Session 3 - EDA

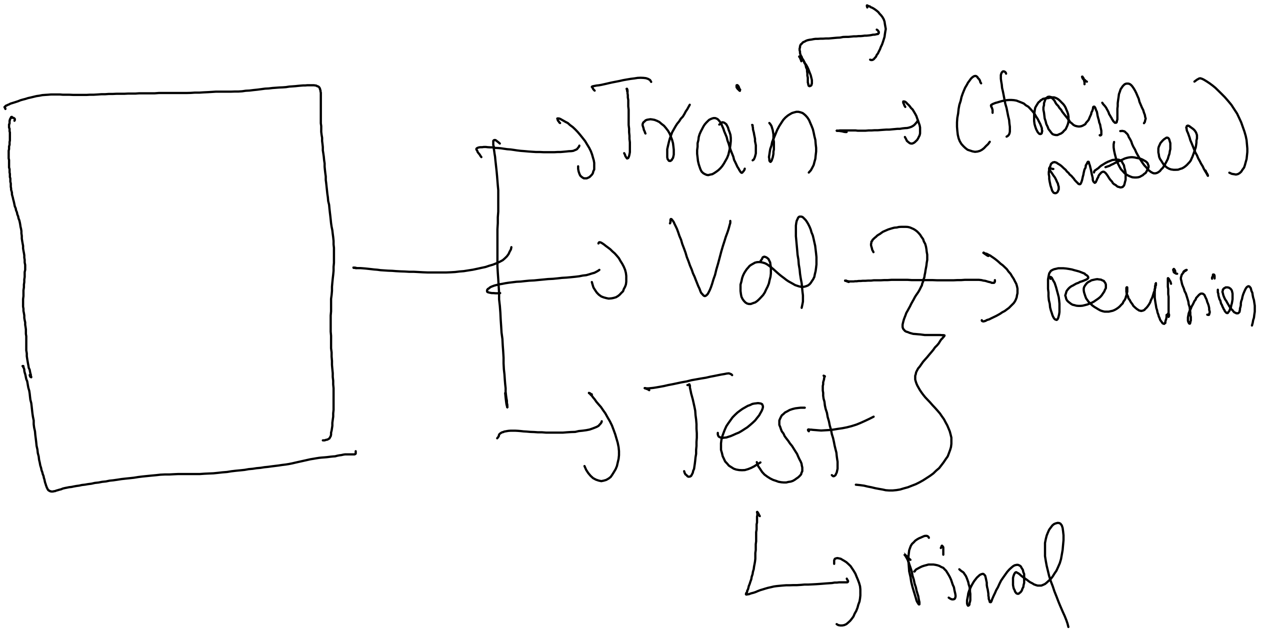
Monday, April 22, 2024 10:44 PM

EDA - Exploratory Data Analysis:

- It's the practice of exploring your dataset, by utilizing visualization tools and statistical measures to understand and extract underlying patterns and information within the data.
- This gives a clearer picture of the data, helps us make data-informed decisions and solve crucial business problems with much fewer assumptions and more facts.
- This step is the backbone of any Data Science project and takes up a major chunk of the project timeline.

Additional Note:

- The term EDA was coined by the late mathematician, John Tukey
- First introduced in his book "*Exploratory Data Analysis*" (1977)
- Mr. Tukey also introduced the "Box Plot"



Val → Tune

↳ Best parameters

Retrain

↳ Best param

↳ Train + Val set

Components

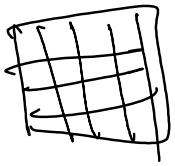
Monday, April 22, 2024 11:03 PM

What is done in EDA stage?

Look at and analyze numbers and plots

- **Descriptive Statistics**
  - Central Tendency ✓
  - Dispersion / Spread ✓
  - Distribution of Variables ✓
  - Symmetry and Shape of variables

Summarizing Data



→ no.s

- **Inferential Statistics**:
  - Strength of Association (correlation) ✓
  - Hypothesis Testing ✓

Sample

↓

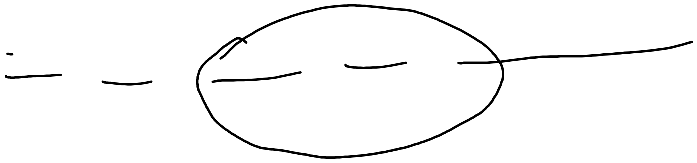
Population

- **Plots / Graphs**: ✓
  - Univariate
  - Bivariate
  - Multivariate

Descriptive Statistics

Monday, April 22, 2024 11:29 PM

1. Measure of Location:



- Mean
- Mode
- Median
- Percentiles
- Quartiles

1st moment of distribution

2. Measure of Spread:

→ reference

- Variance
- Standard Deviation
- Coefficient of Variance
- Mean Absolute Error (MAE)
- Inter Quartile Range (IQR)
- Median Absolute Deviation from Median (MAD)

