Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Центр сопровождения образовательных инициативных проектов

УТВЕРЖДЕНО:

Руководитель сетевой ОПОП
В.В. Кашпур

Рабочая программа дисциплины

Чистка, обработка и исследовательский анализ данных

по направлению подготовки

09.04.03 Прикладная информатика

Направленность (профиль) подготовки:
«Дата-аналитика для бизнеса»

Форма обучения
Очная

Квалификация
Магистр

Год приема
2023

Томск – 2023

1. Цель дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

ОПК 3 - способность анализировать датасеты (наборы данных), выделять в них главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями;

ПК 4 - способность разрабатывать и реализовывать тематические программы с использованием инструментов анализа данных.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК 3.1 – способность использовать принципы, методы и средства решения стандартных задач профессиональной деятельности на основе информационной и библиографической культуры с применением информационнокоммуникационных технологий и с учетом основных требований информационной безопасности.

ИОПК 3.2 – способность применять информационно-коммуникационные технологии решения стандартных задач профессиональной деятельности на основе информационной и библиографической культуры с учетом основных требований информационной безопасности.

ИОПК 3.3 – способность решать стандартные задачи профессиональной деятельности с соблюдением требований информационной безопасности.

ИПК 1.2 – способность осуществлять сбор, анализ и систематизацию информации по проблеме исследования, проводит исследование.

ИПК 4.3 – способность принимать эффективные проектные решения в условиях неопределенности и риска.

2. Задачи освоения дисциплины

- **применять к датасету самые частые способы обработки пропусков**
  - определить причину пропусков (выдвинуть несколько гипотез того, откуда взялись пропуски с учётом решаемой задачи и понимания организации и работы доменной области, отличить случайные от неслучайных пропусков);
  - объяснить разницу между категориальными и количественными значениями
  - классифицировать категориальные и количественные значения в наборе данных
  - применять методы обработки пропусков (заполнять данными исходя из логических связей; заменять пропуски средним, медианой, модой, наиболее частым, нулём с помощью);
  - понимать их применимость в контексте задачи;
- **преобразовывать основные типы данных для применения в различных задачах**
  - использовать стандартные методы pandas (to_datetime, to_numeric, as_type), чтобы преобразовывать основные типы данных
  - работать со значениями даты и времени в строковом виде любого формата (добавлять часы/минуты/дни к дате; получать из дат любые атрибуты (год, час итп);
  - назвать определение и преимущества формата unixtime для хранения дат и вычислений
- **работать с несовершенными реальными наборами данных**
  - перечислить по памяти типы ошибок и способы определения типов ошибок;

- **находить дубликаты в датасете с помощью основных методов**
- **визуализировать свойства данных для понимания их структуры**
    - строить графики hist, boxplot с помощью инструментов [pandas];
    - объяснить, что такое гистограмма и бокс плот и зачем они нужны, отличать гистограмму от бокс-плота;
    - объяснить, что такое квартили, медиана, стандартная ошибка;
- **изучать срезы данных для понимания структуры данных**
    - делать срез данных по смыслу (выборки с нужными условиями);
    - использовать переменную внутри метода query и формировать составные сложные условия с помощью query;
    - приводить к типу datetime в pandas;
    - округлять значения дат (в рамках часов, минут и тд) и добавлять временные интервалы к дате (например, приводить в другой часовой пояс, считать дату доставки и тд)
    - объяснять, как выделять новые признаки;
    - рассчитывать статистики относительно генеральной совокупности (основных данных);
    - объяснять, как подготовить баг-репорт (обобщенно).
- **объединять данные для обогащения со стороны других наборов данных**
    - повторять метод query;
    - фильтровать по вхождению в список;
    - повторять добавление нового столбца и устанавливать столбец, как индекс;
    - объединять две таблицы (вспомогательный dataframe, series);
    - переименовывать столбцы с помощью атрибута;
- **уметь использовать совместное распределение (scatter plot) для изучения взаимосвязей между данными**
    - строить совместное распределение и интерпретировать результаты;
    - объяснить, что такое корреляция и интерпретировать словами результаты подсчета корреляции;
    - считать корреляцию пирсона;
    - строить scatter plot (scatter matrix) с помощью инструментов (pandas)
- **группировать данные из разных источников, чтобы получить обобщенные результаты**

3. Место дисциплины (модуля) в структуре образовательной программы

Дисциплина относится к Блоку 1 «Дисциплины (модули)».

