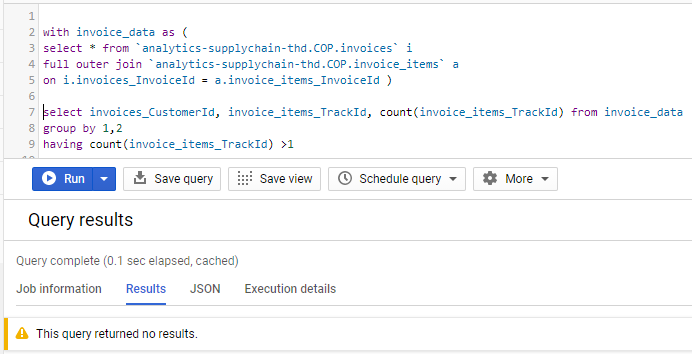
SQL case

5. Your team member discovered some customers were accidentally charged twice for a track. Help them identify who to contact by providing them with the customers’ email address and any invoice numbers related to the duplicate charge.

In general, I **didn’t find** the customers who were charged twice for a track. However, I did conduct detailed research to analyze the data. I will explain briefly how I go through the process and would appreciate if your guys could provide the correct answer.

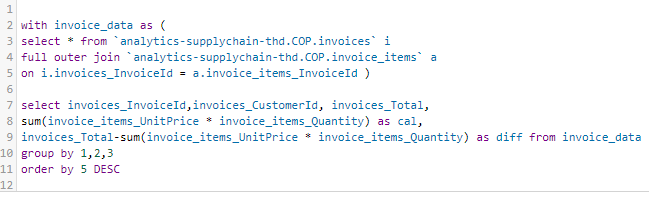
There are 3 circumstances to detect whether customers were charged twice for a track:

1. Customer may purchase 2 same track\_ids (no matter purchased in one invoice\_id or different invoice\_id):



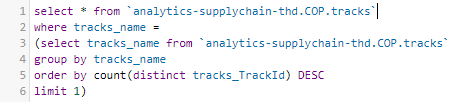
The query returned no results.

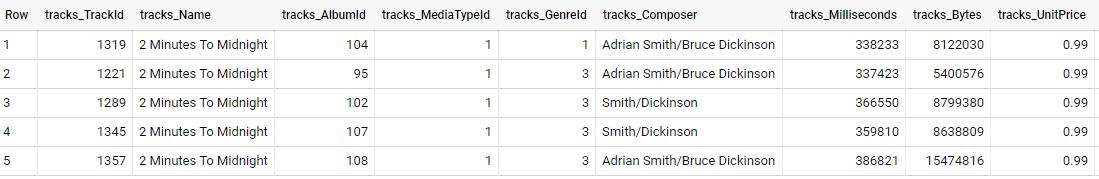
1. One track was calculated twice in invoice\_total but didn’t show in the invoice. To validate, compare the total price from invoice\_total and calculation of unit\*unit price.



The price difference is very close to 0 which means the calculation is correct for total price and there is no duplicate charge.

