



**ÇANKAYA UNIVERSITY FACULTY OF
ENGINEERING COMPUTER ENGINEERING
DEPARTMENT**

Project Report
Version 2

CENG 408

Innovative System Design and Development II

***<Web-based Multifunctional Graphical User Interface for
Data Analysis>***

Alim GEYİK
201411023
Oğulcan BAŞARAN
201411205
Büşra KUTLUER
201511040
Nurseli BAL
201611656

Advisor: Assist. Prof. Dr. Gül TOKDEMİR

Table of Contents

1.	Introduction	6
1.1.	Motivation	6
1.2.	Problem Statement.....	6
1.3.	Related Work.....	6
1.4.	Solution Statement.....	7
2.	Literature Review	7
2.1.	Introduction	8
2.2.	Data Analysis.....	8
2.3.	Data Visualization	10
2.4.	Data Analysis and Visualization Tools	11
2.5.	Conclusion	12
3.	Software Requirements Specification	12
3.1.	Introduction	12
3.1.1.	Purpose	12
3.1.2.	What are we building?.....	13
3.1.3.	Scope of Project	13
3.1.4.	Overview	14
3.1.5.	Development Responsibility	14
3.1.6.	Definitions, Acronyms, and Abbreviations.....	15
3.2.	Overall Description.....	15
3.2.1.	Product Perspective	15
3.2.2.	User Interface	16
3.2.3.	Software Interface	17
3.2.4.	Hardware Interface.....	17
3.3.	Requirements Specification.....	17
3.3.1.	Use-Case.....	17

3.3.2.	Brief Description	18
4.	Software Design Document	18
4.1.	Introduction	18
4.1.1.	Purpose	18
4.2.	Scope of Project.....	19
4.3.	Architecture Design	20
4.3.1.	Data Preprocessing	22
4.3.2.	Data Cleaning	22
4.3.3.	Data Transformation	22
4.3.4.	Data Reduction.....	23
4.3.5.	Classification.....	23
4.3.6.	Clustering	24
4.3.7.	Regression	24
4.3.8.	Analysis.....	25
4.3.9.	Visualization.....	26
4.4.	Interactivity.....	27
4.5.	System UI Design.....	27
4.5.1.	Index Page	33
4.5.2.	Upload Page	33
4.5.3.	Analysis Page	34
5.	Conclusions	35
	Referances	35

Abstract

The data that people create in the virtual world (social media posts, news, shopping bills, emails, etc.) is increasing and accumulating day by day. These data may be unnecessary and meaningless to most of us, but they are vital to an analyst, a company or a business. These data are used in the promotion of a new product, in the design of a store, in an application update or in determining the products to be discounted. Every footprint in the virtual world is truly a treasure for someone interested in the point we step on. Of course, not every data is meaningful and usable. At this point data analysis programs and platforms come into play. It is precisely these applications and platforms that will reveal the meaning of the available data and determine the value of this treasure. WADAV is a web platform that allows users to review their data and access the information they need. Our goal in developing WADAV was to create a simple and understandable interface that would enable all types of users to evaluate the data they had. In this report, we explained why we plan to create WADAV software requirements and how we plan to design our software.

Key words:

WADAV, data, data analysis,

1. Introduction

1.1. Motivation

We are a group of senior students from the computer engineering department interested in data science. Our group members have taken the data mining course being taught at our school and are working in this project because of their interest in this field. Unlike many data analysis applications currently available on the market, we aim to interactive interface to examine data in more detail. For this purpose, we want to enable users to reach a faster experience from anywhere with WADAV platform. In order to further develop ourselves in the field of data analysis and data visualization, we enrolled in courses via the internet. We plan to complete this project by developing ourselves with the help of these courses.

1.2. Problem Statement

Our main problem in this project is that the data has many missing elements or unnecessary information. In order to achieve more consistent and understandable results, we have to clean the data for analysis by passing through the correct preprocessing stages.

1.3. Related Work

We wanted to design and build an accessible platform that would not only allow everybody to uncover the hidden predictive power of data with ease, but also would make the whole experience “enjoyable”. [1] (BigML)

Exploratory's Simple UI experience makes it possible for anyone to use Data Science to Explore data quickly, Discover deeper insights, and Communicate effectively.[2]

1.4. Solution Statement

As a result, there will be a data analysis and data visualization system for WADAV users, which has an interactive interface, works easily and quickly, and can perform the preprocessing stages of data.

2. Literature Review

In this world , as a result of scientific and technological developments, we are in increasing masses of unprocessed information and data traffic. Extracting meaningful information from this data stack is difficult if the data is presented in plain text or traditional tabular format. Therefore, effective graphical representations of data have been developed to activate human visual sensing capabilities. Data visualization is the use of computer-based interactive visual representations of abstract and non-physical data to reinforce human knowledge , defined as visual elements used to reveal hidden patterns within data sets.

Big data is a revolutionary issue, dealing with all aspects of our lives, from businesses to consumers and science to government. Data Analysis and data visualization, which has become a globally focused force over the past decade, has been proposed as a dominant source of innovation, competition and productivity. Health, finance, science and other fields a lot of objects to be connected to the internet , any movement of data into and clicking on very large and diverse new technologies and analyzing this data led to the design of architectures in the name of liability. Worldwide organizations such as DARPA (Advanced Defense Research Projects Agency), National Security Agency(NSA), National Oceanic and Atmospheric Organization (NOAA), U.S. National Aeronautics and Space Agency (NASA) use advanced data analysis systems.

Given the importance of knowledge today, the development of effective solutions by increasing perception with the right methods to be developed for analysis is a sign of a major transformation in many areas. For this reason , we aimed to be able

to contribute with a design in hardware that can meet the requirements of the age. We have covered functional data analysis and visualization approaches as opposed to methods such as linear modeling. Data visualization can be in the form of charts, lines, maps, tables, or various elements, but interactive visualization enables direct operation on a visual by linking multiple data. It covers topics such as extracting valuable information according to the needs of users on different data sets, associating data with each other, grouping on charts, filtering and action realization, and preparing an interactive resume for the user.

2.1. Introduction

Nowadays the biggest treasure is considered information. Companies, businesses, and all people want to have knowledge. About their own businesses, about their customers, about their friends, about everyone else. Social media is the primary proof of this situation. At this point, if we need to address this issue, it is necessary to organize all this information chaos. Data analysis and visualization methods facilitate our work in this field. They enable us to have more readable, predictable, computable and processable data.

In this project, we are trying to analyze and visualize our data in line with our needs with an interactive interface. We will design a website that contains many different analysis methods, tables, and graphs. Our main goal is to provide the user with a user-friendly interface in the fastest way possible. In this process, we will examine many different analysis tools and applications and try to reach the most stable and reliable results. We will try to present the diversity of analysis and visualization at the highest level possible.

2.2. Data Analysis

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusion and supporting decision-making. [4] Data analysis consists of several steps. If we follow these steps in order and correctly, we will have an analysis process in which we can get the most accurate results. These steps are determining the requirements, data collection,

data cleaning, data analysis, data interpretation. And after completing these steps correctly, we will continue with data visualization.

First, we need to determine why we need data analysis. For example, do we own a hotel business where our customers will have a better experience? Do we want to turn our hotel into a more sophisticated complex? If we ask feedback from our customers after determining our aim, we may have collected the necessary data for the roadmap to follow. For our example, we can ask our customers ideas about room comfort, cleanliness, hotel facilities, etc.

