ADVANCED BIG DATA

SAN FRANCISCO CRIME CLASSIFICATION Using SparkML

A Project Report

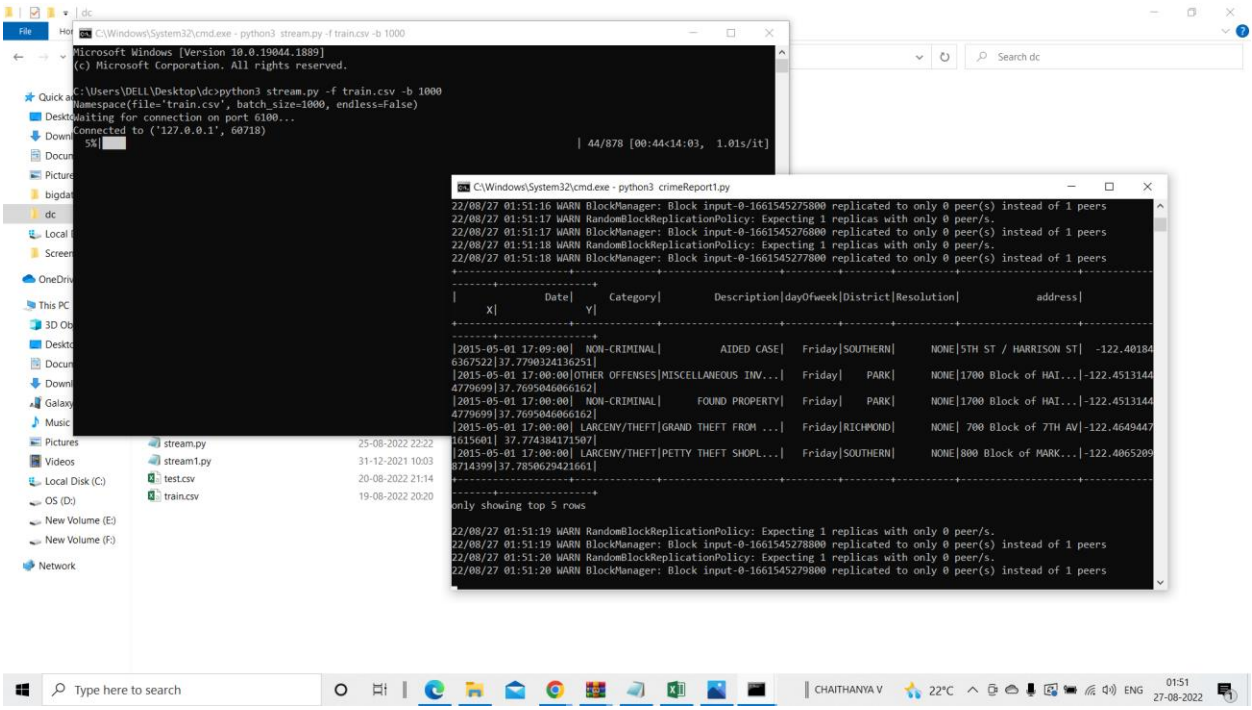
Submitted by

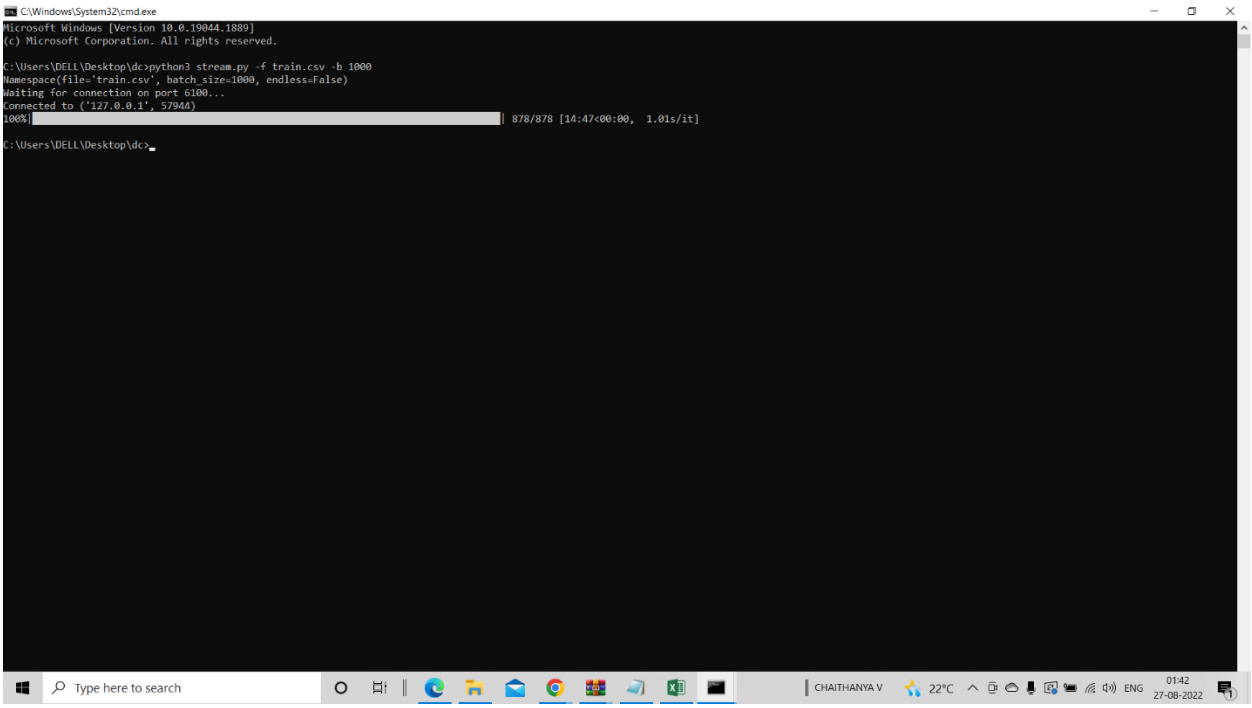
BHOOMIKA S BABU - PES1PG21CS010

CHAITHANYA V - PES1PG21CS011

1. STREAMING DATA:
   Streaming data is data that is generated continuously by thousands of data sources, which is typically send in the data records simultaneously, and in small sizes.

   Converting from RDD to DataFrames:



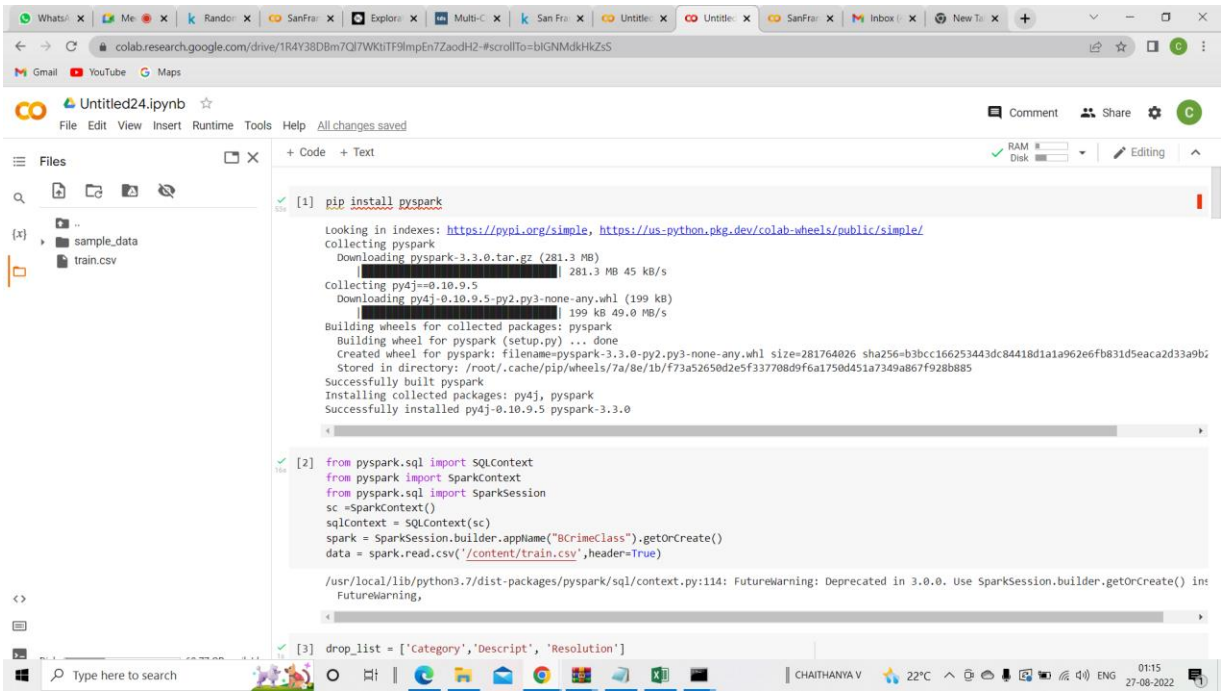


2. Preprocessing Data:
   Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks.
   Step-1: Installing the pyspark
   pip install pyspark
   Step-2: Importing all the modules and loading the datasets into the model
   Importing, → SQLContect – class pyspark.sql.SQLContext(sparkContext, sqlContext=None) Main entry point for Spark SQL functionality. A SQL Context can be used create DateFrame, register DataFrame as tables, execute SQL over tables.