1. Same track name with different trackid : customers were charged twice for a same track

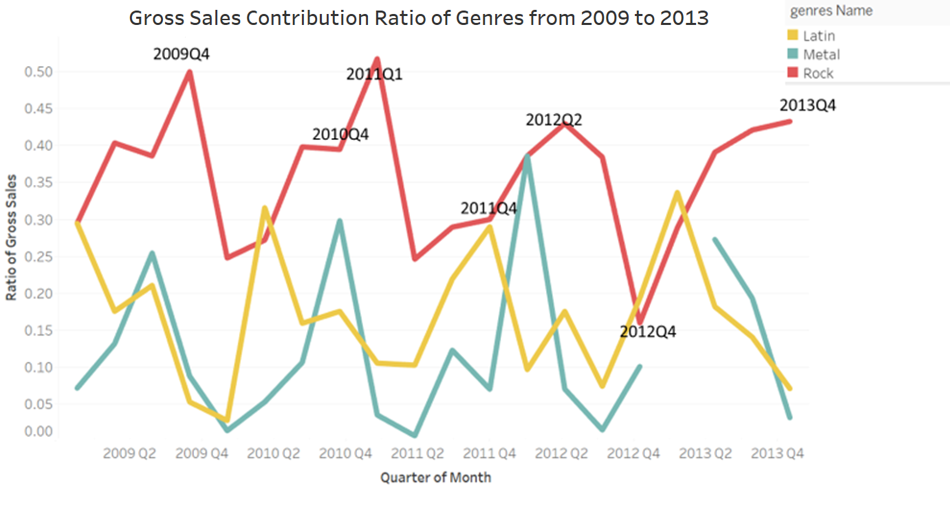


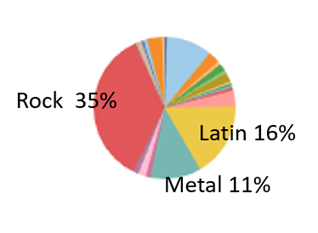


Judged by tracks Milliseconds/Bytes, although they have same name, they are not the same track.

In summary, there are no customers who were charged twice for a track if data is correct. If I got the invoice\_id and customer\_id, I will join with customers data to get their email address.

Data Viz





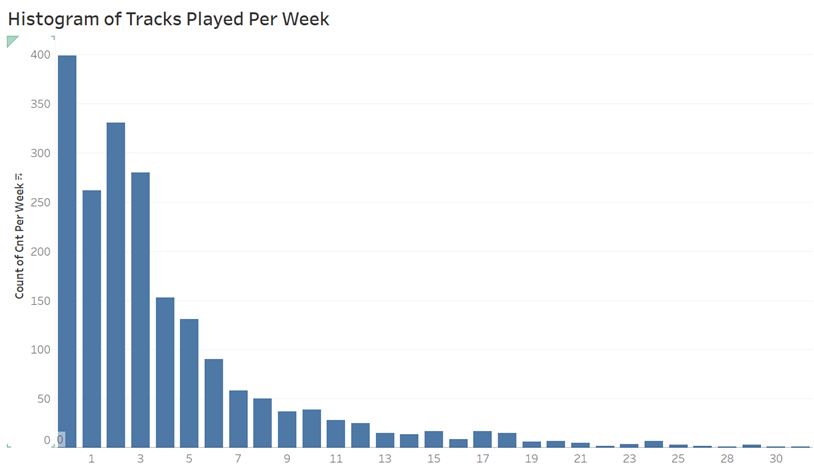


|  |
| --- |
| Answer:  There is no denying that Rock genre contributes most to the gross sales over 2009 to 2013 with 35% percent total contribution. However, I do want to see the sales trend over some top genres. |
| Firstly, I want to define the time batch to analyze the sales trend in time series. The quarter seems to be a better time range than year for its details in seasonality. |
| Also, if set month for the time batch, the plot would have too many data points in x- axis and a lot 0s for genres with fewer sales like World genre. |
| From the line chart, I try to depict the top 3 genres contributed to the gross sales which are Rock, Latin, and Metal. |
| Rock genre has the most sales all over 5 years despite very few exceptions. Latin competes with Rock well in 2009 Q1, 20010Q2 and 2013 Q1. |
| In fact, these two genres seem complementary. Latin sold very good while Rock performed worst in some quarters. |
| Some insights could also be driven from the chart. For example, there is no genre which remains popular all over the year. |
| Rock seems to boost their sales around Q4 but will encounter a large slump in the coming new year. |
| The sales of Metal seem to have a high correlation with Rock before 2013 but didn't get any sales on 2013Q1. |
| In summary, the fluctuation is large for all three genres over years. |

