CSDE 502 2021 **Assignment 1** **Introduction to R Markdown** **Instructor: Phil Hurvitz** **phurvitz@uw.edu**	**My Name: *answer key*** **My UWNetID:**

Due Date: 2021-01-15 09:00

Instructions:

1. Fill out your name and UWNetID at the top of this page.
2. Add your answers to this document.
3. Use the "00Answers" Word style for your answers so they will be clearly discernible from the questions.
4. When you are completed with your work, create a PDF from the Word document.
5. Name your PDF document with the pattern UWNetID_HW_n.pdf where UWNetID is your UWNetID and n is the week number when the homework was assigned. For example, the first assignment by me would be named *phurvitz_HW_01.pdf*. Upload your completed document to Canvas.
6. Upload any specified data or code files to Canvas.

Explanation:

For this assignment, you will be looking in detail at the file week_01.Rmd (http://staff.washington.edu/phurvitz/csde502_winter_2021/files/week_01.html), examining some code that creates a set of files with different names, and suggesting better file names for a list of files.

If you are not familiar with R Markdown format, these resources will be helpful, or others you can find by performing a web search for "R Markdown":

https://rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf
https://www.dataquest.io/blog/r-markdown-guide-cheatsheet/

Within RStudio, you can also see:

Help > Cheatsheets > R Markdown Cheat Sheet
Help > Cheatsheets > R Markdown Reference Guide

Questions:

1. For week_01.Rmd:
   1.1. Explain what the **output** code in the front matter/YAML header of the Rmd file does.

The intention of this question was to get you to find out what options are available in generation of the output from Rmd.

See https://bookdown.org/yihui/rmarkdown/html-document.html for details on the YAML header.

The YAML header includes this text for "output":

```
output:
    html_document:
        number_sections: true
        self_contained: true
        code_folding: hide
        toc: true
        toc_float:
            collapsed: true
            smooth_scroll: false
    pdf_document:
        number_sections: true
        toc: true
        fig_cap: yes
        keep_tex: yes
```

There are two output formats that are specified to have options, HTML and PDF, and the values control how the outputs are configured. For the HTML output:

- **Sections should be numbered. That is, using "#", "##", etc. will produce dot-delimited hierarchical numbers at the left of the section heading.**
- **"self_contained: true" includes all code in a single HTML file; if this is not the case then the output is placed in different files and directories. Self-containment is better for sharing the output.**
- **"code_folding: hide" makes all of the code chunks hidden by default. They can be shown by clicking on the "Code/Hide" toggles.**
- **"toc: true" includes a table of contents (TOC) at the upper left.**
- **"toc_float:" makes the TOC always visible as the file is scrolled.**
- **"collapsed: true" controls what level of TOC hierarchy is shown; levels will expand as necessary.**
- **"smooth_scroll: false" (defaults to TRUE) controls whether page scrolls are animated when TOC items are navigated to via mouse clicks, versus just jumping to the clicked section.**

For the PDF output:

- **"number_sections: true" (same as above)**
- **"toc: true" (same as above)**
- **"fig_cap: yes" controls whether figure captions are printed**
- **"keep_tex: yes" When the Rmd file is rendered to PDF, an intermediate TeX file is created. This controls whether the intermediate file is deleted or not. Can be useful for troubleshooting.**

1.2. Explain how the **kableExtra** package is used in the Rmd file.

The intention of this question was to get you to look at the documentation of the kableExtra package so you can be aware of different ways you can print tables in your output.

See https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html.

The kable_styling for this table is

```
kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
full_width = F, position = "left")
```

This shows that the rows should be striped (i.e., grey and white), and they should highlighted when the mouse pointer hovers over a record. "condensed" has slightly shorter row height than the default. "responsive" allows the table to be scrolled if the window size is decreased. Other options are "full_width = F" so the table will not take up the entire window width if not necessary, and 'position = "left"' makes the table left-justified.

1.3. Describe how captions and cross-references are implemented in this Rmd with the **captioner** package.

Captioner was specified in the setup chunk to use the prefixes "Table" and "Figure" (i.e., one could use "Fig." if desired):

```
table_nums <- captioner(prefix = "Table")
figure_nums <- captioner(prefix = "Figure")
```

Captions are managed by their "name". For example, see the code

```
See `r table_nums(name = "tcap0", display = "cite")`
_`r table_nums(name = "tcap0", caption = "My caption")`_
```

The table caption named "tcap0" is created and will be consistently numbered as "Table 1". In the first line, the "display" option creates a citation, i.e., "Table 1", and in the second line, the full caption is printed as "Table 1: My caption". It is rendered in italics by the underscores at the start and end of the text. Both are printed inline by running the r code as `` `r some_R_code` `` construction.

