Ministry of Education and Science of the Republic of Kazakhstan
Suleyman Demirel University

Rinat Bektemirov

Comparison of Tools for Analyzing Big Data Based on Analysis of Judicial Practice in Kazakhstan

THESIS

Presented in Partial Fulfillment for the
Degree of Master of Science in Computing Systems and Software
(degree code: 6M070400)
Department of Computer Sciences
Faculty of Engineering and Natural Sciences

Supervisor: **Cemil Turan**

Kaskelen, 2019

Abstract

This work introduces a problem of analyzing court practice in Kazakhstan. Focus of this work is to find best solution to introduced problem of analysis. The objective is to review existing solutions in big data analytics, examine them and compare their performance on the tasks of analyzing court practice. To complete this objective many of modern solutions were reviewed and separated by groups, such as frameworks, programming languages and software tools. For comparison Tableau, Orange and RapidMiner software tools were chosen. Furthermore, a program developed to generate sample data for analysis. Results of conducted comparison shows that Tableau is the best software to solve problem of analyzing court practice in Kazakhstan. Measured results fully reflect real solution of analyzing court practice in Kazakhstan and can help Supreme Court to improve their analytical systems.

Аңдатпа

Бұл жұмыс Қазақстандағы сот практиканы талдауын мәселены ұсынады. Бұл жұмыс - талдау үшін тапсырылған ең жақсы шешім табу. Мақсаты көлемде тұрады, үлкен талдауларға талдау жасайды, оларды зерттеуге және оларды практикалық дағдылармен талдауға арналған тиімділіктерді салыстыру. Осы уақытқа дейін көптеген шешімдерді топтарға бөліп, бөлімдерге бөліп, фреймворк, бағдарламалау және бағдарламалық құралдармен бөлісу керек болды. Tableau, Orange және RapidMiner программалық құралдарымен бірге соққыға ұшырады. Сонымен қатар, талдауға арналған іріктеу деректерін генерациялауға арналған бағдарлама әзірленді. Салыстыру нәтижелерімен, бұл Tableau Қазақстандағы практикалық тәжірибе талдауының проблемаларын шешуге арналған бағдарламалық қамтамасыз етудің үздік бағдарламаларын ұсынады. Зерттелген қорытындылар толықтай Қазақстандағы тәжірибе сынағының нақты шешімін шығарып, жоғары сапалы Су жүйесін өздерінің аналитикалық жүйелерін жетілдіре алады.

Аннотация

Данная работа представляет проблему анализа судебной практики в Казахстане. Цель этой работы - найти лучшее решение для поставленной задачи анализа. Цель состоит в том, чтобы рассмотреть существующие решения в области анализа больших данных, изучить их и сравнить их эффективность с задачами анализа судебной практики. Для достижения этой цели многие современные решения были рассмотрены и разделены по группам, таким как фреймворки, языки программирования и программные средства. Для сравнения были выбраны программные средства Tableau, Orange и RapidMiner. Кроме того, разработана программа для генерации выборочных данных для анализа. Результаты проведенного сравнения показывают, что Tableau является лучшим программным обеспечением для решения проблемы анализа судебной практики в Казахстане. Измеренные результаты полностью отражают реальное решение анализа судебной практики в Казахстане и могут помочь Верховному Суду улучшить свои аналитические системы.

Acknowledgements

I would first like to thank my thesis advisor PhD.assist.prof. Cemil Turan of the Faculty of Computer Engineering and Natural Sciences at Suleyman Demirel University. The door to mr. Turan's office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to thank the experts who were involved in the validation survey for this research project: PhD. Bogdanchikov Andrey and prof. Muhamediev Ravil. Without their passionate participation and input, the validation survey could not have been successfully conducted.

Finally, I must express my very profound gratitude to my parents and to my wife for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

To my family

Contents

1. Introduction

1.1 Motivation

The beginning of the complex digitalization of the courts of Kazakhstan is associated with the approval of the Concept of the automated information-analytical system "Judicial Administration". As a result of the implementation of this Concept, by 2007, the Unified Automated Information and Analytical System (UAIAS) was implemented in all courts of Kazakhstan.

Today we ended up having a lot of data gathered by juridical institutions such as, but not limited to, court cases, lawsuits, complaints, etc.

One of the problems of the judicial system of Kazakhstan - digitalization - has already been solved. The second problem is the improvement of the current system by monitoring and optimizing the work of the courts and judges.

Analyzing the amount of decisions of court cases, the burden on judges can be found out, and therefore steps to remedy this situation could be taken. Also, knowing the number of cases canceled for a particular judge, we can find out his competence. Further analyzing the number of convictions and acquittals, we will be able to judge the need for changes in the legislation in general. These examples, of course, do not limit the use of analytics in judicial practice. For example, it can also be useful for parties preparing for a case in a particular court if they can see the number of cases won / lost from a particular judge in a certain category of case.

1.2 Aims and Objectives

Big Data is a huge collection of data that can be gathered, transmitted, accumulated, reserved and analyzed. This reason has made big data a tempting field

for research scholars in the innovative activity of using algorithmic techniques to analyze sophisticated and/or unstructured data pools.

In order to fully evaluate and analyze big data, researchers that work with a data must have a certain kind of awareness to use mighty tools and languages for analysis.

Therefore, studies related to the review and breakdown of available programming tools and applications statistics, analytical programs and software for visualization in the field of big data analysis are relevant.

Object of the study in this paper is big data.

The subject of study is development tools, languages and methods for analyzing big data.

The goal of this work is to compare modern tools for working with big data to show their effectiveness in analyzing court cases in real time.

To achieve this goal it is necessary to solve the following tasks:

- Run a review of software tools for analyzing big data;
- To analyze the process of preparing and processing data;
- Perform selection of software tools for analyzing big data in court practice tasks;
- Generate raw data;
- Create and download data sets for analysis;
- Test and compare data analysis tools.

2. Software tools for big data analysis

There are many analytical tools for big data; they can be divided into three groups: programming languages, statistical solutions and visualization tools [1]. The choice of one or more of them depends on programming experience and knowledge in data analysis.

For example, if you plan to use the R language, you must have good experience in both scientific programming and statistics. On the contrary, when using visualization tools, you can work with large ones without such specific knowledge.

2.1 Cloud platforms

There are many platforms for processing big data on the market, some of them are open source, such as Apache Hadoop and SciDB, while others are proprietary platforms and belong to companies such as Google, IBM, Amazon and Microsoft [1].

Depending on the features of these platforms, many platforms were implemented in the cloud (Google AppEngine, Microsoft Azure, and Amazon EC2); each of these solutions has its own ways of solving big data analysis problems (data storage, analytics, machine learning and implementation).

Table 2.1 presents a comparison between the known cloud platforms for working with big data.

Features	Amazon	Microsoft	Google
Big Data storage	S3	Azure	Google cloud services
Big Data analytics	Elastic	MapReduce (Hadoop)	Hadoop on Azure BigQuery
Relational databases	MySQL or Oracle	SQL Azure	Cloud SQL
NoSQL databases	DynamoDB	Table storage	AppEngine Datastore
MapReduce	Elastic MapReduce (Hadoop)	Hadoop on Azure	AppEngine
Stream processing	No streaming	Streaminsight	Search API
Machine learning	Hadoop+Mahout	Hadoop+Mahout	Prediction API
Data sources	Public datasets	Windows Azure marketplace	Few datasets with examples
Availability	Public	Beta	Beta

Table 2.1: Comparing Big Data Cloud Platforms

2.2 Programming languages

For analyzing big data, several programming languages are used and they can be divided into two groups [2]:

- High level languages;
- Low level languages.

The levels of programming languages are determined by the analytical use of these languages. In high level programming languages there are many functions for solving analytical problems.

2.2.1 Python

Python is one of the most famous programming languages for data analysis. The interactive nature of this language and its scientific system libraries makes it preferable for developing analytical programs and studying hidden facts in data sets [3].

Focusing on the scientific computer community, it is easy to see how the use of the Python language is increasing (since the beginning of 2000) in both

industries: the creation of analytical applications and academic research [4].

Python has its own scientific ecosystem, as well as many useful libraries. Numpy [5] is the base library for basic data structures and the main package in the Python language. Knowing that all input in Python is represented as an array numpy; it is easy to conclude that all the libraries in this language are built on top of this package.

Numpy provides functions:

- Ndarray: efficient and fast multidimensional array.
- A long list of functions for working with arrays using element-by-element calculations with them or providing mathematical operators for working with arrays.
- Tools to read and write array-based data arrays to disk.
- Fourier transform, linear algebra operations, and random number generation.
- Tools for integrating code from other languages (C, C ++ and FORTRAN) into Python.

Pandas [6]: this package that supports scientists and helps solving their problems of data structuring with their preparation and preliminary design.

Matplotlib [7]: package for data visualization. This tool is very popular and effective for the tasks of visualization and graphing, especially for 2D-graphs.

IPython [8]: an environment for interactive computing and development, it is used to maximize performance in both interactive computing and software development. It also includes a multifunctional console with a graphical interface, and a lightweight, fast parallel computing engine.

SciPy [9]: a package with a collection of efficient algorithms for solving linear algebra problems, working with sparse matrix representations, special functions, and basic statistical functions. These packages include:

- scipy.integrate
- scipy.linalg
- scipy.optimize
- scipy.signal

- scipy.sparse
- scipy.special
- scipy.stats
- scipy.weave

Also for standard Fortran-based numeric packages, scipy has bindings for many of them, for example, such as LAPACK.

Cython: Scientists can use Python syntax and high-level operations to improve performance.

2.2.2 R

R is a widespread open source programming language. It is a language for factual investigation and information examination [10]. In R framework, you can do any sort of measurable estimation utilizing useful linguistic structure or program code with ground-breaking troubleshooting apparatuses. This language has interfaces with many programming dialects. Synopsis measurements can be shown utilizing abnormal state graphical instruments R [11].

R has the following features [12]:

- Understandable syntax to speed up data analysis tasks;
- Various formats for downloading and storing data for both local and Internet tasks;
- Ability to perform tasks in memory;
- A set of tools (functions, packages) for analysis tasks, some of them are built-in, the others are open source tools;
- Has various simple ways of presenting the results of data analysis in graphical form with saving these graphs on a disk;
- The ability to automate the analysis and create new functions (R is a programming language);
- Users do not need to reload their data every time because the system saves data between sessions and saves the history of entered commands;
- There are many free GUIs for R:

- RStudio
- R Commander
- StatET
- ESS
- JGR
- Java GUI for R

- Available on all platforms: Windows, Macintosh and Linux.

For all these functions, R provides a wide variety of tools for statistical analysis, machine learning (linear and nonlinear modeling, classical statistical tests, time series analysis, classification, clustering) and graphical methods:

- Data Extraction
- Data Cleansing
- Data Loading
- Data Conversion
- Statistical analysis
- Data Visualization
- Predictive modeling

2.2.3 SAS Software

SAS programming (and programming language) is an outstanding and generally utilized answer for getting to, changing and examining information utilizing an adaptable, extensible web interface [13].

SAS expository stage comprises of an assortment of diagnostic applications that structure the structure of the stage and makes it a helpful device for researchers in the field of enormous information investigation. The fundamental valuable expository applications are:

1. SAS Text Miner: this is a module that can be included for the SAS Enterprise Miner condition, since it encourages the arrangement of issues in content mining. Content Miner can control different wellsprings of printed information:

- Regular nearby content records.
- Text extricated from SAS datasets or other outer databases.
- Files on the Internet.

SAS Text Miner 14.1 incorporates numerous hubs that can be utilized in content examination:

- Text import hub
- Text investigation hub
- Text Filter Node
- Text hub
- Text Cluster Node
- Text Rules Builder

2. SAS Forecast Server: its computerization and adaptability enable associations to settle on increasingly productive and powerful choices dependent on created brilliant estimates. This apparatus improves the productivity of gauges and enables you to pick the most significant estimates and spotlight your endeavors on them.
3. SAS Model Manager: sorts out work on the development of accumulations of systematic models, beginning with the creation, through administration and checking, and finishing with their organization. SAS Model Manager furnishes a leader with an advantageous web condition that encourages client involvement with the board apparatuses and support for the existence cycle of models.

2.2.4 MS SQL Server

MS SQL Server is a very notable answer for customary social databases, and has generally excellent devices for making ERD diagrams, just as for enhancing questions utilizing a graphical apparatus.

MS SQL Server has an incorporated Analysis Service, which has a cozy association with the Tabular Model, Multidimensional Model and the Microsoft BI stack. There are likewise three primary administrations for business knowledge in MS SQL Server:

- SQL Server Integration Service (SSIS) for information gathering.
- SQL Server Analysis Service (SSAS) for information examination.
- SQL Server View Service (SSRS) for survey information (representation).

Microsoft SQL Server has a worked in interface to incorporate with Apache Hadoop (SQL Server Hadoop connector), which is a Sqoop-based interface;

The primary motivation behind the structure of this connector is to give a compelling apparatus to exchanging information between SQL Server and Hadoop [14].

MS SQL Server 2012 bundles give scientists a total arrangement of information incorporation devices, perception critical thinking, a business insight bundle, and the capacity to associate with Apache Hadoop and Hive through an effective extension.

2.2.5 Hadoop Family

Apache Hadoop [15] is an open source programming suite that exploits oSaf PC bunches to store and process a tremendous measure of information. Hadoop comprises of two sections: the archive is HDFS and the handling part is the MapReduce programming worldview.

HDFS [16] is a conveyed, versatile and convenient document framework written in Java. HDFS enables you to store huge records on different machines. It gives unwavering quality by recreating information to other information hubs, which causes excess.

MapReduce [17] enables you to compose applications for parallel preparing of colossal measures of information on huge groups. MapReduce separates work into autonomous pieces, which are handled in parallel by MapReduce errands.

Apache Spark [18] is a brought together investigative motor for huge scale information handling. Sparkle gives an interface to programming whole bunches with verifiable information parallelism and adaptation to internal failure. Sparkle gives a similar adaptability and versatility as MapReduce, yet works quicker for a particular application because of an alternate information reflection and multi-utilitarian API.

2.3 Software tools for statistical analysis

2.3.1 Orange

Orange [19] is a lot of apparatuses for imagining enormous information, AI and information mining with open source. It has a visual programming interface for exploratory information examination and intuitive information representation, and can likewise be utilized as a Python library.

Orange is a segment based visual programming bundle intended for information representation, AI, and information mining.

Orange parts are called widgets and they go from straightforward information perception, choice of subsets and preprocessing to experimental assessment of learning calculations and prescient displaying.

Visual writing computer programs is actualized through an interface wherein work processes are made by connecting pre-characterized or client created widgets, while propelled clients can utilize Orange as a Python library to control information and change widgets.

