Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

КАФЕДРА ШТУЧНОГО ІНТЕЛЕКТУ

Department of Artificial Intelligence

R AND PYTHON PROGRAMMING LANGUAGES IN STATISTICAL COMPUTING

Curriculum (Syllabus)

Course details

Level of higher education	***Second (Master's)***
Field of knowledge	*C - social sciences, journalism, information and international relations*
Specialisation	*C5 Sociology*
Educational programme	*Social Data Analytics*
Status of discipline (code)	*Mandatory, Professional training cycle*
Form of study	*Full-time (day)*
Year of training, semester	*1st year, spring semester*
Scope of the discipline	*4 ECTS credits, 120 hours* *16 hours of lectures, 30 hours of practical classes, 74 hours of independent work.*
Semester assessment/assessment measures	*Test, Modular control work, Calculation assignments*
Class schedule	*https://schedule.kpi.ua/* *1 hour of lectures and 2 hours of computer workshops per week*
Language of instruction	*Ukrainian*
Information about course coordinator/teachers	Lecturer: *Ivan Oleksandrovych Pyshnohraiev, PhD in Physics and Mathematics, Associate Professor, pyshnograiev@wdc.org.ua* Computer workshops: *Ivan Oleksandrovych Pyshnohraiev, PhD in Physics and Mathematics, Associate Professor*
Course location	Google Classroom https://classroom.google.com/c/NjYxNTc2NjUyMDM4

Course programme

1. Description of the course, its purpose, subject matter and learning outcomes

The discipline is a compulsory part of the educational programme. The study of the discipline is aimed at forming, developing and consolidating the following general and professional competences in students:
GC 01 Ability to think abstractly, analyse and synthesise
PC 13 Ability to use the SPSS statistical package and the R and Python programming languages for social data analysis.

As a result of studying the course, students should be able to demonstrate the following learning outcomes:
PRN 04 Apply scientific knowledge, sociological and statistical methods, digital technologies, and specialised software to solve complex problems in sociology and related fields of knowledge,
PRN 14 Apply the programming languages R and Python to analyse social data.

*At the end of the course, students should **know**:*

- *the features of working with the R and Python programming languages;*
- *methods of data processing and analysis;*
- *means of data processing, storage and analysis;*

***be able to**:*

- *analyse data using the R and Python programming languages;*
- *create scripts and programmes for analysing social data.*

2. Prerequisites and post-requisites of the discipline (place in the structural-logical scheme of training under the relevant educational programme)

The discipline is based on the knowledge and skills of related disciplines studied in the previous semester and educational level. This discipline precedes OK PO 05 "OSINT Technologies in Sociological Research" and may be one of the main components of a master's thesis.

3. Content of the academic discipline:

Section 1. Using the R language to solve data analysis problems

Topic 1.1. Basics of the R language.

1. *Characteristics of the R language and programming environment;*
2. *Basic data types and operations on them, writing elementary scripts;*
3. *Basic data structures.*
4. *Basic algorithmic elements in the R language.*
5. *Working with data sources.*

Topic 1.2. Data visualisation using the R language.

1. *Introduction to data visualisation, basic types of graphs;*
2. *Advanced visualisation tools, map building;*
3. *Creating dashboards.*

Topic 1.3. Data analysis in R.

1. *Problem setting, main stages;*
2. *Methods of processing different types of data, formatting, normalisation, etc.;*
3. *Preliminary data analysis, correlation and regression analysis;*
4. *Forecasting methods, building regression models;*
5. *Model evaluation methods.*

Topic 1.4. Examples of using R for socio-economic data analysis.

Section 2. Using Python to solve data analysis problems

Topic 2.1. Basics of the Python language.

1. *Characteristics of the R language and programming environment;*
2. *Basic data types and operations on them, writing elementary scripts;*
3. *Basic data structures.*
4. *Basic algorithmic elements in Python.*
5. *Classes and objects (overview).*

Topic 2.2. Basic Python libraries for working with data sources.

1. *Using Pandas for working with data;*
2. *Using Numpy for working with data structures.*

Topic 2.3. Exploratory data analysis.

