Johan Tokarskij 2024.06.14

Bean Stream

Introduction

I have had the honor of executing this project for the fictional online company Bean Stream, whose main business is the sale of coffee beans, ground coffee, and coffee pods to Scandinavian countries (although the project did not involve a real client, I will write my diary as if it were real, to provide a realistic picture of how the project would be carried out in a real situation). The company collaborates with a few suppliers to provide the finest coffee products on the market. They have existed since April 2022, and now, about two years later, they want to gain insights from their data to make well-informed decisions.

I have chosen to divide this diary into two parts: In part 1, I will briefly touch upon the company's startup phase where I, as a data engineer, model and create a database that the company can use for its sales. Part two will deal with the company's operational issues and challenges, and how the sales data that the company has aggregated during its two years of operation can help answer these questions. Here, I take on the role of data engineer/analyst and try to help the company gain insights from their data through Power BI dashboards and visualizations.

Part 1: Database Modeling

To meet Bean Stream's requirement to be able to record their sales data, I have chosen to model a database of the OLTP type with a relatively simple set of necessary tables so that the database maintains its data integrity while being easy to understand and maintain. I normalized the database schema by splitting tables and defining relationships between them, as well as using appropriate data types and constraints to ensure the consistency of the database. In addition to the usual customer, product, and order-related tables, I have implemented the company's specific requirements for tracking price history, reviews, and returns through the PriceHistory, ProductRating, and OrderReturn tables.

Part 2: Business Analysis with Power BI

Problem Formulation

In this phase, where I create a Power BI dashboard for the company, I tried to apply Design Thinking as much as possible within the project's scope. I aimed to follow the principles of understanding the user's needs, defining the problem, generating ideas, creating prototypes, and testing solutions to ensure that the dashboard was as user-friendly and effective as possible.

Together with Bean Stream, I began by formulating the business questions they wanted to answer using their data. These questions would form the basis for the "storytelling" in the upcoming Power BI dashboard. The following questions and issues were formulated:

1. What are the revenues for each category, product, supplier, and country?
2. What are the costs for each category, product, supplier, and country?
3. What are the profits for each category, product, supplier, and country, both currently and historically over time?
4. What are the profit margins for each category, product, supplier, and country, both currently and historically over time?
5. How many orders do we have, and what are the best-selling categories, products, and suppliers?
6. How does the price compare to the cost per unit for each product and supplier over time?
7. What are our most popular products and suppliers based on the number of units sold?
8. What are our best and worst-rated products and suppliers, and what are their countries of origin?
9. Which products and suppliers generate the most returns, what are the most common reasons, and is there any correlation between the rating and the number of returns?
10. Who are our active customers (with at least one order), and what is the demographic distribution of these customers?
11. Bean Stream wants to find a solution to a common business problem for online stores: how to identify and calculate customer churn. They have observed that customers who stop shopping regularly often do not close or delete their accounts, meaning that the simple solution of using the 'AccountDeleted' column in the 'Customers' table would lead to misleading results. Is there another solution?

12. Who are the best and worst customers based on revenue?
13. What is customer satisfaction like?
14. What is the geographical distribution of our customers who make returns?

These questions provided a clear roadmap for the development of the Power BI dashboard, ensuring that it would deliver actionable insights tailored to Bean Stream's specific business needs and challenges.

Storytelling

To create a product that meets expectations, I needed to identify the intended end-user of the dashboard. I determined that the company's management is the end-user, whose goal is to gain insights into their sales to make informed decisions. I observed a clear pattern in management's questions, which I could categorize into three groups: sales, products, and customers. Therefore, I conceptually divided the questions between different pages ("Overview," "Products," "Customers") in my dashboard and created a separate aggregated KPI for each question. This approach ensures a clear and structured presentation.

I categorized the determined questions as follows and created the following KPIs to answer them:

- **Overview**:
  - **Net Revenue**: What are the revenues for each category, product, supplier, and country?
  - **COGS**: What are the costs for each category, product, supplier, and country?
  - **Gross Profit**: What are the profits for each category, product, supplier, and country, both currently and historically over time?
  - **Gross Profit Margin**: What are the profit margins for each category, product, supplier, and country, both currently and historically over time?
  - **Total Orders**: How many orders do we have, and what are the best-selling categories, products, and suppliers?
- **Products**:
  - **Total Products from N Suppliers**: I chose to create a "landing page" to provide an overview of the company's products.
  - **Profit Margin Over Time**: How does the price compare to the cost per unit for each product and supplier over time?
  - **Units Sold**: What are our most popular products and suppliers based on the number of units sold?