Orange is an open source programming bundle discharged under the GPL permit. Forms up to 3.0 incorporate real parts in C++ with Python wrappers that are accessible on GitHub. Beginning with rendition 3.0, Orange uses freely accessible Python libraries for logical figuring, for example, numpy, scipy, and scikit-learn, and the graphical UI works in a cross-stage Qt condition. Orange3 has its own different github.

The default setting incorporates various AI, preprocessing and information representation calculations in 6 sets of widgets (information, perception, arrangement, relapse, assessment and control). Extra highlights are accessible as additional items (bioinformatics, information total and content examination).

Orange is bolstered on MacOS, Windows and Linux, and can likewise be introduced from the Python Package Index archive (pip introduce Orange3).

As of May 2018, the steady form 3.13 keeps running on Python 3, and the obsolete variant 2.7 on Python 2.7 is as yet accessible.

The program gives a stage to choosing tests, proposal frameworks, and prescient demonstrating and is utilized in biomedicine, bioinformatics, genomic research, and preparing. In science, it is utilized as a stage for testing new

AI calculations and for presenting new techniques in hereditary qualities and bioinformatics. In training, it was utilized for encouraging AI and information mining to science understudies, biomedicine and software engineering.

2.3.2 RapidMiner

RapidMiner [20] is a product level for coping with big facts created by a corporation of a similar name that offers an incorporated domain to records readiness, AI, top to backside making ready, content research and prescient examination. It is applied for enterprise and non-business programs, just as for research, practise, making ready, quick prototyping and alertness development, and supports all phases of the AI process, together with statistics planning, representation of outcomes, model approval and streamlining. RapidMiner relies upon on an open middle model. RapidMiner Studio loose version, that is limited to at least one coherent processor and 10,000 strains of facts, is obtainable under the AGPL allow.

RapidMiner, some time ago known as YALE (another learning condition), was created in 2001 by Ralph Klinkenberg, Ingo Miersva and Simon Fisher from the Artificial Intelligence Division of Dortmund Technical University. Since 2006, it has been driven by Rapid-I, an organization established by Ingo Miersva and Ralph Klinkenberg around the same time. In 2007, the name of the product was changed from YALE to RapidMiner. In 2013, the organization was renamed from Rapid-I to RapidMiner.

RapidMiner utilizes a customer/server model with a server offered in both neighborhood and open or private cloud foundation. As per Bloor Research, RapidMiner gives 99% of cutting edge scientific arrangements dependent on format conditions that speed conveyance and diminish mistakes, nearly dispensing with the requirement for composing code. RapidMiner gives information mining and AI systems, including: information stacking and change (extraction, change, stacking (ETL)), information pre-preparing and perception, prescient investigation and factual displaying, assessment and organization. RapidMiner is written in the Java programming language. RapidMiner gives a graphical interface to creating and executing scientific work processes. These work processes are classified "forms" in RapidMiner and comprise of a few "administrators". Every administrator performs one assignment inside the procedure, and the

yield of every administrator shapes the info information of the following. On the other hand, the instrument can be called from different projects or utilized as an API. Separate capacities can be called from the direction line. RapidMiner gives preparing plans, models and calculations and can be extended utilizing R and Python contents.

The usefulness of RapidMiner can be stretched out with extra modules that are accessible through the RapidMiner Marketplace. RapidMiner Marketplace furnishes engineers with a stage for making information examination calculations and distributing them in the network [21].

In 2018, Gartner put RapidMiner in the initiative part of its Magic Quadrant for information and AI stages. The report takes note of that RapidMiner gives profound and expansive demonstrating capacities for computerized start to finish model advancement. In the yearly programming overview in 2018, KDnuggets perusers perceived RapidMiner as a standout amongst the most well known information investigation programming. Review respondents demonstrated that this product bundle is the apparatus they use. RapidMiner has gotten a huge number of downloads and has more than 400,000 clients, including BMW, Intel, Cisco, GE and Samsung as paid clients. RapidMiner cases to be the market head in programming for information preparing stages against contenders, for example, SAS and IBM.

Around 50 designers worldwide are associated with the open source improvement of RapidMiner, the vast majority of who are workers of RapidMiner. The organization, which creates RapidMiner, got financing of Series C in the measure of $16 million with the investment of funding organizations Nokia Growth Partners, Ascent Venture Partners, Longworth Venture Partners, Earlybird Venture Capital and OpenOcean. OpenOcean accomplice Michael "Monty" Widenius is the organizer of MySQL.

2.3.3 Tableau Software

Tableau Software is a product improvement organization headquartered in Seattle, Washington, USA, that produces business knowledge items for intelligent information representation. Initially created to market thinks about that was led at the Faculty of Computer Science at Stanford University from 1999 to 2002. Tableau items inquiry social databases, intuitive investigative informa-

tion shapes, cloud databases and spreadsheets, and after that produce a few kinds of charts. Items can likewise extricate information, just as store and recover it from the information system in memory [22].

Tableau has a mapping capacity and can show scope and longitude, organize and associate with spatial records, for example, Esri Shapefiles, KML and GeoJSON, to show client topography. Installed geocoding enables you to show managerial areas (nation, state/territory, locale/region), postal divisions, US Congress regions, US CBSA/MSA, region codes, airplane terminals, and measurable zones of the European Union (NUTS codes). There are five different ways to get to Tableau items: Desktop (for both expert and individual versions), Server, Online (which supports a large number of clients), Reader and Public, the last two are free. Vizable, a versatile shopper information representation application, was discharged in 2015. The 6th item, known as Data Prep, for information readiness work processes, was discharged in May 2018 [23].

2.3.4 SPSS

SPSS is a software product of the Statistical Package for the Social Sciences (SPSS). The current version of this software is called "IBM SPSS Statistics". The full version of this software product contains the following set of features and add-ons [24]:

- Basic statistics
- Advanced statistics
- Exact Tests
- Categories
- Samples
- Custom tables
- Data preparation
- Decision Trees
- Prediction
- Neural networks
- Regression

3. Research design and methodology

3.1 Selection of software tools for analyzing big data in court practice tasks

Before analysis data should be created and stored somewhere. For purpose of this research PostgreSQL relational database was chosen. Reason for choice was that originally Supreme Court uses this database to store their data. Also PostgreSQL is easy to setup and use, which is great advantage over other RDMS software.

For testing purposes RapidMiner, Orange and Tableau were chosen. Despite of undeniable advantages of big data analytics frameworks such as Apache Lucene or Apache Hadoop; or programming languages like Python or R; chosen software tools offer big convenience over other methods - simplicity. Proprietary software offer easy to understand GUI and integration with most of the data formats. This facts made choice obvious.

3.2 Formation of source data

For data processing and analysis, in this work, it is proposed to use the database of court cases in Kazakhstan. This database contains a description of court cases, a list of participants and the results of decisions made. The data scheme is presented in Figure 3.1.

Consider the description of all tables of the submitted database:

- cases: a table with a description of all cases considered in the courts of

Kazakhstan;

- courts: table-reference with a list of vessels;
- regions: table reference with the regions to which the courts belong;
- case_categories: table-reference category of the case;
- result_types: table reference with a description of the results of consideration of cases;
- instances: reference table with instances. Cases can be first instance, appeal or cassation. The work deals only with cases of first instance in order to simplify the analysis;
- case_types: reference table with case type. Civil, administrative or criminal;
- case_participants: the table is linked to the cases table. It stores the participants of the case, and their role in the court case;
- participant_roles: reference table with roles in the case. For example, a judge, a lawyer, etc.;
- Participants: a table of reference with all participants who participated in the affairs;
- participant_types: participant type can be individual or company.

Consider the description of the database tables, which is used for analysis. Tables A.1 - A.8 provide a description of all the database tables.

To create relationships between data tables, it is necessary to describe key fields and relationships between tables. Table A.9 provides a description of the relationships between database tables.

As next step data has to be generated. For data generation Java program was written to fill tables and make relations with them.

Data generation begins with creating dictionaries such as roles, courts, regions. This information was taken from official website of Supreme Court. Other dictionaries like participants were created dinamically using opensource data on internet.

After all dictionaries were created and loaded process of generating cases started. When new case created all its values was generated randomly from

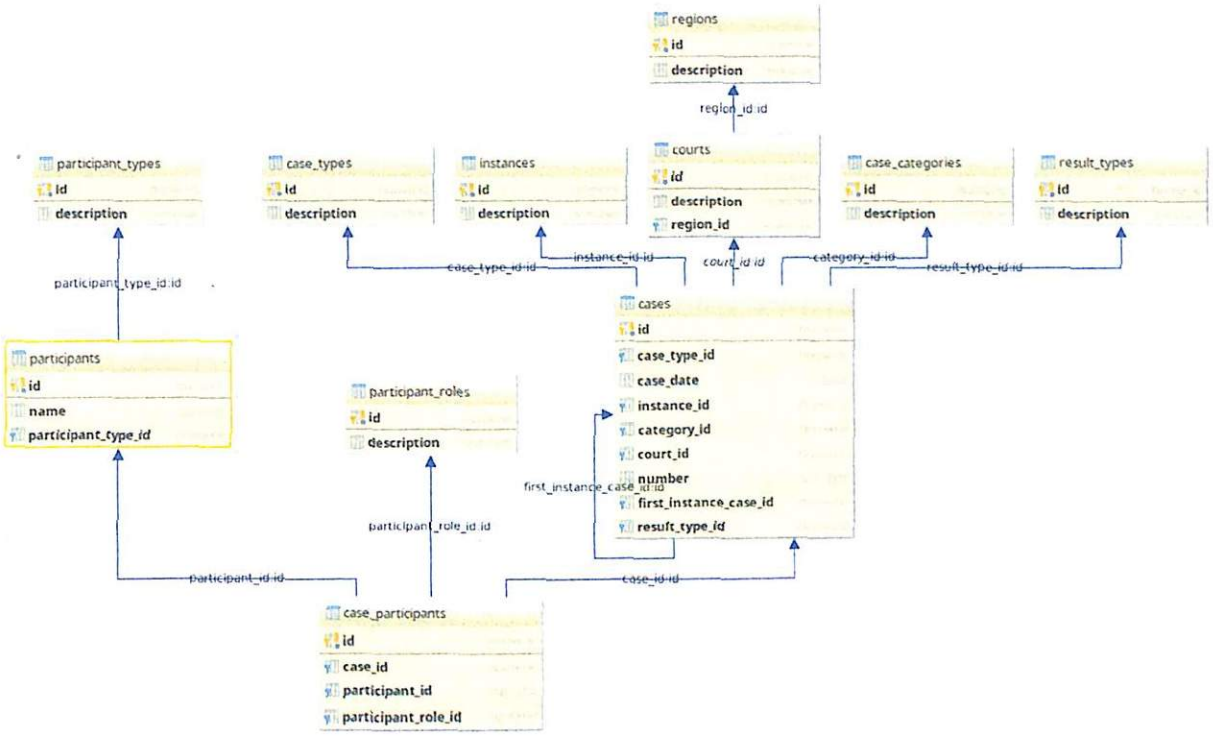


Figure 3.1: Data schemas for the source database

dictionaries. For testing purposes were created approximately 11M cases for each case type. Although this amount exceeds real numbers, but amount was chosen taken into account future cases.

3.3 Creating and loading datasets for analysis

Before start testing every tool on analysis, all necessary data must be loaded in each tool. This process differs from tool to tool. In some cases like Tableau can connect directly to relational database and download all data with relations. Other tools like Orange and RapidMiner need csv files to work.

Below is the process how data loading looks for every tool separetely.

3.3.1 Tableau

Consider the process of creating and uploading datasets to the Tableau online platform. After creating an online solution for data analysis, you need to create the site and activate it (Fig. B.1).

After creating the site, the user goes to the main page where he can connect to the distributed database (Fig. B.2).

After selecting the Connect to data item, you need to select the type of DBMS to which you are connecting and specify the connection parameters.

The work uses a distributed database that is loaded into Microsoft Azure and is available at: p-court.postgres.database.azure.com. To connect to a remote database, you must fill in the connection fields, as shown in Figure B.3.

After connecting to the database, all tables will be available and displayed in the Database-Table field. For further data analysis, it is necessary to create data sets (data sets) that will provide the necessary data sets for analysis.

Consider the process of creating the core data sets that will be used for analysis. It is proposed to use the following kits:

- Aggregation by type of cases;
- Selection and sorting according to the results (decisions made) in the considered cases;
- Selection and sorting of cases, according to the courts of a particular region;
- Aggregation by region and by category of cases;

Consider the process of creating a data set for analyzing court decisions by category of cases. The description of the database tables given in section 3.2 allows us to define a set of tables that need to be combined in a given data set. Since the aggregation of data by types of court cases will be performed, the data set must include data from the cases table (Figure B.4) and data from the case_types table (Figure B.5).

Next, a data set is created that combines data from the cases and cases_types tables (Figure B.6)

Consider the process of creating a data set for the selection and sorting by results (decisions) in the cases considered. The description of the database tables given in section 3.2 allows us to define a set of tables that need to be combined in a given data set. Since the aggregation of data on the results of court cases will be performed, the data set must include data from the cases table (Figure 3.2) and data from the result_types table (Figure B.7).

Next, a data set is created that combines data from the cases table and result_types (Figure B.8)

View Data: cases

1000 • rows

case_date	case_type_id	category_id	court_id	first_instance_case_id	id	in
2013-07-13	3	138	69	null	142320233	
2015-02-16	3	33	129	null	142320239	
2013-08-25	3	20	276	null	142320245	
2013-12-20	3	41	163	null	142320252	
2016-04-12	3	132	171	null	142320258	
2016-11-01	3	144	187	null	142320263	
2012-03-16	3	13	5	null	142320270	
2015-11-17	3	100	20	null	142320277	
2013-07-28	3	120	46	null	142320283	
2012-08-01	3	28	358	null	142320290	
2012-03-05	3	145	250	null	142320295	
2012-12-02	3	78	158	null	142320301	

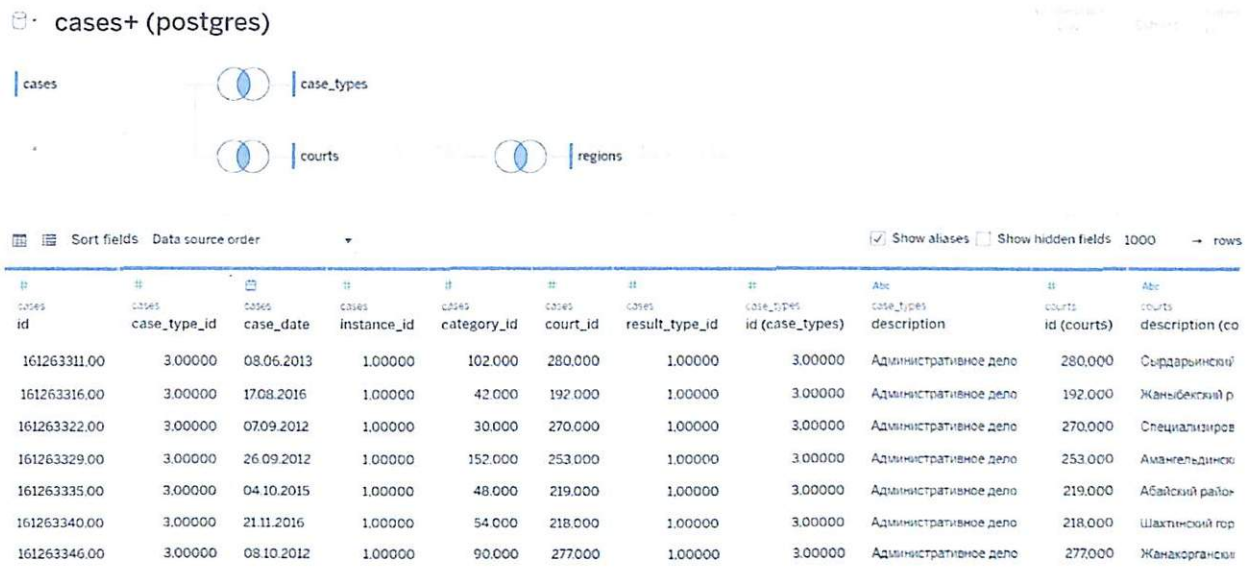
Figure 3.2: The cases table data

Consider the process of creating a data set for the selection and sorting of cases by the courts of a particular region. Since aggregation of data by vessels for a particular region will be performed, the data set must include data from the cases table and data from the courts tables (Figure B.9) and regions (Figure B.10).

