Higher School of Economics 2018

Утверждена Академическим советом
образовательной программы
«5» сентября 2018 г., № протокола_1
Академический руководитель
образовательной программы

Д.А. Щербаков

Data Culture

Part 1. Course Information

Instructor: Maria Korosteleva
Office: Room 316, 17 Malaya Ordynka, Moscow
Office Hours: by appointment
E-mail: mkorosteleva@hse.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.

Prerequisites

Mathematics for Economics and Business is a formal prerequisite for students wishing to enroll in Data Culture.

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:

Zhao, Yanchang. R and data mining: Examples and case studies. Academic Press, 2012.
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.	Use of library resources	4	2	6
3.	Introduction to R	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.	Classification and clustering	2	4	6
9.	Correlation and regression	4	6	10
10.	Content analysis in texts	2	6	8
11.	Data on international trade and investment	2	2	4
12.	Time series analysis	4	6	10
13.	Trade composition and intra-industry trade index	2	4	6
14.	Finding the FDI determinants	2	6	8
15.	Qualitative analysis: news digest and case studies	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) Participation. 10 percent of your final grade will be based on your class preparedness and in-class discussion. The more you talk in class, the better your participation score will be; (3) Midterm Exam. In-class midterm exam will constitute 30 percent of the final grade; (4) Group work. Presentation of group research projects will bring you additional 30 percent of the final grade; (5) Home assignments. The remaining 30 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**
Participation	**10%**
Group work	**30%**
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 R software environment. 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 R and Excel software. 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. Use of library resources
- Weeks 4, 5. Introduction to R
- Week 6. Data collection
- Weeks 7, 8. Graphical analysis
- Weeks 9, 10. Data description
- Week 11. Midterm exam
- Week 12. Classification and clustering
- Week 13, 14. Correlation and dependence
- Week 15. Semantic analysis in texts
- Week 16. Data on international trade and investment
- Weeks 17, 18. Time series analysis
- Week 19. Trade composition and intra-industry trade index
- Week 20. Finding the FDI determinants
- Week 21. Qualitative analysis: news digest and case studies
- 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
3. Limitations of data analysis

Weeks 2, 3. Use of library resources

(1) Learning Objectives

After this session, students should be able to:

- Use search engine query languages
- Find relevant scientific literature in online libraries
- Understand the meaning of research metrics
- Understand the advantages and disadvantages of different search engines

(2) Session Outline

1. Query languages in different search engines
2. Use of online libraries (HSE resources)
   a. Logical operators
   b. Search by field in publication
   c. Filters to narrow the search field
   d. Finding relevant search results using research metrics
3. Use of Google Scholar: advantages and disadvantages.
4. Use of reference management software (Zotero, Mendeley, etc.)
   a. Quick download, changing location and renaming of electronic files with scientific publications
   b. Extracting bibliographic information from pdf-metadata
   c. Citation and bibliography formatting
   d. Creating own library and its export
   e. Advantages and disadvantages of reference management software.

Weeks 4, 5. Introduction to R

(1) Learning Objectives

After this session, students should be able to:

- Understand why we use data analysis software
- Have an idea of the variety of data analysis software and their degree of popularity
- Acknowledge the main features of R language
- Perform simple manipulations in R

(2) Session Outline

1. Data analysis software
   a. Application in real life
   b. The variety of data analysis software
   c. R environment: advantages and the scope of use
2. R language: syntax

3. Simple manipulations with different types of R objects:
    a. Numbers and vectors
    b. Factors
    c. Arrays and matrices
    d. Lists and data frames
    e. Functions
4. R packages

(3) Recommended Readings

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

Venables, W. N., and D. M. Smith. An Introduction to R: Notes on R: A Programming Environment for Data Analysis and Graphics (Version 3.4.4). 2018. Chapters 1-6, 10, 13.
https://cran.r-project.org/doc/manuals/R-intro.pdf

Week 6. Data collection

(1) Learning Objectives

After this session, students should be able to:

- Understand what open data is
- Import data in R from electronic files (.txt, .csv, .xls)
- Import data in R 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 R
    a. From textual files
    b. From Excel files
    c. Directly from ODBC sources using API (examples include World Bank, UN Comtrade)

