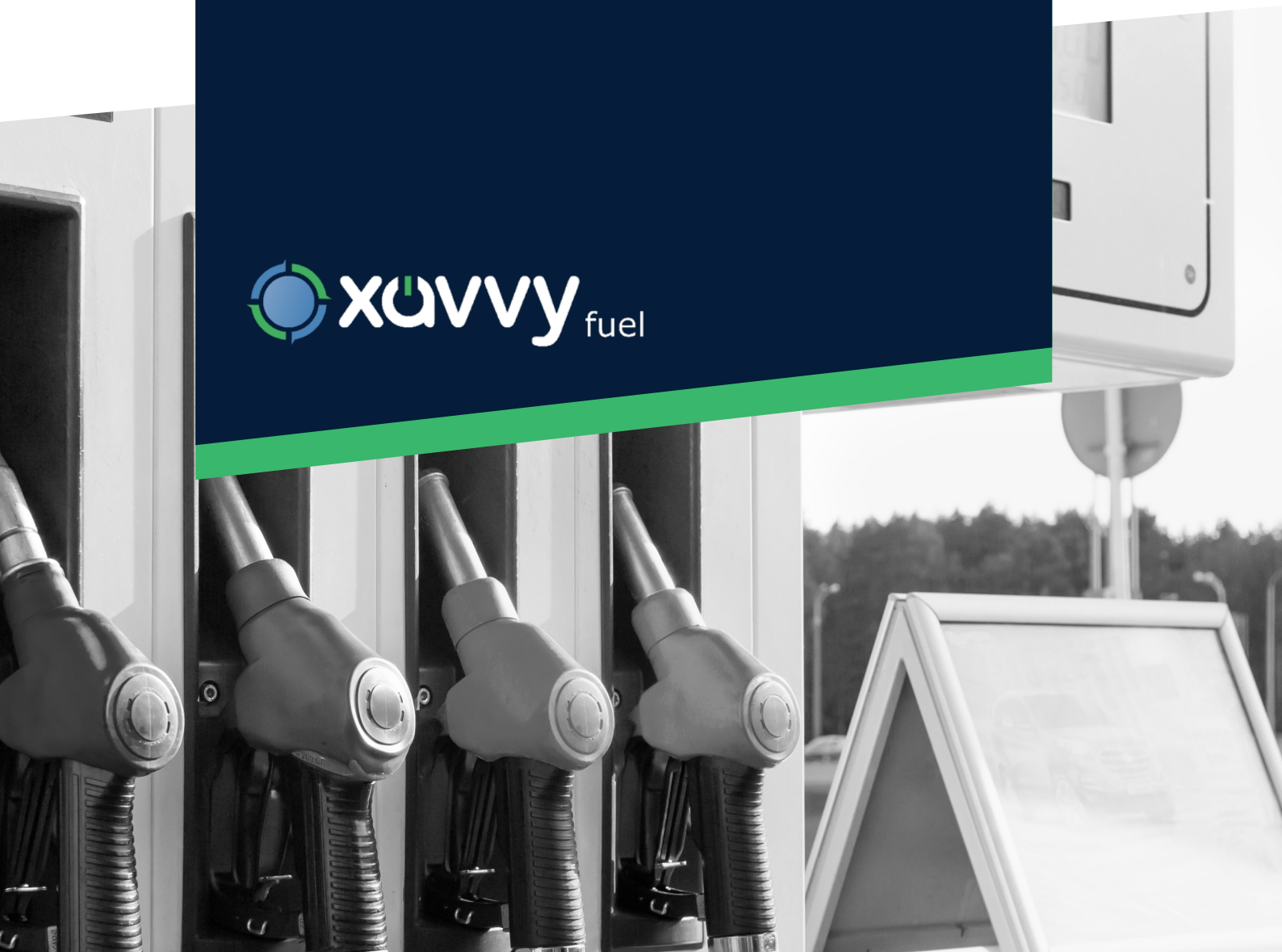
Enhancing Data Quality in Big Data Applications

xavvy fuel's approach for gas station data

xavvy fuel



Introduction

2

Don‘t let bad data leave you stranded: xavvy fuel is your trusted partner for enhancing data quality of all kinds. Get in touch with us today to learn more.

Big Data applications are highly complex and diverse, posing unique challenges when it comes to data quality. These applications are typically used to process and analyze large volumes of data from different sources, making it difficult to ensure that the data being used is accurate and consistent. Common quality problems include data inconsistencies, duplication, missing data, and inaccurate data entries. **These quality problems can significantly impact the reliability and credibility of the insights and decisions derived from Big Data**, underscoring the need for robust data quality enhancement techniques.

