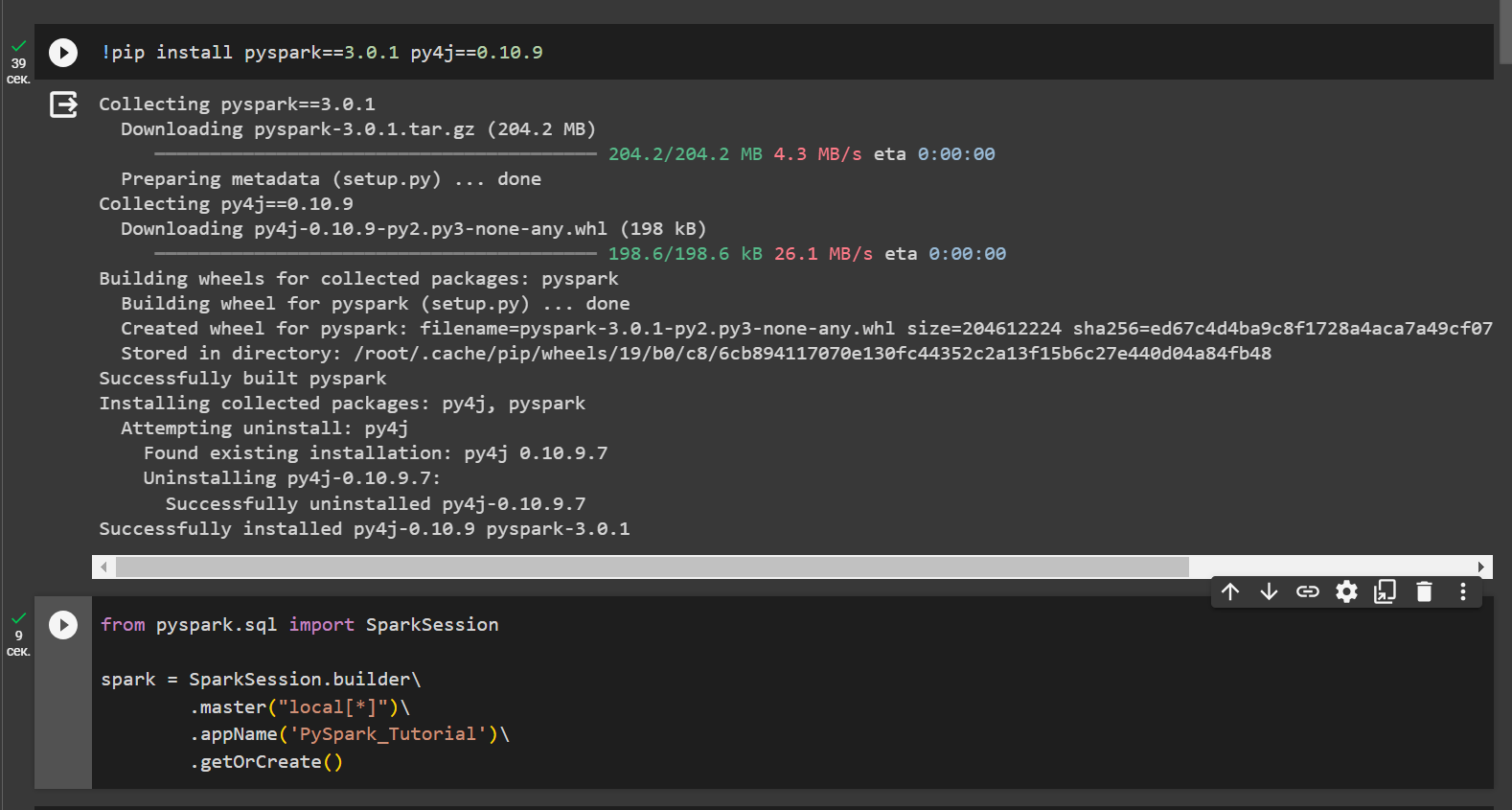
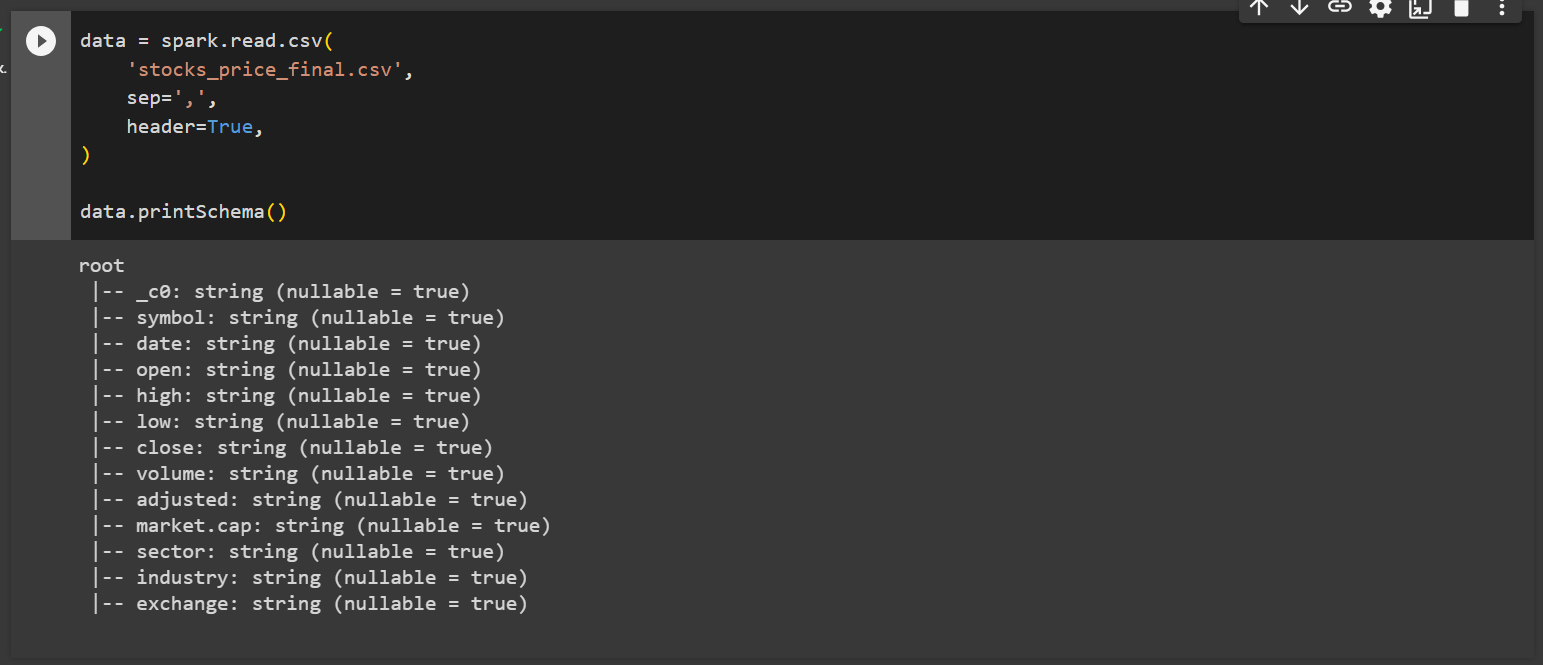
