

# COM 3590 Data Cleaning and Transformation

Adrian Caciula,  
Belfer Hall, Rm 1124  
adrian.caciula@yu.edu  
Office: 212-960-5440



Yeshiva University®

# Course Motivation

---

An article in the New York Times [1] reported that the whole process of data wrangling could account up to 80% of the time in the analysis cycle:

“Yet far too much handcrafted work — what data scientists call “data wrangling,” “data munging” and “data janitor work” — is still required. Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.”

- Give a few reasons for why data exploration is so time consuming.
- Are there ways to cut down on the time of data exploration?



# Course Motivation (cont.)

---

- Data Scientists can analyze big datasets to resolve some real problems using Machine Learning techniques.
  - These techniques are applied to huge amounts of information, to learn the relationships between its features.
- Machine Learning algorithms use all the values of the dataset. If we have a “**dirty**” dataset with a lot of mistakes and issues, these algorithms will not learn as effectively.
- **It’s therefore necessary to fix the issues first**



# Course Outcomes

---

## **By the end of this course, you'll be able to:**

- apply descriptive statistics to explore a data set
- use data visualization tools to understand and explain the characteristics of a data set
- write programs to clean data sets
- derive from existing data sets new data sets that are ready for analysis, via transformation, integration, and sampling



# Textbooks

---

Hellerstein, J. M. (2008). [Quantitative data cleaning for large databases](#). *United Nations Economic Commission for Europe (UNECE)*.

## Required:

- “Principles of Data Integration.” Doan, Halevy, and Ives.
- “Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.” 2<sup>nd</sup> Ed. McKinney.

## Recommended:

- “Practical Statistics for Data Scientists: 50 Essential Concepts.” Bruce, Bruce.
- “Data Wrangling with Python: Tips and Tools to Make Your Life Easier.” Kazil, Jarmul.
- “The Visual Display of Quantitative Information.” 2<sup>nd</sup> Ed. Tufte.



# Recommended Reading

---

1. Barnett V, Lewis T (1994) Outliers in statistical data, 3rd edn. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley, Chichester/New York [zbMATH](#), [Google Scholar](#)
2. Crown (2015) Data science report. <http://visit.crowdfunder.com/2015-data-scientist-report.html>
3. Dasu T, Johnson T (2003) Exploratory data mining and data cleaning, vol 479. Wiley, NY: [zbMATH](#), [CrossRef](#), [Google Scholar](#)
4. Data science report 2015: <http://visit.crowdfunder.com/2015-data-scientist-report.html>  
2018: [https://visit.figure-eight.com/WC-2018-Data-Scientist-Report\\_.html](https://visit.figure-eight.com/WC-2018-Data-Scientist-Report_.html)
5. Doan A, Halevy A, Ives Z (2012) Principles of data integration. Morgan Kaufmann, Waltham. [Google Scholar](#)
6. For Big-Data Scientists (2014) 'Janitor Work' Is Key Hurdle to Insights. <http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html? r=0&module=ArrowsNav&contentCollection=Technology&action=keypress&region=FixedLeft&pgtype=article><http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html? r=0&module=ArrowsNav&contentCollection=Technology&action=keypress&region=FixedLeft&pgtype=article> (The NY T article by Steve Lohr). [Google Scholar](#)
6. García S, Luengo J, Herrera F (2015) Data preprocessing in data mining. Springer, Cham. [CrossRef](#), [Google Scholar](#)
7. Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier, Burlington. [zbMATH](#), [Google Scholar](#)
8. Müller H, Freytag J-C (2005) Problems, methods, and challenges in comprehensive data cleansing. Professoren des Inst. Für Informatik, Berlin. [Google Scholar](#)
9. Pyle D (1999) Data preparation for data mining. Morgan Kaufmann, San Francisco, [Google Scholar](#)
10. Tukey JW (1977) Exploratory data analysis, pp 2–3. [Google Scholar](#)
11. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, Amsterdam/Boston. [zbMATH](#), [Google Scholar](#)



# Week 1 Objectives

---

- Introduction: Understand the importance of data cleaning. Garbage In, Garbage Out
  - Data Cleanup and Transformation
  - Dealing with Missing Values
  - Dealing with Outliers
  - Adding and Removing Variables
- Sources of errors in data and their telltale signs in data sets.



