What is Data Analytics?

Data Analytics – sample definitions

the process of evaluating data with the purpose of drawing conclusions to address business questions.

a way to search through large structured and unstructured data to discover unknown patterns or relationships.

involves the technologies, systems, practices, methodologies, databases, statistics, and applications used to analyze diverse business data to give organizations the information they need to make sound and timely business decisions.

aims to transform raw data into knowledge to create value.

Data Analytics Methodology: IMPACT Model

1. **Identify** the questions
2. **Master** the data
3. **Perform** the test plan
4. **Address** and refine results
5. **Communicate** insights
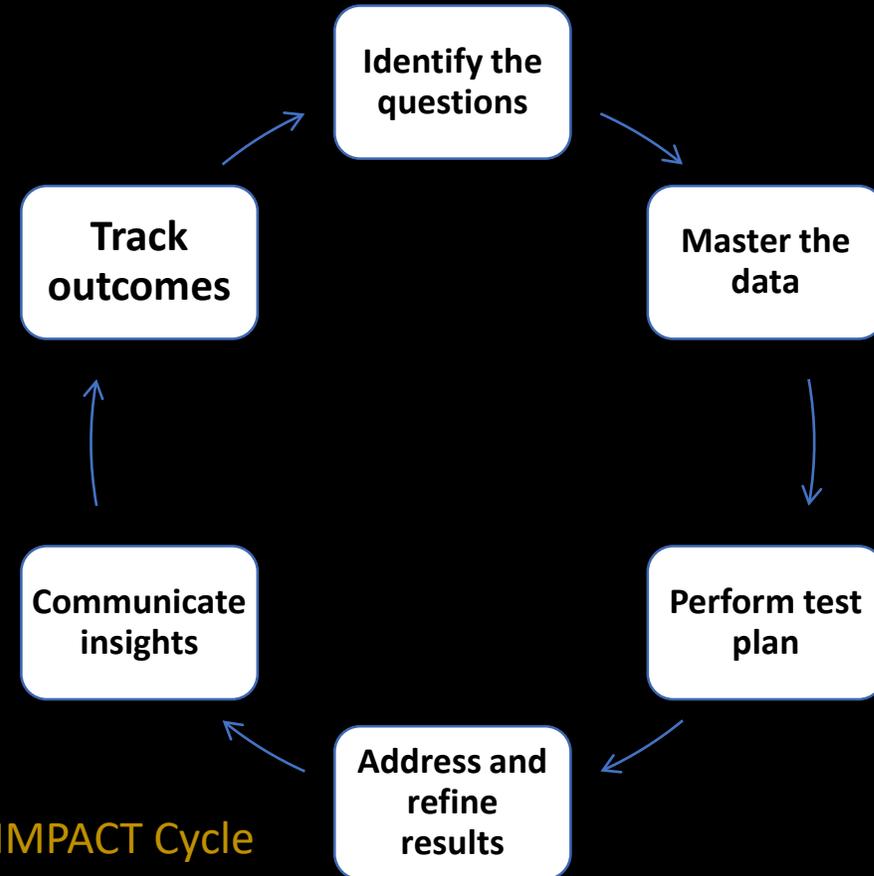6. **Track** outcomes

Exhibit 1-1 The IMPACT Cycle

Possible Questions

Are employees circumventing internal controls over payments?

Are there any suspicious travel and entertainment expenses?

How can we increase the amount of add-on sales of additional goods to our customers? Are our customers paying us in a timely manner?

How can we predict the allowance for loan losses for our bank loans?

How can we find transactions that are risky in terms of accounting issues?

Who authorizes checks above $100,000? How can errors be identified?

How to Determine Your Data Analytics Approach

Exploratory versus Confirmatory?

Descriptive, Diagnostic, Predictive, Prescriptive?

Exploratory versus Confirmatory

Exploratory Data Analysis
First step when you receive a new dataset
Summary/descriptive statistics
Look for errors, anomalies
You might not have a question yet
If you have a generic question, you explore when you haven't yet formed a hypothesis – you have noticed a problem and need to explore the data to form a hypothesis

Confirmatory
You have a hypothesis
Traditional statistical tools– use tests of significance to test your hypothesis

Descriptive, Diagnostic, Predictive, Prescriptive

Exploratory:

- Descriptive: procedures that summarize existing data to determine what has happened in the past.