Immediately after this stage, we have to clean the data set. Our data set may have incomplete data, or if we don't have the data set ourselves, it may have data that we don't need. The cleaning phase is one of the most important steps for data analysis because incomplete or unnecessary information can remove us from our purpose. After collecting, clearing and processing the data, we are ready for analysis.

After all these steps, we may have the information we need or we may need to collect more data. At this stage, we can use data analysis tools and software to help us understand, interpret, and achieve results based on requirements.[5] And it's time to interpret our results. We can choose the way to express or transmit our data analysis result and come to the end of a successful analysis process.

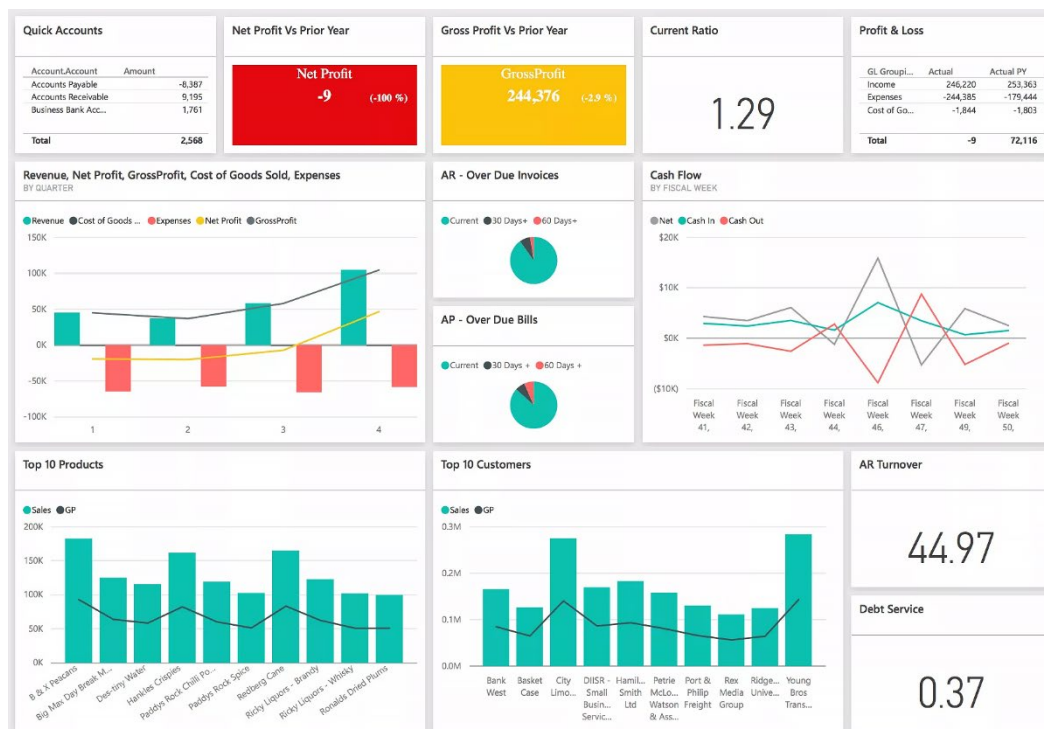


2.3. Data Visualization

Data visualization transfers data and information to the user using visualization elements. Tables, charts, graphs, maps etc. items are tools used to visualize data. These tools make the data more understandable and easy to read. Data visualization is widely used today's world. They make it easier for us to understand the data we have in surveys, stock market analysis, social media charts and more. They help our brain process data more easily. We can use data visualization to put our data, which we have completed the latest data analysis process, into a form that is understandable for everyone. In this way, our results can be better examined and evaluated and the most accurate roadmap can be determined.

If we talk about a few simple data visualization methods, these are histogram graphs, scatter plots, pie charts (pizza charts) and line graphs. Scatter plots are generally used in graphs with an axis dependent on time. Another ubiquitous chart is pie chart. It is a visualization element for comparing the percentages of a kind of data. However, there are several disadvantages. We need many labels to identify and understand chart

elements. And another example, histogram charts, gives a lot of information like pie charts but visualize them in a simpler and more understandable way.



2.4. Data Analysis and Visualization Tools

Today, data analysis and visualization tools are really popular. Many companies and businesses (Google, Facebook, Instagram, Apple, etc.) use these tools. In our literature study, we examined a few of these which are popular. We have examined web-based systems such as BigML, Kaggle and desktop programs such as Exploratory and Orange. These applications and tools help us to analyze and visualize the data we have. They include many visualization methods and analysis methods. They support the programming languages (Python, R, JavaScript, etc.) that we need during analysis and visualization, helping us to get more detailed results. BigML is a machine-based data analysis and visualization web-interface. You can upload your own data as well as open-source data sets. It has many methods of data visualization while facilitating data cleaning, editing, and analysis. Kaggle is a website that allows Google's data analysts and data scientists to review and write articles. It contains many data sets and analyzes. Supports many programming languages. It also creates discussion environments for data analysts, such as a blog page. Exp is a desktop application developed for data analysis

and visualization. They cooperate with many large companies to provide support. They have R language support and SQL support for dataset editing.

2.5. Conclusion

The general concept of data analysis and visualization is to bring a library layout to this dispersed pool of information, to reveal key points and necessary information. In this growing data complex, companies and individual employees find it hard to find their way. Our Project serves them as a GPS in this data chaos. It's a system that shows users how to access the actual information they need and search through all that chaos.

3. Software Requirements Specification

3.1. Introduction

This document is the software requirement specification document for the Web-based Multifunctional Graphical User Interface for Data Analysis and Visualization-WMGUIDAV Project, briefly Web App Data Analysis and Visualization - WADAV. This system allows users to analyze their data without using any desktop application. The graphic representation of the data, which does not make sense when examined alone, is called data visualization. Using visualization to understand the data that we have acquired more quickly is one of the best methods. In data visualization, we can interpret the patterns of our data, the correlations between them, the outliers, etc. Data visualization, which is a rapidly growing data interpretation technique in the world, is used by many people.

3.1.1. Purpose

According to the World Economic Forum, the world generates 2.5 quintillion bytes of data every single day, and 90% of all data has been generated in the last two years. When we look at the number, it is big enough and none of the people's capabilities are enough to interpret and understand those amount of data. We can make it faster by using data visualization. We can understand that this product is very useful.

This document describes the general concept and explains the required features of the system. It describes the interfaces, user characteristics, functionalities, and constraints of the system to make the specifications confirmed and refer to the development phase of the system.

3.1.2. What are we building?

We are building an interactive web-based data analysis system. that is useful for visualizing the data sets which are numeric. The datasets that we will be dealing with are completely numeric. Therefore, we don't need to apply future engineering on the step of the data preprocessing. We will just focus on the general data preprocessing.

3.1.3. Scope of Project

In this project, we aimed to present statistical and variable information units in a different and creative way by isolating them from classical schematic forms. One of the most important goals of data visualization is to make complex and confusing data normally presented in the classic format easy to understand with easy-to-understand graphical interfaces. We needed to present the data in a comprehensible way and get the right design in line with the overall objectives of the project. The content of this design covers many fields such as tools, IDEs, protocols, back-end, front-end, etc. parts used in the project.

Firstly, we have decided to use Python as the language. Python is a powerful, dynamic programming language used in a wide range of fields. The Pandas library offers high-performance, easy-to-use data structures, and data analysis features. Thus, without the need for statistical programs such as R or Stata, we can do data analysis and modeling. Combined with tools and libraries such as IPython, statmodels, and sci-kit-learn, a powerful data analysis environment can be achieved in terms of performance and productivity.

