**Санкт-Петербургский филиал федерального государственного   
автономного образовательного учреждения высшего   
образования "Национальный исследовательский университет**

**"Высшая школа экономики"**

Факультет Санкт-Петербургская школа экономики и менеджмента Национального исследовательского университета «Высшая школа экономики»

**Рабочая программа дисциплины**

«Программирование для анализа данных и воспроизводимые исследования»

для майнора «Обработка и анализ данных»

уровень бакалавриат

2 курс (1-2 модуль)

Разработчики программы:

Мусабиров И.Л., MA (Digital Media and Information Society), магистр по направлению «Информатика и вычислительная техника», старший преподаватель департамента информатики, [ilya@musabirov.info](mailto:ilya@musabirov.info)

Суворова А.В., кандидат физико-математических наук, доцент департамента информатики, [asuvorova@hse.ru](mailto:asuvorova@hse.ru)

Бахитова А.А., преподаватель департамента информатики, [abakhitova@hse.ru](mailto:abakhitova@hse.ru)

Сироткин А.В., кандидат физико-математических наук, доцент департамента прикладной математики и бизнес-информатики, [avsirotkin@hse.ru](mailto:avsirotkin@hse.ru)

Утверждена академическим руководителем майнора

Сироткин А.В. \_\_\_\_\_\_\_\_\_\_\_

«30» августа 2018 г.

Санкт-Петербург, 2018

*Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения кафедры-разработчика программы.***Область применения и нормативные ссылки**

Настоящая рабочая программа дисциплины устанавливает минимальные требования к знаниям и умениям студента, а также определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов, изучающих дисциплину «Программирование для анализа данных и воспроизводимые исследования» в рамках майнора «Обработка и анализ данных».

Рабочая программа дисциплины разработана в соответствии с:

- Объединенным учебным планом майнора, утвержденным в 2018 г.

1. **Цели освоения дисциплины**

Целями освоения дисциплины «Программирование для анализа данных и воспроизводимые исследования» являются изучение особенностей различных методов сбора и агрегации данных, формирование навыков планирования сбора и обработки данных, изучение пакетов ориентированных на основные этапы подготовки данных и их анализа.

1. **Компетенции обучающегося, формируемые в результате освоения дисциплины**

В результате освоения дисциплины студент должен:

* Знать
  + ключевые понятия и задачи анализа данных, в том числе идеи “обучения с учителем” и “обучение без учителя”;
  + основные методы эксплораторного анализа данных;
  + примеры задач классификации и регрессии;
  + понятие воспроизводимого исследования и то, как достигается воспроизводимость исследований
* Уметь
  + применять на практике основные структуры данных в R;
  + производить простейшие статистические операции с помощью языка R;
  + выполнять эксплораторный анализ данных;
  + выбирать метод анализа, соответствующий исследовательской задаче;
  + визуализировать результаты анализа;
  + интерпретировать полученные результаты;
* Иметь навыки (приобрести опыт):
  + чтения и анализа академической литературы по анализу данных,
  + использования современного инструментария, применяемого в анализе данных.

Уровни формирования компетенций:

**РБ** - ресурсная база, в основном теоретические и предметные основы (знания, умения)

**СД** - способы деятельности, составляющие практическое ядро данной компетенции

**МЦ** - мотивационно-ценностная составляющая, отражает степень осознания ценности

компетенции человеком и готовность ее использовать

В результате освоения дисциплины студент осваивает следующие компетенции:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Компетенция | Код по ОС ВШЭ | Уровень формирования компетенции | Дескрипторы – основные признаки освоения (показатели достижения результата) | Формы и методы обучения, способствующие формированию и развитию компетенции | Форма контроля уровня сформированности компетенции |
| Способен учиться, приобретать новые знания, умения, в том числе в области, отличной от профессиональной | УК-1 | РБ | Способен на основе рекомендаций и самостоятельного поиска осваивать новые методы анализа данных | Практические и самостоятельные занятия по написанию программ для сбора и обработки информации | Контрольная работа  Домашнее задание |
| Способен вести исследовательскую деятельность, включая анализ проблем, постановку целей и задач, выделение объекта и предмета исследования, выбор способа и методов исследования, а также оценку его качества | УК-6 | РБ | Способен выделять постановки задач для решения с использованием различных методов анализа данных, осознанно выбирать методы и инструментальные средства | Самостоятельная деятельность по анализу результатов проведенных исследований | Контрольная работа  Домашнее задание |
| Способен работать с информацией: находить, оценивать и использовать информацию из различных источников, необходимую для решения научных и профессиональных задач (в том числе на основе системного подхода) | УК-5 | РБ | Владеет навыками написания программ для автоматизированного сбора и анализа информации из различных источников в глобальных компьютерных сетях | Практические занятия, программный проект | Контрольная работа  Домашнее задание |

1. **Место дисциплины в структуре образовательной программы**

Настоящая дисциплина относится к блоку дисциплин дополнительного профиля (майнора) «Обработка и анализ данных», обеспечивающих бакалаврскую подготовку.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин майнора.

1. **Тематический план учебной дисциплины**

ОБЪЕМ ДИСЦИПЛИНЫ - 5 зачетных единиц

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Название раздела | Всего часов | Аудиторные часы | | | Самостоятельная работа |
| Лекции | Практические работы | Семинары |
| 1 | Воспроизводимые исследования | 54 | 2 | 12 |  | 40 |
| 2 | Эксплораторный анализ данных | 63 | 3 | 18 |  | 42 |
| 3 | Статистическое обучение. Введение в клиентскую аналитику | 73 | 3 | 18 |  | 52 |
| ИТОГО | | **190** | 8 | 48 |  | 134 |

1. **Формы контроля знаний студентов**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Тип контроля | Форма контроля | 1 год | | | | Параметры |
| 1 | 2 | 3 | 4 |
| Текущий | Контрольная работа | \* | \* |  |  | Письменные работы на 65 минут |
| Домашнее задание | \* | \* |  |  | Программный проект |
| Подготовка к практикумам и работа на них | \* | \* |  |  | Прохождение онлайн-упражнений  Работа на практикумах |

1. **Критерии оценки знаний, навыков**

**Критерии оценки программного проекта**

• визуальная составляющая и содержательная наполненность отчета

• хорошо структурированная организация отчёта, ориентированность на заказчика

• использование навыков оформления графического материала, полученного в ходе обучения на настоящем курсе

• корректность применения методов агрегации и анализа данных

**Критерии оценки подготовки к практикумам и работы на них**

Практическая направленность курса предполагает высокую вовлечённость в индивидуальную и групповую работу на практикумах. Её оценка предполагает два компонента:

*Выполнение онлайн-упражнений* на закрепление и проверку усвоения материала на онлайн-платформе Stepik (stepik.org). Их прохождение играет формативную роль: количество попыток не ограничивается.

*Работа на практикумах* оценивается по: наличию активности в виртуальном образовательном окружении в компьютерной аудитории во время практикума (исключения и технические ошибки должны быть доведены до сведения преподавателя или учебного ассистента и отмечаются ими в электронной ведомости), ответам на вопросы и выполнению текущих заданий на практикуме.

Оценка за подготовку к практикумам и работу на них: зачтено/не зачтено.

Оценка “зачтено” ставится при одновременном выполнении следующих условий:

* прохождение суммарно более 50% заданий в установленные и отображаемые на онлайн-платформе Stepik дедлайны, без ограничения числа попыток (компонент Выполнение онлайн-упражнений)
* и активность не менее, чем на 50% практикумов в 1 и 2 (компонент Работа на практикумах).

При получении оценки “зачтено”, в формулу накопленной оценки выставляется 10, при получении оценки “не зачтено” -- 0.

