

2.3 добав вш столбец, join merge

- 1 Срез по данным из вш словаря;
 - 2 Созд новый столбец с вш данными
 - 3 Присваивать столбцы по порядку строк(list) или по совпадению индексов(series)
 - 4 join() и merge() - слияние столбцов.
-

case

чистим от max-min аномалий

медиана очищенных данных(там медиана медианы)

Срез по данным из вш словаря

query in @list уже было

вместо list можно и другие типы данных

in @dict - чекает ключи

@df.index

@df.column

Добав вш столбец df2 в df1

строки заполняются по совпадению индексов

a df2 не имеет индексов df1 - такие строки NaN

b df2 имеет свои уник - такие не добавятся

c df2 повторы индекса - ошибка

= **left join по индексу** (только в sql повторы индекса рожают cross join в подгруппе)

4 join() и merge()

merge:

suffixes=('_x',

'_y') data_x data_y

join:

можно больше двух

default - left

default on-по индексу

case structure:

срезы

1 good_ids = too_fast < 0.5

2

good_data = query id in good_ids +
time_spent [60, 1000]

группировки и джойны

3 good_stations_stat = good_data

gpb(id) - time_spent: median

4 stat = **data**

gpb(name) - time_spent: mean

~~5 good_stat = good_data~~

~~gpb name - time_spent median~~

6 id_name = good_data

gpb(id) - **name**: first count

1stname это первая сеть для id(заправка) =

= сеть этой заправки

7 stations_stat_full =

= 6id_name Join 3good_stations_stat

7 = 6 J 3 можно было сделать одним pivot?

нет, left выкинул строки уникальные для good_stations

а они были? нет

получается можно было)

7 cols:

id 1stname ~~count~~ time_median

8 stations_stat_full

gpb(1stname) - time_median: median count

9 = 8 J 4(name time_mean)

join on name=1stname

result:

time_mean(id) vs

time_median(id)_median(1stname)

среднее заправки(плохие данные) vs медиана медиан заправок для компании(хорошие данные)