2023-01-23

# # # Please note every technique applied are basic library of R and no additional package or library is required

# Statistical Techniques and Application

Our first dataset is Lake Information. Importing the file to R

lake = read.csv(file.choose())

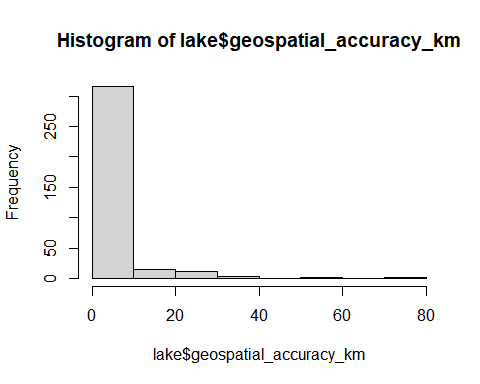
Taking a quick look at the summary of our dataset

summary(lake)

## siteID Lake\_name Other\_names lake\_or\_reservoir   
## Min. : 1.00 Length:348 Length:348 Length:348   
## 1st Qu.: 87.75 Class :character Class :character Class :character   
## Median :174.50 Mode :character Mode :character Mode :character   
## Mean :174.55   
## 3rd Qu.:261.25   
## Max. :349.00   
##   
## location region latitude longitude   
## Length:348 Length:348 Min. :-50.23 Min. :-160.520   
## Class :character Class :character 1st Qu.: 27.00 1st Qu.: -81.382   
## Mode :character Mode :character Median : 43.06 Median : 11.895   
## Mean : 31.43 Mean : -3.304   
## 3rd Qu.: 49.86 3rd Qu.: 49.830   
## Max. : 74.40 Max. : 176.420   
## NA's :1   
## geospatial\_accuracy\_km elevation\_m mean\_depth\_m max\_depth\_m   
## Min. : 0.130 Min. :-404.0 Min. : 0.50 Min. : 1.0   
## 1st Qu.: 2.000 1st Qu.: 61.5 1st Qu.: 4.53 1st Qu.: 10.5   
## Median : 2.000 Median : 202.9 Median : 10.00 Median : 27.0   
## Mean : 4.901 Mean : 446.8 Mean : 43.35 Mean : 109.6   
## 3rd Qu.: 3.625 3rd Qu.: 417.5 3rd Qu.: 27.10 3rd Qu.: 92.0   
## Max. :75.000 Max. :4743.0 Max. :730.00 Max. :1637.0   
## NA's :21 NA's :16   
## surface\_area\_km2 volume\_km3 source sampling\_depth   
## Min. : 0.0 Min. : 0.00 Length:348 Length:348   
## 1st Qu.: 14.2 1st Qu.: 0.09 Class :character Class :character   
## Median : 600.5 Median : 3.27 Mode :character Mode :character   
## Mean : 5774.1 Mean : 933.50   
## 3rd Qu.: 1965.0 3rd Qu.: 35.80   
## Max. :378119.3 Max. :78200.00   
## NA's :22   
## sampling\_time\_of\_day time\_period contributor   
## Length:348 Length:348 Length:348   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##

Performing EDA on our dataset to give us a preview

hist(lake$geospatial\_accuracy\_km)

