Программа учебной дисциплины «Введение в науку о данных»

Утверждена
Академическим советом ОП
Протокол № 15 от «28» июня 2018 г.

Автор	Г. А. Мороз
Число кредитов	3
Контактная работа (час.)	6
Самостоятельная работа (час.)	108
Курс	3
Формат изучения дисциплины	с использования онлайн-курса

1. Цель, результаты и пререквизиты освоения дисциплины

Целями освоения дисциплины «Введение в науку о данных» является знакомство:
с основами работы в R и RStudio
с основными типами данных (таблицы, тексты, изображение с текстом);
с основными методами сбора, обработки и трансформации данных;
с основными методами визуализации и представления данных;
с основными методами статистического анализа;
с основными методами регрессионного анализа;

В результате освоения дисциплины студент должен:

знать
- особенности работы R, основные особенности анализа различных типов данных;
- познакомиться с основами методами регрессионного анализа данных

уметь
- преобразовывать и визуализировать данные;

владеть
- базовыми навыками самостоятельного анализа данных, а также критической интерпретации анализа данных, представленной в научных работах;

У курса нет пререквизитов.

Формат изучения: лекции, семинары, с использованием онлайн курса.

2. Содержание учебной дисциплины

Тема (раздел дисциплины)	**Объем в часах**	**Планируемые результаты обучения (ПРО), подлежащие контролю**	**Формы контроля**
Тема 1. Введение в Data Science	лк 2 см 4 ср 0	Знать чем наука о данных отличается от машинного обучения и статистики.	Один из вопросов из домашней работы

Тема 2. Введение в R: основные элементы, функции, циклы	лк 0 см 0 ср 11	Владеть основами R	Один из вопросов из домашней работы
Тема 3. Продвинутая обработка данных: пакеты tidyr и dplyr	лк 0 см 0 ср 11	Владеть методами обработки данных	Один из вопросов из домашней работы
Тема 4. Работа со строками: строки в R, регулярные выражения	лк 0 см 0 ср 11	Владеть методам анализа строк	Один из вопросов из домашней работы
Тема 5. Визуализация данных: base R vs. ggplot2	лк 0 см 0 ср 11	Владеть методами визуализации данных	Один из вопросов из домашней работы
Тема 6. Лингвистические пакеты	лк 0 см 0 ср 11	Знать лингвистические пакеты на R	Один из вопросов из домашней работы
Тема 7. Введение в статистику: основы фриквентисткой статистики, формулировка гипотез	лк 0 см 0 ср 11	Применять основные фриквентисткие тесты	Один из вопросов из домашней работы
Тема 8. Корреляция и линейная регрессия	лк 0 см 0 ср 11	Применять корреляционный и регрессионный анализы	Один из вопросов из домашней работы
Тема 9. Логистическая и мультиномиальная регрессия Тема 10. Критерии согласия	лк 0 см 0 ср 11 лк 0 см 0 ср 10	Применять метод логистической регресии Применять методы логистической регресии	Один из вопросов из домашней работы
Часов по видам учебных занятий:	лк 3		
	см 3		
	ср 108		
Итого часов:	304		

3. Оценивание

Итоговая оценка за курс состоит из оценок за домашние работы и экзамен.

$$O_{итоговый} = 0.6 * O_{промежуточный\ тесты} + 0.4 * O_{финальный\ тест}$$

Оценки выставляются по 10-балльной шкале. Способ округления оценок: арифметический.
Оценки выставляются по 10-балльной шкале. Способ округления оценок: арифметический.
Оценка 10 ставится за абсолютно верный ответ, содержащий элементы нетривиального подхода к анализу материала.
Оценка 9 ставится за абсолютно верный ответ, не обладающий нетривиальными особенностями.
Оценка 8 ставится за абсолютно верный ответ с незначительными погрешностями при условии их самостоятельного исправления в процессе диалога с преподавателем.