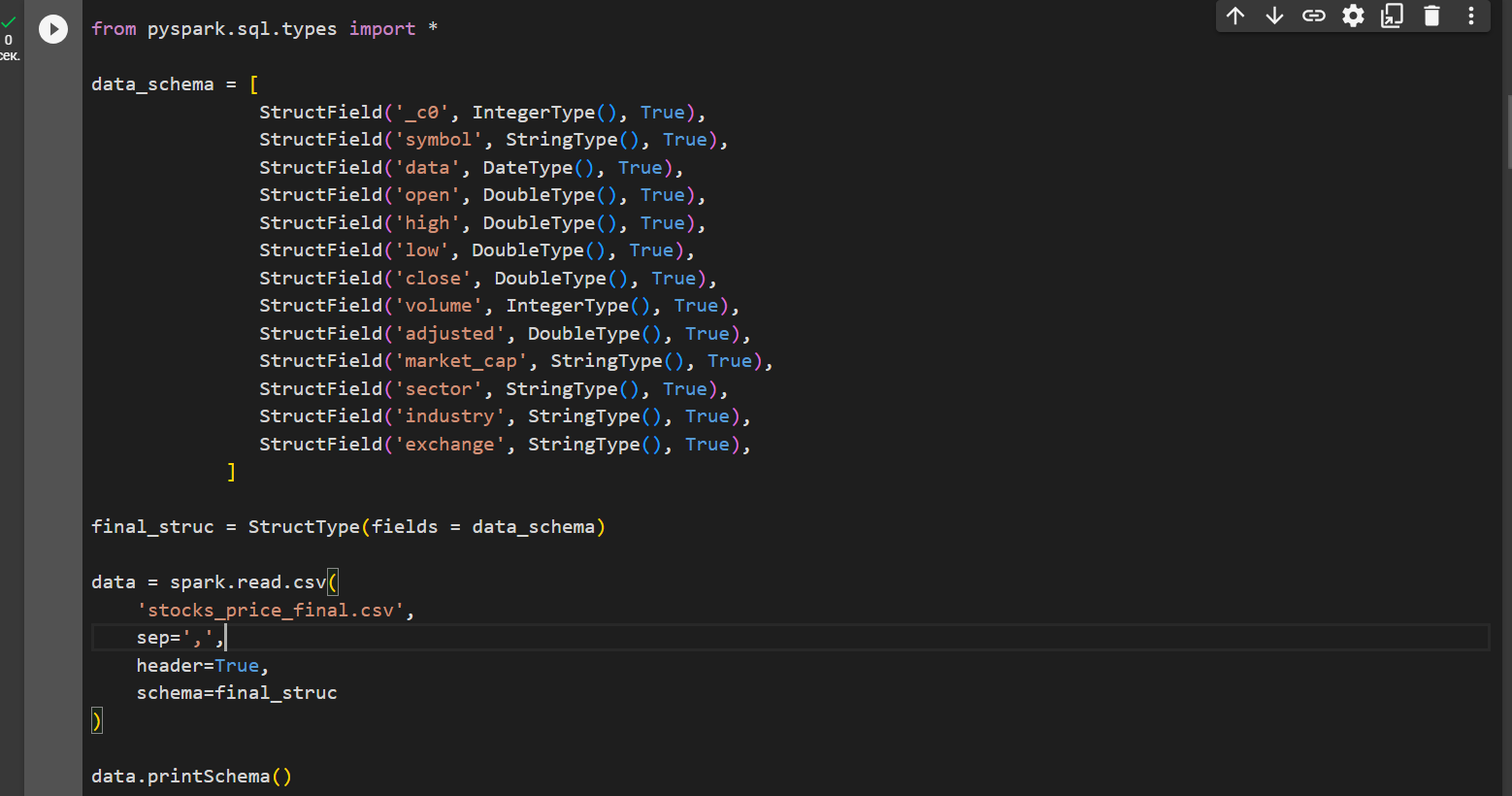
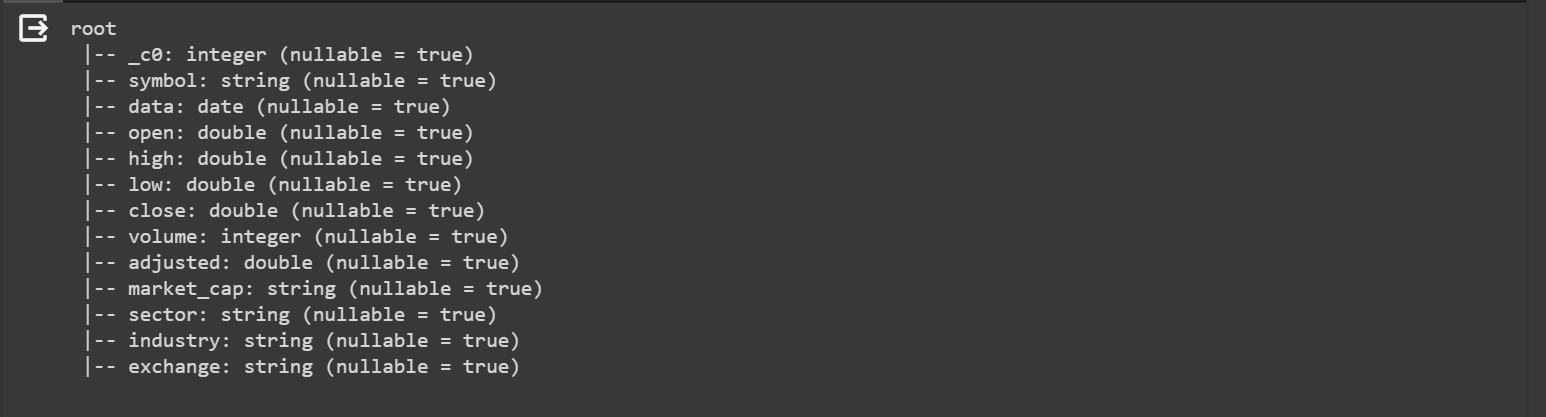
