CCADD
Center for Convergence Approaches in Drug Development
서울대학교
융합과학기술대학원

[Book Reading Session #3]

Exploratory Data Analysis

2025. 01. 10. 구나영

Exploratory Data Analysis (EDA)

- the process of observing and understanding the collected data from various perspectives
- Purposes:
  - To understand the phenomena represented by the data and identify potential issues within it
  - To uncover patterns that might not have been considered during the problem definition stage
- Based on these findings, we can revise existing hypotheses or formulate new ones.

Exploratory Data Analysis (EDA)

<Observation> + Why?

- EDA highlights **creativity** to generate a **large quantity of questions in good quality.**
- You have to keep asking "**why**" throughout the whole process.

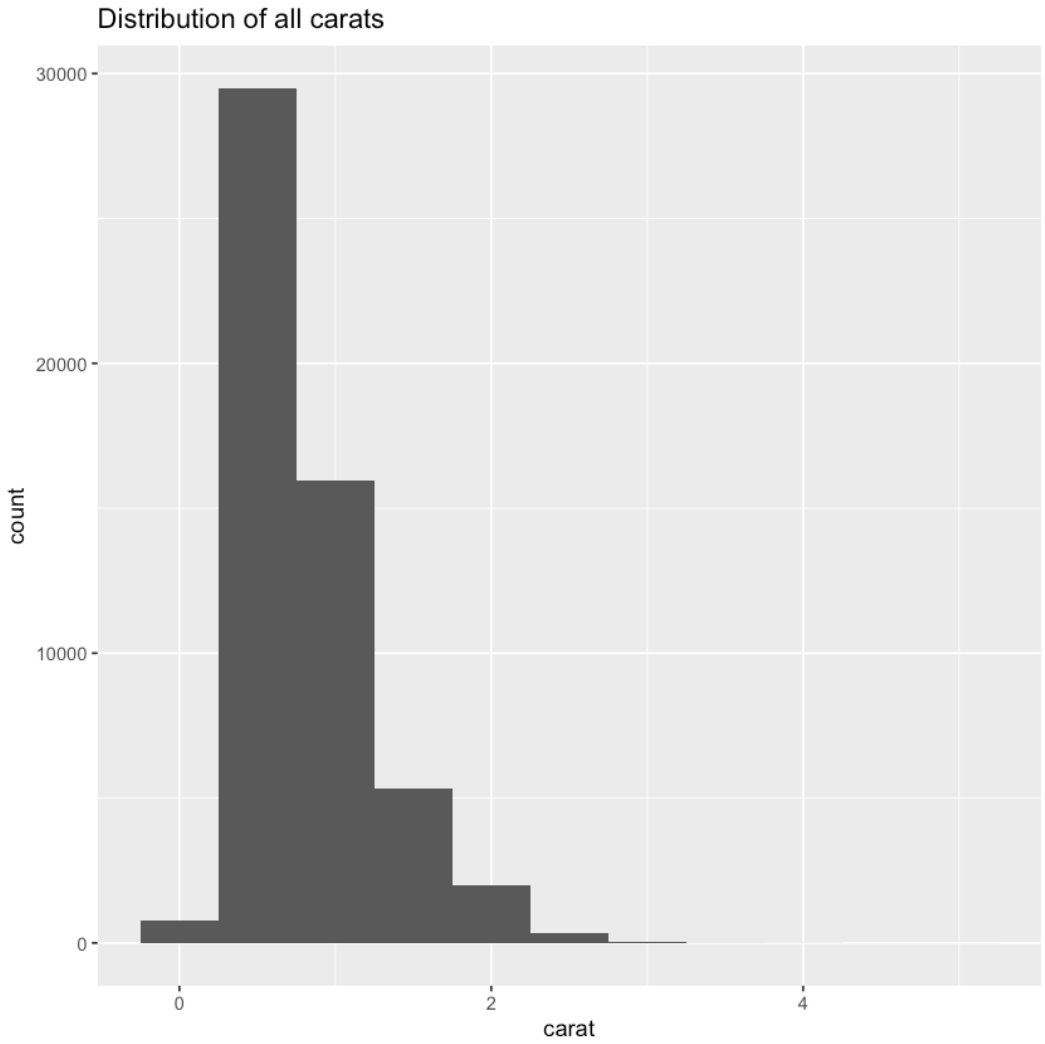
Exploratory Data Analysis (EDA)