В исключительных документально подтвержденных учебным офисом ОП студента уважительных случаях длительного отсутствия студента на занятиях оценка за подготовку к практикумам и работу за них может быть выставлена студенту только по онлайн-компоненте, при более чем 90% выполнении онлайн-заданий.

**Критерии заданий, включающих элементы анализа данных (в составе домашних и контрольных работ)**

• корректность применения методик анализа (в рамках знаний, полученных в курсе, смежных дисциплинах, домашнем чтении) и интерпретации результатов;

• уверенность использования языковых средств и структур данных, методов преобразования и агрегации данных в организации потока анализа данных, их ввода и вывода.

**Критерии оценки активности на форуме**

Учитывается полезность вопросов и ответов (полезные вопросы отмечаются модераторами –учебными ассистентами и преподавателями дисциплины), их количество и содержательность наполнения. В качестве руководства для студентов можно ориентироваться на советы StackExchange: https://stackoverflow.com/help/how-to-answer

При подведении итогов семестра публикуется упорядоченный по убыванию рейтинг активности студентов и соответствующие ему бонусные оценки.

1. **Содержание дисциплины**

**Раздел 1. Воспроизводимые исследования**

Область науки о данных: главные методы, техники, темы и прикладные возможности. Сравнение науки о данных и вычислительной социальной науки. Воспроизводимые исследования: возможности и ключевые цели. Создание отчетов и язык разметки RMarkdown. Введение в язык программирования R.

Количество часов аудиторной работы – 14

Общий объем самостоятельной работы – 40, из них выполнение заданий текущего контроля 20, подготовка к практическим занятиям 20.

Формы и методы проведения занятий по разделу: практические занятия в компьютерном классе.

Литература по разделу

* Roger D. Peng and Elizabeth Matsui. The Art of Data Science. A Guide for Anyone Who Works with Data. (гл. 1, 2, 10) <https://leanpub.com/artofdatascience>
* Skiena, Steven. 2017. *The Data Science Design Manual*. Springer. //www.springer.com/gp/book/9783319554433. (гл. 1)

**Раздел 2. Эксплораторный анализ данных**

Визуализация и агрегация данных: фильтрация, объединение, сортировка данных. Работа с датами и текстовыми данными.

Количество часов аудиторной работы – 21

Общий объем самостоятельной работы – 42, из них выполнение заданий текущего контроля 21, подготовка к практическим занятиям - 21.

Формы и методы проведения занятий по разделу: практические занятия в компьютерном классе.

Литература по разделу

* Roger D. Peng and Elizabeth Matsui. The Art of Data Science. A Guide for Anyone Who Works with Data. (гл. 4)

<https://leanpub.com/artofdatascience>

* Roger D. Peng. Exploratory Data Analysis with R. (гл. 4-7)

<https://leanpub.com/exdata>

* Skiena, Steven. 2017. *The Data Science Design Manual*. Springer. http://www.springer.com/gp/book/9783319554433 (гл. 1.3, 2.2, 5.1)

**Раздел 3. Статистическое обучение. Введение в клиентскую аналитику**

Вычислительная статистика и статистическое обучение в бизнесе, социологии, экономике. Возможности применения статистического обучения в информационных системах. Статистическое обучение с учителем и без учителя. Классификационные и регрессионные деревья. Тестирование статистических гипотез. Нахождение зависимостей в данных. Клиентская аналитика.

Количество часов аудиторной работы – 21

Общий объем самостоятельной работы – 52, из них выполнение заданий текущего контроля 21, подготовка к практическим занятиям 31.

Формы и методы проведения занятий по разделу: практические занятия в компьютерном классе.

