NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science

Bachelor's Programme in Data Science and Business Analytics

TERM PAPER

Research Project

"Interactive digital technology dashboard"

Prepared by the student of Group 191, in Year: 3,

Ivanova Daria Nikolaevna

Term Paper Supervisor:

Leading expert, PhD, Privorotskaya Sofia Grigorievna

Moscow

2022

Contents

1. Abstract

2. Key words

3. Introduction

3.1 Description of subject area

3.2 Relevance of the problem

3.3 Project tasks and objectives

3.4 Main result

4. Literature review

4.1 Comparative analysis of sources and Analogous materials

5. Theoretical material

5.1 Key terms and definitions

5.2 Overview of the situation now in AI analytics

5.3 Measuring digital technologies

5.4 Data visualization

6. Methodology

6.1 Concept of the dashboard

6.2 Data search and selection

6.3 Choosing the BI instrument

6.4 Choosing the diagram types for the data

6.5 Structure of the dashboard

7. Technical part

7.1 Key instruments

7.2 Data processing and storage

7.3 Implementation in BI instruments

8. Description of the dashboard

8.1 Overview

8.2 Research in AI

8.3 Publications and patent activity in Russia

8.4 AI players and activities

8.5 AI talent

8.6 Investments in AI

8.7 Investments in startups

9. Building a model

9.1 The aim of the model

9.2 Dataset description

9.3 Data Preparation and Implementation

9.4 Result description

10. Conclusion

11. Bibliography

12. Appendix

1. Abstract

As a result of my research, I was able to examine the field of Artificial intelligence and its applications, collect data, process, prepare, and organize it for subsequent visualizations using Tableau. My goal was to present AI data in a format suitable for business user decision making. A comprehensive picture was formed using different types of analysis: publication and patent analysis, rankings, indices, venture market analytics, and different sources: Eurostat, OECD, European Commission, MacroPolo. I published a story including seven pages, each of which discusses a different aspect of

Artificial intelligence and provides concise yet detailed information. Overall, I concluded that there is a data shortage for Artificial intelligence and AI related technologies, as well as a lack of end-to-end analytics and AI data, as well as a general framework for AI data. The regression model was built as part of the other component of my work. The AI-related startup or company's overall investment amount is predicted by the model. The model performed well overall and may be used by a business to obtain the expected fundings.

2. Key words

AI technologies
Artificial intelligence
BI systems
Data analysis
Data visualization
Digital technologies
Digital transformation

3. Introduction

3.1 Short description of subject area

Artificial intelligence is an ability of an intelligent system to perform those functions and tasks that are usually characteristic of intelligent beings. This may be a manifestation of some kind of creative abilities, a tendency to reason, generalization, learning based on previous experience. Over the past ten years, the number of annual publications on Artificial intelligence increased more than twice, and investments in AI increased more than 20 times. The world is changing fast and new technologies appear daily in order to be applied in various fields and for a huge value of needs. Artificial intelligence now can be used in every company, by an individual, or by the government. It simplifies many work processes, a lot of

which can be done using AI. AI technologies are applicable instruments in many industries. “Mobility and autonomous vehicles” industry has invested over $60 billion in Artificial intelligence over the past five years, “Healthcare, drugs and biotechnology” and “Media, social platforms, marketing” industries have invested over $40 billion.
Artificial intelligence is a very wide sphere. There are several technology groups: Computer vision, Electronic component base of AI, NLP, perspective models, recommendation system and intelligent decision support systems, speech recognition and synthesis. AI includes scientific theories as well as specific technological practices, i.e. creating programs that behave like a human would do, but n times faster. The goal in the development of AI is to simplify the execution of tasks that are based on a large number of variable factors, are not easy to understand, imply a complex solution and are quite difficult to algorithmize manually.

3.2 Relevance of the problem

Artificial intelligence is the foundation of a new generation of digital technologies, and it's the base of digital transformation in nearly every industry. The current stage in the development of Artificial intelligence is associated with the development and adaptation of AI products and services for a wide range of applied tasks. Automation of regular jobs, the development of new work formats, the introduction of new business models, and the exploration of new market niches are all conceivable with AI-based solutions.
The potential of widespread acceptance depends on enterprises' willingness to modify not only the technology foundation, but also business processes and data culture. AI is the new area for statistical accounting. Practice of receiving and using data is being developed. The development of AI is impossible in the expected future without comprehensive assessment.

3.3 Project tasks and objectives

The aim of the project is to analyze the field of AI, to create an analytical base for decision-making in the field of AI and use BI instruments to illustrate the key trends in different AI-related spheres.

1. Steps of analysis:
- Choose the main topic
- Define the target audience for the dashboard
- State the value for future users
- Explore the relevant resources, data and reports
- Create the structure and logic of the report
- Explore the platforms and the ways a report may be created and decide on the system in which the report will be done.

2. The draft/plan of dashboards – a graphic file, representing the draft visualization of future dashboards should be done

3. Select the content for dashboards (quantitative metrics, lists of promising developments and leading competence centers, etc.)

4. Develop and publish interactive dashboards

5. Defend the concept and present the work of online dashboards.

3.4 Main results

As a result of the project the dashboard was made with the help of Tableau service and published online. The dashboard consists of 7 pages that include a wide range of information on AI such as Research in AI, Investments, Technology groups etc.

All the project tasks were done:

- Concept of the dashboard was created and worked out
- The draft and plan of the dashboard were made
- The data sources and analogous materials were analyzed

- Data was collected, processed and analyzed
- Dashboard was developed

4. Literature review

4.1 Comparative analysis of sources and Analogous materials

It is essential for the future progress in AI technologies and services to develop and maintain AI-related services and portals, collect AI data in a format suitable for comprehensive analysis, building analytical models and decision making. As I mentioned before, AI is the next stage in the development in statistical accounting. Data reception and use practice are being developed, and it is important to keep track of the progress made not only by publishing papers, but also by creation of a single database of AI-related information. AI development will be impossible in the conceivable future without a thorough assessment. There are several major bases of AI knowledge that I would like to comment on, also dashboards that represent key trends in AI can be found there.

OECD.AI

OECD. AI brings together resources from the OECD, its partners, and all stakeholders. In the areas where AI has the most influence, the OECD.AI encourages conversation amongst stakeholders while also providing multidisciplinary, evidence-based policy analysis.This knowledge base[1] may be considered as a useful one.

There is a dashboard with live data on AI. There are several sections that include a variety of graphs.

First of all, I would like to mention that this report has live data and updates daily, which is a very good point for such a source of information.

[1] https://oecd.ai/en/

The first page “AI news” shows the real-time news on the AI topic and the country it was released at, which I find quite interactive and engaging. The next pages “AI research” and “Investments in AI” have pretty much data visualizations, which is a good point. One more good point here is that there is a variety of settings and filters so one can adjust the layout for himself.
There are also several problems regarding the selected methods of the data representation: the line charts are hard to read somewhere because of the amount of the lines and because it is not a BI system, when pointing on a specific area (point on a graph), tooltips are rarely shown, which makes the graph unreadable. Moreover, some of the graphs are not representable as they have too much information/lines/data points on them. Also in each section there are over 10 subsections, which make it harder to use and I feel like some of them may be combined.
I would say that this AI knowledge base establishes a hub for the gathering and exchange of AI evidence.
Live data report is a good one because of the amount of the data it contains and the overall structure of the pages and dashboards, but as in the previous one the data representation in some places are not satisfying. I would suggest remaking this report using BI system and I'm pretty sure it will be more user-friendly and demonstrative.

AI WATCH

AI Watch[2] tracks industrial, technological, and research capabilities, legislative initiatives in Member States, AI adoption and technical breakthroughs, and their influence on the economy, society, and public services. It contains a variety of analyses that are required to track and aid the implementation of the European AI Strategy.
AI watch contains a big number of research publications, useful topics and tools to investigate and analyze the AI sphere. There are also several data sets, news

[2] https://ai-watch.ec.europa.eu/

and events sections. There is an AI Watch landscape dashboard. The AI watch has 7 pages with the graphs. The first one is the most useful one in my opinion.
As we open the first page of the dashboard we see that there are 2 unrepresentative pie charts. I gave them these characteristics as it is difficult to read the pie chart when it has more than 5 sections (and there exists a rule not to use pie charts for more than 5 fields or to unite some in the "Other" section). Also the titles of the sections are not readable. The map is also not quite representative, there's no explanation of what the colors mean or data labels. All the barcharts are ok even though they lack data labels. The last graph on the first page (Revealed Comparative Advantage) I didn't quite understand as all the values there are the same. All the other pages of the dashboard have the same problems with the data representation.
Overall, I can say that AI Watch is a good service that provides data and information on the AI topic, there are many useful links, sources and publications.

