Final Group Project
CIS 467

Class section 23 team members:
Yutong Shen yshen50@simon.rochester.edu
Ruohong Li rli66@ur.rochester.edu
Yuxiao Yao yyao41@ur.rochester.edu
Yijia Liu yliu304@ur.rochester.edu
Saivarshini Ravichandran sravich5@simon.rochester.edu

We built this warehouse to analyze customers' data by analyzing customer's personal information, number of invoices, total spent amount, number of tracks, number of genres and number of artists they favor. From the data warehouse, we can get the information about customers to use the conclusion to help company make decisions to improve their performance.

Part 1: Data Warehouse

```
CREATE OR REPLACE VIEW warehouse AS
SELECT c.CustomerId,
c.FirstName,c.LastName,c.State,c.Country,c.Email,Number_of_Invoices ,
Total_amount,Number_of_tracks, Number_of_genre,Number_of_artists
FROM Customer c
        JOIN (SELECT i.CustomerId,COUNT(DISTINCT(i.InvoiceId)) AS
Number_of_Invoices,
   COUNT(DISTINCT(Track.TrackId)) AS Number_of_tracks,
COUNT(DISTINCT(Genre.GenreId)) AS Number_of_genre,
  COUNT(DISTINCT(Artist.ArtistId)) AS Number_of_artists
                        FROM Invoice i
                                JOIN Invoiceline ON i.InvoiceId = Invoiceline.InvoiceId
           JOIN Track ON Track.TrackId = Invoiceline.TrackId
           JOIN Album ON Track.AlbumId = Album.AlbumId
           JOIN Genre ON Genre.GenreId = Track.GenreId
           JOIN Artist ON Artist.ArtistId = Album.ArtistId
                        GROUP BY i.CustomerId)
        AS Infor
        ON c.CustomerId = Infor.CustomerId
        JOIN (SELECT i.CustomerId, SUM(i.Total) AS Total_amount
                FROM Invoice i
      GROUP BY i.CustomerId) AS Expense
      ON Expense.CustomerId = c.CustomerId
ORDER BY CustomerId;
SELECT * from warehouse LIMIT 25;
```

```
mysql> select * from warehouse LIMIT 25;
+------------+-----------+-------------+-------+----------------+-----------------------------+-------------------+--------------+--------------------+-------------------+--------------------+
| CustomerId | FirstName | LastName    | State | Country        | Email                       | Number_of_Invoices | Total_amount | Number_of_tracks | Number_of_genre | Number_of_artists |
+------------+-----------+-------------+-------+----------------+-----------------------------+-------------------+--------------+--------------------+-------------------+--------------------+
|          1 | Luís      | Gonçalves   | SP    | Brazil         | luisg@embraer.com.br        |                 7 |        39.62 |                 38 |                 8 |                 15 |
|          2 | Leonie    | Köhler      | NULL  | Germany        | leonekohler@surfeu.de       |                 7 |        37.62 |                 38 |                 7 |                 17 |
|          3 | François  | Tremblay    | QC    | Canada         | ftremblay@gmail.com         |                 7 |        39.62 |                 38 |                10 |                 21 |
|          4 | Bjørn     | Hansen      | NULL  | Norway         | bjorn.hansen@yahoo.no       |                 7 |        39.62 |                 38 |                 8 |                 18 |
|          5 | František | Wichterlová | NULL  | Czech Republic | frantisekw@jetbrains.com    |                 7 |        40.62 |                 38 |                 8 |                 14 |
|          6 | Helena    | Holý        | NULL  | Czech Republic | hholy@gmail.com             |                 7 |        49.62 |                 38 |                 9 |                 13 |
|          7 | Astrid    | Gruber      | NULL  | Austria        | astrid.gruber@apple.at      |                 7 |        42.62 |                 38 |                 9 |                 11 |
|          8 | Daan      | Peeters     | NULL  | Belgium        | daan_peeters@apple.be       |                 7 |        37.62 |                 38 |                 4 |                 13 |
|          9 | Kara      | Nielsen     | NULL  | Denmark        | kara.nielsen@jubii.dk       |                 7 |        37.62 |                 38 |                 5 |                 15 |
|         10 | Eduardo   | Martins     | SP    | Brazil         | eduardo@woodstock.com.br    |                 7 |        37.62 |                 38 |                 7 |                 14 |
|         11 | Alexandre | Rocha       | SP    | Brazil         | alero@uol.com.br            |                 7 |        37.62 |                 38 |                 6 |                 16 |
|         12 | Roberto   | Almeida     | RJ    | Brazil         | roberto.almeida@riotur.gov.br |                 7 |        37.62 |                 38 |                 5 |                 16 |
|         13 | Fernanda  | Ramos       | DF    | Brazil         | fernadaramos4@uol.com.br    |                 7 |        37.62 |                 38 |                 7 |                 20 |
|         14 | Mark      | Philips     | AB    | Canada         | mphilips12@shaw.ca          |                 7 |        37.62 |                 38 |                10 |                 19 |
|         15 | Jennifer  | Peterso     | BC    | Canada         | jenniferp@rogers.ca         |                 7 |        38.62 |                 38 |                 8 |                 17 |
|         16 | Frank     | Harris      | CA    | USA            | fharris@google.com          |                 7 |        37.62 |                 38 |                 7 |                  9 |
|         17 | Jack      | Smith       | WA    | USA            | jacksmith@microsoft.com     |                 7 |        39.62 |                 38 |                10 |                 13 |
|         18 | Michelle  | Brooks      | NY    | USA            | michelleb@aol.com           |                 7 |        37.62 |                 38 |                 6 |                 15 |
|         19 | Tim       | Goyer       | CA    | USA            | tgoyer@apple.com            |                 7 |        38.62 |                 38 |                 9 |                 16 |
|         20 | Dan       | Miller      | CA    | USA            | dmiller@comcast.com         |                 7 |        39.62 |                 38 |                 7 |                 17 |
|         21 | Kathy     | Chase       | NV    | USA            | kachase@hotmail.com         |                 7 |        37.62 |                 38 |                 9 |                 16 |
|         22 | Heather   | Leacock     | FL    | USA            | hleacock@gmail.com          |                 7 |        39.62 |                 38 |                 8 |                 16 |
|         23 | John      | Gordon      | MA    | USA            | johngordon22@yahoo.com      |                 7 |        37.62 |                 38 |                 9 |                 19 |
|         24 | Frank     | Ralston     | IL    | USA            | fralston@gmail.com          |                 7 |        43.62 |                 38 |                10 |                 19 |
|         25 | Victor    | Stevens     | WI    | USA            | vstevens@yahoo.com          |                 7 |        42.62 |                 38 |                 8 |                 15 |
+------------+-----------+-------------+-------+----------------+-----------------------------+-------------------+--------------+--------------------+-------------------+--------------------+
25 rows in set (0.02 sec)
```

