BUSINESS ANALYTICS

TERM PROJECT

RACHIT MAHESHWARI (BA025-21)

RFM ANALYSIS OF CUSTOMER SEGMENTATION, USING PYTHON & POWER BI

- INTRODUCTION
- METHODOLOGY
- EDA ANALYSIS

- RFM SCORES
- SEGMENTATION
- POWER BI

- KEY OBSERVATIONS
- MANAGERIAL SUGGESTIONS
- LEARNINGS

INTRODUCTION

The aim of this project is to create a useful tool for sales managers that can aid in increasing sales and customer retention by identifying high-priority customers for outreach.

To accomplish this goal, I utilized RFM analysis, a commonly used technique in direct marketing and database marketing, especially in the retail industry.

The study focuses on customer segmentation through data visualization using Power BI, where RFM analysis was performed on the sales data of the company.

METHODOLOGY

1 **EDA ANALYSIS**

Data Preparation, Data Cleaning, Data Exploration

2 **DATA MODELLING**

Transforming the data to obtain RFM values.
For this, RFM scores will be calculated

3 **SEGMENTATION**

Segregation of customers based on the above RFM scores into various different 11 categories

4 **VISULIZATION IN POWER BI**

Designing Power BI dashboard for better visualization

DATASET DESCRIPTION

	A	B	C	D	E
1	country	id	week.year	revenue	units
2	KR	702234	3.2019	808,08	1
3	KR	702234	6.2019	1606,80	2
4	KR	3618438	8.2019	803,40	1
5	KR	3618438	9.2019	803,40	1
6	KR	3618438	9.2019	803,40	1
7	KR	3618438	13.2019	2376,42	3
8	KR	3618438	12.2019	1198,74	1
9	KR	702234	16.2019	797,82	1
10	KR	3618438	18.2019	399,54	1

1. ***country*** *– Country name codes. (Nominal)*
2. ***id*** *– Customer id (Numeric - int)*
3. ***week.year*** *– Transaction date (Date)*
4. ***revenue*** *– Revenue from a particular order (Numeric – float)*
5. ***units*** *– Number of units bought (Numeric – int)*

 # Rows - 235574
 # Columns - 5

EDA ANALYSIS

Step 1: Data Preparation & Data Cleaning

a) Importing the data

```
df1 = pd.read_csv('https://raw.githubusercontent.com/RaxRachit/Final-Year-Project/main/sales_asia.csv')
```

b) Splitting the data

```
# Splitting 'week.year' column on '.' and creating 'week' and 'year' columns

df1['week'] = df1['week.year'].astype(str).str.split('.').str[0]
df1['year'] = df1['week.year'].astype(str).str.split('.').str[1]
```

c) Formatting the data

```
# Converting year and week into date, using Monday as first day of the week

df1['date'] = pd.to_datetime(df1['year'].map(str) + df1['week'].map(str) + '-1', format='%Y%W-%w')
```

EDA ANALYSIS

Step 1: Data Preparation & Data Cleaning

d) Removing unnecessary columns

```
# Removing unnecesary columns

df2 = df1.drop(['week.year', 'week', 'year'], axis=1)
```

e) Renaming columns

```
#Rename columns

df2.rename({'revenue': 'monetary'}, axis="columns", inplace=True)
```

f) Checking null values

```
df2.isnull().sum()
```

EDA ANALYSIS

	country	id	monetary	units	date
0	KR	702234	808.08	1	2019-01-21
1	KR	702234	1606.80	2	2019-02-11
2	KR	3618438	803.40	1	2019-02-25
3	KR	3618438	803.40	1	2019-03-04
4	KR	3618438	803.40	1	2019-03-04

EDA ANALYSIS

Step 2: Raw Data Description

a) Basic statistical details

	id	monetary	units
count	2.355740e+05	2.355740e+05	235574.000000
mean	3.193118e+06	2.840211e+03	8.599642
std	7.371744e+06	2.247532e+04	602.939290
min	6.000180e+05	-1.061539e+05	-150000.000000
25%	2.214396e+06	3.994800e+02	1.000000
50%	3.140856e+06	1.150320e+03	1.000000
75%	3.892650e+06	2.216160e+03	2.000000
max	2.419308e+08	2.415857e+06	150000.000000

- *235,574 transactions and 5 columns.*
- *The largest transaction in terms of units was 150,000.*
- *There was also a return of the same amount, resulting in a negative 150,000 units.*
- *The costliest purchase 2.41 million.*

EDA ANALYSIS

Step 2: Raw Data Description

b) Examining the number of countries in which sales were made

```
# Let's explore in how many different countries we have sales in that period

df2['country'].unique()
```

```
array(['KR', 'PK', 'MM', 'VN', 'IN', 'SA', 'PH', 'AF', 'CN', 'BD', 'ID',
       'TH', 'IQ', 'MY', 'JP', 'IR', 'TR', 'UZ'], dtype=object)
```

```
df2['country'].nunique()
```

```
18
```

EDA ANALYSIS

Step 2: Raw Data Description

c) Examining the number of countries in which sales were made

```
# Transforming country codes into full country names with clean_country function
# from dataprep library

clean_country(df2, "country")['country_clean'].unique()
```

```
Country Cleaning Report:
	235574 values cleaned (100.0%)
Result contains 235574 (100.0%) values in the correct format and 0 null values (0.0%)

array(['South Korea', 'Pakistan', 'Myanmar', 'Vietnam', 'India',
       'Saudi Arabia', 'Philippines', 'Afghanistan', 'China',
       'Bangladesh', 'Indonesia', 'Thailand', 'Iraq', 'Malaysia', 'Japan',
       'Iran', 'Turkey', 'Uzbekistan'], dtype=object)
```

