Udacity: Predictive Analytics for Business
Problem Solving with Analytics

Cross Industry Standard Process for Data Mining (CRISP-DM)

1. business understanding - what is decision, what info needed, type of analysis
2. data understanding - what data needed, available, type.
3. data preparation - gather, cleanse, format, blend, sample
4. analysis/modeling - methodology, identify important factors
5. validation
6. presentation/visualization
   - custom to audience
     - walk audience through analysis.
     - tell story
     - reference data sources

Methodology Map:

Business Problem

Predict Outcome				Data Analysis
Data Rich			Data Poor	GeoSpatial
Numeric		Classification	A/B testing	Segmentation
Continuous	Time Based	Binary / Non-Binary		Aggregation
				Description
- Linear Regression - Decision Tree - Forest Model - Boosted Model	- ARIMA - ETS	- Logistic Regression - Decision Tree	- Forest Model - Boosted Model	

Linear regression: $y = mx + b$

$$m = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

coefficient of determination, $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

sum of squared total, $SST = \Sigma_i (y_i - \bar{y})^2$

sum of squared regression, $SSR = \Sigma_i (\hat{y} - \bar{\hat{y}})^2$

sum of squared error, $SSE = \Sigma_i (y - \hat{y})^2$

multiple linear regression,

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots$$
$$= b_0 + \Sigma_i b_i x_i$$

Data Wrangling: Understanding Data

3 types of data structure
- structured - organized into rows / columns
- unstructured - no structure
- semi-structured - some structure, e.g. log, require parsing.

Data sources
- computer files - e.g. xls, xml, csv
- database - e.g. mongo DB, mySQL
- web-based - salesforce, twitter, ..

Data Types
- string - alpha numeric
- numeric - integer, float,
- date time - date, time
- boolean - True / False
- special objects - images, maps, pdf, audio files, video files, ...

Data Issues:

Dirty data - missing
- duplicate
- unreadable format - not parsed properly
  - extra character / missing character
- incomplete data

Parsing Data - csv
- delimiter - ',', '\t', '.' } text to Column

Extra characters - " ", $, ^M
- Formula - Replace
  Left - length of character

Duplicate - Unique
Sum (groupby, Count) + Filter
" " + Sort

Mining - some algos can't work with Missing data
adds bias
- delete wimy Filter is Not Null
- explore wimy Field Summary
- impute - replace with mean, median, mode
  - Summary + Append + Formula
    if, else
  or use Imputation

- others - multiple imputat - create model to predict
  - full informati maximum likelihood - factor missing values into model

- number of missing w significant of values

Outlier - incorrect data or abnormal but correct data
- box-whisker plot - >1.5x interquartile range beyond 1st, 3rd quartile
- build models w/ w/o outlier + compare
- drop only if obvious outlier
- truncation - limit to max.

Data Formatting:
Transpose - Transpose tool
Aggregate - Sum tool
Cross tabulation - join 2 tables

Data Blending:
Union - appends data
- Union tool

Join - join tables
- Join tool - left, join, right

Fuzzy matching - Jaro - measures characters in common
Levenshtein - counts number of edits to convert one string to another.
- match 2 to 3 columns for duplicates, repeat for all permutations

Spatial Matching
- ~~lines~~ points
- lines - strings of latitude, longitude
- polygon - vertices of "

Spatial Blending
- create Points, trade area tools, spatial match tool

Predictn Variables:
matrix of correlation
1. Pearson correlatn coefficient
   - normalized covariance by standard deviations
   - $\frac{cov(X,Y)}{\sigma_X \sigma_Y}$ when applied to population
   - $r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$ when applied to sample

2. Spearman's rank-order correlatn
   - $\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$

3. Hoeffding's independence test

Heat plot to visualize correlations - positive, negative.
- Association tool

Preparing data - EDA
- Field Summary tool — interactive — Browse tool
- Null/missing values — filter out (low % missing)
  remove field (high % missing)
  impute records (if important)

- Frequency tool — interactive — Browse tool

Classificatn Models
Binary classificatn - Logistic Regression
multiple linear regression, $y = b_0 + b_1 x_1 + b_2 x_2 + \dots b_n x_n$

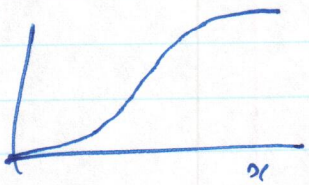
logistic regression, $\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x + b_2 x_2 + \dots + b_n x_n$
p = probability of outcome

$$p = \frac{e^{(b_0 + b_1 x_1 + \dots b_n x_n)}}{1 + e^{(b_0 + b_1 x_1 + \dots b_n x_n)}}$$

