ILLINOIS
Gies College of Business

Introduction to Business Analytics

Ashish Khandelwal

ILLINOIS
Gies College of Business

Introduction to Module 3

Introduction to Business Analytics


CLOUDY
with a chance of
MEATBALLS
IN 3D

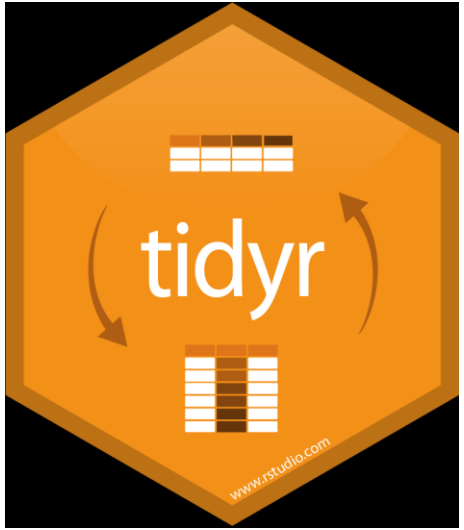

CLOUDY
MEATBALLS
IN 3D

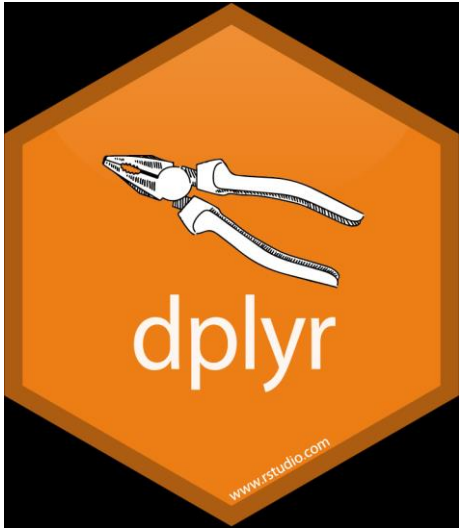
Module 3 Objectives

- Data quality

Module 3 Objectives

- Data quality
- Planning for data preparation


tidyr
www.rstudio.com


dplyr
www.rstudio.com


stringr
www.rstudio.com

ILLINOIS
Gies College of Business

Data Quality

Introduction to Business Analytics

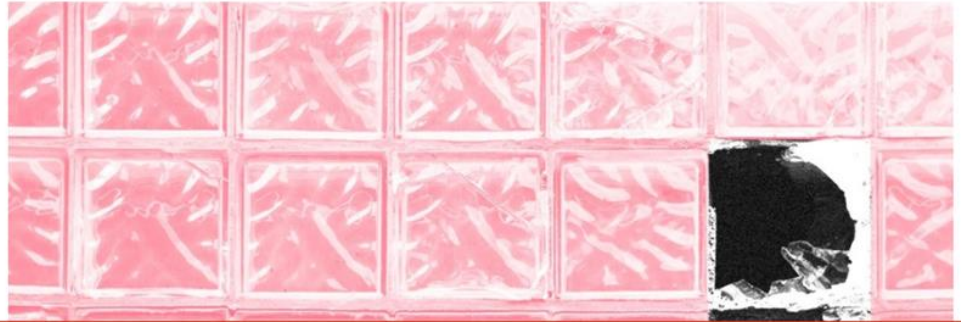
DATA

Only 3% of Companies' Data Meets Basic Quality Standards

by Tadhg Nagle, Thomas C. Redman, and David Sammon

September 11, 2017

Summary Save Share 12 Comment Print $8.95 Buy Copies



https://hbsp.harvard.edu/product/H03VU5-PDF-ENG?Ntt=Only+3%25+of+Companies%E2%80%99+Data+Meets+Basic+Quality+Standards&itemFindingMethod=Search

amazon

Uber

NETFLIX

2020 Global data management research

Our 2020 Global data management benchmark report launches February 18th. Sign up to access our exclusive industry insights and learn how you can gain control over your data.