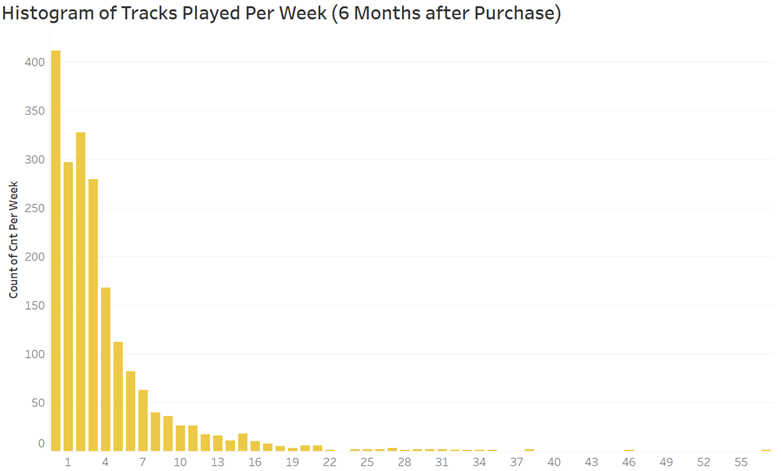
2(1)

To calculate the frequency of tracks played after purchasing, I use the total sum of tracks played between the invoice date and the last date of play data which is ”2013-12-31” divided by the weeks difference between the invoice date and “2013-12-31”. Then I could generate a list of how many times individual track will be played by customer weekly. The data could then be used in building a histogram to analyze the distribution of track playing frequency.

To understand whether tracks were listened from time to time or not played quite often after purchase, **I first decided to define 6 months as period of long time after purchase.** Therefore, I need to calculate the tracks playing frequency between 6 months after the invoice date and “2013-12-31”. There are a lot of details to be aware of, to ensure the data balance before distribution comparison, **I filter out the invoice date which is later than “2013-07-01”( 6 month before “2013-12-31”), if InvoiceDate >= "2013-07-01", then it is meaningless to add 6 months as long time to compare** Also, I need to consider the situation when customer didn’t play any individual track after purchasing. (left join play data)



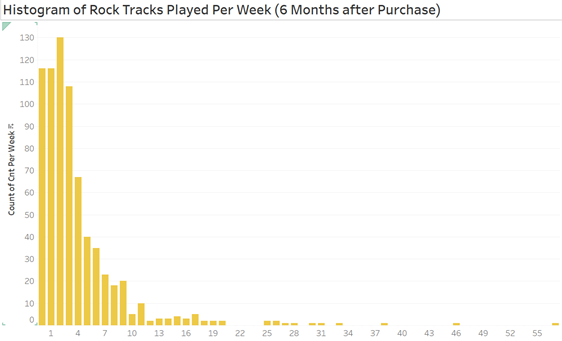
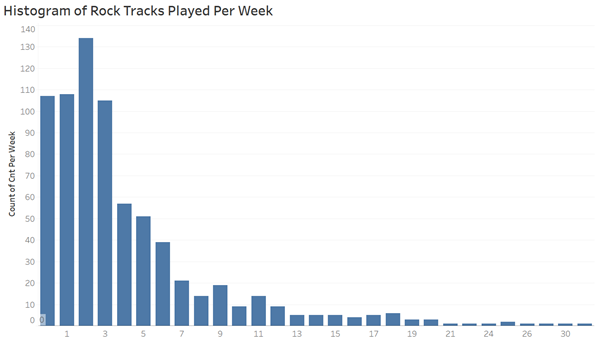
Tracks played count per week

