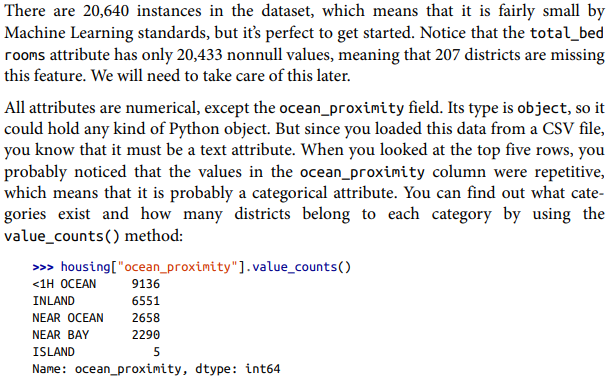
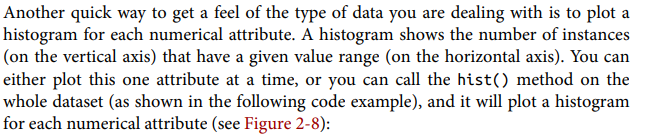
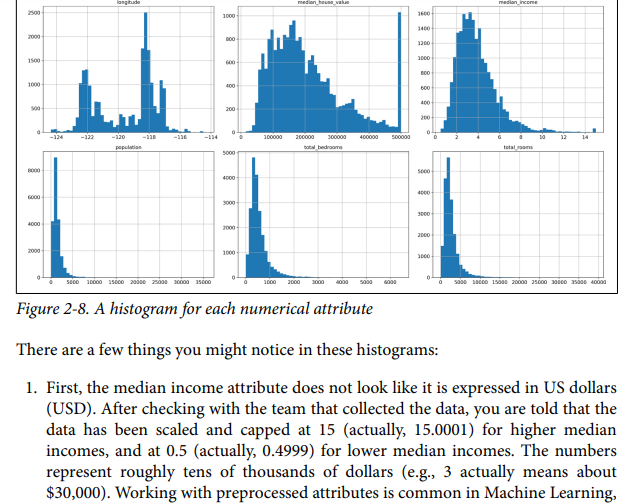
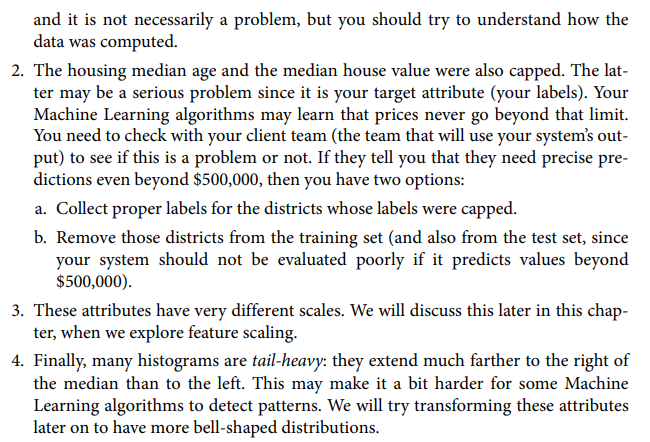
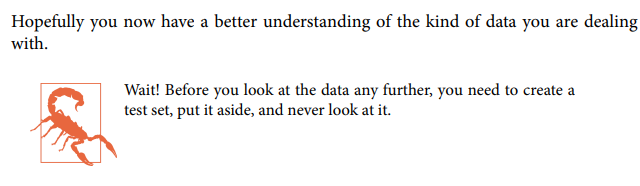
Housing code chapter 2 hands on machine

