

Methodological guidelines.

Fundamentals of Pandas. Data filtering

STORYLINE:

The business concept defined in the third module: **Pitching** (a short and structured project presentation intended for investors). The presentation is aimed at demonstrating the unique character of the product based on the available data and inspiring the investor to invest in its development. Studying the module, students will apply the data analysis methods for doing market research and for preparing the pitching mini-presentation.

SUMMARY:

The lesson goal is to motivate students to study the new Data Analysis module and master the basic techniques of working with the main object of the Pandas library, DataFrame.

The students will get acquainted with the concept of Data Analysis, learn about the Series and DataFrame structures; they'll learn how to work with CSV files, get general information about a dataset, calculate statistical indicators, and perform data filtering.

LINKS AND ACCESSORIES:









- [presentation for the lesson](#);
- exercises in VSC and on the platform: [part 1](#), [part 2](#).

EDUCATIONAL OUTCOMES OF THE LESSON

<i>After the lesson, the students:</i>	<i>The result is achieved when the students:</i>
know: <ul style="list-style-type: none"> • data characteristics: Series and DataFrame; • purpose and syntax of methods; • data filtering syntax; understand: <ul style="list-style-type: none"> • the value of statistical indicators: arithmetic mean, standard deviation, median, quartiles; • what data filtering is and in what cases it is necessary to use it; can: <ul style="list-style-type: none"> • create DataFrame based on a CSV file; • calculate the average, min, and max values; • perform data filtering by one or more conditions. 	<ul style="list-style-type: none"> • have participated in the discussion about the amount of available information and asked clarifying questions; • have coped with the tasks given on the platform (Fundamentals of Pandas); • have tried to correct the errors occurred without any assistance; • have phrased questions for the dataset based on the examples given.

RECOMMENDED LESSON STRUCTURE

Time	Stage	Stage aims
------	-------	------------

10 min 	Discussion: Data analysis	<ul style="list-style-type: none"> ❑ Introduce students to a letter from the World of Code incubator representatives. ❑ Discuss the main tasks of the Data Analysis module. ❑ To present Data Science as the science for data processing and extraction of useful knowledge.
15 min 	New topic: Getting acquainted with the Pandas Library	<ul style="list-style-type: none"> ❑ Analyze the Pandas data structures: <ul style="list-style-type: none"> ❑ Series; ❑ DataFrame. ❑ Demonstrate the creation of DataFrame based on a CSV file. ❑ Analyze the info(), head(), tail(), describe() functions; ❑ Get to know the following statistical indicators: arithmetic mean, standard deviation, percentile, median.
15 min 	VSC + Platform: Fundamentals of Pandas	<ul style="list-style-type: none"> ❑ Organize the execution of the Fundamentals of Pandas task on the platform.
5 min 	Break	<ul style="list-style-type: none"> ❑ Help students regain concentration.
5 min 	Discussion: Data analysis tools	<ul style="list-style-type: none"> ❑ Discuss the role of the Problem Statement stage in the process of data analysis.
15 min 	New topic: Data Filtering	<ul style="list-style-type: none"> ❑ Phrase "simple" questions for the dataset together with the students. ❑ Learn the syntax of the min(), max(), mean(), median () methods. ❑ Together with the students, phrase questions for the dataset with an additional condition. ❑ Learn the syntax of filtering by one or more conditions.
20 min 	VSC + Platform. Data filtering	<ul style="list-style-type: none"> ❑ Arrange the execution of the Data Filtering task on the platform.
5 min 	Wrapping up the lesson. Reflection	<ul style="list-style-type: none"> ❑ Congratulate the students with their first steps completed in data analysis, namely, problematization and extraction. ❑ Review the new material and ask the students' impression of the classes. ❑ Announce the topic of the next lesson.

Discussion: Data analysis

(10 min)

The Data Analysis module is aimed at developing students' skills of goal-directed data analysis to extract ideas and knowledge from that data necessary for the implementation of projects, market research, identifying patterns, and making the right decisions.

❏ 3 min Introduce students to a letter from the World of Code representatives.