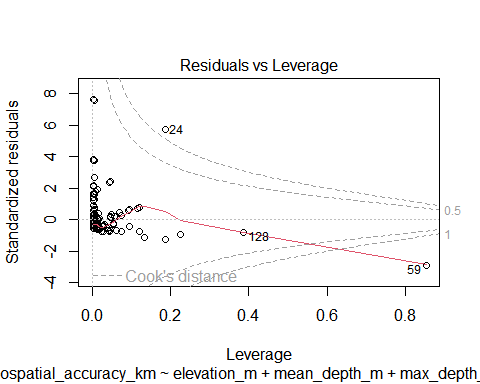
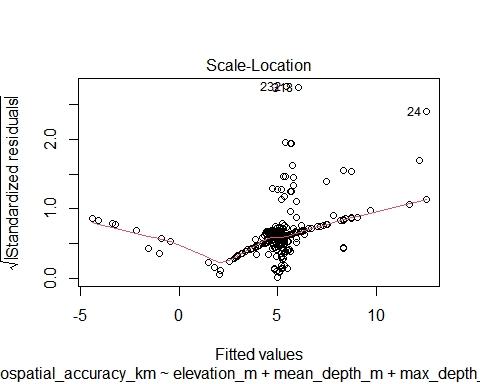
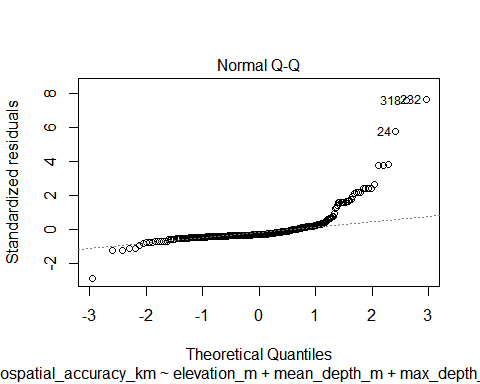
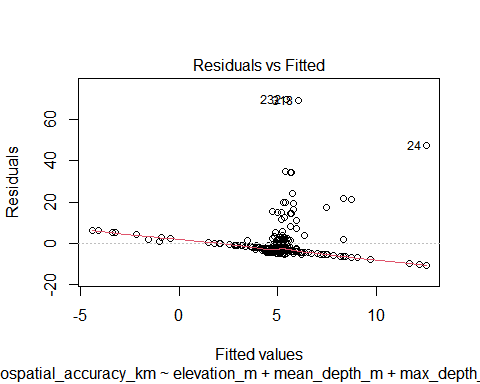
 The above diagram shows 10% of our sample collected have high geospatial accuracy, we intend to investigate the causes of factors affecting the result. Therefore our dependent variable is geospatial\_accuracy\_km and other variable are identified as independent variable

Fitting a linear model to our dataset using geospatial\_accuracy\_km as our dependent variable

Lake\_model = lm(geospatial\_accuracy\_km ~ elevation\_m+mean\_depth\_m+max\_depth\_m+surface\_area\_km2+volume\_km3, data = lake)

Checking if our dataset is normally distributed

plot(Lake\_model)



comment: our data is not normally distributed

Showing summary of our linear model

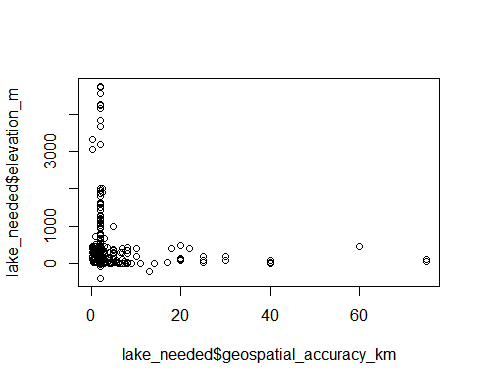
summary(Lake\_model)

##   
## Call:  
## lm(formula = geospatial\_accuracy\_km ~ elevation\_m + mean\_depth\_m +   
## max\_depth\_m + surface\_area\_km2 + volume\_km3, data = lake)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.497 -3.786 -2.881 0.061 69.576   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.184e+00 6.790e-01 7.635 2.78e-13 \*\*\*  
## elevation\_m -2.038e-03 7.155e-04 -2.849 0.00468 \*\*   
## mean\_depth\_m 2.121e-03 1.782e-02 0.119 0.90534   
## max\_depth\_m 6.247e-03 8.499e-03 0.735 0.46286   
## surface\_area\_km2 4.525e-05 5.220e-05 0.867 0.38673   
## volume\_km3 -2.169e-04 2.884e-04 -0.752 0.45261   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.153 on 312 degrees of freedom  
## (30 observations deleted due to missingness)  
## Multiple R-squared: 0.04572, Adjusted R-squared: 0.03043   
## F-statistic: 2.99 on 5 and 312 DF, p-value: 0.01186

At 0.05 level of significance, the above result shows that elevation has significant effect on our dependent variable (geospatial\_accuracy) having a low p-value while other variables does not really show significant effect. In order to back up our result, we therefore have to investigate cities with high elevation to check for their geospatial accuracy if it corresponds with our result

# # Filtering out our needed variable (region, deviation and geospatial accuracy)

lake\_needed = lake[c(6,10,9)]  
plot(lake\_needed$geospatial\_accuracy\_km, lake\_needed$elevation\_m)



comment:Based on our dataset and hypothesis testing, we can assume that countries with high elevation where the lake or reservoir presides will probably have significant ecosystem effect if there ought to be a defiling agent to their reservoir and lakes .

