Unit 2

Data Munging

Why data munging?

- Garbage In, Garbage Out
  - Incorrect/Inconsistent data leads to false conclusions
- Quality data trumps fancy algorithms – every single time
- Real world data is dirty

How good is the data? – Data Quality

- We need to be able to quantify how good the data is
- One good way to measure quality of data would be check the data against a predefined set of constraints/criteria.
- The checks should be customized based on the business needs.
  - Same dataset might be good enough to answer a given question while it might be bad to answer another question

Data Quality

Typical areas of checks for data quality:

- Validity
- Accuracy
- Completeness
- Consistency
- Uniformity

Data Quality - Validity

- If the data meets the criteria set forth by business needs
- Can be sub-categorized as:
  - Data Type Constraints
  - Range Constraints
  - Mandatory Constraints
  - Uniqueness Constraints
  - Set Membership Constraints
  - Others: Foreign-key constraints, Regex/Patterns, Cross-field validation

Data Quality - Accuracy

- Measure of how close the given values are to actual values
- ‘Valid’ doesn’t imply accuracy

Data Quality - Completeness

- Degree to which all required data is known

Data Quality - Consistency

- Do the fields agree with each other ?

Data Quality - Uniformity

- Standard values, units

How to ensure data quality – Data Cleanup

- Data cleanup must be a continuous and iterative process
- Typically, it would involve:
  - Inspect
  - Clean
  - Verify
  - Report/Audit

Data Cleanup - Inspect

- Identify issues with data
- Can be done by:
  - Profiling
  - Visualization
  - Other automated tools

Data Cleanup - Clean

- Fix the issues with data
- Each issues type should be dealt differently, based on the business needs:
  - Irrelevant Data: Drop/Remove
  - Duplicates: Drop/Remove
  - Type Mismatch: Type conversion
  - Non-Standard: Scaling/Transformation/Normalization
  - Missing Values: Drop, Impute, Flag
  - Outliers: Remove/Keep

Data Cleanup – Clean Irrelevant Data

- Data that’s not needed for the scope of analysis
- Typically, domain expert would be needed to identify if the data is irrelevant
- Irrelevance might be on Variable(Columns) or Observations(Rows)
- Once deemed irrelevant – the data can be removed/deleted for the analysis

Data Cleanup – Clean Duplicates

- Duplicates can occur as:
  - Data is combined from various sources
  - User error/mistake
- Duplicates should be removed/deleted

Data Cleanup – Clean Type Mismatch

- Non-numerical values in numerical fields
- Date vs Timestamps
- Syntax errors: Whitespaces, typos

Data Cleanup – Clean Non-Standard Values

- Mismatached units
- Date formats
- Sometimes scaling/normalization might be needed

Data Cleanup – Clean Missing Values

- Can be handled as:
  - Drop
  - Impute
  - Flag

Data Cleanup – Clean Outliers

- Innocent until proven guilty
- Typically removed if we have enough data points

Data Cleanup - Verify

- Check if the cleanup resulted in good enough data

Data Cleanup - Report

- Report on what was fixed/changed and how good the current data is

Data Cleanup - Summary

- Data Cleanup is an iterative process, involving
  - Inspect
  - Clean
    - Irrelevant Data: Drop/Remove
    - Duplicates: Drop/Remove
    - Type Mismatch: Type conversion
    - Non-Standard: Scaling/Transformation/Normalization
    - Missing Values: Drop, Impute, Flag
    - Outliers: Remove/Keep
  - Verify
  - Report/Audit