→ outliers

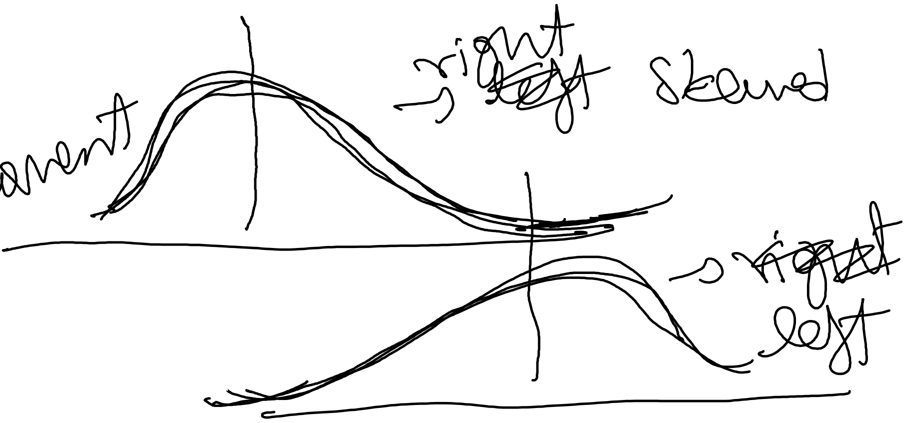
2nd moment

→ robust

3. Measure of Symmetry:

- Skewness

3rd moment



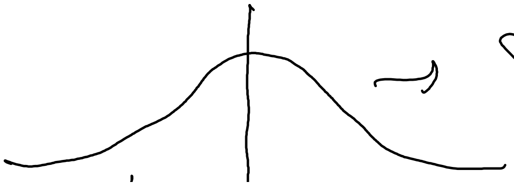
→ right ~~left~~ skewed

→ ~~right~~ left

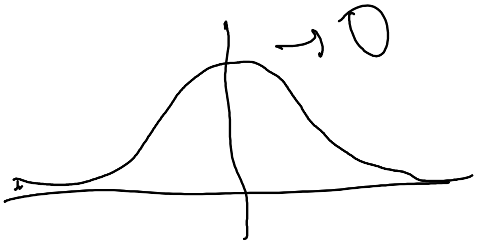
4. Measure of Shape:

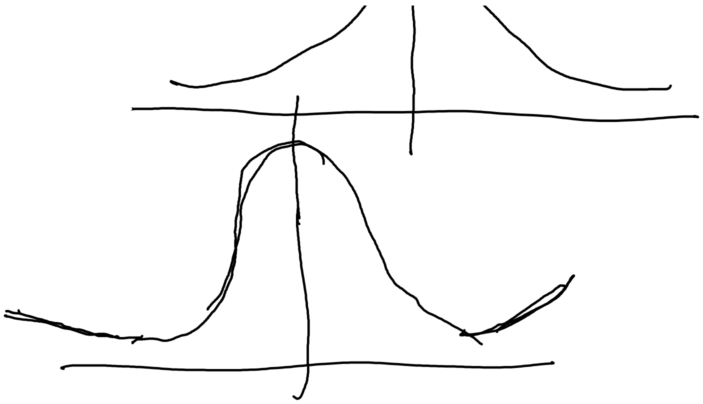
- Kurtosis

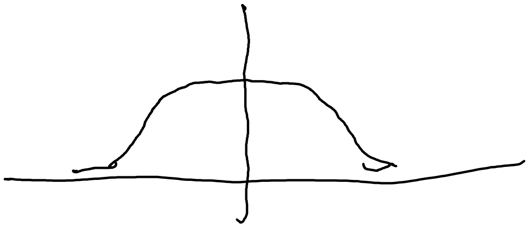
4th moment



→ symmetric







Inferential Statistics

Monday, April 22, 2024 11:30 PM

X1 X2

1. **Strength of Association:**
   - ○ Pearson's Correlation ✓
   - ○ Spearman's Rank Correlation ✓
   - ○ Cramer's V ✓

2 categorical variables

2. **Hypothesis Testing:**
   - ○ ***Test for Normality:***
     - ▪ Shapiro-Wilk Test
     - ▪ Anderson-Darling Test

   } Normality → numeric

   - ○ ***Test for Association:***
     - ▪ Numeric Variables: ✓
       - □ Pearson's Test
       - □ Spearman's Test
     - ▪ Categorical Variables:
       - □ Chi-Square Test ✓
     - ▪ Numeric - Categorical Variable:
       - □ One-way ANOVA Test ✓
       - □ Kruskal-Wallis Test ✓

   ANOVA

   cat2 m1

   cat2 m2

   cat3 m3

   - ○ ***Steps Involved:***

cat3 m3

- ○ ***Steps Involved:***

  - State the hypotheses:

    - □ *Null Hypothesis* Ho num. var
    - □ *Alternate Hypothesis* HA

  - Determine significance level (alpha ~ 5%) α

  - Determine which test to perform Test

  - Collect necessary data (sample) Data

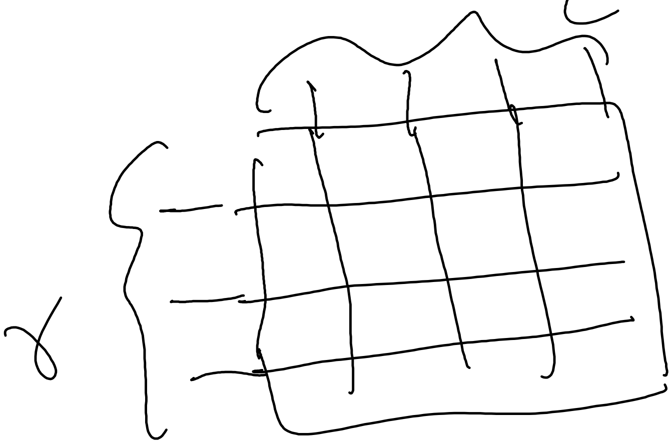
  - Obtain Critical values → (based on α)

  - Compute Test Statistic (and p-value) → (sample)

  - Compare:

    - □ *Significance level vs p-value; or*
    - □ *Critical values vs Test Statistic*

  - State conclusions ✓



Frequency Table

↓

Chi-Square Test

r.

↓

χ^2 Test-Statistic

Frequency Table → Chi-Square Test → χ^2 Chi-square statistic

↓

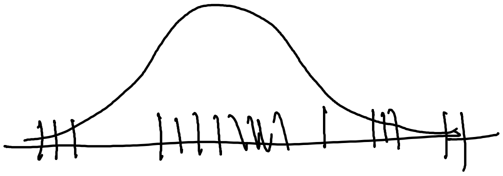
Cramer's V

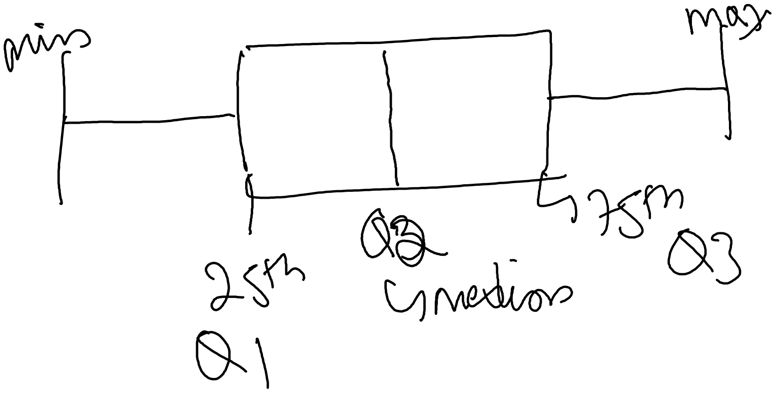
Plots

Monday, April 22, 2024 11:32 PM

1. **Univariate:**

- ***Numeric:***
  - Histrogram
  - KDE Plot → (smoothed Histogram)
  - Rug Plot
  - Box Plot
  - Violin Plot (box plot + kde plot)
  - Q-Q Plot





- ***Categorical:***
  - Count Plot
  - Pie Chart

- ***Time-related:***
  - Line Plot
  - Aggregated Line Plot

Q3+1.5×IQR

Q1 − 1.5×IQR

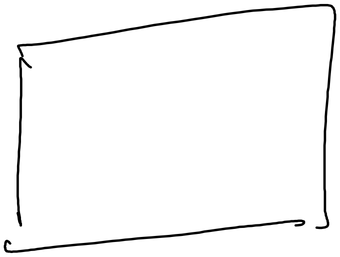
2. **Bivariate:**

- ***Numeric - Numeric:***
  - Scatter Plot
  - Hexagonal Bin Plot
  - Contour Density Plot