# SECOND DATA SET

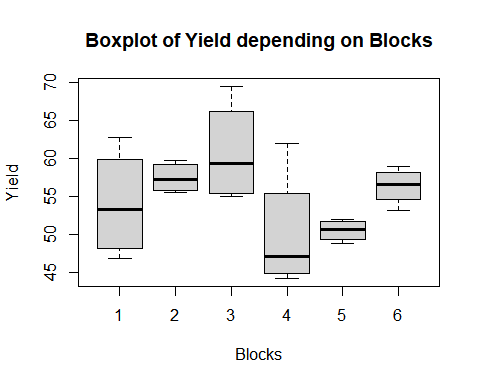
# N,P,K FACTORIAL EXPERIMENT DATASET

Our dataset is an inbuilt R dataset

data(npk)

Performing EDA on our dataset

boxplot(yield ~ block, data = npk, xlab = 'Blocks', ylab = 'Yield',   
 main = 'Boxplot of Yield depending on Blocks')



comment:The boxplot diagram shows that block 3 have the highest yield with the minimun treatment used while block 4 have the lowest yield with maximun treatment while other blocks have similar yield and treatments

Fitting our dataset into ANOVA model

model <- aov(yield ~N\*P\*K, data = npk)   
anova(model)

## Analysis of Variance Table  
##   
## Response: yield  
## Df Sum Sq Mean Sq F value Pr(>F)   
## N 1 189.28 189.282 6.1608 0.02454 \*  
## P 1 8.40 8.402 0.2735 0.60819   
## K 1 95.20 95.202 3.0986 0.09746 .  
## N:P 1 21.28 21.282 0.6927 0.41750   
## N:K 1 33.14 33.135 1.0785 0.31448   
## P:K 1 0.48 0.482 0.0157 0.90192   
## N:P:K 1 37.00 37.002 1.2043 0.28870   
## Residuals 16 491.58 30.724   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

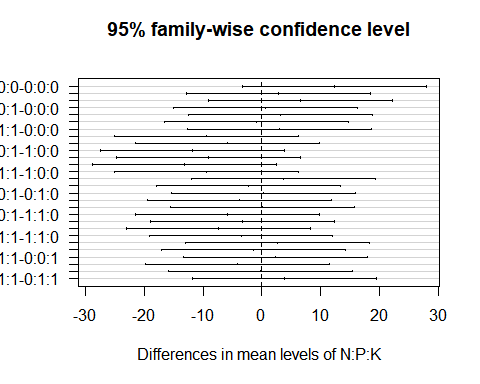
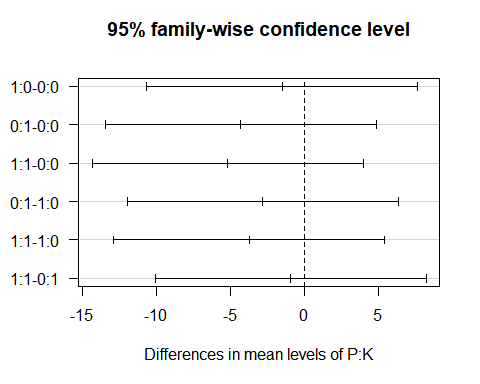
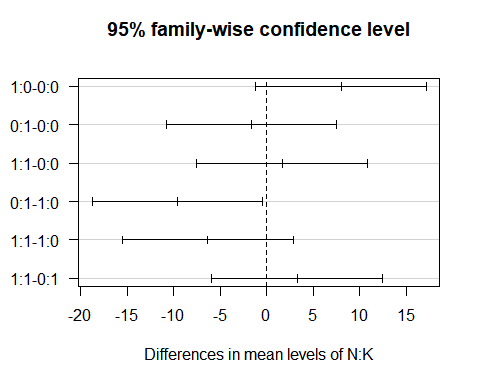
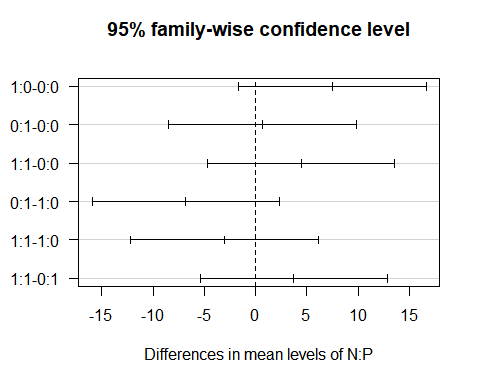
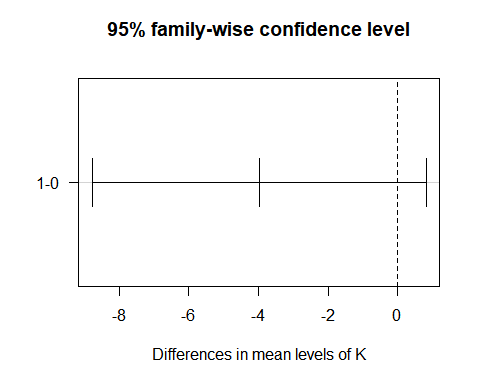
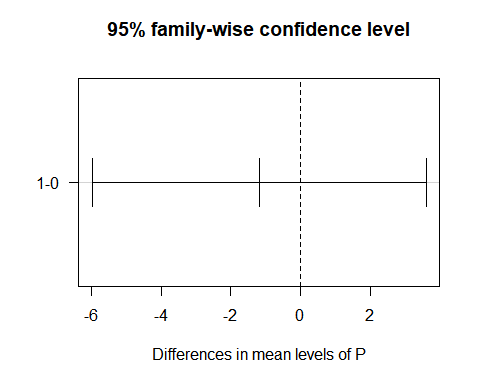
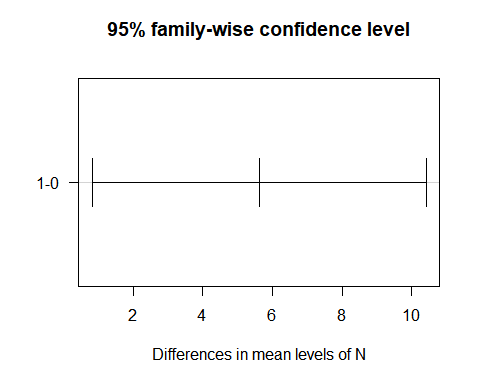
Performing tukey test on our model

