Зеленым обозначено условие для первого варианта, желтым для второго. Задача состоит в том, чтобы построить регрессионную модель для прогнозирования 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. Можно использовать методы, рассмотренные в рамках курса (линейная регрессия, lars/lasso, pls/pcr, логнормальная, GLM, сплайны, loess, нелинейные регрессии с кастомизированным уравнением, а также их комбинации, например, можно фильтровать переменные одним методом, а модель строить другим по отобранным переменным, или оценивать выбросы одной моделью, потом их удалять из тренировочного набора (из тестового нельзя удалять наблюдения), а строить модель другой процедурой по отфильтрованному набору. Для оценки качества разработанной модели использовать оценку MAPE = (1/n) * Σ(|Original – Predicted| / |Original|). Для оценки качества модели можно использовать любые подходы: информационные критерии, статистические оценки на основе кросс-валидации, бутстреппинга, проверочных выборок и т.д. Финальная модель будет проверяться одним из подходов, параметры которого вы не знаете (чтобы не было подгонки), и вам будет сообщаться результат оценки (MAPE на проверочной выборке или прицедуре). Ваша модель должна работать с любым подмножеством набора данных для обучения и для тестирования (не использовать предположение, что такое-то наблюдение есть в тестовом или в тренировочном наборе). Результирующее MAPE на закрытой проверочной процедуре должно быть меньше или равно0.065. Рекомендуется использовать техники:

- Преобразования категориальных переменных (группировка значений) с использованием дисперсионного анализа (не забывайте делать эти преобразования и в тестовом наборе)
- Отбор значимых переменных с помощью пошаговых методов или регуляризации
- Преобразование входных переменных и добавление полиномиальных членов в уравнение регрессии (не забывайте делать эти преобразования и в тестовом наборе)
- Использовать нелинейные уравнения зависимости в том числе и делать свою «нейросеть» и обучать ее как нелинейную регрессию с помощью nls.
- Преобразование отклика или использование обобщенных линейных (или нелинейных) моделей с разными распределениями ошибки и функциями связи (не забывайте делать правильный пересчет отклика после прогноза).

Для получения требуемого качества модели обычно достаточно использовать две из перечисленных выше техник.

На всем наборе данных построить 3D график (контурный график) зависимости отклика от пары наиболее важных отобранных переменных финальной модели с равномерной сеткой 20 на 20 точек (сетку в наборе данных сгенерировать самостоятельно), значение остальных переменных, не вошедших в пару (если есть), при построении графика усреднять.

Весь функционал реализовать в виде класса, поддерживающего методы:

- model<-**fit(train_data)** – строит модель model на основе train_data (поднабора по строкам исходного набора cars), включает всю необходимую предобработку данных и построение финальной модели
- **predict(model, test_data)** – выдает вектор прогнозов для test_data (поднабора по строкам исходного набора cars), метод должен включать всю предобработку данных, что делал fit (кроме удаления выбросов, исправление признаков наблюдений-выбросов без удаления самого наблюдения возможно)
- **plot(model)** – строит 3D или контурный график