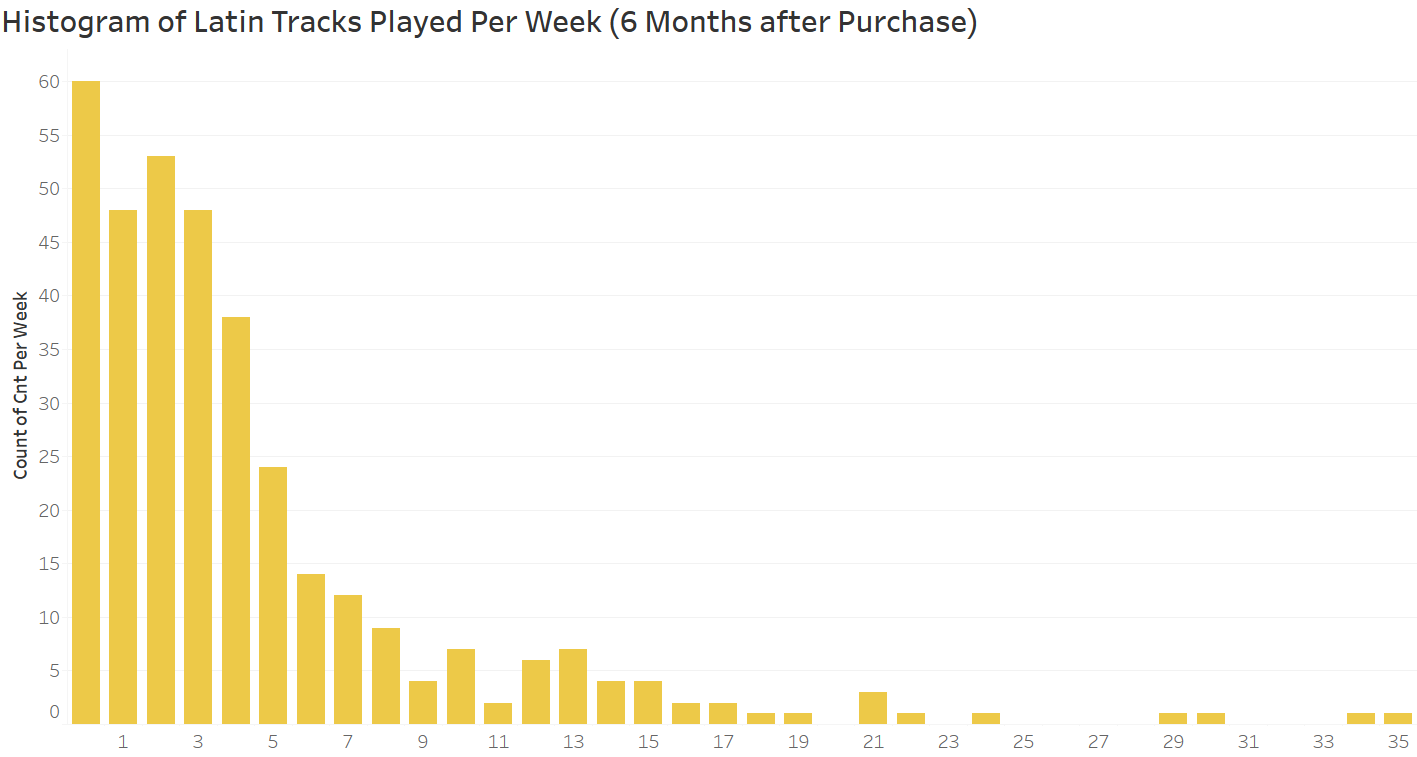
