2.6 case

case
сети азс, конкр азс, время заправки

1базовая проверка и графики
пришлось прикинуть количество машин на заправке в день
ср заправка, топ longest, число заправок по сетям

2найдите аномально быстрые и сверхдолгие заправки. на каких АЗС
Определить границу, после которой можно считать заправку «слишком долгой»
hist(), boxplot(), describe()
выявили зависимость продолжительности заправки от времени заезда
нашли долю аномально быстрых и долгих заправок на разных станциях.
пометили их индикаторным столбцем

3Избавьтесь от сверхбыстрых и аномально медленных заправок
Составьте рейтинг АЗС по времени заправки.
structure of gpbs and joins in 2.3
3медиана по всем заправкам, среднее по сетям АЗС(плох данные), 6заправка-сеть
6джойн3 чтобы дабавить сеть в 3 (можно было пайвот)
таблица с средним по сетям(плохие данные) vs медиана очищенных данных**(медиана медианы по заправкам затем по сетям)**
зачем медиана медианы понятно(рейтинг сети это рейтинг их азс), но почему среднее? среднее, чтобы показать выбросы, среднее среднего изза выбросов слишком исказилось бы

4**Убедитесь в том, что избавление от аномалий не повлияло на природу распределения исходных данных.**
тем что корреляции не очень сильно изменились на совместных распределениях

выявили взаимосвязи медианного (по заправкам) времени и числа (скаттер)
посчитали коэфф корр между числом заездов на АЗС и (медианным по заправкам) временем заправки.
1 матрица диаграмм рассеяния по исходным данным vs по отфильтрованным(аном макс мин)
[ср доля быстрых, ср доля медленных, ср время плохие данные(все), ср время хорошие данные(все)]
Самый большой коэффициент корреляции: 0,8 между 'too_slow' и 'time_spent'. Это заметно и на соответствующей диаграмме рассеяния.
3ная одно значение, можно предсказать другое.
Если бы мы не отрезали слишком долгие заезды, они бы сказались на среднем времени. Поэтому так важно было их отбросить и вместо среднего значения взять медиану.

ВИДНО, что данный пункт искусственный для обучения,
можно было бы использовать медиану и не смотреть корреляции
2 снизили влияние подозрительных данных на итоговый результат.
корреляции между данными слиш быстрых и медленных с нужными упали

кст мы Убрали не очень много долгих заправок. Возможно, ошибка не в данных,
а на АЗС действительно не торопятся заправлять

5**Перепроверьте реалистичность данных, чтобы подготовить финальный отчёт для руководства.**
мелкие сети в одну группу чтобы средние на барплоте отобразить нормально
У многих сетей АЗС явно обнаруживается аномальный пик на коротких поездках.
Но в основном распределение ожидаемой формы, а значит медиана(пункт 3) хорошо передаёт характерное время заправки.
Вы действительно выявили много долгих заправок в самых медленных сетях АЗС.
Причём **форма распределения достаточно плавная: не походит на явную аномалию** на продолжительных временах заправки.