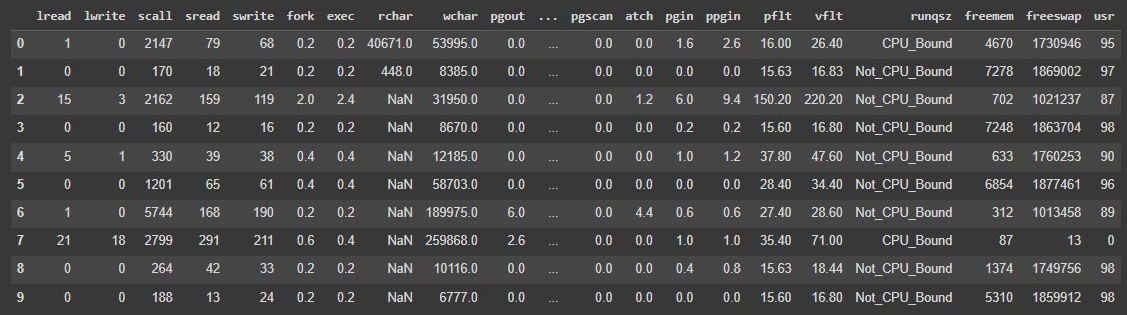
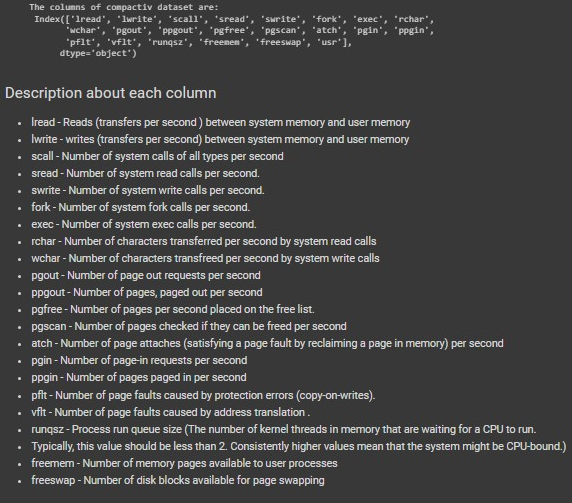
**Report**

**Problem-1**

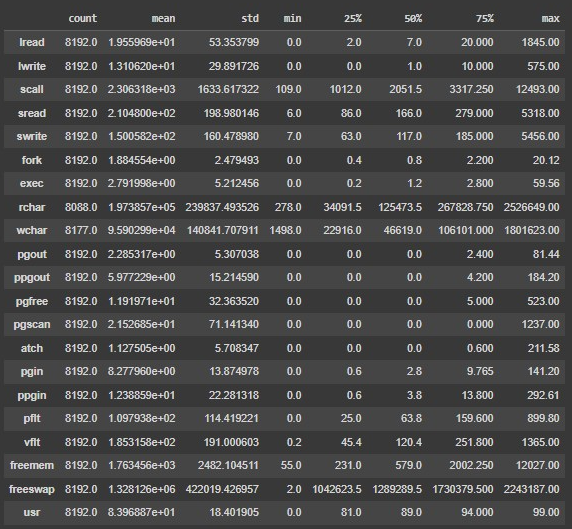
Solution 1.1:



* The compactive dataset have 8192 rows and 22 columns
* The compactive dataset contains 180224 data points



Five Number Summary:



Basic information about data:

* There are 3 types of data format present in compactive dataset.
* int64 (8 columns)
* flaot64 (13 columns)
* object (1 column)
* The dataset occupies 1.4+ MB of memory space
* There are missing values in compactive dataset

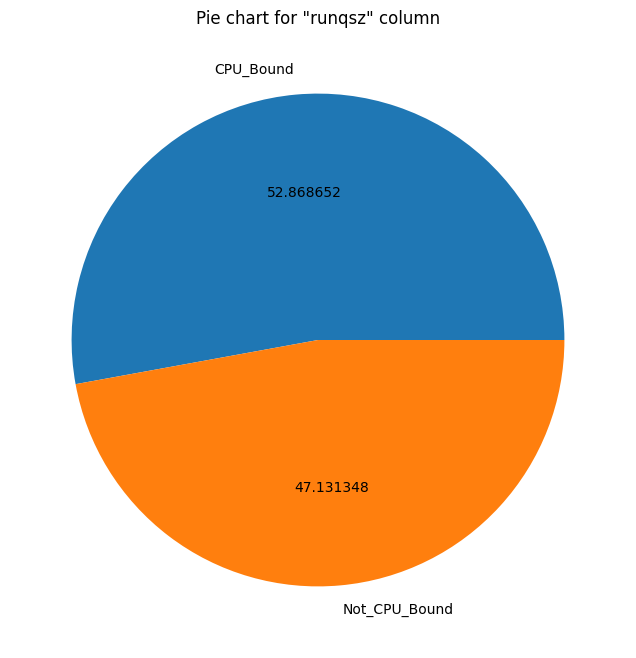
Missing values in dataset:

* There are 104 missing values in "rchar" column.
* There are 15 missing values in "wchar" column.

