# BUSINESS REPORT

# Predictive Modelling

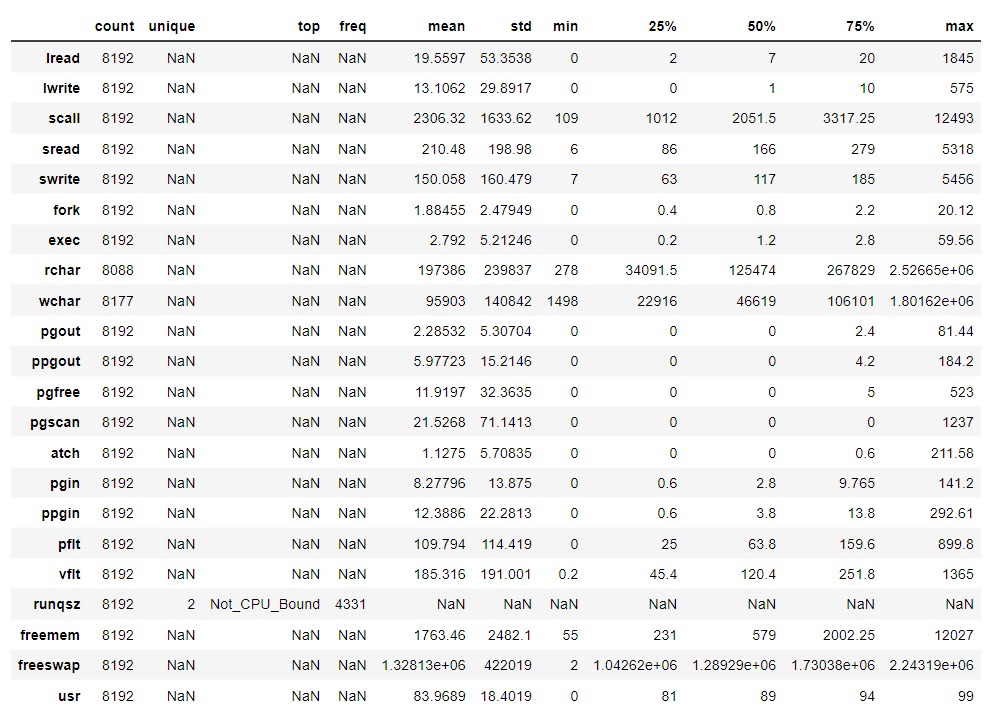
Problem Statement 1: Linear Regression The comp-activ databases is a collection of a computer systems activity measures . The data was collected from a Sun Sparcstation 20/712 with 128 Mbytes of memory running in a multi-user university department. Users would typically be doing a large variety of tasks ranging from accessing the internet, editing files or running very cpu-bound programs. As you are a budding data scientist you thought to find out a linear equation to build a model to predict 'usr'(Portion of time (%) that cpus run in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

Data Description

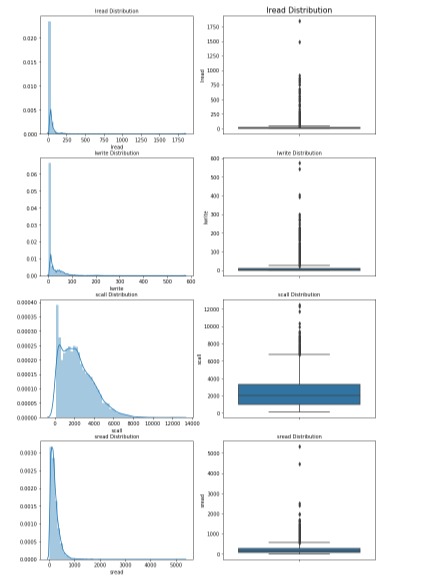
lread - Reads (transfers per second ) between system memory and user memory  
lwrite - writes (transfers per second) between system memory and user memory  
scall - Number of system calls of all types per second  
sread - Number of system read calls per second .  
swrite - Number of system write calls per second .  
fork - Number of system fork calls per second.  
exec - Number of system exec calls per second.  
rchar - Number of characters transferred per second by system read calls  
wchar - Number of characters transfreed per second by system write calls  
pgout - Number of page out requests per second  
ppgout - Number of pages, paged out per second  
pgfree - Number of pages per second placed on the free list.  
pgscan - Number of pages checked if they can be freed per second  
atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second  
pgin - Number of page-in requests per second  
ppgin - Number of pages paged in per second  
pflt - Number of page faults caused by protection errors (copy-on-writes).  
vflt - Number of page faults caused by address translation .  
runqsz - Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run.  
Typically, this value should be less than 2. Consistently higher values mean that the system might be CPU-bound.)  
freemem - Number of memory pages available to user processes  
freeswap - Number of disk blocks available for page swapping.  
usr - Portion of time (%) that cpus run in user mode

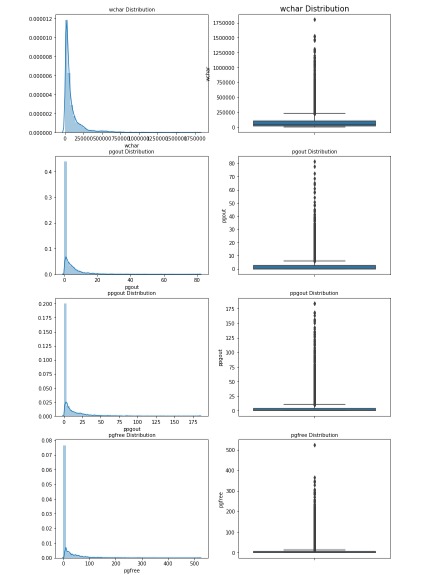
* 1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the Data types, shape, EDA, 5 point summary). Perform Univariate, Bivariate Analysis, Multivariate Analysis.