Tell the students about the new module: Data Analysis. Start with a letter from the representatives of the World of Code incubator. It will facilitate a smooth switching from the Mobile Development module to the Data Analysis module. Focus on the following aspect: a startup is always a story about a product that should generate profit. And then the **question** arises: How to make a minimally viable product a profitable one? **Answer:** it is necessary to correctly assess the target audience of the product, think about the sources of income, and use all this data while positioning, advertising, and delivering the product to the consumer. All these problems could be solved with data analysis. For example, data analysis has allowed the WhatsApp startup to become profitable without inviting significant investor funds and come up with a new model of funding startups.

Ask the students: 1. What questions should an entrepreneur answer to make a product profitable? 2. How could they find answers to these questions? **Answer:** 1. Which market should be chosen for promoting the product? 2. Who will be the main consumer of the product? What financial model will be the optimum for such a product? 3. What competitive advantages does the product have?


<p>Now you need to find the answer to the question: how can you make a minimum viable product profitable?</p> <p><small>Discussion: News about you and for you</small></p>	<p>Key aspects influencing the success of a future product</p> <ul style="list-style-type: none"> • In which market should we promote the product? • Who will be the main consumer of our product? • Which financial model will be best for this kind of product? • What competitive advantages does the product have? <p><small>Discussion: Data analysis</small></p>	<p>In this module, in order to find the answers to the essential questions, you will learn how to:</p> <ul style="list-style-type: none"> • Analyze structured and unstructured data using the Pandas library. • Generate new business hypotheses and test existing ones using data analysis. • Come up with projects/startups and, based on data analysis, determine the direction of their growth, financial models, competitive advantages, etc. • Present the ideas behind business projects, whose relevance is proven by scientific data, to investors. <p><small>Discussion: News about you and for you</small></p>
--	---	---

❏ 3 min Discuss the main tasks of the Data Analysis module.

After you have told the students what business needs data analysis could accommodate, tell them what tasks they will have to solve in this module. Please accentuate the fact that it is data analysis that can turn an assumption/opinion into a fact.


❏ 4 min To present Data Science as the science for data processing and extraction of useful knowledge out of this data.

Tell the students about the Data Scientist profession: what knowledge is required for this work, what stages data analysis consists of. Represent Data Science as the science for data processing and extraction of useful knowledge out of these data. Give some examples of the data analysis application in different business areas. For example, Netflix determines the genre first for some projects, and then selects directors and actors for these projects based on big data analysis. So, Kevin Spacey and David Fincher were invited to the political thriller House of Cards as the featured actor and the director, respectively, without any trial or other tests. In the final part, focus on the tasks the students will complete while studying the module.

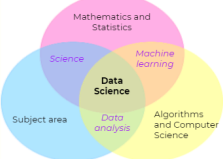


MODULE 3. LESSON 1


Discussion.
Data analysis



Very large science



Discussion, Data analysis



What are we going to do with the data in this module?

Let's produce a **general project** with real data from **Google Play** to master the basic operations:


- Filtering
- Grouping
- Clean up
- Visualization

This will help us to learn to:

- create and test business hypotheses;
- come up with projects, determine the direction of their growth and financial models;
- represent ideas to investors.

To realize an **individual project** of your own choice and present it to investors as a **pitch**.

Discussion, Data analysis



New topic: Getting acquainted with the Pandas Library

(15 min)

3 min Analyze the Pandas data structures.

Tell the students that all the functionality necessary for data analysis is contained in the Pandas library that they will study over the next 6 lessons. Emphasize that the library name is usually abbreviated during its import. Explain to the children that it is possible not to use the proposed abbreviation, but analysts all over the world use exactly the abbreviated version in their code.

Analyze the features of Series and DataFrame — the two main data structures that the students will work with.


The large Pandas Library

The library provides the following features for data analysis in Python:

- Reading and recording files of different formats.
- Calculation of the main statistical indicators.
- Inserting and deleting columns with data.
- Merging datasets.
- Filtering and grouping of data.
- Creating pivot tables.
- Creating diagrams.

[Link to the official website of the library](#)

New topic: Data Science



Pandas data structures. Series

A series is a data structure containing elements and their indexes.

Index	Data
01	TikTok
02	YouTube
00	Instagram

The index is the sequence number in a line. It is created automatically


Index	Data
SE00-16	TikTok
SE23-08	YouTube
OC06-10	Instagram

Index — the application release date (month date-year)

Index (id)	Data
835599320	TikTok
544007654	YouTube
389801252	Instagram

The index is the application ID in stores

New topic: Data Science




Pandas data structures. DataFrame

DataFrame is a data structure containing sets of elements and their indexes.

