Math 664: Hw 1

Feb 2022

Name: Yaksh Patel

ID: 31536823

Q1_____

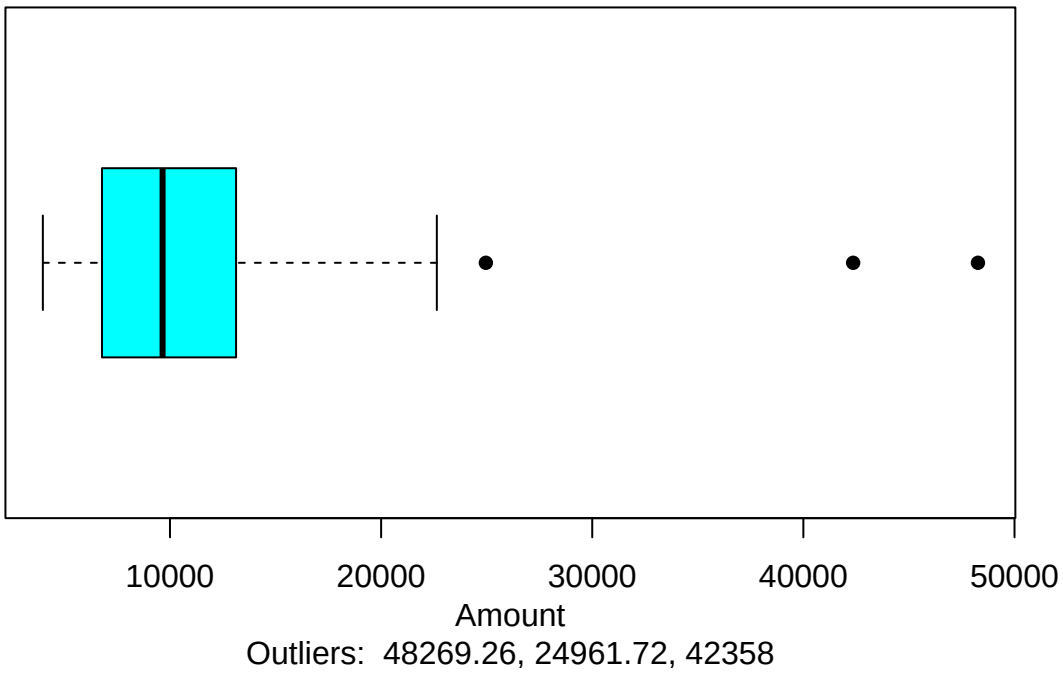
Importing the Data set

Checking and Removing NA values

Summary for quick glance

```
##       Week           Date               Amount
##  Min.   : 1.00   Length:51          Min.   : 3978
##  1st Qu.:14.50   Class :character   1st Qu.: 6779
##  Median :27.00   Mode  :character   Median : 9650
##  Mean   :26.84                      Mean   :11697
##  3rd Qu.:39.50                      3rd Qu.:13130
##  Max.   :52.00                      Max.   :48269
```

Box-plot



Stats of Box-plot

```
##                          Stats Stat_Values
## 1                Lower Whisker    3978.000
## 2                           Q1    6779.125
## 3                 Q2 (Median)    9650.140
## 4                           Q3   13130.270
## 5                Upper Whisker   22639.000
## 6 IQR (Inter Quartile Range)    6351.145
```

From the Summary, Box-plot (and its stats) of the cleaned data set, we infer that: -

There 51 valid records, 9th entry was Na.

