

**p.6 #3.** An automobile assembly line is manned by two shifts a day. The first shift accounts for two-thirds of the overall production. Quality control engineers want to compare the average number of nonconformances per car in each of the two shifts.

(a) Describe the population(s) involved.

The population consists of employees from two different work shifts, first shift and second shift for an automobile assembly line.

(b) Is (are) the population(s) involved hypothetical or not?

They are not hypothetical because all members of the population are available for census and study is on what has been done already and contains no future unknown data.

(c) What is the characteristic of interest?

The characteristic of interest is the nonconformances per car

*Cars are the pop.*  
*All cars are included at?*  
*Not employees (cars)*

**p10. #6.** A service agency wishes to assess its clients' views on quality of service over the past year. Computer records identify 1000 clients over the past 12 months, and a decision is made to select 100 clients to survey.

(a) Describe a procedure for selecting a simple random sample of 100 clients from last year's population of 1000 clients.

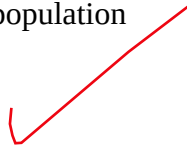
Using R's library tidyverse, a dataframe with 1000 rows can be created with one column named "ethnicity". This column will be populated with a vector containing 800 elements of the string "Caucasian-Americans", 150 elements of the string "African-Americans", and 50 elements of the string "Hispanic-Americans". This vector can be achieved by using the concatenate function `c()` with the function `replicate()`. Lastly in order to populate the vector in a random fashion, use the function `sample()` with the vector and set the size to 1000, `repeat = F`. Now that the data frame has been constructed, again use the function `sample()` with the column "ethnicity" in the data frame with a sample size of 100.

(b) The population of 1000 clients consists of 800 Caucasian-Americans, 150 African-Americans and 50 Hispanic-Americans. Describe an alternative procedure for selecting a representative random sample of size 100 from the population of 1000 clients.

(c) Give the R commands for implementing the sampling procedures described in parts (a) and (b).

```
> df1 = data.frame("Ethnicity"=sample(c(replicate(800,"Caucasian-American"),
+                                       replicate(150,"African-American"),
+                                       replicate(50,"Hispanic-American")),1000))
>
> # Perform sanity check to be sure data frame is correctly populated
> sum(df1[["Ethnicity"]]=="Caucasian-American")
[1] 800
> sum(df1[["Ethnicity"]]=="African-American")
[1] 150
> sum(df1[["Ethnicity"]]=="Hispanic-American")
[1] 50
>
```

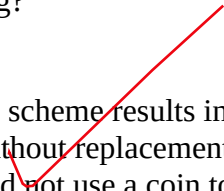
```
> # Take random sample of 100 population units from entire population
> #dat = df1[["Ethnicity"]]
> #sample(dat,100)
> s100 <- sample(df1[["Ethnicity"]],100)
```



**p.10 #8.** A particular product is manufactured in two facilities, A and B. Facility B is more modern and accounts for 70% of the total production. A quality control engineer wishes to obtain a simple random sample of 50 from the entire production during the past hour. A coin is flipped and each time the flip results in heads, the engineer selects an item at random from those produced in facility A, and each time the flip results in tails, the engineer selects an item at random from those produced in facility B.

Does this sampling scheme result in simple random sampling?  
Explain your answer.

No it does not. Using a coin toss to choose as describe in the scheme results in sampling with replacement. Simple random sampling involves sampling without replacement., and the scheme fails to mention this critical element. Simple random sampling would not use a coin toss at all, and a completely random sampling without replacement is required.




**p.13 #3.** At the final assembly point of BMW cars in Graz, Austria, the car's engine and transmission arrive from Germany and France, respectively. A quality control inspector, visiting for the day, selects a simple random sample of  $n$  cars from the  $N$  cars available for inspection, and records the total number of engine and transmission nonconformances for each of the  $n$  cars.

(a) Is the variable of interest univariate, bivariate or multivariate?

Univariate because he is looking at the engine and transmission conformances together

(b) Is the variable of interest quantitative or qualitative?

The variable of interest is quantitative because it is recorded as a number.



(c) Describe the statistical population.

The statistical population consists of the  $N$  cars available and the total number of engine and transmission nonconformities for those  $n$  cars.

(d) (d) Suppose the number of nonconformances in the engine and transmission are recorded separately for each car. Is the new variable univariate, bivariate, or multivariate?

The new variable is bivariate since it now has exactly two characteristics, engine nonconformances and transmission nonconformances.

