Truemeds - Asst - Business Analyst

Q1. SQL Query:

-Retention Rate:
Refers to the percentage of customers who continue paying for a product over a given timeframe.
-So here we want the quaterly data and the order status is also fixed as “55” ,so first we will be grouping customers based on these quarters and order status.
-Then query calculates customer retention by joining to find instances where a customer who made a purchase in a given quarter (of the previous year) made another purchase in a subsequent quarter (Retention CTE). This allows the query to track customer retention over consecutive quarters.
-The final statement aggregates and counts the number of unique customers retained across these quarters, providing insights into customer retention trends over time. This analysis can help businesses understand patterns in repeat purchases.

```
WITH Orders AS (
   SELECT
      CustomerID,
      YEAR(CreatedOn) AS Yr,
      QUARTER(CreatedOn) AS Qtr
   FROM OrderDetails
   WHERE OrderStatus = '55'
   GROUP BY CustomerID, YEAR(CreatedOn), QUARTER(CreatedOn)
),

Retention AS (
   SELECT
      o1.CustomerID,
      o1.Yr AS JoinYr,
      o1.Qtr AS JoinQtr,
      o2.Yr AS RetainYr,
      o2.Qtr AS RetainQtr
   FROM Orders o1
   JOIN Orders o2
      ON o1.CustomerID = o2.CustomerID
      AND (o2.Yr > o1.Yr OR (o2.Yr = o1.Yr AND o2.Qtr > o1.Qtr))
   WHERE o1.Yr = YEAR(CURDATE()) - 1
)

SELECT
   JoinYr,
   JoinQtr,
   RetainYr,
   RetainQtr,
   COUNT(DISTINCT CustomerID) AS RetainedCust
FROM Retention
GROUP BY JoinYr, JoinQtr, RetainYr, RetainQtr
ORDER BY JoinYr, JoinQtr, RetainYr, RetainQtr;
```

So after we got the table which is in format as in pic below:

So after regaining the table from here,we can use Bi tools such as Powerbi ,Tableau or Simple Excel pivot tables for visualization,as we don’t have the visualization availble in Mysql. I will be mainly using powerbi for this purpose and will create visualization .

Here apart from making the visualization using this data,we can also add slicers and interactions to make our graphs make more appealing and can get detailed insights from it.

Q2.Business analysis & insight generation:

-*Data Preparation and cleaning :*

Initially, as we examine the types of data in the dataset, we discover that the column "Parameters" has JSON data format.
I have thus used the ***Excel power query*** to extend these columns.
Additionally, the dataset is currently in the format seen in the image:

AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS
Device Model	has_coupon_code	selling_price_total_amount	discount_amount	no_of_items	is_switch_added	af_currency	packaging_charge_amount	is_addons_added	af_revenue	is_core_customer	mrp_total_amount	estimated_payable_amount	reposr
OPPO::CPH2127	FALSE	2307.41	816.79	17	TRUE	INR	11	TRUE	2318.41	FALSE	3124.2	2318.41	1.5E+07
Redmi::Redmi 8A Dual	FALSE	401.8	445.2	1	TRUE	INR	11	FALSE	412.8	TRUE	847	412.8	1.5E+07
samsung::SM-A042F	TRUE	239.4	240.6	1	FALSE	INR	11	FALSE	289.4	TRUE	480	289.4	1.5E+07
samsung::SM-M136B	FALSE	559.61	230.98	6	FALSE	INR	11	TRUE	570.61	TRUE	790.59	570.61	1.5E+07
samsung::SM-A505F	TRUE				FALSE	INR		FALSE		FALSE			1.5E+07
vivo::vivo 1938	FALSE	594.4	148.6	1	FALSE	INR	11	FALSE	605.4	FALSE	743	605.4	1.5E+07
OnePlus::EB2101	FALSE	274.2	64.8	4	FALSE	INR	11	TRUE	324.2	FALSE	339	324.2	1.5E+07
lge::LM-G850	FALSE	1065.6	503.4	5	TRUE	INR	11	FALSE	1047.65	TRUE	1569	1047.65	1.5E+07
OPPO::CPH2251	FALSE	257.09	147.91	3	TRUE	INR	11	TRUE	307.09	TRUE	405	307.09	1.5E+07
HONOR::REA-NX9	FALSE	4179.6	2914.44	5	TRUE	INR	11	TRUE	4090.6	TRUE	7094.04	4090.6	1.5E+07
Redmi::220333QBI	FALSE	2352.36	588.12	3	FALSE	INR	11	FALSE	2216.34	TRUE	2940.48	2216.34	1.5E+07
samsung::SM-A336E	FALSE	785.08	196.28	3	FALSE	INR	11	FALSE	796.08	TRUE	981.36	796.08	1.5E+07
samsung::SM-F415F	FALSE	315	295.3	3	TRUE	INR	11	FALSE	365	TRUE	610.3	365	1.5E+07
POCO::211033MI	FALSE	39.05	8	1	FALSE	INR	11	TRUE	99.05	FALSE	47.05	99.05	1.5E+07
xiaomi::Redmi Note 7	FALSE	1353.38	338.34	4	FALSE	INR	11	TRUE	1278.96	TRUE	1691.72	1278.96	1.5E+07
samsung::SM-A305F	FALSE	1435.82	386.07	9	TRUE	INR	11	TRUE	1431.5	TRUE	1821.89	1431.5	1.5E+07
POCO::2201116PI	FALSE	778.34	244.66	3	TRUE	INR	11	TRUE	789.34	FALSE	1023	789.34	1.5E+07
samsung::SM-G998B	TRUE	1746.16	436.54	7	FALSE	INR	11	FALSE	1648.02	TRUE	2182.7	1648.02	1.5E+07
vivo::V2130	FALSE	162.3	40.57	1	FALSE	INR	11	FALSE	212.3	FALSE	202.87	212.3	1.5E+07
OPPO::CPH2269	FALSE	515.15	125.83	3	FALSE	INR	11	TRUE	575.15	FALSE	640.98	575.15	1.5E+07
iQOO::I2017	TRUE	1996.68	667.05	3	TRUE	INR	11	FALSE	1908.27	TRUE	2663.73	1908.27	1.5E+07
OnePlus::CPH2423	FALSE	45.7	11.42	1	FALSE	INR	11	FALSE	95.7	FALSE	57.12	95.7	1.5E+07
realme::RMX1992	FALSE	500.03	125.01	3	FALSE	INR	11	FALSE	511.03	TRUE	625.04	511.03	9430302
realme::RMX3261	FALSE	153.32	38.33	2	FALSE	INR	11	TRUE	203.32	TRUE	191.65	203.32	1.5E+07
OPPO::CPH2269	FALSE	515.15	125.83	3	FALSE	INR	11	TRUE	575.15	FALSE	640.98	575.15	1.5E+07
Nothing::AIN065	FALSE	692.56	164.44	2	FALSE	INR	11	TRUE	703.58	TRUE	857	703.58	1.5E+07
Redmi::M2010J19SI	FALSE	1067.91	202.09	2	FALSE	INR	11	TRUE	1067.66	FALSE	1270	1067.66	1.5E+07
xiaomi::Redmi Note 7 Pro	FALSE	430.56	107.64	1	FALSE	INR	11	FALSE	441.56	FALSE	538.2	441.56	1.4E+07
Redmi::22120RN86I	FALSE	777.6	780.9	2	FALSE	INR	11	FALSE	788.6	TRUE	1558.5	788.6	1.4E+07
samsung::SM-A236E	FALSE	840.4	210.1	1	FALSE	INR	11	TRUE	851.4	FALSE	1050.5	851.4	1.5E+07
realme::RMX3771	FALSE	577.52	144.38	2	FALSE	INR	11	FALSE	588.52	TRUE	721.9	588.52	1.5E+07
Redmi::23076RN4BI	FALSE	195.05	39.95	1	FALSE	INR	11	TRUE	245.05	FALSE	235	245.05	1.5E+07
POCO::2201117PI	FALSE	452.7	357.3	2	TRUE	INR	11	FALSE	463.7	TRUE	810	463.7	1.5E+07
motorola::motorola edge 30	TRUE	509.66	380.99	2	TRUE	INR	11	TRUE	520.66	TRUE	890.65	520.66	1.5E+07
vivo::vivo 1803	FALSE	976.24	244.06	3	FALSE	INR	11	FALSE	987.24	TRUE	1220.3	987.24	1.5E+07
Redmi::23076RN4BI	FALSE	1180.55	392.45	3	FALSE	INR	11	TRUE	1153.48	TRUE	1573	1153.48	1.5E+07
OPPO::CPH2527	FALSE	399.68	99.92	1	FALSE	INR	11	FALSE	449.68	TRUE	499.6	449.68	1.5E+07
Redmi::M2101K7AI	FALSE	156	39	1	FALSE	INR	11	FALSE	216	FALSE	195	216	1.5E+07

