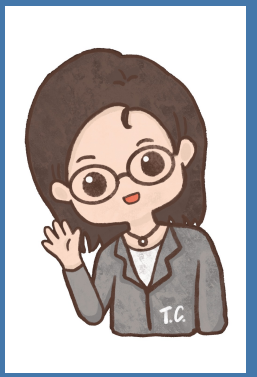
MTH 208 Exploratory Data Analysis

Lesson 01: Introduction to EDA

Ying-Ju Tessa Chen, PhD

Associate Professor
Department of Mathematics
University of Dayton

@ying-ju
ying-ju
ychen4@udayton.edu



Learning Objectives

- What is EDA, and why is it important?
- Overview of the data analysis pipeline
- Differences between EDA, confirmatory analysis, and predictive modeling

What is EDA?

Exploratory Data Analysis (EDA) is a foundational approach in the field of data science, essential for analyzing, summarizing, and understanding data sets. At its core, EDA is about uncovering the underlying structure, patterns, anomalies, and insights in data, often before formal modeling commences. It is instrumental in identifying the most appropriate statistical techniques and methods for further analysis.

Why is EDA important?

EDA combines a variety of statistical techniques with visual methods to facilitate a deeper understanding of the data. This approach allows data scientists and analysts to:

- **Discover Patterns**: Uncover underlying patterns and relationships hidden within the data.This is essential for understanding the data set before applying complex models or statistical techniques.
- **Spot Anomalies**: Identify outliers or unusual data points that might indicate errors in data collection or potential areas of interest for further investigation.
- **Test Hypotheses**: Formulate and test hypotheses about the data, paving the way for more in-depth analysis.
- **Check Assumptions**: Verify the assumptions underlying - statistical analyses, ensuring their validity and appropriateness.

Why is EDA important (Continue)?

Exploratory Data Analysis (EDA) is distinguished by its flexible and exploratory nature. Unlike rigid, hypothesis-driven methods, EDA encourages open-ended exploration and intuitive data understanding, making it an essential first step in data analysis and subsequent statistical testing and modeling.

The concept and practice of EDA were significantly shaped and popularized by American mathematician **John Tukey** in the 1970s. Tukey's work in this area was groundbreaking; he brought together a range of methods and an overarching philosophy that emphasized the exploratory aspect of analyzing data. It's important to note, however, that while Tukey played a pivotal role in organizing and advocating for these methods, the practice of exploratory analysis of data has been a part of statistics since its inception.

Today, EDA remains a critical tool in the data discovery process. Its techniques, continually evolving with advances in computing and data visualization technologies, are fundamental to data science, enabling practitioners to make informed, data-driven decisions.

John Wilder Tukey
(June 16, 1915 – July 26, 2000)

Four principles of EDA

We will introduce various methods for exploring data, each embodying the core characteristics that define the philosophy of EDA. Tukey established key principles that guide the exploratory process. These principles are fundamental to understanding data and are integral to the EDA methodology.

- **Revelation**: This principle is about uncovering the underlying structure of the data. The idea is to reveal hidden patterns, trends, and relationships that are not immediately obvious. This often involves visualizing the data in various ways to uncover these insights.
- **Resistance**: This refers to the robustness of the analysis methods. The techniques used in EDA should be resistant to outliers and not be overly influenced by a few data points. This ensures that the analysis gives a true representation of the data as a whole.
- **Reexpression**: This principle involves transforming or reexpressing data to make it easier to understand or reveal hidden aspects. This can include applying logarithmic transformations, scaling, or other mathematical manipulations to make patterns more evident or data more amenable to analysis.
- **Residuals**: The focus here is on examining the residuals - the differences between observed and predicted values in a model. Analyzing residuals can reveal whether the chosen model is appropriate or if there are patterns in the data that the model isn't capturing.

$$\text{Residual} = \text{Observed Value (Data)} - \text{Predicted Values (Fitted by a model)}$$