This is particularly important for the use case of xavvy fuel and its customers of gas station data. xavvy fuel is a Big Data application that aggregates data from over 230 heterogeneous sources related to gas stations. With more than 2 million raw data points, this platform generates a comprehensive dataset of more than 270,000 unique gas stations, for which 500 million price reports are registered daily. The accuracy and completeness of this data is critical for several use cases, such as competitor analysis of fuel retailers or enrichment of mobile apps offered by fuel card providers.

In this article, we first use the example of gas station data to show the effects that poor data quality can have. Gas station data comprises of various attributes such as brand name, geocoordinates, address-information, available fuel types and amenities, etc.

We then discuss various approaches to improving data quality in Big Data.

The Consequences of Poor Data Quality

3

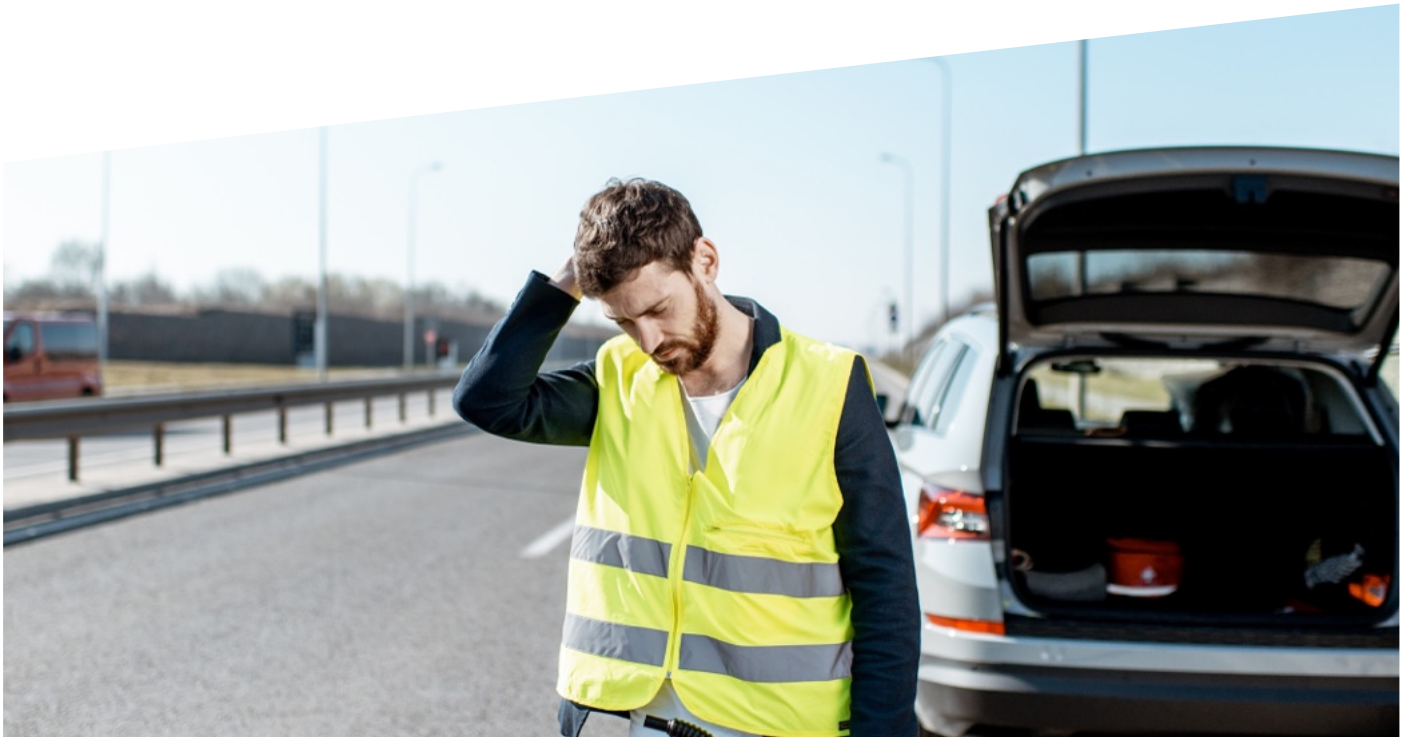
Once upon a time, there was a car driver named John Doe who was tasked for an important customer appointment in Lyon, France. Being a tech-savvy driver, he relied on his trusty mobile app to find gas stations on his route from Cologne, Germany to Lyon. However, he quickly encountered a few problems that left him scratching his head.

