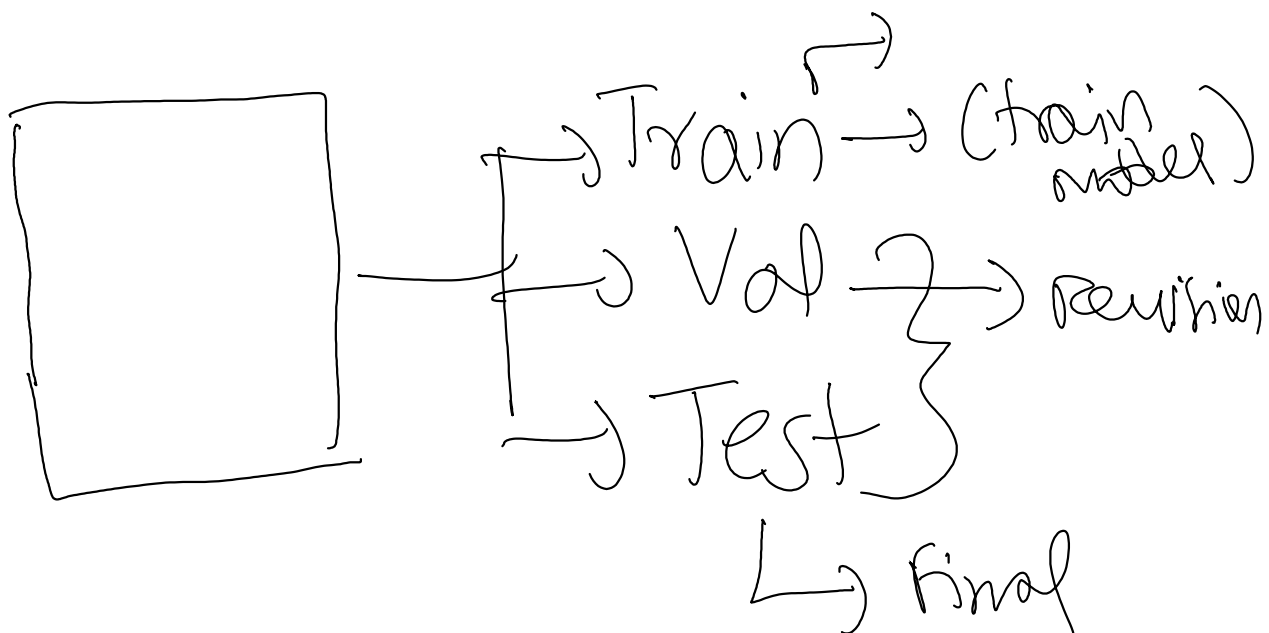


EDA - Exploratory Data Analysis:

- 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

What is done in EDA stage?

Look at and analyze numbers and plots

- **Descriptive Statistics**

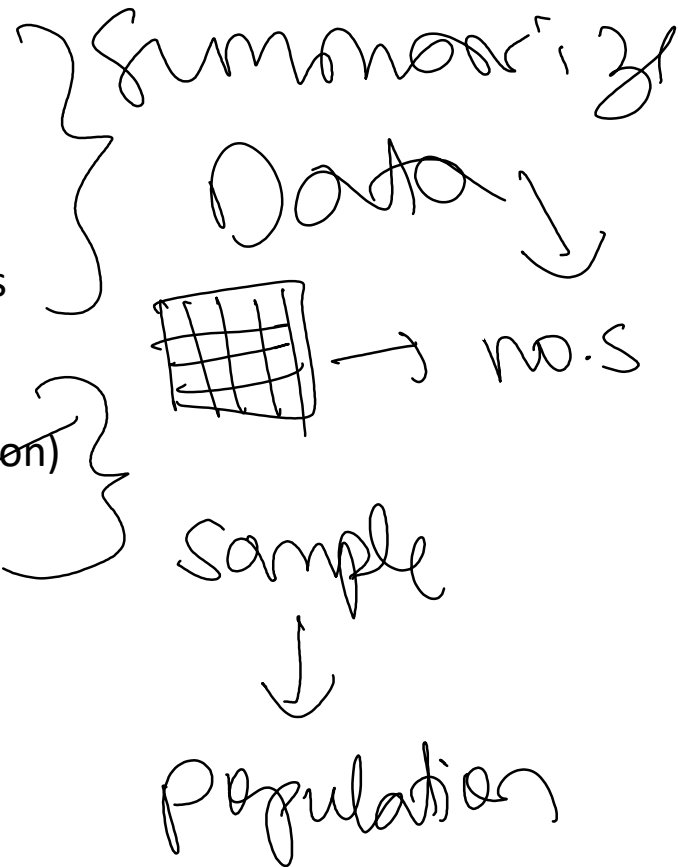
- Central Tendency ✓
- Dispersion / Spread ✓
- Distribution of Variables ✓
- Symmetry and Shape of variables ✓

