**SQL FINAL CLASS PROJECT**

Diagram

Description automatically generated

Please put all of your work into **this** **single Word doc and also submit a Tableau Workbook file .twb and the Excel file of your Data Warehouse that you used for Tableau visualizations**. Please see instructions for Tableau below in the question 3.

1. **(160 points)** Design and create a data warehouse for the provided database. The decisions about which fields to include and how to aggregate the data are left to you. You do not need to include every single data point from the 11 tables given. Use your judgement as to what will be interesting/useful for the organization. But please make sure that you pull (combine) data from **at least six tables** and compute relevant aggregate statistics. Please compute relevant aggregate statistics for each table that you join. In your queries later in part 2, you may join your Data Warehouse with other tables to answer useful questions. Please see many examples from class lectures and you may adapt those codes for your purpose (for this dataset).

**Submit a screenshot of the first 25 rows of your data warehouse (paste into this Word document) and the SQL code that you used to create it. Please copy and paste your SQL code into this Word document. If you PC does not show 25 rows of data, please submit what you have (i.e., rows you can see on a screenshot) with a comment that you cannot show 25 rows of data. Please add a description of what your Data Warehouse will be tracking for a company.**

**Screenshot:**电脑萤幕

低可信度描述已自动生成

ArtistId: ID number for each artist

Name: Name of each Artist

Album\_num: how many albums each artist has

Genre\_num: How many genres each artist has

Frequency: how many times the artist’s song has been sold

FreqGroup: Artists are divided into 5 groups according to the frequency their tracks are sold. Those with most tracks sold are in group 1, those with least tracks sold are in group 5.

Last\_sold: Date the song was last sold

LastGroup: After sorting Last\_sold, divide them into 5 groups, where group 1 contains the artists, whose works have been purchased recently, and group 5 contains the artists whose works haven’t been purchased for a long time.

Distinct\_customers: How many customers have bought this artist’s songs

Count\_playlist: how many playlists this artist’s songs have been included in

Revenue: total revenue this artist made

Track\_num: how many songs the artist has

RevGroup: Artists are divided into 5 groups according to the revenue earned by selling their tracks. Those with most revenues are in group 1, those with least revenues are in group 5.

Genre\_explore: Indicates whether one artist with more than one album explores more genres. Limited Album means this artist only has one album. Multiple Genres means this artist has released more than 1 album and covers more than one genre. One Genre means this artist has released more than 1 album but focuses on only one genre.

**SQL code:**

USE Final\_project;

CREATE OR REPLACE VIEW final\_dw\_dm AS

SELECT a.ArtistId, a.Name , al\_num.album\_num,

genre\_num.genre\_num, freq.frequency, freq.FreqGroup, freq.last\_sold, freq.LastGroup,

tot\_cust.distinct\_customers, playlist.count\_playlist, artist\_revenue.revenue, track\_count.track\_num, artist\_revenue.RevGroup,

CASE WHEN genre\_num.genre\_num > 1 THEN 'Multiple Genre' WHEN al\_num.album\_num = 1 THEN 'Limited Album' ELSE 'One Genre' END AS genre\_explore

