**R AND RSTUDIO INSTALLATION AND TUTORIAL**

Answer key for instructors and teaching assistants!

**EXPLORING R**

(Slide 5 - *Helpful Commands #1*)  
Test your console by typing print("Hello World!") in the console and pressing Enter.  
"Hello World!"

(Slide 10 - *Viewing Data*)  
air\_quality\_data\_full – lists the entire dataframe for you.  
How many entries are there in total? Find line 43 and paste it below, what day and month is it? What is the wind speed on that day (with units)?

| 153 lines  Line: 43 NA 250 9.2 92 6 12  June 12  9.2 mph |
| --- |

colnames(air\_quality\_data\_full)  
Paste your output below. What character surrounds the column names?

| "Ozone" "Solar.R" "Wind" "Temp" "Month" "Day"  quotes |
| --- |

str(air\_quality\_data\_full)  
Paste your output below. There is an additional column here that shows the words “int” or “num” for each variable, what do “int” and “num” mean?

| 'data.frame': 153 obs. of 6 variables:  $ Ozone : int 41 36 12 18 NA 28 23 19 8 NA ...  $ Solar.R: int 190 118 149 313 NA NA 299 99 19 194 ...  $ Wind : num 7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...  $ Temp : int 67 72 74 62 56 66 65 59 61 69 ...  $ Month : int 5 5 5 5 5 5 5 5 5 5 ...  $ Day : int 1 2 3 4 5 6 7 8 9 10 …  Type of values in that column/associated to that variable. Int means integer and num means numeric, a floating point value (this second line of the answer isn’t required) |
| --- |

head(air\_quality\_data\_full).  
How many rows does the head function show? Paste your output below.

| 6 rows  1 41 190 7.4 67 5 1  2 36 118 8.0 72 5 2  3 12 149 12.6 74 5 3  4 18 313 11.5 62 5 4  5 NA NA 14.3 56 5 5  6 28 NA 14.9 66 5 6 |
| --- |

length(air\_quality\_data\_full$Ozone) – $ notation denotes columns.  
Paste your output, what does this number correspond to?

| 153  The number of rows in the dataset with Ozone (includes NA) |
| --- |

(Slide 12 - *Verifying Data*)  
Make sure you have **2** variables/columns (**Ozone** and **Month**) and **153** lines of data. Sometimes a line gets lost due to programming or file formatting errors; always double check that your data are correct before proceeding.  
  
In the *Environment* tab/window/panel, double-click on air\_quality\_data.  
  
This will open up the data sheet in the Viewer panel. In addition, what command(s) that you already learned can you use to show that you only have Ozone and Month now?

| colnames(air\_quality\_data)  Also true:  air\_quality\_data  str(air\_quality\_data)  head(air\_quality\_data) |
| --- |

(Slide 14 - *Plot 1: Histogram*)  
Let’s check what hist can do for us. Execute this line:  
  
help(hist)  
  
hist can take many different arguments, but at least one is required: *x*. The other arguments that you should know are *breaks*, *main*, *xlab*, *ylab*, *xlim*, and *ylim*. From the help function, summarize what these arguments represent and used for:  
x:   
breaks:   
main, xlab, ylab:

xlim, ylim:

| From help(hist): x: a vector of values for which the histogram is desired.  breaks: one of:   * a vector giving the breakpoints between histogram cells, * a function to compute the vector of breakpoints, * a single number giving the number of cells for the histogram, * a character string naming an algorithm to compute the number of cells (see ‘Details’), * a function to compute the number of cells.   main, xlab, ylab: main title and axis labels.  xlim, ylim: the range of x and y values with sensible defaults. |
| --- |

(Slide 14 - *Plot 1: Histogram*)  
Create a histogram for ALL of the months data on ozone, using the default options:  
  
hist(air\_quality\_data$Ozone)  
  
Your plot will appear in the Plots panel. Paste your plot → oh! Next slide or next step in the full assignment will explain how to save/copy plots.