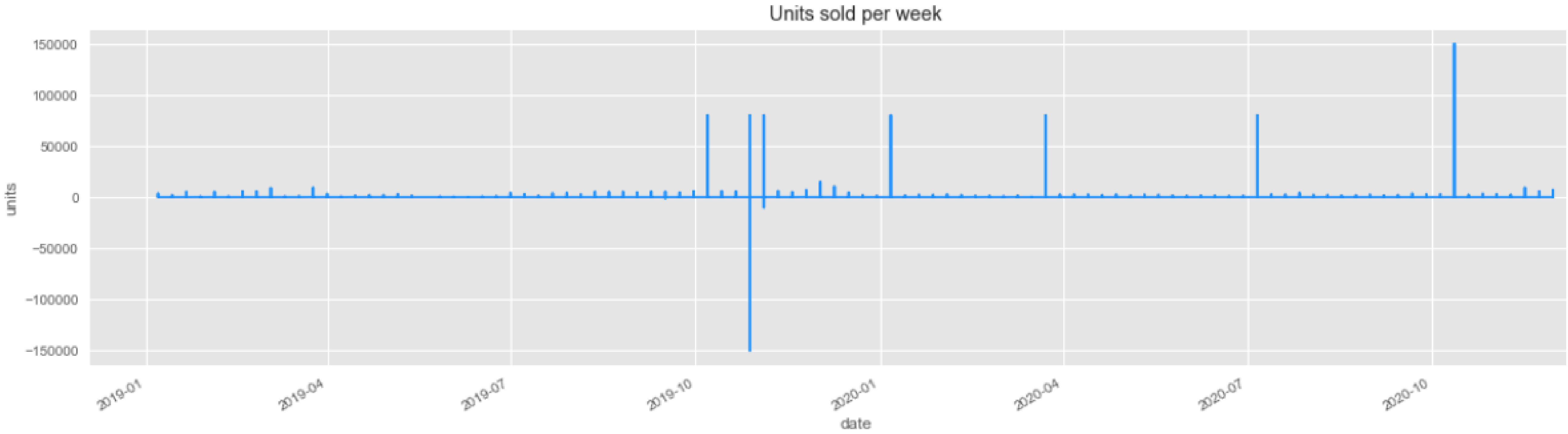
d) The total count of customers across all countries

```
df2['id'].nunique()
```