Part two: Query Questions

1. **How much did the top 20% of customers spent compared to the rest 80% of customers?**

```
SELECT  sum(Total_amount) as total_spent, "top 20" as GROUP_number
FROM (select  Total_amount,
NTILE(5) OVER(
ORDER BY Total_amount ) group_name From warehouse) sub
WHERE group_name = 5
Union
SELECT  sum(Total_amount) as total_spent, "other groups" as GROUP_number
FROM (select  Total_amount,
NTILE(5) OVER(
ORDER BY Total_amount ) group_name From warehouse) sub
WHERE group_name < 5;
```

```
+-------------+--------------+
| total_spent | group_name   |
+-------------+--------------+
|      492.82 | top 20       |
|     1835.78 | other groups |
+-------------+--------------+
2 rows in set (0.05 sec)
```

From the result, the top 20% of customers spent $492.82, and the rest 80% of customers spent $1835.78.

2. **What are the numbers of customers per state in the USA?**

```
SELECT COUNT(CustomerID) AS Number_of_Customers, State
FROM warehouse
GROUP BY State, Country
 HAVING Country = 'USA';
```

```
+--------------------+-------+
| Number_of_Customers | State |
+--------------------+-------+
|                  3 | CA    |
|                  1 | WA    |
|                  1 | NY    |
|                  1 | NV    |
|                  1 | FL    |
|                  1 | MA    |
|                  1 | IL    |
|                  1 | WI    |
|                  1 | TX    |
|                  1 | AZ    |
|                  1 | UT    |
+--------------------+-------+
11 rows in set (0.02 sec)
```

The result shows that 3 customers are located in CA and 1 customer for each state of WA, NY, NV, FL, MA, IL, WI, TX, AZ, and UT.