Литература по разделу

* Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2014. *Introductory Statistics with Randomization and Simulation*. 1 edition. CreateSpace Independent Publishing Platform.
* Skiena, Steven. 2017. *The Data Science Design Manual*. Springer. //www.springer.com/gp/book/9783319554433 (гл. 1.4, 5.3)
* Roger D. Peng and Elizabeth Matsui. The Art of Data Science. A Guide for Anyone Who Works with Data. (гл. 7-9)
* <https://leanpub.com/artofdatascience>
* James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer. (гл. 1, 2, 8)
  + <http://www-bcf.usc.edu/~gareth/ISL/>
* Provost, Foster, and Tom Fawcett. 2013. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. 1 edition. Sebastopol, Calif.: O’Reilly Media.

1. **Образовательные технологии**

Преподавание языковых средств R и концепций эксплораторного анализа данных и статистического обучения осуществляется с использованием современных абстракций, с упором на понимание на концептуальном уровне и формальным введением по мере необходимости.

* 1. **Методические указания студентам по освоению дисциплины**

Для обеспечения необходимого уровня уверенного владения инструментальными средствами (языком R и средой RStudio) предусмотрен сквозной компьютерный практикум по всем разделам курса. Кроме того, рабочая среда с веб-доступом позволяет прозрачно переносить работу между практикумом и самостоятельной работой студента. Поэтому для успешного освоения дисциплины студент должен пользоваться возможностью самостоятельной работы, дополнительными ресурсами, указанными в программе и на форуме курса. **Пользуйтесь возможностью самостоятельной работы!**

Свои знания по части материалов курса можно проверить и дополнить используя онлайн-тесты. Хотя порог их прохождения на оценку достаточно низок, **мы советуем вам проходить все тесты и задавать вопросы на форуме.** Ориентируйтесь на здравый смысл и рекомендации https://stackoverflow.com/help/how-to-answer, формулируя ответы, старайтесь их делать полезными для многих.

Учебные ассистенты, преподаватели, ваши однокурсники и старшекурсники часто делятся дополнительными ресурсами на форуме курса. **Пользуйтесь этой возможностью, задавайте и отвечайте на вопросы других.** Помимо возможного бонуса к оценке, это позволяет более глубоко усвоить материал.

* 1. **Учебно-методическая литература для самостоятельной работы студентов**
* Grolemund, Garrett, and Hadley Wickham. 2016. *R for Data Science*. <http://r4ds.had.co.nz/>
* Roger D. Peng and Elizabeth Matsui. The Art of Data Science. A Guide for Anyone Who Works with Data. <https://leanpub.com/artofdatascience>
* Roger D. Peng. Exploratory Data Analysis with R. <https://leanpub.com/exdata>
* Skiena, Steven. 2017. *The Data Science Design Manual*. Springer. //www.springer.com/gp/book/9783319554433.
* Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2014. *Introductory Statistics with Randomization and Simulation*. 1 edition. CreateSpace Independent Publishing Platform.
* James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer. <http://www-bcf.usc.edu/~gareth/ISL/>
* Provost, Foster, and Tom Fawcett. 2013. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. 1 edition. Sebastopol, Calif.: O’Reilly Media.
* Gandrud, C. (2013). Reproducible Research with R and R Studio. CRC Press.
* Davenport, Thomas H. 2014. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Boston Massachusetts: Harvard Business Review Press.
* Siegel, Eric, and Thomas H. Davenport. 2013. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. 1 edition. Hoboken, N.J: Wiley.

