Question 1

**(1)Describe what constitutes the populations and samples in the case study, and give a critical commentary on the sampling technique used by Fairfield officials.**

The population is 943 water reservoirs

The sample size is 122 reservoirs

**Critical comments:**

The amount of fish sampled is too small to be statistically significant!

There is also a issue with Mercury levels being averaged from a small amount of fish being analysed per water source.

From a Correlation standpoint, there is no relationship between any of the variables in this dataset.

Most of the individual data variables are not normal Distributions.

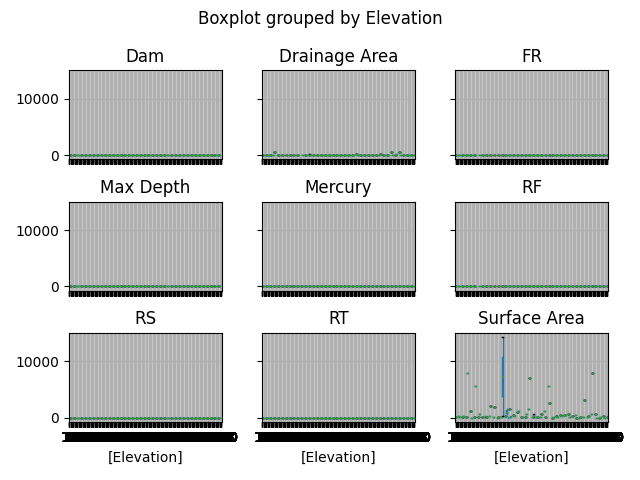
Question 2

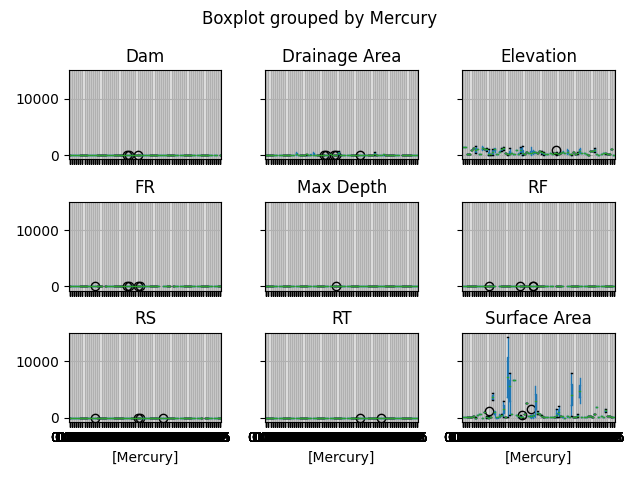
Present and analyse the distribution of each variable in the dataset.