FROM artist AS a

JOIN ( SELECT art.ArtistId, COUNT(AlbumId) AS album\_num

FROM album AS al JOIN artist AS art

ON al.ArtistId= art.ArtistId

group by ArtistId

) AS al\_num

ON a.ArtistId= al\_num.ArtistId

JOIN( SELECT art.ArtistId, COUNT(DISTINCT GenreId) AS genre\_num

FROM artist AS art JOIN album AS al

ON art.ArtistId= al.ArtistId JOIN track AS t

ON al.AlbumId= t.AlbumId

group by art.ArtistId

) AS genre\_num

ON a.ArtistId= genre\_num.ArtistId

JOIN( SELECT art.ArtistId, COUNT(il.InvoiceLineId) AS frequency,

NTILE(5) OVER (ORDER BY COUNT(il.InvoiceLineId) DESC) AS FreqGroup,

MAX(InvoiceDate) AS last\_sold,

NTILE(5) OVER (ORDER BY MAX(InvoiceDate) DESC) AS LastGroup

FROM artist AS art JOIN album AS al

ON art.ArtistId= al.ArtistId LEFT JOIN track AS t

ON al.AlbumId= t.AlbumId LEFT JOIN invoiceline AS il

ON t.TrackId= il.TrackId LEFT JOIN invoice AS i

ON il.InvoiceId= i.InvoiceId

GROUP BY art.ArtistId

) AS freq

ON a.ArtistId= freq.ArtistId

JOIN( SELECT art.ArtistId, COUNT(DISTINCT c.CustomerId) AS distinct\_customers

FROM artist AS art JOIN album AS al

ON art.ArtistId= al.ArtistId LEFT JOIN track AS t

ON al.AlbumId= t.AlbumId LEFT JOIN invoiceline AS il

ON t.TrackId= il.TrackId LEFT JOIN invoice AS i

ON il.InvoiceId= i.InvoiceId LEFT JOIN customer AS c

ON i.CustomerId= c.CustomerId

GROUP BY art.ArtistId

) as tot\_cust

ON a.ArtistId= tot\_cust.ArtistId

JOIN ( SELECT art.ArtistId, COUNT(DISTINCT PlaylistId) AS count\_playlist

FROM artist AS art JOIN album AS al

ON art.ArtistId= al.ArtistId JOIN track AS t

ON al.AlbumId= t.AlbumId JOIN playlisttrack AS p

ON t.TrackId= p.TrackId

GROUP BY art.ArtistId

) AS playlist

ON a.ArtistId = playlist.ArtistId

JOIN (SELECT art.ArtistId, SUM(il.UnitPrice \* Quantity) AS revenue,

NTILE(5) OVER (ORDER BY SUM(il.UnitPrice \* Quantity) DESC) AS RevGroup

FROM artist art JOIN album al

ON art.ArtistId = al.ArtistId

LEFT JOIN track t ON al.AlbumId = t.AlbumId

LEFT JOIN invoiceline il ON t.TrackId = il.TrackId

GROUP BY art.ArtistId) AS artist\_revenue

ON a.ArtistId = artist\_revenue.ArtistId

JOIN (SELECT art.ArtistId, COUNT( DISTINCT t.trackId) AS track\_num

FROM artist AS art JOIN album AS al

ON art.ArtistId= al.ArtistId JOIN track AS t

ON al.AlbumId= t.AlbumId

GROUP BY art.ArtistId) AS track\_count

ON a.ArtistId = track\_count.ArtistId

GROUP BY ArtistId

ORDER BY ArtistId;

SELECT \* FROM final\_dw\_dm;

2. **(140 points)** Create **eight** SQL queries **on your data warehouse** (not on the original dataset) that answer interesting questions. At least **6** queries should be more complex queries. For example, more complex queries could include Joins, a Group By, UNION elements or a subquery or use some aggregate functions and summary calculations (see examples in the class lectures’ slides).

**Submit a copy of each query SQL code (paste into this Word document), and the screenshot of each query results (or the first 25 rows if it is longer or how many rows you can get on your PC) and full description of the question your SQL code was addressing and what you found in the results. The question that each query answers should be useful for a company to make decisions and act upon.**

1. **The company wants to know for each artist how many times each customer purchased his/her songs.**

For example, from the screenshot, Bjor Hanse purchased Artist 1’s song 4 times while Daa peeters purchased Artist 1’s song for 2 times. The graph is useful for the company to locate which customer is the top fan of which Artist. If the company wants to give back the fan benefits, they can choose a list from this, depending on how many fans from each artist they want to pick up.