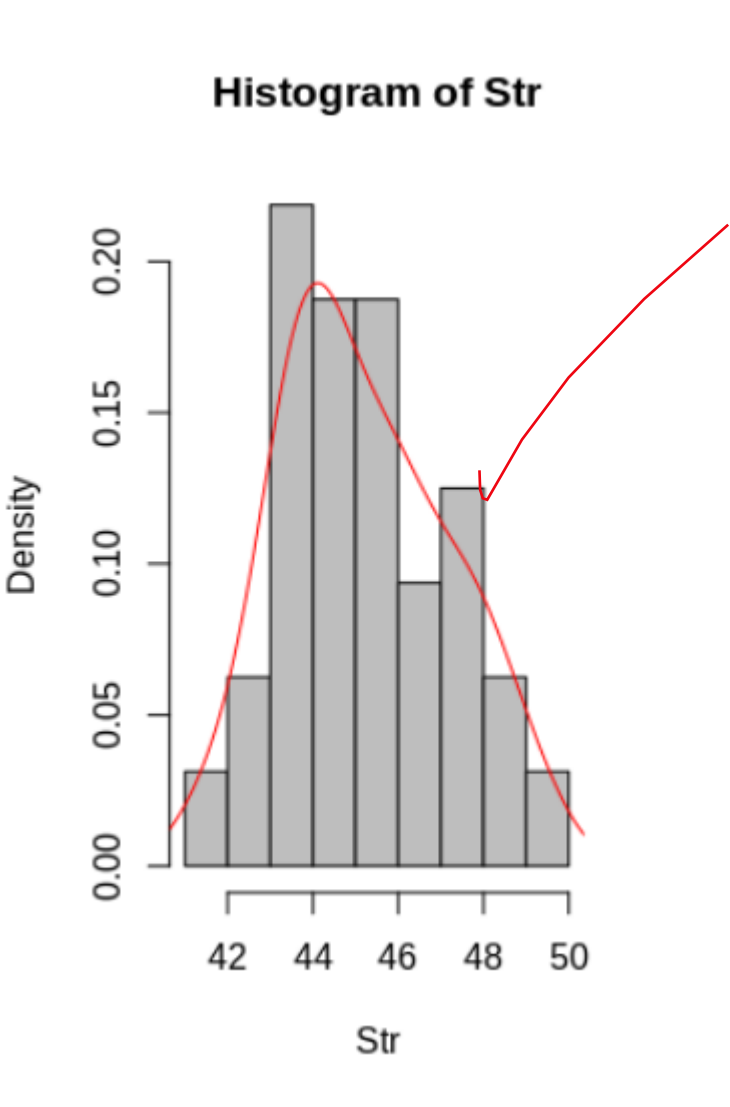
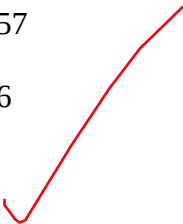
**pp.20-21 #1.** Use `cs = read.table("Concr.Strength.1s.Data.txt", header = T)` to read into the R object `cs` data on 28-day compressive-strength measurements of concrete cylinders using water/cement ratio 0.4. 8 Then use the commands `attach(cs)`; `hist(Str, freq = FALSE)`; `lines(density(Str))`; `stem(Str)` to produce a histogram with the smooth histogram superimposed, and a stem and leaf plot.

```
> path = "/home/scott/Documents/STAT401/akritas_datasets/"
> setwd(path)
> cs = read.table("Concr.Strength.1s.Data.txt", header = T)
```

```
attach(cs); hist(Str, freq=FALSE,col='gray'); lines(density(Str),col='red'); stem(Str)
```

The decimal point is at the |

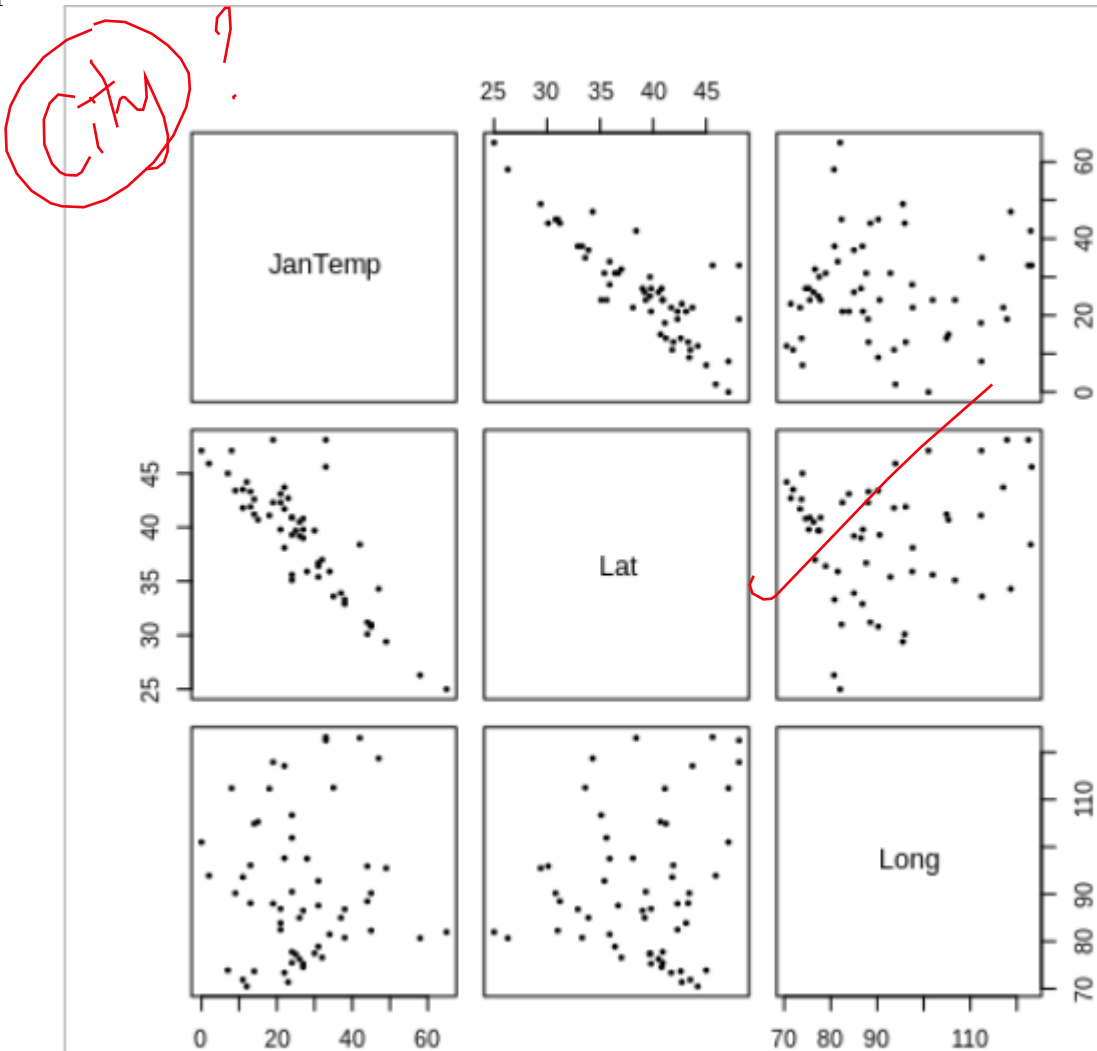
```
41 | 5
42 | 39
43 | 1445788
44 | 122357
45 | 1446
46 | 00246
47 | 3577
48 | 36
49 | 3
```



**pp.20-21 #4.** The data in Temp.Long.Lat.txt give the average (over the years 1931 to 1960) daily minimum January temperature in degrees Fahrenheit with the latitude and longitude of 56 US cities.

