

AUTOMATING THE DATA PREPARATION PROCESS USING R PROGRAMMING

ZACHARY WEST - BOOZ ALLEN HAMILTON

AGENDA

-
- ▶ INTRODUCTION
 - ▶ WHAT IS R PROGRAMMING?
 - ▶ UTILIZING R PROGRAMMING
 - ▶ AUTOMATION EXAMPLES
 - ▶ PROCESS OF AUTOMATING DATA PREPARATION
 - ▶ BENEFITS & LIMITATIONS
 - ▶ PROPER DOCUMENTATION IS KEY
 - ▶ R PROGRAMMING RESOURCES
 - ▶ QUESTIONS
-

INTRODUCTION

Objective:

- Enable cost estimators to start thinking about adding automation into their day-to-day activities, by providing examples on automating the data preparation process

Purpose:

- Data Analysts spend a significant amount of their time gathering, cleaning, and organizing their data
- Automation provides a different viewpoint on how recurring reports can be gathered and compiled

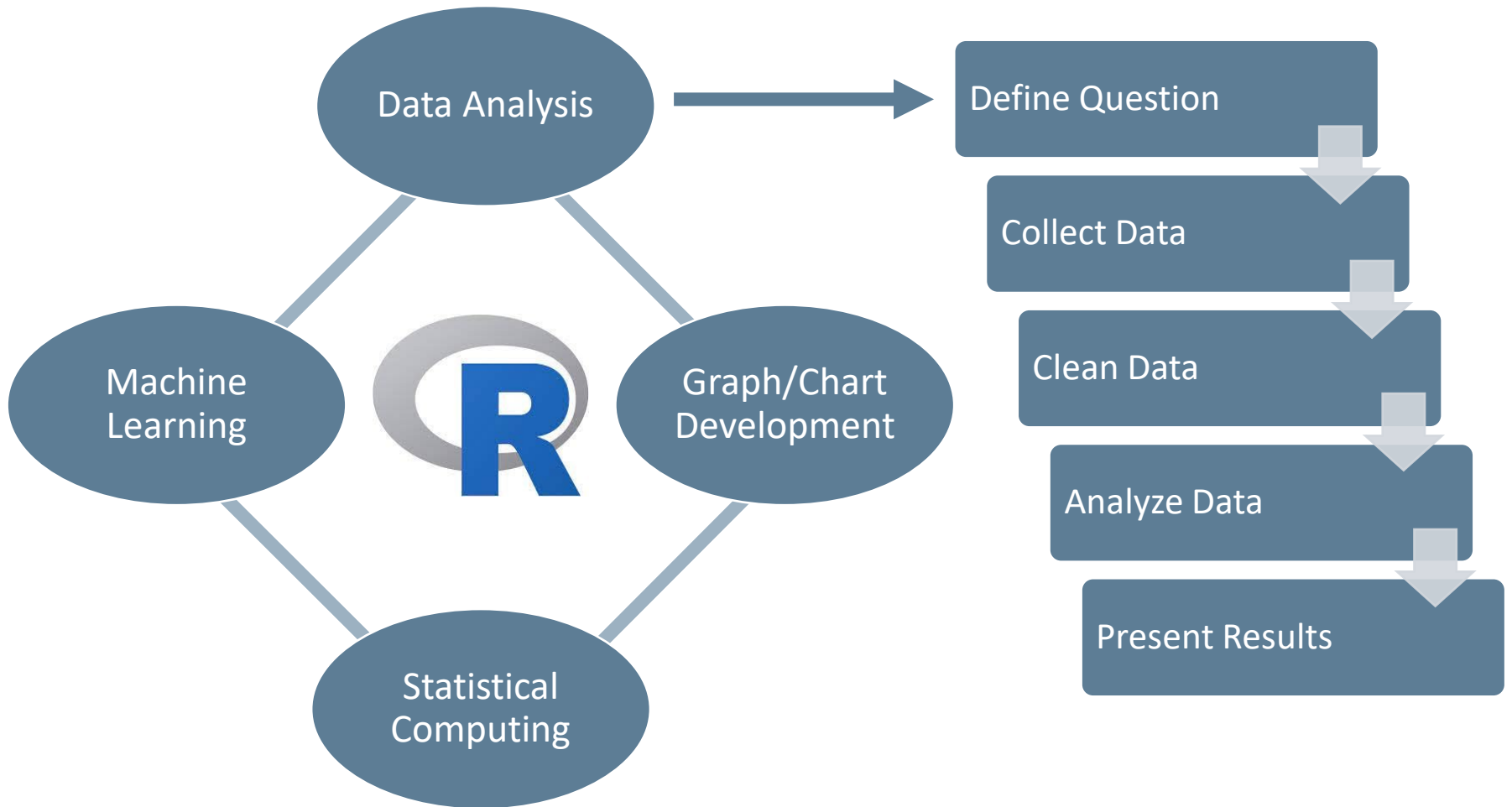
How could you automate a monotonous task using R Programming?

