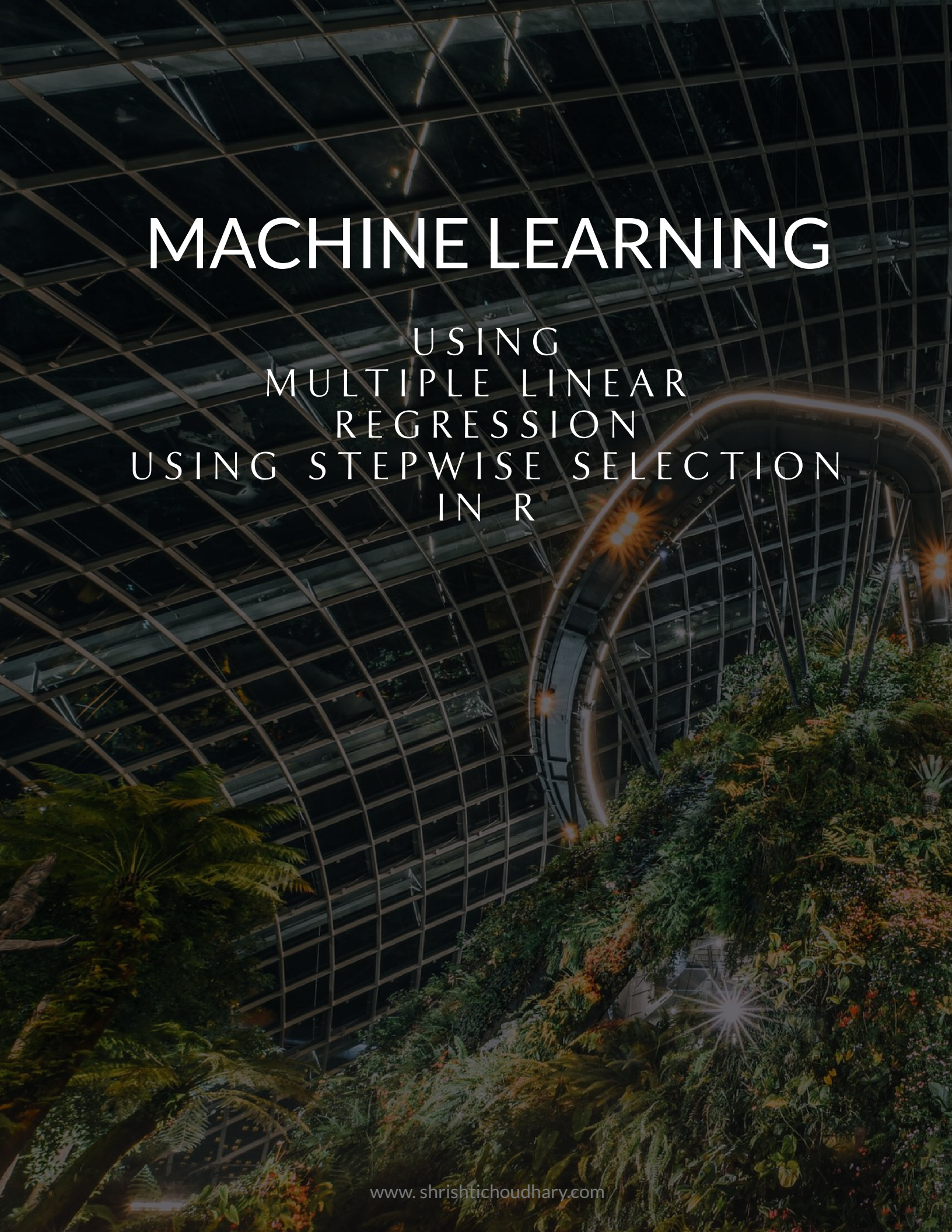


MACHINE LEARNING

USING MULTIPLE LINEAR REGRESSION USING STEPWISE SELECTION IN R

www. shrishtichoudhary.com

TABLE OF CONTENT

Data Exploration

The first step before fitting in a model is to clean the data to perform calculations effectively. In the given data set, “FMarea_intercondyle”- the Proxy for body mass, consisted of one NA which had to be omitted.
The dataset consisted of categorical and continuous variables. Hence the continuous data has been
summarised as shown in figure 1 based on Measures of Central tendency. The analysis of data has
been further facilitated by a graphical representation using a histogram. The individual histograms,
figure 2 will give us an insight into the data structure which helps us understand the symmetry and
skewness of each of the predictor variable independently. Furthermore, we have analysed the
relationship of the predictor variables to the response variable Mean Orbital Volume to understand
the linearity before performing the regression.

```
> summary(L)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-35.630   1.360  28.510  24.860  48.420  65.000
> summary(AL)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.020   8.535  28.510  29.023  48.420  65.000
> summary(FI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 434.6   548.0   615.2   628.0   699.8   890.6
> summary(CC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1100    1290    1350    1373    1465    1700
> summary(MI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1862   25411   64565   62270  101165  112202
> summary(MT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
14.390   0.000   7.780   6.708  16.670  26.670
```

Figure 1: Descriptive statistics showing measures of central tendency

Histograms

The following histograms represent the data set of each predictor variable that are independent of each other. The graph shows that the data set of Latitude is bimodal with zero outliers. Cranial Capacity data set is quite close to symmetrical without any outliers. The graph of Absolute Latitude can be classified as approximately uniform with a slight variation due to skewness at the beginning (positively skewed). The graph of Minimum Illuminance follows a random distribution. Finally, the graphs of FM area intercondyle and Minimum Temperature in Celsius consists of outliers in the end and the beginning of their data set respectively, making it a skewed data set with no apparent distribution

