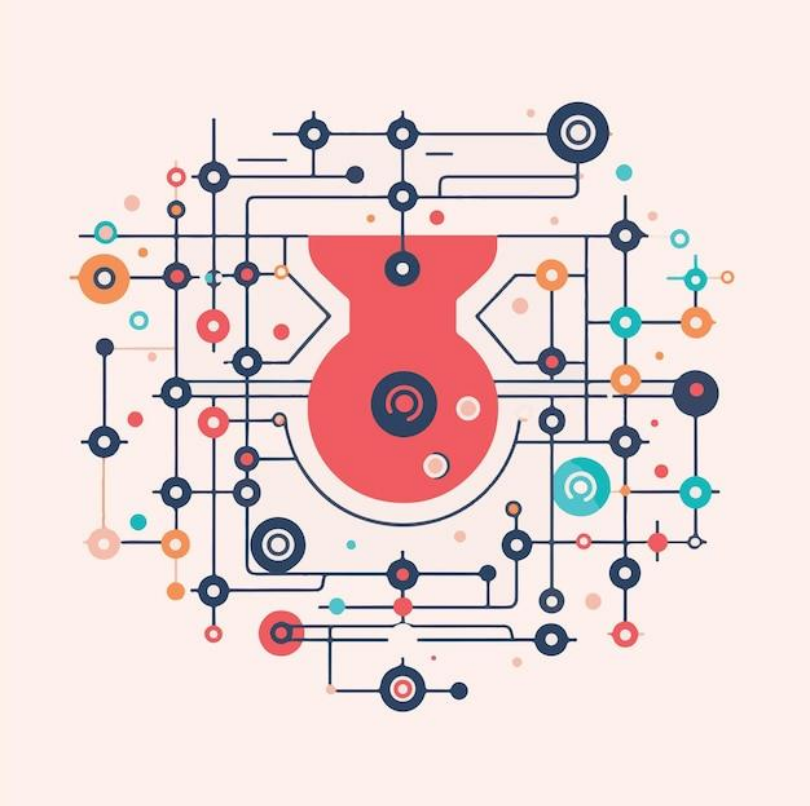

PREDICTIVE MODELLING PROJECT

By Kurt Warren Mario Gilby On June 09th 2024

Submitted to

Great Learning
POWER AHEAD

As a part of the requirements for completion of PGP-DSBA offered in affiliation with

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Table of Contents

List of Figures

List of Tables

List of Equations

No table of figures entries found.

Introduction

While doing the **"Predictive Modelling course"**, we have seen and practised some tools and techniques that fall under the supervised learning criteria, which means that we look at all the independent and dependant features, and try and find mathematical relationships between them.

We do this by generating and selecting the most important features which minimizes the mean squared error in the data "Linear and Logistic Regression", Or/And Create Decision Trees which would help us classify/predict the value of variable based on threshold values or one/more independent variables "Decision Trees Classifier/Regressor" and use these methods to predict "Y" given a set of values of "X".

I will in the following pages use the said tools and techniques to review the problem sets given and answer the questions posed. In the following pages of this document, I have given the following sections:

- **Overview**: High level overview of the problem statement/case study
- **Dataset Description**: Definition of the dataset provided
- **Objective**: detail list of the steps that will be taken when answering the questions asked.
- **Questions Asked**: List of all the questions asked with their answers and supporting material like Figures Tables etc.

The first problem set here showcases two regression techniques "Linear Regression" and "Decision Tree Regressor". While the second problem set looks at the dataset given and showcases the "Logistic Regression Classifier" and "Decision Tree Classifier".

Along with using and showcasing the above I will also showcase statistical techniques and good practices learnt in pervious courses, such as Exploratory Data Analysis (EDA), and Data Preprocessing.

Problem 1

Overview

We are given data from a workstation that is operating in a multi-user university department, the data has activity measures of the computers systems, Users use the workstation for tasks like internet access, file editing and CPU-intensive Programs, using the data set given with the system measures, we need to find a linear equation which can best predict the "Proportion of time that the CPU runs in user mode, given the input parameters that are made up of the activates listed in the data.

Objective

Using the data provided will perform the following steps:

1. Define the problem
2. Explore the data
3. Get the statistical summary of the data.
4. Perform data preprocessing
5. Perform Linear Regression
6. Perform Decision Tree Regression
7. Compare the Models
8. Derive Actionable Insights and Recommendations

Dataset Description

This is the Definition of the data provided in the below table:

Sl. No	Column Name	Column Description
1	**lread**	Reads (transfers per second) between system memory and user memory
2	**lwrite**	writes (transfers per second) between system memory and user memory
3	scall	Number of systems calls of all types per second
4	sread	Number of systems read calls per second
5	smriti	Number of systems write calls per second
6	fork	Number of system fork calls per second
7	exec	Number of system exec calls per second
8	rchar	Number of characters transferred per second by system read calls
9	wchar	Number of characters transferred per second by system write calls
10	pgout	Number of pages out requests per second
11	ppgout	Number of pages, paged out per second
12	pgfree	Number of pages per second placed on the free list
13	pgscan	Number of pages checked if they can be freed per second
14	atch	Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second
15	pgin	Number of page-in requests per second
16	ppgin	Number of pages paged in per second
17	pflt	Number of page faults caused by protection errors (copy-on-writes)
18	vflt	Number of page faults caused by address translation
19	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.)
20	freemem	Number of memory pages available to user processes
21	freeswap	Number of disk blocks available for page swapping
22	usr	Portion of time (%) that CPUs run in user mode

Table 1 Workstation Activities Data Definitions

Questions Asked

Define the problem and perform Exploratory Data Analysis.

Problem definition

We are given data from a workstation that is operating in a multi-user university department, the data has activity measures of the computers systems, Users use the workstation for tasks like internet access, file editing and CPU-intensive Programs, using the data set given with the system measures, we need to find a linear equation which can best predict the "Proportion of time that the CPU runs in user mode, given the input parameters that are made up of the activities listed in the data.

To do this we will apply linear regression models to the data set to find the best fit line which gives us the linear equation of the independent features to the dependant features which in this case is "usr".

Check Shape, Data Types, Statistical Summary

- The Data has 8192 observations and 21 independent features/attributes and 1 dependant attribute
- The Data has 13 features with data type : float64, 8 features with data type : int6, 1 feature with data type : object.
- "runqsz" is of data type "object", we need to check this for values counts and convert to int or float.
- "Null" values are seen in "rchar" and "wchar", we need to check this and treat accordingly
- Statistical Summary:

Attribute	Statistics	Observations
lread	count: 8192, mean: 19.56, stdev: 53.354, min: 0, 25th_Per: 2.0, 50th_Per: 7.0, 75th_Per: 20.0, max: 1845	no missing values-potential outliers-high variance
lwrite	count: 8192, mean: 13.106, stdev: 29.892, min: 0, 25th_Per: 0.0, 50th_Per: 1.0, 75th_Per: 10.0, max: 575	no missing values-potential outliers-high variance-25th Percentile have 0 value.
scall	count: 8192, mean: 2306.318, stdev: 1633.617, min: 109, 25th_Per: 1012.0, 50th_Per: 2051.5, 75th_Per: 3317.25, max: 12493	no missing values-potential outliers-high variance.
sread	count: 8192, mean: 210.48, stdev: 198.98, min: 6, 25th_Per: 86.0, 50th_Per: 166.0, 75th_Per: 279.0, max: 5318	no missing values-potential outliers-high variance.
swrite	count: 8192, mean: 150.058, stdev: 160.479, min: 7, 25th_Per: 63.0, 50th_Per: 117.0, 75th_Per: 185.0, max: 5456	no missing values-potential outliers-high variance.
fork	count: 8192, mean: 1.885, stdev: 2.479, min: 0.0, 25th_Per: 0.4, 50th_Per: 0.8, 75th_Per: 2.2, max: 20.12	no missing values-potential outliers-high variance.
exec	count: 8192, mean: 2.792, stdev: 5.212, min: 0.0, 25th_Per: 0.2, 50th_Per: 1.2, 75th_Per: 2.8, max: 59.56	no missing values-potential outliers-high variance.
rchar	count: 8088, mean: 197385.728, stdev: 239837.494, min: 278.0, 25th_Per: 34091.5, 50th_Per: 125473.5, 75th_Per: 267828.75, max: 2526649.0	missing values-potential outliers-high variance.
wchar	count: 8177, mean: 95902.993, stdev: 140841.708, min: 1498.0, 25th_Per: 22916.0, 50th_Per: 46619.0, 75th_Per: 106101.0, max: 1801623.0	missing values-potential outliers-low variance.
pgout	count: 8192, mean: 2.285, stdev: 5.307, min: 0.0, 25th_Per: 0.0, 50th_Per: 0.0, 75th_Per: 2.4, max: 81.44	no missing values-potential outliers-high variance-50 percent records with zero
ppgout	count: 8192, mean: 5.977, stdev: 15.215, min: 0.0, 25th_Per: 0.0, 50th_Per: 0.0, 75th_Per: 4.2, max: 184.2	no missing values-potential outliers-high variance-50 percent records with zero
pgfree	count: 8192, mean: 11.92, stdev: 32.364, min: 0.0, 25th_Per: 0.0, 50th_Per: 0.0, 75th_Per: 5.0, max: 523.0	no missing values-potential outliers-high variance-50 percent records with zero

pgscan	count: 8192, mean: 21.527, stdev: 71.141, min: 0.0, 25th_Per: 0.0, 50th_Per: 0.0, 75th_Per: 0.0, max: 1237.0	no missing values-potential outliers-high variance-75 percent records with zero
atch	count: 8192, mean: 1.128, stdev: 5.708, min: 0.0, 25th_Per: 0.0, 50th_Per: 0.0, 75th_Per: 0.6, max: 211.58	no missing values-potential outliers-high variance-50 percent records with zero
pgin	count: 8192, mean: 8.278, stdev: 13.875, min: 0.0, 25th_Per: 0.6, 50th_Per: 2.8, 75th_Per: 9.765, max: 141.2	no missing values-potential outliers-high variance
ppgin	count: 8192, mean: 12.389, stdev: 22.281, min: 0.0, 25th_Per: 0.6, 50th_Per: 3.8, 75th_Per: 13.8, max: 292.61	no missing values-potential outliers-high variance
pflt	count: 8192, mean: 109.794, stdev: 114.419, min: 0.0, 25th_Per: 25.0, 50th_Per: 63.8, 75th_Per: 159.6, max: 899.8	no missing values-potential outliers-high variance
vflt	count: 8192, mean: 185.316, stdev: 191.001, min: 0.2, 25th_Per: 45.4, 50th_Per: 120.4, 75th_Per: 251.8, max: 1365.0	no missing values-potential outliers-high variance
freemem	count: 8192, mean: 1763.456, stdev: 2482.105, min: 55, 25th_Per: 231.0, 50th_Per: 579.0, 75th_Per: 2002.25, max: 12027	no missing values-potential outliers-high variance
freeswap	count: 8192, mean: 1328125.96, stdev: 422019.427, min: 2, 25th_Per: 1042623.5, 50th_Per: 1289289.5, 75th_Per: 1730379.5, max: 2243187	no missing values-low variance
usr	count: 8192, mean: 83.969, stdev: 18.402, min: 0, 25th_Per: 81.0, 50th_Per: 89.0, 75th_Per: 94.0, max: 99	no missing values-low variance

Table 2 Workstations Activity data statistical summary

Exploratory Data Analysis (Univariate and Bivariate)

Univariate Analysis

We perform **"Univariate"** Analysis on the all the numerical variables by plotting boxplots and histograms of these variables and these are the findings.

