Занятие № 6

Работа с пропусками

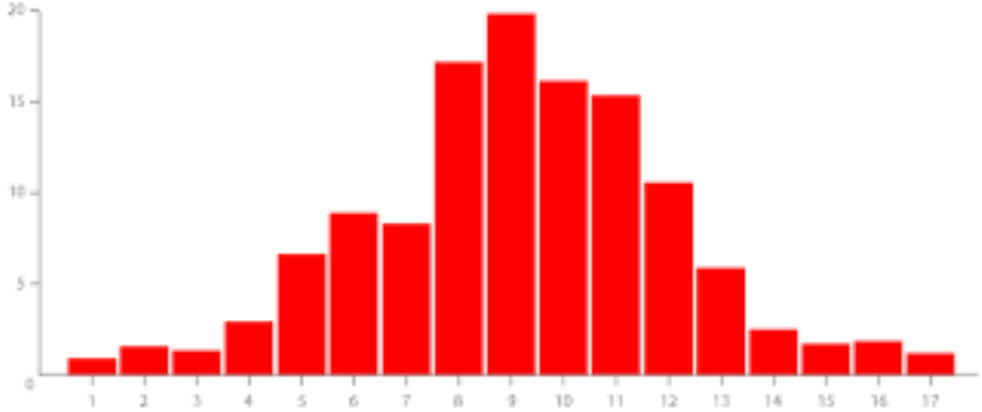
Сапрыкин Артур
Data Scientist

Содержание

Первичный анализ данных. Визуальный анализ данных

EDA - это критический важный процесс первоначального исследования данных с помощью сводной статистики и визуализаций с 4 основными целями:

- Выявить паттерны/зависимости
- Заметить аномалии
- Сформировать гипотезы
- Проверить первичные предположения



Первичный анализ данных. Визуальный анализ данных

1. Как собираются данные?
2. Сколько и каких переменных?
3. Что обозначает каждая переменная, какие единицы измерения и как она собирается?
4. Есть ли пропущенные значения и как они появились?
5. Есть ли аномалии в распределениях?
6. Есть ли корреляции и другие зависимости?

Обработка нулевых значений

Не стоит делать в большинстве случаев:

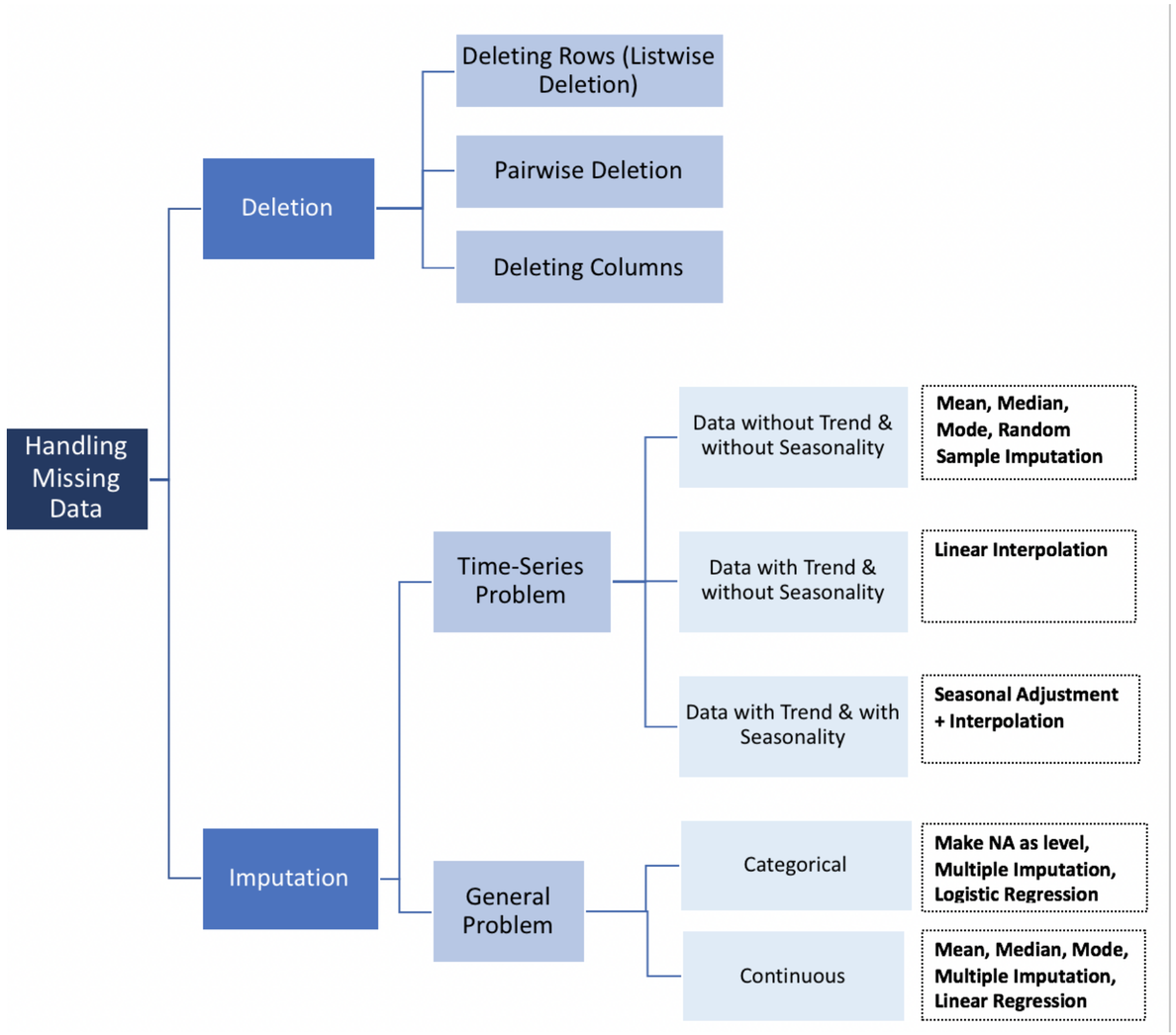
- Удалять столбец содержащий нулевое значение (потеря информации)
- Удалять строки, в которых атрибут равен нулевому значению (потеря информации)

Обработка нулевых значений

Что же делать?

- Заменять на среднее значение, медиану, моду
- Indicator Method - замена пропущенных значений нулями и создание новой переменной индикатора (где она принимает значение 1 при наличие пропуска и
0 в остальных случаях)
- Повторить результат последнего наблюдения (среднее между
- Восстановление пропусков на основе моделей

Обработка нулевых значений



ПРАКТИКА

Спасибо за внимание!

Сапрыкин Артур
Data Scientist

fb.com/asaprykin92

asaprykin92@gmail.com