# Introduction

---

- Before data can be analyzed they must be organized into an appropriate form.
- Data preparation is typically an iterative process of manipulating raw data, which is often unstructured and messy, into a more structured and useful form that is ready for further analysis.
- The whole preparation process consists of a series of major activities (or tasks) including data profiling, cleansing, integration and transformation.





# Major Tasks in Data Preparation

---

- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results



# Data Quality: Definitions

---

## **Definition #1: “Fitness for Use”**

- *Degree to which data can be used for its intended purpose*

## **Definition #2: “Real-World Representation”**

- *Degree to which data accurately represents the real world*

**We can try to satisfy BOTH definitions!**



# Understanding Input Data

---

- As the saying goes: garbage in, garbage out.
- **High-quality input data is fundamental to producing reliable models and datasets.**
- If the input data used to build and implement them are bad—incomplete, outdated, biased or otherwise inaccurate—the resulting predictions or datasets have little chance of being reliable.
- **Data is subject to how, where, when, and from whom they were captured. Any of these aspects can be a source of bias or error.**



# Data Cleaning Checklist (Partial)

---

- Do character variables have valid values?
- Are numeric variables are within range?
- Are there missing values?
- Are there duplicate values?
- Are values unique for some variables (e.g., ID variables)?
- Are the dates valid?
- Do we need to combining multiple data files?

*More questions [here](#)*



**Q: At what stage(s) of Data Exploration would you address missing values in a data set?**

Select all that apply:

- Context understanding
- Data transformation
- Data clean-up
- Data reduction



# Tools for Data Cleanup

---

- OpenRefine (formerly Google Refine): <http://openrefine.org>
- Data Wrangler: <http://vis.stanford.edu/wrangler/>
- Many other open source and commercial tools  
<https://cloud.google.com/dataprep/>  
<http://www.datapreparator.com/>



# Data Transformation: Centering and Scaling

---

- **Data Transformation** ~ applying a mathematical function to each data value.
- The most common data transformation is **centering and scaling** of a single variable:
  - Calculate the “z-score” of each observed value
  - Generally improves numerical stability
  - Drawback: loss of interpretability
- Centering and scaling is often required or recommended for some modeling tools, such as clustering, principal component analysis, and neural networks.
- The main drawback is that the data becomes harder to interpret: data value after centering and scaling **measures the number of standard deviations between each data point and the mean**, and its uniqueness.



## **Why would one want to center and scale a set of data?**

- a. To make data easier to interpret
- b. To remove duplicates
- c. So multiple variables in the data set are on a common scale
- d. To make all data values positive





# Data Transformation

---

- Common transformations: log, square, square root, or inverse

Polynomial Transformations  
*(because they involve polynomial terms of the original data value)*

Transformation	Transformed Value
Log	$\log(x)$
Square	$x^2$
Square root	$\sqrt{x}$
Inverse	$1/x$

- Different transformations are appropriate for different problem contexts. Sometimes we have to experiment to figure out the right one to use.



# Box and Cox Transformation

- Estimate  $\lambda$  from data such that:

$$x^* = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases}$$

- Covers all common data transformations

Lambda can be estimated from the data itself, and can take many different values.

$\lambda$	Transformation	Transformed Value
0	Log	$\log(x)$
2	Square	$\frac{x^2 - 1}{2}$
1/2	Square root	$2(\sqrt{x} - 1)$
-1	Inverse	$1 - 1/x$



# Data Reduction

---

- Acts on multiple variables
- Data reduction: reduce the data by generating a smaller set of variables that seek to capture a majority of the information in the original variables

## **Principle Component Analysis (PCA)**

- Find linear combinations (i.e., weighted averages) of the variables, known as principal components (PCs), which capture the most possible variance
  - These linear combinations are uncorrelated (no information overlap), and only a few of them contain most of the original information.
- Important to first apply single variable transformations so all original variables are on the same scale



# Assessment

---

**Q: What must be done to variables in a data set before applying principal component analysis and why?**

A. You must take the square root of all data values to reduce the overall magnitudes of the data set

B. You must scale the variables so that principal components are not dominated by variables of much larger scale

C. You must make all variables negative to work with values of the same sign

D. You must scale the variables so that only outliers are considered as principal components



# Week 1 Objectives

---

- Understand the importance of data cleaning. Garbage In, Garbage Out
  - Data Cleanup and Transformation
  - **Dealing with Missing Values**
  - Dealing with Outliers
  - Adding and Removing Variables
- To be able to identify errors in data and artifacts.
- Sources of errors in data and their telltale signs in data sets.