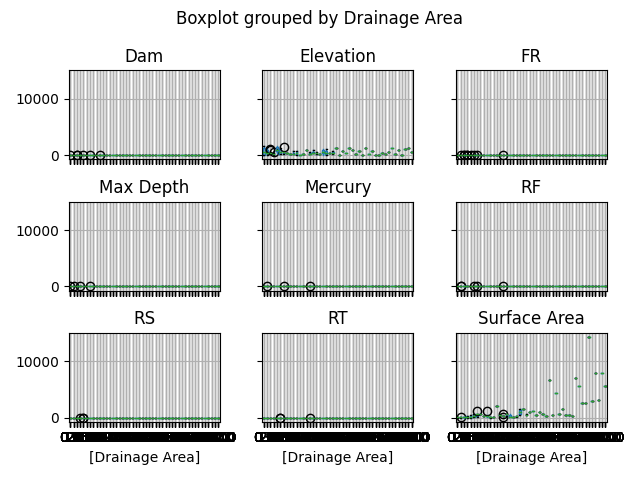
|  |  |
| --- | --- |
| ***Mercury***  Describe the distribution  count 122.000000  mean 0.492287  std 0.353569  min 0.025000  25% 0.250000  50% 0.410000  75% 0.667500  max 2.500000  Name: Mercury, dtype: float64  Central Tendancy CALC:Measure the Mean  0.49228688524590164  Central Tendancy CALC:Measure the Median  0.41  Central Tendancy CALC:Measure the Mode  0 0.41  dtype: float64  SPREAD CALC: Quartiles  [0.025 0.25 0.41 0.6675 2.5 ]  SPREAD CALC: Quantiles  [0.025 0.25 0.41 0.6675 2.5 ]  SPREAD CALC: Deciles  [0.025 0.142 0.22 0.283 0.364 0.41 0.48 0.577 0.746 0.899 2.5 ]  SPREAD CALC: Interquartile Range  0.4175000000000001  Distribution is Skewed Positively (Skewed Right) | |
| |  |  | | --- | --- | |  |  | |
| ***Elevation***  count 122.000000  mean 569.049180  std 431.745292  min 15.000000  25% 251.000000  50% 416.500000  75% 840.500000  max 1700.000000  Name: Elevation, dtype: float64  Central Tendancy CALC:Measure the Mean  569.0491803278688  Central Tendancy CALC:Measure the Median  416.5  Central Tendancy CALC:Measure the Mode  0 204  1 205  2 298  3 311  4 312  5 328  6 446  7 683  dtype: int64  SPREAD CALC: Quartiles  [ 15. 251. 416.5 840.5 1700. ]  SPREAD CALC: Quantiles  [ 15. 251. 416.5 840.5 1700. ]  SPREAD CALC: Deciles  [ 15. 123.9 205. 296.6 335. 416.5 515.4 706.8 927.6 1243.1  1700. ]  SPREAD CALC: Interquartile Range  589.5  Distribution is Skewed Positively (Skewed Right) |
| |  |  | | --- | --- | |  |  | |
| **Drainage Area**  count 119.000000  mean 41.747899  std 116.082176  min 0.000000  25% 2.000000  50% 6.000000  75% 20.000000  max 762.000000  Name: Drainage Area, dtype: float64  Central Tendancy CALC:Measure the Mean  41.747899159663866  Central Tendancy CALC:Measure the Median  6.0  Central Tendancy CALC:Measure the Mode  0 1.0  dtype: float64  SPREAD CALC: Quartiles  [nan nan nan nan nan]  SPREAD CALC: Quantiles  [nan nan nan nan nan]  SPREAD CALC: Deciles  [nan nan nan nan nan nan nan nan nan nan nan]  SPREAD CALC: Interquartile Range  Nan  Distribution is Exponentially skewed |
| |  |  | | --- | --- | |  |  | |
| ***Surface Area***  count 122.000000  mean 888.778689  std 1971.466906  min 12.000000  25% 76.250000  50% 230.500000  75% 623.750000  max 14340.000000  Name: Surface Area, dtype: float64  Central Tendancy CALC:Measure the Mean  888.7786885245902  Central Tendancy CALC:Measure the Median  230.5  Central Tendancy CALC:Measure the Mode  0 24  1 30  2 41  3 44  4 45  5 47  6 102  7 164  dtype: int64  SPREAD CALC: Quartiles  [1.2000e+01 7.6250e+01 2.3050e+02 6.2375e+02 1.4340e+04]  SPREAD CALC: Quantiles  [1.2000e+01 7.6250e+01 2.3050e+02 6.2375e+02 1.4340e+04]  SPREAD CALC: Deciles  [1.2000e+01 3.5100e+01 4.9800e+01 1.0410e+02 1.7780e+02 2.3050e+02  3.3580e+02 5.1460e+02 7.1040e+02 2.0138e+03 1.4340e+04]  SPREAD CALC: Interquartile Range  547.5  Distribution is Exponentially skewed |