WHAT IS R PROGRAMMING?

- R (Programming) is “a language and environment for statistical computing and graphics”
- R is a universal programming language compatible with Windows, Macintosh, UNIX, and Linux platforms
- R is open-source, and no fees or licenses are needed
- R keeps evolving and growing, increasing in capability through industry updates
- R-Studio provides a user-friendly, integrated development environment (IDE) for R Programming



UTILIZING R PROGRAMMING

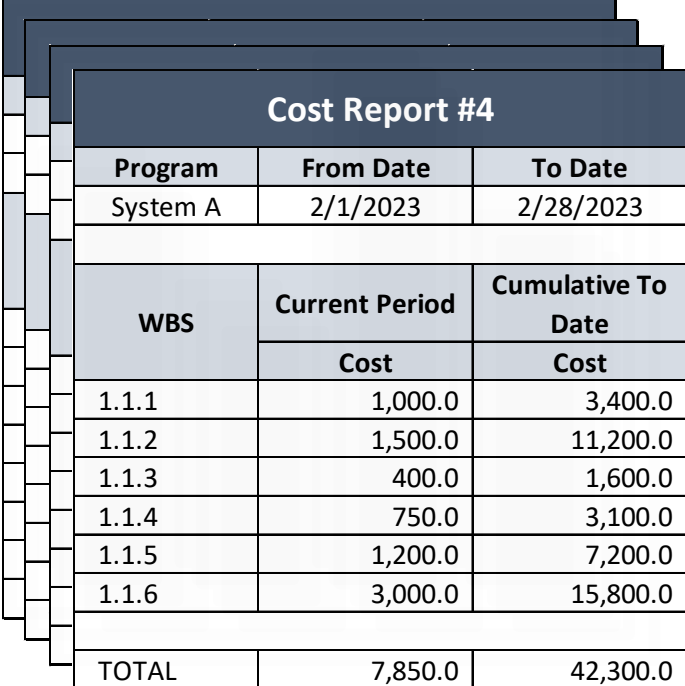


References:

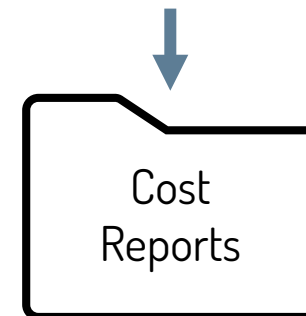
- https://www.simplilearn.com/what-is-r-article#what_is_r_used_for
- <https://careerfoundry.com/en/blog/data-analytics/the-data-analysis-process-step-by-step/>

EXAMPLE – AUTOMATE GATHERING

- I recently found monthly reports with relevant cost and schedule data, but it is tedious, and time consuming to gather and then combine all data into one report
- **Q: How can I gather these reports efficiently month after month?**
- **A:** By setting a standard file directory, my R script will gather all files found in the folder, no matter how many there are at the time

