Федеральное государственное автономное образовательное учреждение высшего образования
"Национальный исследовательский университет
"Высшая школа экономики"

Факультет социальных наук
Кафедра высшей математики

Рабочая программа дисциплины «Основы программирования в R»

для направления 41.03.04 «Политология» подготовки бакалавра

Разработчик программы:
Тамбовцева А.А., ассистент, atambovtseva@hse.ru

Одобрена на заседании кафедры высшей математики
«___»_____ 2018 г.
Зав. Кафедрой, к.ф.-м.н., проф. Макаров А.А. _____

Утверждена Академическим советом образовательной программы
«___»_____ 2018 г., № протокола _____

Академический руководитель образовательной программы
Локшин И.М. _____

Москва, 2018

1 Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает требования к образовательным результатам и результатам обучения студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих дисциплину «Основы программирования в R», учебных ассистентов и студентов направления подготовки 41.03.04 «Политология», обучающихся по образовательной программе «Политология».

Программа учебной дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ;
- Образовательной программой «Политология»;
- Объединенным учебным планом университета по образовательной программе «Политология», утвержденным в 2018г.

2 Цели освоения дисциплины

Целями освоения дисциплины «Основы программирования в R» являются овладение навыками программирования на языке R, овладение методами обработки, визуализации и анализа качественных и количественных данных для решения политологических и социально-экономических задач.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент осваивает 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	МЦ	Владеет навыками программирования в R	Практические задания на семинарах	Регулярные домашние задания, экзамен
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	СД	Находит необходимую статистическую информацию в сети Интернет, демонстрирует навыки работы с базами данных	Лекции и семинары	Домашнее задание, указанное в РУП
Способен осуществлять поиск, сбор, обработку, анализ и хранение информации для решения поставленных задач	ПК-4	СД	Владеет навыками обработки и визуализации качественных и количественных данных в R	Лекции и практические задания на семинарах	Проверочные работы, домашние задания, экзамен

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина читается на 3 курсе бакалавриата образовательной программы «Политология» и является дисциплиной по выбору.

Изучение данной дисциплины базируется на следующих дисциплинах:

- Математика и статистика

Для освоения учебной дисциплины студенты должны владеть следующими знаниями и компетенциями:

- Базовые навыки работы с персональным компьютером
- Знания математики и статистики в рамках обязательного курса «Математика и статистика»
- Знание английского языка, достаточное для чтения учебной литературы и документации

Основные положения дисциплины должны быть использованы в дальнейшем при изучении дисциплин:

- Основы программирования в Python
- Анализ больших данных в социальных науках
- Анализ социальных сетей

5 Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы		Самостоятельная работа
			Лекции	Семинары	
1	Установка R и RStudio. Особенности интерфейса RStudio. Язык разметки markdown. Базовые объекты R: переменные, векторы, списки, матрицы.	14	4	4	6
2	Форматы данных. Загрузка данных в R. Работа с текстовыми файлами в R.	8	2	2	4
3	Основы работы с базами данных.	10	2	2	6
4	Базовые графики в R. Визуализация количественных и качественных данных в R. Разведывательный анализ данных в R.	16	4	4	8
5	Корреляционный и регрессионный анализ в R. Множественная регрессия в R. Работа с пространственно-временными данными в R.	16	4	4	8
6	Управляющие конструкции в R. Циклы for и while. Функции в R.	18	4	4	10
7	Парсинг веб-страниц средствами R. Библиотека rvest.	16	6	4	8
8	Работа с API средствами R.	16	4	4	8
	ИТОГО	**152**	**30**	**28**	**94**

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год		Кафедра/подразделение	Параметры
		3	4		
Текущий	Домашнее задание		*		Выполненное студентом самостоятельно исследование с использованием методов обработки, визуализации и анализа данных в R.
Итоговый	Экзамен		*		Письменная работа, 180 минут

7 Критерии оценки знаний, навыков

Домашнее задание представляет собой небольшое исследование, которое предполагает несколько этапов: выбор базы данных, описание выбранной базы данных, первичная обработка данных, визуализация и анализ данных. В рамках выполнения домашнего задания от студента требуется подготовить текстовый файл с описанием выбранной базы данных и файл, содержащий код на R, используемый для выполнения задания, а также необходимые комментарии. Домашнее задание оценивается по 10-ти балльной шкале.

Экзамен представляет собой набор задач по пройденным темам, которые выполняются на компьютере, в R (RStudio). Экзамен оценивается по 10-ти балльной шкале.