Our new research will help you unlock the true power of your organization's data, helping you transform your business. We spoke with more than 1,000 global professionals to uncover today's top challenges in leveraging trusted data and tips for how you can drive a data-driven culture.

https://www.edq.com/global-data-management-research-2020/

Data Debt

Data asset that isn’t necessarily fit for the purpose or has high degree of inaccuracy

Experian Reports

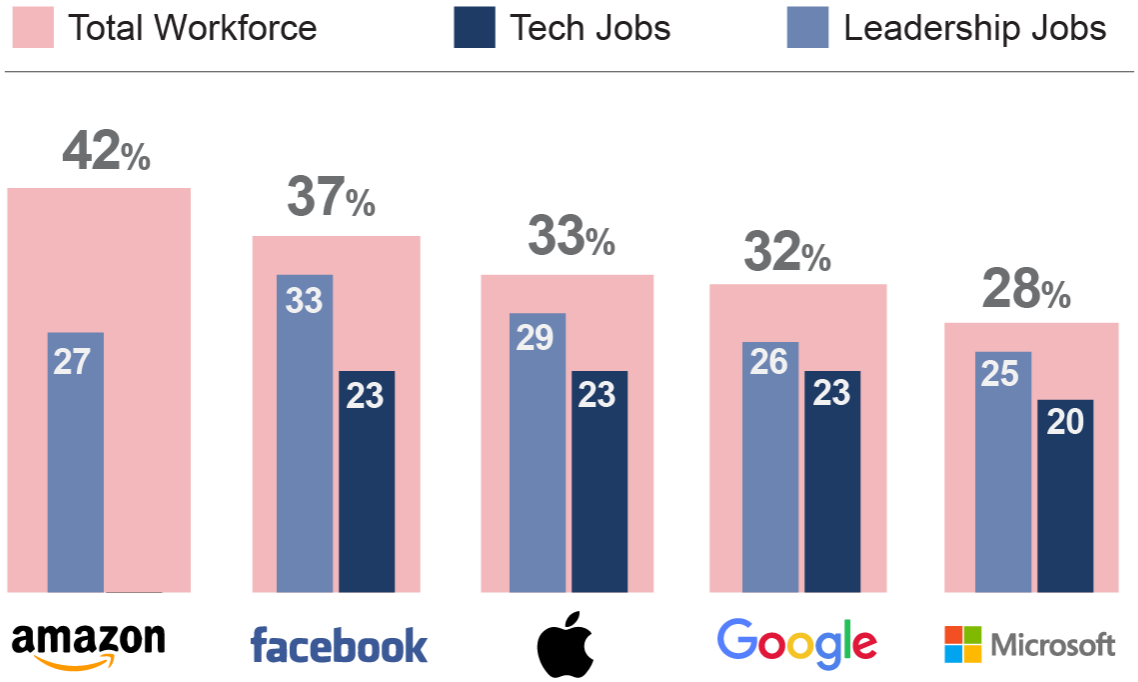
78% of organizations suffer from data debt.

Data Debt Hurts

- Lack of trust on insights from the data
- Poor ROI on the data and tech investments
- Barrier to being a data-driven company

https://www.edq.com/global-data-management-research-2020/

Proportion of Female in Tech Industry



Capital One

ILLINOIS
Gies College of Business

Data Structure Based on the Business Problem Part 1

Introduction to Business Analytics

The real data

The data that goes for modeling

https://pixabay.com/photos/people-kid-child-dirty-ice-cream-2588030/

https://pixabay.com/photos/boy-handsome-child-young-man-male-773630/

https://pixabay.com/photos/black-boy-child-costume-dressed-72856/

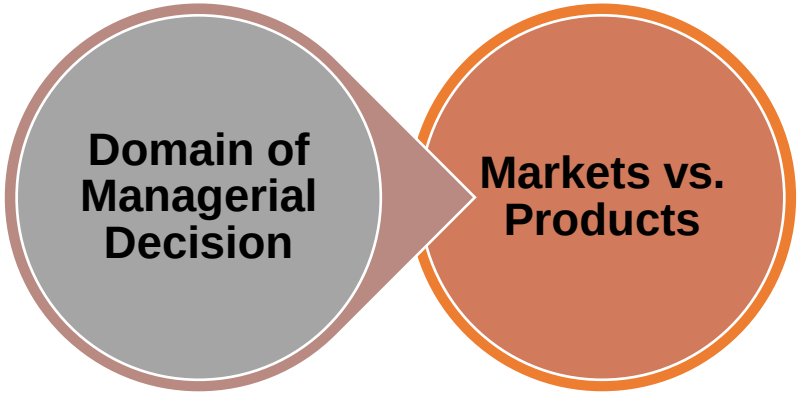
https://pixabay.com/photos/boy-handsome-child-young-man-male-773630/

Business Analytics

Managerial Domain of Action/Decision

Analysis

Data


Domain of Managerial Decision
Markets vs. Products

Market Focus

Problem Focus

Which markets to focus more on?