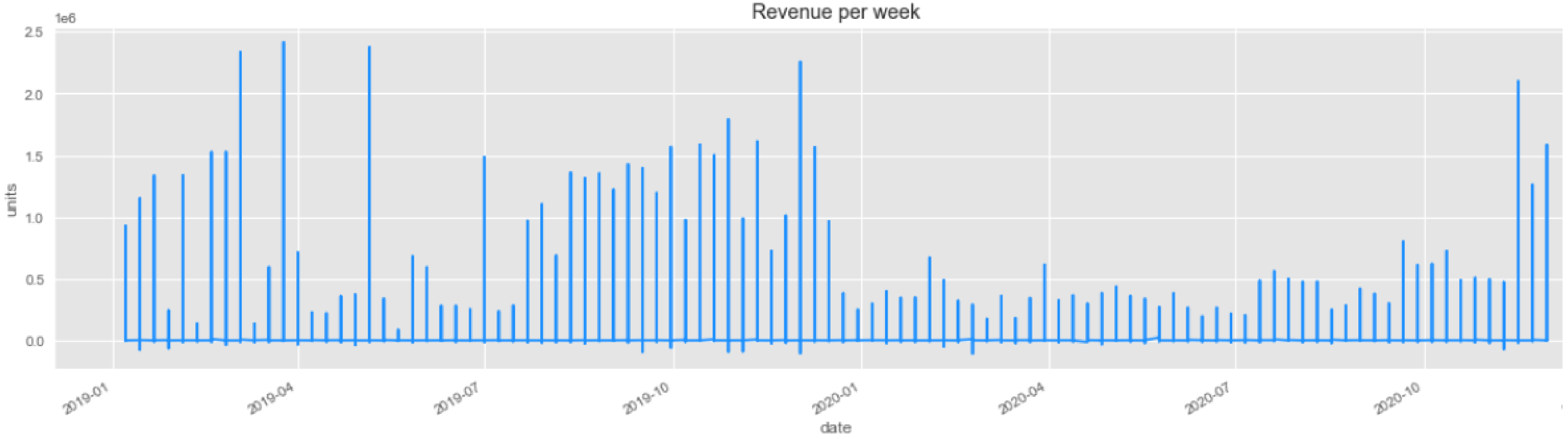
```
21837
```

EDA ANALYSIS

Step 3: Data Exploration



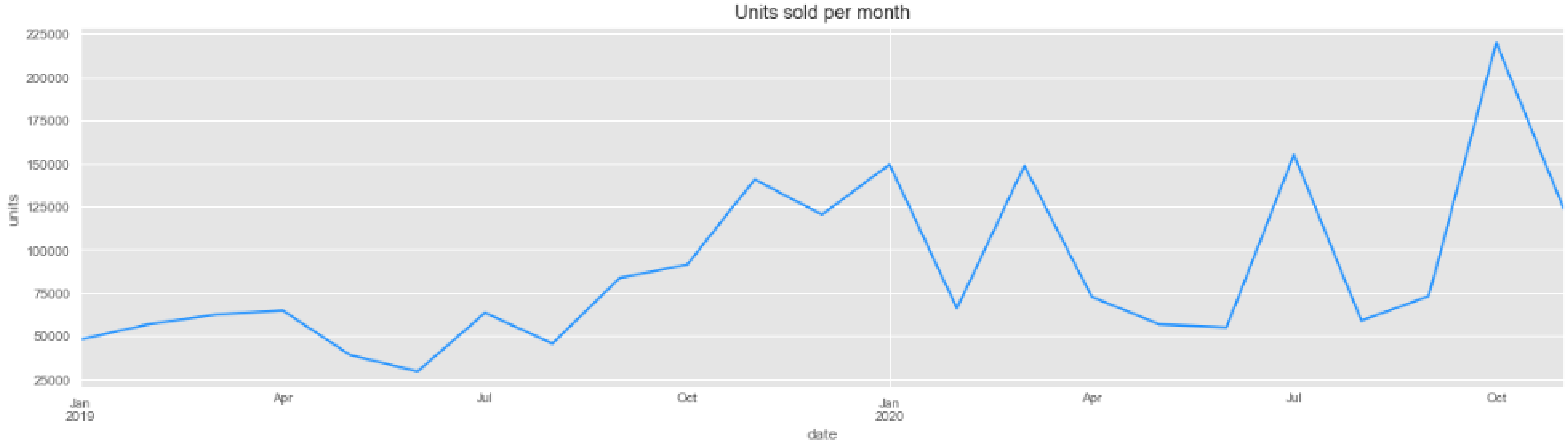
Units Chart



Revenue Chart

EDA ANALYSIS

Step 3: Data Exploration



Units Chart

Revenue Chart

DATA MODELLING

Step 4: Transforming data to get RFM values

a) Narrowing our focus to sales made within the past 365 days

```
period = 365
date_N_days_ago = df2['date'].max() - timedelta(days=period)
```

b) Eliminating the rows with dates that precede 365 days ago

```
df2 = df2[df2['date']> date_N_days_ago]
```

```
df2.reset_index(drop=True, inplace=True)
```

c) Setting the NOW date as one day after the date of the last sale

```
NOW = df3['date'].max() + timedelta(days=1)
NOW
```

DATA MODELLING

Step 4: Transforming data to get RFM values

d) There are customers with the same 'id' in several countries. This causes errors in the monetary values. We will solve this by creating a new feature: a unique 'id+' identifier that combines country code and customer id

```
df3 = df2.copy()

df3['id+'] = df3['country'].map(str) + df3['id'].map(str)
```

e) 'days_since_last_purchase' calculates the number of days between the purchase date and the latest date

```
df3['days_since_purchase'] = df3['date'].apply(lambda x:(NOW - x).days)
```

DATA MODELLING

Step 4: Transforming data to get RFM values

	country	id	monetary	units	date	id+	days_since_purchase
0	KR	4375152	773.58	1	2019-12-16	KR4375152	351
1	KR	705462	337.26	1	2019-12-09	KR705462	358
2	KR	705462	337.26	1	2019-12-23	KR705462	344
3	KR	705462	421.56	2	2019-12-16	KR705462	351
4	KR	706854	391.50	1	2019-12-09	KR706854	358

DATA MODELLING

Step 4: Transforming data to get RFM values

*a) The **'recency'** feature will be determined by finding the minimum value of 'days_since_last_purchase' for each customer.*

```
aggr = {
    'days_since_purchase': lambda x:x.min(),
    'date': lambda x: len([d for d in x if d >= NOW - timedelta(days=period)])
}
```

*b) The **'frequency'** feature will be calculated by counting the total number of orders made by each customer during a specific period.*

```
rfm = df3.groupby(['id', 'id+', 'country']).agg(aggr).reset_index()
rfm.rename(columns={'days_since_purchase': 'recency',
                    'date': 'frequency'},
           inplace=True)
```

DATA MODELLING

Step 4: Transforming data to get RFM values

*c) The **'monetary'** feature will be calculated by summing up the total value of all purchases made by each customer during the same period.*

```
df3[df3['date'] >= NOW - timedelta(days=period)]\
    .groupby('id+')['monetary'].sum()
```

	id	id+	country	recency	frequency	monetary
0	600018	CN600018	CN	29	7	21402.78
1	600060	CN600060	CN	155	1	1201.14
2	600462	CN600462	CN	211	2	2033.64
3	600888	CN600888	CN	8	3	2335.80
4	601014	CN601014	CN	225	1	230.52

DATA MODELLING

Step 4: Transforming data to get RFM values

d) Calculating the revenue generated by each customer within the last 365 days.

```
df3[df3['date'] >= NOW - timedelta(days=period)]\
     .groupby('id+')['monetary'].sum()
```

e) Verifying if customers belonging to different countries have distinct monetary values by examining the data of customer with id 3790218

```
rfm[rfm['id']==3790218]
```

	id	id+	country	recency	frequency	monetary
11057	3790218	AF3790218	AF	309	1	9706.08
11058	3790218	BD3790218	BD	176	4	7267.38
11059	3790218	CN3790218	CN	1	60	716199.60
11060	3790218	ID3790218	ID	260	9	49154.22
11061	3790218	IQ3790218	IQ	176	1	1243.08
11062	3790218	MM3790218	MM	183	3	7110.60

DATA MODELLING

Step 5: Calculating R, F, M Scores

Rate the customers' R, F & M value factors on a scale of 1 to 5.
We'll split each characteristic into groups with 20% of the samples using the quintiles method.
Recency scores will be lower numbers, while frequency and monetary value scores will be higher.

```
quintiles = rfm[['recency', 'frequency', 'monetary']].quantile([.2, .4, .6, .8]).to_dict()
quintiles
```

	id	country	recency	frequency	monetary	r	f	m	rfm_score
0	600018	CN	29	7	21402.78	4	4	5	445
1	600060	CN	155	1	1201.14	2	1	2	212
2	600462	CN	211	2	2033.64	2	2	2	222
3	600888	CN	8	3	2335.80	5	3	3	533
4	601014	CN	225	1	230.52	2	1	1	211

DATA MODELLING

Step 5: Calculating R, F, M Scores

We can use the R, F, and M scores to create 125 customer segments with these values. However, we can reduce the number of segments by combining F and M scores, resulting in 11 segments

$$fm = (f+m)/2$$

	id	country	recency	frequency	monetary	r	f	m	rfm_score	fm
0	600018	CN	29	7	21402.78	4	4	5	445	4
1	600060	CN	155	1	1201.14	2	1	2	212	1
2	600462	CN	211	2	2033.64	2	2	2	222	2
3	600888	CN	8	3	2335.80	5	3	3	533	3
4	601014	CN	225	1	230.52	2	1	1	211	1

DATA MODELLING

Step 5: Calculating R, F, M Scores

We create a segment map of only 11 segments based on only two scores: 'r' and 'fm'
This code block is mapping the RFM scores to customer segments using regular expressions. It creates a dictionary called segment_map that defines the segment names based on the combination of R, F, and M scores.

```
segment_map = {
    r'22': 'hibernating',
    r'[1-2][1-2]': 'lost',
    r'15': 'can\'t lose',
    r'[1-2][3-5]': 'at risk',
    r'3[1-2]': 'about to sleep',
    r'33': 'need attention',
    r'55': 'champions',
    r'[3-5][4-5]': 'loyal customers',
    r'41': 'promising',
    r'51': 'new customers',
    r'[4-5][2-3]': 'potential loyalists'
}

rfm['segment'] = rfm['r'].map(str) + rfm['fm'].map(str)
rfm['segment'] = rfm['segment'].replace(segment_map, regex=True)
rfm.head()
```

DATA MODELLING

Step 5: Calculating R, F, M Scores

For example, customers with an R score of 5, an F score of 5, and an M score of 5 will have an RFM score of "555". This value will match the regular expression "55" in the segment_map dictionary, and the corresponding segment name "champions" will be assigned to these customers in the rfm['segment'] column.

	id	country	recency	frequency	monetary	r	f	m	rfm_score	fm	segment
0	600018	CN	29	7	21402.78	4	4	5	445	4	loyal customers
1	600060	CN	155	1	1201.14	2	1	2	212	1	lost
2	600462	CN	211	2	2033.64	2	2	2	222	2	hibernating
3	600888	CN	8	3	2335.80	5	3	3	533	3	potential loyalists
4	601014	CN	225	1	230.52	2	1	1	211	1	lost

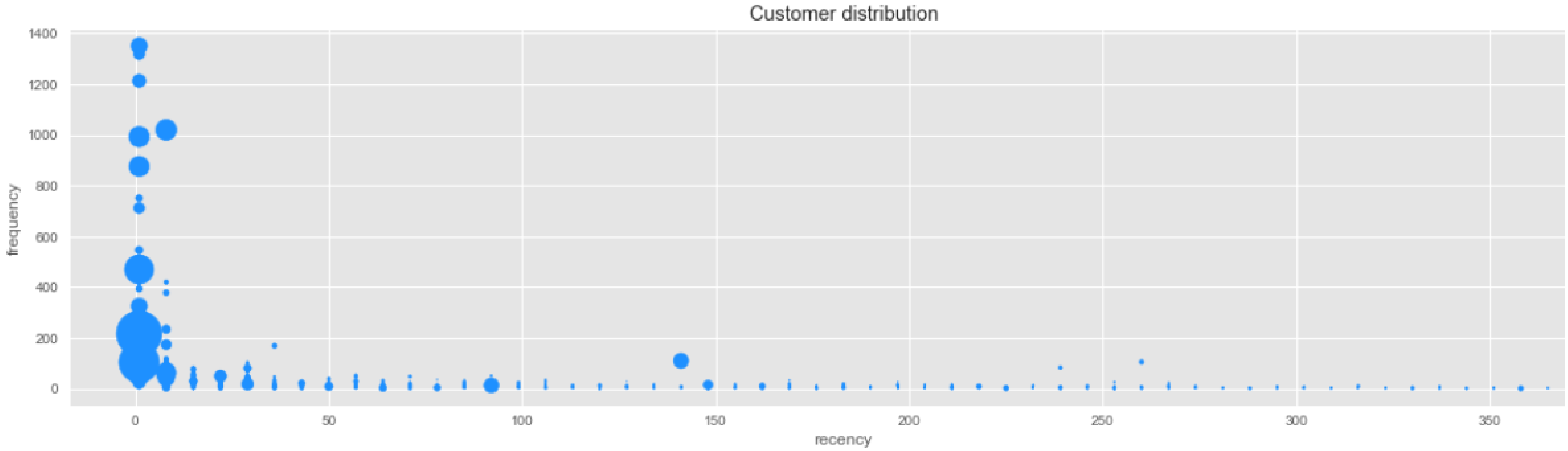
DATA MODELLING

Step 6: Segmentation of 11 categories of customers

- ***Champions:*** *Bought recently, buy often and spend the most*
- ***Loyal Customers:*** *Buy on a regular basis. Responsive to promotions.*
- ***Potential Loyalists:*** *Recent customers with average frequency.*
- ***Recent Customers:*** *Bought most recently, but not often.*
- ***Promising:*** *Recent shoppers, but haven't spent much.*
- ***Customers Needing Attention:*** *Above average recency, frequency and monetary values. May not have bought very recently though.*
- ***About To Sleep:*** *Below average recency and frequency. Will lose them if not reactivated.*
- ***At Risk:*** *Purchased often but a long time ago. Need to bring them back!*
- ***Can't Lose Them:*** *Used to purchase frequently but haven't returned for a long time.*
- ***Hibernating:*** *Last purchase was long back and low number of orders.*
- ***Lost:*** *Purchased long time ago and never came back.*