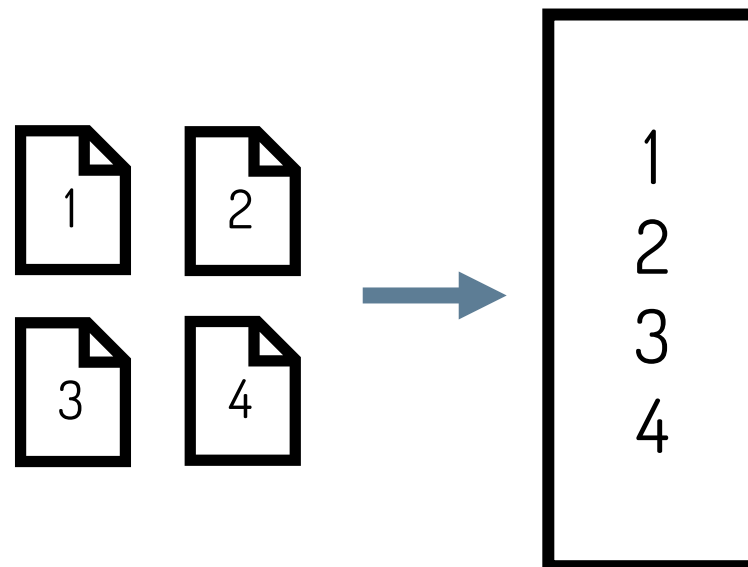


Cost Report #4		
Program	From Date	To Date
System A	2/1/2023	2/28/2023
WBS	Current Period	Cumulative To Date
	Cost	Cost
1.1.1	1,000.0	3,400.0
1.1.2	1,500.0	11,200.0
1.1.3	400.0	1,600.0
1.1.4	750.0	3,100.0
1.1.5	1,200.0	7,200.0
1.1.6	3,000.0	15,800.0
TOTAL	7,850.0	42,300.0



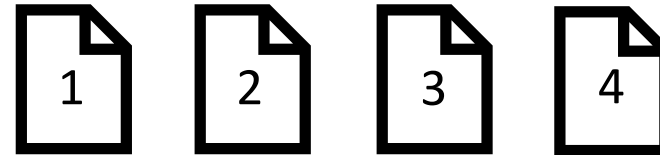
EXAMPLE – AUTOMATE COMBINING

- R has now located and is familiar with all cost reports to date, but I don't have the time or resources to combine them all one by one
- **Q: How can I efficiently combine all these reports?**
- A: By using names, dates, and the length of a report as indicator variables, my R script can stack the reports on top of one another in order by date to create one large data frame



EXAMPLE – AUTOMATE CLEANING

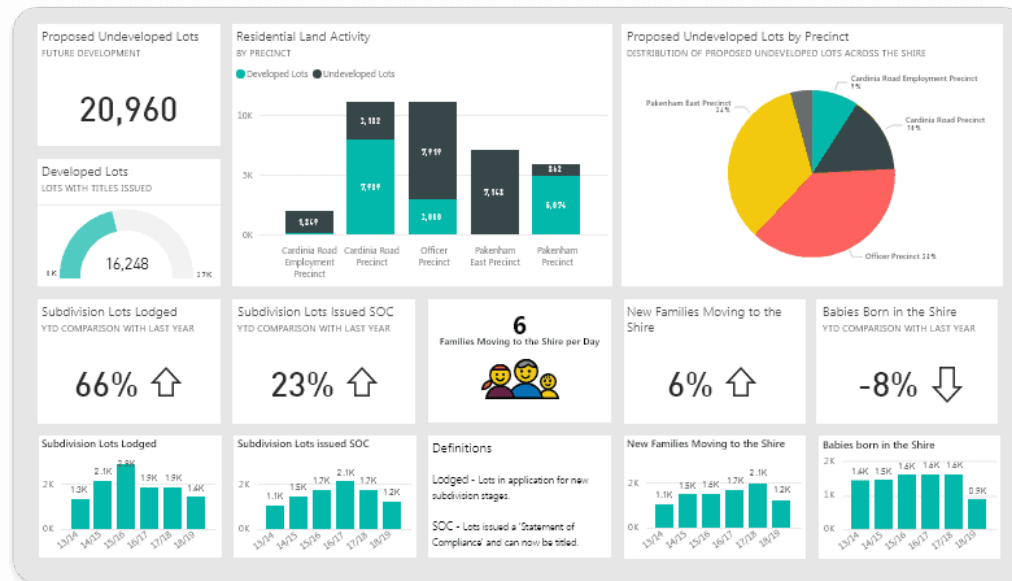
- Now that I’ve combined all the reports, I’m left with a large data frame of disorganized data
- **Q: How can I clean this data to get it to a desirable state?**
- A: By using simple R data frame manipulation, I can clean and structure the data based on how I want to present the results in the data dashboard