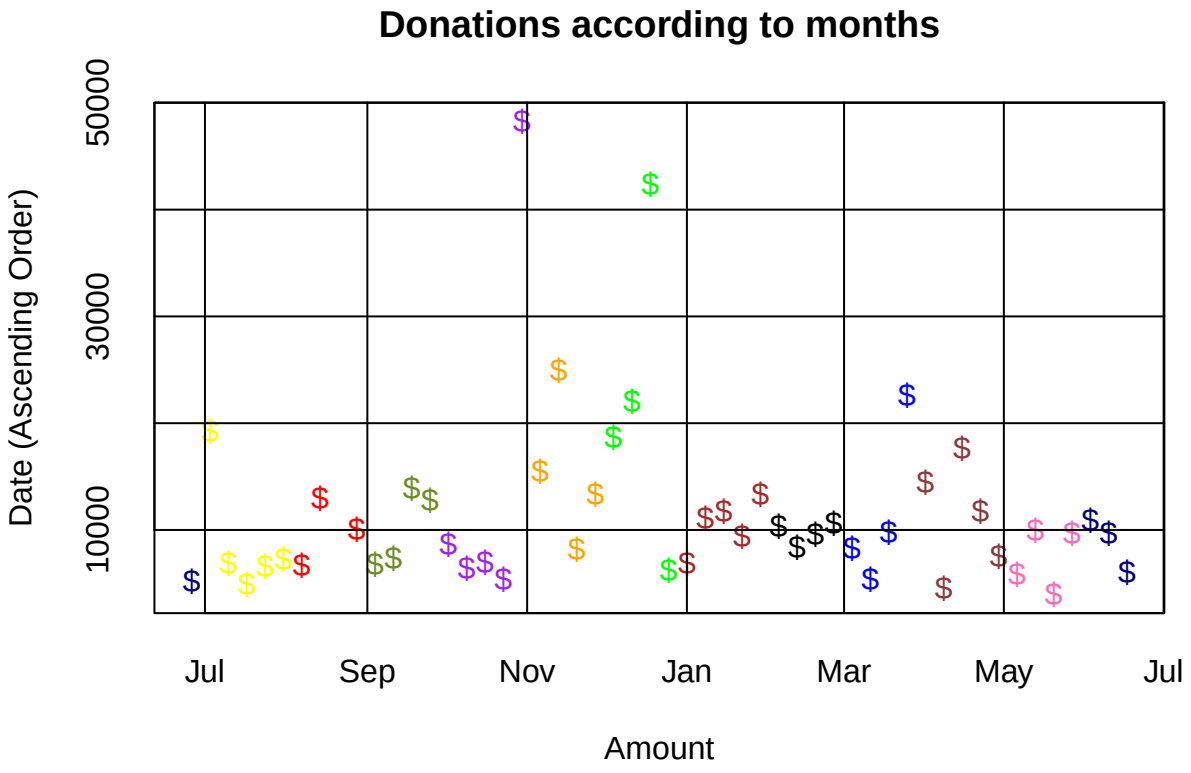
From the summary, we get the minimum and maximum amount.

We also get the information about quartiles and the mean.

Median < Mean for the data, thus, it is mildly-skewed right (positive skew).

There are 3 outliers present in the Amount of the data set which lie out of the Upper and Lower Fence.

Scatter-plot



The Data points are color-coded for each month so that it is easier to identify the values, frequencies, anomalies, etc. for the corresponding weeks of the month. It is in the chronological order starting from June 26, 2011 to June 17, 2012. Highest single donation of $48269 is in the 19th week, October 2012.

Checking donations per month and ordering them in decreasing order.

```
##        months      sum count Amount_Per_Donation
##  1:  December 89249.46     4           22312.365
##  2:   October 75747.04     5           15149.408
##  3:  November 61992.72     4           15498.180
##  4:     April 55838.75     5           11167.750
##  5:   January 52467.67     5           10493.534
##  6:     March 46048.00     4           11512.000
##  7:      July 44814.70     5            8962.940
##  8: September 40732.01     4           10183.003
##  9:  February 38857.56     4            9714.390
## 10:      June 31696.04     4            7924.010
## 11:    August 29655.29     3            9885.097
## 12:       May 29464.00     4            7366.000
```

This table demonstrates: -

Amount of total donation of the month. Min: May ($29464); Max: December ($89249).

Frequency of donations for that month. Min: August (3); Max: October, April, January, July (5).

Mean amount per donation in the month. Min: May ($7366); Max: December ($22312).