Analysis

Compare markets for profitability potential

Data Structured

For market level analysis

Product Focus

Problem Focus

Which products to focus more on?

Analysis

Compare products for profitability potential

Data Structured

For product level analysis

Data prepared for Market level decision

Data prepared for Product level decision

Characteristics of each row/Markets

Market ID	Market Feature1	Market Feature2	Market Feature3	Market Feature4	Profitability Score
Market1					
Market2					
Market3					

Markets

Characteristics of each row/Product

Product ID	Product Feature1	Product Feature2	Product Feature3	Product Feature4	Profitability Score
Product 1					
Product 2					
Product 3					

Products

ILLINOIS
Gies College of Business

Data Structure Based on the Business Problem Part 2

Introduction to Business Analytics

Levels within the Product Domain

Department

Category

Lineitem

Data Structured for Lineitem Level Analysis

Characteristics of each row/Lineitem

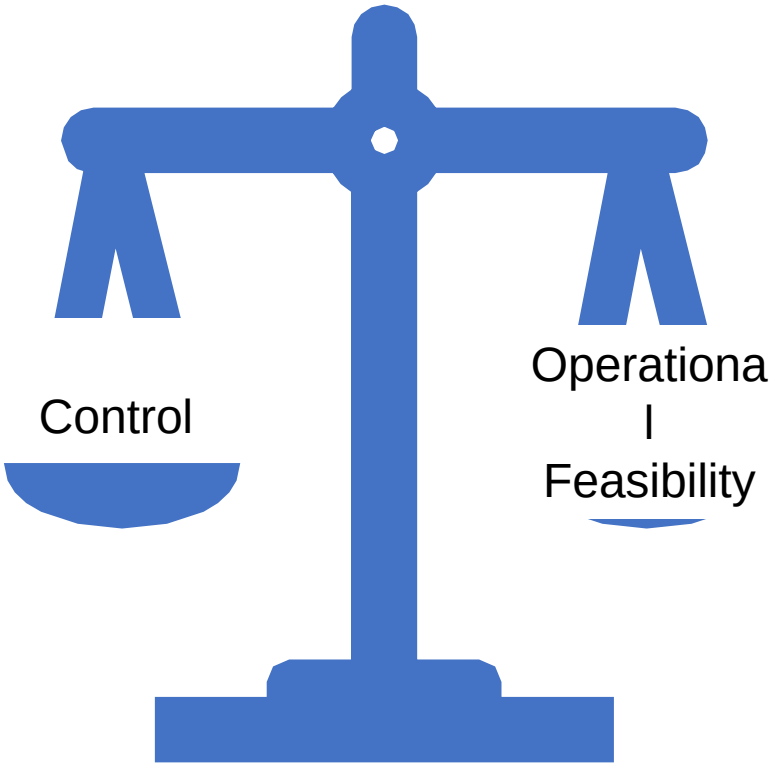
Market ID	Cost	Price	Sales	Profitability Score
Lineitem1				
Lineitem2				
Lineitem3				

Lineitems

Levels within the Domain of Time

https://www.showbizcinemas.com/promotions/

http://onekkom.blogspot.com/2019/05/happy-hour-enjoy-50-discount-


Control
Operationa
l
Feasibility

ILLINOIS
Gies College of Business

Useful Operators for Data Manipulation

Introduction to Business Analytics

Two useful functions in R

- %>%
- %in%

Pipe operator

%>%

- Filter “ities” data where *Operation Type is SALE* and then select columns Cost and Price

- ***Code using %>%***

- df_piped <- df %>% filter(Operation_Type == “SALE”) %>% select(Cost, Price)

Code without Using %>%

Two steps:

Intermediate object

a) df1 <- filter(df, OperationType = “SALE”)

b) df2 <- select(df1, Cost, Price)

Passed as data argument

Code without using %>%

```
df2 <-  select(filter(df, OperationType = “SALE”), Cost, Price)
```

Piped code

```
df_piped  <-  df  %>%  filter(OperationType == “SALE”)  %>%
select(Cost, Price)
```

- The output returned by %>% is always a dataframe

- The output returned by %>% is always a valid input (data argument) to any tidyverse function such as filter, select, ggplot, join, etc.

