# Homework 4

Mariia Nikitash

Prepare your answers as a **single PDF file**.
**Group work**: You may work in groups of 1-3. Include all group member names in the PDF file. You may work with students in both sections (375-01, -02). Only one person in the group should submit to Canvas.
**Due**: check on Canvas.

P.S. All text in pink I made for myself for the future references

**1.** Load the `nycflights13` library (will have to install the `nycflights13` package first) which contains flight arrival and departure data in a table called `flights`. Apply the tidyverse's data wrangling verbs to answer these questions. For each question, **give only the (one line as a data pipeline) code beginning with `flights %>%` ….**

1. Add a new variable that indicates the total delay (both departure and arrival delay).
   flights %>% mutate(total_delay = (dep_delay+arr_delay)) %>% View()

2. Show only the origin, destination, and total delay of 10 most delayed flights
   flights %>% mutate(total_delay = dep_delay + arr_delay) %>% arrange(-total_delay) %>% slice(1:10) %>% select(origin, dest, total_delay) %>% View()

**2.** Consider the two tables shown below called *bands* and *instruments*.

**bands**:

| *name* | *lastname* | *band* | *year* |
|---|---|---|---|
| Mick | Jagger | Stones | 1962 |
| John | Lennon | Beatles | 1960 |
| Paul | McCartney | Beatles | 1960 |
| Paul | McCartney | Wings | 1971 |

**instruments**:

| *artist* | *artistname* | *plays* | *model* |
|---|---|---|---|
| John | Lennon | guitar | Gibson |
| Paul | McCartney | bass | Hofner |
| Keith | Richards | guitar | Fender |
| Paul | McCartney | bass | Hofner |

Draw the output table from the following operations (you should be able to calculate the output by hand though you may use R to check your answers).

```
bands <- data.frame(name= c("Mick", "John", "Paul", "Paul"), lastname= c("Jagger", "Lennon", "McCartney","McCartney"), band=c("Stones", "Beatles","Beatles","Wings"), year= c(1962, 1960, 1960, 1971))
```

```
instruments <- data.frame(artist= c("John", "Paul", "Keith", "Paul"), artistname= c("Lennon", "McCartney", "Richards","McCartney"), plays=c("guitar", "bass","guitar","bass"), model= c("Gibson", "Hofner", "Fender", "Hofner"))
```

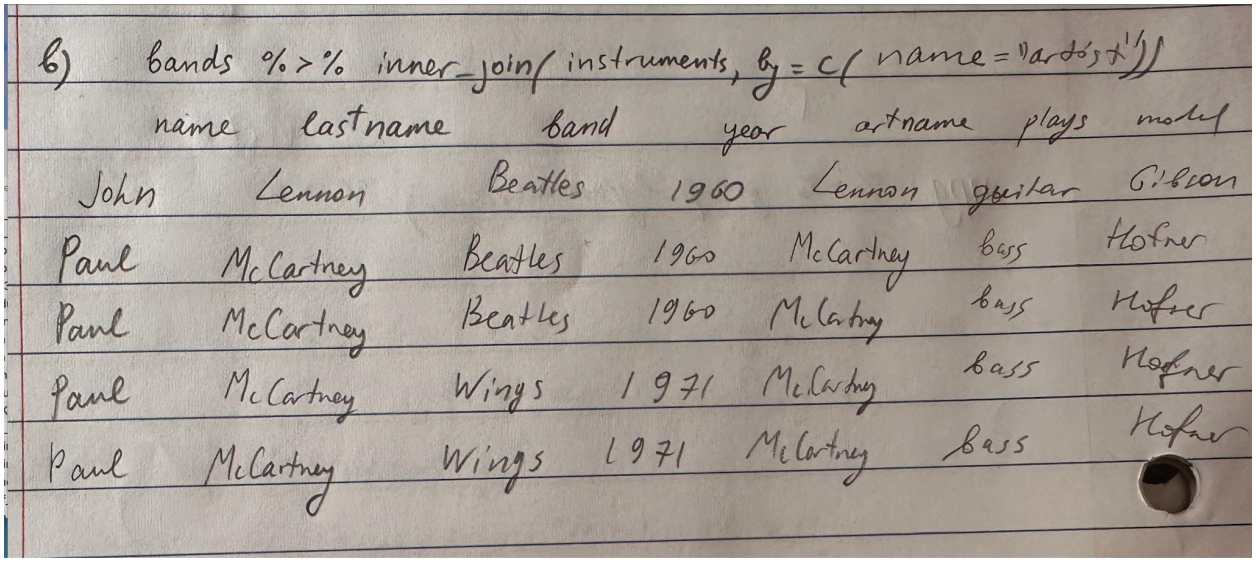
a) `bands %>% inner_join(instruments)` [Hint: this is a trick question!]

There is no common column between “bands” and “instruments”, so it is impossible to join these two tables, it will give an error.

