DATA BANK

BY ANUSH BHARATHWAJ L

weekly_sales

weekly_sales.csv

week_date	region	platform	segment	customer_type	transactions	sales
2020-08-31	ASIA	Retail	C3	New	120631	3656163
2020-08-31	ASIA	Retail	F1	New	31574	996575
2020-08-31	USA	Retail	null	Guest	529151	16509610
2020-08-31	EUROPE	Retail	C1	New	4517	141942
2020-08-31	AFRICA	Retail	C2	New	58046	1758388
2020-08-31	CANADA	Shopify	F2	Existing	1336	243878
2020-08-31	AFRICA	Shopify	F3	Existing	2514	519502
2020-08-31	ASIA	Shopify	F1	Existing	2158	371417
2020-08-31	AFRICA	Shopify	F2	New	318	49557
2020-08-31	AFRICA	Retail	C3	New	111032	3888162
2020-08-31	USA	Shopify	F1	Existing	1398	260773
2020-08-31	OCEANIA	Shopify	C2	Existing	4661	882690
2020-08-31	SOUTH AMERICA	Retail	C2	Existing	1029	38762
2020-08-31	SOUTH AMERICA	Shopify	C4	New	6	917
2020-08-31	EUROPE	Shopify	F3	Existing	115	35215
2020-08-31	OCEANIA	Retail	F3	Existing	551905	30371770
2020-08-31	ASIA	Shopify	C3	Existing	1969	374327
2020-08-31	AFRICA	Retail	F1	Existing	97604	5185233
2020-08-31	OCEANIA	Retail	C2	New	111219	2980673
2020-08-31	USA	Retail	F1	New	11820	463738
2020-08-31	SOUTH AMERICA	Retail	F3	Existing	1363	65730
2020-08-31	AFRICA	Retail	C3	Existing	284971	14430196
2020-08-31	ASIA	Retail	F2	New	70496	2176980
2020-08-31	AFRICA	Shopify	F1	Existing	2678	478756
2020-08-31	USA	Shopify	C4	New	22	3319
2020-08-31	CANADA	Retail	F3	Existing	94274	5306746
2020-08-31	ASIA	Retail	F1	Existing	94287	4511841
2020-08-31	EUROPE	Retail	null	New	3064	134249
2020-08-31	EUROPE	Shopify	F1	New	7	1579
2020-08-31	SOUTH AMERICA	Retail	C4	New	329	11451
2020-08-31	SOUTH AMERICA	Retail	F1	Existing	854	31589
2020-08-31	EUROPE	Shopify	C2	Existing	180	53567
2020-08-31	EUROPE	Shopify	F2	New	15	4023
2020-08-31	AFRICA	Retail	C2	Existing	112361	4768214
2020-08-31	ASIA	Shopify	C2	Existing	2269	396909
2020-08-31	AFRICA	Shopify	C4	New	58	8562

DuckDB 13 ms (1 hour ago) 7 columns · 17,117 rows

i) Data Cleansing Steps

In a single query, perform the following operations and generate a new table in the data_mart schema named clean_weekly_sales

1. **Add a week_number as the second column for each week_date value, for example any value from the 1st of January to 7th of January will be 1, 8th to 14th will be 2, etc.**
2. **Add a month_number with the calendar month for each week_date value as the 3rd column.**
3. **Add a calendar_year column as the 4th column containing either 2018, 2019 or 2020 values.**
4. **Add a new column called age_band after the original segment column using the following mapping on the number inside the segment value.**

segment	age_band
1	Young Adults
2	Middle Aged
3 or 4	Retirees

5. **Add a new demographic column using the following mapping for the first letter in the segment values:**

segment | demographic |

C | Couples |

F | Families |

6. **Ensure all null string values with an "unknown" string value in the original segment column as well as the new age_band and demographic columns.**

7. **Generate a new avg_transaction column as the sales value divided by transactions rounded to 2 decimal places for each record.**