Apart from this we can achive this task using ***python-pandas*** also but doing it using excel is more efficient and easier.
In pandas you ***expand_json*** function takes a DataFrame with a JSON column, ***parses*** the JSON data, ***normalizes*** it into a flat table, and then combines this new data with the original DataFrame, excluding the original JSON column.

So the columns in the expanded dataset will be:

```
Columns in the expanded DataFrame:

Index(['Attributed Touch Time', 'Install Time', 'Event Time', 'Event Name',
       'Event Revenue', 'Cost Model', 'Cost Value', 'Partner', 'Media Source',
       'Channel', 'Campaign ID', 'Country Code', 'State', 'City', 'Operator',
       'Carrier', 'Language', 'Unnamed: 18', 'Unnamed: 19', 'Device Category',
       'Platform', 'OS Version', 'App Version', 'SDK Version', 'App ID',
       'App Name', 'Is Retargeting', 'Retargeting Conversion Type',
       'Is Primary Attribution', 'Reengagement Window', 'Original URL',
       'Device Model', 'has_coupon_code', 'selling_price_total_amount',
       'discount_amount', 'no_of_items', 'is_switch_added', 'af_currency',
       'packaging_charge_amount', 'is_addons_added', 'af_revenue',
       'is_core_customer', 'mrp_total_amount', 'estimated_payable_amount',
       'reposr', 'delivery_charge_amount', 'coupon_discount_amount',
       'coupon_applied', 'tm_reward_amount', 'tm_credit_amount',
       'product_code', 'savings_amount', 'no_of_item', 'customer_id',
       'subs_source'],
      dtype='object')
```

Now we will be doing analysis on this data using the python ,where we will be doing data cleaning and transforming data to get meaningful insights.

Initially there are null values in almost all columns:

```
dt.isnull().sum()
```

```
Attributed Touch Time            0
Install Time                     0
Event Time                       0
Event Name                       0
Event Revenue                 4161
Cost Model                   84218
Cost Value                   84218
Partner                      84218
Media Source                     0
Channel                          0
Campaign ID                      0
Country Code                     0
State                            0
City                             0
Operator                     46431
Carrier                      47048
Language                     46241
Unnamed: 18                  84218
Unnamed: 19                  84218
Device Category              46241
Platform                         0
OS Version                   46241
App Version                      0
SDK Version                  46241
App ID                           0
App Name                     46241
Is Retargeting                   0
Retargeting Conversion Type  84218
Is Primary Attribution           0
Reengagement Window          84218
Original URL                 84218
Device Model                 46241
has_coupon_code                  0
```

There are columns which have amount related data,which we can fill will **0** and some categorical values which we can fill with **unknown** and similarly some columns filled with **none** for coupons data which we have no idea about.And others cols we can leave it like that.

```
fill_zero_cols = [
    'Event Revenue', 'selling_price_total_amount', 'discount_amount',
    'no_of_items', 'packaging_charge_amount', 'af_revenue', 'mrp_total_amount',
    'estimated_payable_amount', 'delivery_charge_amount', 'coupon_discount_amount',
    'tm_reward_amount', 'tm_credit_amount',"savings_amount"
]

fill_unknown_cols = [
    'Operator', 'Carrier', 'Language', 'Device Category', 'App Name',"OS Version",
    'SDK Version', 'Device Model', 'af_currency'
]

fill_none_cols = ['coupon_applied']

dt[fill_zero_cols] = dt[fill_zero_cols].fillna(0)
dt[fill_unknown_cols] = dt[fill_unknown_cols].fillna('unknown')
dt[fill_none_cols] = dt[fill_none_cols].fillna('none')
dt.isnull().sum()
```

```
3]: Attributed Touch Time            0
    Install Time                     0
    Event Time                       0
    Event Name                       0
    Event Revenue                    0
    Media Source                     0
    Channel                          0
    Campaign ID                      0
    Country Code                     0
    State                            0
    City                             0
    Operator                         0
    Carrier                          0
    Language                         0
    Device Category                  0
    Platform                         0
```