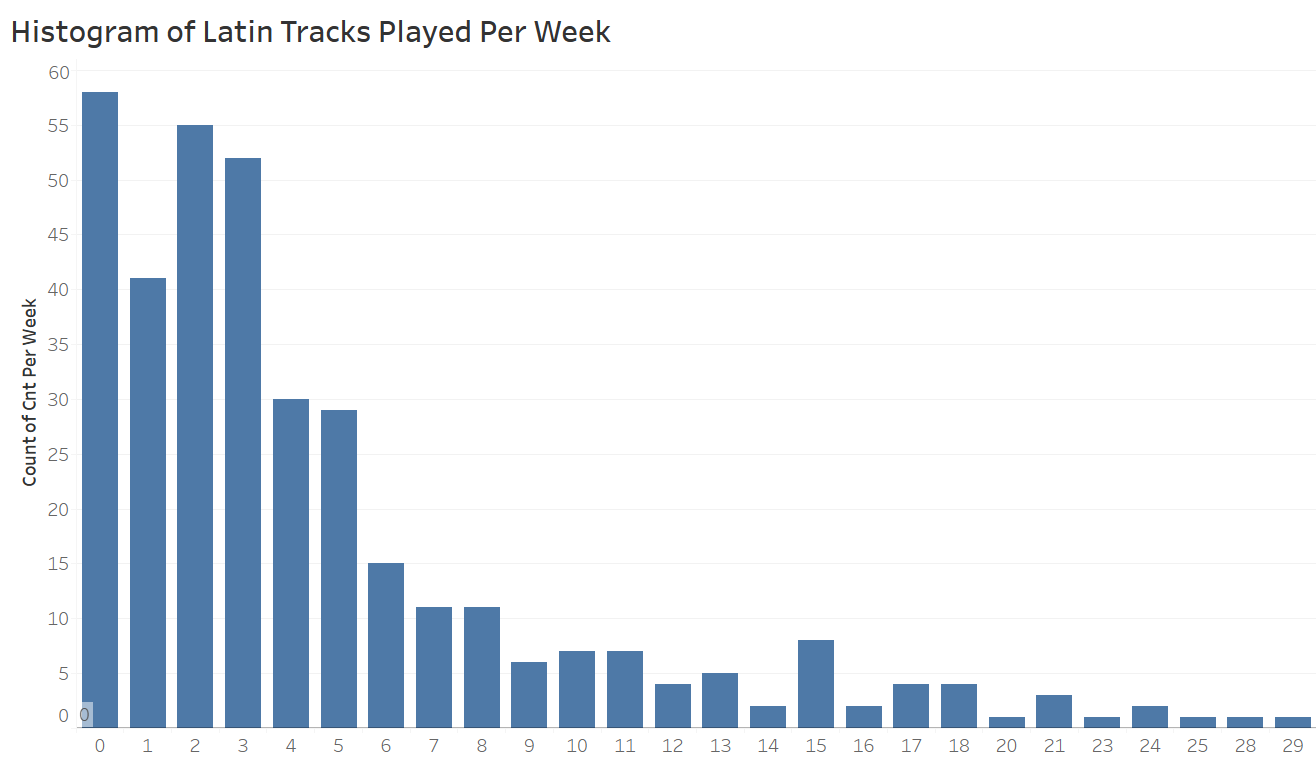


