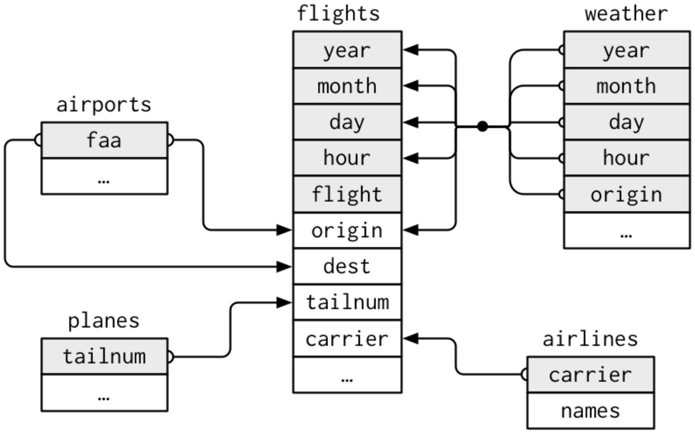
Week 2 Tasks

Tasks

An airline industry measure of a passenger airline's capacity is the available seat miles, which is equal to the number of seats available multiplied by the number of miles or kilometers flown. So for example say an airline had 2 flights using a plane with 10 seats that flew 500 miles and 3 flights using a plane with 20 seats that flew 1000 miles, the available seat miles would be 2 × 10 × 500 + 3 × 20 × 1000 = 70,000 seat miles.

Using the data sets included in the `nycflights13` package, compute the available seat miles for each airline sorted in descending order. After completing all the necessary data wrangling steps, the resulting data frame should have 16 rows (one for each airline) and 2 columns (airline name and available seat miles). Here are some hints:

1. Take a close look at all the data sets using the `View`, `head` or `glimpse` functions: `flights`, `weather`, `planes`, `airports`, and `airlines` to identify which variables are necessary to compute available seat miles.
2. This diagram (from the **Joining section**) will also be useful.



3. Consider the data wrangling verbs in the table above as your toolbox!

If you want to work through it **step by step**, here are some hints:

Step 1: To compute the available seat miles for a given flight, we need the `distance` variable from the `flights` data frame and the `seats` variable from the `planes` data frame, necessitating a join by the key variable `tailnum`. To keep the resulting data frame easy to view, we'll `select` only these two variables and `carrier`.

Step 2: Now for each flight we can compute the available seat miles `ASM` by multiplying the number of seats by the distance via a `mutate`.

Step 3: Next we want to sum the `ASM` for each carrier. We achieve this by first grouping by `carrier` and then summarising using the `sum` function.

Step 4: However, if it was the case that some carriers had certain flights with missing `NA` values, the resulting table above would also return `NA`'s (NB: this is not the case for this data). We can eliminate these by adding the `na.rm = TRUE` argument to `sum`, telling R that we want to remove the `NA`'s in the sum.

Step 5: Finally, `arrange` the data in `desc`ending order of `ASM`.

Further Tasks

Further Task 1

In this task we will work with the data set analysed and reported in the 2016 article from FiveThirtyEight.com entitled Some People Are Too Superstitious To Have A Baby On Friday The 13th. The data set is called `US_births_2000_2014` and is within the `fivethirtyeight` package.

1. Create an object called `US_births_2013` which focuses only on data corresponding to 2013 births.
2. By only choosing birth data for the years 2010, 2011, 2012, and 2014 create a new data frame called `US_births_small` and check that this resulting data frame has 1461 rows. Note that there are many different ways to do this, but try and come up with three different ways using:

- the "or" operator `|`
- the `%in%` operator
- the "not" operator `!`

or combinations of them.

