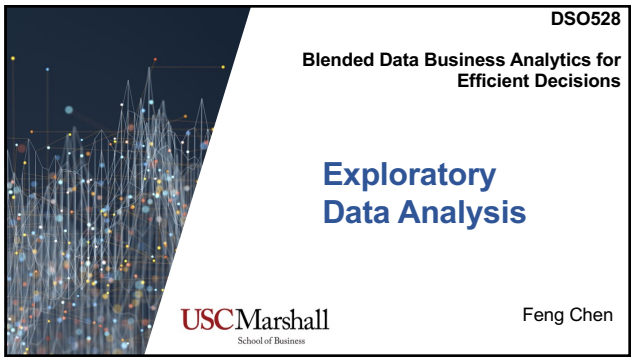


DSO528

Blended Data Business Analytics for Efficient Decisions

Exploratory Data Analysis

USC Marshall School of Business

Feng Chen

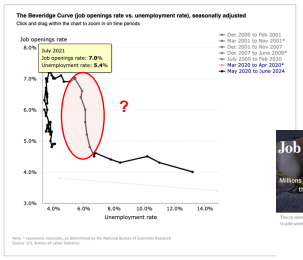
1

Learning Objectives

- Analytical thinking
- Describe variables in the data set
- Explain basic measures for qualitative and quantitative data
- Load and view data in JMP and Colab
- Develop basic measures different types of variables
- Build simple graphs to visualize data and explain findings

2

Analytical thinking



https://www.bls.gov/charts/job-openings-and-labor-turnover/b-beveridge-curve.htm

https://www.wsj.com/articles/job-openings-are-at-record-highs-why-arent-unemployed-americans-filling-them-11594302000

3

Analytics in a Nutshell

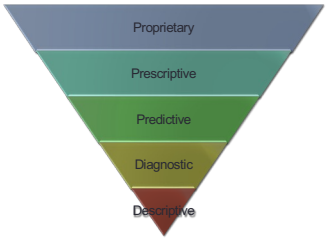
Proprietary – Building Models/Algorithms for Specific Domain Based Problems

Prescriptive – Recommending actions and strategies to optimize business processes and achieve desired outcomes.

Predictive - Forecasting future outcomes and trends based on models.

Diagnostic - Determining why something has happened by analyzing data to identify causes and correlations.

Descriptive - Understanding what has happened in the past and gaining insights into historical data patterns and trends.



4

Exploratory Data Analysis (EDA, Descriptive Analytics)

*"In statistics, **exploratory data analysis** (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing. Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments."*

John Tukey

https://en.wikipedia.org/wiki/Exploratory_data_analysis

5

Exploratory Data Analysis (EDA, Descriptive Analytics)

1. Understand the business context and project objectives.
2. Confirm the availability of data and recognize any limitations of data sources.
3. Propose treatments for missing values and errors, if applicable.
4. Clarify the definitions and explanations of all variables (pay attention to units).
5. Identify outliers and propose methods for handling them.
6. Visualize the data and document key findings.
7. Transform the data, if necessary.
8. Summarize findings and insights.

6

DSO528 7

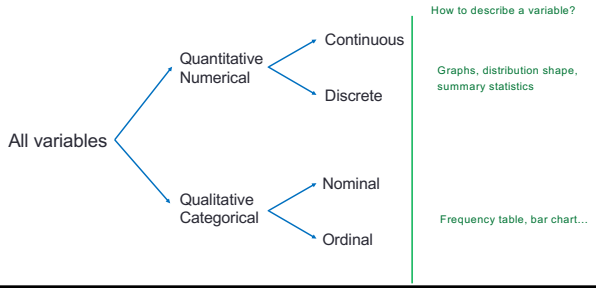
Basic measures in descriptive analysis

- Five-number summary
  Minimum, 25th percentile (first quantile), 50th percentile (second quantile or median), 75th percentile (third quantile), and the maximum.
- Interquartile range (IQR = $Q_3 - Q_1$)
  Measures how spread out the middle 50% of values are.
  IQR corresponds to the length of the box in the boxplot.
- Outliers (IQR method)
  Values in the dataset that fall far outside the range of "ordinary" value.
  For boxplot, values below $Q_1 - 1.5*IQR$ or above $Q_3+1.5*IQR$
- Distribution of a random variable
  How frequently different values of a variable occur.

7

DSO528 8

Variable types and measures



8

DSO528 9

Learning Objectives

- Analytical thinking
- Describe variables in the data set
- Explain basic measures for qualitative and quantitative data
- Load and view data in JMP and Colab
- Develop basic measures different types of variables
- Build simple graphs to visualize data and explain findings

9