Missing Values -

MCAR - Missing Completely at Random

MNAR - Missing Not at Random

Ways -

KNN - K nearest

Mice - Package R -

- Pmm - predicted mean matching

- logreg – logistic regression imputation

- polyreg – polytomous regression imputation for unordered categorical data

- polr – proportional odds model

Speech Recognition –

Sphinx – Speech recognition system CMU

Google speech recognition - gTTS

Houndify

Collecting data from PDF – pdfMiner

hOCR – xml markup for OCR

* ocr\_par, ocr\_line, ocrx\_word --

Bar codes – pyzbar

Pdftools – based on libpoppler

Extract text, convert pdf to image

Tabulizer – Extract tables from pdf

Face eye detection –

Rapid Object Detection using a Boosted Cascade of Simple Features

OpenCV – open source computer vision

Convert for grey scale and back to color

Let’s quickly organize common secondary data sources:

Internal sources External sources

Records

Systems

(e.g., ERP,

CRM, SCM

etc.)

Website,

apps etc.

Public Proprietary

Web

Govt.

Etc. Syndicated

Data

Providers

1. Profile data (aka basic demographics)

2. CRM data - Purchase history, site browsing history, brand loyalty,

payment methods etc.

3. Environmental data - real-time weather, temp, geolocation, date &

time...

4. Social media data - trends, likes and preferences, topics discussed of late

5. Real-time events - usually big sports or entertainment events around

which some clever messaging, copy or campaigning can be built.

• E.g., "Go, Mumbai Indians!“, "Buy your Valentine a Swatch."

• 6. Site/cookie data - first party cookies on consumer websites

• E.g., “Traveling? Try the new VIP series 6 suitcases…”

• 7. Search data - SEO etc. But advertising based on this is harder now that

Google has moved to secure search and hides the referrer URL's search

term...

• 8. Contextual data - media section based. E.g., is user browsing finance

section? sports? lifestyle? health?

* Access nodes – html\_nodes(‘a’)
* Access attributes - html\_attr('href')
* Access class – <a href="..” class="storylink">..</a>  --- example -- html\_nodes('a.storylink')
* Access text content – html\_text()