Duplicate values in dataset:

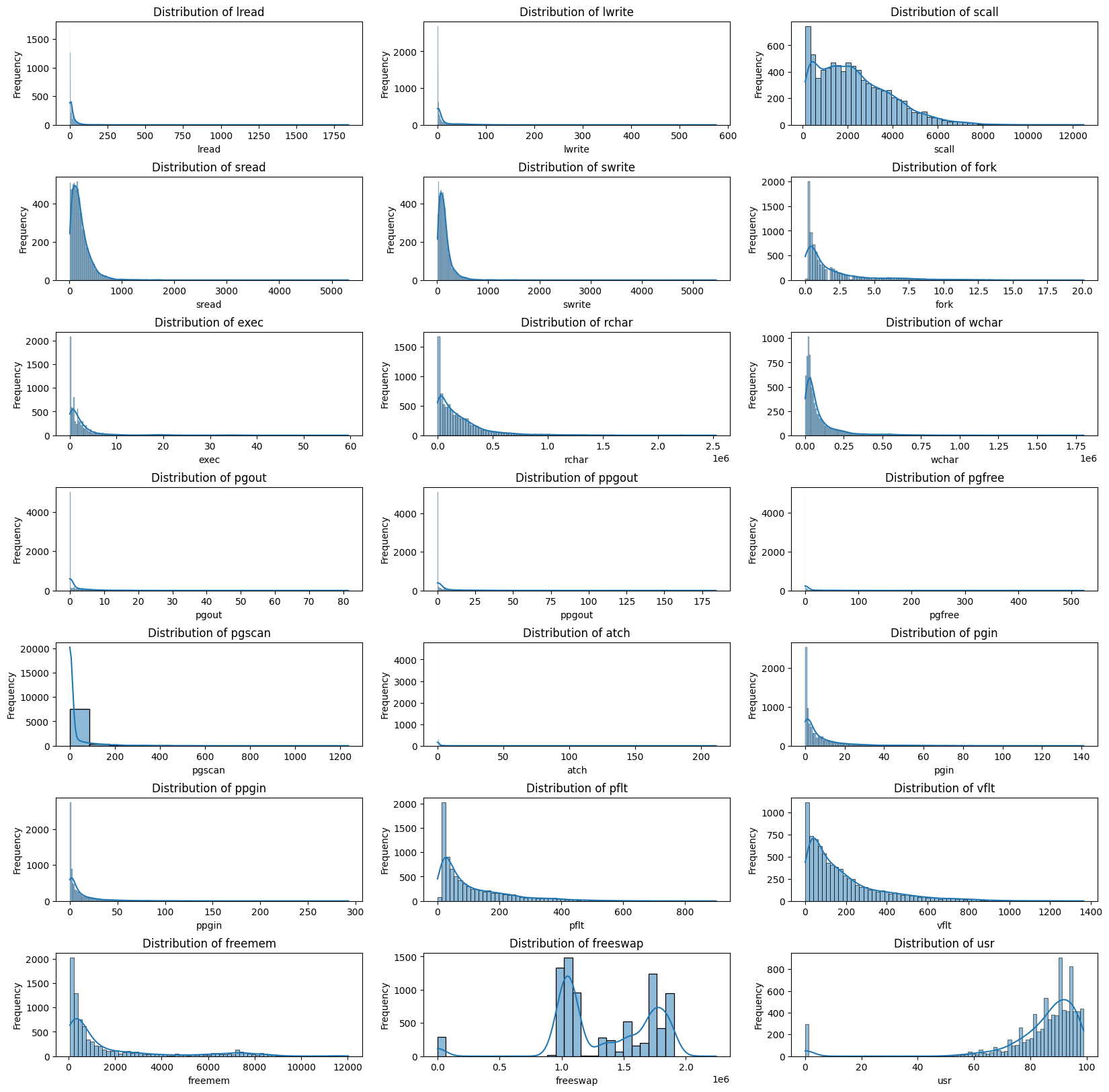
* There is no duplicated values in compactive dataset.

Analysis of categorical column:

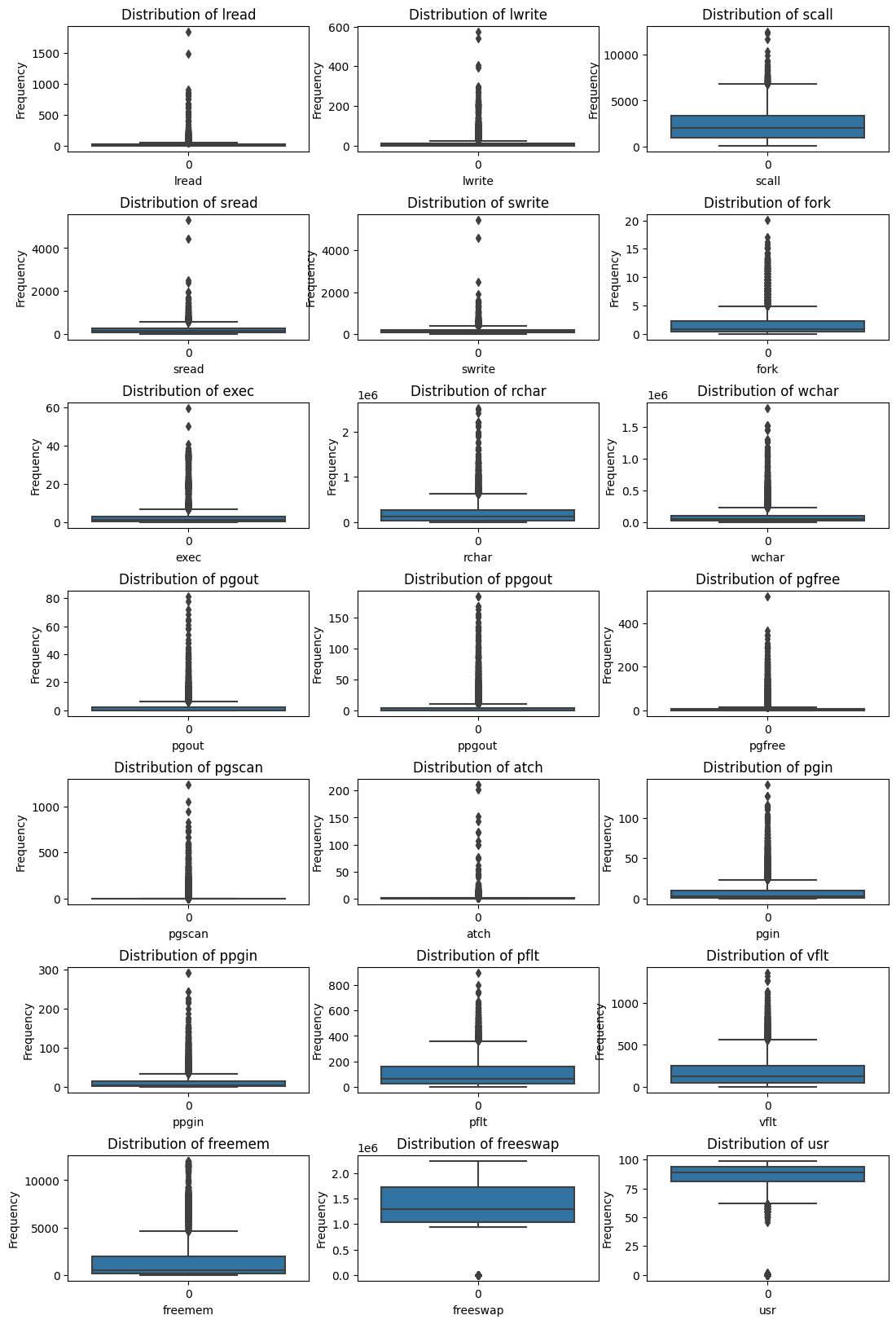


Analysis of numerical columns:

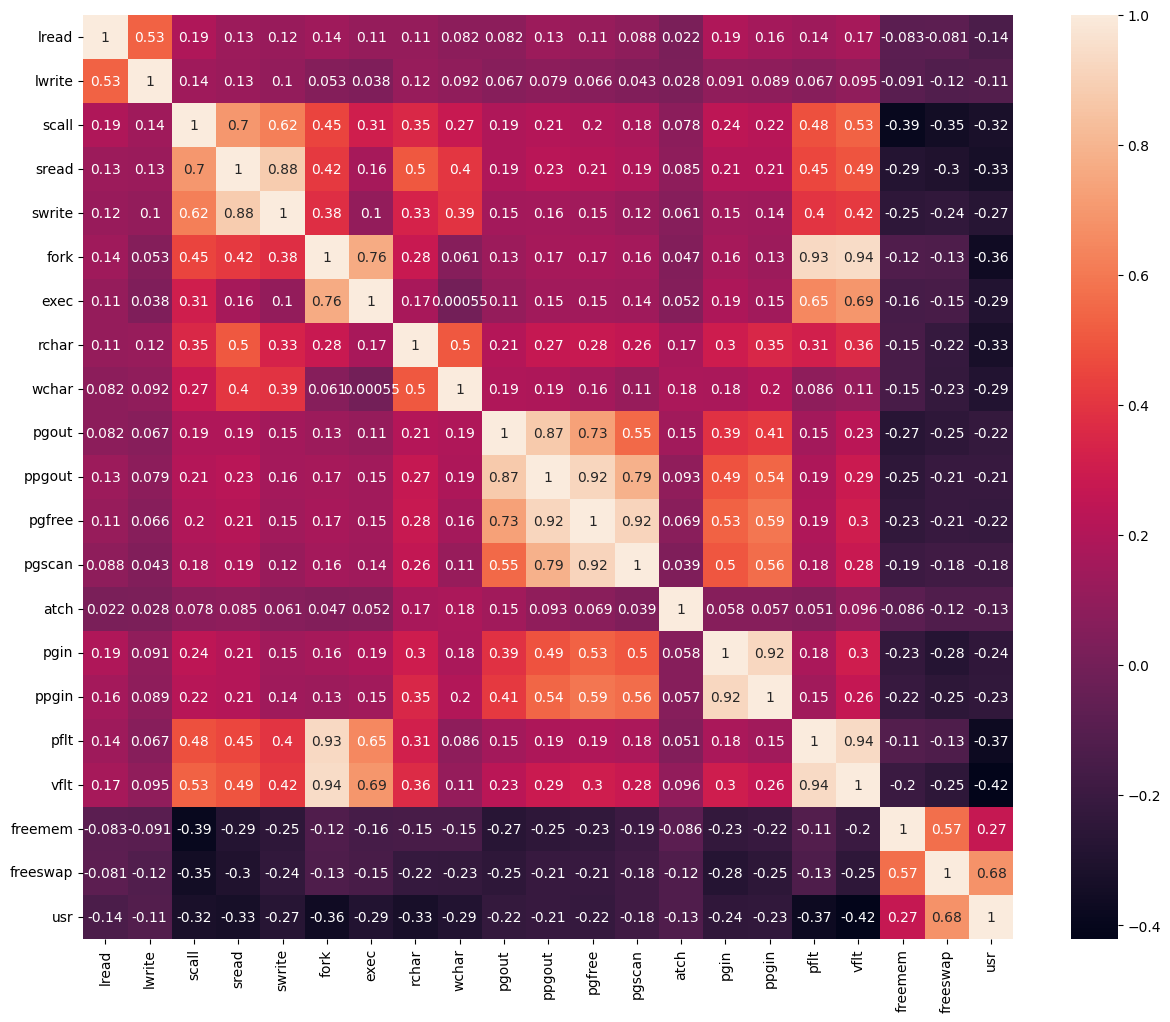
* Distribution plot for numerical columns:



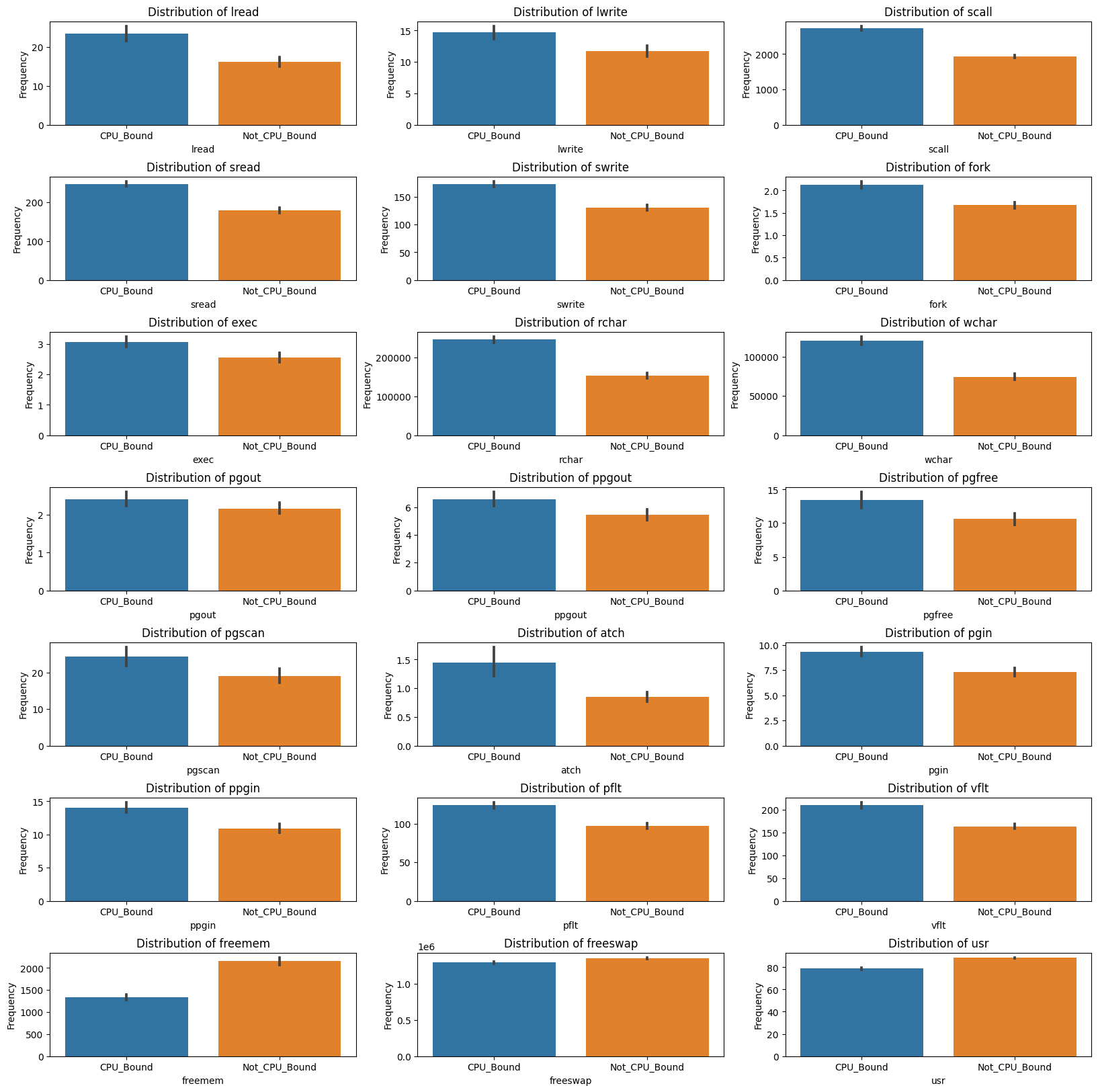
* Box plot for numerical columns and looking at outliers:



* We can observe from above box plots that almost every column have outliers.
* Heatmap that show correlation between features:



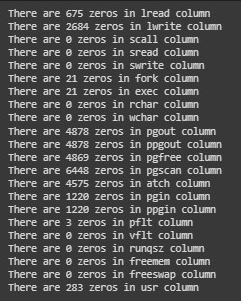
* Here we can understand the correlation between each feature.
* value in (-0.25,0) or (0,0.25) --> Weak correltion between features.
* value in (-0.75,-0.25) or (0.25,0.75) --> intermediate correltion between features.
* value in (-1,-0.75) or (0.75,1) --> strong correltion between features.
* Based of categorical variable visualizing the relation between other variables:



Solution:1.2:

* There are 104 missing values in "rchar" column.
* There are 15 missing values in "wchar" column.
* And since both columns have outliers we will used median value of respective columns to fill the missing values.

Analyzing the zero’s in each columns:



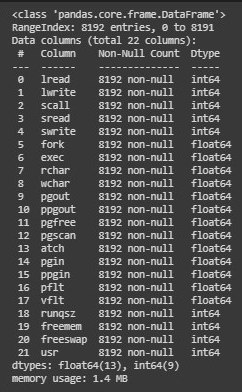
* We can keep the Zero's in our dataset for further analysis as we have some features in our dataset which can be Zero if the system is not performing any task.

Solution 1.3:

Handling categorical variable:

* From above analysis we can see that we have only one categorical features i.e "runqsz" it has two categories.
* CPU\_Bound
* Not\_CPU\_Bound
* We will encode it as:
* CPU\_Bound : 1
* Not\_CPU\_Bound : 0

After encoding:

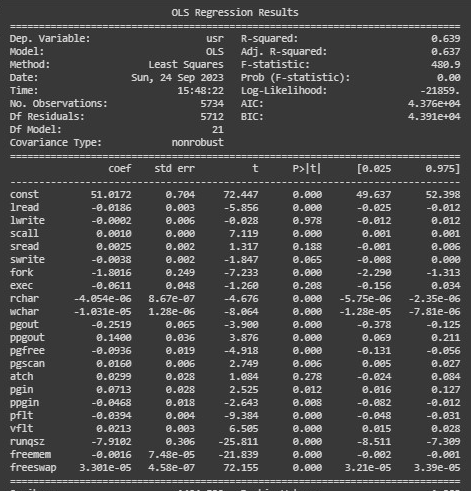


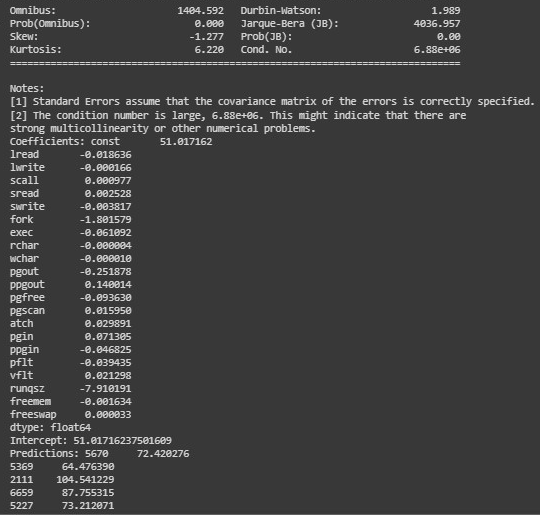
Metrics of Linear Regression model from sci-kit library:



* We get 64% accuracy.

Using OLS for Linear Regression from stats\_model:





* We get R2\_squared accuracy of 63.90%.

From model-1 that is sci-kit Linear Regression we have r2\_squared accuracy tov be: 64.06%

From model-2 that is stats-model Linear Regression we have r2\_squared accuracy tov be: 63.9%

There we select model-1 for prediction.