DATA MODELLING

Step 7: Distribution of customers



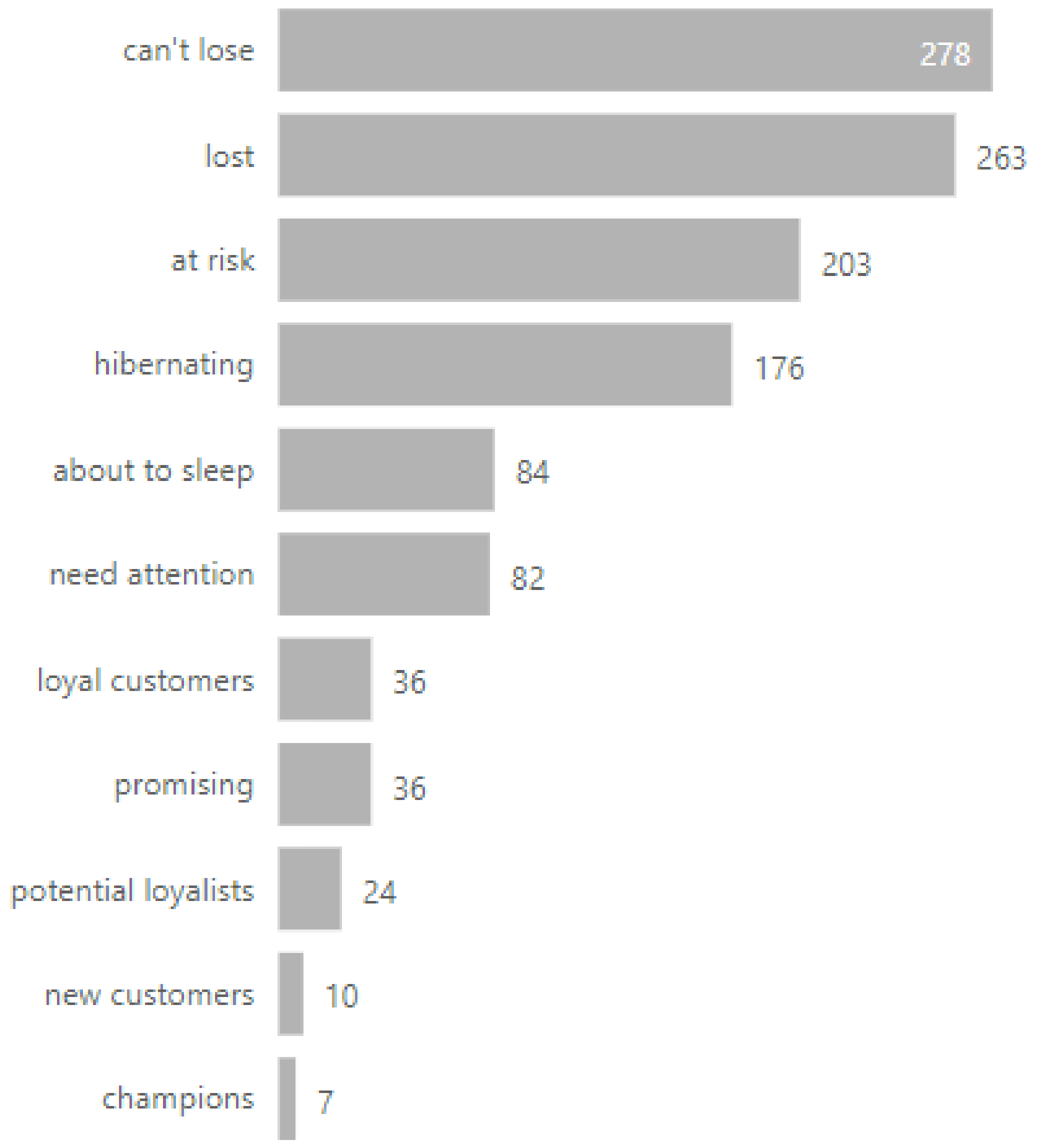
It can be observed that customers who spend the most also tend to purchase more frequently & more recently.