1. *Descriptive statistics;*
2. *Correlation and regression analysis.*

Topic 2.4. Building models using Python.

1. *Linear and multivariate linear regression;*
2. *Polynomial regression.*
3. *Model quality assessment*

Topic 2.5. Examples of using Python for analysing socio-economic data.

List of computer workshops:

1. *Creating scripts in R for data analysis and visualisation.*
2. *Solving social data analysis problems using the R language.*
3. *Creating scripts in Python for data analysis and visualisation.*
4. *Solving social data analysis problems using Python.*

3. Training materials and resources

Basic:

1. *R for Data Science by Hadley Wickham, Garrett Grolemund [en]. URL: https://r4ds.had.co.nz/ .*
2. *Advanced R by Hadley Wickham [en]. URL: https://adv-r.hadley.nz/index.html*
3. *Maiboroda R.E. Computer Statistics. Professional Start. Tutorial. Kyiv University, 2020. – 482 p. http://probability.univ.kiev.ua/userfiles/mre/compsta1.pdf*
4. *Methods and Models of Intelligent Data Analysis. Practicum [Electronic resource]: textbook for students majoring in 122 "Computer Science", educational programme "Systems and Methods of Artificial Intelligence" / N. I. Nedashkivska; Igor Sikorsky KPI. – Kyiv: Igor Sikorsky Kyiv Polytechnic Institute, 2019. – 71 p. https://ela.kpi.ua/handle/123456789/53764*
5. *Tsesliv, O. V. Programming for Analytical Research [Electronic resource]: a textbook for bachelor's degree students in the educational programme Economic Analytics, speciality 051 Economics. Electronic network publication / O. V. Tsesliv; Igor Sikorsky Kyiv Polytechnic Institute. – Electronic text data (1 file: 1.55 MB). – Kyiv: Igor Sikorsky Kyiv Polytechnic Institute, 2024. – 238 p. https://ela.kpi.ua/handle/123456789/66102.*

Supplementary:

1. *Data analysis. Laboratory workshop [Electronic resource]: textbook for students majoring in 113 "Applied Mathematics" / N. M. Kussul, A. Yu. Shelestov, S. A. Tarasenko, G. O. Yailimova; Igor Sikorsky Kyiv Polytechnic Institute. – Electronic text data (1 file: 582.97 KB). – Kyiv: Igor Sikorsky Kyiv Polytechnic Institute, 2022. – 28 p. https://ela.kpi.ua/handle/123456789/50425*
2. *Novotarsky, M. A. Fundamentals of Programming in the Python Algorithmic Language [Electronic resource]: textbook for students of the educational programme "Computer Systems and Networks" speciality 123 "Computer Engineering" / M. A. Novotarsky; Igor Sikorsky Kyiv Polytechnic Institute. – Electronic text data (1 file: 17.93 MB). – Kyiv: Igor Sikorsky Kyiv Polytechnic Institute, 2022. – 701 p. https://ela.kpi.ua/handle/123456789/49913*
3. *Data Mining Tutorial. https://www.geeksforgeeks.org/data-mining/*
4. *Python Data Mining Quick Start Guide, published by Packt. https://github.com/PacktPublishing/Python-Data-Mining-Quick-Start-Guide*
5. *R programming // Electronic resource. Access mode: https://coderlessons.com/tutorials/mashinnoe-obuchenie/r-programmirovanie/r-programmirovanie*
6. *Virtualisation Technology // Electronic resource. Access mode: https://www.sciencedirect.com/topics/computer-science/virtualization-technology*
7. The R Manuals. https://cran.r-project.org
8. *Practical Data Mining with Python. https://dzone.com/refcardz/data-mining-discovering-and*

Educational content

4. Methodology for mastering the academic discipline (educational component)

The Google Classroom will contain assignments with detailed instructions and the necessary materials, which must be completed on time.

Section 1. Using the R language to solve data analysis problems

Lecture 1. *Basics of the R language: syntax and data structures*

History and philosophy of the R language. Overview of the RStudio development environment. The concept of packages and the CRAN repository. Basic data types: numeric, integer, character, logical. Vector arithmetic. Basic data structures: vectors, matrices, lists, and data frames. Control structures (if-else, loops) and writing your own functions. Importing and exporting data (CSV, Excel).