tukey\_test <- TukeyHSD(model)  
tukey\_test

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = yield ~ N \* P \* K, data = npk)  
##   
## $N  
## diff lwr upr p adj  
## 1-0 5.616667 0.8195759 10.41376 0.0245421  
##   
## $P  
## diff lwr upr p adj  
## 1-0 -1.183333 -5.980424 3.613757 0.6081875  
##   
## $K  
## diff lwr upr p adj  
## 1-0 -3.983333 -8.780424 0.8137574 0.0974577  
##   
## $`N:P`  
## diff lwr upr p adj  
## 1:0-0:0 7.500000 -1.655822 16.655822 0.1294203  
## 0:1-0:0 0.700000 -8.455822 9.855822 0.9961506  
## 1:1-0:0 4.433333 -4.722489 13.589156 0.5257140  
## 0:1-1:0 -6.800000 -15.955822 2.355822 0.1873570  
## 1:1-1:0 -3.066667 -12.222489 6.089156 0.7743737  
## 1:1-0:1 3.733333 -5.422489 12.889156 0.6554324  
##   
## $`N:K`  
## diff lwr upr p adj  
## 1:0-0:0 7.966667 -1.189156 17.1224889 0.0999349  
## 0:1-0:0 -1.633333 -10.789156 7.5224889 0.9554188  
## 1:1-0:0 1.633333 -7.522489 10.7891555 0.9554188  
## 0:1-1:0 -9.600000 -18.755822 -0.4441778 0.0382331  
## 1:1-1:0 -6.333333 -15.489156 2.8224889 0.2364506  
## 1:1-0:1 3.266667 -5.889156 12.4224889 0.7399770  
##   
## $`P:K`  
## diff lwr upr p adj  
## 1:0-0:0 -1.466667 -10.62249 7.689156 0.9670135  
## 0:1-0:0 -4.266667 -13.42249 4.889156 0.5562955  
## 1:1-0:0 -5.166667 -14.32249 3.989156 0.3986648  
## 0:1-1:0 -2.800000 -11.95582 6.355822 0.8176375  
## 1:1-1:0 -3.700000 -12.85582 5.455822 0.6615998  
## 1:1-0:1 -0.900000 -10.05582 8.255822 0.9919305  
##   
## $`N:P:K`  
## diff lwr upr p adj  
## 1:0:0-0:0:0 12.33333333 -3.335528 28.002195 0.1841773  
## 0:1:0-0:0:0 2.90000000 -12.768862 18.568862 0.9975401  
## 1:1:0-0:0:0 6.50000000 -9.168862 22.168862 0.8280891  
## 0:0:1-0:0:0 0.56666667 -15.102195 16.235528 1.0000000  
## 1:0:1-0:0:0 3.23333333 -12.435528 18.902195 0.9952141  
## 0:1:1-0:0:0 -0.93333333 -16.602195 14.735528 0.9999987  
## 1:1:1-0:0:0 2.93333333 -12.735528 18.602195 0.9973597  
## 0:1:0-1:0:0 -9.43333333 -25.102195 6.235528 0.4627120  
## 1:1:0-1:0:0 -5.83333333 -21.502195 9.835528 0.8902999  
## 0:0:1-1:0:0 -11.76666667 -27.435528 3.902195 0.2249895  
## 1:0:1-1:0:0 -9.10000000 -24.768862 6.568862 0.5045183  
## 0:1:1-1:0:0 -13.26666667 -28.935528 2.402195 0.1303264  
## 1:1:1-1:0:0 -9.40000000 -25.068862 6.268862 0.4668300  
## 1:1:0-0:1:0 3.60000000 -12.068862 19.268862 0.9909550  
## 0:0:1-0:1:0 -2.33333333 -18.002195 13.335528 0.9993808  
## 1:0:1-0:1:0 0.33333333 -15.335528 16.002195 1.0000000  
## 0:1:1-0:1:0 -3.83333333 -19.502195 11.835528 0.9870208  
## 1:1:1-0:1:0 0.03333333 -15.635528 15.702195 1.0000000  
## 0:0:1-1:1:0 -5.93333333 -21.602195 9.735528 0.8819049  
## 1:0:1-1:1:0 -3.26666667 -18.935528 12.402195 0.9949094  
## 0:1:1-1:1:0 -7.43333333 -23.102195 8.235528 0.7204567  
## 1:1:1-1:1:0 -3.56666667 -19.235528 12.102195 0.9914323  
## 1:0:1-0:0:1 2.66666667 -13.002195 18.335528 0.9985452  
## 0:1:1-0:0:1 -1.50000000 -17.168862 14.168862 0.9999671  
## 1:1:1-0:0:1 2.36666667 -13.302195 18.035528 0.9993214  
## 0:1:1-1:0:1 -4.16666667 -19.835528 11.502195 0.9793323  
## 1:1:1-1:0:1 -0.30000000 -15.968862 15.368862 1.0000000  
## 1:1:1-0:1:1 3.86666667 -11.802195 19.535528 0.9863677

Visualizing tukey test

plot(tukey\_test, las = 1)



# Third dataset

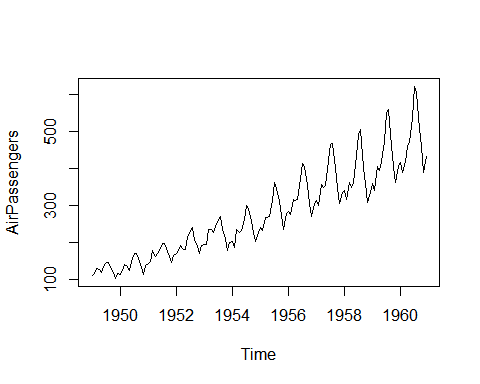
# MONTHLY AIRLINE PASSENGERS

Our dataset is an inbuilt R dataset

data("AirPassengers")

Performing EDA on our dataset

plot(AirPassengers)



comment:After plotting our dataset. It appears our dataset contains trends and the trend to be identical. This tells us our data is not stationary which shows Air passangers increases as time goes on , we check for Ljung Test to see if the test statistics (X¬-squared) gets larger with time. We intend to check residual are independently distributed

Performing Ljung Box test on our data

model = ARMAacf(AirPassengers)  
Box.test(model, lag = 12, type = 'Ljung-Box')

##   
## Box-Ljung test  
##   
## data: model  
## X-squared = 0.0043445, df = 12, p-value = 1