2. Download and run the file http://staff.washington.edu/phurvitz/csde502_winter_2021/files/file_naming.R and briefly comment on different approaches to naming files.

When files are named using dates, if the pattern 'yyyymmdd' or 'yyyy-mm-dd' or 'yyyy_mm_dd' is not used, sorting by file name does not sort file names in temporal order.

3. Download the file http://staff.washington.edu/phurvitz/csde502_winter_2021/files/example_filenames.csv and produce a table with the original file name and your suggestion for a better file name.

The file names are a hodgepodge using no standard. Problems with the file names include:

- **internal dots, e.g., "v.20"**
- **specification by version number or date used inconsistently**
- **no date at all, but some with versions that seem to have been modified on the same date**
- **spaces and underscores used inconsistently**
- **dates using multiple patterns, e.g., "Nov 30", "December 10", "12_11_2020", "1215". This is particularly problematic for Dec < Nov in alphabetic order but 12 > 11 in numeric order.**
- **mixed upper and lower case**

Additionally, using directories might be worthwhile to store the different locations' data. That might be overkill with this relatively small set of files, but if the list of files is expected to grow, a directory structure would be suggested.

"Fixing" these file names (Table 1) may not truly be possible, particularly without knowing more about the file contents. At least those files with the same first part of the name (e.g., "pitt_phone_numbers_") would sort properly (see Table 2 and Table 3). This list of poorly named files should serve as an example of how using a consistent file naming strategy is a good idea.

Table 1: File names with suggested corrections

date	file_name	suggested_name
2020-10-13	Instrument_v.20 - bilingual.docx	instrument_v20_bilingual_2020-10-13.docx
2020-11-30	Master_Durham.csv	master_durham_2020-11-30.csv
2020-11-30	Master_Durham.xlsx	master_durham_2020-11-30.xlsx
2020-12-04	Durham emails_Nov 30.xlsx	durham_emails_2020-11-30.xlsx
2020-12-04	NJ emails_Nov 30.xlsx	nj_emails_2020-11-30.xlsx
2020-12-07	Durham tasks_Nov 30.xlsx	durham_tasks_2020-11-30.xlsx
2020-12-07	Master_Durham_v2.csv	master_durham_v2_2020-12-07a\.csv
2020-12-07	Master_Durham_v2.xlsx	master_durham_v2_2020-12-07.xlsx
2020-12-07	NJ emails_December 7.xlsx	nj_emails_2020-12-07.xlsx
2020-12-07	NJ phone numbers_Nov 30.xlsx	nj_phone_numbers_2020-11-30.xlsx
2020-12-07	Pitt phone numbers_December 7.xlsx	pitt_phone_numbers_2020-12-07.xlsx
2020-12-07	Pitt phone numbers_Nov 30.xlsx	pitt_phone_numbers_2020-11-30.xlsx
2020-12-11	Fraud and quota report_v1.xlsx	fraud_and_quota_report_v1_2020-12-11.xlsx
2020-12-11	Fraud and quota report_v2.xlsx	fraud_and_quota_report_v2_2020-12-11.xlsx
2020-12-11	Rutgers_people_to_pay_12_11_2020.xlsx	rutgers_people_to_pay_2020-12-11.xlsx
2020-12-15	Durham emails_Dec 14.xlsx	durham_emails_2020-12-14.xlsx
2020-12-15	Durham phone numbers_Dec 14.xlsx	durham_phone_numbers_2020-12-14.xlsx
2020-12-15	Export_1215.7z	export_2020-12-15.7z
2020-12-15	NJ emails_December 14.xlsx	nj_emails_2020-12-14.xlsx
2020-12-15	NJ phone numbers_December 14.xlsx	nj_phone_numbers_2020-12-14.xlsx
2020-12-15	VR_Dec1420_Durham.xlsx	vr_durham_2020-12-14.xlsx
2020-12-15	VR_NJ_Dec1420.xlsx	vr_nj_2020-12-14.xlsx
2021-01-05	transfer_0105.zip	transfer_2021-01-05.zip
2021-01-07	Fraud and quota report_v3.xlsx	fraud_and_quota_report_v3_2021-01-07.xlsx
2021-01-07	Fraud and quota report_v4.xlsx	fraud_and_quota_report_v4_2021-01-07.xlsx
2021-01-08	test1.dta	test1_2021-01-08.dta