# Missing Values

- *Missing values*: no data value is available for a variable in an observation

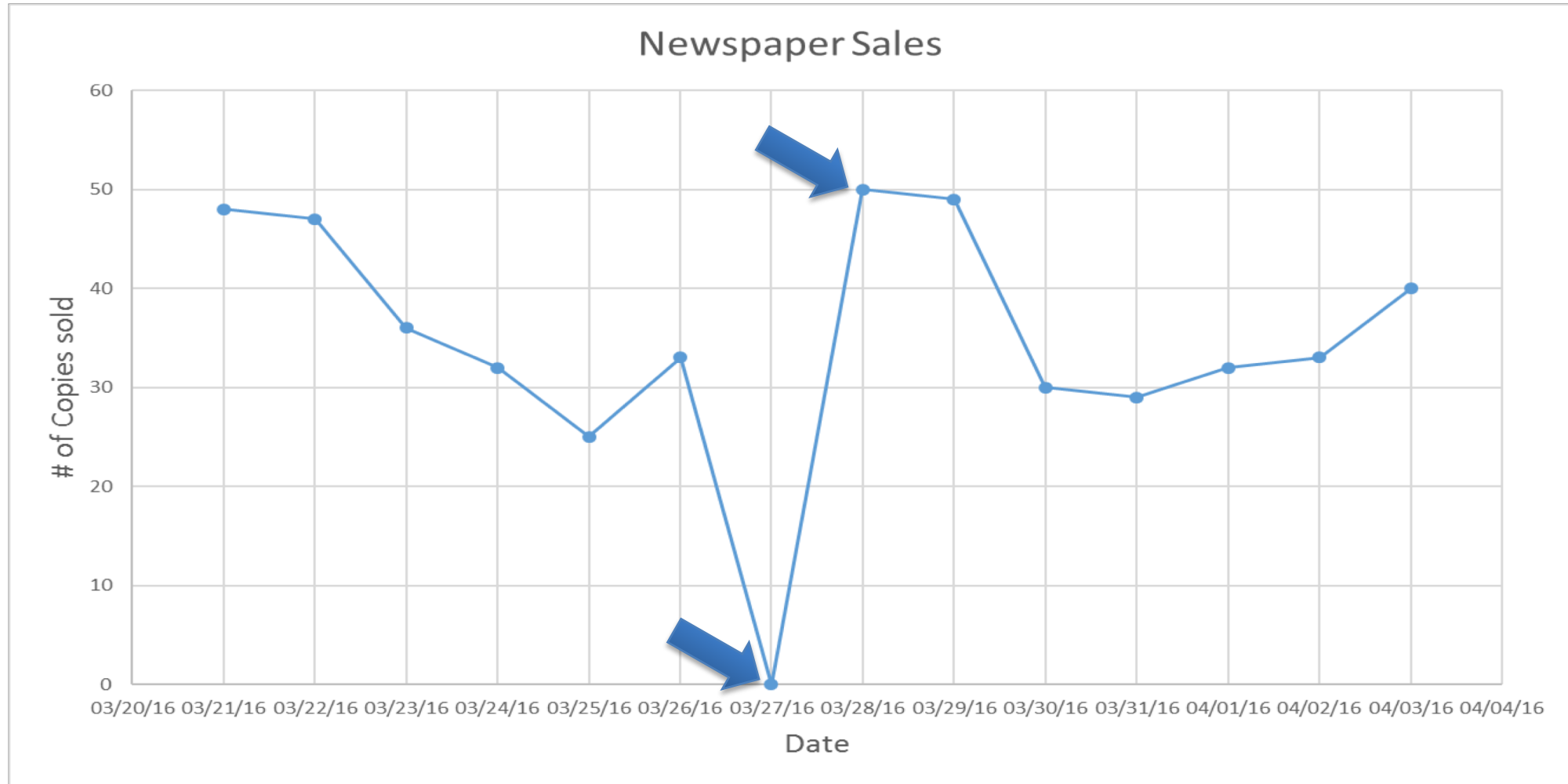
## Newspaper Sales Data

Date	Sales	Column 3	Column 4	Column 5	Column 6
03/21/16	48	xxx	xxx		xxx
03/22/16	47	xxx	xxx	xxx	xxx
03/23/16	36		xxx	xxx	xxx
03/24/16	32		xxx	xxx	xxx
03/25/16	25	xxx	xxx	xxx	xxx
03/26/16	33	xxx		xxx	xxx
03/27/16		xxx	xxx	xxx	
03/28/16	50	xxx	xxx	xxx	xxx
03/29/16	49	xxx	xxx		xxx
03/30/16	30	xxx	xxx	xxx	xxx
03/31/16	29	xxx	xxx	xxx	
04/01/16	32	xxx	xxx	xxx	xxx
04/02/16	33	xxx	xxx	xxx	xxx
04/03/16	40	xxx	xxx	xxx	xxx