- **Inferential Statistics:**

- Strength of Association (correlation) ✓
- Hypothesis Testing ✓

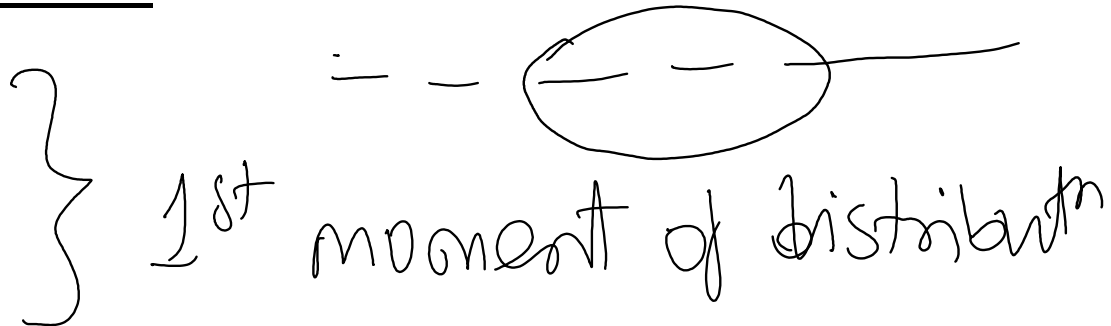
- **Plots / Graphs:** ✓

- Univariate
- Bivariate
- Multivariate



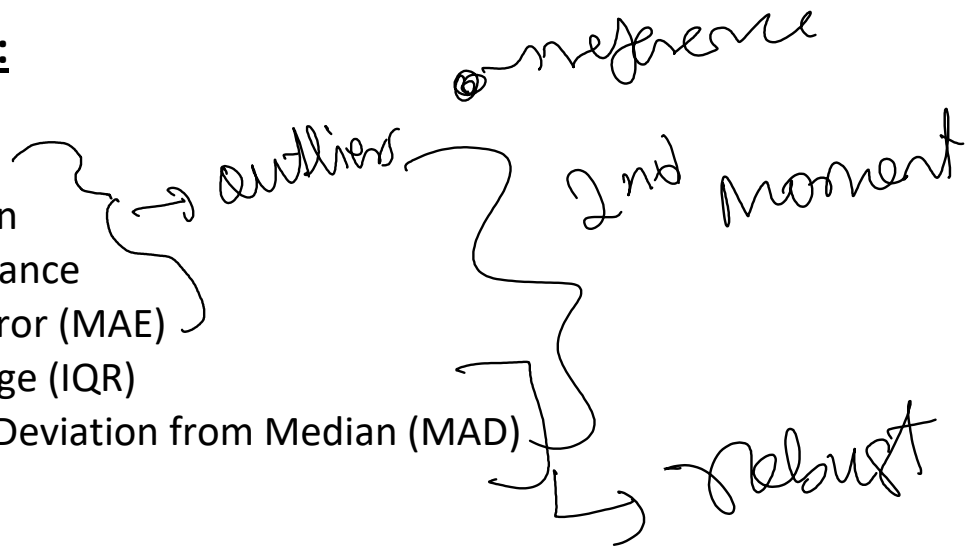
1. Measure of Location:

- Mean
- Mode
- Median
- Percentiles
- Quartiles



2. Measure of Spread:

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



3. Measure of Symmetry:

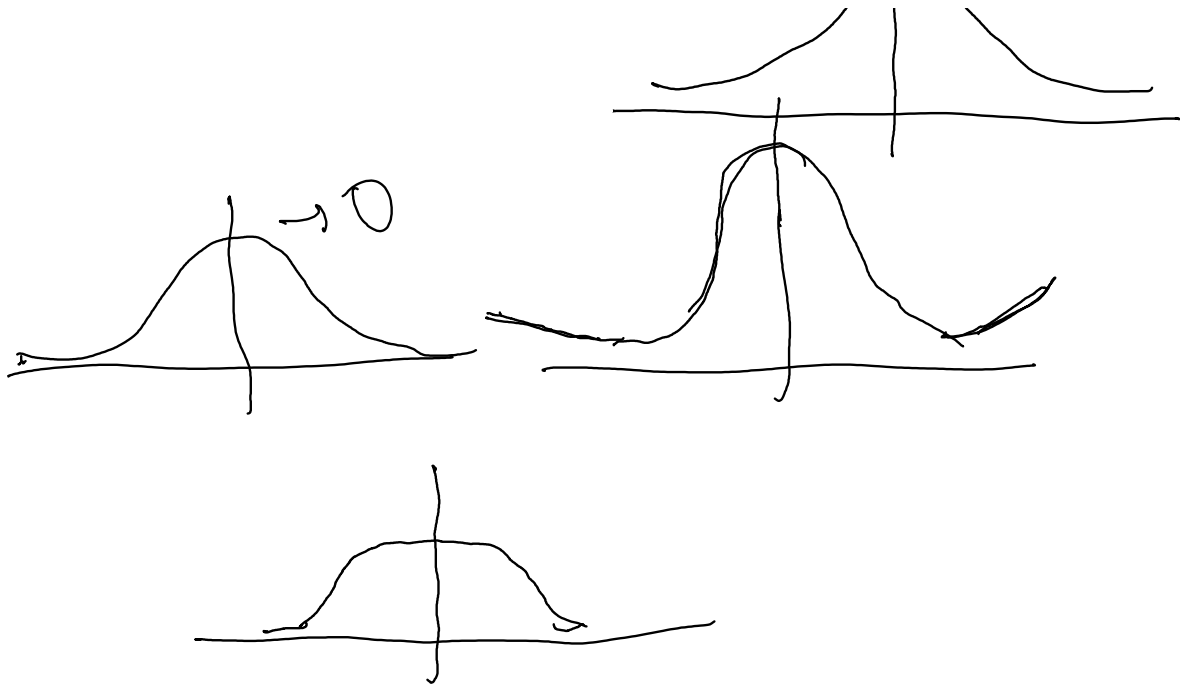
- Skewness



4. Measure of Shape:

- Kurtosis





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 ✓

○ Steps Involved:

ANOVA

cat 1	<u>m1</u>
cat 2	<u>m2</u>
cat 3	<u>m3</u>

○ Steps Involved:



with
cat 3

m3

▪ State the hypotheses:

□ Null Hypothesis

H_0

num. var

□ Alternate Hypothesis

H_A

▪ Determine significance level (alpha ~ 5%)

2

▪ Determine which test to perform

Test

▪ Collect necessary data (sample)

Data

▪ Obtain Critical values

→ (based on (2))

▪ Compute Test Statistic (and p-value)

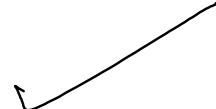
→ (sample)

▪ Compare:

□ Significance level vs p-value; or

□ Critical values vs Test Statistic

▪ State conclusions



2 {

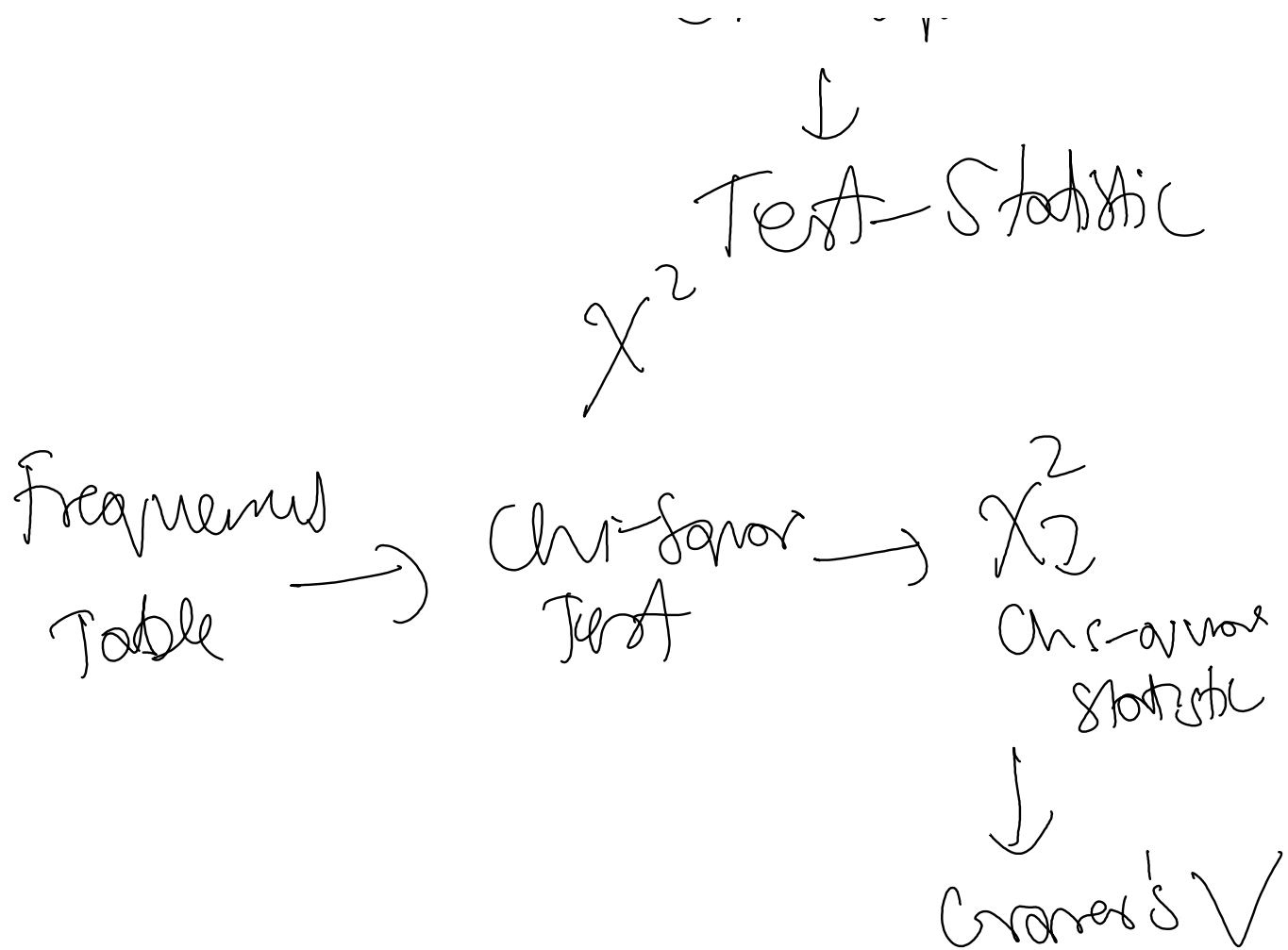
C

Contingency Table



Chi-Square Test

↑



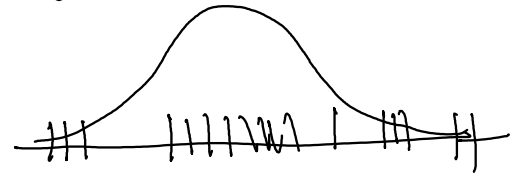
Plots

Monday, April 22, 2024 11:32 PM

1. Univariate:

• Numeric:

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

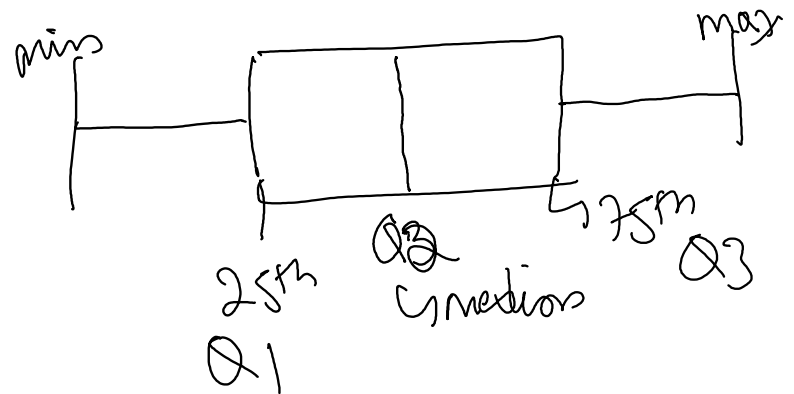


• Categorical:

- Count Plot
- Pie Chart

• Time-related:

- Line Plot
- Aggregated Line Plot



$$Q3 + 1.5 \times IQR$$

$$Q1 - 1.5 \times IQR$$

2. Bivariate:

• Numeric - Numeric:

- Scatter Plot
- Hexagonal Bin Plot
- Contour Density Plot

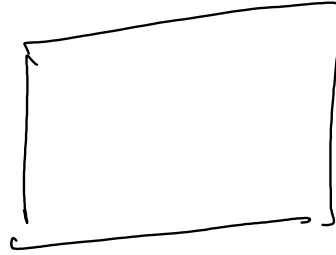
• Numeric - Categorical:

- Bar Plot

- Box Plot
- KDE Plot
- Violin Plot

- **Categorical - Categorical:**

- Bar Plot
- Stacked Bar Plot
- Frequency Heatmap



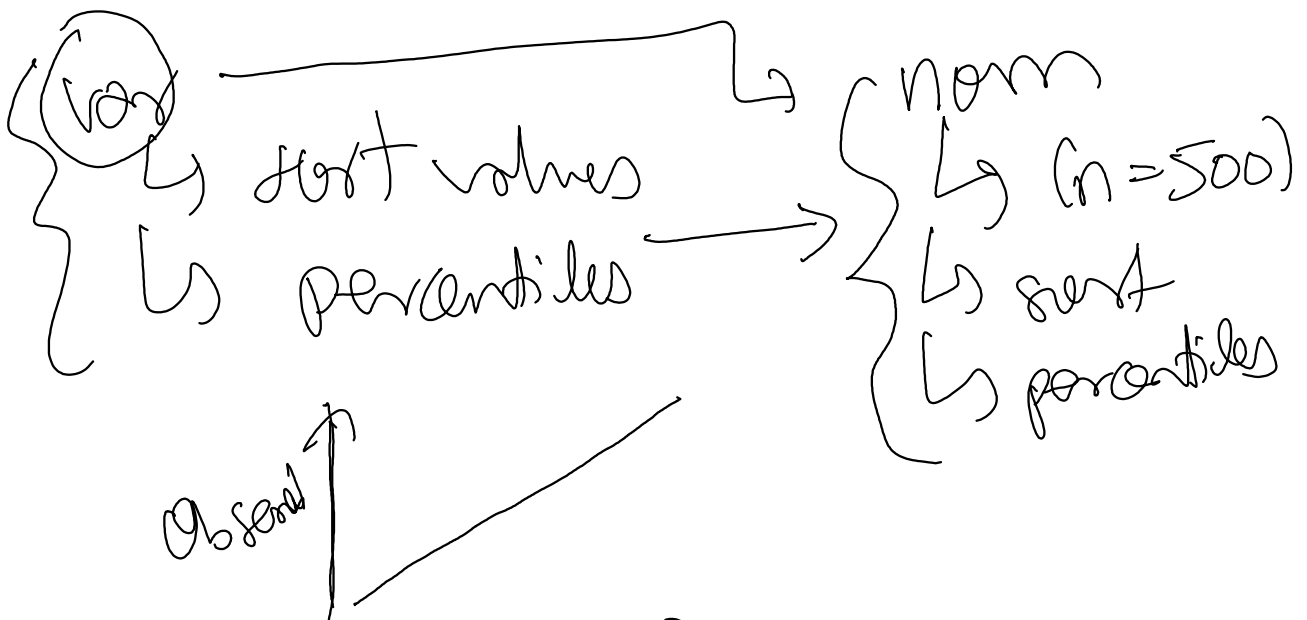
3. **Multivariate:**

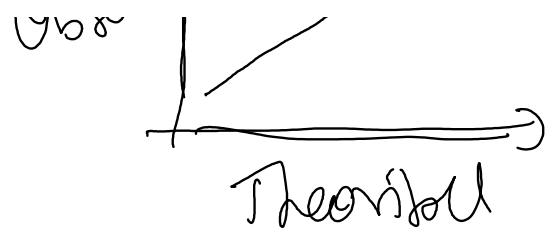
- **Pair Plots**

- **Correlation Heatmap:**

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

- **Facet Grid (Seaborn)**





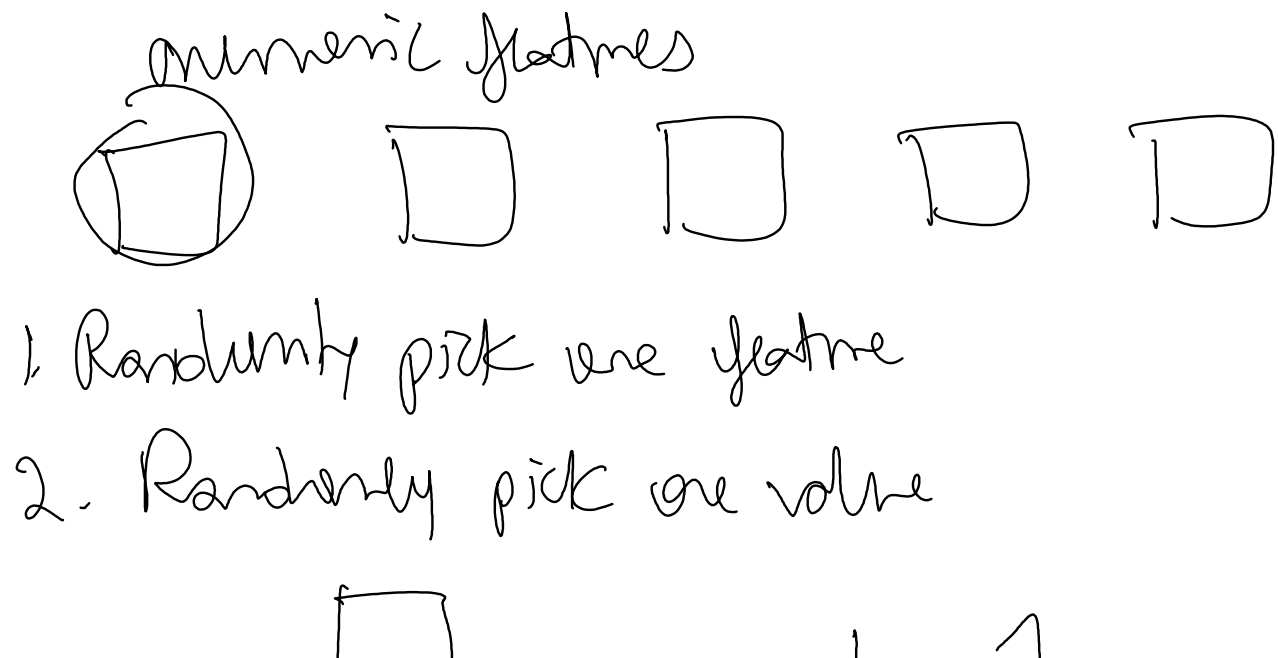
Steps

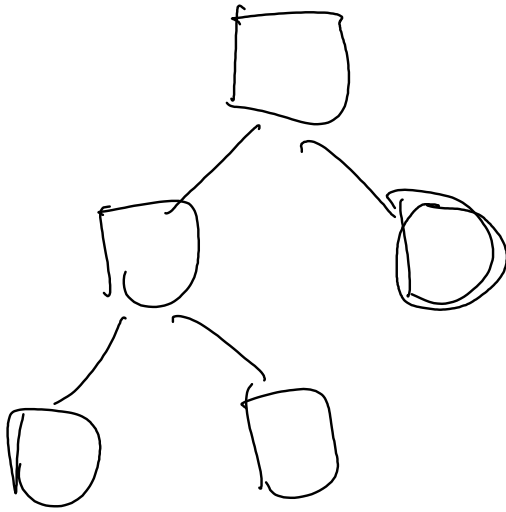
Tuesday, April 23, 2024 12:30 AM

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:

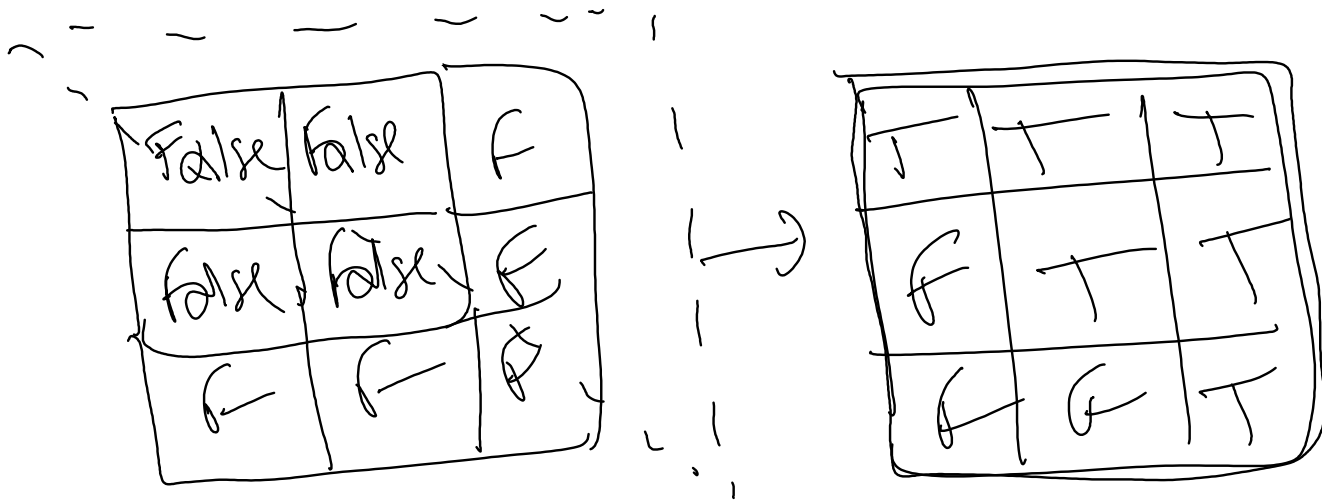
- Summary
 - Univariate plots
 - Bivariate plots (w.r.t. the target variable)
 - Hypothesis Testing (normality, strength of association)
 - Multivariate plots
 - Inspect missing values and extreme values in-depth
 - Filter for necessary subsets
 - Inspect values of other features (plots, summary stats)
 - Note observations
- Feature Engineering
 - Create new features
 - Repeat above steps for newly created features
 - Repeat above steps iteratively
 - Note all observations



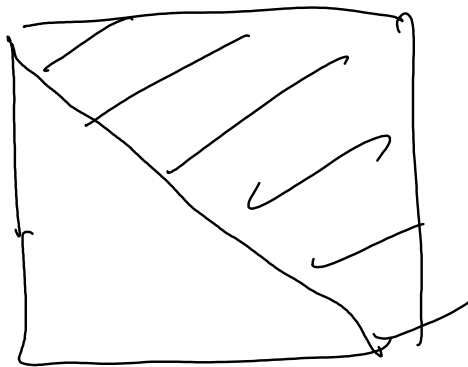


-1, 1

	dr	sr	pr
dr	X	X	X
sr		X	X
pr			X



copy()



Automated EDA

Tuesday, April 23, 2024 12:52 AM

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