Analysis :

1) Which media source has the biggest delta between install time & event time? What is the average time from install to the 3 events? Given the three events and funnel shared, can you provide a reasoning for this delay from install?

```
df['Delta'] = (df['Event Time'] - df['Install Time']).dt.total_seconds() / 60

media_delta = df.groupby('Media Source')['Delta'].mean().reset_index()
max_media_delta = media_delta.loc[media_delta['Delta'].idxmax()]

avg_times = df.groupby('Event Name')['Delta'].mean().reset_index()

print(max_media_delta)
print(avg_times)
```

```
Media Source    Rocketship
Delta           4481.21172
Name: 0, dtype: object
         Event Name        Delta
0  app_order_placed  2887.527582
1      box_verified  3813.078501
2   order_delivered  6803.697385
```

From this we can get insights such as :

Placed App Order (2888 minutes):

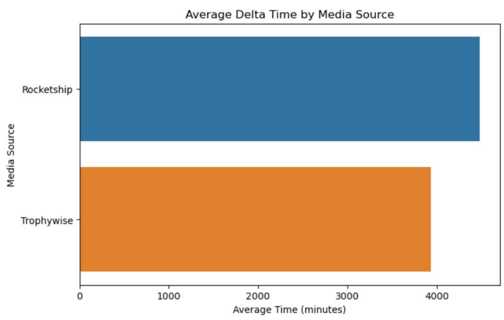
Customers take their time browsing the app before deciding what to buy.
Anticipating sales or special offers may prolong the wait.
Establishing confidence with the app might potentially impede placing the first order.

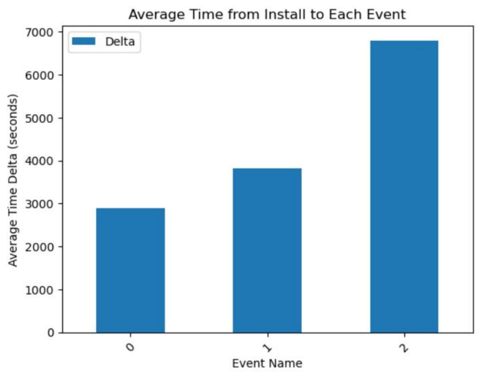
Verified Box in 3813 minutes:

Order packing and processing require time.
The delay is increased by the logistics of picking up and confirming orders.
There might be more delays if the user confirms the order.

Order Fulfilled in 6804.1 Minutes:

It takes a long time to ship and get to the user's location.
Delivery delays can be caused by unanticipated events and geographic reasons.
The user's availability to accept the shipment might cause the delivery time to be further extended.



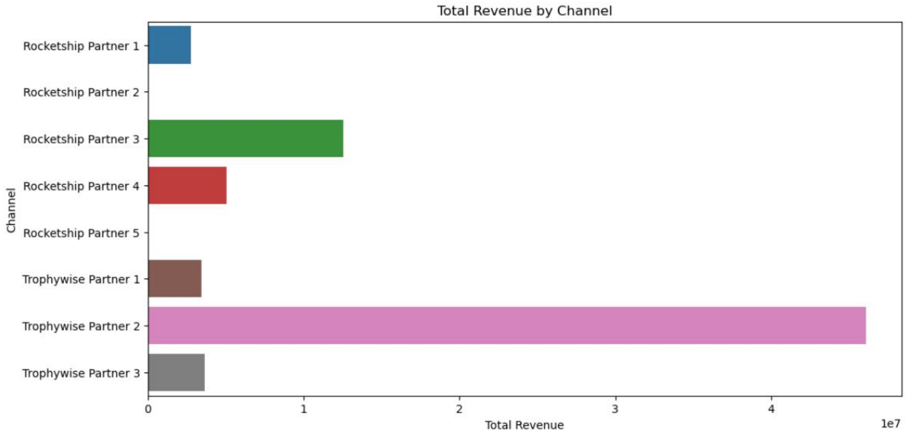


From these visualizations ,we can clearly see that the Rocketship have the the highest average delta time it could be due to various factors including the media source's effectiveness in driving immediate engagement or the quality of user experience which increases the delta time.