Answer check:

```
Error in `inner_join()`:
! `by` must be supplied when `x` and `y` have no common variables.
i Use `cross_join()` to perform a cross-join.
```

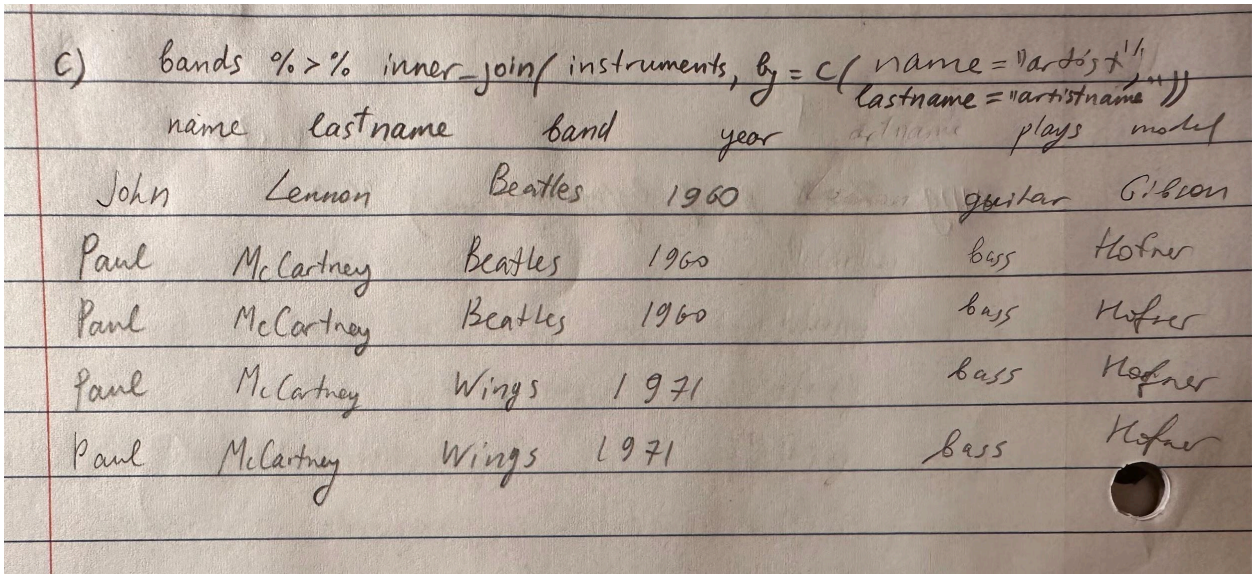
b) `bands %>% inner_join(instruments, by=c(name="artist"))`




Answer check:

```
bands %>% inner_join(instruments, by=c(name="artist")
name  lastname    band year artistname  plays  model
1 John    Lennon Beatles 1960     Lennon guitar Gibson
2 Paul McCartney Beatles 1960  McCartney   bass Hofner
3 Paul McCartney Beatles 1960  McCartney   bass Hofner
4 Paul McCartney   Wings 1971  McCartney   bass Hofner
5 Paul McCartney   Wings 1971  McCartney   bass Hofner
```

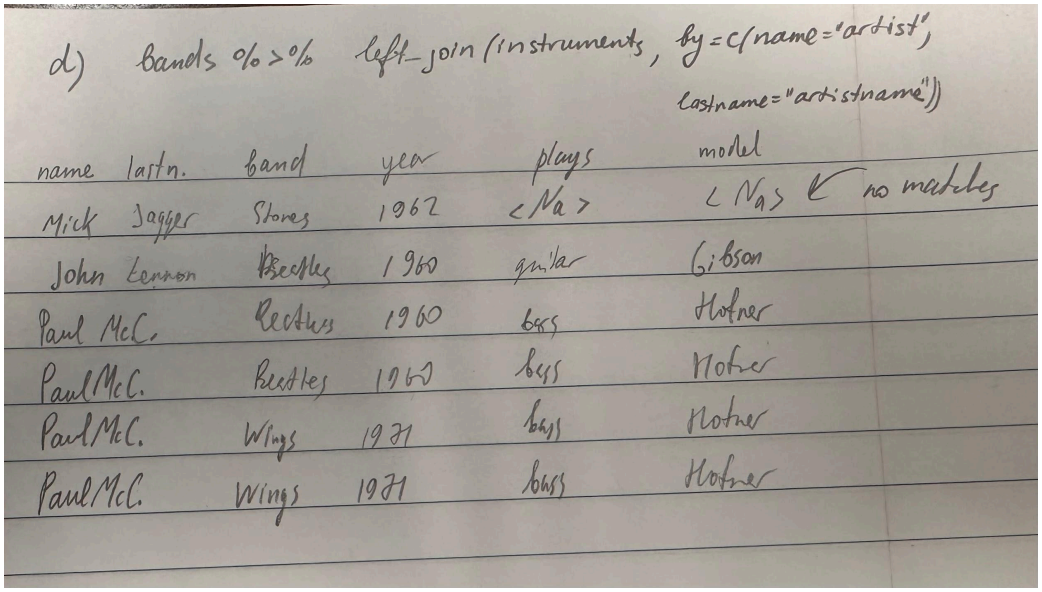
```
c) bands %>% inner_join(instruments, by=c(name="artist",
   lastname="artistname"))
```