Secondly, the protocol can be used as Web Distributed Authoring and Versioning (WebDAV). Is a hypertext transfer (HTTP) protocol that allows you to manage files and documents stored on web servers.

With WebDAV, you can connect to the cloud environment like Google Drive, OneDrive, OwnCloud, and do a lot of things like deleting, writing, copying, moving, and you can do it using your internet browser without installing any applications.

As a storage system, we have thought of a platform where large amounts of data can be securely held, quick access to information, the integrity of information, and access to multiple users simultaneously.

Instead of stacking all the data in one place, we thought of using MySQL, which stores it in different tables and databases. The control, optimization, and repair of the tables are done quickly.

Finally, we thought of using HTML5 and CSS3 to visualize the project (website). It has advantages such as working together using the same database common, making it easy to use for everyone.

3.1.4. Overview

WADAV is a web-based application that enables users of all levels to create an analysis project via interactive user interface. Users will upload their own data and decide how to analyze and visualize it on a chart or graph. As a user builds a project, he or she interacts with the all system elements via a live preview. Once finished, download the graphs and report of data analysis.

3.1.5. Development Responsibility

- Different data import option.
- Mapping capability.
- On visualization, complex topics are easy to understand.
- Compare data by time intervals.
- Ability to obtain snapshots of data.
- Highly customizable and extensible.
- Embedding graphics into websites and applications should be easy.
- In order to compare and contrast the data, it is the best to use stacked graph.
- By using scatter plot, we can represent the data trends.
- An effective way to show trends and comparisons is to use line charts.
- Especially, to show more points to color coded, scatter plots are useful to show a multitude of data.
- Relationships between group of entities is shown by chord diagrams.

- Using dashboards will help us numerous data visualizations edge to edge.
- Bubble charts simultaneously, can showcase the numerous data points.
- When multi-axis line charts are annotated, they're better.

3.1.6. Definitions, Acronyms, and Abbreviations

- Analysis – Detailed examination of the elements or structure of something.
- Data Set - A data source or data subset.
- Data Source - A collection of data, such as the contents of a CSV or JSON file.
- Event - A user interaction (click, hover, etc.) that triggers an action.
- EU - End-User
- Visualization - Any technique for creating images, diagrams to communicate a message.

3.2. Overall Description

The following section and subsections present an overall description of the WADAV. In particular, the product has been put into perspective through a detailed assessment of the system, user, software and hardware interfaces.

3.2.1. Product Perspective

WADAV is an interface that processes data to enable people, companies or businesses to develop or Express themselves better. It is an interface where they can simply load data, analyze and visualize the loaded data. It does not require membership, it provides fast and free use. Just load the data and select the process you want to do.

3.2.2. User Interface

This section describes the path a user will take after connecting to the website. After logging in to the web site, you reach the welcome page. And you can see the interface diagram of the web site below.

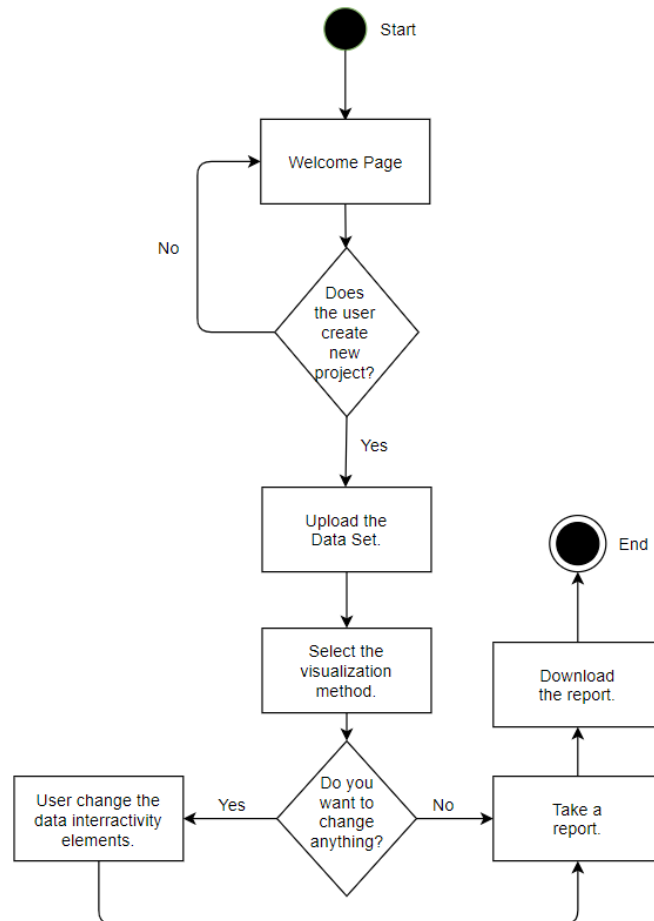


Fig1. User Interface Mapping

The welcome page will Show you instructions. According to these guidelines, you must first create a Project. After you name the Project, you have to upload the dataset that you want to examine. The system then prepares the data set for analysis and you will only have to select the visualization method you want to see. If you want to make changes in visualization methods, you can change the values you want with the interactive interface and see the results of the changes from the graphs and diagrams.

After the process is completed, you can download the report and graphics for your project.

3.2.3. Software Interface

The loaded dataset is prepared for visualization. At this stage, the system applies certain analysis methods to data. Classification, Regression, Analyze and Clustering methods are prepared for data visualization.

3.2.4. Hardware Interface

It can be accessed from all devices with CPU and operating system with Internet access.

3.3. Requirements Specification

3.3.1. Use-Case

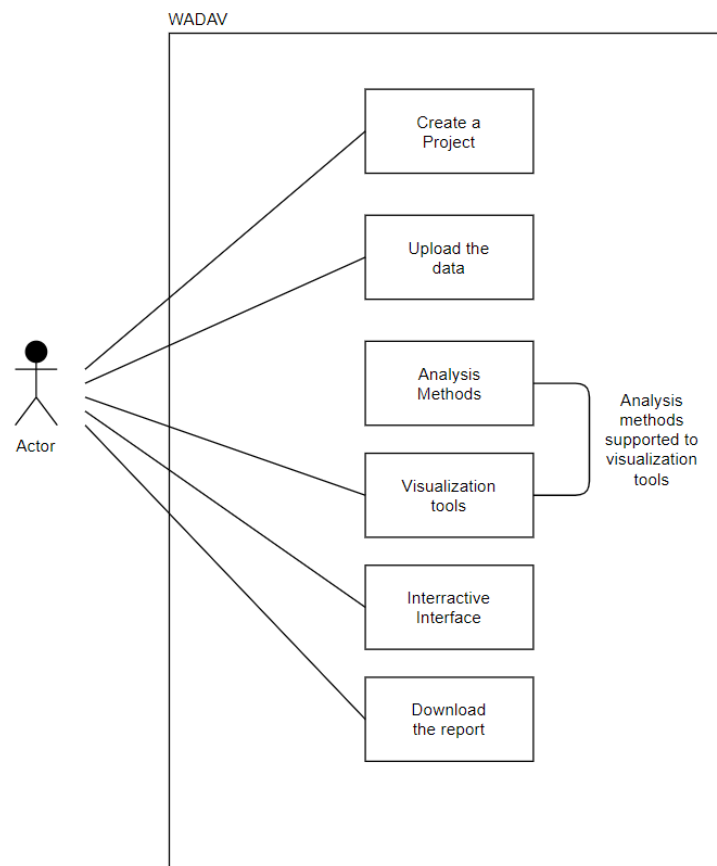


Fig2. Use-Case Diagram

- Create Project
- Upload the data
- Analysis method
- Visualization Tools
- Interactive Interface
- Download the Report

3.3.2. Brief Description

The figure shows the interaction between the actor and the system. The users can see the visualized form of data by uploading their data. Interactive Interface provides users to visualize the graphics associated with the chosen attribute. Users also can change the type of graphics. After all the operations are done, they can download the visualization and analysis report.