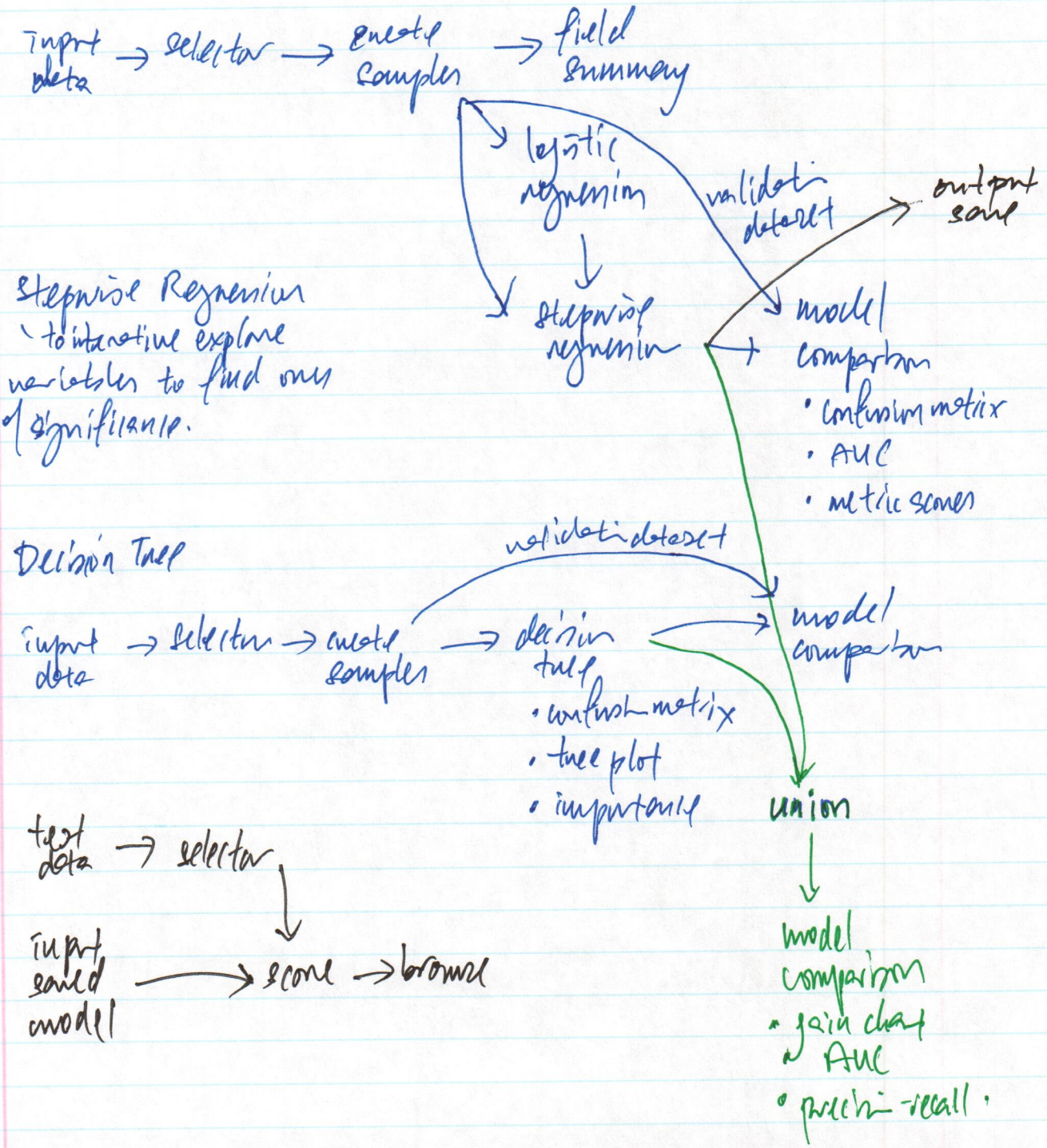
logit: $\eta(p) = \ln\left(\frac{p}{1-p}\right)$

probit: $\eta(p) = \Phi^{-1}(p)$

log-log: $\eta(p) = \ln(-\ln(1-p))$



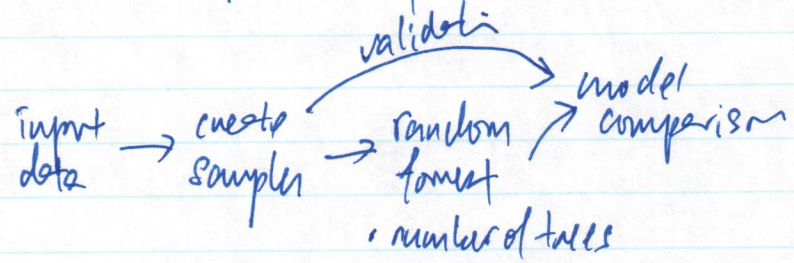
Alteryx workflow:



Non-Binary Classificat Model:

- Decisi tree - tends to overfit training dataset
  - need to validate with validati test set

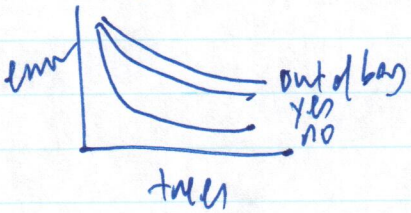
- Random forest - creates many trees to form ensemble of decisi trees
  - each tree created by different randomly selected subsets of original data → bootstrap
  - looks at result as a whole to make prediction
  - overfitting gets averaged out by all trees.
  - majority vote over all terminal nodes of each tree.
    probability averaged over all trees.



- Out of Bag Error Rate - indicates model performance with own-validati set in training dataset.

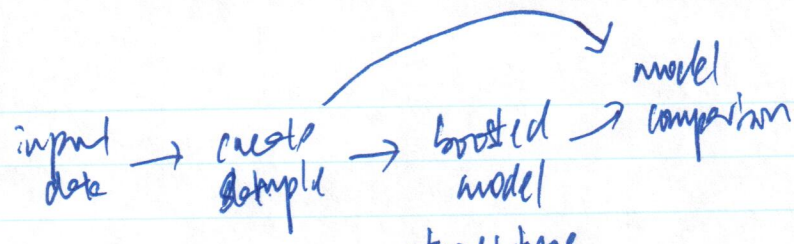
- Confusion matrix - true/false positive/negative

- Percentage Error for Trees - number of trees needed to minimize error i.e. convergence



