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.

```r
library(tidyverse)
library(knitr)
library(kableExtra)
library(broom)

table1 <- tibble(
  artist = c(
    "Taylor Swift", "Drake", "Adele",
    "Radiohead", "The Smile"
  ),
  song = c(
    "Blank Space", "Hotline Bling", "Easy On Me",
    "The National Anthem", "Thin Thing"
  ),
  star_rating = c(2, 3, 4, 5, 5)
)

kable(table1)
```

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

```r
table2 <- tibble(
  artist = c(
    "Taylor Swift", "Drake", "Adele",
    "Radiohead", "The Smile"
  ),
  lead_performer = c(
    "Taylor Swift", "Drake", "Adele",
    "Thom Yorke", "Thom Yorke"
  ),
  genre = c("Pop", "Hip-Hop", "Pop", "Rock", "Rock")
)

kable(table2)
```

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

```
table3 <- tibble(
  artist = c("Taylor Swift", "Drake", "Adele"),
  song = c("Blank Space", "Hotline Bling", "Easy On Me"),
  streams_2015 = c(500, 700, NA),
  streams_2024 = c(600, 900, 1000)
)

kable(table3)
```

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

```
table4 <- tibble(
  lead_performer = c(
    "Taylor Swift", "Drake", "Adele", "Thom Yorke"
  ),
  birth_country = c(
    "United States", "Canada", "England", "England"
  )
)

kable(table4)
```

Table 4: R object name: `table4`

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

```r
table5 <- tibble(
  id = c(13, "Drizzy", 1988, 15),
  mUSICAL_aRTIST = c(
    "TSwift", "Drake (Aubrey Drake Graham)",
    "Adele (born 1988)",
    "Radiohead and also The Smile"
  ),
  BILLBOARDno1YEARSsince2018 = list(
    c(2020, 2022, 2023, 2024),
    c(2018, 2020),
    c(2021),
    "No number one hits"
  ),
  gEnRe = c("pop", "hip-Hop", "Pop", "Rock'n'roll")
)

kable(table5) |>
  kable_styling(bootstrap_options = "striped") |>
  row_spec(2, background = "#fde725") |>
  row_spec(3, background = "#a0da39") |>
  row_spec(4, background = "#4ac16d")
```

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

```r
table6 <- tibble(
  song = c(
    "Blank Space", "Hotline Bling", "Easy On Me",
    "The National Anthem", "Thin Thing"
  ),
  minutes = c(3, 4, 3, 5, 4),
  seconds = c(51, 27, 44, 51, 30)
)

kable(table6)
```

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:

```
table2 |>
  left_join(table1, by = "artist")
```

```
# 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"
  )
```

```
# A tibble: 6 x 4
  artist       song           year streams
  <chr>        <chr>         <int>   <dbl>
1 Taylor Swift Blank Space    2015     500
2 Taylor Swift Blank Space    2024     600
3 Drake        Hotline Bling  2015     700
4 Drake        Hotline Bling  2024     900
5 Adele        Easy On Me     2015      NA
6 Adele        Easy On Me     2024    1000
```

4.3 Songs by English people

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

```
table1 |>
  left_join(table2 |> select(artist, lead_performer), by = "artist") |>
  left_join(table4, by = "lead_performer") |>
  filter(birth_country == "England")
```

```
# 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.

1. mixed values in ids
2. non-numeric ids
3. nonstandard artist name representation
4. non-snake_case column names
5. multiple values per cell in artist and billboard years
6. mismatched genre labels (e.g., “pop” and “Pop”)
7. row highlights use color for meaning
8. multiple datatypes per column

4.5 Similar joins

Draw the output of the following code chunks.

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

```
# A tibble: 3 x 5
  artist       song          star_rating streams_2015 streams_2024
  <chr>        <chr>               <dbl>        <dbl>        <dbl>
1 Taylor Swift Blank Space             2          500          600
2 Drake        Hotline Bling           3          700          900
3 Adele        Easy On Me              4           NA         1000
```

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

```
# A tibble: 3 x 3
  artist       song          star_rating
  <chr>        <chr>               <dbl>
1 Taylor Swift Blank Space             2
2 Drake        Hotline Bling           3
3 Adele        Easy On Me              4
```

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)
```

```
table1 |>
  left_join(table6, by = "song") |>
  mutate(
    length = minutes * 60 + seconds,
    long_song = if_else(length > 240, TRUE, FALSE)
  ) |>
  select(artist, song, length, 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

0. *power* is not distributed equally in the world
1. Use data to create *more just, equitable, and livable futures*
2. Recognize that data *is never neutral or objective*
3. Make *labor* visible

4.8 Normalize Table 2

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

