Data Wrangling – Part 4

Data Summary: Using `group_by()` and `summarise()` together allows you to efficiently compute summary statistics, aggregations, or any other computations of data based on different groups defined by one or more variables.

`group_by()`: is used to group a data frame by one or more variables. This creates a "grouped" data frame where subsequent operations are performed within each group separately. This works best with categorical variables or factor variables. Using `group_by()` in it's own doesn't change the "look" of the data.

`summarise()`: is used to compute summary statistics or other values for each group. It condenses the grouped data into a single row per group, summarizing the specified variables.

Command Illustrations

```
new_dataframe_name <- dataframe_name %>%
     summarise(new_column_name = function_name(column_name2))
```

If using group_by:

```
new_dataframe_name <- dataframe_name %>%
     group_by(column_name1) %>%
     summarise(new_column_name = function_name(column_name2))
```

For the illustration examples, assume the dataframe is the following:

Illustration_Data

Name	Age	total_Income	var_1	var_2	var_3	zipcode	honesty	cat_total
Val	18	18000	apple	carrots	elephant	60001	agree	0
Derek	25	25000	grapes	carrots	tiger	60073	disagree	1
Whitney	30	30000	bananas	carrots	lion	60109	disagree	2
Sasha	40	40000	peaches	carrots	rabbit	60111	disagree	1
Daniella	45	45000	bananas	carrots	shark	60155	agree	1

Example 1: Compute mean and median of total_Income

```
example_1 <- Illustration_Data %>%
      summarise(mean_income = mean(total_Income),
                median_income = median(total_Income))
```

Output will be a dataframe that looks like:

Example 2: Compute mean and median of total_Income, grouped by favorite fruit (var_1)

```
example_2 <- Illustration_Data %>%
     group_by(var_1) %>%
     summarise(mean_income = mean(total_Income),
               median_income = median(total_Income))
```

Output will be a dataframe that looks like:

Example 3: Compute sum and mean of total_Income, grouped by honesty (var_1) and cat_total

```
example_3 <- Illustration_Data %>%
     group_by(honesty, cat_total) %>%
     summarise(sum_income = sum(total_Income),
               mean_income = mean(total_Income))
```

Output will be a dataframe that looks like:

Example 4: Compute the number of pro dancers per honesty

```
example_4 <- Illustration_Data %>%
      group_by(honesty) %>%
      summarise(Num_Pro = n())
```

Output will be a dataframe that looks like:
