Summer 2022: Data Analysis
Homework I
Due: TBA
Submit through Canvas

Instructions: Provided solutions to these questions using this template. Include graphics with your solutions. Put all code an appendix to this homework. Use the *verbatim* command to leave code unchanged.

1. This question considers the *airquality* data set in R.

a) Provide a summary of the data. That is, what are the variables. Provide summary statistics for each variable.

Solution:
Below are the summary stats for the variables of *airquality:*

```
     Ozone           Solar.R           Wind
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400
 Median : 31.50   Median :205.0   Median : 9.700
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500
 Max.   :168.00   Max.   :334.0   Max.   :20.700
 NA's   :37       NA's   :7
      Temp           Month            Day
 Min.   :56.00   Min.   :5.000   Min.   : 1.0
 1st Qu.:72.00   1st Qu.:6.000   1st Qu.: 8.0
 Median :79.00   Median :7.000   Median :16.0
 Mean   :77.88   Mean   :6.993   Mean   :15.8
 3rd Qu.:85.00   3rd Qu.:8.000   3rd Qu.:23.0
 Max.   :97.00   Max.   :9.000   Max.   :31.0
```

b) Make a time series plot for two variables in the *airquality* data set. Make sure to label each plot carefully. Extra credit for labeling the months of the data on the x-axis.
Solution:

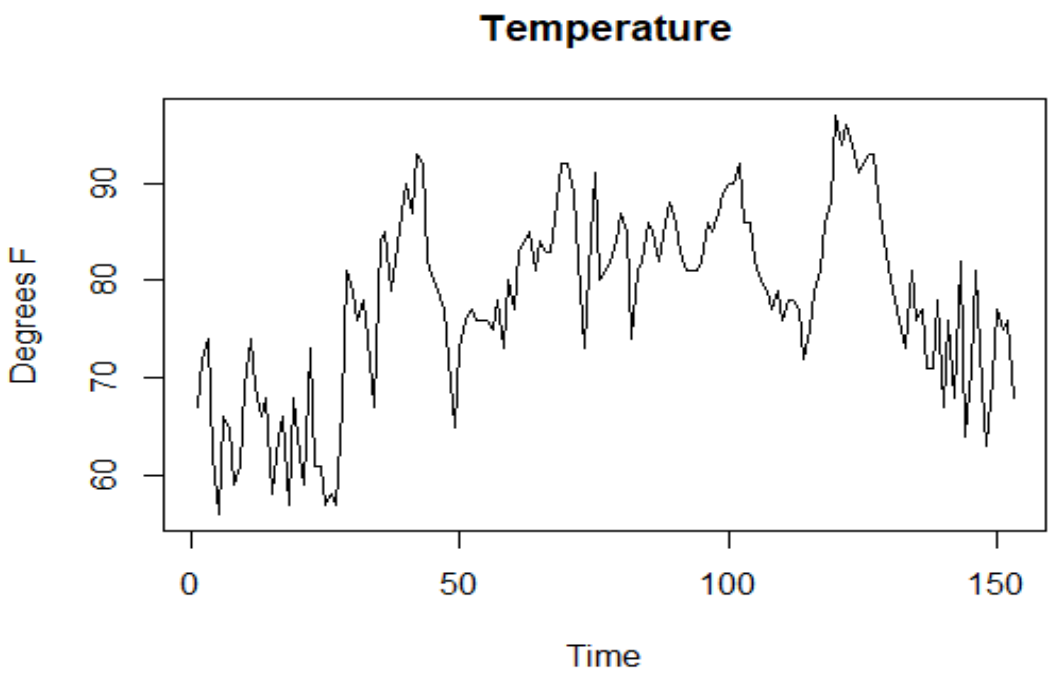

Figure 1: This Figure represents air temperature in New York City

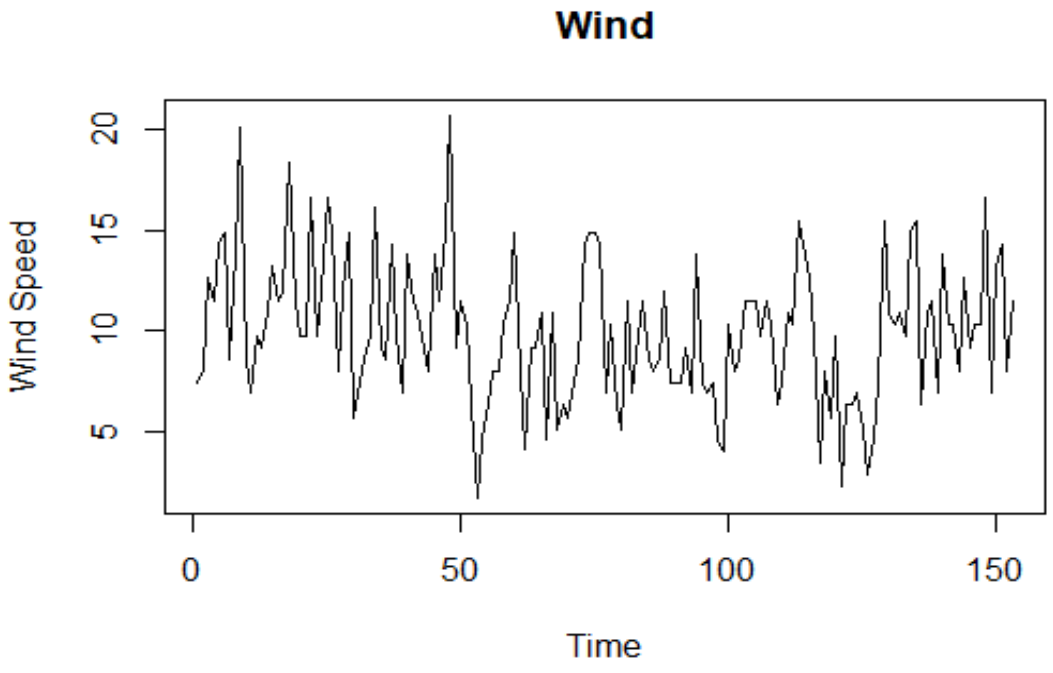

Figure 2: This Figure represents wind speed in New York City

c) Make side by side boxplots for the ozone data for each month. Do you believe the distribution of the ozone is the same for each month.

Solution:

The ozone layer distribution seems to be effected by the changing of months. This can be due to either the months changing themselves or other correlation months have, such as; Temperature, Wind Speed, etc.