3. **What is the total amount of sales for each month and year?**

```
SELECT MONTHNAME(InvoiceDate) AS Month_name, YEAR(InvoiceDate) as Year_Sales,
Total_Amount
FROM warehouse wh
JOIN Invoice iv
ON wh.customerid = iv.customerid
GROUP BY MONTHNAME(InvoiceDate), YEAR(InvoiceDate)
ORDER BY  YEAR(InvoiceDate), MONTH(InvoiceDate);
```

Month_name	Year_Sales	Total_amount
January	2009	37.62
February	2009	37.62
March	2009	39.62
April	2009	39.62
May	2009	37.62
June	2009	37.62
July	2009	49.62
August	2009	37.62
September	2009	37.62
October	2009	37.62
November	2009	39.62
December	2009	40.62
January	2010	42.62
February	2010	39.62
March	2010	39.62
April	2010	39.62
May	2010	37.62
June	2010	39.62
July	2010	37.62
August	2010	37.62
September	2010	39.62
October	2010	37.62
November	2010	37.62
December	2010	39.62
January	2011	47.62

25 rows in set (0.02 sec)

From the result, the total_amount is calculated by the sum of sales within each month and year. The largest total amount in 2009 is $49.62, which happened in July. The largest total sales amount happened in January 2010 which is $42.62.

4. **How to classify customers based on the RFM model?**

```
Select * , CASE
                    WHEN RFM_score = 111 THEN "Best customers"
                    WHEN RFM_score = 444 THEN "Churn customers"
                    WHEN Mv_score = 1 THEN "Highest Paying Customers"
                    WHEN F_score = 1 THEN "Loyal Customers"
                    WHEN R_score = 1 AND F_score = 4 THEN "Newest customers"
                    WHEN R_score = 2 AND F_score = 4 THEN "Once loyal, now gone"
                    ELSE
                    "Normal"
                    END AS Customer_segment
FROM(

Select customerid, DATEDIFF(max_invoice,most_recent) as
Recency,Frequency,Monetary_value,
R_score, F_score, MV_score,R_score * 100 + F_score * 10 + MV_score AS RFM_score
From (
(Select max(invoicedate) as max_invoice, '1' as temp From invoice) sub1
Join
```

```
(SELECT wh.CustomerID, '1' as temp, Total_amount AS Monetary_value,
                    Max(InvoiceDate) AS most_recent,
       Number_of_invoices AS Frequency,

       NTILE(4) OVER (ORDER BY DATEDIFF(MAX(InvoiceDate), InvoiceDate)) AS
R_score,
       NTILE(4) OVER (ORDER BY COUNT(InvoiceID) DESC) AS F_score,
       NTILE(4) OVER (ORDER BY Total_amount DESC) AS MV_score

FROM warehouse wh
JOIN Invoice iv
ON wh.customerid = iv.customerid
GROUP BY CustomerID
ORDER BY Total_amount DESC
)  sub2
on sub1.temp=sub2.temp
) )as casesub;
```

customerid	Recency	Frequency	Monetary_value	R_score	F_score	MV_score	RFM_score	Customer_segment
6	39	7	49.62	3	3	1	331	Highest Paying Customers
26	261	7	47.62	2	2	1	221	Highest Paying Customers
57	434	7	46.62	2	2	1	221	Highest Paying Customers
46	48	7	45.62	4	4	1	441	Highest Paying Customers
45	155	7	45.62	3	3	1	331	Highest Paying Customers
37	202	7	43.62	4	4	1	441	Highest Paying Customers
24	124	7	43.62	2	2	1	221	Highest Paying Customers
28	217	7	43.62	2	2	1	221	Highest Paying Customers
25	17	7	42.62	4	4	1	441	Highest Paying Customers
7	186	7	42.62	2	2	1	221	Highest Paying Customers
44	8	7	41.62	3	3	1	331	Highest Paying Customers
5	230	7	40.62	1	1	1	111	Best customers
43	199	7	40.62	1	1	1	111	Best customers
48	101	7	40.62	3	3	1	331	Highest Paying Customers
1	137	7	39.62	1	1	1	111	Best customers
4	80	7	39.62	4	4	2	442	Normal
42	49	7	39.62	4	4	2	442	Normal
34	447	7	39.62	1	1	2	112	Loyal Customers
17	509	7	39.62	2	2	2	222	Normal
22	168	7	39.62	2	2	2	222	Normal
20	31	7	39.62	2	2	2	222	Normal
3	93	7	39.62	3	3	2	332	Normal
51	385	7	38.62	1	1	2	112	Loyal Customers
39	106	7	38.62	1	1	2	112	Loyal Customers
15	372	7	38.62	3	3	2	332	Normal