Assignments for independent study:

1. *Install R and RStudio, set up a working directory.*

2. *Create a script that loads an external data file, checks variable types, and outputs basic statistics (summary).*

Lecture 2. *Data visualisation in the R ecosystem*

Base R vs. ggplot2 graphics system. The concept of "Grammar of Graphics". Building basic types of graphs: histograms, scatter plots, boxplots. Configuring aesthetics, geometries (geoms) and faceting. Working with geospatial data: sf and leaflet packages for building interactive maps. Introduction to creating web applications and dashboards using R Shiny.

Homework assignments:

1. *Familiarise yourself with the R Graph Gallery.*

2. *Build a complex graph in ggplot2 using at least three layers (data, geometry, theme) and colour grouping.*

Lecture 3. *Data analysis methodology in R*

The concept of "Tidy Data" and the use of dplyr and tidyr packages for data manipulation (filtering, aggregation, table joining). Exploratory data analysis (EDA): identifying outliers, checking distribution. Correlation analysis: calculating Pearson/Spearman coefficients, constructing correlograms. Building a linear regression (lm function). Model diagnostics and interpretation of results. Quality metrics: R-squared, RMSE.

Assignments for SRC:

1. *Study the cheat sheet for working with dplyr.*

2. *Clean up the "dirty" dataset: rename columns, replace missing values with the mean or median.*

Lecture 4. *R in socio-economic research*

Specifics of working with survey data (sample weighting, survey package). Analysis of categorical data and contingency tables. Working with Likert scales. Using APIs to obtain economic data (e.g., WDI packages for World Bank or Eurostat data). Examples of visualising demographic changes and economic trends.

Homework assignment:

1. *Find an R package that provides access to open government data (Open Data) or data from international organisations.*

2. *Download a set of socio-economic indicators for one country for the last 10 years.*

Section 2. Using Python to solve data analysis problems

Lecture 5. *Introduction to Python for data analysis*

Python as a universal programming language. Jupyter Notebook and Google Colab environments. Basic data types: int, float, str, bool. Data structures: lists, tuples, dictionaries, sets. Basics of algorithmisation: loops (for, while), conditional operators, list comprehensions. Basics of object-oriented programming: concepts of class, object, method, and attribute. Homework assignment:

1. *Write a Python function that processes a list of numbers (for example, filters even numbers and squares them).*

2. *Familiarise yourself with the differences in array indexing in R (from 1) and Python (from 0).*

Lecture 6. *NumPy and Pandas libraries: the foundation of Data Science*

NumPy library: the concept of ndarray, vectorisation of calculations, broadcasting. Pandas library: Series and DataFrame objects. Loading data from different sources. Indexing, slicing, and filtering data (loc, iloc). Grouping and aggregating data (groupby, pivot_table). Handling missing values and duplicates.

Homework assignment:

1. *Perform a merge operation on two dataframes using a common key.*

2. *Compare the speed of mathematical operations with standard Python lists and NumPy arrays.*

Lecture 7. *Exploratory data analysis (EDA) using Python*

Descriptive statistics using Pandas (describe, value_counts). Visualisation of distributions and dependencies: Matplotlib and Seaborn libraries. Construction of histograms, boxplots, and heatmaps for correlation matrices. Statistical hypotheses: scipy.stats library (t-test, normality test).

Homework assignment:

1. *Build a correlation matrix for a multidimensional dataset and visualise it using a heatmap.*

2. *Identify and visualise outliers in the data using a boxplot.*

Lecture 8. *Modelling and machine learning in Python. Python in applied social research*

Introduction to the scikit-learn and statsmodels libraries. Linear regression: simple and multiple. Polynomial regression as a way to work with nonlinear relationships. Splitting data into training and test samples (train_test_split). The problem of overfitting. Metrics for evaluating the quality of regression models: MAE, MSE, R2. Specific features of libraries for text analysis (NLTK, spaCy) in sociology. Working with time series in economics (statsmodels library, trend decomposition). Obtaining data from social networks and web scraping (BeautifulSoup, Requests). Automation of statistics collection.