DATA VISUALIZATION IN POWER BI

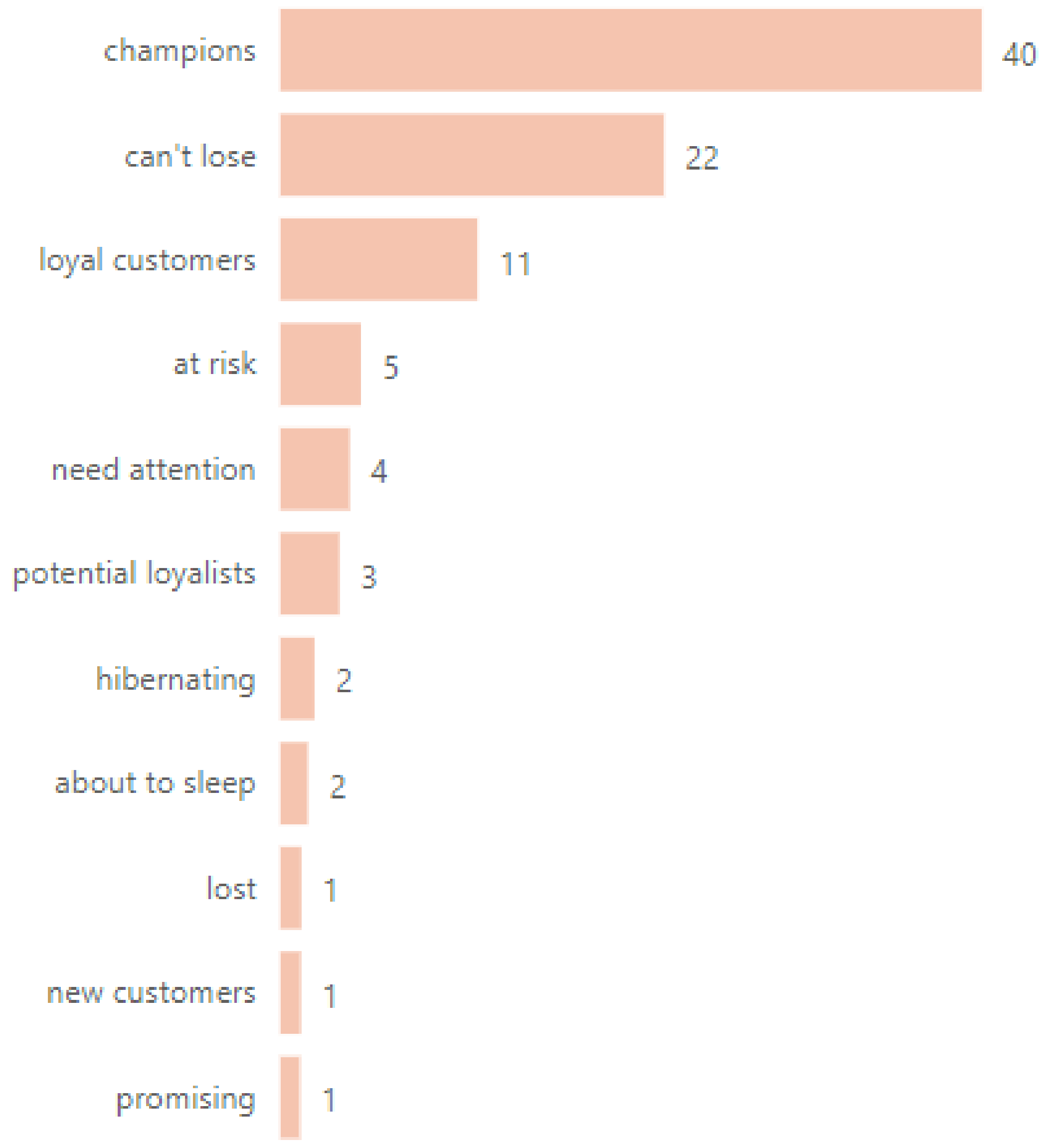
RFM Analysis

country
- AF
- BD
- CN
- ID

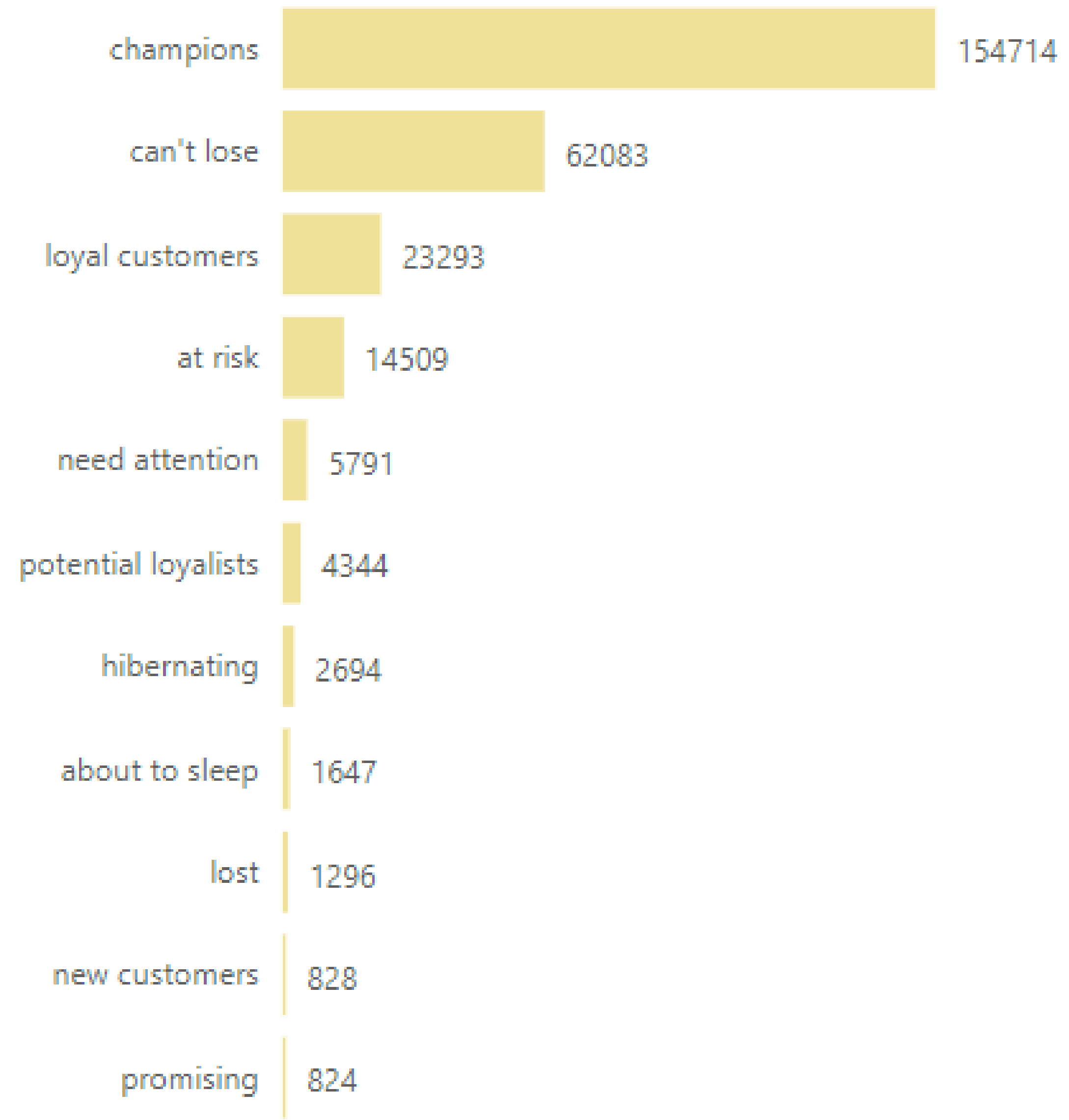
Days since last purchase (Avg by Recency)



NUmber of orders in last 365 days (Avg by Frequency)

