National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"

Department of Artificial Intelligence

R AND PYTHON PROGRAMMING LANGUAGES IN STATISTICAL COMPUTING

Working program of the academic discipline (Syllabus)

Details of the academic discipline

Level of higher education	***Second (master)***
Branch of knowledge	*05 Social and behavioral sciences*
Specialty	*054 Sociology*
Educational program	*Social Data Analytics*
Discipline status (code)	*Compulsory, Professional Training Cycle*
Form of education	*full-time*
Year of training, semester	*1st year, spring semester*
Scope of the discipline	*4 ECTS credits, 120 hours* *Lectures: 18 hours, practical classes: 36 hours, independent work: 66 hours.*
Semester control/ control measures	*Examination, modular control work, calculation work*
Lessons schedule	*rozklad.kpi.ua* *1 hour of lectures and 2 hours of computer workshops per week*
Language of teaching	*English*
Information about head of the course / teachers	Lecturer: PhD, Associate Professor Ivan Pyshnograiev, pyshnograiev@wdc.org.ua Computer workshops: PhD, Associate Professor Ivan Pyshnograiev
Placement of the course	Google classroom https://classroom.google.com/c/NjYxNTc2NjUyMDM4

Program of educational discipline

1. Description of the educational discipline, its purpose, subject of study and learning outcomes

Discipline is normative in the educational program. The study of the academic discipline is aimed at the formation, development and consolidation of the acquirers of the following general and professional competencies:
ЗК 01 Ability to abstract thinking, analysis and synthesis,
ФК 13 Ability to use SPSS statistical package and programming languages R and Python to analyze social data.

Upon completion of the course, the student should be able to demonstrate the following program learning outcome of the Educational and Scientific Program:
ПРН 04 Apply scientific knowledge, sociological and statistical methods, digital technologies, specialized software to solve complex problems of sociology and related fields of knowledge,
ПРН Apply programming languages R and Python to analyze social data.

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

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

***be able**:*

- *to analyze data using the R and Python programming languages;*
- *to create scripts and programs for analyzing social data.*

2. Pre-requisites and post-requisites of the discipline (place in the structural and logical scheme of training according to the relevant educational program)

The discipline is based on the knowledge and skills of related disciplines studied in the previous semester and educational level. This discipline precedes ПО 05 «OSINT technologies in sociological research» and can 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 actions 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 visualization using the R language.

1. *Introduction to data visualization, basic types of graphs;*
2. *Advanced visualization tools, map construction;*
3. *Creation of dashboards.*

Topic 1.3. Data analysis in R.

1. *Statement of the problem, main stages;*
2. *Methods of data processing of various types, their formatting, normalization, etc.;*
3. *Preliminary data analysis, correlation and regression analysis;*
4. *Forecasting methods, construction of regression models;*
5. *Methods of evaluating models.*

Topic 1.4. Examples of using R to analyze socio-economic data.

Section 2. Using the Python language 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 actions on them, writing elementary scripts;*
3. *Basic data structures.*
4. *Basic algorithmic elements in the Python language.*
5. *Classes and objects (overview).*

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

1. *Using Pandas to work with data;*
2. *Using Numpy to work with data structures.*

Topic 2.3. Exploratory data analysis.

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

Topic 2.4. Building models using Python.

1. *Linear and multivariate linear regression;*
2. *Polynomial regression.*
3. *Evaluation of the quality of models*

Topic 2.5. Examples of using Python to analyze socio-economic data.

List of computer workshops:

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

4. Educational 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. *Майборода Р.Є Комп'ютерна статистика. Професійний старт. Навчальний посібник. Київський університет», 2020. – 482 с. http://probability.univ.kiev.ua/userfiles/mre/compsta1.pdf*
4. *Методи і моделі інтелектуального аналізу даних. Практикум [Електронний ресурс] : навчальний посібник для студентів, які навчаються за спеціальністю 122 «Комп'ютерні науки», освітньої програми «Системи і методи штучного інтелекту» / Н. І. Недашківська ; КПІ ім. Ігоря Сікорського. – Київ : КПІ ім. Ігоря Сікорського, 2019. – 71 с. https://ela.kpi.ua/handle/123456789/53764*
5. *Цеслів, О. В. Програмування для аналітичних досліджень [Електронний ресурс] : навчальний посібник для здобувача ступеня бакалавра за освітньою програмою Економічна аналітика, спеціальності 051 Економіка. Електронне мережеве видання / О. В. Цеслів ; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 1,55 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2024. – 238 с. https://ela.kpi.ua/handle/123456789/66102.*

Additional:

1. *Аналіз даних. Лабораторний практикум [Електронний ресурс] : навч. посіб. для студ. спеціальності 113 «Прикладна математика» / Н. М. Куссуль, А. Ю. Шелестов, С. А. Тарасенко, Г. О. Яйлимова ; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 582.97 Кбайт). – Київ: КПІ ім. Ігоря Сікорського, 2022. – 28 с. https://ela.kpi.ua/handle/123456789/50425*
2. *Новотарський, М. А. Основи програмування алгоритмічною мовою Python [Електронний ресурс] : навч. посіб. для студ. освітньої програми «Комп'ютерні системи та мережі» спеціальності 123 «Комп'ютерна інженерія» / М. А. Новотарський ; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 17.93 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2022. – 701 с. 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 програмування // Електронний ресурс. Режим доступу: https://coderlessons.com/tutorials/mashinnoe-obuchenie/r-programmirovanie/r-programmirovanie*
6. *Virtualization Technology // Електронний ресурс. Режим доступу: 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

5. Methods of mastering an educational discipline (educational component)

Assignments with detailed instructions and necessary materials will be posted in Google Classroom and must be completed on time.

Section 1. Using the R Language for Solving Data Analysis Tasks

Lecture 1. *Fundamentals of the R Language: Syntax and Data Structures*

History and philosophy of the R language. Overview of the RStudio development environment. Concepts of packages and the CRAN repository. Basic data types: numeric, integer, character, logical. Vector arithmetic. Main data structures: vectors, matrices, lists, and data frames. Control structures (if-else, loops) and writing custom functions. Data import and export (CSV, Excel).

Self-study assignments:

1. *Install R and RStudio, set up the working directory.*
2. *Create a script that loads an external data file, checks variable types, and outputs basic statistics (summary).*

Lecture 2. *Data Visualization in the R Ecosystem*

Base R vs. ggplot2 graphics systems. The «Grammar of Graphics» concept. Building basic graph types: histograms, scatter plots, boxplots. Setting up 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.

Self-study assignments:

1. *Familiarize yourself with the R Graph Gallery.*
2. *Build a complex plot in ggplot2 using a minimum of three layers (data, geometry, theme) and grouping by color.*

Lecture 3. *Data Analysis Methodology in R*
The «Tidy Data» concept and using dplyr and tidyr packages for data manipulation (filtering, aggregation, joining tables). Exploratory Data Analysis (EDA): detecting outliers, checking distributions. Correlation analysis: calculating Pearson/Spearman coefficients, building correlograms. Building linear regression (lm function). Model diagnostics and result interpretation. Quality metrics: R-squared, RMSE. Self-study assignments:

1. *Study the dplyr cheat sheet.*
2. *Clean a «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 data or eurostat). Examples of visualizing demographic changes and economic trends.
Self-study assignments:

1. *Find an R package that provides access to Open Data or data from international organizations.*
2. *Download a set of socio-economic indicators for one country for the last 10 years.*

Section 2. Using the Python Language for Solving Data Analysis Tasks
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. Fundamentals of algorithmization: loops (for, while), conditional operators, list comprehensions. Fundamentals of Object-Oriented Programming: concepts of class, object, method, and attribute.
Self-study assignments:

1. *Write a function in Python that processes a list of numbers (e.g., filters even numbers and squares them).*
2. *Familiarize yourself with the differences in array indexing between R (starts at 1) and Python (starts at 0).*

Lecture 6. *NumPy and Pandas Libraries: The Foundation of Data Science*
The NumPy library: the concept of ndarray, vectorization of calculations, broadcasting. The Pandas library: Series and DataFrame objects. Loading data from various sources. Indexing, slicing, and filtering data (loc, iloc). Data grouping and aggregation (groupby, pivot_table). Handling missing values and duplicates.
Self-study assignments:

1. *Perform a merge operation of two dataframes by a common key.*
2. *Compare the execution speed of mathematical operations using standard Python lists versus NumPy arrays.*

Lecture 7. *Exploratory Data Analysis (EDA) using Python*
Descriptive statistics using Pandas (describe, value_counts). Visualization of distributions and dependencies: Matplotlib and Seaborn libraries. Building histograms, boxplots, heatmaps for correlation matrices. Statistical hypotheses: scipy.stats library (t-test, checking distribution normality).
Self-study assignments:

1. *Build a correlation matrix for a multidimensional dataset and visualize it using a heatmap.*
2. *Detect and visualize outliers in data using a box-and-whisker plot.*