| |  |  | | --- | --- | |  |  | |
| **Max Depth**  count 121.000000  mean 41.165289  std 29.733636  min 5.000000  25% 21.000000  50% 36.000000  75% 50.000000  max 158.000000  Name: Max Depth, dtype: float64  Central Tendancy CALC:Measure the Mean  41.16528925619835  Central Tendancy CALC:Measure the Median  36.0  Central Tendancy CALC:Measure the Mode  0 38.0  dtype: float64  SPREAD CALC: Quartiles  [nan nan nan nan nan]  SPREAD CALC: Quantiles  [nan nan nan nan nan]  SPREAD CALC: Deciles  [nan nan nan nan nan nan nan nan nan nan nan]  SPREAD CALC: Interquartile Range  Nan  Distribution is skewed right or positively skewed. |
| |  |  | | --- | --- | |  |  | |
| **RF**  count 115.000000  mean 0.550348  std 0.099789  min 0.060000  25% 0.510000  50% 0.560000  75% 0.610000  max 0.760000  Name: RF, dtype: float64  Central Tendancy CALC: Measure the Mean  0.5503478260869564  Central Tendancy CALC: Measure the Median  0.56  Central Tendancy CALC: Measure the Mode  0 0.51  1 0.56  dtype: float64  SPREAD CALC: Quartiles  [nan nan nan nan nan]  SPREAD CALC: Quantiles  [nan nan nan nan nan]  SPREAD CALC: Deciles  [nan nan nan nan nan nan nan nan nan nan nan]  SPREAD CALC: Interquartile Range  Nan  Distribution is negatively skewed |
| |  |  | | --- | --- | |  |  | |
| **FR**  count 115.000000  mean 6.366087  std 11.532448  min 0.100000  25% 0.900000  50% 2.100000  75% 5.800000  max 64.100000  Name: FR, dtype: float64  Central Tendancy CALC:Measure the Mean  6.366086956521738  Central Tendancy CALC:Measure the Median  2.1  Central Tendancy CALC:Measure the Mode  0 1.1  dtype: float64  SPREAD CALC: Quartiles  [nan nan nan nan nan]  SPREAD CALC: Quantiles  [nan nan nan nan nan]  SPREAD CALC: Deciles  [nan nan nan nan nan nan nan nan nan nan nan]  SPREAD CALC: Interquartile Range  Nan  Distribution is Exponentially skewed |
| |  |  | | --- | --- | |  |  | |
| **Dam**  count 119.000000  mean 0.605042  std 0.490909  min 0.000000  25% 0.000000  50% 1.000000  75% 1.000000  max 1.000000  Name: Dam, dtype: float64  Central Tendancy CALC:Measure the Mean  0.6050420168067226  Central Tendancy CALC:Measure the Median  1.0  Central Tendancy CALC:Measure the Mode  0 1.0  dtype: float64  SPREAD CALC: Quartiles  [nan nan nan nan nan]  SPREAD CALC: Quantiles  [nan nan nan nan nan]  SPREAD CALC: Deciles  [nan nan nan nan nan nan nan nan nan nan nan]  SPREAD CALC: Interquartile Range  Nan  Distribution is Bimodel skewed |