5. Theoretical material

5.1 Key terms and definitions

Artificial intelligence[3] is a complex of technological solutions that allow simulating human cognitive functions (including self-learning and search for solutions without a predetermined algorithm) and get when executing specific tasks results comparable, at least with the results of intellectual human activities. It covers information and communication infrastructure, software (including those that use machine learning methods), data processing processes and services and finding solutions.

Nowadays, there are 7 types of AI[4]:

[3] Развитие отдельных высокотехнологичных направлений. Белая книга. НИУ ВШЭ
[4] https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-Artificial-intelligence/

1. **Reactive machines**. AI systems that do not have memory and they are able to solve only certain tasks. They cannot form memories and use previous experiences to perform their functions.
2. **Limited memory.** These are systems with memory based on experience. But this experience is not being stored or compiled in the AI Information Library.
3. **Theory of mind**. These types of systems can understand human emotions and intentions, they also have social intelligence and are able to participate in teamwork.
4. **Self-awareness.** AI systems of this type form an idea of themselves, and therefore they completely imitate human intelligence.
5. **Artificial Narrow Intelligence (ANI).** ANI is the use of Artificial intelligence (AI) technology to develop a versatile system that mimics and possibly surpasses human intelligence for a single purpose.
6. **Artificial General Intelligence (AGI)**. AGI systems will be able to build various competences on their own, as well as make linkages and generalizations across domains, significantly reducing training time.
7. **Artificial Superintelligence (ASI)**. ASI will have a level of intelligence that is higher than that of an individual or the entire human race. Currently, all expressions of Artificial intelligence are restricted.

Technologies that are primarily used to create AI products and services[5]:

1. **Machine learning.** Machine learning is the ability of a computer with an AI system to make decisions based on the results of data processing, without adhering to clear patterns and rules.
2. **Deep learning.** It is a subsection of machine learning that allows one to detect patterns in Big Data. Data processing in deep learning technology is carried out by Artificial neural networks (ANNs), created by analogy with biological neural networks.

[5] Развитие отдельных высокотехнологичных направлений. Белая книга. НИУ ВШЭ

3. **Natural language processing (NLP)**. Compiling software for transforming any data into a natural language that a computer understands and uses to respond to a person.
4. **Computer vision**[6] allows computers and systems to extract useful information from digital photos, videos, and other visual inputs, as well as execute actions or make recommendations based on that data. If Artificial intelligence allows computers to think, computer vision allows them to see, watch, and comprehend.
5. **Speech recognition and synthesis**[7]. Speech recognition is the process of automatically transforming a voice signal into digital data (e.g. text data). Speech synthesis is the inverse problem, it is everything that is connected with the Artificial production of human speech.
6. **Intelligent support systems decision making**[8] is an automated computer network that performs objective data analytics by building a mathematical model of how events will unfold in the future. The goal of such solutions is to assist people who are faced with challenging management decisions.

Now there are two directions of AI development:

- Solving problems related to the approximation of specialized AI systems to human capabilities, and their integration, which is realized by human nature.
- The creation of Artificial intelligence, which would be the integration of already created AI systems into a single one, that will be capable of solving the issues of the society.

5.2 Overview of the situation now in AI analytics

The digitization of numerous sectors of the economy and human life includes changes to both nature and the amount of data. There is a rise in demand for statistical data. New opportunities are emerging due to the development of digital communications and the

[6] https://www.ibm.com/topics/computer-vision
[7] https://en.wikipedia.org/wiki/Speech_recognition
[8] https://fisgroup.ru/blog/fis-dss/

appearance of Big data suitable for statistical processing and analysis. Furthermore, statistical measurement of the digital economy's progress is required, but international standards in this field have not yet been established.[9]

There are several bases of knowledge that collect and organize AI publications and AI-related topics. For example, OECD.AI is one of the biggest bases that process the data on AI.

5.3 Measuring digital technologies[10]

Digital transformation is a set of qualitative changes in business processes or ways of carrying out economic activities as a result of the introduction of digital technologies, leading to significant socio-economic effects. The base of digital transformation is an ecosystem of interdependent digital technologies, the constant development of which stimulates economic and social changes.

One of the criteria for assessing the achievement of digital transformation is the level of digital maturity of industries and governments. Several methods for accessing the digital maturity are being developed for:

1) Comparing the achieved level with the target
2) Cross-industry comparison of the level of digital technology spread.

For example, OECD assesses the digital maturity of business sector organizations based on these factors[11]:

- ICT opportunities (training employees in digital skills, availability of ICT specialists, introduction of digital technologies);
- Advanced ICT functions (information security, business management software adaptation, own developments);
- Web maturity (availability of a website with the possibility of e-commerce, online advertising).

To assess the digital maturity indicators of the governments OECD use such indicators:

[9] Цифровая трансформация: Ожидание и реальность. Доклад НИУ ВШЭ
[10] Что такое цифровая экономика? Тренды, компетенции, измерение. Доклад НИУ ВШЭ
[11] https://oecd.ai/en/

1) Availability of digital platforms
2) Data use and storage
3) Openness
4) User control

For a complete assessment of digital maturity, indicators that characterize the use in industries of the most significant digital technologies and specialized software such as AI, Big Data analysis, cloud services etc. should be taken into account.

5.4 Data visualization

Data visualization goal is the realization of the main idea of information, this is what the selected data needs to be shown for, what effect needs to be achieved - identifying relationships in information, showing data distribution, composing or comparing data.Data visualization is the presentation of data in a form that will be the most convenient for a person to work with its analysis. Data visualization is widely used in scientific and statistical research: in forecasting, data mining, business analysis, in instructional design for teaching and testing, in summaries and survey results.
Visual information is better perceived and allows one to quickly and effectively convey the huge amounts of the data, own thoughts and ideas to the viewer. Physiologically, the perception of visual information is fundamental for a person. That is why it is so important now when it is the era of Big Data.

The success of visualization directly depends on the correctness of its application, namely on the choice of the type of graph, its correct use and design. The graph allows one to express the idea that the data carries in the most complete and accurate way, so it is very important to choose the right type of a chart. The choice can be made according to the algorithm: define the goal of the visualization -> define the data type -> choose the suitable diagram.

Types of the diagrams:

- Line chart
- Bar chart & histogram
- Pie chart
- Radar
- Scatter plot
- Bubble chart
- A map
- Tree map

Data visualization tools:

In this list examples of BI tools and market leaders are presented. Their advantages and flaws are also listed here and discussed.

1. **Power BI**[12]

This service helps to combine and compare data from different sources. There are a lot of visualization galleries (built-in and a lot of additional ones). It works well with many sources of the data, has an easily understandable interface and is overall user-friendly. It is also possible to integrate BI into its own visualization applications. But it does not always work well with Big data.

2. **Tableau**[13]

The largest and most simplified platform for the user, specializing in data analysis and visualization. Allows one to build spectacular graphics. A flexible control panel interface that allows a user to combine and overlay the necessary elements. Interface is easy and understandable. Tableau is best suited for corporations that need data visualization solutions without having to manually configure them.

[12] https://powerbi.microsoft.com/
[13] https://tableau.com/

3. **Google data studio**[14]

This service is free and intuitive. It integrates well with Google products. There's a possibility to customize your own templates. However, it has a small set of visual tools, the ability to work with calculated fields is limited compared to other popular visualization services.

4. **Plotly**[15]

It is a platform for creating graphs, charts, presentations and unique dashboards. Allows one to download data, select a visual, customize the result. A user can create a visualization in which you can edit almost everything: legend, labels, line thickness, color, size. The gallery has unique diagrams that are not available in other services. You can create a visual, save it as a vector graphic or a png image, and embed it on a website in html code format.

