💵

актуализация зп

https://colab.research.google.com/drive/1zrqxixeyJLXbC0g1pDHd3aziA8YSUpmM?usp=sharing

Проводилось дополнительное преобразование данных. А именно необходимо было провести актуализацию заработных плат, чтобы модель работала корректно. Пример, когда он могла бы работать не корректно: есть слесарь без навыков и данные по нему за 2022 год, а есть слесарь с каким-то навыком но данные именно с такой вакансией есть только за 2018 год, тогда могло получится так что профессия с каким-то навыком стоила бы меньше чем без него, из-за разницы в годах, поэтому необходимо было актуализировать деньги.

- Идея, лежащая в основе актуализации данных

  Актуализация зарплаты происходила следующим способом:
  Ищем для каждого региона среднюю зарплату, сортируем по возрастанию и разделяем на группы, начиная от наименьшей средней зарплаты, постепенно добавляя стандартное отклонение (именно такое значение величины прироста обеспечивает не слишком большой разброс зарплат для регионов, входящих в одну группу). Таким образом в наших данных появился новый столбец group_region.

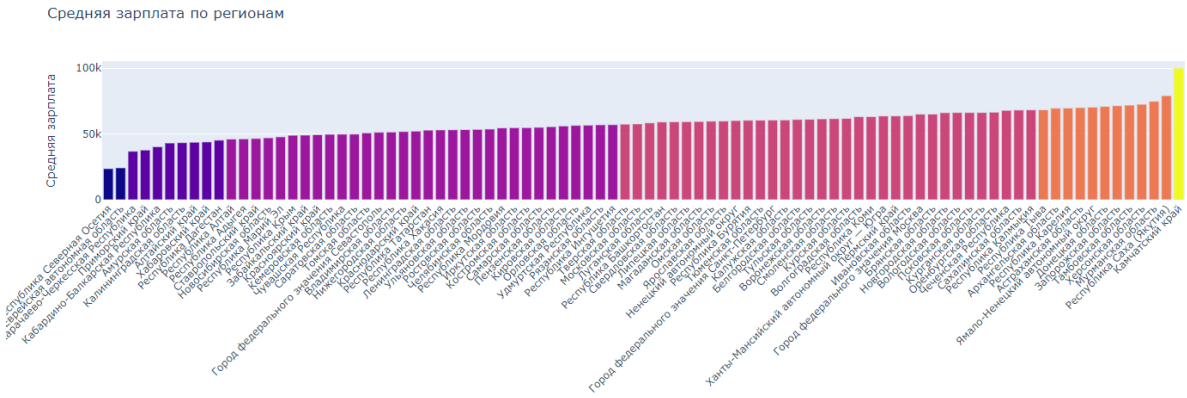
  Общий вид того, как выглядят границы для групп:

  [min_mean_salary : min_mean_salary + std) - первая группа

  [min_mean_salary + std : min_mean_salary + 2*std) - вторая группа

  [min_mean_salary + std*(n-1) : min_mean_salary + n*std] - для n-ой группы

  Ниже представлена визуализация такого деления регионов на группы по зарплате. (Синий цвет - первая группа, фиолетовый - вторая и т.д.)



Таким образом в наших данных появился новый столбец group_region. Далее группируем данные относительно опыта работы и нового столбца group_region и для каждого года вычисляем коэффициент равный отношению средней зарплаты за 2023 и средней зарплаты за рассматриваемый год. Именно на этот коэффициент умножаем текущую зп.

В циклах перебираю уникальные значения опыта работы, группы региона и доступные года. Для каждого такого района нахожу среднюю заработную плату.

Коэффициент = среднее 2023 / среднее за рассматриваемый год

столбец new_salary = Коэффициент * столбец salary_from_rub

в коде написаны подробные комментарии по работе каждой из функций

Данный способ был выбран, потому что в целом изменения в зп, связаны с инфляцией, но просчитывать как именно инфляция влияет сложно, потому что каждая компания для каждой профессии это делает по разному. Они индексируют только какую-то часть зп (какую именно не известно). Поэтому решено было смотреть на изменение рынка определённой профессии в целом (среднем) и так уже актуализировать данные.