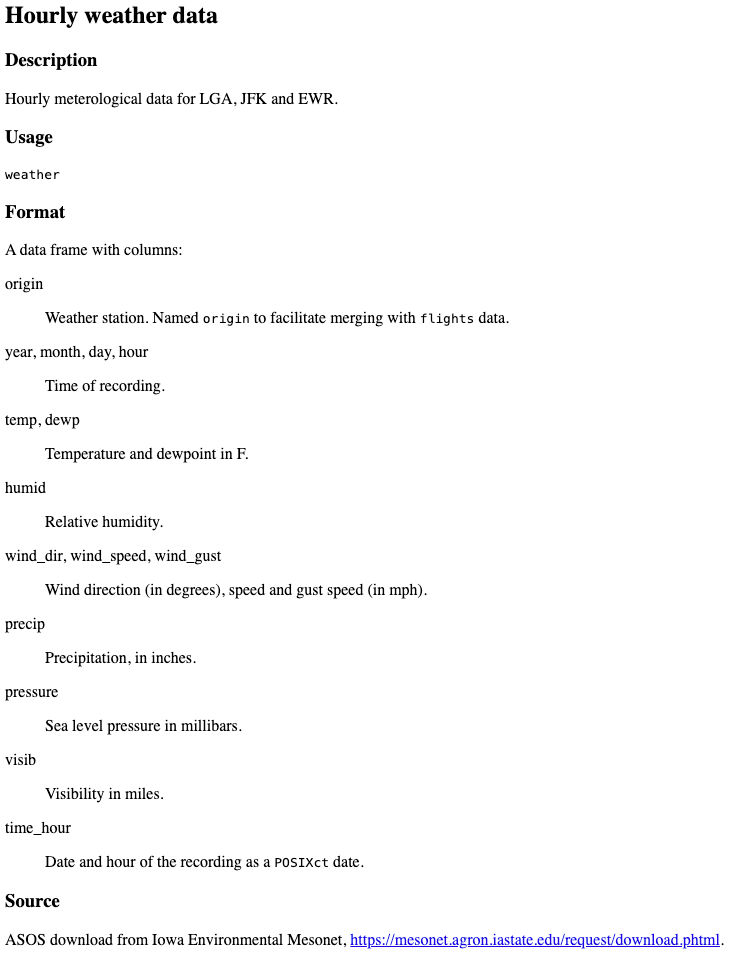
# **R for Data Science – Chapters 1 to 3 Summative Lab**

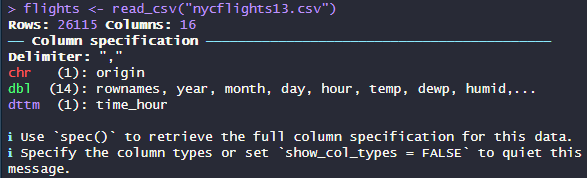
For this lab, we’ll be using a dataset containing information on weather at the airports in New York City. The data was found in an excellent dataset aggregator at <https://vincentarelbundock.github.io/Rdatasets/articles/data.html>. This dataset is set 281 on that page and is similar to the one used in the R for Data Science chapter 3 tutorial. That dataset focused on the flight attributes whereas this focuses on the weather at the airports.

Documentation explaining the columns is below:

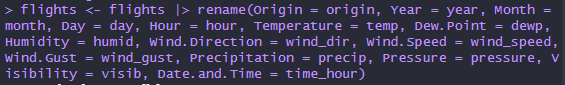


For each exercise below, paste the code and any outputs that it creates below each question.

1. First, download the data from the linked page above. Next, read it into R. We haven’t covered file imports yet, but it’s not difficult. Use chapter 7 of the R for Data Science book as a reference.



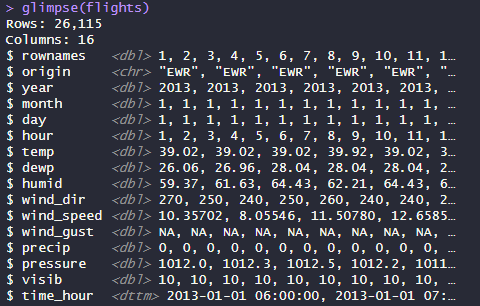
1. Rename the columns using more descriptive but still succinct terms. For example, “temp” to “Temperature” and “wind\_speed” to “Wind.Speed”. This will ensure that the subsequent graphs look better.

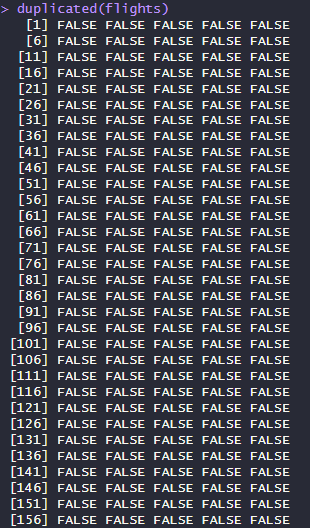


1. Create a new column called “Temp.Celsius” with the appropriate formula to convert “Temperature” to its Celsius equivalent. Save this new column to the existing dataframe.

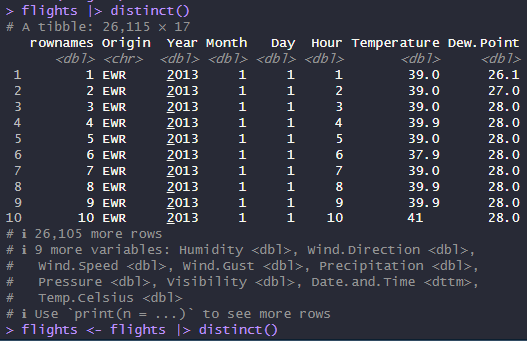


1. Determine whether there are any duplicate rows. If there are, remove them and save the change to the existing dataframe.



