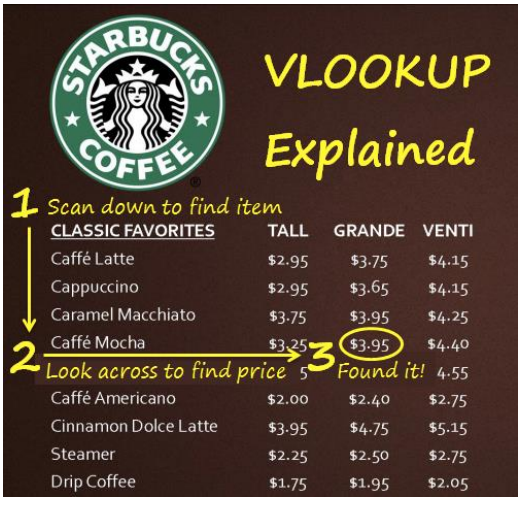
DATA ANALYTICS
Business • Statistics • Coding

Session 1

Session 1 - Dataset 1 - Student Performance.csv

1. Open the .csv file and save it as an .xlsx file.
2. (Bold, color, freeze top row) Before you work on a new table, do the following. **Ctrl+Shift+right** arrow to select the first row. Bold it. Fill Color black background. Font Color white text. View > Freeze Top Row.
3. (**Auto fit columns) Ctrl+A or Ctrl+Shift+right arrow+down** arrow to select the entire table, or click select all button at top-left corner to select the entire sheet, hover over between any two columns, double click.
4. Data > Filter, check unique values of each column.
5. Data > Sort by Categorical variable from A to Z or from Z to A, **by Numerical variable from Smallest to Largest or from Largest to Smallest.**
6. Select column F:H, Home > Conditional Formatting > New Rules > Icon Sets > 3 Traffic Lights > Display green when value is >= 70 Number, yellow when value is <70 and >= 50 Number, red when value < 50. Option: Show icon only.
7. Average I2=ROUND(AVERAGE(F2:H2),0), Ctrl+Shift+! Number, No decimal. Min J2=MIN(F2:H2). Max K2=MAX(F2:H2). Select cell I2:K2, a Fill Handle (+) will appear as a small square at bottom-right corner of the selected cell, click it! and copy formula down to the last row. Format Painter copies format from column F and applies it to column I-K. Select column I. Ctrl+Shift+! to format number no decimal.
8. Insert a column between B:C and name it group. Functions UPPER(), LOWER(), PROPER(), TRIM(). C2=TRIM(PROPER(B2)).
9. Insert a Sheet and rename it Metrics. SUMIFS(sum_range, criteria_range, criteria). COUNTIFS(count_range, criteria_range, criteria). AVERAGEIFS(average_range, criteria_range, criteria). Difference between relative and absolute cell references is that relative cell references move when you copy them, but absolute references don't. Use F4 to toggle through 4 combinations A1, A1, A$1, $A1. Copy > Past Special, Formula > Ctrl+H replace "AVERAGE" with "MIN" or "MAX". Always add a title "Student Performance Metrics", black background, white text, Merge & Center, font 12. File > Page Setup, Landscape, 1 page wide 1 page tall. Save selection as PDF.
10. Review > Protect Sheet or Protect Workbook > Password. Business Example: Send data that includes personal identifiable information (PII) in password protected Excel Workbook, effective in Protect Sheet. How to protect cells? Select the cells you want to lock > Format Cells > Protection > Locked, Review > Protect Sheet.
11. VLOOKUP or vertical lookup is a function that **looks up the value in one column** and **returns the corresponding value from another column**. It's like ordering coffee at Starbucks: I scan down to find my Caffe Mocha, then I look across to find its price, there it is, I found it! =VLOOKUP("Caffe Mocha",A:D,3,false)=VLOOKUP(lookup **value**, lookup **range**, **column#** that contains **return value**, exact or approximate match). (Column#, Row#) = C5 = $3.95.