- Visualization
- Transformation
- Modeling

1. Variation

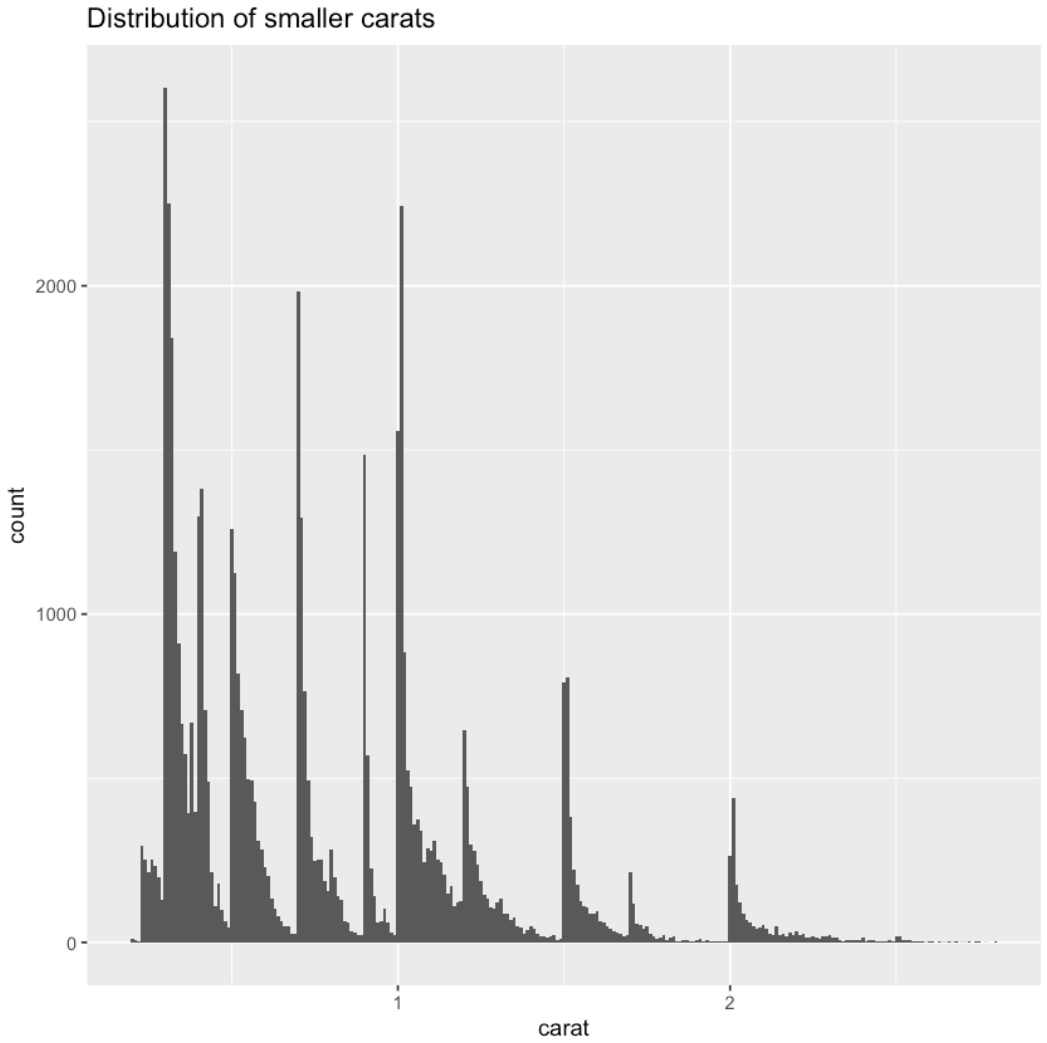
- the behavior *within* values of a variable.
- it helps identify patterns, anomalies, and the overall distribution of data.
- How:
  - Distribution Analysis
  - Identifying Anomalies
  - Data Transformation