- Feature importance by mean decrease in Gini

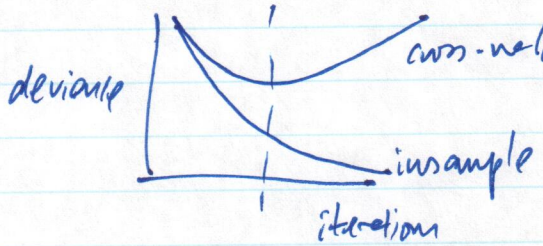
- Boosted Model - calculate errors from trees
  modify tree to reduce error
  repeats



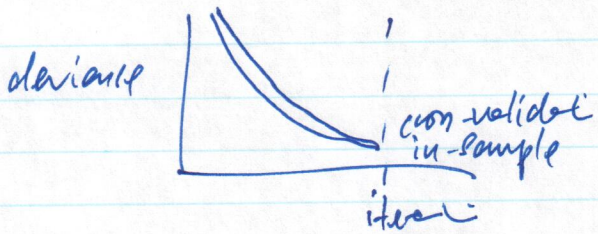
- target type
  - continuous,
    binary
    multinominal – 3 or more categories
- max trees
- cross-validation – k folds

→ ~~imp~~ feature importance plot - which is important

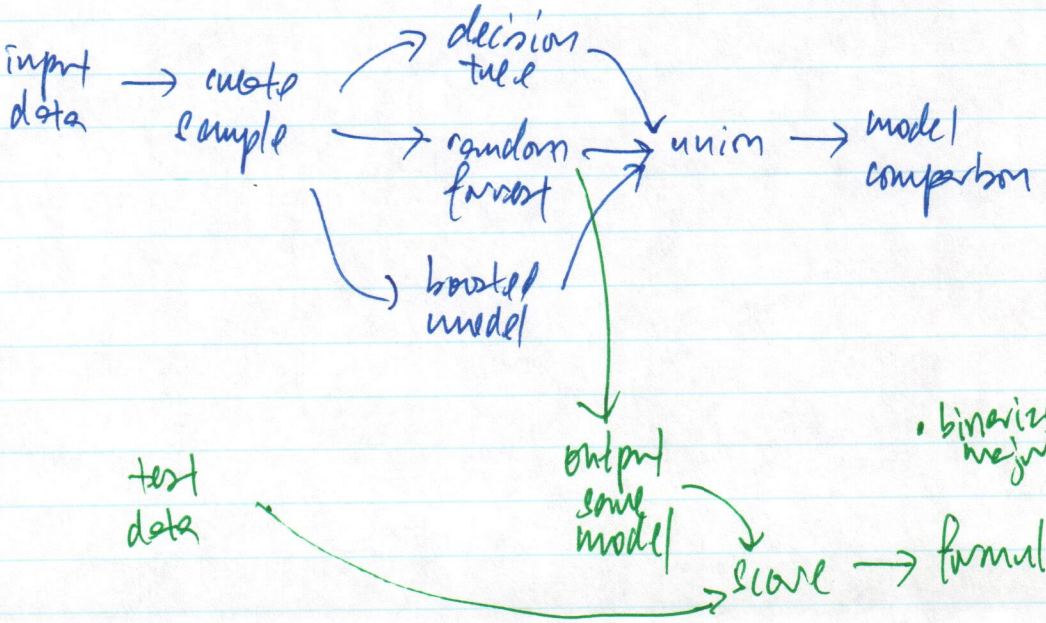
→ number of iteration assessment plot



- minimum number of iteration needed
- large differences between in sample (training) + cross-validati indicate poor performance with validati data.



- similarity between cross-validat + in-sample, good predicti with validation data. performance



A/B Testing

Unit - person
store
production line
website

Treatment Group:
- collection of units getting treatment
- treatment = change

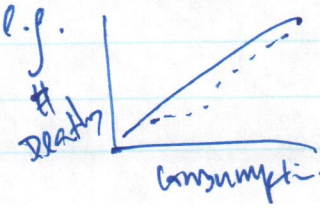
Control Group:
- baseline comparison
- no treatment applied
- control units

Experiment Variable
- treatment change to see affect on target variable.
- independent

Control Variables
~~predictors~~ sales
- customer demographics
  location
- makes sure treatment units + control units are similar to make comparison on experiment variable.

Steps to identify control variables
1. list of potential variables
2. Is data available
3. logical connection between control + target variable
4. Test ~~correct~~ correlation between control + target variables
5. Test correlation between control variables.

Lurking variable / confounding variable
- highly correlated to target, controlling variable
- cause to include controlling variable that really is not important.
- e.g.



no cause-effect, but lurking is temperature.
use logic to avoid.

Experiment Design
- Randomized Design
  - treatment, control units selected randomly on the fly.
  - used when hard to control variables, volume + velocity of data is high enough to avoid bias.
  - e.g. website, phone-based experiments.

- Matched Pair
  - volume of observation is low
    concern for bias = high
    cost per observation = high
  - treatment unit matched to control unit
  - pair analysis, aggregate to single result.

Length of experiment
- minimum 1 full cycle of data to reduce chance of bias in responses.

Randomized Design Example:
- Treatment Control
- target variable - click on link - binary Yes - No
- control variables - customer demographics
  membership existing
- unit of diversion - how to assign units to control + treatment groups.
- population - people with app
- duration } ensure control + treatment groups represent population
- size
  - e.g. 1 week
    unique visitors tracked by cookies or IP.
  - e.g. 2 servers - control
    1 server - treatment

test of means analysis
- test of whether mean of treatment, control groups are same
  - p = likelihood difference between means = ϕ
    p < 0.05 statistically significant.
- Test of Means tool

Matched Pair Design Example:
1. identify outliers for treatment units.
2. need > 10 treatments
3. randomly selected treatment units.

