Federal State Autonomous Educational Institution
for Higher Professional Education
National Research University Higher School of Economics

Утверждена Академическим советом
образовательной программы
«30» августа 2019 г., № протокола 7
Академический руководитель
образовательной программы

Д.А. Щербаков _____

Data Culture
Course Syllabus

Part 1. Course Information

Instructors: Anna Kuzina
Office: TBA
Office Hours: By appointment
E-mail: av.kuzina@yandex.ru

Course Description

This is a required course for students of all undergraduate programs at the HSE. The course provides students with basic knowledge of machine learning with special focus on the application of modern analytic tools in area studies and international studies.

The course consists of three parts. In the first part, students learn how to collect data from various sources. The second part introduces some concepts from statistics and methods of data analysis. In the third part of the course, students are expected to conduct their own research and analyze data on bilateral trade and investment. Students achieve excellent results by doing a considerable number of practical exercises in class, completing individual assignments and taking part in group research projects.

Learning Outcomes

After completion of this course, students will be able to:

- Collect statistical data from open databases
- Represent data using graphs, charts and plots
- Use key concepts from statistics to describe data
- Analyze numeric data using clustering, correlation methods
- Use content analysis for textual data
- Analyze time series
- Collect and interpret qualitative data
- Use Microsoft Excel software at an advanced level
- Use R software environment at an elementary level
- To organize work in groups

Textbook & Course Materials

Required texts:
Jake VanderPlas, Python Data Science Handbook, O'Reilly Media, 2016. Available online at: https://jakevdp.github.io/PythonDataScienceHandbook/
Bluman, Allan G. Elementary statistics. McGraw Hill, 2013.
Grimwade, Nigel. International trade: new patterns of trade, production and investment. Routledge, 2003.

Other necessary materials, including lecture PPT files and related articles, will be provided in form of electronic files.

Teaching & Learning Strategies

The sessions of the course take place every week. We do not distinguish between lectures and seminars. Each session includes practical guidelines and plenty of exercises. Some sessions introduce theoretical framework.

BYOD (Bring Your Own Device)

Our school has a blended learning approach to learning and teaching where students learn through seamless integration of technology-enhanced strategies and face-to-face activities. The blended learning approach requires students to use their own devices (tablets and/or laptops that can be connected to school Wi-Fi) to access learning resources and to participate in class activities. Students are encouraged to install all the necessary apps to support their learning and bring their own devices to their classes.

Group work

The material in the course is designed to develop analysis skills, rather than traditional memory techniques. In the third part of the course, students are expected to form mini-groups (3-4 persons) for conducting research projects. There are several reasons for

that. Firstly, work in groups enables dealing with large amounts of information successfully. Secondly, work in groups generates synergetic effects when students share their knowledge. Meeting with fellow students to discuss the subject matter has proven to be highly effective.

Lecture/Seminar/Homework Hours

No	Topic	Contact hours	Home work	Hours total
1.	Introduction and course overview	2	2	4
2.	Introduction to Python: data structures, loops	4	2	6
3.	Introduction to Python: pandas and numpy	4	4	8
4.	Data collection	2	4	6
5.	Graphical analysis	4	4	8
6.	Data description	4	6	10
7.	Midterm exam	2	0	2
8.	Line fitting, standard errors, and covariance	2	4	6
9.	Correlation and regression	4	6	10
10.	Applications of regression analysis	2	6	8
11.	Creating Indices	2	2	4
12.	Exploratory Data Analysis	4	6	10
13.	Data and methodological tools for group projects	2	4	6
14.	Quantitative analysis	2	6	8
15.	Qualitative analysis	2	6	8
16.	Presentation of group research projects	4	6	10
	Total	46	68	114

Part 2. Grading Policy

Each student is required to attend every session. The grade for this course is based on the following components:
(1) Attendance. Three absences are excused throughout the course. No documentation is required. Each additional absence beyond the allowed number will lower your final grade by one point (e.g., 8 to 7);
(2) Midterm Exam. In-class midterm exam will constitute 30 percent of the final grade;
(3) Group work. Presentation of group research projects will bring you additional 30 percent of the final grade;
(4) Home assignments. The remaining 40 percent will be graded by two individual assignments. Note that there will be no final exam at the end of the course.