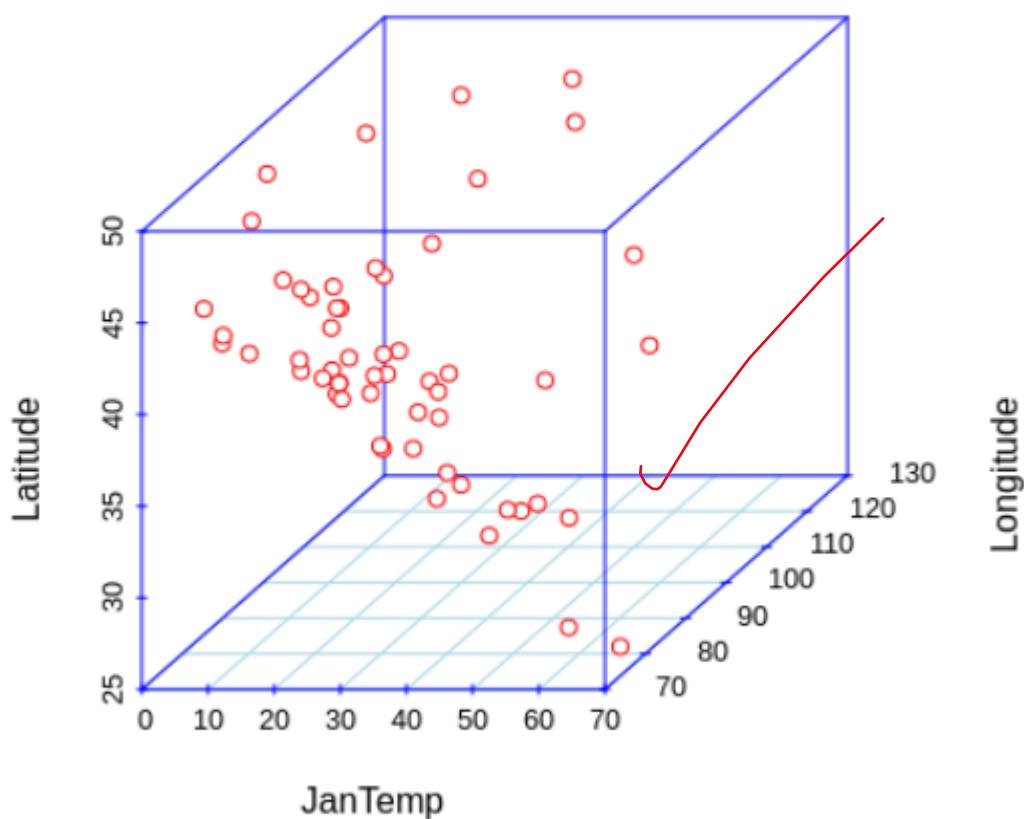
(a) Construct a scatterplot matrix of the data. Does longitude or latitude appear to be the better predictor of a city's temperature? Explain in terms of this plot.

As found in the scatterplot matrix below, the relationship between Lat and JanTemp is a lot more linear than with Long and JanTemp which is all over the place. Latitude is the better predictor of a city's temperature.



(b) Construct a 3D scatterplot of the data. Does longitude or latitude appear to be the better predictor of a city's temperature? Explain in terms of this plot.

Shown below in the 3D scatter plot, latitude still appears to be the better predictor. Similarly to the scatterplot matrix, latitude and temperature appear to have a linear relationship, and longitude does not.



**pp.20-21 #7.** Read the data on the average stopping times (on a level, dry # stretch of highway, free from loose material) of cars and trucks at various speeds into the data frame `bd` by `bd = read.table("SpeedStopCarTruck.txt", header = T)`. Then, use commands similar to those for Figure 1-6, given in Section 1.5.2 to plot the data using colors to differentiate between cars and trucks.

# Add a legend to the plot.

```
bd = read.table("SpeedStopCarTruck.txt", header = T)
```

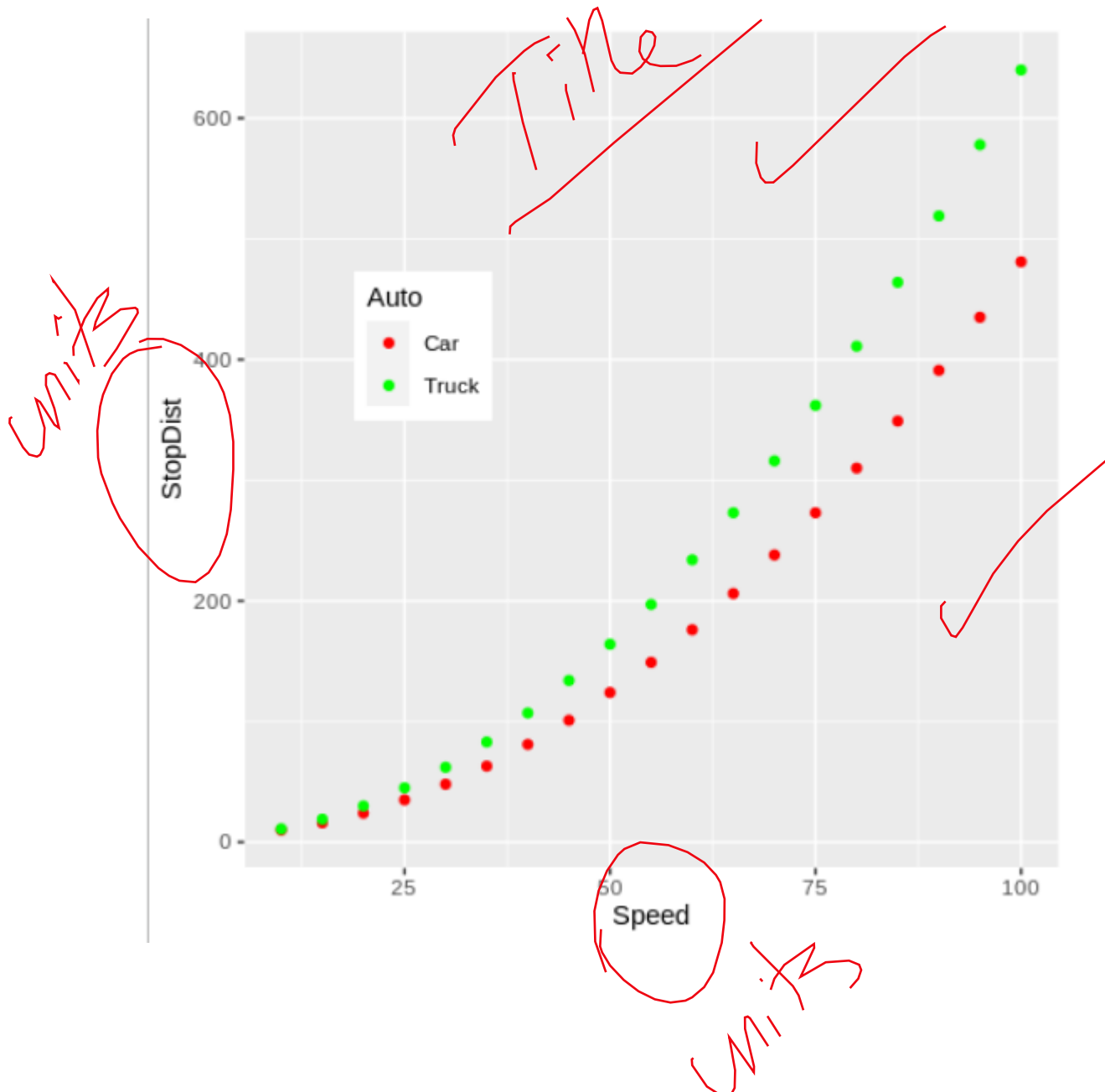
```
bd %>%
```

```
  ggplot(data=., aes(x=Speed, y=StopDist, color = Auto)) +
```

```
  geom_point() +
```

```
  scale_color_manual(values = c("red", "green")) +
```

```
  theme(legend.position = c(22/max(bd$Speed), 400/max(bd$StopDist)))
```



