Exam 02 Closed Notes

DSST 289: Introduction to Data Science

1 Honor

You may only use a pen/pencil and scratch paper on this exam.

> "I pledge that I will neither give nor receive unauthorized assistance during the completion of this work."

Name_____

Signature_____

Section start time_____

2 Exam

Please write neatly.

If you cannot solve a problem, write what you do know about the question to maximize partial credit.

Your code will be graded on its quality, which includes both accuracy and proper formatting.

3 Data

We will use tables about music for this exam.

Table 1: R object name: `table1`

artist	song	star_rating
Taylor Swift	Blank Space	2
Drake	Hotline Bling	3
Adele	Easy On Me	4
Radiohead	The National Anthem	5
The Smile	Thin Thing	5

Table 2: R object name: `table2`

artist	lead_performer	genre
Taylor Swift	Taylor Swift	Pop
Drake	Drake	Hip-Hop
Adele	Adele	Pop
Radiohead	Thom Yorke	Rock
The Smile	Thom Yorke	Rock

Table 3: R object name: `table3`

artist	song	streams_2015	streams_2024
Taylor Swift	Blank Space	500	600
Drake	Hotline Bling	700	900
Adele	Easy On Me	NA	1000

Table 4: R object name: `table4`

lead_performer	birth_country
Taylor Swift	United States
Drake	Canada
Adele	England
Thom Yorke	England

Table 5: R object name: `table5`

id	mUSICAL_aRTIST	BILLBOARDno1YEARSsince2018	gEnRe
13	TSwift	2020, 2022, 2023, 2024	pop
Drizzy	Drake (Aubrey Drake Graham)	2018, 2020	hip-Hop
1988	Adele (born 1988)	2021	Pop
15	Radiohead and also The Smile	No number one hits	Rock'n'roll

Table 6: R object name: `table6`

song	minutes	seconds
Blank Space	3	51
Hotline Bling	4	27
Easy On Me	3	44
The National Anthem	5	51
Thin Thing	4	30

4 Questions

4.1 Write input

Write code to reproduce the table below using the tables defined in the Data section:

```
# A tibble: 5 x 5
  artist       lead_performer genre   song                star_rating
  <chr>        <chr>          <chr>   <chr>                     <dbl>
1 Taylor Swift Taylor Swift   Pop     Blank Space                   2
2 Drake        Drake          Hip-Hop Hotline Bling                 3
3 Adele        Adele          Pop     Easy On Me                    4
4 Radiohead    Thom Yorke     Rock    The National Anthem           5
5 The Smile    Thom Yorke     Rock    Thin Thing                    5
```

4.2 Draw output

Draw the output of the following code chunk:

```
table3 |>
  pivot_longer(
    cols = starts_with("streams_"),
    names_to = "year",
    names_prefix = "streams_",
    names_transform = as.integer,
    values_to = "streams"
  )
```

4.3 Songs by English people

Write code to reproduce the table below using the tables defined in the Data section:

```
# A tibble: 3 x 5
  artist    song                star_rating lead_performer birth_country
  <chr>     <chr>                     <dbl> <chr>          <chr>
1 Adele     Easy On Me                    4 Adele          England
2 Radiohead The National Anthem           5 Thom Yorke     England
3 The Smile Thin Thing                    5 Thom Yorke     England
```

4.4 Tidy up Table 5

Name tidy data principles that Table 5 violates and how to fix them.

4.5 Similar joins

Draw the output of the following code chunks.

```
table1 |>
  inner_join(table3, by = c("artist", "song"))
```

```
table1 |>
  semi_join(table3, by = c("artist", "song"))
```

4.6 Song length

Fill in the blanks in the code below such that it produces the following table.

Rewrite the code in the blank part of the page if need be.

Nota bene: The number of blanks does *not* necessarily correspond to the number of characters in the blanked out field.