    - **Average Product Rating**: What are our best and worst-rated products and suppliers, and what are their countries of origin? I also created a tooltip page "Product Rating Details" that pops up and shows all reviews for the selected product when hovering over the bar chart.
    - **Total Returns**: Which products and suppliers generate the most returns, what are the most common reasons, and is there any correlation between the rating and the number of returns?
- **Customers**:
    - **Number of Customers with at least 1 order**: Who are our active customers (with at least one order), and what is the demographic distribution of these customers?
    - **Inactive Customers**: The highlighted issue in question 11 about customer churn.
    - **Net Revenue for N Highest/Lowest Buyers**: Who are the best and worst customers based on revenue? I created a tooltip page "Customer Order Details" that provides an overview of the customer's orders. Additionally, there is a custom drill-through page "Customer Order Details" that shows all details about orders.
    - **Customer Satisfaction for N Most/Least Satisfied**: What is customer satisfaction like? There is also a link to a tooltip page "Product Rating Details" that shows all customer reviews here.
    - **Total Returns**: What is the geographical distribution of our customers who return products?

Additionally, I noticed that certain questions can be applied to different subcategories. For example, the question "What are the revenues?" can be broken down into subcategories such as "for each product, supplier, category, country" (which is also applicable to several other questions). Moreover, there are narratives such as "best/worst" and "highest/lowest." These patterns led me to consider finding a solution where, instead of creating different visualizations for each question, I could send dynamic data to a few visualizations, creating a clearer structure. This would then form the basis for how my UX/UI should be designed (more on this in the sections "UX/UI" and "Practical Implementation").

EDA

Before proceeding to test my ideas about dynamic content, I wanted to investigate whether the data and structure owned by Bean Stream would be sufficient to answer the aforementioned questions. Since an analysis of certain KPIs over time

was requested, I followed best practices and created a separate date table (with all dates for the period and without any gaps) to ensure that the analysis over time would be accurate and consistent.

The data owned by Bean Stream was deemed insufficient to address the issue highlighted in question 11. Therefore, I suggested that instead of using the 'AccountClosed' column, a logic could be applied to identify inactive customers based on how many days they have not made a purchase. This way, it would be possible to determine whether a customer was inactive/lost, even if they had not deleted their account.

UX/UI

My overall idea for UX/UI was to create clear, informative, and aesthetically pleasing data visualizations by minimizing chart overload and eliminating all distractions by limiting what the user sees on the screen. This is to facilitate information processing. By applying dual process theory, which describes how our thinking occurs through two different processes – one fast, intuitive and one slow, analytical – I aimed to design a user experience that supports fast and intuitive thinking. By reducing unnecessary visual elements, users can quickly and easily grasp important information without cognitive strain.

By applying principles of nudging, I also wanted to encourage users to focus on relevant information and make well-informed decisions. This means that the design would not only present data effectively but also discreetly guide the user towards important insights and actions.

To ensure that the data visualizations were both effective and user-friendly, I followed Edward Tufte's “Six Principles of Graphical Integrity”:

- Data-Ink Ratio: I chose minimalist visualization types without unnecessary effects or details that might distract the user.
- Chartjunk: I removed unnecessary axes, gridlines, and other visual elements to keep visualizations simple and clean.
- Data Density: Instead of creating different visualizations for different KPIs, these were logically grouped and categorized. Through various methods and techniques, I created visualizations that would allow the user to dynamically change the content.
- Small Multiples: I used the same pattern that the user would recognize between different visualizations of KPIs.

- Sparklines: I used aggregated KPIs for different categories of visualizations to provide context.
- Narrative Structure: I logically grouped visualizations and different pages with relevant names.