Combined Report				
Report	1	2	3	4
From Date	11/1/2022	12/1/2022	1/1/2023	2/1/2023
To Date	11/30/2022	12/31/2022	1/31/2023	2/28/2023
Cost_Current				
1.1.1	400.0	1,200.0	800.0	1,000.0
1.1.2	4,000.0	3,500.0	2,200.0	1,500.0
1.1.3	400.0	500.0	300.0	400.0
1.1.4	600.0	850.0	900.0	750.0
1.1.5	1,000.0	3,200.0	1,800.0	1,200.0
1.1.6	4,300.0	4,400.0	4,100.0	3,000.0
Cost_Cumulative				
1.1.1	400.0	1,600.0	2,400.0	3,400.0
1.1.2	4,000.0	7,500.0	9,700.0	11,200.0
1.1.3	400.0	900.0	1,200.0	1,600.0
1.1.4	600.0	1,450.0	2,350.0	3,100.0
1.1.5	1,000.0	4,200.0	6,000.0	7,200.0
1.1.6	4,300.0	8,700.0	12,800.0	15,800.0

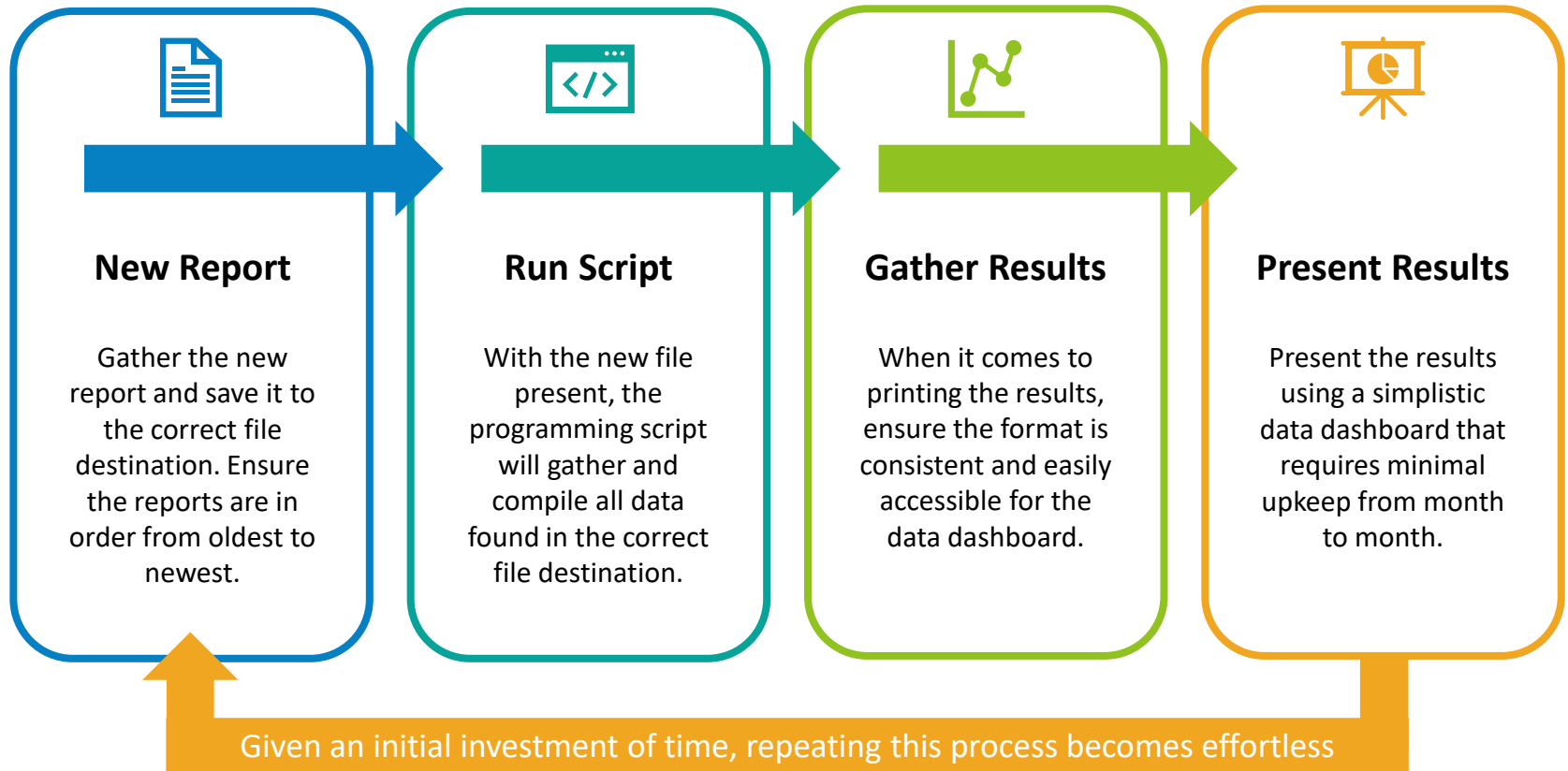
EXAMPLE – AUTOMATE PRESENTATION

- Now the data is in a clean and structured data frame ready to be used in analysis
- **Q: How can I efficiently present this data in a dashboard?**
- **A:** By utilizing software like R Markdown, Tableau, Power BI, and Microsoft Excel, I can create a dashboard where my data can be easily inserted every month



Source: <https://www.syskit.com/blog/power-bi-dashboards-vs-reports/>

AUTOMATING DATA PREPARATION



BENEFITS & LIMITATIONS

Benefits



- Saves a substantial amount of time and effort in the long run