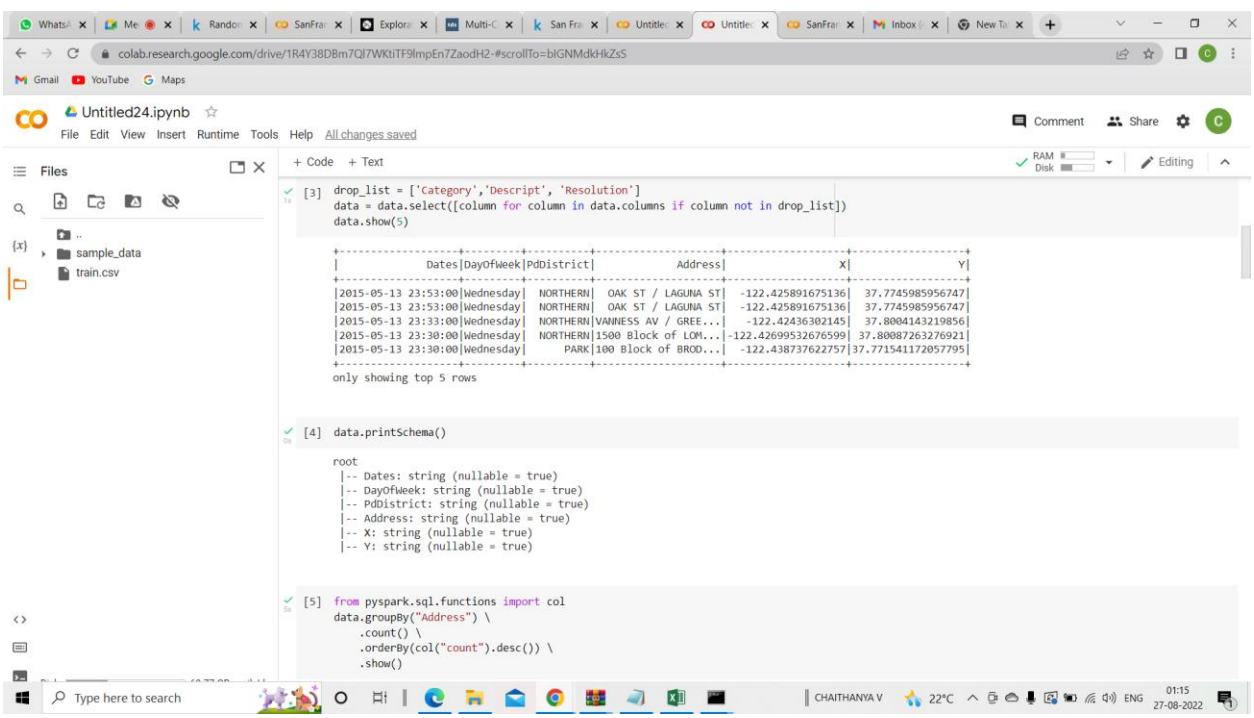
→SparkContext -- A SparkContext represents the connection to a Spark cluster, and can be used to create RDD and broadcast variables on that cluster.

→SparkSessions – SparkSession(sparkContext, jsparkSession=None)[source] The entry point to programming Spark with the DataSet and DataFrame API. A SparkSession can be used create DataFrame, register DataFrame as tables, execute as tables, execute SQL over tables.

