Лабораторная работа 1 «Подготовка данных. Разведочный анализ данных»

1. Загрузить набор данных в Python (желательно использовать Google Colab). Название блокнота «Номер группы, ФИО, ЛР 1.
2. Ознакомиться с набором данных. Исключить неинформативные столбцы (обосновать их исключение (текстовая запись в блокноте)). Проверить форматы данных. Найти и исключить дубликаты.
3. Найти и заполнить пропуски. Реализуем метод, сравниваем описательную статистику показателя до и после заполнения пропусков, формулируем выводы: как повлияла процедура замены на распределение показателей, применимость или неприменимость метода в данном случае. Методы заполнения пропусков:
   1. Удаление строк / столбцов с пропущенными значениями
   2. Замена пропущенных значений константой
   3. Параметры центральной тенденции: среднее арифметическое значение, медиана, мода
   4. Внутригрупповые параметры центральной тенденции: среднее арифметическое значение, медиана, мода
   5. Линейная / нелинейная регрессия
   6. Стохастическая линейная / нелинейная регрессия
   7. Метод k-ближайших соседей (k-NN)

Выбрать наилучший метод. В качестве эталонного метода взять k-NN. Сравнить среднее относительное отклонение 67 восстановленных значений, полученных методами b-f, от эталонных значений.

1. Работа с выбросами. Определить выбросы по каждому показателю. Провести содержательный анализ выбросов: какие вузы, в какой период, является ли выброс случайным или закономерным (используйте динамику по годам). Как следует обработать выбросы? Подберите метод и приведите обоснование. Как изменился ряд после обработки выбросов, анализ до / после.
2. Описательный анализ показателей. Выводы
3. Графический анализ показателей. Выводы о структуре данных.
4. Корреляционный анализ показателей. Выводы о взаимосвязях. Что на что влияет? Сформулировать модель «белый ящик» взаимосвязи показателей, модель результативности деятельности вуза.

**Описание набора данных**

Набор содержит основные показатели результативности деятельности всех российских организаций высшего образования (за исключением негосударственных вузов и филиалов) за 2013-2017 годы. Эти данные собираются и публикуются в рамках мониторинга эффективности деятельности организаций высшего образования Минобрнауки РФ, но в рамках этого датасета были приведены к более удобному для машинной обработки формату (панельные данные в формате плоской таблицы).

Единица наблюдения в наборе (строка): организация высшего образования (университет) за определенный год. Для каждого наблюдения представлены следующие атрибуты:

федеральный округ, регион, ОКАТО региона, к которому относится организация высшего образования;

показатели, отражающие результативность образовательной, научно-исследовательской деятельности вуза, а также некоторые характеристики его инфраструктуры.

Набор охватывает все доступные сведения за 2013-2017 гг. Всего набор содержит 2 564 наблюдений по 27 атрибутам.

