Q&A: Accenture Dragonyte Use Case

**Subcategory**

1. In Subcategories\_csv, there is a subcategory of “Beer” inside the category “Beer”, can we merge this with another subcategory so that it is clearer / more representative with an actual subcategory? Or how would we interpret this subcategory that bears the same name as its category? -  Em, Ghent
2. In the subcategory file : you do have some beer subcategories   sub cat 1 to 12  and then a beer category ( nb 13) what is inside that category compare the the 12 others sub-beers category above -  Sylvain, Bouman7

* *After merging the categories & subcategories together, join these with the channel volume, company share & market size file. Use the information you have in the other columns to determine the exact use of category & subcategories, the “hierarchy level” column together with the “lowest level” column will give you better insight into the actual use of the categories & subcategories.*

**Hierarchy Levels:**

1. What exactly does ‘Hierarchy\_Level’ mean? - Bouman7, Brussels
2. In Channel\_Volume, what do "Hierarchy\_Level" and "Outlet\_Hierarchy" represent?- What does"hierarchy" mean here and what is the difference between these two features? What would Level 1, the highest Level in a Hierarchy as explained by the Dictionary, actually mean? - Em, Ghent (edited)

* *As mentioned, level 1 is the highest hierarchy. This allows us to sort the categories & subcategories. I would, as a data analyst, try to understand what the impact is when “playing” with the data. Let’s try to create an overview of how the subcategories and categories are intertwined i.e.:*

*Category A*

*Subcategory A1*

*Subcategory A2*

*…*

*Category B*

*…*

*As a tip: use the Market Share file, get a total volume grouped by category, subcategory (additionally lowest level & hierarchy level).*

* *For Outlet Hierarchy, do the same as for hierarchy level but with the channel volume file, here it will be more complex as the “lowest outlet level” column does not exist.*

**Channels**

1. In the Channel\_volume.csv, can we interpret the channels as the outlet through which the drinks are sold (e.g. restaurants vs. grocery stores)? - Alfiya, Ghent
   * *Yes*
2. How should we interpret hierarchy mean in this database? - Ariana, Ghent
   * *See exercise done above.*
3. In the Channel\_volume.csv, what does the volume data represent? For example, does the volume 2613,845506 translate to ~2613.84  millions of liter?
   * *First standardize your data, what would be the impact if you round your data on million of liters? It is always possible to show the client a rounded value but store the full value in the dataset.*
4. In the Channel\_volume.csv, since the unit of volume different (millions of liter & 1000 liters) across categories do you suggest we should make different analyses across categories? - Ariana, Ghent
   * *One of the first steps you need to do when working with data is cleaning up your data. For this question, standardization. After you have done this, you will be able to have the same analysis on rows with both unit type.*
5. In ‘Channel\_Volume.csv’ columns ‘Category’ and ‘Subcategory’ contain the same information. Is it possible to update this file so that we can use it to analyze across subcategories? - Polina, Brussels
   * *The client only have the data as given. My database contains these rows and I am not able to go into the past to collect & split this data.*
   * *From a consulting point of view this insight is very interesting, as we can now propose to the client to change their data collection. We can implement this for the client and the client will have more insight into the market and their products.*
6. In ‘Channel\_Volume.csv’ we have information for 2007-2021, in ‘Company\_Share\_GBO\_unit.csv’ for 2016-2021 and in ‘Market\_Sizes.csv’ for 2016-2026. Is there any specific reason for that? - Polina, Brussels
   * *Client has collected data from 2007-2021 for channel volumes, but only “recently” has started collecting data for the Company share & Market Size. They have also already started with a prediction for market size.*
7. In Channel\_Volume, what do "Year\_text" and "Year\_date" represent? The year of what, and what is the difference between these two features?
   * *Each file/table contains data for a specific timeframe. Data cleanup is key, do you as the analyst think that it is required to keep both?*
8. In Channel\_Volume we only have one unique value for the "Data\_Type" feature, namely "Off-trade Volume". Is this because:
9. Dragonyte does not do on-trade sales (does not sell its products directly in bars, restaurant, pubs)
10. it does do on-trade but this data is not available at the moment
11. it does do on-trade and it's covered by some of the categories inside off-trade outlets? - Em, Ghent
    * *b, data for this is not present in the files.*
12. We were not able to reconcile the channel volumes with either the market file  volume or the company share file volumes  (on comparabe perimeter). We wonder if the channel file volumes relate to the market or to the company sales? Sylvain - Bouman7
    * I would advise keeping the 3 main files as separate. It is possible to connect channel volume & market size, but for this exercise having a solid overview & analysis for the 3 separate files is more than enough.

**Market Sizes:**

1. the Market\_Sizes.csv, the data in the columns Year\_minues\_2016 and Year\_minus\_2022 do not actually equal year-2016 and year-2022, but year-2012 and year-2018 respectively. Are we supposed to assume the data is correct or the labeling? - Jean, Brussels
   1. What is the use of this column? Do we need this column to analyse our data? This column is a prime example of bad data management from the client, they have a full column dedicated to a calculation. Which as you pointed out, is not even correct as they probably changed the name of the column but forgot to update the actual calculation. Or even worse, have old data that is calculated on 2012 and new data that is calculated on 2016. As a data analysis you first clean up the dataset before you analyse.
2. In Market\_Sizes.csv some Year\_date is not matching Year. Are we supposed to assume that in Year is the correct information? - Polina, Brussels
   1. Try to use other columns to determine the correct year, if not possible let the client know that some of its data was not possible to analyse on & exclude from dataset.
3. In Market\_size, how do we interpret Current\_Constant, Currency\_Conversion, RSP, Year\_minus\_2016, Year\_minus\_2022, Edition(2022)? - Daryoush, Ghent
   1. Current / Constant: <https://datahelpdesk.worldbank.org/knowledgebase/articles/114942-what-is-the-difference-between-current-and-constan>
   2. Currency conversion: <https://en.wikipedia.org/wiki/Fixed_exchange_rate_system>
   3. RSP lookup on google
   4. Year minus… see point 10
   5. Edition is the edition of the file
4. What is the real unit for the volume columns as it can not be in liters as specified in the  data dictionary. Indeed when you divide the RSP in million by the volume you get too high average price per liter - Sylvain, Bouman7
   1. Million liters

**Locations**

1. In ‘Locations.csv’ columns ‘Region’ and ‘Country’ have the same information. Is it possible to update this file so that we can’t use it to analyze across countries? - Polina, Brussels
2. In the Location file, country and region respective names are same and not accessible on map in Tableau - Archana, Brussels
   1. Data is not present on in the dataset, all data that the client has is shared.

**Predictions:**

1. Are we expected to employ a machine learning model in Python, or utilize trend analysis within Tableau, for the predictive modeling aspect? - Bouman7, Brussels
   1. It is up the team to decide which model is used but you need to be able to explain the reason why you used a specific model.
2. We are assuming that the profit margin is the same for all the categories/subcategories. Is this correct?  - Bouman7, Brussels
   1. Profit margin is not present in the data, it is therefor impossible to do any analysis or prediction on this. I would propose to predict on the data that is present i.e. volume & rsp