%in%

%>%

ILLINOIS
Gies College of Business

Creating New Variables using Mutate function

Introduction to Business Analytics

- Mutate - To create a new column
- Distinct - To extract the unique value

ILLINOIS
Gies College of Business

Data Aggregation using Summarize and Group_By functions

Introduction to Business Analytics

Dplyr Functions for Aggregating Data

1. summaries
2. group by

Aggregating Data

Customer Level Insight

Which are the best customers?

Customer Level Analysis

Scoring each customer

Data Structure

Aggregated at the customer level

Transaction Data Aggregated at the Customer Level

Transaction Level Data

TransactionNumber	CustomerCode	Price	Quantity
00RPUIR153361	CWM11331L8O	16.19	1
004GU2B121842	CWM11331L8O	12.02	1
0039X3L74243	CWM11331L8O	7.81	1
00YYWNB173227	CXP4593H7E	2.88	1

Customer Level Data

CustomerCode	Avg_Price	Avg_Quantity
CWM11331L8O	12.84133333	1
CXP4593H7E	2.88	1

Summary Measures

- Mean
- Sum
- Median

Dplyr Functions for Aggregating Data

1. summarise
2. group_by

ILLINOIS
Gies College of Business

Handling Missing Values

Introduction to Business Analytics

Identifying and Handling Missing Values in R

Missing values are saved as “na”

is.na()

Returns list of TRUE and FALSE

Handling Missing Values

- Mean imputation
- Median imputation (when there are outliers)
- Model-based approach
- Dropping the rows with missing values

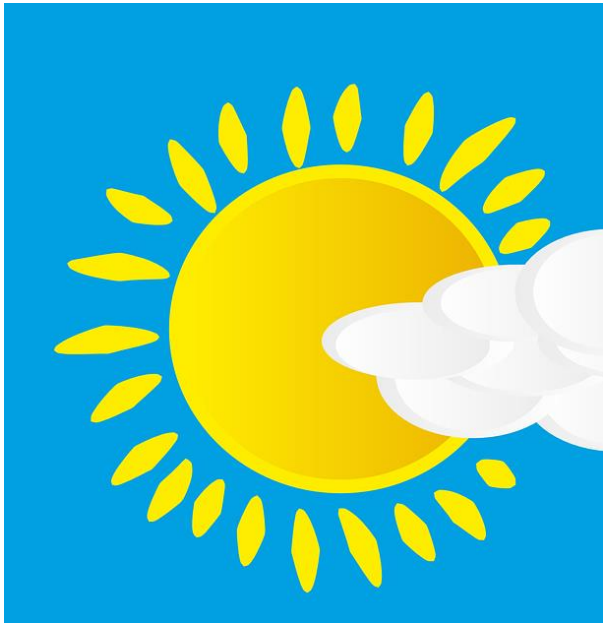
ILLINOIS
Gies College of Business

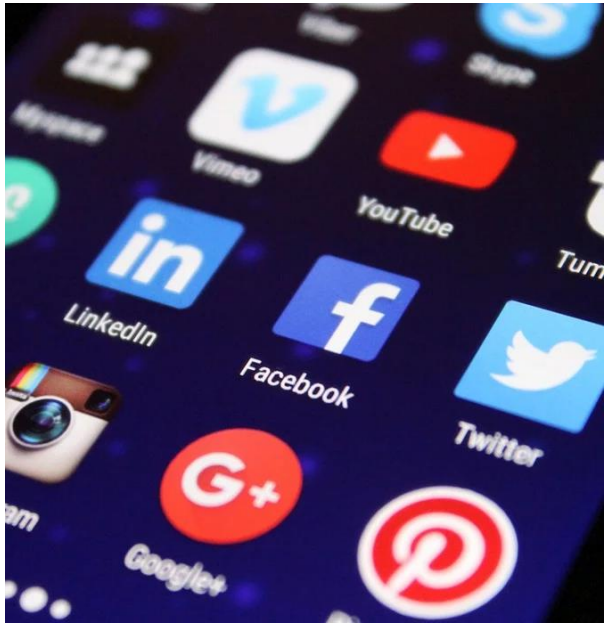
Data Join

Introduction to Business Analytics

Joining Datasets Using Dplyr Package

Add More Context to Our Analysis





https://pixabay.com/vectors/cloud-weather-forecast-weather-sky-346710/
https://pixabay.com/photos/football-playing-field-corner-1419954/
https://pixabay.com/photos/media-social-media-apps-998990/

