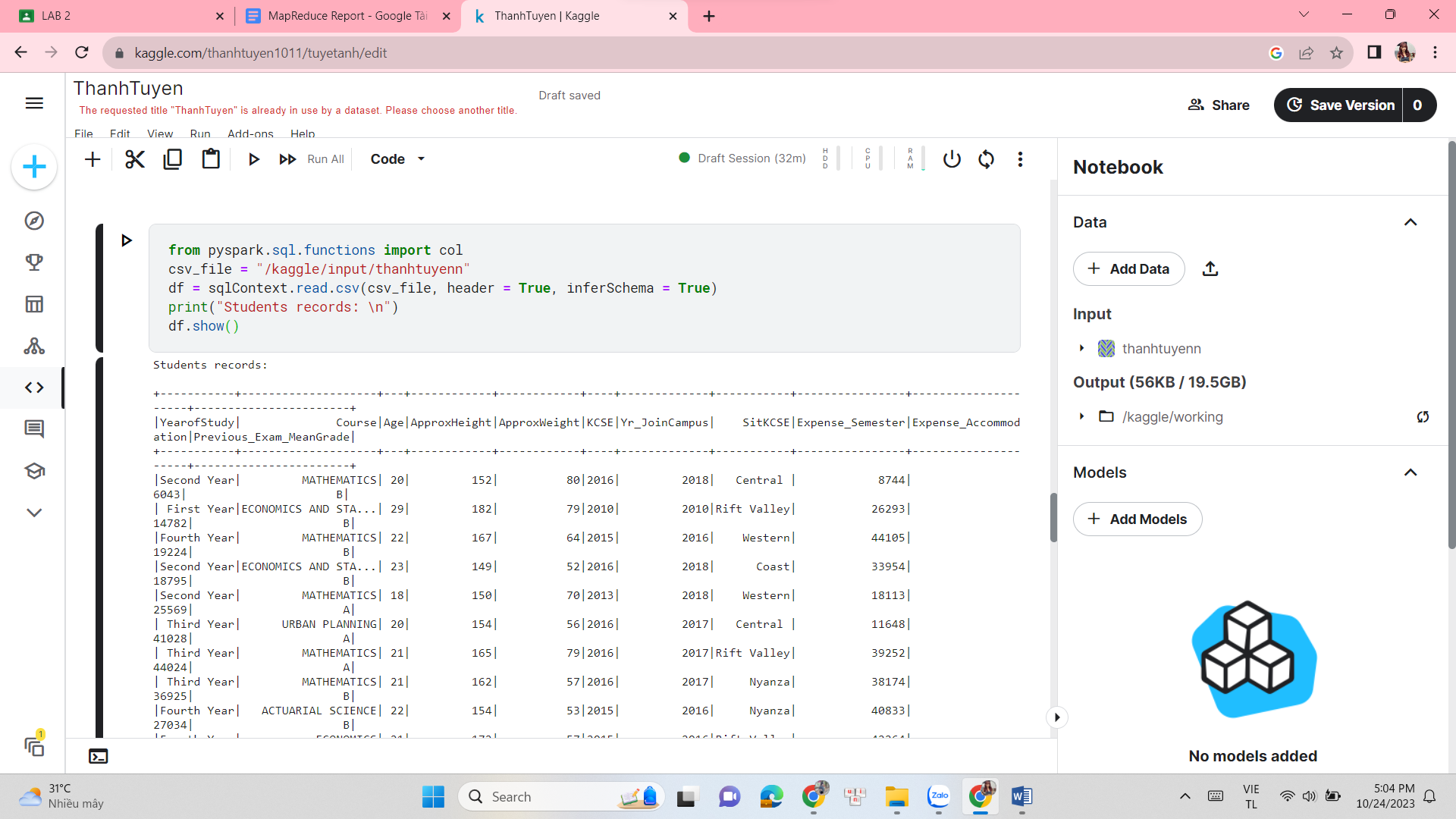
csv\_file = "/kaggle/input/thanhtuyenn"