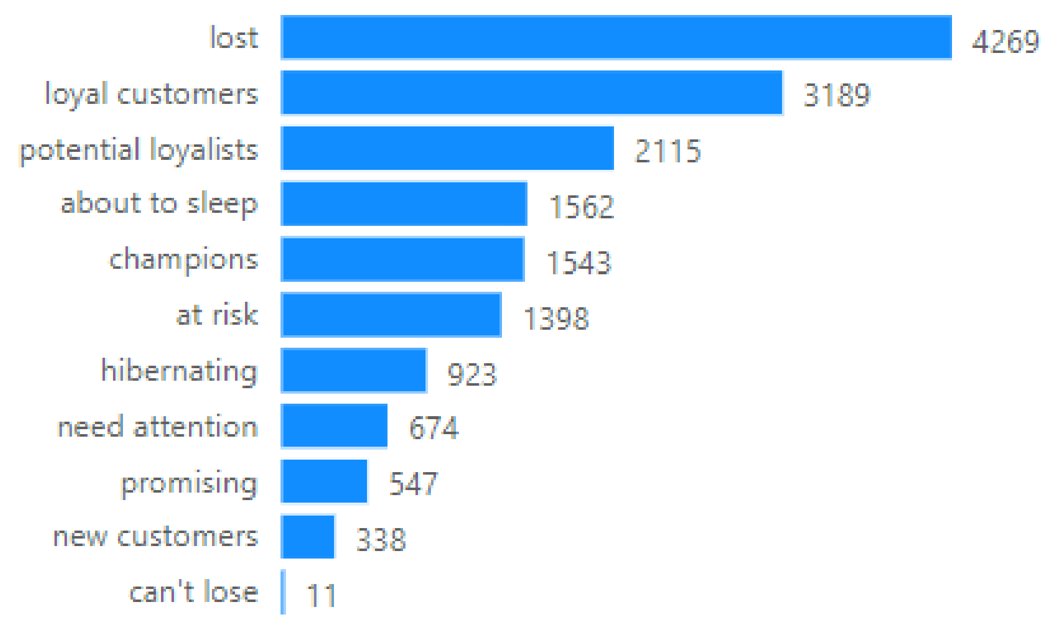


Revenue in last 365 days (Avg by Monetary)