6. Methodology

There were several stages of creating a report. It was necessary to develop a concept of the dashboard, study the literature on the related topics, and conduct research on the situation in AI analytics. The methodology describes the methods that were used to collect and process the data, methods of plotting and visualization, creating a report.

6.1 Concept of the dashboard

The concept of the dashboard is to reflect the key trends in the development of Artificial intelligence and AI technologies. Dashboard covers many spheres such as research in AI, AI activities, investments in AI, etc. The target audience of the dashboard are

- people who are interested in AI area
- business
- public sector representatives

[14] https://datastudio.google.com/
[15] https://plotly.com/

- decision makers
- people who study AI and want to see key trends and sub-spheres in order to continue their research or choose the right institution
- emerging startups

One of the goals of the dashboard is to show that the AI field develops fast and there is a lack of end-to-end analytics and data on Artificial intelligence. The dashboard demonstrates main streams of AI development and investments. The future users of the report will be able to create an idea about AI in general, AI technologies, trends, see what countries and institutions expand the AI environment. After full review and study of the dashboard, the user will have a complete and capacious impression of the development of digital technologies in the context of Artificial intelligence.
The main sources of data when creating a report is statistical data collected from various statistical accounting services. The latest and most representative sources of information are used to fully reflect the situation in the development of Artificial intelligence technologies at the moment. The structure of the report is such that each page represents a separate category of information related to Artificial intelligence and the report as a whole makes up a general picture. Each page provides a concise and clear presentation of data on each of the areas of AI technologies. The report is developed on a BI system Tableau.

6.2 Data search and selection

One of the main struggles that I faced during the process of my project was to find the relevant data sets, extract the necessary fields from them and arrange the tables.
The first thing I did was just search the web for anything AI related to find some articles or statistics. That was a pretty tough task as there is plenty of useless information online, especially on the trending AI topic right now, so it was hard to find anything that will correspond to my needs. The key objective was to collect correct, accurate and most relevant data, so all the sources were double checked.

For the convenience of future research I created a google spreadsheet in order to store all the sources of the data I will find. There are several pages, the most important of them was structured in a way that was easy to understand what this data is about. Here is the header of the table.

Data source	Link	Organization	Can the data be downloaded (yes/no)	Time period of the data	Description of the dataset	Key words	How it can be used in the dashboard	Comments

Using this table simplified the process of distribution of the data, understanding whether it will be useful for the dashboard or not.
Mainly, all the resources had csv of xlsx files, so I just downloaded them.
I was collecting the data from a variety of sources:

- Official statistics portals (Eurostat, European commission, Росстат)
- Reports
- Articles
- Dashboards (OECD.AI live data, AI Watch Landscape)
- Research publications
- AI-related websites

Then all the data was color-coded according to whether it will be used in the report. The next step was to organize all the data sources into groups by topic and time period.
The last step was creation of the page in the google spreadsheet with preliminary names of the dashboard pages and the list of the file names that will be used there.

6.4 Choosing the BI instrument

Earlier in my work, I stated the presence of several tools that can reproduce the report needed.As I previously worked with a number of BI systems, used power query language and overall analyzed data with BI instruments, the key choice was made based on the

user-friendliness, future convenience for the dashboard users and overall appearance. After careful selection and creating draft in Power BI, I settled on choosing Tableau for several reasons:

- Connection to many different data sources, including various databases, Google Sheets, the same Excel or text files
- Convenient connection of several tables into one data source via drag-and-drop, as well as the ability to easily collect data from tables with different levels of detail;
- Intuitive interface and building visualizations via drag-and-drop
- Beautiful visualizations with numerous settings to adjust for a convenient use
- The ability to add interactive

6.5 Choosing diagram types for the data

The success of visualization directly depends on the correctness of its application, namely on the choice of the type of graph, its correct use and design. So the important part of my work was to make the correct visual representation of the data. I started with stating the aim of every graph I wanted to make.

Relationships in data are how they depend on each other. With the help of relationships we identify the presence or absence of dependencies between variables. If the main idea of the information contains the phrases “refers to”, “decreases / increases with”, then we are going to show what relationship there is. For example, I show how the number of employees and average investment range are related.

The **distribution** of data is how they are located relative to something, how many objects fall into certain consecutive areas of numerical values.

Data **composition** is the combination of data in order to analyze the big picture as a whole, comparing components that make up a percentage of a certain whole. I have many examples of these shown in my dashboard because it is one of the best ways to show the real picture of a topic, what is more important and what is less, where more money goes and where less, what country made a bigger contribution to AI and so on.

Comparison of data - combining data, in order to compare some indicators, identifying how objects relate to each other. It is also a comparison of components that change over

time. For example, I have several graphs that represent value's change through the time (number of research publications and investments).

So at this point I had all my future visualizations divided into the groups by purpose. The next step of my work was to define the datatype needed for each of the visualizations. The simple way of distinguishing the data structures is: continuous, numerical and time series, discrete data, geographical and logical data.

- Continuous numerical data contains information about the dependence of one numerical value on another.
- Time series contains data about events occurring in any period of time. In my dashboard I used several time series datasets in order to explicitly show the trends in various AI-related spheres over the time.
- Discrete data may contain dependencies of categorical values. For example, in my work discrete data structure is used for the number of firms that are engaged in AI in different countries, for the number of research publications that were made on a certain topic.

The next step was to choose the most suitable chart for a visualization. It is a known fact that the simpler the better. So my goal here was to choose the right graph so it would correctly and clearly show the data, convey the correct idea and was simple to read and understand.

There is a table that I used as my guide in choosing the correct diagram for my dashboard. It is taken from the book of Gene Zelezlny "Say it with charts", which I read before starting to work on the project.[16]

The idea of the visualization / data type	**Relationships**	**Distribution**	**Composition**	**Comparison**
continuous numeric data	line, area, scatter, bubble	scatter, bubble	line, area, radar	stacked line, stacked area

[16] Say It With Charts: The Executive's Guide to Visual Communication. Gene Zelezny

continuous time data	line, area, radar, scatter, bubble	time line, gantt, waterfall, radar	time line, gantt, waterfall, radar	stacked line, stacked area, gantt
discrete data	bar, scatter, bubble	bar, scatter, bubble	bar, pie, doughnut	pie, doughnut, stacked bar.
geographical data	map, line, area	map, scatter	map, bar	map, stacked bar

After all the graphs have been chosen for the visualizations it was the time for their enhancement:

- adjustment of the scale
- choosing the appropriate colors for grouping objects
- naming the axes
- adjust the size, direction of the lines and bars
- select the appropriate data labeling
- setting legends and captions

At this point all the graphs for the dashboard were done and ready to be assembled.

6.3 Structure of the dashboard

The main task of the dashboard is to provide complete information about the current trends in the development of digital technologies, in particular Artificial intelligence. That is why the dashboard is structured so that the first page provides basic information about the players in the AI field, about companies and institutions that are engaged in research, as well as the geography of research.
Further the dashboard is divided into several pages that fully describe each of the sections. Dashboard pages are:

1. AI research

2. Publications activity
3. AI players and activities
4. AI talent
5. Investments in AI
6. Investments in startups.

7. Technical part

7.1 Key instruments

1. Tableau (for the final report)
2. Power BI (for the draft report layout)
3. Google spreadsheets/ MS Excel (data)
4. Pycharm/Jupyter Notebook
5. Github

7.2 Data processing and storage

All the datasets that were used to create the visualization for the report were collected from various Internet resources in csv or xlsx format. Each file has been converted to a single format and all files have been uploaded to a folder on Google drive for convenience and accessibility. Further, all the files were processed using Jupyter Notebook and Pandas to bring all columns and rows, all names of countries and enterprises to a single format so that the report structure and logic were not violated. After carrying out all these procedures, all files in the folder were updated, also each file name corresponded to the diagram in which in the future it will be used; this was also done for ease of use.

