## **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)
df = df.withColumn("order date", to timestamp(col("order date")))
df = df.filter(~hour(col("order date")).between(0, 5))
def get time of day(hour):
    if 5 <= hour < 12:
    elif 12 <= hour < 18:
        return 'afternoon'
    elif 18 <= hour < 24:
time of day udf = udf(get time of day, StringType())
time of day udf(hour(col('order date'))))
```

```
# 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()
```

							↑ ↓	· c> 🗏		
order_date order	_id  product	product_id	category	purc	hase_address	quantity_ordered	rice_each	+  cost_pri	.ce turno	ver  margi
2023-01-22 21:25:00  141  2023-01-28 14:15:00  141	234  iPhone  235 Lightning Chargin	5638008983335  5563319511488 A			nut St, B le St. Po	1	700.0 14.95			0.0  469. .95  7.47
2023-01-17 13:33:00  141  2023-01-05 20:33:00  141					ıms St, Sa h St, Los	2	11.99 149.99			.98   11.9 .99   52.496
2023-01-25 11:59:00 141						1	11.99			.99 5.99
only showing top 5 rows	+	+-						+	+	+
<b></b>		· <del>+</del>	<del>+</del>	+			+	+		<del>-</del>
order_date order	_id  product	t  product_id -+	categor +	ry  p	ourchase_addr	ess quantity_order +	red price	_each co: +	t_price	turnover
		e 5638008983335			Walnut St, B			700.0	231.0	
	235 Lightning Chargin 236  Wired Headphone				Maple St, Po Adams St, Sa			14.95  11.99	7.475 5.995	
	237 27in FHD Monito				10th St, Los			19.99  19.99	97.4935	
	238 Wired Headphone							11.99	5.995	
2023-01-29 20:22:00 141	239 AAA Batteries (4						1	2.99	1.495	
2023-01-26 12:16:00 141	240 27in 4K Gaming Mo	.   5173670800988	Vêtement	ts   "979	Park St, Los		1 3	39.99	28.6967	389.99
2023-01-05 12:04:00 141	241 USB-C Charging Cable	e 8051736777568	vêtement	ts "181	6th St, San		1 :	11.95	5.975	11.95
2023-01-01 10:30:00 141	242 Bose SoundSport H	.   1508418177978	électroniqu	ue "867	Willow St, L		1 !	99.99	49.995	99.99
2023-01-22 21:20:00 141	243 Apple Airpods Hea	.  1386344211590					1 :	150.0	97.5	150.0
	244 Apple Airpods Hea				Walnut St, S			150.0	97.5	
	245  Macbook Pro Laptop				6th St, San		1 1	700.0	561.0	
	246 AAA Batteries (4				7th St, Los		3	2.99	1.495	
	247 27in FHD Monito				Wilson St, S			19.99	97.4935	
	249 27in FHD Monito							19.99	97.4935	
	250 Vareebadd Phone							100.0	132.0	
	251 Apple Airpods Hea							150.0	97.5	
	252 USB-C Charging Cable				9th St, Los		1	11.95	5.975	
	254 AAA Batteries (4						1	2.99	1.495	
2023-01-09 20:55:00 141	255 USB-C Charging Cable	2 7739134543383	alimentatio	on   764	lith St, Los	••••	1	11.95	5.975	11.95
only showing top 20 rows		- <del>,</del>								
mry snowing top 20 rows										

Lightning Chargin   5563319511488   alimentation   "185 Maple St, Po   1   14.95   7.475   14.95   7.475   afternoon   OR 976   Wired Headphones   2113973395220   vētements   "538 Adams St, Sa   2   11.99   5.995   23.98   11.99   afternoon   CA 946   27in FHD Monitor   306915679167   sports   "738 10th St, Los   1   149.99   97.4935   149.99   52.4965   evening   CA 946   Wired Headphones   9692680938163   électronique   "387 10th St, Aus   1   11.99   5.995   11.99   5.995   morning   TX 733   AAA Batteries (4   2953868554188   alimentation   "775   willow St, S   1   2.99   1.495   2.99   1.495   evening   CA 946   27in 4K Gaming Mo   5173670800988   vētements   "379 Park St, Los   1   389.99   128.6967   389.99   261.2933   afternoon   CA 946   27in 4K Gaming Mo   5173670800988   vētements   "181 6th St, San   1   11.95   5.975   11.95   5.975   afternoon   CA 946   27in 4K Gaming Mo   1504818177978   électronique   "867   willow St, L   1   99.99   49.995   99.99   49.995   morning   CA 946   Apple Airpods Hea   1386344211590   électronique   "867   willow St, L   1   150.0   97.5   150.0   52.5   evening   CA 946   Apple Airpods Hea   1386344211590   électronique   "657   Johnson St,   1   150.0   97.5   150.0   