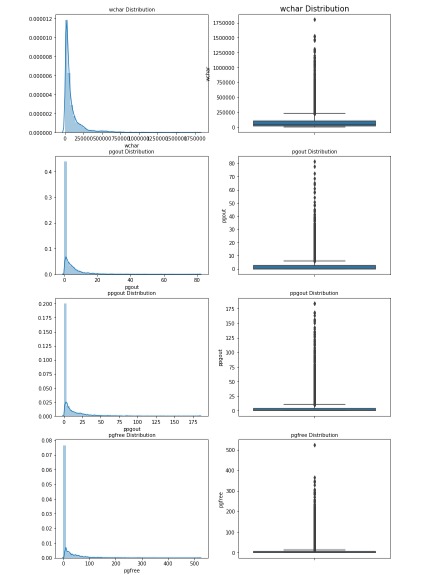
EDA The data is imported, and the following are the observations: • The data has 8192 rows and 22 columns. There is 1 object type data types and rest are float 7 int data types.



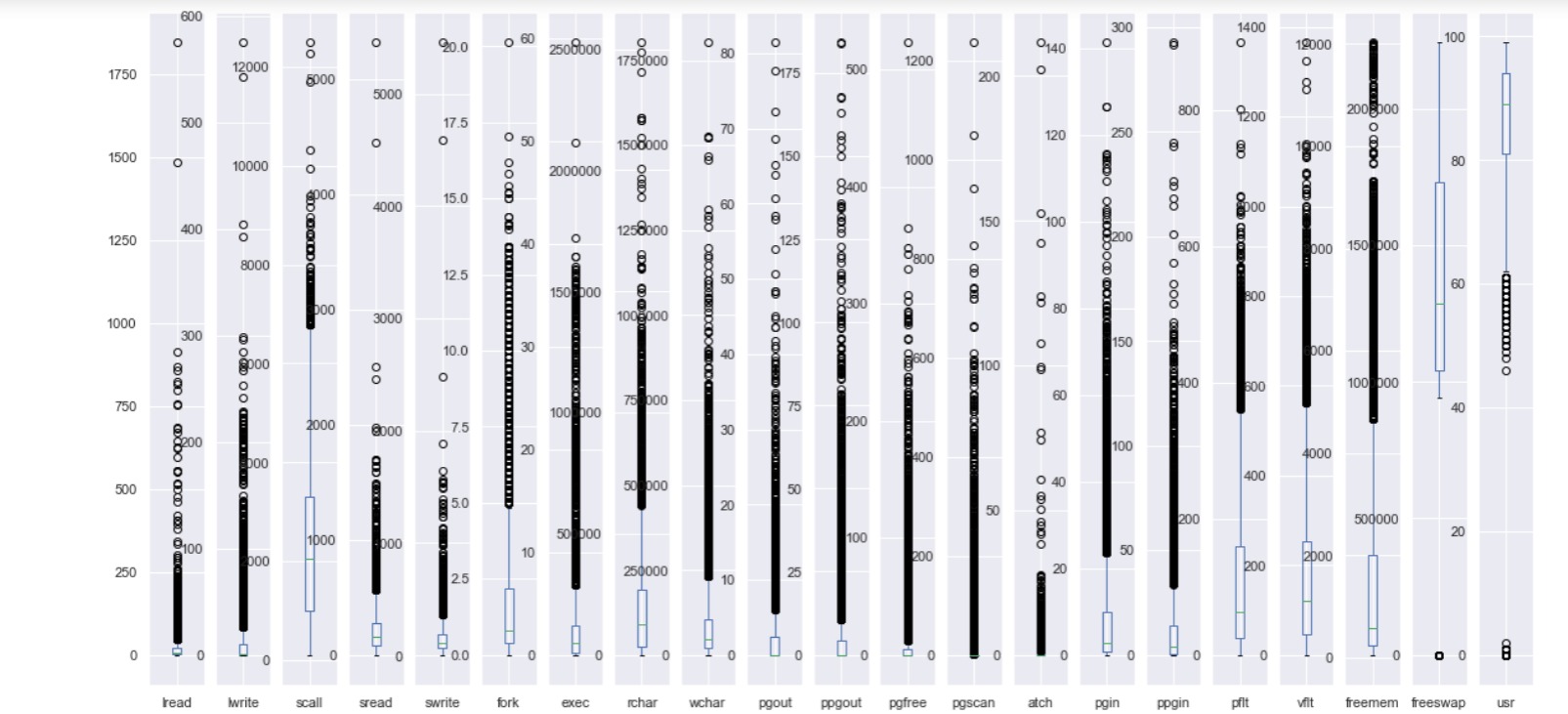
* Majority of the times the process run queue size was Not CPU Bound.
* On an average 83.9% of times the cpus run in user mode.
* There are no duplicate rows present in the data.
* There are few missing values in variables ‘rchar’ & ‘wchar’. These will be treated later.
* A new feature can be calculated which is the ‘System Read-Write rate’ by the features Number of system read and write calls per second. So these features can be replaced by the newly created one.