Lecture 8. *Modeling and Machine Learning in Python*
Introduction to scikit-learn and statsmodels libraries. Linear regression: simple and multiple. Polynomial regression as a way to work with non-linear relationships. Splitting data into training and test sets (train_test_split). The problem of overfitting. Metrics for evaluating regression model quality: MAE, MSE, R2.
Self-study assignments:

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.*

Lecture 9. *Python in Applied Social Research*
Specifics of text analysis libraries (NLTK, spaCy) in sociology. Working with time series in economics (statsmodels library, trend decomposition). Obtaining data from social networks and web scraping (BeautifulSoup, Requests). Automating statistics collection.
Self-study assignments:

1. *Write a simple script to obtain data from an open API (e.g., weather, currency exchange rates, or news).*
2. *Interpret the results of a regression model in the context of a social phenomenon (e.g., the impact of education on income level).*

Practical classes (computer workshops)

Computer Workshop 1. *Creating R Scripts for Data Analysis and Visualization (8 hours)*

Goal: To master basic R syntax, learn to import «raw» data, standardize their format, and create informative visualizations.

Plan

1. *Import of a CSV file. Using the dplyr package to filter records, select columns, and create new variables.*
2. *Grouping data by a categorical feature and calculating summary statistics (mean, median, standard deviation).*
3. *Using ggplot2. Building a distribution histogram and a scatter plot with a trend line.*
4. *Saving obtained results to a new file and graphs in PNG/PDF format.*

Computer Workshop 2. *Solving Social Data Analysis Tasks Using R (8 hours)*
Goal: To apply R statistical tools for hypothesis testing and finding relationships in real social data.

Plan

1. *Building a correlation matrix for a set of social indicators (e.g., unemployment rate and crime rate). Visualization of a correlogram.*
2. *Building a linear model. Analysis of coefficients, p-values, and R-squared. Checking model residuals for normality.*
3. *Mapping indicators on a map of Ukrainian regions or the world (choropleth map) using the sf or tmap libraries.*
4. *Creating a short report with conclusions in R Markdown.*

Computer Workshop 3. *Creating Python Scripts for Data Analysis and Visualization (8 hours)*
Goal: To learn how to use Pandas and Matplotlib/Seaborn libraries for manipulating tabular data and their graphical representation.

Plan

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

Computer Workshop 4. *Solving Social Data Analysis Tasks Using Python (8 hours)*
Goal: To implement the full cycle of building a predictive model based on socio-economic data.

Plan

1. *Selection of the target variable and predictors. Splitting into X_train, X_test, y_train, y_test.*
2. *Training a linear regression model (LinearRegression from scikit-learn). Training a polynomial regression model for comparison.*
3. *Calculating prediction errors (MSE, MAE) on the test set. Comparing model performance.*
4. *Analysis of model weights (coefficients). Identifying which social factors most influence the target indicator. Visualizing the regression line against the real data.*

Modular Control Work (2 hours)

Exam (2 hours)

6. The student's independent work

Individual assignments consist of preparation for computer workshops, study of lecture material, and completion of the calculation work.
The calculation work involves creating a personal mini-project on social data analysis using any programming language. The student chooses the topic independently and approves it with the lecturer. If necessary, the lecturer can suggest a topic of their choice.
The student's independent work includes:

- *preparation for classroom sessions – 22 hours;*
- *preparation for the Modular Control Work – 4 hours;*
- *preparation of the calculation work – 10 hours;*
- *preparation for the exam – 30 hours.*

Total – 66 hours.

Policy and control

7. Policy of academic discipline (educational component)

Students must attach all assignments in their personal Google Classroom account. Deadlines for each assignment are indicated in the tasks within Google Classroom. Assignments must be completed in compliance with academic integrity. The policy and principles of academic integrity, and the ethical behavior of students, are defined in the Honor Code https://kpi.ua/code. The lecturer may suggest that students take online courses on the Coursera platform. Additionally, certificates from these courses may be partially credited in accordance with the Regulations https://document.kpi.ua/files/2020_7-124.pdf.
The topics of the assignments are aimed at deepening the understanding of the lecture material. During computer workshops, problems and exercises related to the lecture topics are solved.