Answer check:

```
name  lastname    band year  plays  model
1 John    Lennon Beatles 1960 guitar Gibson
2 Paul McCartney Beatles 1960   bass Hofner
3 Paul McCartney Beatles 1960   bass Hofner
4 Paul McCartney   Wings 1971   bass Hofner
5 Paul McCartney   Wings 1971   bass Hofner
```

```
d) bands %>% left_join(instruments, by=c(name="artist",
lastname="artistname"))
```




Answer check:

```
name  lastname    band year  plays  model
1 Mick    Jagger  Stones 1962   <NA>   <NA>
2 John    Lennon Beatles 1960 guitar Gibson
3 Paul McCartney Beatles 1960   bass Hofner
4 Paul McCartney Beatles 1960   bass Hofner
5 Paul McCartney   Wings 1971   bass Hofner
6 Paul McCartney   Wings 1971   bass Hofner
```

e) bands %>% inner_join(instruments, by=c(name="artist", lastname="artistname", year="plays"))

I cant combine string and character

Answer check:

```
Error in `inner_join()`:
! Can't join `x$year` with `y$plays` due to incompatible types.
i `x$year` is a <double>.
i `y$plays` is a <character>.
```

If i change character in bands to a string I still wont be able to do the join, since non of the values are matching. It will give an empty table

Answer check:

```
[1] name     lastname band     year     model
<0 rows> (or 0-length row.names)
```

**3. a.** Create a table called enrollment (of class “tibble”) with the enrollment information of CPSC 375 over the last few years given below: [Give code only]

**enrollment**:

| *Year* | *Semester* | *Enrolled* |
|---|---|---|
| 2024 | Spring | 74 |
| 2023 | Fall | 26 |
| 2023 | Summer | 34 |
| 2023 | Spring | 85 |
| 2022 | Summer | 21 |
| 2022 | Spring | 79 |
| 2021 | Fall | 42 |
| 2021 | Summer | 34 |
| 2021 | Spring | 79 |
| 2020 | Fall | 44 |
| 2020 | Spring | 57 |
| 2019 | Fall | 36 |
| 2019 | Spring | 70 |
| 2018 | Fall | 45 |

Hint: `tibble(Year=c(2024,2023, …), Semester=c("Spring", "Fall", …), …)`

```
tibble(Year=c(2024,2023, 2023, 2023, 2022, 2022, 2021, 2021, 2021,
2020, 2020, 2019, 2019, 2018), Semester=c("Spring", "Fall", "Summer",
"Spring", "Summer", "Spring", "Fall", "Summer" , "Spring", "Fall",
"Spring", "Fall", "Spring", "Fall"), Enrolled= c(74, 26, 34, 85, 21,
79, 42, 34, 79, 44, 57, 36, 70, 45))
```

**b.** Load the data collected from the anonymous survey given at the beginning of the course and from previous semesters into a variable called “survey”. The dataset can be downloaded from the Datasets module on Canvas. Use the function **`read_csv()`** [Give code only]

```
survey <- read_csv("surveydataSpring2024.csv")
```

**c.** Write code (as a pipeline of tidyverse functions) to merge the two tables above and create a single table with the total number of responses in every semester as below (the order of rows/columns is not important): [Give code only]

| ***Year*** | ***Semester*** | ***Responses*** | ***Enrolled*** |
|---|---|---|---|
| 2024 | Spring | 63 | 74 |
| 2023 | Fall | 27 | 26 |
| … | … | … | … |

1. I made a new variable for the table in 3a

```
table <- tibble(Year=c(2024,2023, 2023, 2023, 2022, 2022, 2021, 2021, 2021, 2020, 2020, 2019, 2019, 2018), Semester=c("Spring", "Fall", "Summer", "Spring", "Summer", "Spring", "Fall", "Summer" , "Spring", "Fall", "Spring", "Fall", "Spring", "Fall"), Enrolled= c(74, 26, 34, 85, 21, 79, 42, 34, 79, 44, 57, 36, 70, 45))
```

2. I made a new variable for survey after I did some functions with the table: grouped by year and semester and sumarrized the responses

```
surv <- survey %>% group_by(Year, Semester) %>% summarise(Responses=(n()))
```

3. I combined two tables

```
full_join(table, surv)
```

4. Assigned new variable to full_join and removed the NA

```
ful <- full_join(table, surv)
ful %>% filter(!is.na(Responses))
```

Answer check

```
Year Semester Enrolled Responses
  <dbl> <chr>       <dbl>     <int>
 1  2024 Spring        74        63
 2  2023 Fall          26        27
 3  2023 Summer        34        31
 4  2023 Spring        85        86
 5  2022 Summer        21        22
 6  2022 Spring        79        78
 7  2021 Fall          42        40
 8  2021 Summer        34        34
 9  2021 Spring        79        40
10  2020 Fall          44        48
11  2020 Spring        57        60
12  2019 Fall          36        26
13  2019 Spring        70        69
```