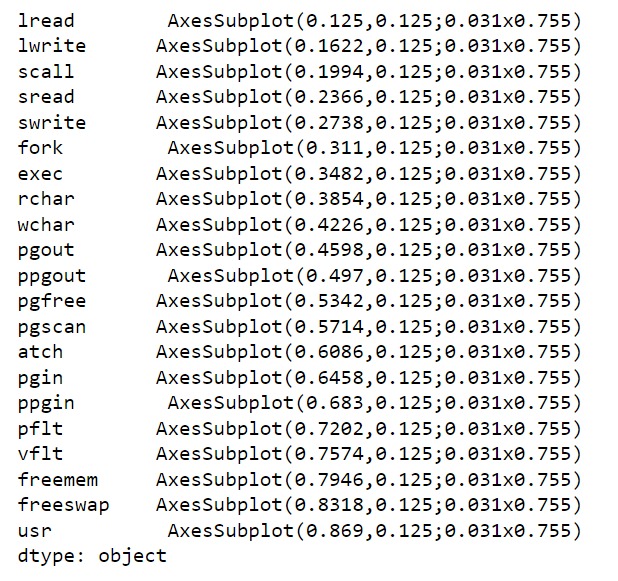


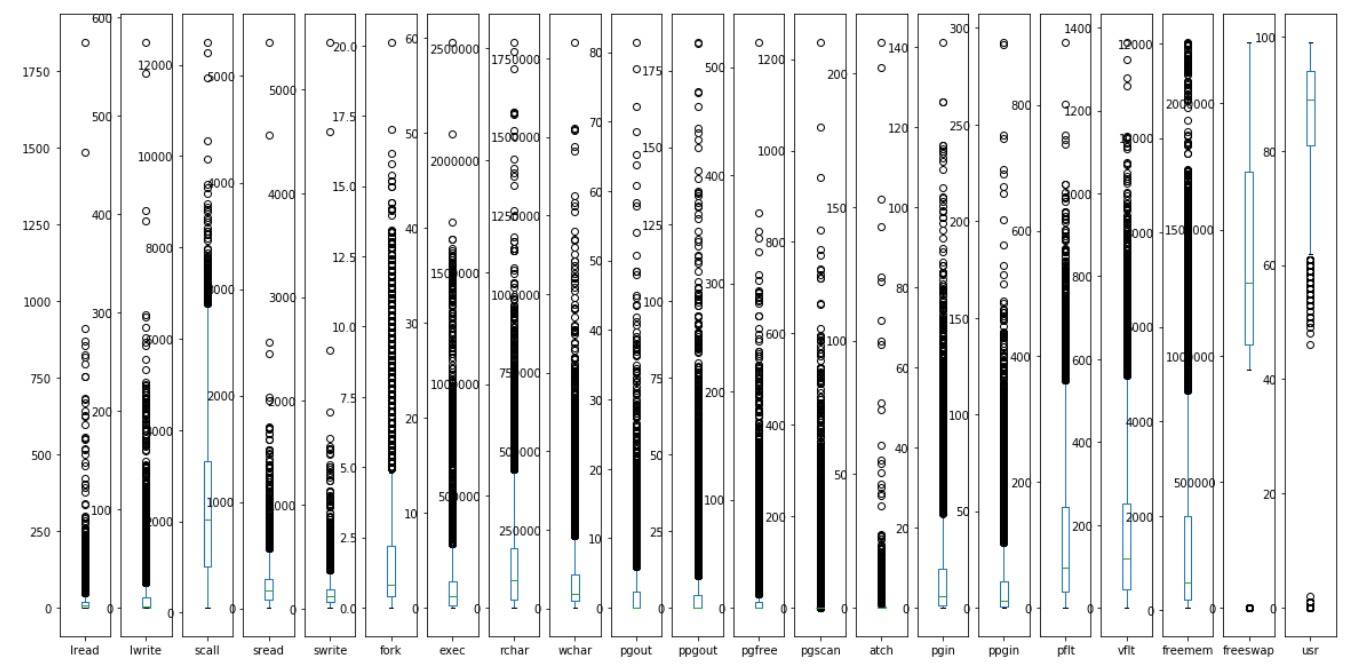




The transfers per seconds for read and write is pretty fast as majority of the transfers are happening quickly. The System read-write rate is also quick and majority of the transactions happen to be under 5%. It seems that there are not many activities that are happening



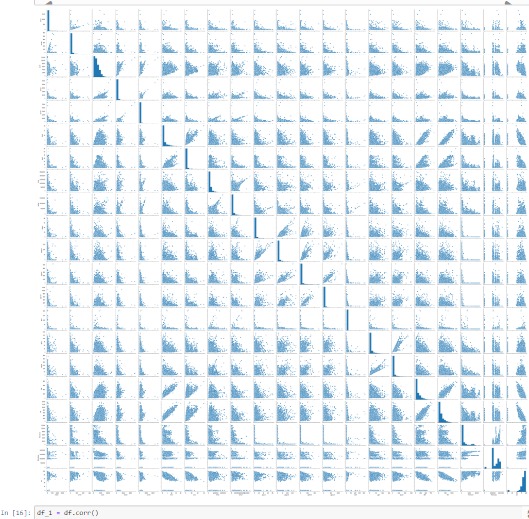




Shown all perfectly.

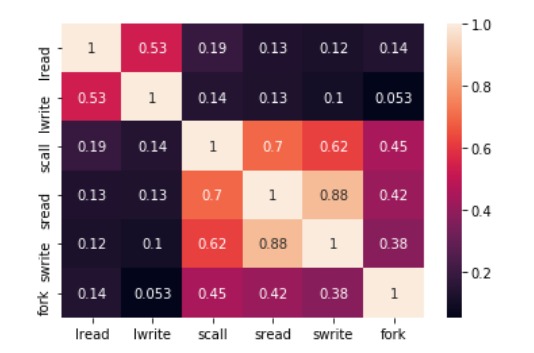
From the pairplot shown below, we can see correlation between a few variables: o Linear correlation can be seen between ‘vflt’, ‘pflt’ & ‘fork’. If the fork calls increase, number of page faults also tend to increase.

Similarly, Number of page out requests per second is also highly correlated to the number of pages, paged out per second.