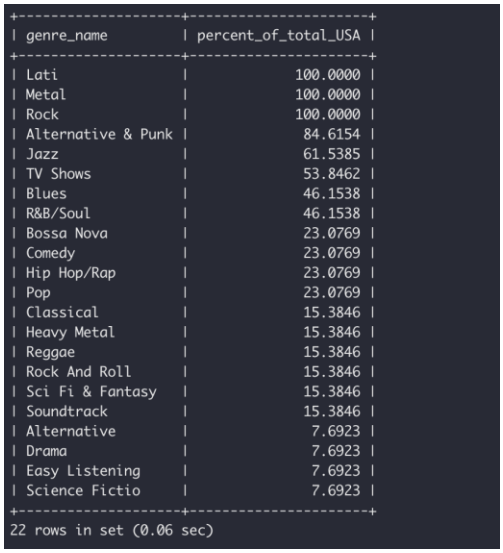
25 rows in set (0.03 sec)

In this query, customers were classified based ona their RFM_score. According to the score, customers were divided into Highest Paying Customers, Best Customers, Normal Customers, Churn Customers, and Loyal Customers. In the result, there has 12 Highest Paying Customers, 3 Best Customers, 3 Loyal Customers, and 7 Normal Customers.

5. **What is the percentage of customers based on each Genre in the USA?**

```
Select genre_name,(count(distinct(Customerid))/(SELECT Count(Distinct(Customerid))
 From warehouse
 WHERE country = 'USA')) * 100 as percent_of_total_USA
 From (
 SELECT Genre.Name as genre_name, wh.customerid,country
FROM warehouse wh
JOIN Invoice On wh.customerid = Invoice.customerid
JOIN Invoiceline ON Invoice.InvoiceId = Invoiceline.InvoiceId
JOIN track ON Invoiceline.TrackId = track.TrackId
JOIN Genre ON track.GenreId = Genre.GenreId
```

```
group by wh.customerid,Genre.Name
)AS two
 where country='USA'
 group by genre_name
 ORDER BY (count(distinct(Customerid))/(SELECT Count(Distinct(Customerid))
 From warehouse
 WHERE country = 'USA')) * 100 desc;
```



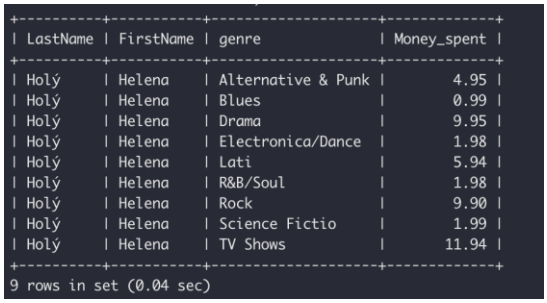
genre_name	percent_of_total_USA
Lati	100.0000
Metal	100.0000
Rock	100.0000
Alternative & Punk	84.6154
Jazz	61.5385
TV Shows	53.8462
Blues	46.1538
R&B/Soul	46.1538
Bossa Nova	23.0769
Comedy	23.0769
Hip Hop/Rap	23.0769
Pop	23.0769
Classical	15.3846
Heavy Metal	15.3846
Reggae	15.3846
Rock And Roll	15.3846
Sci Fi & Fantasy	15.3846
Soundtrack	15.3846
Alternative	7.6923
Drama	7.6923
Easy Listening	7.6923
Science Fictio	7.6923

22 rows in set (0.06 sec)

From the result, Lati, metal, and rock genres are the most popular genres of USA customers purchased in albums by 100 percent respectively. Alternative & Punk genre is purchased by USA customers by 84.6154 percent. The jazz genre is purchased by USA customers by 61.5385 percent. TV Shows genre is purchased by USA customers by 53.8462 percent. Blues and R&B/Soul genres are purchased by USA customers by 46.1538 percent respectively. Bossa Nova, Comedy, Hip Hop/Rap, and pop genres are purchased by USA customers by 23.0769 percent respectively. Classical, heavy metal, reggae, rock and roll, sci fi & fantasy, and soundtrack purchased by USA customers by 15.3846 percent respectively. Alternative, drama, easy listening, and science fictio genres are the least popular genres of USA customers purchased by 7.6923 percent respectively.

