# DATA PREPARATION AND CLEANING

## Remove Duplicates:

* Duplicate case are overweighed and bias ML model training
* Identify duplicate cases
  + By unique ID
  + By value – with caution!
* Removal strategies
  + Keep most recent (or oldest)
  + Keep first
  + Keep Last

## Missing Values

* Use exploration to detect
* How are missing values coded?
  + NULL
  + A string or number, -99999, 0, NA, etc.
* Treatment strategies
  + Remove column with mostly missing values - little information.
  + If few rows remove
  + Forward or backward fill
  + Impute mean, median, trend, etc.

## Errors, outliers

* Use exploration to detect
  + Erroneous values
  + Vs. Important data; keep!
* Identify cases
  + Statistics
  + Visualization
* Treatment strategies
  + Limit to min-max range.
  + Same as missing values.

## Scaling

* Improper scaling biases ML model training
  + Features with large numeric range can dominate training
  + Large mean values skew model parameters
* Z-Score
  + Mean = 0, standard deviation = 1
  + If distribution is approximately normal
* Min-Max
  + If distribution far from normal
  + Fail with outliers!
* Treat errors and missing values before scaling!

## Splitting

* Split data before training
  + Independent training, test and evaluation subsets
  + Failure to split causes information leakage
  + Using same data results optimistic bias
* Bernoulli sampling
  + Random sampling avoids bias
* Cross Validation
  + Resampling method

## SUMMARY

* Data preparation is key to success with ML
* Iterative process
  + Data exploration used to identify problems and test results
  + Find problems when training and evaluation ML models
  + Treatment of problems

**EXPECT TO ADDRES DATA ISSUE MULTIPLE TIME IN REAL/WORLD PROJECTS!**

# FEATURE ENGINEERING

* GOAL: Develop highly predictive features
* Raw features are often not best

**GOOD FEATURES ALLOW SIMPLE MACHINE LEARNING ALGORITHMS TO WORK WELL**

**POOR FEATURES RESULT IN POOR PERFORMANCE FROM EVEN THE BEST MACHINE LEARNING ALGORITHMS**

## FEATURE ENGINEERING STEPS

* Exploration to understand data relationships
* Transform features
* Compute interaction terms
* Visualization to check results
* Test with ML model
* Repeat above as needed

## TRANSFORM FEATURES

* Why transform features?
  + Improve distribution properties
  + More covariate with label
* Common transformations
  + Log, exponential, square, square root, variance, etc.
  + Difference, cumulative sum, etc.
  + Nonlinear transformed features are not colinear.

## INTERACTION TERMS

* Example:
  + Predict number of rider on bus route
  + Depends on time of day **and** holiday
  + Time of day by holiday is an **interaction term**
* Compute interaction terms
  + Mean, median, etc.

## Summary

* Predictive features are key to success with ML
* Iterative process
  + Data exploration used to identify and test ideas
  + Test multiple ideas

**ITERATIVE PROCESS: TTRY LOTS OF IDEAS, TEST, FAIL FAST, KEEP WHAT WORKS.**