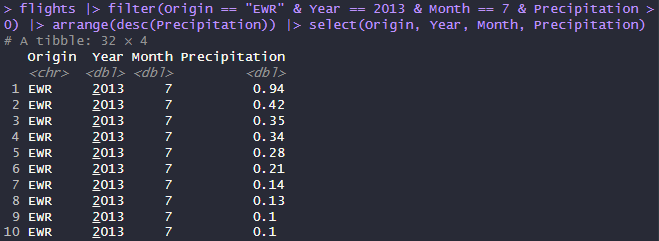
 before removing duplicates

(26,115 rows)

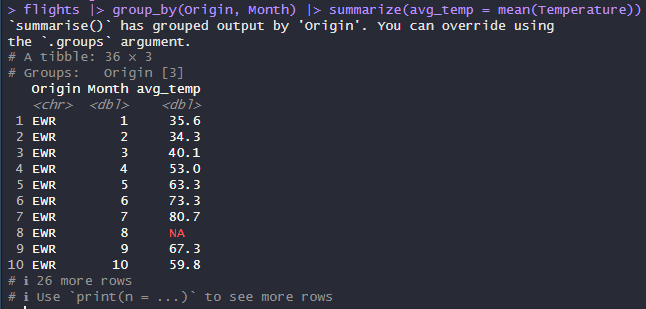


after removing duplicates (still 26,225 rows; there are no duplicates in the dataset)

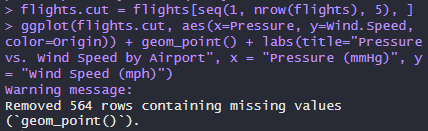
1. Write a filter that only returns the rows for Newark Airport in July 2013 in which there was precipitation that hour. Sort the rows in descending precipitation. Do not save the modifications in this question to the dataframe. Use it only for this question.

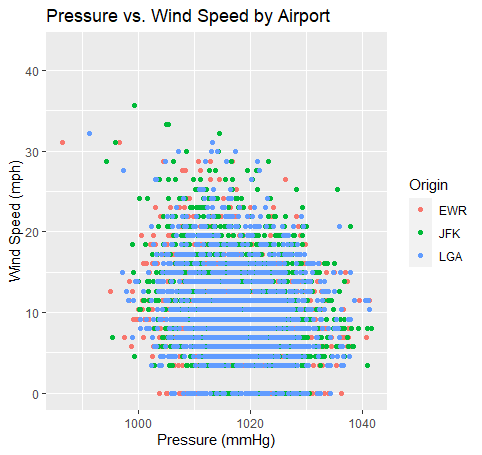


1. Find the average Fahrenheit temperature for each airport in each month.

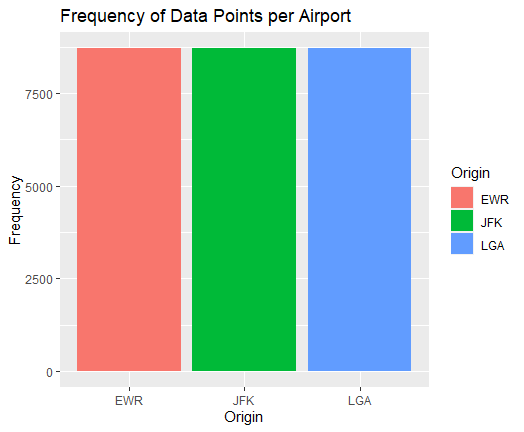


1. Use ggplot to create the following graphs. Add a title and axis labels to each:
   1. A scatter plot of pressure vs. wind speed, colored by airport. You may want to take a sample of the dataset for this graph. See <https://stackoverflow.com/questions/23279550/select-every-nth-row-from-dataframe>

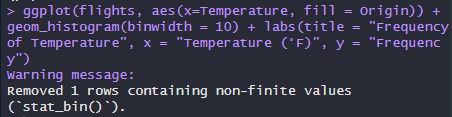


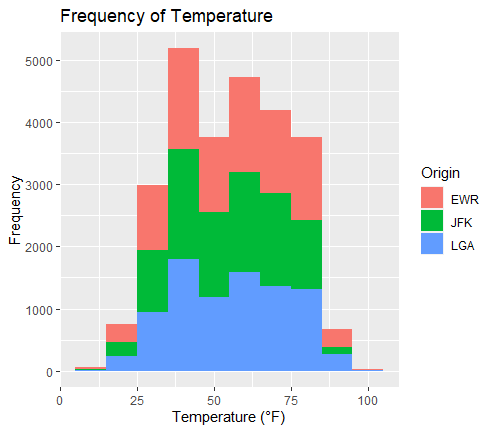


* 1. A bar plot of the number of data points for each airport in the dataset.

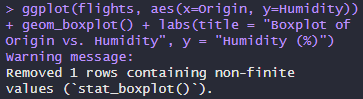
 

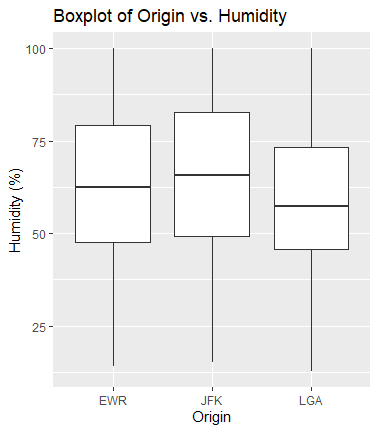
* 1. A histogram of the temperatures across all airports. Use 10 bins.





* 1. A box plot of the humidity for each airport.





* 1. A facet wrap with a separate scatter plot for each airport showing the temperature vs. wind speed.