When creating my report, I used the connection from the scoreboard service to Microsoft Excel files, but since I plan to develop my report further, keep it working and make it more usable and updated, I decided to create a database that will store all the information, all the datasets I used. The presence of a single database in which all my tables will be

stored in the future will greatly simplify the procedure of updating the data. I created a database using python and the sqlite3 library. In the future, I plan to expand the capabilities of my database and I will transfer it to the service PostgreSQL because it is more convenient for me.
One of the important tasks in the creation of my report was data processing, because it had to be done correctly so that all the connections that exist between a large number of data sets were made accurately in order not to violate the logic of the report.

7.3 Implementation in BI instruments

After the concept of the dashboard was thoroughly thought out, all the data was collected, processed and analyzed, the analytics tool was selected, the final part of the work - creating a report in Tableau service started.
I gradually connected to the files in Microsoft Excel, extracted data from them and after the data was received, the creation of each individual diagram began. Since the structure of the report and the types of charts were already selected by me earlier in my work, creating the charts themselves in the Tableau service was much easier for me. After all the diagrams were made, I created the primary pages of my report and placed graphs in each of them in correspondence to the topic. During the process of creating a dashboard with a good visual picture with a semantic load, some visual elements were changed in order to satisfy the needs of the report.
My goal was to show the relationship between various metrics and find cross-cutting sections such as countries, regions, technology groups, topics and fields so that the report looks complete and combines data from different resources together, creating a full informative dashboard.
One of the important parts in creating visual elements was the correct selection of colors for the labels on the data, the correct name of the x and y axes for a full understanding of the idea of the chart. To create the best visual pictures I chose the same colors for each individual page so that there is no visual noise. Also all numerical data was formatted in accordance with the requests of certain charts so that the user has a sensible idea of what data is displayed on the chart.

One of the key parts of every report is the ability to filter by multiple parameters. On every page of my report there are filters, also some of the graphs can be used as filters. This is done so that each user can customize the reports to suit their needs.
After all the charts were created and distributed on separate report pages, all the pages were collected into a single story and this story was published in Tableau online.

8. Description of the dashboard

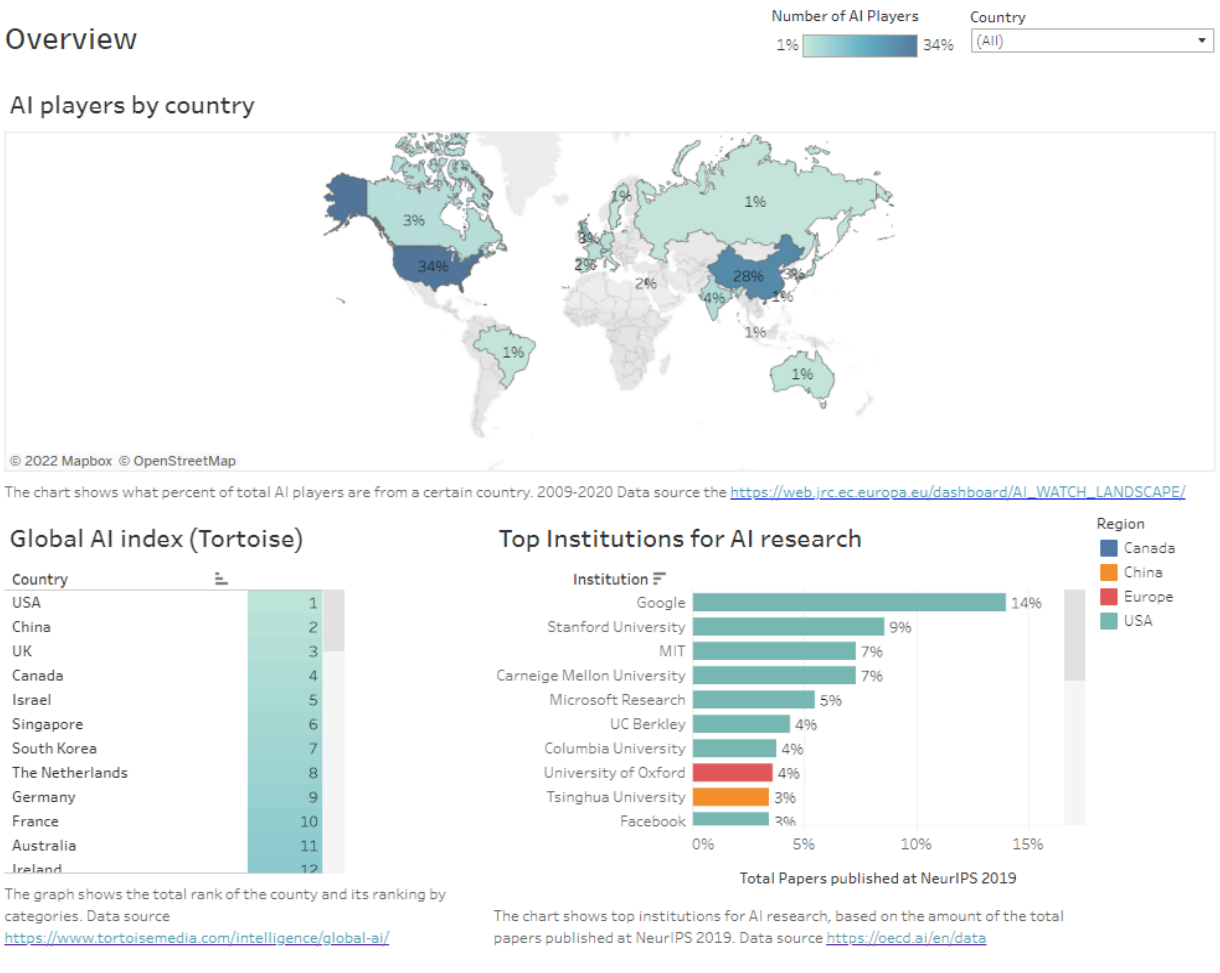
The dashboard itself consists of 7 pages that include up-to-date information on various fields of Artificial intelligence, represents key trends in AI-related topics, displays the investment flow in AI and gives an overall picture on the current situation with AI, its services and technologies around the world.
Dashboard combines various data sources in one, allowing a user to perceive the overall picture of the AI area, delving deeper into every field related to AI.

8.1 Overview

On the first page of my report you can find the distribution of AI players by country players. AI players are companies, government and research institutions that are engaged in research in the field of Artificial intelligence and apply AI technologies in their work. We see that the players in Artificial intelligence around the world are unevenly distributed, most of them are representatives from the United States of America and China.

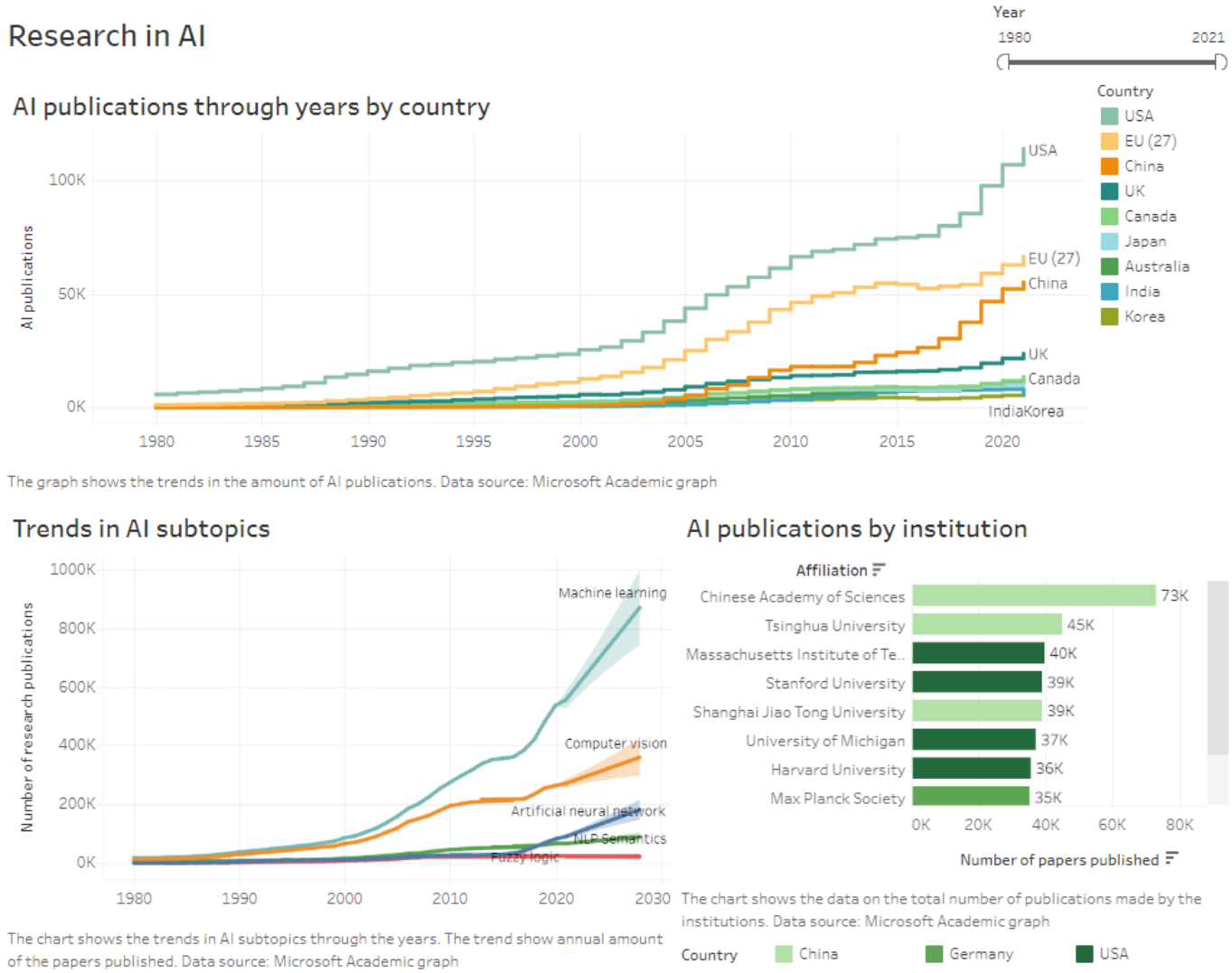
Also on this page one can see the total country ranking in AI. And the ranking was made by separate services that evaluate the contribution and development of countries in the field of Artificial intelligence. Further there is a graph that shows top institutions by their contribution to scientific publications on Artificial intelligence.

In general, this page gives a general idea of which countries are involved in the research of Artificial intelligence and which are the most interesting and popular institutions for research and how the situation is with the development of Artificial intelligence in the world.



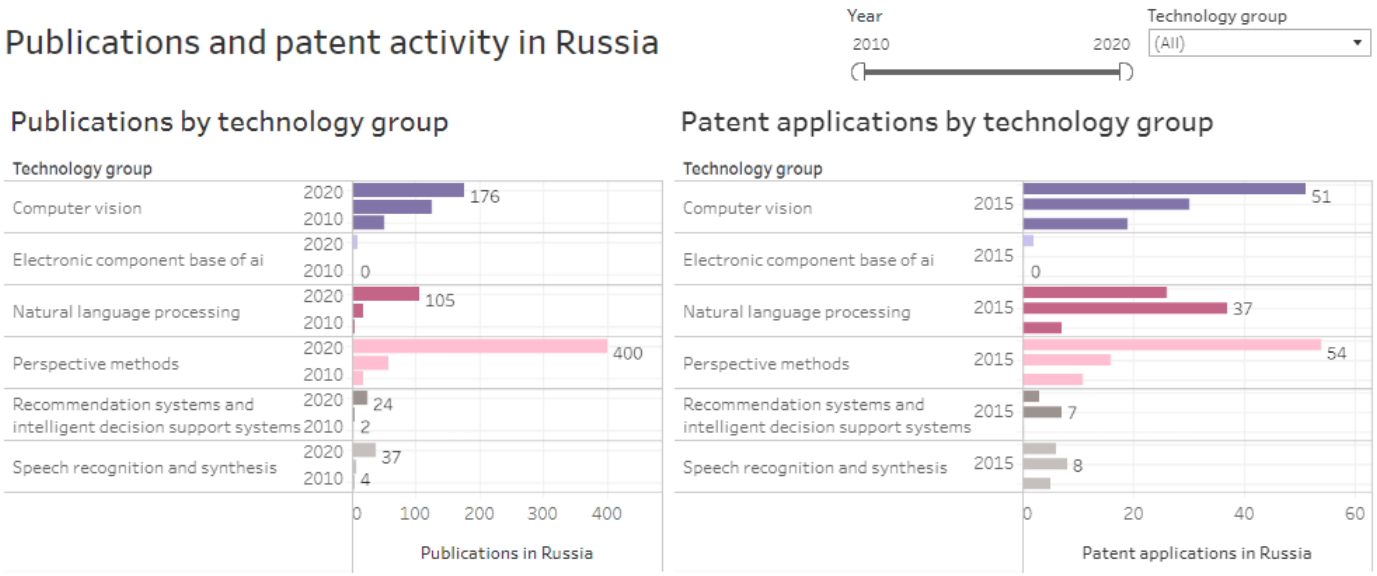
8.2 Research in AI

The next page of my dashboard is called “Research in AI". On the first graph one can see how countries are distributed by the number of publications from 1980 to the present time. We see that over the past few years there has been a strong jump in the number of publications for each of the countries and the United States all the time is in the lead. Further we see a graph that shows the main trends in the subtropics of Artificial intelligence also from 1980 to the present. Here we can see what are the dominant subtopics in Artificial intelligence research. This graph also contains a forecast for the number of publications in certain subgroups up to 2030 made using analytical tools. In order to use these graphs in a more convenient way there is a filter on the data. The last graph on this page represents the number of publications by different institutions over time.



8.3 Publication and patent activity in Russia

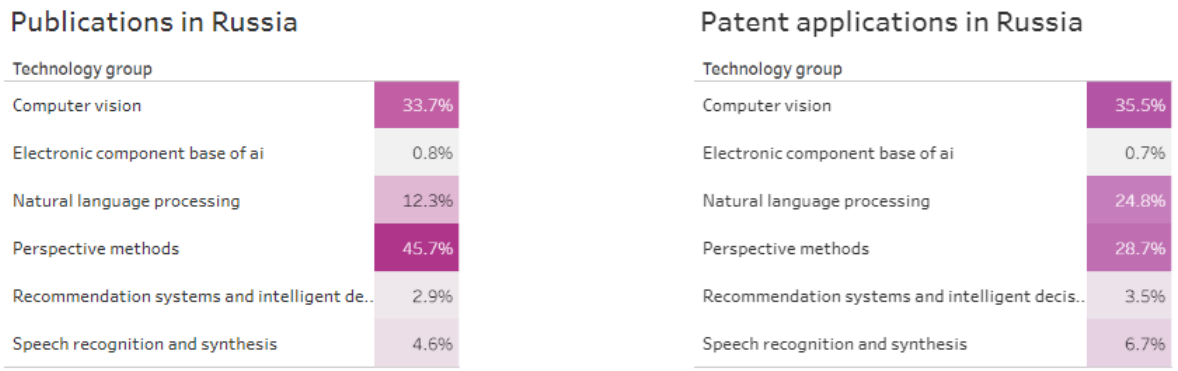
The next dashboard is dedicated to publication and patent activity in Russia. We can see a graph that clearly shows the number of publications and patents that have been made in the last 10 years in Russia divided into technology groups. Also the two graphs below show the percentage of certain technology groups in which publications and patents were made from the total number. All these graphs can be filtered by date and the graphs themselves can be used as filters.