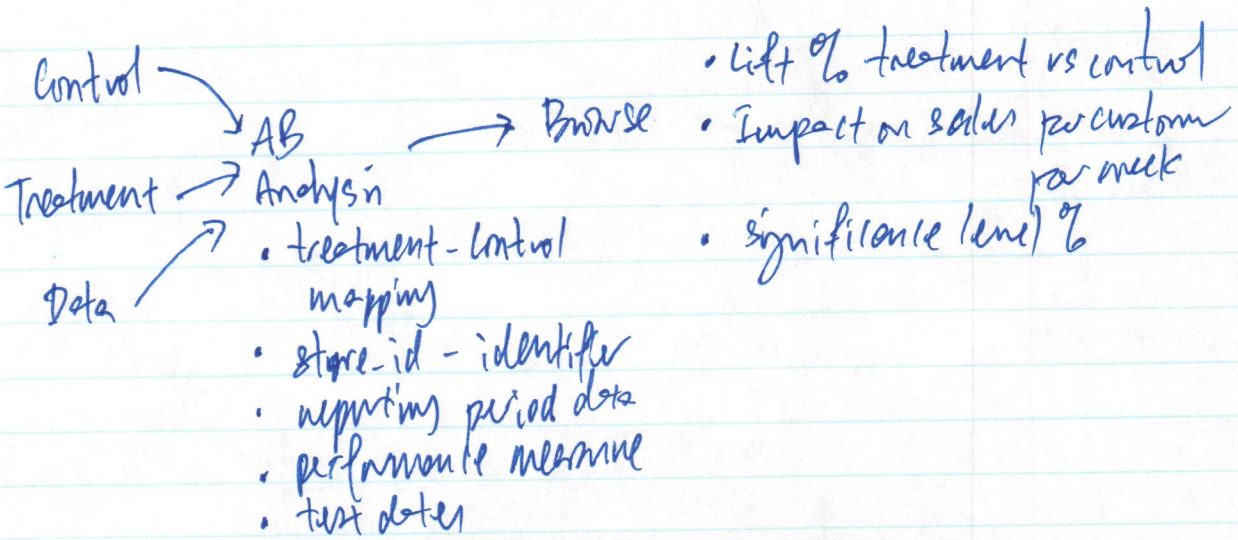
Control ~~stores~~ units
- ~~baseline~~ serve as baseline
- match to treatment unit based on control variables
  - sales volumes
  - products sold
  - location - e.g. state, city
- AB Controls tool
  - KD tree
    metric - distance
- balance number of control units and average distances

input control stores.csv → control → AB Controls → browse

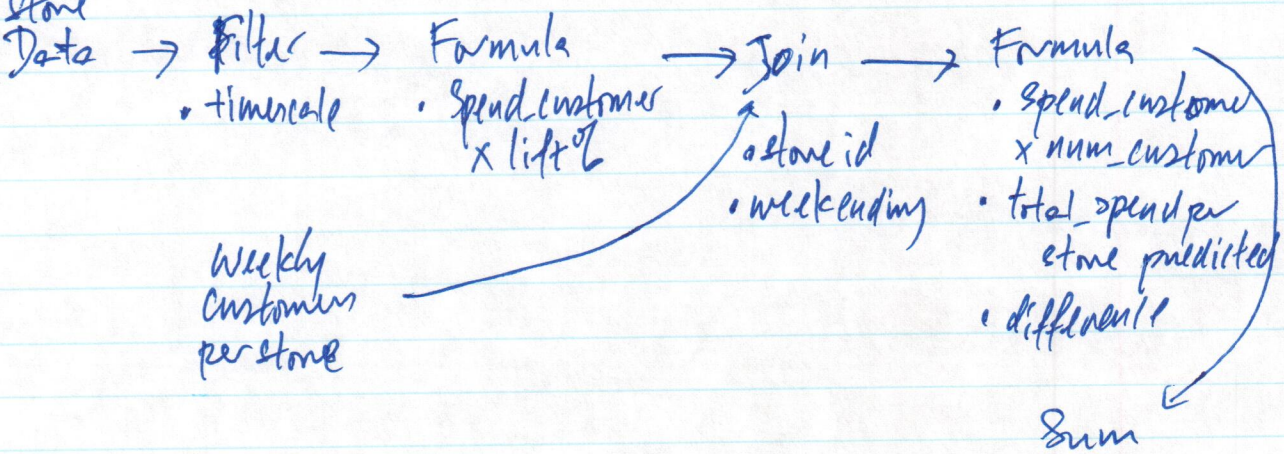
input treatment stores.csv → treatment → AB Controls

