High Level Design (HLD)

Neuro Data Engineering

Exploratory Data Analysis

**Document Version Control**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date Issued** | **Version** | **Description** | **Author** |
| 11/08/2021 | 1.0 | Initial EDA HLD – V1.0 | Supreeth Shetty |

Abstract

Chapter 15

Exploratory Data Analysis

Matthieu Komorowski, Dominic C. Marshall, Justin D. Salciccioli

and Yves Crutain

Learning Objectives

•Why is EDA important during the initial exploration of a dataset?

•What are the most essential tools of graphical and non-graphical EDA?

15.1 Introduction

Exploratory data analysis (EDA) is an essential step in any research analysis. The

primary aim with exploratory analysis is to examine the data for distribution,

outliers and anomalies to direct speciﬁc testing of your hypothesis. It also provides

tools for hypothesis generation by visualizing and understanding the data usually

through graphical representation

Chapter 15

Exploratory Data Analysis

Matthieu Komorowski, Dominic C. Marshall, Justin D. Salciccioli

and Yves Crutain

Learning Objectives

•Why is EDA important during the initial exploration of a dataset?

•What are the most essential tools of graphical and non-graphical EDA?

15.1 Introduction

Exploratory data analysis (EDA) is an essential step in any research analysis. The

primary aim with exploratory analysis is to examine the data for distribution,

outliers and anomalies to direct speciﬁc testing of your hypothesis. It also provides

tools for hypothesis generation by visualizing and understanding the data usually

through graphical representation

Chapter 15

Exploratory Data Analysis

Matthieu Komorowski, Dominic C. Marshall, Justin D. Salciccioli

and Yves Crutain

Learning Objectives

•Why is EDA important during the initial exploration of a dataset?

•What are the most essential tools of graphical and non-graphical EDA?

15.1 Introduction

Exploratory data analysis (EDA) is an essential step in any research analysis. The

primary aim with exploratory analysis is to examine the data for distribution,

outliers and anomalies to direct speciﬁc testing of your hypothesis. It also provides

tools for hypothesis generation by visualizing and understanding the data usually

through graphical representation

Exploratory data analysis (EDA) is an essential step in any research analysis. The

primary aim with exploratory analysis is to examine the data for distribution,

outliers and anomalies to direct

.

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

Table of Contents

[**Document Version Control………………………………………………………………………………………………………**..2](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.gjdgxs)

[**Abstract**………………………………………………………………………………………………………………………………………..3](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.30j0zll)

[**1** **Introduction**…………………………………………………………………………………………………….……………….5](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.1fob9te)

[1.1 Why this High-Level Design Documents…………………………………………………………………………….5](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.3znysh7)

[1.2 Scope…………………………………………………………………………………………………………………………………5](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.2et92p0)

[1.3 Definitions………………..……………………………………………………………………………………………………….5](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.tyjcwt)

[**2** **General Description**………………………………………………………………………………………………………….6](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.3dy6vkm)

[2.1 Product Perspective…………………………………………………………………………………………………………..6](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.1t3h5sf)

[2.2 ProblemStatement……………………………………………………………………………………………………………6](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.4d34og8)

[2.3 ProposedSolution……………………………………………………………………………………………………………..6](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.2s8eyo1)

[2.4 Data Requirement..……………………………………………………………………………………………………………6](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.lnxbz9)

[2.5 Tools Used…………………………………………………………………………………………………………………………6](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.lnxbz9)

[2.8 Process flow………………………………………………………………………………………………………………………7](https://docs.google.com/document/d/1vQWYiPew8rO24KScWpBZyVKQJ2uEtBly/edit#heading=h.35nkun2)

# Introduction

## **Why this High-Level Design Documents?**

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

This HLD will:

* Present all of the design aspects and define them in detail
* Describe the user interface being implemented
* Describe the hardware and software interfaces
* Describe the performance requirements
* Include design features and the architecture of the project
* List and describe the non-functional attributes like:
  + Security
  + Reliability
  + Maintainability
  + Portability
  + Reusability
  + Application compatibility
  + Resource utilization
  + Serviceability

## **Scope**

The HLD documentation presents the structure of the system, such as the database architecture, application architecture(layers), application flow (Navigation), and technology architecture. The HDL uses non-technical to mildly-technical terms which should be understandable to administrators of the system.

**Definitions**

|  |  |
| --- | --- |
| Term | Descriptions |
| EDA | Exploratory Data Analysis |
| DB | Database |
| IDE | Integrated Development Environment |

# 2. General Description

## **Product Perspective**

A web based application to perform exploratory data analysis (EDA), which help users to analyze their dataset and get meaningful insights from the datasets

#### **Problem statement**

To create a web based application to perform extensive EDA(exploratory data analysis) on a dataset provided from the user, the platform should also be flexible to drag and drop the dataset

Application should be help us perform following functionalities :

* Exhibit head, tail, shape of dataset
* Univariate visualization of each field
* Multivariate visualization
* Describing the numerical features
* Describing the categorical features
* Identify missing values in every feature
* Identify the outliers in every feature
* Identify duplicate data

#### **Proposed Solution**

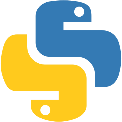
The solution proposed here is “**Neuro Data Engineering”** tool, which helps the user to perform EDA(exploratory data analysis) on the dataset uploaded by the user.The tool can implement each and every EDA technique which is generally used.

#### **Data Requirement**

Data Is provided by the user and it can be in any format

#### **Tools Used**

Python programming language and frameworks such as NumPy, Pandas, Flask



#### **Process Flow**

For identifying the different types of anomalies, we will use a deep learning base model. Below is the process flow diagram is as shown below.

Perform EDA

Output

Read dataset

Save dataset to db

Load dataset

#### **Technical Requirements**

This document addresses the requirements for creating web app application to perform Data cleaning, Feature engineering, and EDA..

We will be using **Python 3.8.** **1**

The web application will be built using Django web framework

Templates will be built using HTML, CSS, Javascript

Cassandra online Database will used as database

**\*User must login to the account.**

We can use Django inbuilt authentication system, which reduces our work on validating valid emails, and password hashing etc.

(refrence : <https://docs.djangoproject.com/en/3.2/topics/auth/default/>

**Users must be able to drag and drop the existing component at UI to perform any operation.**

Not sure how to implement this please help

Logging should be implemented using logging modules

Exception handling should be implemented

\*User should be able to load a dataset from source(File, Database, Cloud Storage).

System should provide flexiliblity to user in terms of loading dataset, i.e user should be able load the dataset from their local system or, from any database, or from any cloud storage

Not sure how to do this

The system should be provide the three main features i.e Exploratory Data analysis, Data Preprocessing, Feature Engineering which can be performed on loaded dataset

So we will be using libraries like numpy, pandas, scikit learn, plotly

Server requirements

Server (AWS, AZURE and Heroku)