|  |
| --- |

(Slide 16 - *Plot 1: Histogram (con’t)*)  
This code includes an x-axis label, but deletes the graph’s title because it won’t be needed once you write a good caption.  
  
hist(air\_quality\_data$Ozone, xlab="ppb", main="")  
hist(air\_quality\_data$Ozone, xlab="ug/L", main="")  
  
Run both commands and paste your plots. Which argument deleted the title?

| main=”” deleted the title |
| --- |

(Slide 16 - *Plot 1: Histogram (con’t)*)  
To make the notation prettier you can correspond the expression (value) to a variable name, and use it later.  
  
hist(air\_quality\_data$Ozone, xlab=expression(paste(mu,"g/",L)), main="")  
  
Run the command and paste your plot. The expression command creates an expression object and assigns it to xlab, so what does the paste function do?

| Paste function: Concatenate vectors after converting to character. Also acceptable: attaches the comma separated characters together into one line |
| --- |

(Slide 17 - *Plot 1: Histogram (con’t)*)  
Change the number of bins in a histogram. R determines the number of bins (or breaks) to put in a histogram that makes it look "prettiest" using something called the Sturges method. Alternatively, you can tell R how many bins you would like it to create:  
  
hist(air\_quality\_data$Ozone, xlab="ppb", main="", breaks=22)  
hist(air\_quality\_data$Ozone, xlab="ppb", main="", breaks=5)  
  
Choose a number of bins/breaks you think is most appropriate, paste that plot, and explain why that number is most appropriate.

| 22 breaks:    5 breaks:    <flexible and up to you>, you probably want enough bins to show the secondary peak most visible with 22 breaks, but also visible with the default. Reasoning could mention the scale, the secondary peak, not making any bin too relatively small, etc. |
| --- |

(Slide 18 - *Plot 1: Histogram (con’t)*)  
Notice that you can change the look of the histogram by changing the range of values shown on the x-axis:  
  
hist(air\_quality\_data$Ozone, xlab="ppb", main="", breaks=22, xlim=c(0, 200))  
  
Choose an xlim that you think is best, explain why, and paste your plot.

| For xlim=c(0, 200):    <flexible and up to you>, reasoning may include having a round number, but also including all of the data, etc. |
| --- |

(Slide 18 - *Plot 1: Histogram (con’t)*)  
Play around with the other options (and more that you find in the help documentation for hist()): create a plot that you think is best, paste it below, and describe the distribution you see (i.e., shape, center, spread)**.**

Hint: hist() has lots of options, so look at its help page for details.

| <flexible and up to you>, should ideally describe that the pollutant level is concentrated at left, with a secondary peak in the middle, continuously falling off towards higher pollutant levels. Though detail may vary |
| --- |

(Slide 19 - *Refining and Segmenting Data*)  
Let’s create a histogram on ozone for each month.  
  
Brute-force method:  
hist(air\_quality\_data$Ozone[1:31], xlab="ppb", main="May")  
hist(air\_quality\_data$Ozone[32:61], xlab="ppb", main="June")  
hist(air\_quality\_data$Ozone[62:92], xlab="ppb", main="July")  
hist(air\_quality\_data$Ozone[93:123], xlab="ppb", main="August")  
hist(air\_quality\_data$Ozone[124:153], xlab="ppb", main="September")  
  
Logical selection way:  
hist(air\_quality\_data$Ozone[air\_quality\_data$Month==5], xlab="ppb", main="May")

What goes inside the square brackets after air\_quality\_data$Ozone (looking for a general answer)? Think about what the code in the brackets does.

| The range or selection, in rows, of data that you would like to use. Note, brackets index values are inclusive. |
| --- |

Run air\_quality\_data$Month==5 alone, what does this return (describe the output)?

| Returns an array of True/False (same size as the dataset), representing the Boolean value of each indice/row given the expression.  In this case, it returns an array of True/False where each index represents if that row in the air\_quality\_data has a Month value of 5. |
| --- |

Now, explain how the logical selection method works, and why this might be better than a brute-force approach.

| The hist function will create a histogram plot based on the data whose indice in the above expression produced True.  This is a more flexible approach – does not require explicit knowledge of the data, and any changes in the data (more rows, deletion, reordering edits) will not affect the output. |
| --- |

Paste a plot for each month and describe the distribution (i.e., shape, spread, outliers) of each plot.

| <flexible and up to you> |
| --- |

(Slide 23 - *Segmenting with lattice*)  
Describe at least one similarity and one difference between the new histogram function and the old hist function.  
  