52.5   evening   CA 946   Apple Airpods Hea   4332898380865   vētements   "322 6th St, San   1   150.0   97.5   150.0   52.5   evening   CA 946   Apple Airpods Hea   4346184749366   vētements   "618 7th St, Los   1   1700.0   561.0   1700.0   1139.0   morning   CA 946   Apple Airpods Hea   4343684749366   vētements   "512   wilson St, S   1   149.99   97.4935   149.99   52.4965   evening   CA 946   27in FHD Monitor   976434283080795   alimentation   "440   Cedar St, Po   1   149.99   97.4935   149.99   52.4965   evening   CA 946   Apple Airpods Hea   2700099961823   alimentation   "441   Walnut St, B   1   150.0   97.5   150.0   52.5   morning   CA 946   Apple Airpods   160.0   160.0   160.0   160.0   160.0   160.0	product	+   product_id	+    category	ا ا	purchase_address	++  quantity_ordered	price_each	+  cost_price	turnover	margin	time_of_day	purchase	e_state
Wired Headphones   2113973395220   vêtements   "538 Adams St, Sa   2   11.99   5.995   23.98   11.99   afternoon   CA 946   271n FHD Monitor   3069156795167   sports   "738 10th St, Los   1   149.99   97.4935   149.99   52.4965   evening   CA 966   APAB Batteries (4   2953868554188   alimentation   "775 Willow St, S   1   2.99   1.495   2.99   1.495   evening   CA 946   271n AK Gaming Mo   5173670800988   vêtements   "979 Park St, Los   1   389.99   128.6967   389.99   261.2933   afternoon   CA 946   271n AK Gaming Mo   5173670800988   vêtements   "979 Park St, Los   1   389.99   128.6967   389.99   261.2933   afternoon   CA 946   271n AK Gaming Mo   5173670800988   vêtements   "979 Park St, Los   1   11.95   5.975   11.95   5.975   afternoon   CA 946   271n AK Gaming Mo   1   11.95   5.975   11.95   5.975   afternoon   CA 946   271n AK Gaming Mo   1   11.95   5.975   11.95   5.975   afternoon   CA 946   271n AK Gaming Mo   1   11.95   5.975   11.95   5.975   afternoon   CA 946   271n AK Gaming Mo   1   150.0   97.5   150.0   52.5   evening   CA 946   271n FHD Monitor   7313825995563   vêtements   "322 6th St, San   1   150.0   97.5   150.0   52.5   evening   CA 946   271n FHD Monitor   9643428300795   alimentation   "440 Cedar St, Po   1   149.99   97.4935   149.99   52.4965   evening   CA 946   271n FHD Monitor   9643428300795   alimentation   "440 Cedar St, Po   1   149.99   97.4935   149.99   52.4965   evening   CA 946   271n FHD Monitor   9643428300795   alimentation   "441 Walnut St, B   1   140.0   132.0   400.0   268.0   morning   CA 946   271n FHD Monitor   9643428300795   alimentation   "441 Walnut St, B   1   140.0   97.5   150.0   52.5   morning   CA 946   400.0	iPhone	+  5638008983335	+   vêtements	"944	Walnut St, B	+	700.0	+   231.0	700.0	469.0	evening	MA	02215"
27in FHD Monitor   3069156759167   sports   "738 10th St, Los   1   149.99   97.4935   149.99   52.4965   evening   CA 906   Wired Headphones   9692680938163   électronique   "387 10th St, Aus   1   11.99   5.995   11.99   5.995   11.99   5.995   morning   TX 733   AAA Batteries (4  2953868554188   alimentation   "775 Willow St, S   1   2.99   1.495   2.99   1.495   evening   CA 906   27in 4K Gaming Mo   5173670800988   vētements   "979 Park St, Los   1   389.99   128.6967   389.99   261.2933   afternoon   CA 906   USB-C Charging Cable   8051736777588   vētements   "181 6th St, San   1   11.95   5.975   11.95   5.975   afternoon   CA 906   CA 90	Lightning Chargin	5563319511488	alimentation	"185	Maple St, Po	j 1	14.95	7.475	14.95	7.475	afternoon	OR	97035"
Wired Headphones   9692680938163   électronique   "387 10th St, Aus	Wired Headphones	2113973395220	vêtements	"538	Adams St, Sa	2	11.99	5.995	23.98	11.99	afternoon	CA	94016"