2) What is the most revenue driving channel? Put a case forward for where you can accurately visualise the revenue driven vs quality factors (‘core customers’ who accepts the ‘switch’, ‘does not use coupon’ can be considered as quality metrics)

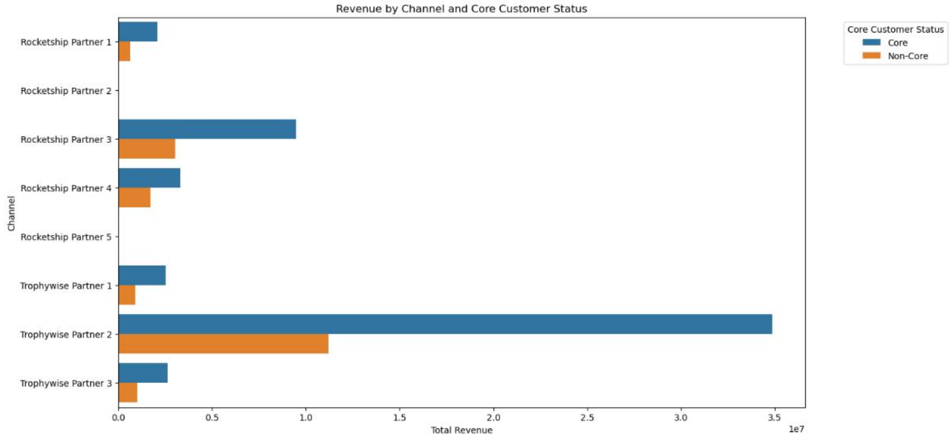
```
channel_revenue = df.groupby('Channel')['Event Revenue'].sum().reset_index()
most_revenue_channel = channel_revenue.loc[channel_revenue['Event Revenue'].idxmax()]

print("Most revenue-driving channel:")
print(most_revenue_channel)
```

```
Most revenue-driving channel:
Channel            Trophywise Partner 2
Event Revenue               46087751.42
Name: 6, dtype: object
```

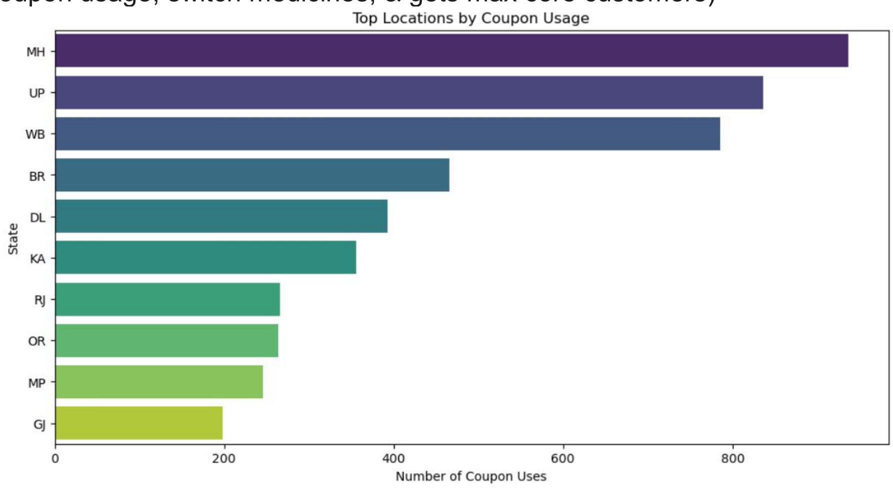


In above code and visualization also ,you can clearly see that the *Trophywise Partner 2* is the most revenue driving channel .