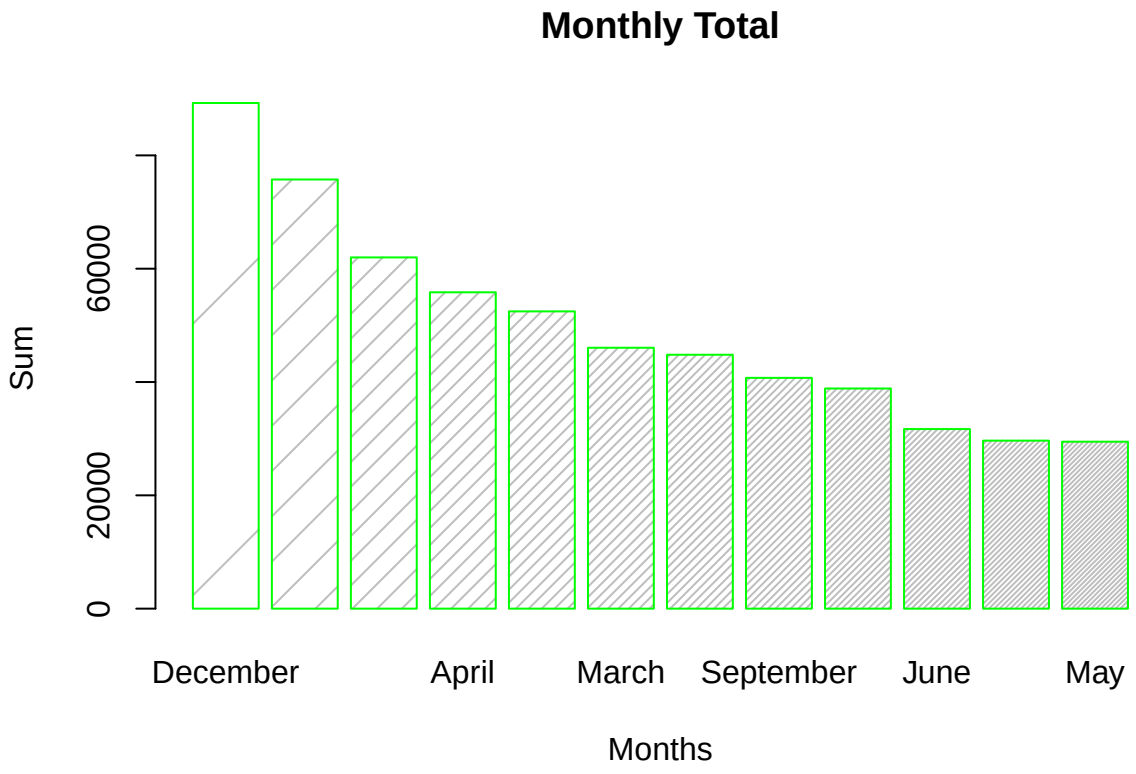
Data is sorted in the Descending order of the sum of the total donation Amount.

The month with the highest mean donation is also the month with the highest total donation.

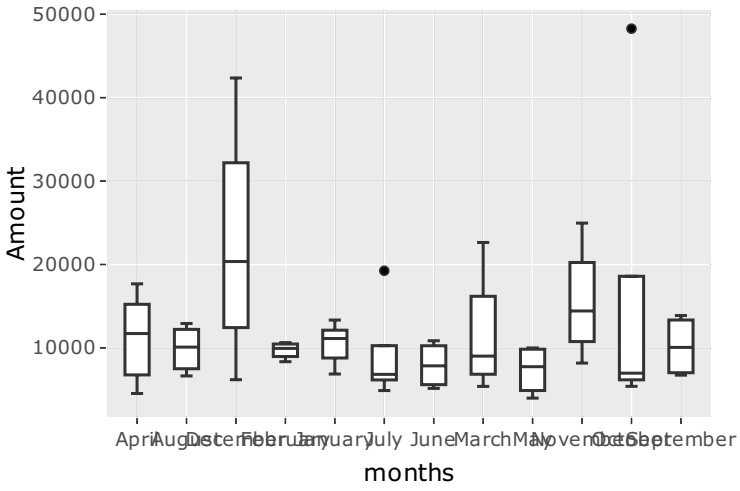
The month with the lowest mean donation is also the month with the lowest total donation.

Reason for December being highest could be the due the season of Christmas and New Year. On the other hand, October and November also have high amounts of donations, probably due to the festivals of other religions, say Hinduism in that month.