SELECT fdd.ArtistId, count(c.CustomerId) as count, c.FirstName, c.LastName

FROM final\_dw\_dm as fdd join album as al on al.ArtistId = fdd.ArtistId

join track as t on t.AlbumId = al.AlbumId

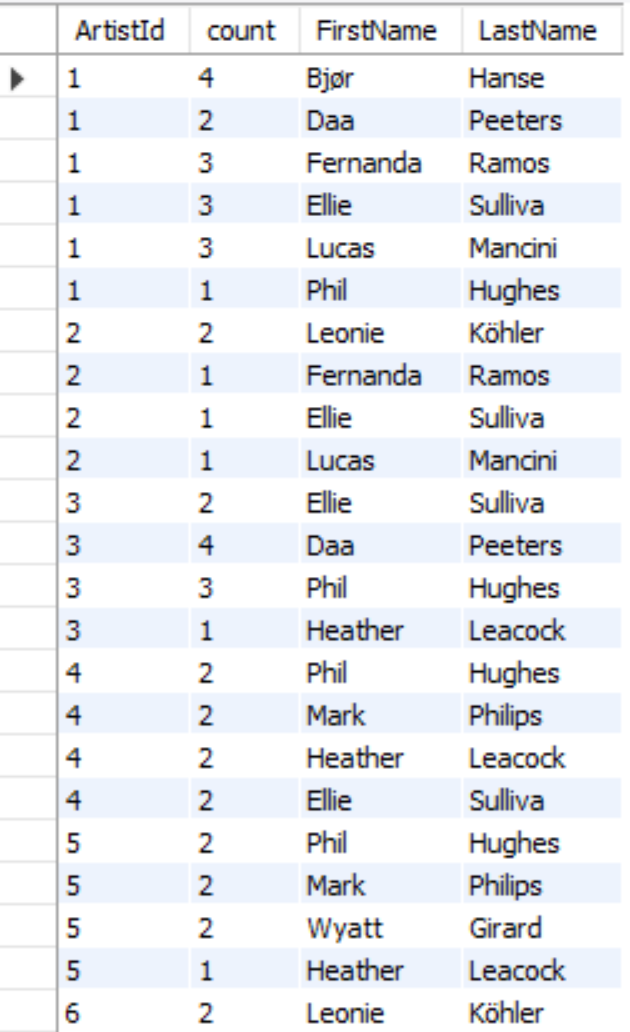
join invoiceline as il on t.TrackId = il.TrackId

join invoice as i on i.InvoiceId = il.InvoiceId

join customer as c on c.CustomerId = i.CustomerId

group by c.CustomerId, fdd.ArtistId

order by fdd.artistId;

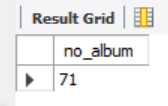


1. **How many artists do not have an album**

SELECT (COUNT(artist.ArtistId)- COUNT(art.ArtistId)) AS no\_album

FROM artist AS artist LEFT JOIN final\_dw\_dm AS art

ON artist.ArtistId=art.ArtistId;



As we are tracking everything by artist, while making the warehouse we realized not all the artists have albums, this shows the difference between the artist in the warehouse and in the total table, which 71

1. **What are the top 10 artists with most percentage of customers in USA in Revenue Group 1?**

SELECT f.ArtistId, f.Name, COUNT(DISTINCT c.CustomerId) AS USA\_customer, distinct\_customers,

COUNT(DISTINCT c.CustomerId)/distinct\_customers AS USA\_customer\_percent

FROM final\_dw\_dm AS f JOIN album AS al

ON f.ArtistId= al.ArtistId LEFT JOIN track AS t

ON al.AlbumId= t.AlbumId LEFT JOIN invoiceline AS il

ON t.TrackId= il.TrackId LEFT JOIN invoice AS i

ON il.InvoiceId= i.InvoiceId LEFT JOIN customer AS c

ON i.CustomerId= c.CustomerId

WHERE c.Country = 'USA' AND RevGroup = 1

GROUP BY f.ArtistId

ORDER BY USA\_customer\_percent DESC, USA\_customer DESC LIMIT 10;

表格

描述已自动生成

From the outcome, the 10 artists in the most earning group have the most percentage of customers in USA. Among them, Caetano Veloso and Amy Winehouse have the most percentage of customers which is 75%, followed by The Office, whose number is 66.67%. Those artists are very potential since they are the most earning ones, and they have large proportion of fans in USA local.

1. **Do artists whose tracks are sold more tend to cover more genres? Do their tracks tend to appear in more playlists?**

SELECT FreqGroup, AVG(genre\_num) AS avg\_genre\_num, AVG(count\_playlist) AS avg\_playlist

FROM final\_dw\_dm

GROUP BY FreqGroup

ORDER BY FreqGroup;

表格

描述已自动生成

It is true that those artists with higher sales in their tracks tend to cover more genres on average. The first frequency groups on average covers 1.59 genre, while the last 3 groups of artists only focus on one genre. However, this phenomenon does not hold when it comes to playlist. Those most popular artists’ tracks don’t show up in more playlists than those unpopular artists.

1. **What is the average revenue for each LastGroup (the 5 groups for each artists’ track last sold date)?**

SELECT LastGroup, AVG(revenue) AS avg\_revenue

FROM final\_dw\_dm

GROUP BY LastGroup

ORDER BY LastGroup;

表格

描述已自动生成

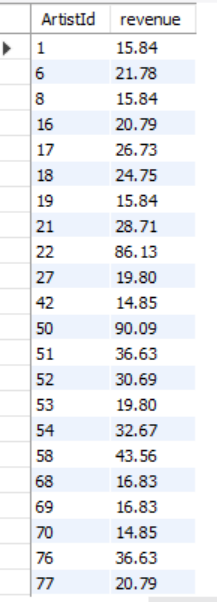
The group of artists whose tracks were sold most recently and those whose tracks were sold second most recently have the largest average revenue per artist and their number is close. Those artists in the 4th and 5th LastGroups have relatively low average revenue. This shows artists whose tracks are still sold recently tend to be more profitable.

