Stage_2_Prepare

Jakub Jędrych

2024-04-11

In this stage we need to answer to some questions about our data. Its important to know what data is that ? Does this data ROCCC? From what source is that data and who are responsible for optimizing and manage this data (data engineer)?

This is really simple but important stage because we need to ask more and more questions if we do not so our data can include a lot of bias and i can be really make hard to create professional process of analysis a data

- **Where is the data stored?**
  FitBit Fitness Tracker Data (CC0: Public Domain) is stored in Kaggle-Linked
- **How is the data organized? Is it in long or wide format?**
  The data is organized in structured format tables with rows. It follows a long format due to repeated values in columns like the ID column, where data is recorded at regular intervals (e.g., hourly).
- **Are there issues with bias or credibility in this data? Does this data ROCCC?**
  Data was created by Mobius.
  Metadata about our data says:
  The data was collected by Mobius and made available on Kaggle. Metadata about our data indicates that the datasets were generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Variation between outputs represents the use of different types of Fitbit trackers and individual tracking behaviors/preferences.
- **How are you addressing licensing, privacy, security, and accessibility?** The data is provided under the CC0: Public Domain license, ensuring it can be used freely for any purpose. However, it's crucial to consider privacy and security concerns, especially when dealing with personal health data. Proper anonymization techniques should be applied to protect users' identities and sensitive information. Accessibility can be ensured by providing clear documentation and data formats that are easily accessible by various analytical tools.
- **How did you verify the data's integrity?** Data integrity can be verified through various means, such as cross-referencing with other reliable sources, checking for inconsistencies or anomalies, and validating against known patterns or benchmarks. Additionally, conducting exploratory data analysis (EDA) can help identify any irregularities or data quality issues.
- **How does it help you answer your question?** The FitBit Fitness Tracker Data provides minute-level output for physical activity, heart rate, and sleep monitoring, allowing for in-depth analysis of users' daily habits. By analyzing this data, we can gain insights into patterns, trends, and correlations related to users' behaviors and health metrics.
- **Are there any problems with the data?** Potential issues with the data may include missing values, outliers, inaccuracies, or biases. It's essential to address these issues through data preprocessing techniques such as data cleaning, normalization, and imputation to ensure the reliability and accuracy of the analysis results.