Next, a data set is created that combines data from the cases and case_types, courts and regions tables (Figure 3.3)

The online platform Tableau not only allows you to connect to data sources, but also generates data sets and allows you to export these data sets for use in other programs. Consider the process of exporting a data set to a csv file. As an example of a data set, choose data by region and category of cases. After creating the dataset, you must specify the measurements and the facts for which then will be analyzed. In our case, this is the name of the type of case and the name of the region in which the case was reviewed. The facts are considered the number of cases of a certain type, aggregated by region. An example of setting up measurements and facts for a data set is presented in Figure B.11.

After creating and configuring a dataset, you can export a dataset. To do



cases id	cases case_type_id	cases case_date	cases instance_id	cases category_id	cases court_id	cases result_type_id	case_types id (case_types)	case_types description	courts id (courts)	courts description (co
161263311.00	3.00000	08.06.2013	1.00000	102.000	280.000	1.00000	3.00000	Административное дело	280.000	Сырдарьинский
161263316.00	3.00000	17.08.2016	1.00000	42.000	192.000	1.00000	3.00000	Административное дело	192.000	Жаныбекский р
161263322.00	3.00000	07.09.2012	1.00000	30.000	270.000	1.00000	3.00000	Административное дело	270.000	Специализиров
161263329.00	3.00000	26.09.2012	1.00000	152.000	253.000	1.00000	3.00000	Административное дело	253.000	Амангельдинск
161263335.00	3.00000	04.10.2015	1.00000	48.000	219.000	1.00000	3.00000	Административное дело	219.000	Абайский райо
161263340.00	3.00000	21.11.2016	1.00000	54.000	218.000	1.00000	3.00000	Административное дело	218.000	Шахтинский гор
161263346.00	3.00000	08.10.2012	1.00000	90.000	277.000	1.00000	3.00000	Административное дело	277.000	Жанакорганский

Figure 3.3: A set of data for analyzing data by type of cases

this, click the export button and select the data saving format (Fig. B.12). After selecting the storage format, the platform will generate and display aggregated data on the selected dimensions and facts (Fig. B.13). After viewing the data, you can download it as a csv file that can be used in other data analysis programs.

3.3.2 RapidMiner

Consider the process of loading data sets that were obtained on the Tableau platform. To work with RapidMiner, you need to start the application and create a new project. After creating the project, a designer window will open in which you need to select the Data-File tool (Fig. B.14).

Figure 3.4: Selecting a tool to import a file with a set of data

After selecting a file and importing data, the user has the ability to upload

data to RapidMiner and perform preliminary processing and analysis of this data set. To process a dataset, you must start the process in RapidMiner and then view the results of the work in the Results window (Fig.B.15). When you select the Statistics tab, you can view the metadata on the downloaded data (Fig. 3.4).

3.3.3 Orange

Consider the process of creating a project in Orange, as well as the process of importing a data set on the example of a csv file with data previously created in Tableau. To create a new project, select the menu item File-New. After that, you must specify the name of the project (Fig.B.16).

Next, in the project window, select the File widget and configure the file access parameters with the data set (Figure B.17). To display data, unlike Tableau and RapidMiner, in Orange you need to use the Table widget. You must select the Data Table widget and connect it to the File widget, specifying exactly which data will be displayed in the Data Table widget (Fig.3.5).

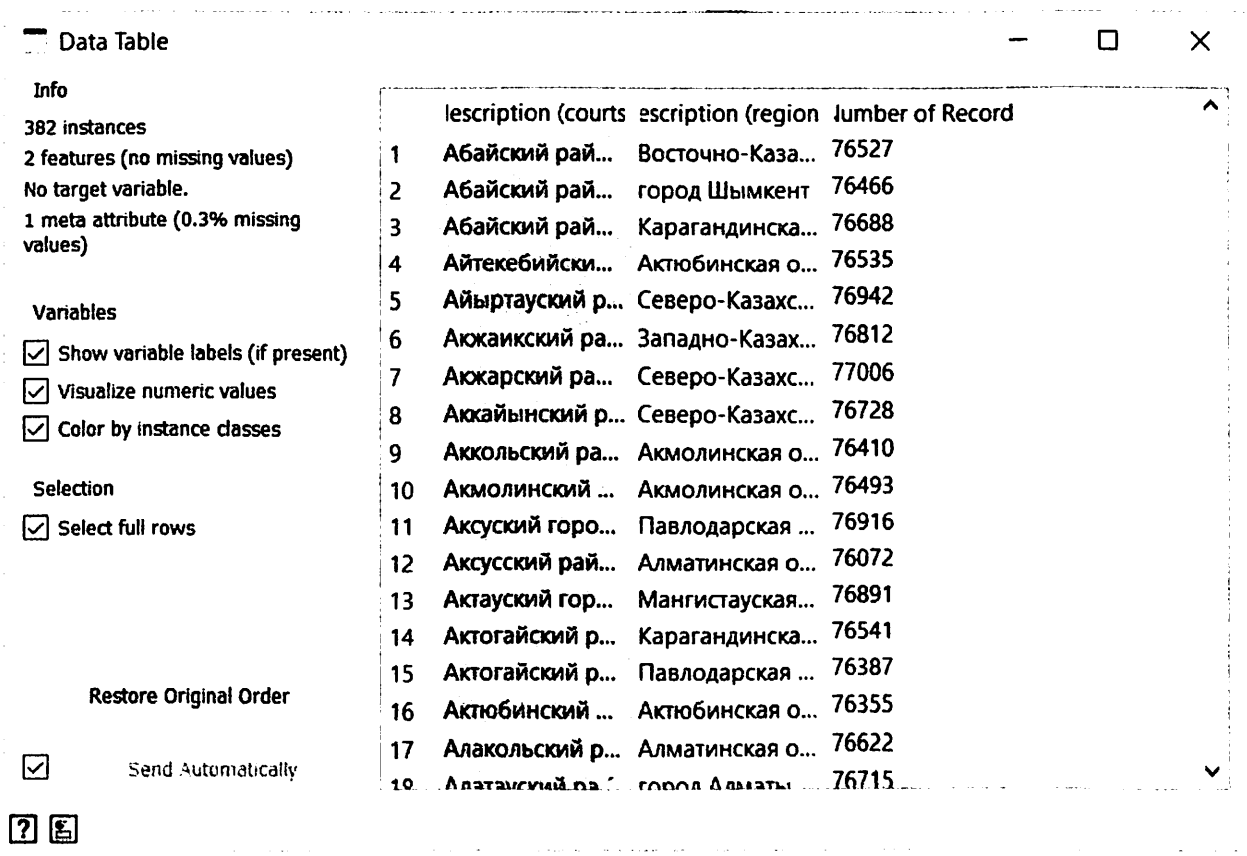


Figure 3.5: Working with the Data Table widget

4. Testing and results

To test and compare big data analysis tools, consider the analysis and visualization of data for previously formed data sets. Consider the process of visualizing a dataset that contains aggregated data on considered court cases in the courts of Kazakhstan.

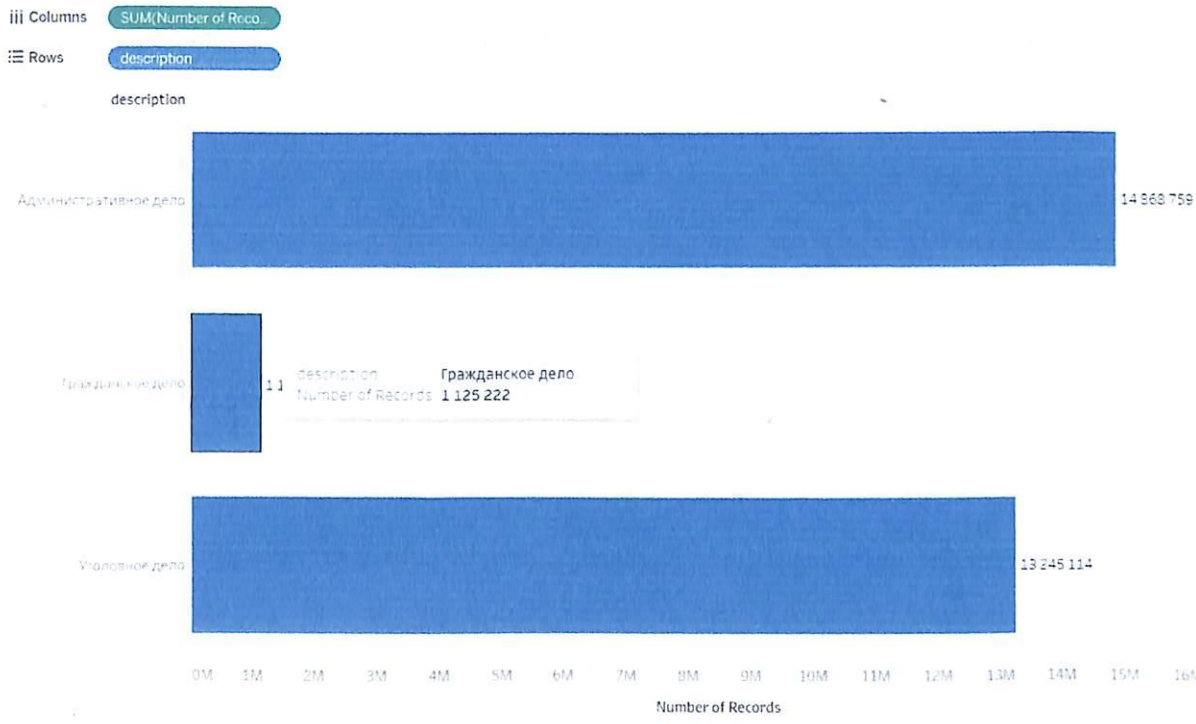


Figure 4.1: Visualization of data on cases heard in court

To start working with a dataset, you must select the measurements on which the visualization will be performed. The fields are selected in the Dimension field, which contains a list of all available fields from the selected data set (Figure B.18).

To visualize the aggregated data on cases that were considered in the courts of Kazakhstan, the description fields (case_types) result_type_id are selected. Grouping is carried out by the field type of the case (description), and aggregation is performed by the field result_type_id (Figure B.19).

After selecting the fields for visualization, the Tableau platform will perform the necessary calculations, build and display a graph of the total number of cases that have been listened to, grouped by type of cases (Fig. 4.1).

An example of the same graph is presented in Figures B.20 and B.21 for RapidMiner and Orange tools, respectively.

Tableau platform tools provide an opportunity to conduct primary statistical data analysis. To carry out such an analysis, go to the Analytics tab (Figure B.22).

The tools on the Analytics tab allow you to calculate and add to the chart: a line with a constant value, a line of average value, medians and quartiles, calculate grand totals, etc. Consider an example of adding a total and average for a previously constructed graph. To do this, select the Totals tool in the Analytics panel and drag it onto the chart. The line for the average value is constructed similarly (Fig.4.2).

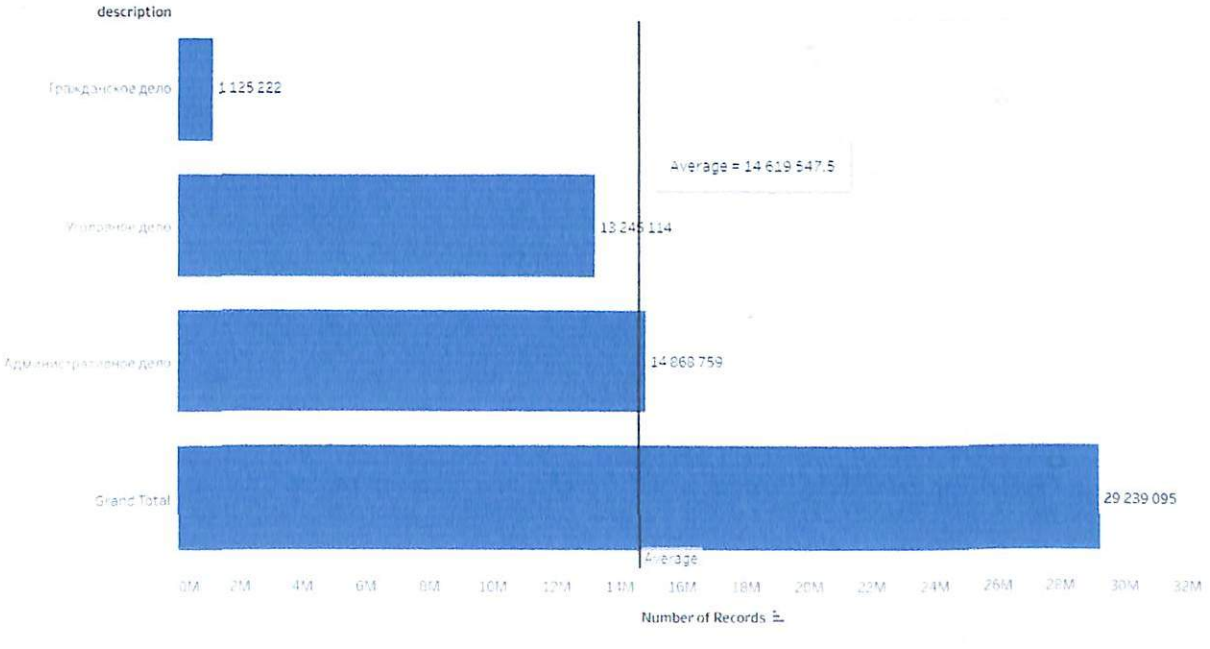


Figure 4.2: Totals tool in the Analytics

And RapidMiner analyst is presented on a separate tab Statistics (Fig.B.23). In Orange, this tool is missing.