You may use the following function call to help out.  
? histogram

| Sim: produces histograms…  Diff: histogram (from lattice) allows you to easily produce co-plots (conditional plots), i.e., a collection of histograms |
| --- |

(Slide 23 - *Segmenting with lattice*)  
The following arguments will be needed for histogram plots: *x*, *data*, and *type*.

From the previous step’s help type function, summarize what these arguments represent and used for:  
*x*:   
*data*:   
*type*:

| From ? histogram:  x: The object on which method dispatch is carried out.  data: For the formula method, an optional data source (usually a data frame) in which variables are to be evaluated.  type: A character string indicating the type of histogram that is to be drawn. |
| --- |

(Slide 23 - *Segmenting with lattice*)  
Execute the following lines *one at a time*:  
  
histogram(~ Ozone | Month, data=air\_quality\_data)  
histogram(~ Ozone | Month, data=air\_quality\_data, type="count")  
  
Notice the difference between the two histograms? Paste your plots and describe the difference.

| The default “type” or y-axis was Percent Total, which changes the tick mark values on the y-axis. The new plot has the count type and so lists the pure count of readings in that bar. |
| --- |

(Slide 24 - *Custom Graphics*)  
We will be using the par function (feel free to check it out: ? par ).  
  
Now we will set some graphical parameters ourselves. Build the figure we want:  
  
par(mfrow=c(5,1), mar=c(4,2,2,2)+0.01, oma=c(2,2,0,0))  
hist(air\_quality\_data$Ozone[1:31], xlab="", main="May", xlim=c(0, 200))  
hist(air\_quality\_data$Ozone[32:61], xlab="", main="June", xlim=c(0, 200))  
hist(air\_quality\_data$Ozone[62:92], xlab="", main="July", xlim=c(0, 200))  
hist(air\_quality\_data$Ozone[93:123], xlab="", main="August", xlim=c(0, 200))  
hist(air\_quality\_data$Ozone[124:153], xlab="", main="September", xlim=c(0, 200))  
mtext("ppb", 1, outer=TRUE, cex=1.5)  
mtext("Frequency", 2, outer=TRUE, cex=1.5)  
  
Paste your final plot and explain at least two pros and two cons of this plot compared to the earlier ones.

| Pros:  All plots have their x-axis parallel with each other for better evaluation vertically.  For some this is more visually appealing.  Cons:  Difficult to compare each month on y-axis (frequency).  Y-axis (frequency) ranges are squished to fit the same height – which can lead to misleading analysis or conclusions.  In general, it can look squished/tight for some people. |
| --- |

(Slides 25-26 - *Box Plots*)  
Create box plots for the whole dataset and for each month.  
  
boxplot(air\_quality\_data$Ozone, xlab="ppb", main="")  
boxplot(air\_quality\_data$Ozone ~ air\_quality\_data$Month, xlab="ppb", ylab="Month", main="")  
  
*Trouble?* If you are having trouble plotting the boxplot due to the previous par window formatting, run this before the boxplots:  
par(mfrow=c(1,1), mar=c(4,2,2,2)+0.01, oma=c(2,2,0,0))

Paste your plots. Describe what you see in the box plots, and how this is similar or different from the histograms.

| Still shows distribution of value/variable, but highlights the numerical summaries (minimum, first quartile, median, third quartile, maximum, and outliers (circles)). |
| --- |

Which months are captured well by the histogram? By the boxplot? Which are captured poorly by either or both of these months? Is there perhaps a third option worth pursuing?

| <flexible and up to you>, but a reasonable attempt should be made to answer all parts of this question. Scatter plot may be the best answer for the third option, but not the only answer. |
| --- |

(Slide 27 - *Numeric Summaries*)  
Numeric summaries – mean & SD vs. 5-number summary – can be important for discussing the distribution of each variable, if appropriate.