Boxplots observations

- The features are on different scales, but the scale for rchar, wchar and freeswap is very different, plot these separate than the others.
- Also, the features scall, sread, swrite, pgfree, pgscan, pflt, vflt and freemem are on a medium scale, plot these separate.
- all the features have outliers
- all the feature are skewed

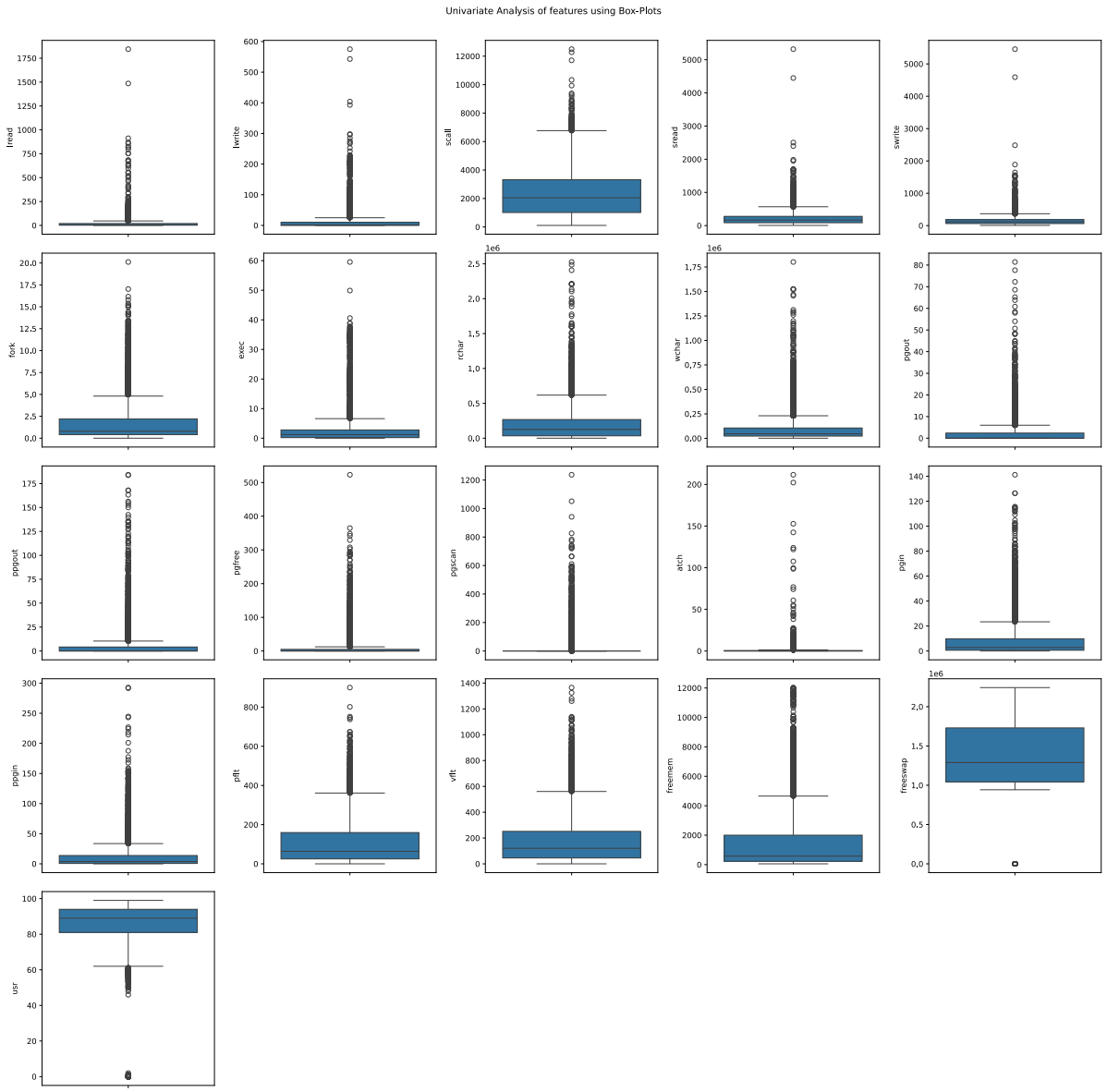

Figure 1 Workstation Data Numerical Features Distribution

KDE Plots Observations

- lread follows bell curve, a lot of small values, small IQR, long right tail.
- lwrite follows bell curve, a lot of small values, small IQR, long right tail.
- scall follows bell curve- two peaks seen, most values between 0 and 7500, spread IQR, long right tail.
- sread follows bell curve, a lot of small values, small IQR, long right tail.
- swrite follows bell curve, a lot of small values, small IQR, long right tail.
- fork follows bell curve, most values between 0 and 10, spread IQR, long right tail.
- exec follows bell curve-small bumps around values 20 and 35, a lot of small values, small IQR, long right tail.
- rchar follows bell curve, most values between 0.0 and 1.0, spread IQR, long right tail.
- wchar follows bell curve, most values between 0.0 and 0.5, spread IQR, long right tail.
- pgout follows bell curve, a lot of small values, small IQR, long right tail.
- ppgout follows bell curve, a lot of small values, small IQR, long right tail.
- pgfree follows bell curve, a lot of small values, small IQR, long right tail.
- pgscan follows bell curve, a lot of small values, small IQR, long right tail.
- atch follows bell curve, a lot of small values, small IQR, long right tail.
- pgin follows bell curve, most values between 0 and 50, spread IQR, long right tail.
- ppgin follows bell curve, most values between 0 and 50, spread IQR, long right tail.
- pflt follows bell curve, most values between 0 and 400, spread IQR, long right tail.
- vflt follows bell curve, most values between 0 and 750, spread IQR, long right tail.
- freemem follows bell curve-small bump around 7500, most values between 0 and 7500, spread IQR, long right tail.
- free swap follows bell curve-bumps/peaks around [0.0,1.0,2.0], most values between 1.0 and 2,0, spread IQR, long left tail.
- usr follows bell curve-bumps/peaks around [0,100], most values between 60 and 100, spread IQR, long left tail.

distribution of runqsz

- The proportion for observations for "Not-CPU-Bound" to "CPU-Bound" is ~53% to 47%.