1. **ARTISTS BRINGING IN MORE THAN AVERAGE REVENUE AND HOW MUCH?**

SELECT ArtistId, revenue

FROM final\_dw\_dm

WHERE revenue> (SELECT AVG(revenue) FROM final\_dw\_dm);



These are artists bringing in more than average of the total revenue and the amount of their revenue is also shown. The company should more focus on the artists with above average revenues like artist 50 whose average revenue reaches $90.

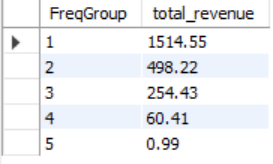
1. **sum of revenue brought in by each freqgroup**

SELECT FreqGroup, SUM(revenue) AS total\_revenue

FROM final\_dw\_dm

GROUP BY FreqGroup

ORDER BY FreqGroup;



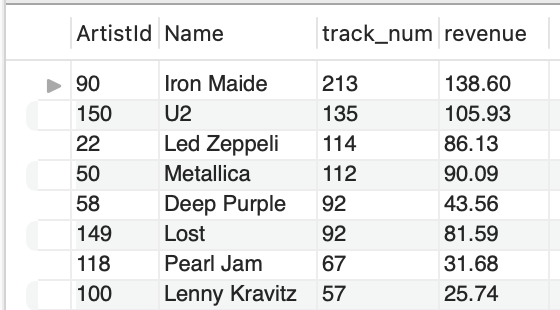
We can see from the table that groups with higher buying frequency have a higher total revenue. And, group-1, containing the artists with highest buying frequency, have a total revenue that is way higher than all the other groups. And the total revenue for group-5 is extremely low. Which might indicate that the company should give up on group-5 artists and collaborate more with group-1 artists.

1. **Who are the five most productive artists? What are the revenue they bring?**

SELECT ArtistId, Name, track\_num, revenue

FROM final\_dw\_dm

ORDER BY track\_num DESC;