Case Study

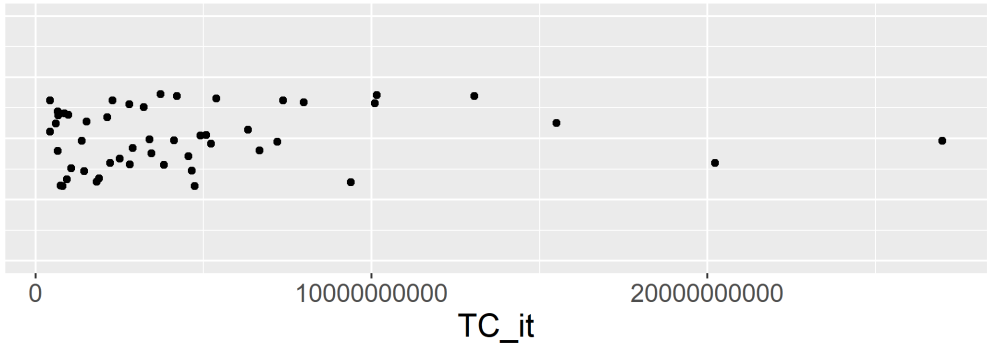
We will study data obesity_cost_full from R package obcost. This data set gives the name of each state in the United States, a Year variable from 1996 to 2018, and some variables related to medical costs on obesity. We will focus on the average total economic cost of obesity by state from 1996 to 2018. The following table shows a few rows of the data with the variables of interest.

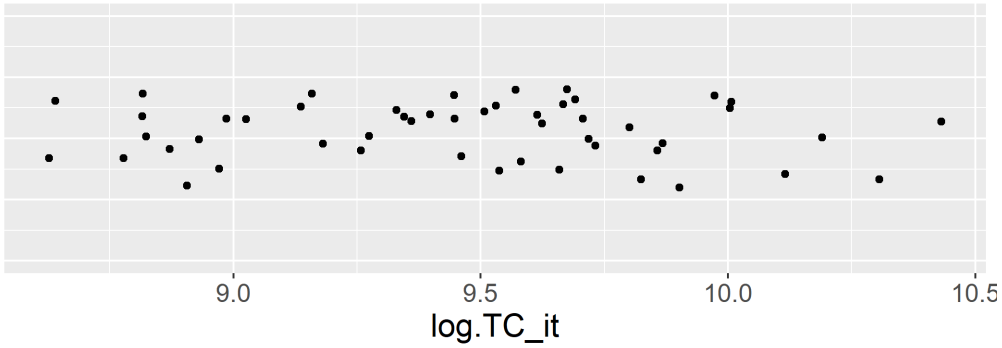
State	TC_it
Alabama	3723234610
Alaska	654101733
Arizona	4206639898
Arkansas	2137036028
California	26992318816

Here are some questions that might interest us.

- What states have contributed to the most average medical cost on obesity?
- What is a typical number of medical cost on obesity?
- Is Ohio contributing more or less medical cost than Michigan?
- Are southeastern states contributing less medical cost than western states?
- In relation to population size, which state contributes the most average medical costs related to obesity?

Case Study (Continue)





We can see that the majority of the average total prices are between 0 and 10 billions, with only a few states having substantial average expenses. This data is heavily right-skewed.

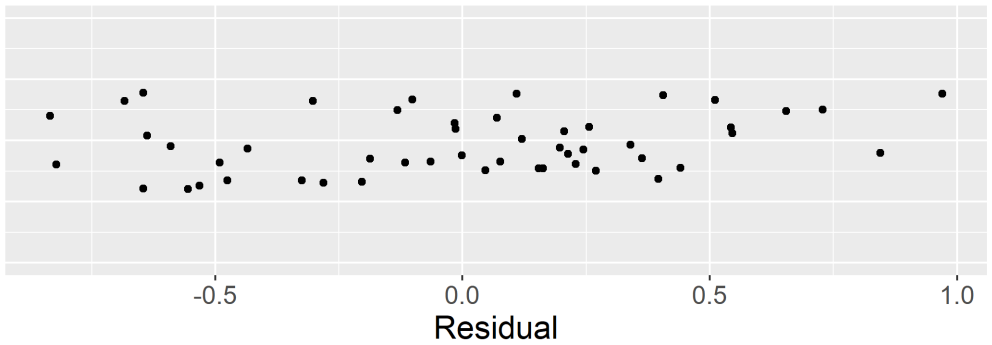
Consider the variable $\log_{10}(\text{TC\_it})$. We can improve the presentation of the data by reexpressing the average cost by taking logs.

This graphical display is more appropriate for viewing the data. The log average costs are evenly spread out between -0.15 and 0.25 and we can see more interesting structure in the data.

