UNIT 1: UNDERSTANDING DATA
DAY 1: WELCOME TO DATA ANALYTICS

TAKEAWAYS

- Data ranks among one of the most valuable resources in business today. Data analysis is commonly used to increase revenue, decrease risk, and justify investments.
- The Data Analytics Workflow is used to answer data-driven questions.
    - Frame: Develop a hypothesis-driven approach to your analysis.
    - Prepare: Select, import, explore, and clean your data.
    - Analyze: Structure, visualize, and complete your analysis.
    - Interpret: Make recommendations and business decisions from your data.
    - Communicate: Present insights from your data to different audiences
- The five Vs of data are volume, velocity, variety, veracity, and value.
    - Volume: The scale of the data.
    - Velocity: The data source, timing, and flow.
    - Variety: The forms and types required to answer questions.
    - Veracity: The quality, accuracy, and reliability of source(s).
    - Value: The metrics or measures for desired outcomes.
- Color theory and industry principles for visualization design can help create more effective presentations.
    - Clear: Easily seen and sharply applied.
    - Clean: Thorough, complete, and unadulterated.
    - Concise: Brief but comprehensive.
    - Captivating: Attracts and holds attention by beauty or excellence.
- Data narratives are much like traditional stories, with characters (data), plot (your process), etc. Consider the story map of your presentation as you share your results. The presentation canvas can be a good tool for shaping this story.
- Consider how Q&A will affect your presentation. Use leading questions during your presentation. Limit interruptions by telling your audience beforehand that there will be Q&A time at the end.
- Data cleaning and exploratory analysis help you organize and understand your data set, allowing you to work with it in more advanced ways. Here are some best practices:
    - Keep a copy of raw, untouched data.
    - Document your steps through cleaning and preparation:
        - Create new columns for converted data.
        - Use color to highlight prepared columns.
        - Use conditional formatting to identify outliers.
    - Create a summary sheet with “metadata” including:
        - A directory of other sheets.
        - An explanation of analysis.
        - A sample summary of results.

UNIT 1: UNDERSTANDING DATA
DAY 1: WELCOME TO DATA ANALYTICS

VOCAB

KEY TERM	DEFINITION
Data-Driven Decision	*An approach to business and stakeholder decision-making that values choices that can be backed by verifiable data. The success of this approach is reliant on the quality of the data gathered and the effectiveness of the analysis and interpretation.*
Color Theory	*A body of practical guidance on specific color combinations. An understanding of color theory can help create effective visual data presentations.*
Function	*A keyword that takes parameters to simplify calculations, such as SUM(), AVERAGE(), PRODUCT(), STDEV.P, STDEV.S, etc.*
String	*A text entry surrounded by quotes (e.g., "this").*
Delimited	*Values separated by a common indicator (e.g., space or comma).*
Bar/Column Charts	*While in practice, bar charts can be horizontal (bar) or vertical (column). Excel uses specific terminology for each.*
Workbook	*A standard Excel file.*
Worksheet	*The area of a workbook where you create and manage data. You can have multiple worksheets in a workbook.*
Spreadsheet	*The primary feature of a worksheet, made up of columns and rows.*
Column	*Vertical sets of cells, labeled alphabetically from left to right.*
Row	*Horizontal sets of cells, labeled numerically from top to bottom.*
Cell	*The most fundamental element in Excel, each with a unique name based on column and row (starting with A1 at the top-left corner).*
Range	*A group of cells, only one of which can be active at a time.*
Parameter/Argument	*The elements passed to (plugged into) a function.*
Conditional Statement	*Any statement with a TRUE or FALSE value (AND, NOT, OR).*

SYNTAX

KEY ACTION	SPREADSHEET SYNTAX
IF	*=IF (conditional, value_if_true, value_if_false)*
AND	*=AND (condition1, condition2...) [Only returns TRUE if all values are true]*
OR	*=OR (condition1, condition2...) [Only returns FALSE if all values are false]*
NOT	*=NOT (condition) [Returns the opposite of the condition]*
MID	*=MID (text, num1, num2) [Returns characters from num1 to num2 in text]*
LEFT/RIGHT	*=LEFT (text, number) [Returns leftmost n characters from the text string]* *=RIGHT (text, number) [Returns rightmost n characters from the text string]*
LOWER/UPPER	*=LOWER (text) [Changes all characters in string to lowercase]* *=UPPER (text) [Changes all characters in string to uppercase]*
CONCATENATE	*=CONCATENATE (text, text...) or =text&text [Adds strings together]*
CLEAN	*=CLEAN (text) [Removes non-printing characters such as line breaks]*
SUBSTITUTE	*=SUBSTITUTE (range, text_to_find, text_to_replace)*
TRIM	*=TRIM (range) [Removes whitespace at beginning or end of a string]*
LEN	*=LEN (cell) [Returns the number of characters in a string]*
Also Useful	*Text to Columns, Remove Duplicates*