Attendance	**N/A**
Group work	**40%**
Midterm Exam	**30%**
Home assignments	**30%**

Midterm Exam

We will have a midterm exam in Week 11. The midterm exam will cover topics 1-6. If you are absent at the midterm exam, you will get 0 grade. You will have an opportunity to take the midterm exam on another date only if your absence is due to medical reasons (confirming documents required).

Home assignments

Two assignments will be administered through the course. They are intended as an opportunity for revision of the course material. The first home assignment is to be submitted to the instructor's email at the end of Week 14. The second home assignment is to be submitted at the end of Week 15. These two assignments cover topics 9 and 10 respectfully. You should analyze real data using appropriate methods and techniques in Python. In fairness to students who complete assignments on time, late assignments will be given 0 grade. Students who do not complete assignments will get 0 grade.

Group work

Each group consists of 3-4 students who conduct mini-research together. Each group chooses two countries, one of which should be from the Asian region. The purpose of their research is to describe the current state of trade and investment ties between these two countries, find the determinants of the bilateral economic relations and predict how the relations will develop in a medium-term perspective. Students are encouraged to analyze data from official sources with the methods studied in this course using Python. At the final sessions, each group will present the results of their

research. All group members will get the same grade, which was given for their presentation.

Part 3. Topic Outline/ Schedule

Weekly Schedule:

- Week 1. Introduction and course overview
- Weeks 2, 3. Introduction to Python: data structures, loops
- Weeks 4, 5. Introduction to Python: pandas and numpy
- Week 6. Data collection
- Weeks 7, 8. Graphical analysis
- Weeks 9, 10. Data description
- Week 11. Midterm exam
- Week 12. Line fitting, standard errors, and covariance
- Week 13, 14. Correlation and regression
- Week 15. Applications of regression analysis
- Week 16. Creating indices
- Weeks 17, 18. Exploratory Data Analysis
- Week 19. Data and methodological tools for group projects
- Week 20. Quantitative analysis
- Week 21. Qualitative analysis
- Weeks 22, 23. Presentation of group research projects

Session Outlines

Week 1. Introduction and course overview

(1) Learning Objectives

After this session, students should be able to acknowledge:

- The goal of the course
- What machine learning is
- The advantages and limitations of modern analytic tools
- The requirements and grading policy
- The course schedule

(2) Session Outline

1. Machine learning and its application
2. Data analysis
    a. Why do we need data analysis?
    b. Types of data
    c. Data sources
    d. Measurement problems

Weeks 2, 3. Introduction to Python: data structures, loops

(1) Learning Objectives

After this session, students should be able to:

- Understand basic data structures in python
- Use IPython environment, to:
  - Create new variables
  - Use if, elif, else structures
  - Perform arithmetic calculations
  - Use for loop
  - Create functions

(2) Session Outline

1. Advantages of python for data analysis
2. Introduction to IPython environment: Anaconda, Google Colaboratory
3. Simple manipulations with different types of R objects:
   a. Numbers and vectors
   b. Lists
   c. Functions
4. Loops: for, while

(3) Recommended Readings

Muenchen, Robert A. The Popularity of Data Science Software.
http://r4stats.com/articles/popularity/

Jake VanderPlas, Python Data Science Handbook, Chapter 1
https://jakevdp.github.io/PythonDataScienceHandbook/01.00-ipython-beyond-normal-python.html

Weeks 4, 5. Introduction to Python: Pandas and Numpy

(1) Learning Objectives

After this session, students should be able to:

- Use the main feature of numpy and pandas packages in python

(2) Session Outline

1. Numpy:
   a. Arrays
   b. Calculations with arrays, broadcasting
   c. Indexing
2. Pandas:
   a. Series and DataFrame objects
   b. Indexing and data selection
   c. Aggregation and grouping
   d. Combining datasets: merge, join, concat

(3) Recommended Readings

Jake VanderPlas, Python Data Science Handbook, Chapters 2-3
https://jakevdp.github.io/PythonDataScienceHandbook/