**pp.20-21 #13.** The R data set `airquality` contains daily ozone measurements, temperature, wind, solar radiation, month, and day. Use the command `pairs(airquality[1:5])` to produce a scatterplot matrix for the first five variables of this data set, and answer the following questions:

```
air = airquality;
```

```
pairs(airquality[1:5])
```

# (a) Is it more likely to have higher ozone levels on hot days?

Yes it is.

# (b) Is it more likely to have higher ozone levels on windy days?

No it is not.

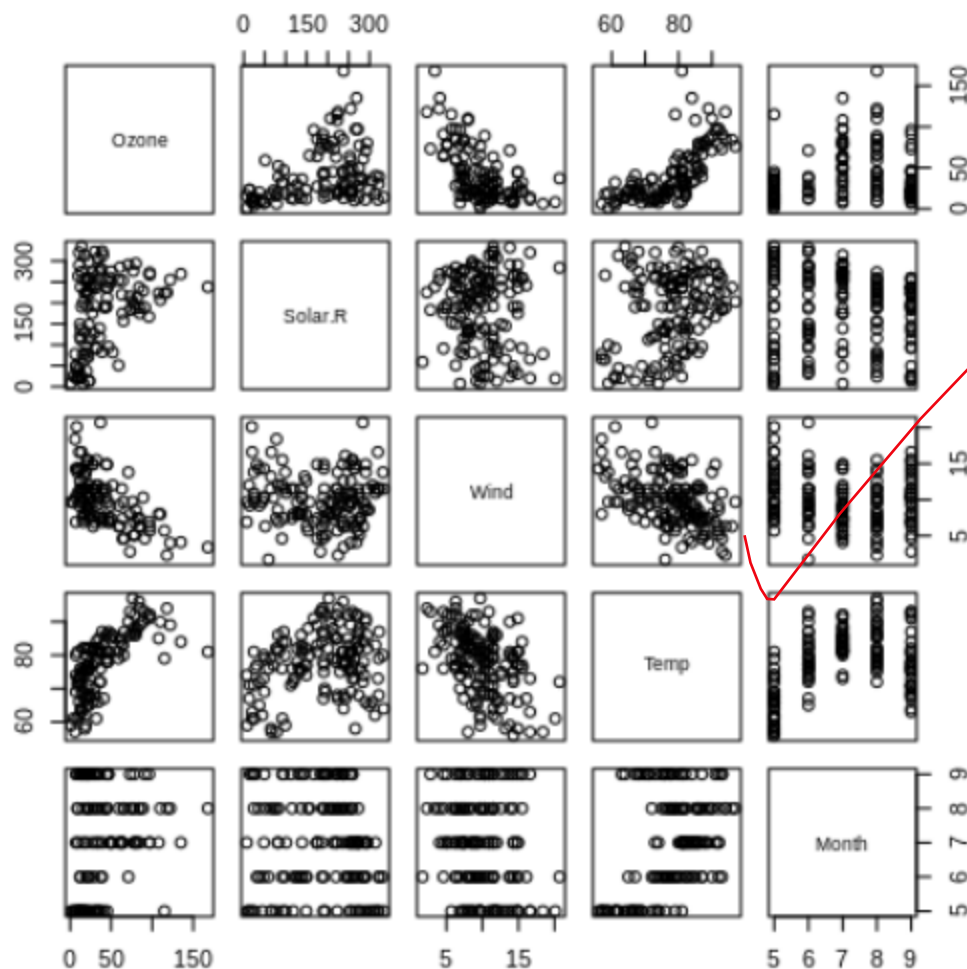
# (c) What seems to happen to ozone levels when there is increased solar radiation?

