Зеленым обозначено условие для первого варианта, желтым для второго. Задача состоит в том, чтобы провести подготовку данных для моделирования зависимости расхода бензина на трассе (MPG_highway) (или в городе (MPG_city) от числовых переменных Length Weight Wheelbase Horsepower Invoice EngineSize Cylinders и категориальных переменных Origin и Type. В задаче необходимо:

1. Построить регрессию МНК на всех переменных, оценить $R^{2,}$, информационный критерий AIC (BIC), а также скорректированный R^2 (MSE=SSE/df_{error}). Построить графики зависимости остатков от прогноза и гистограмму распределения остатков. Выберите наиболее значимую и наиболее не значимую переменные по критерию Стюдента (при совпадении P-value можно выбрать любую). Что можно сказать о выполнении предположений линейной регрессии для вашей постановки задачи с этими данными?
2. Провести поиск и исключение выбросов с использованием статистики CookD (DFBETA). Необходимо построить графики этих статистик (значения для всех наблюдений выборки) с отметкой 3 наиболее аномальных. Удалите их из выборки, перестройте регрессию МНК. Как изменились все статистики, графики и важность переменных из пункта 1. Прокомментируйте эти изменения.
3. Для выборки после пункта 2 проведите группировку уровней категориальных переменных с помощью попарного t-тест. Перестройте регрессию МНК. Как изменились все статистики, графики и важность переменных из пункта 1. Прокомментируйте изменения t-test для категориальных переменных до и после преобразования.
4. Для выборки после пункта 3 проведите пошаговый отбор любым методом, рассмотренным на лекции (прямой, обратный, комбинированный, LARS), постройте график трассы коэффициентов и график зависимости качества модели на каждом шаге по критерию AIC(BIC). Обратите внимание, что LARS отбирает категориальные переменны по уровням, а вам нужно «целиком». Как изменились все статистики, графики и важность переменных из пункта 1. Прокомментируйте эти изменения.
5. Для выборки после пункта 3 постройте серию регрессий PCR (PLS), перебрав все возможные числа компонент, постройте график зависимости AIC(BIC) от числа компонент, выберите лучшую модель по этому критерию. Как изменились все статистики, графики и важность переменных из пункта 1, пункта 4 и пункта 5. Прокомментируйте эти изменения.

Весь функционал реализовать в виде класса, поддерживающего методы:

- model<-**fit(train_data, step=1)** – строит модель model на основе train_data, включает всю необходимую предобработку данных и построение модели для шага step.
- **summary (model)** – выводит все необходимые статистики для модели с учетом шага, на котором она была построена
- **plot(model)** – строит все необходимые графики модели с учетом шага, на котором она была построена