• Similar correlations can be observed from the heatmap

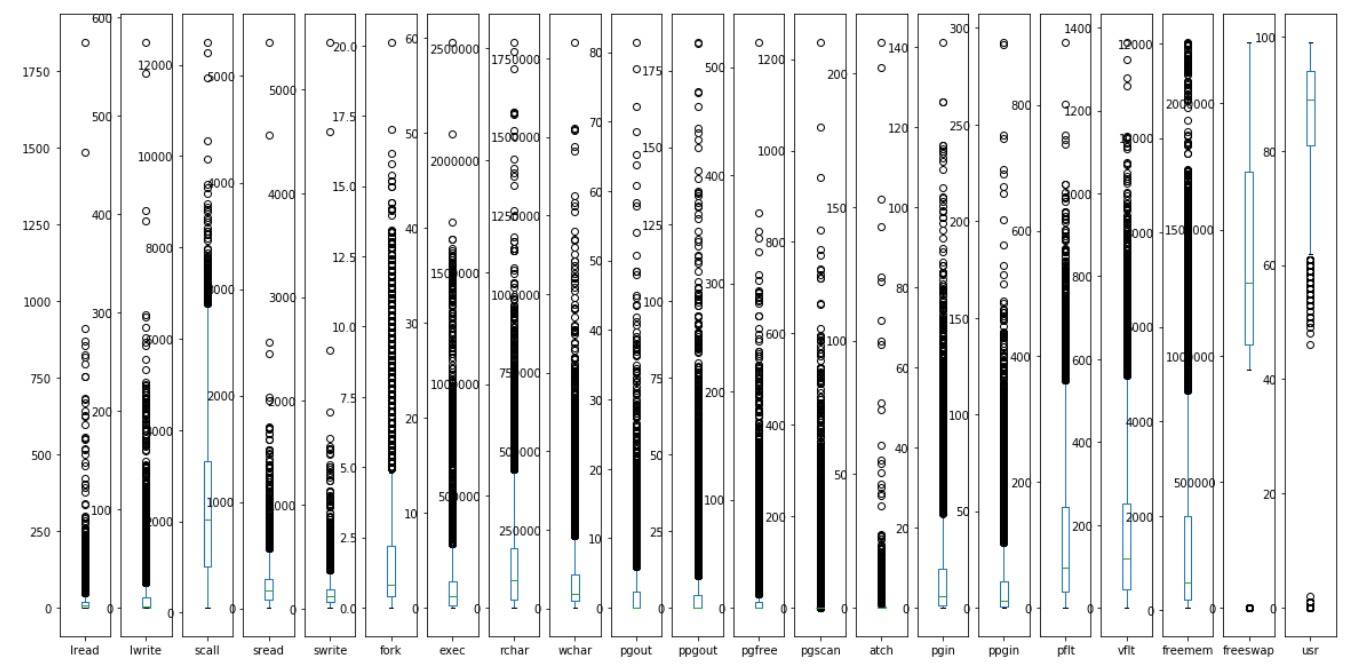
•Majority of the times ‘pgscan’ i.e the number of pages checked if they can be freed per second is 0.



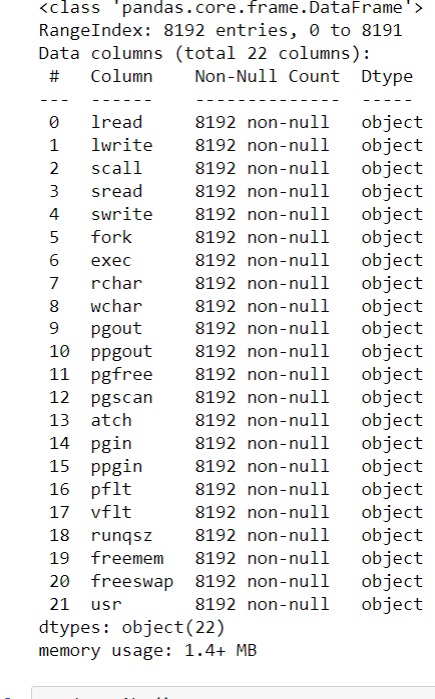
• The black dots boxplots show that there is presence of outliers in all the variables.

• Majority of the variables are highly skewed as well.

• All the outliers are treated by adjusting them to the lower and upper bound values calculated by the IQR value.



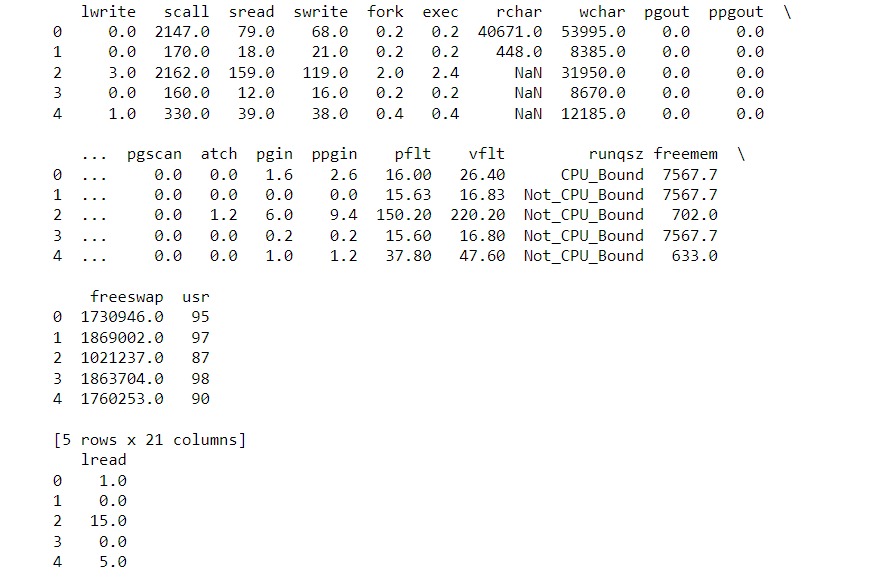
**1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of creating new features if required. Also check for outliers and duplicates if there.**