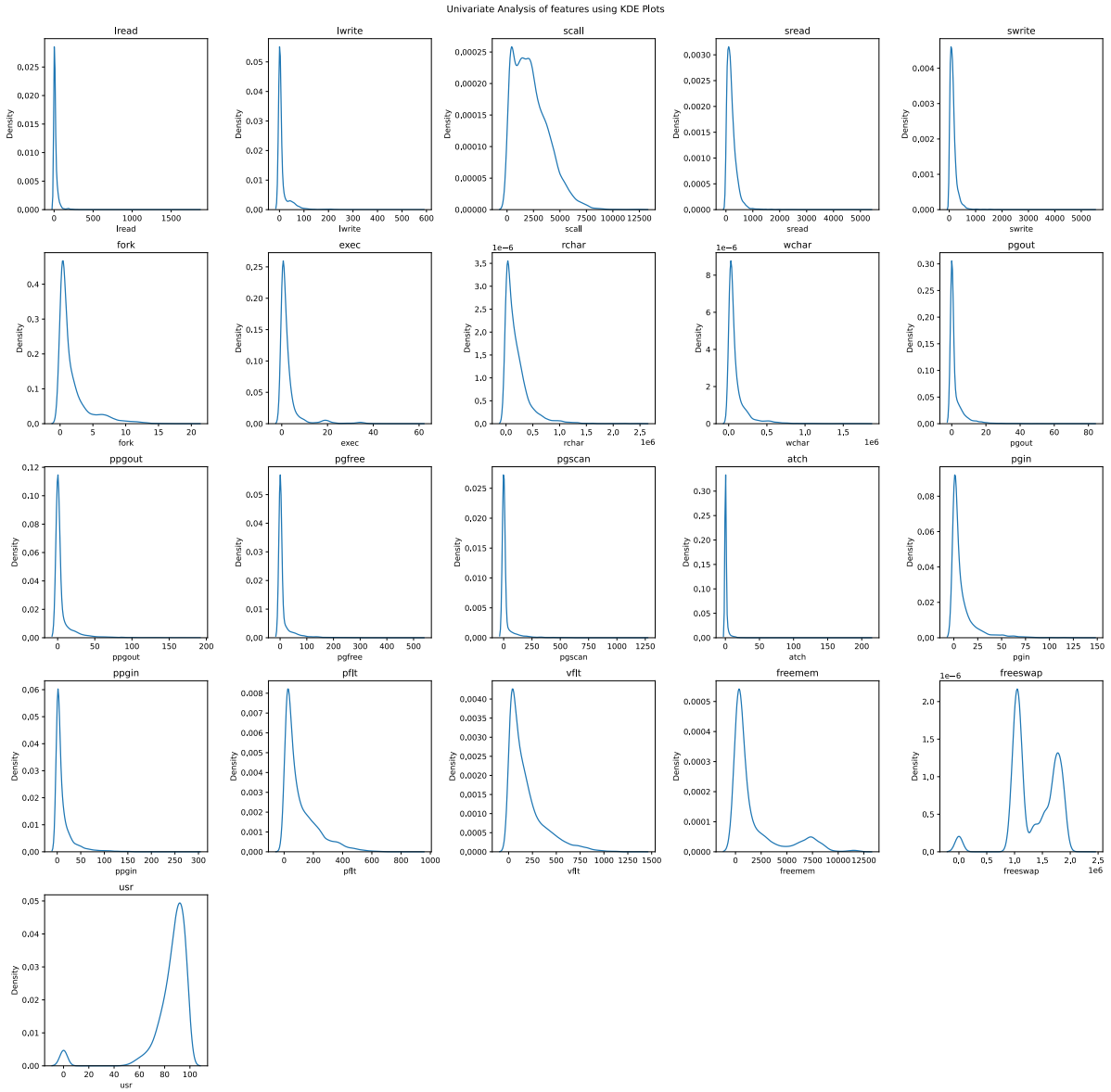

Figure 2 Workstation Activities Data Numerical Features KDE Plots

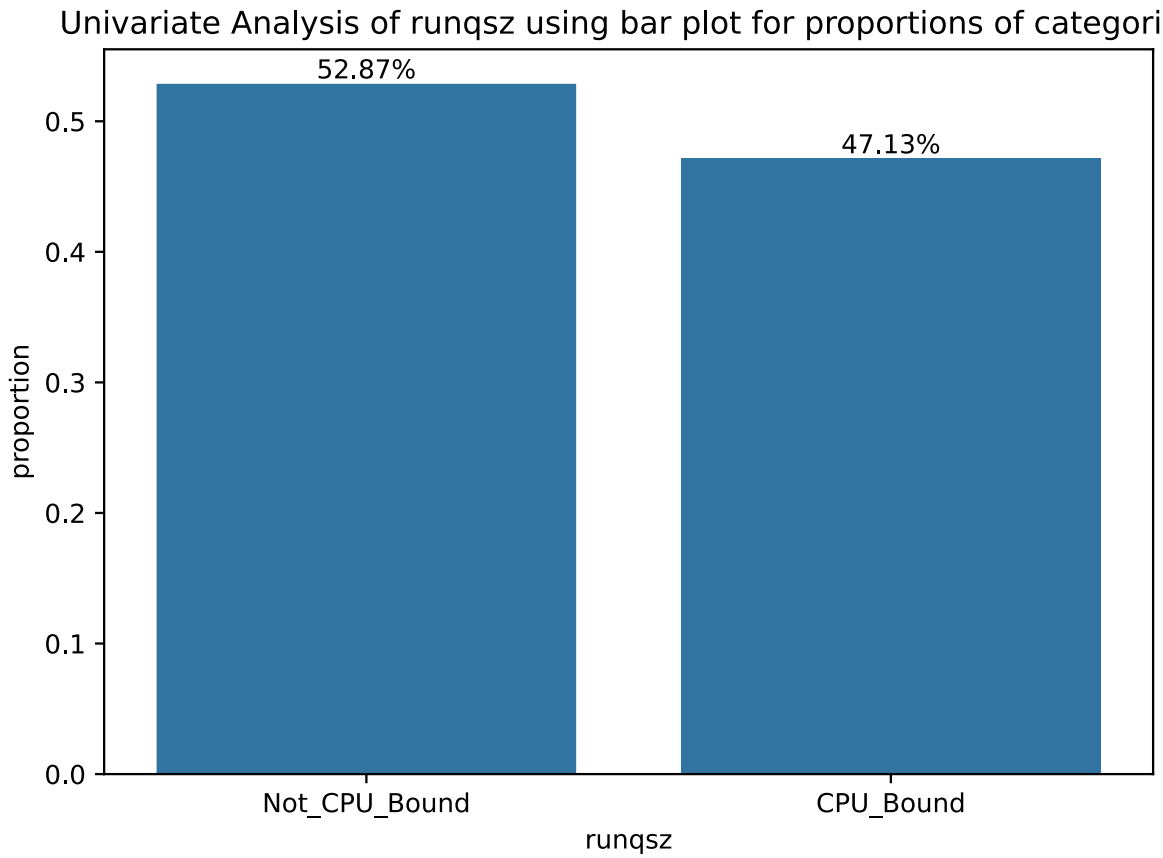

Figure 3 Proportion of runqsz

Multivariate Analysis

Using a **boxplot, KDE plot split by runqsz, pair plot** and **heatmap** for all the numerical variables.

boxplots split on runqsz

- There is difference seen in the distribution between for "Not-CPU-Bound" vs "CPU-Bound" in the following features
    - scall, sread, fork, exec, rchar, wchar, pgin, ppgin, pflt, vflt, freemem, freeswap and usr

KDE Plots split on runqsz

- There is significant difference seen in the distribution between for "Not-CPU-Bound" vs "CPU-Bound" in the following features
    - scall, rchar, freemem, freeswap, usr

Pair Plots

- Reviewing the pair plot we see there are potential correlations between
    - lread and lwrite
    - scall and sread and swrite
    - swrite and rchar and wchar
    - fork and exec and pflt and vflt
    - exec and pflt and vflt
    - rchar and wchar and atch
    - pgout and ppgout and pgfree
    - pgfree and pgscan
    - pgin and ppgin
    - usr and fork and pflt and vflt

heatmap

- Reviewing the heatmap plot we see their strong correlations between
    - lread and lwrite
    - scall and sread and swrite and fork and pflt and vflt
    - sread and swrite and rchar and pflt and vflt
    - fork and scall and exec and pflt and vflt
    - exec and fork pflt and vflt
    - rchar and sread and wchar
    - wchar and rchar
    - pgout and ppgout and pgfree and pgscan
    - ppgout and pgout and pgfree and pgscan and pgin and ppgin
    - pgfree and pgout and ppgout and pgscan and pgin and ppgin
    - pgscan and pgout and ppgout and pgfree and pgin and ppgin
    - pgin and ppgout and pgfree and pgscan and ppgin
    - ppgin and ppgout and pgfree and pgscan and pgin
    - pflt and scall and sread and fork and exec and vflt
    - vflt and scall and sread and fork and exec and pflt
    - freemem and freeswap
    - freeswap and freemem and usr
    - usr and freeswap

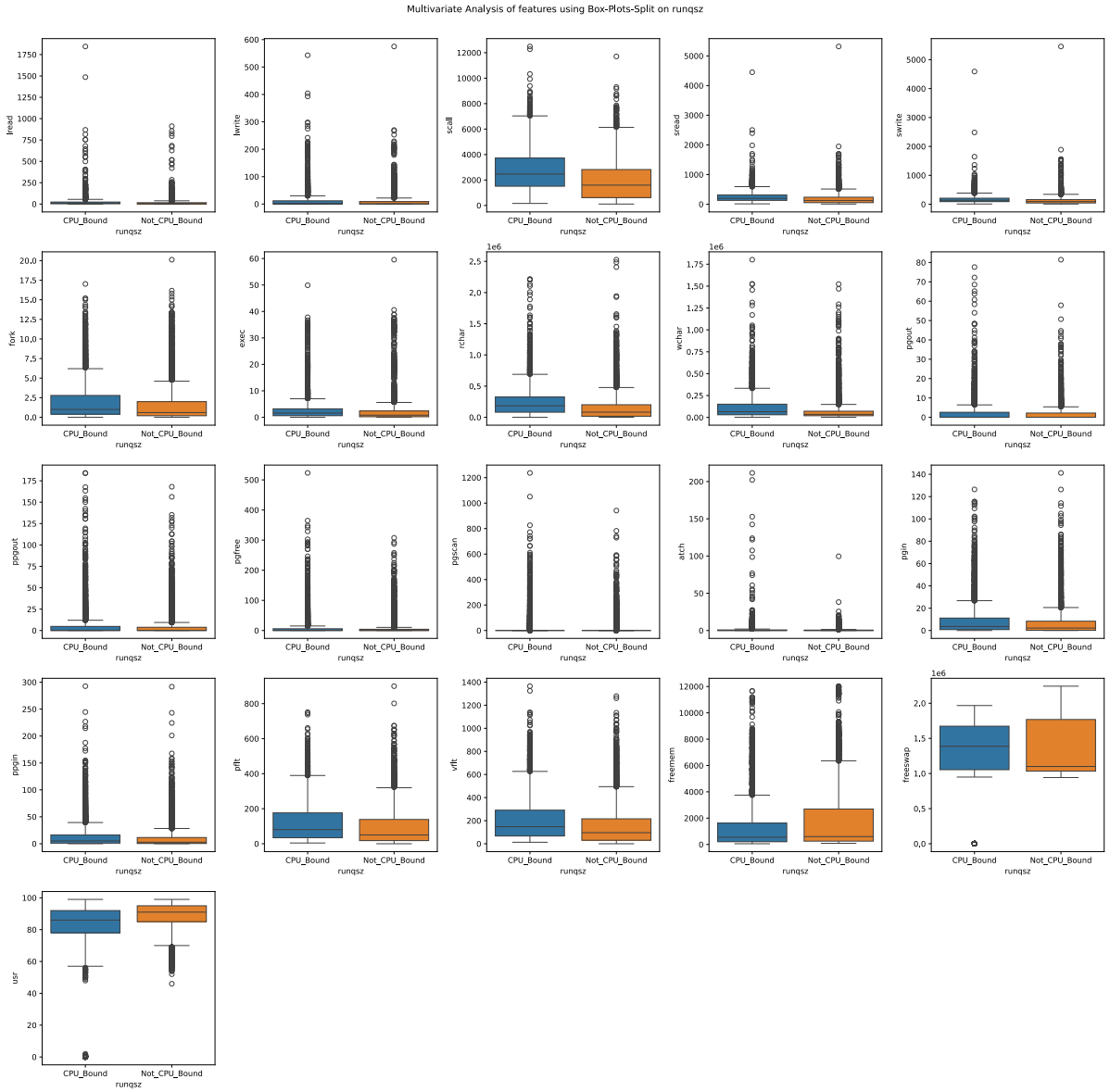

Figure 4 Workstation data Numerical Features Box Plot Split by "runqsz"