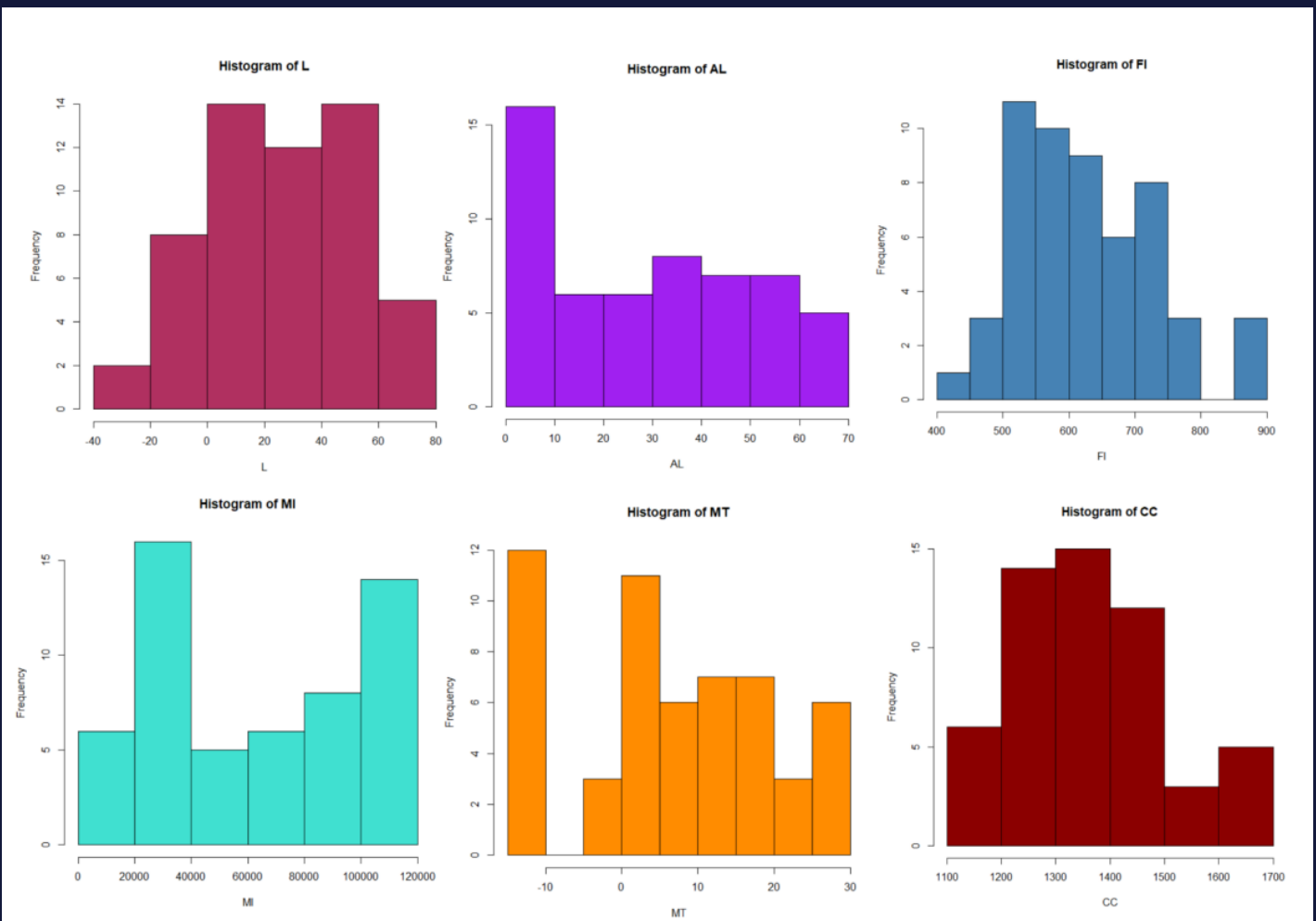


Figure 2: Histograms that show the distribution of the Predictor variable dataset

Scatterplot Matrix

The scatterplot matrix is an important tool that consists of the response variable (Moving Orbital Volume – MOV) at the top followed by all the predictor variables in the order of Gender, Population, Latitude, Absolute Latitude, Cranial Capacity, FM area intercondyle, Minimum Illuminance, Minimum Temperature in Celsius respectively.

The plot establishes a positive linear relationship between the response variable Mean Orbital Volume and Absolute Latitude, Cranial Capacity and a weak positive relationship with FM area intercondyle respectively.
The plot also reflects on a very weak negative relationship between the response variable Mean Orbital Volume and Minimum Illuminance, Minimum Temperature respectively.

The plot shows a scattered relationship (no relationship) with both the categorical variables, that is Gender and Population. Since the programming language R is equipped to handle the categorical variables directly, there was no use of any Dummy variables to plot the linear relationship.

A further analysis in the later parts of this report would reflect on whether either of these categorical variables are significant in having an impact on the response variable.

