**Задание 1 (**реализуйте на Python**)**

Условия.

Данные для задачи находяться в файле **"bikes\_rent.csv"**

Для каждого дня проката известны следующие признаки (как они были указаны в источнике данных):

* *season*: 1 - весна, 2 - лето, 3 - осень, 4 - зима
* \_yr\_: 0 - 2011, 1 - 2012
* *mnth*: от 1 до 12
* *holiday*: 0 - нет праздника, 1 - есть праздник
* *weekday*: от 0 до 6
* *workingday*: 0 - нерабочий день, 1 - рабочий день
* *weathersit*: оценка благоприятности погоды от 1 (чистый, ясный день) до 4 (ливень, туман)
* *temp*: температура в Цельсиях
* *atemp*: температура по ощущениям в Цельсиях
* *hum*: влажность
* *windspeed(mph)*: скорость ветра в милях в час
* *windspeed(ms)*: скорость ветра в метрах в секунду
* *cnt*: количество арендованных велосипедов (это целевой признак, его мы будем предсказывать)

Задание:

* Постройте графики зависимостей целевого признака от остальных.
* Каков характер зависимости числа прокатов от месяца?
* От какого признака число прокатов зависит линейно?
* Оцените уровень линейной зависимости между признаками и целевой переменной.
* Оцените масштаб разброса значений признаков.
* Проведите дополнительные мероприятия, которые считаете нужными, по предобработке данных и обучите линейную регрессию.
* Обучите регрессоры Ridge и Lasso с параметрами по умолчанию.
* Как меняются веса при увеличении коэффициента регуляризации?
* Какой из регуляризаторов подойдет для отбора неинформативных признаков?
* Проведите кросс валидацию и оцените итоговый результат.

**Задание 2**

Есть датасет от агрегатора новостей (файл news.xlsx) Новости содержат упоминания партнеров или контрагентов Сбербанка.

Цель – оставить только релевантные новости, то есть только те новости, которые могут прямо или косвенно повлиять на решения ЛПРов (лиц, принимающих решения) в Сбербанке.

Задача - предложите **алгоритм действий** по обработке датасета для достижения цели.

**Задание 3**

Есть набор данных, описывающих 4950 цифр (csv\_x.csv). Каждая цифра представляется в виде 400 признаков. 400 признаков – это матрица 20х20, характеризующая степень серости пикселя, описывающего цифру.

Есть набор целевых переменных, то есть набор тех самых цифр (4950), которые описывают признаки (csv\_y.csv).

Надо построить нейросеть, которая будет по датасету определять, о какой цифре идет речь. То есть будет классифицировать набор новых входных данных на 10 классов (0, 1, 2…9)

Качество модели будет оцениваться на тестовом наборе данных на следующей встрече.