4. Software Design Document

4.1. Introduction

This document is the Software Design Document for the Web-based Multifunctional Graphical User Interface for Data Analysis and Visualization-WMGUIDAV Project, briefly Web App Data Analysis and Visualization - WADAV. This system allows users to analyze their data without using any desktop application. In this document, the project will be detailed.

4.1.1. Purpose

The purpose of this Software Design Document is to provide the details of the project titled Web-Based Multifunctional Graphical User Interface for Data Analysis.

The target audience is people who would like to visualize, interpret their data. This visual methods of data analysis generally are employed by engineers, analysts and scientists. In this project, we will create a Web-based application that is used for data visualization. Data visualization is among many business intelligence and is crucial for advance analytics. Data visualization is to make analysis easier and faster. It offers the ability to see important topics at a glance. Elements for visualization as charts, tools, graphs are used to provide an accessible way to see the data. That is to say, we can see the outliers, patterns and trends. However, before we start to visualize the given data, in the real world, data is not always good. It may be inconsistent, incomplete or contains many errors or outliers, we firstly preprocess the data. Therefore, preprocessing data is one of the important aims of our project.

4.2. Scope of Project

In this project, we aimed to present statistical and variable information units in a different and creative way by isolating them from classical schematic forms. One of the most important goals of data visualization is to make complex and confusing data normally presented in the classic format easy to understand with easy-to-understand graphical interfaces. We needed to present the data in a comprehensible way and get the right design in line with the overall objectives of the project. The content of this design covers many fields such as tools, IDEs, protocols, back-end, front-end etc. parts used in the project. Firstly, we have decided to use Python as the language. Python is a powerful, dynamic programming language used in a wide range of fields. The Pandas library offers high-performance, easy-to-use data structures and data analysis features. Thus, without the need for statistical programs such as R or stata, we can do data analysis and modeling. Combined with tools and libraries such as IPython, statsmodels, and scikit-learn, a powerful data analysis environment can be achieved in terms of performance and productivity. Secondly, The protocol can be used as Web Distributed Authoring and Versioning (WebDAV). Is a hyper text transfer (HTTP) protocol that allows you to manage files and documents stored on web servers. With WebDAV, you can connect to the cloud environment like Google Drive, OneDrive, OwnCloud, and do a lot of things like deleting, writing, copying, moving, and you can do it using your internet browser without installing any applications. As a storage system, we have

thought of a platform where large amounts of data can be securely held, quick access to information, integrity of information, and access to multiple users simultaneously. Instead of stacking all the data in one place, we thought of using MySQL, which stores it in different tables and databases. The control, optimization and repair of the tables are done quickly. Finally, we thought of using HTML5 and CSS3 to visualize the project (website). It has the advantages such as working together using the same database common, making it easy to use for everyone.

4.3. Architecture Design

In this process, during which the project report was prepared, we examined many similar applications and platforms to determine our development methods and the way we will take. We have examined the methods of data preprocessing, data cleaning data analysis, and data visualization of these applications and platforms. As a result of these investigations, we have determined the paths we want to follow.

Sütun1	Big ML	Exploratory.io	Orange
Classification	Decision Tree - Deepnets	Decision Tree - Random Forest	Decision Tree - Random Forest - kNN
Regression	Linear Regression - Logistic Regression - Deepnets	Linear Regression Logistic Regression Random Forest	Linear Regression - Logistic Regression - SVM
Analysis	Cluster Analysis - Time Series -	Correlation - Distance - Market	Distance - ROC Analysis - Time

	Anomaly Detection	Basket - Principal Component	Series - Correspondence Analysis
Clustering	K-Mean & G- Mean	K-Mean	K-Mean - Hierarchical Clustering - Louvain Clustering
Visualization	Scatter & Histogram Plot - Pie Chart - Cluster & Tree Visualization	Pivot - Table - Bar - Line - Area - Density - Scatter - Histogram - Word Cloud - Heatmap	Scatter - Histogram - Line - Box Plot - Mosaic Sieve - Venn Diagram - Pythagorean Tree
Interactivity	Create a tables by selected element for graphics.	Select the part of data, analysis & visualize the selected part.	Instruct visualization to send out a data subset that corresponds to the selected part of visualization
User-Interface	Web - Based	Desktop Application	Desktop Application
Data Type	CSV - TSV - ZIP	CSV - Excel - JSON - Log File - SAS/SPSS/STATA	
Extras		Wilconox Test - Kruskal-Wallis Test - Chi-Square Test - Normality Test	

Table of similar applications and methods

4.3.1. Data Preprocessing

The data preprocessing step is one of the most important steps. We basically prepare the dataset we will be using. We make it more useful, convenient. On this data preprocessing step, we will follow the steps as Data Cleaning, Data Transformation, Data Reduction. These are the general steps on Data Preprocessing.

4.3.2. Data Cleaning

We basically start the preprocessing first cleaning the data. While collecting data, data is not always presented to us as we want. During the collection phase, data may be missing, that is, missing values. If we do not have enough data, we can delete or ignore that tuple because we will not have a good analysis. If we have enough data but have a small number of missing values then we can fill the gaps. On this step, we have to decide how to fill the gaps. There are certain ways. We can fill the gaps with the mean of the attribute or with the most possible value in the attribute values.

Again, the data may not be in the range we want during the collection phase. At this point, the machine may not be able to detect this. There are also certain methods to handle these kinds of data.

4.3.2.1. Binding Method

On sorted data, we can use this method to be able to have smooth data. The data in the attribute is divided into equal parts and each part of the dataset is handled individually.

4.3.2.2. Regression

To make the data smooth, it is fitted to a regression function. The regression function may be linear or non-linear.

4.3.3. Data Transformation

This is another good way to prepare our data. In this step, data is transformed into appropriate forms. There are also several types of doing this step.

4.3.3.1. Normalization

It used to scale the data in a specific range. Attribute Selection: sometimes it is necessary to add a new attribute except for the attribute used in the dataset. On the other hand, we may have to delete an attribute/attribute.

4.3.3.2. Diczetion

I divide the range of a continuous attribute into intervals.

4.3.4. Data Reduction

Data may be repeated in a huge amount of data. This data reduction handling method, we deal with data reduction. Dimensionality reduction: It removes unnecessary and unimportant attributes. After all of the data preprocessing steps, we can play with our data with better results. We will apply certain algorithms to cluster or to classify our data. Then we can visualize the uploaded data.

4.3.5. Classification

4.3.5.1. Decision Tree

This supervised learning algorithm is one of the most used and one of the most important algorithm. On the decision tree, each node will represent the attributes of the dataset. Besides, each branch represents a rule and each of the leaves represent the our outcomes.

The user will have the right to visualize the built tree and the path of the data will be visualized on the tree. In this Project, the path of the data will be visualized on the tree.