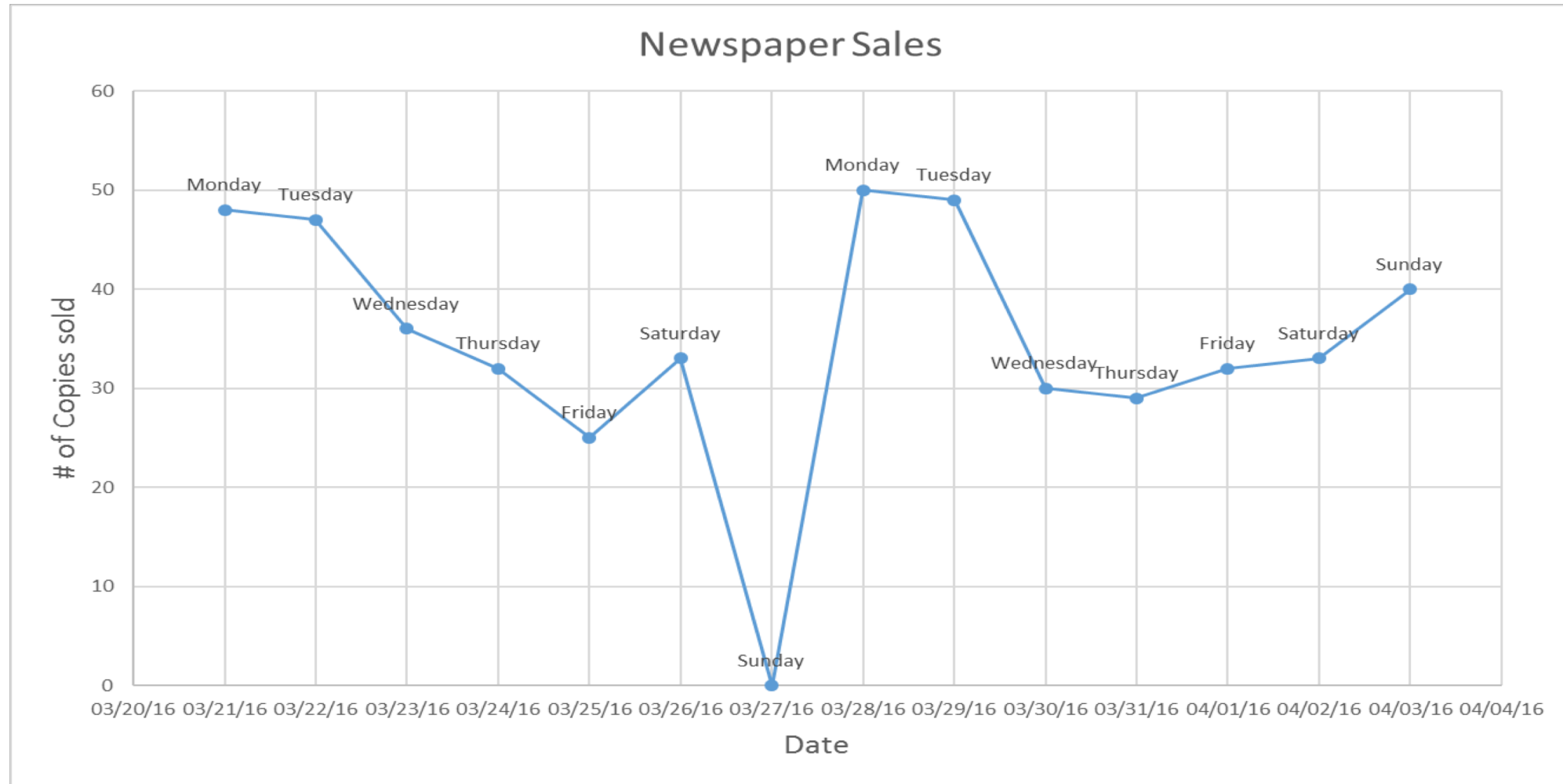
Date	Sales
03/21/16	48
03/22/16	47
03/23/16	36
03/24/16	32
03/25/16	25
03/26/16	33
03/27/16	0
03/28/16	50
03/29/16	49
03/30/16	30
03/31/16	29
04/01/16	32
04/02/16	33
04/03/16	40



