Avinash Bunga

Information Systems and Business Analytics, Park University

CIS611HOS1P2024 - Introduction to Business Analytics

Professor: Timur Rakhimov

Feb 1, 2024

Unit 4: Discussion

Exploratory Data Analysis (EDA) and Data Quality Assurance: Understanding Their Roles

When working with data analytics, two critical steps help us make sense of our data: Exploratory Data Analysis (EDA) and Data Quality Assurance. These steps are different but very important for analyzing data correctly and making good decisions.

Exploratory Data Analysis (EDA) is like being a detective with your data. It is when you look at your data differently using charts, graphs, and numbers to find patterns, unusual things, or trends. You do not start with a specific guess or question; instead, you explore the data to see what stories it might tell you. This can help you develop ideas or questions you had not thought of before (IBM, n.d.; JMP, n.d.).

Data Quality Assurance is ensuring your data is in good shape before you analyze it deeply. This means checking for mistakes, making sure everything is where it should be, and fixing any problems. This ensures that when you start looking for insights, you are working with data you can work with (Knight, 2023).

How Are EDA and Data Quality Assurance Different?

EDA is the process of exploring data to find insights. It is creative and open-ended. You might make graphs, calculate some basic statistics, or look for patterns in the data. It is about understanding the data and finding interesting things that could lead to more questions or insights.

Data Quality Assurance ensures the data is correct and ready for analysis. This involves fixing errors, ensuring no duplicates, and dealing with missing information. It is crucial to ensure that any findings or insights are based on reliable data (IBM, n.d.; JMP, n.d.;Tate, 2023).

Where Do EDA and Data Quality Assurance Overlap?

Even though EDA and Data Quality Assurance have different purposes, they often come together during the data analysis. For example:

- While doing EDA, you might find some errors or inconsistencies in the data that need to be fixed. This is part of Data Quality Assurance.
- On the other hand, while cleaning your data (Data Quality Assurance), you might notice some interesting patterns or trends you want to explore further through EDA (EPA, n.d.; IBM, n.d.; JMP, n.d.;Tate, 2023).

Difference in Activities and Goals

Activities

- **For EDA:** This includes making charts and graphs, finding trends, and looking at data in different ways to uncover new insights.

- **For Data Quality Assurance:** This involves correcting mistakes in the data, removing any duplicate information, and ensuring all the data you will use is consistent and reliable.

Goals/Aims

- **The goal of EDA** is to dive deep into the data and develop new ideas, questions, or insights that could be useful for further analysis or decision-making.

- **The goal of Data Quality Assurance** is to prepare a solid foundation of data that's accurate and reliable. This ensures that any analysis done afterward is based on good-quality data, making the insights or conclusions you draw much more trustworthy (EPA, n.d.; IBM, n.d.; JMP, n.d.;Tate, 2023).

Detailed Examples Using the Uber Rides Dataset

Dataset Overview for Clarity:
Our dataset comprises records of Uber rides, including columns for Ride ID, Date, Start Time, End Time, Distance (km), and Fare (USD). We focus on a subset of this data to address data quality issues and then proceed with exploratory analysis.

Data Quality Assurance Detailed Example:

- **Identifying and Correcting Zero Distance Entries:** Initially, we discover entries where rides have a distance of '0' km, which likely indicates canceled rides or input errors. Additionally, we ensure the fare column is consistent with these findings.

Action Taken:

- **Zero Distance Entries Correction:**

Ride_ID	Date	Distance	Original_Fare_USD	Action Taken	Updated_Fare_USD
R123	2024-01-05	0	5	Marked as canceled	N/A
R124	2024-01-06	12	20	No action needed	20

Resolution: Entries with '0' km were assessed and marked as canceled, removing them from the dataset to maintain its integrity for analysis. Fares associated with these rides were also examined and adjusted accordingly.

Exploratory Data Analysis (EDA) Detailed Example:

With a cleaned dataset, we explore two key areas:

- **Ride Demand by Day of the Week:** This analysis aims to uncover patterns in ride frequency across different days, indicating potential demand fluctuations.
- **Distance and Fare Relationship:** We investigate how distance correlates with fare, expecting longer rides to have higher fares.

Findings and Insights:

- **Ride Demand by Day of the Week:**

Day of the Week	Number of Rides	Average Fare (USD)
Monday	200	18
...	...	...
Friday	350	22

Insight: Fridays experience the highest demand, with an increase in both ride numbers and average fares, likely due to higher demand during weekends.

- **Distance and Fare Relationship:** A scatter plot analysis of distance versus fare reveals a positive correlation, with longer distances typically incurring higher fares. This relationship is critical to understanding pricing strategies and customer preferences.

- **Visualization Insight:** The scatter plot illustrates that fares increase with distance, highlighting a predictable fare structure that could inform pricing adjustments or promotional offerings.

Conclusion:

Through rigorous Data Quality Assurance, we ensured our dataset was accurate and reliable, setting a solid foundation for meaningful EDA. The insights from analyzing demand patterns and the distance-fare relationship provide actionable intelligence for operational and strategic decision-making. This process exemplifies the critical interplay between data quality and exploratory analysis, demonstrating how each step enhances the value and utility of data in real-world applications (EPA, n.d.; IBM, n.d.; JMP, n.d.;Tate, 2023).

References:

EPA (n.d.). *Exploratory Data Analysis*. https://www.epa.gov/caddis-vol4/exploratory-data-analysis#:~:text=Exploratory%20Data%20Analysis%20(EDA)%20is,step%20in%20any%20data%20analysis.

IBM (n.d.). *What is exploratory data analysis?* https://www.ibm.com/topics/exploratory-data-analysis

JMP (n.d.). *Exploratory Data Analysis*. https://www.jmp.com/en_us/statistics-knowledge-portal/exploratory-data-analysis.html

Knight, M. (2023, August 1). *What Is Data Quality? Dimensions, Benefits, Uses*. Dataversity. https://www.dataversity.net/what-is-data-quality/

Tate, A. (2023, October 26). *The Importance of Data Cleaning in EDA*. HEX. https://hex.tech/blog/data-cleaning-exploratory-data-analysis/#:~:text=The%20efficacy%20of%20exploratory%20data,data%20to%20improve%20its%20quality.