Работа студентов оценивается по следующим критериям: соответствие решения поставленной задаче, работоспособность и корректность кода программы (код должен запускаться без ошибок и выдавать ожидаемый результат), использование рассмотренных в курсе средств и методов, качество выполнения заданий (наличие требуемых заголовков, подписей и легенды к графикам, комментариев к коду), соответствие требованиям преподавателя (формат и срок сдачи заданий).

Задачи, для решения которых предоставлен неработающий код (код, который невозможно запустить из-за наличия грубых ошибок/опечаток), не засчитываются, даже если при этом зафиксирован верный результат.

8 Содержание дисциплины

Раздел 1. Установка R и RStudio. Особенности интерфейса RStudio. Установка библиотек. Базовые объекты R: переменные, векторы, списки, матрицы.

Установка R и RStudio. Консоль R. Знакомство с интерфейсом RStudio.

Язык разметки Markdown. Создание файлов Rmarkdown. Публикация кода на Rpubs.

Установка и загрузка библиотек в R. Знакомство с документацией R.

R как калькулятор. Переменные в R. Типы данных: числовой, целочисленный, логический, текстовый. Преобразование типов. Факторы (factor vectors) и уровни.

Базовые объекты в R: векторы, списки, матрицы. Векторы: создание, доступ к элементам, изменение, добавление и удаление элементов, выбор элементов, сортировка. Матрицы и списки: создание, доступ к элементам, изменение, добавление и удаление элементов.

Раздел 2. Форматы данных. Загрузка данных в R. Работа с текстовыми файлами в R.

Разнообразие форматов данных: таблицы Excel (xls, xlsx), таблицы Stata и SPSS (dta, sav), текстовые файлы (txt, csv), json-файлы.

Загрузка данных в R. Открытие, изменение и запись файлов в R.

Регулярные выражения.

Раздел 3. Основы работы с базами данных.

Выбор, добавление и удаление переменных. Преобразование типов переменных. Фильтрация, добавление и удаление наблюдений.

Создание базы данных. Преобразование матриц и списков в объект *data.frame*. Объединение баз данных.

Раздел 4. Базовые графики в R. Визуализация количественных и качественных данных в R. Разведывательный анализ данных в R.

Базовые графики в R. Функция *plot()*. Построение графиков математических функций.

Генерирование данных в R. (Псевдо)случайные значения.

Визуализация количественных данных в R: гистограммы, графики плотности распределения, ящики с усами.

Визуализация качественных данных в R: таблицы сопряженности, столбчатые и круговые диаграммы.

Описательные статистики: среднее арифметическое, среднеквадратичное отклонение, медиана, квантили, квартили и процентили.

Разведывательный анализ данных: типы и распределения данных. Проверка данных на нормальность в R: нормальная вероятностная бумага, критерий Колмогорова-Смирнова и критерий Шапиро-Уилка.

Раздел 5. Корреляционный и регрессионный анализ в R. Множественная регрессия в R. Работа с пространственно-временными данными в R.

Коэффициенты корреляции Пирсона и Спирмена в R. Корреляционные матрицы в R. Визуализация корреляций между переменными в R: heatmaps.

Множественная регрессия в R: построение модели, интерпретация выдачи R. Визуализация результатов регрессионного анализа с помощью библиотеки *sjPlot*. Выгрузка необходимой информации из R в текстовые редакторы.

Перекрестные данные (cross section data), временные ряды (time series data) и пространственно-временные данные (time series cross section). Регрессионные модели для пространственно-временных данных: модель с фиксированными эффектами, модель со случайными эффектами.

Раздел 6. Управляющие конструкции в R. Циклы for и while. Функции в R.

Операторы в R. Условные операторы *if* и *else*. Множественные и разветвленные условия.

Циклы в R. Циклы for и while: достоинства и недостатки. Циклы vs векторные операции в R.

Устройство функций в R. Основные функции в R. Функция *assign()*. Написание простейших функций в R.

Раздел 7. Парсинг веб-страниц средствами R. Библиотека rvest.

Структура html-файлов.

Извлечение информации из html-файлов. Выгрузка текста из html-файлов.

Функционал библиотеки rvest.

Раздел 8. Работа с API средствами R.

Знакомство с API.
Работа с API ВКонтакте и API Twitter.
Библиотеки vkR, Rvk и twitteR.

9 Образовательные технологии

Занятия по курсу включают лекции и семинарские занятия.

10 Оценочные средства для текущего контроля и аттестации студента

Примеры задач практических заданий, домашних заданий и экзамена

1. Создайте Rmd-файл *„my_first_rmarkdown.Rmd“*. Воспроизведите Rmd-файл, предложенный преподавателем: напечатайте такой же текст с соответствующей разметкой, включите блоки, содержащие код на R. Сохраните полученный файл в формате html и опубликуйте его на RPubs.