- common store ID
- variables - product sold
  sales
- number of controls

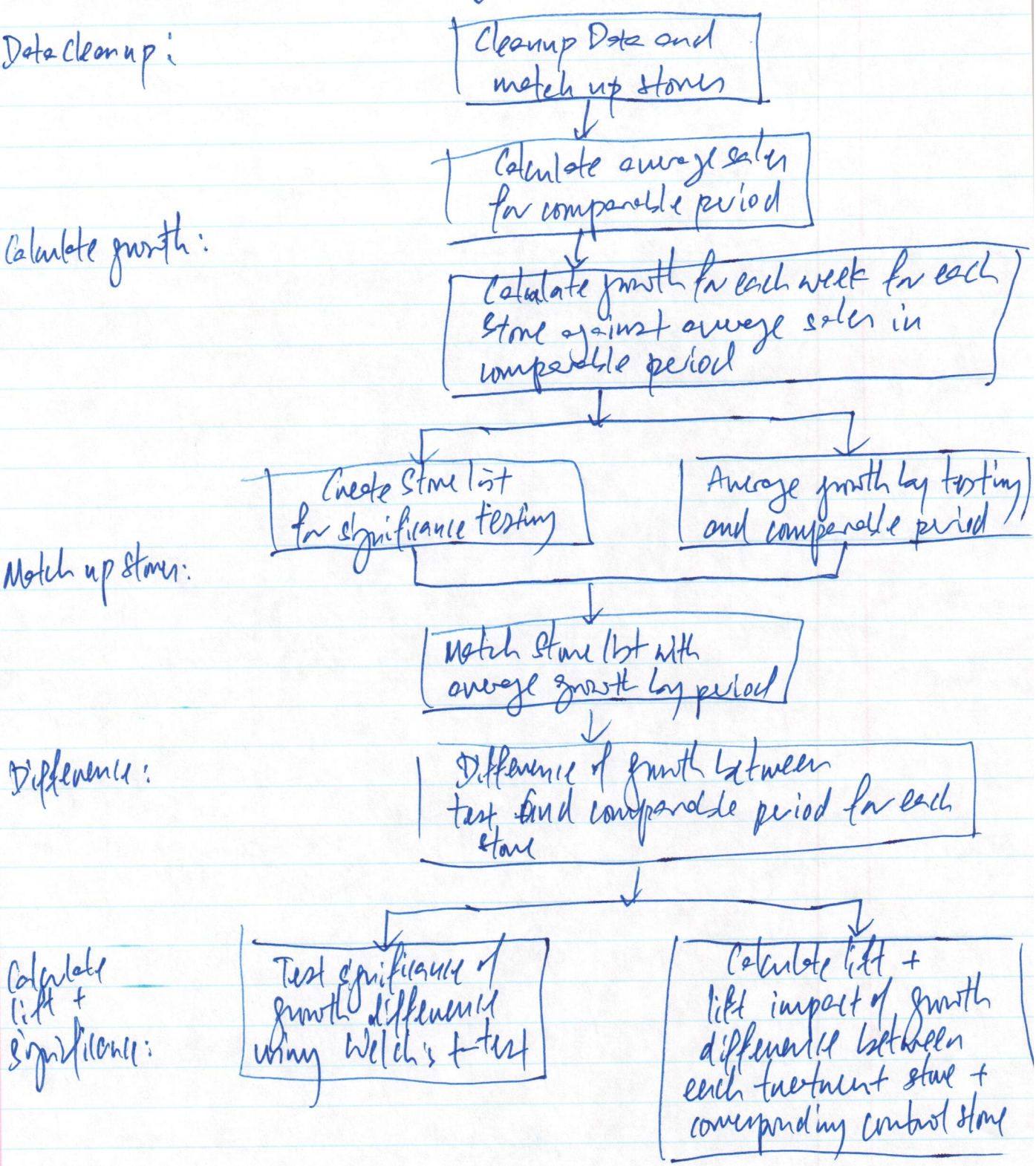
- mean + variance of differences between each control + treatment pair differ from ϕ



Estimated New Sales from change:



A/B Test Flow to Calculate Lift + Significance for Store Sales.



Match Pair Practice:
Price elasticity of demand
- responsiveness of quantity demanded of a goods/service to a change in its price.

A/B testing analyzes whether a change applied to an experimental group is statistically different from a control group.

Control - spa-facial - $88-99
Treatment - -$76-99) treatment group
$87-99)

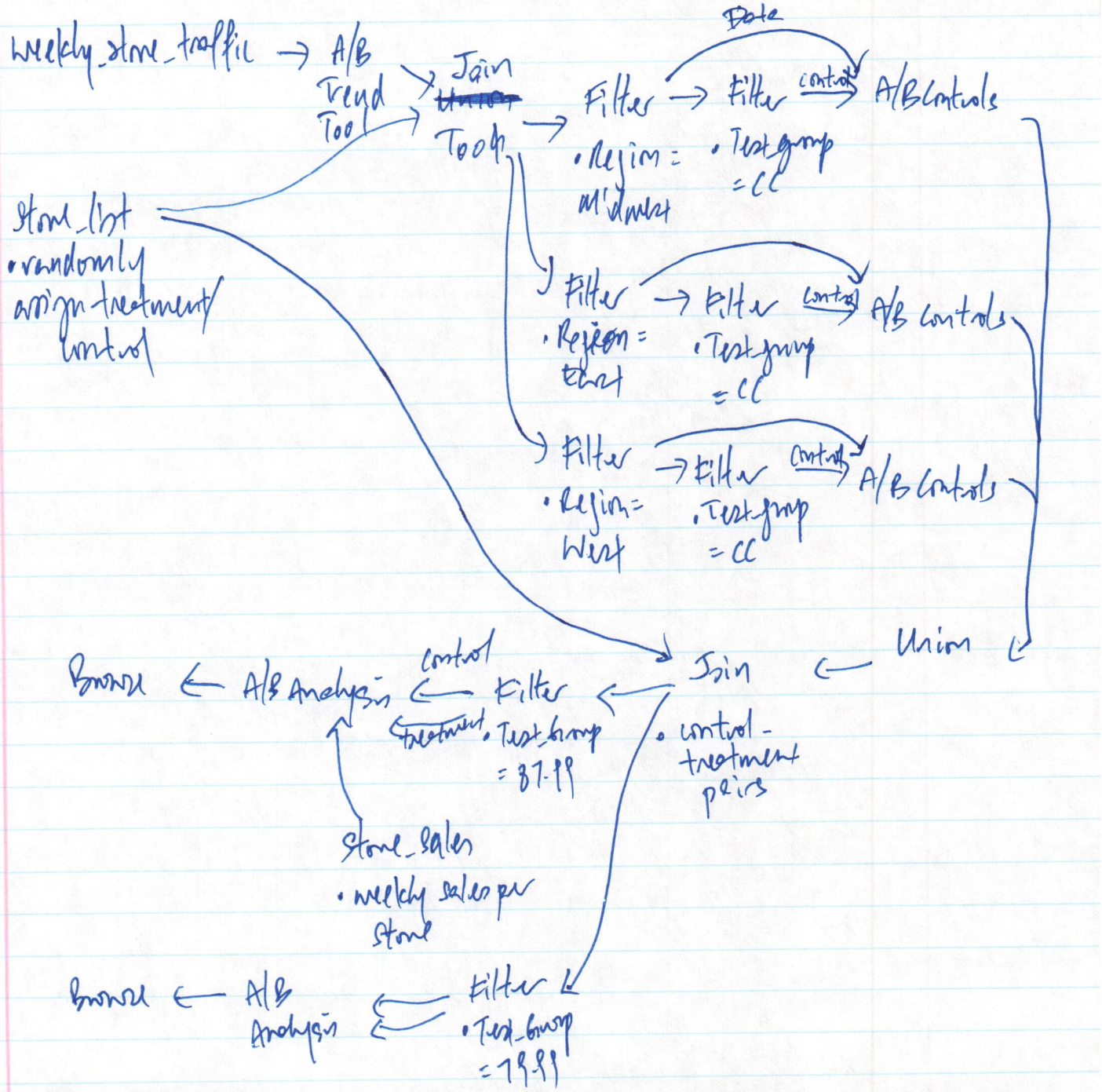
Control variables
- average sales per store per month
- average invoices per store per month
- average number of facials by SKU per store per month
- average sales of " " " " "