Assignments for independent study:

1. *Build a linear regression model to predict real estate prices (or another indicator) on a test dataset.*

2. *Compare quality metrics for linear and polynomial models.*

3. *Write a simple script to retrieve data from an open API (e.g., weather, exchange rates, or news).*

4. *Interpret the results of a regression model in the context of a social phenomenon (e.g., the impact of education on income levels).*

Practical classes (computer workshops)

Computer workshop 1. *Creating scripts in R for data analysis and visualisation (6 hours)*

Objective: *to master the basic syntax of R, learn how to import raw data, bring it into a uniform format, and create informative visualisations.*

Plan

1. *Importing a CSV file. Using the dplyr package to filter records, select columns, and create new variables.*

   Grouping data by category and calculating summary statistics (mean, median, standard deviation).

2. *Use ggplot2. Build a distribution histogram and scatter plot with a trend line.*
3. *Saving the results to a new file and graphs in PNG/PDF format.*

Computer workshop 2. *Solving social data analysis problems using the R language (8 hours)*

Objective: *to apply the R statistical toolkit to test hypotheses and search for correlations in real social data.*

Plan

1. *Construction of a correlation matrix for a set of social indicators (e.g., unemployment rate and crime rate). Visualisation of the correlogram.*
2. *Construction of a linear model. Analysis of coefficients, p-values, and R-squared. Testing model residuals for normality.*
3. *Displaying indicators on a map of regions of Ukraine or the world (chloropleth map) using the sf or tmap libraries.*
4. *Formation of a short report with conclusions in R Markdown.*

Computer workshop 3*. Creating Python scripts for data analysis and visualisation (8 hours)*

Objective: *to learn how to use the Pandas and Matplotlib/Seaborn libraries for manipulating tabular data and its graphical representation.*

Plan

1. *Creating a DataFrame, loading data. Using the .info(), .describe(), .value_counts() methods. Working with indexes and slices.*
2. *Finding and replacing missing values. Encoding categorical variables (get_dummies).*
3. *Building a subplot (grid of graphs) using Matplotlib. Creating a boxplot in Seaborn to compare distributions in different groups.*
4. *Converting the date column to datetime format, building a graph of the indicator dynamics over time.*

Computer workshop 4. *Solving social data analysis problems using Python (6 hours)*

Objective: *to implement a complete cycle of building a predictive model based on socio-economic data.*

Plan

5. *Selecting the target variable and predictors. Splitting into X_train, X_test, y_train, y_test.*
6. *Training a linear regression model (LinearRegression from scikit-learn). Training a polynomial regression model for comparison.*
7. *Calculating prediction errors (MSE, MAE) on a test sample. Comparing model effectiveness.*
8. *Analysis of model weights (coefficients). Which social factors have the greatest impact on the target indicator. Visualisation of the regression line against the background of real data.*

Modular control work *(2 hours)*

9. Independent work by students

Individual assignments consist of preparation for computer workshops, studying lecture material, and completing computational work.

The calculation work consists of creating a personal mini-project on the analysis of social data using any programming language. The student chooses the topic independently and approves it with the teacher. If necessary, the teacher may suggest a topic to choose from.

Independent work includes:

- preparation for classroom sessions – 54 hours;

- calculation assignments – 10 hours;

- preparation for the final exam – 4 hours;

- preparation for the test – 6 hours.

Total – 74 hours.

Policy and control

10. Academic discipline policy (educational component)

Students must attach all their work to their personal Google Classroom account. The deadlines for each assignment are indicated in the Google Classroom assignments. Work must be completed in accordance with academic integrity. The policy and principles of academic integrity and ethical behaviour of students are defined in the Code of Honour https://kpi.ua/code. The lecturer may offer students to take online courses on the Coursera platform. Certificates for these courses may also be partially credited in accordance with the Regulations.
The topics of the assignments are aimed at deepening the material covered in the lectures. In the computer lab classes, students solve problems and exercises related to the lecture topics.

11. Types of assessment and the learning outcomes assessment rating system (LOAS)

*Semester assessment: **test**.*
The student's semester rating for the discipline is given by the lecturer and consists of points awarded for:

~ completion of the Modular control work;

~ completion of 4 computer workshops;

~ completion of the calculation assignments.