DATA ANALYTICS
Business • Statistics • Coding

Session 2

Session 1 - Dataset 3 - Google Play Store.csv

1. **Look at data the way a detective examines a crime scene.**
2. Data: summary or detailed? Detailed. This table has the most granular level of details (LOD) about Google Play apps. It has 10841 rows (observations) and 11 columns (variables). Each row represents an app and 10 attributes associated with it. After you clean the data, you aggregate (summarize) the rows. Depending on data type of the columns, you group the rows by "dimensions" columns and aggregate "measures".
3. Excel data types: General, Text, Date; categorical and numerical; discrete and continuous; dimensions vs measures.
4. Find and clean anomalies in each column: App (#NAME), Category (1.9), Rating (NaN, 19), Reviews (3.0M), Size (Varies with device), Installs (Free), Type (0, NaN), Price (Everyone), Content Rating (Blanks), Genres (11-Feb-18), Last Updated (1.0.19). NaN means "not a number". Obviously, Row ??? was mis-aligned. Move B:J to C:K.
5. Any variable with too many distinct values isn't suitable for grouping. Type has 4 distinct value: 1 (1 row), Free (10039 rows), NaN (1 row), Paid (800 rows). You can delete 1 (1 row) and NaN (1 row) because 2 out of 10841 rows are insignificant for analysis.
6. Also look for consistency: when Type = "Free", Price = 0
7. Column B (Category). Replace "_" with space. Insert a column between B and C. C2=TRIM(PROPER(B2)). Click Fill Handle. Copy column C and paste value column B. Delete column C.
8. Column E (Size). Ctrl+H, find what, replace with, select "Find entire cells only"

a. Select column E > Ctrl+H, find "M" replace with 000000, replace all

b. Select column E > Ctrl+H, find "K" replace with 000, replace all

c. L2 = IF(E2<=10,E2*1000000,E2), click Fill Handle to copy formula down. Copy column L and Paste Special > Values column E.

9. Column F (Installs). Replace "+" with "".
10. Column J (Genres). Insert a column to its right. Data > Text to Columns > Delimited > Semicolon > Text, Text > Finish. Delete column L. Now Genres have a lot less distinct values.
11. Column K (Date Updated). Right click > Format Cells > Custom > yyyy-mm-dd.
12. Create a PivotTable. Filters: Type ="Paid". Rows: Category. Values: average(Rating), average(Review), average(Size), average (Installs), average(Price), max(Date_Updated) --- format to number with 1/0/0/0/2 decimals and yyyy-mm-dd, respectively. Place cursor anywhere in PivotTable, click Design > Pivot Style, Row/Column Headers, Banded Rows/Columns.
13. We usually group data by a "categorical" variable. But we can group continuous values of a "numerical" variable. Group Reviews, insert a column to right of column D, do nested IF's statement: E2=IF(D2>100000,">100K",IF(D2>10000,">10K",IF(D2>1000,">1K","<=1K"))).
14. Create a PivotChart. Filters: Type = "Paid". Rows: Category. Values: average(Price). PivotChart: select a bar chart, title "Average Price by Category". Select Print Area > Page Setup > Landscape, Fit to 1 Page Wide by 1 Tall

Session 3

Session 2 - Dataset 2 - Kickstarter Report (2018-01).xlsx

1. Macro: file > Option > Customize Ribbon > Developer, check!
2. Macro name=Clean_Data, Store macro in This Workbook, Shortcut key: Option+Cmd+Q, Save AS Macro-Enabled Workbook, Visual Basic to debug
3. Delete column B (name) and C (category)
4. TRIM() column A (ID), B (main_category), C (currency), TRIM(PROPER()) column H (state)
5. Text to Columns column F (Deadline) > Date: YMD

6. Ctrl+Shift+$ to reformat column K (usd pledged) to currency
7. Add column L (Country Full Name): INDEX(A:A, MATCH(J2, B:B,0))
8. 1-way: MATCH() returns row# in a column. INDEX() returns value in another column based on same row#. A5 → C5 = $3.95
9. 1-way: MATCH() returns column# in a row. INDEX() returns value in another row based on same column#. C1 → C5 = $3.95
10. 2-way: First MATCH() returns column#, second MATCH() returns row#. INDEX() returns value of a cell in a range based on same (Column#, Row#). C5 = $3.95
11. Heading: a descriptive title. As of: date
12. Put your cursor in slicer, right click > Size and Properties > Format Slicer > Position > Disable resizing & Moving
13. **(Dashboard)** At first, I gave my manager a static report and he asked me: What about this? What about that? Then I spent all day producing different reports to answer different questions. I'm like: wait a minute, all these reports are the same measures for different combinations of dimensions. So instead of giving him many reports, I gave him a dynamic, self-serve dashboard. As long as it is simple, intuitive and straightforward, he was happy to use it. It answered most of his questions and also freed up my time!