1. **Оценочные средства для текущего контроля и аттестации студента**
   1. **Вопросы для оценки качества освоения дисциплины**
2. Что такое воспроизводимое исследование и какими средствами достигается воспроизводимость?
3. В чем принципиальное различие задач классификации и регрессии? Приведите содержательные примеры задач классификации и регрессии.
4. Назовите типы переменных, которые Вам известны. Можно ли перекодировать номинальную переменную в порядковую? Наоборот?
5. Рассмотрим график “ящик с усами” (boxplot). Что означает:
   1. линия посередине “ящика”?
   2. верхняя и нижняя границы ящика?
   3. точки, не соединяющиеся с графиком?
6. Представьте, что вам нужно сравнить количественный показатель для двух различных категорий (уровней номинальной переменной). Например, сравнить, как отличается заработная плата у мужчин и женщин в Твери. Какой график вам поможет справиться с этой задачей?
7. Требуется узнать, наблюдается ли зависимость между уровнем ВВП на душу населения и уровнем детской смертности в разных странах. Какой тип графика подойдет для этого больше всего?
8. В чем принципиальная разница между столбчатой диаграммой (bar chart) и гистограммой (histogram)? В каких случаях применяются данные графики?
9. Назовите известные Вам статистические тесты. На какие характеристики необходимо обращать внимание для интерпретации результатов? Приведите примеры задач, для решения которых необходимо использовать данный тест.
10. Как алгоритмом CART выбираются разбиения на каждом шаге? Какие существуют альтернативные способы выбора разбиений? В чем преимущества и недостатки этих вариантов?
    1. **Пример заданий, входящих в тесты и контрольные работы**

**Что из перечисленного является обязательным для исследования (анализа данных), чтобы его можно было считать воспроизводимым (reproducible)?**

1. В итоговом отчете должны быть библиографические ссылки на все упомянутые/процитированные научные работы

2. Все данные и код, использованные для анализа, должны быть публично доступны

3. Процесс анализа должен быть произведен на одном из вариантов операционной системы Unix

4. Анализ данных должен быть сделан с помощью R

**Какая из описанных задач является "предсказательной" (predictive)?**

1. Приводит ли большое количество времени, проведенного за видеоиграми к преждевременному старению мозга?

2. Связаны ли количество книг дома у школьника и его успеваемость?

3. Какова вероятность, что студент пропустивший первые две пары майнора, закончит его с итоговой оценкой выше 8?

4. Как широко распространено использование образовательных онлайн-платформ в вузах РФ?

5. Связано ли количество входящих писем на личной электронной почте с доходом человека?

На основе набора данных о всех пассажирских авиарейсах, вылетевших из аэропортов Нью-Йорка в 2013 году, ответьте на следующий вопрос: Сколько самолетов, совершавших рейсы из международного аэропорта La Guardia (LGA), вылетели с опозданием более чем на 30 минут?

Нарисуйте ящик с усами (box plot), показывающий, какую дистанцию преодолевают самолеты, с разбиением по аэропорту вылета. Напишите 2-4 предложения с содержательной интерпретацией получившегося графика. Какой(-ие) вывод(-ы) можно из него сделать? Получившуюся картинку прикрепите к решению.

* 1. **Примерное задание для программного проекта**

***## Разведочный анализ данных и постановка задачи***

Мы с вами поговорили про основы разведочного анализа данных и про основные методы обработки и агрегации данных. Еще одна часть, которую мы подробно рассмотрим чуть позже -- это проверка гипотез с помощью статистических тестов. Сегодня вы попробуете применить ваши знания к реальным данным, про которые вы почти ничего не знаете. Это будет частью вашей домашней работы.

Загрузим необходимые данные с которыми вам предстоит работать:

*## Подгружаем библиотеки*

library(dplyr)

library(ggplot2)

x = read.csv("~/shared/minor2\_2017/1-Intro/data/Customer\_demographics.csv", sep = ";")

y = read.csv("~/shared/minor2\_2017/1-Intro/data/Product\_sales.csv", sep = ";")

df = inner\_join(x, y, by="Customer.ID")

Данные связанны между собой, возможно, вам стоит объединить их в одну табличку с помощью \*join\*, по полю \*Customer.ID\*. Ваша задача рассказать нам про эти данные, то что вы сможете в них увидеть. Сегодня вы должны понять что за данные перед вами. Построить ряд графиков, которые помогут вам охарактеризовать данные. И поставить вопросы о взаимосвязи разных переменных из набора данных. Работу нужно сдать с помощью gist.