AN

UNIT 1: UNDERSTANDING DATA
DAY 2: WORKING IN EXCEL

TAKEAWAYS

- Four primary strategies for dealing with nulls:
  - Delete them (with caution).
  - Ignore them (some have meaning).
  - Impute values (medians or zeroes).
  - Find the value (reference the resources).
- Lookup tables are useful for referencing other data. HLOOKUP/VLOOKUP use simple numbers to represent indexes, so they don't update automatically. INDEX/MATCH solves this problem by using actual column/cell references.
- PivotTables allow you to quickly create aggregations, slices, filters, and more using your data as a source — without having to write your own formulas.
  - They have four main components: Filters, Rows, Columns, and Values.
  - They work best when each row defines a single observation (data are not aggregated) and variables are categorical.
- Conditional formatting allows us to apply complex formatting rules to one or more columns of data, such as bolding or highlighting certain values or ranges.
- Keep plots and graphs simple for effective visualization:
  - Scatterplots are created by graphing numeric values.
  - Line graphs are effective for plotting a value over time.
- Dashboards can be used to display many visualizations at once. They are useful for showing key performance indicators (KPIs).

VOCAB

KEY TERM	DEFINITION
Referencing	*Pulling the value of one cell into the value of another.*
Lookup Table	*A column or table that references and "looks up" values from another column or table and returns them.*
Aggregate Functions	*Allows you to quickly summarize data using formulas in Excel.*
PivotTables	*Tables with manipulable filter and arrangement options that allow you to look at data grouped in different ways.*

SYNTAX

ACTION	SPREADSHEET SYNTAX
Lookup	*=VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])* *=HLOOKUP(lookup_value, table_array, row_index_num, [range_lookup])*
Index	*=INDEX(range, row_or_column)* *[MATCH argument can replace row_or_column]*
Match	*=MATCH(lookup_value, lookup_range, match_type)*
Aggregate Functions	*Calculations: MIN(values), MAX(values), SUM(values), and AVERAGE(values).* *Counts: COUNT(values) [counts numeric only], COUNTA(values) [counts all non-blank cells], COUNTBLANKS(range), and COUNTIF(range, criteria) / COUNTIFS(range1, criteria1, range2, criteria2).*

RESOURCES

Want to dig deeper? Check out these resources:

- Data Cleaning Steps: https://www.siop.org/tip/backissues/Jan05/PDF/423_089to096.pdf
- Using INDEX/MATCH: http://www.randomwok.com/excel/how-to-use-index-match/
- Using INDEX/MATCH/MATCH: https://www.deskbright.com/excel/using-index-match-match/
- Excel Documentation From Microsoft: https://support.office.com/en-us/excel

UNIT 1: UNDERSTANDING DATA
DAY 3: FUNDAMENTALS OF DATA AND STATISTICS

TAKEAWAYS

- There is a vast amount of data available today because of cheap, accessible digital storage and the fact that nearly all activities can create data in some form.
- On its own, a data set doesn't tell us much. An analyst's job is to tell the data's story. This information is so valuable it has become a business in itself.
- Data analysis follows a general workflow, but it is not strictly linear. Some steps will be revisited and every project involves a different procedure.
    - Frame: Develop a hypothesis-driven approach to your analysis.
    - Prepare: Select, import, explore, and clean your data.
    - Analyze: Structure, visualize, and complete your analysis.
    - Interpret: Make recommendations and business decisions from your data.
    - Communicate: Present insights from your data to different audiences.
- Data visualization is useful because people are better at recognizing patterns in visual displays than in raw numbers. It can come before analysis, help with understanding during analysis, or be included in the presentation of findings.
- No single chart or graph is the best — each can be the best in a certain situation.

	Single Variable	**Multiple Variables**
Continuous Variable	Histogram or Line Graph	Scatterplot
Categorical Variable	Bar Chart, Pie Chart, or Line Graph	Layer other charts

- Linear regressions predict a continuous, numeric-dependent variable from one or more independent variables by using ordinary least squares to produce the line that minimizes the squared errors of the data to that line.
- After the mean, standard deviation is the most prevalent statistic in data analysis, measuring the data's spread from the mean or the variance in a distribution.
- Correlations measure relationships in data, with 1 being a perfect, positive correlation, -1 being perfectly negative, and 0 representing no discernible pattern.
- The R^2 value tells us the proportion of variance that is explained by our model. The higher the value, the better the model. For models with many independent variables, the adjusted R^2 penalizes extra variables that do not add value.
- Regression also gives a p value — a level of confidence about our coefficients. It tells us the probability that the coefficient was found to be not equal to zero by chance.
- Linear regression models take the form $y = b+ax$ (y = dependent variable, x = independent variable, a = regression coefficient, and b = y intercept).
- We need to be careful of statistical pitfalls when examining correlation, including "correlation does not imply causation" (sometimes a hidden variable can explain the relationship), multiple comparisons, selection bias, nonlinear data, etc.