To calculate all of the numeric summaries by month:  
  
by(air\_quality\_data$Ozone, air\_quality\_data$Month, summary)

Write the output in the table below. Also, highlight the row (month) that has the greatest potential error. (Hint: what do the NAs in the data mean? Look at your data in a way that shows you NAs to learn more about this.)

| **Month** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** | **NA’s** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 5 | 1.00 | 11.00 | 18.00 | 23.62 | 31.50 | 115.00 | 5 |
| 6 | 12.00 | 20.00 | 23.00 | 29.44 | 37.00 | 71.00 | 21 |
| 7 | 7.00 | 36.25 | 60.00 | 59.12 | 79.75 | 135.00 | 5 |
| 8 | 9.00 | 28.75 | 52.00 | 59.96 | 82.50 | 168.00 | 5 |
| 9 | 7.00 | 16.00 | 23.00 | 31.45 | 36.00 | 96.00 | 1 |

We highlight month 6 because there are 21 NAs, more than half of the month! This could cause an inaccurate reading for that month.

(Slide 28 - *Removing NAs*)  
At times, calls/functions might not work properly with your dataset as it includes NA’s. To clean or remove NA’s from datasets:

nas <- is.na(air\_quality\_data) # Returns a logical array of TRUEs where NAs are in the original set. New array is same size as data  
  
air\_quality\_data\_cleaned <- air\_quality\_data[complete.cases(air\_quality\_data), ] # Note: we need the comma and space in the brackets to select all rows and columns to check  
  
Check this cleaned up dataset – what is the number of the last row? Are there that many rows?

| 153 20 9  No, only 116 rows |
| --- |

(Slide 29)  
Tidyverse package and examples.  
  
To install and load:  
  
install.packages("tidyverse")  
install.packages("ggplot2")  
**library**(tidyverse)  
**library**(ggplot2)  
  
Go through each example, paste your plot, and in one sentence describe the possible purpose of designing such a plot.

1. Histograms
   1. (Slide 30) Ex. 1:  
      qplot(air\_quality\_data$Ozone,  
       geom = "histogram", # graph type  
       binwidth = 3, # bin size  
       main = "Histogram of Ozone", # figure title  
       xlab = "ppb", # label for x axis  
       fill = I("green"), # fill color  
       col = I("black"), # outline color  
       alpha = I(.5), # fill transparency (between 0-see through and 1-opaque)  
       xlim = c(0, 200)) # x axis range

| <flexible and up to you> |
| --- |

* 1. (Slide 31) Ex. 2:  
     ggplot(data = air\_quality\_data, aes(air\_quality\_data$Ozone)) +  
      geom\_histogram(breaks = seq(0, 200, by = 3),  
      col = "black",  
      aes(fill = ..count..)) +  
      scale\_fill\_gradient("Count", low = "green", high = "red") +  
      labs(title = "Histogram of Ozone", x = "ppb", y = "Count")

| <flexible and up to you> |
| --- |

* 1. (Slide 32) Ex. 3:  
     multi <- air\_quality\_data %>% #tidyverse's "piping" functionality with "%>%"  
      ggplot(aes(x = Ozone, color = Month, fill = Month)) +  
      geom\_histogram(alpha = 0.5, binwidth = 3) +  
      theme(legend.position = "none") +  
      xlab("") +  
      ylab("ppb") +  
      facet\_wrap(~Month)  
     multi

| <flexible and up to you> |
| --- |

1. Boxplots
   1. (Slide 33) Ex. 1:  
      outliboxplot <- air\_quality\_data %>%  
       ggplot(aes(x = Month, y = Ozone, fill = Month)) +  
       geom\_boxplot(outlier.color = "black", outlier.shape = 8, outlier.size = 2) + # outlier boxplot  
       stat\_summary(fun = mean, geom = "point", shape = 23, size = 4) # add mean diamond  
      outliboxplot

