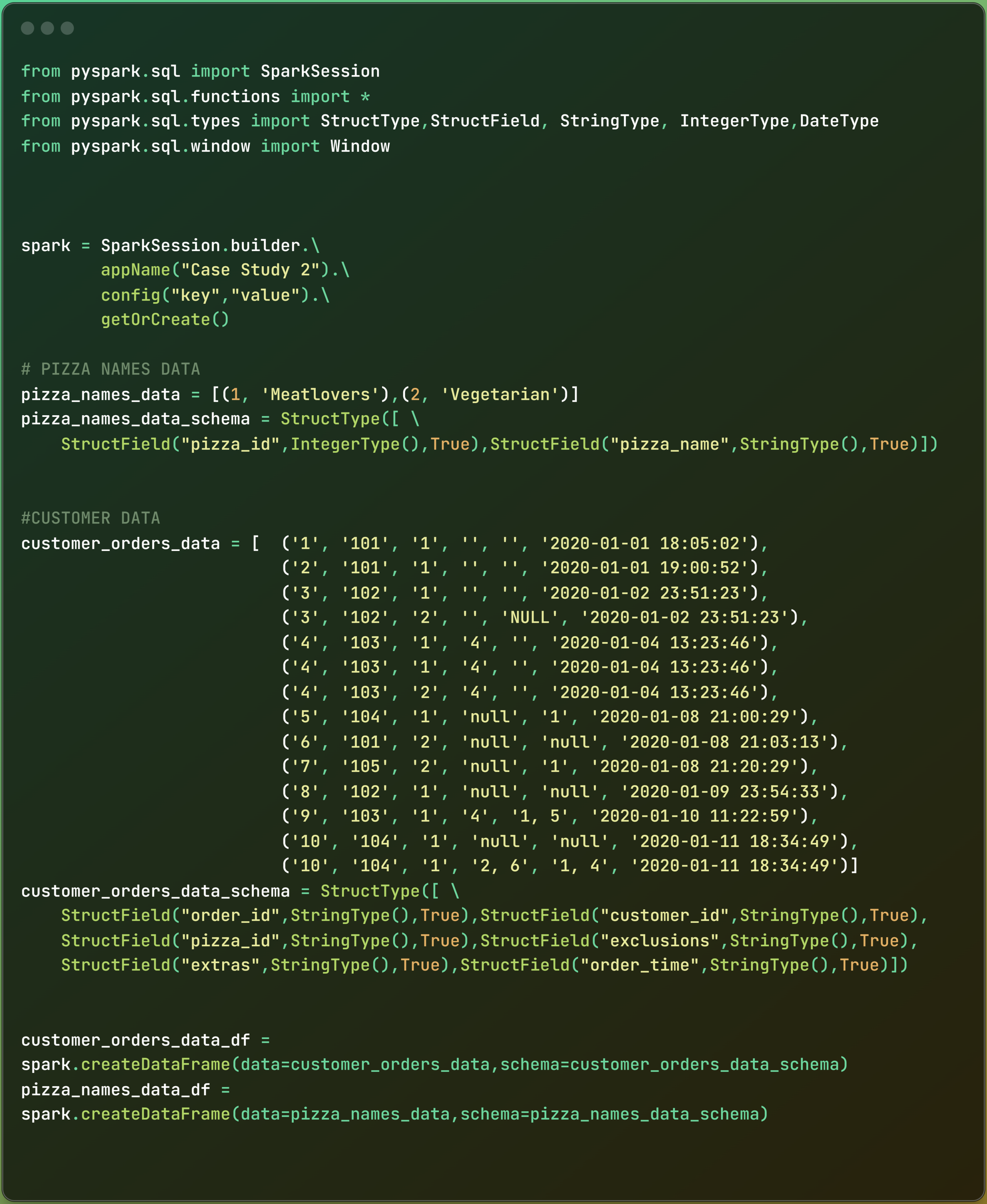
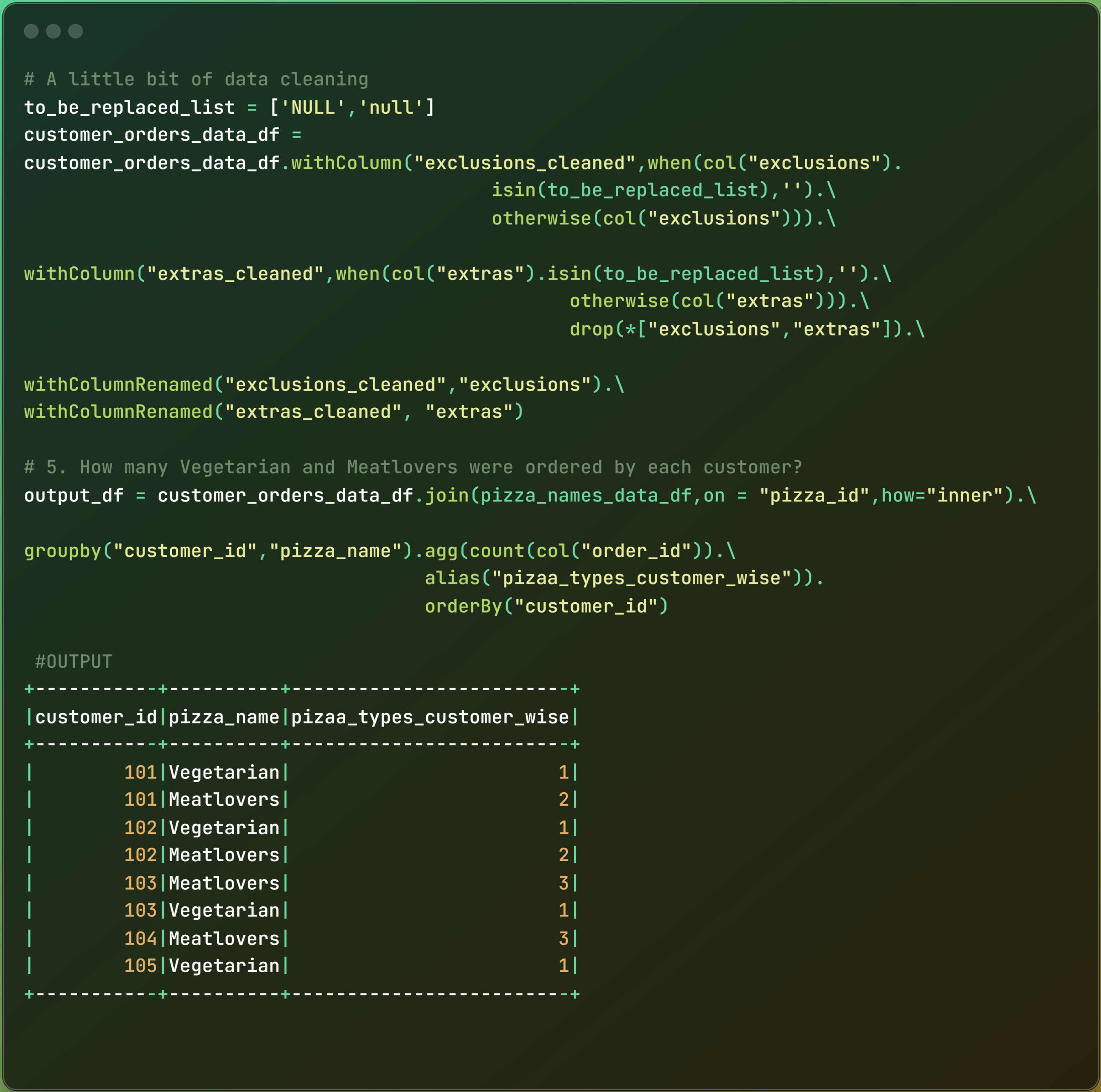


```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
from pyspark.sql.types import StructType,StructField, StringType, IntegerType,DateType
from pyspark.sql.window import Window



spark = SparkSession.builder.\
        appName("Case Study 2").\
        config("key","value").\
        getOrCreate()

# PIZZA NAMES DATA
pizza_names_data = [(1, 'Meatlovers'),(2, 'Vegetarian')]
pizza_names_data_schema = StructType([ \
    StructField("pizza_id",IntegerType(),True),StructField("pizza_name",StringType(),True)])


#CUSTOMER DATA
customer_orders_data = [  ('1', '101', '1', '', '', '2020-01-01 18:05:02'),
                          ('2', '101', '1', '', '', '2020-01-01 19:00:52'),
                          ('3', '102', '1', '', '', '2020-01-02 23:51:23'),
                          ('3', '102', '2', '', 'NULL', '2020-01-02 23:51:23'),
                          ('4', '103', '1', '4', '', '2020-01-04 13:23:46'),
                          ('4', '103', '1', '4', '', '2020-01-04 13:23:46'),
                          ('4', '103', '2', '4', '', '2020-01-04 13:23:46'),
                          ('5', '104', '1', 'null', '1', '2020-01-08 21:00:29'),
                          ('6', '101', '2', 'null', 'null', '2020-01-08 21:03:13'),