The Graph below is a visual representation of the donations made every month.



Comparing how th donations differ monthly



This graph compares the distribution of donation amounts across all the months.

It is evident from the table that all the outliers are the ones which are above the upperfence and none are below the lower fence.

The highest donations of July and October are outliers for them, making their box smaller, adding to that, their upper whisker and the 3rd Quartile coincide each other. Both of them are positive-skewed (right).

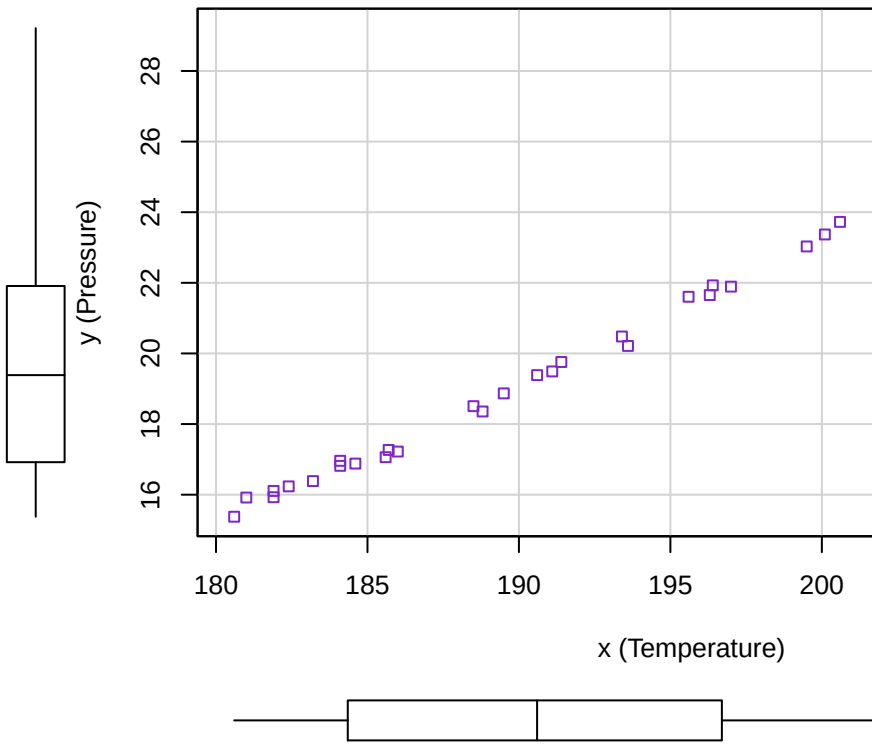
February's smallest box represents that the values don't vary highly for that month.

December's highest donation was an outlier when overall donations of the year were considered previously, however, here for itself, it seems to be in the allowed range.

Q2_____

Importing the Data and its summary

```
##       Temp          Pressure
##  Min.   :180.6   Min.   :15.38
##  1st Qu.:184.3   1st Qu.:16.92
##  Median :190.6   Median :19.39
##  Mean   :191.8   Mean   :20.03
##  3rd Qu.:196.7   3rd Qu.:21.91
##  Max.   :210.8   Max.   :29.21
```



