# Which BVG routes are most pleasant and why

**Task**

Brainstorming how I might approach this question.   
Which data sources I might be able to use.   
How they are likely to look and be structured  
How to link them all together and make the into useful structured data that could be used to spot correlations between a range of factors and overall customer satisfaction.

**Data Source 1: BVG Customer Survey**

The BVG does periodically survey its customers and has some data around customer satisfaction that it monitors. It is likely that some of this data includes individual route ids.  
So hope for some structured data with fields like: Route-id, Overall Satisfaction Score, Common Issues.

This table would be top in the hierarchy as I want to use the satisfaction score as a core metric to compare across the other data sets (so this would be the left most table in any joining process.

**Data Source 2: BVG Social Channels – Scrape and Analyse comments**

The resulting list of comments would likely contain some references to individual routes that could be pulled out and listed against those ids. Columns being: Route Mentions, Sentiment, Keyword, full comment.

Route mentions could be converted to match Route-id from Table 1

**Data Source 3. BVG usage Data**

BVG will have data on no of users and punctuality of service by route: Route\_id, daily no of users, punctuality.

**Data Source 4. BVG other info**

Probably other information that you could manually link to the routes, such as: Rolling Stock type, Frequency of announcements, works that took place on the line etc.