Advantages:

- Model Interpretability is high which means it is easy to understand.
- Model maintainability is low which means modifying the model is easy.
- Classification for classifying an unseen record is low.

4.3.5.2. Random Forest

This algorithm is another supervised algorithm that will be used. The algorithm can do the task of classification for categorical values.

4.3.5.3. KNN

This is another classification algorithm that we are going to use.

The aim of this study is to examine the closeness of the new individual to the k from the previous individuals. The k -nearest neighborhood (KNN) algorithm is easy-to-implement supervised learning algorithms. Although it is used to solve both classification and regression problems, it is mostly used to solve classification problems.

Advantages:

- Modifying the model is easy.
- Model training cost is low.
- It is easy to understand.

4.3.6. Clustering

4.3.6.1. K-Means Clustering

This machine learning algorithm is one of the simplest and popular unsupervised algorithm. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K .

Advantages:

- The algorithm is easy to implement.
- It scales to large data sets.
- Adaptivity is high to new data.

4.3.6.2. Hierarchical Clustering

One of the worst aspects of the knn algorithm is the difficulty of deciding on the number of clusters. On Hierarchical Clustering algorithm we don't have to decide the number of clusters.

4.3.7. Regression

4.3.7.1. Linear Regression

It is a statistical method to examine the mathematical relationship between two or more quantitative variables. When we have only one predictor variable, then we call

the prediction method as simple regression. It finds the best line between the data points. This line is called as a regression line.

4.3.7.2. Logistic Regression

Logistic regression is another technique used by machine learning. The probability of the default class is modeled by logistic regression. If the variable permission type we want to guess is categorical, we can use Logistic Regression.

4.3.7.3. Support Vector Machine (SVM)

Support Vector Machines are basically used to separate the data of two classes in the most appropriate way. For this, decision boundaries or in other words, hyper planes are determined.

SVMs are used in many classification problems from face recognition systems to voice analysis.

4.3.8. Analysis

4.3.8.1. Cluster

Cluster analysis is the process of dividing information into groups according to certain proximity criteria. Each of these groups is called "cluster". Cluster analysis is briefly called "clustering". In clustering, the similarity of the elements with clusters is high, and the similarity between clusters is less. Clustering falls from data mining techniques to descriptive models, namely unattended classification. In unattended classification, the aim is to cluster the data so that it is a cluster that was originally given and not yet classified, for meaningful subsets. Clustering is done completely according to the properties of the incoming data.

4.3.8.2. Time Series

Analyzing and transforming events and processes that develop over time is an important technique to understand historical effects. Implications to be obtained from events that change with time effect and marking these effects are very useful for establishing causality and meaning relationship.

- Time series data: This kind of data is a set of data on the values that at different times, in the data set, a variable takes different values.

- Cross-sectional data: If data is collected at the same point in time, then it is called cross-sectional data.
- Pooled data: Both time series data and cross-sectional data are combined here.

4.3.8.3. Principal Component Analysis (PCA)

Principal component analysis (PCA) is the mathematical technique of explaining information in a multivariable numerical data set with fewer variables but with minimal loss of information. The information in a data set is explained by the total variability here. PCA reduces dimensionality in large datasets. Based on the spectral analysis methods of symmetric matrices, PCA requires users to have advanced knowledge of mathematics, statistics and linear algebra.

PCA generally consists of 5 basic steps:

1. Prepare the data.
2. Create the Covariance / Korel final matrix. Both are the same in scaled data.
3. Calculate the eigenvalue and eigenvectors of the covariance / Korean final matrix.
4. Select Principal Components. Eigenvalues are sorted from large to small and there are corresponding eigenvectors.
5. Calculate the new dataset.

4.3.9. Visualization

The table shows the methods that used the most stable and functional applications and platforms (BigML, Exploratory and Orange). We aim to adapt these basic methods to our Project and use them.

Visualization Methods	Interactivity Method	Extra
-----------------------	----------------------	-------

Scatter Plot	Creates a table showing the information of the hovered point.	When preparing data for analysis, this graph is created for each attribute.
Histogram Chart	Have a scrolling interface for the appropriate data types.	When preparing data for analysis, this graph is created for each attribute.
Pie Chart	Shows the percentage values of the pie chart by hovering.	
Area Chart	A graph of a single variable or a graph showing the interaction of multiple variables.	
Heat Map	Presents correlation between attributes in a table supported by colors. Values can be viewed by hover.	
Design Tree	After hover shows the path of the point.	
Box Plot	Value tables of selected points are displayed.	

Table of Visualization Methods

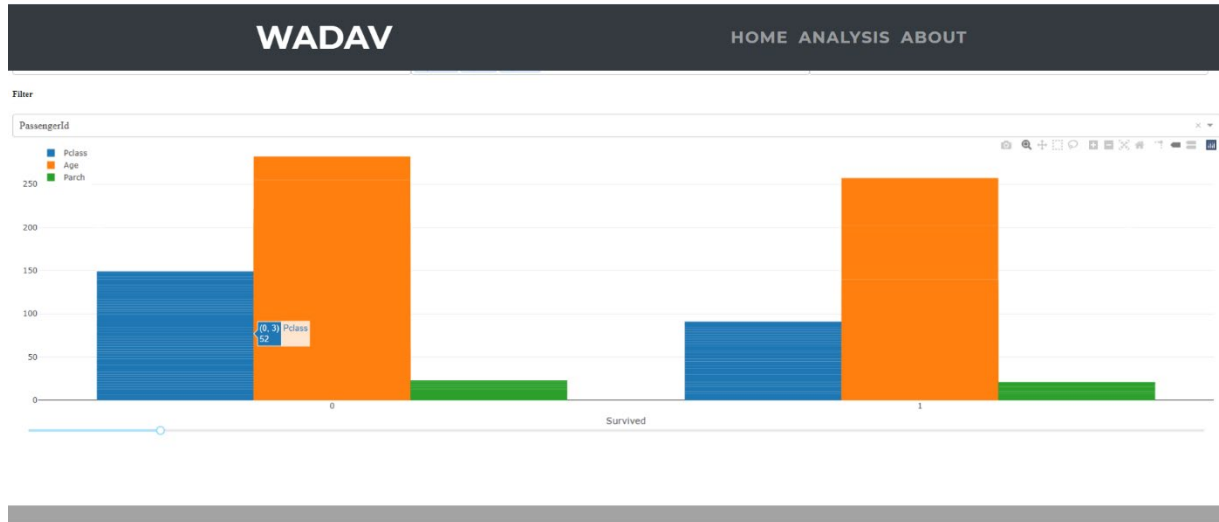
4.4. Interactivity

On this step, users will be able to use the interactivity. They will be able to select a time interval if the data has a relationship between time. If there is an index, they can choose index interval.

The user also will be able to select a part of the data types that has high value of correlations.

4.5. System UI Design

In this part you can find User Interface with project screenshots.



Şekil 1 Index Page


When users connect project page, index page welcome them.

The figure shows the 'ABOUT' page of the WADAV application. The header is the same as the index page. The main content area has a grey background and features the word 'ABOUT' in large white letters, followed by a star icon. Below this, there is a paragraph of text on the left and a list of names on the right. At the bottom center, there is a GitHub logo and the text 'GitHub'.

ABOUT

Wadav analyze and visualize your data with interactive interface. It is totally free for everyone. You can upload your data and start to analyze your data. Whatever you need for data analysis, it's all at hand with Wadav. For more information visit GitHub page.