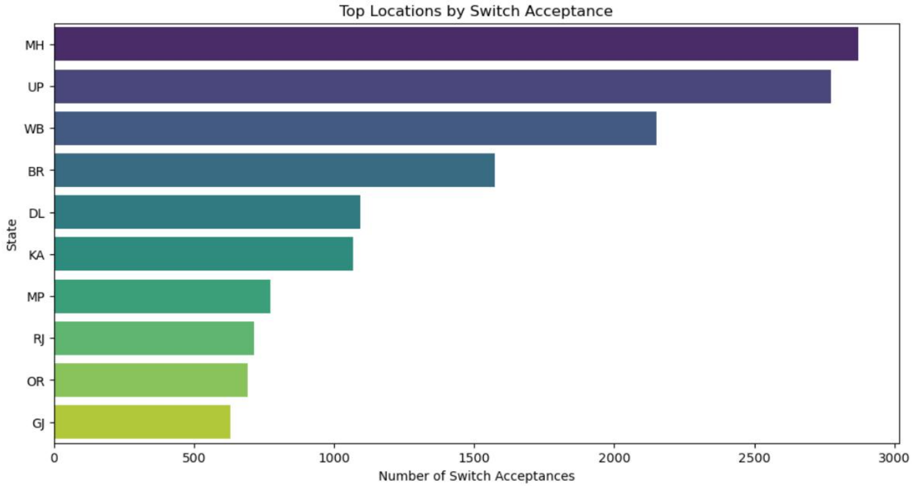
For Core customers and non core :



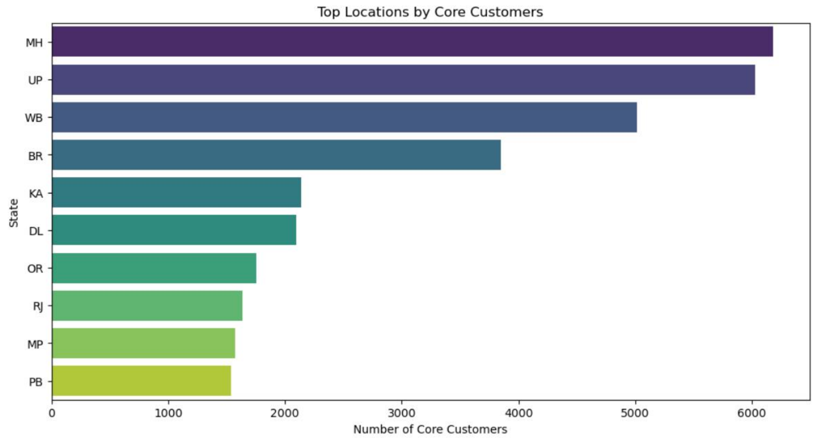
3)Location level analysis: Given the current dataset, which are the top locations in terms of coupon usage, switch medicines, & gets max core customers)



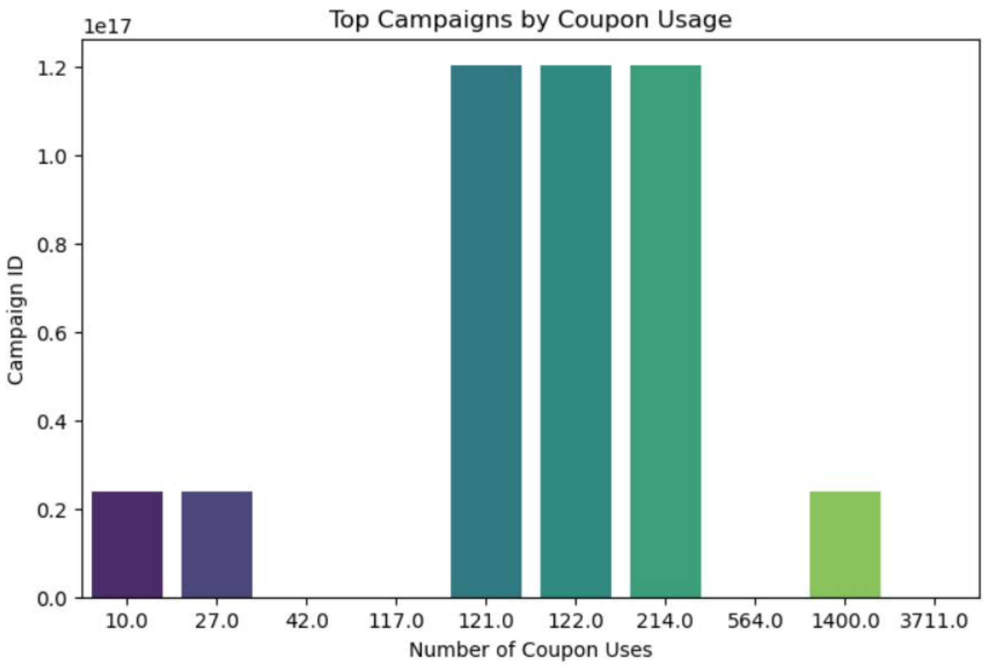
From here we can clearly see that “**MH**” has the highest usage of coupons followed by “**UP”.**