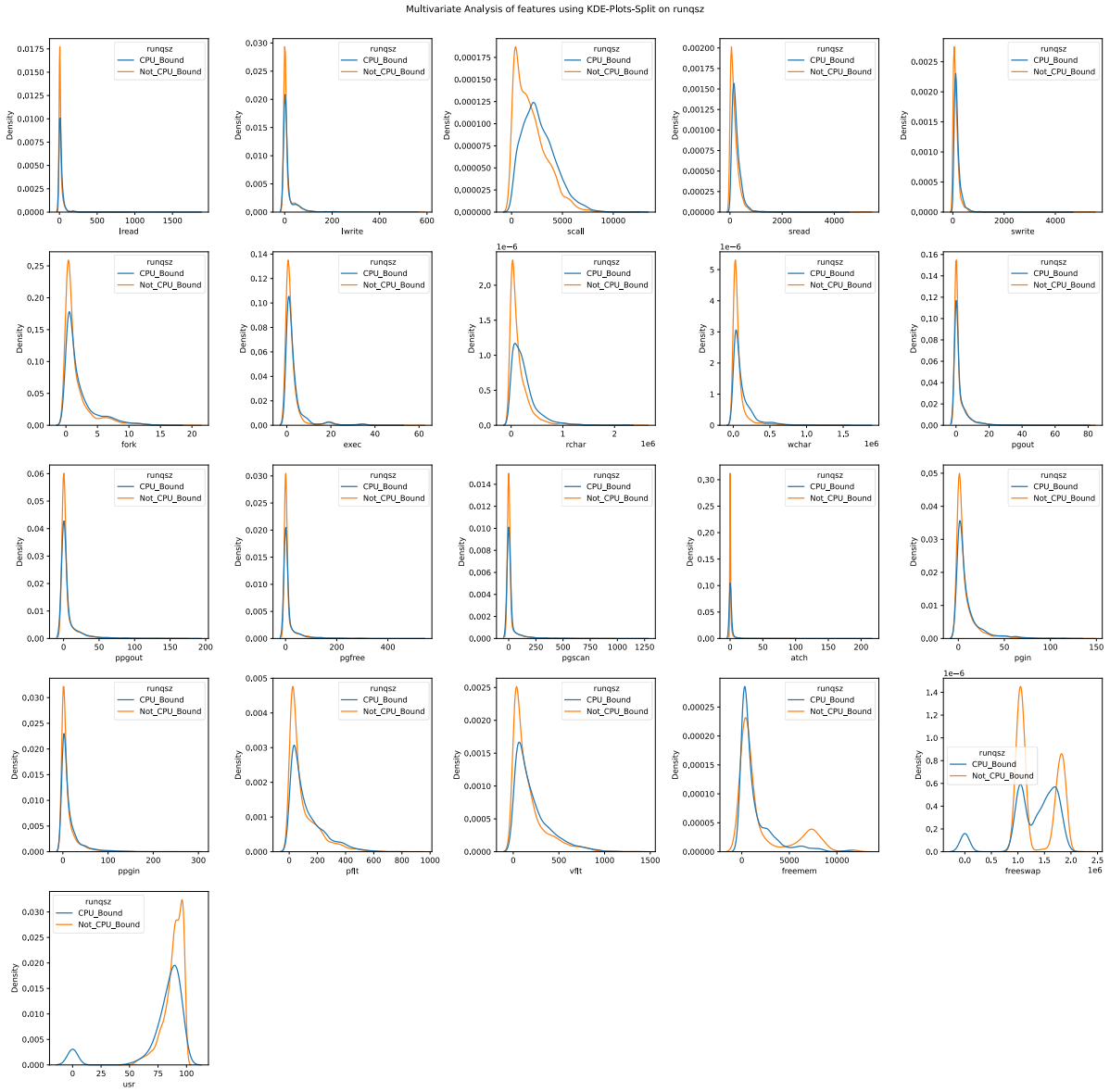

Figure 5 Workstation Activity Data Numerical Features KDE Plots Split by "runqsz"

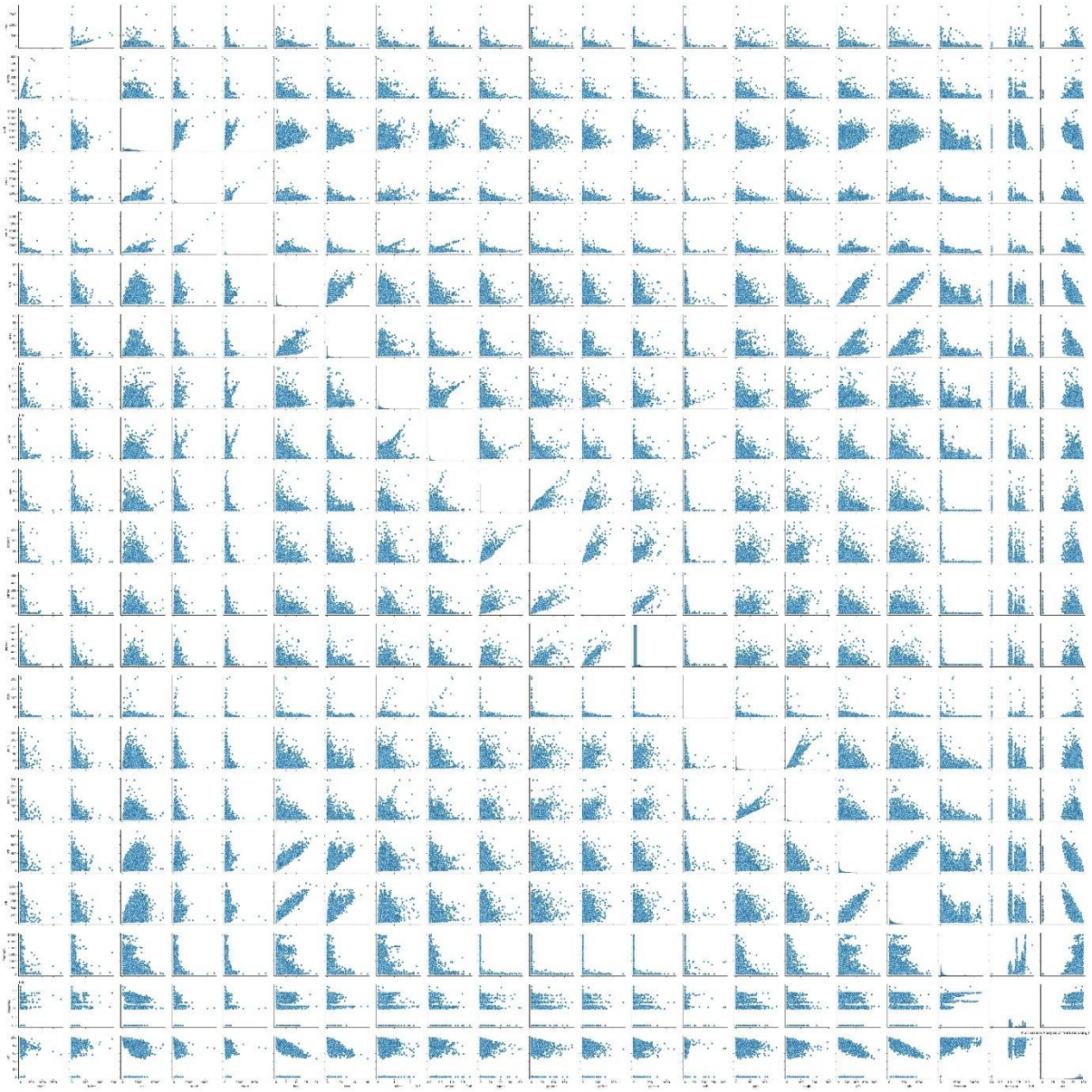

Figure 6 Workstation Activity Data Numerical Features Pair Plot

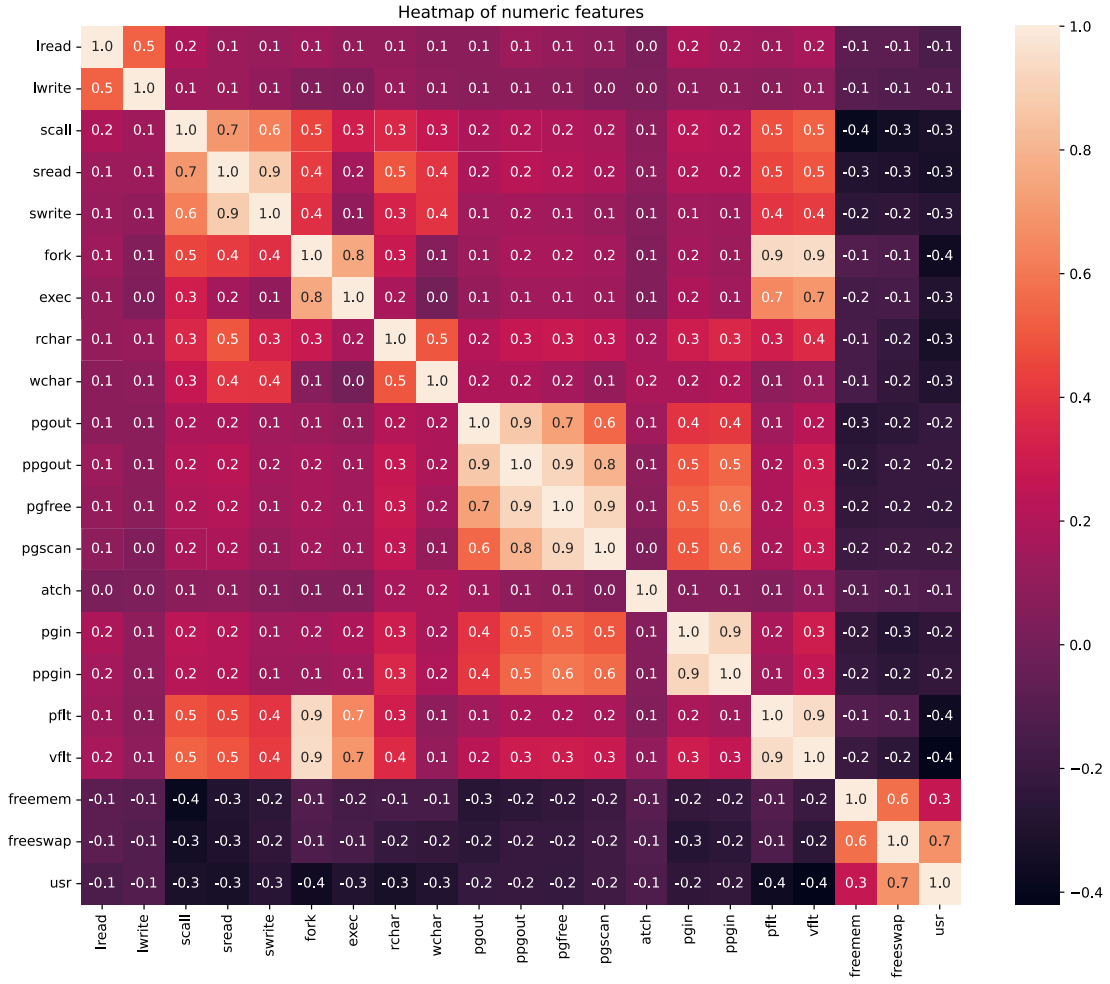

Figure 7 Workstation Data Numerical Features Correlation Matrix Heat Map

Key Observations

- There are Null values that need to be treated
- The Scaling of the Data is not consistent and we would need to do scaling.
- The all of the numeric features have a bell curve distribution with outliers.
- skewness is present in all the numerical features; we should try and reduce the skewness.
- There are few that have multiple peaks, this is due to distribution difference due to "Not-CPU-Bound" to "CPU-Bound" states in the "runqsz", we need to take runqsz when running the model.
- There is high correlation in the dataset -.5 or .5 and above between multiple features, we would use techniques like "VIF" to reduce the impact of this.

Data Preprocessing

Prepare the data for modelling:

Missing Values Treatment:

- Values are missing in "rchar" and "wchar", since these states theatrically possible but highly unlikely we will replace these values with the minimum values in the series.
- Check for duplicates.

Outlier Detection (treat, if needed):

- Outlier values are present but these are values that genuinely captured from the system, we will not do any Outlier Treatments.

Encoding:

- "runqsz" is Label encoded, and treated as a categorical variable i.e., we will not do any scaling or Transformation with this column.
- This is Encoded as {'CPU_Bound': 1, 'Not_CPU_Bound': 0}

Feature Engineering:

- ***These steps will be run post splitting the data into "Train" and "Test" this is to ensure that the Transformation, Scaling of the Train data is not influenced by the Test data.***
- All the features are highly skewed, we check transformations to check if reduce a lot of this skewness, exploring various transformations:
- The Boxcox transformation would be done to the following features
  - lread, scall, sread, swrite, fork, exec, rchar, wchar, pgout, ppgout, pgfree, pgscan, atch, pflt, vflt, freemem and freeswap
- The Cube root transformation would be done to the following features
  - lwrite, pgin and ppgin
- We will not transform the target variable of "usr"
- We will scale the data to get all the attribute to the same scale.
- Since all of the features follow a bell curve, we will scale using zscore method.

Train-Test Split:

- We do a split of the data 70-30 Train-Test.
- Apply the transformation and scaling (described in Feature Engineering) to the Train and Test post the Split.
- Post the application using the heat map on the Train data we see:
  - a lot features have values greater than and equal to 0.5 and less than and equal to -0.5
  - this indicates that moderate and strong correlations between features