- ***Numeric - Categorical:***
  - Bar Plot

- ○ Box Plot
- ○ KDE Plot
- ○ Violin Plot

- ***Categorical - Categorical:***



- ○ Bar Plot
- ○ Stacked Bar Plot
- ○ Frequency Heatmap

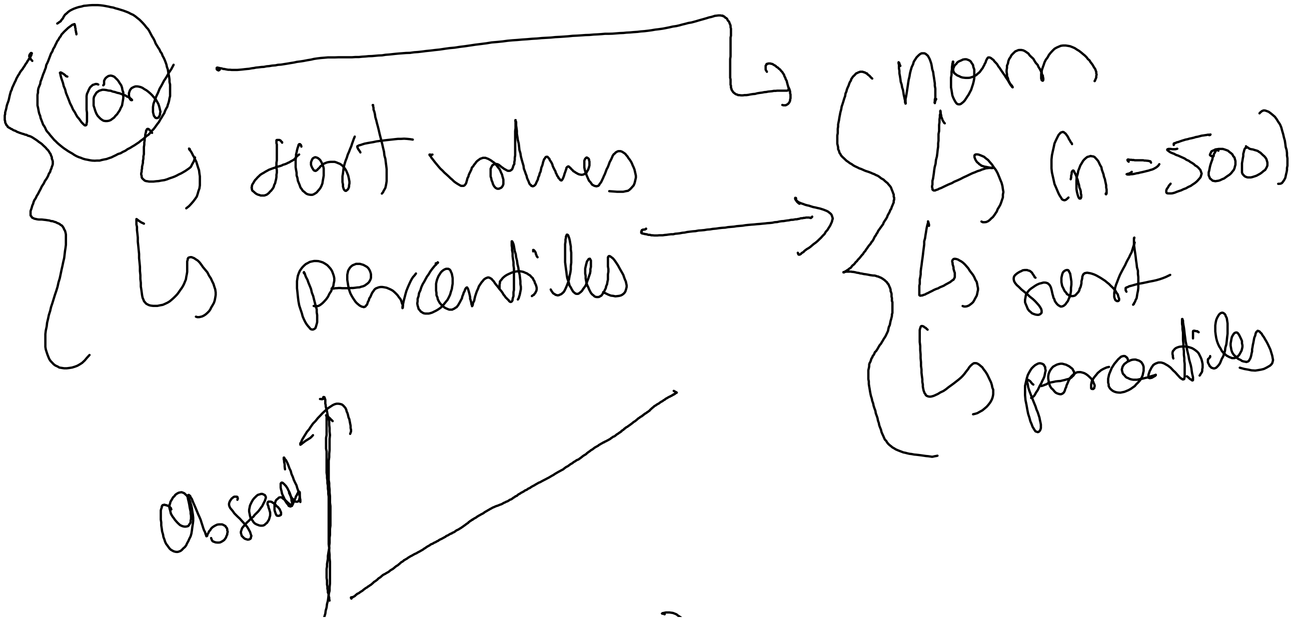
3. Multivariate:

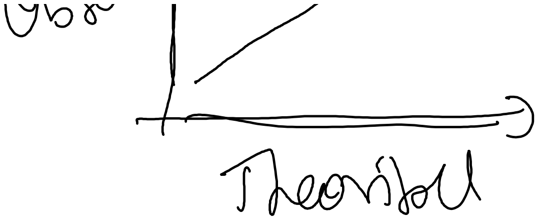
- ***Pair Plots***

- ***Correlation Heatmap:***

  - ○ Pearson
  - ○ Spearman's Rank
  - ○ Cramer's V

- ***Facet Grid (Seaborn)***




Obs
Theorital

Steps

Tuesday, April 23, 2024 12:30 AM

Sequence of Steps:

- Import libraries ✓
- Read the '*training*' data subset ✓
  - Check data types
  - Fix data types (if applicable) ✓
- Gather high-level summary of the data ✓
  - .info() method
  - .describe() method on numeric and categorical features
- High-level analysis of missing values: ✓
  - Bar plot
  - Count plot
  - Missingno
- High-level analysis of outliers: ✓
  - Isolation Forest
- Pair plots ✓
- Correlation Analysis (heatmaps) ✓
  - Numeric (Pearson's / Spearman's)
  - Categorical (Cramer's V)
- Detailed Analysis of each Feature: ✓
  - Summary

  - Univariate plots
  - Bivariate plots (w.r.t. the target variable)
  - Hypothesis Testing (normality, strength of association)
  - Multivariate plots
  - Inspect missing values and extreme values in-depth
    - Filter for necessary subsets
    - Inspect values of other features (plots, summary stats)
  - Note observations