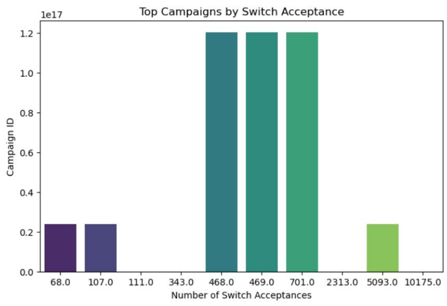


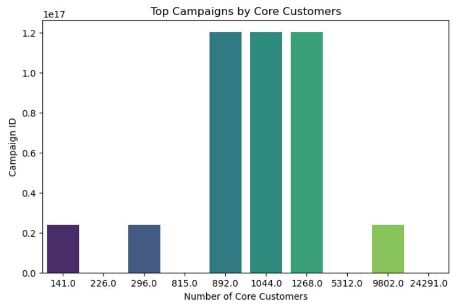
There have been a similar trend for switch acceptance and core customers ,ie,”**MH”** topping the charts.

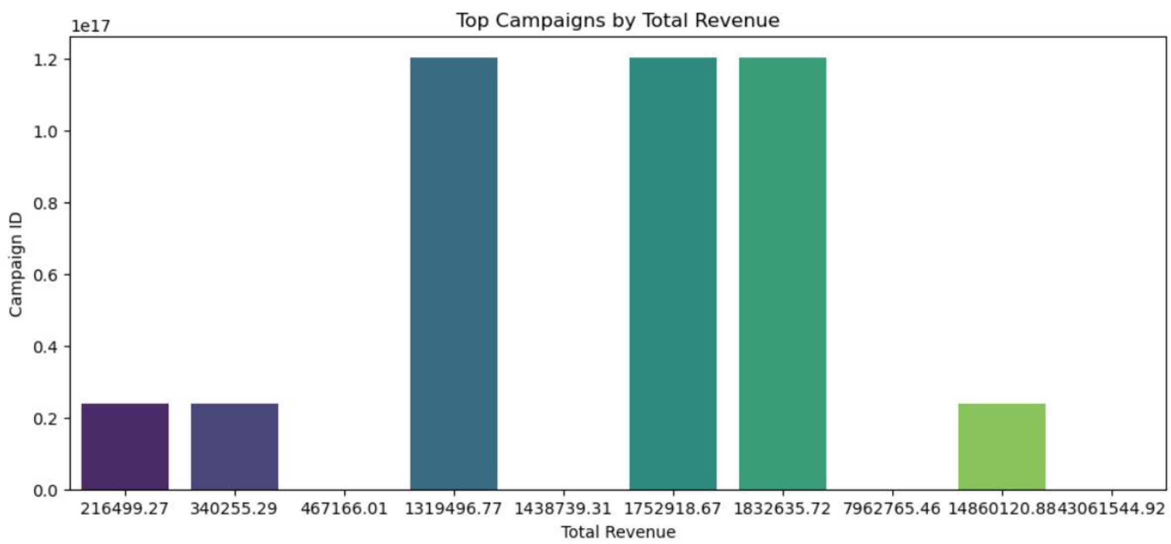


4)If you were to advise the marketing team to double down on spending on such campaigns, which are the top campaigns to increase spending and why?