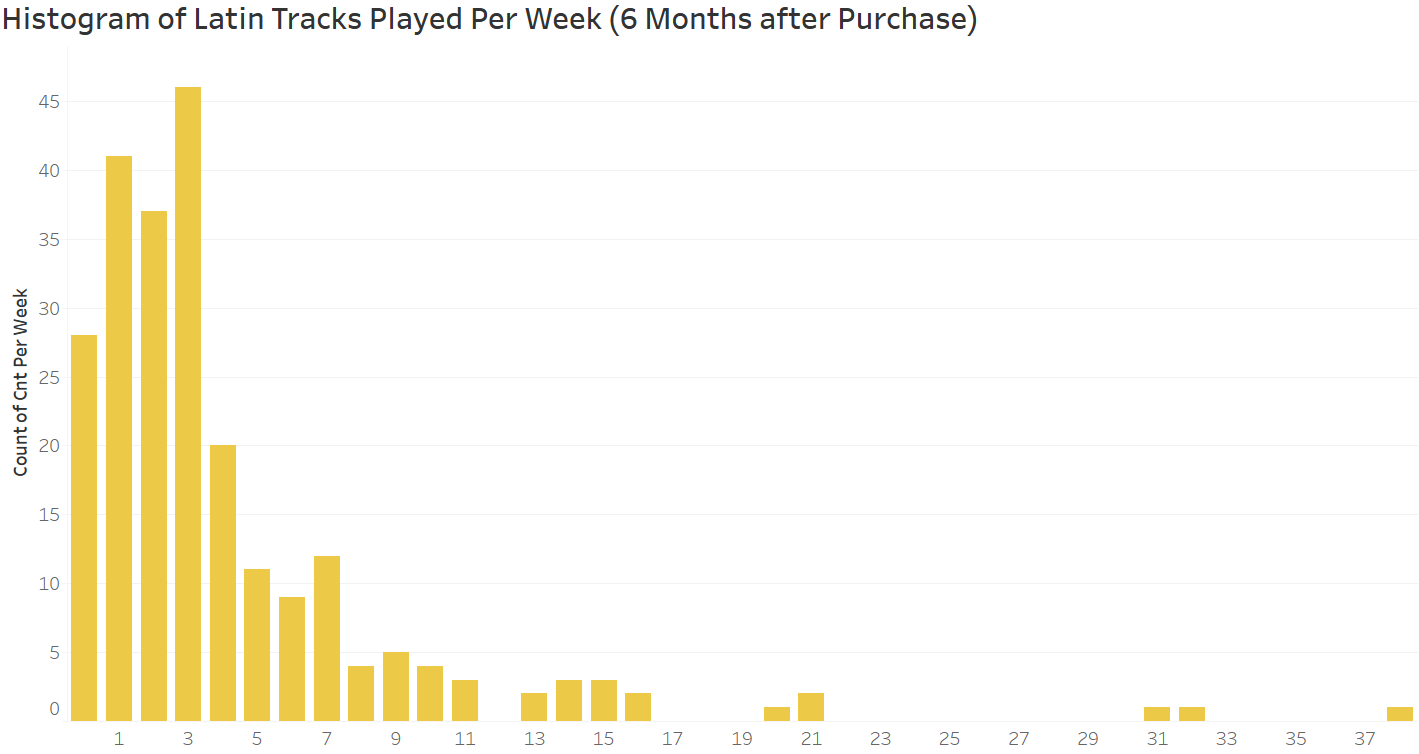
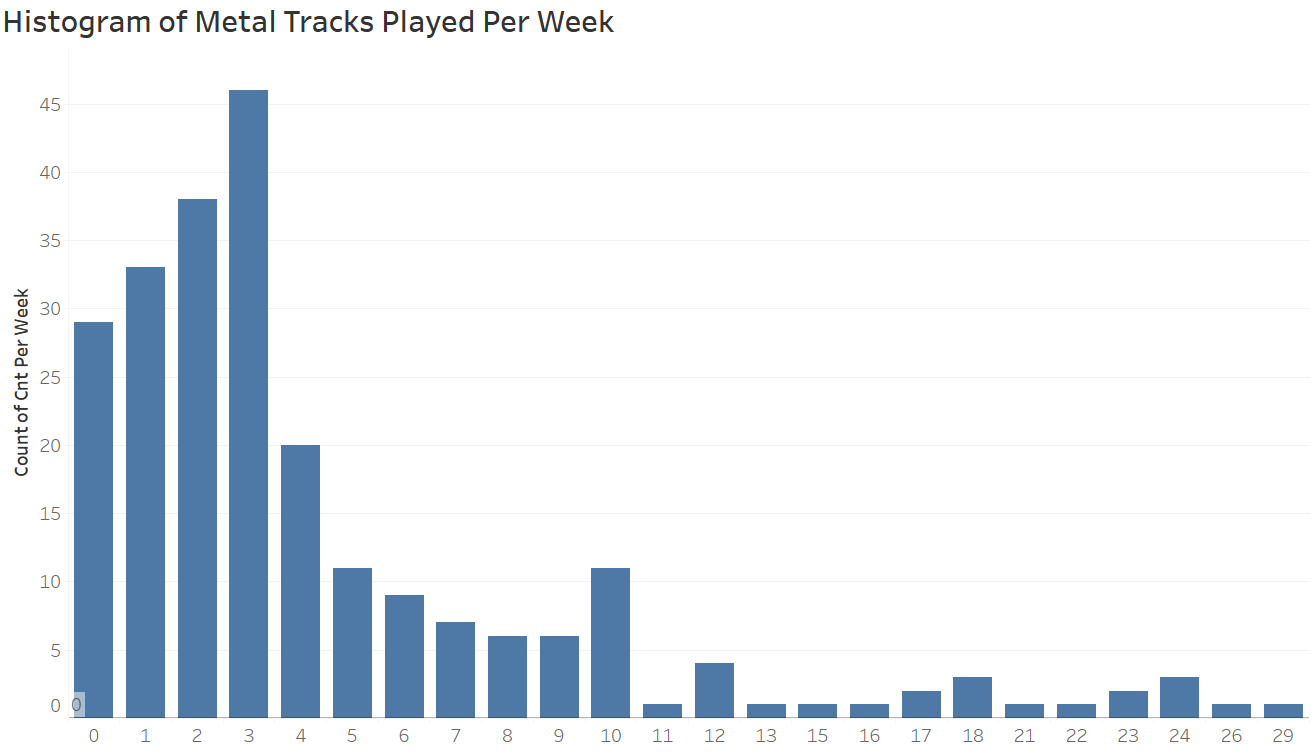
Tracks played count per week