The first problem arose when John Doe wanted to wash his car before leaving Cologne in the morning. He found a gas station on the app a few minutes away, but to his dismay, **the station did not offer any car washing services.** Since he was already late, he couldn‘t seek out an alternative, so he had to abandon his plan of driving a sparkling clean car and start his journey in a dirty one.

The second problem occurred when John reached the border to Belgium and needed to fuel up his car. The app suggested a gas station that offered especially favorable cheap prices compared to the area. However, when he arrived, he realized **that the prices were much higher than those indicated** on the app. Feeling cheated, John reluctantly filled up his car and continued his journey.

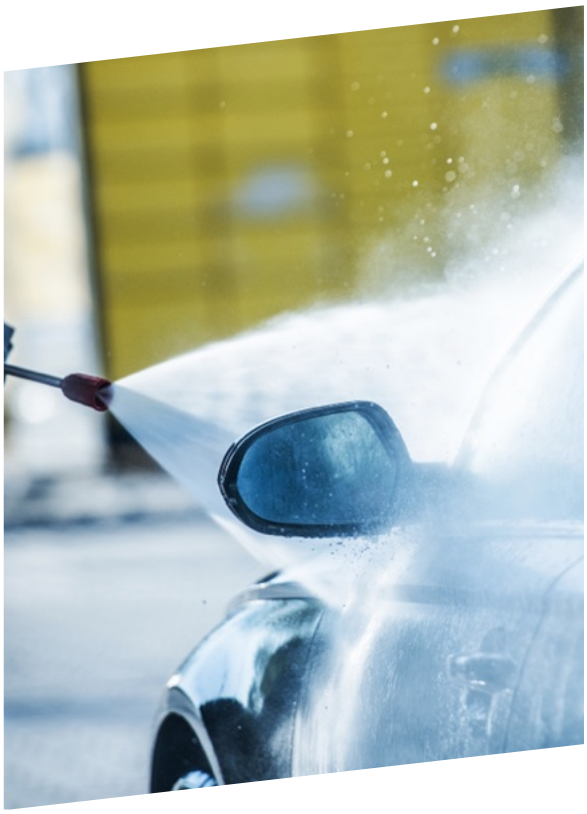
As luck would have it, John faced another problem when he had to refuel again in France just outside Lyon. The app directed him to a gas station that was supposedly nearby, but the **coordinates turned out to be completely wrong.** John Doe ran out of fuel and was stuck in the middle of nowhere with no choice but to cancel the important client appointment.

In the end, John lost an important customer and learned a valuable lesson: **data quality can make the difference between a company‘s success and failure**. Without high-quality data, decision-making can be flawed and misguided, leading to poor business performance, missed opportunities, and costly mistakes.



Approaches to enhance data quality

4



When it comes to enhance data quality, there is a wide variety of methods and techniques that can be applied. In the previous section, we saw the impact that poor data quality can have. In this section, we will explore several approaches to data quality enhancement, with a focus on problems that John faced before regarding gas station data.

It's important to note that the **choice of method and parameters** for data quality enhancement will vary depending on the specific use case and dataset. What works well for one type of data may not be the best approach for another. However, by understanding the various methods available and their strengths and weaknesses, we can make informed decisions about which techniques to apply to improve data quality in our particular use case.

Multi-Source Evaluation and Scoring

John Doe's first problem was that the app showed a "car wash" feature for a gas station, even though the station did not offer that service. This example shows very nicely: Sometimes no information is better than wrong information of low quality sources.

xavvy fuel addresses this problem on multiple levels. One of them is the consideration of multiple sources combined with scoring methods in data fusion to **determine the most reliable and accurate data points.** Therefore, data quality does not depend on a single source. This involves assigning a score to each piece of data based on factors such as source credibility, data completeness, and data consistency. Scores are used to weight the importance of each data point when combining multiple sources into a single dataset. Scoring methods help to improve the accuracy and reliability of the final dataset by giving greater weight to data from more credible sources. In our case, an overall score is then determined for each equipment feature.

Using the example of the gas station that John Doe stopped at to find a car wash, about 2 providers with comparatively low quality indicate that there is a car wash, while 1 provider has no information about it and 3 providers of high quality indicate that there is no car wash. Based on the score xavvy fuel determines that there is no car wash service available.

