Data Exploration

HI 743

Ryan Gallagher

Department of Health Informatics and Administration
Zilber College of Public Health
University of Wisconsin - Milwaukee

February 13, 2025

Overview

Data Exploration

- **Data exploration** is the initial step in analyzing a dataset.
  - It involves summarizing key characteristics, detecting anomalies, and understanding distributions.
- Helps identify potential **data quality issues** before model building.
- Essential for effective **feature selection** and **data preprocessing**.

Relationship to CRISP-DM Methodology

- Falls under the **Data Undestanding** & **Data Preparation**.
- A critical step. Poor data quality at this stage leads to unreliable models.

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Goals of Data Exploration

- Understand the **structure** of the dataset:
  - Number of features (columns) and instances / observations(rows).
  - Types of variables: categorical vs. continuous.
  - Summary statistics: mean, median, standard deviation, missing values, etc..
- **Detect potential issues**:
  - Missing or inconsistent values.
  - Outliers or extreme values.
  - Data entry errors, duplicate records.
- Guide data **preprocessing decisions**:
  - Whether normalization or standardization is needed.
  - Features selection & alignment with objective.

The Data Quality Report

- A structured summary of dataset characteristics.
- Helps identify **inconsistencies, missing data, and errors** before modeling.
- Serves as a **documentation tool** for tracking dataset changes.
- Provides insights into whether additional **data cleaning or preprocessing** is needed.

Components of a Data Quality Report

- **Tabular Summaries for Features**
  - Overview of numerical and categorical variables.
  - Summary statistics:
    - **Continuous variables**: mean, median, standard deviation, min, max, quartiles.
    - **Categorical variables**: unique values, mode, frequency distribution.

Statistical Measures for Data Quality

- **Missing values**: Count and percentage of missing data.
- **Cardinality**: Number of unique values in categorical features.
- **Outliers**: Extreme values outside expected ranges.
- **Correlations**: Detecting redundant features.

Data Visualizations for Exploration

- **Histograms**: Visualize feature distributions.
- **Box Plots**: Identify outliers and spread of data.
- **Bar Charts**: Categorical feature distributions.
- **Scatter Plots**: Detect relationships between numerical variables.

Case Study: Motor Insurance Fraud Detection

- **Example scenario**: Predict fraudulent motor insurance claims.
- **Initial data review**:
  - Identify missing or inconsistent claim details.
  - Assess frequency distributions of claim types.
  - Detect patterns in high-claim amounts and fraudulent cases.
- **Goal**: Guide data preprocessing for better fraud detection models.

Case Study: Motor Insurance Fraud Detection

ID	Type	Inc.	Marital Status	Num. Clmnts.	Injury Type	Hospital Stay	Claim Amt.	Total Claimed	Num Claims	Num. Soft Tiss.	% Soft Tiss.	Claim Amt. Rcvd.	Fraud Flag
1	ci	0		2	soft tissue	no	1,625	3,250	2	2	1.0	0	1
2	ci	0		2	back	yes	15,028	60,112	1		0	15,028	0
3	ci	54,613	married	1	broken limb	no	-99,999	0	0	0	0	572	0
4	ci	0		4	broken limb	yes	5,097	11,661	1	1	1.0	7,864	0
5	ci	0		4	soft tissue	no	8,869	0	0	0	0	0	1
			⋮			⋮					⋮		
300	ci	0		2	broken limb	no	2,244	0	0	0	0	2,244	0
301	ci	0		1	broken limb	no	1,627	92,283	3	0	0	1,627	0
302	ci	0		3	serious	yes	270,200	0	0	0	0	270,200	0
303	ci	0		1	soft tissue	no	7,668	92,806	3	0	0	7,668	0
304	ci	46,365	married	1	back	no	3,217	0	0		0	1,653	0
			⋮			⋮					⋮		
458	ci	48,176	married	3	soft tissue	yes	4,653	8,203	1	0	0	4,653	0
459	ci	0		1	soft tissue	yes	881	51,245	3	0	0	0	1
460	ci	0		3	back	no	8,688	729,792	56	5	0.08	8,688	0
461	ci	47,371	divorced	1	broken limb	yes	3,194	11,668	1	0	0	3,194	0
462	ci	0		1	soft tissue	no	6,821	0	0	0	0	0	1
			⋮			⋮					⋮		
496	ci	0		1	soft tissue	no	2,118	0	0	0	0	0	1
497	ci	29,280	married	4	broken limb	yes	3,199	0	0	0	0	0	1
498	ci	0		1	broken limb	yes	32,469	0	0	0	0	16,763	0
499	ci	46,683	married	1	broken limb	no	179,448	0	0		0	179,448	0
500	ci	0		1	broken limb	no	8,259	0	0	0	0	0	1

