****Московский государственный технический университет им. Н.Э. Баумана  
Кафедра «Системы обработки информации и управления»

Лабораторная работа №2

по дисциплине «Методы машинного обучения»

на тему«Обработка пропусков в данных»

Выполнил:  
студент группы： ИУ5-23М  
Аимань Мухэяти

Москва — 2021 г.

1.Цель лабораторной работы: изучение продвинутых способов предварительной обработки данных для дальнейшего формирования моделей

### 2.Задание

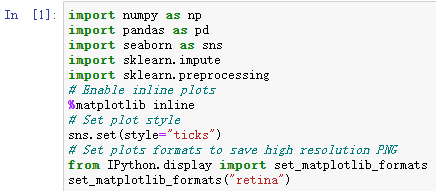
1.Выбрать набор данных (датасет), содержащий категориальные и числовые признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)

2.Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:

* устранение пропусков в данных;
* кодирование категориальных признаков;
* нормализацию числовых признаков.

### 3. Ход выполнения работы

Подключим все необходимые библиотеки



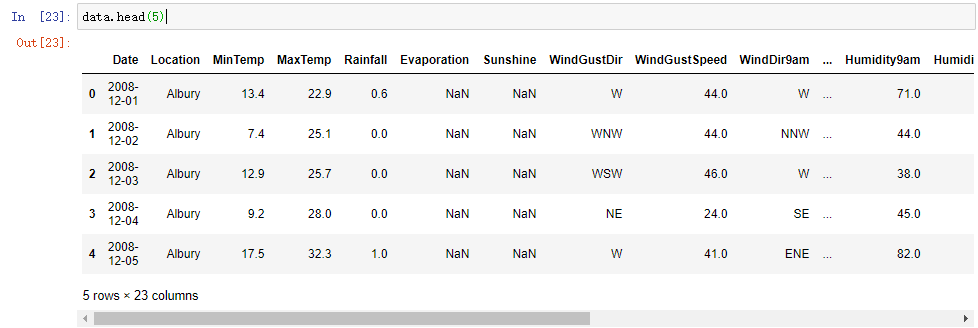
Зададим ширину текстового представления данных, чтобы в дальнейшем текст в отчёте влезал на А4 :

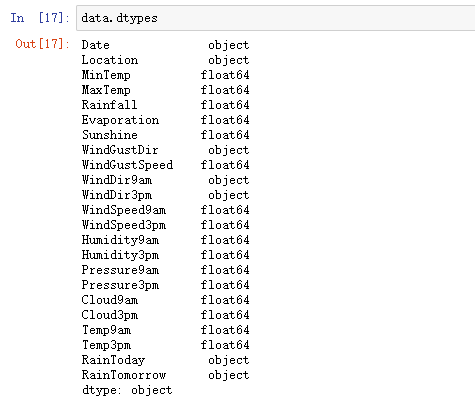


Для выполнения данной лабораторной работы возьмём набор данных по приложениям в waetherAUS ( Погода в Австралии)



Посмотрим на эти наборы данных:

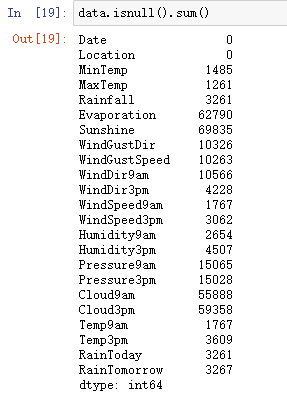




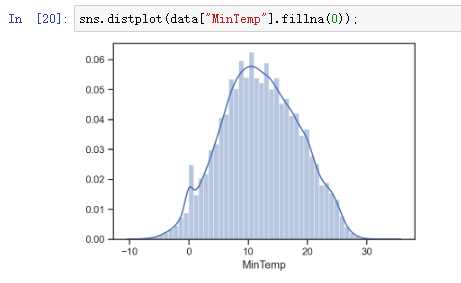


### 3.1. Обработка пропусков в данных

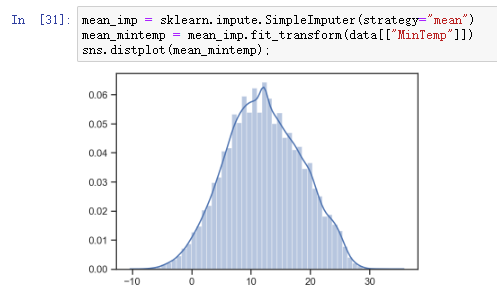
Найдем все пропуски в данных:



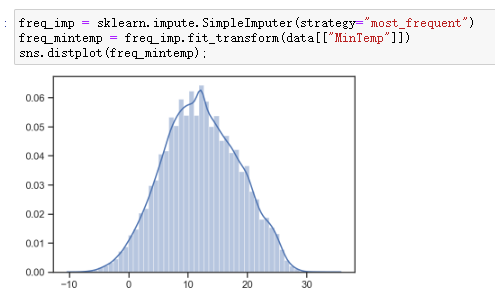
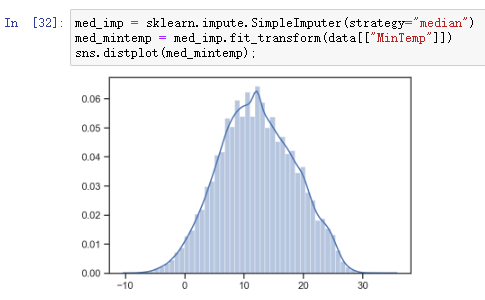
Очевидно, что мы будем работать с колонкой MinTemp. Самый простой вариант — заполнить пропуски нулями:



Видно, что в данной ситуации это приводит к выбросам. Логичнее было бы приложениям без рейтинга присваивать средний рейтинг:



Попробуем также медианный рейтинг и самый частый рейтинг:

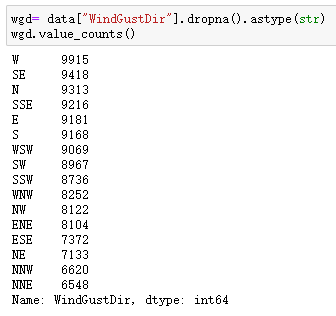


Как видно, эти три значения очень близки к нормальному распределению , поэтому выбираем любое из них на линии, здесь выбираем среднее

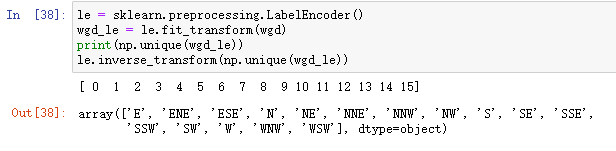


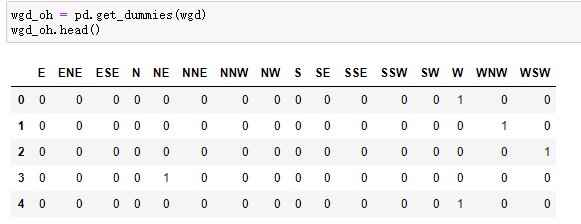
### 3.2. Кодирование категориальных признаков

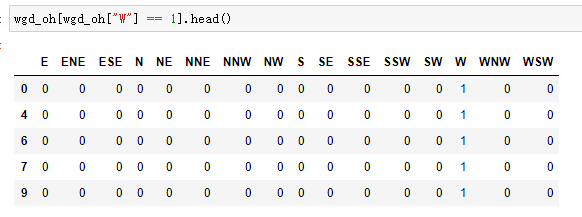
Рассмотрим колонку WindGustDir:



Выполним кодирование категорий целочисленными значениями:

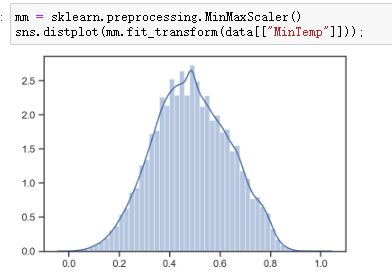
Выполним кодирование категорий наборами бинарных значений:



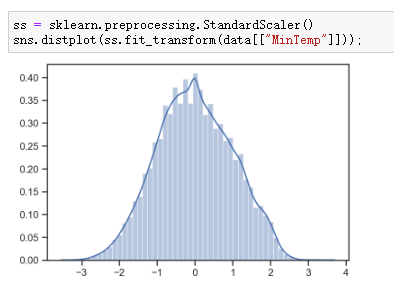


### 3.3. Масштабирование данных

Для начала попробуем обычное MinMax-масштабирование:



Результат вполне ожидаемый и вполне приемлемый. Но попробуем и другие вари-анты, например, масштабирование на основе Z-оценки:



Также результат ожидаемый, но его применимость зависит от дальнейшего использования.

### Список литературы

## [1] Гапанюк Ю. Е. Лабораторная работа «Обработка признаков» [Электронный ресурс] https://github.com/ugapanyuk/ml\_course\_2021/wiki/LAB\_MMO\_\_FEATURES