**R AND RSTUDIO INSTALLATION AND TUTORIAL**

**TEMPLATE: Please copy content/questions to your own document.**

**Your name:**

**INSTALLATION**

No responses needed.

**EXPLORING R**  
air\_quality\_data\_full – lists the entire dataframe for you.  
**Paste your output.**

|  |
| --- |

colnames(air\_quality\_data\_full) – lists the column names for you.  
**Paste your output.**

|  |
| --- |

str(air\_quality\_data\_full) – tells you about the structure of your dataset; gives # of rows, # of columns (variables), and the object (modeling) type of each variable.  
**Paste your output.**

|  |
| --- |

head(air\_quality\_data\_full) – shows the first 6 rows of data.  
**Paste your output.**

|  |
| --- |

length(air\_quality\_data\_full$Ozone) – gives # of rows, and $ notation denotes columns.  
**Paste your output.**

|  |
| --- |

Make sure you have **2** variables/columns (**Ozone** and **Month**) and **153** lines of data. **Alternatively, what command(s) that you already learned can you use to show that you only have ozone and month now?**

|  |
| --- |

**HISTOGRAMS**

help(hist)  
**Summarize the information that is provided about the hist function.**

|  |
| --- |

hist(air\_quality\_data$Ozone)  
**Paste your plot.**

|  |
| --- |

hist(air\_quality\_data$Ozone, xlab="ppb", main="")  
hist(air\_quality\_data$Ozone, xlab="ug/L", main="")  
**Paste your plots.**

|  |
| --- |

hist(air\_quality\_data$Ozone, xlab=expression(paste(mu,"g/",L)), main="")  
micrograms\_per\_liter <- expression(paste(mu,"g/",L))  
hist(air\_quality\_data$Ozone, xlab=micrograms\_per\_liter, main="")  
**Paste your plots.**

|  |
| --- |

hist(air\_quality\_data$Ozone, xlab="ppb", main="", breaks=22)  
hist(air\_quality\_data$Ozone, xlab="ppb", main="", breaks=5)  
**Choose a bin/breaks number that makes most sense to you, explain why, and paste your plot.**

|  |
| --- |

hist(air\_quality\_data$Ozone, xlab="ppb", main="", breaks=22, xlim=c(0, 200))  
**Paste your plot.**

|  |
| --- |

**Play around with the options, like the number of bars used, for example, to see how that affects the interpretations of the graph. Choose your favorite plot, paste it below, and describe the distribution (i.e., shape, center, spread).**

|  |
| --- |

hist(air\_quality\_data$Ozone[1:31], xlab="ppb", main="May")

…

OR  
hist(air\_quality\_data$Ozone[air\_quality\_data$Month==5], xlab="ppb", main="May")

…

**Paste your plots and describe the distribution (i.e., shape, spread, outliers) of each month’s plot.**

|  |
| --- |

install.packages("lattice") **library**(lattice)  
? histogram  
**Summarize the similarities and differences between this function and the previous hist function.**

|  |
| --- |

histogram(~ Ozone | Month, data=air\_quality\_data)  
histogram(~ Ozone | Month, data=air\_quality\_data, type="count")  
  
**Paste your plots and describe the differences.**

|  |
| --- |

**CUSTOM GRAPHICS**

? 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 the pros/cons of this plot compared to the earlier ones.**

|  |
| --- |

**BOX PLOTS**  
  
boxplot(air\_quality\_data$Ozone, xlab="ppb", main="")  
boxplot(air\_quality\_data$Ozone ~ air\_quality\_data$Month, xlab="ppb", ylab="Month", main="")  
# you may need par(mfrow=c(1,1), mar=c(4,2,2,2)+0.01, oma=c(2,2,0,0))  
OR  
points(air\_quality\_data$Month, air\_quality\_data$Ozone, col="red", cex=0.5)

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

|  |
| --- |

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

|  |
| --- |

**NUMERIC SUMMARIES**

by(air\_quality\_data$Ozone, air\_quality\_data$Month, summary)

**Write the output in the table below, and try to note where and when summaries are relevant or not relevant to include in reports.**

| **Month** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** | **NA’s** |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

air\_quality\_data\_cleaned <- air\_quality\_data[complete.cases(air\_quality\_data), ]  
  
**Check this cleaned up dataset – are there any changes or differences?**

|  |
| --- |

**TIDYVERSE AND GGPLOT**

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.**   
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

|  |
| --- |

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")

|  |
| --- |

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

|  |
| --- |

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

|  |
| --- |

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

|  |
| --- |

? 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), ]

air\_quality\_data\_cleaned %>%  
 group\_by(Month) %>%  
 summarise(count = n(),  
 df = count-1,  
 mean = mean(Ozone),  
 sd = sd(Ozone))

|  |
| --- |

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))

|  |
| --- |

**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. 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*)