Wadav created by Çankaya University Computer Engineering students.
Alim GEYİK
Nurseli BAL
Oğulcan BAŞARAN
Büşra KUTLUER

 GitHub

Şekil 2 Index Page

When users scrool down or click “About” , they will see project summary and member of project group..

If users click to “GitHub” button, they can connect Project GitHub Page.

UPLOAD YOUR DATA

Name: File: Dosya Seç Dosya seçilmedi

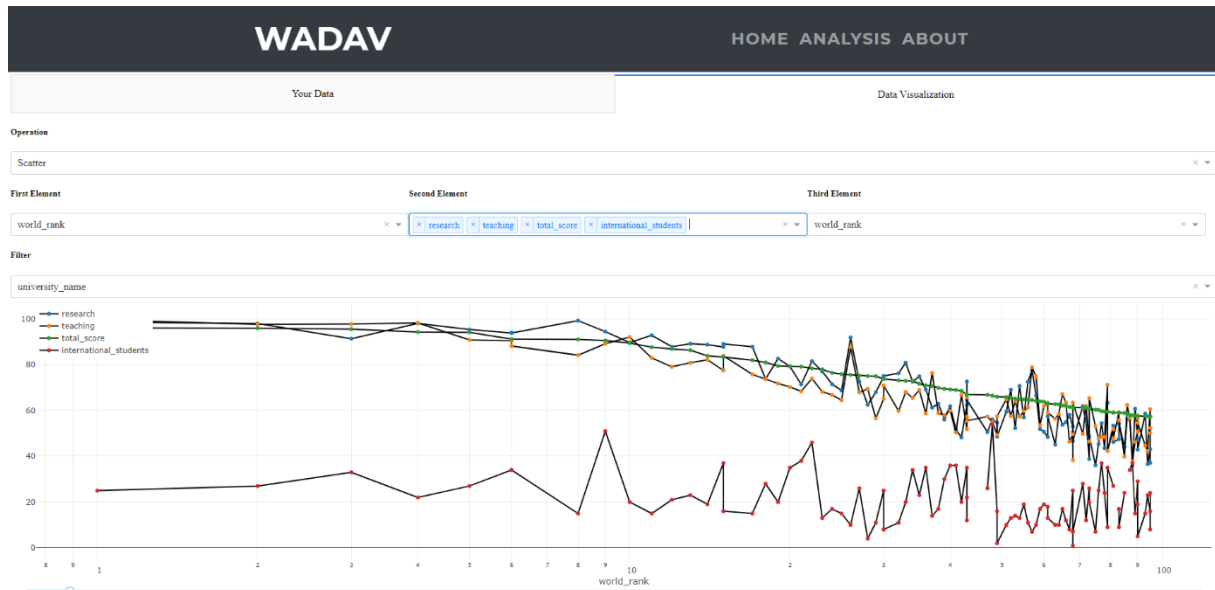
Şekil 3 Data Upload Page

When users click “Analysis” button, they will be directed to the “Upload Page”. In this page they can upload their data to system for analyze it. After upload the data via “Upload File” button, they automaticly connect Analysis Page.

Your Data					Data Visualization									
world_rank	university_name	country	teaching	international	research	citations	income	total_score	num_students	student_staff_ratio	international_students	female_male_ratio	year	
1	Harvard University	United States of America	99.7	72.4	98.7	98.8	34.5	96.1	20,152	8.9	25%		2011	
2	California Institute of Technology	United States of America	97.7	54.6	98	99.9	83.7	96.0	2,243	6.9	27%	33 : 67	2011	
3	Massachusetts Institute of Technology	United States of America	97.8	82.3	91.4	99.9	87.5	95.6	11,074	9	33%	37 : 63	2011	
4	Stanford University	United States of America	98.3	29.5	98.1	99.2	64.3	94.3	15,586	7.8	22%	42 : 58	2011	
5	Princeton University	United States of America	98.9	70.3	95.4	99.9	-	94.2	7,929	8.4	27%	45 : 55	2011	
6	University of Cambridge	United Kingdom	98.5	77.7	94.1	94	57.0	91.2	18,812	11.8	34%	46 : 54	2011	
6	University of Oxford	United Kingdom	88.2	77.2	93.9	95.1	73.5	91.2	19,919	11.6	34%	46 : 54	2011	
8	University of California, Berkeley	United States of America	84.2	39.6	99.3	97.8	-	91.1	36,186	16.4	15%	50 : 50	2011	
9	Imperial College London	United Kingdom	89.2	90.0	94.5	88.3	92.9	90.6	15,060	11.7	51%	37 : 63	2011	
10	Yale University	United States of America	92.1	59.2	89.7	91.5	-	89.5	11,751	4.4	20%	50 : 50	2011	
11	University of California, Los Angeles	United States of America	83	48.1	92.9	93.2	-	87.7	38,206	10.3	15%	52 : 48	2011	
12	University of Chicago	United States of America	79.1	62.8	87.9	96.9	-	86.9	14,221	6.9	21%	42 : 58	2011	
13	Johns Hopkins University	United States of America	88.9	58.5	89.2	92.3	100.0	86.4	15,128	3.6	23%	50 : 50	2011	
14	Cornell University	United States of America	82.2	62.4	88.8	88.1	94.7	83.9	21,424	10.2	19%	48 : 52	2011	
15	ETH Zurich - Swiss Federal Institute of Technology Zurich	Switzerland	77.5	93.7	87.8	83.1	-	83.4	18,178	14.7	37%	31 : 69	2011	
15	University of Michigan	United States of America	83.9	53.3	89.1	84.1	59.6	83.4	41,786	9	16%	48 : 52	2011	
17	University of Toronto	Canada	75.8	-	87.9	82.2	-	82.0	66,198	19.5	15%		2011	
18	Columbia University	United States of America	73.8	90.9	73.8	92.6	-	81.0	25,855	5.9	28%		2011	

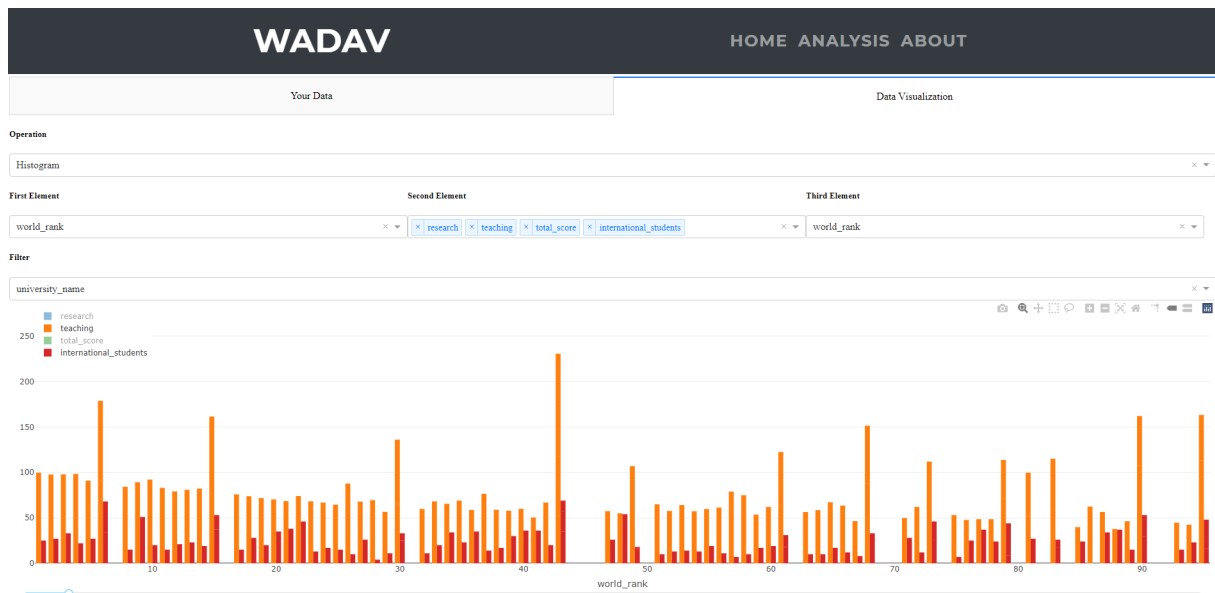
Şekil 4 First look to data in system

When users connect analysis page, they will have 2 tabs (Your Data & Data Visualization). Under the first tab (Your Data), they can look at details of their data.



