Вопросы на понимание

Упражнение 1 (10 баллов). Ответьте на следующие вопросы:

1. Что можно сказать про случайные величины X и Y, если $\mathrm{Corr}(X, Y) = 1$?
2. Чем отличаются коэффициенты корреляции Пирсона и Спирмена? В каких случаях лучше пользоваться коэффициентом корреляции Спирмена?
3. Что такое коэффициент детерминации? Что происходит с коэффициентом детерминации, при добавлении признаков в регрессионную модель? А при удалении?
4. В каких предположениях работают критерии Стьюдента и Фишера о значимости коэффициентов регрессии?
5. Какую гипотезу проверяет критерий Фишера (F-тест), когда мы вызываем `summary` при построении линейной регрессии в Python или R?

Задачи

Следующие задания — практические. В них необходимо попытаться применить все возможные инструменты, которые мы изучали, чтобы ответить на заданный в задаче вопрос. Не забывайте об одном из самых важных инструментов — визуализации. Ваши решения необходимо сопроводить краткими комментариями и выводами, которые Вы сделали на основе анализа.

Упражнение 2 (20 баллов). В файле `lifeline.xls` содержатся 50 пар наблюдений из исследования докторов Л. Матера и М. Уилсона. В нем рассматривались следующие переменные: X — длина «линии жизни» на левой руке в сантиметрах (с точностью до 0.15 см) и Y — продолжительность жизни человека (округленная до ближайшего целого года). Изучите корреляцию X и Y. Верно ли, что X и Y связаны линейной регрессионной зависимостью?

Упражнение 3 (20 баллов). В файле `metal.xls` содержатся данные о 13 металлургических компаний. Выясните, какие из указанных признаков влияют на цену предприятия. Предварительно исследуйте данные и исключите «выбросы».

Упражнение 4 (25 баллов). В файле `auto.csv` находятся данные, которые описывают характеристики 392 автомобилей. А именно, в данных вы найдете следующие столбцы:

1) `mpg` — расход топлива (миль/галлон);
2) `cylinders` — количество цилиндров мотора;
3) `displacement` — объем мотора (куб. дюйм);
4) `horsepower` — мощность мотора (в лошадиных силах);
5) `weight` — вес (тысяч фунтов);
6) `acceleration` — время, за которое автомобиль разгоняется до 60 mph;
7) `year` — год автомобиля (по модулю 100);
8) `origin` — место производства (1 — Америка, 2 — Европа, 3 — Япония);
9) `name` — название автомобиля.

Какие признаки влияют на расход топлива? Постройте (хорошую) линейную регрессионную модель для предсказания расхода топлива по этим признакам. Интерпретируйте значения коэффициентов регрессии (влияние каждого признака на отклик). Не забудьте о преобразовании признаков, если они будут нужны.

Упражнение 5 (25 баллов). В файле `homes.csv` содержатся данные о цене продажи 50 домов. В файле содержатся следующие признаки:

1) `Sell` — цена продажи;
2) `List` — запрашиваемая цена;
3) `Living` — жилая площадь;
4) `Rooms` — количество комнат;
5) `Beds` — количество спален;
6) `Baths` — количество ванных комнат;
7) `Age` — возраст жилого помещения;
8) `Acres` — площадь жилого помещения;
9) `Taxes` — налоги, которые должен платить владелец жилого помещения.

Что можно сказать о связи первых двух признаков? Постройте (хорошую) линейную регрессию для признака Sell с и без признака List. Не забудьте о преобразовании признаков, если они будут нужны.