(3) Recommended Readings

Venables, W. N., and D. M. Smith. An Introduction to R: Notes on R: A Programming Environment for Data Analysis and Graphics (Version 3.4.4). 2018. Chapter 7.
https://cran.r-project.org/doc/manuals/R-intro.pdf

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
- Include graphs constructed in R in textual reports
- 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 the graphs in Excel
    c. How to construct graphs in R
3. googleVis package for R
    a. Intensity maps and geo charts
    b. Motion charts
4. Creating reports in R: how to combine text and code
5. Misleading graphs

(3) Recommended Readings

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

Venables, W. N., and D. M. Smith. An Introduction to R: Notes on R: A Programming Environment for Data Analysis and Graphics (Version 3.4.4). 2018. Chapter 8, 12. https://cran.r-project.org/doc/manuals/R-intro.pdf

Gesmann, Markus, and Diego de Castillo. Introduction to googleVis 0.6.2. 2017. https://cran.r-project.org/web/packages/googleVis/vignettes/googleVis.pdf

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) show how the values changed over time using motion chart; (5) summarize the data for different periods using boxplots; (6) identify the percentile ranks of Russia and the country of specialization; (7) 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 a half. Students are allowed to use their materials, such as notes and previously done exercises.

Week 12. Classification and clustering

(1) Learning Objectives

After this session, students should be able to:

- Understand the features of classification and clustering methods and their scope of use
- Understand the advantages and disadvantages of different clustering algorithms
- Cluster data in R using different techniques

(2) Session Outline

1. Data classification and differentiated approach
   a. Why do we need to classify data?
   b. When data grouping makes sense
   c. How to classify data
   d. Clustering and multidimensional clustering
2. Clustering algorithms
   a. K-means and k-medoid clustering
   b. Hierarchical clustering
   c. Density based clustering

   d. Advantages and disadvantages of each algorithm

3. Clustering techniques in R

(3) Recommended Readings

Chitra, K., and D. Maheswari. A Comparative Study of Various Clustering Algorithms in Data Mining. International Journal of Computer Science and Mobile Computing, vol. 6, issue 8, August 2017.
http://www.ijcsmc.com/docs/papers/August2017/V6I8201725.pdf

Zhao, Yanchang. R and data mining: Examples and case studies. Academic Press, 2012.

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
- Test the null hypothesis
- Compute the coefficient of determination
- Find a prediction interval

(2) Session Outline

1. Correlation
   a. How to find a relationship between two variables?
   b. Pearson correlation coefficient: assumptions and properties
   c. Correlation and causation
2. Regression
   a. Ordinary least squares (OLS) method in R
   b. Multiple regression
   c. Testing the hypothesis of the insignificance of a coefficient
   d. Comparing models
   e. Predicting using OLS: assumptions, prediction interval

(3) Recommended Readings

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

Venables, W. N., and D. M. Smith. An Introduction to R: Notes on R: A Programming Environment for Data Analysis and Graphics (Version 3.4.4). 2018. Chapter 11.
https://cran.r-project.org/doc/manuals/R-intro.pdf

Week 15. Content analysis in texts

(1) Learning Objectives

After this session, students should be able to:

- Create a document-term matrix
- Compare the lexical composition of different text documents

(2) Session Outline

1. Qualitative and quantitative content analysis in texts
2. Creating a document-term matrix in R
    a. Cleaning texts from meaningless symbols
    b. Removing stop words
    c. Stemming
    d. Weighting terms using term frequency-inverse document frequency
3. Comparing document-term matrices for different documents

Week 16. Data on international trade and investment

(1) Learning Objectives

After this session, students should be able to:

- Understand how trade and investment flows are recorded in national statistics
- Understand and use key terms to describe data on international trade and investment
- Acknowledge possible causes of discrepancies between data from different sources
- Interpret data on international trade and investment
- Use appropriate data sources