| |  |  | | --- | --- | |  |  | |
| **RT**  count 121.000000  mean 2.190083  std 0.722427  min 1.000000  25% 2.000000  50% 2.000000  75% 3.000000  max 3.000000  Name: RT, dtype: float64  Central Tendancy CALC:Measure the Mean  2.190082644628099  Central Tendancy CALC:Measure the Median  2.0  Central Tendancy CALC:Measure the Mode  0 2.0  dtype: float64  SPREAD CALC: Quartiles  [nan nan nan nan nan]  SPREAD CALC: Quantiles  [nan nan nan nan nan]  SPREAD CALC: Deciles  [nan nan nan nan nan nan nan nan nan nan nan]  SPREAD CALC: Interquartile Range  Nan  This distribution is Bi modal as it has more than 1 peak |
| |  |  | | --- | --- | |  |  | |
| **RS**  count 120.000000  mean 0.558333  std 0.498668  min 0.000000  25% 0.000000  50% 1.000000  75% 1.000000  max 1.000000  Name: RS, dtype: float64  Central Tendancy CALC:Measure the Mean  0.5583333333333333  Central Tendancy CALC:Measure the Median  1.0  Central Tendancy CALC:Measure the Mode  0 1.0  dtype: float64  SPREAD CALC: Quartiles  [nan nan nan nan nan]  SPREAD CALC: Quantiles  [nan nan nan nan nan]  SPREAD CALC: Deciles  [nan nan nan nan nan nan nan nan nan nan nan]  SPREAD CALC: Interquartile Range  Nan  This distribution is Bi modal |
| |  |  | | --- | --- | |  |  | |
|  |
|  |
| Python Code used  **import** pandas **as** pd **import** numpy **as** np **import** matplotlib.pyplot **as** plt **from** statsmodels.graphics.gofplots **import** qqplot  df = pd.read\_csv(**r'C:\temp\deadfishies2.csv'**) df.head()  *# boxplot = df.boxplot(fontsize=8, column=['Mercury', 'Elevation', 'Drainage Area', 'Surface Area','Max Depth','RF','FR','Dam','RT','RS'''])* col1 = df[**'Mercury'**] col2 = df[**'Elevation'**] col3 = df[**'Drainage Area'**] col4 = df[**'Surface Area'**] col5 = df[**'Max Depth'**] col6 = df[**'RF'**] col7 = df[**'FR'**] col8 = df[**'Dam'**] col9 = df[**'RT'**] col10 = df[**'RS'**] *# df.query("A == 2 & B > 1 & C != 'z'").shape[0]* print(**"Describe the distribution"**) print(col10.describe()) print(**"Central Tendancy CALC:Measure the Mean"**) print(col10.mean()) print(**"Central Tendancy CALC:Measure the Median"**) print(col10.median()) print(**"Central Tendancy CALC:Measure the Mode"**) print(col10.mode()) print(**"SPREAD CALC: Quartiles"**) print(np.quantile(col10, [0, 0.25, 0.5, 0.75, 1])) print(**"SPREAD CALC: Quantiles"**) print(np.quantile(col10, np.linspace(0, 1, 5))) print(**"SPREAD CALC: Deciles"**) print(np.quantile(col10, np.linspace(0, 1, 11))) print(**"SPREAD CALC: Interquartile Range "**) IQR = np.quantile(col10, 0.75) - np.quantile(col10, 0.25) print(IQR) |