Table 2: File names with suggested corrections, sorted by original file name

date	file_name	suggested_name
2020-12-15	Durham emails_Dec 14.xlsx	durham_emails_2020-12-14.xlsx
2020-12-04	Durham emails_Nov 30.xlsx	durham_emails_2020-11_30.xlsx
2020-12-15	Durham phone numbers_Dec 14.xlsx	durham_phone_numbers_2020-12-14.xlsx
2020-12-07	Durham tasks_Nov 30.xlsx	durham_tasks_2020-11-30.xlsx
2020-12-15	Export_1215.7z	export_2020-12-15.7z
2020-12-11	Fraud and quota report_v1.xlsx	fraud_and_quota_report_v1_2020-12-11.xlsx
2020-12-11	Fraud and quota report_v2.xlsx	fraud_and_quota_report_v2_2020-12-11.xlsx
2021-01-07	Fraud and quota report_v3.xlsx	fraud_and_quota_report_v3_2021-01-07.xlsx
2021-01-07	Fraud and quota report_v4.xlsx	fraud_and_quota_report_v4_2021-01-07.xlsx
2020-10-13	Instrument_v.20 - bilingual.docx	instrument_v20_bilingual_2020-10-13.docx
2020-11-30	Master_Durham.csv	master_durham_2020-11-30.csv
2020-11-30	Master_Durham.xlsx	master_durham_2020-11-30.xlsx
2020-12-07	Master_Durham_v2.csv	master_durham_v2_2020-12-07.csv
2020-12-07	Master_Durham_v2.xlsx	master_durham_v2_2020-12-07.xlsx
2020-12-15	NJ emails_December 14.xlsx	nj_emails_2020-12-14.xlsx
2020-12-07	NJ emails_December 7.xlsx	nj_emails_2020-12-07.xlsx
2020-12-04	NJ emails_Nov 30.xlsx	nj_emails_2020-11-30.xlsx
2020-12-15	NJ phone numbers_December 14.xlsx	nj_phone_numbers_2020-12-14.xlsx
2020-12-07	NJ phone numbers_Nov 30.xlsx	nj_phone_numbers_2020-11-30.xlsx
2020-12-07	Pitt phone numbers_December 7.xlsx	pitt_phone_numbers_2020-12-07.xlsx
2020-12-07	Pitt phone numbers_Nov 30.xlsx	pitt_phone_numbers_2020-11-30.xlsx
2020-12-11	Rutgers_people_to_pay_12_11_2020.xlsx	rutgers_people_to_pay_2020-12-11.xlsx
2021-01-08	test1.dta	test1_2021-01-08.dta
2021-01-05	transfer_0105.zip	transfer_2021-01-05.zip
2020-12-15	VR_Dec1420_Durham.xlsx	vr_durham_2020-12-14.xlsx
2020-12-15	VR_NJ_Dec1420.xlsx	vr_nj_2020-12-14.xlsx