Case Study (Continue)

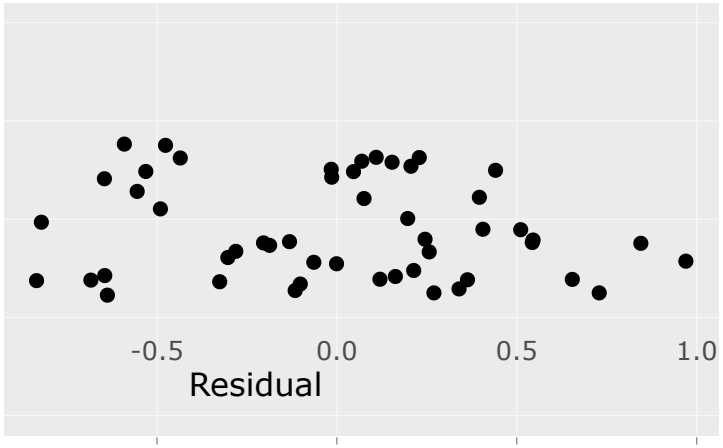
A typical log average total cost can be seen to be about 9.46 (How did I estimate a typical value of the data here?). We can summarize the data by the typical value of 9.46. Then, the residuals can be computed by

$$\text{Residual} = \log(\text{TC\_it}) - 9.46.$$



On the log scale, two states have average total costs that are 0.75 less than the average of 9.46, while another two states exceed the average by the same margin. The residual plot shows the variation of these log average costs from the overall average.

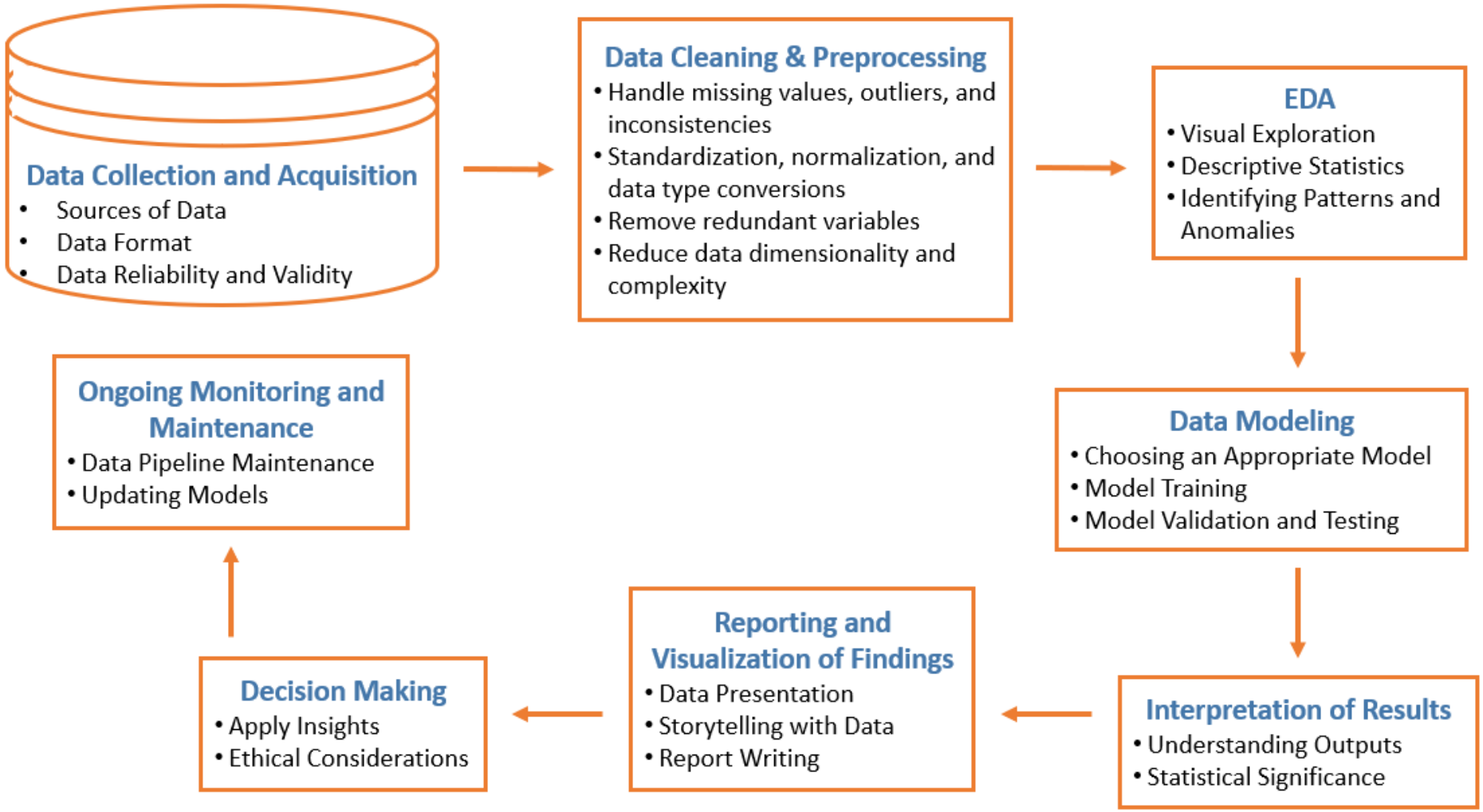
Case Study (Continue)