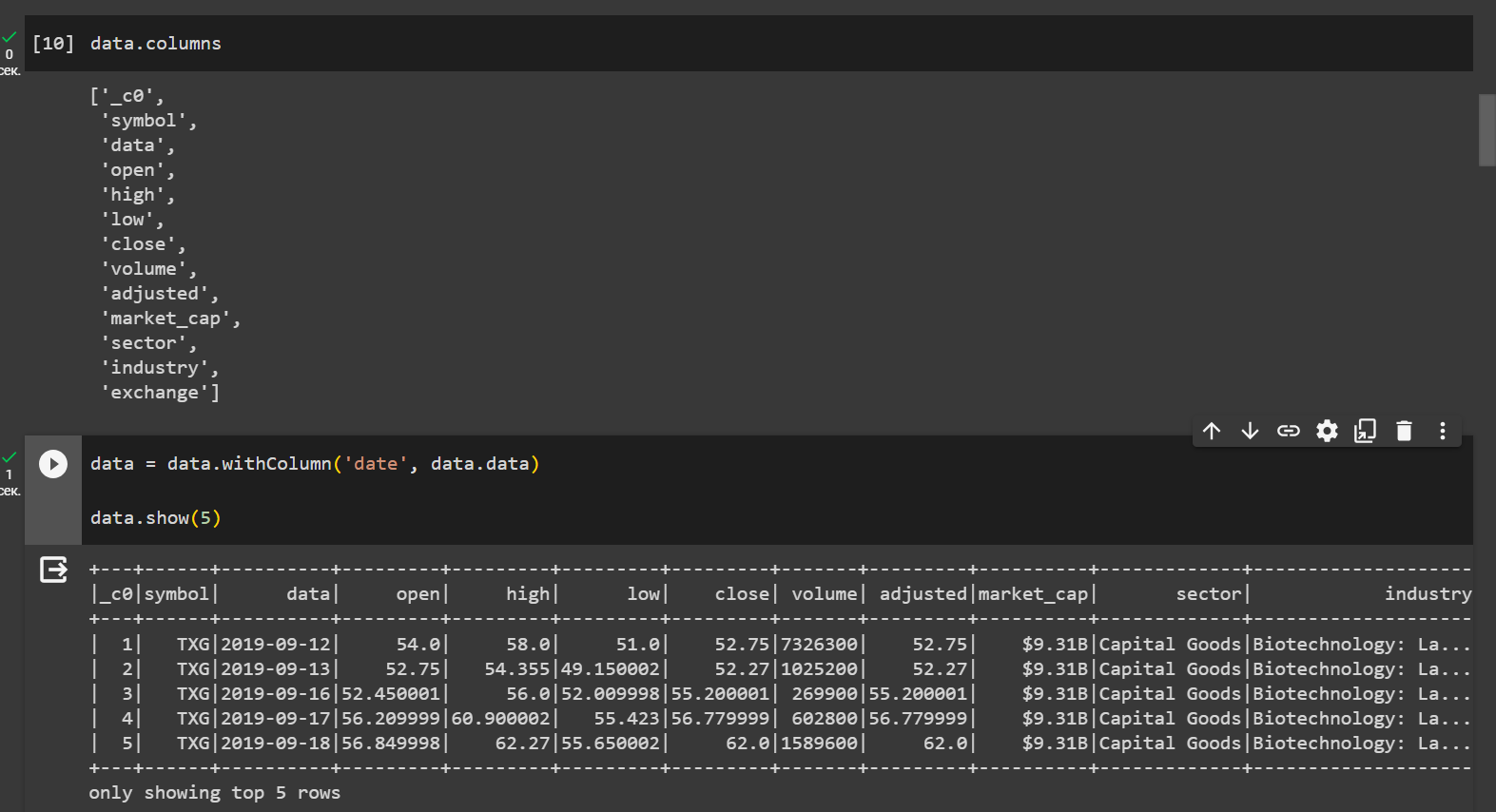
Часть 1 - Goggle Colab