# Line Plot of the Data



# Line Plot with Weekday Information





# Line Plot with Weekday Information



# What to do about it?

---

- Remove the corresponding row/column
- Impute (“estimate”) a value
  - With zero
  - With average
  - With similar data points (“interpolation”)
- Make “missing” its own category



## Imputing a sales value for March 27

<i>Date</i>	<i>Sales</i>
03/21/16	48
03/22/16	47
03/23/16	36
03/24/16	32
03/25/16	25
03/26/16	33
03/27/16	0
03/28/16	50
03/29/16	49
03/30/16	30
03/31/16	29
04/01/16	32
04/02/16	33
04/03/16	40

- Impute (“estimate”) a value
  - ~~With zero~~
  - With average → 37.23
  - With similar data points
    - Using sales on Sunday, April 3 as an estimate → 40



**Q: What are the risks of replacing a missing value with a guess?**

Select all that apply :

- Falsifying results
- None, the database is capable of correcting input mistakes
- Distorting the data set
- Introducing biases



# Missing Values vs. Censored Values

---

- Only 50 copies are available for sale each day.
- The sales on Monday, March 28<sup>th</sup> is 50.
- Sales data on that day is “censored.”



# Missing Data Solutions

---

- No single accepted solution.
- Consider context.
- Try not to induce biases or distortions.
- Need enough data to remain for meaningful analysis.
- Pattern of missing values can itself carry information.



# Week 1 Objectives

---

- Understand the importance of data cleaning. Garbage In, Garbage Out
  - Data Cleanup and Transformation
  - Dealing with Missing Values
  - **Dealing with Outliers**
  - Adding and Removing Variables
- To be able to identify errors in data and artifacts.
- Sources of errors in data and their telltale signs in data sets.



# Outliers of a Single Variable

---

- **Outlier:** An observation “far away” from other observations.
- No consensus definition of “far away”
  - More than 3 standard deviations away from the mean.
  - Other definitions based on statistical tests, nearest neighbors, quartile ranges.





## **Q: What are the characteristics of an outlier?**

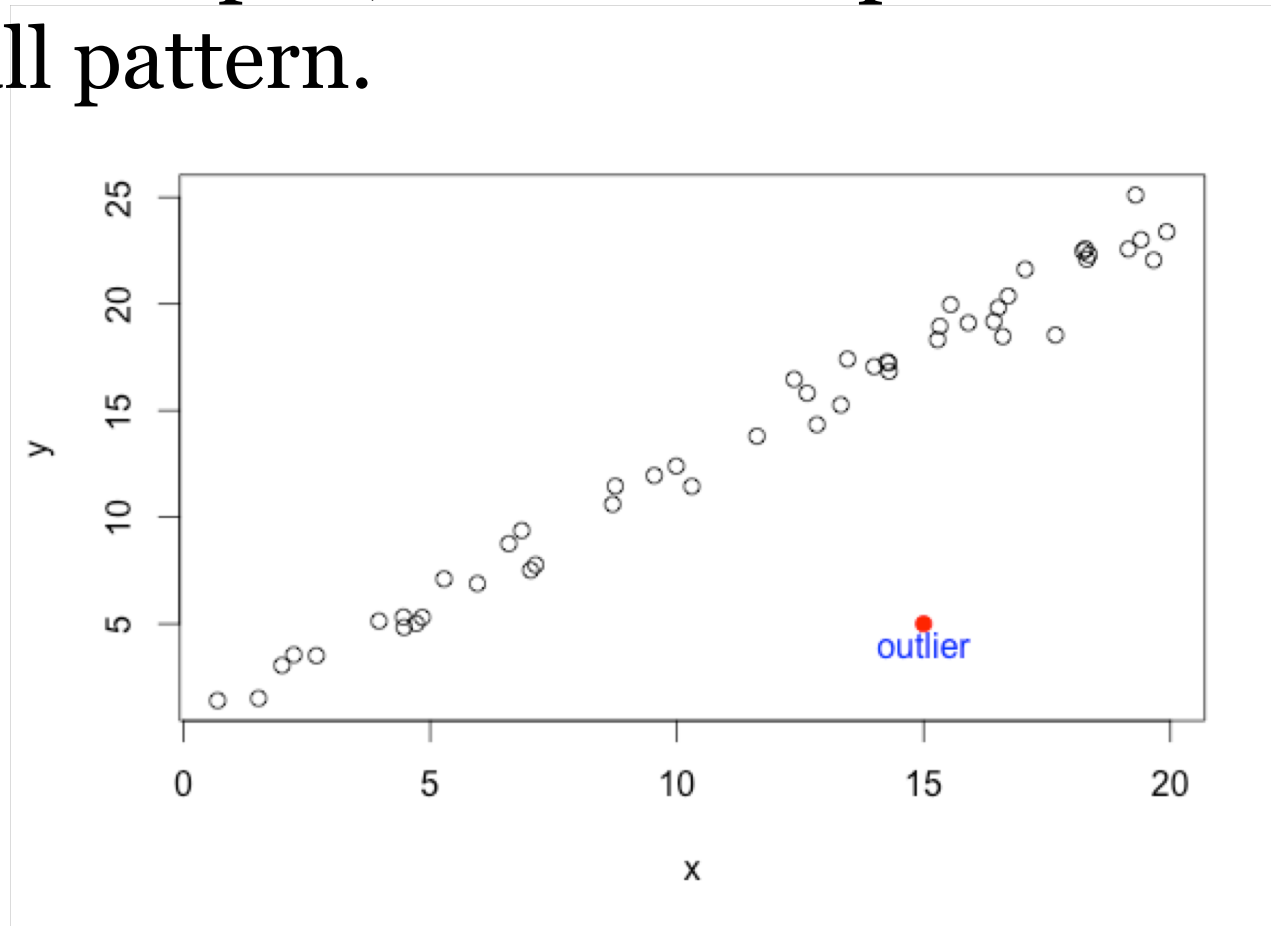
Select all that apply :