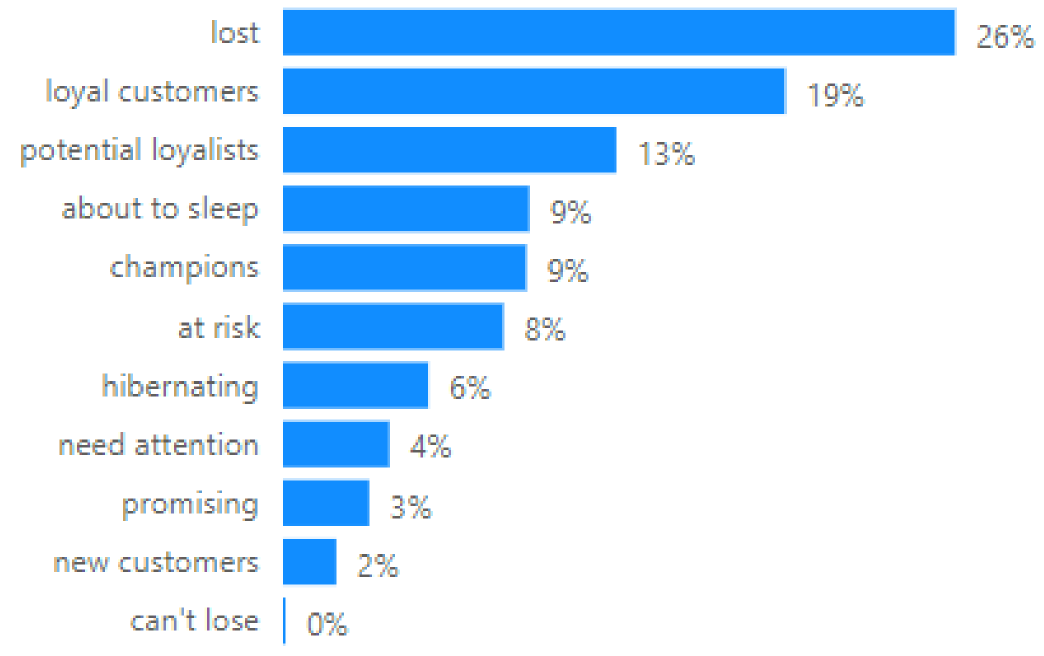


DATA VISUALIZATION IN POWER BI

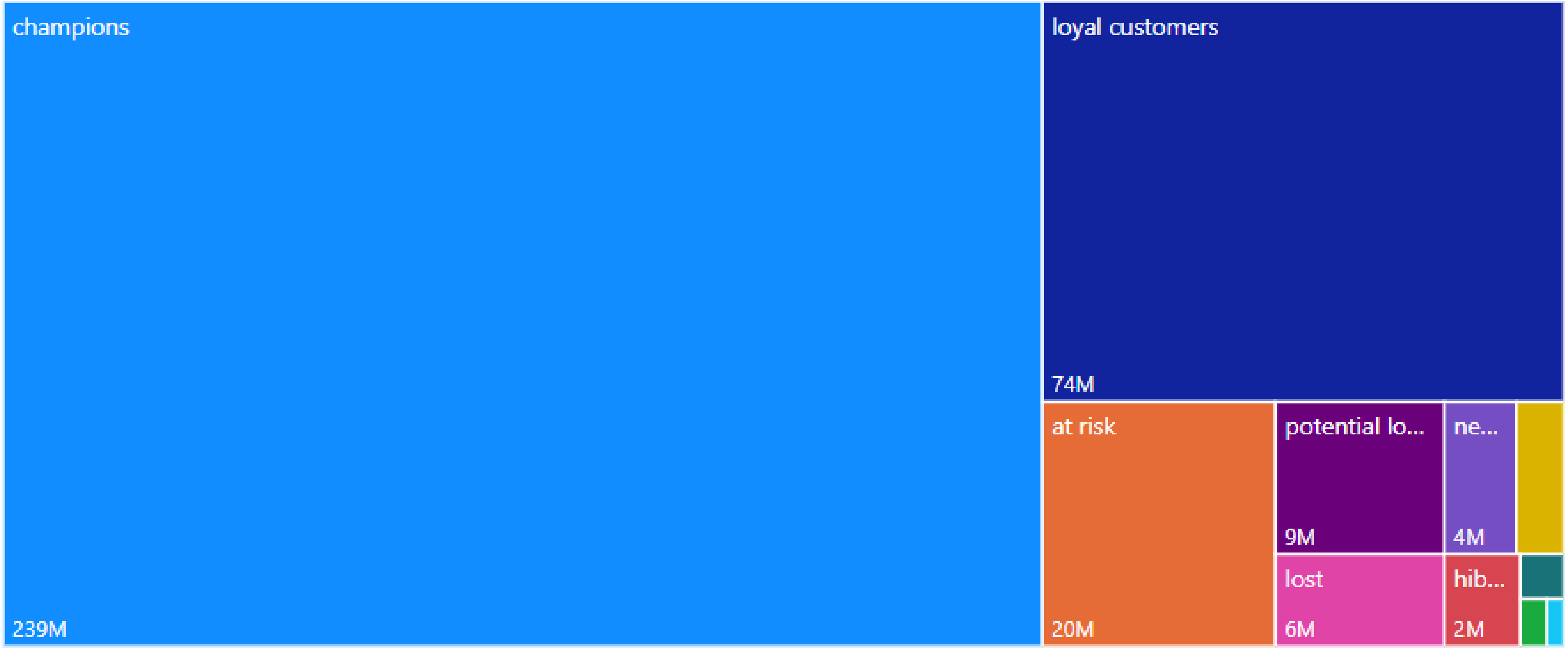
Number of Customers per segment



% of Customers per segment



Revenue (monetary values) of last 365 days per segment



country

- AF
- BD
- CN
- ID

country	id	rfm_score	monetary
CN	638544	555	2,14,82,332.56
CN	4424580	555	1,69,12,322.46
TR	4341960	555	1,65,50,997.90
ID	3929094	555	87,48,884.64
JP	3520734	555	62,07,519.96
TR	4494150	555	48,74,668.14
KR	3618438	555	46,15,660.08
PH	4245048	555	43,58,515.98
IN	2111100	555	42,70,717.80
PH	3894492	555	41,06,366.22
PH	1145142	555	35,24,879.46
ID	3857664	555	30,27,573.60
JP	4540974	555	29,97,013.62
TR	3249114	555	27,31,448.04
VN	792522	555	26,77,778.52
JP	2115414	255	24,24,168.66
TR	4422780	355	23,15,341.14
ID	3721002	555	21,09,053.76
TR	4564152	555	21,05,041.20
TH	2195970	555	18,11,210.04
JP	4377870	555	16,16,119.20
ID	4052706	555	16,01,799.24
JP	2030526	455	15,19,339.86

KEY OBSERVATIONS

- ***32% of the customers are 'lost' or 'hibernating'*** *(meaning they have a few orders from long ago) which comprises almost 1/3rd of the total customers*
  - *Design campaigns, offer relevant products with special discounts*

- ***28% of the customers are either 'champions' or 'loyal customers'*** *(meaning they visit frequently and spend the most)*
  - *Almost 87% revenue is contributed by these two segments*
  - *Reward them, can be early adopters of new products, upsell high value products, reviews*

- ***'Can't Lose' customers*** *frequently made the biggest purchases, but they have not returned for a long time. Though the percentage of these customers is very low, less than* ***1%****, but we can not afford to lose them.*
  - *Talk directly & win them back via renewals or new products*

- ***16% of the customers are either 'potential loyalists' or 'promising'*** *(recent customers who have purchased from us)*
  - Convert these in short term: 'Loyal customers' & in long term; 'Champions'
  - Loyalty program, recommendation of products, free trials to create brand awareness

- ***8% of the customers are 'at risk' customers*** *who are 3rd among all the segment in terms of revenue generated by them which is around 6% of the total revenue*
  - *Personalized emails to reconnect with them, offer renewals & helpful resources*

MANAGERIAL SUGGESTIONS

Customer Segment	Activity	Actionable Tip
ABOUT TO SLEEP	Below average recency, frequency and monetary values. Will lose them if not reactivated.	Share valuable resources, recommend popular products / renewals at discount, reconnect with them.
AT RISK	Spent big money and purchased often. But long time ago. Need to bring them back!	Send personalized emails to reconnect, offer renewals, provide helpful resources.
CAN'T LOSE	Made biggest purchases, and often. But haven't returned for a long time.	Win them back via renewals or newer products, don't lose them to competition, talk to them.
CHAMPIONS	Bought recently, buy often and spend the most!	Reward them. Can be early adopters for new products. Will promote your brand.
HIBERNATING	Last purchase was long back, low spenders and low number of orders.	Offer other relevant products and special discounts. Recreate brand value.
LOST	Lowest recency, frequency and monetary scores.	Revive interest with reach out campaign, ignore otherwise.
LOYAL CUSTOMERS	Spend good money with us often. Responsive to promotions.	Upsell higher value products. Ask for reviews. Engage them.
NEED ATTENTION	Above average recency, frequency and monetary values. May not have bought very recently though.	Make limited time offers. Recommend based on past purchases. Reactivate them.
NEW CUSTOMERS	Bought most recently, but not often.	Provide on-boarding support, give them early success, start building relationship.
POTENTIAL LOYALISTS	Recent customers, but spent a good amount and bought more than once.	Offer membership / loyalty program, recommend other products.
PROMISING	Recent shoppers, but haven't spent much.	Create brand awareness, offer free trials.

LEARNINGS FROM THE PROJECT

1. IMPORTANCE OF CUSTOMER SEGMENTATION
2. RFM ANALYSIS
3. DATA PRE-PROCESSING
4. DATA VISUALIZATION
5. BUSINESS INSIGHTS
6. USAGE OF LANGUAGE/TOOLS - PYTHON & POWER BI

REFERENCES

1. https://www.analyticsvidhya.com/blog/2021/07/customer-segmentation-using-rfm-analysis/
2. https://medium.com/@ugursavci/customer-segmentation-using-rfm-analysis-in-python-218a3255f714
3. https://towardsdatascience.com/implementing-customer-segmentation-using-rfm-analysis-with-pyspark-3aed363f1d53
4. https://ploiitubsamon.medium.com/rfm-analysis-for-customer-segmentation-with-power-bi-5d2f5bd62038#:~:text=To%20determine%20the%20customer%20segmentation,is%20the%20latest%20purchase%20date.

THANK YOU