AAA Batteries (4   2953868554188   alimentation   "775 Willow St, S   1   2.99   1.495   2.99   1.495   evening   CA 946   27in FHD Monitor   9643428300795   alternation   "476 Cedar St, Po   1   2.99   1.495   2.99   1.495   evening   CA 946   27in FHD Monitor   9643428300795   alternation   "471 Center St, L   1   2.99   1.495   2.99   1.495   evening   CA 946   27in FHD Monitor   96128722121   sports   "220 9th St, L   1   2.99   1.495   2.99   1.495   evening   CA 946   27in FHD Monitor   97.5150.0   52.51   evening   CA 946   27in FHD Monitor   97.5150.0   52.51   evening   CA 946   27in FHD Monitor   97.5150.0   52.51   evening   CA 946   27in FHD Monitor   97.5150.0   97	27in FHD Monitor	3069156759167	sports	"738	10th St, Los	j 1	149.99	97.4935	149.99	52.4965	evening	CA	90001"
27in 4K Gaming Mo   5173670800988   vêtements   "979 Park St, Los   1   389.99   128.6967   389.99   261.2933   afternoon   CA 906   USB-C Charging Cable   8051736777568   vêtements   "181 6th St, San   1   11.95   5.975   11.95   5.975   11.95   5.975   afternoon   CA 946   Soss SoundSport H   1508418179798   électronique   "657 Johnson St,   1   150.0   97.5   150.0   52.5   evening   CA 946   Apple Airpods Hea   1386344211590   électronique   "657 Johnson St,   1   150.0   97.5   150.0   52.5   evening   CA 946   Apple Airpods Hea   4332898830865   vêtements   "492 Walnut St, S   1   150.0   97.5   150.0   52.5   evening   CA 946   AAA Batteries (4   4436184749366   vêtements   "512 Wilson St, S   1   1700.0   561.0   1700.0   1139.0   morning   CA 946   APPLE AIRPORTING   APPL	Wired Headphones	9692680938163	électronique	"387	10th St, Aus	1	11.99	5.995	11.99	5.995	morning	TX	73301"
USB-C Charging Cable 8051736777568   vêtements "181 6th St, San   1   11.95   5.975   11.95   5.975   afternoon   CA 946   Bose SoundSport H   1508418177978   électronique   "867 Willow St, L   1   99.99   49.995   99.99   49.995   morning   CA 946   Apple Airpods Hea   4332898830865   vêtements   "492 Walnut St, S   1   150.0   97.5   150.0   52.5   morning   CA 946   Macbook Pro Laptop   1169379570345   vêtements   "322 6th St, San   1   1700.0   561.0   1700.0   1139.0   morning   CA 946   AAA Batteries (4   4436184749366   vêtements   "618 7th St, Los   3   2.99   1.495   8.97   4.485   evening   CA 946   27in FHD Monitor   731382599563   vêtements   "512 Wilson St, S   1   149.99   97.4935   149.99   52.4965   evening   CA 946   27in FHD Monitor   9643428300795   alimentation   "440 Cedar St, Po   1   149.99   97.4935   149.99   52.4965   afternoon   OR 976   Vareebadd Phone   6721780072847   alimentation   "471 Center St, L   1   400.0   132.0   400.0   268.0   morning   CA 946   Apple Airpods Hea   27000909961823   alimentation   "414 Walnut St, B   1   150.0   97.5   150.0   52.5   morning   CA 946   AAA Batteries (4   8219536039183   électronique   "238 Sunset St, S   1   11.95   5.975   11.95   5.975   morning   CA 946   AAA Batteries (4   8219536039183   électronique   "238 Sunset St, S   1   2.99   1.495   2.99   1.495   morning   WA 981	AAA Batteries (4	2953868554188	alimentation	"775	Willow St, S	1	2.99	1.495	2.99	1.495	evening	CA	94016"
Bose SoundSport H   1508418177978   électronique   "867 Willow St, L   1   99.99   49.995   99.99   49.995   morning   CA 906   Apple Airpods Hea   1386344211590   électronique   "657 Johnson St,   1   150.0   97.5   150.0   52.5   evening   CA 946   Apple Airpods Hea   4332898838865   vêtements   "392 Walnut St, S   1   150.0   97.5   150.0   52.5   morning   CA 946   Macbook Pro Laptop   1169379570345   vêtements   "322 6th St, San   1   1700.0   561.0   1700.0   1139.0   morning   CA 946   AAA Batteries (4   4436184749366   vêtements   "618 7th St, Los   3   2.99   1.495   8.97   4.485   evening   CA 946   27in FHD Monitor   7313825995563   vêtements   "512 Wilson St, S   1   149.99   97.4935   149.99   52.4965   evening   CA 946   27in FHD Monitor   9643428300795   alimentation   "440 Cedar St, Po   1   149.99   97.4935   149.99   52.4965   afternoon   OR 976   Vareebadd Phone   