**11 Порядок формирования оценок по дисциплине**

**Накопленная оценка по дисциплине** рассчитывается с помощью взвешенной суммы оценок за отдельные формы текущего контроля знаний следующим образом:

О*накопленная* = 0.2\**Отекущий 2* + 0.15*\*Отекущий 1* + 0.15\**Отекущий 3* + 0.2\**Отекущий 4* + 0.3\**Отекущий 5*, где

*Отекущий 2*– оценка за контрольную работу №1,

*Отекущий 1*– оценка за домашнее задание - программный проект №1,

*Отекущий 3*– оценка за домашнее задание - программный проект №2,

*Отекущий 4*– оценка за контрольную работу №2,

*Отекущий 5*– оценка за прохождение онлайн-упражнений и работа на практикумах.

Способ округления накопленной оценки текущего контроля: арифметический.

**Результирующая оценка по дисциплине** рассчитывается следующим образом:

*Орезульт = Онакопл,* где

*Онакопл* – накопленная оценка по дисциплине. Способ округления результирующей оценки: арифметический.

Получивший неудовлетворительную оценку за контрольную работу студент может в специально отведенное время на сессии 2 модуля переписать контрольную работу 1 или контрольную работу 2.

Студент может получить дополнительное задание для компенсации оценок текущего контроля (за исключением домашних работ).

Активность на форуме курса приносит бонус к результирующей оценке до 20%.

1. **Учебно-методическое и информационное обеспечение дисциплины**
   1. **Основная литература**

* Grolemund, Garrett, and Hadley Wickham. 2016. R for Data Science. <http://r4ds.had.co.nz/>
* Roger D. Peng and Elizabeth Matsui. The Art of Data Science. A Guide for Anyone Who Works with Data. <https://leanpub.com/artofdatascience>
* Roger D. Peng. Exploratory Data Analysis with R. <https://leanpub.com/exdata>
* Skiena, Steven. 2017. *The Data Science Design Manual*. Springer. [www.springer.com/gp/book/9783319554433](http://www.springer.com/gp/book/9783319554433)
* Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2014. *Introductory Statistics with Randomization and Simulation*. 1 edition. CreateSpace Independent Publishing Platform.
* James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer. <http://www-bcf.usc.edu/~gareth/ISL/>
  1. **Дополнительная литература**
* Provost, Foster, and Tom Fawcett. 2013. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. 1 edition. Sebastopol, Calif.: O’Reilly Media.
* Gandrud, C. (2013). Reproducible Research with R and R Studio. CRC Press.
* Davenport, Thomas H. 2014. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Boston Massachusetts: Harvard Business Review Press.
* Siegel, Eric, and Thomas H. Davenport. 2013. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. 1 edition. Hoboken, N.J: Wiley.
  1. **Ресурсы информационно-телекоммуникационной сети “Интернет”**

1. Интерактивный учебник по основам R <http://tryr.codeschool.com/>
2. Онлайн-курс Анализ данных в R <https://stepik.org/s/A45JQdR5>
3. Онлайн-курс Программирование в R <https://stepik.org/s/yqm30W4N>

Большое количество онлайн-курсов и дополнительных ресурсов по R и анализу данных во всем его многообразии можно найти в интернет и на основных площадках образовательных курсов. Их всегда можно обсудить с одно и старшекурсниками, учебными ассистентами и преподавателями на форуме.

Часть материалов курса доступна в виде интерактивных упражнений на платформе Степик.

* 1. **Программные средства**

RStudio. Пакет MS Office/OpenOffice.org

1. **Материально-техническое обеспечение дисциплины**

Аудитория с проектором для лекций, компьютерные аудитории с современными версиями браузеров согласно требованиям RStudio Server к клиентам. Клиент ssh для консольного доступа к серверу. Сервер для работы студентов (спецификация в зависимости от количества записавшихся на майнор) с Ubuntu Linux, RStudio Server и другими пакетами, в зависимости от специфики проектов.