- It is above or below 3 standard deviations of the mean
- It falls far outside the overall data pattern
- It is the pivot point for the overall pattern that the data follows
- It is the data point most proximal to the mean



# Outliers in Relationships

- In a scatterplot, outliers are points that fall outside of the overall pattern.



# Assessment

---

True or False?

A data point is not considered an outlier unless it deviates dramatically on either the x-axis or the y-axis?

- True
- False



# Key Questions about Outliers

---

- Is the outlier a mistake or legitimate point?
- Is the outlier part of the population of interest?



# Outliers: What to do?

---

- Correct (if outlier results from a mistake)
- Remove (if outlier is outside the population of interest)
- Data transformation
- Treat as missing data
- Use tools robust to outliers
  - Mean vs. Median
  - “Robust” modeling
  - Cross validation



## **Q: Why do outliers exist?**

Select all that apply :

- a. Legitimate but odd observations
- b. Entropy of a system
- c. Distortion of time
- d. Data recording errors



# Outliers as the main focus

---

- One can argue that all successful startups are outliers
  - as the success rate of startups is very low
  - however, one can be interested only in successful startups
- Data transformation can sometimes eliminate outliers
- It is also possible to treat outliers as missing data



# Week 1 Objectives

---

- Understand the importance of data cleaning. Garbage In, Garbage Out
  - Data Cleanup and Transformation
  - Dealing with Missing Values
  - Dealing with Outliers
  - **Adding and Removing Variables**
- To be able to identify errors in data and artifacts.
- Sources of errors in data and their telltale signs in data sets.





# Adding Variables

---

- Can data be collected on other variables?
- Creating additional variables from data
  - Dummy variables
  - Transformed variables (e.g., polynomial terms)
  - Interactions



# Creating Dummy Variables


---

- A single categorical variable with  $m$  categories is typically transformed into  **$m-1$**  dummy variables
- Each dummy variable takes the values 0 or 1
  - 0 = “no”, 1 = “yes”



# Dummy Variable Example

---

Property type		D1	D2
Single family home		1	0
Townhouse		0	1
Condo		0	0

- D1: Is it a single family home?
- D2: Is it a townhouse?
- $D1 = D2 = 0$  indicates a condo.



# Dummy Variable. Cons.

---

- When there are a large number of categories, we can end up with too many dummy variables
- Solution: Reduce by combining categories that are close to each other



# Dummy Variable Example Revisited

---

Property type
Single family home
Townhouse
Condo
Coop
Multi-family
Mobile home

Property type
Single family home
Townhouse
Others



- Is it necessary to keep all categories?