Interview Questions

1. **What's VLOOKUP?** Lookup a value in one column, go to the other column, bring back a corresponding value.
2. **What's your favorite Excel function?** SUMIFS and COUNTIFS.
3. **What would you use PIVOTTABLE for?** To summarize measures by dimensions, drill down and roll up, slide and dice, sort and rank, chart and slicer and dashboard.
4. **What's the difference between VLOOKUP and INDEX MATCH?** Why INDEX MATCH is better Than VLOOKUP? VLOOKUP breaks easily. INDEX MATCH doesn't break. You don't get incorrect result when a column is inserted or deleted with INDEX MATCH, and you can look up to the left.
5. **What is VLOOKUP and how does it work?** VLOOKUP or vertical lookup is a function that looks up the value in one column and returns the corresponding value from another column. It's like order coffee at Starbucks: scan down to find my coffee latte, look across to find price, found it! VLOOKUP(lookup value, lookup range, number of column to the right that contains return value, exact or approximate match).
6. **What is the difference between a histogram and a bar chart?** A histogram shows frequency distribution of a continuous (numeric) variable. A bar chart compares a continuous (numeric) variable between a discrete (categorical) variable. The histogram has no gap between bars. The bar chart has gaps between bars. The histogram has only one variable. The bar chart has two variables.
7. **What is the difference between a line chart and a scatter plot?**
8. Excel is the most widely used software in business. It is also the largest and most successfully implemented CMDB in the world.
9. **While many people claim to "know" Excel, our experience has shown that most are barely scraping the surface of Excel's capabilities.**
10. Heeral teaches you 20% essential skills to solve 80% common problems e.g. data prep.
11. **Data in a table have one variable per column, one observation per row, one value per cell.**
12. Teach me an EXCEL trick! This is a good one. How to repeat value in blank cells until next value? Select a range (A1:A100). Press Ctrl+G. Go To Special dialog box appears. Select Blanks. Type in Formula Bar address of the first non-blank cell (=A2), and press Ctrl-Enter. Ta-Da!

13. **Functions: Pivot Table/Pivot Chart/Slicer, VLookup / HLookup / Match / Index / Offset, Indirect, SumIfs / CountIfs, Count / Counta, Conditional Formatting, Data Validation, Filter, Sort, Nested If, Iserror, Iferror, Text to Columns, What-If Analysis, Goal Seek, Scenario Manager, Data Table, Solver**
14. Statistical Analysis: Moving Average, Hypothesis Testing, Anova, Covariance, Correlation, Regression, and Normal Distribution
15. **(5-minute Excel challenge)** 1) create a mortgage amortization schedule in Excel to calculate monthly payment based on interest rate, amortization period, and mortgage amount; 2) create 4 data tables to show impact of changing interest rate, amortization period, and mortgage amount on monthly payment.
16. **(Sales Variance Analysis using Excel)** Answer the following questions using data in the "Raw Data" tab: 1) Summarize in a table actual & budgeted margins (in $ and %) by category, product, customer. 2) Calculate volume & price variance and find root cause(s) by product and customer. 3) Based on your analysis what's your recommendation to improve Product 3 sales performance?
17. When my son was 10, we used to go to Browsers Den to buy magic tricks. Browers Den is Toronto's oldest magic shop and has been selling magic since 1975. You pay 20 bucks, they give you a pen and how you how to make a pen penetrate a dollar bill...
18. Looped square (⌘), place of interest sign, command key symbol.
19. Number of worksheets in a workbook is "limited only by available memory".
20. Total number of rows and columns on a worksheet: 1,048,576 rows by 16,384 columns.