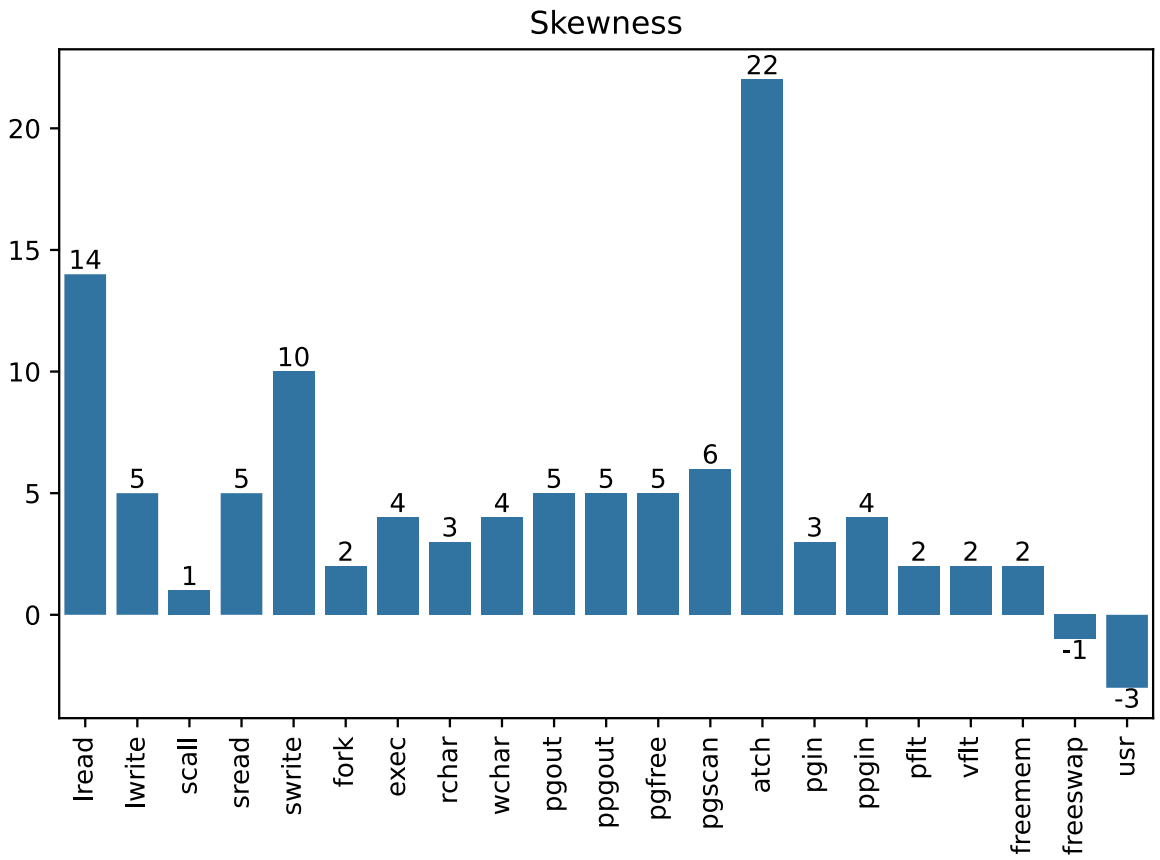

Figure 8 Skewness of Numerical Features of Workstation Activity Data

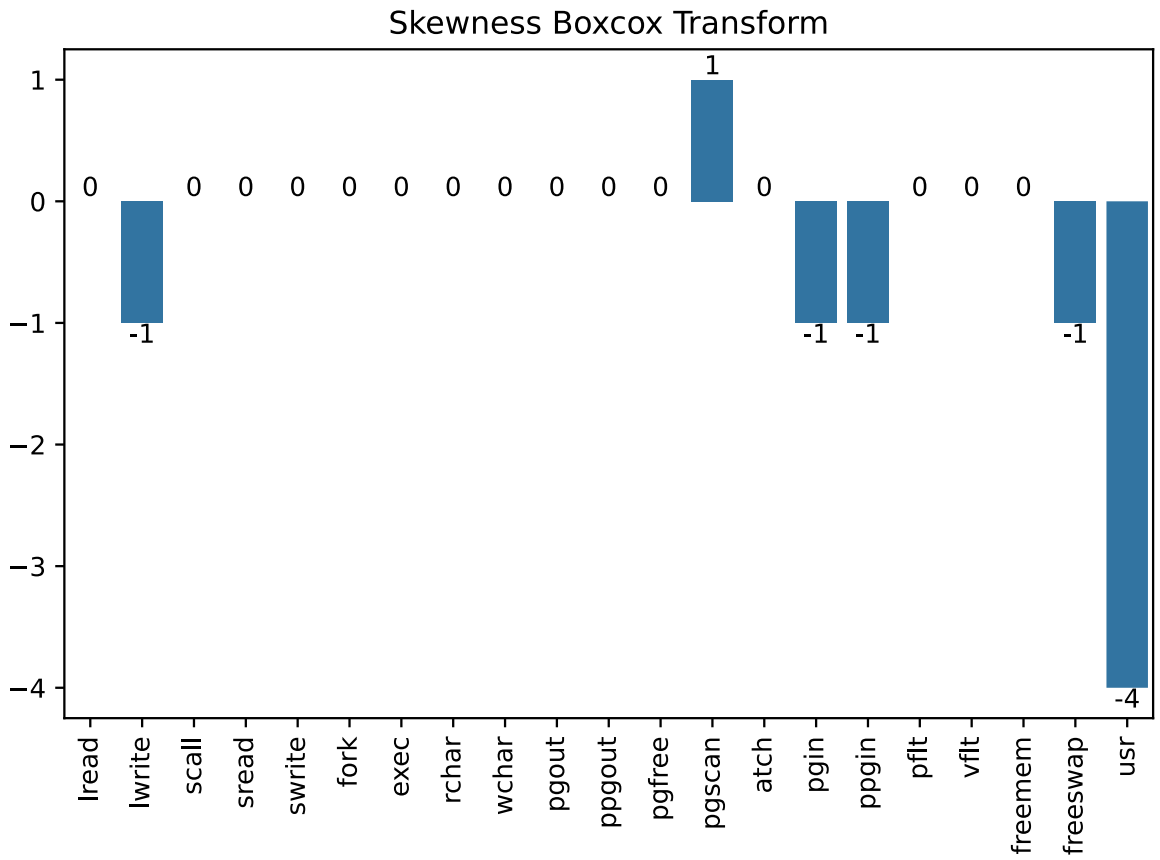

Figure 9 Skewness post Box-Cox Transformation

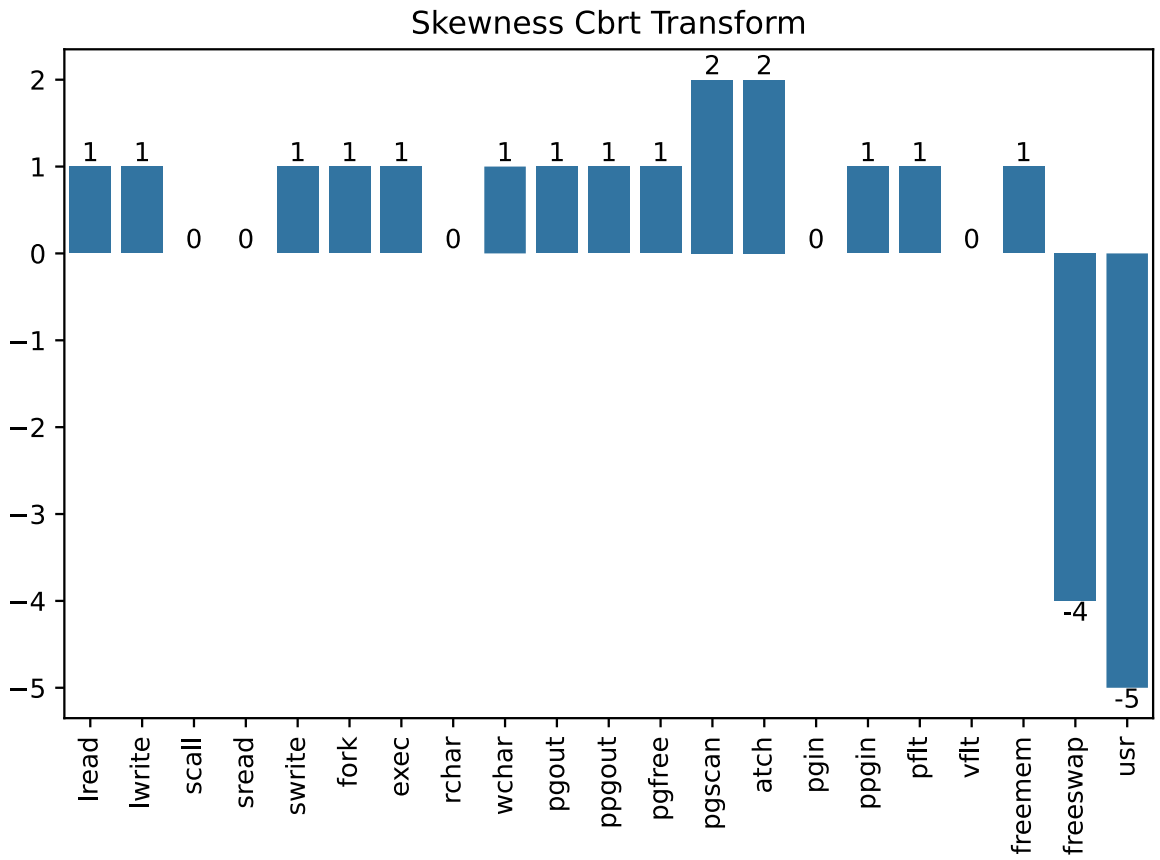

Figure 10 Skewness post cube root transformation

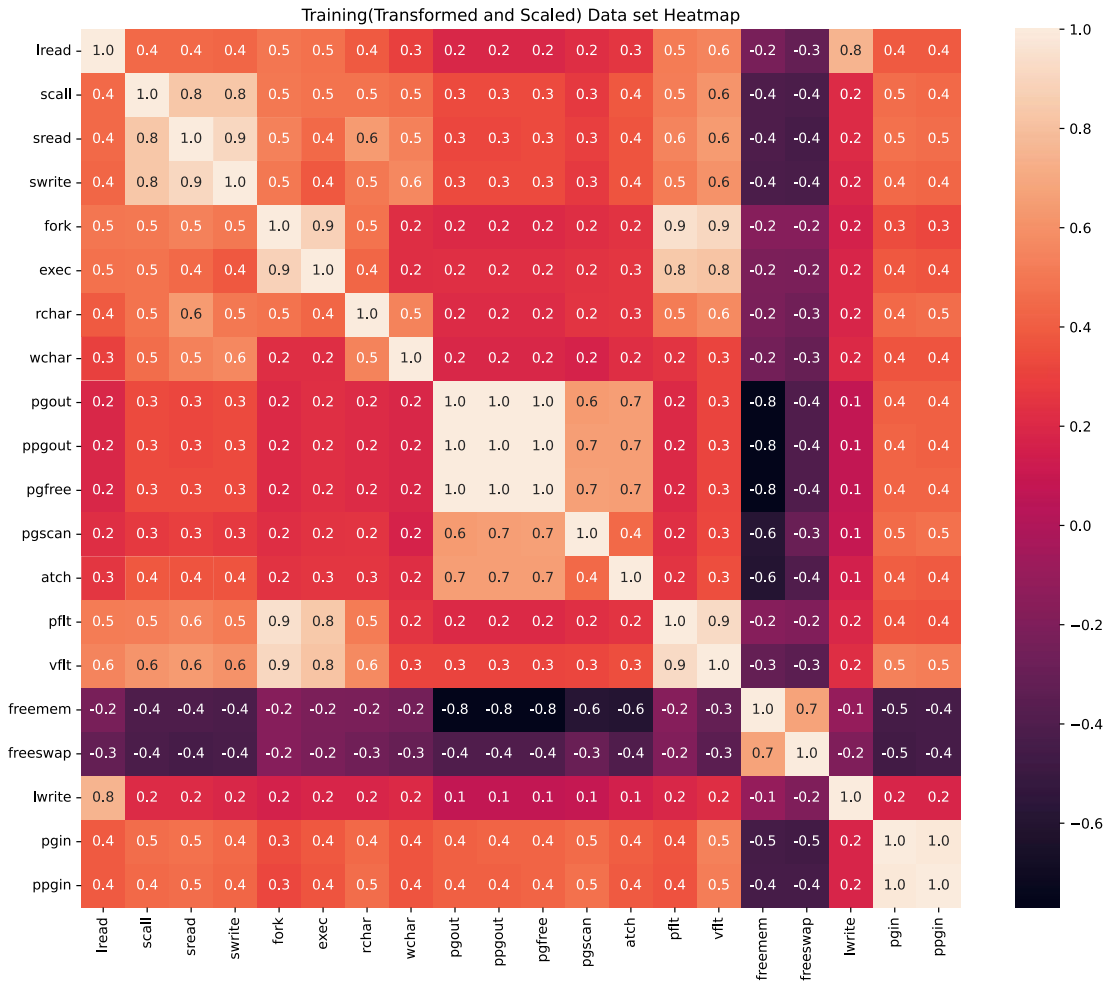

Figure 11 Heat Map of Correlation Matrix of Training data transformed and scaled

Model Building - Linear regression:

Apply linear Regression using Sklearn:

- We run Linear Regression using sklearn.linear_model
- Fit the same to the transformed and scaled training data.
- running the model with all the features and using Sklearn we get
  - The MSE in training is ~124 and in test it is 133, there is more error in test.
  - By the R2 we see that in the training the model explains ~62.7% of the variance in training and ~62.4% in testing.

Using Statsmodels Perform checks for significant variables using the appropriate method:

- We would check to run VIF as we know that there is multicollinearity so we would like to run this to get and remove features with Vif more than 5.
- Using Statsmodels we would like to run the model see what are the results and review the summary to get the p value of the features low p values means significant.
- Results with all features:
  - MSE: 124.11929490749988
  - R-squared: 0.627402382682

- ppgout has the highest vif (5951.473458) and high p>|t| value (0.808) dropping the same and rerunning, vif and Statsmodels
- pgout has the high vif (350.108379) and high p>|t| value (0.056) dropping the same and rerunning, vif and Statsmodels
- swrite has a high p>|t| value (0.280) dropping the same and rerunning, vif and Statsmodels
- pgin has a high vif (27) dropping the same and rerunning, vif and Statsmodels
- ppgin has a high p>|t| value (0.07) dropping the same and rerunning, vif and Statsmodels
- exec has a high p>|t| value (0.658) dropping the same and rerunning, vif and Statsmodels
- pgscan has a high p>|t| value (0.481) dropping the same and rerunning, vif and Statsmodels
- wchar has a high p>|t| value (0.157) dropping the same and rerunning, vif and Statsmodels
- Results after removing the features on the test data:
    - Test MSE: 134.73827114064625
    - Test R-squared: 0.6197317031796956
- ***Even though the "R-squared" decreases with the removal of the features, it is preferred as it removes multicollinearity and non-significant features***

Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare.:

- We would run the multiple models after dropping the features due to multicollinearity and non-significance:
    - ['ppgout','pgout','swrite','pgin','ppgin','exec','pgscan','wchar']
- checking 2 models using Decision Tree Regressor and Linear Regression and comparing the RMSE to find the best model
- The "Decision Tree Regressor" is over fitted we need to prune the Decision Tree Regressor
-

Model	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	11.207289	11.607682	0.621495	0.619732
Decision Tree Regressor	0.000000	3.884469	1.000000	0.957414

-

Table 3 Workstation Data Linear Regression and Decision Tree Regressor RMSE and R2 Score compare.

- After pruning the Decision Tree Regressor, we get a model which is working well on the Training and Test.

Model	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	11.207289	11.607682	0.621495	0.619732
Decision Tree Regressor	2.952173	3.238304	0.973736	0.970404

Table 4 Workstation Linear Regression and Decision Tree Regressor RMSE and R2 Score compare after pruning Decision Tree Regressor

Business Insights & Recommendations

Comment on the Linear Regression equation from the final model and impact of relevant variables (at least 2) as per the equation:

- ***Linear Equation: y = 0.0*const + 0.674*lread + 1.86*scall + 1.052*sread + -5.876*fork + -2.281*rchar + 1.254*pgfree + 1.086*atch + -3.383*pflt + 4.882*vflt + 1.802*freemem + 12.987*freeswap + -1.004*lwrite + -6.91*runqsz + 87.252***
- ***coefficients sorted absolute : {'freeswap': 12.987, 'runqsz': -6.91, 'fork': -5.876, 'vflt': 4.882, 'pflt': -3.383, 'rchar': -2.281, 'scall': 1.86, 'freemem': 1.802, 'pgfree': 1.254, 'atch': 1.086, 'sread': 1.052, 'lwrite': -1.004, 'lread': 0.674, 'const': 0.0}***
- As per the equation freeswap has most impact on usr and runqsz has the most inverse impact on usr
- As per the equation lread has least impact on usr and lwrite has the least inverse impact on usr

Conclude with the key takeaways (actionable insights and recommendations) for the business:

- if we run the "Linear Regression" in its current state it is not very useful, in order to strengthen the model, ***we would like to perform tools Like PCA for further feature engineering to try and get a more robust "Linear Regression"-***
- In the current state we do have the ***option of going with the "Decision Tree Regressor"*** which though may not have a linear relationship with the independent and dependant variables, still seems to a very good fit given the current training and test.
- This model however ***may need to be trained move frequently on live data and tested in real case scenarios*** to make sure of its viability

Problem 2

Overview

We are given the demographic and socio-economic survey data of married females who were either not pregnant or were uncertain of their pregnancy status during the survey, using logistic regression methods, we need to predict whether these women opt for a contraceptive method or not.