2. Создайте переменную *m* и присвойте ей значение 3.1415. Определите тип переменной. Проверьте, является ли данная переменная целочисленной, текстовой, логической. Округлите переменную *m* до второго знака после запятой и сохраните полученный результат в переменную *p*.

3. Дана строка *S*:

"регион: 50 45 52 39 3 75 10 91;

ВРП: 2551284 165150 925832 277362 177692 229782 175975 NA;

тип региона: E A E E A A E NA;"

Используя функции *strsplit()* и *unlist()*, получите три вектора из строки *S*: *n_reg*, *reg_grp*, *reg_type* (номер региона, ВРП, тип региона – в европейской или азиатской части России). Скорректируйте типы полученных векторов (первые два вектора должны быть числовыми, последний – текстовым).

4. С помощью функции *paste()* создайте вектор *my_groups*, содержащий следующие элементы: „group_161“, „group_162“, „group_163“, „group_164“, „group_165“. Замените во всех элементах „_“ на одинарный пробел и сохраните изменения в исходном векторе *my_groups*.

5. а) Скачайте базу данных, содержащую результаты выборов в Государственную Думу 2016 года по городу Москве (по ссылке). Ознакомьтесь с ее описанием (codebook.pdf).

б) Загрузите данные в R. Проверьте, есть ли в базе данных пропущенные значения. Если есть, удалите их из базы.

в) Создайте переменную *turnout* (явка на выборы). Напоминание: явка определяется как сумма действительных и недействительных бюллетеней на выборах.

г) Создайте переменную *turnout_perc* (процент явки на выборы). Напоминание: процент явки на выборы считается как показатель явки, деленный на число

зарегистрированных избирателей. Убедитесь в том, что переменная *turnout_perc* является числовой, но не является целочисленной.

д) Определите, на каких избирательных участках процент явки а) превышает значение 35; б) превышает медианное значение процента явки по Москве.

е) Выберите из базы избирательные участки, процент явки на которых не превышает медианное значение по Москве. Сохраните выбранные наблюдения в отдельную базу данных *data_low*. Сохраните базу *data_low* в *xlsx*-файл *„elections_low_turnout.xlsx“*.

6. Загрузите файл *persson_tabellini2003.csv* в R. Постройте гистограмму для переменной *logyl* (объем выпуска на одного рабочего, натуральный логарифм). Наложите на гистограмму график плотности нормального распределения с соответствующими параметрами. Добавьте название графика, подпишите оси. Сохраните его как *‘logyl.png’*. Проверьте с помощью формального критерия, является ли распределение переменной *logyl* нормальным (используйте, например, критерий Шапиро-Уилка).

Задачи повышенной сложности из продвинутого блока домашних заданий могут выходить за рамки изученного на предшествующих занятиях материала (требовать более высокого уровня владения R, установки дополнительных библиотек, самостоятельного изучения документации по инструкции, предложенной преподавателем).

Примеры вопросов проверочных работ

1. Укажите, к какому типу данных (интервальные, ординальные, номинальные) относятся следующие переменные:

а) пол респондента

б) место студента в рейтинге

в) ВВП на душу населения

г) образование респондента (1 – начальное, 2 – среднее общее, 3 – среднее профессиональное, 4 – высшее профессиональное)

д) степень согласия респондента с утверждением (1 – полностью несогласен, 2 – не согласен, 3 – скорее согласен, чем не согласен, 4 – согласен, 5 – полностью согласен)

е) доля людей старше трудоспособного возраста в регионе

2. Какие из перечисленных ниже графиков *не* могут быть использованы для визуализации качественных (номинальных) данных?

а) столбчатая диаграмма

б) круговая диаграмма

в) гистограмма

г) ящик с усами

д) мозаичный график

е) диаграмма рассеяния

3. Вам необходимо построить столбиковую диаграмму, которая показывала бы, сколько в базе данных респондентов-женщин, а сколько мужчин. Переменная *gender* содержит значения *„male“* и *„female“*. Что необходимо сделать перед построением диаграммы, если для этого Вы собираетесь использовать базовую функцию *plot()*?

4. Вася в рамках своего исследования связи экономического развития и типа политического режима решил посчитать коэффициент корреляции Пирсона между переменными «Рост ВВП на душу населения в год (в процентах)» и «Тип политического режима по Freedom House (1 – Free, 2 – Partly Free, 3 – Not Free)». Прокомментируйте идею Васи.