Implementation of Data Cleaning

Cleansing The Data

```
create table
  clean_weekly_sales as
select
  week_date,
  week (week_date) as week_number,
  month (week_date) as month_number,
  monthname (week_date) as month_name,
  year (week_date) as calendar_year,
  case
    when segment = null then Unknown
    else segment
  end as segment,
  case
    when Right(segment, 1) = 1 then Young Adults
    when Right(segment, 1) = 2 then Middle Aged
    when Right(segment, 1) in (3, 4) then Retirees
    else Unknown
  end as age_band,
  case
    when Left(segment, 1) = C then Couples
    when Left(segment, 1) = F then Families
    else Unknown
  end as demographic,
  platform,
  region,
  Round(sales / transactions, 2) as avg_transaction,
  transactions,
  sales
from
  weekly_sales;
```

Display

```
select * from clean_weekly_sales;
```

clean_weekly_sales

clean_weekly_sales.csv

week_date	week_number	month_number	month_name	calendar_year	segment	age_band	demographic	platform	region	avg_transaction	transactions	sales
2020-08-31	35	8	August	2020	C3	Retirees	Couples	Retail	ASIA	30.31	120631	3656163
2020-08-31	35	8	August	2020	F1	Young Adults	Families	Retail	ASIA	31.56	31574	996575
2020-08-31	35	8	August	2020	Unknown	Unknown	Unknown	Retail	USA	31.2	529151	16509610
2020-08-31	35	8	August	2020	C1	Young Adults	Couples	Retail	EUROPE	31.42	4517	141942
2020-08-31	35	8	August	2020	C2	Middle Aged	Couples	Retail	AFRICA	30.29	58046	1758388
2020-08-31	35	8	August	2020	F2	Middle Aged	Families	Shopify	CANADA	182.54	1336	243878
2020-08-31	35	8	August	2020	F3	Retirees	Families	Shopify	AFRICA	206.64	2514	519502
2020-08-31	35	8	August	2020	F1	Young Adults	Families	Shopify	ASIA	172.11	2158	371417
2020-08-31	35	8	August	2020	F2	Middle Aged	Families	Shopify	AFRICA	155.84	318	49557
2020-08-31	35	8	August	2020	C3	Retirees	Couples	Retail	AFRICA	35.02	111032	3888162
2020-08-31	35	8	August	2020	F1	Young Adults	Families	Shopify	USA	186.53	1398	260773
2020-08-31	35	8	August	2020	C2	Middle Aged	Couples	Shopify	OCEANIA	189.38	4661	882690
2020-08-31	35	8	August	2020	C2	Middle Aged	Couples	Retail	SOUTH AMERICA	37.67	1029	38762
2020-08-31	35	8	August	2020	C4	Retirees	Couples	Shopify	SOUTH AMERICA	152.83	6	917
2020-08-31	35	8	August	2020	F3	Retirees	Families	Shopify	EUROPE	306.22	115	35215
2020-08-31	35	8	August	2020	F3	Retirees	Families	Retail	OCEANIA	55.03	551905	30371770
2020-08-31	35	8	August	2020	C3	Retirees	Couples	Shopify	ASIA	190.11	1969	374327
2020-08-31	35	8	August	2020	F1	Young Adults	Families	Retail	AFRICA	53.13	97604	5185233
2020-08-31	35	8	August	2020	C2	Middle Aged	Couples	Retail	OCEANIA	26.8	111219	2980673
2020-08-31	35	8	August	2020	F1	Young Adults	Families	Retail	USA	39.23	11820	463738
2020-08-31	35	8	August	2020	F3	Retirees	Families	Retail	SOUTH AMERICA	48.22	1363	65730
2020-08-31	35	8	August	2020	C3	Retirees	Couples	Retail	AFRICA	50.64	284971	14430196
2020-08-31	35	8	August	2020	F2	Middle Aged	Families	Retail	ASIA	30.88	70496	2176980
2020-08-31	35	8	August	2020	F1	Young Adults	Families	Shopify	AFRICA	178.77	2678	478756
2020-08-31	35	8	August	2020	C4	Retirees	Couples	Shopify	USA	150.86	22	3319
2020-08-31	35	8	August	2020	F3	Retirees	Families	Retail	CANADA	56.29	94274	5306746
2020-08-31	35	8	August	2020	F1	Young Adults	Families	Retail	ASIA	47.85	94287	4511841
2020-08-31	35	8	August	2020	Unknown	Unknown	Unknown	Retail	EUROPE	43.81	3064	134249
2020-08-31	35	8	August	2020	F1	Young Adults	Families	Shopify	EUROPE	225.57	7	1579
2020-08-31	35	8	August	2020	C4	Retirees	Couples	Retail	SOUTH AMERICA	34.81	329	11451
2020-08-31	35	8	August	2020	F1	Young Adults	Families	Retail	SOUTH AMERICA	36.99	854	31589
2020-08-31	35	8	August	2020	C2	Middle Aged	Couples	Shopify	EUROPE	297.59	180	53567
2020-08-31	35	8	August	2020	F2	Middle Aged	Families	Shopify	EUROPE	268.2	15	4023
2020-08-31	35	8	August	2020	C2	Middle Aged	Couples	Retail	AFRICA	42.44	112361	4768214
2020-08-31	35	8	August	2020	C2	Middle Aged	Couples	Shopify	ASIA	174.93	2269	396909
2020-08-31	35	8	August	2020	C4	Retirees	Couples	Shopify	AFRICA	147.62	58	8562
2020-08-31	35	8	August	2020	F3	Retirees	Families	Retail	USA	61.05	142898	8723663
2020-08-31	35	8	August	2020	C3	Retirees	Couples	Shopify	OCEANIA	203.69	4703	957939