The plot shows the number of publications in Russia through years by technology group. 2010-2020. Data source: ИСИЭЗ НИУ ВШЭ

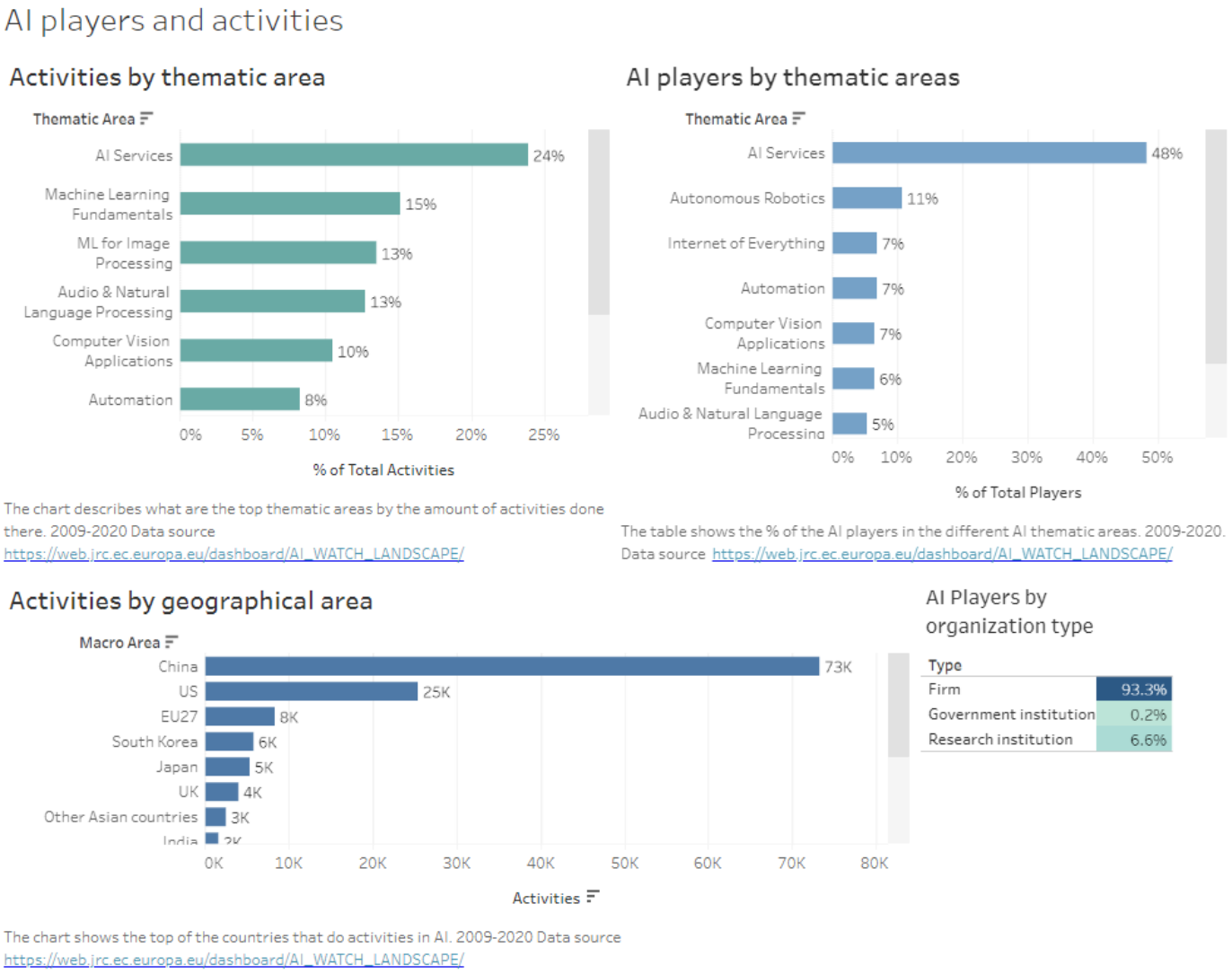
The plot shows the number of patent applications in Russia through years by technology group. 2010-2019. Data source: ИСИЭЗ НИУ ВШЭ

Total publications and petent appications by technology group



8.4 AI players and activities

This page defines the main directions and thematic areas of AI players and activities. We can see how the players in the field of Artificial intelligence and activities are distributed by thematic area. We see that AI services occupy a leading position on both points, then the activities are distributed approximately equally. Then there is a graph that shows the number of activities carried out in each geographical area. We see that again, the United States and China are in the lead, then the European Union follows them. One can also note that despite the fact that most of the players in the field of AI are in the United States, more Artificial intelligence activities are carried out in China. The last graph on this page shows how the players in the field of Artificial intelligence are distributed by type of organization. We see that the majority belong to firms, followed by research institutes and less than 0,5% are government institutions.



The chart describes what are the top thematic areas by the amount of activities done there. 2009-2020 Data source https://web.jrc.ec.europa.eu/dashboard/AI_WATCH_LANDSCAPE/

The table shows the % of the AI players in the different AI thematic areas. 2009-2020. Data source https://web.jrc.ec.europa.eu/dashboard/AI_WATCH_LANDSCAPE/

The chart shows the top of the countries that do activities in AI. 2009-2020 Data source https://web.jrc.ec.europa.eu/dashboard/AI_WATCH_LANDSCAPE/

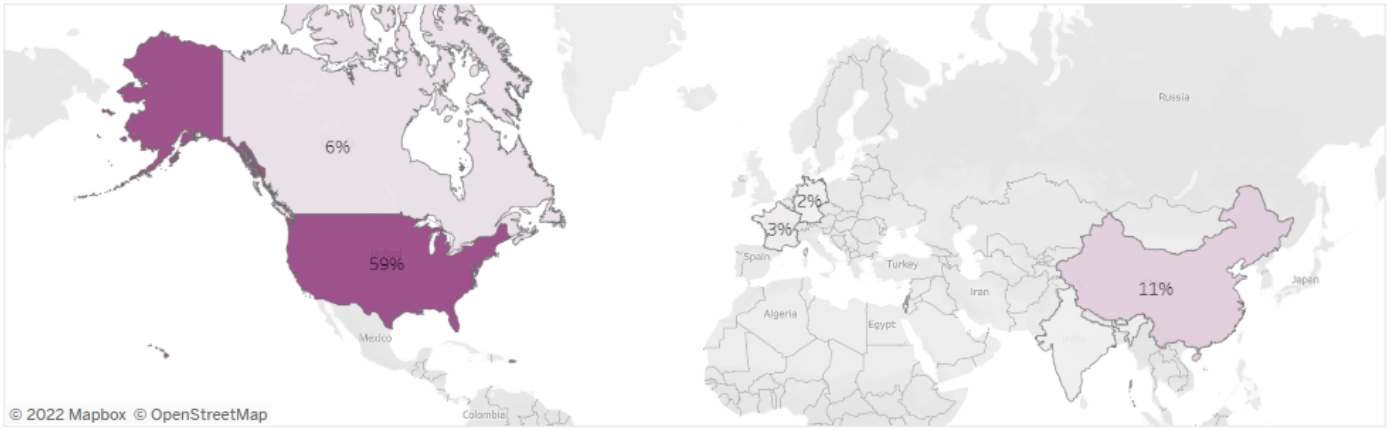
8.5 AI talent

Page “AI talent” helps the user to understand what is the geography of the top researchers in the field of Artificial intelligence. We see that now most of the elite researchers are working in the USA, in China, Canada and some EU countries. Most of the researchers came from China and America, but there’s a huge percent of specialists who are not from a top country. Also we can see where elite researchers received undergraduate degrees.

AI talent

Country
(All)

Where do elite researchers work now?



The map shows where elite researchers work now. Data from https://macropolo.org/digital-projects/the-global-ai-talent-tracker/

Where do top-tier AI researchers come from?