# radiation?

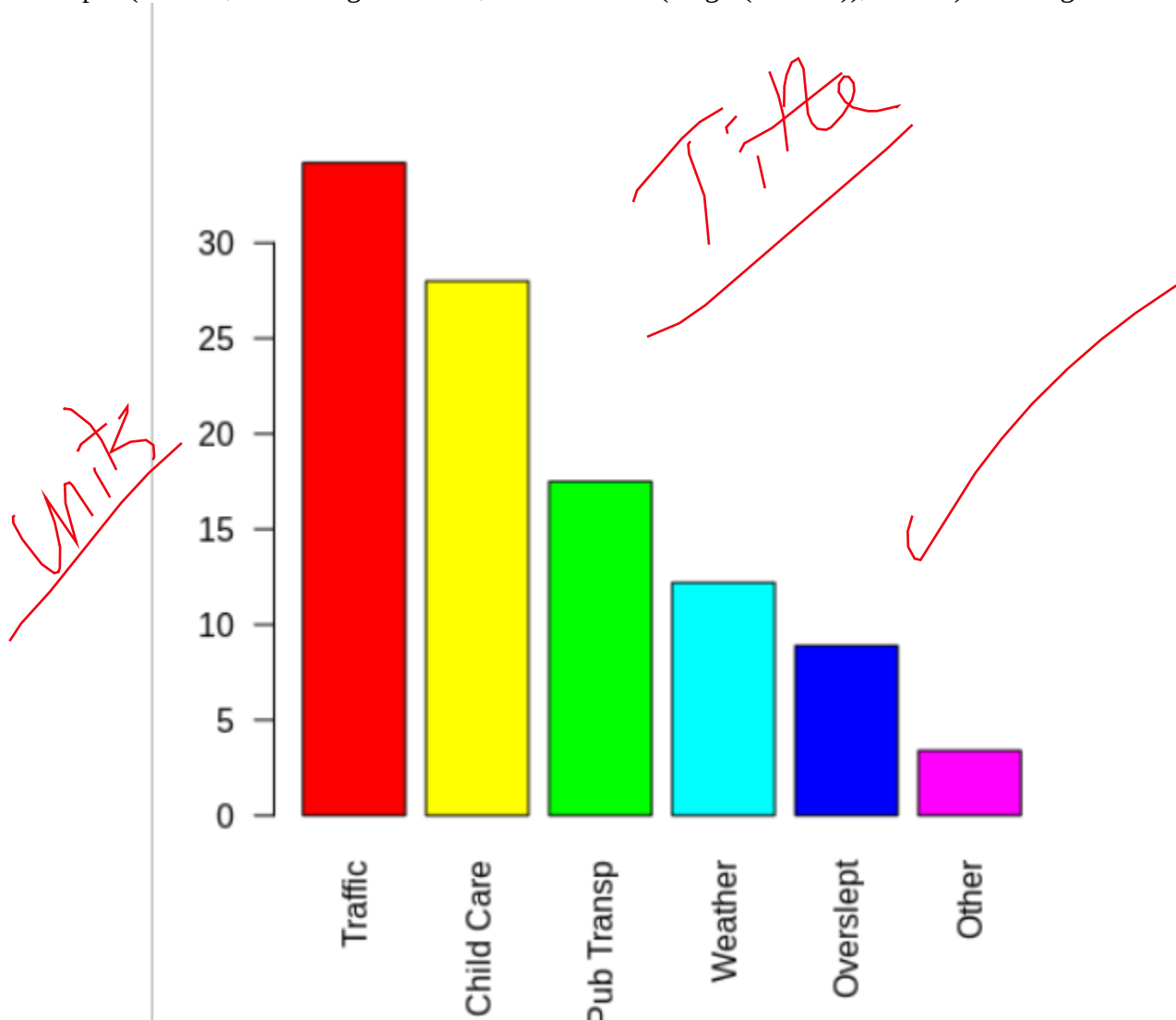
Ozone levels rise until the solar radiation level reaches approximately 175 then they start to decrease as solar radiation levels continue to rise

# (d) Which month seems to have the highest ozone levels?

The months of July and August seem to have the highest ozone levels.



**pp.20-21 #17.** Read the projected percents and reasons why people in the Boston area are late for work into the data frame `lw` using the command  
`lw = read.table("ReasonsLateForWork.txt", sep = ",", header = T).`  
 # (a) Construct a bar graph and a pie chart for this data using the commands given  
 # in Section 1.5.3.  
`barplot(Percent, names.arg = Reason, col = rainbow(length(Percent)), las = 2) # for Figure 1-11`