Plotting the points for overall idea of the scattering

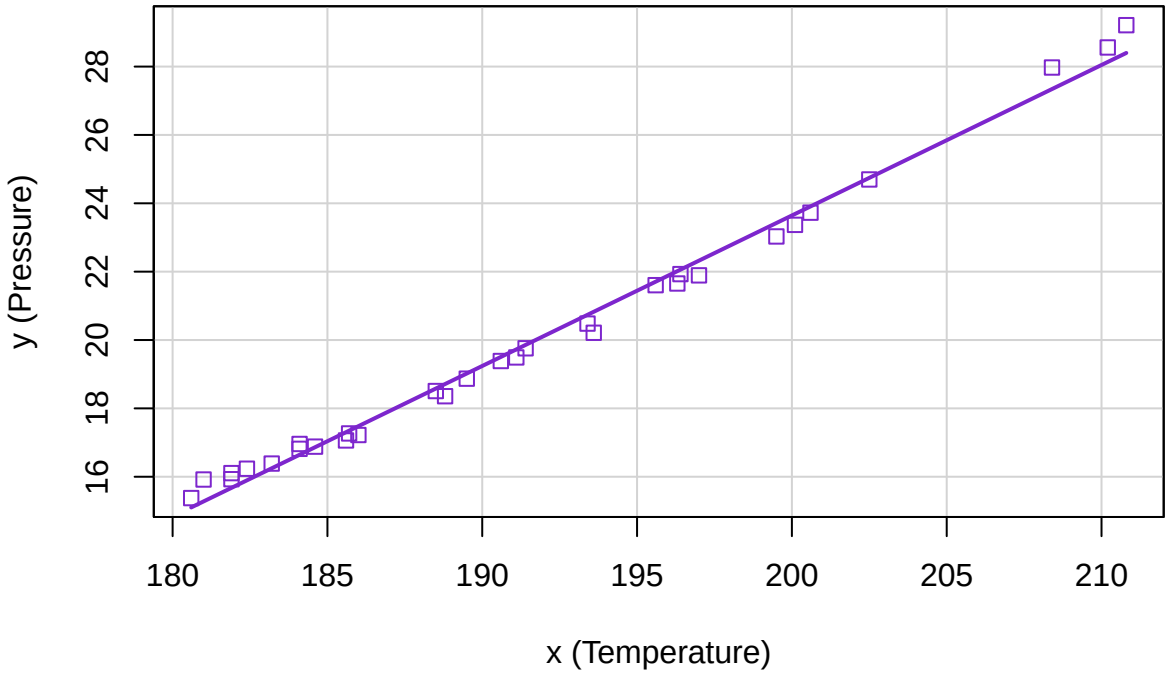
Initializing data structure for predicted values of y, generating regression function and its summary

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61383 -0.24968 -0.09921  0.26365  0.81232
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -64.412751   1.429165  -45.07   <2e-16 ***
## x             0.440282   0.007444   59.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3563 on 29 degrees of freedom
## Multiple R-squared:  0.9918, Adjusted R-squared:  0.9915
## F-statistic:  3498 on 1 and 29 DF,  p-value: < 2.2e-16
```

Analysis of Variance

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 444.17  444.17  3497.9 < 2.2e-16 ***
```

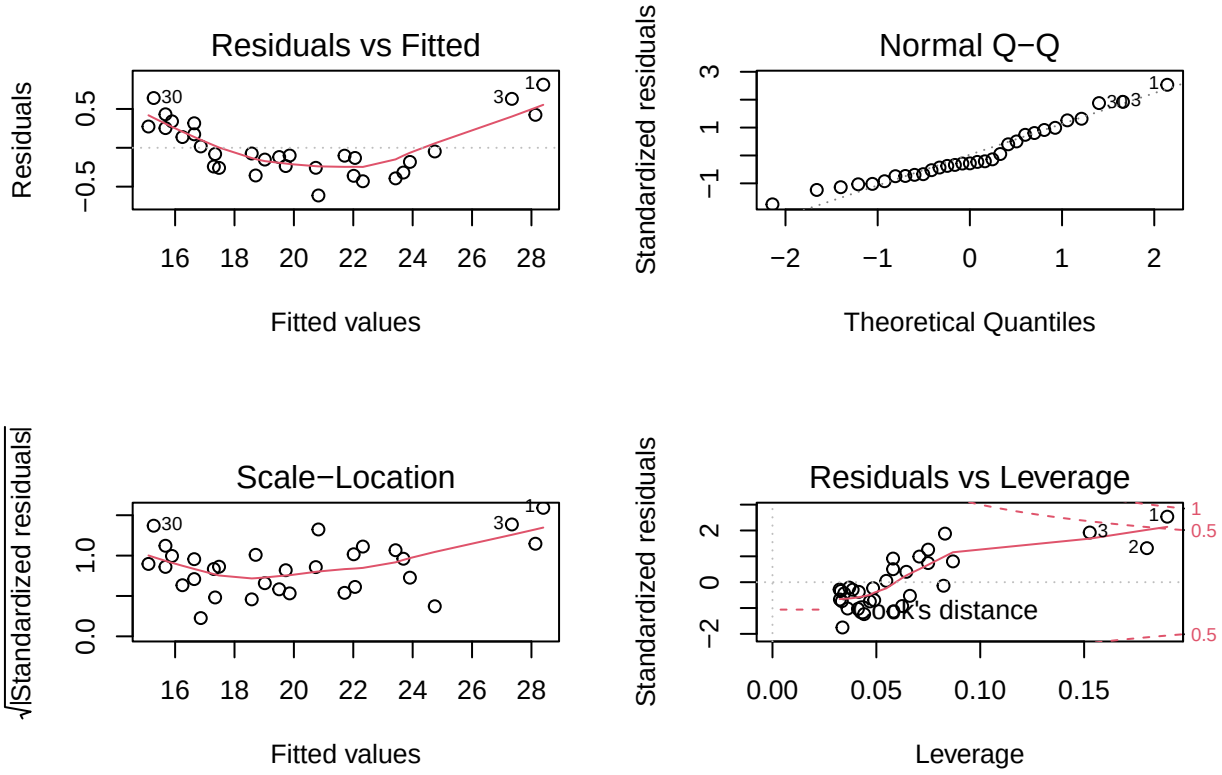
```
## Residuals 29   3.68    0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Plot of the line

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##    -64.4128       0.4403
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma::gvlma(x = reg, alphalevel = 0.05)
##
##                      Value   p-value                   Decision
## Global Stat        28.8599 8.347e-06 Assumptions NOT satisfied!
## Skewness            1.7889 1.811e-01    Assumptions acceptable.
## Kurtosis            0.2381 6.256e-01    Assumptions acceptable.
## Link Function      25.1342 5.348e-07 Assumptions NOT satisfied!
## Heteroscedasticity  1.6987 1.925e-01    Assumptions acceptable.
```

