

educational_projects:

This repository is intended for projects that I did in the course of training on the course ["Data Analyst"](#) from **Yandex Practicum**

The course program is aimed at a smooth study of the skills necessary for the work of an analyst, a consistent increase in the complexity of the topics studied and the tools used. This course prepared me to enter the path of data analysis, helped me master the skills necessary for work: upload, transform and clean data using SQL queries, calculate key company performance metrics and evaluate their significance, run A/B tests to test hypotheses, help businesses make data-driven decisions, create dashboards and presentations.

The following libraries were used in the projects:

numpy pandas scipy sqlalchemy matplotlib warnings plotly seaborn datetime sys io requests
math sklearn itertools

Part 1. Introduction to the Data Analyst profession. Fundamentals of Python and Data Analysis

Topics covered:

1. Variables and data types. Data output and arithmetic operations.
2. Lines.
3. Lists.
4. The for loop.
5. Nested lists.
6. Conditional operator. while loop.
7. Functions.
8. Dictionaries.
9. The pandas library.
10. Data preprocessing.
11. Data analysis and presentation of results.
12. Jupyter Notebook - a notebook in a cell.

Project: music of big cities:

The study was conducted on real Yandex.Music data. The data were checked and the behavior of users of the two capitals was compared.

During the project were:

-
- omissions were identified and processed;
 - data types are replaced with corresponding ones;
 - removed duplicates;
 - carried out logical indexing, grouping, sorting.

Part 2. Data preprocessing

Topics covered:

1. Work with passes.
2. Determination of abnormal values.
3. Conversion of data types.
4. Basic methods for finding duplicates.
5. Working with imperfect real datasets.

Project: Borrower Reliability Study - Analysis of Banking Data:

Based on the data of the credit department of the bank, the influence of marital status, number of children, income level and purpose of the loan on the fact of repayment on time was studied.

During the project were:

-
- omissions were identified and processed;
 - data types are replaced with appropriate ones;
 - removed duplicates
 - categorized data;
 - one dataframe decomposed into three.

Part 3: Exploratory Data Analysis

Topics covered:

1. Data visualization with histograms and boxes with whiskers.
2. Exploring data slices.
3. Finding relationships between different parameters in the data.
4. Merging tables.
5. Getting conclusions from grouped data.

Project: sale of apartments in St. Petersburg - analysis of the real estate market:

Based on the Yandex.Realty service data, the market value of various types of real estate objects, typical parameters of apartments, depending on distance from the center.

During the project were:

-
- data preprocessing was carried out;
 - added new data;
 - built histograms, boxplots, scatterplots.

Part 4: Statistical Data Analysis

Topics covered:

1. The study of objects and their relationships by statistical methods.
2. Samples and statistical significance.
3. Detection and processing of anomalies.

Project: determination of a favorable tariff for a federal mobile operator:

Based on the data of the mobile operator's customers, the behavior of customers was analyzed and the search for the optimal tariff was carried out (a preliminary analysis of the use of tariffs on a sample of customers was carried out, the behavior of customers when using the operator's services was analyzed, and optimal sets of services for users were recommended).

During the project were:

-
- data preprocessing and analysis;
 - tested hypotheses about the difference in the revenue of subscribers of different tariffs and the difference in the revenue of subscribers from Moscow and other regions.

Part 5: Building Project 1: Explore the patterns that determine successful games:

The parameters that determine the success of the game in different regions of the world are revealed. Based on this, a report was prepared for a computer game store for planning advertising campaigns.

During the project were:

-
- data preprocessing, analysis;
 - selected current period for analysis;
 - portraits of users of each region were compiled
 - hypotheses were tested: the average user ratings of the Xbox One and PC platforms are the same; the average user ratings of the Action and Sports genres are different (in the analysis, I used Student's t-test for independent samples).

Part 6: Business Performance Analysis

Topics covered:

1. Metrics and funnels.
2. Cohort analysis.

3. Unit economics.
4. Custom metrics.

Project: Analysis of the effectiveness of attracting customers:

A challenge for a marketing analyst for the Procrastinate Pro+ entertainment app. Despite huge investments in advertising, the company has been losing money for the past few months. The task is to understand the reasons and help the company to become a plus.

During the project were:

-
- data preprocessing was carried out;
 - analysis of data from ProcrastinatePRO+;
 - various metrics were calculated, cohort analysis was used: LTV, CAC, retention rate, etc. (the previously written functions for calculating metrics were used);
 - Conclusions are drawn from the data obtained.

Part 7: Business Decision Making

Topics covered:

1. Methods and tools for testing hypotheses.
2. Design of experiments. Seasonality.
3. Cohort analysis.
4. A/B testing.

Project: Testing hypotheses to increase revenue in an online store - evaluate the results of an A / B test:

Using the data of the online store, hypotheses were prioritized, and the results of A / B testing were evaluated using various methods.

During the project were:

-
- prioritization of hypotheses by ICE and RICE frameworks was carried out;
 - an analysis of the results of the A / B test was carried out, graphs of cumulative revenue, average bill, conversion by groups were built;
 - the statistical significance of differences in conversions and average checks was calculated for raw and cleaned data;
 - based on the analysis of the test results, making a decision on the advisability of further testing.

