BWT Task-09 Exercise

Submitted By: ABRAR SAEED

PySpark:

Code & Output:

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, to_date, udf, hour, lower,
to_timestamp
from pyspark.sql.types import StringType

spark = SparkSession.builder \
    .appName("Data Cleaning") \
    .getOrCreate()

df = spark.read.csv('/content/drive/MyDrive/data.csv', header=True,
inferSchema=True, quote='"', escape='"')

print("Initial Data:")
df.show(5)

# DATA Cleaning operations
# 1. Remove orders placed between 12am and 5am and convert timestamp to
date
df = df.withColumn("order_date", to_timestamp(col("order_date")))

# Apply the hour filter to exclude times between 0 and 5 AM
df = df.filter(~hour(col("order_date")).between(0, 5))

# 2. Adding new column time_of_day
def get_time_of_day(hour):
    if 5 <= hour < 12:
        return 'morning'
    elif 12 <= hour < 18:
        return 'afternoon'
    elif 18 <= hour < 24:
        return 'evening'
    else:
        return 'night'

time_of_day_udf = udf(get_time_of_day, StringType())
df = df.withColumn('time_of_day',
time_of_day_udf(hour(col('order_date'))))
```

```python
# 3. Remove rows containing "TV" in the product column
df = df.filter(~df.product.contains("TV"))

# 4. Ensure all product categories are in lowercase
df = df.withColumn("category", lower(col("category")))

# 5. Adding new column purchase_state
def extract_state_and_zip(address):
    parts = address.split(',')
    if len(parts) >= 3:
        return parts[-1].strip()
    else:
        return None

extract_state_and_zip_udf = udf(extract_state_and_zip, StringType())
df = df.withColumn('purchase_state',
extract_state_and_zip_udf(col('purchase_address')))