SCATTERPLOT MATRIX PLOTS

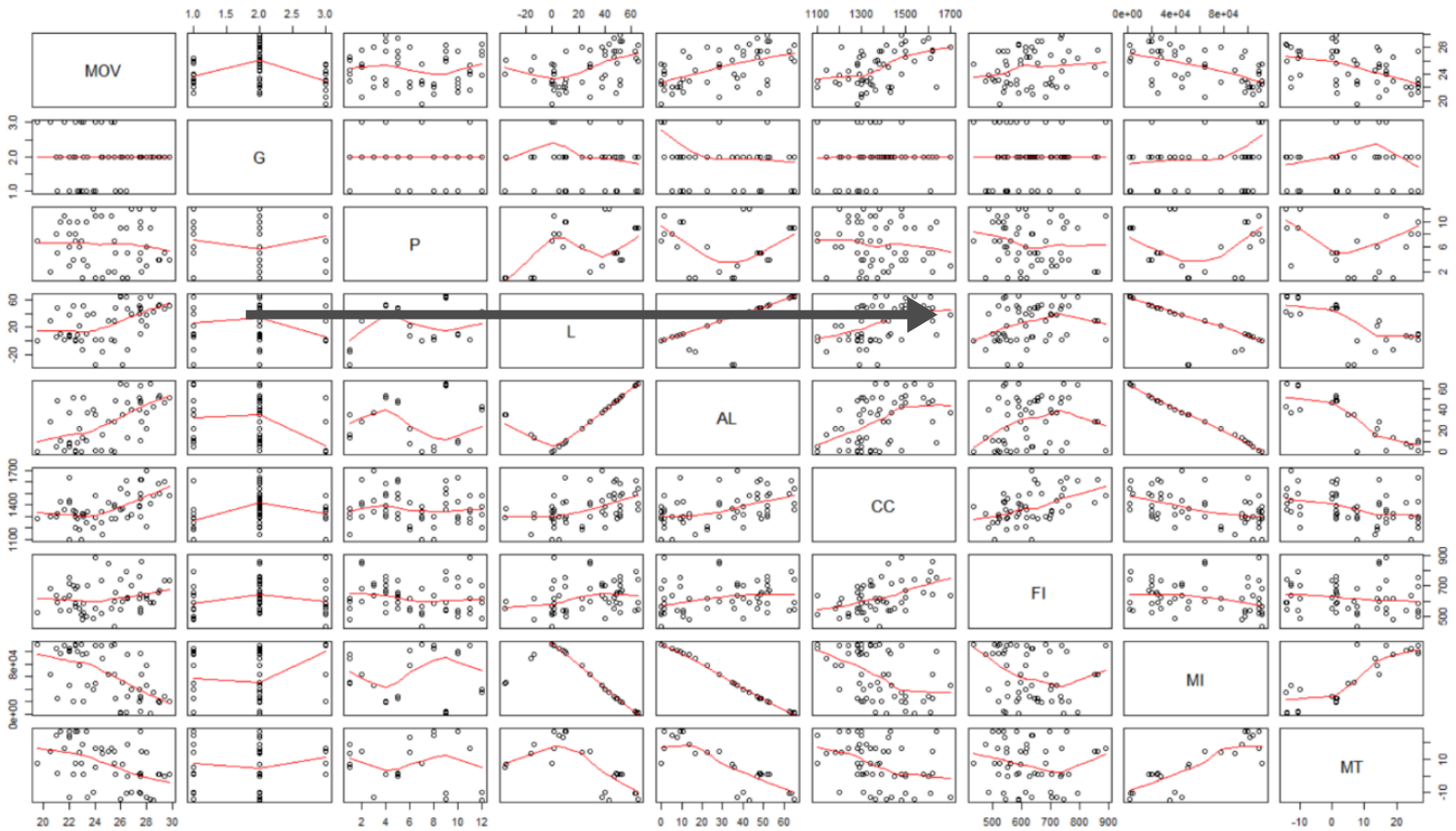


Figure 3: Scatterplot of all the response and predictor variables

	Latitude	Abs.Latitude	Cranial.Cap	F.intercondyle	MeanOrbitalVolume	Min.Illuminance	MinTemp.C
Latitude	1.0000000	0.8336693	0.4596949	0.3242741	0.4423110	-0.8393263	-0.6837520
Abs.Latitude	0.8336693	1.0000000	0.4522449	0.2001993	0.5433619	-0.9979144	-0.7825639
Cranial.Cap	0.4596949	0.4522449	1.0000000	0.5070895	0.4897965	-0.4556977	-0.3563430
F.intercondyle	0.3242741	0.2001993	0.5070895	1.0000000	0.1526362	-0.2123179	-0.1103763
MeanOrbitalVolume	0.4423110	0.5433619	0.4897965	0.1526362	1.0000000	-0.5432854	-0.4500650
Min.Illuminance	-0.8393263	-0.9979144	-0.4556977	-0.2123179	-0.5432854	1.0000000	0.7988645
MinTemp.C	-0.6837520	-0.7825639	-0.3563430	-0.1103763	-0.4500650	0.7988645	1.0000000

MODEL ASSUMPTIONS

Establish a linear relationship between the response variable and the predictor variables: As seen in the scatterplots most of the predictor variable showed a linear relationship with the response variable but varied on the intensity of the relationship as some were strong positive, strong negative, weak positive, weak negative.

The categorical variable showed a weak curvilinear relationship as well. Normality assumption: Under this assumption, the residuals obtained from the data is Normally Distributed (figure 4).

Multicollinearity must not exist: The predictor variables must not be highly correlated with one another. When this occurs it undermines the statistical significance of predictor variables independently. Thus when the standard error of its regression coefficient is very large then the predictor variable will be less statistically significant.
Homoscedasticity: Under this assumption, we know that the variance of error terms may be similar across all the predictor variables.