6. **How much did the top customer (based on the money spent) spend across all genres?**

```
SELECT wh.lastname, wh.firstname, g.name as genre,
  SUM(IL.UNITPRICE) as Money_spent
  FROM warehouse wh
  JOIN invoice I ON I.customerid = wh.customerid
  JOIN invoiceline IL ON IL.invoiceid = I.invoiceid
  JOIN track T ON T.trackid = IL.trackid
  JOIN genre G ON G.genreid = T.genreid
  WHERE wh.Total_amount = (Select max(Total_amount) from warehouse)
  GROUP BY wh.lastname, wh.firstname, g.name
  ORDER BY wh.lastname;
```



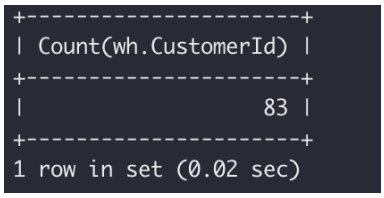
LastName	FirstName	genre	Money_spent
Holý	Helena	Alternative & Punk	4.95
Holý	Helena	Blues	0.99
Holý	Helena	Drama	9.95
Holý	Helena	Electronica/Dance	1.98
Holý	Helena	Lati	5.94
Holý	Helena	R&B/Soul	1.98
Holý	Helena	Rock	9.90
Holý	Helena	Science Fictio	1.99
Holý	Helena	TV Shows	11.94

9 rows in set (0.04 sec)

From the result, the top customer, Helena Holy, spent 4.95 on the alternative & punk genre, 0.99 on the blues genre, 9.95 on the drama genre, 1.98 on the electronica/dance genre, 5.94 on the Lati genre, 1.98 on the R&B/Soul genre, 9.90 on the Rock genre, 1.99 on the Science Fictio, and 11.94 on the TV Shows.

7. **How many customers ordered in 2009?**

```
Select Count(wh.CustomerId)
FROM warehouse wh
JOIN Invoice ON wh.CustomerID = Invoice.CustomerID
WHERE  YEAR(InvoiceDate) = 2009
GROUP BY YEAR(InvoiceDate);
```



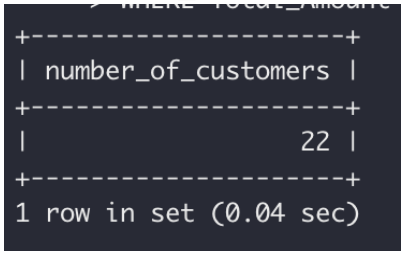
Count(wh.CustomerId)
83

1 row in set (0.02 sec)

From the result, there are 83 customers ordered in 2009.

8. **How many customers spent more than the average amount?**

```
SELECT COUNT(customerid) AS number_of_customers
FROM warehouse
WHERE Total_Amount > (Select AVG(Total_Amount) from warehouse)
```



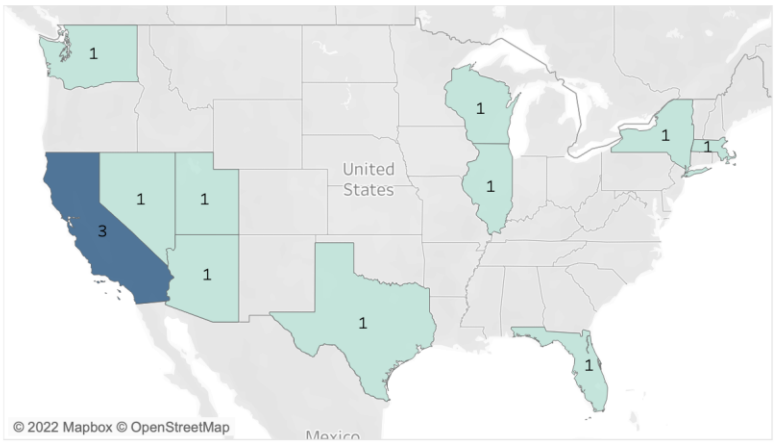
number_of_customers
22

1 row in set (0.04 sec)

