Week-4 : Python for Data Management

Learning Objectives

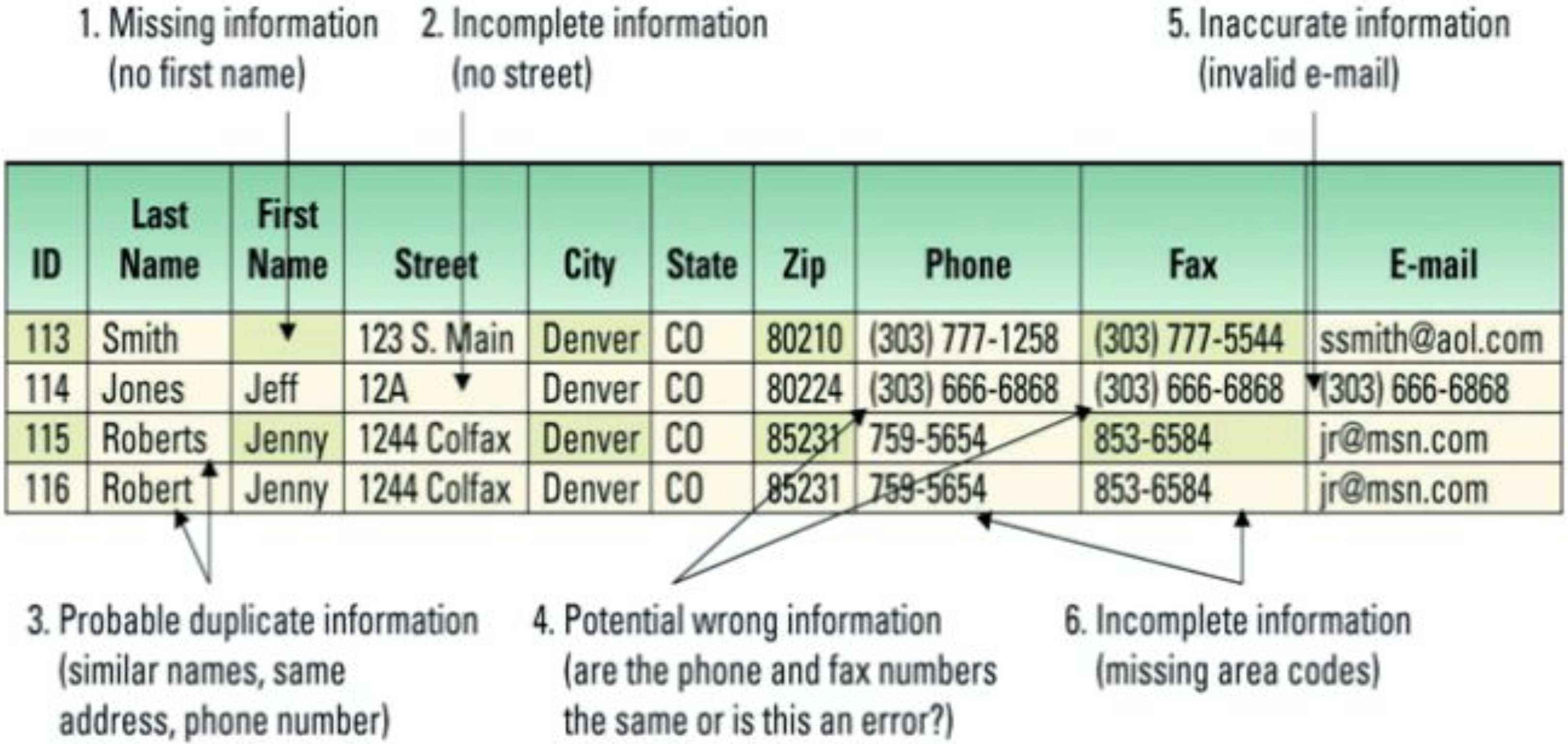
- Understand why managing data quality matters and how it affects decision-making.
- Learn how to deal with missing data so that your datasets are complete and accurate.
- Know how to validate data to make sure it meets the standards you set.
- Learn to spot outliers in your data that could mess up your analysis.
- Understand data profiling to see what your data looks like and identify any issues.
- Learn techniques to clean up and prepare your data for analysis, including transforming features, scaling them, encoding them, selecting the important ones, and reducing dimensionality.
- Understand the importance of data governance, data catalog and metadata management

Data Quality Management

Data Quality Management

- **What** -
  - Accuracy
  - Completeness
  - Consistency
  - reliability and whether it's up to date.
- **Why** -
  - it directly impacts the accuracy and reliability of information used for decision-making.
  - Quality data is key to making accurate, informed decisions.
  - From a financial standpoint, maintaining high data quality levels enables organizations to reduce the cost of identifying and fixing bad data in their systems