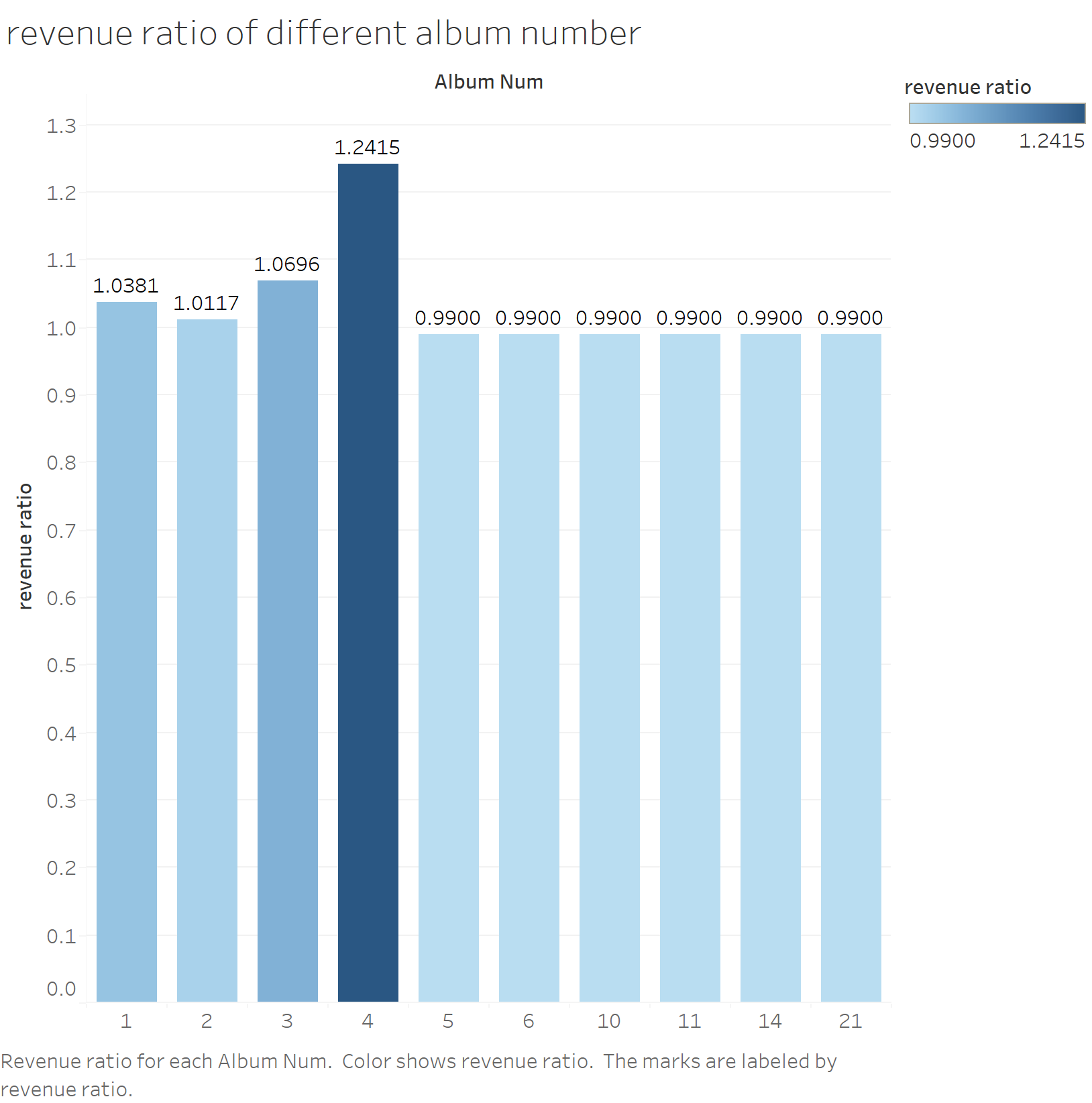
Iron Maide, U2, Led Zeppeli, Metallica and Deep Purple are the most productive five artists. And their revenue are shown in the screen shot above.

3. (**100 points**) Create **five** Tableau individual visualizations (graphs) **on your data warehouse** with valuable information to present findings to senior management of the company. Save each visualization as a png file (as we will practice in the lab 5) and paste each individual visualization png file **into this Word** document with the full explanation of what the visualizations show, how they are useful to a company and how company management could make decisions based on what you show. Finally, combine those **five** visualizations into one **Dashboard** (as we will practice in the lab 5), and save this Dashboard as a png file and **paste the Dashboard into this Word** document.

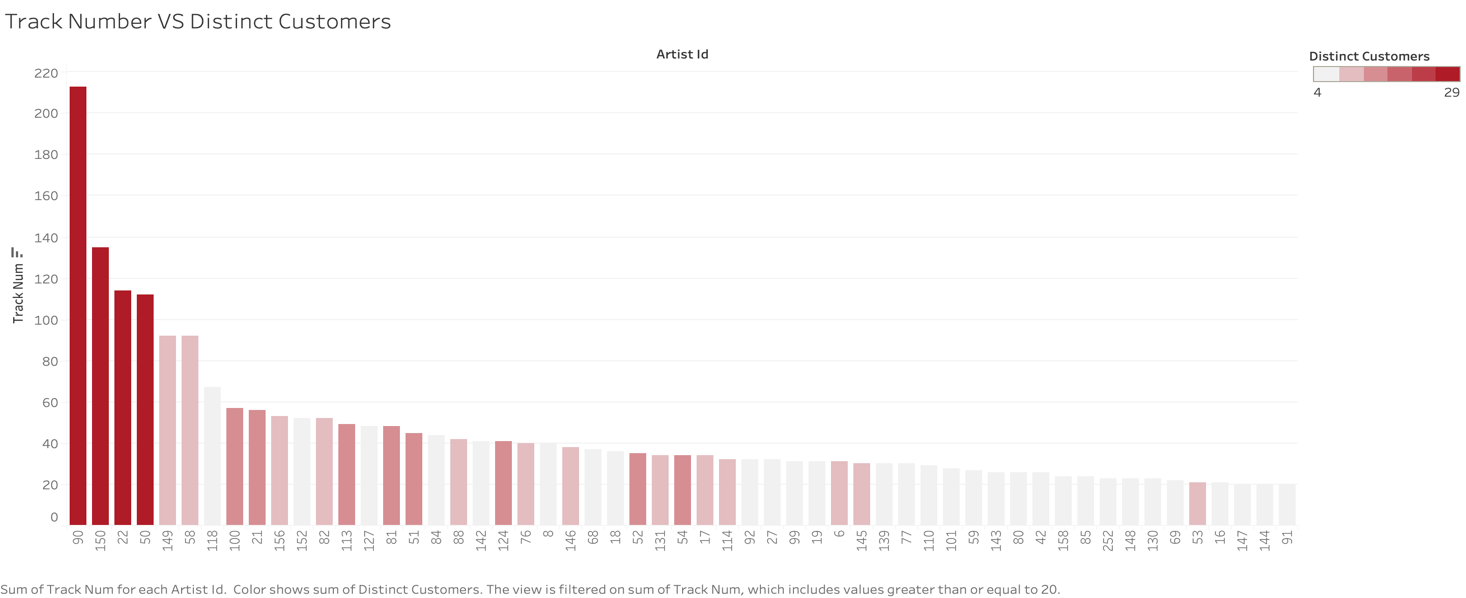
**Please also save the whole Tableau project as a Tableau Workbook file .twb (In Tableau use File - Save as) and submit to the Final Team Project folder on Blackboard together with this Word document and together with the Excel file of your Data Warehouse which you uploaded to Tableau and used for visualizations. If you cannot attach the Tableau Workbook .twb file and the Excel file for your Data Warehouse to the Final Team Project folder on Blackboard, please email the Tableau Workbook .twb file and the Excel file for your Data Warehouse to me at** [**mlysyako@simon.rochester.edu**](mailto:mlysyako@simon.rochester.edu) **indicating your class section and your team name from Blackboard and all members in the email.**