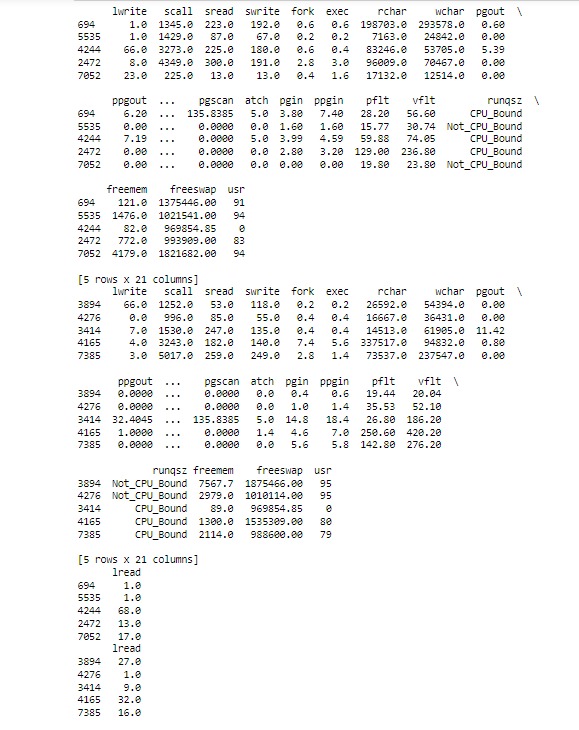
****

The data is imported, and the following are the observations:

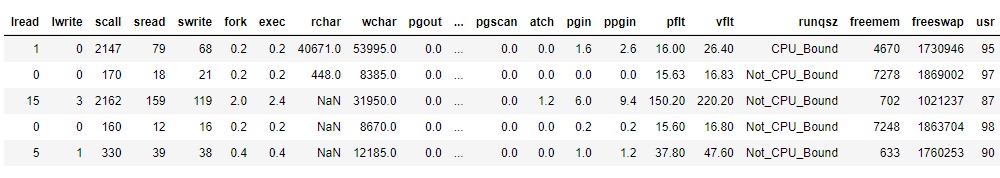
* As observed before, there were null values present in the variables ‘rchar’ & ‘wchar’.
* Since there are only a few values missing, these null values are replaced with the median value.
* It has also been observed that there are 0s present for many dimensions. These are all valid values as a it is related to the activities in the computer system.
* No ordinal variables are available in the data hence an option to combine the sub-ordinal variables is not there.
* Instead, as described above; a new feature is generated i.e. ‘srw\_rate’ (System Read-Write rate) which will be useful in model building and reducing multi-collinearity in the data.
* New features - number of page rate & page requests rate have also been created with the variables pgin, pgout, ppgin & ppgout.
* However these new features are not giving any significant output as majority of the values are 0 or inf.

**1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.**





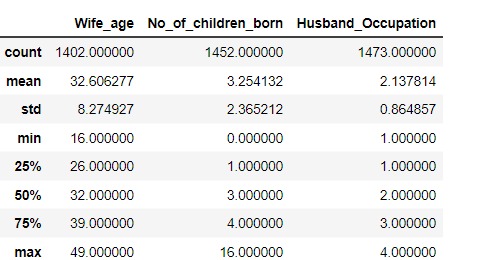
One Hot encoding is done on the only ‘Object’ types variable i.e ‘runqsz’. A new column is created, with 1 indicating that variable as True and 0 as False and this is how the extended variable’s data looks.



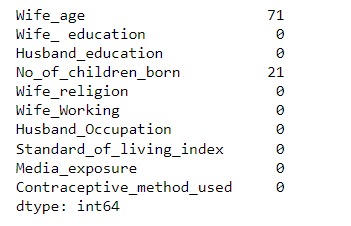
**2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, check for duplicates and outliers and write an inference on it. Perform Univariate and Bivariate Analysis and Multivariate Analysis.**

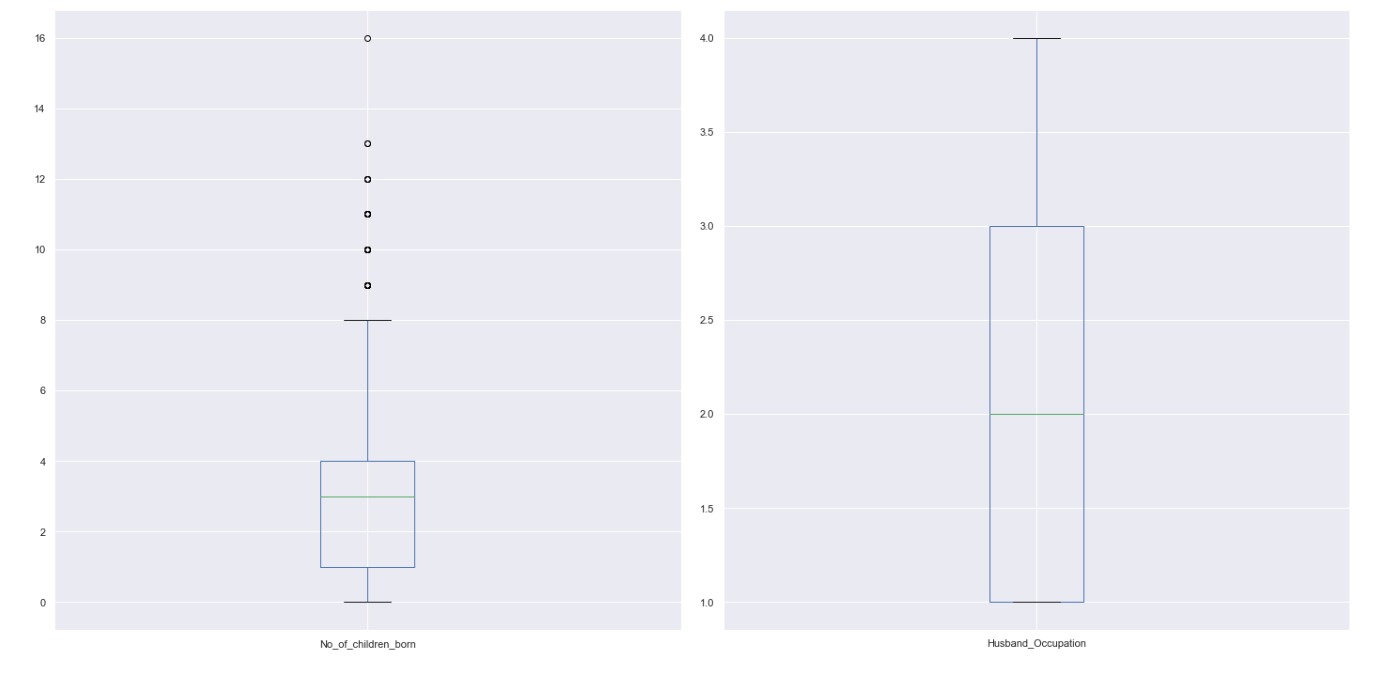
EDA The data is imported, and the following are the observations:

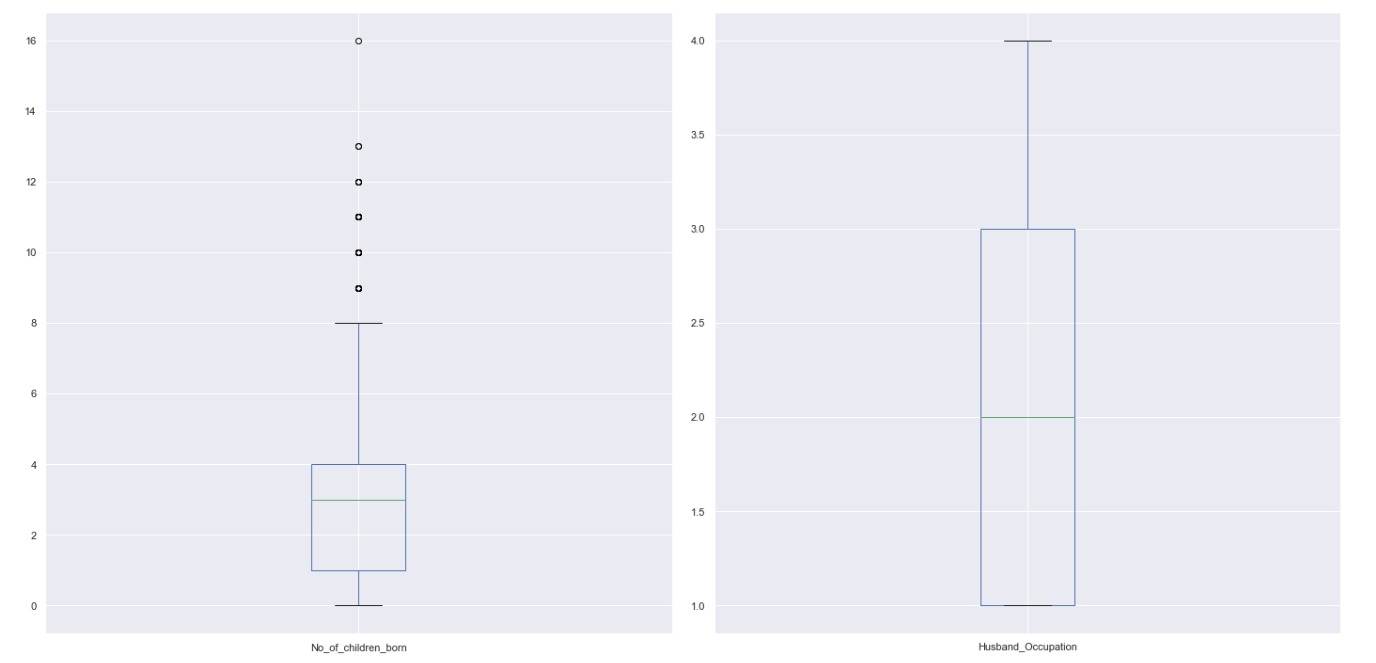
* The data has 1473 rows and 10 variables. There are 5 variables that have object data types and rest are int data types.
* There are a few missing values in the dataset in the variables ‘wife\_age’ and ‘No\_of\_children\_born’. These are replaced by the median values to remove the null entries.
* There are 80 duplicate rows which can be dropped from the dataset. The number for rows is 1393 now.
* The variable ‘Husband\_Occupation’ has been also changed to Object data type as it is a categorical variable.

