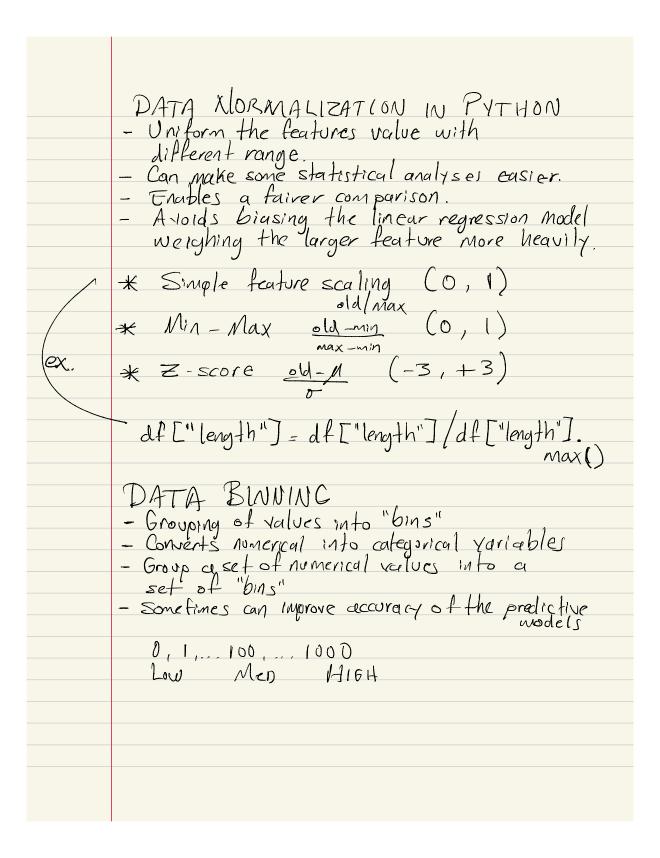
DATA PRE-PROCESSING Converting or mapping raw data to another format, in order to prepare it for further analysis. Also called DATA CLEANING or DATA WRANGLING
PHIT WICANGLING
Missing Valves Deta Formatting Deta Normalization (centering / scaling) Deta Binning (numerical to Cartegorical) Categorical to Numeric Variables
To make statistical groups of data modeling easier DEALING WATH MISSING VALUES
1. Try to find out the corresponding value, going to the source of the data Z. Remove the data where the MISSING Value is found> REDUCING THE DATA
3. Replacing data -> UNACTURACY RISKS -> Average -> Most common (categorical) -> buessing based on unowledge from the gatherer

PYTHON LIBS FOR REMOVINS -> MISSING VALUES
→NUSSING VALUES dropna dxis=0 → remove rows
inplace = True -> affects df
-> REPLACING VALUES MEAN = df["column"]. mean() df["column"]. replace (np. nan, mean)
DATA FORMATTING Put in a standard of expression that will allow users make meaningful comparisons—more clear
- casy to aggregate - casy to compare ex. df ["column"] = Z35 [df ["column"]
df. rename (columns = { "column 1": "new name"}, inplace = True)
IN CORRECT DATA TYPES of ["column"] = of ["column"]. astype ("int")



Example bins = np.linespace (min (df ["price"]),

max(df ["price"]),

4)

group-names = {"Low", "Medium", "High"} df["price-binned"]=pd.cut(df["price"],bins, labels=groyo_names, include_lowest=True) We can use Histograms to visualize the distribution after theire been divided into bins. CATEGORICAL -> NUMERIC - Most statistical models connot take in the objects strings as input pd.get-dummies (df ["foel"])