The points appear random and the line quite pretty flat, without increasing or decreasing trend. So, the condition of homoscedasticity can be accepted. Thus, Homoscedasticity (variance of the dependent variable being same across the data) assumption is satisfied. Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution.



Time-series analysis for independence of Residuals (trend in the data based on previous instances). It is to check whether residual are independent of one another.

```
##
##  Durbin-Watson test
##
## data:  reg
## DW = 0.62441, p-value = 1.067e-06
## alternative hypothesis: true autocorrelation is greater than 0
```

Confidence Interval (Point estimate, lower and upper boundary of Pressure for Temp = 185'F)

```
##        fit      lwr      upr
## 1 17.03941 16.87264 17.20618
```

Prediction Interval (Point estimate, lower and upper boundary of Pressure for Temp = 185'F)

```
##        fit      lwr      upr
## 1 17.03941 16.29177 17.78705
```

Since the beginning, the data of Pressure and Temperature seems quite linear from the plot of the points. We found out the data summary including mean, median and quartiles. Though the boxes look symmetrical but whiskers of the box-plot are unequal. Hence, we can say there is variability outside the range of 1st and 3rd quartile (Right skewed). The intercept is -64.4 and the slope is 0.44. The errors of residuals are magnificently low (almost 0 for the slope) with good values of F-statistics to satisfy the hypothesis condition. Multiple

R-squared and adjusted R-squared values are impressive (.992 and .991 respectively). Other conditions are fulfilled.

However, there might be a mild curvature in the relationship. As few of the high-pressure and high-temperature values along with low-pressure and low-temperature values, seem to be slightly above the line. And some of the points near the central region seem to be quite below te line. Therefore, I tried to go for another approach using log of Pressure against Temperature below. It gives more accurate and precise results. It helps in reducing the error value of Intercept significantly (by 97.65%). Same goes for the slope, although, its error was already negligible. The multiple and adjusted R-squared values are almost close to .998. Thus, the 2nd model is little more preferable.