Index (id)	Name	Rating	Number of reviews
835599320	TikTok	4.3	38 330 273
544007654	YouTube	4.6	117 208 38
389801252	Instagram	3.8	122 104 613
460177396	Twitch	4.4	4 423 232

New topic: Data Science



1 min Demonstrate the creation of DataFrame based on a CSV file.

Tell the students about the CSV file format. Explain that tabular data containing hundreds of thousands or even millions of rows could be more conveniently stored in text format because, in this case, the file will not take up much disk space. Show the kids how to create a data frame from a CSV file.

5 min Analyze the info(), head(), tail(), describe() functions

Talk more specifically about the methods of working with data frames. Demonstrate to the students how to display the first and last few lines on the screen, get generalized information about the data amount and types, and calculate the basic statistical indicators for quantitative attributes.

6 min Get to know the following statistical indicators: arithmetic mean, standard deviation, percentile, median.

Especially direct the children's attention to the difference between the arithmetic mean and the median. Show by example that the median is always a real (or close to real) value of an attribute that divides a series of numbers into two absolutely equal parts. When explaining concepts such as standard deviation and percentiles/quartiles, be guided by the general level of mathematical background the children in the group have. If this material is difficult for students, you can skip it in this lesson because using these indicators is not mandatory for completing tasks and projects.

General information about data

Method	Purpose
<code><name_DataFrame>.info()</code>	Returns brief information about each column and the data frame as a whole
<code><name_DataFrame>.head(n)</code>	Returns the first n lines (5 lines by default) of the data frame
<code><name_DataFrame>.tail(n)</code>	Returns the last n lines (5 lines by default) of the data frame
<code><name_DataFrame>.describe()</code>	Returns generalized statistical indicators based on the data stored in the dataframe

describe()

Indicator	Explanation
Count	The number of non-empty values
Mean	Arithmetic mean
Std	Standard deviation
min	The minimum value of the attribute
25%	25th percentile (the first quartile)
50%	50th percentile (median)
75%	75th percentile (third quartile)
max	The maximum value of an attribute

Median

Sorting the numbers in ascending order:

10 14 1 12 40 6 24 5 5 5 5 6 3 4 5 29 31

1 3 4 5 5 5 5 6 6 10 12 14 24 29 31 40

Indicator	Value	Explanation
Median or 50%	6	The value of the attribute that is greater than 50% of the values in a sample, and less than 50% of the remaining values in the sample

VSC + Platform: Fundamentals of Pandas

(15 min)

Arrange the students' work with the dataset. Note to the students that they should write the code in VSC and then transfer the received answers to the platform where the tasks with automatic verification are uploaded. If you see that children have started their work with uncertainty and are facing difficulties, analyze one or two tasks together with them commenting on their actions and inviting them to ask questions.

You can find the answers to the tasks and the reference code for obtaining them at the last pages of the methodological guidelines.

Break

(5 min)

Switch the students off their computers. The purpose of the break is to shift attention and warm up. Arrange one of the [suggested physical activities](#).

Discussion: Data Analysis tools

(5 min)

Return to the Stages of Data Analysis scheme. Refer to the concept of Data Extraction. Notice to the children that they already know how to extract data from a file but have not considered it a significant stage of the problem statement yet. Tell the students about the problem structure, draw their attention to the concept of a **hypothesis**. Analyze the problem structure using the example of the fitness application developed in the previous module.

MODULE 3. LESSON 1

Discussion. Data analysis tools

Perhaps the problem should be formulated first?

Discussion. Data analysis tools

The structure of the problem

The problem	— this is "realization of ignorance" or a complex, uncertain situation that requires examination.
Research question	— this is the answer that should be obtained as a result of the solution to a problem.
Working hypothesis	— a conventionally accepted verifiable assumption based on research questions. The hypothesis is proposed according to the following model: "If... then...".

Discussion. Data analysis tools

Invite students to voice their **suggestions**: 1. What are the main characteristics of your application? 2. Assume for whom such a product would be suitable.

Principally, these questions are aimed at bringing the students to understanding that their answers are only assumptions, and the objective facts can be only found by analyzing data on real or potential users of an application.

Tell the students about the four main ways of generating income from a mobile application.