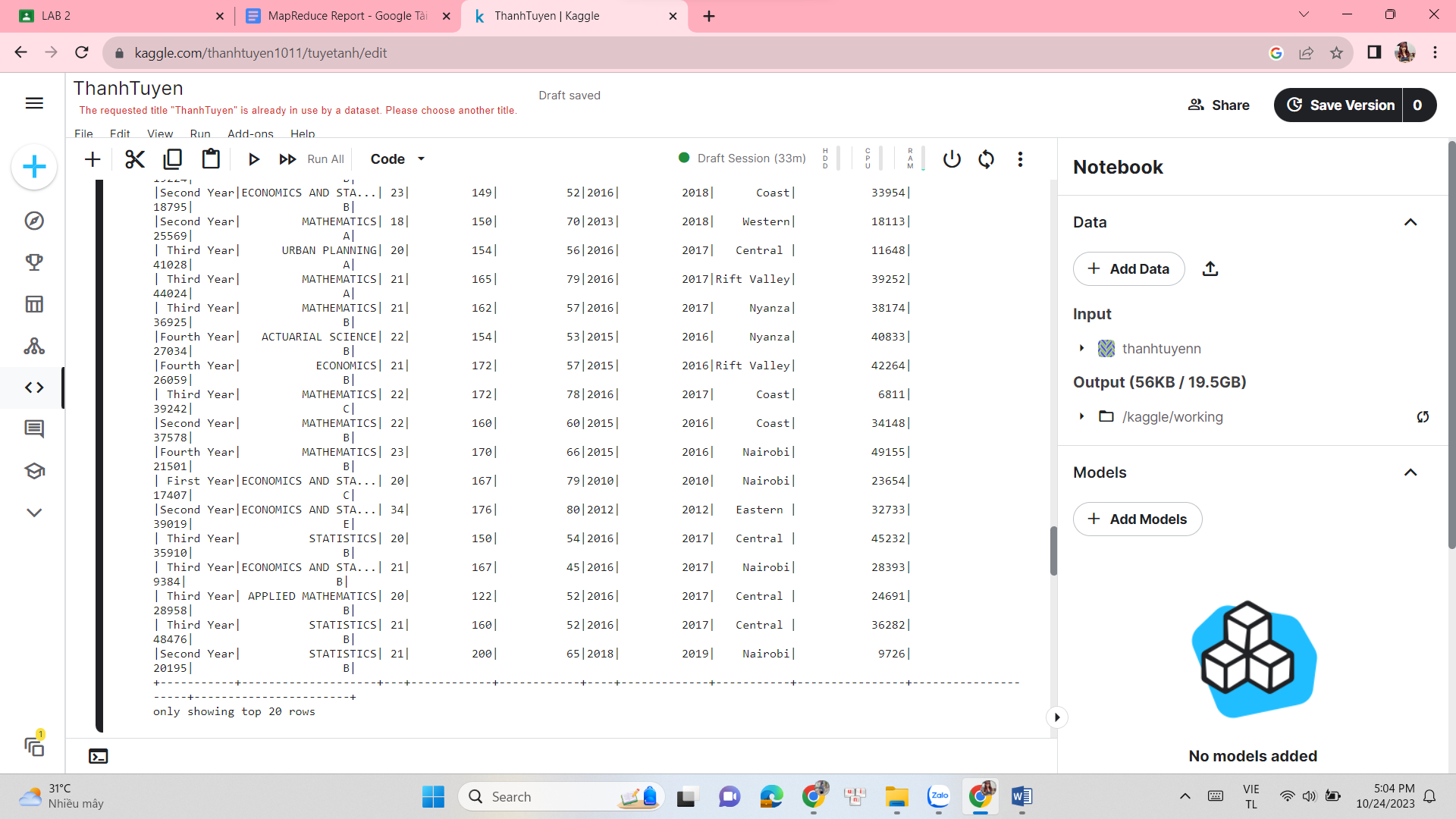
df = sqlContext.read.csv(csv\_file, header = True, inferSchema = True)

print("Students records: \n")

df.show()