(2) Session Outline

1. Data on international trade
    a. Trade in goods and trade in services: technical and statistical differences
    b. CIF and FOB values in customs statistics
    c. Possible causes of discrepancies between data from different sources
    d. Data sources
2. Data on foreign direct investment (FDI)
    a. Investment values shown in the balance of payments
    b. Investment flows and investment stock
    c. Fallacies in interpreting data on FDI
    d. Data sources

Weeks 17, 18. Time series analysis

(1) Learning Objectives

After this session, students should be able to:

- Describe a time series
- Explore a time series and identify trends, seasonality and random behavior
- Check if a time series is stationary and stationarize the series
- Construct ARIMA models and make predictions

(2) Session Outline

1. Description of a time series
    a. Absolute and relative growth rate
    b. Average growth rate
    c. Indices
2. Stationary series and random walks. Stationarity tests
3. How to identify trends and seasonal effects in time series
4. ARMA time series modelling
    a. Auto-regressive time series model

   b. Moving average time series model
   c. Identifying the type of stationary series using ACF & PACF plots
5. ARIMA model
   a. How to make a time series stationary
   b. Constructing ARIMA model
   c. Making predictions

Week 19. Trade composition and intra-industry trade index

(1) Learning Objectives

After this session, students should be able to:

- Understand the features and scope of use of international trade classifications
- Differentiate between classical and new trade theories
- Measure the level of intra-industry trade between two countries
- Perform simple calculations using formulas in Excel and R

(2) Session Outline

1. Classifications for international trade and their scope of use
2. Inter-industry trade and intra-industry trade
   a. Two theories of international trade: Heckscher-Ohlin theory and intra-industry trade theory
   b. Intra-industry trade: the problems of definition and measurement
   c. Forms of intra-industry trade
   d. The determinants of intra-industry trade
   e. Computing Grubel-Lloyd index in Excel and R

(3) Recommended Readings

Grimwade, Nigel. International trade: new patterns of trade, production and investment. Routledge, 2003. Chapters 2, 3.

Lloyd, Peter J. How Intra-Industry Trade Changed Our Perception of the World Economy. Singapore Economic Review, vol. 49, no. 1, April 2004, pp. 1–17.

Week 20. Finding the FDI determinants

(1) Learning Objectives

After this session, students should be able to:

- Understand the reasons underlying international capital flow
- Use rankings to assess investment attractiveness of countries
- Find correlations between investment attractiveness factors and the FDI flows
- Identify the determinants of FDI outflows of a given country

(2) Session Outline

1. Causes for international capital flow
2. Factors of FDI attractiveness of countries
   a. Investment attractiveness rankings (World Bank, Economist Intelligence Unit, Ernst&Young, Milken Institute, etc.)

   b. Methodologies of the rankings: measuring economic fundamentals, the performance of institutes, the quality of business services, the results of investments policies and the business perception
   c. Advantages and disadvantages of the rankings
3. Constructing models to evaluate the influence of different investment attractiveness factors on the FDI outflow from a reporting country (for one year only, not a time series)

Week 21. Qualitative analysis: news digest and case studies

(1) Learning Objectives

After this session, students should be able to:

- Understand the features of qualitative research
- Use case study as a research method
- Interpret the results of a case study

(2) Session Outline

1. Qualitative research
   a. Special features
   b. Advantages and disadvantages
   c. Methods
2. Case studies
   a. Selecting the cases
   b. Collecting data
   c. Evaluating and analyzing the results

Weeks 22, 23. Presentation of group research projects

Each group chooses two countries, one of which should be from the Asian region. The goal 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 R and Excel software. At the final sessions, each group will present the results of their research. The contents of the presentation should be organized as follows: (1) a brief description of the raw data on the bilateral trade and investment; (2) ARIMA model for the bilateral trade and a medium-term forecast; (3) trade composition analysis, evaluation of the bilateral trade pattern (inter- or intra-industry trade); (4) determinants of the FDI outflows by the reporting country, recommendations for the partner country: what measures can be taken to enhance the investment attractiveness; (5) cases to illustrate the motives of investors from the reporting country. After each presentation there will be a Q&A session. The time limit for a presentation + Q&A session is 20 minutes.