Regarding the UI-specific part, I chose to use a monochromatic color palette with an accent color in line with Bean Stream's logo. Just as Bean Stream uses this color for their motto 'Brew, Sip, Connect' and to highlight important design elements like the saucer and heart, I used the accent color for Call To Action elements in the user interface, such as buttons for slicers, menus, and various input fields.

I also followed the principles of contrast, repetition, alignment, and proximity (C.R.A.P.) to ensure that my Power BI dashboard is both aesthetically pleasing and functional:

- Contrast: I used contrasting colors to highlight key data, with the accent color from Bean Stream's logo for important Call To Action elements and metrics. I used the Adobe Color tool to verify that the contrast ratio is satisfactory.
- Repetition: By consistently using the same colors, fonts, and graphic elements, I created a cohesive design that helps users quickly understand the interaction.
- Alignment: All elements are aligned to create an orderly layout, making the dashboard easier to read and interpret.
- Proximity: Related information is grouped together to clearly show connections and reduce cognitive load, facilitating quick and well-informed decisions.

I used hierarchical principles for different elements in the dashboard. The company's logo was placed at the top left to create brand awareness and provide a clear reference point. By using different sizes and colors for text, I created a visual hierarchy that helps users quickly identify and focus on the most important information. The aggregated KPIs are prominently displayed at the top, while detailed information and interactive elements are placed further down – this, together with visual separations and spacing to define sections, improves the user experience.

To enhance user experience, I have also integrated a dark mode using a toggle button inside the “hidden menu”.

Practical Implementation

In line with my overall idea for UX/UI, I created a “hidden” menu for slicers and the button to switch to “dark mode.” The menu opens with a button designed with Call To Action colors and a recognizable menu icon, keeping user-friendliness in mind. The choice of dimensions for the slicers is based on providing users the ability to further filter the data they wish to analyze (e.g., specific category, supplier, product, country, and/or time period). I also chose to use images of categories, supplier and product logos, and flags in the slicers to make them more visually appealing.

My overall idea with dynamic visualizations is to allow the user to break down each aggregated KPI into relevant visualizations and different dimensions – thus, the user would only see the relevant (associated with the selected KPI) visualizations on their screen. Each aggregated KPI became a button that controls what is displayed on the dashboard. Due to the practical limitations of Power BI, the implementation varied on different pages:

- Aggregated KPIs on the “Overview” page can be broken down further into different dimensions (categories, suppliers, products, supplier countries, and customer countries) and displayed with two visualizations – a bar chart categorized by the selected dimension, and development over time with a column chart. The practical implementation involved using Fields Parameters for both the bar chart and the column chart. This makes a DAX measure with the selected KPI dynamically sent on the x-axis and a user-selected dimension on the y-axis.
- On the “Products” and “Suppliers” pages, I chose to use different types of visualizations for different aggregated KPIs to better convey the message. Considering the limitations in Power BI that prevent dynamically switching between different types of visualizations, my workaround was to use a combination of bookmarks, Fields, and Numeric Range Parameters. On the “Products” page, I implemented bookmarks to switch between different types of visualizations and Fields Parameters to dimension them to Products or Suppliers. On the “Customers” page, I used a combination of bookmarks to switch between KPIs and both Fields and Numeric Range Parameters to dynamically select a number of most/least satisfied customers and the customers who have bought the most/least.

The practical implementation of question 11 (KPI “Inactive Customers”) involved a Fields Parameter that allows the user to choose the number of days a customer has not made a purchase to consider the customer “inactive.”

Technical Limitations

Smart Narratives - there are a number of documented limitations by Microsoft for the functionality of Smart Narratives (https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-smart-narrative):

- Using dynamic values and conditional formatting (for example, data-bound titles)
- Summaries of visuals whose columns are grouped by other columns and for visuals built on a data group field
- Cross-filtering out of a visual
- Summaries of visuals that contain on-the-fly calculations like QnA arithmetic, complex measures such as percentage of grand total, and measures from extension schemas.

Since my dashboard uses bookmarks as well as Numeric Range and Fields Parameters, the functionality of Smart Narratives is affected – the output of Smart Narratives in my dashboard became somewhat limited and may not be entirely representative of the entire dataset. To address the issue, my solution was to write a custom narrative and add some variables that Smart Narratives could not handle correctly.