Consider the process of data set visualization, which contains aggregated data on the considered court cases in the regions of Kazakhstan.

To start working with a dataset, you must select the measurements on which the visualization will be performed.

To visualize the aggregated data for the regions of Kazakhstan, the description (regions) result_type_id fields are selected. The grouping is carried out by the region (description) field, and the aggregation is performed by the result_type_id field (Figure B.24).

After selecting the fields for visualization, the Tableau platform will perform the necessary calculations, build and display a graph of the total number of cases heard with their grouping by regions of Kazakhstan (Fig.4.3).

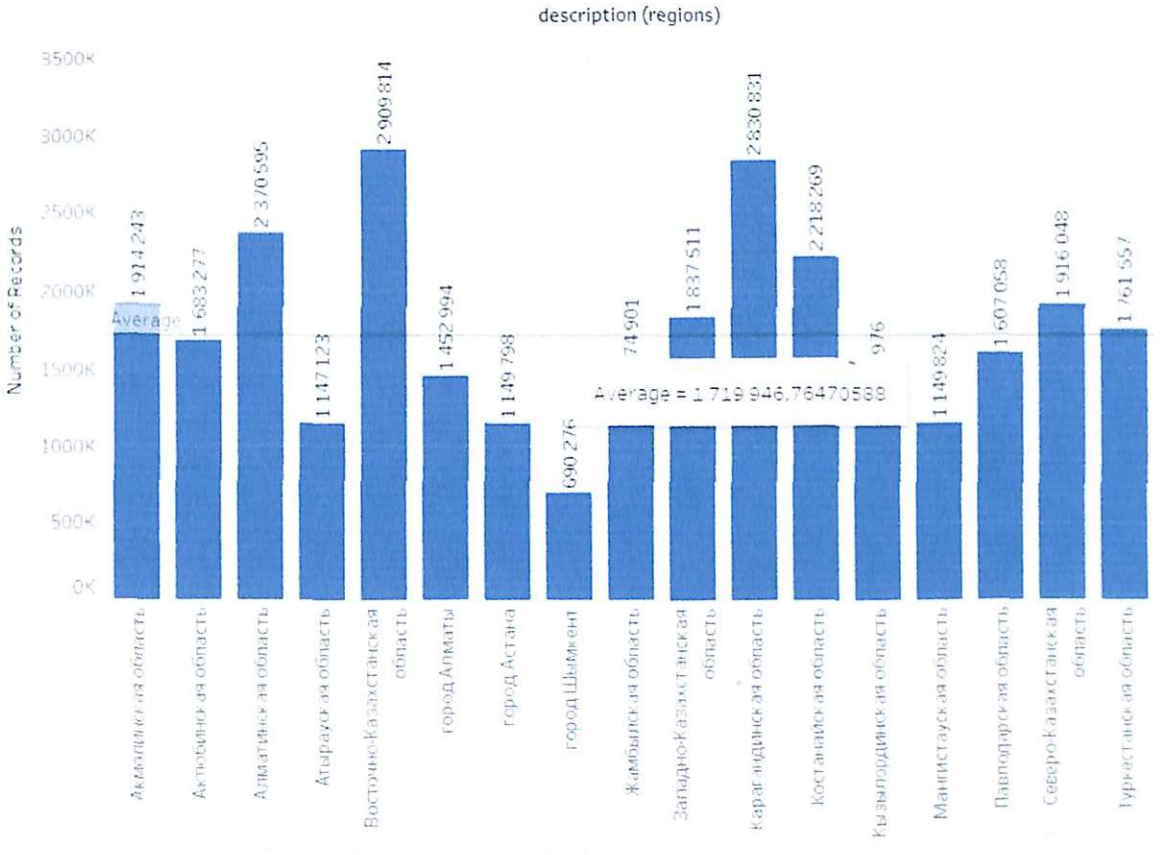


Figure 4.3: Visualization of data on cases heard in court by regions of Kazakhstan

An example of visualization of the same data in RapidMiner is presented in Figure B.25.

Tableau platform tools provide the ability to sort data in ascending and descending order. Data on cases heard by regions of Kazakhstan, sorted in ascending order, is presented in Figure B.26. This sorting allows you to identify regions with the smallest and largest number of cases considered.

In rapidMiner and Orange there are also data sorting tools (Fig.B.27-B.28).

Consider the process of visualizing a dataset that contains aggregated data on the considered court cases in the regions of Kazakhstan, grouped by type of

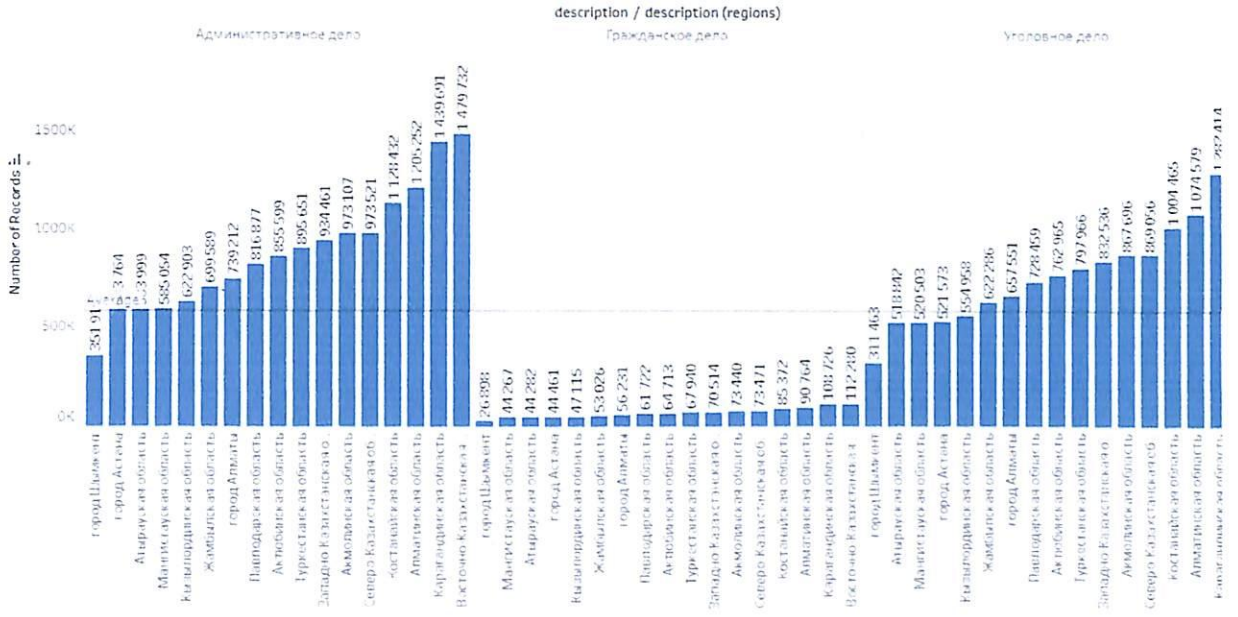


Figure 4.4: Visualization of data on cases heard in court by regions of Kazakhstan, grouped by type of case

case.

To start working with a dataset, you must select the measurements on which the visualization will be performed.

For visualization of the aggregated data, the fields description (regions), description (case_types) and result_type_id are selected. The grouping is carried out by the region (description (regions)), the type of the case (description), and the aggregation is performed by the result_type_id field (Figure B.29).

After selecting the fields for visualization, the Tableau platform will perform the necessary calculations, build and display a graph of the total number of cases heard with their grouping by regions of Kazakhstan and by types of cases (Fig.4.4).

Features	Tableau	RapidMiner	Orange
Rendering speed	High	Above average	Good
Data integration	Excellent	Good	Requires additional widgets
Modeling and analytics	Excellent	Good	Excellent
Display	Excellent	Average	Good

Table 4.1: Comparison of big data analysis and visualization tools

Visualization tools RapidMiner and Orange make it impossible to group data on the same graph. It is possible to build separate graphs, as shown in

Figures B.30-B.16.

Comparison of visualization tools is presented in Table 4.1.

Analysis shows that Tableau public dominates over other in all aspects. This means that Tableau public might be preferable tool to analyze Kazakhstan court practice.

5. Conclusion

The thesis contains relevant research related to the review and analysis of available programming languages and statistical tools, analytical solutions and visualization applications in the field of big data analysis.

The object of the study was big data. The subject of the study was programming languages and tools for analyzing big data.

The aim of the work was to compare modern tools for working with big data, in order to compare their effectiveness in analyzing court cases in real time.

To achieve this goal, in the work, the following tasks were formulated and solved:

- Reviewed software tools for analyzing big data;
- The analysis of the process of preparing and processing data;
- The selection of software tools for the analysis of big data in the tasks of judicial practice;
- Formed sets of baseline data for the analysis of court cases in Kazakhstan;
- Created and loaded data sets for analysis;
- Completed testing and comparison of data analysis tools.

In the first chapter of the work, a review of programming languages and software tools for analyzing and visualizing big data was made. The chapter describes various types of tools for analyzing big data in different areas (programming languages, statistical solutions and visualization tools). The analysis made it possible to determine which of them is more popular than others. It was noted that R is a common programming language for use in data analysis, SPSS is easy to use for statistical analysis by non-statisticians, and Tableau Public is an ideal visualization tool for presenting data and analyzing it graphically.

Second chapter describes architecture of data, how cases are created stored and related to each other. Method of generation sample data for analysis also mentioned in this chapter.

In the third chapter various tests were conducted on sample data using analytical software. Results show that Tableau is far ahead of Orange and RapidMiner in all tests. Here can be safely assumed that Tableau software can be used to analyze court practice in Kazakhstan.

However, market of analytical software is too big to say that Tableau is the best software for big data analysis. Comparison of other software existing now can be a reason for further research.

Furthermore, in this research rather simple tests were conducted to show aggregation, sorting and filtering capabilities of analytical software. Some more sophisticated tests could be performed to check if results change.

Gained results of this research can be used by Supreme Court as base to seek improvement of their analytical systems.

A. Appendix A

Field name	Description	Data type	Key type
Id	Case Type Identifier	Numeric	Primary key
Description	Type Name	varchar	

Table A.1: Description of the participiant_types table

Field name	Description	Data type	Key type
Id	Case Type Id	Numeric	Primary key
Description	Name of case type	varchar	

Table A.2: Description of the case_types table

Field name	Description	Data type	Key type
Id	Region ID	Numeric	Primary key
Description	Name of the region	varchar	

Table A.3: Description of regions table

Field name	Description	Data type	Key type
Id	Case Category Id	Numeric	Primary key
Description	Case Category Name	varchar	

Table A.4: Description of case_categories table

Field name	Description	Data type	Key type
Id	Case Participant Id	Numeric	Primary key
Name	Name of the participant in the case	varchar	
Participiant_type_id	Member Type Id	Numeric	Foreign key

Table A.5: Description of participants table

Field name	Description	Data type	Key type
Id	Identifier of the role of the case participant	Numeric	Primary key
Description	Role of a case participant	varchar	

Table A.6: Description of participiant_role table

Field name	Description	Data type	Key type
Id	Identifier of the role of the case participant	Numeric	Primary key
Case_id	Case ID	Numeric	Foreign key
Participiant_id	Participiant ID	Numeric	Foreign key
Participiant_role_id	Identifier of the role of the participant in the case	Numeric	Foreign key

Table A.7: Description of case_participants table

Field name	Description	Data type	Key type
Id	Case ID	Numeric	Primary key
Case_type_id	Case Type Id	Numeric	Foreign key
instance_id	Instance ID	Numeric	Foreign key
category_id	Case Category Id	Numeric	Foreign key
Court_id	Court ID	Numeric	Foreign key
First_instance_case_id	First instance case ID	Numeric	Foreign key
Result_type_id	Decision ID	Numeric	Foreign key
Case_date	Date of decision	Date	

Table A.8: Description of cases table

Parent table	Child table	Relation name	Relation type
Regions	Courts	Region_id:id	1:M
Participiant_types	perticipiants	Participiant_type_id:id	1:M
case_types	cases	case_type_id:id	1:M
instances	cases	instances_id:id	1:M
courts_types	cases	courts_id:id	1:M
case_categories	cases	category_id:id	1:M
result_types	cases	result_type_id:id	1:M
Participiant_roles	Case_participiants	Participiant_role_id:id	1:M
Cases	Case_participiants	case_id:id	1:M
Participiants	Case_participiants	Participiant_id:id	1:M