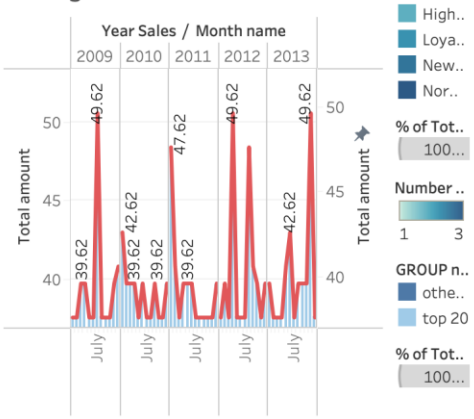
From the result, there are 22 customers spent more than the average amount.

Part 3: Visualizations

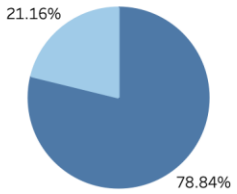
Number of Customers Per State in USA



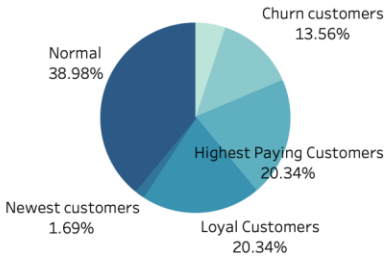
Each Month's Sale Amount Change



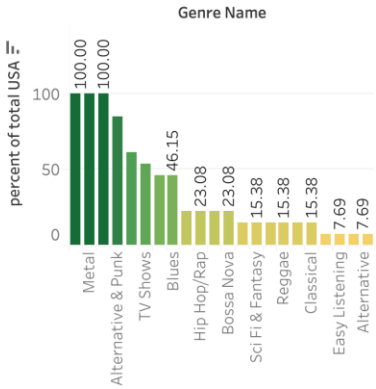
20% Customers Spent Compared To Rest Customers Spent