Ask them: which financial model seems the most promising? How to turn an assumption into a fact? Inform the students that, in this part of the lesson, they will master the basic techniques of professional analysts: they will learn to pose research questions and get reliable answers with the Pandas commands.

New topic: Data Filtering

(15 min)

❑ 2 min Phrase "simple" questions for the dataset together with the students.

Show students examples of questions that an analyst may be interested in answering. Begin with simple tasks for calculating fundamental statistical indicators for separate columns; then, move on to more complex queries that require data filtering.

❑ 6 min Learn the syntax of filtering by one condition.

Together with the students, analyze the syntax of filtering data by one condition. Make sure that the students understand the syntax logic and will be able to apply it upon solving similar problems. Instead of demonstrating slides, you can open VSC and type the code right there showing to students the sequential steps towards the result. Another way of arranging this lesson stage is to show students the code on slides and invite them to type this code in VSC at the same time.

❑ 7 min Learn the syntax of filtering by several conditions.

If the background of students in the group is good enough, analyze a complicated task that requires filtering by two conditions to solve it. Note to the students that each condition should be taken in parentheses, and, to describe logical operations, the symbols & (and) and | (or) should be used instead of words 'and' and 'or' familiar to them.

General information about data		Research questions with the condition	Syntax									
<table border="1"> <thead> <tr> <th>Method</th> <th>Purpose</th> </tr> </thead> <tbody> <tr> <td><name_DataFrame>[<column_name>].min()</td> <td>Returns the minimum value of an attribute</td> </tr> <tr> <td><name_DataFrame>[<column_name>].max()</td> <td>Returns the maximum value of an attribute</td> </tr> <tr> <td><name_DataFrame>[<column_name>].mean()</td> <td>Returns the mean value of an attribute</td> </tr> <tr> <td><name_DataFrame>[<column_name>].median()</td> <td>Returns the median value of an attribute</td> </tr> </tbody> </table>	Method	Purpose	<name_DataFrame>[<column_name>].min()	Returns the minimum value of an attribute	<name_DataFrame>[<column_name>].max()	Returns the maximum value of an attribute	<name_DataFrame>[<column_name>].mean()	Returns the mean value of an attribute	<name_DataFrame>[<column_name>].median()	Returns the median value of an attribute	<ul style="list-style-type: none"> How does the average rating of paid and free apps differ? What is the minimum price for a paid application with a rating greater than 4.5? What is the maximum number of downloads of applications from the ART_AND_DESIGN category? Etc. 	<p>Step 2</p> <p>After the dataframe name, in square brackets the data we are interested in should meet the following conditions:</p> <pre>df[df['Rating'] > 4.9]</pre>
Method	Purpose											
<name_DataFrame>[<column_name>].min()	Returns the minimum value of an attribute											
<name_DataFrame>[<column_name>].max()	Returns the maximum value of an attribute											
<name_DataFrame>[<column_name>].mean()	Returns the mean value of an attribute											
<name_DataFrame>[<column_name>].median()	Returns the median value of an attribute											

VSC + Platform. Data filtering

(20 min)

Arrange the students' work with the data. Like in the previous task, the children would have to type the code in VSC and transfer the received answers to the platform where the tasks with automatic verification are uploaded.

If the group is dominated by weak or unassertive students, suggest solving the first 1-2 tasks in front-end mode: run VCS on your computer, share the screen with students, and start creating the code commenting each command and parameter with them. As soon as the students feel confident and ready to code independently, they can start working on tasks on their own at a pace that is comfortable for them.

Wrapping up the lesson. Reflection*(5 min)*

Invite the students to share their impressions of the lesson. Ask each student to choose one question and answer it. Tell the children that in the next lesson they will learn techniques allowing them to reveal hidden trends and correlations in the data.

Answers to tasks

Task 1. Fundamentals of Pandas

✓ Part 1. My correct solution

Attempt: 1/200

Fill in the blanks in the text

Use the data that you managed to get from the dataset when working with the code in VSC.

What is the name of the first application in the dataset? Copy the title from the console and paste it into the text field unchanged.

Photo Editor & Candy Camera &

Which category does the last application in the dataset belong to? Copy the title from the console and paste it into the text field unchanged.