Figure: Portion of Motor Insurance Claim Data

Case Study: Motor Insurance Fraud Detection

(a) Continuous Features

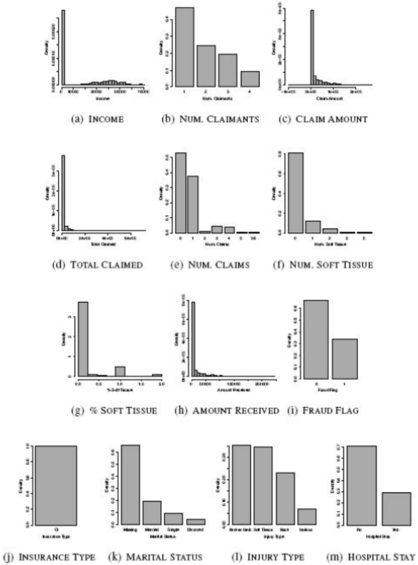
Feature	Count	% Miss.	Card.	Min	1^{st} Qrt.	Mean	Median	3^{rd} Qrt.	Max	Std. Dev.
INCOME	500	0.0	171	0.0	0.0	13,740.0	0.0	33,918.5	71,284.0	20,081.5
NUM. CLAIMANTS	500	0.0	4	1.0	1.0	1.9	2	3.0	4.0	1.0
CLAIM AMOUNT	500	0.0	493	-99,999	3,322.3	16,373.2	5,663.0	12,245.5	270,200.0	29,426.3
TOTAL CLAIMED	500	0.0	235	0.0	0.0	9,597.2	0.0	11,282.8	729,792.0	35,655.7
NUM. CLAIMS	500	0.0	7	0.0	0.0	0.8	0.0	1.0	56.0	2.7
NUM. SOFT TISSUE	500	2.0	6	0.0	0.0	0.2	0.0	0.0	5.0	0.6
% SOFT TISSUE	500	0.0	9	0.0	0.0	0.2	0.0	0.0	2.0	0.4
AMOUNT RECEIVED	500	0.0	329	0.0	0.0	13,051.9	3,253.5	8,191.8	295,303.0	30,547.2
FRAUD FLAG	500	0.0	2	0.0	0.0	0.3	0.0	1.0	1.0	0.5

(b) Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2^{nd} Mode	2^{nd} Mode Freq.	2^{nd} Mode %
INSURANCE TYPE	500	0.0	1	ci	500	1.0	–	–	–
MARITAL STATUS	500	61.2	4	married	99	51.0	single	48	24.7
INJURY TYPE	500	0.0	4	broken limb	177	35.4	soft tissue	172	34.4
HOSPITAL STAY	500	0.0	2	no	354	70.8	yes	146	29.2

Figure: Data Quality Report for Motor Insurance Claim Data

Case Study: Motor Insurance Fraud Detection



(a) Income (b) Num. Claimants (c) Claim Amount

(d) Total Claimed (e) Num. Claims (f) Num. Soft Tissue

(g) % Soft Tissue (h) Amount Received (i) Fraud Flag

(j) Insurance Type (k) Marital Status (l) Injury Type (m) Hospital Stay

Understanding and Exploring the Data

- Identify the types and distributions of features in the dataset.
- Differentiate between **continuous** and **categorical** features.
- Assess skewness, central tendency, and variability.
- Use statistical summaries and visual tools (histograms, bar charts, box plots) to examine data characteristics.

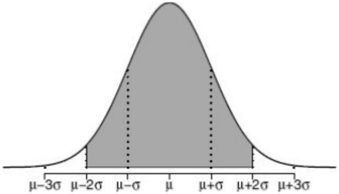
Feature Types and Statistical Measures

- **Continuous Features**:
  - Examples: Age, Income, Temperature.
  - Use histograms, box plots, and summary statistics (mean, median, standard deviation) for insights.
- **Categorical Features**:
  - Examples: Gender, Product Category.
  - Use frequency tables, bar plots, and pie charts to explore distributions.
  - Detect class imbalances that may impact predictive modeling.