- `length` gives the length of the song in seconds.
- `long_song` indicates whether the song is more than four minutes long.

```
________ |>
  ______join(________, by = __________) |>
  ______(
    length = _________________________________,
    long_song = if_else(_______________________________)
  ) |>
  ______(______, _____, ___________, long_song)
```

```
# A tibble: 5 x 4
  artist       song                length long_song
  <chr>        <chr>                <dbl> <lgl>
1 Taylor Swift Blank Space            231 FALSE
2 Drake        Hotline Bling          267 TRUE
3 Adele        Easy On Me             224 FALSE
4 Radiohead    The National Anthem    351 TRUE
5 The Smile    Thin Thing             270 TRUE
```

4.7 Principles of data feminism

Fill in the blanks in the following statements.

Nota bene: If you write statements that are similar to those that have been blanked out, you can still receive substantial credit.

Principles of data feminism

"The starting point for data feminism is something that goes mostly unacknowledged in data science: _____ is not distributed equally in the world."

Principles of data feminism

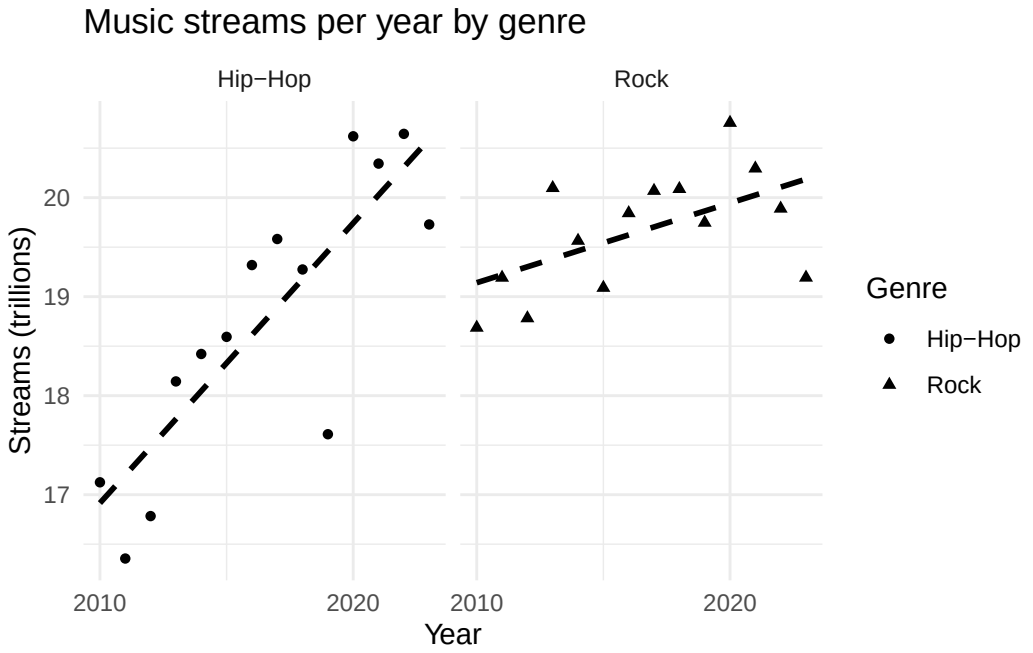
1. Use data to create_____
2. Recognize that data is_____
3. Make _____ visible

4.8 Normalize Table 2

Draw tables demonstrating how to normalize Table 2 to the highest available normal form.

4.9 Interpret trends

Interpret the trends in the faceted plot below. How do the trends differ by genre? What do these trends suggest about the past and future of these genres? Why do you think these values differ in the ways that they do?



4.10 Interpret `tidy()`

The following tables contain the output of `tidy()` for the linear models shown above.

Hip-hop:

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept) -551.     107.         -5.14 0.000245
2 year           0.283    0.0532      5.32 0.000183
```

Rock:

```
# A tibble: 2 x 5
  term         estimate std.error statistic p.value
  <chr>           <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept) -143.       69.7        -2.05  0.0632
2 year           0.0804    0.0345      2.33  0.0381
```

1. Explain what the `estimate` for the `term` `year` means in each table. How and why do they differ?
2. Both estimates for the `term` `year` are positive. Which real-world phenomena about music streaming might explain this? Identify at least two possibilities.