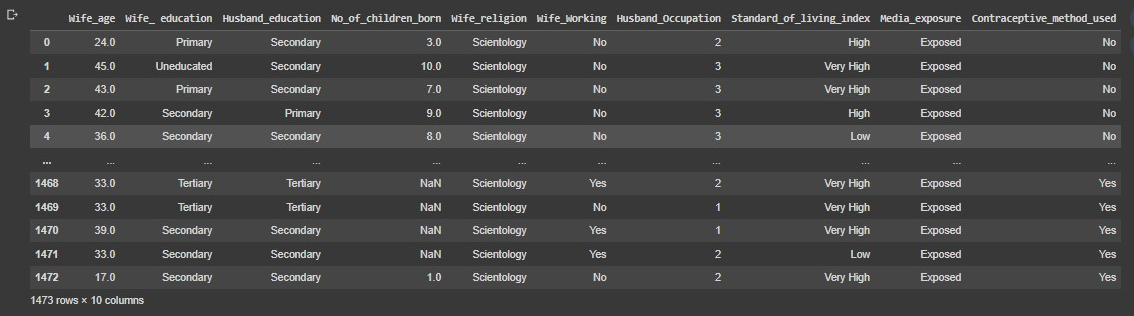
Solution 1.4:

* Bussiness insights and recommendation
* The model suggests that various independent variables have different impacts on the dependent variable 'usr'.
* Focus on the statistically significant variables that have a meaningful impact on 'usr'.
* For negatively correlated variables, consider strategies to optimize or reduce them if they are costly or undesirable.
* For positively correlated variables, consider strategies to increase them if they are beneficial to your business.

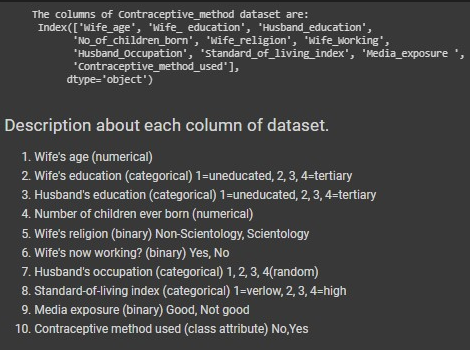
**Problem-2**

Solution 1.1:

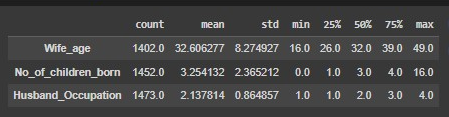
Dataset:



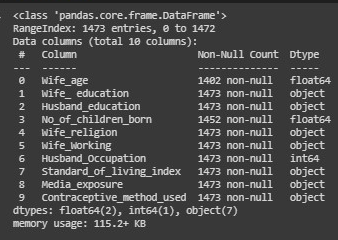
* The Contraceptive\_method dataset have 1473 rows and 10 columns
* The Contraceptive\_method dataset contains 14730 data points



Five number summary of data:



Basic information about datset:



\* There are 3 types of data present in given dataset.

>> \* int64 (1 column)

>> \* float64 (2 columns)

>> \* object (7 columns)

\* File size is 115.2+ KB

\* There are some missing values in "Wife\_age" and "No\_of\_children\_born" column

Missing values:

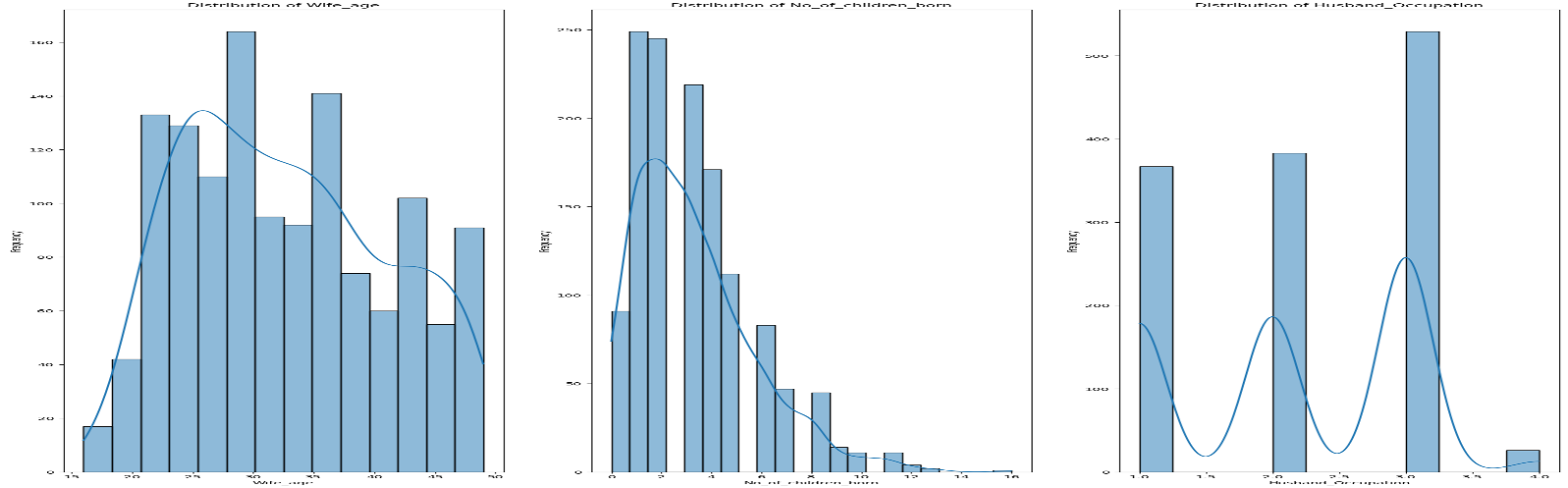
* There are 71 missing values in "Wife\_age" column.
* There are 21 missing values in "No\_of\_children\_born" column.
* We drop the missing values