1. The company's operations are thinking about whether to encourage its artists to release more albums to create more receivables. This graph provides useful information on decision making. Looking at the chart, artists with more than five albums did not earn higher earnings. In the sample, artists with 4 albums had the highest earnings, while artists with only 1-3 albums all had an earnings ratio of more than 1. which means that more albums don't increase revenue.

Profit ratio = sum{[revenue]}/ sum{[frequency]}



1. Is it true that more productive artists are welcomed by more people? In other words, do artists with higher track\_num also have more distinct customers? (With track\_num larger than 20)



We can see from the chart that, for the first few most productive artists, this statement is true. However, for those artists who have less than 92 tracks, higher productivity doesn’t always mean more welcomed by customers.

Each track that is on sale causes its own part of operation costs. The company needs to decide wisely for choosing which artists’ works to sale, productivity shouldn’t be the only criterial, so that the cost-revenue ratio can be improved.

1. Do artists who have released several albums and explored more than one genres on average earn more than those who only focus on one single genre though having released several albums?

图表, 条形图

描述已自动生成

This visualization shows the average revenue for artists who explore multiple genres and focus on one genre in corresponding album number level. From the graph, it holds true that those artists who explore more genres earn more revenue on average than those only focus on one single genre at every corresponding level of album number. This means artists with diversified works can be more profitable than others.