1. Variation



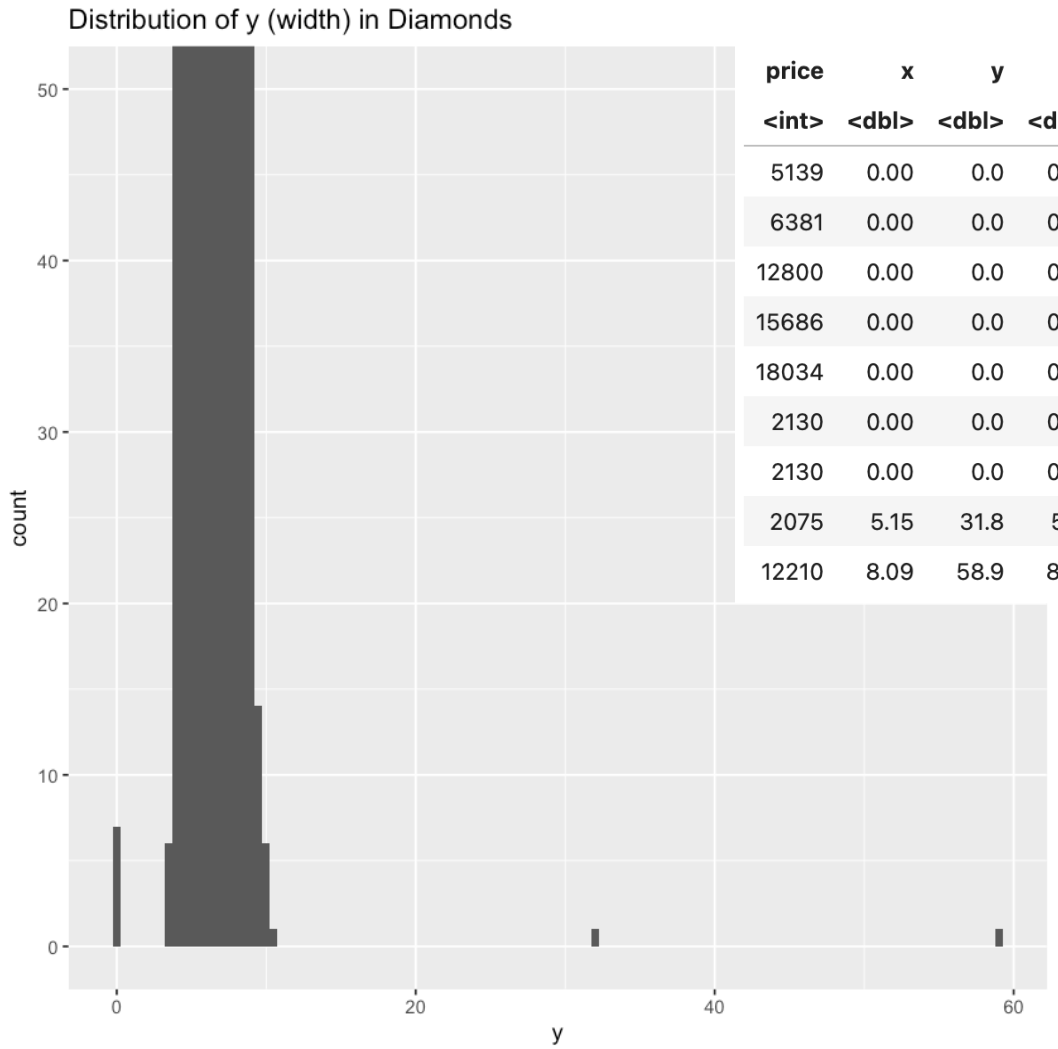
- Follow-up questions:
  - Which values are the most common? + WHY?
  - Which values are rare? + Does that match your expectations?
  - Are there any unusual patterns? + What might explain them?

1. Variation – Typical Values



- #subgroups
- Follow-up questions:
  - Why are there more diamonds at whole carats (i.e., 1.0, 2.0) and common fractions of carats (i.e., 0.25, 0.5, 1.5)
  - Why are there more diamonds slightly to the right of each peak?
  - How can you explain or describe the clusters?