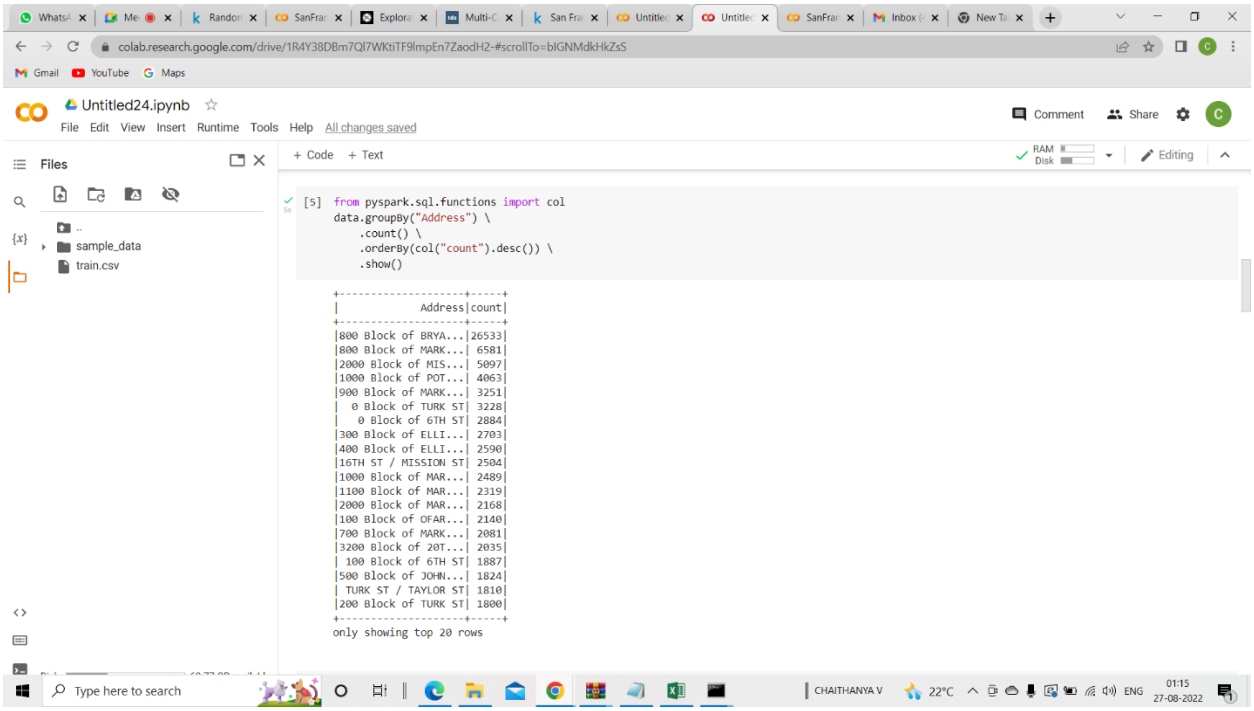
Loading the DataSet(csv file) → spark.read.csv('/content/train.csv', header=True)