Week 6. Data collection

(1) Learning Objectives

After this session, students should be able to:

- Understand what open data is
- Import data in Python from electronic files (.txt, .csv, .xls)
- Import data in Python directly from World Bank and UN Comtrade databases

(2) Session Outline

1. Open data
    a. Definition
    b. Term of use
    c. Databases
2. Import data in Python
    a. From textual files
    b. From Excel files
    c. Directly from ODBC sources using API (examples include World Bank, UN Comtrade)

Weeks 7, 8. Graphical analysis

(1) Learning Objectives

After this session, students should be able to:

- Organize data using a frequency distribution
- Represent data using histograms, bar graphs, Pareto charts, time series graphs, pie graphs, dotplots, scatterplots
- Compare values across geographical regions/countries using map charts
- Represent historical data using dynamic charts
- Create reports in IPython notebooks
- Interpret graphs

(2) Session Outline

1. Frequency distribution
    a. How to construct a frequency distribution
    b. Drawing a histogram
    c. Distribution shapes

2. Other graphs: bar graphs, Pareto charts, time series graphs, pie graphs, dotplots, scatterplots
    a. In which cases we use them
    b. How to construct graphs in Python
3. Creating reports in IPython notebooks: how to combine text and code
4. Misleading graphs

(3) Recommended Readings

Bluman, Allan G. Elementary statistics. McGraw Hill, 2013. Chapter 2.

Jake VanderPlas, Python Data Science Handbook, Chapter 4
https://jakevdp.github.io/PythonDataScienceHandbook/

Weeks 9, 10. Data description

(1) Learning Objectives

After this session, students should be able to:

- Summarize data, using measures of central tendency, such as the mean, median, mode, midrange, etc.
- Describe data, using measures of variation, such as the range, variance, and standard deviation
- Identify the position of a data value in a data set, using various measures of position, such as percentiles, deciles, and quartiles
- Use the techniques of exploratory data analysis, including boxplots, to discover various aspects of data

(2) Session Outline

1. Measures of central tendency
    a. The mean, midrange and geometric mean
    b. The median
    c. The mode
    d. The weighted mean
    e. In which cases we use each measure of central tendency
2. Measures of variation
    a. Range
    b. Population variance and standard deviation
    c. Sample variance and standard deviation
    d. Coefficient of variation
3. Measures of position
    a. Percentiles, deciles and quartiles
    b. How to deal with outliers
4. Exploratory data analysis: boxplots

(3) Recommended Readings

Bluman, Allan G. Elementary statistics. McGraw Hill, 2013. Chapter 3.

Week 11. Midterm exam

At the midterm exam, students are required to get data from open database, represent the data graphically and describe the data. To complete the task students need to take the following steps: (1) import a data set indicated by the instructor from World Bank database; (2) show the frequency distribution for the latest period using histogram; (3) show how the values vary across countries using map chart; (4) summarize the data for different periods using boxplots; (5) identify the percentile ranks of Russia and the country of specialization; (6) compile the results in a HTML document, add comments and/or conclusions. The task should be completed within the time limit of one hour and 20 minutes. Students are allowed to use their materials, such as notes and previously done exercises.

Week 12. Line fitting, standard errors, and covariance

(1) Learning Objectives

After this session, students should be able to:

- Understand the basics of the line fitting process
- Be able to estimate variation in the data using the analysis of residuals
- Be able to assess the normality of residuals
- Understand the concept and calculation of covariance
- Use Python software for estimating covariance and display residual plots

(2) Session Outline

1. Line-fitting
    a. Graphical techniques for displaying relationship between two numerical variables
    b. Plotting data with the fitted line

2. Standard errors
    a. Calculation of residuals
    b. Residual analysis using a histogram, residual plot, and Q-Q plot
    c. Detecting heteroscedasticity and outliers

3. Covariance
    a. Covariance of two random variables
    b. Calculation of sample and population covariance
    c. Interpreting the sign of covariance

(3) Recommended Readings

Diez, David M., Christopher D. Barr, and Mine Cetinkaya-Rundel. OpenIntro Statistics. OpenIntro, 2012. Chapter 7.