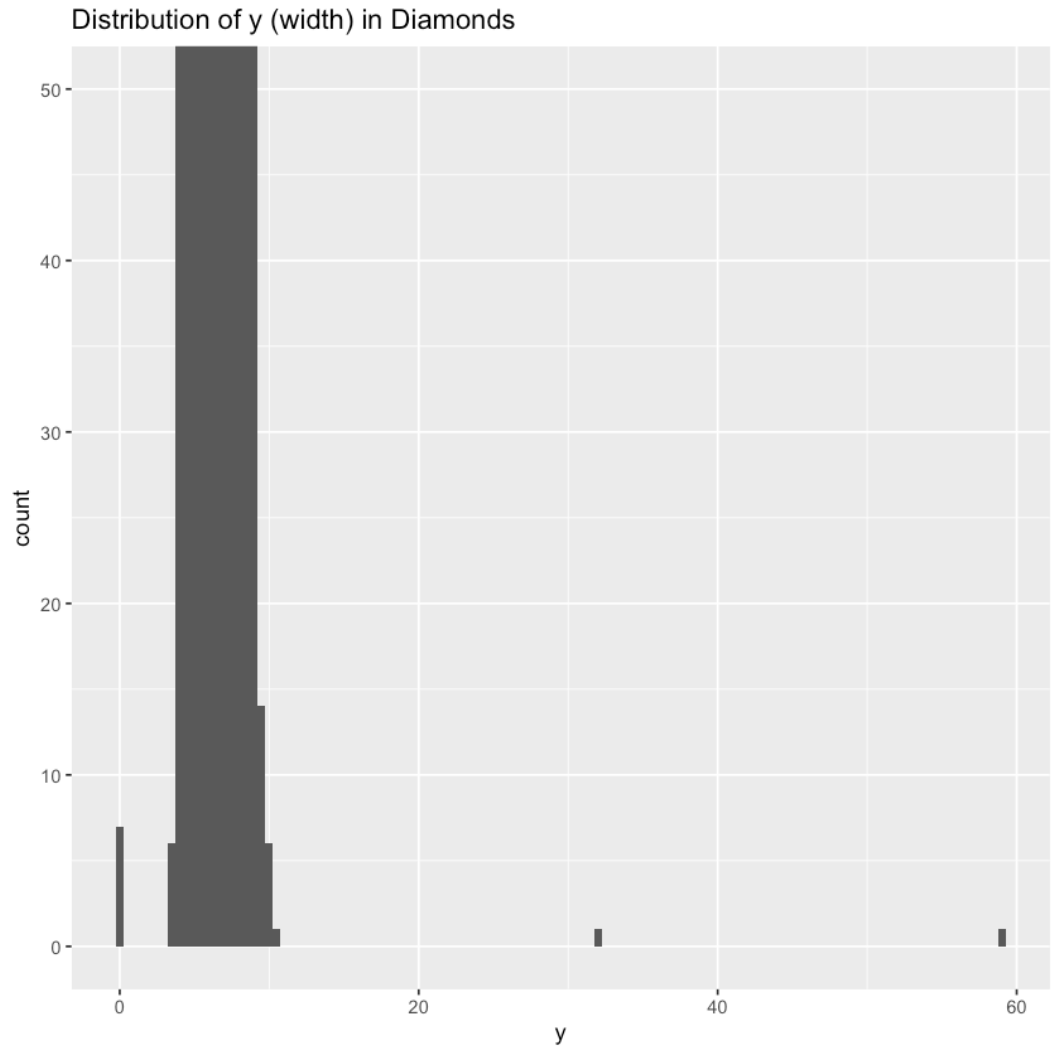
1. Variation – Unusual Values



price	x	y	z
<int>	<dbl>	<dbl>	<dbl>
5139	0.00	0.0	0.00
6381	0.00	0.0	0.00
12800	0.00	0.0	0.00
15686	0.00	0.0	0.00
18034	0.00	0.0	0.00
2130	0.00	0.0	0.00
2130	0.00	0.0	0.00
2075	5.15	31.8	5.12
12210	8.09	58.9	8.06

- Outliers: unusual observations
  - They can be entry errors, extremes, or new discoveries
- Follow-up questions:
  - We know that diamonds cannot have a width of 0mm, so these values must be incorrect. → missing values