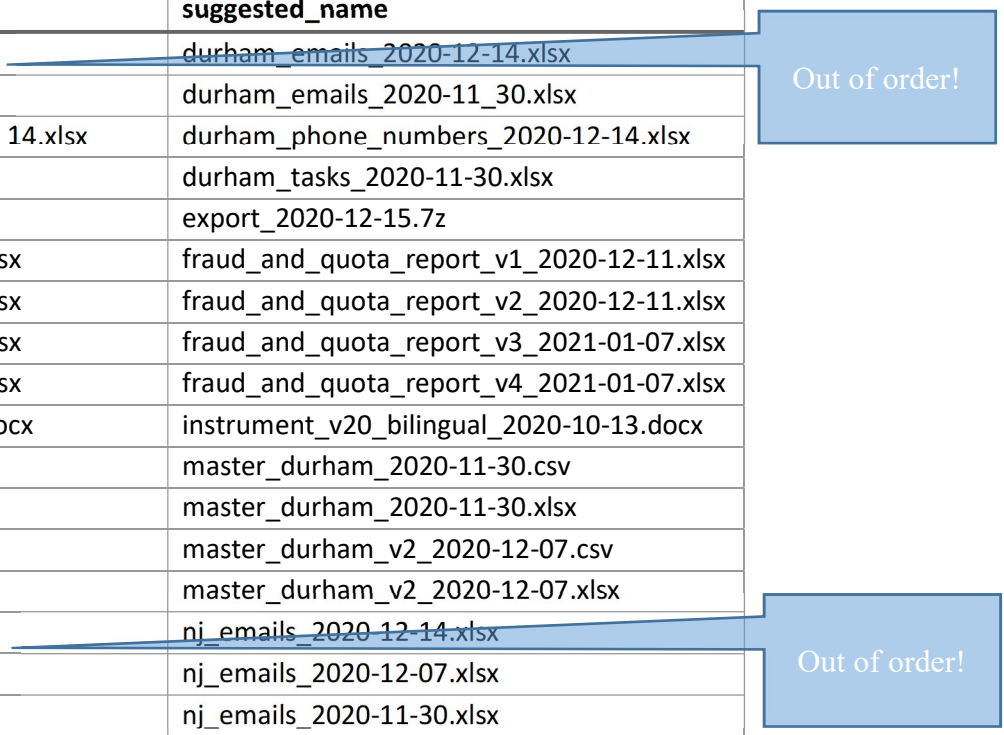

Table 3: File names with suggested corrections, sorted by suggested name

date	file_name	suggested_name
2020-12-04	Durham emails_Nov 30.xlsx	durham_emails_2020-11_30.xlsx
2020-12-15	Durham emails_Dec 14.xlsx	durham_emails_2020-12-14.xlsx
2020-12-15	Durham phone numbers_Dec 14.xlsx	durham_phone_numbers_2020-12-14.xlsx
2020-12-07	Durham tasks_Nov 30.xlsx	durham_tasks_2020-11-30.xlsx
2020-12-15	Export_1215.7z	export_2020-12-15.7z
2020-12-11	Fraud and quota report_v1.xlsx	fraud_and_quota_report_v1_2020-12-11.xlsx
2020-12-11	Fraud and quota report_v2.xlsx	fraud_and_quota_report_v2_2020-12-11.xlsx
2021-01-07	Fraud and quota report_v3.xlsx	fraud_and_quota_report_v3_2021-01-07.xlsx
2021-01-07	Fraud and quota report_v4.xlsx	fraud_and_quota_report_v4_2021-01-07.xlsx
2020-10-13	Instrument_v.20 - bilingual.docx	instrument_v20_bilingual_2020-10-13.docx
2020-11-30	Master_Durham.csv	master_durham_2020-11-30.csv
2020-11-30	Master_Durham.xlsx	master_durham_2020-11-30.xlsx
2020-12-07	Master_Durham_v2.csv	master_durham_v2_2020-12-07.csv
2020-12-07	Master_Durham_v2.xlsx	master_durham_v2_2020-12-07.xlsx
2020-12-04	NJ emails_Nov 30.xlsx	nj_emails_2020-11-30.xlsx
2020-12-07	NJ emails_December 7.xlsx	nj_emails_2020-12-07.xlsx
2020-12-15	NJ emails_December 14.xlsx	nj_emails_2020-12-14.xlsx
2020-12-07	NJ phone numbers_Nov 30.xlsx	nj_phone_numbers_2020-11-30.xlsx
2020-12-15	NJ phone numbers_December 14.xlsx	nj_phone_numbers_2020-12-14.xlsx
2020-12-07	Pitt phone numbers_Nov 30.xlsx	pitt_phone_numbers_2020-11-30.xlsx
2020-12-07	Pitt phone numbers_December 7.xlsx	pitt_phone_numbers_2020-12-07.xlsx
2020-12-11	Rutgers_people_to_pay_12_11_2020.xlsx	rutgers_people_to_pay_2020-12-11.xlsx
2021-01-08	test1.dta	test1_2021-01-08.dta
2021-01-05	transfer_0105.zip	transfer_2021-01-05.zip
2020-12-15	VR_Dec1420_Durham.xlsx	vr_durham_2020-12-14.xlsx
2020-12-15	VR_NJ_Dec1420.xlsx	vr_nj_2020-12-14.xlsx