***Проект по работе с реальными данными по продажам автомобилей в Казахстане в 2019 году. В данных есть все проблемы, с которыми встречаются аналитики в повседневной работе: “шум”, неверные типы данных и неявные дубли в признаках.***

***Основной целью работы является доведения датасета до состояния, в котором с ним можно работать. По очищенным данным мы проведем EDA, проанализируем рынок и позицию на рынке одного из автоцентров.***

***На выходе мы получим готовый скрипт для очистки последующих данных, понимание особенностей рынка Казахстана, понимание где находится автоцентр, в котором работает аналитик.***

***1. Загружаем данные (decimal =’,’, sep=’;’)***

* проверяем загрузку (после экселя стандартная проблема, что разделитель разрядов - запятая и из-за этого на этапе загрузки данных возникают проблемы, я думаю, что с параметром decimal студенты еще не сталкивались)
* удаляем столбцы: дилер, тип платежа, сегмент, сегмент eng, локализация, тип клиента и модификация (они либо бесполезны, либо много пропусков)
* приводим столбцы к английским названиям для удобства (по желанию)
* удаляем NaN и дубликаты

***2. Чистим данные***

* кодируем страны в столбце страна производства в alpha3 (RUS, KAZ, USA и т.д.)
* кодируем тип топлива в (F - fuel, D - diesel, E - electro, HYB - hybrid или по желанию студента)
* удаляем неявные дубликаты в столбце тип привода
* приведем столбцы стоимость, количество и итоговая стоимость к числовым типам данным (количество минус 1 - это возврат, 100 проданных шкода - это оптовик-таксист судя по комплектации машин - минимальный двигатель и механика-коробка)
* создадим из столбцов год и месяц единый столбец дата продажи (число месяца ставим последний день - 30 или 31)
* удаляем неявные дубликаты в столбце название компании (названия на английском и русском)
* приведем в порядок столбец объем двигателя (заменим запятую на точку, уберем букву L, разберемся с мотором 400 литров у ягуара (там перепутано объем и мощность двигателя - тем кто это найдет и определит нормальный литраж двигателя можно значок давать за въедливость, а можно просто удалить)) + есть неправильные объемы двигателя для Шевроле Нива - 21.6, 20.6 и так далее
* столбцы регион и область - привести к нижнему регистру с заглавной буквы
* столбец тип трансмиссии - убрать неявные дубликаты (для простоты оставим два типа трансмиссии: механика и автомат)
* удалим столбцы месяц и год
* дата продажи к типу дата
* столбцы тип топлива, тип трансмиссии, тип привода, сегмент 2013 и класс 2013 можно перевести в категориальные

***3. Проводим исследовательский анализ данных***

* по столбцам

***4. Анализируем рынок***

***5. Считаем показатели эффективности меркур авто***

***6. Делаем выводы / рекомендации***

**Regex для чайников:**

[Регулярные выражения, пособие для новичков. Часть 1](https://habr.com/ru/articles/115825/)

**Функции и методы строк:**

[Строки. Функции и методы строк | Python 3 для начинающих и чайников](https://pythonworld.ru/tipy-dannyx-v-python/stroki-funkcii-i-metody-strok.html)

[Строки в Python: методы и функции](https://proglib.io/p/samouchitel-po-python-dlya-nachinayushchih-chast-4-metody-raboty-so-strokami-2022-10-24)

**Замена по словарю:**

[Pandas Difference Between map, applymap and apply Methods - Spark By {Examples}](https://sparkbyexamples.com/pandas/pandas-difference-between-map-applymap-and-apply-methods/)

[Карта Pandas DataFrame](https://bestprogrammer.ru/programmirovanie-i-razrabotka/karta-pandas-dataframe)