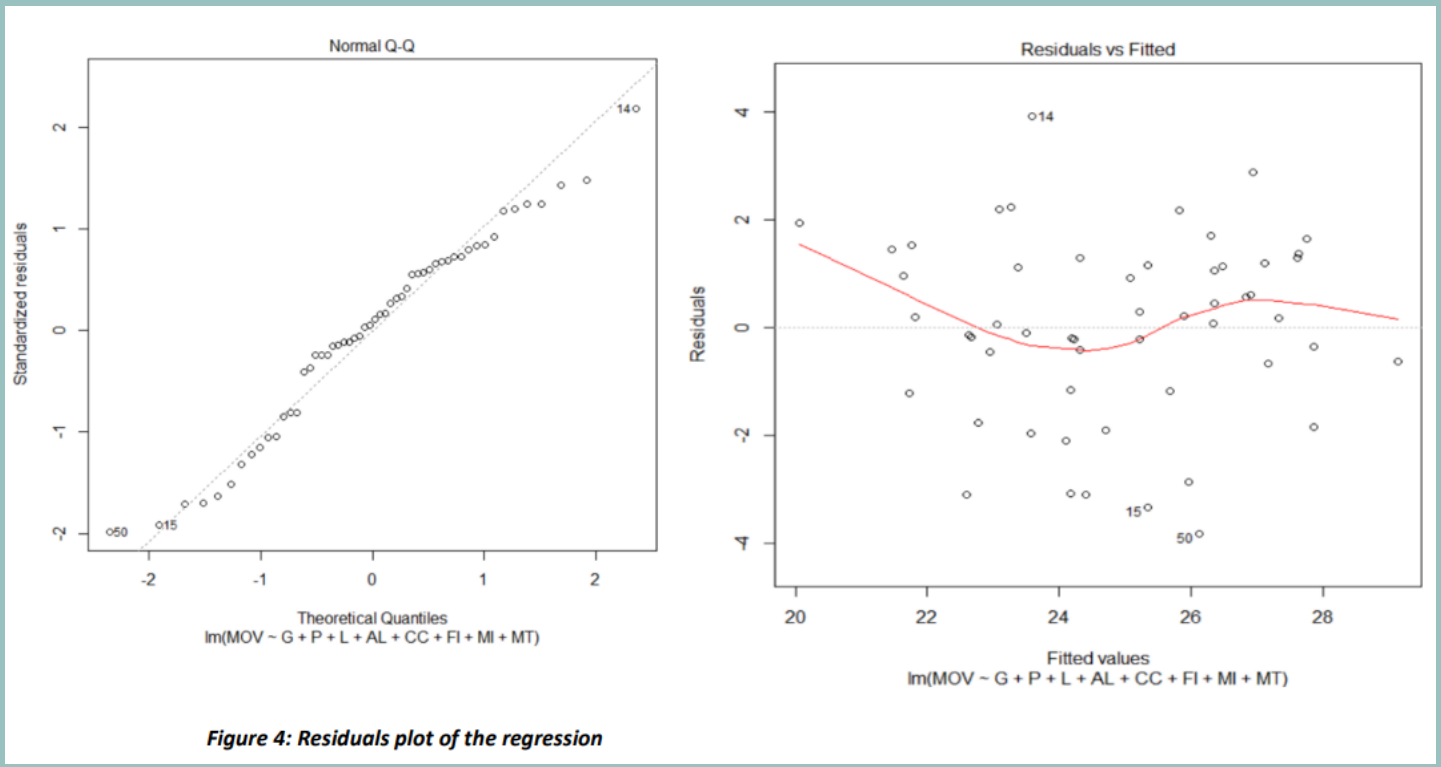


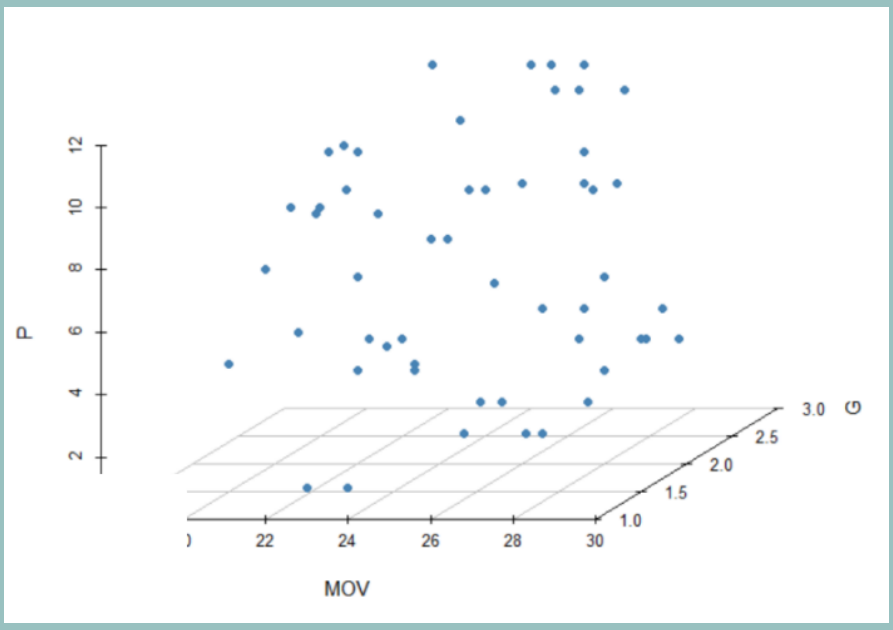
Figure 4: Residuals plot of the regression