Continuous trends
- Growth in store traffic
  └ invoices per week
- Seasonal patterns in sales volume
  └ total sales per store per week

A/B Trend tool
└ needs 1 yr of data + minimum 12 periods in addition to year if measuring weekly.

e.g. 1 customer every 10 weeks
→ length of experiment = 10 weeks
→ comparison period = 10 weeks, same as testing period
→ length historical data = 1 yr + 12 periods
52 weeks + 12 weeks = 64 weeks

Test period, 10 weeks
aggregate weekly sales, invoices
→ weekly per store invoices/traffic ⇒ A/B Trend tool
→ Store list data matches of control to treatment stores ⇒ A/B Controls tool.
→ store sales, gross margin per week per store ⇒ A/B Analysis tool.

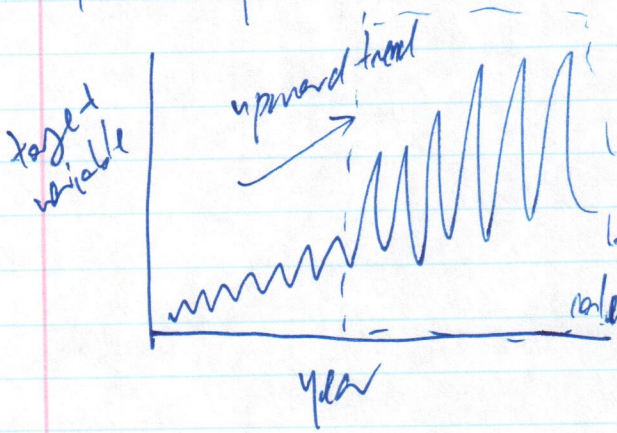


Time Series Forecasting

- Order matters
- continuous time interval
- equal spacing between 2 consecutive measurements

) time series attributes

Steps
- investigate + clean data
- determind trend, seasonal components
- apply findn to ARIMA model
- forecast future

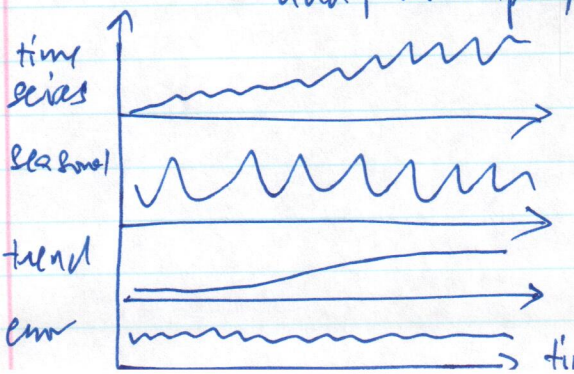


→ seasonal pattern – repeating pattern at fix intervals of time.

calender year

cyclical pattern – upward/downward trend
- longer than seasonal
- much harder to predict
- e.g. stock market

Exponential Smoothing Models
time series - error, trend, seasonality
add, multiple, left out

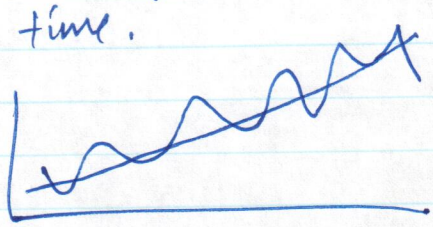


Decompose time series data in seasonal, trend and remainder error components.

