Программа учебной дисциплины
«Анализ и визуализация данных»

Утверждена
Академическим советом ООП
Протокол № 4 от «21» июня 2018 г.

Автор	Деркач Денис Александрович
Число кредитов	4
Контактная работа (час.)	46
Самостоятельная работа (час.)	106
Курс	Бакалавриат, 2 курс
Формат изучения дисциплины	без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Цель освоения дисциплины “ Анализ и визуализация данных ” является ознакомление студентов с основными понятиями и принципами статистического анализа данных, а также обучение корректному представлению и визуализации таких данных.

В результате освоения дисциплины студент должен знать:

- Принципы корректного сбора и интерпретации статистических данных.
- Основные дескриптивные метрики для количественных данных (меры центральности и размаха).
- Основные правила информативного дизайна для визуализации количественной информации.

В результате освоения дисциплины студент должен уметь:

- Описывать выборки и генеральные совокупности на основе имеющихся дескриптивных статистик.
- Интерпретировать результаты простых экспериментальных и корреляционных исследований.
- Находить ошибки в неверной интерпретации дескриптивных статистик и плохой визуализации количественных данных.
- Выбрать корректный тип визуализации для определенной задачи представления информации.
- Строить разные виды графиков.
- Выделять потенциальные зависимые и независимые переменные в наборе данных для проведение простого корреляционного анализа.

Изучение данной дисциплины базируется на следующих дисциплинах:

- “Цифровая грамотность”

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

- знать математику в объеме средней школы

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

- Научно-исследовательский семинар

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Введение в анализ данных. Кейсы применения анализа данных и визуализаций в современных редакциях. Источники данных (открытые и закрытые базы, скрейпинг данных из всемирной сети, Google Trends, краудсорсинг и онлайн-эксперименты). Обсуждение кейсов подтасовки данных.

Тема 2. Введение в статистику. Понятия выборки и генеральной совокупности. Введение в теорию вероятности. Методы эксперимента и наблюдения. Виды переменных (категориальные, порядковые, непрерывные, дискретные). Главные вопросы при работе с данными: как были собраны данные, насколько надежна информация. Понятия концептуализации и валидности.

Тема 3. Описываем данные: распределение, меры центральности, дисперсия, корреляция.

Что нам говорят опросы? Введение в наблюдательные исследования. Оцениваем качество выборки (случайная или неслучайная выборка, методы случайной выборки). Оценочные статистики выборки и их обобщение на генеральную совокупность.

Тема 4. Анализируем данные и оцениваем чужие исследования: тестирование гипотез и p-величина, виды ошибок при тестировании гипотез, оценка силы корреляции и причинно-следственная связь. Логистическая и линейная регрессии.

Тема 5. Введение в экспериментальные исследования. Т-критерий Стьюдента. Интерпретация результатов экспериментальных исследований на основе оценки качества эксперимента, выборки, анализа статистических результатов. A/B тестирование. Этические аспекты экспериментальных исследований.

Тема 6. Избегаем ложных умозаключений: разбираемся с самыми распространенными когнитивными ошибками, касающихся презентации данных (якорение, доступность, репрезентативность, фрэйминг, статус-кво).

Тема 7. Как правильно интерпретировать визуально представленную информацию. Основные виды графиков и особенности их применения (круговая диаграмма, гистограмма, тренд, линейный график, диаграмма рассеяния, пузырьковый график и т.д.). Основы сетевого анализа и интерпретация визуализаций сетей.

Тема 8. Основные принципы хорошей визуализации и представления данных: как отличить хорошую визуализацию от плохой, как не ввести потребителя медиа в заблуждение и не обмануться самому. Обзор бесплатных инструментов для создания статических и интерактивных визуализаций.

III. ОЦЕНИВАНИЕ

Результирующая оценка по дисциплине рассчитывается по формуле

$$O\_итог = 0.8\ O\_накопл + 0.2\ O\_экз$$

Накопленная и итоговая оценки округляются арифметически.

- Накопленная оценка складывается из оценок за самостоятельные работы (практические работы, контрольные работы, домашнее задание, тесты) и проект. Максимальная накопленная оценка – 120 баллов (затем переводится в 10-балльную шкалу путём деления на 12 и округления в большую сторону по математическим правилам):
- *Максимальная оценка за проект (О_проект) - 30 баллов*
- *Максимальная оценка за домашнее задание (О_дз) - 20 баллов*
- *Максимальная оценка за каждую контрольную работу (О_кр) - 13 баллов*
- *Максимальная оценка за каждое практическое задание (О_практич) - 10 баллов*
- *Максимальная оценка за один тест (О_тесты) - 3 балла*

$$O\_накопл = (O\_проект + O\_кр + O\_дз + O\_практич + O\_тесты)/12$$

Оценки за все виды работ суммируются. Оценка за тест выставляется по результатам тестирования, проводимого после каждой лекции. Количество баллов за разные задания внутри работ может различаться в зависимости от их сложности. Все промежуточные оценки (за домашние, самостоятельные и коллоквиум) могут быть не целыми. Накопленная и итоговая оценки округляются математически.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства для текущего контроля студента

Примерные вопросы самостоятельных работ:

1. Найдите среднее, медиану и моду для следующих данных.
2. Дайте определение среднеквадратичного отклонения и напишите формулу.
3. Почему корреляция не всегда означает причинно-следственную связь. Приведите пример.
4. Проинтерпретируйте график с линией регрессии

Примеры домашних заданий

1. Презентация конструктивного анализа плохой визуализации. Предложение по альтернативной визуализации.

2. Визуализация данных выполненная в MS Excel в соответствии с принципами хорошего дизайна и восприятия информации.

Оценочные средства для промежуточной аттестации

Вариант проекта: соберите данные из социальной сети или с новостного сайта для последующего текстового анализа. Проведите разведывательный анализ данных. Постройте регрессию. Визуализируйте данные.

V. РЕСУРСЫ

5.1 Основная литература

1. Thomas, Seemon. Basic Statistics, Alpha Science Internation, 2014. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=5190782

5.2 Дополнительная литература

1. Башмакова, Е. И. Умный Excel. Экономические расчеты : учеб. пособие [Электронный ресурс]/ Б33 Е. И. Башмакова. — М.: Издательство Московского гуманитарного университета, 2014. — 176 с. - ISBN 978-5-906768-21-6. ББК 99.99. Режим доступа: https://elibrary.ru/item.asp?id=25043042

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа
3.	Microsoft Windows 7 Professional RUS Microsoft Windows 10 Microsoft Windows 8.1 Professional RUS	Из внутренней сети университета (договор)
4.	Microsoft Office Professional Plus 2010	Из внутренней сети университета (договор)

5.4 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;
- мультимедийный проектор с дистанционным управлением.

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