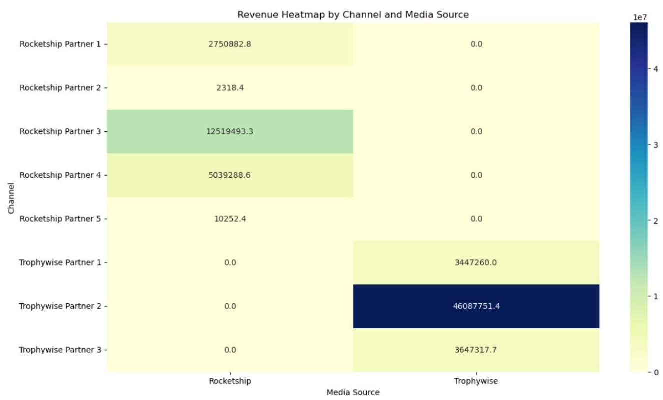




These are few graphs where we get the data about the Campaigns,

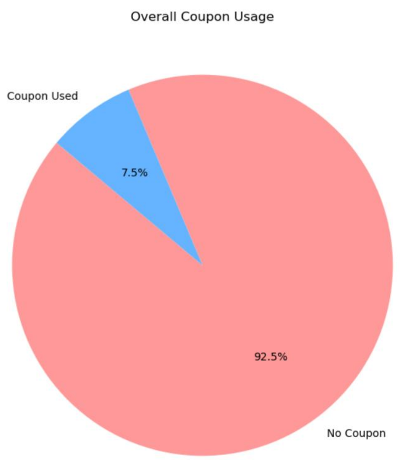
To advise on increasing campaign spending, focus on:
Revenue: Higher revenue indicates profitability.
Coupon Usage: High usage shows strong engagement.
Switch Acceptance: High acceptance suggests effective upselling.

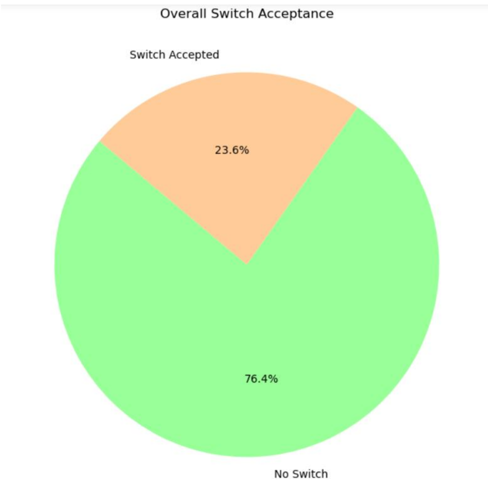
5) Summarise key learning and business insight in brief.



From this graph we can know that after Trophywise Partner 2,Rocketship Partner 3 takes the lead,in terms of revenue.

Also here are some of piecharts,which show the percentage of coupons and Switch acceptance:





From here we can see that than acceptance,no switch and coupon is higher and has most effect.

Final Insights and Conclusions

Revenue Distribution

Channels: Focus marketing efforts on high-revenue channels to optimize return on investment.
Media Sources: Prioritize media sources that generate the most revenue for advertising and partnerships.

Coupon Usage

Top Countries: Target countries with higher coupon usage for future promotions.
Overall Usage: Assess overall coupon usage to refine promotional strategies.

Switch Acceptance

Top States: Tailor marketing to states with higher acceptance of switch offers.
Overall Acceptance: Use general acceptance rates to improve upsell strategies.

Core Customers

Top States: Strengthen loyalty programs in states with more core customers.
Overall Core Customers: Understand core vs. non-core customer proportions to enhance retention strategies.

RECOMMENDATIONS :

- Pay Attention to High-Revenue Media Sources and Channels: Give media outlets and channels that bring in the most money additional funding and resources.
- Choose Regions with High Engagement: In places and nations where switch acceptance and coupon utilisation are greater, intensify promotional activities.
- Improve Loyalty Programs: To further increase client retention, reinforce loyalty programs in areas with a large number of core consumers.
- Improve Your Upsell Techniques: Examine and duplicate the effective components of campaigns with elevated switch acceptance rates to enhance total sales and enhance client retention.

Github link:
Here is the link of all code files added to github:

(https://github.com/Harshinimallipeddi/Truemed_Analysis)