6721780072847   alimentation   "471 Center St, L   1   400.0   132.0   400.0   268.0   morning   CA 946   Apple Airpods Hea   2706099961823   alimentation   "414 Walnut St, B   1   150.0   97.5   150.0   52.5   morning   CA 946   AAA Batteries (4   8219536039183   électronique   "238 Sunset St, S   1   2.99   1.495   2.99   1.495   morning   WA 981   Apple   Airpods Hea   27060393183   électronique   "238 Sunset St, S   1   2.99   1.495   2.99   1.495   morning   WA 981   Apple   Airpods Hea   27060393183   électronique   "238 Sunset St, S   1   2.99   1.495   2.99   1.495   morning   WA 981   Apple   Airpods Hea   27060393183   électronique   "238 Sunset St, S   1   2.99   1.495   2.99   1.495   morning   WA 981   Apple   Airpods Hea   2706039183   électronique   "238 Sunset St, S   1   2.99   1.495   2.99   1.495   morning   WA 981   Apple   Airpods Hea   2706039183   électronique   "238 Sunset St, S   1   2.99   1.495   2.99   1.495   morning   WA 981   Apple   2.99   2.495   2.495   2.495   2.495   2.495   2.495   2.495   2.495   2.495   2.495	27in 4K Gaming Mo	5173670800988	vêtements	"979	Park St, Los	1	389.99	128.6967	389.99	261.2933	afternoon	CA	90001"
Apple Airpods Hea   1386344211590   électronique   "657 Johnson St,         1   150.0         97.5         150.0         52.5         evening   CA 946         Apple Airpods Hea   4332898380865   vêtements   "492 Walnut St, S         1   150.0         97.5         150.0         52.5         evening   CA 946         Macbook Pro Laptop   1169379570345   vêtements   "322 6th St, San         1   1700.0         561.0         1700.0         131.9         morning   CA 946         AAA Batteries (4   4436184749366   vêtements   "512 Wilson St, S         3   2.99         1.495         8.97         4.485         evening   CA 946         27in FHD Monitor   7313825995563   vêtements   "512 Wilson St, S         1   149.99         97.4935         149.99         52.4965         evening   CA 946         27in FHD Monitor   96434283800795   alimentation   "440 Cedar St, Po         1   149.99         97.4935         149.99         52.4965         evening   CA 946         Vareebadd Phone   6721780072847   alimentation   "471 Center St, L         1   400.0         132.0         400.0         268.0         morning   CA 946         Apple Airpods Hea   2700099961823   alimentation   "414 Walnut St, B         1   150.0         97.5         150.0         52.5         morning   MA 022         USB-C Charging Cable   3692435523121   soprits   "320         360394852121         soprits   "3209	USB-C Charging Cable	8051736777568	vêtements	"181	6th St, San	1	11.95	5.975	11.95	5.975	afternoon	CA	94016"
Apple Airpods Hea   4332898830865   vêtements   "492 Walnut St, S   1   150.0   97.5   150.0   52.5   morning   CA 946   Macbook Pro Laptop   1169379570345   vêtements   "322 6th St, San   1   1700.0   561.0   1700.0   1139.0   morning   CA 946   AAA Batteries (4   4436184749366   vêtements   "512 Wilson St, S   3   2.99   1.495   8.97   4.485   evening   CA 946   27in FHD Monitor   7313855995563   vêtements   "512 Wilson St, S   1   149.99   97.4935   149.99   52.4965   evening   CA 946   27in FHD Monitor   9643428300795   alimentation   "440 Cedar St, Po   1   149.99   97.4935   149.99   52.4965   afternoon   OR 976   Vareebadd Phone   6721780072847   alimentation   "471 Center St, L   1   400.0   132.0   400.0   268.0   morning   CA 906   APPLE Airpods Hea   2700099961823   alimentation   "414 Walnut St, B   1   150.0   97.5   150.0   52.5   morning   MA 022   USB-C Charging Cable   3692435232121   sports   "220 9th St, Los   1   11.95   5.975   11.95   5.975   morning   CA 906   AAA Batteries (4   8219536039183   électronique   "238 Sunset St, S   1   2.99   1.495   2.99   1.495   morning   WA 981   400.0   40	Bose SoundSport H	1508418177978	électronique	"867	Willow St, L	1	99.99	49.995	99.99	49.995	morning	CA	90001"