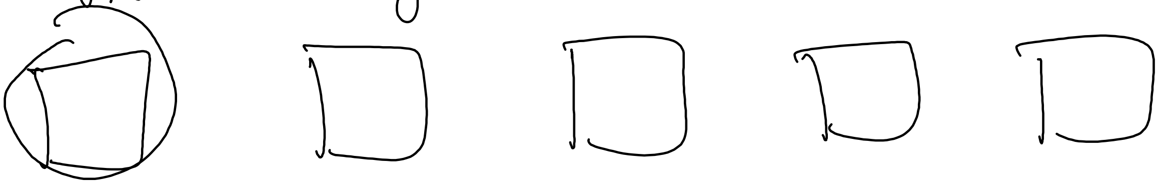
- Feature Engineering

  - Create new features
  - Repeat above steps for newly created features

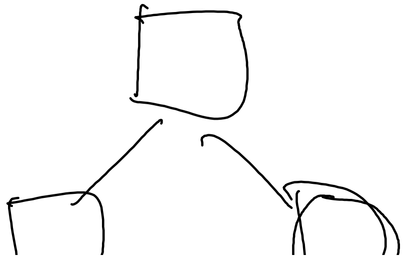
- Repeat above steps iteratively

- Note all observations

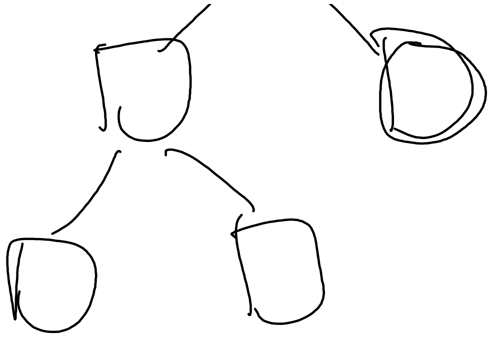
numeric features

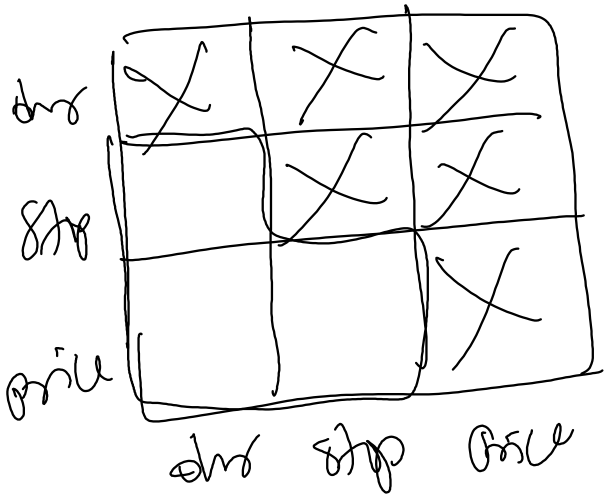


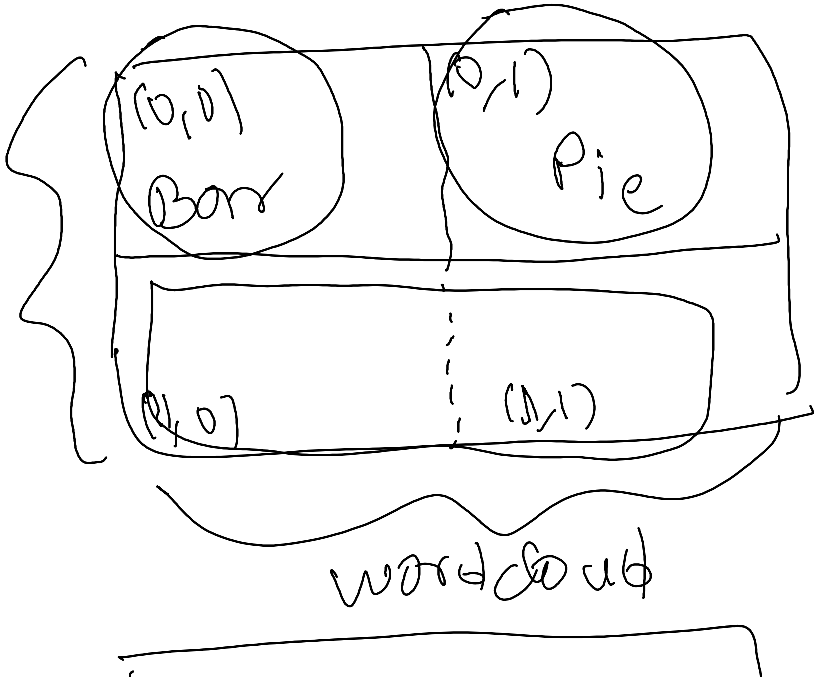
1. Randomly pick one feature
2. Randomly pick one value



-1, 1







color

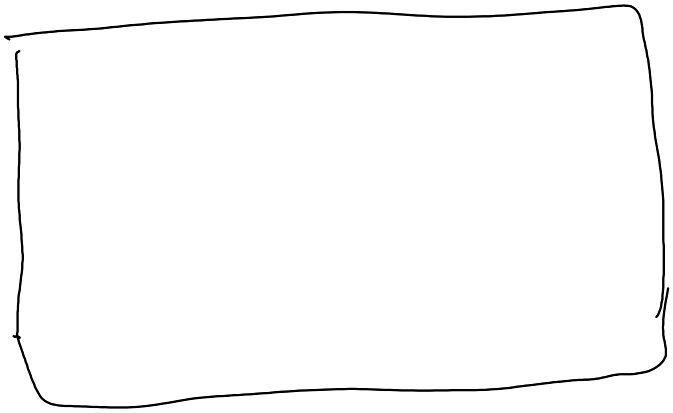
$\underbrace{(r, g, b)}$

(0 - 1)

