CRISP-DM Data Analysis Guide

Housing

1	13300000	7420	4	2
2	12250000	8960	4	4
3	12250000	9960	3	2
4	12215000	7500	4	2
5	11410000	7420	4	1
6	10850000	7500	3	3
	price	area	bedrooms	bathrooms

Using the dataset I provided can you do analyze it using CRISP-DM methodology. Make it presentable for a group of high school students. You want everyone to understand and follow along with the process. Do not use code but explain everything in reasonable steps.

Go step by step with and EDA analysis to educate the students. Begin with data understanding using EDA and the importance of visualization, pattern detection, and summary statistics. Show how you would perform data cleaning, pre processing, handle outliers and missing values. Then data transformation.

Continue to explain the feature selection by reducing dimensionality. You can do data clustering to show if there are any hidden patterns and potentially find outliers. Next start to build predictive models such as regression models to help students see what they do. You can integrate more advanced ML models so students can learn to compare and evaluate models. Use metrics to compare and rank models for accuracy. Also explain for each step how to account for limited compute resources by chunking.

Ensure that you follow the CRISP-DM procedure and clearly explain things steps by step. Summarize each step and outline the next step so students can follow along. Make it detailed yet reasonable for students to understand each concept.

Step-by-Step Data Analysis Using CRISP-DM for High School Students

Let's walk through the CRISP-DM methodology in a way that everyone can easily understand. CRISP-DM stands for **Cross-Industry Standard Process for Data Mining**. It's a step-by-step process that helps us explore, clean, and use data to make predictions. We'll break it down using a dataset related to housing, which could contain information like house prices, square footage, location, etc.

1. Business Understanding

Before we jump into the data, we need to understand **why** we are analyzing this dataset. Let's say we are interested in **predicting house prices**. For this, we need to find patterns in our dataset that influence prices, such as location, number of rooms, or size.

---

2. Data Understanding: Exploratory Data Analysis (EDA)

Now that we know our goal, let's look at the data to understand it better.

- **Step 1: Load the Data**
  We have a dataset (Housing.csv), and we begin by loading it to see what's inside. In the file, we expect columns like "Price," "Size (sq. ft)," "Location," and other relevant features.

- **Step 2: Visualizing the Data**
  Visualizing data is like looking at pictures of it. We can create:
  - **Histograms** to see the distribution of house prices (Are prices mostly low, or do we have expensive houses?).
  - **Scatter Plots** to see the relationship between two variables like house size vs. price (Do bigger houses usually cost more?).

- **Step 3: Summary Statistics**
  We would calculate the **mean** (average), **median** (middle value), and **standard deviation** (how spread out the data is) for the features in our dataset. These statistics help us get a quick overview:
  - For example, the **average price of a house** or the **average number of bedrooms**.
  - Summary statistics will help us spot if there are any strange values (like a house priced at $1, which might be an error).

- **Why is this important?**
  EDA gives us a sneak peek into the data, helping us understand **patterns**, **outliers**, and possible errors before diving deeper.

---

3. Data Preparation (Cleaning and Preprocessing)

The data we get is often **messy**, meaning it may contain missing values, errors, or outliers (extreme values). We need to clean it up!

- **Step 1: Handling Missing Values**
  - Sometimes data is missing (maybe a house doesn't list its size). We can:
    - **Fill in missing values** with averages.
    - **Remove rows** with missing information if the dataset is big enough to afford that.

- **Step 2: Handling Outliers**
  - Outliers are data points that are very different from the rest. For example, if most houses cost around $300,000 but one is listed at $10 million, that’s an outlier.

    - We can **remove outliers** or **analyze them separately**.

- **Step 3: Preprocessing (Normalizing/Scaling)**
  Some features, like house prices, might have a huge range of values (from $100,000 to $1 million), while others, like the number of rooms, might have smaller ranges. We can **scale the data** to ensure each feature is on a similar scale, making analysis easier.

---

4. Data Transformation

Once the data is clean, we may need to **transform** it for easier analysis.

- **Feature Engineering**
  This is when we create new features from existing ones. For instance, if we have the **age** of a house and the **year built**, we can create a new feature called **"Age of House"** (current year - year built).

- **Dimensionality Reduction**
  If we have too many features (like 100 different columns), it can get overwhelming for analysis. We can use techniques like **PCA (Principal Component Analysis)** to reduce the number of features while keeping the most important information.

---

5. Modeling

Now that we've cleaned and prepared our data, we can build models to **predict house prices**.

- **Simple Models (Linear Regression)**
  The easiest model to understand is **Linear Regression**. It tries to fit a line through the data, where one feature (like house size) predicts the price. We can see if the line fits well and how close our predictions are to the actual values.

- **Advanced Models (Decision Trees, Random Forests)**
  We can also use more complex models like **Decision Trees** and **Random Forests**, which break down decisions (like predicting house prices) into simpler steps, based on different features (like location, size, etc.).

---

6. Evaluation

After building the models, we need to check how good they are.

- **Metrics (Accuracy, RMSE)**
  To compare models, we use metrics like **Accuracy** or **Root Mean Squared Error (RMSE)**. RMSE tells us, on average, how far off our predictions are from the actual values.

- **Compare and Rank Models**
  We can compare different models and rank them based on their performance. The goal is to pick the one that predicts house prices with the least amount of error.

---

7. Deployment

Once we've built and evaluated our model, we can deploy it. This means putting it into action so that it can predict prices on new data.

Conclusion: Overview of Steps and Next Steps

1. **Business Understanding** – We focused on predicting house prices.
2. **Data Understanding** – We explored the dataset using EDA, visualization, and summary statistics.
3. **Data Preparation** – We cleaned and preprocessed the data by handling missing values, outliers, and scaling.
4. **Data Transformation** – We created new features and reduced dimensionality to simplify the data.
5. **Modeling** – We built models like Linear Regression and Decision Trees to predict house prices.
6. **Evaluation** – We compared and ranked models using metrics like RMSE.
7. **Deployment** – After evaluating, we can deploy our best model for real-world use.

Dealing with Limited Compute Resources

Since we may not have powerful computers to handle huge datasets or complex models, we can:

- **Chunk the data**: Instead of processing everything at once, we split the data into smaller parts (chunks) and analyze each part separately.
- **Simplify the models**: Start with simpler models like Linear Regression before moving to complex ones like Random Forests, which need more computing power.

This is a quick walkthrough of the CRISP-DM process. By following these steps, we can transform raw data into useful insights or predictions while keeping things manageable for our computers!