Country	
Canada	5%
China	29%
France	5%
Germany	3%
India	8%
Israel	3%
Others	27%
USA	20%

The chart shows the countries from which most of the top AI researchers came from. Data from https://macropolo.org/digital-projects/the-global-ai-talent-tracker/

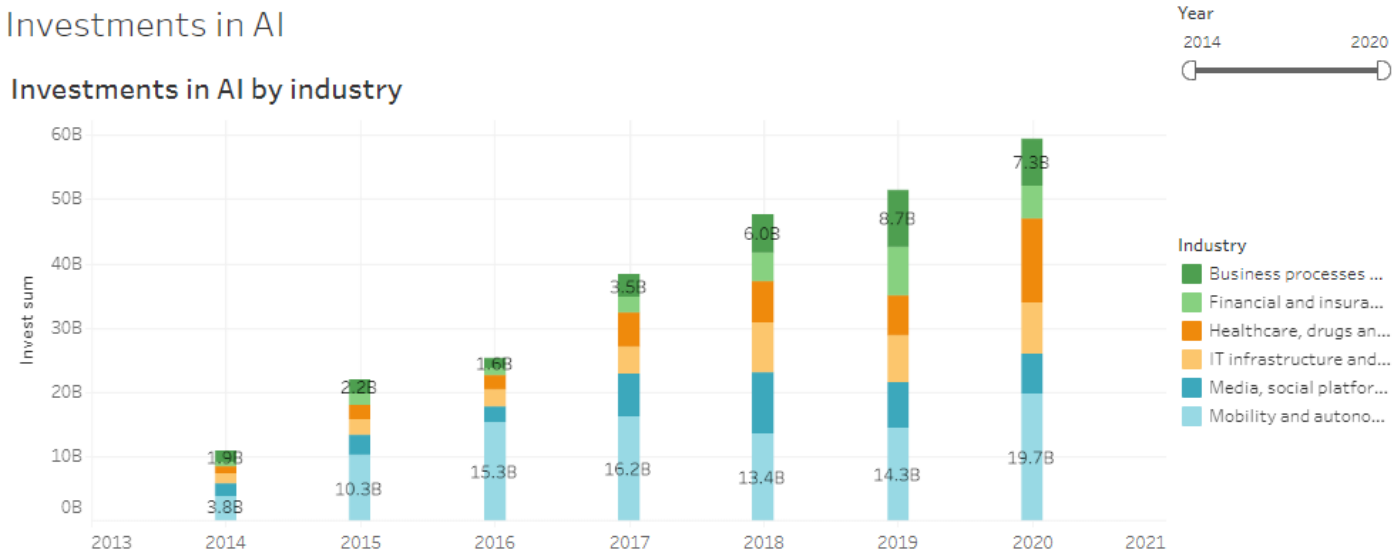
Where do top AI researchers recieve undergraduate degree?

Country	
Canada	5%
China	10%
France	6%
Germany	3%
India	12%
Israel	7%
Others	22%
USA	35%

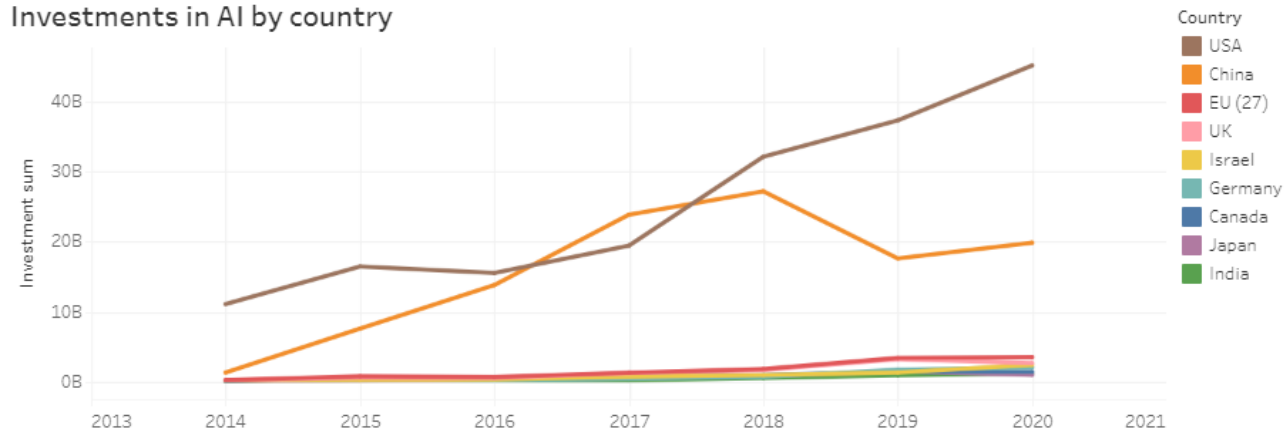
The graph shows the countries where elite researchers received their degree. Data from https://macropolo.org/digital-projects/the-global-ai-talent-tracker/

8.6 Investments in AI

This dashboard is dedicated to investments in AI. We see that there is a graph representing overall investments by year and industry. “Mobility and autonomous vehicles” is the leading industry of investment over the years, with more than 20B investment sum in 2020. The next graph shows countries that invest in AI, leaders are again USA and China.



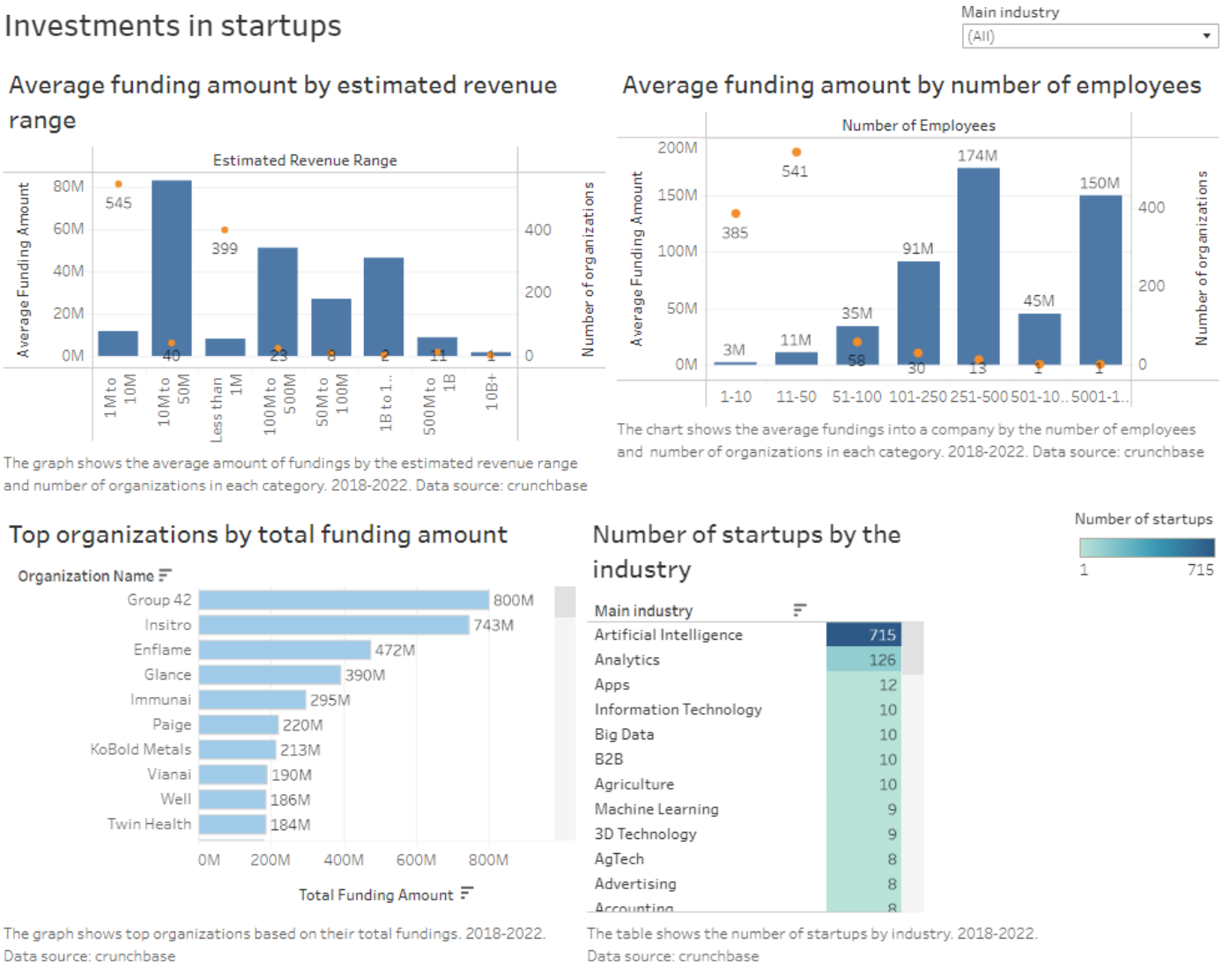
The plot shows the investments in AI by industry in billion dollars. Data source: https://oecd.ai/en/data



The plot shows the investments in AI by country in billion dollars. Data source: https://oecd.ai/en/data

8.7 Investments in startups

On this page we can see investments in startups that are related to Digital technology and in particular to Artificial intelligence. At the top of the page we see two graphs that show the average investment into a certain organization by expected revenue and by the number of employees in this organization. Then there are the top organizations which have the biggest investment sum and the number of startups by the industry. This page gives a basic idea of which projects, which companies and industries are currently getting more investments and are estimated to earn more.