```
artists <- tibble(
  artist_id = 1:5,
  artist = c(
    "Taylor Swift", "Drake", "Adele",
    "Radiohead", "The Smile"
  )
)

artists
```

```
# A tibble: 5 x 2
  artist_id artist
      <int> <chr>
1         1 Taylor Swift
2         2 Drake
3         3 Adele
4         4 Radiohead
5         5 The Smile
```

```
performers <- tibble(
  performer_id = 1:4,
  lead_performer = c("Taylor Swift", "Drake", "Adele", "Thom Yorke")
)

performers
```

```
# A tibble: 4 x 2
  performer_id lead_performer
         <int> <chr>
1            1 Taylor Swift
2            2 Drake
3            3 Adele
4            4 Thom Yorke
```

```
genres <- tibble(
  genre_id = 1:3,
  genre = c("Pop", "Hip-Hop", "Rock")
```

```
)

genres
```

```
# A tibble: 3 x 2
  genre_id genre
     <int> <chr>
1        1 Pop
2        2 Hip-Hop
3        3 Rock
```

```
artist_lead <- tibble(
  artist_id = 1:5,
  performer_id = c(1, 2, 3, 4, 4)
)

artist_lead
```

```
# A tibble: 5 x 2
  artist_id performer_id
      <int>        <dbl>
1         1            1
2         2            2
3         3            3
4         4            4
5         5            4
```

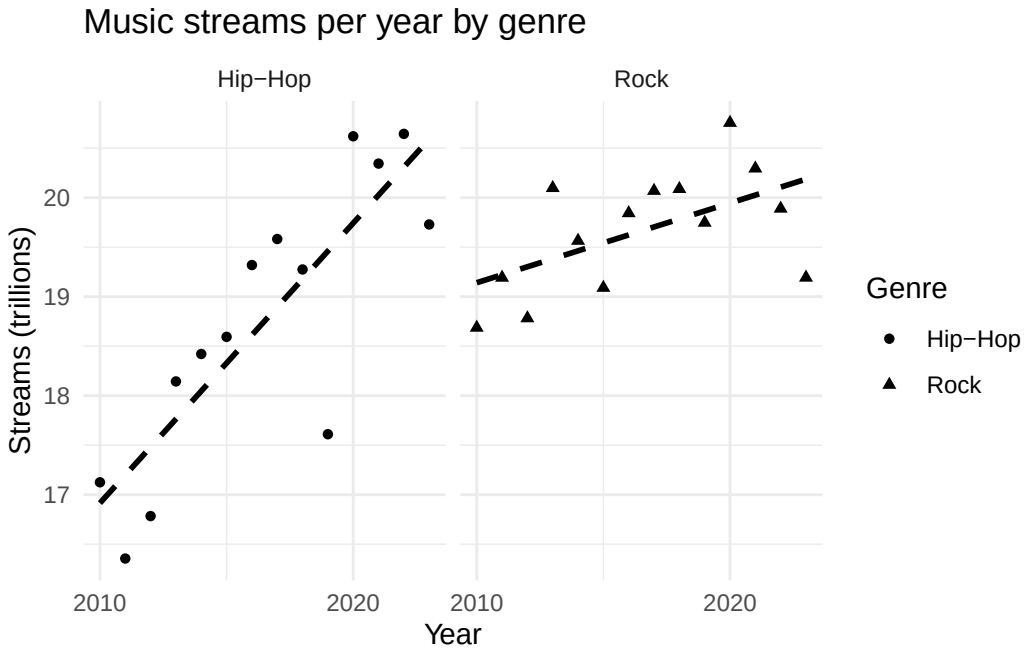
```
artist_genre <- tibble(
  artist_id = 1:5,
  genre_id = c(1, 2, 1, 3, 3)
)

artist_genre
```

```
# A tibble: 5 x 2
  artist_id genre_id
      <int>    <dbl>
1         1        1
2         2        2
3         3        1
```

4	4	3
5	5	3

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.

4.11 Question values

Question Title	Points
1. Write input	3
2. Draw output	4
3. Songs by English people	4
4. Tidy up `table5`	6
5. Similar joins	6
6. Song length	5
7. Principles of data feminism	4
8. Normalize `table2`	6
9. Interpret trends	6
10. Interpret `tidy()`	6