- Improves analyst's technical skills
- Shifts focus from developing reports to generating decisions

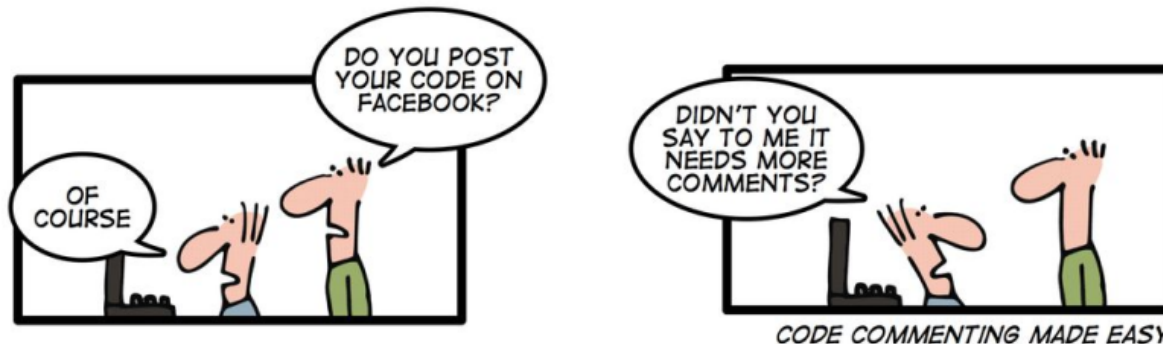
Limitations



- Demands an up-front time investment
- Requires the documents to follow a consistent structure

PROPER DOCUMENTATION IS KEY

- Proper documentation:
 - Provides a good knowledge transfer to both new and experienced programmers
 - Limits the program's maintenance efforts needed, saving time in the long run
- Documentation Examples:
 - README files
 - How-to Guides
 - Various in-code comments (notes, explanations, etc.)



Source: <https://github.com/ybouz2/project-tech/wiki/Coding-standart>

R PROGRAMMING RESOURCES

- Downloading R: <https://cran.r-project.org/bin/windows/base/>
 - Downloading R Studio: <https://posit.co/download/rstudio-desktop/>
 - (Free) Courses:
 - [R Programming](#) – Johns Hopkins University
 - [Statistics with R Specialization](#) – Duke University
 - [Introduction to R](#) – Datacamp
 - [Learn R](#) – Codecademy
 - [Top 100 R Tutorials](#) – Listen Data
 - [Getting Started with R Programming](#) – Data Flair
 - Google
 - YouTube
-

QUESTIONS?