DATASCI 510 Lesson 4

EDA of Numeric Tables

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Reflection

Statistics is not about numbers, it's about understanding the underlying processes

John Tukey, 1915-2000

Lesson 4 Agenda

- Announcements
- Outlier Removal and Replacement Review
- Break
- Preliminary EDA and Pandas Indexing: Lesson-04_student.ipynb (part 1)
- EDA (Slides)
- Break
- EDA and Visualizations: Lesson-04_b_student.ipynb (part 2)
- Basic EDA and Data Cleaning: Lesson-04_b_student.ipynb (part 3)
- Break
- Machine Learning (MachineLearningDataFlowAndSchema.pdf)
- Data as Sparse Multi-Dimensional Matrices (DataAsASparseMultiDMatrix.pdf)

Announcements

- **Create Study Groups**
- **Use of Ed Discussion will be graded!**
  - Please submit at least two posts each week
  - The topics must be related to data science
  - Best topics pertain to the lecture or homework

Outlier Removal and Replacement Review

Remove and Replace Patterns Review

Outlier/Missing Value Replacement Rules:

- To replace the outliers or missing values, the left-hand side of the assignment must use the BadFlag Boolean indexer because we replace bad values not good values: **Variable[BadFlag] =**
- To determine a mean or median the right-hand side of the assignment determines a good average.
  - That means, the arithmetic mean should be determined from values specified by a GoodFlag boolean indexer: = mean(Variable[GoodFlag])
  - The median can be determined using all values because outliers have little effect on medians: = median(Variable)

Outlier Removal Rules:

- To remove the outliers, the left hand side of the assignment does not need a Boolean indexer because we overwrite the variable: **Variable =**
- To remove the outliers, the right hand side of the assignment are the remaining good values: **= Variable[GoodFlag]**

Break

Preliminary EDA and Pandas Indexing

Open: Lesson_04_student.ipynb

EDA and Data Visualization

Review: Steps in the Data Science Process

1. Data Generation and Persistence
2. Extract Transform Load (ETL)
3. Data Preparation
   - Data Cleaning
   - Exploratory Data Analysis (EDA)
   - Data Shaping
4. Create a Model (Machine Learning)
   - Feature Engineering
   - Training
   - Model Selection
5. Use the Model (Predicting, Scoring)

Exploratory Data Analysis (EDA)

- Explore Content
  - Run checks on the data for structure, volume, and content of data. Data scientists refer to "garbage in, garbage out" when the content of the data does not support a machine learning algorithm. In contrast, engineers and developers only need to consider the structure and volume of the data.
- The first goal is to know your data and fix any irregularities
  - Adjust Data Types: The data type should match the content of the data. Data types provide context and reveal the nature of the data and some of its structure. Data scientists often must change them to fit the planned analysis.
  - Make sure missing values are properly flagged.
    - Impute missing values, correct outliers
    - Remove records and attributes with outliers or missing values
  - Make sure the distribution of columns match what we expect
- The second goal is to understand the data
  - Use data visualizations to describe trends, distributions, relationships, etc.
  - Statistical descriptions (Descriptive Analytics) like histograms, standard deviation and average
  - Check for statistical correlations between variables (attributes, columns, features)

Table Column Types: EDA

- data scientists are usually not in charge of storing and managing data
- data is often read from a variety of sources with varying schema structure
- actual column type may be inappropriate
- apparent column type may be misleading – never presume a column type
- desired column type depends on how it's used by us in our analysis

Table Column Types: Data Preparation

- a timestamp column by default inherits the object type because it contains non-numeric characters: This column needs to be converted to datetime explicitly (using pd.to_datetime)
- a numeric column with non-numeric entries is loaded as a type object and needs to be explicitly converted to a numeric data type
- a categorical is often encoded using integers, which means pandas treats it a numerical column (int) unless the column is converted to object or category

Visualizations: Histograms

- Histograms can represent a distribution
- Histograms provide a quick way to view the essentials of a single numeric variable
- Histograms can identify/highlight outliers
- Outliers can make the rest of the histogram hard to interpret
- https://en.wikipedia.org/wiki/Histogram

Visualizations: Scatterplot Matrix

- A single scatter plot helps find a relationship between two numerical variables
- A scatter plot matrix is a quick way to explore all numeric columns at once
- A scatter plot matrix may take a long time to render. Use a small sample of the data if data size is large
- We may need to transform columns with extreme values, so plots aren't skewed, by trimming or using a log transform
- https://en.wikipedia.org/wiki/Scatter_plot
- https://en.wikipedia.org/wiki/Scagnostics

Visualizations: Correlation Matrix

- Correlation Matrix can provide a quick overview of linear correlations and redundant variables
- https://en.wikipedia.org/wiki/Correlation
- https://muthu.co/understanding-correlations-and-correlation-matrix/

Break

EDA and Data Visualization

Continue: Lesson_04_student.ipynb

EDA and Data Cleaning

Continue: Lesson_04_student.ipynb

Break

Machine Learning

Open: MachineLearningDataFlowAndSchema.pdf

Data as Sparse Multi-Dimensional Matrices

Open: DataAsASparseMultiDMatrix.pdf