From the histogram of tracks play frequency, I can have a general understanding of whether customer continue to listen to the same track over and over or tracks that were purchased long time ago not played as often. From the first chart, which is the normal tracks play frequency, I can see that there are 400 out of 2012 records who listened less than 1 track per week, the max tracks played per week is 35 and the median is 3.05 tacks. From the second chart which is the frequency of 6 months after purchasing the track, there are 412 out of 1995 records who listened less than 1 track per week. The median fells to 2.85 tracks played per week. However, the max tracks played goes up to 57 which maybe due to those customers who bought tracks close to “2013-07-01”. Both two histograms **are right-skewed distribution** which large number of data values occur on the left side with a fewer number of data values on the right side. **Judged by the median and histograms, track play frequency is very similar to the frequency of 6 months later after purchase which means most customer do listen to one track at comparatively same frequency.**

2(2)







As for frequency based on each genre, I select the frequency histogram of top three genre which is Rock, Latin and Metal because they have much data available for creating histogram. It seems like the top 3 genre type doesn’t really affect the track playing frequency of two time range judged by the histogram distribution as well as median. There may be some unpopular genre which influences the frequency histogram a lot. However, from my perspective, we should focus more on popular genre due to data size and validation.

|  |  |  |  |
| --- | --- | --- | --- |
| Genres | Metrics | after purchase | 6 months after purchase |
| Rock | Median track cnt | 3.1 | 3 |
| count of 0 tracks per week | 107 | 116 |
| Latin | Median track cnt | 3.54 | 3.33 |
| count of 0 tracks per week | 58 | 60 |
| Metal | Median track cnt | 3.35 | 3.19 |
| count of 0 tracks per week | 12 | 11 |

3.

Your product partners are planning to develop a tool to encourage existing customers to explore new genres. They have come to you for help to better understand what opportunities exist for genre diversification and a sizing of how many customers would be targeted. Thinking beyond the data that you currently have at your disposal, what other kinds of external datasets, research, or experiments could help the stakeholders have a better understanding of the topic area in order to make more informed business decisions? How would you go about solving a problem like this?

Since the intention is to analyze benefits of new genre exploration, I will first measure the correlation between numbers of genres played by individual customer and final sales to understand whether listening to more genres will produce more sales. Sales is like the outcome of funnel data. Also, individual genre could be categorized as dummy variable in a **linear regression** to test the significance of each genre type in uplifting the money spent or other output metrics like counts of “add to favorite”. If the result shows positive outcome, then it gives us reasons to explore new opportunities and develop new tools.

Judged by the histogram, it is known that some genre will boost the track played frequency for the customers and may lead to better gross sales. So it is important to understand the logic of recommendation system and try to recommend genres like Rock, Metal or Latin which is cash cow to those targeted customers who share similar genre listening behavior.

For example, if Rock and Jazz share a high **cosine similarity** in listening pattern by each customer which means those who listened to Rock plays the same list of tracks as those listening Jazz. **User based collaborative filtering** could be applied so we can recommend Rock to those jazz customers in the new tool.

After the company built out the recommendation tool, I will build an **ab test** to test the new tool. The company can first split targeted customers (like frequent app users) into two groups, one group with no changes at all and the other group has the new recommendation function in app. As for the output metrics, apart from money spent in the app, a lot more external metrics are helpful to measure the ab test result including the counts of “add to favorite", the time stays in the app or the sign-up rate. A two independent sample t test could be used to analyze the assumption. To save the cost and avoid novelty effect, the customer groups must be selected, and the test should take at least one month to analyze.