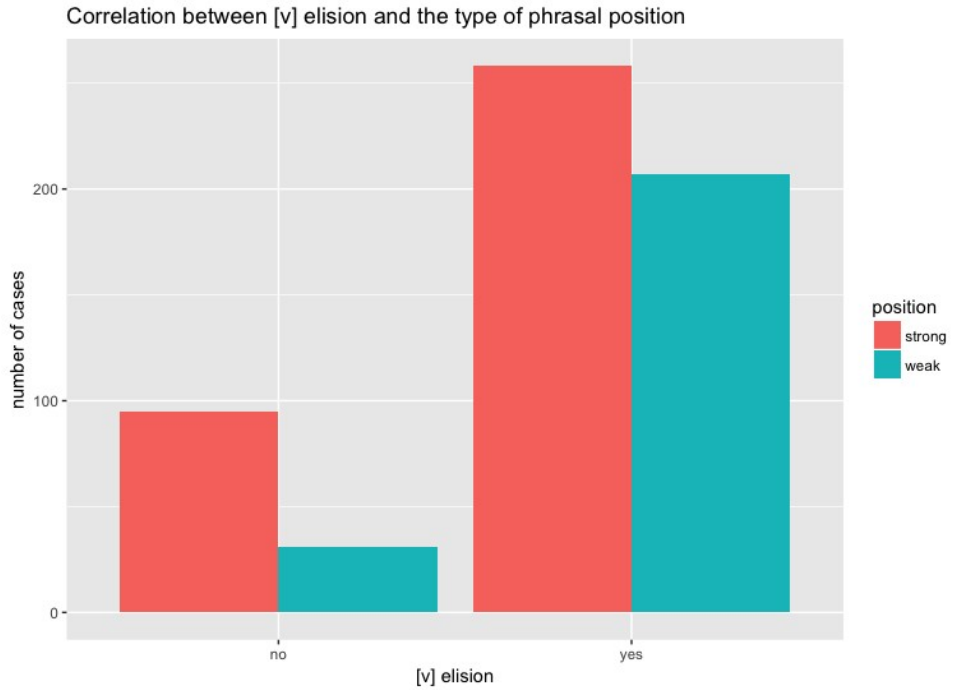
Ни одна из оценок не является блокирующей.

Все элементы контроля подлежат пересдаче в виде 2-ух часовой контрольной работы по всем темам, во время которой можно пользоваться любыми материалами. Время проведения

устанавливается факультетом гуманитарных наук. Тематический состав КИМ-ов для пересдач не отличается от тематического состава КИМ-ов текущего контроля и промежуточной аттестации.

4.Примеры оценочных средств

Постройте следующий график на основании следующего датасета:



Посчитайте значение корреляции двух переменных встроенного датасета ggplot2::diamonds: depth и price.

5. Ресурсы

5.1 Рекомендуемая основная литература

- Wickham, Hadley, and Garrett Grolemund. *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.", 2016. доступна онлайн: https://r4ds.had.co.nz/
- Bivand, R. S. Applied spatial data analysis with R / R. S. Bivand, E. J. Pebesma, V. Gomez-Rubio. – New York: Springer, 2008. – 374 с. – (Use R!) . – На англ. яз. - ISBN 978-0-387-78170-9.
- Wickham, H. R for data science: import, tidy, transform, visualize, and model data / H. Wickham, G. Grolemund. – Sebastopol: O'Reilly, 2017. – 492 с. – На англ. яз. - ISBN 9781491910399: 2002.55.

5.2 Рекомендуемая дополнительная литература

• Xie, Y. Dynamic documents with R and knitr / Y. Xie. – Boca Raton; London; New York: CRC Press, 2014. – 190 с. – (The R series) . – На англ. яз. - ISBN 978-1-482-20353-0.
• Spector, P. Data manipulation with R / P. Spector. – New York: Springer, 2008. – 152 с. – (Use R!). – На англ. яз. - ISBN 978-0-387-74730-9.
• Wickham, H. ggplot2: elegant graphics for data analysis / H. Wickham. – Dordrecht: Springer, 2009. – 212 с. – (Use R!). – На англ. яз. - ISBN 978-0-387-98140-6.
• Wickham, H. Advanced R / H. Wickham. – Boca Raton [etc.]: CRC Press, 2014. – 456 с. – (Chapman & Hall/CRC. The R Series) . – На англ. яз. - ISBN 978-1-466-58696-3.

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	R	*Распространяется бесплатно*
2.	RStudio	*Распространяется бесплатно*
3.	Git	*Распространяется бесплатно*

5.4 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Github	https://github.com/
2.	Онлайн курс Программирование (язык R)	https://openedu.ru/course/hse/RLING/

5.5 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

• ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

• мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены ПЭВМ, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.

6. Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося) а для инвалидов также в соответствии с индивидуальной программой реабилитации инвалида, могут предлагаться следующие варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:

i. *для лиц с нарушениями зрения:* в печатной форме увеличенным шрифтом; в форме электронного документа; в форме аудиофайла (перевод учебных материалов в аудиоформат); в печатной форме на языке Брайля; индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания и консультации.

ii. *для лиц с нарушениями слуха*: в печатной форме; в форме электронного документа; видеоматериалы с субтитрами; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания и консультации.

iii. *для лиц с нарушениями опорно-двигательного аппарата*: в печатной форме; в форме электронного документа; в форме аудиофайла; индивидуальные задания и консультации.