Levels of the Datasets to Join Must Match

Date Level

	date	PRCP	SNOW	TMAX	TMIN
2138	2017-04-03	0.0000000	0.0000000	60.08	33.08
2137	2017-04-02	0.0000000	0.0000000	57.02	32.00
2136	2017-04-01	0.1181102	0.0000000	57.02	33.98
2135	2017-03-31	4.2125984	0.0000000	62.06	33.98
2134	2017-03-30	0.0000000	0.0000000	64.94	46.94
2133	2017-03-29	0.0000000	0.0000000	68.00	42.98
2132	2017-03-28	0.0000000	0.0000000	69.08	39.92
2131	2017-03-27	0.0000000	0.0000000	64.94	30.92
2130	2017-03-26	1.4960630	0.0000000	57.92	39.92

Transaction Level – many rows for one date

	Time	OperationType	BarCode
392387	2017-04-03T17:08:00Z	SALE	*
392388	2017-04-03T17:08:00Z	SALE	*
392390	2017-04-03T16:52:00Z	SALE	*
392396	2017-04-03T16:49:00Z	SALE	*
392402	2017-04-03T16:46:00Z	SALE	*
392401	2017-04-03T16:46:00Z	SALE	*
392406	2017-04-03T16:26:00Z	SALE	*
392410	2017-04-03T15:36:00Z	SALE	*
392416	2017-04-03T14:47:00Z	SALE	*

Needs to be aggregate at the "date" level

Joining Datasets

Left

df_ities_date

Right

df_weather

Datasets Are Merged/Joined Based on a Common Column

Common column – primary key

- Unique values
- Non-missing values

Four Types of Join

- Left Join
- Right Join
- Full Join
- Inner Join

ILLINOIS
Gies College of Business

Long vs Wide Format for Data

Introduction to Business Analytics

Two Formats for Representing Data

- Long Format
- Wide Format

Long Format

LineItem	CashierName	Quantity	CustomerSatisfaction
Aubergine and Chickpea Vindaloo	Katherine Roth	1.362319	0.990235803
Aubergine and Chickpea Vindaloo	Rachael Price	1.305556	0.990235803
Aubergine and Chickpea Vindaloo	Trinidad Johnson	1.000000	0.990235803
Aubergine and Chickpea Vindaloo	Vincent Ball	1.506024	0.990235803
Beef and Apple Burgers	Katherine Roth	1.011364	0.325210194
Beef and Apple Burgers	Rachael Price	1.000000	0.325210194
Beef and Apple Burgers	Trinidad Johnson	1.000000	0.325210194
Beef and Apple Burgers	Vincent Ball	1.015432	0.325210194
Beef and Apple Burgers - Illini Bhangra	Katherine Roth	1.335723	0.527040931
Beef and Apple Burgers - Illini Bhangra	Rachael Price	1.259542	0.527040931
Beef and Apple Burgers - Illini Bhangra	Trinidad Johnson	1.229141	0.527040931
Beef and Apple Burgers - Illini Bhangra	Vincent Ball	15.000000	0.527040931

Wide Format

LineItem	Quantity_Katherine Roth	Quantity_Rachael Price	Quantity_Trinidad Johnson	Quantity_Vincent Ball	CustomerSatisfaction
Aubergine and Chickpea Vindaloo	1.362319	1.305556	1.000000	1.506024	0.990235803
Beef and Apple Burgers	1.011364	1.000000	1.000000	1.015432	0.325210194
Beef and Apple Burgers - Illini Bhangra	1.335723	1.259542	1.229141	15.000000	0.527040931

Long → Wide

- The number of rows decreases
- The number of columns increases

Wide → Long

- A few columns are dropped
- The number of rows increases

ILLINOIS
Gies College of Business

Manipulating Strings

Introduction to Business Analytics

Str_*verb/ noun*

- *Replace*
- *Detect*
- *Extract*
- *Length*

Length of LineItem Name

- *Correlation(LineItem_length, Quantity) = -0.09*
- *Correlation(LineItem_length, Price) = 0.037*
- *Correlation(LineItem_length, Cost) = 0.045*