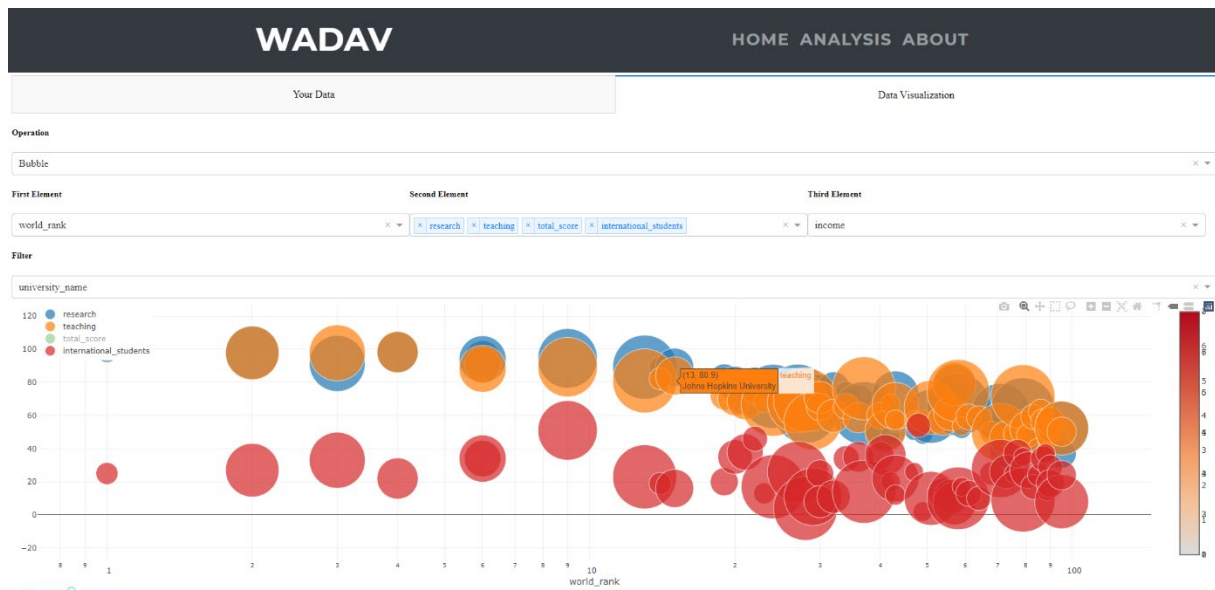
Şekil 5 Scatter Plot Operation

Under the second tab (Data Visualization) users have 5 dropdown menus and one slider. Under the Operation dropdown select “Scatter” and you need 2 elements for axis(x-axis and y-axis). If users want, they can choose more than one elements for y-axis. Filter dropdown menu shows choosen element when users hover on the markers.



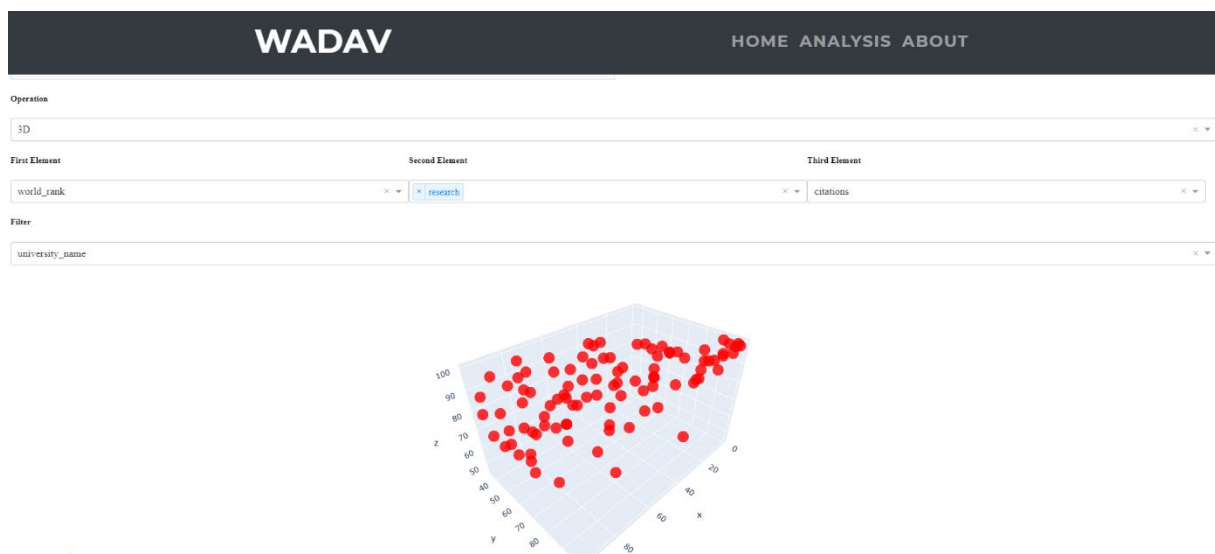
Şekil 6 Histogram Graph

This method works like Scatter method, its need same element.



Şekil 7 Bubble Chart

The Bubble Chart use third element dropdown to change bubbles size. Other elements have same work like other methods.

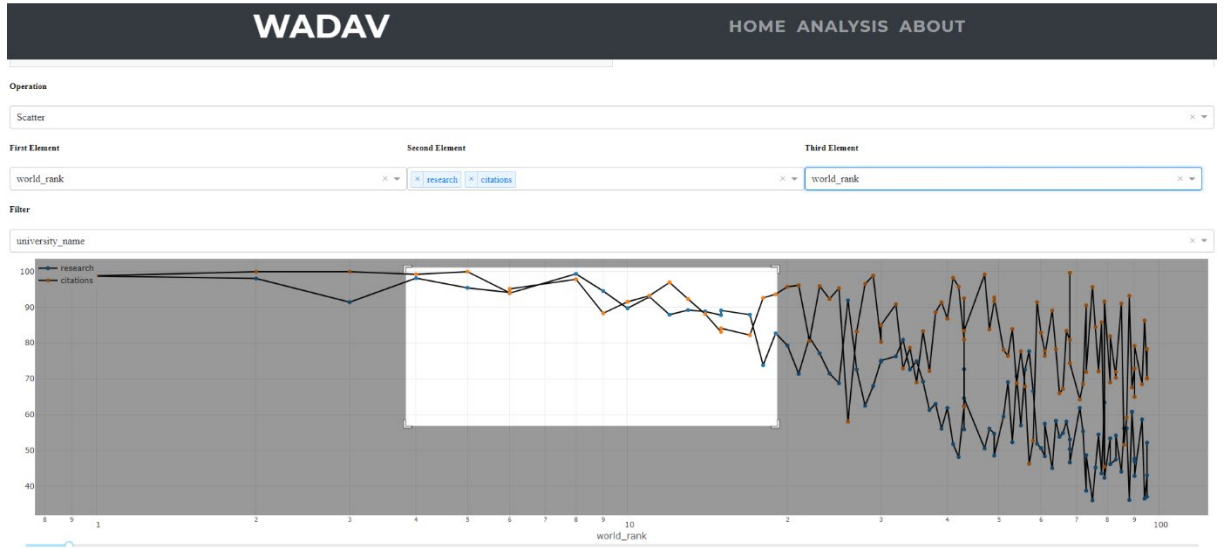


Şekil 8 3D Plot

When users select the 3D operation, users can analyze their data under three dimentionts at the same time. Users can see one attribute's three values, when hover to on it.