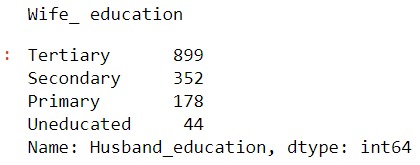


Null values

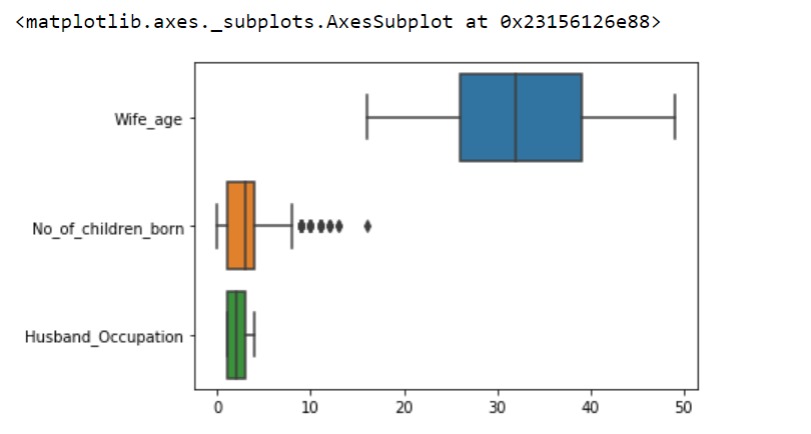




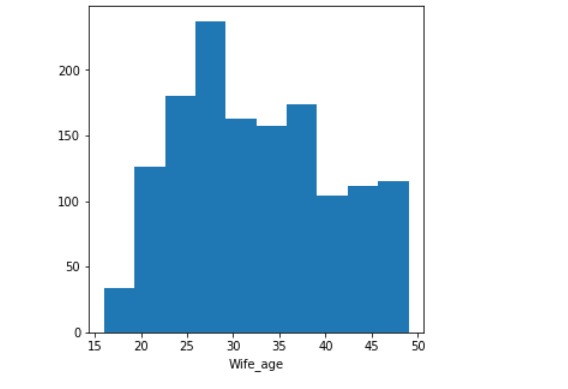




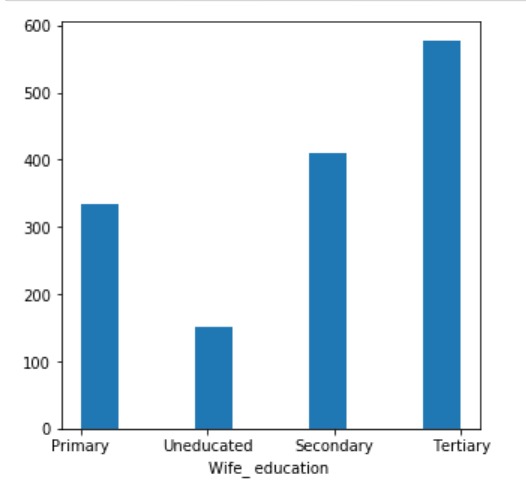
Checked all outliers

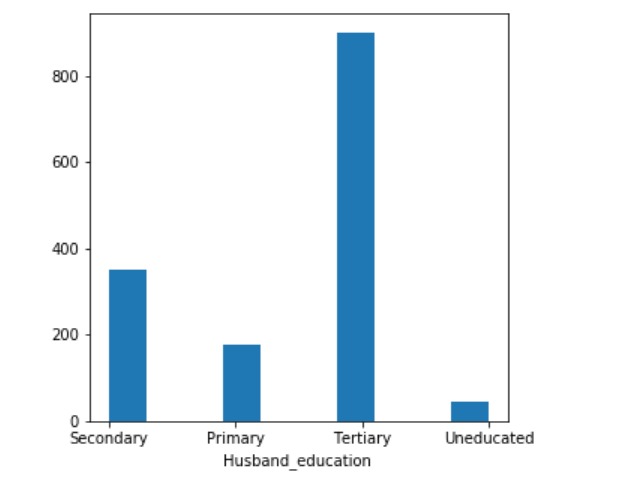


Compared between Wife age , No of children born and husband Occupation.

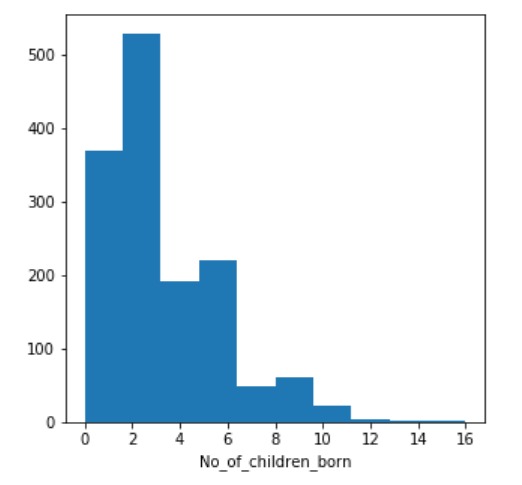


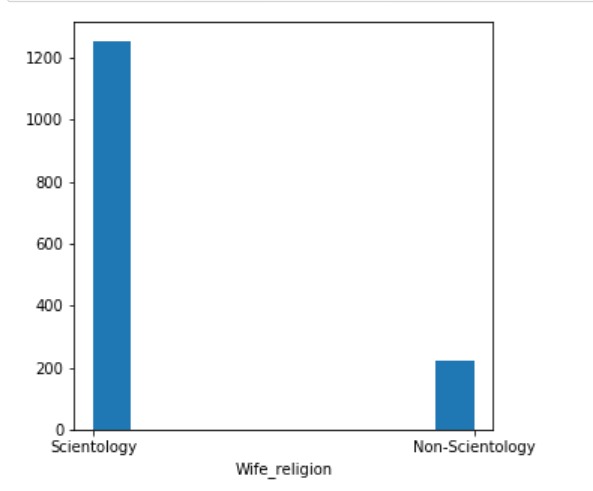
* The age of the wives ranges from 18 to 49 years where most of them are in their 30’s and mid 20’s early 50’s.
* Majority of the people had 1 or 2 children but a few have more than 15 children as well.



