В 1987 году был проведён опрос посетителей гипермаркета в районе залива Сан-Франциско. В анкете было 502 вопроса и опрошено 9409 посетителей. Предлагаем поработать с урезанным вариантом этих данных — с ответами на 14 вопросов, касающихся демографии.

**Описание данных**: <https://web.stanford.edu/~hastie/ElemStatLearn/datasets/marketing.info.txt>

**Сами данные**: <https://web.stanford.edu/~hastie/ElemStatLearn/datasets/marketing.data>

Подробнее про задачу и возможные подходы можно почитать в книге The Elements of Statistical Learning <https://web.stanford.edu/~hastie/Papers/ESLII.pdf> (страницы 371 и 492), но брать за основу подходы и визуализации из книги не обязательно, в частности, в книге уделяется мало внимание EDA, что очень важно для нашего задания. Этот пример также рассмотрен в и других статьях и книгах.

Задача: в jupyter-notebook провести EDA данных и построить две модели: 1) прогнозирование уровня дохода по ответам на остальные 13 вопросов; 2) прогнозирование рода занятости по ответам на остальные 13 вопросов.

Оформление решения:

1. Результат (jupyter-notebooks) выложить на github для проверки. Язык программирования Python.
2. Нужен чистовик (jupyter-notebooks), в котором содержится оформленное решение, и черновик (jupyter-notebook), в котором содержатся в том числе и подходы, которые были попробованы, но не сработали.
3. Чистовик должен содержать не только код и результат его выполнения, но и небольшие комментарии к коду и текстовые описания находок и результатов. Комментарии и текст в черновике не нужны.
4. При нежелании, черновик можно не выкладывать, но чистовик должен быть чистовиком.

Что хочется видеть в решении:

1. **EDA**
   1. Работа с данными с использованием библиотеки pandas.
   2. Построение визуализаций (например, с использованием matplotlib).
   3. Обзор данных и находки, которые могут пригодится при построении моделей.
   4. Визуализации и находки, которые могут быть интересны и понятны бизнесу (например, владельцам гипермаркета). Условно, хочется иметь несколько визуализаций, которые можно вставить в презентацию.
2. **Моделирование**
   1. Работа с пропусками (NA в данных).
   2. Работа с категориальными признаками.
   3. Метрики.
   4. Бейзлайн.
   5. Подбор метапараметров моделей.
   6. Кросс-валидация.
   7. Некоторый анализ важности фичей.