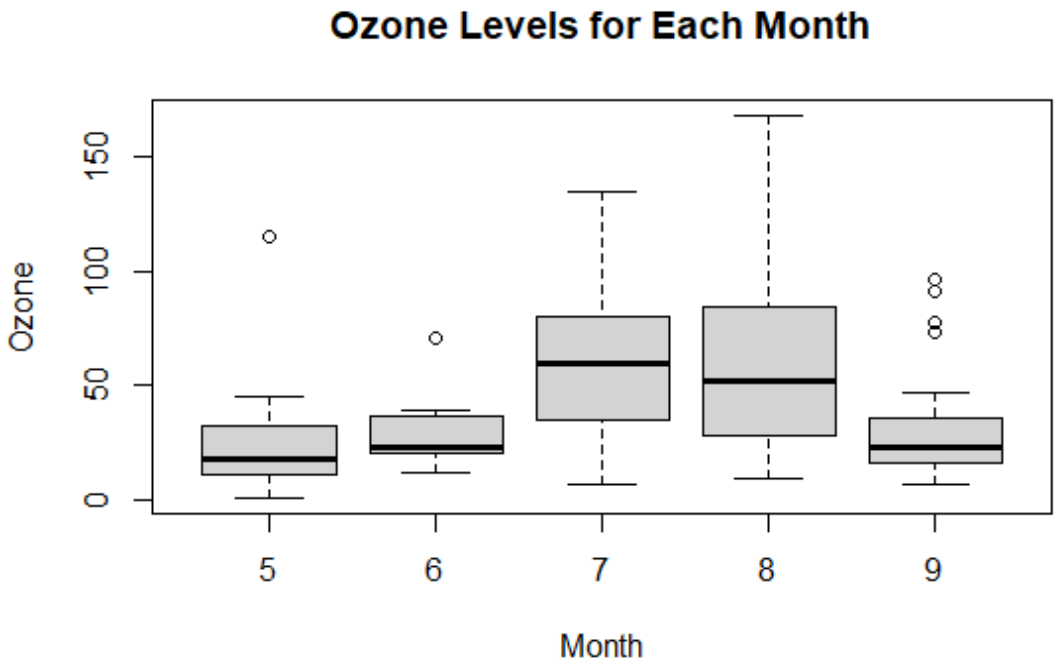

Figure 3: This Figure represents ozone levels each month in New York City

d) Compute the correlation and covariance matrix for the variables in the *airquality* data set.

Solution:

Below are the summary stats for the variables of

	Ozone	Solar.R	Wind	Temp	Month	Day
Ozone	NA	NA	NA	NA	NA	NA
Solar.R	NA	NA	NA	NA	NA	NA
Wind	NA	NA	12.4115385	-15.272136	-0.8897532	0.8488519
Temp	NA	NA	-15.2721362	89.591331	5.6439628	-10.9574303
Month	NA	NA	-0.8897532	5.643963	2.0065359	-0.0999742
Day	NA	NA	0.8488519	-10.957430	-0.0999742	78.5797214

	Ozone	Solar.R	Wind	Temp	Month	Day
Ozone	1	NA	NA	NA	NA	NA
Solar.R	NA	1	NA	NA	NA	NA
Wind	NA	NA	1.0000000	-0.4579879	-0.178292579	0.027180903
Temp	NA	NA	-0.4579879	1.0000000	0.420947252	-0.130593175
Month	NA	NA	-0.1782926	0.4209473	1.000000000	-0.007961763
Day	NA	NA	0.0271809	-0.1305932	-0.007961763	1.000000000

Figure 4: This is the Correlation and Covariance matrix of the *airquality* data set

Use this equation to solve questions below:

$$f(x) = \frac{70}{69x^2} \tag{1}$$

2. Get the CDF
Solution:

$$CDF = F(t) = \int_1^t f(x)\,dx \tag{2}$$

$$F(t) = \int_1^t \frac{70}{69x^2}\,dx \tag{3}$$

$$F(t) = [-\frac{70}{69t} + \frac{70}{69}] \tag{4}$$

3. Find the chances package weighs 20+ lbs
Solution:
I use 1 minus the CDF since we are looking for packages above 20.

$$F(t) = 1 - [-\frac{70}{69t} + \frac{70}{69}] \tag{5}$$

$$F(20) = 1 - [-\frac{70}{69(20)} + \frac{70}{69}] \tag{6}$$

$$F(20) = .036 = 3.6\% \tag{7}$$

4. Get μ and σ^2
Solution:

$$\mu = \int_a^b x f(x)\, dx \tag{8}$$

$$\mu = \int_1^{70} \frac{70}{69x^2}\, dx \tag{9}$$

$$\mu = \int_1^{70} \frac{70}{69x}\, dx \tag{10}$$

$$\mu = [\frac{70}{69} ln(70) - \frac{70}{69} ln(1)] \tag{11}$$

$$\mu = 4.31 \tag{12}$$

$$\sigma^2 = \int_a^b (x^2 - x) f(x)\, dx \tag{13}$$

$$\sigma^2 = \int_1^{70} x^2 \frac{70}{69x^2}\, dx - \mu \tag{14}$$

$$\sigma^2 = [\frac{70}{69} x]_1^{70} - \mu \tag{15}$$

$$\sigma^2 = 65.61 \tag{16}$$

5. If Shipping cost \$5 per lbs
Solution:
Hint: Get E[Shipping Cost]
If E[x] is just μ then my assumption for this problem is to multiply \$5 by μ. This would be \$21.55 per box shipped on average.