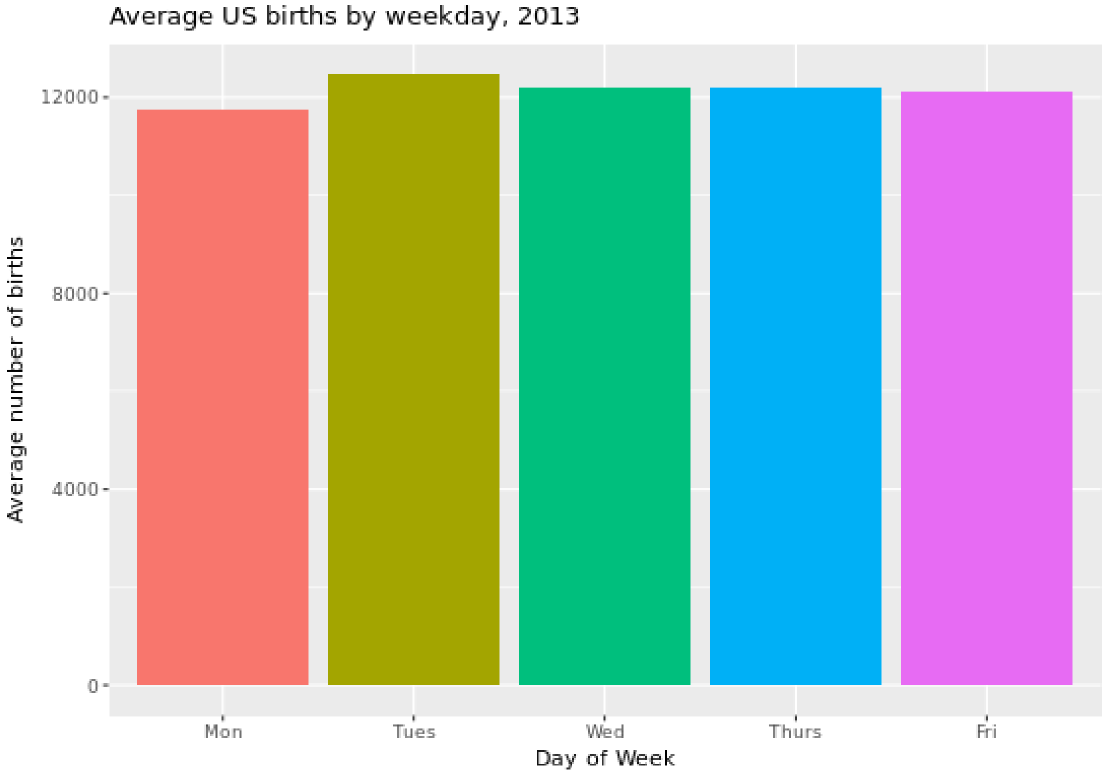
3. Suppose we are interested in choosing rows for only weekdays (not Saturdays or Sundays) for `day_of_week` in year 2013. Write the code to do so and give the name `US_births_weekdays_2013` to the resulting data frame. Note that you may want to run `US_births_2000_2014 |> distinct(day_of_week)` to identify the specific values of `day_of_week`.

4. Using what you covered in Week 1, produce an appropriate plot looking at the pattern of births on all weekdays in 2013 coloured by the particular day of the week.

5. The plot in the previous task has shown there are some outliers in the data for US births on weekdays in 2013. We can use the `summarize` function to get an idea for how these outliers may affect the shape of the births variable in `US_births_weekdays_2013`. Write some code to calculate the mean and median values for all weekday birth totals in 2013. Store this aggregated data in the data frame `birth_summ`. What do these values suggest about the effects of the outliers?

6. Instead of looking at the overall mean and median across all of 2013 weekdays, calculate the mean and median for each of the five different weekdays throughout 2013. Using the same names for the columns as in the `birth_summ` data frame in the previous exercise, create a new data frame called `birth_day_summ`.

7. Using the aggregated data in the `birth_day_summ data` frame, produce this barplot.



Further Task 2

In this task we will work with the data set analysed and reported in the 2014 article from FiveThirtyEight.com entitled 41 Percent Of Fliers Think You're Rude If You Recline Your Seat. The data set is called `flying` and is within the `fivethirtyeight` package.

1. Write code to determine the proportion of respondents in the survey that responded with **Very** when asked if a passenger reclining their seat was rude. You should determine this proportion across the different levels of `age` and `gender` resulting in a data frame of size 8 x 3. Assign the name `prop_very` to this calculated proportion in this aggregated data frame.

> Hint 1
>
> We can obtain proportions using the `mean` function applied to logical values. For example suppose we want to count the proportion of "heads" in five tosses of a fair coin. If the results of the five tosses are stored in
> ```
> tosses <- c("heads", "tails", "tails", "heads", "heads")
> ```
> then we can use `mean(tosses == "heads")` to get the resulting answer of 0.6.

> Hint 2
>
> Including the function `na.omit(TRUE)` in the 'pipe' (`|>`) removes all entries that are not complete whereas including the argument `na.rm=TRUE` in the `mean` function removes just those entries where the relevant variable value is missing.

2. Using the aggregated data you've created, produce two bar plots (one stacked, the other side-by-side) to show the differences between the sexes of the proportion of people who believe reclining your seat is 'very' rude, within each age group. Also, consider

   - What stands out to you as you review these proportions?
   - What gender and age-range pairings have the highest and lowest proportions thinking reclining airline seats is very rude in this survey?