Table A.9: Description of relations between database tables

B. Appendix B

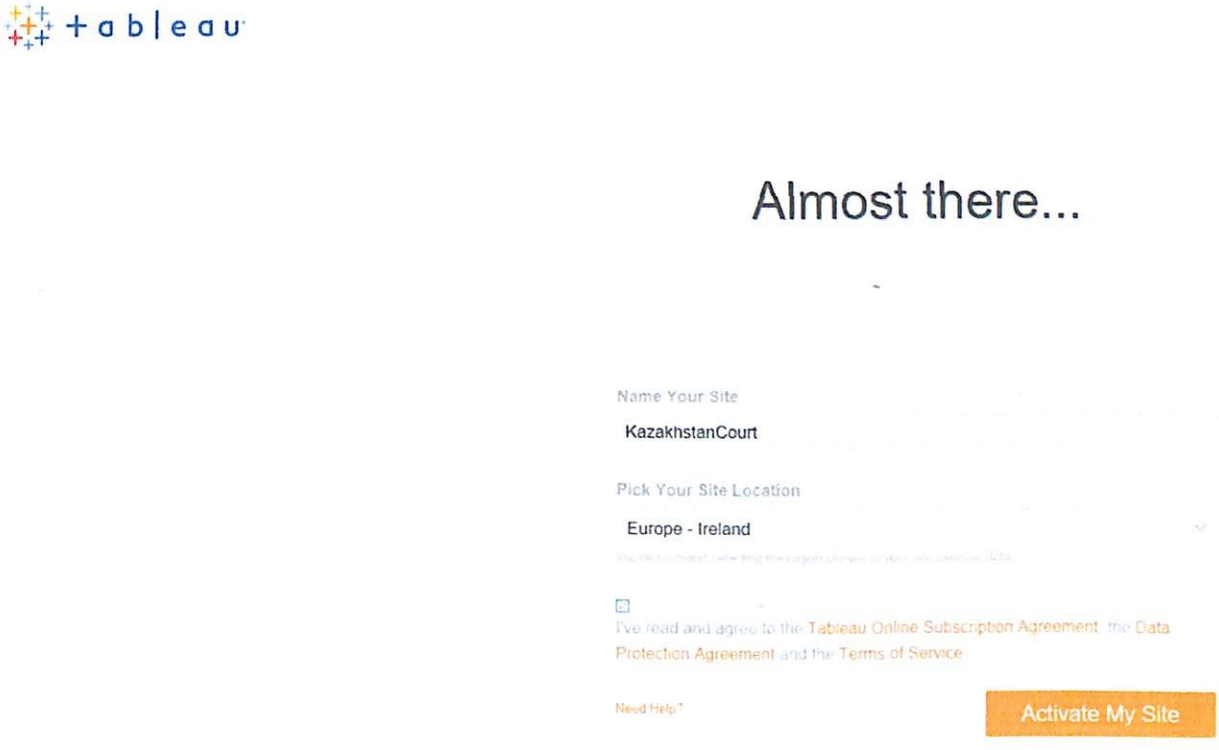


Figure B.1: Creating and activating a site for online data analysis

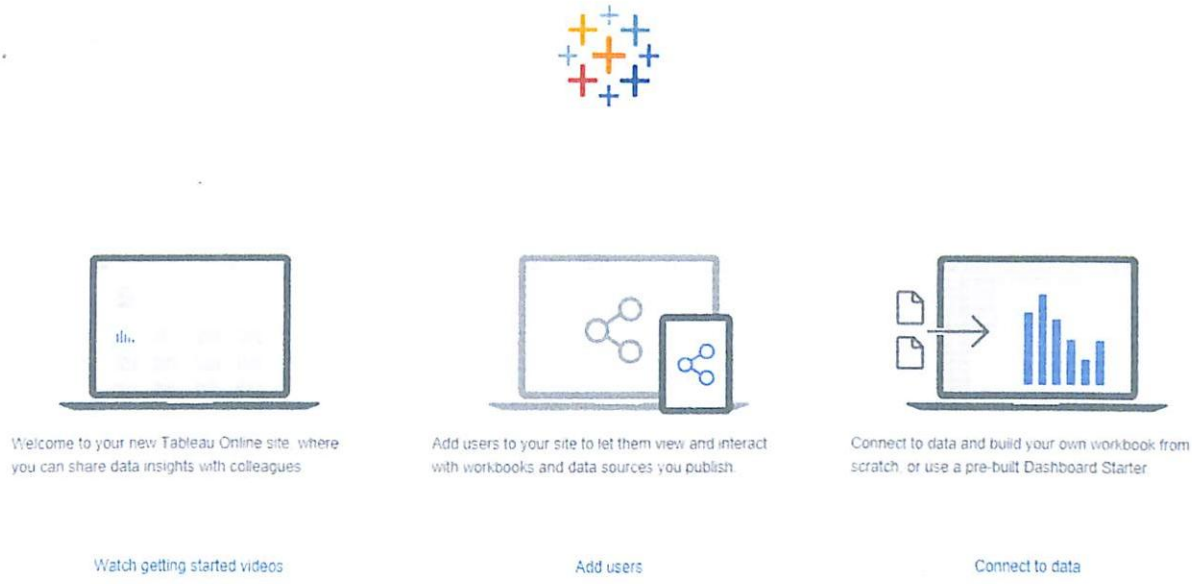


Figure B.2: Tableau start window

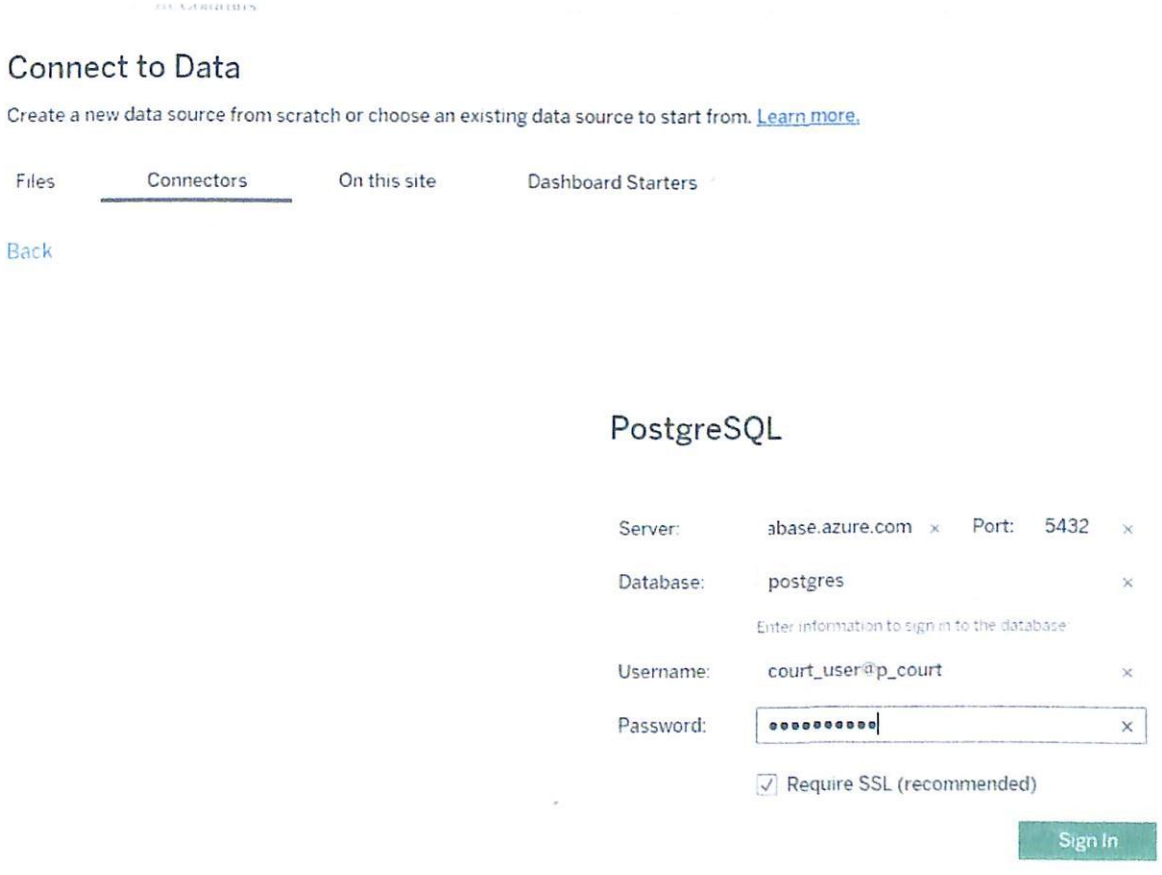


Figure B.3: Settings for connecting to a remote database

View Data: cases

1000 rows

case_date	case_type_id	category_id	court_id	first_instance_case_id		id	in
2013-07-13	3	138	69	null		142320233	
2015-02-16	3	33	129	null		142320239	
2013-08-25	3	20	276	null		142320245	
2013-12-20	3	41	163	null		142320252	
2016-04-12	3	132	171	null		142320258	
2016-11-01	3	144	187	null		142320263	
2012-03-16	3	13	5	null	null	142320270	
2015-11-17	3	100	20	null		142320277	
2013-07-28	3	120	46	null		142320283	
2012-08-01	3	28	358	null		142320290	
2012-03-05	3	145	250	null		142320295	
2012-12-02	3	78	158	null		142320301	

Figure B.4: The cases table data

View Data: case_types

3 rows

description	id
Гражданское дело	1
Уголовное дело	2
Административное дело	3

Figure B.5: The cases_types table data

cases+ (postgres)

cases | case_types

Sort fields Data source order

cases id	cases case_type_id	cases case_date	cases instance_id	cases category_id	cases court_id	cases result_type_id	case_types id (case_types)	case_types description
142320233.00	3.00000	13.07.2013	1.00000	138.000	69.000	1.00000	3.00000	Административное дело
142320239.00	3.00000	16.02.2015	1.00000	33.000	129.000	1.00000	3.00000	Административное дело
142320245.00	3.00000	25.08.2013	1.00000	20.000	276.000	1.00000	3.00000	Административное дело
142320252.00	3.00000	20.12.2013	1.00000	41.000	163.000	1.00000	3.00000	Административное дело
142320258.00	3.00000	12.04.2016	1.00000	132.000	171.000	1.00000	3.00000	Административное дело
142320263.00	3.00000	01.11.2016	1.00000	144.000	187.000	1.00000	3.00000	Административное дело
142320270.00	3.00000	16.03.2012	1.00000	13.000	5.000	1.00000	3.00000	Административное дело
142320277.00	3.00000	17.11.2015	1.00000	100.000	20.000	1.00000	3.00000	Административное дело

Figure B.6: A set of data for analyzing data by type of cases

View Data: result_types

2 → rows

description	id
иск (заявление, жалоба) удо...	1
отказано в удовлетворении и...	2

Figure B.7: Data table result_types

result_types+ (postgres)

result_types | cases

Sort fields Data source order ▾ Show aliases

result_types id	result_types description	cases id (cases)	cases case_type_id	cases category_id	cases court_id	cases first_instance_case_id	cases result_type_id
1.00	иск (заявление, жалоба) удо...	32658392.00	2.00000	49.000	235.000	null	1.00000
1.00	иск (заявление, жалоба) удо...	32658398.00	2.00000	62.000	143.000	null	1.00000
1.00	иск (заявление, жалоба) удо...	32658405.00	2.00000	69.000	320.000	null	1.00000
1.00	иск (заявление, жалоба) удо...	32658412.00	2.00000	165.000	160.000	null	1.00000
1.00	иск (заявление, жалоба) удо...	32658419.00	2.00000	150.000	136.000	null	1.00000
1.00	иск (заявление, жалоба) удо...	32658426.00	2.00000	116.000	268.000	null	1.00000
1.00	иск (заявление, жалоба) удо...	32658434.00	2.00000	32.000	42.000	null	1.00000
1.00	иск (заявление, жалоба) удо...	32658440.00	2.00000	137.000	178.000	null	1.00000
1.00	иск (заявление, жалоба) удо...	32658447.00	2.00000	44.000	367.000	null	1.00000
1.00	иск (заявление, жалоба) удо...	32658454.00	2.00000	49.000	241.000	null	1.00000

Figure B.8: A set of data for analyzing data by type of cases

View Data: regions

17 rows

description	id
город Астана	2
город Алматы	3
Акмолинская область	4
Актюбинская область	5
Алматинская область	6

Figure B.9: A set of data for analyzing data by type of cases

View Data: courts

383 → rows

description	region_id	id
Ерейментауский районный су...	4	46
Карагандинский областной суд	11	244
Сандыктауский районный су...	4	52
Целиноградский районный су...	4	53
Верховный Суд Республики К...	2	1
Алматинский районный суд г...	2	2
Сарыаркинский районный су...	2	3
Районный суд №2 Алматинск...	2	4
Районный суд №2 Сарыарки...	2	5
Специализированный межра...	2	6
Специализированный межра...	2	7
Специализированный межра...	2	8

Figure B.10: Courts table data

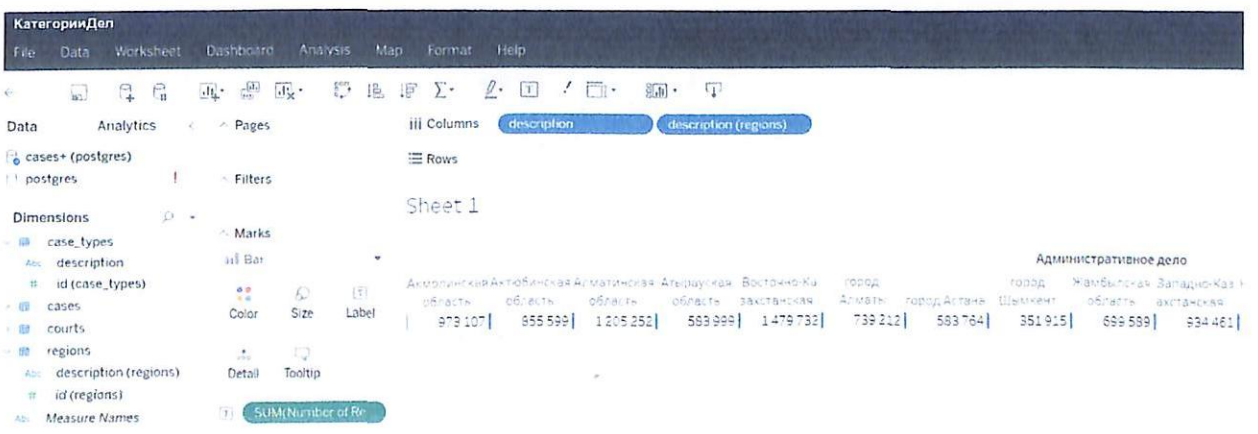


Figure B.11: Set up a data set for export

View Data - Google Chrome

https://eu-west-1a.online.tableau.com/vizql/t/kazakhstancourt/w/_0/v/She

Summary Full data

Showing first 51 rows.
Download all rows as a text file

description	description (regions)	SUM(Number of Records)
Административное дело	Туркестанская область	895 651
Административное дело	Северо-Казахстанская область	973 521
Административное дело	Павлодарская область	816 877
Административное дело	Мангистауская область	585 054
Административное дело	Кызылординская область	622 903
Административное дело	Костанайская область	1 128 432
Административное дело	Карагандинская область	1 439 691
Административное дело	Западно-Казахстанская область	934 461
Административное дело	Жамбылская область	699 589
Административное дело	город Шымкент	351 915
Административное дело	город Астана	583 764
Административное дело	город Алматы	739 212
Административное дело	Восточно-Казахстанская область	1 479 732

Figure B.12: Data Export

View Data - Google Chrome

https://eu-west-1a.online.tableau.com/vizql/t/kazakhstancourt/w/_0/v/She

Summary Full data

Showing first 51 rows.

Download all rows as a text file

description	description (regions)	SUM(Number of Records)
Административное дело	Туркестанская область	895 651
Административное дело	Северо-Казахстанская область	973 521
Административное дело	Павлодарская область	816 877
Административное дело	Мангистауская область	585 054
Административное дело	Кызылординская область	622 903
Административное дело	Костанайская область	1 128 432
Административное дело	Карагандинская область	1 439 691
Административное дело	Западно-Казахстанская область	934 461
Административное дело	Жамбылская область	699 589
Административное дело	город Шымкент	351 915
Административное дело	город Астана	583 764
Административное дело	город Алматы	739 212
Административное дело	Восточно-Казахстанская область	1 479 732

Figure B.13: Data prepared for export

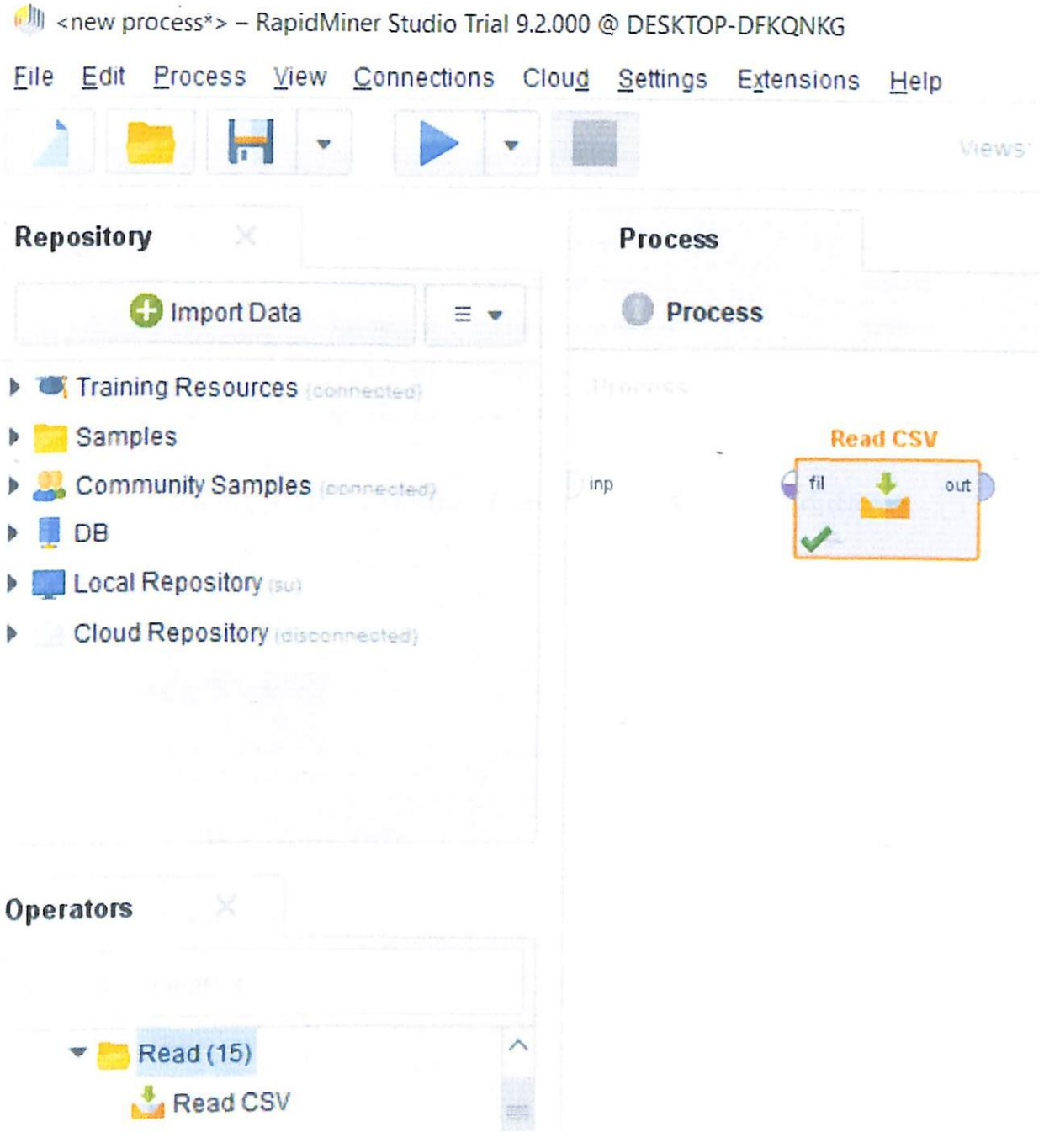


Figure B.14: Selecting a tool to import a file with a set of data

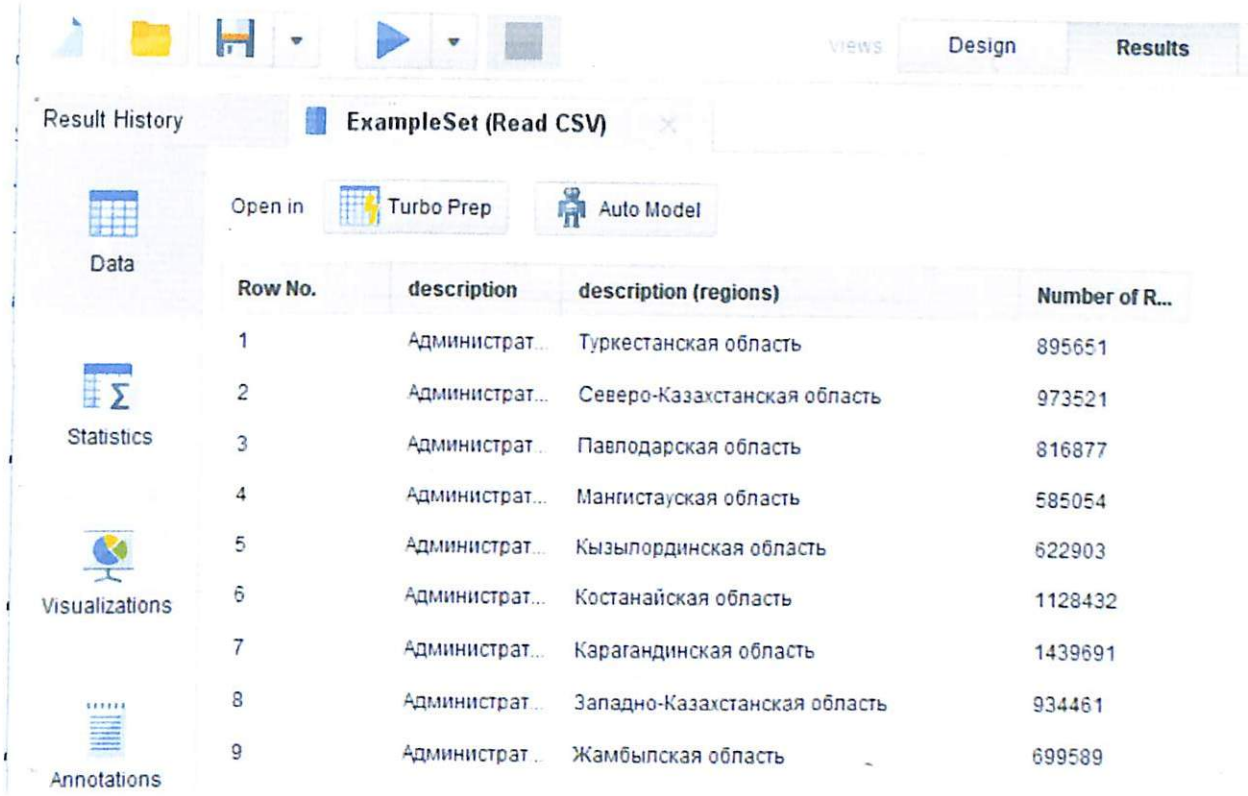


Figure B.15: Results of processing imported data