# Parquet file
df.write.parquet('/content/drive/MyDrive/data.parquet')

df.show()
spark.stop()
```

Initial Data:

order_date	order_id	product	product_id	category	purchase_address	quantity_ordered	price_each	cost_price	turnover	margi
2023-01-22 21:25:00	141234	iPhone	5638008983335	Vêtements	"944 Walnut St, B...	1	700.0	231.0	700.0	469.
2023-01-28 14:15:00	141235	Lightning Chargin...	5563319511488	Alimentation	"185 Maple St, Po...	1	14.95	7.475	14.95	7.47
2023-01-17 13:33:00	141236	Wired Headphones	2113973395220	Vêtements	"538 Adams St, Sa...	2	11.99	5.995	23.98	11.9
2023-01-05 20:33:00	141237	27in FHD Monitor	3069156759167	Sports	"738 10th St, Los...	1	149.99	97.4935	149.99	52.496
2023-01-25 11:59:00	141238	Wired Headphones	9692680938163	Électronique	"387 10th St, Aus...	1	11.99	5.995	11.99	5.99

only showing top 5 rows

order_date	order_id	product	product_id	category	purchase_address	quantity_ordered	price_each	cost_price	turnover
2023-01-22 21:25:00	141234	iPhone	5638008983335	vêtements	"944 Walnut St, B...	1	700.0	231.0	700.0
2023-01-28 14:15:00	141235	Lightning Chargin...	5563319511488	alimentation	"185 Maple St, Po...	1	14.95	7.475	14.95
2023-01-17 13:33:00	141236	Wired Headphones	2113973395220	vêtements	"538 Adams St, Sa...	2	11.99	5.995	23.98
2023-01-05 20:33:00	141237	27in FHD Monitor	3069156759167	sports	"738 10th St, Los...	1	149.99	97.4935	149.99
2023-01-25 11:59:00	141238	Wired Headphones	9692680938163	électronique	"387 10th St, Aus...	1	11.99	5.995	11.99
2023-01-29 20:22:00	141239	AAA Batteries (4-...	2953868554188	alimentation	"775 Willow St, S...	1	2.99	1.495	2.99
2023-01-26 12:16:00	141240	27in 4K Gaming Mo...	5173670800988	vêtements	"979 Park St, Los...	1	389.99	128.6967	389.99
2023-01-05 12:04:00	141241	USB-C Charging Cable	8051736777568	vêtements	"181 6th St, San ...	1	11.95	5.975	11.95
2023-01-01 10:30:00	141242	Bose SoundSport H...	1508418177978	électronique	"867 Willow St, L...	1	99.99	49.995	99.99
2023-01-22 21:20:00	141243	Apple Airpods Hea...	1386344211590	électronique	"657 Johnson St, ...	1	150.0	97.5	150.0
2023-01-07 11:29:00	141244	Apple Airpods Hea...	4332898830865	vêtements	"492 Walnut St, S...	1	150.0	97.5	150.0
2023-01-31 10:12:00	141245	Macbook Pro Laptop	1169379570345	vêtements	"322 6th St, San ...	1	1700.0	561.0	1700.0
2023-01-09 18:57:00	141246	AAA Batteries (4-...	4436184749366	vêtements	"618 7th St, Los ...	3	2.99	1.495	8.97
2023-01-25 19:19:00	141247	27in FHD Monitor	7313825995563	vêtements	"512 Wilson St, S...	1	149.99	97.4935	149.99
2023-01-05 17:20:00	141249	27in FHD Monitor	9643428300795	alimentation	"440 Cedar St, Po...	1	149.99	97.4935	149.99
2023-01-10 11:20:00	141250	Vareebadd Phone	6721780072847	alimentation	"471 Center St, L...	1	400.0	132.0	400.0
2023-01-24 08:13:00	141251	Apple Airpods Hea...	2700099961823	alimentation	"414 Walnut St, B...	1	150.0	97.5	150.0
2023-01-30 09:28:00	141252	USB-C Charging Cable	3692435232121	sports	"220 9th St, Los ...	1	11.95	5.975	11.95
2023-01-08 11:51:00	141254	AAA Batteries (4-...	8219536039183	électronique	"238 Sunset St, S...	1	2.99	1.495	2.99
2023-01-09 20:55:00	141255	USB-C Charging Cable	7739134543383	alimentation	"764 11th St, Los...	1	11.95	5.975	11.95

only showing top 20 rows

product	product_id	category	purchase_address	quantity_ordered	price_each	cost_price	turnover	margin	time_of_day	purchase_state
iPhone	5638008983335	vêtements	"944 Walnut St, B...	1	700.0	231.0	700.0	469.0	evening	MA 02215"
Lightning Chargin...	5563319511488	alimentation	"185 Maple St, Po...	1	14.95	7.475	14.95	7.475	afternoon	OR 97035"
Wired Headphones	2113973395220	vêtements	"538 Adams St, Sa...	2	11.99	5.995	23.98	11.99	afternoon	CA 94016"
27in FHD Monitor	3069156759167	sports	"738 10th St, Los...	1	149.99	97.4935	149.99	52.4965	evening	CA 90001"
Wired Headphones	9692680938163	électronique	"387 10th St, Aus...	1	11.99	5.995	11.99	5.995	morning	TX 73301"
AAA Batteries (4-...	2953868554188	alimentation	"775 Willow St, S...	1	2.99	1.495	2.99	1.495	evening	CA 94016"
27in 4K Gaming Mo...	5173670800988	vêtements	"979 Park St, Los...	1	389.99	128.6967	389.99	261.2933	afternoon	CA 90001"
USB-C Charging Cable	8051736777568	vêtements	"181 6th St, San ...	1	11.95	5.975	11.95	5.975	afternoon	CA 94016"
Bose SoundSport H...	1508418177978	électronique	"867 Willow St, L...	1	99.99	49.995	99.99	49.995	morning	CA 90001"
Apple Airpods Hea...	1386344211590	électronique	"657 Johnson St, ...	1	150.0	97.5	150.0	52.5	evening	CA 94016"
Apple Airpods Hea...	4332898830865	vêtements	"492 Walnut St, S...	1	150.0	97.5	150.0	52.5	morning	CA 94016"
Macbook Pro Laptop	1169379570345	vêtements	"322 6th St, San ...	1	1700.0	561.0	1700.0	1139.0	morning	CA 94016"
AAA Batteries (4-...	4436184749366	vêtements	"618 7th St, Los ...	3	2.99	1.495	8.97	4.485	evening	CA 90001"
27in FHD Monitor	7313825995563	vêtements	"512 Wilson St, S...	1	149.99	97.4935	149.99	52.4965	evening	CA 94016"
27in FHD Monitor	9643428300795	alimentation	"440 Cedar St, Po...	1	149.99	97.4935	149.99	52.4965	afternoon	OR 97035"
Vareebadd Phone	6721780072847	alimentation	"471 Center St, L...	1	400.0	132.0	400.0	268.0	morning	CA 90001"
Apple Airpods Hea...	2700099961823	alimentation	"414 Walnut St, B...	1	150.0	97.5	150.0	52.5	morning	MA 02215"
USB-C Charging Cable	3692435232121	sports	"220 9th St, Los ...	1	11.95	5.975	11.95	5.975	morning	CA 90001"
AAA Batteries (4-...	8219536039183	électronique	"238 Sunset St, S...	1	2.99	1.495	2.99	1.495	morning	WA 98101"
USB-C Charging Cable	7739134543383	alimentation	"764 11th St, Los...	1	11.95	5.975	11.95	5.975	evening	CA 90001"