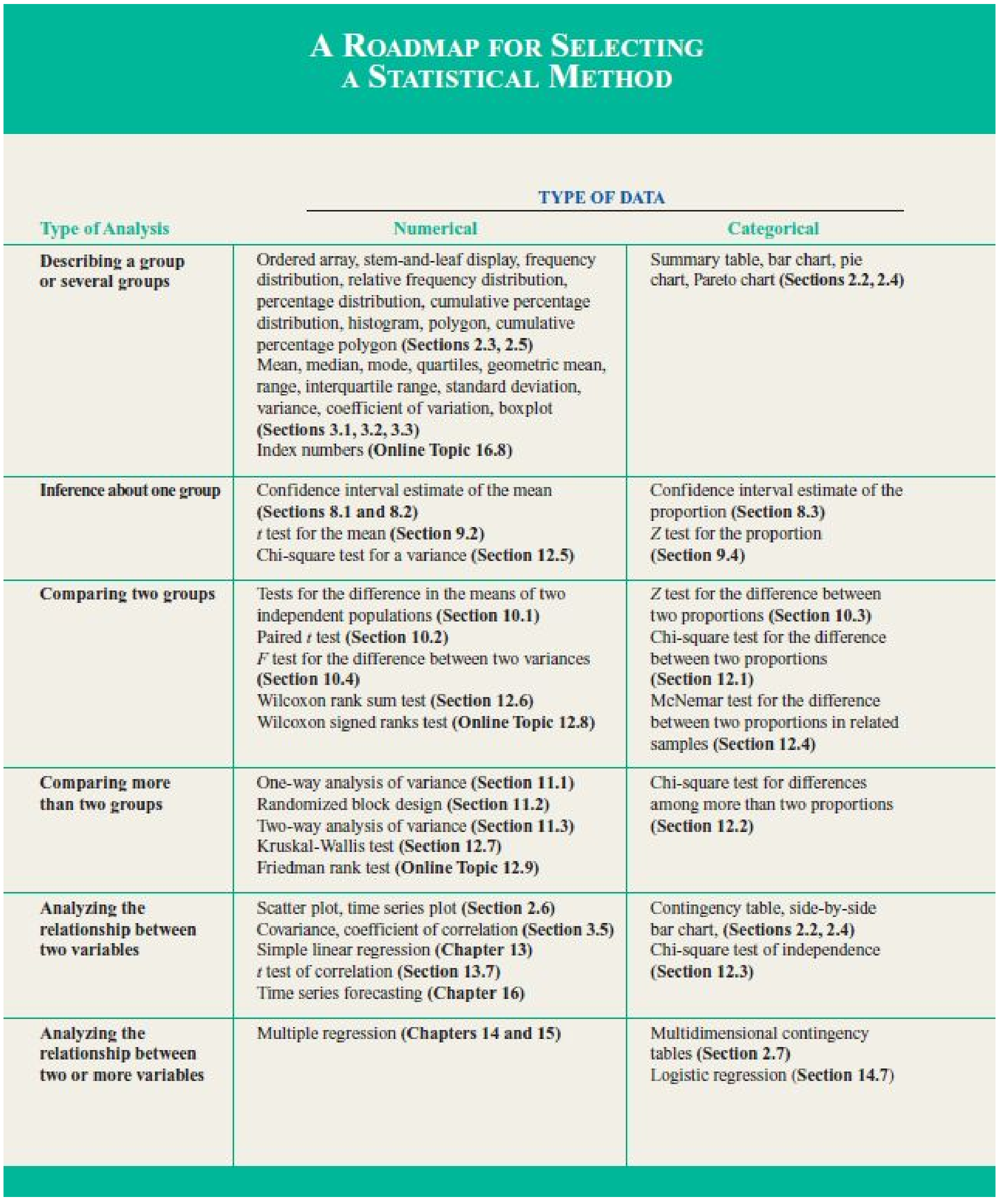
UNIT 1: UNDERSTANDING DATA
DAY 3: FUNDAMENTALS OF DATA AND STATISTICS

VOCAB

KEY TERM	DEFINITION
Dependent Variable	*A variable that is dependent on the values of other variables. Usually the variable we are trying to predict.*
Independent Variable	*A variable that is not dependent on the other variables.*
Exploratory Data Analysis (EDA)	*The process of investigating data. The goal is to obtain a rough understanding of our data and identify potential relationships of interest, which we will analyze later.*
Categorical Variables	*Can only take one of some limited number of values (categories); they typically do not have a natural order. Also called discrete variables.*
Continuous Variables	*Can take an infinite number (or even just a lot) of different values. They have a natural order. The variables are almost always numeric, although dates and times are also possible.*
Histogram	*A bar chart in which data points are grouped into discrete bins. Turns continuous data into categorical data through grouping.*
Statistic	*Any number that describes sample data (rather than population data). Examples include mean, median, max, and min.*
Parameter	*A number that describes the entire population.*
Statistical Inference	*Use of sample statistics to estimate population parameters.*
Standard Deviation	*A measure of variance in a distribution.*
Variance	*A measure of how far away data is from the mean.*
Distribution	*The arrangement of the data. Normal distributions look like a bell curve, while bimodal distributions have two peaks.*
Correlation	*Correlation measures how much two variables tend to have similar values on a scale ranging from 1 to -1.*
Causation	*The idea that changes in one variable cause changes in another (a causal relationship).*
Selection Bias	*The idea that, when selecting sample data from a population, the method of selection may be introducing bias to the data.*
Correlation Matrix	*Table of correlations between all variables in our data set.*
R^2 value	*The proportion of the variance that is explained by our model.*
p-value	*Probability that the coefficient was found to be not equal to zero by chance.*
Homoscedastic	*Data in which the distance between the data and the trendline does not increase as the independent variable increases (as opposed to heteroscedastic).*

UNIT 1: UNDERSTANDING DATA
DAY 3: FUNDAMENTALS OF DATA AND STATISTICS

REFERENCE



A Roadmap for Selecting a Statistical Method

Type of Analysis	TYPE OF DATA Numerical	 Categorical