(0-255, 0-255, 0-255)

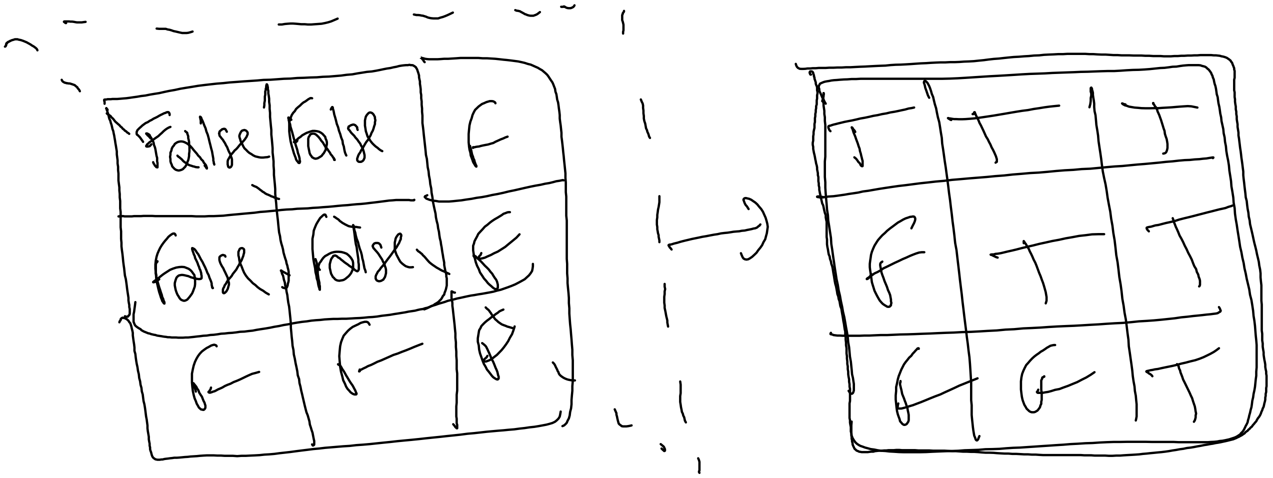
plt.figure()

axes → ax



plt.figure()

axes → ax

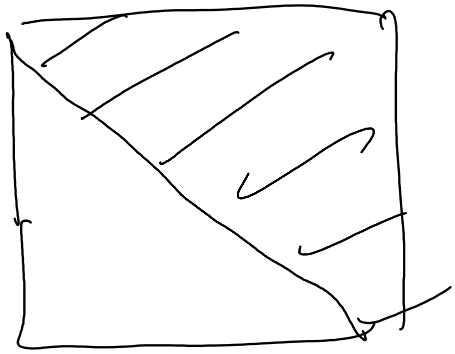


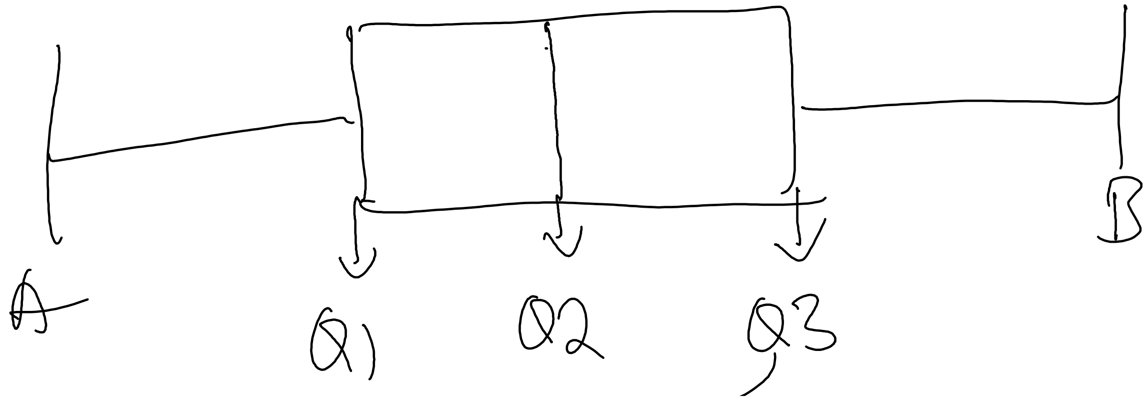
False	False	F
False	False	F
F	F	F

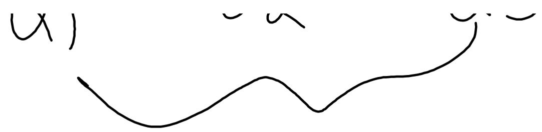
→

T	T	T
F	T	T
F	F	T

.corr()