Keller, Gerald. Statistics for Management and Economics, Abbreviated. Cengage Learning, 2015. Chapter 4.

Weeks 13, 14. Correlation and regression

(1) Learning Objectives

After this session, students should be able to:

- Compute the correlation coefficient
- Compute the equation of the regression line

(2) Session Outline

1. Correlation
    a. How to find a relationship between two variables?
    b. Pearson product-moment correlation coefficient: assumptions and properties
    c. Correlation and causation
2. Regression
    a. Ordinary least squares (OLS) method
    b. Simple bivariate regression

(3) Recommended Readings

Bluman, Allan G. Elementary statistics. McGraw Hill, 2013. Chapter 10.

Week 15. Applications of regression analysis

(1) Learning Objectives

After this session, students should be able to:

- Interpret the linear regression's coefficients
- Understand the concept of control variables
- Test the null hypothesis
- Compute the coefficient of determination
- Find a prediction interval

(2) Session Outline

1. Multivariate regression
2. Control variables in multivariate regression
3. Testing the hypothesis of the insignificance of a coefficient

4. Comparing models and selecting the best model
5. Prediction using OLS: assumptions, prediction interval

(3) Recommended Readings

Keller, Gerald. Statistics for Management and Economics, Abbreviated. Cengage Learning, 2015. Chapter 16.

Diez, David M., Christopher D. Barr, and Mine Cetinkaya-Rundel. OpenIntro Statistics. OpenIntro, 2012. Chapter 8.

Week 16. Creating indices

(1) Learning Objectives

After this session, students should be able to:

- Understand the methodologies underlying index creation process
- Explain the components of an index
- Understand the limitations of any given index

(2) Session Outline

1. The Hobbes Index
2. The Human Development Index (HDI)
3. The Cingranelli Richards (CIRI) Human Rights Index
    a. Coding scores
    b. Decomposing the CIRI index
    c. The limitations of the CIRI index

(3) Recommended Readings

De Mesquita, Bruce Bueno, et al. The Logic of Political Survival. MIT press, 2005. Pages 461-465.
https://books.google.ru/books?id=1PlRlcgQdpMC&pg=PA463&lpg=PA463&dq=Logic+of+political+survival+hobbes+index&source=bl&ots=1guXLzNB_F&sig=ACfU3U0E1pF5UO6NsiIzRBilsllHAGptRA&hl=ru&sa=X&ved=2ahUKEwiHxa_chaLhAhWLyKYKHWaKAvIQ6AEwAnoECAcQAQ#v=onepage&q=Logic%20of%20political%20survival%20hobbes%20index&f=false

Human Development Indices and Indicators: 2018 Statistical Update. Technical Notes Calculating The Human Development Indices - Graphical Presentation.
http://hdr.undp.org/sites/default/files/hdr2018_technical_notes.pdf

Cingranelli, David L. and David L. Richards. 2010. "The Cingranelli and Richards (CIRI) Human Rights Data Project." Human Rights Quarterly 32.2: 401-424.

Cingranelli, David L. and David L. Richards. 2014. The Cingranelli-Richards (CIRI)

Human Rights Data Project Coding Manual Version 5.20.14.
http://www.humanrightsdata.com/p/data-documentation.html

Cingranelli, David L., Mikhail Filippov and Skip Mark. The "CI-RIGHTS" data project.
https://docs.google.com/document/d/1ZxuitRNKpTcEpwi_1p8u0IRnib7ggtKF9TQcf_jbmo0/edit?usp=sharing

Weeks 17, 18. Exploratory Data Analysis

(1) Learning Objectives

After this session, students should be able to:

- Analyze datasets to summarize their main characteristics
- Learn the necessary steps to understand the underlying structure of data and obtain insights
- Apply key EDA techniques to discover patterns in data and formulate business hypotheses

(2) Session Outline

1. Overview of differences between EDA and traditional research designs
2. The key EDA techniques for understanding the structure of a dataset
3. How to formulate a business hypothesis and put it to test?
4. The basic principles behind effective data visualization and storytelling with data

(3) Recommended Readings

Zhao, Yanchang. R and data mining: Examples and case studies. Academic Press, 2012.
Bluman, Allan G. Elementary statistics. McGraw Hill, 2013. Chapter 3.