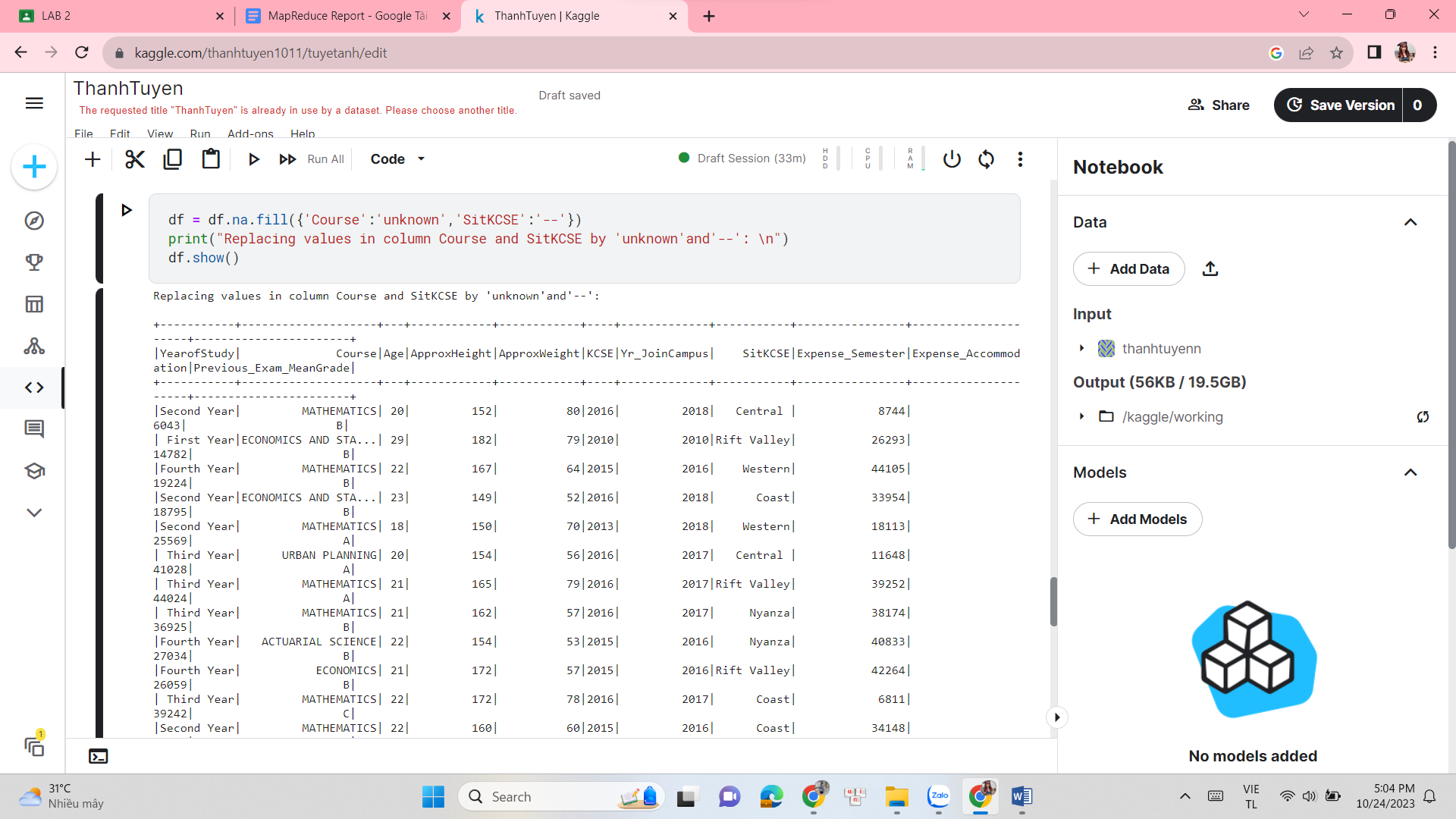


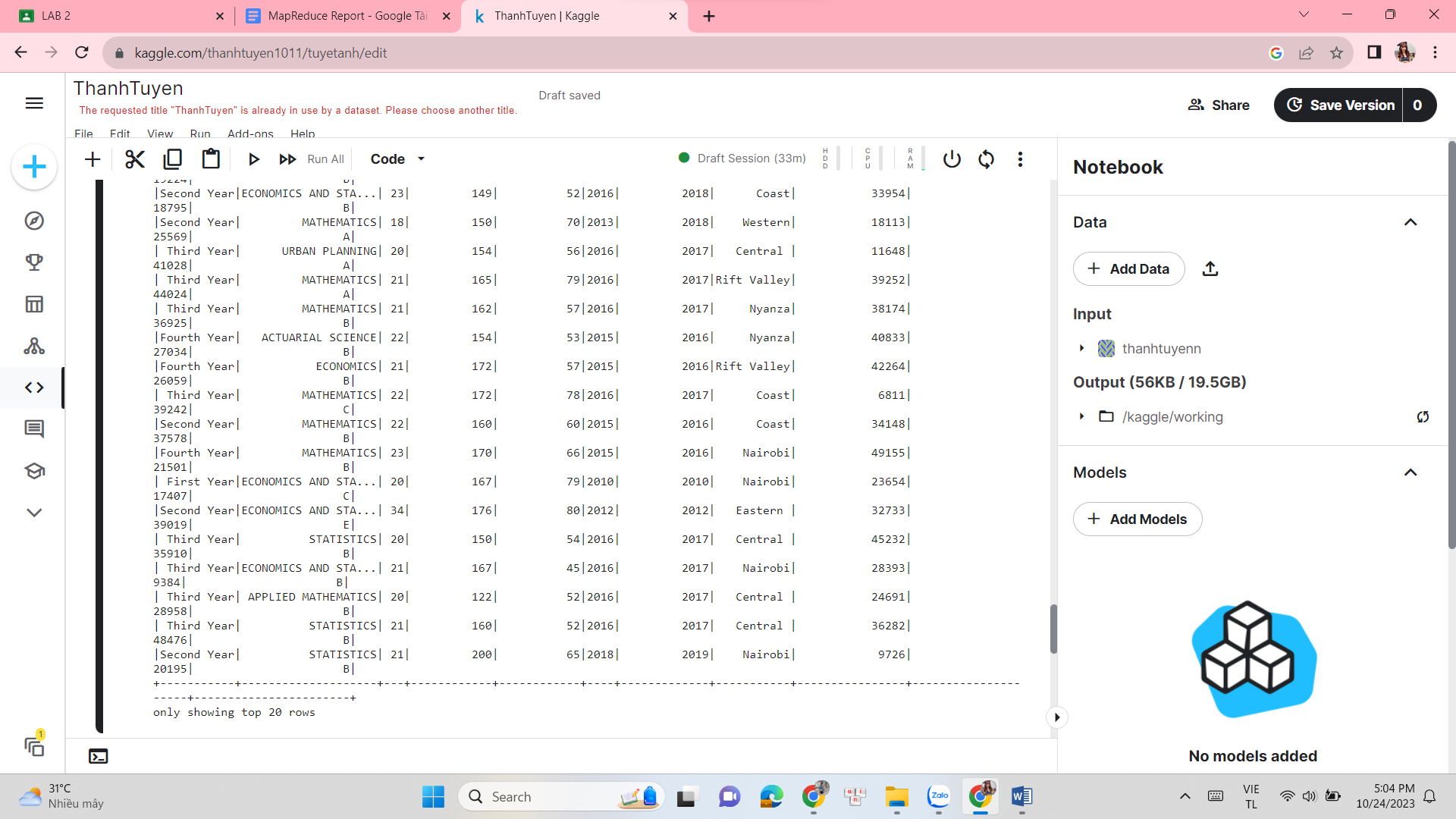


df = df.na.fill({'Course':'unknown','SitKCSE':'--'})

print("Replacing values in column Course and SitKCSE by 'unknown'and'--': \n")

df.show()





from pyspark.sql.functions import mean, col

from pyspark.sql.functions import UserDefinedFunction

from pyspark.sql.types import TimestampType

from dateutil import parser

avg = df.select(mean(col('Age')).alias('mean')).collect()

df = df.na.fill(avg[0]['mean'],subset=['Age'])

df = df.na.fill({'Course': 'Big data','SitKCSE':'--'})

import matplotlib.pyplot as plt

pt = df.toPandas()['Age']

plt.hist(pt,bins=20)

plt.show()