| <flexible and up to you> |
| --- |

* 1. (Slide 34) Ex. 2:  
     dotnboxplot <- air\_quality\_data %>%  
      ggplot(aes(x = Month, y = Ozone)) +   
      geom\_boxplot() +   
      geom\_jitter(color = "black", size = 0.9, alpha = 0.8) +  
      theme(panel.border = element\_blank(),  
      panel.background = element\_blank(),  
      panel.grid.major = element\_blank(),  
      panel.grid.minor = element\_blank(),  
      axis.line = element\_line(color = "black"))  
     dotnboxplot

| <flexible and up to you> |
| --- |

1. Numeric Summaries  
     
   ? summarise  
     
   The following calls/functions will not perform correctly if there are NA’s in your dataset. Please remove them beforehand:  
     
   air\_quality\_data\_cleaned <- air\_quality\_data[complete.cases(air\_quality\_data), ]
   1. (Slide 35) Ex. 1:  
      air\_quality\_data\_cleaned %>%  
       group\_by(Month) %>%  
       summarise(count = n(),  
       df = count-1,  
       mean = mean(Ozone),  
       sd = sd(Ozone))

| # A tibble: 5 × 5  Month count df mean sd  <int> <int> <dbl> <dbl> <dbl>  1 5 26 25 23.6 22.2  2 6 9 8 29.4 18.2  3 7 26 25 59.1 31.6  4 8 26 25 60.0 39.7  5 9 29 28 31.4 24.1 |
| --- |

* 1. (Slide 36) Ex. 2:  
     air\_quality\_data\_cleaned %>%  
      group\_by(Month) %>%  
      summarise(count = n(),  
      df = count-1,  
      min = min(Ozone),  
      Q1 = quantile(Ozone, 0.25),  
      M = median(Ozone),  
      Q3 = quantile(Ozone, 0.75),  
      max = max(Ozone))

| # A tibble: 5 × 8  Month count df min Q1 M Q3 max  <int> <int> <dbl> <int> <dbl> <dbl> <dbl> <int>  1 5 26 25 1 11 18 31.5 115  2 6 9 8 12 20 23 37 71  3 7 26 25 7 36.2 60 79.8 135  4 8 26 25 9 28.8 52 82.5 168  5 9 29 28 7 16 23 36 96 |
| --- |

(Slide 37 - *Final Plot: Your Choice!*)  
Finally, see if you can create a figure (possibly with multiple panels) comparing the distributions, this time, of Wind or Temperature. The exact nature of the figure is up to you; if you have multiple panels, just be sure to label the separate panels. You should also be sure to include descriptions of the distribution (i.e., shape, spread, outliers) of each month or note any observations of the data that you had.

| <flexible and up to you> |
| --- |

Ask your TA for help if needed.

**R RESOURCES**

* [R Resources maintained by UCLA](https://stats.oarc.ucla.edu/r/)
* [R Bootcamp](https://www.jaredknowles.com/r-bootcamp/)
* [Quick-R](https://www.statmethods.net/) (*reference guide*)
* [RStudio Cheatsheets](https://www.rstudio.com/resources/cheatsheets/)
* [R Graph Gallery](https://www.r-graph-gallery.com/index.html) *(Tidyverse and ggplot2 help*)
* [GGPlot Color Tricks](https://www.datanovia.com/en/blog/ggplot-colors-best-tricks-you-will-love/) (*for color-blind-friendly graph-making*)
* [Stack Overflow](https://stackoverflow.com/questions/tagged/r) (*public forum for script-writing questions*)
* [R for Data Science](https://r4ds.had.co.nz/) (*by Hadley Wickham*)
* [Kickstarter guide](https://paulvanderlaken.com/2017/10/18/learn-r/) (*for new R users*)
* [Curated list of packages and tools](https://github.com/qinwf/awesome-R/blob/master/README.md) (*on Github*)