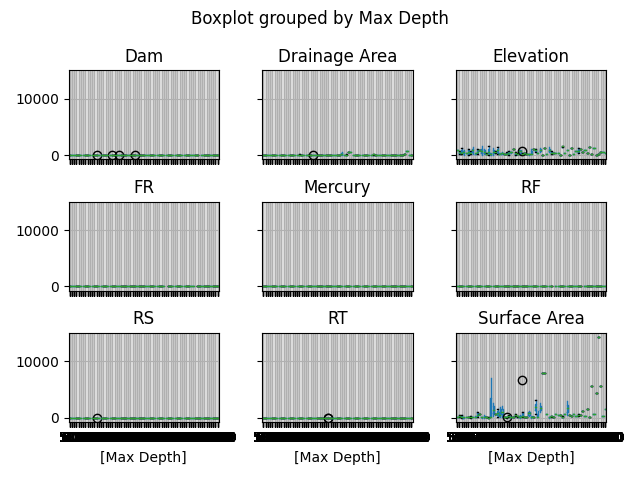
And not to be out down

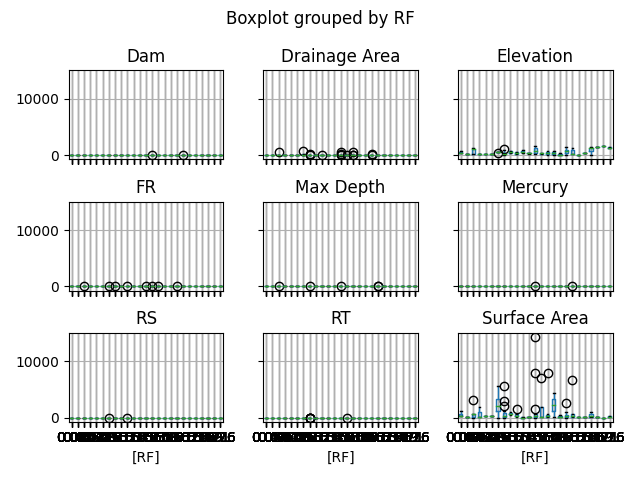
Here are the variables in BOXPLOT

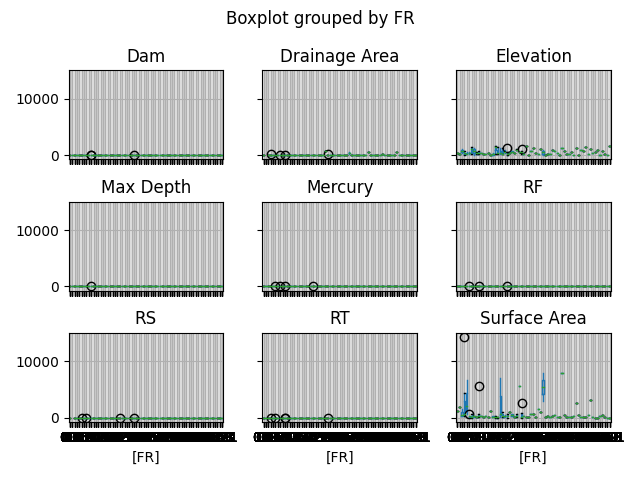


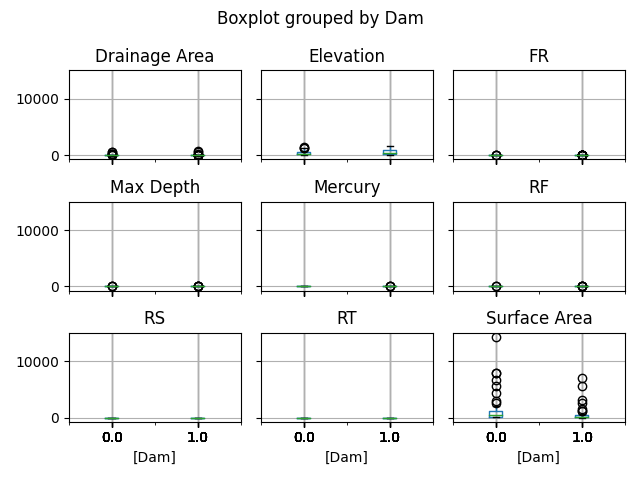


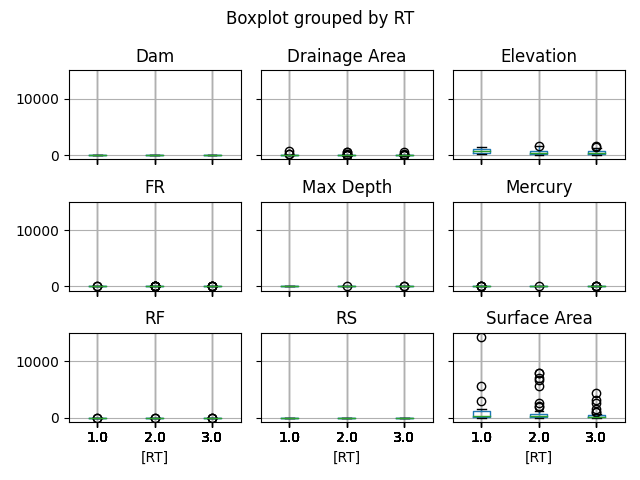


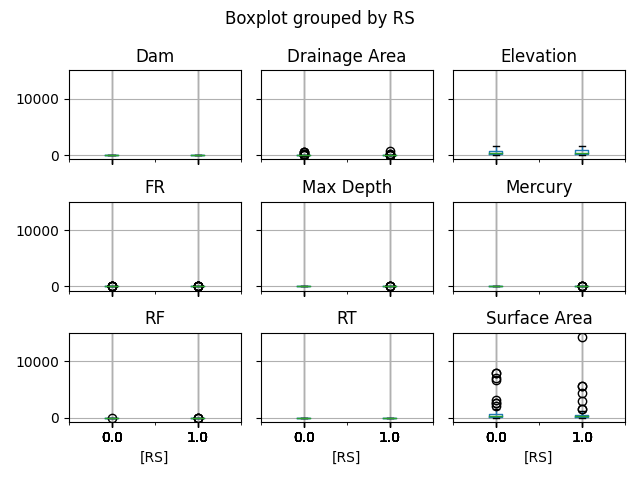


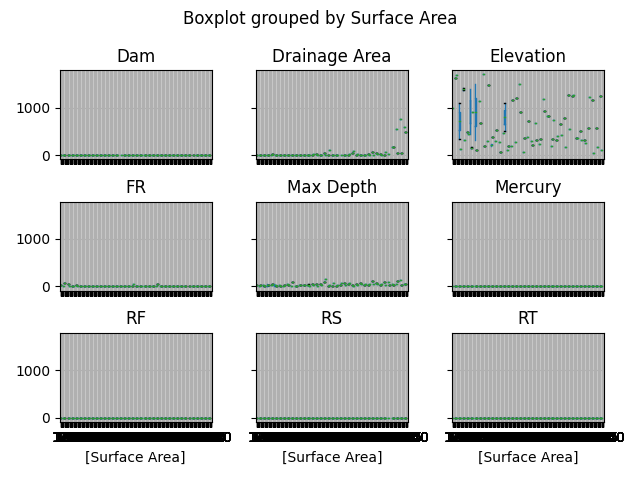












QUESTION 3:

Based on the given dataset, estimate the overall risk level of all reservoirs in Fairfield. Should the local authority take any action?

**Regards to question 3**.

Using the data given by Fairfeild local environmental agency authority on fish samples to see if the fish samples levels exceed the safety limit for human consumption.

We are using a variety of imports function to calculate our mean, standard deviation and size sample. We have used the data frame statement to read the data from the file and then we are just getting the values of our Desired column Mercury excluding the header name. As we are using stats model to calculate our Confidence interval we need standard error, confidence level and significance level as well as we are using t-statistics we need degree of freedom as well. So, we need our Lower C.Interval statmodel and upper C.Interval statmodel to calculate our C.Interval of the mean at 99% confidence level.

As we can see the Confidence level of the mean at 99% confidence level is between 0.41 and 0.57 which is significantly higher from Mercury level provided by Fairfields Local environmental agency authority. So, yes they should need to take action on it to reduce the risk level.