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, является обязательной для изучения.

4. **Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине**

Семестр 1, зачет с оценкой.

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования.

6. Язык реализации

Русский

7. Объем дисциплины (модуля)

Общая трудоемкость дисциплины составляет 4 з.е., 144 часа, из которых:

– лекции: 72 ч.;
– семинарские занятия: 0 ч.
– практические занятия: 62 ч.;
– лабораторные работы: 10 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины (модуля), структурированное по темам

Тема 1. Введение в предобработку данных.
Цели и задачи чистки и предобработки данных.

Тема 2. Работа с пропусками.
Различать категориальные и количественные данные. Выявлять пропуски — случайные и неслучайные. Заменять пропуски на значения – расчётные, дефолтные, характерные.

Тема 3. Изменение типов данных.
Переводить строку в форматы даты и времени; Превращать строковые значения в числовые методами `to_numeric()` и `astype()`; Обрабатывать ошибки связкой `try-except`; Соединять таблицы методом `merge()`; Создавать сводные таблицы методом `pivot_table()`.

Тема 4. Поиск дубликатов.
Изучить способы ручного поиска дубликатов; обрабатывать дубликаты с учётом регистра.

Тема 5. Категоризация данных.
Выделять из данных словарь категорий; Разделять данные на категории по числовому признаку; Писать функции для обработки сразу нескольких ячеек в строке.

Тема 6. Системное и критическое мышление в работе аналитика.
Природа ошибок. Системное мышление в поиске данных. Что делать со своими ошибками?
Тема 7. Введение в исследовательский анализ данных.
Цели и задачи исследовательского анализа данных.

Тема 8. Первые графики и выводы.
Строить гистограммы методом `hist()` **и диаграммы размаха методом **`boxplot()`. Определять типы распределений: нормальное и Пуассона. Получать числовое описание данных методом `describe()`.

Тема 9. Изучение срезов данных.
Получать срезы данных вручную и методом `query()`.Округлять время и переводить его в другие часовые пояса. Строить графики методом `plot()`. Составлять правильные баг-репорты.

Тема 10. Работа с несколькими источниками данных.
Делать срез по данным из внешнего словаря. Создавать новый столбец по данным из другого датафрейма, списка и `Series`. Присваивать столбцы по порядку строк или по совпадению индексов. Объединять данные из двух таблиц. Применять методы `join()` и `merge()` для слияния столбцов.

Тема 11. Взаимосвязь данных.
Строить диаграмму рассеяния. Считать коэффициент корреляции Пирсона. Находить совместное распределение для множества величин.

Тема 12. Валидация результатов.
строить столбчатые графики и круговые диаграммы; выборочно изменять значения в списке методом `where()`; писать цикл, который строит сразу много гистограмм.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем проведения тестов по лекционному материалу (включены в электронный учебник), практических заданий на применение изученного и самостоятельных проектов. Результаты фиксируются в форме контрольной точки не менее одного раза в семестр.

Оценочные материалы текущего контроля размещены на сайте ТГУ в разделе «Информация об образовательной программе» – https://www.tsu.ru/sveden/education/eduop/.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Зачет с оценкой в первом семестре проводится в письменной форме по результатам защиты итоговых проектов по теме «Введение в предобработку данных» и «Введение в исследовательский анализ данных».

Оценочные материалы для проведения промежуточной аттестации размещены на сайте ТГУ в разделе «Информация об образовательной программе» – https://www.tsu.ru/sveden/education/eduop/.

Пример описания проекта по модулю «Введение в предобработку данных»:

Проект выполняется на основе представленных данных от банка – статистики о платёжеспособности клиентов. Заказчик — кредитный отдел банка. Проанализируйте данные и оцените, влияет ли семейное положение и количество детей клиента на факт погашения кредита в срок.
Результаты исследования будут учтены при построении модели *кредитного скоринга* — специальной системы, которая оценивает способность потенциального заёмщика вернуть кредит банку.