Knaflic, C. Storytelling with data: A data visualization guide for business professionals. John Wiley & Sons, 2015. Chapter 2.

Week 19. Overview of data and methodological tools for group projects

(1) Learning Objectives

After this session, students should be able to:

- Understand the contents and structure of the datasets, which will be used for group projects
- Apply the methods learned previously in the course to the research task

(2) Session Outline

1. Overview of the CIRI and TIES datasets
2. Instructions on merging datasets based on Correlates of war (COW) country codes using Excel and R software
3. Overview of the methods to be used in group research projects
4. Discussion of desired end results of group project work

(3) Recommended Readings

Morgan, Clifton T., Navin Bapat, and Yoshi Kobayashi. 2014. "The Threat and Imposition of Sanctions: Updating the TIES dataset." Conflict Management and Peace Science 31(5): 541-558.

Morgan, Clifton T., Navin Bapat, and Yoshiharu Kobayashi. "Threat and Imposition of Sanctions (TIES) Data 4.0 Users' Manual Case Level Data."
Chapel Hill: University of North Carolina (2013).
http://sanctions.web.unc.edu/

Cingranelli, David L., Mikhail Filippov and Skip Mark. The "CI-RIGHTS" data project.
https://docs.google.com/document/d/1ZxuitRNKpTcEpwi_1p8u0IRnib7ggtKF9TQcf_jbmo0/edit?usp=sharing

Cingranelli, David L. and David L. Richards. 2014. The Cingranelli-Richards (CIRI) Human Rights Data Project Coding Manual Version 5.20.14.
http://www.humanrightsdata.com/p/data-documentation.html

Additional datasets:
COW Country Codes, http://www.correlatesofwar.org/data-sets/cow-country-codes

Week 20. Quantitative Analysis

(1) Learning Objectives

After this session, students should be able to:

- Analyze longitudinal human rights trends in country-year datasets
- Consult with datasets' coding manuals
- Map several datasets using a common key
- Present the data in visually informative format and refine plots for effective presentation

(2) Session Outline

1. Case selection for quantitative analysis
2. Preparing the data for analysis
3. Performing variable transformations
4. Applying the method to a data sample
5. Presenting the results using graphical techniques

(3) Recommended Readings

Carnegie, Allison, and Nikolay Marinov. "Foreign aid, human rights, and democracy promotion: Evidence from a natural experiment." American Journal of Political Science 61.3 (2017): 671-683.

Week 21. Qualitative analysis

(1) Learning Objectives

After this session, students should be able to:

- Understand the features of qualitative research
- Supplement quantitative findings with qualitative evidence
- Apply case study as a research method

(2) Session Outline

1. Qualitative research
    a. Advantages and limitations of qualitative methods
    b. An overview of qualitative research tools
2. Case studies
    a. Finding the qualitative evidence for group research project
    b. Integrating qualitative findings with the results of statistical analysis

(3) Recommended Readings

Human Rights Reports | U.S. Department of State
https://www.state.gov/j/drl/rls/hrrpt/

Reports | Freedom House
https://freedomhouse.org/reports

Country Profiles | Amnesty International
https://www.amnesty.org/en/countries/

Weeks 22, 23. Presentation of group research projects

Each group should choose a country (ideally, in Asia) that had been under sanctions during a time frame for which the data are available (1984-2015). The goal is to describe human rights trends before the episode of sanctions and after their implementation. Students must perform data analyses using the sources mentioned in the recommended readings and use the methods studied in this course. All data transformations, data analyses, and data visualizations must be performed in Excel and R software.

During the final sessions, each group will present the results of their team work. The structure of the presentation is to be organized as follows: (1) a brief description of the data; (2) description of the sample and a brief country background; (3) description of the method; (4) data visualization of human rights trends; (5) qualitative evidence to explain the chart and the model's results. After each presentation there will be a Q&A session. The time limit for a presentation + Q&A session is 20 minutes.