- Diagnostic: procedures that explore the current data to determine why something has happened the way it has, typically comparing the data to a benchmark.

Typically use Excel summary formulas, Data Analytics tools for descriptive statistics, Excel PivotTables or Tableau data visualization for exploring the data

For larger, more complex datasets, Python, R, or SQL can be used

Confirmatory:

- Predictive: procedures used to generate a model that can be used to determine what is likely to happen in the future. Examples of predictive analytics include regression analysis and forecasting.

- Prescriptive: procedures that model data to enable recommendations for what should be done in the future. These typically include developing more advanced machine learning and artificial intelligence models to recommend a course of action based on a current problem.

For smaller datasets, Excel data analysis tools can work. For larger datasets, more robust statistical programs can be helpful, such as Python, R, or SAS.

Now what? Obtain and Master the Data!

Sometimes, it's as simple as opening a file

More often…..
Data is stored in one format (maybe a relational database or text file) and you want to explore it in Excel or Tableau, so ETL or ELT:
Extract it
Transform/Clean it
Load it into your preferred tool

But today we'll start simple with a clean data file and a question ready to answer. (We'll get more complex soon, though!)

Accounts Receivable Aging

Open file Accounts Receivable Aging.xlsx

AR Aging – Exploratory Analysis to start

Create a PivotTable (Insert > PivotTable > OK)

Place a checkmark in Customer
This provides a unique customer listing

Place a checkmark in InvoiceAmount

Sort the data in descending order (Data > Z-A)

Filter to view only those customers who owe more than $2,500 (on PivotTable, click Row Labels > Value Filters > Greater Than > 2,500)

Add in the Due Date by dragging it to the Columns in the Field List

Continue exploring the data in the PivotTable

AR Aging – Exploratory Analysis Part 2

Return to Raw Data

We'll calculate some summary statistics

Add in the Data Analysis ToolPak

File > Options > Add Ins > Go > Analysis ToolPak > OK

Data tab in the ribbon > Data Analysis ToolPak > Descriptive Statistics

Select the Invoice_Amount, include a checkbox for Labels, and include a checkbox for Summary Statistics. Click OK.

AR Aging – more specific analysis

Open the file Kristy sent you titled "AR Aging – PivotTables"

Work through the steps

This is an example of some simple analysis you can work through with introductory students to help them understand two main things:

How to work with Excel

A greater understanding of aging accounts receivables

Repeating the exercise with Tableau

Start with exploratory analysis

Open Tableau and connect to the Excel spreadsheet

Add the table to the Data Source

Switch to Sheet 1, and add Invoice Amount (double-click)

Double-click Customer to see the detail

Sort the data

Switch to a box plot – great way to quickly detect outliers

AR Aging – repeated in Tableau

Open a new sheet in Tableau

Work through the steps in "AR Aging– Tableau"

Discussion

Why have your students do the same exercise in both tools?

Which did you prefer?   If it "depends" …. Why? On what?

Moving into more 'fun' analysis

What happens when we don't have clean data to work with?

College Scorecard Data

From the Dropbox: Open the **College Scorecard Data Dictionary**. (PDF file)

Look over the attributes and discuss them with your neighbors. What stands out as interesting to you? What do you think your students would want to explore?

College Scorecard Data Analysis

Load and Transform the data in Excel

Exploratory Analysis: gather summary statistics

Confirmatory Analysis: Does the average SAT average help predict college completion rates? (correlation, then regression)

Master the College Scorecard Data

We'll start by digging into SAT_AVG and C150_4
What do those attributes describe?

Excel > Data tab > Get Data (in Excel 2016, this says "New Query")> From File > From Text > CollegeScorecard_RawData

Transform Data

Change those two attributes to 'Decimal Number," then filter out the Nulls

Close & Load

Master the College Scorecard Data

Copy the two attributes (SAT_AVG and C150_4) to a new spreadsheet so that the two columns are isolated and contiguous

Exploratory/Descriptive Analysis

Run descriptive statistics analyses from the Data Analysis ToolPak to analyze the two isolated attributes (SAT_AVG and C150_4)

Assess your results – does anything stand out as abnormal?

Confirmatory/Predictive Analysis

Run a correlation between the two variables

    =CORREL()

Run a regression test

    Data Analysis ToolPak
    y variable: C150_4
    x variable: SAT_AVG