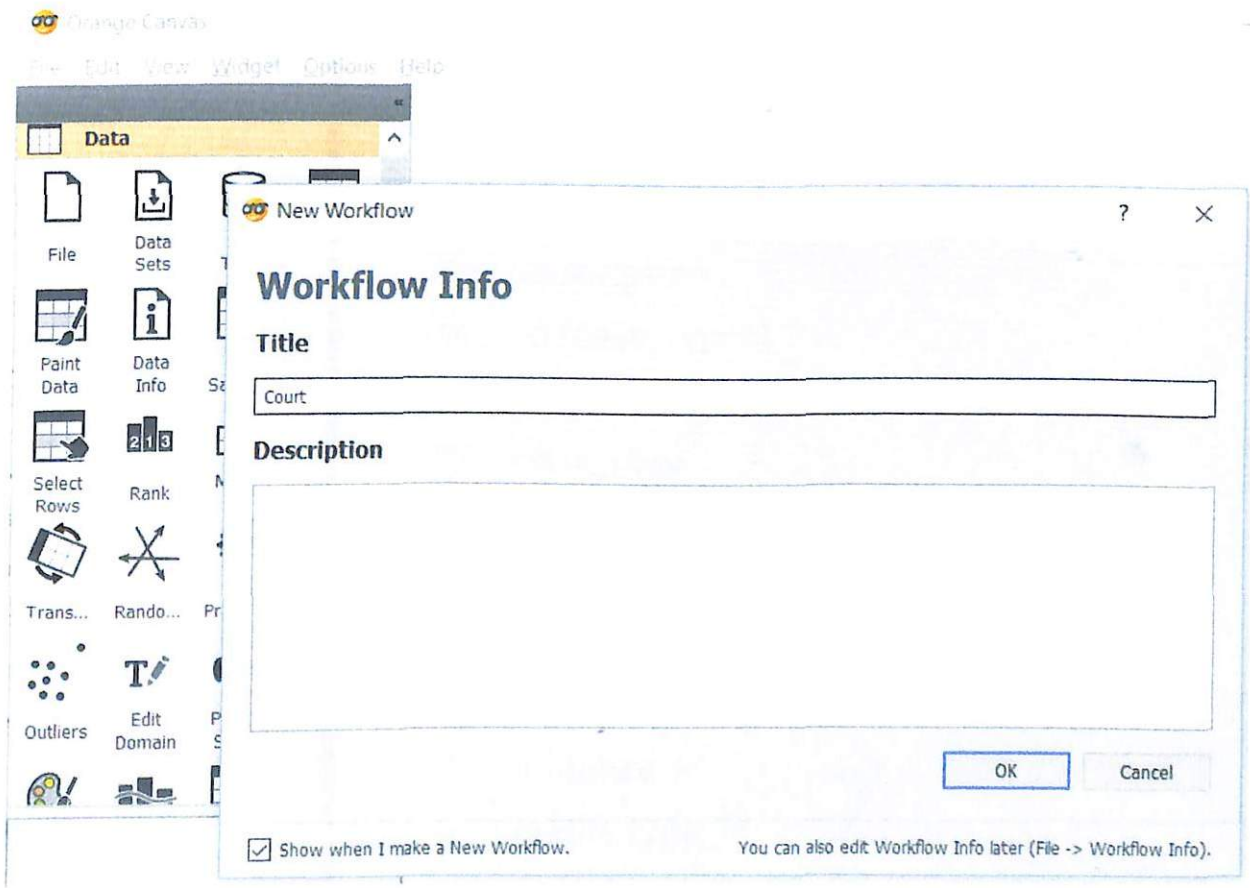


Figure B.16: Creating a new Orange project

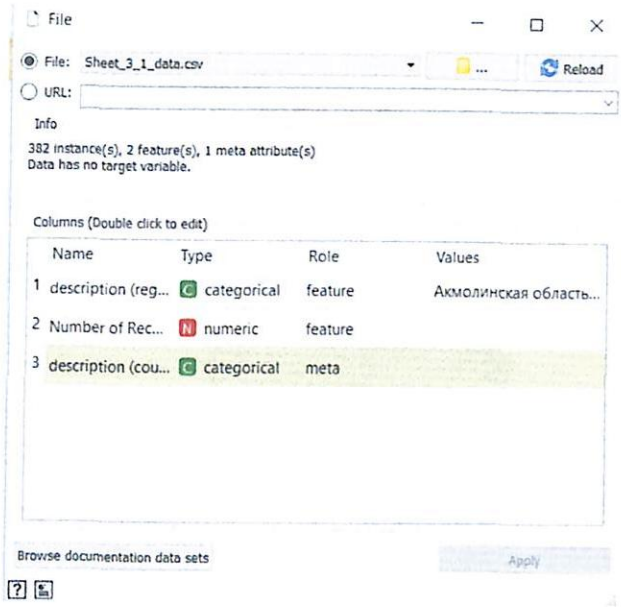


Figure B.17: Setting the File widget

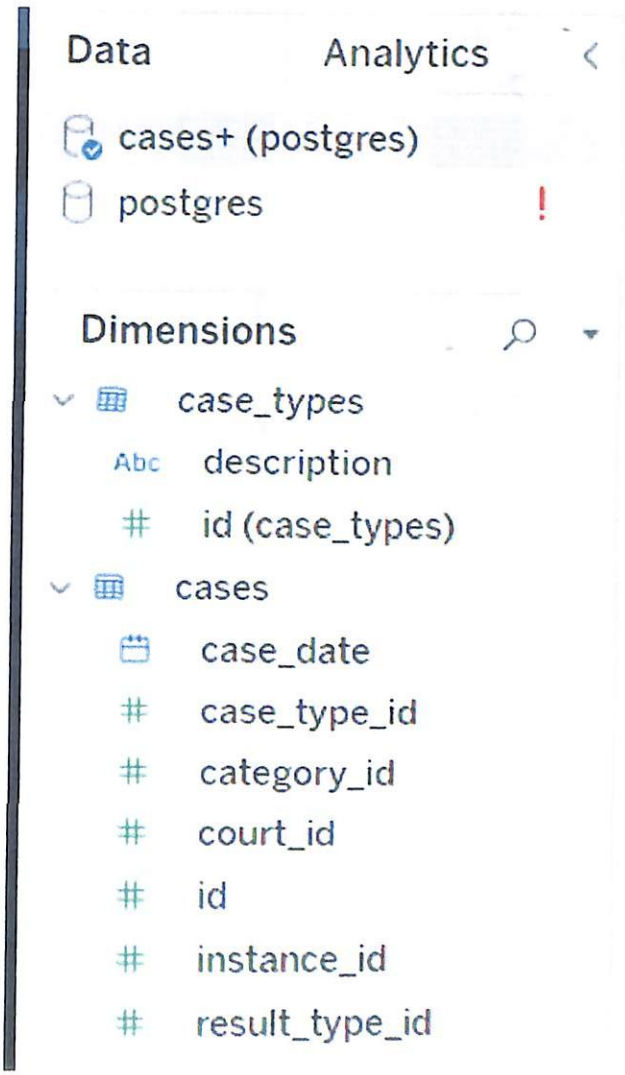


Figure B.18: Selecting dimensions from the provided data set

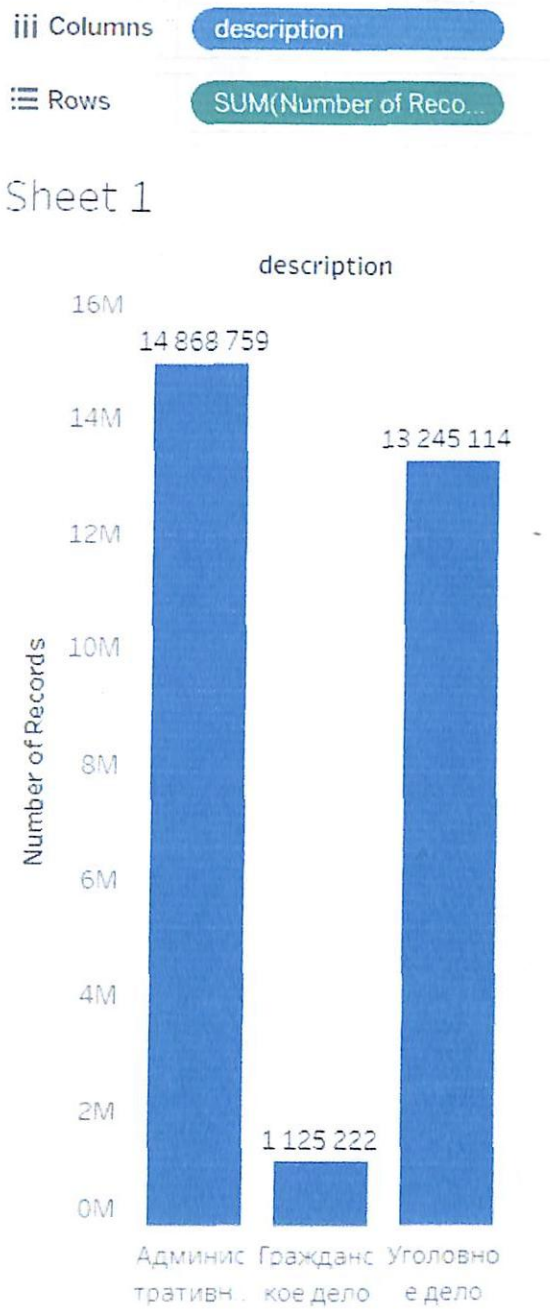


Figure B.19: Selection of fields for visualization

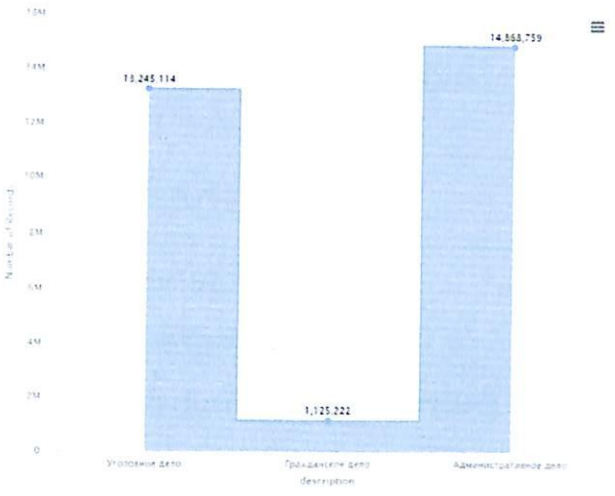


Figure B.20: Visualization of data on cases heard in court RapidMiner

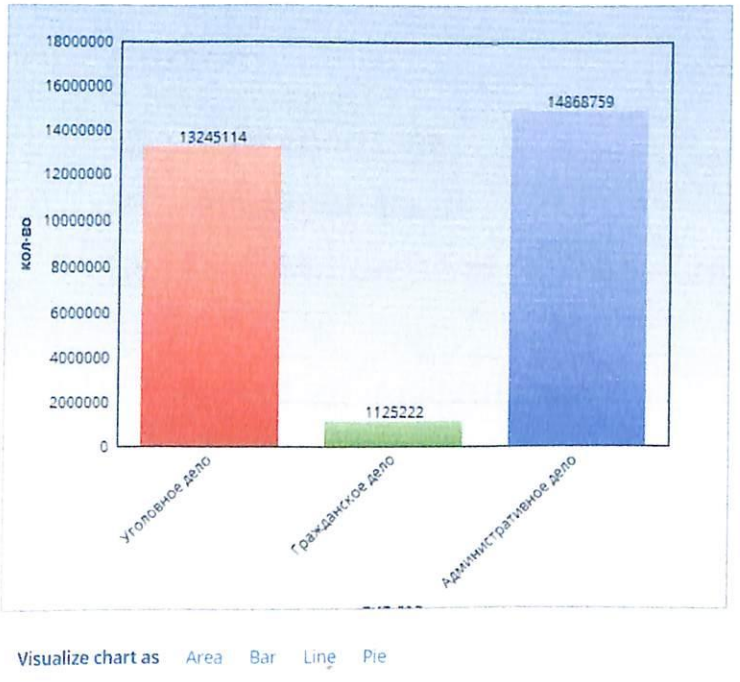


Figure B.21: Visualization of data on cases heard in court Orange

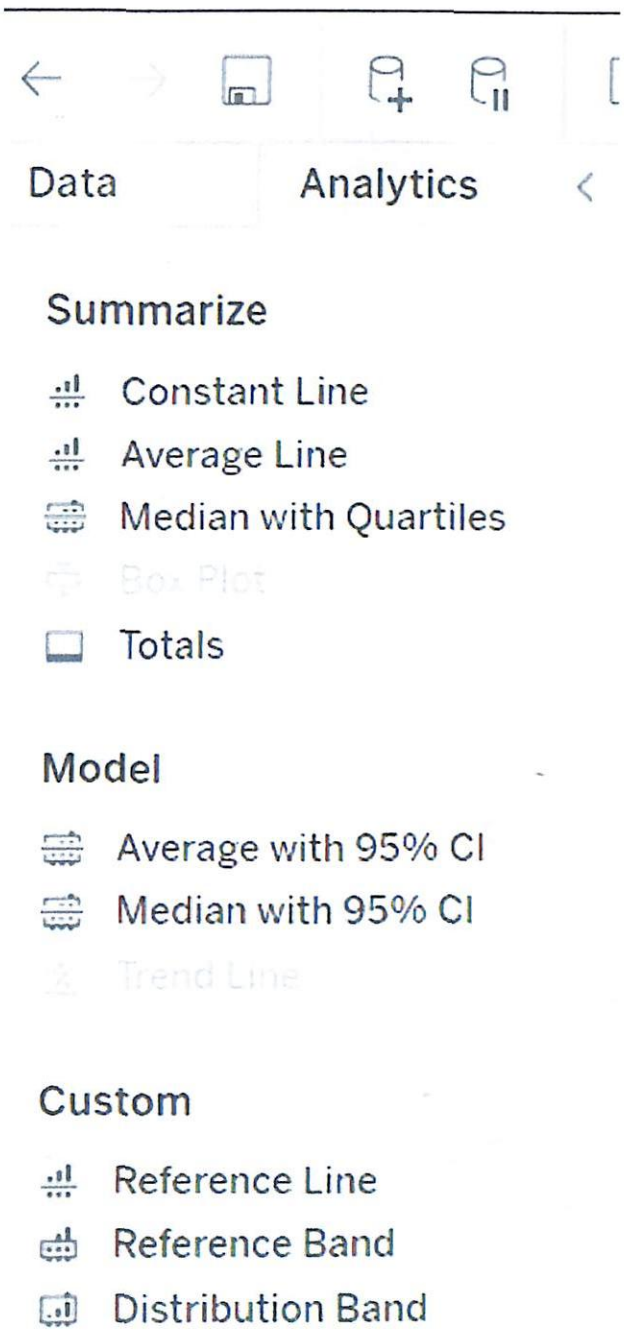


Figure B.22: Analytics tab

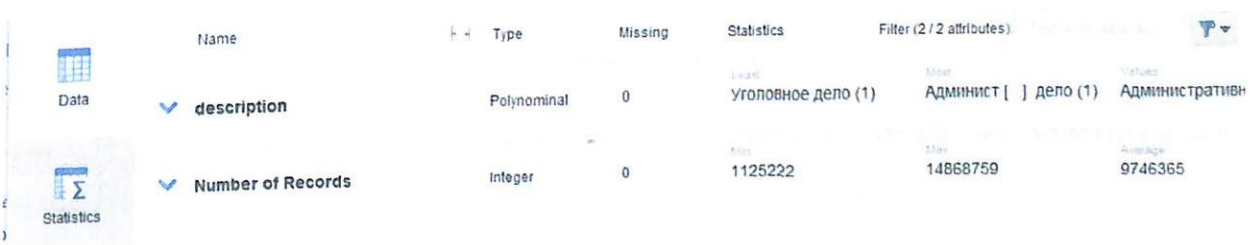


Figure B.23: Statistics Tool in RapidMiner

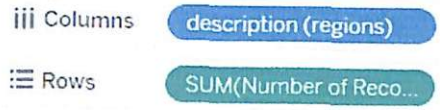


Figure B.24: Selection of fields for visualization

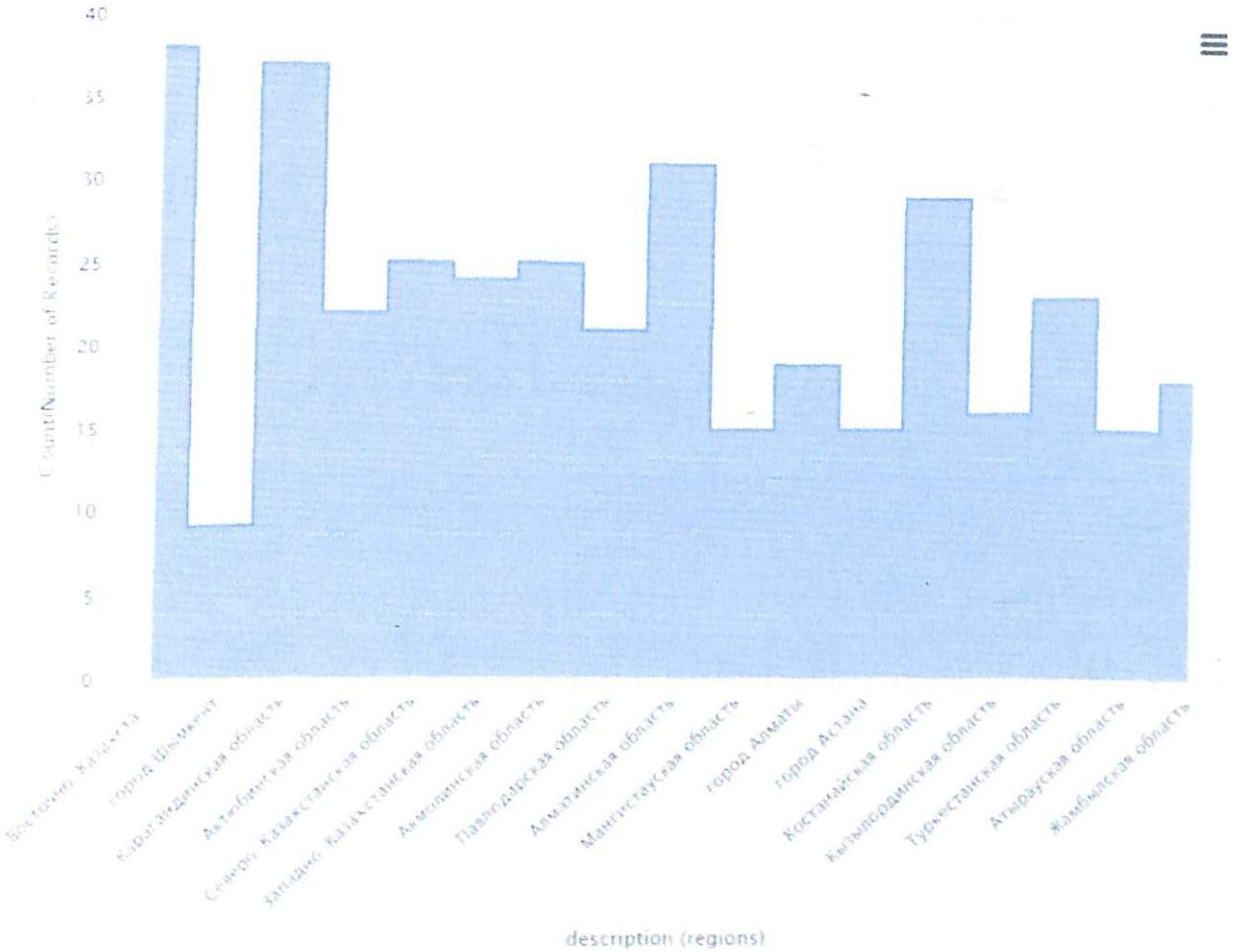


Figure B.25: Visualization of data on cases heard in court by regions of Kazakhstan RapidMiner