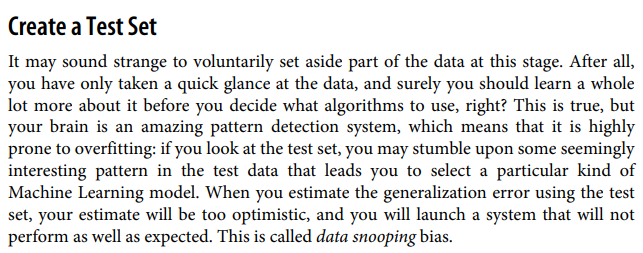




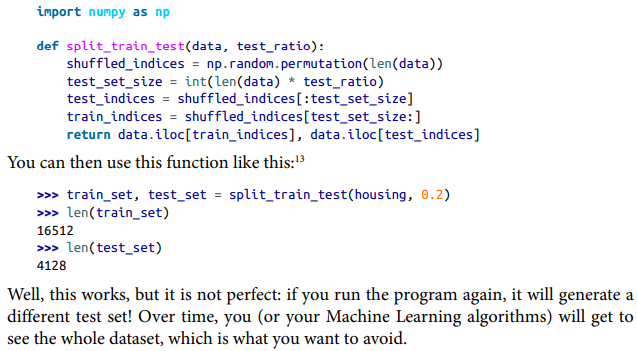


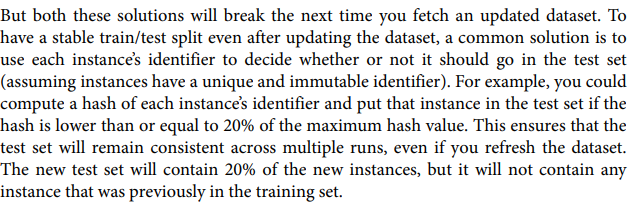


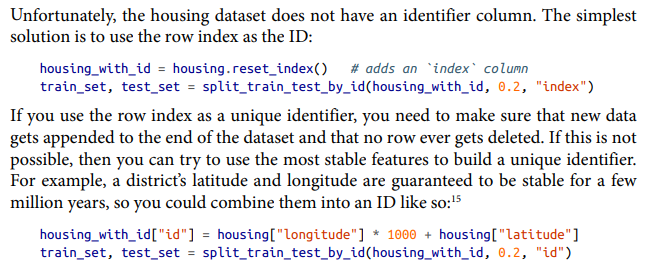


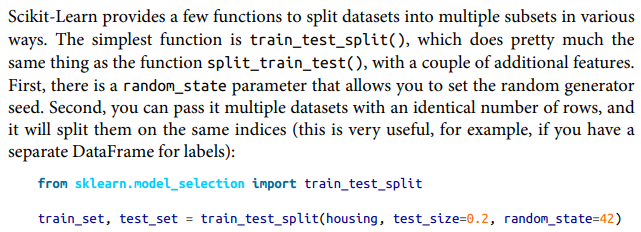


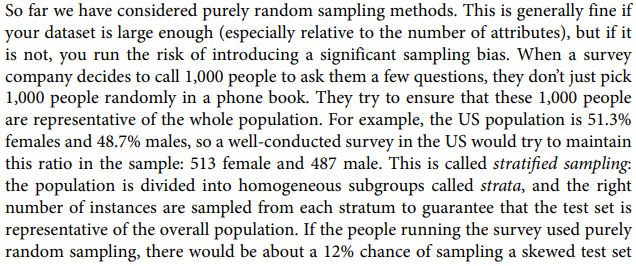


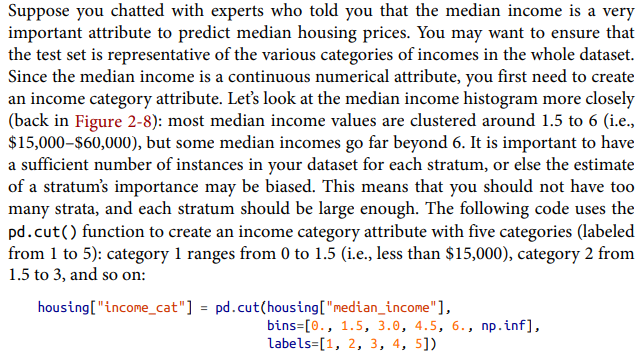












En el libro *"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow"*, se crea la variable income\_cat para abordar un problema específico relacionado con la distribución desigual de las categorías en la variable income del dataset de California Housing. Este proceso es una forma de realizar una **estratificación** en la división del conjunto de datos, lo cual es crucial para asegurar que las proporciones de las categorías de una variable importante (en este caso, income) estén bien representadas en ambos conjuntos: entrenamiento y prueba.

Aquí te explico por qué se crea income\_cat y cómo se utiliza:

1. **Problema de Distribución Desigual**

El dataset de California Housing tiene la variable income, que puede tener una distribución desigual, con muchas instancias en ciertos rangos de ingresos y menos en otros. Cuando dividimos los datos en conjuntos de entrenamiento y prueba, existe el riesgo de que algunas categorías de income se representen de manera desproporcionada en uno de los conjuntos, lo que puede llevar a una evaluación sesgada del modelo.

**2. Creación de income\_cat**

Para resolver este problema, se crea una nueva variable categórica llamada income\_cat que divide el rango de valores de income en categorías discretas (bins). Esto permite realizar una estratificación en la división del dataset, asegurando que cada conjunto tenga una representación proporcional de cada categoría de income.

Después de la división, la variable income\_cat ya no es necesaria para el entrenamiento del modelo, por lo que se puede eliminar del conjunto de datos.