DuckDB · 11 ms (53 minutes ago) · 13 columns · 17,117 rows

ii) Data Exploration

A) Which week numbers are missing from the dataset?

Sequence 52

```
create table
  seq100 (x int auto_increment primary key);
insert into
  seq100
values
  (),(),(),(),(),(),(),(),(),();
insert into
  seq100
values
  (),(),(),(),(),(),(),(),(),();
insert into
  seq100
values
  (),(),(),(),(),(),(),(),(),();
insert into
  seq100
values
  (),(),(),(),(),(),(),(),(),();
insert into
  seq100
values
  (),(),(),(),(),(),(),(),(),();
insert into
  seq100
values
  (),(),(),(),(),(),(),(),(),();
insert into
  seq100
select
  x + 50
from
  seq100;
select
  *
from
  seq100;
create table
  seq52 as
select
  x
from
  seq100
limit
  52;
select
  *
from
  seq52;
select
  x as 'Miss_week_numbers'
from
  seq52
where
  x not in (
    select distinct
      week_number
    from
      clean_weekly_sales
  );
```

Missing Week Number

seq52.csv

x
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44

DuckDB · 1 ms (31 minutes ago) · 1 column · 52 rows

B) How many total transactions were there for each year in the dataset?

Total Transaction By Year

```
select
  calendar_year as Year,
  sum(avg_transaction) as Total_Transactions
from
  clean_weekly_sales
group by
  Year;
```

Year	Total_Transactions
2018	657630.280000003
2019	626260.68
2020	615332.540000001

DuckDB 9 ms (25 minutes ago) 2 columns · 3 rows

C) What are the total sales for each region for each month?

Regional Sales Performance (Monthly)

```
select month_name , month_number ,region ,sum(sales) from clean_weekly_sales group
by month_name,
month_number,region;
```

month_name	month_number	region	sum(sales)
August	8	ASIA	1663320609
August	8	USA	712002790
August	8	EUROPE	122102995
August	8	AFRICA	1809596890
August	8	CANADA	447073019
August	8	OCEANIA	2432313652
August	8	SOUTH AMERICA	221166052
July	7	AFRICA	1960219710
July	7	CANADA	477134947
July	7	USA	760331754
July	7	EUROPE	136757466
July	7	OCEANIA	2563459400
July	7	SOUTH AMERICA	235582776
July	7	ASIA	1768844756
June	6	OCEANIA	2371884744
June	6	USA	703878990
June	6	SOUTH AMERICA	218247455
June	6	EUROPE	122813826
June	6	ASIA	1619482889
June	6	CANADA	443846698
June	6	AFRICA	1767559760
May	5	EUROPE	109338389
May	5	USA	655967121
May	5	SOUTH AMERICA	201391809

DuckDB 12 ms (24 minutes ago) 4 columns · 50 rows

D) What is the total count of transactions for each platform?