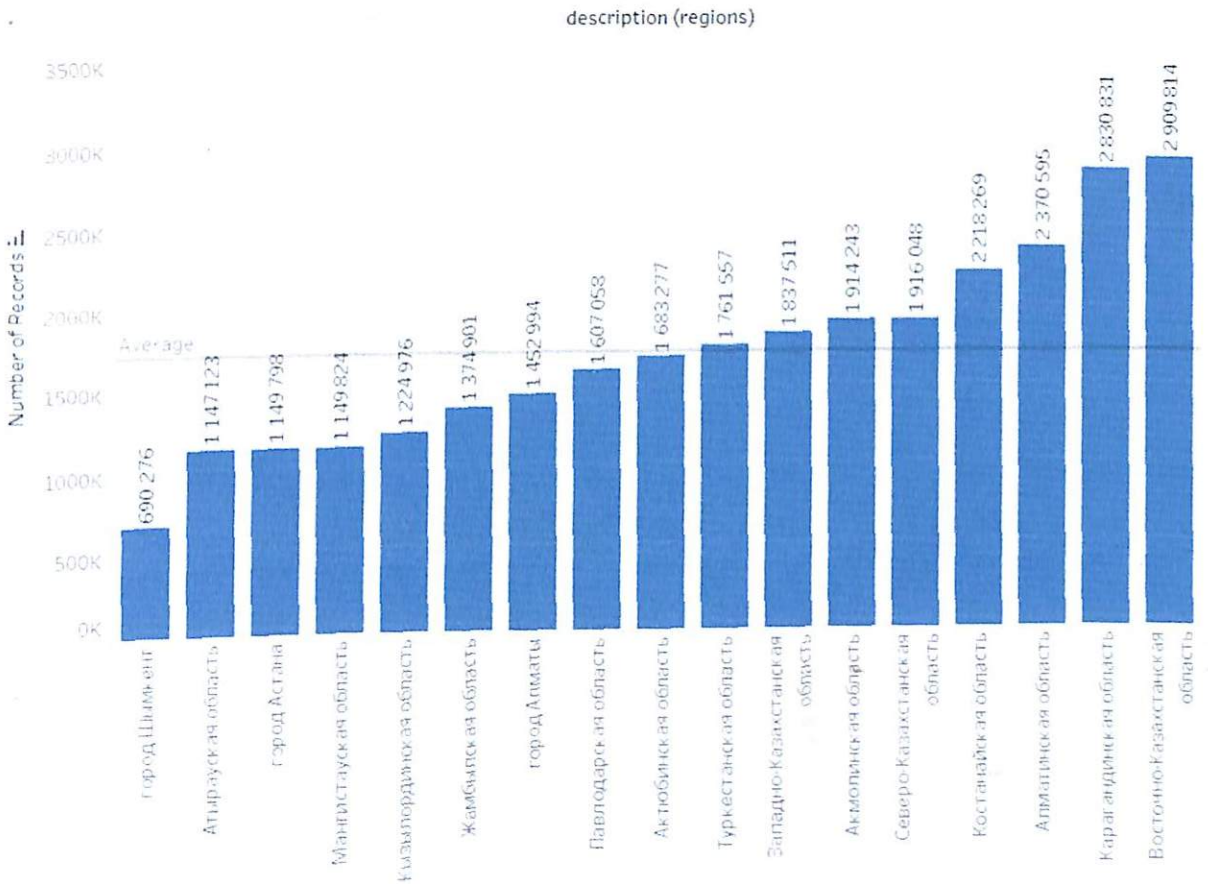


Figure B.26: Sorting data on the chart

Figure B.27: Tools Sorting RapidMiner Data

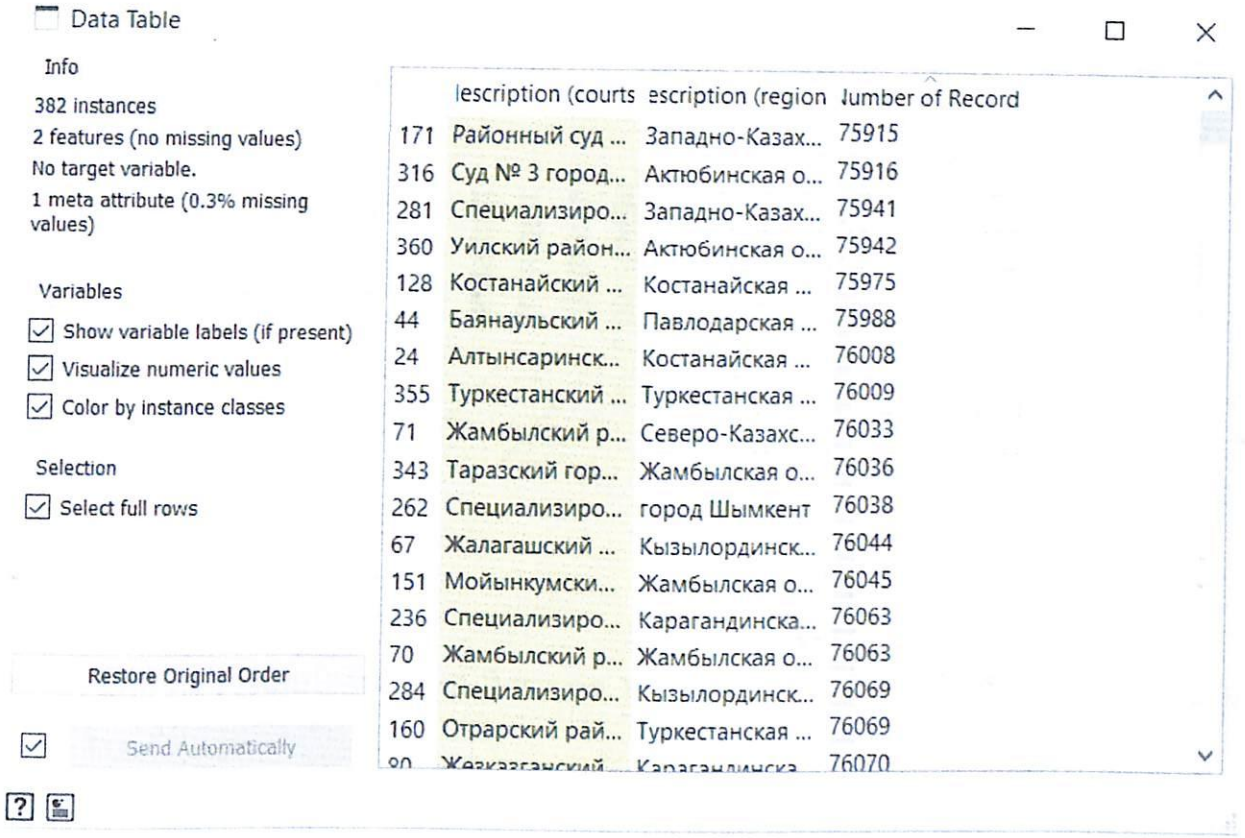


Figure B.28: Sorting Orange Data

Figure B.29: Selection of fields for visualization of aggregated data

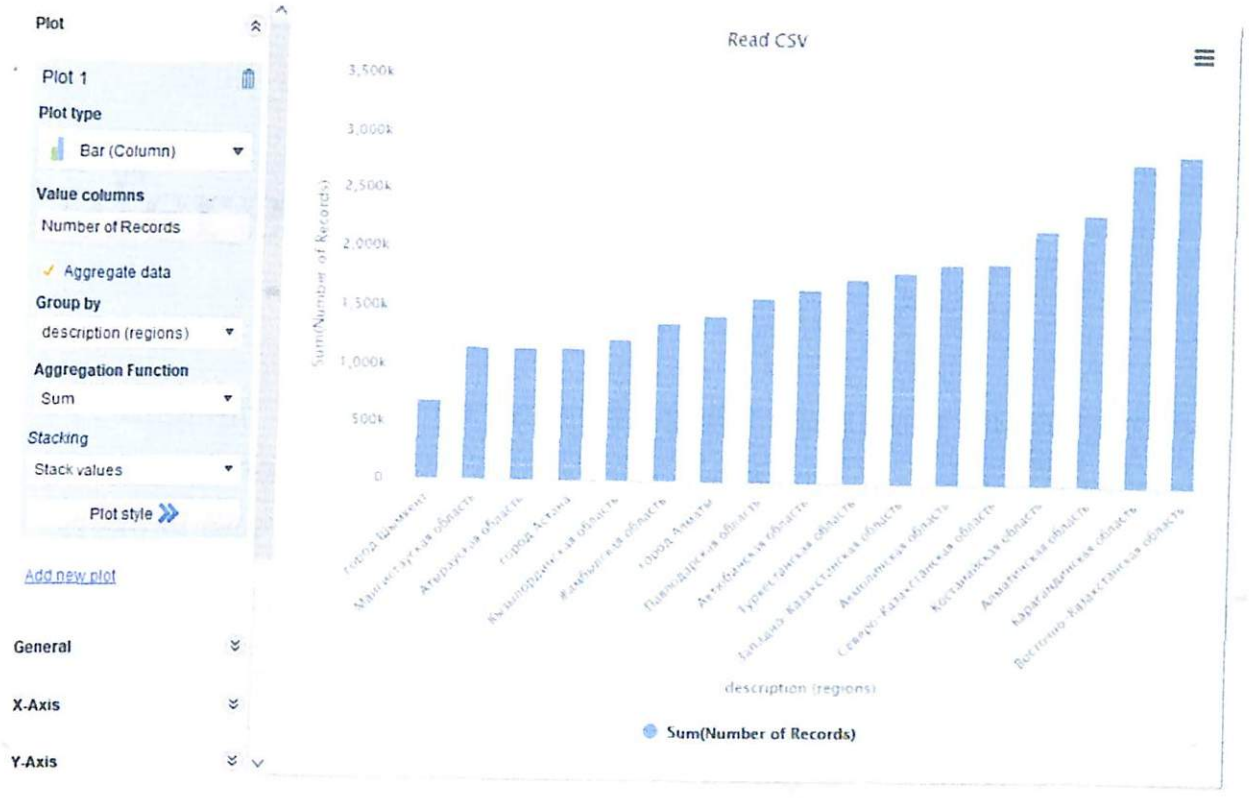


Figure B.30: Visualization of data on the court in the regions of Kazakhstan

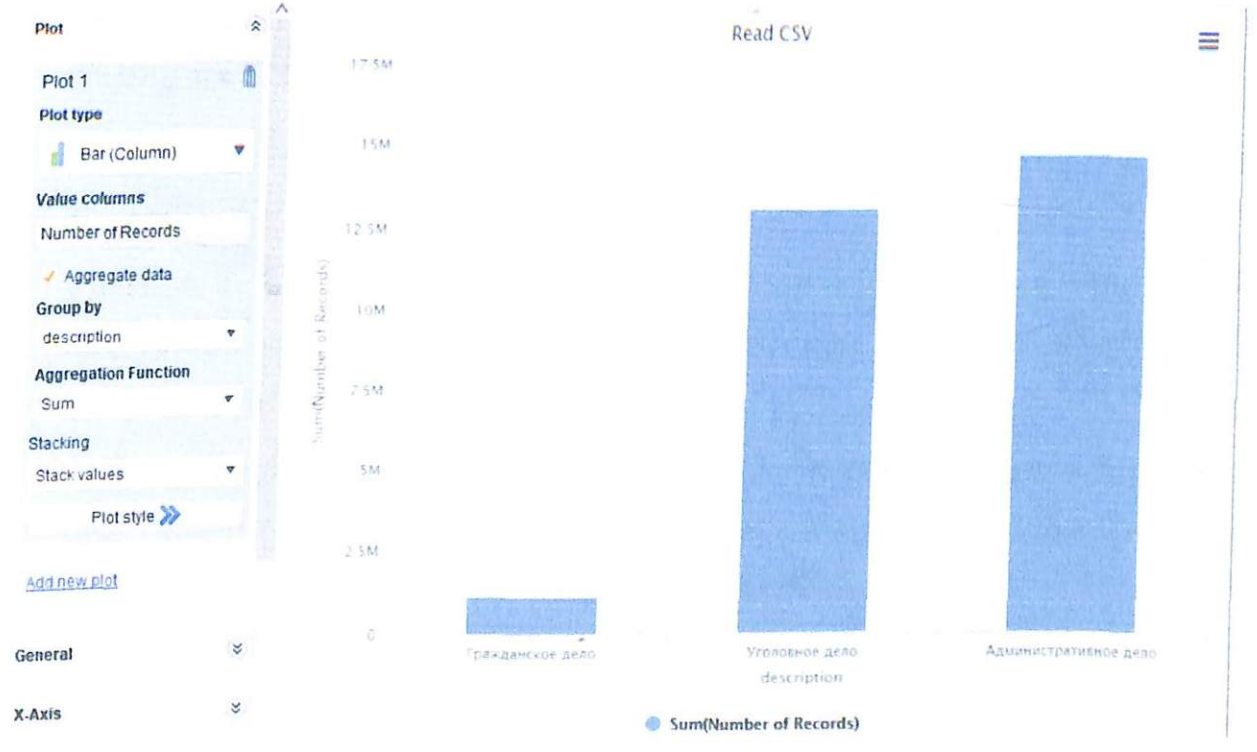


Figure B.31: Visualization of data on cases heard in court by type of case

References

[1] T. Siddiqui and M. Al Kadri. «Big data analytics on the cloud». In: *International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)* (2015), pp. 61–66.

[2] Robert W. Sebesta. *Concepts of programming languages.* Pearson Education, Inc, 2009, p. 21.

[3] C. L. P. Chen and C. Zhang. «Data-intensive applications, challenges, techniques and technologies : a survey on big data». In: *Information Sciences* 275 (2014), pp. 314–347.

[4] G. Varoquax F. Pedregosa. «Scikit-learn: machine learning in python». In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[5] M. Aivazis K. J. Millman. «Python for scientists and engineers». In: *Computing in Science & Engineering* 13.2 (2011), pp. 9–12.

[6] C. Ebert P. Louridas. «Embedded Analytics and Statistics for Big Data[J]». In: *Computing in Science & Engineering* 30.6 (2013), pp. 33–39.

[7] Patricia J. Crossno John R Harger. «Comparison of open-source visual analytics toolkits». In: 8294 (2012). DOI: 10.1117/12.911901. URL: https://doi.org/10.1117/12.911901.

[8] F. Perez and B. E. Granger. «IPython: A System for Interactive Scientific Computing». In: *Computing in Science Engineering* 9.3 (May 2007), pp. 21–29. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.53.

[9] T. E. Oliphant. «Python for Scientific Computing». In: *Computing in Science Engineering* 9.3 (May 2007), pp. 10–20. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.58.

[10] D. Rotolo and L.Leydesdorff. «Matching medline / pubmed data with web of science: a routine in r language». In: *Journal of the Association for Information Science and Technology* 66.10 (2015), pp. 2155–2159.

[11] N. Matloff. *Art of r programming.* No Starch Press, Inc., 2011, p. 373.

[12] D. Toomey. *R for Data Science.* Packt Publishing, 2014, p. 347.

[13] E.Derclaye D. Gervais. «The scope of computer program protection after sas : are we closer to answers ?» In: *European Intellectual Property Review* 34 (2012), pp. 565–572.

[14] D. Sarkar. *Microsoft Sql Server 2012 with Hadoop.* Packt Publishing, 2013, p. 83.

[15] P. Chandarana and M. Vijayalakshmi. «Big Data analytics frameworks». In: (Apr. 2014), pp. 430–434. DOI: 10.1109/CSCITA.2014.6839299.

[16] Alfredo Cuzzocrea, Il-Yeol Song, and Karen C. Davis. «Analytics over Large-scale Multidimensional Data: The Big Data Revolution!» In: DOLAP '11 (2011), pp. 101–104. DOI: 10.1145/2064676.2064695. URL: http://doi.acm.org/10.1145/2064676.2064695.

[17] Boduo Li et al. «A Platform for Scalable One-pass Analytics Using MapReduce». In: SIGMOD '11 (2011), pp. 985–996. DOI: 10.1145/1989323.1989426. URL: http://doi.acm.org/10.1145/1989323.1989426.

[18] Shyam R. et al. «Apache Spark a Big Data Analytics Platform for Smart Grid». In: *Procedia Technology* 21 (2015). SMART GRID TECHNOLOGIES, pp. 171–178. ISSN: 2212-0173. DOI: https://doi.org/10.1016/j.protcy.2015.10.085. URL: http://www.sciencedirect.com/science/article/pii/S2212017315003138.

[19] Janez Demšar et al. «Orange: Data Mining Toolbox in Python». In: *Journal of Machine Learning Research* 14 (2013), pp. 2349–2353. URL: http://jmlr.org/papers/v14/demsar13a.html.

[20] Ralf Klinkenberg Eds. Markus Hofmann. *RapidMiner Data mining use cases and business analytics applications.* CRC Press, 2016, p. xx.

[21] S. Dwivedi, P. Kasliwal, and S. Soni. «Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime)». In: (Mar. 2016), pp. 1–8. DOI: 10.1109/CDAN.2016.7570894.

[22] L. Zhang et al. «Visual analytics for the big data era — A comparative review of state-of-the-art commercial systems». In: (Oct. 2012), pp. 173–182. DOI: 10.1109/VAST.2012.6400554.

[23] Sarah Anne Murphy. «Data Visualization and Rapid Analytics: Applying Tableau Desktop to Support Library Decision-Making». In: *Journal of Web Librarianship* 7.4 (2013), pp. 465–476. DOI: 10.1080/19322909.2013.825148. eprint: https://doi.org/10.1080/19322909.2013.825148. URL: https://doi.org/10.1080/19322909.2013.825148.

[24] K. McCormick and J. Salcedo. *SPSS statistics for dummies.* John Wiley & Sons, Inc, 2015, p. 370.