# «Создание модели прогнозирования заболеваемости населения»

Сорока Дмитрий

# Постановка задачи.

В результате EDA и плотной работы с предлагаемыми данными, было принято решение рассматривать задачу в разрезе класса задач называемых “Временными рядами”.

В данных четко прослеживаются объекты (характеризуются признаками ПОЛ, БОЛЕЗНЬ, ГОРОД, ВОЗРАСТНАЯ КАТЕГОРИЯ) и динамика/история развития/изменения объектов по месяцам, начиная с января 2018 по март 2022.

# Работа с данными

В процессе работы с данными оказалось удобным преобразовать исходный датафрейм, train, в вид удобный для восприятия/понимания и дальнейшей работы.

Датафрейм train был преобразован к классическому виду, для работы с TimeSeries, недостающие (NaN) значения представляем не как отсутствие наблюдений, а как число больных в этот месяц = 0.

# Выбор модели

ML модель ARIMA отлично справляется с задачами прогнозирования на данных подобных тем, которыми мы располагаем.

В наших данных более 39 тысяч объектов и подбирать гиперпараметры ARIMA для каждого объекта будет слишком затратно по времени, попробуем кластеризовать наши данные и обучать модели покластерно.

# Кластеризация

В процессе работы над задачей очень неплохо показала себя кластеризация временных рядов с TimeSeriesKMeans, но соизмеримый результат, в более простом исполнении, дала кластеризация наших объектов по признаку количества “пропущенных” значений или нулей в объекте.

Так и поступили - кластеризовали данные по признаку “полноты” или “заполненности”.

# Score = 0.940999

Подбор гиперпараметров методом минимизации MSE позволил покластерно обучить модель ARIMA  и получить высокий результат.