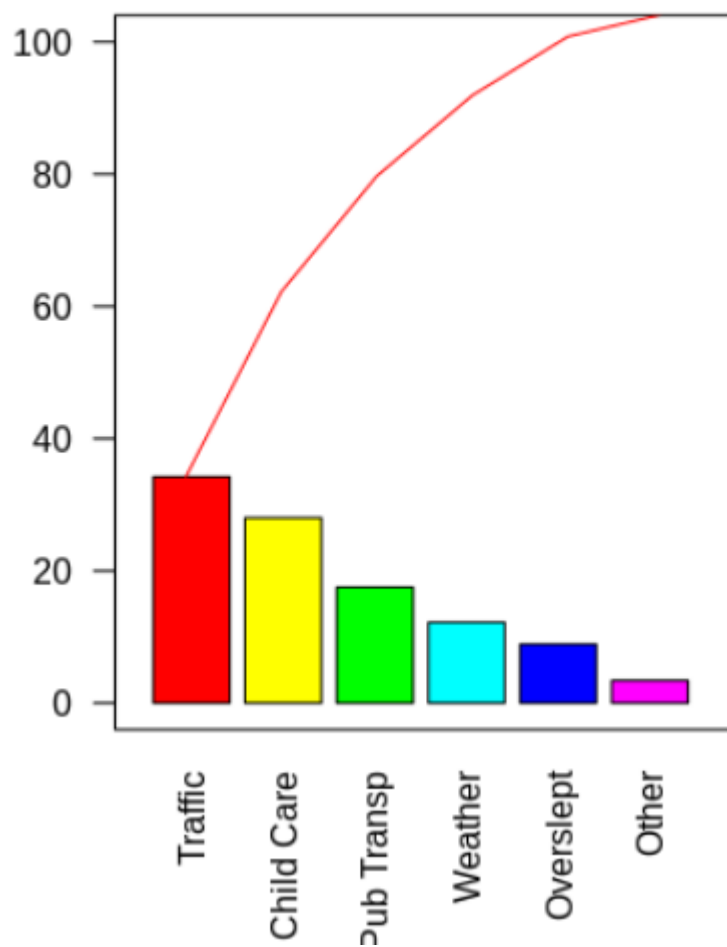


# (b) The bar graph constructed in part (a) above is actually a Pareto chart,  
 # since the bar heights are arranged in decreasing order. Use the commands  
`# attach(lw); plot(c(0, 6), c(0, 100), pch = " ", xlab = " ", ylab = " ", xaxt =`  
`# "n", yaxt = "n");`  
`# barplot(Percent, width = 0.8, names.arg = Reason, col = rainbow(length(Percent)))`  
`# , las = 2, add = T);`  
`# lines(seq(0.5, 5.5, 1), cumsum(Percent), col = "red")`  
 # to construct a variation of the Pareto chart, which also displays the cumulative  
 # percentages.

`attach(lw); plot(c(0, 6), c(0, 100), pch = " ", xlab = " ", ylab = " ", xaxt =`  
`"n", yaxt = "n");`



```
barplot(Percent, width = 0.8, names.arg = Reason, col = rainbow(length(Percent))  
      , las = 2, add = T);  
lines(seq(0.5, 5.5, 1), cumsum(Percent), col = "red")
```



**pp. 29-31 #1.** A polling organization samples 1000 adults nationwide and finds that the average duration of daily exercise is 37 minutes with a standard deviation of 18 minutes.

(a) The correct notation is for the number 37 is (choose one): (i)  $x$ , (ii)  $\mu$ .

(i)  $\bar{x}$

(b) The correct notation is for the number 18 is (choose one): (i)  $S$ , (ii)  $\sigma$ .

(i)  $S$

(c) Of the 1000 adults in the sample 72% favor tougher penalties for persons convicted of drunk driving. The correct notation for the number 0.72 is (choose one): (i)  $\hat{p}$ , (ii)  $p$ .

(i)  $\hat{p}$

**pp. 29-31 #4.** Use `cs = read.table("Concr.Strength.1s.Data.txt", header = T)` to read into the R object `cs` data on 28-day compressive strength measurements of concrete cylinders using water/cement ratio 0.4 (see footnote 5 in Exercise 1 in Section 1.5). Then use the commands `attach(cs)`; `sum(Str <= 44)/length(Str)`; `sum(Str >= 47)/length(Str)` to obtain the proportion of measurements that are less than or equal to 44, and greater than or equal to 47. What do these sample proportions estimate?

The first sample proportions estimates 0.3125, or 31.25% of cylinders of the sample population have a compressive strength less than or equal to 44. The second sample proportion estimates 0.21875, or 21.875% of cylinders of the sample population have a compressive strength greater than or equal to 47.

**pp. 29-31 #6.** Use R commands to obtain a simple random sample of size 50 from the statistical population of productivity ratings given in Example 1.6-3, and calculate the sample mean and sample variance. Repeat this for a total of five times, and report the five pairs of  $(\bar{x}, S^2)$ .

```
> POP = sample(c(rep(1,300),rep(2,700),rep(3,4000),rep(4,4000),rep(5,1000)))
```

```
> cN = data.frame("POP"=POP)
```

```
> s50 = sample(cN[["POP"]],50)
```

```
> m50 = mean(s50); v50 = var(s50);
```

```
> print(paste0("The sample has a mean of ",m50))
```

```
[1] "The sample has a mean of 3.52"
```

```
> print(paste0("The sample has a variance of ",v50))
```

```
[1] "The sample has a variance of 0.622040816326531"
```

**pp. 29-31 #9.** The outcome of a roll of a die is a random variable  $X$  that can be thought of as resulting from a simple random selection of one number from 1, ..., 6.

(a) Compute  $\mu_X$  and  $\sigma_X^2$ , either by hand or using R.

```
> muX = mean(X); sig2X <- var(X)*(N-1)/N
```

```
> print(paste0("The population has a mean of ",muX))
```

```
[1] "The population has a mean of 3.5"
```

```
> print(paste0("The population has a variance of ",sig2X))
```

```
[1] "The population has a variance of 2.91666666666667"
```

(b) Select a sample of size 100 with replacement from the finite population 1, ..., 6, and compute the sample mean and sample variance. The R commands for doing so are:

```
> mux = mean(x); sig2x <- var(x)
```

```
> print(paste0("The sample has a mean of ",mux))
```

```
[1] "The sample has a mean of 3.69"
```

```
> print(paste0("The sample has a variance of ",sig2x))
[1] "The sample has a variance of 3.10494949495"
```

Comment on how well the sample mean and variance approximates the true population parameters. Considering the sample size, the sample mean and variance do approximate the true population mean and variance. However, I think the sample mean and variance could better approximate the true population parameters by increasing the sample size to 1E4. This larger sample size will give a much better approximation of the true population mean and variance.

(c) Use the R command `table(x)/100`, where `x` is the sample of size  $n = 100$  obtained in part (b), to obtain the sample proportions of 1, . . . , 6. Are they, all reasonably close to  $1/6$ ?

```
> 1/6
[1] 0.1666667
> prop.table(table(x))
x
 1  2  3  4  5  6
0.16 0.15 0.12 0.19 0.17 0.21
```

Again, considering the sample size, the sample proportions are reasonably close to  $1/6$  but I would suggest increasing the sample size to get sample proportions closer to the population proportions of  $1/6$ .