Describing a group or several groups	Ordered array, stem-and-leaf display, frequency distribution, relative frequency distribution, percentage distribution, cumulative percentage distribution, histogram, polygon, cumulative percentage polygon **(Sections 2.3, 2.5)** Mean, median, mode, quartiles, geometric mean, range, interquartile range, standard deviation, variance, coefficient of variation, boxplot **(Sections 3.1, 3.2, 3.3)** Index numbers **(Online Topic 16.8)**	Summary table, bar chart, pie chart, Pareto chart **(Sections 2.2, 2.4)**
Inference about one group	Confidence interval estimate of the mean **(Sections 8.1 and 8.2)** *t* test for the mean **(Section 9.2)** Chi-square test for a variance **(Section 12.5)**	Confidence interval estimate of the proportion **(Section 8.3)** *Z* test for the proportion **(Section 9.4)**
Comparing two groups	Tests for the difference in the means of two independent populations **(Section 10.1)** Paired *t* test **(Section 10.2)** *F* test for the difference between two variances **(Section 10.4)** Wilcoxon rank sum test **(Section 12.6)** Wilcoxon signed ranks test **(Online Topic 12.8)**	*Z* test for the difference between two proportions **(Section 10.3)** Chi-square test for the difference between two proportions **(Section 12.1)** McNemar test for the difference between two proportions in related samples **(Section 12.4)**
Comparing more than two groups	One-way analysis of variance **(Section 11.1)** Randomized block design **(Section 11.2)** Two-way analysis of variance **(Section 11.3)** Kruskal-Wallis test **(Section 12.7)** Friedman rank test **(Online Topic 12.9)**	Chi-square test for differences among more than two proportions **(Section 12.2)**
Analyzing the relationship between two variables	Scatter plot, time series plot **(Section 2.6)** Covariance, coefficient of correlation **(Section 3.5)** Simple linear regression **(Chapter 13)** *t* test of correlation **(Section 13.7)** Time series forecasting **(Chapter 16)**	Contingency table, side-by-side bar chart, **(Sections 2.2, 2.4)** Chi-square test of independence **(Section 12.3)**
Analyzing the relationship between two or more variables	Multiple regression **(Chapters 14 and 15)**	Multidimensional contingency tables **(Section 2.7)** Logistic regression **(Section 14.7)**

Source: Basic Business Statistics, https://goo.gl/N4z2VQ

RESOURCES

Want to dig deeper? Check out these resources:

- Using Excel's Analysis ToolPak Add-In: http://www.nvc.vt.edu/rmajor/bit5724/AnalysisToolPakGuide.pdf
- Statistical Analysis Handbook by Dr. M J de Smith (Found in your Student Resource folder)
- Basic Business Statistics, Concepts and Applications by Berenson, Levine, Krehbiel. 12th Edition. (Available for free download online https://goo.gl/N4z2VQ)
- An Introduction to Statistical Learning with applications in R https://goo.gl/6AUuno