Duplicate values:

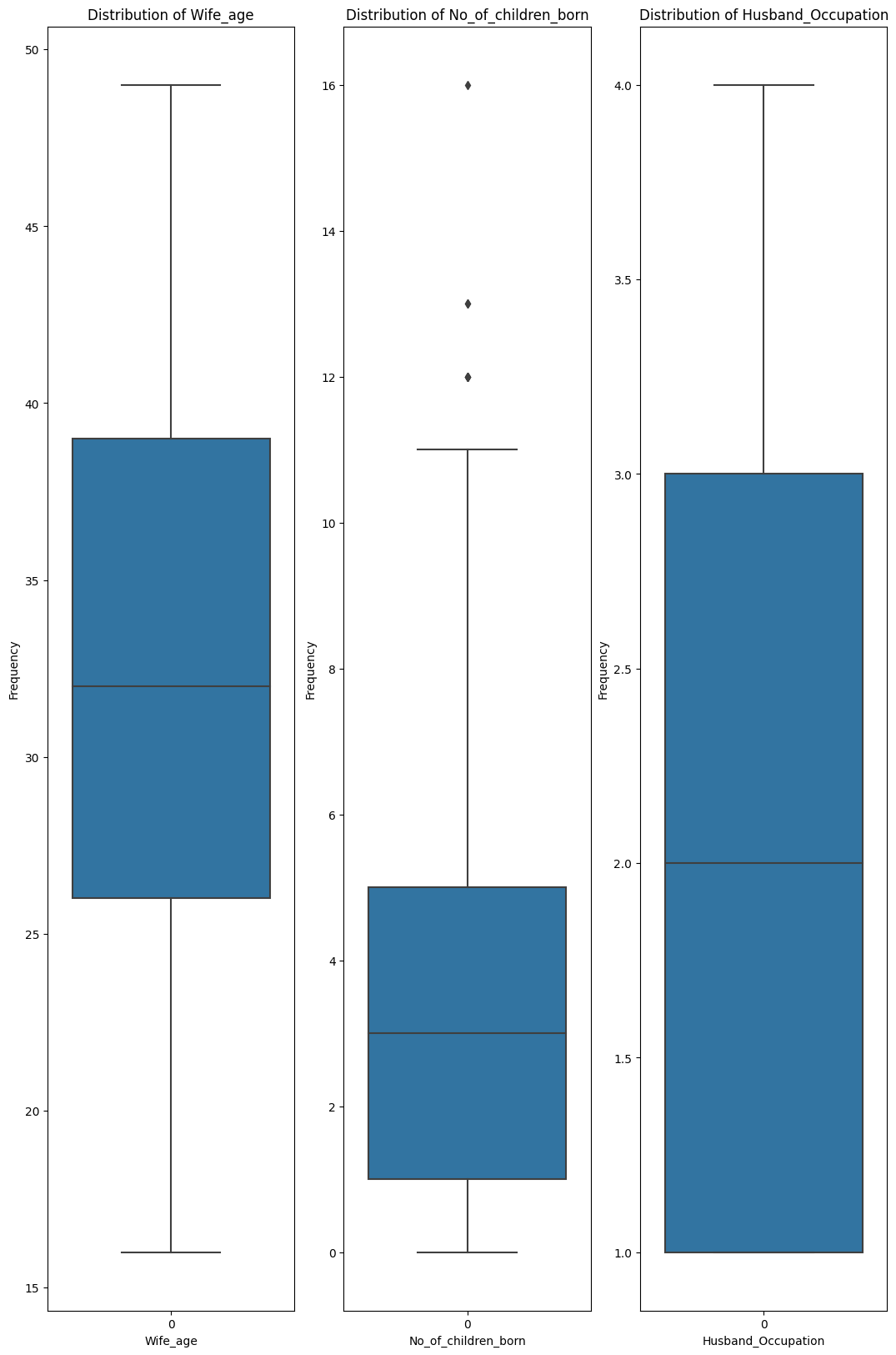
* There are 80 duplicate values in given dataset.
* We delet the duplicated values keeping the firstf one.

Analysing numerical variables:

* Distribution of the data:

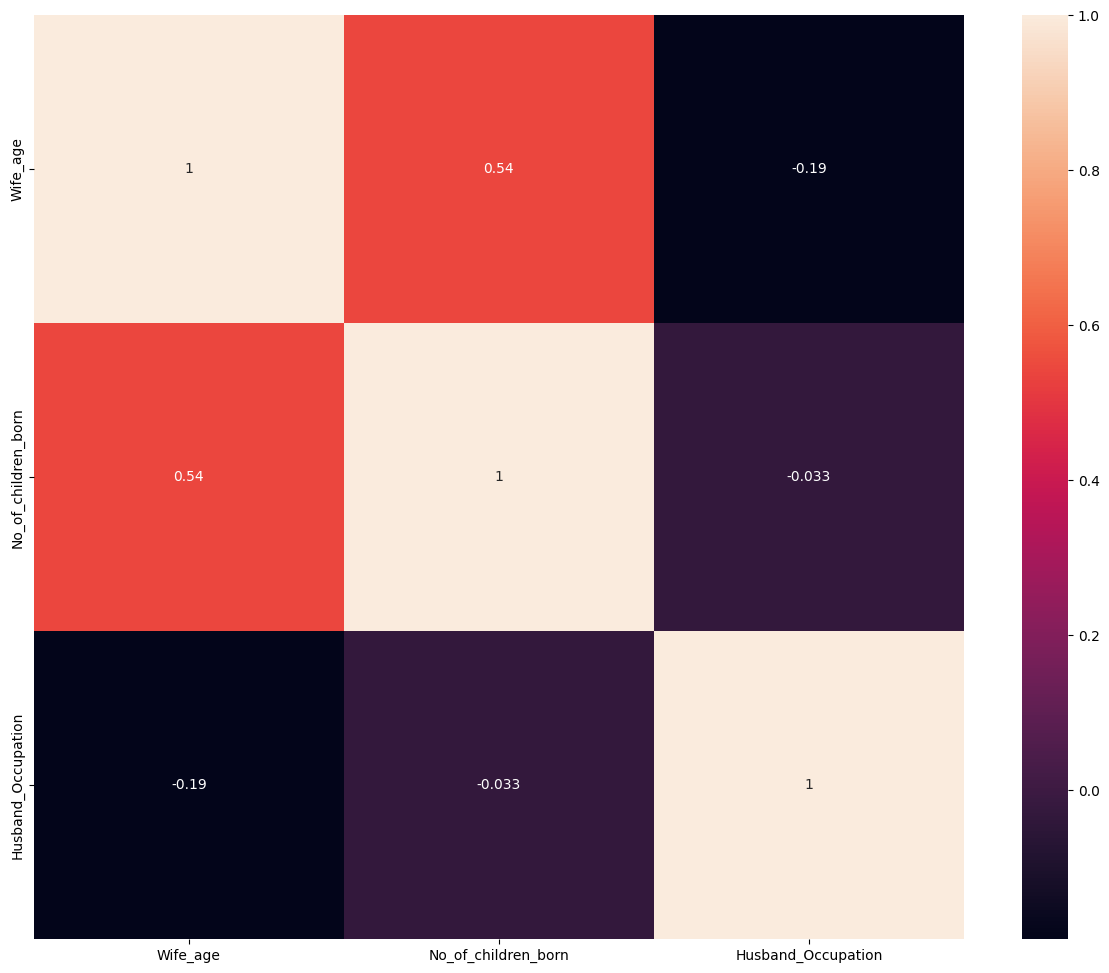


* Box plot of numerical features:



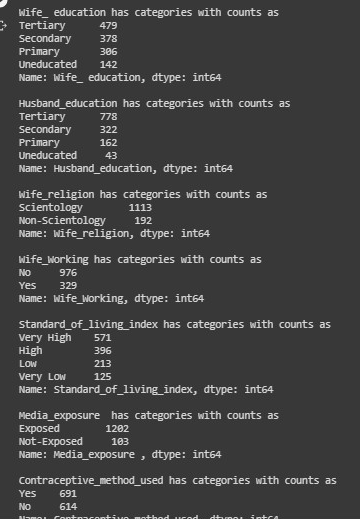
* There are few outliers in No\_of\_children columns

Heatmap to observe the correlation between numerical variables:



* There is intermediate correlation between No\_of\_children and Wife\_Age

Analysis of categorical variables:

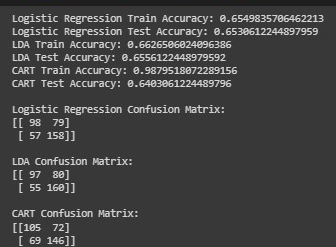


Solution 2.2:

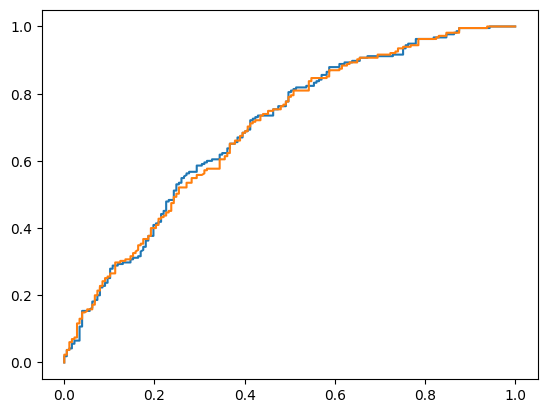
* We encoded the dependent variable “Contraceptive\_method\_used” as:
* No as 0
* Yes as 1
* We used one hot encoding technique for other categorical variable.

Solution 2.3:

Performance metrics:



ROC curve:



Solution2.4:

Based on our predictions, we can draw several insights and recommendations. For instance, we can identify the demographic and socioeconomic characteristics that are most strongly associated with the use of contraceptives among married women. We can also make recommendations to the Indonesian Ministry of Health on how to promote the use of contraceptives among married women, based on the characteristics of the target population.