The graph shows the average amount of fundings by the estimated revenue range and number of organizations in each category. 2018-2022. Data source: crunchbase

The chart shows the average fundings into a company by the number of employees and number of organizations in each category. 2018-2022. Data source: crunchbase

The graph shows top organizations based on their total fundings. 2018-2022. Data source: crunchbase

The table shows the number of startups by industry. 2018-2022. Data source: crunchbase

9. Building a model

Here the key instruments used for research are listed:

- Python, Jupyter Notebook
- Python libraries - numpy, pandas, matplotlib.pyplot, sklearn.preprocessing, sklearn.linear_model, sklearn.model_selection, sklearn.feature_extraction.text, sklearn.metrics
- Microsoft Excel
- GitHub

9.1 The aim of the model

There is a data on AI fundings of different companies and startups. There are several features that define a company and based on which the funding is made. So the purpose of the model is to predict the funding amount for a certain company using some input data.

9.2 Dataset descriprion

The **target** variable is Total Funding Amount Currency (in USD). And the **explanatory** variables are (some of them were dropped while data processing and feature selection): Organization Name, Organization Name URL, Actively Hiring, IPO Status, Industries, Headquarters Location, Description, CB Rank (Company), Headquarters Regions, Estimated Revenue Range, Operating Status, Founded Date, Founded Date Precision, Exit Date, Exit Date Precision, Company Type, Website, Number of Founders, Number of Employees, Number of Funding Rounds, Last Funding Amount, IPqwery - Most Popular Patent Class

In this dataset there is not much numerical data, however there are categorical data as well which is benefitial for a successful modelling. The dataset consists of more than 30000 observations, so we have enough data to build a solid model which will meet the needs.

9.2 Data preparation and Implementation

Firstly, work was carried out on preliminary data processing. The original table contained a lot of undefined values, as well as fields with values that logically do not affect the target. First of all, it was necessary to vectorize various types of the data:

1) Numeric data. These are data of the type 'float64' or 'int32', which can already be used as features, so only undefined values were converted for these fields by replacing them with the corresponding modes for each field.
2) Data containing dates. This field was replaced with 3 numeric ones: day, month and year, which were later also used as features.
3) Categorical data. Data that can take a limited number of values. They have been transformed into one-hot vectors.
4)Text data. Data that is a set of characters that carries some kind of description or keywords. To vectorize this type of data, the tf-idf transformation with preliminary lemmatization and text normalization was used.
5) Rows containing an undefined value in the target field have been deleted.
6) Fields with the number of undefined values of more than half of the observations were deleted.
After data preprocessing, 7 regression models were selected and trained.
GradientBoostingRegressor, ElasticNet, SGDRegressor, SVR, BayesianRidge, LinearRegression, LGBMRegressor

model	explained_variance_score	mean_absolute_error	mean_squared_error	mean_absolute_percentage_error
GradientBoostingRegressor	0.976831	6.64074e+06	2.33281e+15	3.02201
ElasticNet	0.0698223	5.2742e+07	9.36186e+16	52.333
SGDRegressor	-6.90032e+35	2.23836e+25	6.99472e+52	1.81266e+18
SVR	3.10778e-07	5.03124e+07	1.02784e+17	15.465
BayesianRidge	0.0338927	5.04217e+07	9.72332e+16	46.2205
LinearRegression	0.245622	8.13359e+07	7.59258e+16	166.67
LGBMRegressor	0.959637	7.50818e+06	4.06234e+15	48.2047

When comparing the results that were obtained through the process of models' accuracy validation it is seen that GradientBoostingRegressor outperforms other regression models.

model	explained_variance_score	mean_absolute_error	mean_squared_error	mean_absolute_percentage_error
GradientBoostingRegressor	0.992399	6.2788e+06	8.69887e+14	3.4728

9.3 Result description

Metrics that were used:

explained_variance_score calculates the explained regression variance score.

$$explainedvariance(y, \hat{y}) = 1 — \frac{Vary — \hat{y}}{Vary}$$ In our case (for Gradient Boosting) it equals 0.99, which is a good result.

mean_absolute_error calculates the mean absolute error, the risk metric corresponding to the expected value of absolute loss or error 1-normal loss.

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i — \hat{y}_i|.$$

mean_squared_error calculates the mean squared error, the risk metric corresponding to the expected value of the squared (quadratic) error or loss.

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}—1} (y_i — \hat{y}_i)^2.$$

mean_absolute_percentage_error (MAPE), also known as the mean absolute deviation in percent (MOPED), is a metric for evaluating regression problems. The idea of this metric is to be sensitive to relative errors as it is not changed by the global scaling of the target variable.

$$\text{MAPE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \frac{|y_i — \hat{y}_i|}{max(\epsilon, |y_i|)}$$

Based on the results of the validation sample, a gradient boosting model was selected based on the ratio of training time and the quality of the presentation. In the test sample, the selected model gave very similar results for MAPE (3.02% validation, 3.47% test), so the selected model has a good generalizing ability.

10. Conclusion

As a result of my work, I analyzed the field of Artificial intelligence and the application of its technologies, collected data, processed, prepared and organized data for subsequent visualizations using the Tableau. I created a report that consists of 7 pages, each of them describes a specific area related to Artificial intelligence and gives a brief but rich information on this area. The dashboard that I created can serve as a tool for analysis and decision-making in the field of AI development.
Also there are prospects for improvement: adding new data sources, new metrics, larger area of topics coverage etc. I can say that the problem of insufficient data on Artificial intelligence and its technologies exists and there is a lack of end-to-end analytics and Artificial intelligence data and general structure of the AI data.
As fo the other part of my work, the regression model was build. The model was made to predict the total funding amount of a company or a startup in the field of AI. Overall, the model showed good results and may be used for a company to get their estimated fundings.

11. Bibliography

1. Microsoft Research - AI https://www.microsoft.com/en-us/research/search/?q=AI
2. Tortoise Media - Global AI Index https://www.tortoisemedia.com/intelligence/global-ai/
3. OECD.AI https://oecd.ai/en/data?selectedArea=ai-news
4. MarcoPolo - is the Paulson Institute's think tank. https://macropolo.org/digital-projects/the-global-ai-talent-tracker
5. Цифровая трансформация: Ожидание и реальность. Доклад НИУ ВШЭ https://issek.hse.ru/mirror/pubs/share/603838492.pdf
6. OECD library https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classifi

cation-of-ai-systems_cb6d9eca-en;jsessionid=d8IaCqSXccrx1l8GOuKDsHHj.ip-10-240-5-7

7. Forbes https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-Artificial-intelligence/
8. Развитие отдельных высокотехнологичных направлений. Белая книга. НИУ ВШЭ
9. Что такое цифровая экономика? Тренды, компетенции, измерение. Доклад НИУ ВШЭ
10. Finance Information Systems https://fisgroup.ru/blog/fis-dss/
11. Wikipedia https://en.wikipedia.org/wiki/Speech_recognition
12. IBM https://www.ibm.com/topics/computer-vision
13. Plotly https://plotly.com/
14. Data Studio https://datastudio.google.com/
15. Tableau https://tableau.com/
16. Microsoft Power BI https://powerbi.microsoft.com/

12. Appendix

1. Link to the dashboard https://public.tableau.com/app/profile/giorgos2796/viz/Coursework_pt3/AIdashboard?publish=yes
2. Link to the Github repository https://github.com/dalixiv/coursework_dash_bd
3. Link to the Google Spreadsheet where all the data was collected, analyzed and distributed across the sheets of the dashboard https://docs.google.com/spreadsheets/d/1nTj56-G6DqL5uCoair__8EdZr3iDyyaxp0XOSGk-sDQ/edit#gid=770906804