Approaches to enhance data quality

5

Plausibility Checks with Statistical Methods

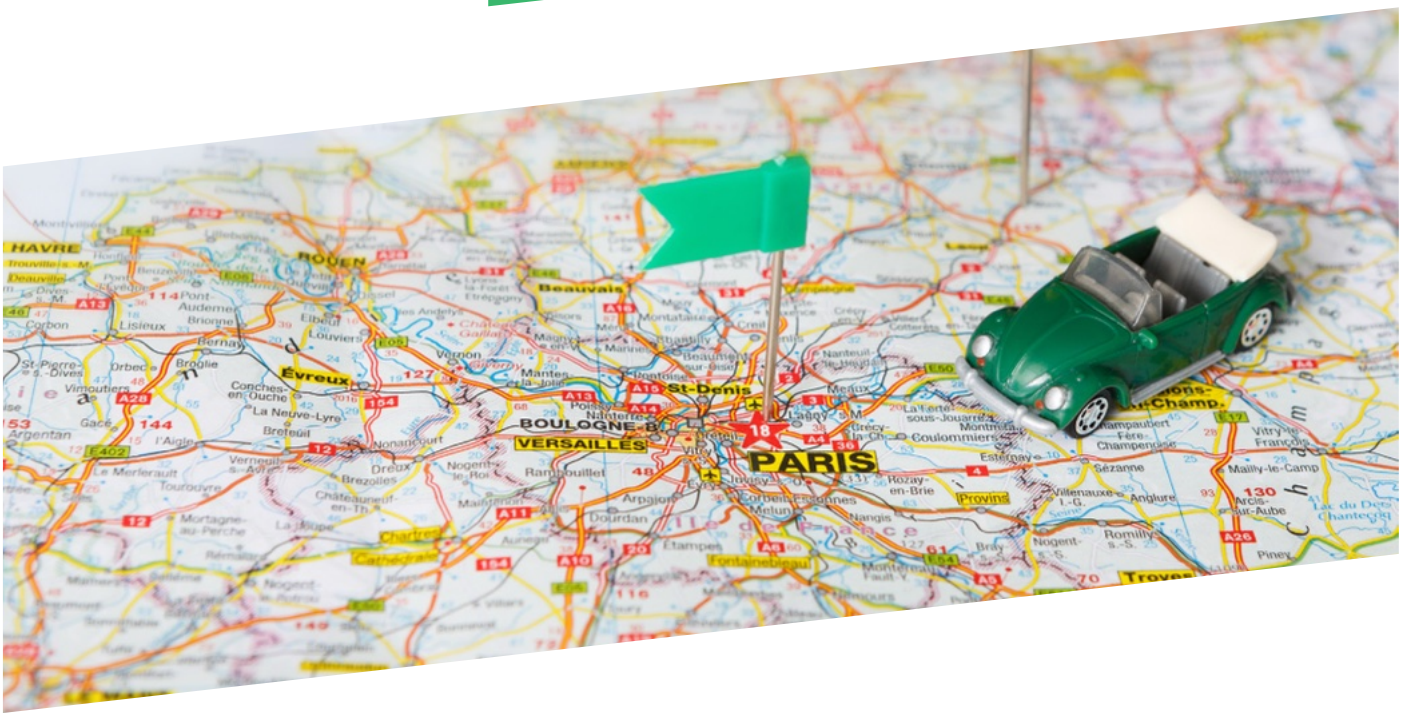
Plausibility checks are a crucial aspect in ensuring data quality. Typically, this means examining data for possible errors or inconsistencies. In the case of John Doe, when he was shown a price that was far too low, plausibility checks could have detected the erroneous price beforehand and saved him trouble.

xavvy fuel relies on **AI-supported outlier detection** that looks at prices in the region and differentiates between price-trends and types of gas stations, such as highway gas stations. This results in regional groups of prices that follow a similar pattern.

Furthermore, statistical methods such as the David-Hartley-Pearson (DHP) test are applied to detect potential outliers in price data in a region. DHP compares the difference between individual data points and the overall mean value, assessing whether the difference is significant enough to be considered an outlier. The DHP test is particularly useful in detecting data entry errors or data that

Approaches to enhance data quality

6



Clustering and Validation

French data sources, for example, tend to „hide“ gas stations for which no geocoordinate has been recorded in the center of Paris. For this reason, the correctness of geocoordinates is validated with forward geocoding. In addition, data sets from different data sources for the supposedly same gas station are clustered with the help of K-means.

K-means clustering is a machine learning technique used to group similar data points together. It works by identifying a set number of clusters in a dataset, then iteratively grouping data points based on their distance from the cluster center. The algorithm continues to refine the clustering until a stable solution is reached. K-means clustering is commonly used in data processing to identify patterns and group data points into meaningful categories, such as in the case of identifying and ignoring incorrect geocoordinate data for gas stations.