**pp. 29-31#16.** The following data show the starting salaries, in \$1000 per year, for a sample of 15 senior engineers:

```
> engx = c(152, 169, 178, 179, 185, 188, 195, 196, 198, 203, 204, 209, 210, 212, 214)
```

# (a) Assuming that the 15 senior engineers represent a simple random sample from the population of senior engineers, estimate the population mean and variance.

```
> print(paste0("The sample has a mean of ",muEngx))
[1] "The sample has a mean of 192.8"
```

```
> print(paste0("The sample has a variance of ",sig2Engx))
[1] "The sample has a variance of 312.314285714286"
```

# (b) Give the sample mean and variance for the data on second-year salaries for the same group of engineers if  
# (i) if each engineer gets a \$5000 raise, and

```
> muEngx5000 = mean(engx5000); sig2Engx5000 <- var(engx5000)
```

```
> print(paste0("The sample has a mean of ",muEngx5000))
[1] "The sample has a mean of 197.8"
```

```
> print(paste0("The sample has a variance of ",sig2Engx5000))
[1] "The sample has a variance of 312.314285714286"
```

> # (ii) if each engineer gets a 5% raise.

```

> engx5p <- engx*1.05

> muEngx5p = mean(engx5p); sig2Engx5p <- var(engx5p)

> print(paste0("The sample has a mean of ",muEngx5p))
[1] "The sample has a mean of 202.44"

> print(paste0("The sample has a variance of ",sig2Engx5p))
[1] "The sample has a variance of 344.3265"

```

**p.35 #2.** Read the data on robot reaction times to simulated malfunctions into the data frame `t` by `t = read.table("RobotReactTime.txt", header = T)`. Read the reaction times of Robot 1 into the vector `t1` by `attach(t); t1 = Time[Robot==1]`, and sort the data (i.e., arrange it from smallest to largest) by `sort(t1)`. Using the sorted data and hand calculations

- (a) estimate the population median and the 25th and the 75th percentiles,
- (b) estimate the population interquartile range, and
- (c) find the percentile of the 19th ordered value.

```

> # t = read.table("RobotReactTime.txt", header = T)

> attach(t); t1 = Time[Robot==1]

> t1 <- sort(t1)

> # (a) estimate the population median and the 25th and the 75th percentiles,
> pimediant1 <- median(t1)

> qt1 = quantile(t1,c(0.25,0.75))

> print("For t1 sorted from smallest to largest:")
[1] "For t1 sorted from smallest to largest:"

> cat("The population median is ",pimediant1)
The population median is 30.55
> print(paste0("The 25th percentile is ",qt1[1]))
[1] "The 25th percentile is 29.6325"

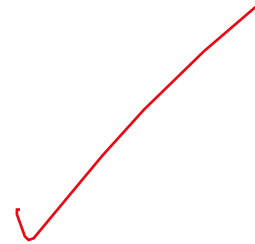
> print(paste0("The 75th percental is ",qt1[2]))
[1] "The 75th percental is 31.375"

> # (b) estimate the population interquartile range, and
> t1IQR <- qt1[2]-qt1[1]

> print(paste0("The population interquartile range is ",t1IQR))
[1] "The population interquartile range is 1.7425"

> # (c) find the percentile of the 19th ordered value.
> 19/length(t1)

```



```
[1] 0.8636364
```

```
> x = 0.75
```

```
> y = 0
```

```
> while (y<=t1[19]){  
+   y = quantile(t1,x); x = x + 0.000001  
+ }
```

```
> x = x - 0.000001
```

```
> sprintf("The 19th element of t1 is approximately the %.3f percentile",x*100)  
[1] "The 19th element of t1 is approximately the 85.714 percentile"
```

**p.35 #3.** The site given in Exercise 2 also gives the reaction times of Robot 2.  
# Use commands similar to those given in Exercise 2 to read the reaction times  
# of Robot 2 into the vector t2.

```
> t2 = t$Time[Robot==2]  
> t2 <- sort(t2)  
> # (a) Use the R command summary given in (1.7.2) to obtain the five number summary  
> # of this data set.  
> t2Summary <- summary(t2)  
> print(t2Summary)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.   Max.     
28.97  29.30  29.94  30.11  30.82  32.23
```

```
> # (b) Use the R command quantile given in (1.7.2) get the sample 90th percentile.
```

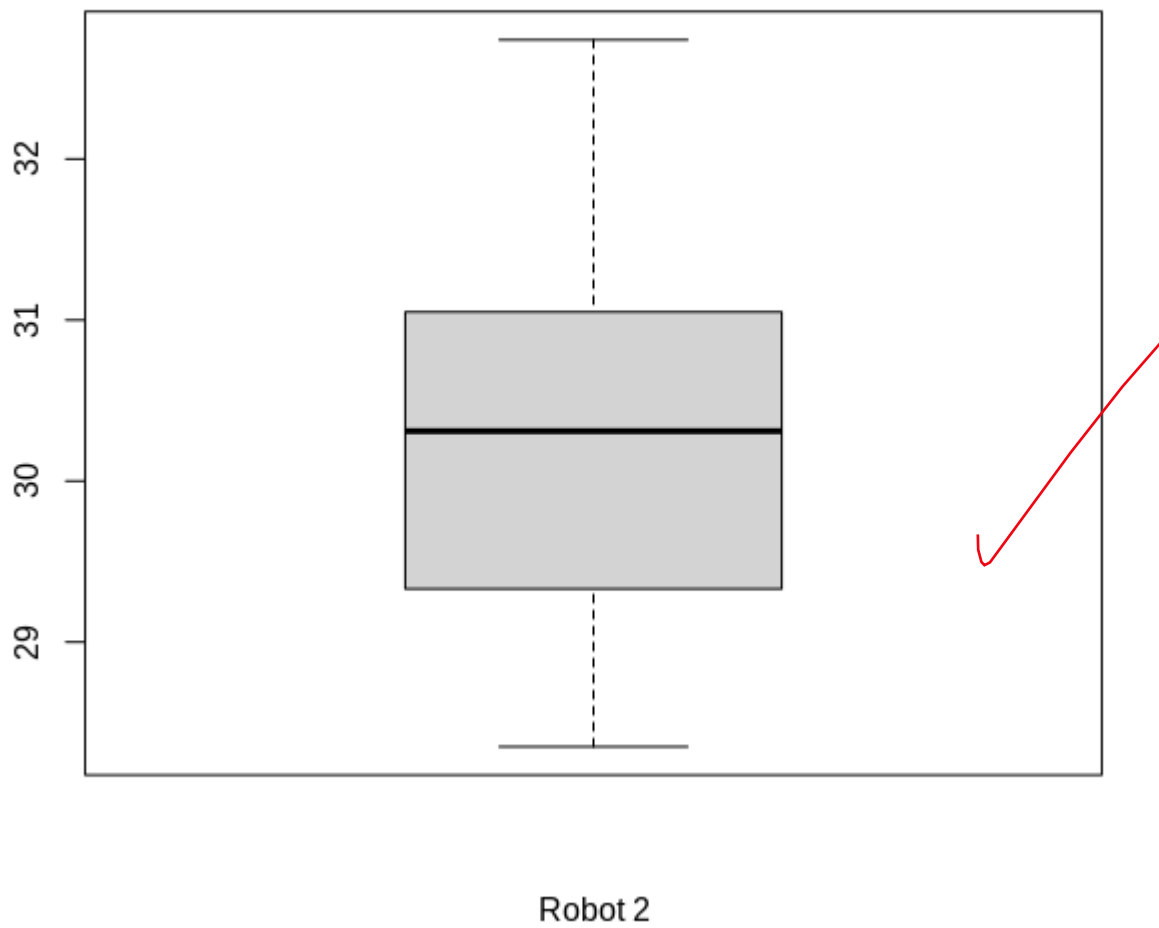
```
> t2q90 = quantile(t2,0.9)  
> print(paste0("The 90th percentile of t2 is ",t2q90))  
[1] "The 90th percentile of t2 is 31.068"
```

```
> # (c) Use the R command "boxplot" given in Example 1.7-3 to construct a boxplot  
> # for the data. Are there any outliers?
```