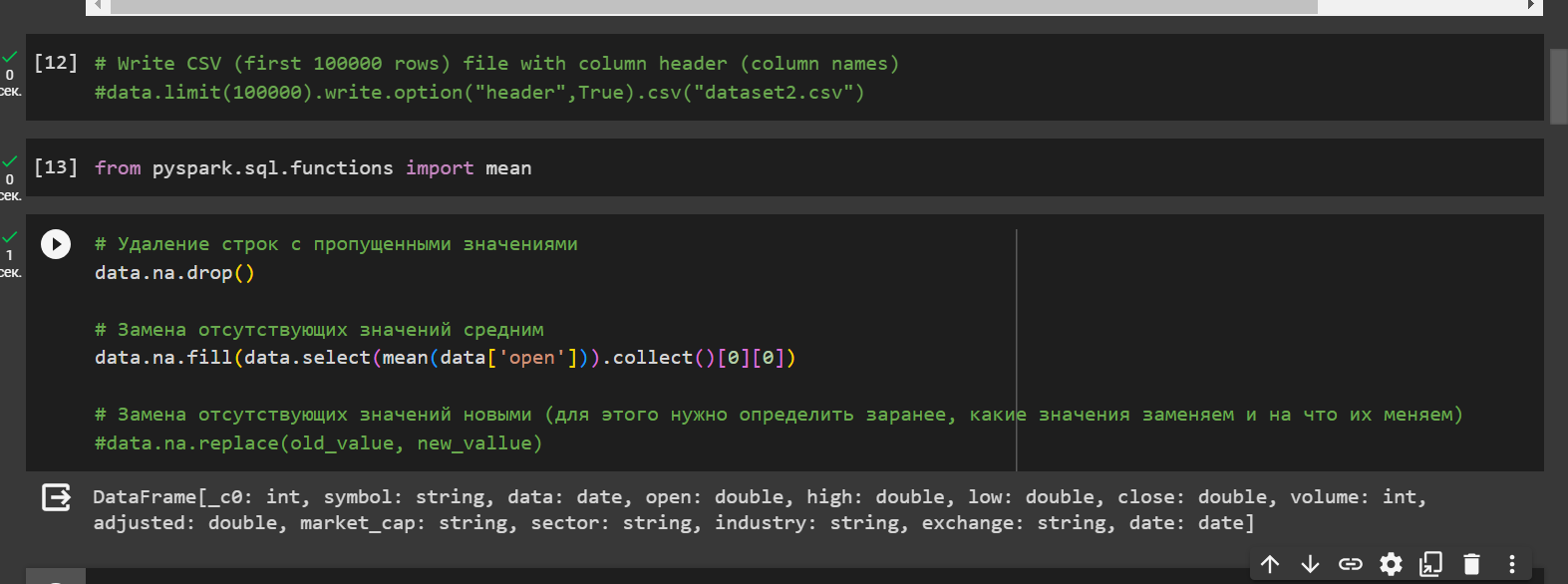




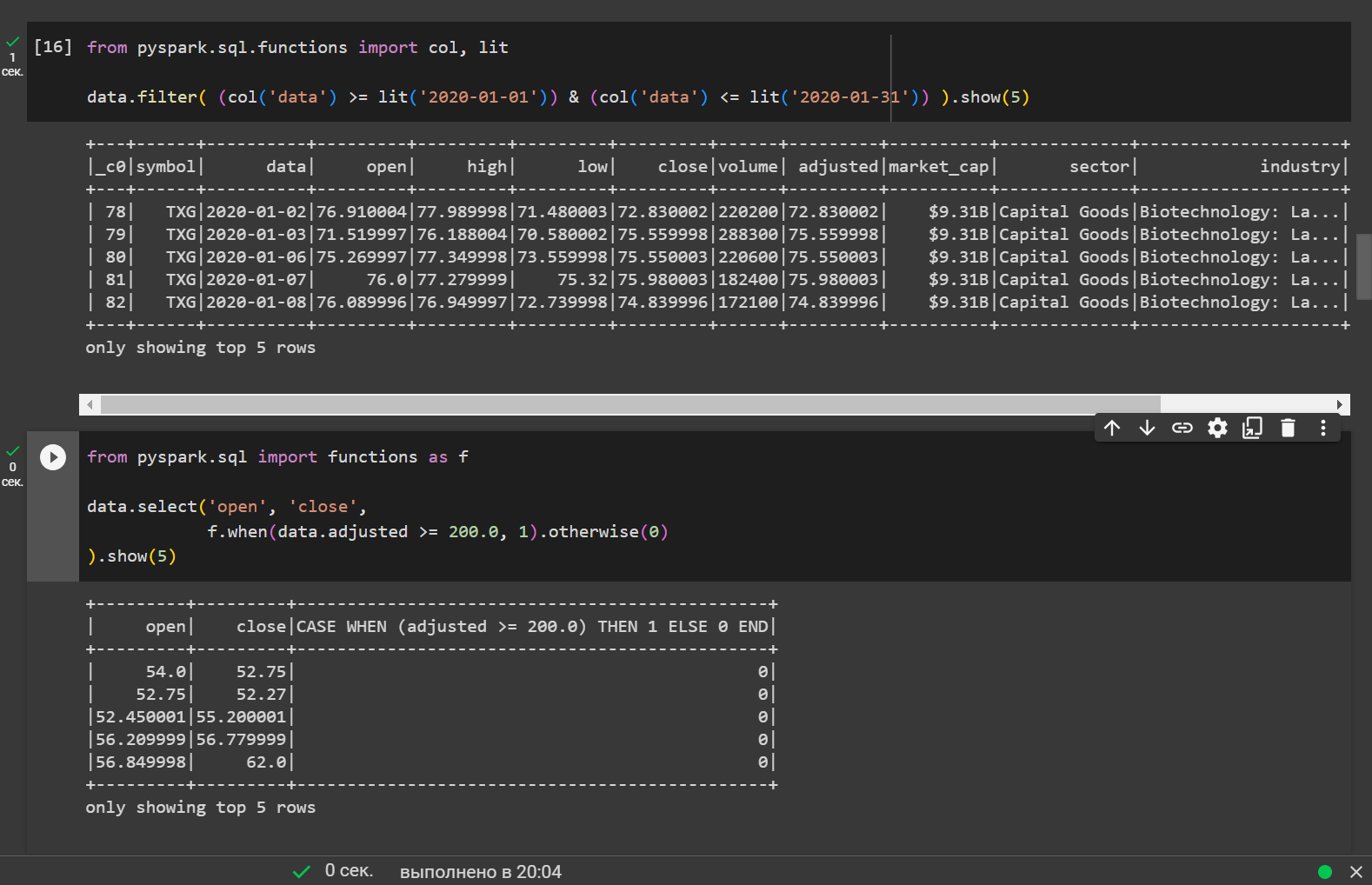


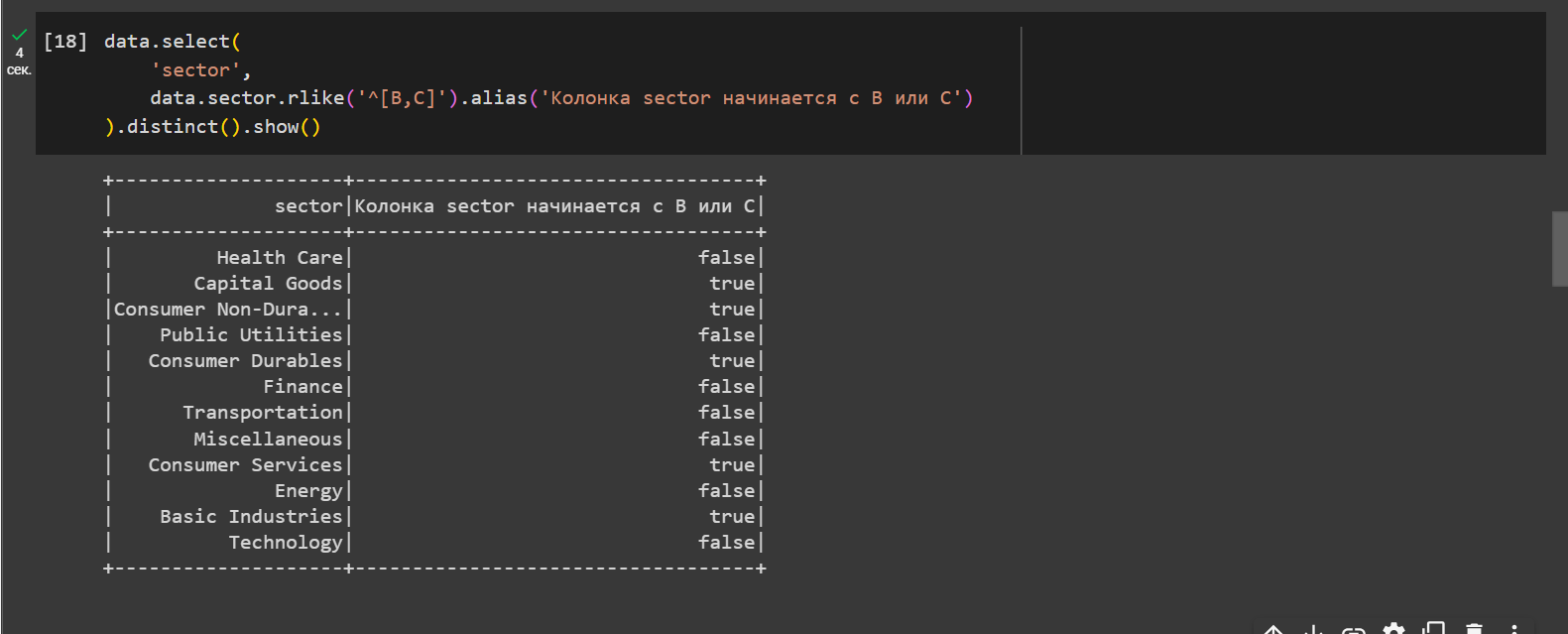


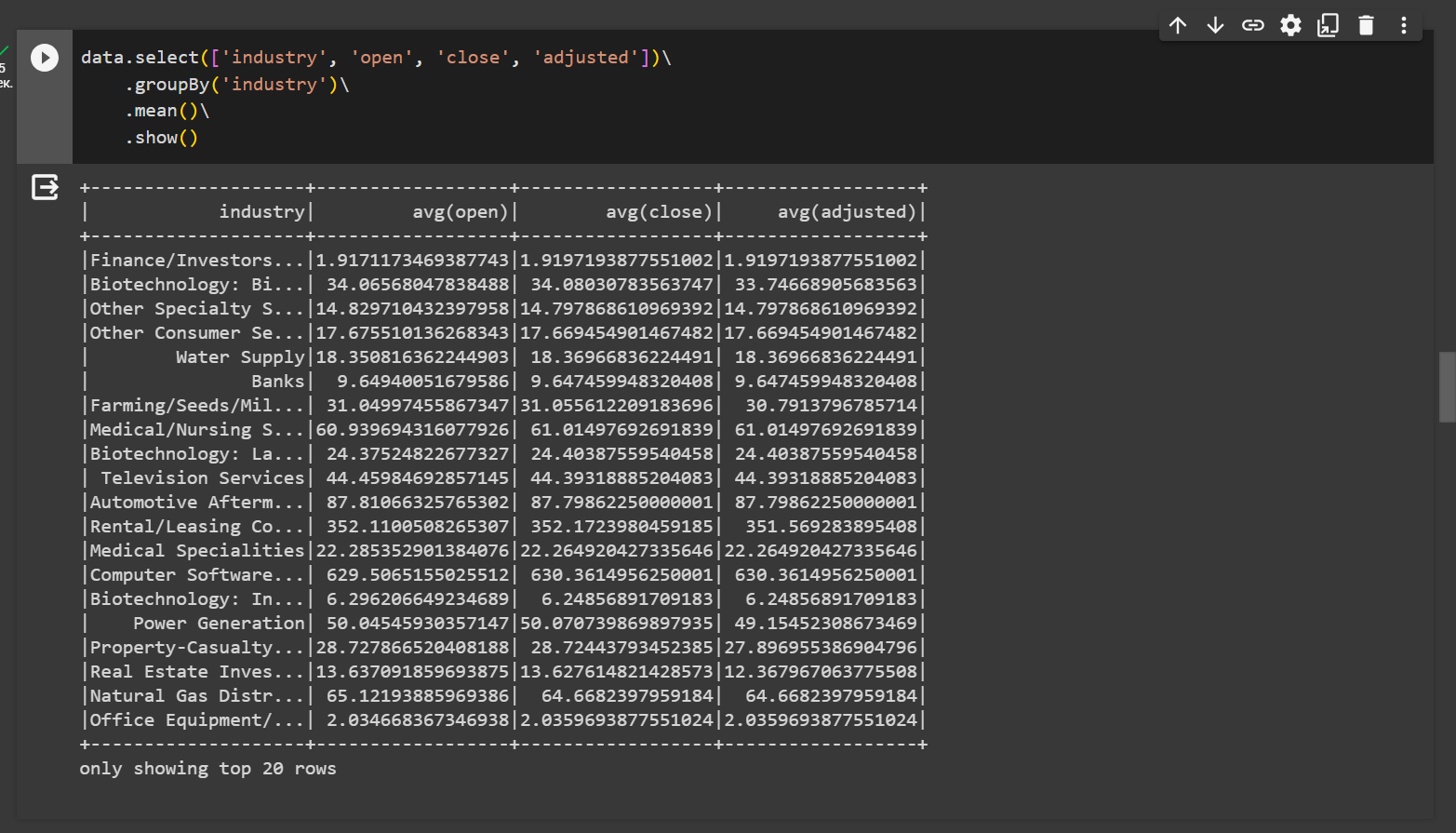


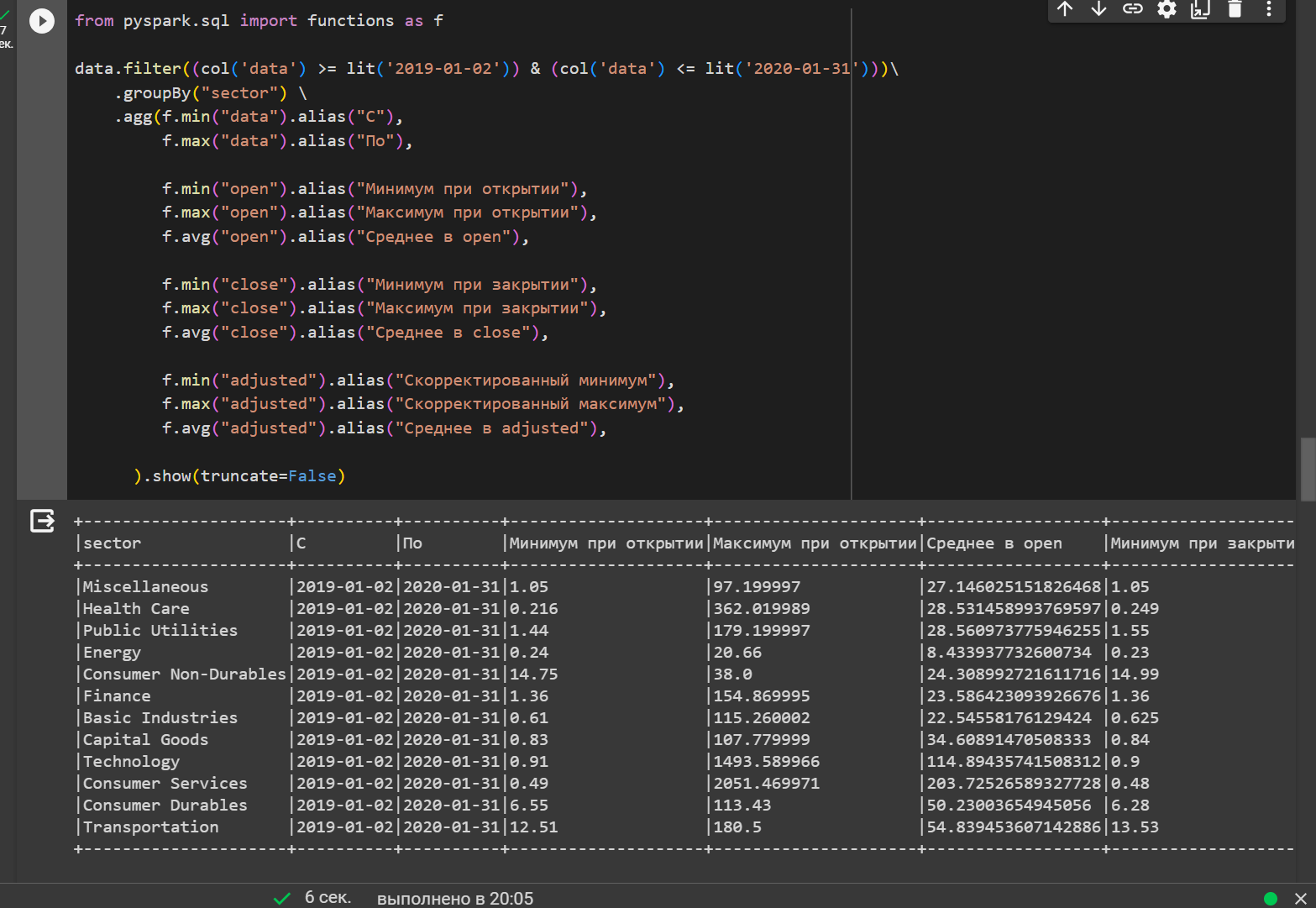


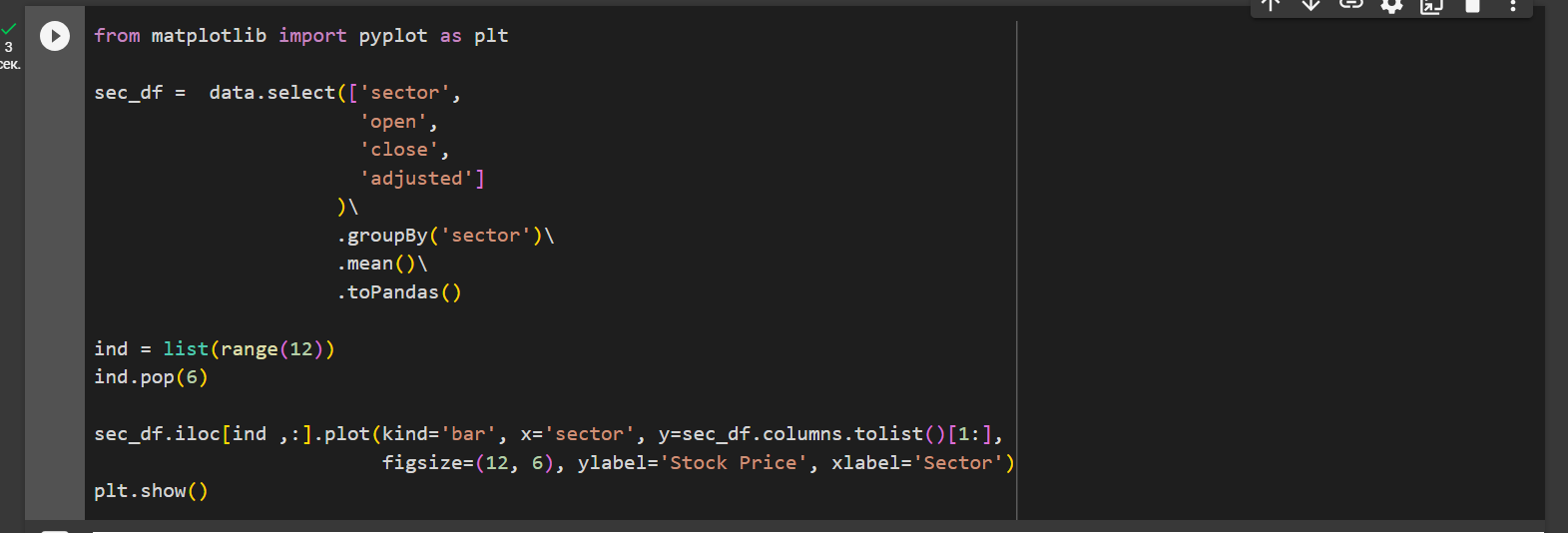


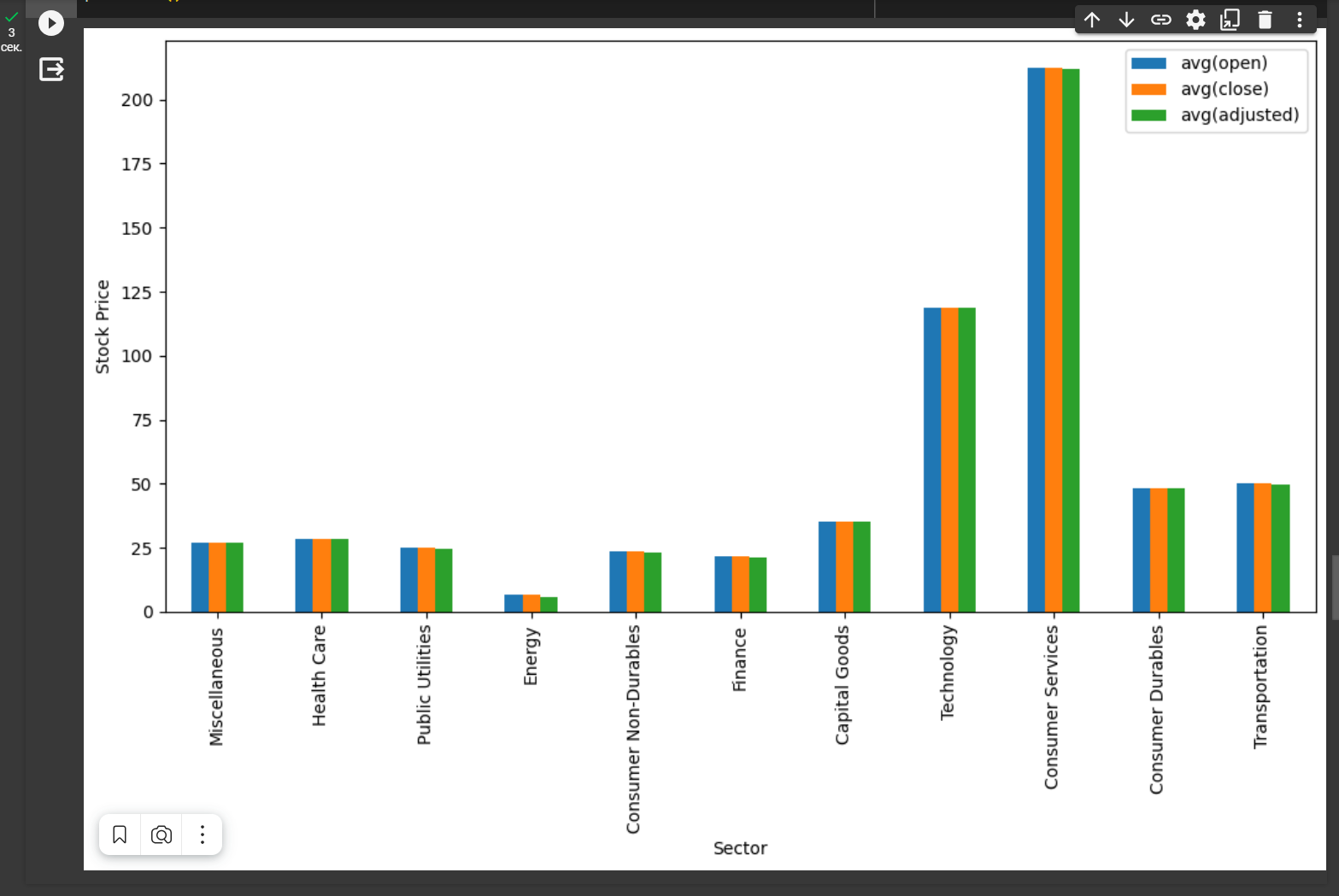


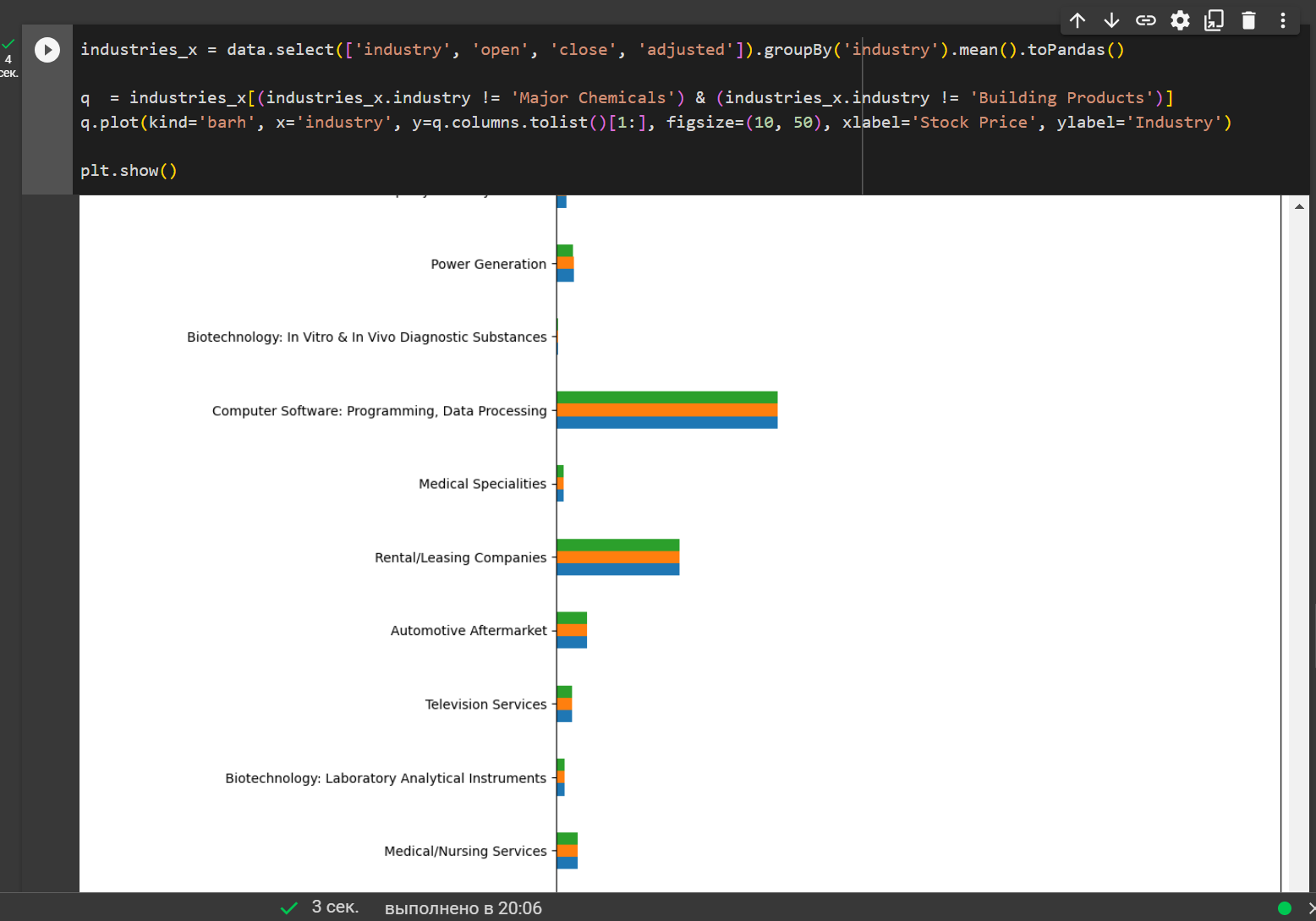


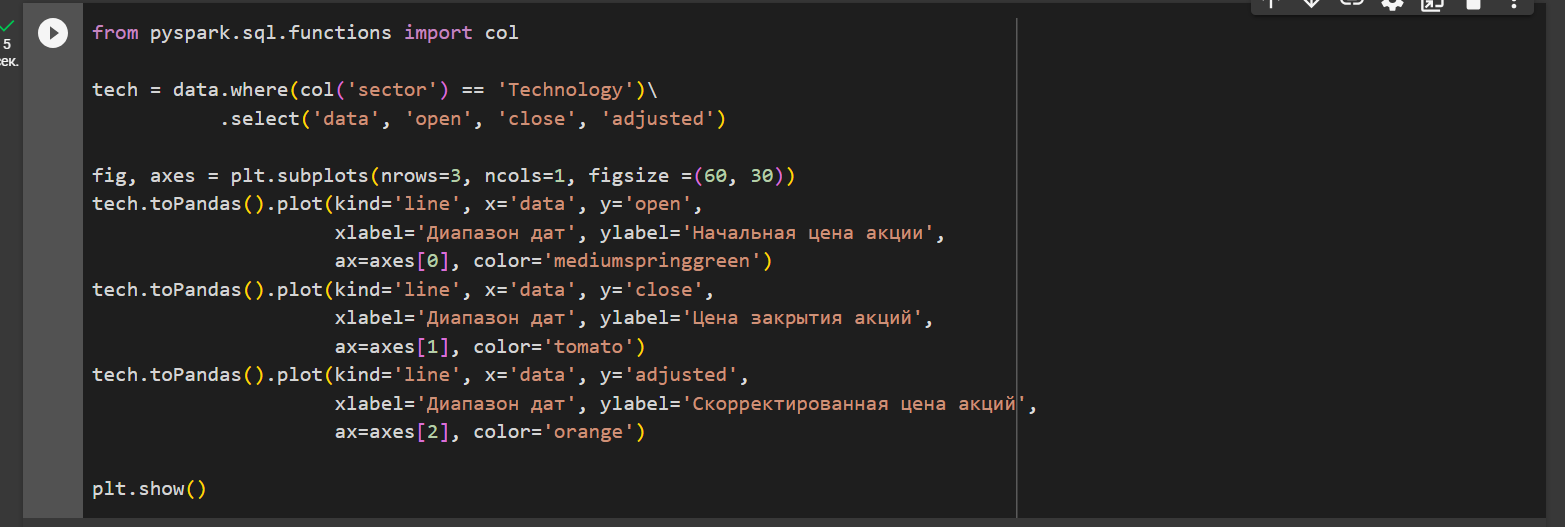


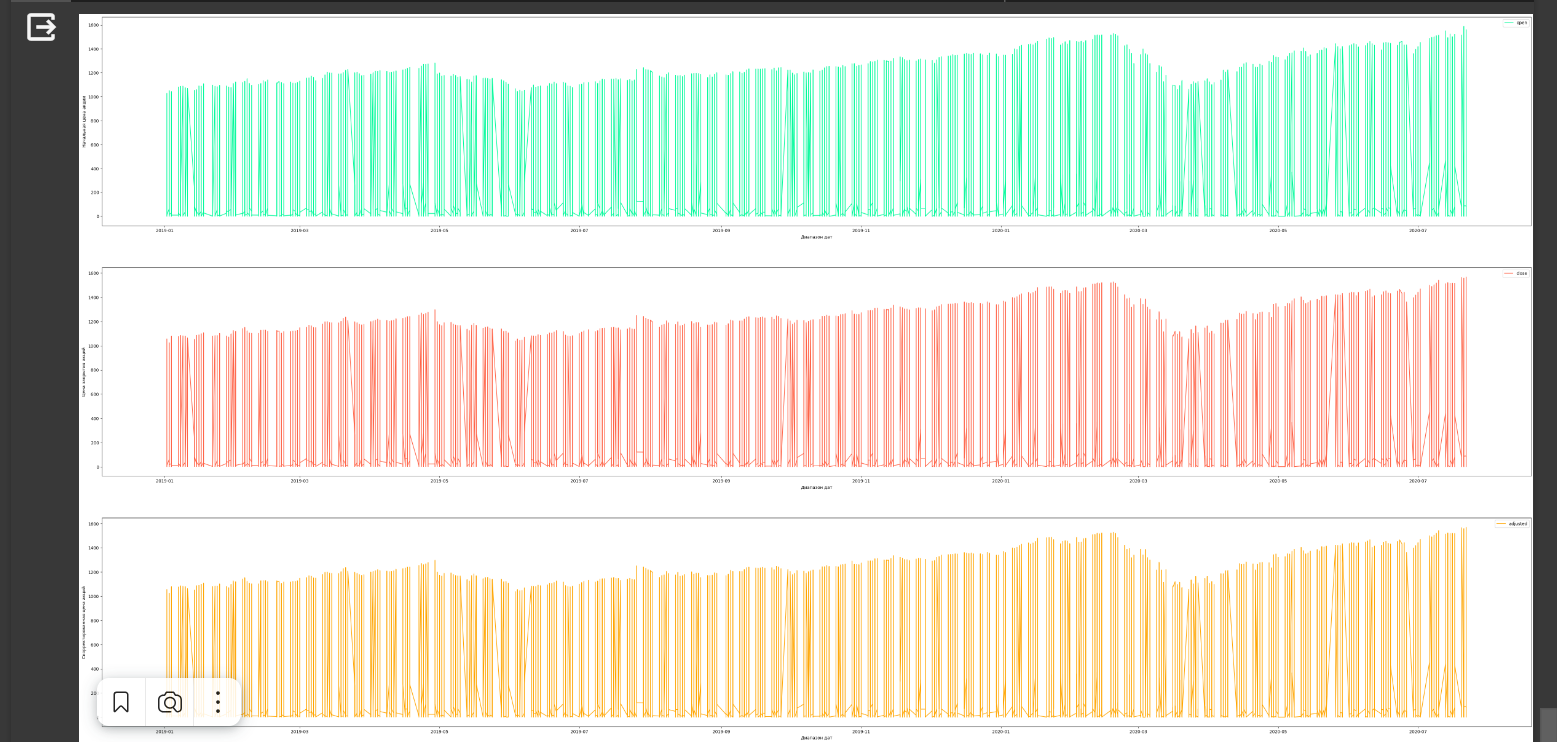














Часть 2 - Zeppelin

Перечень команд

1

import org.apache.spark.sql.SparkSession

val spark = SparkSession.builder().master("local[\*]").appName("PySpark\_Tutorial").getOrCreate()