Objective

Using the data provided will perform the following steps:

1. Define the problem
2. Explore the data
3. Get the statistical summary of the data.
4. Perform data preprocessing
5. Perform Logistic Regression
6. Perform Linear Discriminant Analysis
7. Perform Decision Tree Regression
8. Compare the Models
9. Derive Actionable Insights and Recommendations

Dataset Description

Definition of the Data given is as follows:

Name	Description
Wife_age	The age of the wife given as a number
Wife_education	The education of the wife given as a category with values for e.g. as 'Primary', 'Secondary', 'Uneducated' etc
Husband_education	The education of the husband given as a category with values for e.g. as 'Primary', 'Secondary', 'Uneducated' etc
No_of_children_born	The number of children born to the wife as a number
Wife religion	The religion practiced by the wife as a category
Wife_Working	Is the wife working [No, yes]
Husband_Occupation	What is the category of the husband's occupation random values: [1,2,3,4]
Standard_of_living_index	What is the category of standard of living values:['very low','low','high','very high']
Media_exposure	is the wife exposed to media values [Exposed, not Exposed]
Contraceptive_method_used	does the wife use Contraceptive method values [No, yes]

Table 5 Demographic and Socio-economic Survey Data of Married Females

Questions Asked

Define the problem and perform Exploratory Data Analysis

Problem Definition

We are given the demographic and socio-economic survey data of married females who were either not pregnant or were uncertain of their pregnancy status during the survey, using logistic regression methods, we need to predict whether these women opt for a contraceptive method or not.

To do this we will apply logistic regression models to the data set to find the model which is a best classifies the women into the group of "Contraceptive method used": "No" or "Yes".

we will use classification reports/summaries to identify the models that do the best job in accurately classifying these women.

Check Data Shape, Data Types and Statistical Summary.

- The data has 1473 observations, with 9 features and 1 dependant variable
- There are missing values in "Wife_age" and "No_of_children_born"
- There are 3 numerical features and 7 object features
- statistical summary:
    - Wife_age youngest = 16 years and oldest = 49 years, missing values are present
    - Wife_education 4 unique values with "Tertiary" being the greatest number of observations
    - Husband_education 4 unique values with "Tertiary" being the greatest number of observations
    - No_of_children_born minimum = 1 and maximum = 16, missing values are present.
    - Wife_religion 2 unique values with Scientology being the greatest number of observations
    - Wife_Working 2 unique values with "No" being the greatest number of observations
    - Husband_Occupation has 4 values, needs to be converted to categorical currently numeric
    - Standard_of_living_index has 4 values with "Very High" being the greatest number of observations
    - Media_exposure has 2 values with "Exposed" being the greatest number of observations
    - Contraceptive_method_used has 2 values with "Yes" being the greatest number of observations

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	NaN	NaN	NaN	32.606277	8.274927	16.0	26.0	32.0	39.0	49.0
Wife_ education	1473	4	Tertiary	577	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_education	1473	4	Tertiary	899	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_children_born	1452.0	NaN	NaN	NaN	3.254132	2.365212	0.0	1.0	3.0	4.0	16.0
Wife_religion	1473	2	Scientology	1253	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Wife_Working	1473	2	No	1104	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_Occupation	1473.0	NaN	NaN	NaN	2.137814	0.864857	1.0	1.0	2.0	3.0	4.0
Standard_of_living_index	1473	4	Very High	684	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Media_exposure	1473	2	Exposed	1364	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Contraceptive_method_used	1473	2	Yes	844	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 6 Statistical Summary of the Demographic and Socio-economic data

Univariate analysis:

- Husband_Occupation,Wife_education,Husband_education,Wife_religion,Wife_Working,Standard_of_living_index,Media_exposure,Contraceptive_method_used are categorical so will change these to category
- Summary:
  - Categorical features
    - Husband_Occupation: The category "4" has only ~2% of the observations, the rest is close to a ~40%, ~30% and ~29% for 3,1 and 2 respectively.
    - We may want to merge "4" with "3"
    - Wife_education: Uneducated is only ~10% of the observations, , the rest is close to a ~40%, ~28% and ~23% for Tertiary, Secondary and Primary respectively.
    - Husband_education: Uneducated is only ~3% of the observations, , the rest is close to a ~61%, ~24% and ~12% for Tertiary, Secondary and Primary respectively.
    - Wife_religion: Non-Scientology is ~15% of the observations the rest is Scientology.
    - Wife_Working: Yes is ~25% of the observations the rest is No.
    - Standard_of_living_index: Very Low is only ~9% of the observations, , the rest is close to a ~46%, ~29% and ~16% for Very High, High and Low respectively.
    - Media_exposure: Not-Exposed is ~7% of the observations the rest is Exposed.
    - Contraceptive_method_used: No is 43% of the observations the rest is Exposed.
  - Numerical features
    - Wife_age does not have any outliers
    - No_of_children_born does not have any outliers
    - We may want to change this to a categorical column using binning

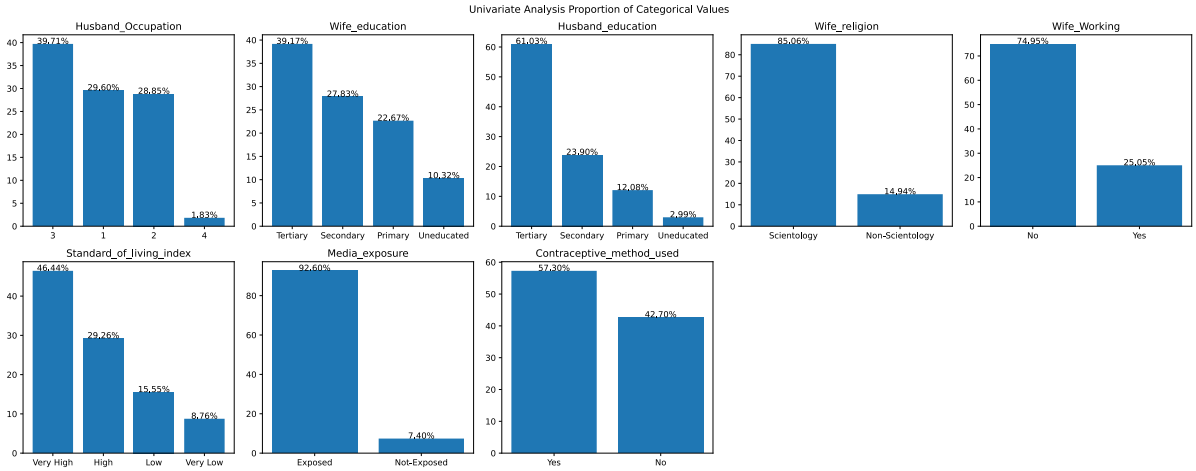

Figure 12 Proportion of values in Categorical Features for the Demo-Socio Survey Data

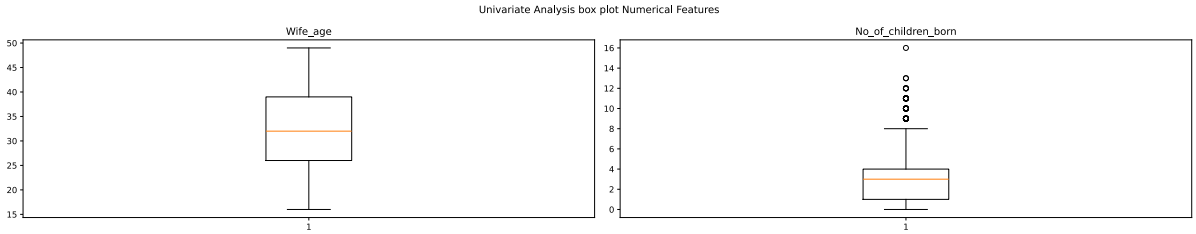

Figure 13 Distribution of the Numerical Features in the Demo-Socio Data

Multivariate analysis:

- Summary:
  - Categorical features
    - Husband_Occupation: all of the categories are close to a 50-50 split vs Contraceptive_method_used for No-Yes, expect "1" which is close a 40-60 split, the mix of splits is different in the categories
    - Wife_education: Secondary and Tertiary have a higher Yes Contraceptive_method_used vs Primary and Uneducated where No Contraceptive_method_used is higher
    - Husband_education: Secondary and Tertiary have a higher Yes Contraceptive_method_used vs Primary and Uneducated where No Contraceptive_method_used is higher
    - Wife_religion: Non-Scientology has a higher Yes Contraceptive_method_used vs Scientology
    - Wife_Working: No has a higher Yes Contraceptive_method_used vs Yes
    - Standard_of_living_index: Low and Very Low have a lower Yes Contraceptive_method_used vs High and Very High
    - Media_exposure: Exposed have a higher Yes Contraceptive_method_used vs Not-Exposed
  - Numerical features:
    - Wife_age: The median age of women of No Contraceptive_method_used is slightly higher than yes, No Contraceptive_method_used
    - Also, there is more spread/variance in the ages of the women not using Contraceptive vs women who use
    - would change this to categorical column using binning in data preprocessing
    - No_of_children_born: The median No of children born is lesser in No Contraceptive_method_used is slightly higher than yes, No Contraceptive_method_used
    - There are outliers in both the data sets