Customer Classification Based On RFM Model

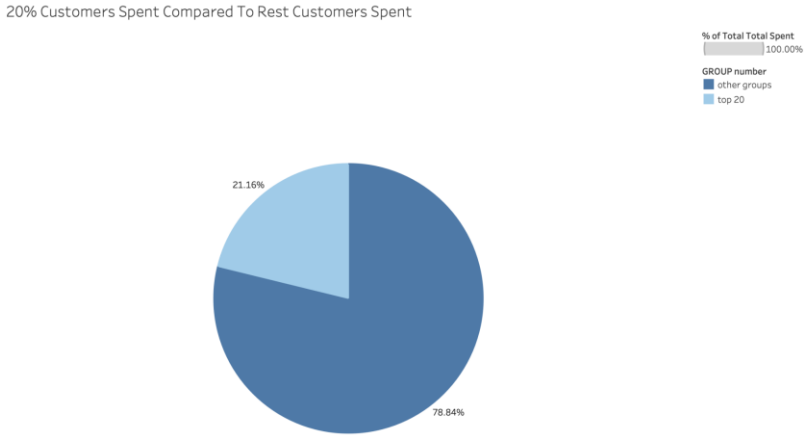


Genre Distribution among USA Customers

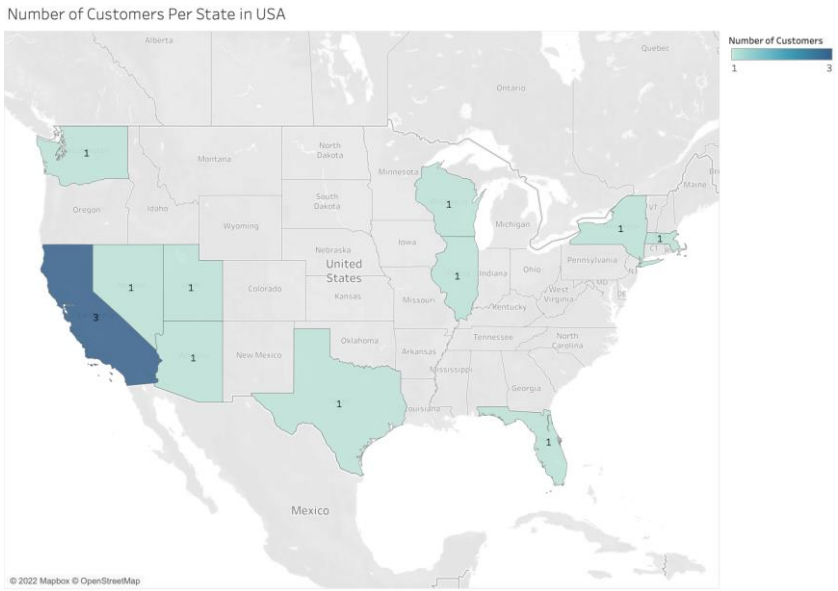


1. According to the pie chart, the spending of the top 20% of consumers accounted for 21.16% of the total, and the spending of the remaining 80% of consumers accounted for 78.84% of the total. Therefore, it can be found that the consumption of the top

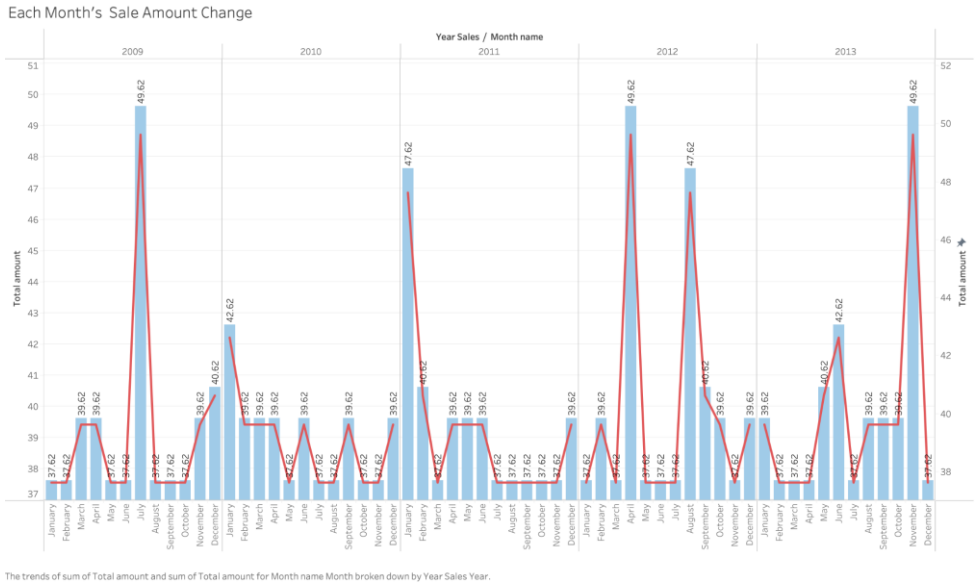
20% of customers is not significantly different from the consumption of the remaining customers because each group counts for about ⅕ of the total. We can conclude if the company wants to increase its sales, the company should increase the number of customers instead of focusing on the top 20% of consumers because they are not aficionados of albums and they won't buy a large number of albums.



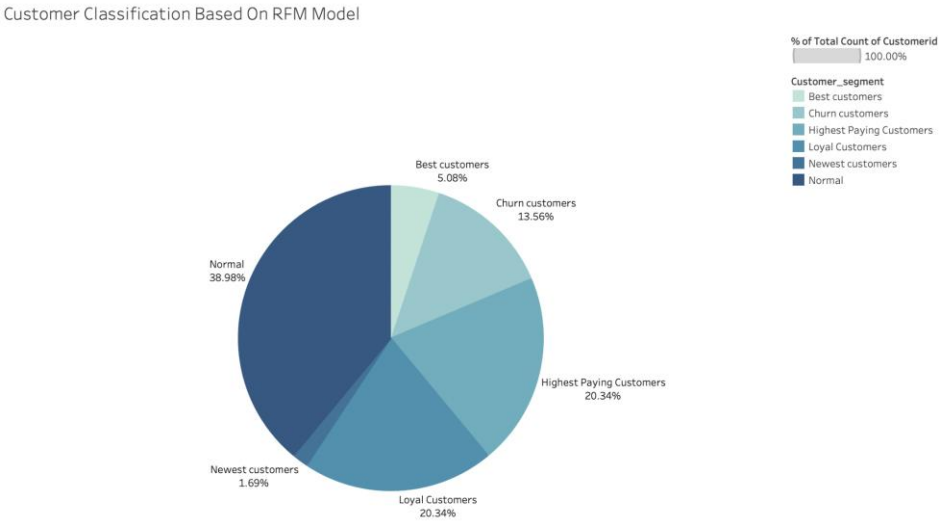
2. According to the map, California had the most customers, with three in all, ten states had one customer, and the rest had none. Therefore, California can be considered a key market for development because it has the most customers.