Example of Low Quality Data



1. Missing information (no first name)

2. Incomplete information (no street)

5. Inaccurate information (invalid e-mail)

ID	Last Name	First Name	Street	City	State	Zip	Phone	Fax	E-mail
113	Smith		123 S. Main	Denver	CO	80210	(303) 777-1258	(303) 777-5544	ssmith@aol.com
114	Jones	Jeff	12A	Denver	CO	80224	(303) 666-6868	(303) 666-6868	(303) 666-6868
115	Roberts	Jenny	1244 Colfax	Denver	CO	85231	759-5654	853-6584	jr@msn.com
116	Robert	Jenny	1244 Colfax	Denver	CO	85231	759-5654	853-6584	jr@msn.com

3. Probable duplicate information (similar names, same address, phone number)

4. Potential wrong information (are the phone and fax numbers the same or is this an error?)

6. Incomplete information (missing area codes)

Impact of Data Quality on Data Lifecycle

- A data lifecycle illustrates how data (in all its various forms and derivatives,
  - data points
  - Datasets
  - Databases
  - data files
  - visualizations, and code
- This life cycle can be split into eight common stages, steps, or phases:-

  Generation , Collection, Processing , Storage ,

  Management, Analysis ,Visualization and Interpretation
- **Impact of Bad Quality data across the data value chain**

Data Cleaning and Preprocessing

Data Cleaning and Preprocessing

- Impact the model's performance.
- A well-cleaned and pre-processed dataset can lead to more accurate and reliable machine learning models
- poorly cleaned and pre-processed dataset can lead to misleading results and conclusions.

Techniques of Data Cleaning and Preprocessing

1. **Data Cleaning Techniques :-**
   a. Handling Missing Values
   b. Removing Duplicates
   c. Data Type Conversion
   d. Outlier Detection
2. **Data Preprocessing Techniques:-**
   a. Encoding Categorical Variables
   b. Feature Scaling
   c. Feature Selection

Missing Value Treatment

- Missing values pose a significant stress as they come in different formats and can adversely impact your analysis or model.

  How?

  1. removing the missing values altogether.
  2. if the number of rows that have missing values is large, then this method will be inadequate.
  3. For numerical data, you can simply compute the mean, median or mode and input i
  4. For categorical data
     a. you can simply use mode and input it into the rows that have   missing values or you may also define a new level in the variable called others.

Removing Duplicates

- Duplicate records can distort your analysis by influencing the results in ways that do not accurately show trends and underlying patterns (by producing outliers).
- A dataset with duplicate records can produce skewed analytical results and false conclusions. A key data-cleaning strategy is to find and eliminate duplicate records

How?

Pandas helps to identify and remove the duplicate values in an easy way by placing them in new variables.

Data type Conversion

- Ensure that your data is in the appropriate format for analysis or modelling.
- Data from various sources are usually messy
- Data types of some values may be in the wrong format
- For example some numerical values may come in 'float' or 'string' format instead of 'integer' format and a mix up of these formats leads to errors and wrong results.

Outlier Detection

- Data points significantly different from the majority of the data
- They can distort statistical measures and adversely affect the performance of machine learning models
- They may be caused by human error, missing NaN values, or could be accurate data that does not correlate with the rest of the data.

Methods of Outlier Detection

There are several methods to identify and remove outliers, they are:

- Remove NaN values.
- Visualize the data before and after removal.
- Z-score method (for normally distributed data).
- IQR (Interquartile range) method for more robust data.
  - The IQR is useful for identifying outliers in a dataset.
  - According to the IQR method, values that fall below Q1−1.5× IQR (Lower Whisker) or above Q3+1.5×IQR (Upper Whisker) are considered outliers.
  - This rule is based on the assumption that most of the data in a normal distribution should fall within this range.
  - You can treat the outliers by capping the data to the lower and the upper whisker

Encoding

Before categorical data can be utilized as input to a machine learning model, it must first be transformed into numerical data. This process of converting categorical data into numeric representation is known as encoding.

There are two types of categorical data: nominal and ordinal.

