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

Факультет коммуникаций, медиа и дизайна
Кафедра высшей математики

# Рабочая программа дисциплины «Программирование для анализа данных»

для образовательной программы «Коммуникации, основанные на данных»
направления 42.04.01 «Реклама и связи с общественностью»
подготовки магистра

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
Грызунова Е.А. _____

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 «Программирование для анализа данных», учебных ассистентов и студентов направления подготовки магистра 42.04.01 «Реклама и связи с общественностью», обучающихся по образовательной программе «Коммуникации, основанные на данных».

Программа учебной дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ по направлению 42.04.01 Реклама и связи с общественностью, квалификация «Магистр» (ред. 2017 г.);
- Образовательной программой «Коммуникации, основанные на данных».
- Объединенным учебным планом университета по образовательной программе «Коммуникации, основанные на данных», утвержденным в 2018г.

# 2 Цели освоения дисциплины

Целями освоения дисциплины «Программирование для анализа данных» являются овладение навыками программирования на языке R и работы в среде RStudio, овладение методами обработки, визуализации и анализа качественных и количественных данных для решения прикладных задач, возникающих в сфере управления интегрированными коммуникациями и маркетинга.

# 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 |
|---|---|---|---|---|---|
| Способность к самостоятельному освоению новых методов исследований, изменению научного и производственного профиля своей деятельности. | УК-3 | МЦ | Владеет навыками программирования в R, находит необходимую статистическую информацию в сети Интернет, демонстрирует навыки работы с базами данных. | Лекции и семинары | Домашние задания, экзамен |
| Способность обрабатывать данные, с целью построения коммуникационной кампании, в том числе используя специальное программное обеспечение | ПК-5 | СБ | Адекватно оценивает корректность использования методов анализа данных, применяемых при решении практических задач, владеет навыками обработки и анализа данных в R. | Лекции и семинары | Домашние задания, экзамен |

# 4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина читается на 1 курсе магистратуры образовательной программы «Коммуникации, основанные на данных» и относится к блоку базовых дисциплин.

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

- Знания математики в рамках школьной программы
- Базовые навыки работы с персональным компьютером
- Знание английского языка, достаточное для чтения документации.

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

- Машинное обучение и анализ больших данных
- Эконометрика

# 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. Установка библиотек. Базовые объекты R: переменные, векторы, списки, матрицы. | 18 | 2 | 4 | 12 |
| 2 | Работа с файлами: открытие, изменение, сохранение. Загрузка данных в R. Основы работы с базами данных (объекты data.frame). Работа с базами данных с использованием библиотеки dplyr. | 28 | 2 | 6 | 20 |
| 3 | Операторы в R. Устройство функций в R. Циклы. | 26 | 2 | 4 | 20 |
| 4 | Разведывательный анализ данных в R. Визуализация количественных и качественных данных в R. Визуализация данных с помощью библиотеки ggplot2. | 42 | 2 | 10 | 30 |
| | **Всего** | **114** | **8** | **24** | **82** |

## 6 Формы контроля знаний студентов

| Тип контроля | Форма контроля | 1 год | | | | Параметры |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | |
| Текущий | Домашнее задание | * | | | | выполненное студентом самостоятельно мини-исследование с использованием методов обработки и визуализации данных в R |
| Итоговый | Экзамен | * | | | | письменный экзамен, 180 мин |

## 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). Во время выполнения экзамена студенту разрешается использовать материалы лекций и семинаров, а также пользоваться Интернет-ресурсами, посвященными программированию и анализу данных в R. Коммуникация с другими студентами и использование социальных сетей во время экзамена не допускается. 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.

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале.

Материалы по данной дисциплине, задачи к семинарам и домашние задания публикуются на странице курса на сайте http://math-info.hse.ru.

## 8 Содержание дисциплины

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

Установка R и RStudio. Особенности интерфейса RStudio. Язык разметки Markdown. Создание файлов Rmarkdown. Публикация кода на Rpubs.

Установка и загрузка библиотек в R. Документация к R и ресурсы, посвященные программированию в R.