The Normal Distribution: Overview

- Defined by two parameters:
  - **Mean (μ)**: The central location of the distribution.
  - **Standard deviation (σ)**: Measures the spread of data.

- Characterized by its symmetric, bell-shaped curve, and common in natural phenomena such as heights, test scores, and financial returns.



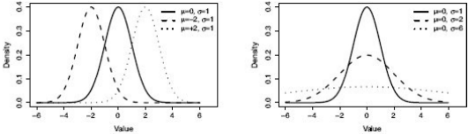
The Normal Distribution: Properties

- **Empirical Rule**:
  - 68% of data falls within 1σ of the mean.
  - 95% within 2σ.
  - 99.7% within 3σ.
- **Statistical Applications**:
  - Basis for parametric statistical tests (e.g., t-tests, ANOVA).
  - Used in probabilistic modeling and hypothesis testing.
- **Central Limit Theorem (CLT)**:
  - Explains why sample means tend to be normally distributed.
  - Justifies using normal-based methods in inferential statistics ($n \geq 30$).

Detecting and Handling Deviations from Normality

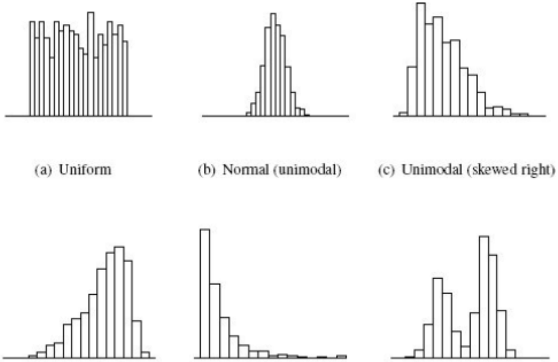
- **Identifying Non-Normality**:
  - Visual methods: Histograms, QQ-plots.
  - Statistical tests: Shapiro-Wilk, Kolmogorov-Smirnov.
- **Implications of Non-Normal Data**:
  - May violate assumptions of statistical models.
  - Can impact performance of machine learning algorithms.
- **Transformations to Normalize Data**:
  - Log transformation for right-skewed data.
  - Box-Cox transformation for general adjustments.
  - Standardization (z-score normalization) for mean-centering.

Different Distributions



(a) Different means

(b) Different standard deviations



(a) Uniform

(b) Normal (unimodal)

(c) Unimodal (skewed right)

(d) Unimodal (skewed left)

(e) Exponential

(f) Multimodal

Identifying Data Quality Issues

- Poor data quality can negatively impact model performance and insights.
- Key data quality issues include:
  - Missing values
  - Irregular cardinality
  - Outliers and anomalies
- Addressing these issues is crucial for reliable data-driven decision making.

The structure of a data quality plan.

Feature	Data Quality Issue	Potential Handling Strategies
_____	___________________	_________________________________
_____	___________________	_________________________________
_____	___________________	_________________________________
_____	___________________	_________________________________

Missing Values: Causes and Consequences

- **Causes of Missing Data**:
  - Human error in data entry.
  - Data corruption or loss during processing.
  - Non-response in surveys or experiments.
- **Types of Missing Data**:
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Missing Not at Random (MNAR)
- **Consequences**:
  - Reduces data usability.
  - Can bias analytical results if not handled properly.

Irregular Cardinality in Categorical Variables

- **Definition**: Cardinality refers to the number of unique values a categorical feature can take.
- **Potential Issues**:
  - High cardinality: Many unique values can cause overfitting and increase model complexity.
  - Low cardinality: May indicate redundancy or poor feature utility.
- **Detection Methods**:
  - Frequency distribution analysis.
  - Visualizing unique values with bar plots.

Outliers and Anomalies

- **Definition**: Outliers are extreme values that deviate significantly from the rest of the data.
- **Causes**:
  - Data entry errors.
  - Genuine rare events.
  - Sensor or measurement errors.
- **Detection Methods**:
  - Statistical techniques: Z-scores, IQR method.
  - Visualization techniques: Box plots, scatter plots.

Handling Data Quality Issues

- Addressing data quality issues improves model reliability and accuracy.
- Common techniques include:
  - Imputing missing values.
  - Managing irregular cardinality in categorical features.
  - Handling outliers effectively.
- How could prediction be used to fill missing fields for imputation? Could correlation help with this?

