Extra Lab: Merging Data

Part 1

Read in the data and use functions of your choice to preview it.

```
library(tidyverse)

crash <- read_csv("https://sisbid.github.io/Data-Wrangling/labs/crashes.csv")
road <- read_csv("https://sisbid.github.io/Data-Wrangling/labs/roads.csv")
```

1. Join data to retain only complete data, (using an inner join) e.g. those observations with road lengths and districts. Merge without using `by` argument, then merge using `by = "Road"`. call the output `merged`. How many observations are there?

```
# Step 1: Inner Join (without specifying by)
merged1 <- inner_join(crash, road)  # matches on all common columns automatically
```

```
## Joining with ‘by = join_by(Road)‘
```

```
nrow(merged1)
```

```
## [1] 88
```

```
# Step 2: Inner Join (specifying key column explicitly)
merged <- inner_join(crash, road, by = "Road")  # safer and more explicit
nrow(merged)
```

```
## [1] 88
```

```
# 88 observations are there
```

2. Join data using a `full_join`. Call the output `full`. How many observations are there?

```
full <- full_join(crash, road, by = "Road")
nrow(full)
```

```
## [1] 111
```

```
# 111 observations are there
```

3. Do a left join of the `road` and `crash`. ORDER matters here! How many observations are there?

```r
# keep all roads, and attach crash info if available
left_join(road, crash, by = "Road")
```

```
## # A tibble: 89 x 6
##    Road          District   Length  Year N_Crashes Volume
##    <chr>         <chr>       <dbl> <dbl>     <dbl>  <dbl>
##  1 Interstate 65 Greenfield    262  1991        25  40000
##  2 Interstate 65 Greenfield    262  1992        37  41000
##  3 Interstate 65 Greenfield    262  1993        45  45000
##  4 Interstate 65 Greenfield    262  1994        46  45600
##  5 Interstate 65 Greenfield    262  1995        46  49000
##  6 Interstate 65 Greenfield    262  1996        59  51000
##  7 Interstate 65 Greenfield    262  1997        76  52000
##  8 Interstate 65 Greenfield    262  1998        90  58000
##  9 Interstate 65 Greenfield    262  1999        95  65000
## 10 Interstate 65 Greenfield    262  2000        95  74000
## # i 79 more rows
```

```r
left <- left_join(road, crash, by = "Road")
nrow(left)
```

```
## [1] 89
```

```r
# 89 observations are there
```

4. Repeat above with a `right_join` with the same order of the arguments. How many observations are there?

```r
right <- right_join(road, crash, by = "Road")
nrow(right)
```

```
## [1] 110
```

```r
# 110 observations are there
```

Bonus Practice

5. Which highways do not have road data? Do this in a "tidy" format. Summarize by the total count of `N_Crashes` per highway. Hint: Use `anti_join()` and `group_by()`.

```r
crash %>%
  anti_join(road, by = "Road") %>%      # only crashes with no matching road info
  group_by(Road) %>%                    # group by highway name
  summarize(total_crashes = sum(N_Crashes, na.rm = TRUE)) %>%
  arrange(desc(total_crashes))
```

```
## # A tibble: 1 x 2
##   Road           total_crashes
##   <chr>                  <dbl>
## 1 Interstate 275           549
```

```
# Rod is interstate 275 with 549 total crashes
```

6. You have an intern who has been pouring over the raw data and found a few mistakes in the `N_Crashes` column of the `crash` dataset. They have made a spreadsheet for you containing only the corrected entries. Modify the original tibble with the following:

- A column (`Corrected`) indicating if a particular entry has a corresponding correction in `corrections`.
- If the row has a correction, take the corrected value
- Keep the original columns (Year, Road, N_Crashes, Volume) plus the column indicating whether the data is corrected or not (`Corrected`).

hint: take a look at the two datasets - are you sure they're joining correctly?

```
corrections <- read_csv("https://sisbid.github.io/Data-Wrangling/labs/crashes_corrections.csv")
```

```
## Rows: 9 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ROAD
## dbl (3): YEAR, N_Crashes, Volume
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
# They are not joining correctly. roblem with `Year` and `Road`
```