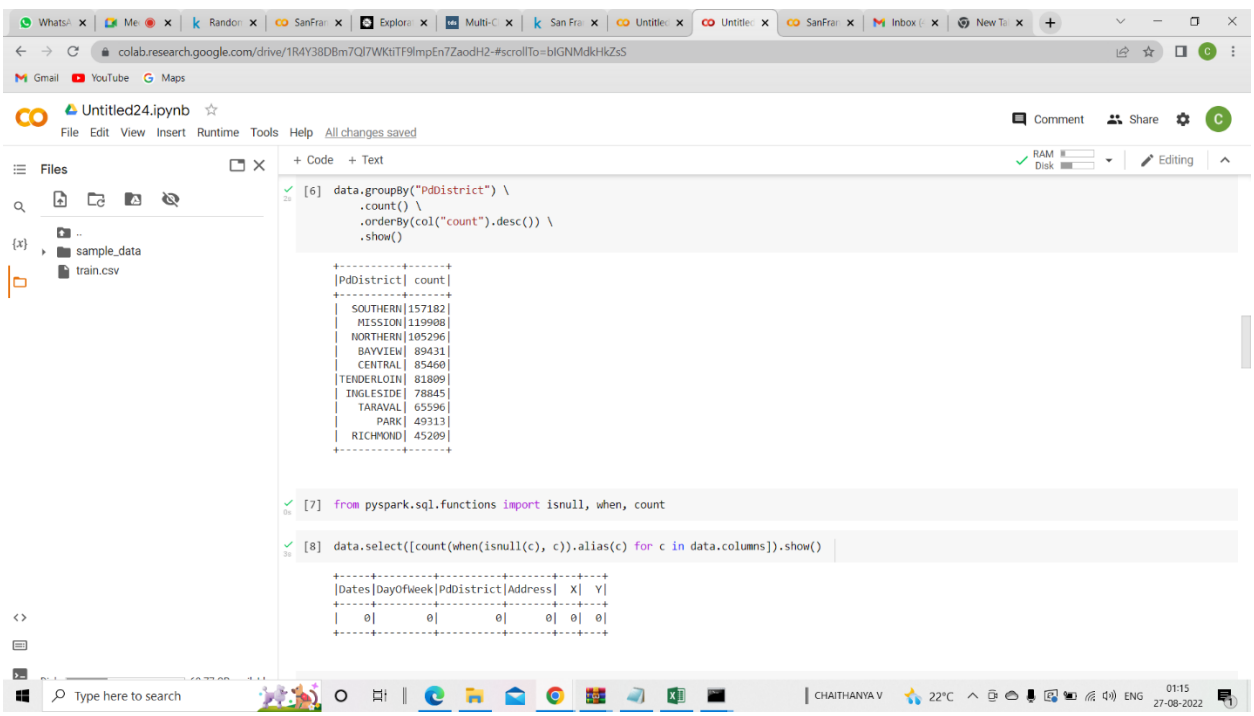


Step-3: Dropping the attributes

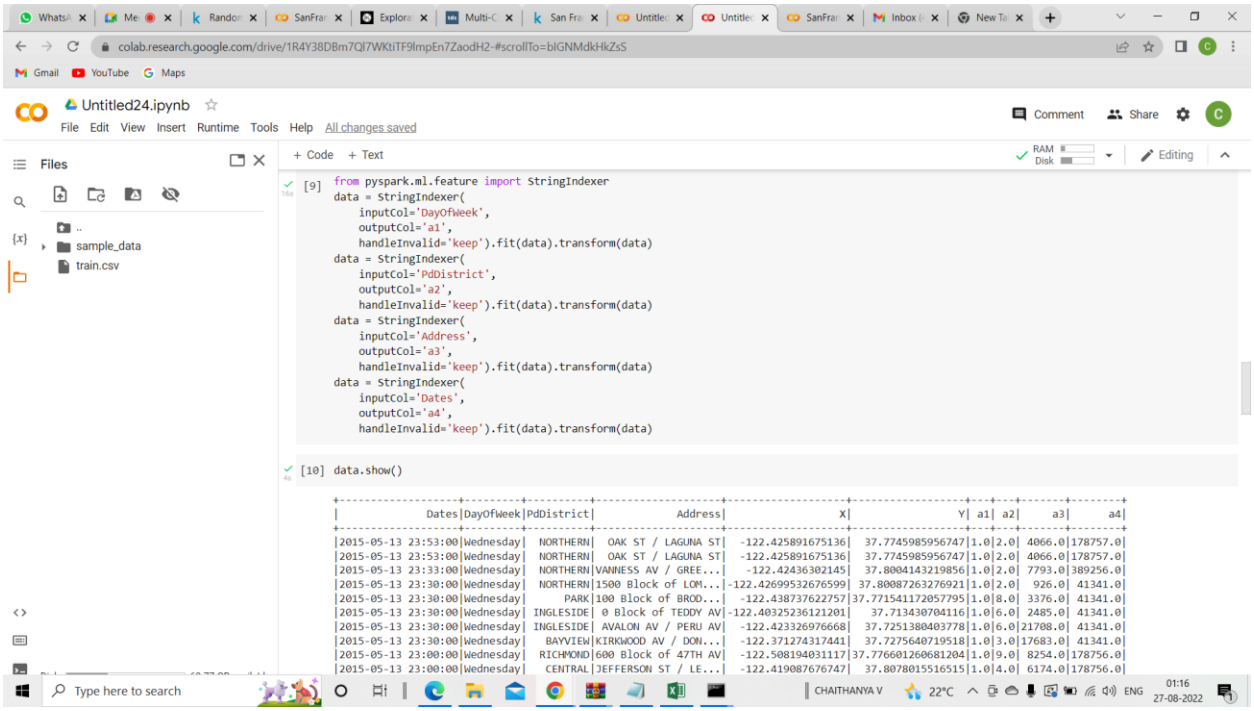


Step-4: Data grouping by the Address and PdDistrict