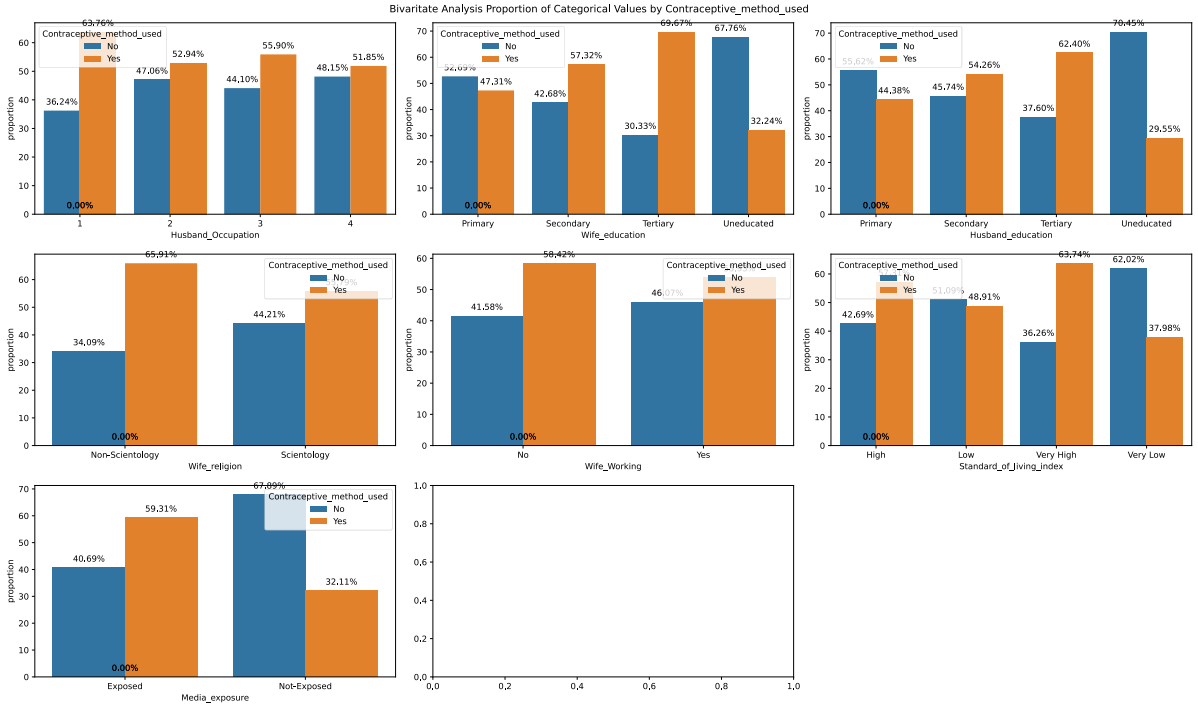

Figure 14 Proportion of values in Categorical Features for the Demo-Socio Survey Data Split by "Contraceptive Method Used"

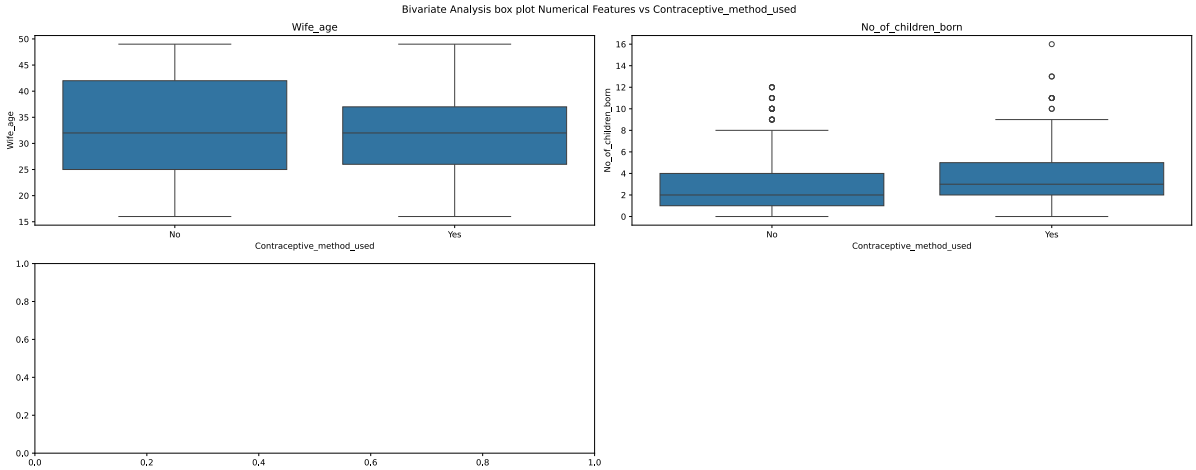

Figure 15 Distribution of the Numerical Features in the Demo-Socio Data Spilt by "Contraceptive Method Used"

Key meaningful observations on individual variables and the relationship between variables:

- ***Husband_Occupation Category "4" does not have enough observations and can be merged with "3".***
- ***Husband_Occupation Category does not seem to have any influence on Contraceptive_method_used.***
- ***The mix of education for both the husband and wife is as follows: Tertiary > Secondary > Primary > Uneducated***
- ***For both Wife_education and Husband_education as the education level increase the mix of Contraceptive_method_used Yes increases from ~30% on the Uneducated end to ~60% and above on the Tertiary end.***
- ***Most of the dataset observations belong to the Scientology religion.(~85%)***
- ***Wife_religion non-scientology women have a ~10% higher observations of Contraceptive_method_used Yes.***
- ***Most of the dataset observations belong to the non-working women.(~75%)***
- ***Wife_Working does not seem to have too much of an impact on Contraceptive_method_used.***
- ***The mix of Standard_of_living_index is as follows: Very High > High > Low > Very Low***
- ***For Standard_of_living_index as the level increases the mix of Contraceptive_method_used Yes increases from ~38% on the Very Low end to ~64% on the Veery High end.***
- ***Media_exposure does have an impact on Contraceptive_method_used.***
- ***The data has a good mix of observations for Contraceptive_method_used, ~57% Yes vs ~43% No***

Data Pre-processing

Prepare the data for modelling :

Missing value Treatment (if needed)

- Wife_age and No_of_children_born are treated for missing values with imputation of median value

Outlier Detection(treat, if needed)

- There are outliers in No_of_children_born but we will not treat these as these are real representations

Feature Engineering (if needed)

- We will not engineer any features
- we would scale the columns Wife_age and number of children.

Encode the data

- Encode Wife_education and Husband_education with values 0,1,2,3 for Uneducated, Primary, Secondary and Tertiary respectively.
- Encode Wife_religion with values 0,1 for Non-Scientology and Scientology respectively.
- Encode Wife_Working with values 0,1 for No and Yes respectively.
- Encode Standard_of_living_index with values 0,1,2,3,4 for Very Low, Low, High and Very High respectively
- Encode Media_exposure with values 0,1 for Not-Exposed and Exposed respectively

Train-Test Split:

- We do a split of the data 70-30 Train-Test.
- Apply scaling (described in Feature Engineering) to the Train and Test post the Split.

Model Building and Compare the Performance of the Models :

Build a Logistic Regression model - Build a Linear Discriminant Analysis model - Build a CART model

- We build Logistic Regression, LDA and a Decision Tree Classifier.
- The results to compare is here:

	Train_accuracy_score	Test_accuracy_score	Training_precision_score	Test_precision_score	Training_recall_score	Test_recall_score	Training_f1_score	Test_f1_score	Training_roc_auc_score	Test_roc_auc_score
Logistic_Regression	0.677983	0.662896	0.684358	0.661238	0.822148	0.818548	0.746951	0.731532	0.651304	0.641233
Decision_Tree_Regressor	0.979631	0.660633	0.989779	0.709402	0.974832	0.669355	0.982249	0.688797	0.98052	0.65942
LDA_Regression	0.678952	0.665158	0.682759	0.66129	0.830537	0.826613	0.749432	0.734767	0.650901	0.642688

Table 7 Compare Scores of Logistic Regression, LDA Regression and Decision Tree Classifier

- f1 score which is the harmonic mean of the Recall and Precision is the highest for the "Logistic Regression", it also has decent recall but precision is a concern.
- Precision score is good for the Decision Tree Classifier, depending on the use case we can got with either.

Prune the CART model by finding the best hyperparameters using Grid Search:

- optimum parameter for the CART Model is : {'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 30, 'min_samples_split': 15}
- we see that with this the CART Model is a good contender as the f1 score is the highest and the recall and precision scores is the highest
- Runing the model with those parameters this is how the performance changes:

	Train_accuracy_score	Test_accuracy_score	Training_precision_score	Test_precision_score	Training_recall_score	Test_recall_score	Training_f1_score	Test_f1_score	Training_roc_auc_score	Test_roc_auc_score
Logistic_Regression	0.677983	0.662896	0.684358	0.661238	0.822148	0.818548	0.746951	0.731532	0.651304	0.641233
Decision_Tree_Regressor	0.733269	0.70362	0.741353	0.702422	0.827181	0.818548	0.781919	0.756052	0.715889	0.687625
LDA_Regression	0.678952	0.665158	0.682759	0.66129	0.830537	0.826613	0.749432	0.734767	0.650901	0.642688

Table 8 Compare Scores of Logistic Regression, LDA Regression and Decision Tree Classifier, after pruning the DTC.

Compare the performance of all the models built and choose the best one with proper rationale:

- ***we see that with this the CART Model is a best as on the test data:***
  - ***f1 score is the highest***
  - ***the recall and precision scores are the highest***
  - ***the roc_auc_score is the highest***

Business Insights & Recommendations:

- if we run the "Logistic Regression" in its current state it is not very useful, in order to strengthen the model, ***we would like to perform tools Like PCA for further feature engineering to try and get a more robust "Logistic Regression"***
- We also need to check if we could ***get more data as the number of observations seems to be less it would be good to get more surveys done*** to get more info.
- we could also ***explore further with one hot encoding techniques*** to see if any goodness is got with that model trained on that data.
- we could ***explore tools and techniques like random forest, Support Vector, KMeans*** clustering too to see if this helps identify patterns in the data.
- In the current state ***we do have the option of going with the "Decision Tree Classifier"*** which though may not have a linear relationship with the independent and dependant variables, still seems to a very good fit given the current training and test. This ***model however may need to be trained move frequently on live data and tested in real case scenarios*** to make sure of its viability