1. Выбрать Dataset (набор данных)

Это может быть ваш собственный датасет, любой набор данных из интернета, в том числе из этих источников:

<https://www.kaggle.com/datasets>  
<https://www.data.gov/> - сайт с американскими открытыми данными (на котором в отличии от open e-gov действительно много крутых открытых датасетов)  
<https://www.ukdataservice.ac.uk/get-data/themes.aspx> - аналогично, но UK  
<https://archive.ics.uci.edu/ml/datasets.html?format=mat&task=&att=&area=&numAtt=greater100&numIns=greater1000&type=mvar&sort=nameUp&view=table> - репозиторий датасетов

Будет большим плюсом, если вы возьмёте несколько таблиц, и совместите их. Вполне может быть, что просто merge/joinом не обойдётся

1. Обработать данные – вам нужно привести все ваши данные к правильному формату (числовые поля сделать int или float, избавившись от текстовых данных, даты и время перевести в datetime, и т.д.), почистить (убрать или объединить «мусорные» категории, избавиться от ошибочных/выбросных данных и т.д.), решить проблему отсутствующих значений (NaN).

Также было бы отлично избавиться от неинформативных колонок и провести Feature Engineering.

Для этого пункта можно и нужно использовать не только технические средства, но и знания о предметной области (domain knowledge)

1. Провести анализ – анализ статистических показателей, визуализация. Также на этом этапе можно выдвинуть и проверить гипотезы о ваших данных, попытаться выявить какие-то insightы/скрытые закономерности
2. Регрессионное моделирование, тестирование и анализ полученной модели. Настройка мета-параметров
3. Выводы, выявить плюсы и недостатки модели, сравнить с вашими гипотезами из пункта 3. Предложить пути для улучшения полученной модели