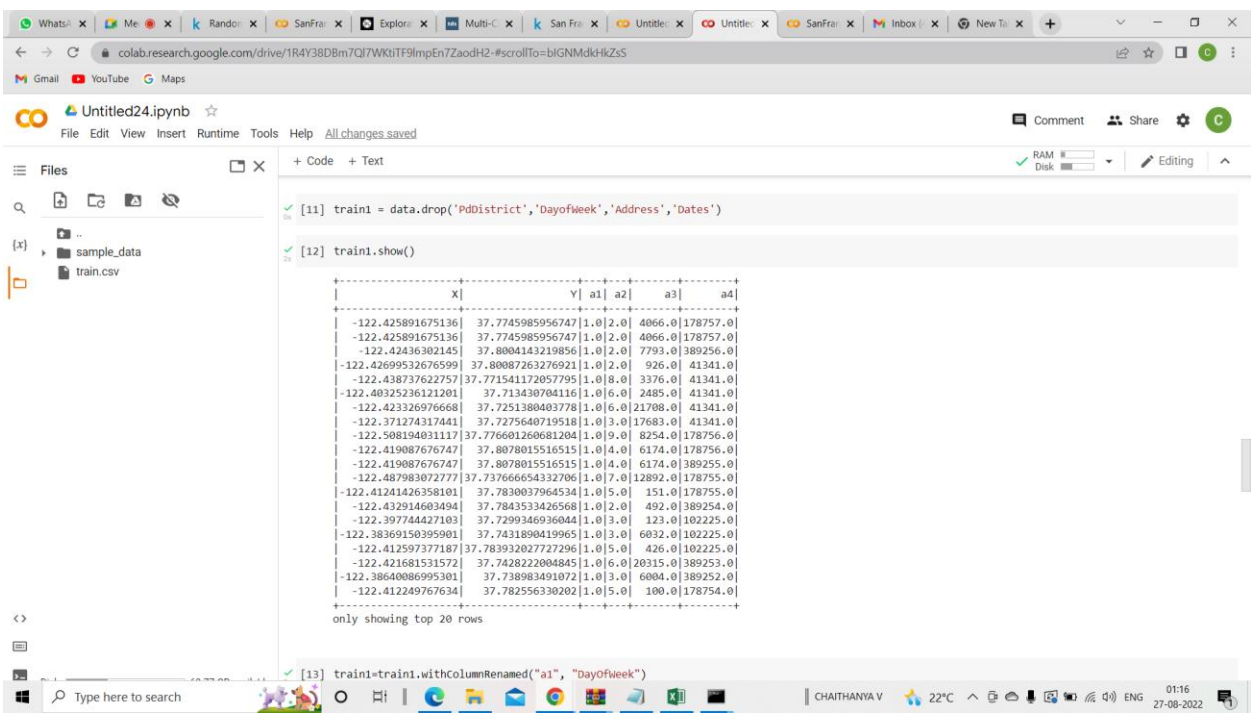


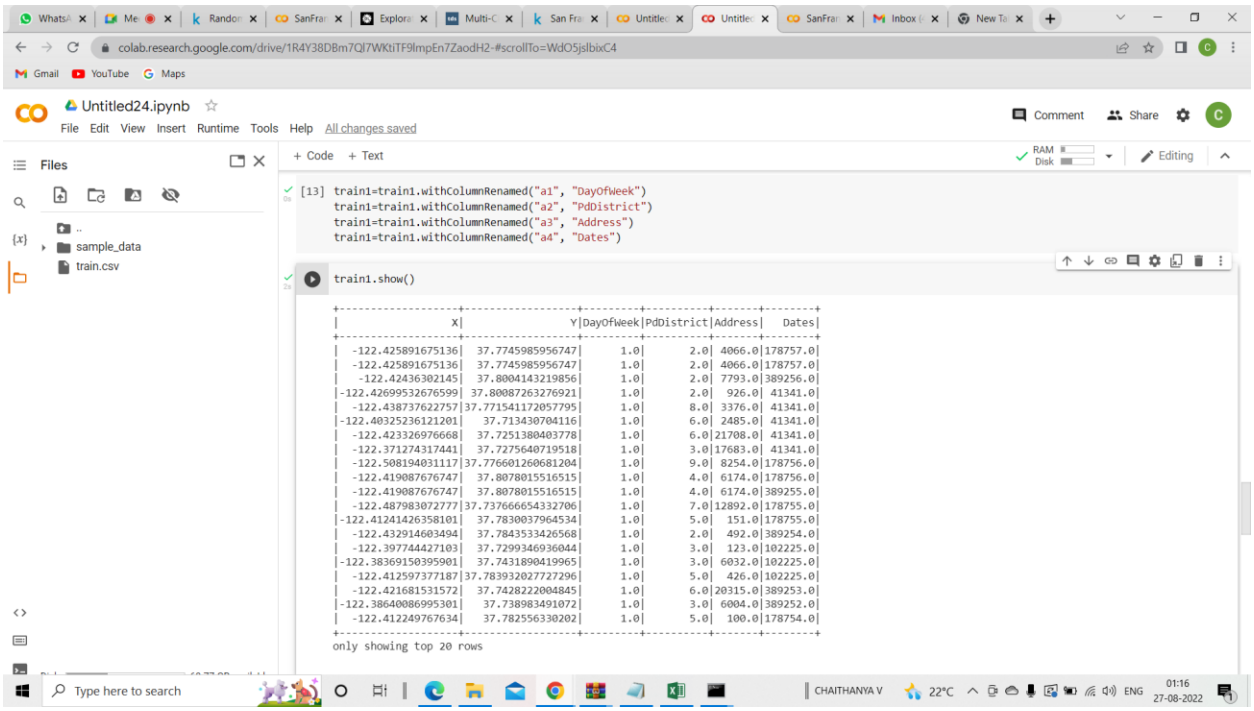


Step-5: Converting the string into numerical



Step-6: Dropping the String rows and re-naming all the converted attributes