Handling Missing Values

- **Strategies for Handling Missing Data**:
  - **Deletion**: Remove rows or columns with excessive missing values.
  - **Imputation**:
    - Mean, median, or mode substitution.
    - Predictive modeling (e.g., KNN imputation, regression imputation).
  - **Indicator Variable Method**: Add a new feature indicating missingness.
- Consider the missing data mechanism (MCAR, MAR, MNAR) before applying a strategy.

Managing Irregular Cardinality

- **High Cardinality Issues**:
  - Increases computational complexity and risk of overfitting.
  - Common in features like zip codes, product IDs, or names.
- **Techniques for Handling High Cardinality**:
  - **Grouping**: Merge rare categories into an "Other" category.
  - **Encoding**:
    - One-hot encoding (useful for small cardinality features).
    - Target encoding or frequency encoding for high cardinality features.

Handling Outliers and Anomalies

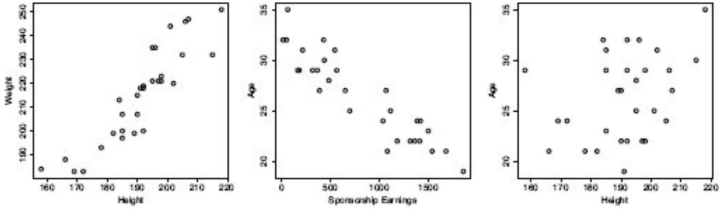
- **Identifying Outliers**:
  - Statistical methods: Z-score, IQR method.
  - Visual methods: Box plots, scatter plots.
- **Strategies for Handling Outliers**:
  - **Truncation**: Cap values within a predefined range.
  - **Transformation**: Apply log transformation to reduce skewness.
  - **Model-based approaches**: Use robust algorithms less sensitive to outliers (e.g., decision trees, random forests).
- Choose the appropriate method based on whether the outliers are data errors or meaningful anomalies.

Advanced Data Exploration

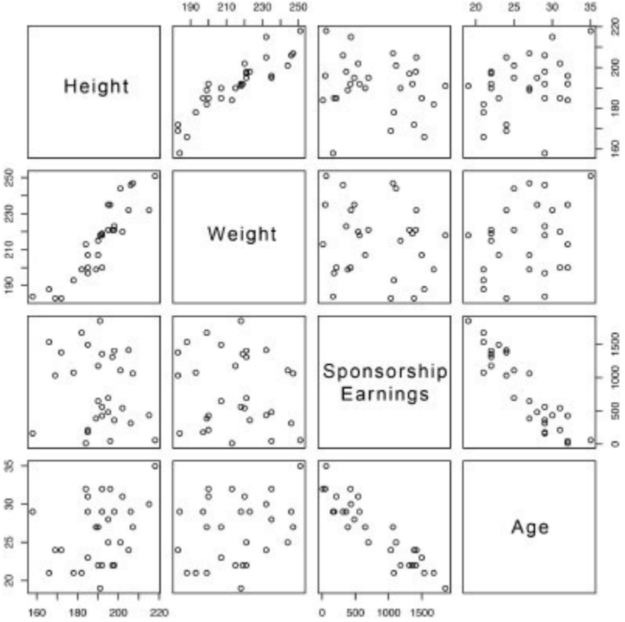
- Data exploration goes beyond basic summaries to uncover deeper patterns that impact model performance.
- Traditional summary statistics provide useful insights, but advanced techniques help:
  - Detect complex relationships between variables.
  - Identify hidden structures in the dataset.
  - Improve feature selection and engineering decisions.
- Why it matters:
  - Machine learning models rely on clean, well-structured input features.
  - Poorly understood data can lead to bias, overfitting, and misleading conclusions.