Структура набора данных

Показатели результативности российских вузов: объединенные данные за 2013-2017 годы (data)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Атрибут** | **Описание** | **Количество пропусков (NaN) (шт/%)** | **Единица измерения** | **ОКЕИ** | **Формат** |
| federal\_district | Федеральный округ (например, «Центральный федеральный округ») | 0 |  |  | cat |
| federal\_district\_short | Сокращенное наименование Федерального округа (например, «ЦФО») | 0 |  |  | cat |
| region\_code | Код субъекта РФ | 0 |  |  | cat |
| region\_name | Субъект РФ (например, «г. Москва») | 0 |  |  | cat |
| okato | ОКАТО | 0 |  |  | cat |
| id | Идентификационный номер образовательной организации | 0 |  |  | integer |
| name | Полное наименование образовательной организации (например, «Федеральное государственное бюджетное образовательное учреждение высшего образования «Юго-Западный государственный университет») | 0 |  |  | text |
| name\_short | Короткое наименование образовательной организации (например, «ФГБОУ ВО «Юго-Западный государственный университет», Юго-Западный государственный университет, ФГБОУ ВО «ЮЗГУ», ЮЗГУ») | 0 |  |  | text |
| year | Год | 0 |  |  | integer |
| e1 | Отношение суммы средних баллов ЕГЭ студентов, принятых на обучение по очной форме обучения по результатам ЕГЭ или по результатам ЕГЭ и дополнительных испытаний за исключением лиц, принятых по результатам целевого приема, по всем направлениям и специальностям программ бакалавриата и специалитета, умноженных на численность таких студентов, обучающихся по соответствующим направлениям и специальностям программ бакалавриата и специалитета, к суммарной численности таких студентов. Результаты студентов, имеющих право на прием без вступительных испытаний, признаются как наивысшие результаты ЕГЭ (100 баллов) по соответствующим общеобразовательным предметам | 0 |  |  | numeric |
| e2 | Отношение общего объема средств, поступивших за отчетный год от выполнения НИОКР к численности НПР. Численность НПР (как общая численность, так и численность всех возможных подмножеств) здесь и далее рассчитывается как приведенная к числу ставок численность работников профессорско-преподавательского состава и научных работников на 01 октября отчетного года, включая работающих на условиях штатного совместительства (внешних совместителей), без работающих по договорам гражданско-правового характера | 0 |  |  | numeric |
| e3 | Отношение приведенного контингента студентов – граждан иностранных государств, к приведенному контингенту студентов (кроме студентов, обучающихся на специальностях, на которые не предусмотрен прием иностранных граждан), выраженное в процентах | 0 | процент |  | numeric |
| e4 | Отношение объема средств организации, поступивших за отчетный год из бюджетных и внебюджетных источников, без учета собственных средств, к численности НПР | 0 |  |  | numeric |
| e5 | Отношение фонда начисленной заработной платы работников профессорско-преподавательского состава, без учета работающих на условиях штатного совместительства (внешних совместителей) и по договорам гражданско-правового характера, к среднесписочной численности таких работников, деленное на 12 и деленное на соответствующую среднемесячную начисленную заработную плату наемных работников в организациях, у индивидуальных предпринимателей и физических лиц (среднемесячный доход от трудовой деятельности) в субъекте Российской Федерации, выраженное в процентах | 0 | процент |  | numeric |
| e6 | Отношение численности трудоустроившихся в течение календарного года, следующего за годом выпуска, выпускников образовательной организации высшего образования, к общей численности выпускников образовательной организации года, предшествующего отчетному, обучавшихся по очной форме обучения по программам бакалавриата, программам специалитета, программам ординатуры/интернатуры и получавших образование данного уровня впервые (первое высшее образование), вне зависимости от основы обучения, за исключением продолживших обучение и (или) являющихся иностранными гражданами, выраженное в процентах | 67 | процент |  | numeric |
| e8 | Отношение приведенной к числу ставок численности работников ППС, имеющих ученую степень доктора наук или кандидата наук, к численности студентов, обучающихся по программам бакалавриата, специалитета и магистратуры | 0 |  |  | numeric |
| ege\_budg | Средний балл ЕГЭ студентов, принятых по результатам ЕГЭ на обучение по очной форме по программам бакалавриата и специалитета за счет средств соответствующих бюджетов бюджетной системы РФ | 0 | балл |  | numeric |
| wos | Число публикаций организации, индексируемых в информационно-аналитической системе научного цитирования Web of Science, в расчете на 100 НПР | 0 | единица |  | numeric |
| scopus | Число публикаций организации, индексируемых в информационно-аналитической системе научного цитирования Scopus, в расчете на 100 НПР | 0 | единица |  | numeric |
| rsci | Число публикаций организации, индексируемых в информационно-аналитической системе научного цитирования РИНЦ, в расчете на 100 НПР | 0 | единица |  | numeric |
| rnd | Общий объем научно-исследовательских и опытно-конструкторских работ | 0 | тысяча рублей |  | numeric |
| total\_income | Доходы вуза из всех источников | 0 | тысяча рублей |  | numeric |
| square | Общая площадь учебно-лабораторных помещений | 0 | квадратный метр |  | integer |
| phd\_share | Удельный вес численности обучающихся (приведенного контингента) подготовки научно-педагогических кадров в аспирантуре (адъюнктуре), ординатуры, ассистентуры-стажировки в общей численности приведенного контингента обучающихся по основным образовательным программам высшего образования | 0 | процент |  | numeric |
| pps | Общая численность ППС (без внешних совместителей и работающих по договорам гражданско-правового характера (ГПХ)) | 0 | человек |  | integer |
| rnd\_private | Доходы от НИОКР (за исключением средств бюджетов бюджетной системы Российской Федерации, государственных фондов поддержки науки) | 0 | тысяча рублей |  | numeric |
| pk | Общая численность студентов (приведенный контингент) | 0 | человек |  | numeric |