1. **Data Exploration and Cleansing**

**2. What is count of records in the fresh\_segments.interest\_metrics for each month year value sorted in chronological order (earliest to latest) with the null values appearing first?**

Graphical user interface, application, table

Description automatically generated

3. **What do you think we should do with these null values in the fresh\_segments.interest\_metrics?**

Graphical user interface, text, application, email

Description automatically generated

Since the percentage of null values is less than 10% we shall drop it

Graphical user interface, text

Description automatically generated with medium confidence

Graphical user interface, text, application, chat or text message

Description automatically generated

**4. How many interest\_id values exist in the fresh\_segments.interest\_metrics table but not in the fresh\_segments.interest\_map** **table? What about the other way around?**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**5. Summarise the id values in the fresh\_segments.interest\_map by its total record count in this table.**

Graphical user interface, text, application, chat or text message

Description automatically generated

1. **What sort of table join should we perform for our analysis and why? Check your logic by checking the rows where interest\_id = 21246 in your joined output and include all columns from fresh\_segments.interest\_metrics and all columns from fresh\_segments.interest\_map except from the id column.**

We should use outer join (left join).Because the interest\_map consist all the interest\_id which is a primary key .

Graphical user interface, text, application

Description automatically generated

1. **Are there any records in your joined table where the month\_year value is before the created\_at value from the fresh\_segments.interest\_map table? Do you think these values are valid and why?**

Graphical user interface, text, application

Description automatically generated

We have 188 rows where the month\_year value is before the created\_at value from the fresh\_segments.interest\_map table. I think these values are valid because months are the same, and the value in the month\_year column has the first day of the month but we do not know the real day of the month as we created this column by combining month and year only.

**Interest Analysis**

1. **Which interests have been present in all month\_year dates in our dataset?**

Graphical user interface, text, application, email

Description automatically generated

1. **Using this same total\_months measure - calculate the cumulative percentage of all records starting at 14 months - which total\_months value passes the 90% cumulative percentage value?**

Graphical user interface, text, application, email

Description automatically generated

1. **If we were to remove all interest\_id values which are lower than the total\_months value we found in the previous question - how many total data points would we be removing?**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

1. **Does this decision make sense to remove these data points from a business perspective? Use an example where there are all 14 months present to a removed interest example for your arguments - think about what it means to have less months present from a segment perspective.**

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

1. **After removing these interests - how many unique interests are there for each month?**

Graphical user interface, application

Description automatically generated

### Segment Analysis

1. **Using our filtered dataset by removing the interests with less than 6 months worth of data, which are the top 10 and bottom 10 interests which have the largest composition values in any month\_year? Only use the maximum composition value for each interest but you must keep the corresponding month\_year**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

1. **Which 5 interests had the lowest average ranking value?**

Graphical user interface, text, application, email

Description automatically generated

1. **Which 5 interests had the largest standard deviation in their percentile\_ranking value?**

Graphical user interface, text, application

Description automatically generated

1. **For the 5 interests found in the previous question - what was minimum and maximum percentile\_ranking values for each interest and its corresponding year\_month value? Can you describe what is happening for these 5 interests?**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated

1. **How would you describe our customers in this segment based off their composition and ranking values? What sort of products or services should we show to these customers and what should we avoid?**

The customers in this segment love to travel, some of them are probably business travellers, they prefer luxury lifestyle and go into sports. We should show the products or services related to luxury travel or luxury lifestyle (furniture, cosmetics, apparel), and avoid budget segment or any product or services related to random interests like computer games or astrology. We also can exclude some topics related to locations that are out of area of the customers' interests like Tampa or Oregon, because the customers possibly has already visited those locations and do not wish to return there. Also we can exclude topics related to some long-term needs and the long-term use products, that the customers have probably already purchased. For example, if a customer had an interest in Luxury Furniture or Gym Equipment, they might have purchased those products and do not have interest in this topic anymore.

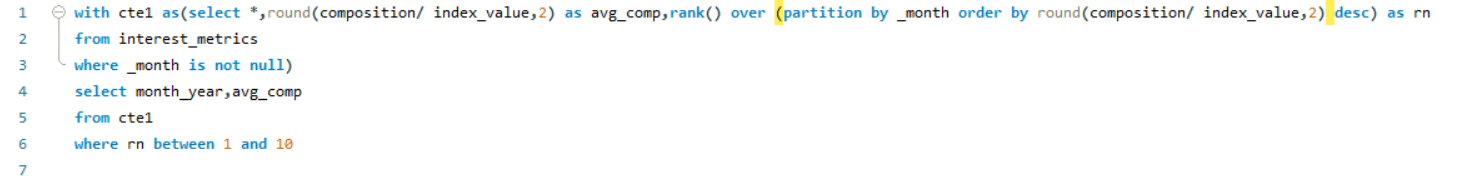
So in general we need to focus on the interests with high composition value but we need to track this metric to define when customers lose their interest in the topic.

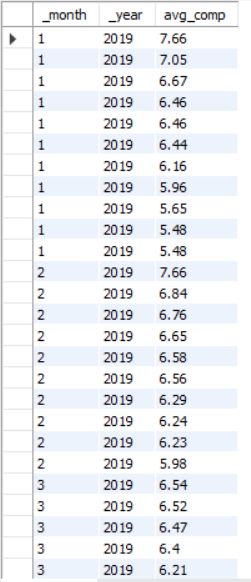
### Index Analysis

**The index\_value is a measure which can be used to reverse calculate the average composition for Fresh Segments’ clients.**

**Average composition can be calculated by dividing the composition column by the index\_value column rounded to 2 decimal places.**

1. **What is the top 10 interests by the average composition for each month?**





1. **For all these top 10 interests - which interest appears the most often?**

Graphical user interface, application

Description automatically generated

1. **What is the average of the average composition for the top 10 interests for each month?**

Graphical user interface, text, application

Description automatically generated