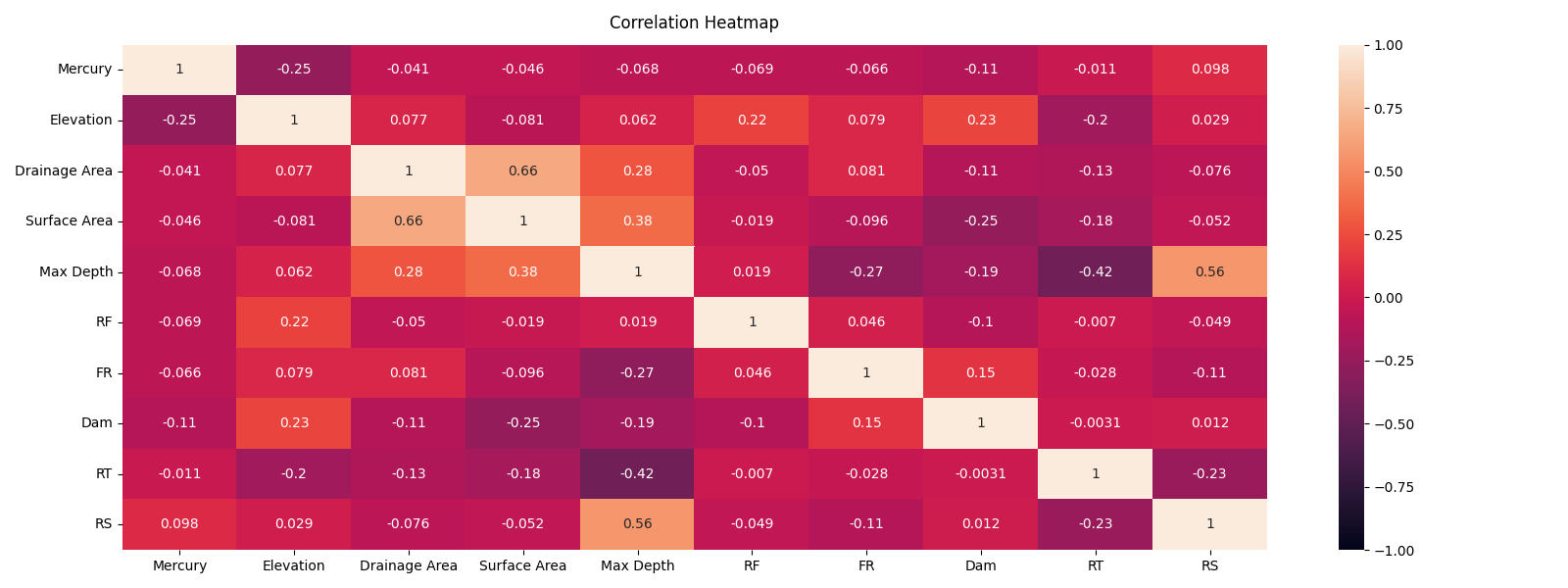
QUESTION 4

There are concerns among industrialists who are benefiting from dams and dam constructions that there will be claims that high mercury levels in fish are related to the presence of dams in the reservoirs drainage.

Does the data support or refute this claim?

Provide some justifications.

Firstly There is no Correlation between Mercury and DAMs !



Correlation analysis in PANDAS shows no significant correlation between Mercury and d

Or any other variable! This is using all 3 CORRELATION methods of : PEARSON, KENDALL TAU and SPEARMAN

Above is HEATMAP by Correlation (CODE BELOW)

Correlation of Mercury and DAM

pearson

Mercury Dam

Mercury 1.000000 -0.112865

Dam -0.112865 1.000000

kendall

Mercury Dam

Mercury 1.000000 -0.110572

Dam -0.110572 1.000000

spearman

Mercury Dam

Mercury 1.000000 -0.134144

Dam -0.134144 1.000000

Secondly the Hypothis Test does not

Code FOR Correlations!

**import** pandas **as** pd  
**import** numpy **as** np  
**import** matplotlib.pyplot **as** plt  
**from** statsmodels.graphics.gofplots **import** qqplot  
  
df = pd.read\_csv (**r'C:\temp\deadfishies.csv'**)  
df.head()  
  
  
*#Elevation  
#Drainage Area  
#Surface Area  
#Max Depth  
  
#Describe the Distribition*print(**"Correlation of Mercury and DAM"**)  
  
data = df[[**'Mercury'**,**'Dam'**]]  
correlation1 = data.corr(method=**'pearson'**)  
correlation2 = data.corr(method=**'kendall'**)  
correlation3 = data.corr(method=**'spearman'**)  
print(**'pearson'**)  
print(correlation1)  
print(**'kendall'**)  
print(correlation2)  
print(**'spearman'**)  
print(correlation3)

Code FOR Heatmap!

**import** pandas **as** pd  
**import** numpy **as** np  
**import** matplotlib.pyplot **as** plt  
**import** seaborn **as** sns  
**from** statsmodels.graphics.gofplots **import** qqplot  
  
df = pd.read\_csv (**r'C:\temp\deadfishies2.csv'** )  
df.head()  
 df[**'Dam'**]  
col9=df[**'RT'**]  
col10=df[**'RS'**]  
  
  
sns.heatmap(df.corr());  
   
plt.figure(figsize=(16, 6))  
*# Store heatmap object in a variable to easily access it when you want to include more features (such as title).  
# Set the range of values to be displayed on the colormap from -1 to 1, and set the annotation to True to display the correlation values on the heatmap.*heatmap = sns.heatmap(df.corr(), vmin=-1, vmax=1, annot=**True**)  
*# Give a title to the heatmap. Pad defines the distance of the title from the top of the heatmap.*heatmap.set\_title(**'Correlation Heatmap'**, fontdict={**'fontsize'**:12}, pad=12);  
   
*#plt.hist(col10, edgecolor='black', bins=20)  
  
# showing the graph*plt.show()