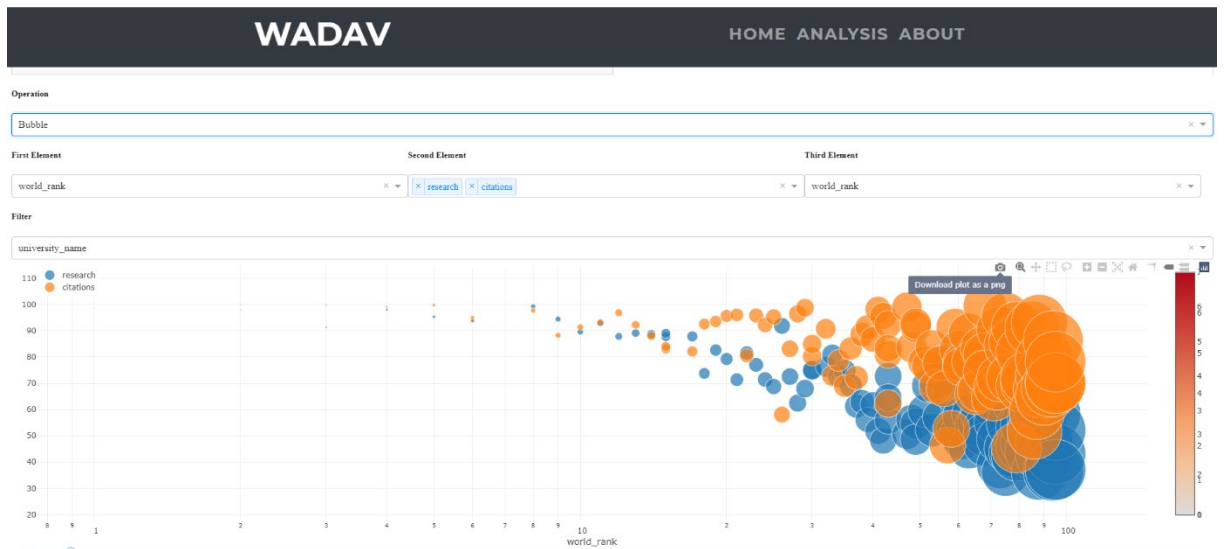
Every plot (operation) has many different interactive elements.

- Hover : Show “Filter & Axis values”
- Legend : Show different elements and activate and deactivate each one.
- Select and Zoom : User can select and zoom any part graph.



Şekil 9 Select and Zoom operation

- Download Graph : Click to PhotographMachine button and download png.



Şekil 10 Download PNG

4.5.1. Index Page

Name: Index Page

PageDescription: There are some buttons to surfing in the Wadav pages (Home – connect to Index page, Analysis – connect to analysis page, About – take an information about Project and group members.

Actor: User

Precondition: None

Operations: Watch a slider to see example of operation and surfing on the web pages via buttons.

Arguments: None

Returns: None

Pre-condition: None

Post-condition: None

Exceptions: None

Basic Sequence:

1. Clicks the button.
2. Scrolling page

Exception: Internet and database connection can be fail.

Post Conditions: None

Priority: High

4.5.2. Upload Page

Name: Upload Page

PageDescription: There is one Text area for give a name to your data. One file select button from your directory on your own computer. And one Upload file button to save your data file to the system.

Actor: User

Precondition: Calling via Analysis button from Index page.

Operations: Select your data from file's directory and give it a name. Upload the data to database.

Arguments: Text and File

Returns: None

Pre-condition: None

Post-condition: None

Exceptions: If you don't upload a file you can not connect the analysis page.

Basic Sequence:

1. Give file name
2. Find your data file from directort
3. Click Upload File button.

Exception: Internet and database connection can be fail. User can upload file other than CSV.

Post Conditions: None

Priority: High

4.5.3. Analysis Page

Name: Analysis Page

PageDescription: There are 5 dropdown menus for select operations, elements of data, filter, and one slider to select part of data.

Actor: User

Precondition: Uploaded File

Operations:

1. Select Operation form Operation Dropdown
2. Select Elements from Elements Dropdown Menus
3. Select Filter to see the name of filter name when you hover on the graph.

Arguments:Elements, operation, filter and slider value

Returns: Graph

Pre-condition: File

Post-condition: None

Exceptions: None

Basic Sequence:

1. Follow the operation steps.
2. Change the graph values interactively.

Exception: Internet and database connection can be fail. Missing selection.

Post Conditions: None

Priority: High

5. Conclusions

This document has discussed the software requirement specification document for the Web-based Multifunctional Graphical User Interface for Data Analysis and Visualization-WMGUIDAV Project, briefly Web App Data Analysis and Visualization - WADAV. This system allows users to analyze their data without using any desktop application.

The data that people create in the virtual world (social media posts, news, shopping bills, emails, etc.) is increasing and accumulating day by day. These data may be unnecessary and meaningless to most of us, but they are vital to an analyst, a company or a business. Of course, not every data is meaningful and usable. At this point data analysis programs and platforms come into play. It is precisely these applications and platforms that will reveal the meaning of the available data and determine the value of this treasure. WADAV is a web platform that allows users to review their data and access the information they need. Data Visualization has been used in many areas to see our data and to interpret it easily. In this report we have represented a Project that is used to visualize our data and the algorithms that we will be using. The purpose of this document is to outline requirements for the WADAV. This document describes the general concept and explains the required features of the system. One of the most important issue here is dealing with the data. We have also discussed Data Preprocessing that is very strong problem in dealing with the data. We have understand it explore it, and make it clear dealing with outliers, noisy data etc.

Referances

- [1]"About BigML.com", *BigML.com - Machine Learning made easy*, 2019. [Online]. Available: <https://bigml.com/about/>. [Accessed: 30- Dec- 2019]
- [2]"Exploratory", *Exploratory.io*, 2020. [Online]. Available: <https://exploratory.io/>. [Accessed: 30- Dec- 2019]
- [3]"Data analysis", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Data_analysis. [Accessed: 04- Nov- 2019]

[4]Guru99.com, 2019. [Online]. Available: <https://www.guru99.com/what-is-data-analysis.html>. [Accessed: 05- Nov- 2019]

[5]"Data visualization beginner's guide: a definition, examples, and learning resources", Tableau Software, 2019. [Online]. Available: <https://www.tableau.com/learn/articles/data-visualization>. [Accessed: 11- Nov- 2019]