LIFESTYLE

How many columns are there in the dataset? 12

Answer: Photo Editor & Candy Camera & Grid & ScrapBook

✓ Part 1. My correct solution

Attempt: 1/200

Match the column names with their categories

Match the data type with its column:

Int64

Float64

Object

Content Rating

Reviews

Rating

✓ Part 1. My correct solution

Attempt: 1/200

Fill in the blanks in the text

Use the data that you managed to get from the dataset when working with the code in VSC.

Enter the average (round to the nearest hundredth) and the median (integer) for the application size (size).

Average: 22.77

Mean: 14

How much does the most expensive app cost? Enter the answer as an integer. 400

✓ Part 1. My correct solution

Attempt: 1/200

Fill in the blanks in the text

Use the data that you managed to get from the dataset when working with the code in VSC.

Enter the average and median for the number of installs of the applications. Enter both figures as integers.

Average: 8662313

Median: 100000

Reference code of the solution:

```
import pandas as pd
df = pd.read_csv('GoogleApps.csv')

# What is the name of the application being the first in the dataset?
print(df.head())

# What is the category (Category) of the application being the last one in the dataset?
print(df.tail())

# How many columns are there in the dataset?
# What is the type of data stored in each of the columns?
print(df.info())
```

```
# Specify the arithmetic mean and the median of the applications size (Size)
# How much does the most expensive app cost?
# * Specify the arithmetic mean and median of the number of application installations
(Installs)
print(df.describe())
```

Task 2. Data filtering

✓ Part 1. My correct solution

Attempt: 1/200

Fill in the blanks in the text

Use the data that you managed to get from the dataset when working with the code in VSC.

How much does the cheapest paid app cost? Copy the answer from the console and paste it into the text field unchanged:

0.99

✓ Part 1. My correct solution

Attempt: 1/200

Fill in the blanks in the text

Use the data that you managed to get from the dataset when working with the code in VSC.

What is the median number of app installs for the ART_AND_DESIGN category? Give the answer as an integer:

100000

✓ Part 1. My correct solution

Attempt: 1/200

Fill in the blanks in the text

Use the data that you managed to get from the dataset when working with the code in VSC.

How much higher is the maximum number of reviews for free apps than the maximum number of reviews for paid apps? Give the answer as an integer:

44703802

Answer: 44703802

✓ Part 1. My correct solution

Attempt: 1/200

Fill in the blanks in the text

Use the data that you managed to get from the dataset when working with the code in VSC.

What is the minimum application size for teenagers? The answer is rounded to the nearest thousandth:

0.315

✓ Part 1. My correct solution

Attempt: 1/200

Fill in the blanks in the text

Use the data that you managed to get from the dataset when working with the code in VSC.

Which category does the app with the most reviews belong to? Copy the title from the console and paste it into the text field unchanged.

GAME

✓ Part 1. My correct solution

Attempt: 1/200

Fill in the blanks in the text

Use the data that you managed to get from the dataset when working with the code in VSC.

What is the average rating for apps that cost over \$20 and have over 10,000 installs? Copy the answer from the console and paste it into the text field:

4.25

Reference code of the solution:

```
import pandas as pd
df = pd.read_csv('GoogleApps.csv')

# What is the price (Price) of the cheapest paid app (Type == 'Paid')?
print(df[df['Type'] == 'Paid']['Price'].min())

# What is the median (median) number of installs (Installs)
# of applications from the "ART_AND_DESIGN" category (Category)?
print(df[df['Category'] == 'ART_AND_DESIGN']['Installs'].median())

# For how much the maximum number of reviews for free apps (Type == 'Free')
# exceeds the maximum number of reviews for paid apps (Type == 'Paid')?
free = df[df['Type'] == 'Free']['Reviews'].max()
```

```
paid = df[df['Type'] == 'Paid']['Reviews'].max()
print(free - paid)

# What is the minimum size (Size) of an app for teenagers (Content Rating == 'Teen')?
print(df[df['Content Rating'] == 'Teen']['Size'].min())

# *What is the category (Category) of an app with the largest number of reviews (Reviews)?
print(df[df['Reviews'] == df['Reviews'].max()]['Category'])

# *What is the mean (mean) rating (Rating) of apps with the price (Price) more than $ 20 and
# with the number of installs (Installs) more than 10,000?
print(df[(df['Price'] > 20) & (df['Installs'] > 10000)]['Rating'].mean())
```