Visualizing Relationships Between Features

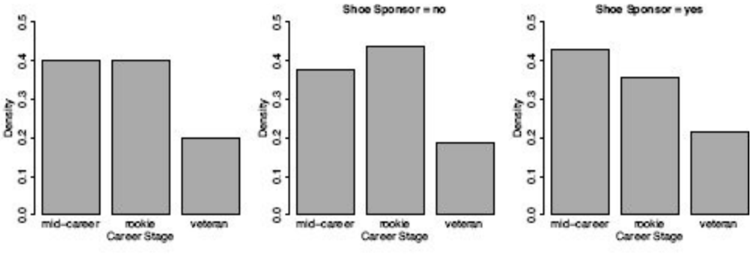
- Graphical representations reveal hidden patterns and dependencies.
- Common visualization techniques:
  - **Scatter Plots**: Show relationships between continuous variables.
  - **Pair Plots**: Matrix of scatter plots for multiple features.
  - **Box Plots**: Compare distributions across categorical groups.
  - **Heatmaps**: Visualize correlation matrices.



(a) (b) (c)



Height
Weight
Sponsorship
Earnings
Age
190 210 230 250
20 25 30 35
160 180 200 220
0 500 1000 1500



(a) Career Stage and Shoe Sponsor

(b) Position and Shoe Sponsor

Measuring Covariance and Correlation

- **Covariance**:
  - Measures the direction of a linear relationship between two variables.
  - A positive covariance indicates both variables increase together.
  - A negative covariance indicates one increases while the other decreases.
- **Correlation**:
  - Standardized measure of the strength and direction of a relationship.
  - Ranges from -1 (strong negative) to $+1$ (strong positive).
  - Pearson, Spearman, and Kendall correlation methods.

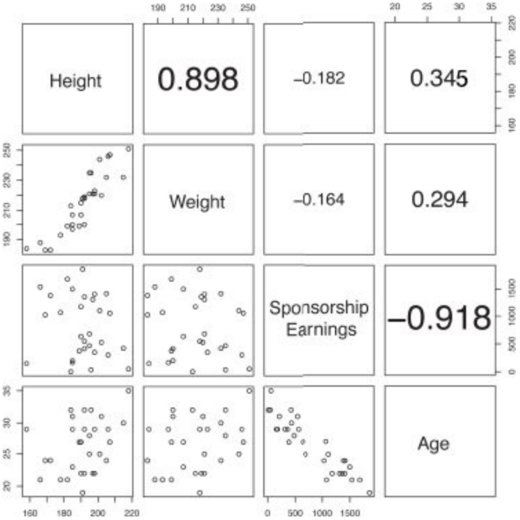


Figure: A scatter plot matrix showing scatter plots of the continuous features from the professional basketball team dataset with correlation coefficients included.

Identifying Multicollinearity

- Multicollinearity occurs when independent variables are highly correlated.
- Issues caused by multicollinearity:
  - Inflates variance in regression coefficients.
  - Reduces model interpretability.
- Detection Methods:
  - Variance Inflation Factor (VIF): Higher VIF values indicate multicollinearity.
  - Eigenvalue decomposition of correlation matrix.
- Handling multicollinearity:
  - Removing highly correlated features.
  - Principal Component Analysis (PCA) for dimensionality reduction.

Lab: Introduction to Git and GitHub

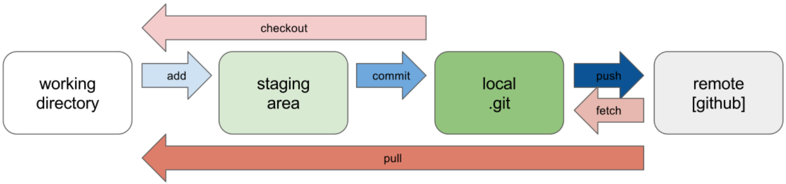
- **Git**: A command-line version control system for tracking changes in files. "Saves Progress" in projects, and allows for version roll-backs.
- **GitHub**: A cloud-based platform for hosting Git repositories. Stores code for personal or public sharing. The standard for sharing projects and code.
- Benefits of Git/GitHub:
    - Enables collaboration on projects by managing user privleges by project.
    - Provides a detailed history of changes.
    - Facilitates code backup and versioning.

Setting Up a Local Repository with GitHub

Let's Set-Up Git/Github for our Projects.

Basic Git Command-Line Workflow

- A simple workflow for tracking changes using Git:
    1. **Initialize a repository**: ‘git init‘
    2. **Check the status**: ‘git status‘
    3. **Stage changes**: ‘git add <file>‘
    4. **Commit changes**: ‘git commit -m ”Commit message”‘
    5. **View commit history**: ‘git log‘
    6. **Push to remote repository**: ‘git push‘



Lab: Data Exploration & Tidyverse

- Worksheet