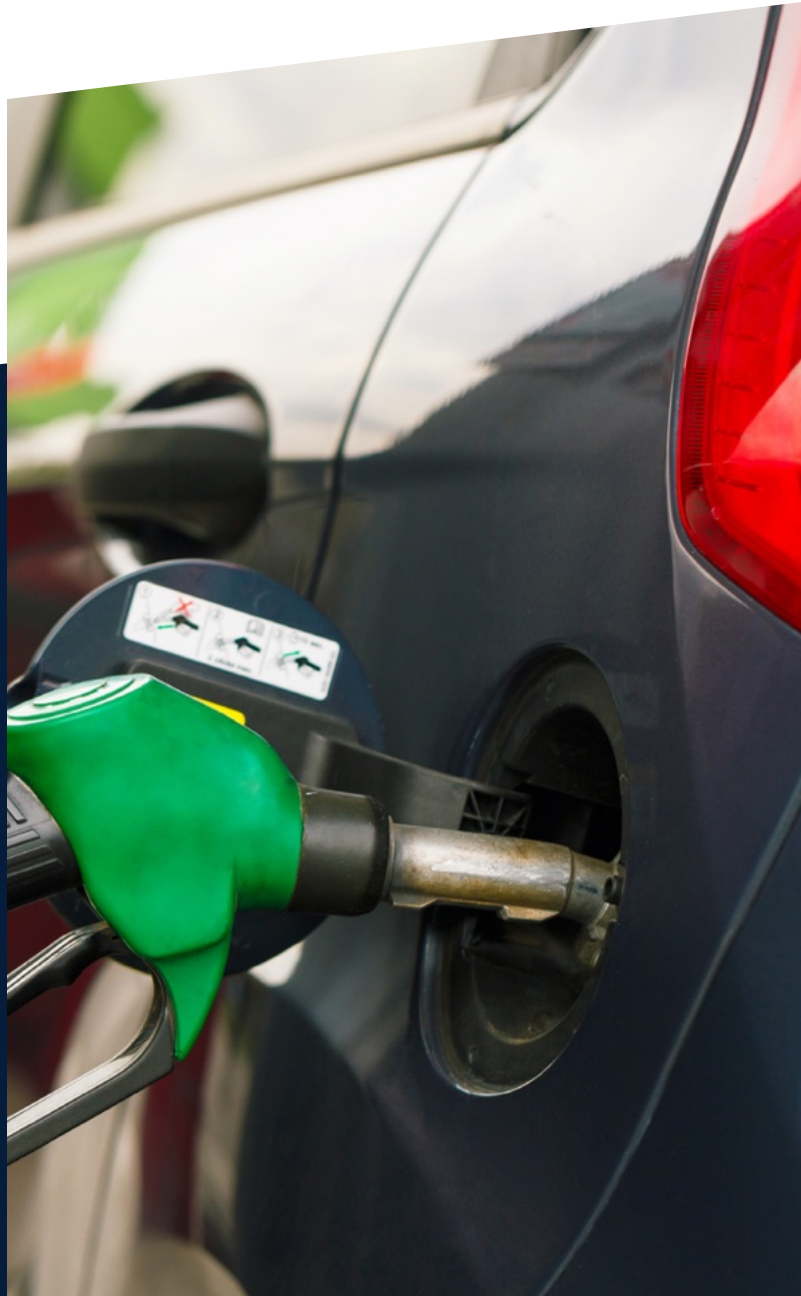
In comparison the calculation of a simple mean value is too inaccurate in the given example, since this approach considers „good“ and „bad“ geocoordinates equally, so that the mean value can still be grossly wrong. While with the help of K-means clustering **also a group of wrong data can be recognized** and ignored.

The Benefits of High-Quality Data for Informed Decisions

7

In this article, a simple example was used to show the effects that poor data quality can have and the means by which the problems could have been avoided. If John Doe had used an app with high data quality on his journey, his story would certainly have turned out differently. In a clean car and with favorable fuel stops John would have appeared in a good mood to his customer appointment, which he could bring to a successful conclusion.

Enhance data quality and leave data issues behind with xavvy. Contact us today.



xavvy fuel

incs Intelligent Coporate Solutions GmbH
Edmund-Rumpler Straße 6a, 51149 Köln

Markus Ruland
mr@incs.org

Tel. +49 2203 20 21 6 - 15
xavvy.solutions/fuel