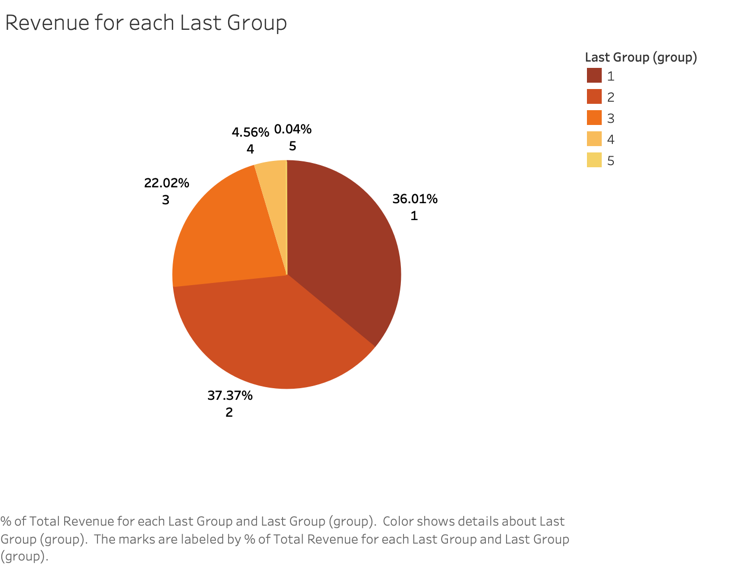
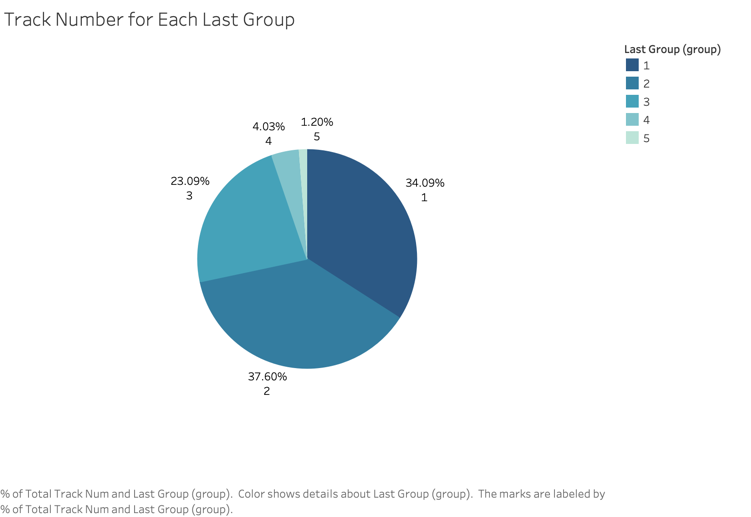
1. Are the fans of artists whose tracks are most recently bought more loyal so that more profitable than the fans of artists whose tracks are not bought recently? What about the fans of the most popular artists?

图表, 条形图

描述已自动生成

This shows the average revenue per customer of each Last Group and Frequency Group. The average revenue per customer shows how profitable one artist’s fans are. This graph shows very clear trend that the more frequently an artist’s track is bought, the more profitable their customers are. Generally, the artists with tracks being more recently bought tend to have more profitable fans, though the first frequency group in the fourth last group has especially high average revenue per customer.

1. Compare the percentage Revenue and Track Number each Last Group take in total?



Comparing these two pie-plots we can see that:

Although the revenue brought by artists from group-5 (whose works haven’t been purchased for a long time) takes only 0.04% in total, the number of tracks sold of group-5 takes 1.20% in total track sales, which is significantly larger than 0.04%.

Then looking at group-1, the revenue percentage (36.01%) is significantly larger than the track number percentage (34.09%).

Intuitively, the company might need to reduce the number of tracks on sale created by group-5 artists, and do promotions on tracks created by group-1 artists since they have higher ability of making profits.

**Please also save the whole Tableau project as a Tableau Workbook file .twb (In Tableau use File - Save as) and submit to the Final Team Project folder on Blackboard together with this Word document and together with the Excel file of your Data Warehouse which you uploaded to Tableau and used for visualizations. If you cannot attach the Tableau Workbook .twb file and the Excel file for your Data Warehouse to the Final Team Project folder on Blackboard, please email the Tableau Workbook .twb file and the Excel file for your Data Warehouse to me at mlysyako@simon.rochester.edu indicating your class section and your team name from Blackboard and all members in the email.**

图表, 条形图

描述已自动生成

**General grading criteria: Your completed work will be evaluated using the criteria below. I encourage you to use your creativity and other business skills (communication, presentation, critical thinking) in addition to the data management concepts and the SQL and Tableau skills that we have covered in CIS467.**

|  |  |  |  |
| --- | --- | --- | --- |
| **High score** | **Score between high and good** | **Good/medium score** | **Low score** |
| All required parts of the final project are complete and technically correct. Queries are useful/interesting and provide valuable information for senior management to act upon. Not just random queries. Tableau visualizations provide interesting useful information based on which senior management of the company can make important decisions. | All required parts of the final project are complete and technically correct (with possibly a few minor errors). Queries are useful/interesting and provide valuable information for senior management to act upon. Not just random queries (with possibly a few minor errors). Tableau visualizations provide interesting useful information based on which senior management of the company can make important decisions (with possibly a few minor errors). | Some required parts of the final project are missing and/or there are more significant errors. Some queries appear random and do not answer any useful/interesting questions. Tableau visualizations are very simple but may still provide interesting useful information based on which senior management of the company can make important decisions. | The final project has large portions missing and/or major conceptual errors. Most/all queries (if any) appear random and do not answer any useful/interesting questions. Tableau visualizations are very simple and **do not** provide interesting useful information based on which senior management of the company can make important decisions. |