Part 8: How to tell a story with data

Topics covered:

1. Presentation of the results of the analytical study.
2. Ways of visual presentation of data.
3. Creation of reports explaining the conclusions of the analyst.
4. Seaborn library.

Project: Research of the catering market in Moscow to make a decision to open a new establishment:

Public catering market research based on open data, preparing a presentation for investors. Will a cafe where guests are served by robot waiters be successful and popular for a long time?

During the project were:

-
- data preprocessing was carried out;
 - Based on the results of the analysis, a presentation was prepared for investors with recommendations.

Part 9: Building Project 2: Analyzing User Behavior in a Mobile App:

Based on the data on the use of a mobile application for the sale of food products, the sales funnel was analyzed, and the results of A/A/B testing were evaluated.

During the project were:

-
- data preprocessing was carried out;
 - Studied the principles of event analytics, built a sales funnel, explored the path of users to purchase;
 - analyzed the results of the A / B-test of the introduction of new fonts;
 - 2 control groups were compared with each other and with the test group;
 - determined whether the new font will affect user behavior.

Part 10: Basic SQL

Topics covered:

1. Introduction to databases.
2. Slices of data in SQL.
3. Aggregating functions. Grouping and sorting data.
4. Relationships between tables. Types of joining tables.

5. Subqueries and temporary tables.

Project: Research data on investments of venture funds in start-up companies:

In the standalone project of this course, you work with a database that stores information about venture capital funds and investments in startup companies. This database is based on the Startup Investments dataset published on the popular data mining competition platform Kaggle.

During the project were:

-
- various data unloading using SQL queries;
 - Calculations of business indicators were made using SQL queries.

Part 10.1: Advanced SQL

Topics covered:

1. Calculation of business indicators.
2. Fundamentals of window functions.
3. Cohort analysis.
4. Installing and configuring the database and database client.
5. Frames in window functions. Additional SQL tools.

Project: Exploring data about questions and answers, posts, etc. hosted on the StackOverflow platform:

In the standalone project of this course, you work with the StackOverflow database, a service for questions and answers about programming. StackOverflow is like a social network - users of the service ask questions, reply to posts, leave comments, and rate other answers. The work was carried out with the version of the database where data on posts for 2008 is stored, but the tables also contain information about later ratings that these posts received.

During the project were:

-
- Various data unloading using SQL queries and analysis of the results were performed.

Part 11: Automation

Topics covered:

1. Automation of data analysis processes.
2. Flow analytical solutions.

3. Registration of events in logs, creation of regular reports.
4. Dashboards. Monitoring.

Project: Creating a custom events dashboard for a news aggregator:

Using Yandex.Zen data, a dashboard was built with metrics for user interaction with article cards.

During the project were:

-
- data was unloaded using SQL queries;
 - the data was analyzed and the file for the dashboard was saved in the .csv format;
 - for data visualization, a dashboard with several filters was created on the Tableau public platform;
 - Based on the results, a presentation was prepared with the resulting graphs.

Part 12: Machine Learning Basics

Topics covered:

1. Tasks of machine learning in business.
2. Machine learning algorithms.
3. The process of solving machine learning problems.

Project: Churn Prediction for Fitness Centers:

Based on data on visitors to the network of fitness centers, the probability of outflow for each client in the next month was predicted, user portraits were formed using clustering.

During the project were:

-
- predicted the probability of outflow (at the level of the next month) for each client;
 - typical portraits of users have been formed: the brightest groups have been identified, their main properties have been characterized;
 - analyzed the main features that most strongly affect the outflow;
 - Suggested recommendations for the strategy of interaction with customers and their retention.

Graduation project:

Based on all the data received in the course, a bootcamp project was completed in the field of "mobile applications":

Part 1. **Identification of underperforming call center agents:**

Based on data containing information about call centers and their operators, the most inefficient / effective operators were found, and also checked statistical hypotheses.

During the project were:

- a review and pre-processing of data was carried out;
- research data analysis was carried out;
- inefficient operators are defined;
- tested statistical hypotheses:
 - outgoing calls of call-cents are more often missed than answered;
 - with incoming calls, due to the long waiting time, they often do not wait for the appointment of an operator.

Part 2. **Testing changes (A/B test) related to the introduction of an improved recommender system in an online store:**

Based on data with user actions, terms of reference and several auxiliary datasets, the results of an A / B test of changes associated with the introduction of an improved recommender system for online store products were evaluated.

During the project were:

- data review and data preprocessing;
- an assessment of the correctness of the test and an exploratory analysis of the data was carried out;
- A/B test analysis was carried out.

Part 3. **SQL. Analysis of the database of the service for reading books by subscription:**

The database containing information about books, publishing houses, authors, as well as user reviews of books was analyzed. Based on these data, value propositions for a new product are formulated.

During the project were:

-
- researched database tables;
 - Various data unloading using SQL queries and analysis of the results were performed.