2

val df = spark.read.format("csv").option("header", "true").load("/home/av/Downloads/stocks\_price\_final.csv")

df.show()

3

val data = df.limit(5)

data.show()

4

data.write.csv("/home/av/Downloads/data.csv")

data.select("date", "open", "close", "adjusted").write.json("/home/av/Downloads/data.json")

5

data.printSchema()

6

import org.apache.spark.sql.types.\_

val data\_schema = Array(

  StructField("\_c0", IntegerType, true),

            StructField("symbol", StringType, true),

            StructField("data", DateType, true),

            StructField("open", DoubleType, true),

            StructField("high", DoubleType, true),

            StructField("low", DoubleType, true),

            StructField("close", DoubleType, true),

            StructField("volume", IntegerType, true),

            StructField("adjusted", DoubleType, true),

            StructField("market.cap", StringType, true),

            StructField("sector", StringType, true),

            StructField("industry", StringType, true),

            StructField("exchange", StringType, true)

  )

7

val final\_struc =StructType(fields = data\_schema )

val df = spark.read.format("csv").option("header", "true").schema(final\_struc ).load("/home/av/Downloads/stocks\_price\_final.csv")

df.printSchema()

8

val tst\_josn = spark.read.json("/home/av/Downloads/data.json")

tst\_josn.show()

9

data.schema

10

data.dtypes

11

data.take(5)

12

data.first()

13

data.first()

14

data.describe().show()

15

data.columns

16

data.count()

17

data.distinct().count()

18

import org.apache.spark.sql.Row

import org.apache.spark.sql.functions.lit

val data\_new = data.withColumn("test\_new", lit("test")) // lit - создает столбец test\_new и заполняет его значениями "test"

data\_new.show(5)

19

data\_new.withColumnRenamed("test\_new","test\_new\_1").show()

20

data\_new.drop("test\_new").show()

21

val data\_null = data.withColumn("close",     when(col("close")===62,null).otherwise(col("close")))

22

val data\_null\_new=data\_null.withColumnRenamed("market.cap","market\_cap")

23

data\_null\_new.na.drop().show()

24

// # Замена отсутствующих значений средним

import org.apache.spark.sql.functions.mean

25

data.select("sector").show()

26

data.select("open", "close","adjusted").show()

27

data.filter( (col("open") >= 54 )).show()