# Combined categories notes

---

- how categories should be combined is not always straightforward and it should be chosen carefully
- It is common to combine sparse categories together, but it may also depend on the problem context.
- Many software packages have built-in functions for this purpose



# Why Remove Variables?

---

- Simplifies model and analysis
- Some variables contain the same information
- More parsimonious and interpretable model
- Some analytical methods can be crippled by **degenerate\*** distribution, causing model performance and stability issues

*\*degenerate random variables – a die having the same number in all 6 faces*



# Degenerative Distribution Example

---

- Variables with zero variance (only one value for all rows) or near zero variance.
  - **zero variance** example, if everybody in a high school class are born in the same year, then the first year is a zero variance variable, and is not a relevant variable in analysis, including it may cause numeric difficulties.
  - **near zero variance** example, if all students in the class are born in the same year except one student, then the first year is a variable with near-zero variance.





**Q: To say a variable is degenerate means which of the following?**

Select all that apply :

- The variable can only take on a single value
- When plotted, the variable is modeled with an exponential decay
- The variable is a zero variance variable
- The variable is immoral and corrupt



# Degenerative variables

---

- Keeping these variables can cause **collinearity** in regression analysis, where coefficient estimates may change erratically in response to small changes to data
- *Collinearity ~ correlation bt. predictor variables (or independent variables), s.t. they express a linear relationship in a regression model. When predictor variables in the same regression model are correlated, they cannot independently predict the value of the dependent variable*



# Week 1 Objectives

---

- Understand the importance of data cleaning. Garbage In, Garbage Out
  - Data Cleanup and Transformation
  - Dealing with Missing Values
  - Dealing with Outliers
  - Adding and Removing Variables
- **Sources of errors in data and their telltale signs in data sets.**



# Sources of errors in data

---

- The most effective way to address data quality issues is to
  - **prevent them from ever happening** in the first place by controlling by how data is captured at the source. How we do this depends on exactly how the data is captured.
- One of the biggest drivers of bad data are **errors introduced via manual data entry** by people, whether their customers, other outside partners or our own employees.
- To help minimize these types of errors organizations might **build in validation mechanisms** or auto populate certain pieces of information.
  - For example, an online form might force you to enter a valid phone number in a specific format, make you use a dropdown box to choose the state where you live, or even prepopulate your city based on the ZIP code you enter. It also might not let you submit or proceed unless all the require fields are filled in.



# Sources of Error in Data (cont.)

---

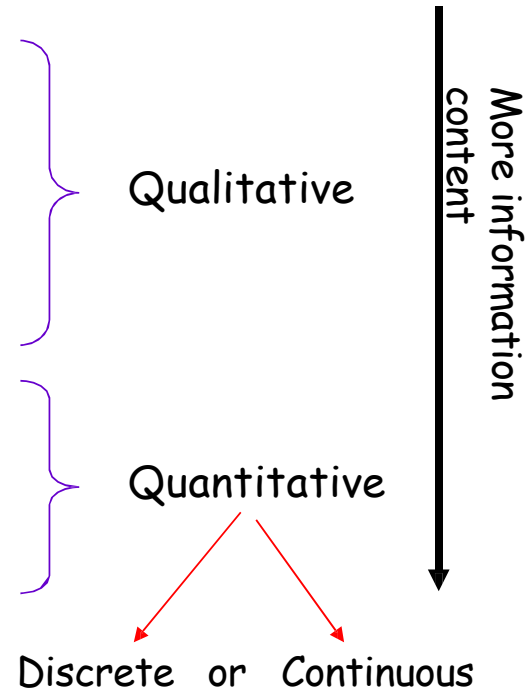
Sources of error in databases categories:

- Data entry errors
- Measurement errors
- Distillation errors
- Data integration errors:



# Types of Measurements

- Nominal scale
- Categorical scale
- Ordinal scale
- Interval scale
- Ratio scale