Criteria for awarding points for the semester:

1) The Modular control work is worth 20 points.
2) Each workshop is worth 14 points - 14x4=56 points.
3) Calculation assignments is worth 24 points.

Criteria for awarding points for tests:

- *"excellent": 95-100% - the applicant has demonstrated comprehensive, systematic and in-depth knowledge of the subject matter; demonstrated the ability to freely perform all tasks required by the programme; mastered the main and additional literature; showed creative abilities in understanding, logical, clear, concise and clear interpretation of the course material; mastered the interconnection of the main concepts of the discipline, their significance for further professional activity*
- *"very good": 85-94% - the applicant has demonstrated systematic knowledge of the subject matter above average; demonstrated the ability to perform all tasks required by the programme well, with minor errors; mastered the main and additional literature; mastered the interrelationship of the main concepts of the discipline and their significance for further professional activity*
- *"Good": 75-84% - the applicant demonstrated generally good knowledge of the course material when performing the tasks provided for in the programme, but made a number of noticeable mistakes; mastered the basic literature; demonstrated systematic knowledge of the discipline; is capable of using and supplementing this knowledge independently in the process of further study and professional activity*
- *"Satisfactory": 65-74% - the applicant demonstrated knowledge of the basic course material to the extent necessary for further study and future professional activity; familiarised themselves with the main literature; has coped with the tasks set out in the programme, but has made a significant number of mistakes or shortcomings in questions during interviews, tests and tasks, etc., the fundamental ones of which can be eliminated independently*
- *"sufficient": 60-64% - the applicant demonstrated knowledge of the basic educational material of the discipline to the minimum extent necessary for further study and future professional activity; familiarised themselves with the basic literature; has basically completed the tasks specified in the programme, but has made mistakes in answering questions during interviews, testing and performing tasks, etc., which he can only correct under the guidance and with the help of a teacher*
- *"unsatisfactory": 30-54% - the applicant had significant gaps in knowledge of the basic course material; made fundamental mistakes in completing the tasks specified in the programme, but is able to independently complete the programme material and prepare to retake the course*
- *"unsatisfactory": 0-29% - the applicant did not have knowledge of a significant part of the course material; made fundamental mistakes in performing most of the tasks specified in the programme or did not perform these tasks*

The condition for the first assessment is a current rating of at least 30% of the planned points for the semester. The condition for the second assessment is a current rating of at least 60% of the planned points.

A prerequisite for admission to the exam is the acceptance of the calculation work and a semester rating of 36 points.
The exam is worth 100 points. The exam is conducted in the form of a written test, which includes three theoretical questions and two practical questions. Each task is scored out of 20 points according to the following criteria:
"excellent", complete answer, at least 90% of the required information, performed in accordance with the requirements for the "skills" level (complete, error-free solution of the task) – 18-20 points;
"good", sufficiently complete answer, at least 75% of the required information, performed in accordance with the requirements for the "skills" level, or there are minor inaccuracies (complete solution of the task with minor inaccuracies) – 14-16 points;
"satisfactory", incomplete answer, at least 60% of the required information, performed in accordance with the requirements for the "stereotypical" level and some errors (task performed with certain shortcomings) – 12 points;
"Unsatisfactory", the answer does not meet the conditions for "Satisfactory" – 0 points.

The sum of the rating points received by the student during the semester is converted to a final grade according to the table.

Table of correspondence between rating points and grades on the university scale:

Points:	***Grade***
100…95	*Excellent*
94	*Very good*
84	*Good*
74…65	*Satisfactory*
64	*Sufficient*
Less than 60	*Unsatisfactory*
Not counted Calculation work or less than 36	*Not admitted*

Work programme for the academic discipline (syllabus):

Compiled by *Associate Professor, Candidate of Physical and Mathematical Sciences, Associate Professor Ivan Oleksandrovych Pyshnohraiev*

Approved by *the AI Department (Minutes No. 14 of 24 June 2025)*

Approved by *the Methodological Commission of the Faculty of Information Systems and Automation (Minutes No. 7 of 25 June 2025)*