Recall the questions we might be interested:

- What states have contributed to the most average medical cost on obesity?
- What is a typical number of medical cost on obesity?
- Is Ohio contributing more or less medical cost than New York?
- Are southeastern states contributing less medical cost than western states?
- In relation to population size, which state contributes the most average medical costs related to obesity?

Overview of the data analysis pipeline



Differences between EDA, confirmatory analysis, and predictive modeling

To understand the various facets of data analysis, it is crucial to know the differences between Exploratory Data Analysis (EDA), Confirmatory Data Analysis, and Predictive Modeling. Each has its own objectives, methodologies, and implications.

- **Exploratory Data Analysis**
  - Objective: To uncover patterns, anomalies, relationships, and trends in data without any prior hypotheses.
  - Approach: Primarily visual and intuitive; involves summarizing the main characteristics of a dataset through visualization and descriptive statistics.
  - Outcome: Generates hypotheses, insights, and a deeper understanding of data's underlying structure and characteristics.
  - Key Characteristics: Flexible, open-ended, often the first step in data analysis.

Differences between EDA, confirmatory analysis, and predictive modeling (Continue)

- **Confirmatory Data Analysis**
  - Objective: To test hypotheses or models that were formulated prior to data analysis.
  - Approach: Uses statistical tools to verify or refute predefined hypotheses. This analysis is usually more structured than EDA and follows a more rigid methodology.
  - Outcome: Provides evidence to support or reject hypotheses, often with a focus on statistical significance.
  - Key Characteristics: Hypothesis-driven, relies on statistical inference, more formal and structured.

Differences between EDA, confirmatory analysis, and predictive modeling (Continue)

- **Predictive Modeling**
  - Objective: To create models that can predict future outcomes based on historical data.
  - Approach: Involves selecting and training models (like regression, classification, or machine learning algorithms) on existing data to forecast unknown or future values.
  - Outcome: Produces a model that can be used for forecasting, along with metrics to gauge its accuracy and reliability.
  - Key Characteristics: Forward-looking, uses machine learning and statistical techniques, focused on prediction accuracy.

Differences between EDA, confirmatory analysis, and predictive modeling (Continue)

- **Scope and Focus**: EDA is about exploring and understanding data without specific aims; confirmatory analysis tests specific theories or hypotheses; predictive modeling forecasts future events based on historical patterns.
- **Methodology**: EDA is more about visualization and intuition, confirmatory analysis relies on statistical tests, and predictive modeling focuses on algorithmic and statistical methods for forecasting.
- **Sequence in Data Analysis**: Often, EDA precedes confirmatory analysis and predictive modeling. Insights gained from EDA can inform the development of hypotheses for confirmatory analysis or feature selection in predictive modeling.
- **Nature of Conclusions**: EDA provides insights and raises questions, confirmatory analysis offers statistical evidence, and predictive modeling delivers actionable predictions.

It is important to note that these methods are not mutually exclusive but are often used in conjunction as part of a comprehensive data analysis strategy. Understanding their differences helps in selecting the appropriate approach based on the objectives of a specific data analysis task.

References

The lectures of this course are based on the ideas from the following references.

- Exploratory Data Analysis by John W. Tukey
- A Course in Exploratory Data Analysis by Jim Albert
- The Visual Display of Quantitative Information by Edward R. Tufte
- Data Science for Business: what you need to know about data mining and data-analytic thinking by Foster Provost and Tom Fawcett
- Storytelling with Data: A Data Visualization Guide for Business Professionals by Cole Nussbaumer Knaflic