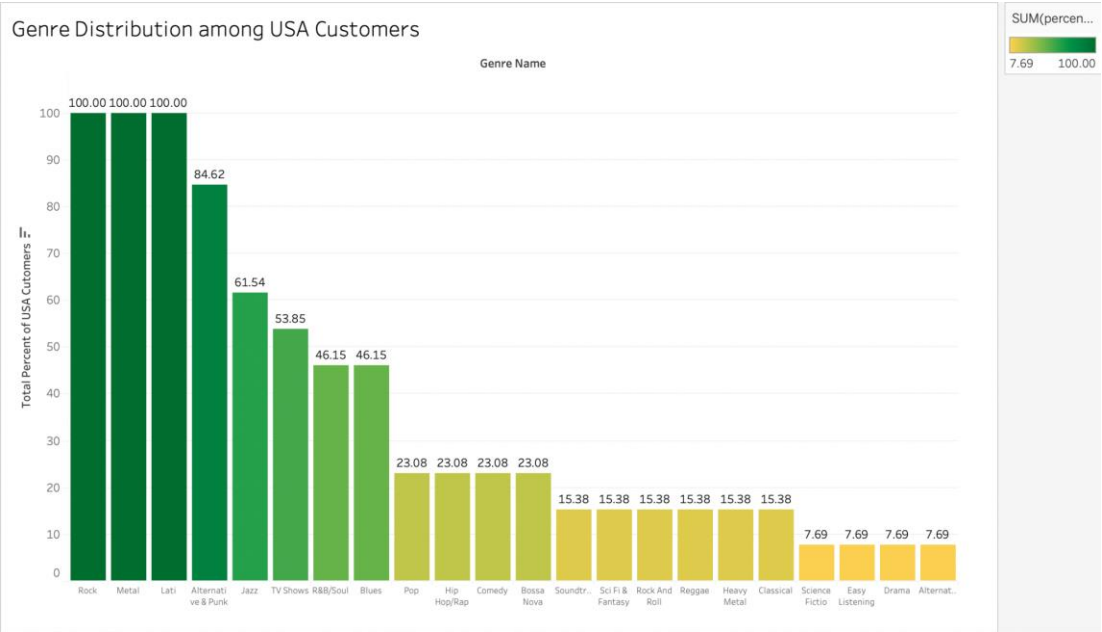
3. According to the fluctuation of the month, we can find that the annual sales amount is relatively stable. Although there is no overall downward trend, there is no overall upward trend, so for the company, it may be necessary to formulate strategies in the future to increase sales to expand its business.



4. Based on the pie chart, 5.08% of customers are Best Customers, 13.56% of customers are Churn Customers, 20.34% of customers are Highest Paying Customers, 20.34% of customers are Loyal Customers, 1.69% of customers are Newest Customers and 38.98% customers are Normal customers. Therefore, most customers are normal and there doesn't have a lot of new customers.



5. From the graph, we can tell the distribution of USA customers in purchasing which types of genres of albums. It transits from green color (the most) to yellow color (the least). The three genres (rock, metal, and lati) are the most popular genres of the albums USA customers bought. The four genres (science ficto, easy listening, drama, and alternative) are the least popular genres of the albums USA customers bought.


Genre Distribution among USA Customers
Genre Name
Total Percent of USA Cutomers
100.00 100.00 100.00 84.62 61.54 53.85 46.15 46.15 23.08 23.08 23.08 23.08 15.38 15.38 15.38 15.38 15.38 15.38 7.69 7.69 7.69 7.69
Rock
Metal
Lati
Alternative & Punk
Jazz
TV Shows
R&B/Soul
Blues
Pop
Hip Hop/Rap
Comedy
Bossa Nova
Soundtr..
Sci Fi & Fantasy
Rock And Roll
Reggae
Heavy Metal
Classical
Science Fictio
Easy Listening
Drama
Alternat..
SUM(percen...
7.69
100.00