JAN 23, 2024

Data Analysis Procedures

Forked from Data Analysis Procedures

Deziray.Howard[1]

[1]PDI CTR - Data Analyst

Deziray.Howard

ABSTRACT

Draft of the procedures and workflow for data analysis within PDI

Protocol status: In development
We are still developing and optimizing this protocol

Created: Jan 19, 2024

Last Modified: Jan 23, 2024

PROTOCOL integer ID: 93841

Introduction

1

Welcome to the PDI Data Analysis Playbook

This document serves as a comprehensive guide to the processes and methodologies followed by our team in data analysis to derive meaningful insights and support informed decision-making. We aim to ensure consistency, reliability, and efficiency in our analytical endeavors by adhering to these standardized procedures.

Purpose

This manual is designed to provide a clear and structured framework for conducting data analysis within our team. It outlines the key steps, best practices, and responsibilities associated with data collection, cleaning, exploratory analysis, statistical modeling, reporting, and more. Whether you are a seasoned analyst or are new to our team, this manual is a valuable resource to help you navigate our data analysis processes.

Governance and Compliance

2

Compliance with FedRAMP

We ensure that all data analysis and transferal programs adhere to FedRAMP (Federal Risk and Authorization Management Program) standards. All selected tools and platforms that we use are verified to have obtained FedRAMP approval and/or are whitelisted.

Endpoint Security

We diligently maintain the currency of our antivirus software and security patches. Utilization of government-issued equipment is ensured when deemed necessary. In instances of teleworking, strict adherence to operating solely on secured and private networks is maintained, with a strict prohibition on conducting work on public Wi-Fi. Reference is made to the USDA Information Security Awareness, in addition to scrupulous adherence to USDA policies and system-specific rules.

Data Collection

3 TBD - Should this be a separate protocol?

Data Cleaning and Preprocessing

4 Data cleaning stands as a pivotal phase in the data preparation process, ensuring the accuracy, consistency, and preparedness of datasets for analysis. The recommended procedures outlined below serve as integral sub-steps to proficiently steer the data cleaning process. During the execution of these steps, it is imperative to annotate corresponding actions to document the applied mechanisms and techniques, encompassing formulas, programs, troubleshooting, and other pertinent aspects.

4.1 Understand Your Data:
- Examine the dataset to understand its structure, variables, and overall content.
- Identify missing values, outliers, and any anomalies in the data.

4.2 Handle Missing Values:

Identify and handle missing values appropriately. Some options include:
- Removing rows with missing values.
- Imputing missing values using mean, median, or mode.
- Using imputation techniques like regression or machine learning.

4.3 Deal with Duplicates:
- Check for and remove duplicate rows in the dataset.
- Ensure that your dataset contains unique observations.

4.4 Correct Inconsistent Data:
- Standardize categorical variables to ensure consistent values.
- Correct any inconsistencies or errors in naming conventions.
- Ensure that data types are appropriate for each column.
- Convert data types if needed (e.g., converting strings to numeric values).
- Identify and handle outliers that might skew your analysis.

4.5 Quality Assurance:
- Perform a final check to ensure that the data is clean and ready for analysis.
- Use descriptive statistics and visualizations to verify data integrity.

Exploration and Analysis

5 Data Exploration and Descriptive Analysis
To insert:
- Exploratory Data Analysis (EDA) Guidelines
- Summary Statistics and Visualizations
- Documentation of Key Findings

Data Modeling

6 TBD