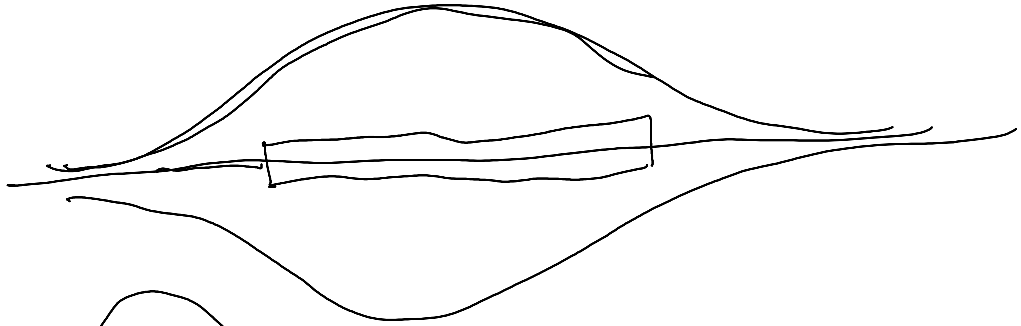


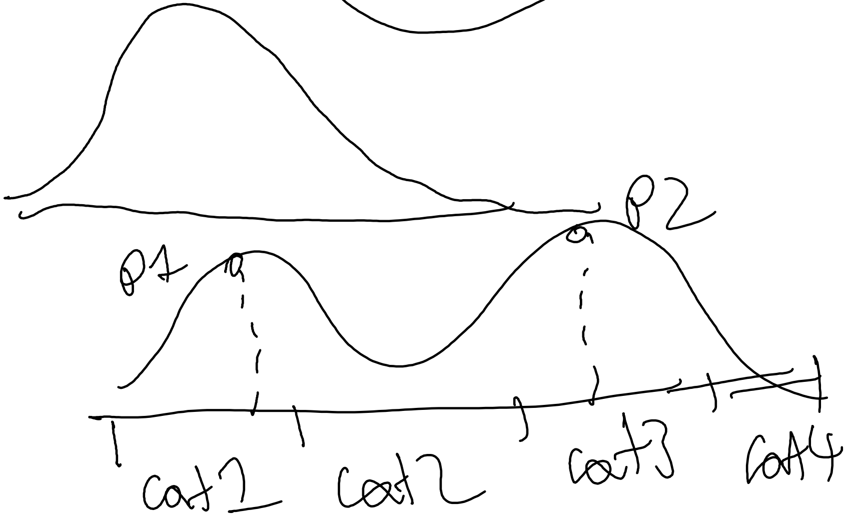


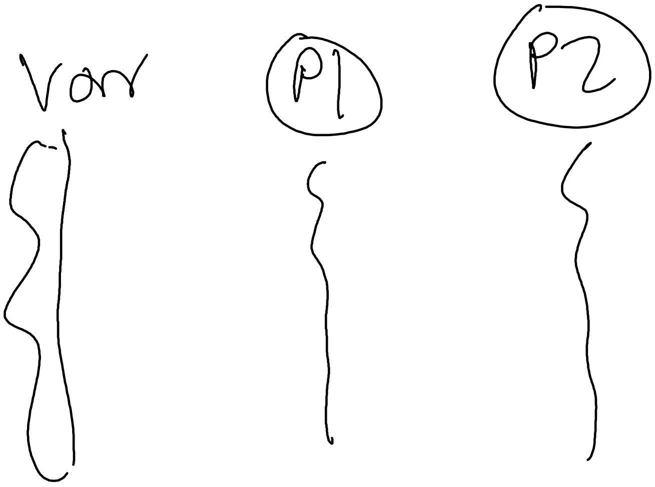
$$IQR = Q3 - Q1$$

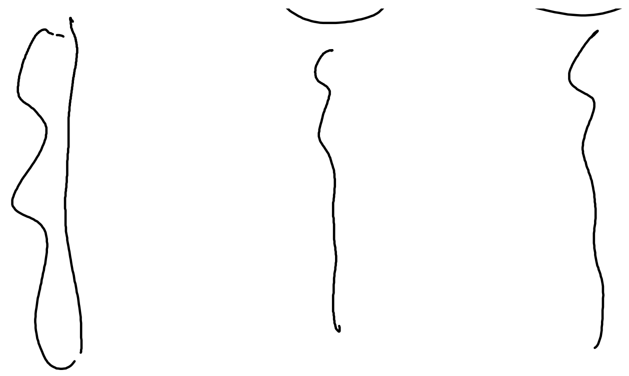
$$B = Q3 + (1.5 \times IQR)$$

$$A = Q1 - (1.5 \times IQR)$$









[cond1, cond2, cond3, ...]