                          ('7', '105', '2', 'null', '1', '2020-01-08 21:20:29'),
                          ('8', '102', '1', 'null', 'null', '2020-01-09 23:54:33'),
                          ('9', '103', '1', '4', '1, 5', '2020-01-10 11:22:59'),
                          ('10', '104', '1', 'null', 'null', '2020-01-11 18:34:49'),
                          ('10', '104', '1', '2, 6', '1, 4', '2020-01-11 18:34:49')]
customer_orders_data_schema = StructType([ \
    StructField("order_id",StringType(),True),StructField("customer_id",StringType(),True),
    StructField("pizza_id",StringType(),True),StructField("exclusions",StringType(),True),
    StructField("extras",StringType(),True),StructField("order_time",StringType(),True)])


customer_orders_data_df =
spark.createDataFrame(data=customer_orders_data,schema=customer_orders_data_schema)
pizza_names_data_df =
spark.createDataFrame(data=pizza_names_data,schema=pizza_names_data_schema)
```



```
# A little bit of data cleaning
to_be_replaced_list = ['NULL','null']
customer_orders_data_df =
customer_orders_data_df.withColumn("exclusions_cleaned",when(col("exclusions").
                                          isin(to_be_replaced_list),'').\
                                          otherwise(col("exclusions"))).\

withColumn("extras_cleaned",when(col("extras").isin(to_be_replaced_list),'').\
                                                 otherwise(col("extras"))).\
                                                 drop(*["exclusions","extras"]).\

withColumnRenamed("exclusions_cleaned","exclusions").\
withColumnRenamed("extras_cleaned", "extras")

# 5. How many Vegetarian and Meatlovers were ordered by each customer?
output_df = customer_orders_data_df.join(pizza_names_data_df,on = "pizza_id",how="inner").\

groupby("customer_id","pizza_name").agg(count(col("order_id")).\
                                    alias("pizaa_types_customer_wise")).
                                    orderBy("customer_id")

 #OUTPUT
+-----------+----------+-------------------------+
|customer_id|pizza_name|pizaa_types_customer_wise|
+-----------+----------+-------------------------+
|        101|Vegetarian|                        1|
|        101|Meatlovers|                        2|
|        102|Vegetarian|                        1|
|        102|Meatlovers|                        2|
|        103|Meatlovers|                        3|
|        103|Vegetarian|                        1|
|        104|Meatlovers|                        3|
|        105|Vegetarian|                        1|
+-----------+----------+-------------------------+
```