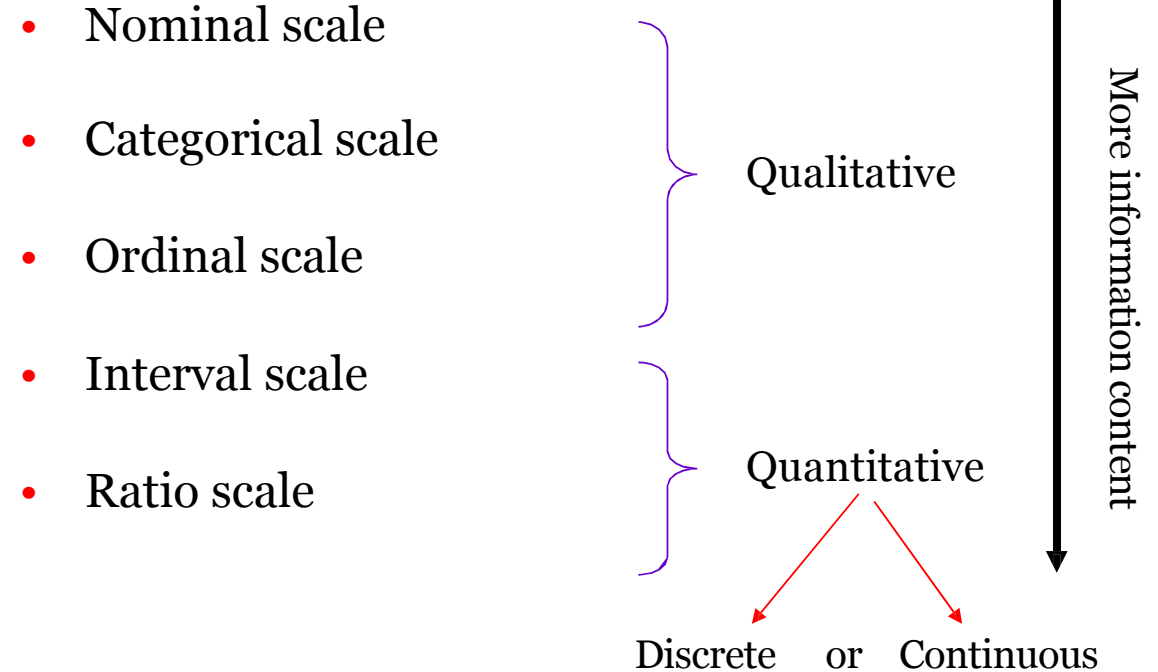


TODO combine this slide



# Types of Measurements: Examples

- Nominal:
  - ID numbers, Names of people
- Categorical:
  - eye color, zip codes
- Ordinal:
  - rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
- Interval:
  - calendar dates, temperatures in Celsius or Fahrenheit, GRE (Graduate Record Examination) and IQ scores
- Ratio:
  - temperature in Kelvin, length, time, counts



# Approaches to Improving Data Quality

---

- The “lifetime” of data is a multi-step and sometimes iterative process involving collection, transformation, storage, auditing, cleaning and analysis.
  
- Approaches suggested for maintaining or improving data quality:
  - Data entry interface design
  - Organizational management
  - Automated data auditing and cleaning
  - Exploratory data analysis and cleaning





# Recap



Yeshiva University®

# Data entry interface design

---

- For human data entry, errors in data can often be mitigated through judicious design of data entry interfaces.
- Specification and maintenance of database integrity constraints:
  - data type checks,
  - bounds on numeric values, and
  - referential integrity (the prevention of references to non-existent data).
- When these integrity constraints are enforced by the database, data entry interfaces prevent data-entry users from providing data that violates the constraints.



# Organizational management

---

- Total Data Quality Management. This work tends to include the use of technological solutions, but also focuses on organizational structures and incentives to help improve data quality.
- These include streamlining processes for data collection, archiving and analysis to minimize opportunities for error; automating data capture; capturing metadata and using it to improve data interpretation; and incentives for multiple parties to participate in the process of maintaining data quality [Huang et al., 1999].



# Exploratory data analysis and cleaning

---

- In many if not most instances, data can only be cleaned effectively with some human involvement. Therefore there is typically an interaction between data cleaning tools and data visualization systems. Exploratory Data Analysis [Tukey, 1977] (sometimes called Exploratory Data Mining in more recent literature [Dasu and Johnson, 2003]) typically involves a human in the process of understanding properties of a dataset, including the identification and possible rectification of errors.
- Data profiling is often used to give a big picture of the contents of a dataset, alongside metadata that describes the possible structures and values in the database. Data visualizations are often used to make statistical properties of the data (distributions, correlations, etc.) accessible to data analysts.



# Enforced Quality

---

- Bring data into a common location
- Despite our best efforts, it's still possible that data has made it's way into our database.
- Or we somehow introduce some errors in moving it from one place to another.
- The best data quality programs use a multi faceted approach that puts quality controls at every step of the process.