1. Variation – Unusual Values



- Dealing with missing values:
  - **Drop** the entire row with strange values
  - **Replace** the unusual values with missing values (Recommended

carat	cut	color	clarity	depth	table	price	x	y	z
<dbl>	<ord>	<ord>	<ord>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>
1.00	Very Good	H	VS2	63.3	53	5139	0.00	NA	0.00
1.14	Fair	G	VS1	57.5	67	6381	0.00	NA	0.00
2.00	Premium	H	SI2	58.9	57	12210	8.09	NA	8.06
1.56	Ideal	G	VS2	62.2	54	12800	0.00	NA	0.00
1.20	Premium	D	VVS1	62.1	59	15686	0.00	NA	0.00
2.25	Premium	H	SI2	62.8	59	18034	0.00	NA	0.00
0.51	Ideal	E	VS1	61.8	55	2075	5.15	NA	5.12
0.71	Good	F	SI2	64.1	60	2130	0.00	NA	0.00
0.71	Good	F	SI2	64.1	60	2130	0.00	NA	0.00

2. Covariation

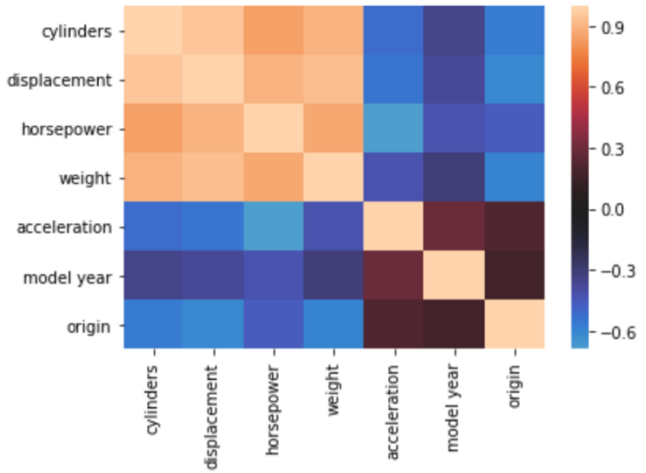
- The behavior *between* variables.
- Understanding covariation is crucial for identifying relationships between variables, which can inform modeling and analysis decisions.

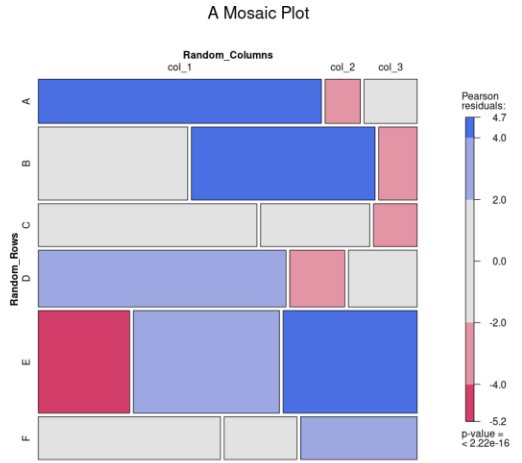
2. Covariation between a Categorical and a Numerical Variable

- Assess how the distribution of a numerical variable varies across the levels of a categorical variable.
- How to visualize:
  - **Boxplots:** Display the distribution of the numerical variable for each category, highlighting the median, quartiles, and potential outliers.
  - **Violin Plots:** Combine aspects of boxplots and density plots to show the distribution's shape for each category.
  - **Frequency Polygons:** Useful when the categorical variable has numerous levels; they display the density of the numerical variable across categories.

2. Covariation between Two Categorical Variables

- Explore the relationship between two categorical variables.
- How to visualize:
  - **Count Tables**: Summarize the frequency of each combination of categories.
  - **Heatmaps**: Graphically represents count tables, where color intensity indicates the frequency of category combinations.
  - **Mosaic Plots:** Display the proportion of each combination of categories, with the area of each tile representing the frequency.





2. Covariation between Two Numerical Variables

- Investigate how two numerical variables are related.
- How to visualize:
  - **Scatter Plots**: Plot individual data points to observe relationships, trends, or clusters.
  - **Hexbin Plots**: For large datasets, hexbin plots bin data points into hexagonal cells, with color intensity representing the number of observations, reducing overplotting.
  - **Contour Plots**: Show density estimates with contour lines, useful for identifying areas of high concentration in the data.