Additive - trend, seasonal variations are relatively constant over time.

linear trend

Multiplicative - trend, seasonal variation increase / decrease over time.



exponential trend

Scenarios

No trend, no seasonal
" , seasonal constant
" , seasonal increasing

Trend-linear, no season
" , season constant
" , season increasing

Trend-exponential, no Seasonal
" , seasonal constant
" , seasonal increasing

ETS Models (Error, Trend, Season)

1. Simple Exponential Smoothing
- no trend, no seasonality
- forecast = $w_1 x_1 + w_2 x_2 + \dots w_t x_t$

$w = \alpha(1-\alpha)^t$

α = range 0 to 1
- more weight place on more recent data

2. Double Exponential Smoothing / Holt's Linear Trend Method
- add trend + level
linear

3. Exponential Trend Method
- multiply trend + level
  (exponentially

4. Damped Trend Methods
- damps trend line into flat line into future
- additive & multiplicative
- damping parameter, Φ

5. Holt-Winters Seasonal Method
- trend, level, seasonal
- add, multiple method
- use with Φ

ARIMA Models:
- Auto Regressive Integrated Moving Average

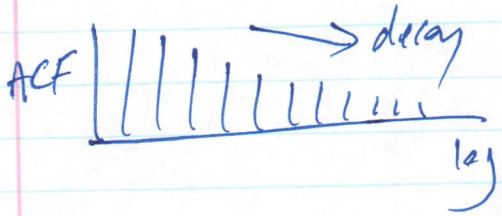
1. Non-Seasonal ARIMA (p, d, q)
   AR - p - periods to lag for in prediction

   I - d - number of transformers used to transform a time series into a stationary one.

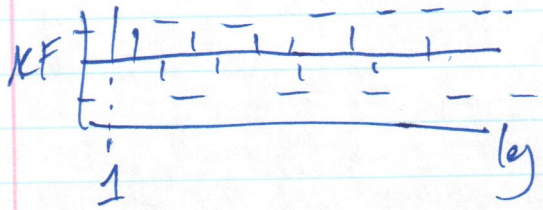
   MA - q - lags of error component (not explained by trend/seasonality)

Stationary time series - mean + variance constant over time.

Auto correlation
- how correlated a time series is with past values.
- Autocorrelation Function (ACF) plot



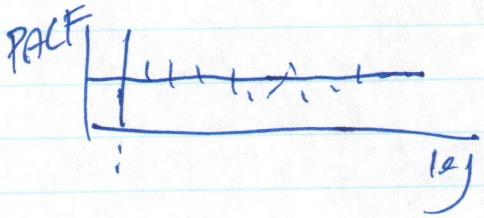
Not a stationary time series.



After lag 1, ACF much lower, stationary series

AR model - positive ACF at lag 1.
MA model - negative ACF at lag 1.
- cuts off sharply after a few lags.

Partial correlation - correlation between 2 variables controlling for values of another set of variables.

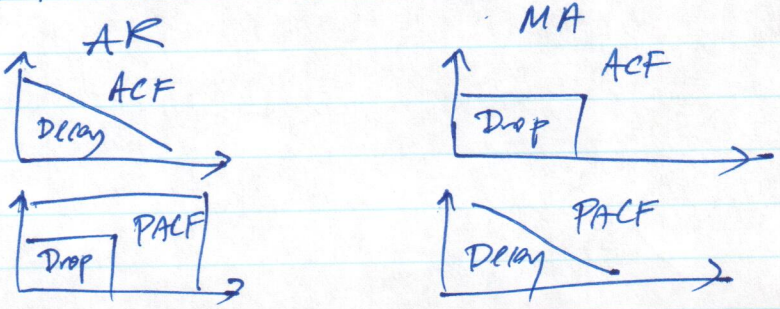


AR model - PACF drops off after lag-k indicates AR-k model
MA model - PACF drops off gradually, indicates MA model

2. Seasonal

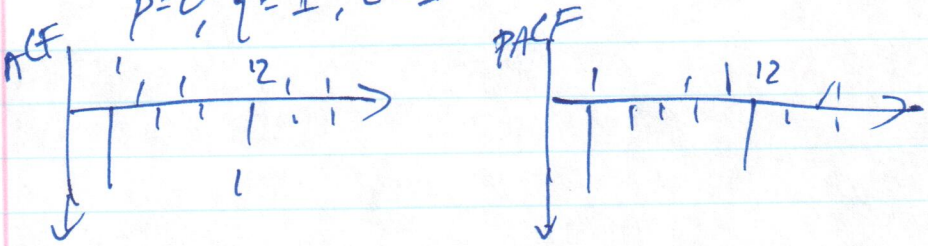
$ARIMA(p,d,q)(P,D,Q)_m$ (seasonal)

- use seasonal lag to calculate differenced series.
- rest is similar to non-seasonal



P=0, Q=1, D=1

p=0, q=1, d=1



p, d, q - look at lags not same as seasonality - 12, 24
P, D, Q - only look at lag at seasonality - 12, 24

Interpreting Results
Hold out sample - subset of data withheld to check accuracy of predictions of model.

Residuals = difference between observed + forecasted value.
Good models:
- Residuals in model uncorrelated
  - plot in ACF
  - adjust for lags

- Residuals means = 0
  - otherwise subtract mean from model

Measures of Error
1. Mean error (ME)
   - average of difference between actual + forecasted.

2. Mean percentage error (MPE)
   - average of percentage difference between actual + forecasted.

3. Root mean squared error (RMSE)
   - sample standard deviation of differences between predicted and observed values.

4. Mean absolute error (MAE)
   - average of absolute error. less sensitive than RMSE to large errors.

5. Mean absolute Percentage error (MAPE)
- compare between 2 different data series.

6. Mean absolute scaled error (MASE)
$$= \frac{\text{mean absolute error of model}}{\text{mean absolute error of 1st difference of series}}$$
- relative reduction in error compared to naive model.

Akaike Information Criterion (AIC)
- measure of relative quality of statistical model
- balance of goodness of fit of model and complexity of model.
- lowest is best

Confidence Interval
- interval within which forecast will lie.

Segment and Clustering
Standardization – treat all units the same through generalization.
- large errors, but easy.

Localization – treat local needs individually
- small errors, but effort needed to differentiate.

Clustering - statistical methodology to group similar objects into clusters based on multidimensional features or attributes.

Distance - to quantify or measure similarity.