8. Types of control and rating system for evaluating learning outcomes (ELO)

Semester control: *Exam.*
The student's semester rating for the discipline is assigned by the lecturer and consists of points received for:

- *~ completion of the modular control work;*
- *~ completion of 4 computer workshops;*
- *~ completion of the calculation work;*
- *~ exam.*

Criteria for awarding points during the semester:

1. *The modular control work is valued at 10 points.*
2. *Each of the workshops is valued at 7 points – 7x4=28 points.*
3. *The calculation work is valued at 12 points.*

Criteria for awarding points for control measures:

- ***«Excellent»: 95-100%*** *– the student demonstrated comprehensive, systematic, and deep knowledge of the educational material of the discipline; demonstrated the ability to freely perform all tasks provided by the program; mastered primary and supplementary literature; showed creative abilities in understanding, and in the logical, clear, concise, and distinct interpretation of the educational material; mastered the relationship of the basic concepts of the discipline and their significance for further professional activity.*
- ***«Very good»: 85-94%*** *– the student demonstrated systematic knowledge of the educational material of the discipline above the average level; demonstrated the ability to perform all tasks provided by the program well, while making minor errors; mastered primary and supplementary literature; mastered the relationship of the basic concepts of the discipline and their significance for further professional activity.*
- ***«Good»: 75-84%*** *– the student demonstrated generally good knowledge of the educational material when performing the tasks provided by the program, but made a number of noticeable errors; mastered primary literature; showed a systematic character of knowledge in the discipline; is capable of their independent use and replenishment in the process of further educational work and professional activity.*
- ***«Satisfactory»: 65-74%*** *– the student demonstrated knowledge of the main educational material of the discipline in the volume necessary for further study and future professional activity; familiarized themselves with the primary literature; coped with the completion of tasks provided by the program, but made a significant number of errors or shortcomings in answers to questions during interviews, testing, and task execution, the principal ones of which they can eliminate independently.*
- ***«Sufficient»: 60-64%*** *– the student demonstrated knowledge of the main educational material of the discipline in the minimum volume necessary for further study and future professional activity; familiarized themselves with the primary literature; mainly completed the tasks provided by the program, but made errors in answers to questions during interviews, testing, and task execution, which they can eliminate only under the guidance and with the help of the teacher.*
- ***«Unsatisfactory»: 30-59%*** *– the student had significant gaps in the knowledge of the main educational material; made principal errors when performing the tasks provided by the program, but is capable of independently reworking the program material and preparing for the retake of the discipline.*
- ***«Unsatisfactory»: 0-29%*** *– the student had no knowledge of a significant part of the educational material of the discipline; made principal errors when performing the majority of the tasks provided by the program or did not perform these tasks.*

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

The condition for the second attestation is a current rating of at least 60% of the planned points.

A necessary condition for admission to the exam is the successful completion of the calculation work and a semester rating of at least 36 points. Students with fewer than 36 points are not admitted to the exam. The exam is valued at 50 points. The exam is conducted in the form of a written paper consisting of two theoretical questions and two practical ones.

Each theoretical task is valued at 10 points according to the following criteria:

- ***«Excellent»**, complete answer, at least 90% of the required information, executed according to the «skills» level requirements (complete, error-free solution of the task) – **9-10 points**;*
- ***«Good»**, sufficiently complete answer, at least 75% of the required information, executed according to the «skills» level requirements or containing minor inaccuracies (complete solution of the task with minor inaccuracies) – **8 points**;*
- ***«Satisfactory»**, incomplete answer, at least 60% of the required information, executed according to the «stereotypical» level requirements and containing some errors (task executed with certain shortcomings) – **7 points**;*
- ***«Unsatisfactory"**, the answer does not meet the conditions for «satisfactory» – **0-6 points**.*

Each practical task is valued at 15 points according to the following criteria:

- ***«Excellent»**, complete answer, at least 90% of the required information, executed according to the «skills» level requirements (complete, error-free solution of the task) – **13-15 points**;*
- ***«Good»**, sufficiently complete answer, at least 75% of the required information, executed according to the «skills» level requirements or containing minor inaccuracies (complete solution of the task with minor inaccuracies) – **11-12 points**;*
- ***«Satisfactory»**, incomplete answer, at least 60% of the required information, executed according to the «stereotypical» level requirements and containing some errors (task executed with certain shortcomings) – **9-10 points**;*
- ***«Unsatisfactory»**, the answer does not meet the conditions for «satisfactory» – **0-8 points**.*

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

Table of correspondence of rating points to grades on the university scale:

Points:	***Rating***
100…95	*Excelent*
94…85	*Very good*
84…75	*Good*
74…65	*Satisfactorily*
64…60	*Enough*
Less than 60	*Unsatisfactorily*
calculation work is not included or less than 36	*Not allowed*

Working program of the academic discipline (syllabus):

Compiled by Ivan Pyshnograiev, Associate Professor, Candidate of Physical and Mathematical Sciences.

Approved by the Department of AI (protocol № 14 from 11.05.2024)

Agreed by the Methodical Commission of the ES IASA (protocol № 10 from 24.06.2024)