Macbook Pro Laptop   1169379570345   vêtements   "322 6th St, San         1   1700.0   561.0   1700.0   1139.0   morning   CA 946         AAA Batteries (4   4436184749366   vêtements   "618 7th St, Los         3   2.99   1.495   8.97   4.485   evening   CA 946         27in FHD Monitor   7313825995563   vêtements   "512 Wilson St, S         1   149.99   97.4935   149.99   52.4965   evening   CA 946         27in FHD Monitor   9643428300795   alimentation   "440 Cedar St, Po         1   149.99   97.4935   149.99   52.4965   afternoon   OR 976         Vareebadd Phone   6721780072847   alimentation   "471 Center St, L         1   400.0   132.0   400.0   268.0   morning   CA 906         Apple Airpods Hea   27000099961823   alimentation   "414 Walnut St, B         1   150.0   97.5   150.0   52.5   morning   CA 906         AAA Batteries (4   8219536039183   électronique   "238 Sunset St, S         1   11.95   5.975   11.95   5.975   morning   WA 981	Apple Airpods Hea	1386344211590	électronique	"657	Johnson St,	j 1	150.0	97.5	150.0	52.5	evening	CA	94016"
AAA Batteries (4 4436184749366  vētements "618 7th St, Los  3  2.99  1.495  8.97  4.485  evening  CA 906	Apple Airpods Hea	4332898830865	vêtements	"492	Walnut St, S	1	150.0	97.5	150.0	52.5	morning	CA	94016"
27in FHD Monitor   7313825995563   vêtements   "512 Wilson St, S         1   149.99   97.4935   149.99   52.4965   evening   CA 946   27in FHD Monitor   9643428300795   alimentation   "440 Cedar St, Po         1   149.99   97.4935   149.99   52.4965   evening   CA 946   27in FHD Monitor   9643428300795   alimentation   "440 Cedar St, Po         1   149.99   97.4935   149.99   52.4965   evening   CA 946   27in FHD Monitor   97.50   149.99   126.90   27in FHD Monitor   140.00   132.0   400.0   268.0   morning   CA 946   27in FHD Monitor   140.00   27in FHD Monitor   150.0   97.5   150.0   52.5   morning   MA 922   140.00	Macbook Pro Laptop	1169379570345	vêtements	"322	6th St, San	1	1700.0	561.0	1700.0	1139.0	morning	CA	94016"
27in FHD Monitor   9643428300795   alimentation   "440 Cedar St, Po         1   149.99   97.4935   149.99   52.4965   afternoon         0R 976   9	AAA Batteries (4	4436184749366	vêtements	"618	7th St, Los	3	2.99	1.495	8.97	4.485	evening	CA	90001"
Vareebadd Phone 6721780072847 alimentation "471 Center St, L  1  400.0  132.0  400.0  268.0  morning  CA 900	27in FHD Monitor	7313825995563	vêtements	"512	Wilson St, S	1	149.99	97.4935	149.99	52.4965	evening	CA	94016"
Apple Airpods Hea   2700099961823   alimentation   "414 Walnut St, B         1   150.0         97.5         150.0         52.5         morning         MA 022         USB-C Charging Cable   3692435232121         sports   "220 9th St, Los         1   11.95         5.975         11.95         5.975         morning         CA 900           AAA Batteries (4   8219536039183   électronique   "238 Sunset St, S         1   2.99         1.495         2.99         1.495         morning         WA 981	27in FHD Monitor	9643428300795	alimentation	"440	Cedar St, Po	1	149.99	97.4935	149.99	52.4965	afternoon	OR	97035"
USB-C Charging Cable 3692435232121   sports 220 9th St, Los   1   11.95   5.975   11.95   5.975   morning   CA 900   AAA Batteries (4   8219536039183   électronique 238 Sunset St, S   1   2.99   1.495   2.99   1.495   morning   WA 981	Vareebadd Phone	6721780072847	alimentation	"471	Center St, L	1	400.0	132.0	400.0	268.0	morning	CA	90001"
AAA Batteries (4 8219536039183 électronique "238 Sunset St, S  1  2.99  1.495  2.99  1.495  morning  WA 981	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"
USB-C Charging Cable 7739134543383 alimentation 764 11th St, Los  1   11.95   5.975   11.95   5.975   evening   CA 900	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"
	<b>}</b>		+			++		+					