При оценке проекта будут учитываться такие критерии, как:

- Как вы описываете найденные в данных проблемы?
- Какие методы замены типов данных, обработки пропусков и дубликатов применяете?
- Категоризируете ли данные? Почему именно таким образом?
- Выводите ли финальные данные в сводных таблицах с помощью метода `pivot_table()`?
- Применяете ли конструкцию `try...except...` для обработки потенциальных ошибок?
- Соблюдаете ли структуру проекта и поддерживаете аккуратность кода?
- Какие выводы делаете?

Результаты зачета с оценкой определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно». Итоговая оценка будет складываться из результатов двух крупных проектов по темам 1-6 и темам 7-12. При успешной сдаче двух проектов будет выставлена итоговая оценка.

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине на платформе Яндекс. Практикума

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине (https://www.tsu.ru/sveden/education/eduop/).

в) План вебинаров и семинаров с преподавателем.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

- Груздев А.В. "Изучаем pandas. Высокопроизводительная обработка и анализ данных в Python" / Груздев А.В. - ДМК Пресс, 2019. - 650 с.
- Пасхавер Борис "Pandas в действии" / Пасхавер Борис - Издательство Питер, 2022. - 512 с.

б) дополнительная литература:

- Маккини Уэс "Python и анализ данных" / Маккини Уэс - ДМК Пресс, 2020. - 540 с.

в) ресурсы сети Интернет:

- Аналитикам: большая шпаргалка по Pandas https://smysl.io/blog/pandas/
- Изучаем pandas. Урок 3. Доступ к данным в структурах pandas https://devpractice.ru/pandas-indexing-part3/
- Открытый курс машинного обучения. Тема 1. Первичный анализ данных с Pandas https://habr.com/ru/companies/ods/articles/322626/
- Открытый курс машинного обучения. Тема 2: Визуализация данных с Python https://habr.com/ru/companies/ods/articles/323210/
- 5 простых способов визуализации данных на Python https://medium.com/nuances-of-programming/5-%D0%BF%D1%80%D0%BE%D1%81%D1%82%D1%8B%D1%85-%D1%81%D0%BF%D0%BE%D1%81%D0%BE%D0%B1%D0%BE%D0%B2-%D0%B2%D0%B8%D0%B7%D1%83%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D0%B8-%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85-%D0%BD%D0%B0-python-%D1%81-%D0%BA%D0%BE%D0%B4%D0%BE%D0%BC-e0053808c83d
- Построение графиков в Python при помощи Matplotlib https://python-scripts.com/matplotlib
- 6 Examples of Correlation in Real Life https://www.statology.org/correlation-examples-in-real-life/
- Обработка пропусков в данных https://loginom.ru/blog/missing
- Руководство по использованию pandas для анализа больших наборов данных https://habr.com/ru/companies/ruvds/articles/442516/
- Нечеткий поиск в словаре с универсальным автоматом Левенштейна https://habr.com/ru/articles/275937/

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

- публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.);
- свободно распространяемый пакет аналитических программ Anaconda Distribution, включающий в себя язык программирования Python, среду

программирования Jupyter и набор базовых библиотек для анализа данных (включая pandas, numpy и другие);
- образовательная платформа Яндекс Практикум.

б) информационные справочные системы:
- Электронный каталог Научной библиотеки ТГУ – http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system
- Электронная библиотека (репозиторий) ТГУ – http://vital.lib.tsu.ru/vital/access/manager/Index
- ЭБС Лань – http://e.lanbook.com/
- ЭБС Консультант студента – http://www.studentlibrary.ru/
- Образовательная платформа Юрайт – https://urait.ru/
- ЭБС ZNANIUM.com – https://znanium.com/
- ЭБС IPRbooks – http://www.iprbookshop.ru/

14. Материально-техническое обеспечение

Занятия по учебной дисциплине проводятся с использованием дистанционных образовательных технологий. Каждый обучающийся обеспечен доступом к образовательной платформе https://practicum.yandex.ru/profile/high-education-data-analyst-magistr/

15. Информация о разработчиках

Литовченко Дмитрий Евгеньевич, Яндекс Практикум, продуктовый аналитик, менеджер образовательных программ "Аналитик данных" / "Аналитик данных Плюс" /"Аналитик данных Bootcamp"

Зотов Вячеслав Анатольевич, кандидат технических наук, старший аналитик данных, старший эксперт программ по анализу данных в Яндекс. Практикум