5. Чем временные ряды отличаются от перекрестных данных?

11 Порядок формирования оценок по дисциплине

Преподаватель оценивает работу студентов на семинарских занятиях: выполнение практических заданий и проверочных работ. Практические задания представляют собой набор небольших задач по теме текущего или прошлого занятия, выполняемых студентом в R (RStudio). Во время выполнения практических заданий студент может пользоваться материалами лекций и семинаров, а также Интернетом. Проверочные работы предполагают ответы на тестовые и открытые вопросы по пройденному материалу. Во время выполнения проверочных работ не допускается использование учебных материалов, компьютер для выполнения проверочных работ не требуется. Оценки за работу на семинарских занятиях преподаватель выставляет в рабочую ведомость. Общая оценка по 10-ти балльной шкале за работу на семинарских занятиях ($O_{аудиторная}$) определяется как округленное до целого среднее арифметическое оценок, полученных на занятиях (способ округления: арифметический).

Преподаватель оценивает самостоятельную работу студентов: текущие домашние задания, не включенные в РУП. Домашние задания предполагают решение задач по программированию по пройденной теме. Оценки за самостоятельную работу студента преподаватель выставляет в рабочую ведомость. Общая оценка по 10-ти балльной шкале за самостоятельную работу ($O_{сам.работа}$) определяется как округленное до целого среднее арифметическое оценок, полученных за домашние работы (способ округления: арифметический).

С целью учета разного уровня подготовки студентов домашние задания состоят из двух блоков: базовый и продвинутый. Студенту на выбор предлагается решить задачи одного из блоков. Если студент хочет решить задачи продвинутого блока, но не уверен, что сделает их полностью верно или в требуемом объеме, возможен следующий вариант. Студент решает 50% задач из базового блока, решает задачи продвинутого блока, и тогда его оценка за домашнюю работу считается как округленное среднее арифметическое оценок, полученных за базовый блок и продвинутый блок (способ округления: арифметический).

Пример. Базовый блок включает 10 задач. Продвинутый блок состоит из одной задачи повышенной сложности. Студент выбирает 5 задач базового блока и предлагает неполное/частично верное решение задачи из продвинутого блока. За базовую часть студент получает оценку 10, за продвинутую часть – оценку 5, итоговая оценка за данное домашнее задание равна 8.

Домашние задания (текущие и включенное в РУП), сданные после срока, оцениваются с использованием понижающих коэффициентов: опоздание в пределах часа – штраф 10% от полученной оценки, в пределах суток – штраф 20%, в пределах недели – штраф 50%. Домашние

задания, сданные через неделю после указанного срока и позже, не принимаются и не оцениваются.

Если при проверке работ (текущий и итоговый контроль, аудиторная и самостоятельная работа) установлен факт нарушения академической этики, студент получает оценку «0» за данную работу. Работа студента, предоставившего свою работу для списывания, также аннулируется.

В случае нарушения правил проведения экзамена студент удаляется с экзамена с оценкой «0». К нарушениям правил проведения экзамена относятся: коммуникация с другими студентами во время выполнения работы, использование социальных сетей/телефона во время экзамена (с любой целью), списывание.

Накопленная оценка по дисциплине рассчитывается по формуле:

$$O_{накопленная} = 0.4 * O_{ДЗ} + 0.2 * O_{аудиторная} + 0.4 * O_{сам.работа},$$

где $O_{ДЗ}$ – оценка за домашнее задание, включенное в РУП (см. пункт 7).

В диплом выставляется результирующая оценка по учебной дисциплине.

$$O_{результирующая} = 0.6 * O_{накопленная} + 0.4 * O_{экзамен}$$

Способ округления результирующей оценки по учебной дисциплине: арифметический.

12 Учебно-методическое и информационное обеспечение дисциплины

12.1 Базовый учебник

А. Б. Шипунов и др. Наглядная статистика. Используем R! М.: ДМК Пресс. 2017.

N.J.Horton, K.Kleinman. Using R for Data Management, Statistical Analysis, and Graphics. CRC Press. 2010.

12.2 Основная литература

A.Gohil. R Data Visualization Cookbook. Packt Publishing. 2015.

12.3 Программные средства

Для успешного освоения дисциплины, студент использует следующие программные средства:

- R (https://cran.r-project.org/)
- RStudio (https://www.rstudio.com/)

13 Материально-техническое обеспечение дисциплины

Лекции и семинары проводятся в компьютерном классе. Студентам во время работы рекомендуется использовать свои ноутбуки. Из программного обеспечения необходимы R и RStudio.