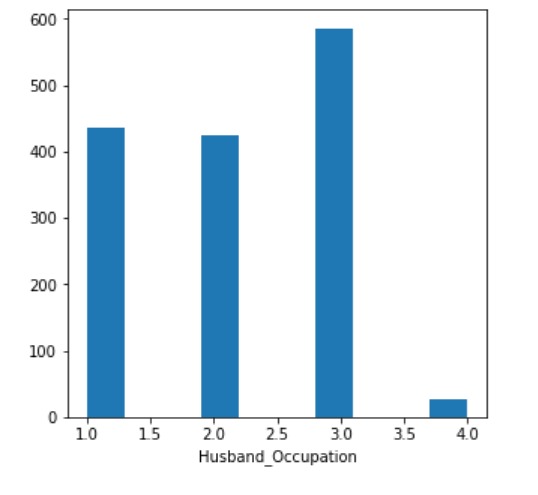


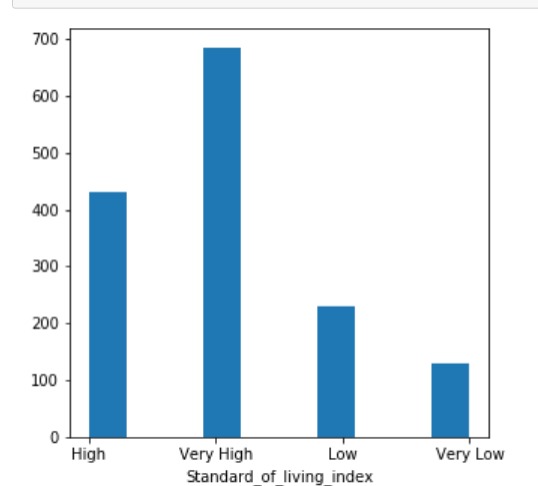
* As mentioned, Tertiary is the most frequent education level of both Husband and Wife.
* Fewer Husbands are uneducated as compared to the wives.



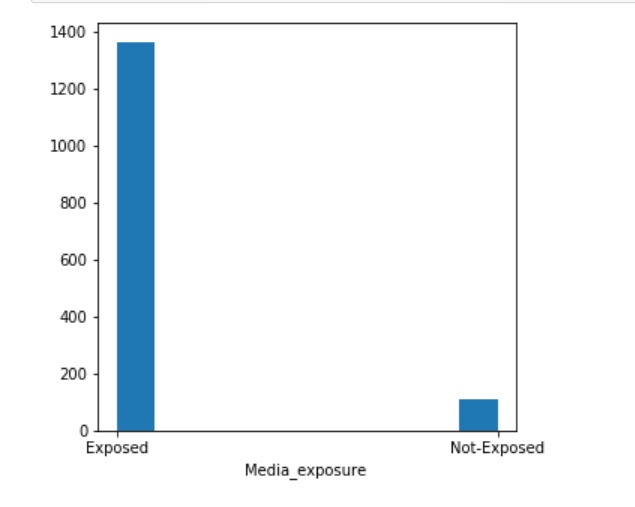


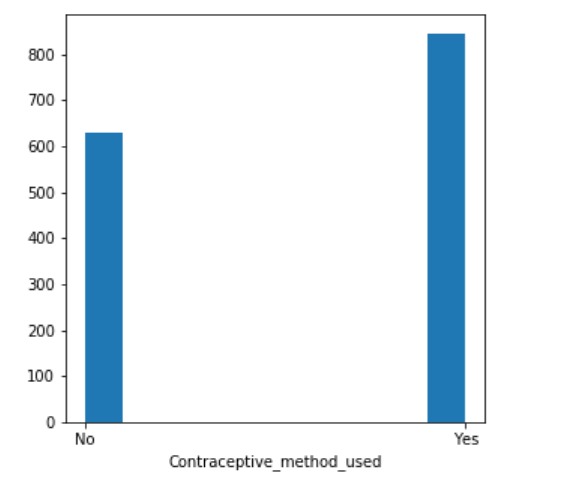
Also, Scientology is followed the most



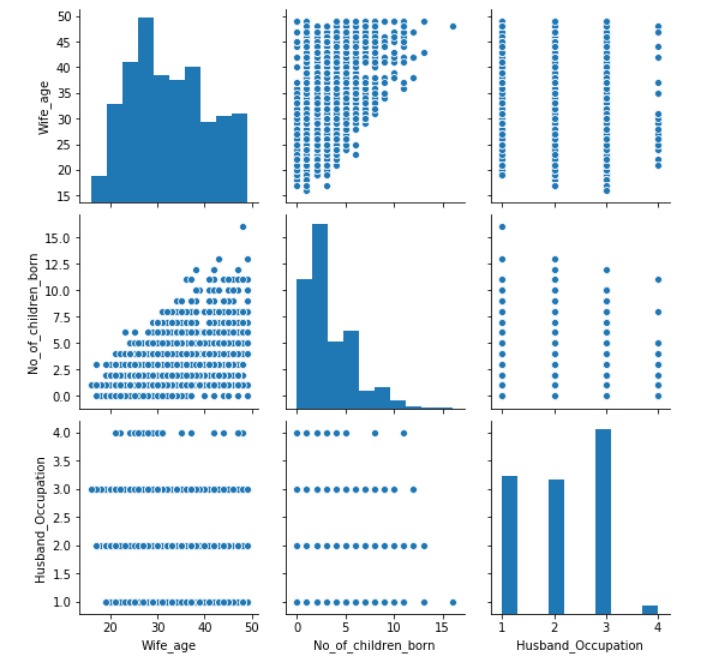


* Major portion of the people are from the areas where the standard of living is Very High and High.
* In total around 350 people are from the areas with Low and Very low standard of living index.





We already know that the majority of the women have used a contraceptive method, however there is a good proportion as well who have not used any



* Major portion of the people are from the areas where the standard of living is Very High and High.
* In total around 350 people are from the areas with Low and Very low standard of living index.

### **2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis)**

* Since the data has string & categorical type variables, these variables must be encoded so that the Machine Learning model understands the data.
* In the target variable, "No" is replaced by 0 and "Yes" is replaced by 1 first. girishchadha.gc@gmail.com JV65UCK2AH This file is meant for personal use by girishchadha.gc@gmail.com only. Sharing or publishing the contents in part or full is liable for legal action. 25
* Similarly, ordinal numbers are given to the values in variables Wife\_ education, Husband\_education & Standard\_of\_living\_index.
* After this dummy encoding us used to encode the data for the rest of the columns.
* The dataset looks like this.