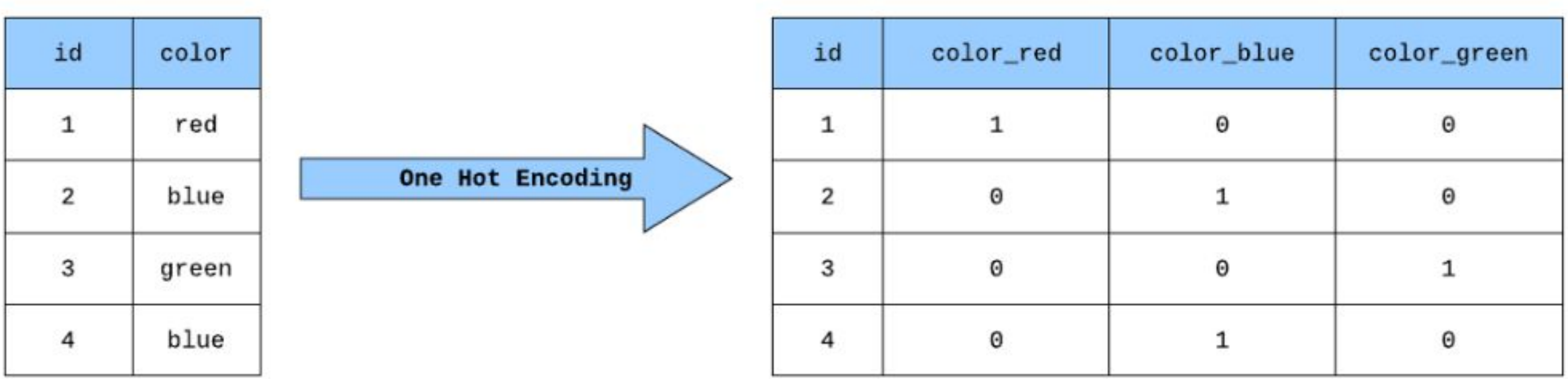
1. Nominal data
2. Ordinal data

There are two types of encoding

- One Hot Encoding
- Label Encoding

One-Hot Encoding

- Representing categorical data as a set of binary values
- where each category is mapped to a unique binary value
- In this representation, only one bit is set to 1, and the rest are set to 0



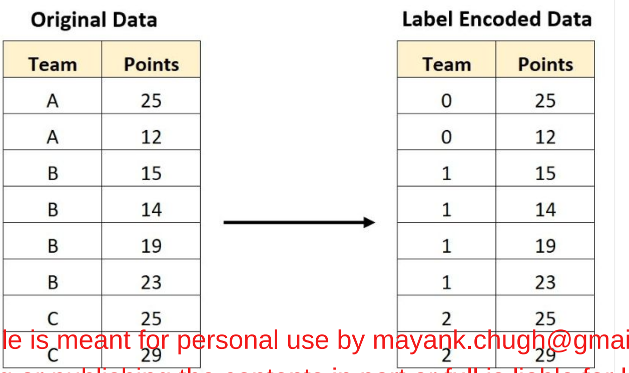
id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Label Encoding

- encoding categorical variables as numeric values, with each category assigned a unique integer
- label encoding introduces an arbitrary ordering of the categories, which may not necessarily reflect any meaningful relationship between them.
- In some cases, this can lead to problems in the analysis, especially if the ordering is interpreted as having some kind of ordinal relationship.



Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

Feature Scaling

1. This means that you're transforming your data so that it fits within a specific scale.
2. Scaling your data before using it for model building can accomplish the following:
   a. Scaling ensures that features have values in the same range
   b. Scaling ensures that the features used in model building are dimensionless
   c. Scaling can be used for detecting outliers as well.

Normalization and Standardization

When data is scaled using **Normalization**, the transformed data can be calculated using this equation.

Where Xmax and Xmin are the maximum and minimum values of the data, respectfully. The scaled data obtained is in the range [0, 1]

$$X_{\text{norm}}^{(i)} = \frac{X^{(i)} - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Standardization should be used when the data is distributed according to the normal or Gaussian distribution. The standardized data can be calculated as follows:

$$X_{\text{std}}^{(i)} = \frac{X^{(i)} - \bar{X}}{\sigma X}$$

Feature Selection

Feature Selection is the process of selecting out the most significant features from a given dataset. In many of the cases, Feature Selection can enhance the performance of a machine learning model as well.

Importance

It enables the machine learning algorithm to train faster.
It reduces the complexity of a model and makes it easier to interpret.
It improves the accuracy of a model if the right subset is chosen.
It reduces Overfitting.

- We can use Pearson's correlation and show heatmap to check the correlation of all features in a dataset
- If the value is **near 1** means those two features are correlated and we can drop any one of them

Data Transformation and Feature Engineering

Data Transformation

- Converting, cleansing, and structuring data into a usable format that can be analysed to support decision making processes
- Data transformation is used when data needs to be converted to match that of the destination system.
- **Type -** Some of the transformation types, depending on the data involved, include
  - Filtering
  - Enriching
  - Splitting
  - Joining data

Feature Extraction

- Reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features).
- Four ways feature extraction enables machine learning models to better serve their intended purpose:
  - Reduces redundant data
  - Improves model accuracy
  - Boosts speed of learning
  - More-efficient use of compute resources

Dimensionality Reduction

Process of reducing the number of input variables or features in a dataset while preserving the most important information

Broadly divided into two categories:

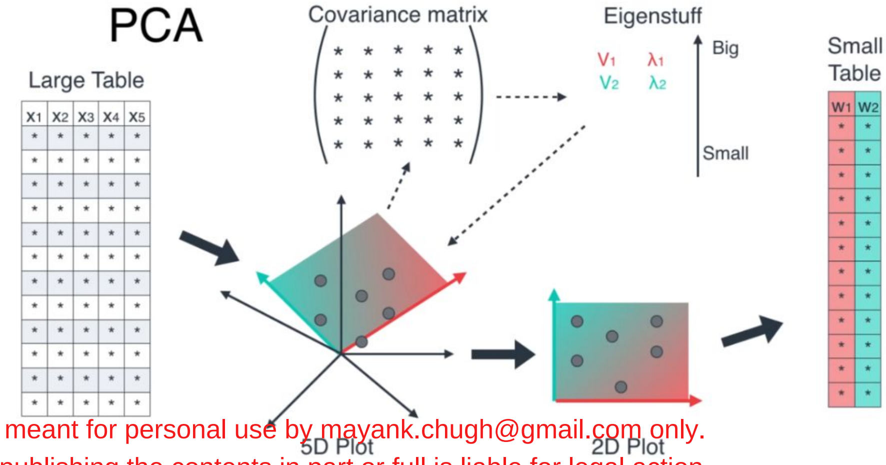
Feature selection:

Feature extraction:

- principal component analysis (PCA),
- linear discriminant analysis (LDA),
- Kernel PCA (K-PCA)
- quadratic discriminant analysis (QCA).

Dimensionality Reduction Most Commonly used Techniques

Principal component analysis performs orthogonal transformations to convert an observation of correlated characteristics into a set of linearly correlated features. The newly changed characteristics are termed ‘principal components’. This statistical method is a key data analysis and predictive modelling technique



Dimensionality Reduction Example

- Digital libraries
- social media content
- Emails
- ecommerce data

Data Governance

Data Governance

- Data governance is everything you do to ensure data is secure, private, accurate, available, and usable. It includes the actions people must take, the processes they must follow, and the technology that supports them throughout the data life cycle.
- Data governance strategies can improve:
  - Data quality
  - Data security
  - Data integrity

Data Governance - Best Practices

1. Develop a data governance strategy
2. Establish data ownership and accountability
3. Implement data catalogs and metadata management
4. Adopt data privacy by design
5. Automate data governance processes
6. Monitor and audit

Strategies of Data Governance

- Creating a data governance strategy that defines your organization and data stewards goals, roles, responsibilities, and processes can help provide a clear roadmap for effective data management in machine learning applications.
- A data governance strategy defines how data is named, stored, processed, and shared. Instead of data being a byproduct of your applications, it becomes a vital company asset. The strategy defines how data will be used efficiently in an organization.
- Typically, data governance departments are handled by a governance manager, governance committee and a team of data stewards and custodians.'

Data Ownership and Accountability

Establish data ownership and accountability

Clearly defining data ownership and assigning responsibilities for data quality, privacy, and compliance can aid in the effective implementation of data governance policies.

Without a clear understanding of who owns the data and who is responsible for its management, organizations run the risk of data breaches, misuse of data, and regulatory non-compliance.

Tips for establishing data ownership and responsibility:

- Clearly communicate the importance of data ownership and responsibility to all employees.
- Regularly review and update data governance policies to adapt to changing regulations and business needs.
- Encourage cross-department collaboration to ensure a holistic approach to data management.
- Invest in data governance tools and technologies to streamline data management processes.

Data Catalog and Metadata Management

Creating a data catalog and maintaining metadata about datasets used in machine learning applications can aid in:

- Understanding data lineage
- Improving data discoverability
- Preserving data quality

A data catalog is a record of an organization’s existing data. It is a library where an organization's’ data is indexed, organized and stored. Most data catalogs contain data sources, data usage information, and data lineage that describes the origin of the data and how it changed to its final form. With a data catalog, organizations can centralize information so that they can identify what data they have, distinguish data based on its quality and source

Data Privacy and Design

Integrating data privacy and security considerations into the due process and design of ML applications and processes can aid in the proactive management of potential risks and compliance with data protection regulations.

Principle 1: Proactive, not Reactive; Preventative, not Remedial

Principle 2: Privacy as a Default Setting

Principle 3: Privacy Embedded into Design

Principle 4:Positive-Sum, not Zero-Sum

Principle 5: End-to-End Security – Full Data Lifecycle Protection

Principle 6: Visibility and Transparency

Principle 7: Respect User Privacy- Keep it User-Centric

Automate, Monitor and Audit

Automate data governance processes

Data governance tasks such as data validation, cleansing, and enrichment can be automated to improve efficiency and maintain high data quality standards.

Monitor and audit

Monitoring and auditing data governance processes on a regular basis can help:

- Identify potential issues
- Maintain data quality
- Ensure compliance with applicable regulations

Using data fabric tools can be especially useful in monitoring and auditing data governance.

Power Ahead!