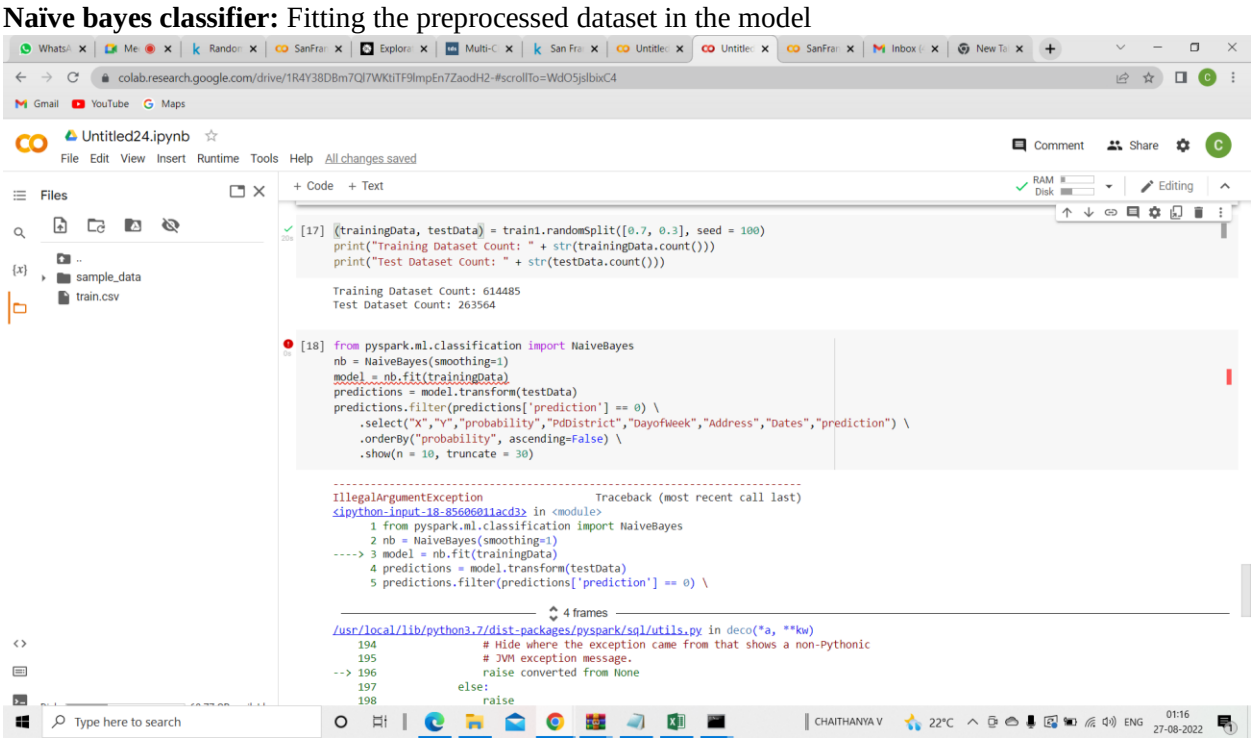


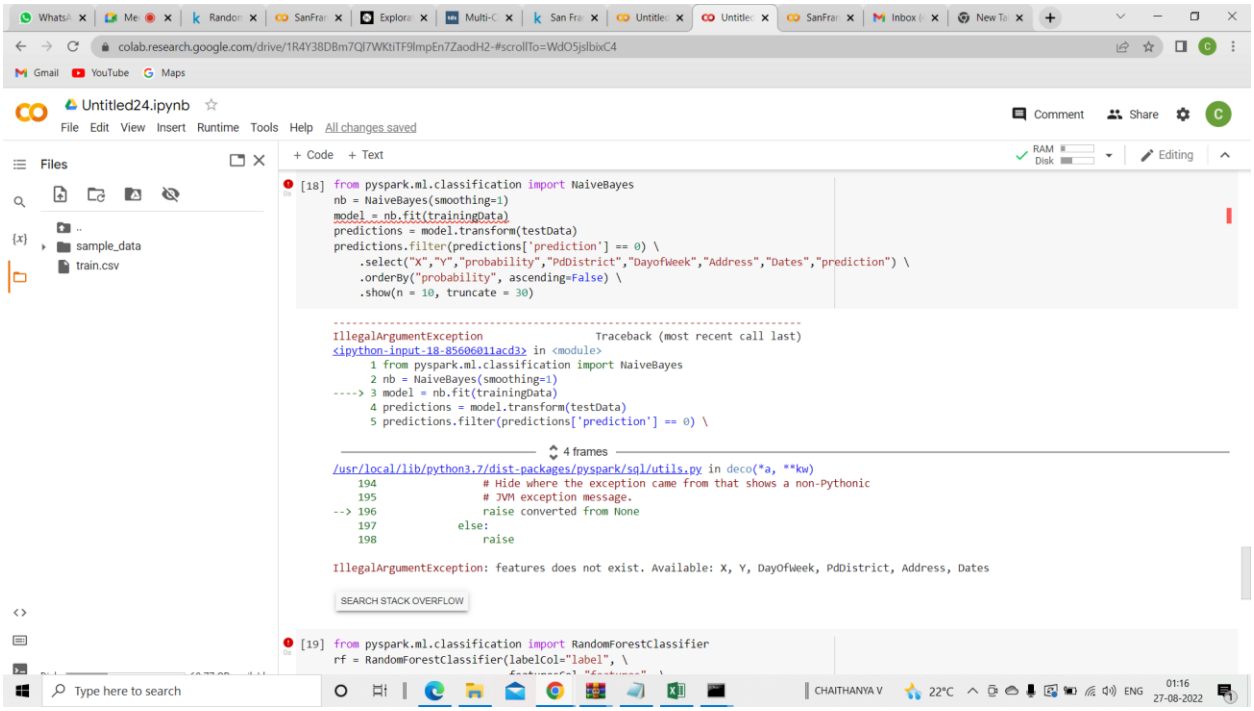


3. Training the model using Random Forest and Naive bayes classifiers:

→Splitting the data into training and testing datasets.

Naïve bayes classifier: Fitting the preprocessed dataset in the model





Random Forest classifier: Fitting the preprocessed dataset in the model