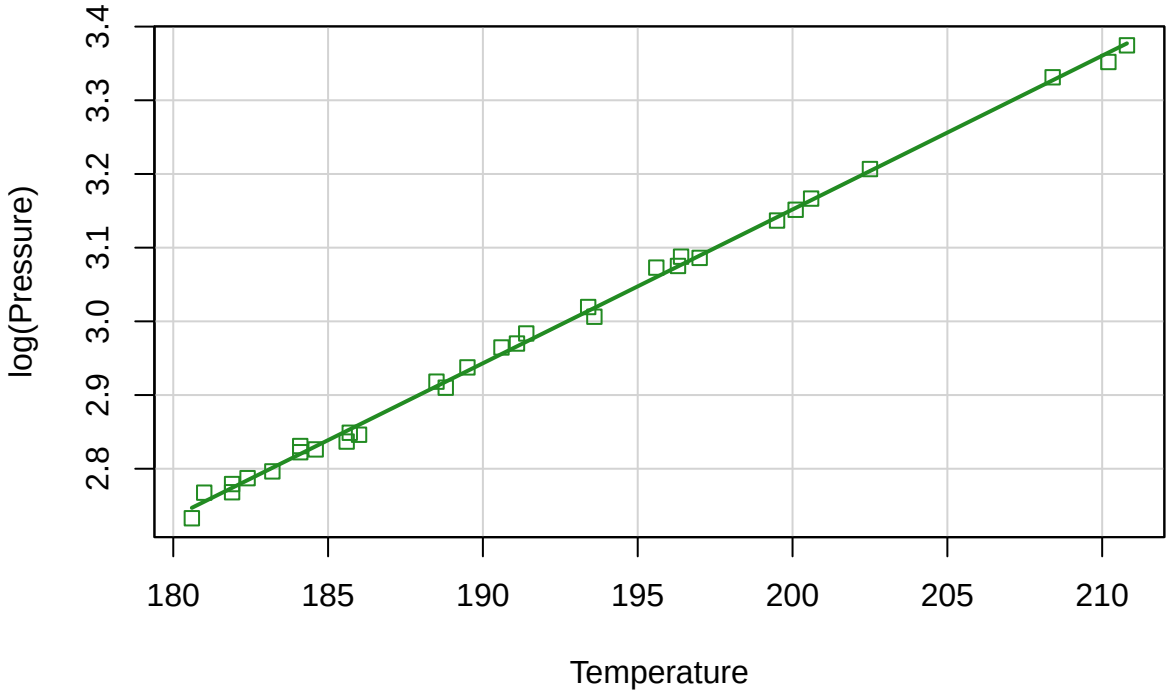
The relationship between the Altitude and Pressure or the Pressure and Temperature may change for different range of the values, i.e.: probably at slightly lower altitudes the relationships between the response and the predictors might not be linear, they might form curvature, etc.

2nd approach using log(Pressure) against Temperature

```
##
## Call:
## lm(formula = logP ~ temp)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.014437 -0.004710  0.002234  0.005247  0.012937
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.0221374  0.0336450  -30.38   <2e-16 ***
## temp         0.0208698  0.0001753  119.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.008389 on 29 degrees of freedom
## Multiple R-squared:  0.998,  Adjusted R-squared:  0.9979
## F-statistic: 1.418e+04 on 1 and 29 DF,  p-value: < 2.2e-16
```

Analysis of Variance

```
## Analysis of Variance Table
##
## Response: logP
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## temp       1 0.99798 0.99798   14181 < 2.2e-16 ***
## Residuals 29 0.00204 0.00007
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



New Confidence Interval for reg2 with logP at Temp=185'F

```
##        fit      lwr      upr
## 1 17.09471 17.02772 17.16195
```

New Prediction Interval for reg2 with logP at Temp=185'F

```
##        fit      lwr      upr
## 1 17.09471 16.79646 17.39825
```

References: -

https://github.com/YKP-The-GREAT/Death_Rate_Model_Comparison/blob/main/x28_Models.R

https://statisticsglobe.com

https://statology.com

https://stackoverflow.com

https://r4ds.had.co.nz/exploratory-data-analysis.html

https://sites.harding.edu/fmccown/r/

https://www.rdocumentation.org

http://www.sthda.com/

https://www.datacamp.com/

https://www.datasciencemadesimple.com/

https://corporatefinanceinstitute.com/resources/knowledge/other/kurtosis/

https://www.learnbymarketing.com/tutorials/linear-regression-in-r/