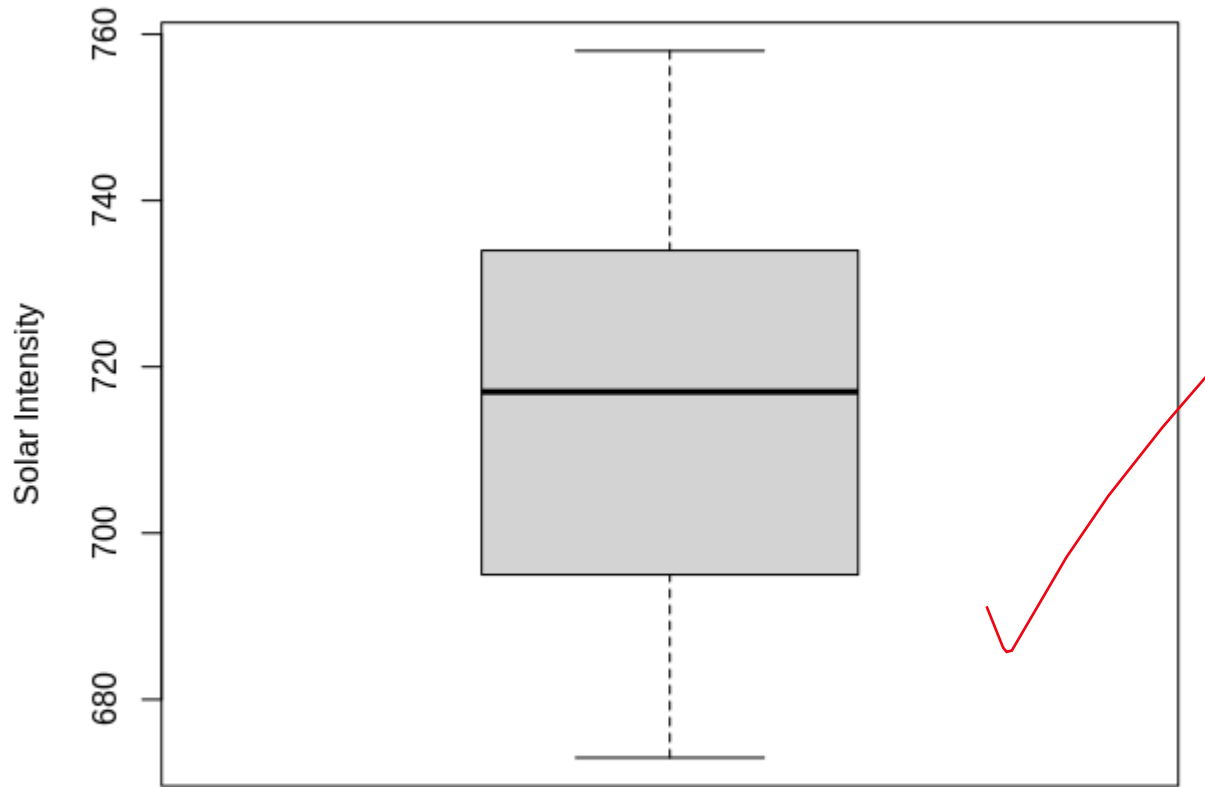
There are no outliers found, the boxplot found below shows none. I also checked using the analytical method and confirmed no outliers.

```
> IQRt2 = t2Summary[[5-2]]  
> q1 = t2Summary[2]  
> q3 = t2Summary[5]  
> t2cutoff = c(q1-IQRt2*1.5,q3+IQRt2*1.5)  
>  
> boxplot(height=Time[Robot==2],Time,xlab='Robot 2',ylab='Time')
```





**p.35 #4a.** Enter the solar intensity measurements of Exercise 1 into the R object si with si = read.table("SolarIntensAuData.txt", header = T). Use R commands to  
(a) construct a boxplot of the solar intensity measurements



Boxplot of Solar Intensity Data

```
> si = read.table("SolarIntensAuData.txt", header = T)
> dat = sort(si$SI)
> fig <- boxplot(dat,xlab='Boxplot of Solar Intensity Data',ylab='Solar Intensity')
> show(fig)
```

\$stats

[,1]

[1,] 673

[2,] 695

[3,] 717

[4,] 734

[5,] 758

attr(,"class")

"integer"

\$n

[1] 40

\$conf

[,1]

[1,] 707.257

[2,] 726.743

\$out

numeric(0)

\$group

numeric(0)

What about  
the 2 writer?  
problems/proofs?



\$names

[1] ""