- Data Preparation

1. Selecting data

Retail example for required data

Benchmarking system:
- square footage
- business hours
- management
- street traffic
- employee + hours
- population
- competition

Pricing rules:
- income
- life stage
- distance to competition
- type of competition
- regulatory zones

Product Allocation + Space Planning:
- area demographics - income, age, ethicity
- product category sales history
- sell-through - compare amount of inventory retailer receives from manufacturer against what is sold to customer.
- product discount trends

Predetermined bias

Retail - e.g. product category sales
- force categorization store in segment of simila stores, despite demographic + social data.
- cluster based on demographics + historical sales to see any bias present.

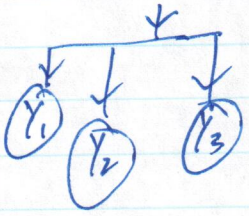
2. Data types - ordinal
- hot encoding

3. Data quality - missing data

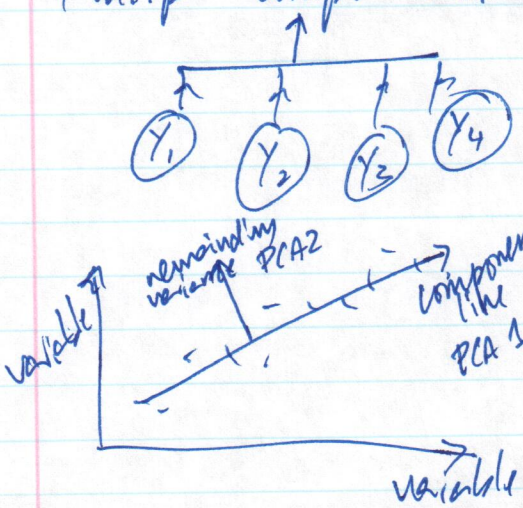
4. Scale data
   - z-score - number of standard deviations from mean
     $$z = \frac{x - \mu}{\sigma}$$
   - standard - rescale to 0 to 1

- Variable Reduction:
  - account for most of variance in all observed variables.

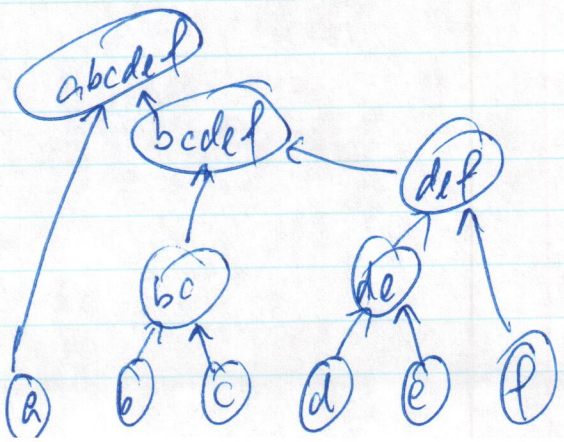
Factor analysis - hidden ~~layer~~ factor underlying variables in question.
- correlation between variables
- find cause between variables.



Principal Component Analysis (PCA) - summarizes all variance within total variables into fewer components, reducing # variables.
- accounts for total variation
- explain as much of variance as possible.
- PCA tool

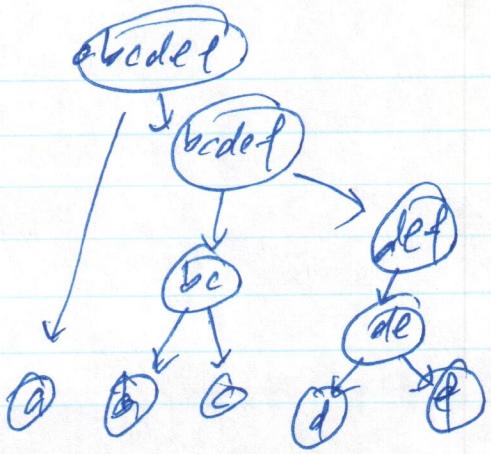


- Clustering Models:
  - Hierarchical Clustering
    1. bottom-up



2. top down

Cluster based on linkage
1. Single - closest point to point distance
2. Complete - farthest point to point distance
3. Average - average distance
4. Centroid - centroid to centroid



Stops when all objects are in its cluster - no need to specify K, # cluster

- K-centroid Clustering
  need to specify k
  initial centroid chosen randomly
  assign objects closest to centroid
  recalculate centroid
  repeat until no or little change in centroid

works better with large number objects
even number of cluster
different density of cluster.

Highly sensitive to outlier, affecting centroid

Validation:
- internal validation
- external validation

evaluate results using criteria of cluster
compare results using an index.

objective of clustering
1. minimized intra-cluster separation → compactness
2. maximized inter-cluster separation → separateness or distinctness

Indices

1. Adjusted Rand Index
   - how similar objects within a clusters are w cluster stability

2. Calinski-Harabasz Index (CH)
   - measures both compactness + distinctness of clusters.

K-~~Centroids~~ids

Input → Diagnostics Tool → Browse

- standardize ← z score unit
- cluster methods ← K-means, K-medians, Neural Gas
- min/max number of clusters

Calculates Adjusted Rand and CH Indices between min and max ranges of # clusters.

AR index - higher, more stability
CH index - higher, more compactness + more distinctness.

→ want high median and interquartile range is small.

- vary # clusters, methods, standardization, index metrics.

input data → k clustroid tool → Append cluster → output data

- Validating Models
  - Internal validation
    - best number of clusters
  - External validation
    - holdout set
    - validation through visualization

Data Visualization in Tableau:

1. Worksheet - 1 visualizat
2. dashboard - multiple visualizati
3. stories - combinatio of 2