R как калькулятор. Переменные в R. Типы данных: числовой, целочисленный, текстовый, логический. 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. Работа с файлами: открытие, изменение, сохранение. Загрузка данных в R. Основы работы с базами данных (объекты data.frame). Работа с базами данных с использованием библиотеки dplyr.**

Загрузка данных в R. Загрузка текстовых файлов (txt, csv), загрузка таблиц Excel (xls, xlsx), загрузка файлов Stata и SPSS (dta, sav). Конвертация файлов в разных форматах.

Обращение к базе данных (объект data.frame). 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.

Загрузка и основной функционал библиотеки dplyr.

**Раздел 3. Операторы в R. Устройство функций в R. Циклы.**

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

Циклы в R, их достоинства и недостатки. Устройство функций в R. Основные функции в R.

Функция assign(). Написание простейших функций в R.

**Раздел 4. Разведывательный анализ данных в R. Визуализация количественных и качественных данных в R. Визуализация данных с помощью библиотеки ggplot2.**

Описание базы данных в R. Описательные статистики: среднее арифметическое, среднеквадратичное отклонение, медиана, квантили, квартили и процентили. Выгрузка необходимой информации из R в текстовые редакторы.

Разведывательный анализ данных: типы и распределения данных. Выявление связей между качественными и количественными переменными.

Базовые графики в R. Визуализация количественных данных в R: гистограммы, графики плотности распределения, ящики с усами, скрипичные диаграммы (violin plots), диаграммы рассеяния, матрицы диаграмм рассеяния. Визуализация качественных данных в R: таблицы сопряженности, столбчатые и круговые диаграммы.

Логика построения графиков с помощью ggplot2: соответствие переменным базы данных используемым визуальным средствам (aes), тип графика (geom), тип статистического преобразования (stat). Работа с форматом графиков: изменение фона, палитра цветов, типы маркеров и линий, редактирование легенды графика.

## 9 Образовательные технологии

Занятия по курсу включают лекции и семинарские занятия. Лекции и семинары проходят в компьютерном классе.

## 10 Оценочные средства для текущего контроля и аттестации студента

**Оценочные средства для оценки качества освоения дисциплины в ходе текущего и итогового контроля**

*Примеры задач по программированию из домашних заданий и экзаменационной работы:*

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

2. Создайте переменную x и присвойте ей значение 2.58. Определите тип переменной. Проверьте, является ли данная переменная целочисленной, текстовой. Округлите переменную x до первого знака после запятой и сохраните полученный результат в переменную y.

3. Создайте вектор v, содержащий следующие числовые значения: 2000, 2500, 15000, 7500, 5200, 8700, 9400, 10200. Отсортируйте его а) по возрастанию; б) по убыванию. Прологарифмируйте значения вектора v (натуральный логарифм) и сохраните полученные значения в вектор v_log.

4. С помощью функции paste() создайте вектор resp, содержащий следующие элементы: „respondent-1“, „respondent-2“, „respondent-3“, „respondent-4“, „respondent-5“. Замените во всех элементах „-“ на „_“ и сохраните изменения в исходном векторе resp.

5. Скачайте базу данных, содержащую результаты опроса покупателей (по ссылке). Ознакомьтесь с ее описанием (codebook.pdf).

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

б) Создайте переменную «возраст в квадрате» (*age_sq*). Убедитесь в том, что переменная *age_sq* является числовой, но не является целочисленной.

в) Выберите из базы покупателей, чей возраст не превышает среднее значение, посчитанное на основе имеющихся данных. Сохраните выбранные наблюдения в отдельную базу данных df2. Сохраните базу df2 в csv-файл.

## 11 Порядок формирования оценок по дисциплине

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. 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).

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:

$$O_{накопленная} = 0,5 * O_{сам.работа} + 0,5 * O_{ДЗ,}$$

где $O_{сам.работа}$ — округленное среднее арифметическое оценок за текущие домашние задания (способ округления: арифметический), $O_{ДЗ}$ — оценка за домашнее задание, включенное в РУП (см. пункт 7).

В диплом выставляется результирующая оценка по учебной дисциплине.

$$O_{результ} = 0,6 * O_{накопленная} + 0,4 * O_{экзамен}$$

Способ округления результирующей оценки по учебной дисциплине: арифметический.

## 12 Учебно-методическое и информационное обеспечение дисциплины

### 12.1 Базовый учебник

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

2. 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/)

### 12.4 Дистанционная поддержка дисциплины

Материалы по курсу выкладываются на странице http://math-info.hse.ru.

## 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.