Transaction Counts by Platform

```
select platform , count(transactions) from clean_weekly_sales group by platform;
```

platform	count(transactions)
Retail	8568
Shopify	8549

DuckDB 11 ms (21 minutes ago) 2 columns · 2 rows

Transaction Sum by Platform

```
select platform , sum(transactions) from clean_weekly_sales group by platform;
```

platform	sum(transactions)
Retail	1081934227
Shopify	5925169

DuckDB 12 ms (20 minutes ago) 2 columns · 2 rows

E) What is the percentage of sales for Retail vs Shopify for each month?

Monthly Sales Share: Retail vs Shopify

```
WITH cte_monthly_sales AS (
  SELECT
    month_number,
    calendar_year,
    platform,
    SUM(sales) AS sales -- Use "sales" instead of "saless" for consistency
  FROM clean_weekly_sales
  GROUP BY month_number, calendar_year, platform
)
SELECT
  month_number,
  calendar_year,
  ROUND(100 * MAX(CASE WHEN platform = 'Retail' THEN sales ELSE NULL END) / SUM
(sales), 2) AS retail_perc,
  ROUND(100 * MAX(CASE WHEN platform = 'Shopify' THEN sales ELSE NULL END) / SUM
(sales), 2) AS shopify_perc
FROM cte_monthly_sales
GROUP BY month_number, calendar_year
ORDER BY month_number, calendar_year;
```

month_number	calendar_year	retail_perc	shopify_perc
3	2018	97.92	2.08
3	2019	97.71	2.29
3	2020	97.3	2.7
4	2018	97.93	2.07
4	2019	97.8	2.2
4	2020	96.96	3.04
5	2018	97.73	2.27
5	2019	97.52	2.48

DuckDB 28 ms (17 minutes ago) 4 columns · 20 rows

F) What is the percentage of sales by demographic for each year in the dataset?

Sales by Demographic (Yearly Breakdown)

```
select calendar_year, demographic , sum(sales) as year_sales, round( 100 * sum
(sales)/sum(sum(sales)) over(partition by demographic),2)as perc from
clean_weekly_sales
group by calendar_year, demographic order by calendar_year, demographic;
```

calendar_year	demographic	year_sales	perc
2018	Couples	3402388688	30.38
2018	Families	4125558033	31.25
2018	Unknown	5369434106	32.86
2019	Couples	3749251935	33.47
2019	Families	4463918344	33.81
2019	Unknown	5532862221	33.86
2020	Couples	4049566928	36.15
2020	Families	4614338065	34.95
2020	Unknown	5436315907	33.27

DuckDB 28 ms (14 minutes ago) 4 columns · 9 rows

G) Which age_band and demographic values contribute the most to Retail sales?

Top Contributors to Retail Sales (Age Band & Demographic)

```
select
  age_band,
  demographic,
  sum(sales) as Sales
from
  clean_weekly_sales
where
  platform = 'Retail'
group by
  age_band,
  demographic
order by
  Sales desc;
```

age_band	demographic	Sales
Unknown	Unknown	16067285533
Retirees	Families	6634686916
Retirees	Couples	6370580014
Middle Aged	Families	4354091554
Young Adults	Couples	2602922797
Middle Aged	Couples	1854160330
Young Adults	Families	1770889293

DuckDB 5 ms (12 minutes ago) 3 columns · 7 rows

Sources And Refrences

Query Dataset: https://drive.google.com/drive/folders/1-dulIJKUcYr8jP4K3hGkcPVH5NO-xKj6

CSV Dataset: https://drive.google.com/drive/folders/1XcnocuoElWEC6w1iNwUOSgw0SARQ-HHx?usp=sharing

My Portfolio: https://anushbharathwaj.w3spaces.com/

Report: https://drive.google.com/file/d/129StXQElmEp_EIxeaps4l9sGlcH4lnxa/view?usp=sharing