[val1, val2, val3, ...]

default = "abc"

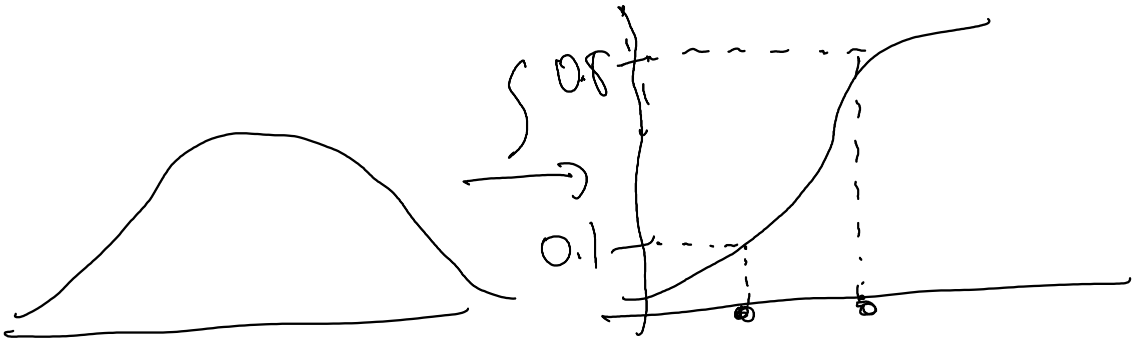
$$[\underbrace{1, 2,}_{5\%}\ \underbrace{10, 50, 80, 100,}_{\downarrow \text{ mean}}\ \underbrace{120}_{5\%})$$

1. Median
2. var — median
3. |var — median|

4. median

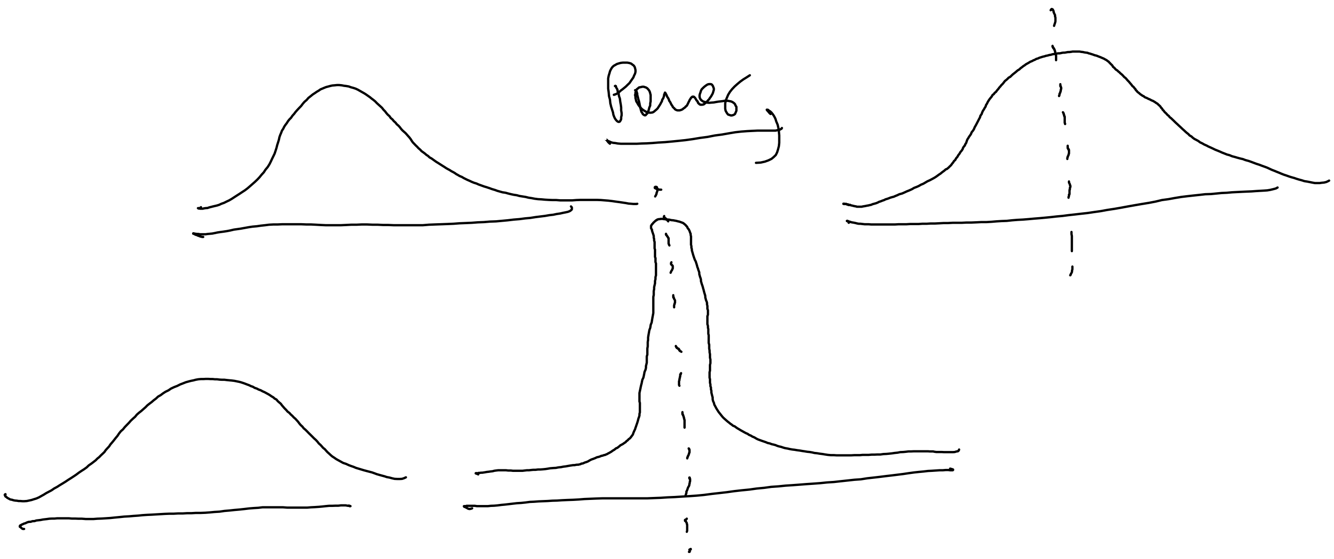
var
sd } impacted outliers

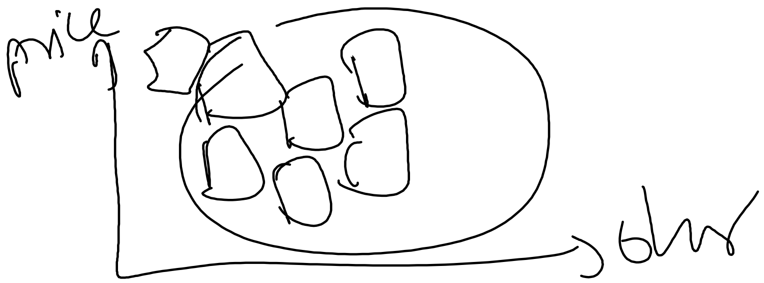
IQR
MAD } not imp. outliers



X

Power(X) = X^λ




price

Automated EDA

Tuesday, April 23, 2024 12:52 AM

- **Pandas Profiling (ydata-profiling)**
- **Sweetviz**
- **Autoviz**
- **D-Tale**