MODEL FITTING & ANALYSIS

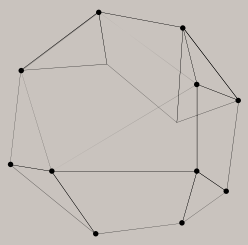
The model was fitted according to the following equation:

$$Y = \beta 0 + \beta 1Xi1 + \beta 2Xi2 + \beta 3Xi3 + \beta 4Xi4 + \beta 5Xi5 + \beta 6Xi6 + \varepsilon$$

The "β" values are the regression weights used to minimise the sum of squared deviations. Our model consists of six predictors. The first term is independent of x and hence is the intercept. It also consists of an error term in the end which is the difference between an expected value at any given point of time and the price that was actually observed.
We also report a sample correlation matrix for the data to understand the relationships between the identified variable. Here the categorical data has been omitted because R cannot compute a correlation matrix for non-numerical data. The relationship is evident in terms of linearity as seen earlier in the scatter plot (Figure3) as well.



The figure is an independent scatter plot of the categorical variables against the response variable. Since it was not possible to compute the correlation value numerically, it was ideal to compare the categorical variables "Gender" and "Population" independently against the response variable Mean Orbit Volume in R. The plot shows a weak positive curvilinear possibility.



MODEL FINDINGS

The model was then fitted under a regression using all the predictor variables. The output summary is:

```
Call:
lm(formula = MOV ~ G + P + L + AL + CC + FI + MI + MT)

Coefficients:
   (Intercept)           GMale        GUnknown  PCanaryIslands          PChina
     1.612e+02        1.444e+00       -1.366e+00       -2.280e+01       -4.152e+01
        PEngland          PFrance           PIndia           PKenya      PMicronesia
    -3.629e+01       -3.663e+01       -1.653e+01       -2.078e+01       -1.534e+01
   PScandinavia         PSomalia          PUganda             PUSA                L
    -4.023e+01       -1.251e+01       -1.694e+01       -3.926e+01        4.142e-01
             AL               CC               FI               MI               MT
    -2.076e+00        6.356e-03       -5.041e-04       -1.077e-03       -4.353e-01
```

The output shows the R sq also known as the coefficient of determination helps in understanding how aligned is the model and the Adjusted R sq, always lower than R sq compares the correlations by considering the predictor variables that are statistically significant in the model and drops the one that are less significant. While the former is adequately high which means that approximately 59.16% of the variation in the Mean Orbital Volume can be explained by the model that is by the use of the predictor variables Gender, Population, Latitude, Absolute Latitude, Cranial Capacity, Minimum Illuminance and Minimum Temperature in Celsius. Adjusted R sq insinuates that there are some predictor variables that can be dropped out of the regression to give the model a better fit. The intercept is the estimated mean value of the Mean Orbital Volume when all the predictor variables are 0.

The Residual Standard Error tells us how far the observed Mean Orbital Volume is from the predicted or fitted Mean Orbital Volume. The F statistic and p-value is the test statistic that checks for overall significance of our model. The null hypothesis is that all of the model coefficients are equal to 0 versus the alternate that it is not 0.

H0 : $\beta1 = \beta2 = \beta3 = \beta4 = \beta5 = \beta6 = 0$
HA : $\beta1 = \beta2 = \beta3 = \beta4 = \beta5 = \beta6 \neq 0$

WHY STEPWISE REGRESSION?

Thus, the overall summary suggests that there is more room for a better fit due to high p values of the individual variables. In multiple linear regression when it becomes difficult to ascertain which combination of variables are a better fit or not, a Stepwise Regression becomes more useful.

Under the Backward Elimination technique, the model starts with all the predictor variables and drops down significantly in the process based on the AIC values. The AIC values play a crucial role because a lower AIC value of the model suggests that the model is a good fit and predictors are statistically significant.

```
Call:
lm(formula = MOV ~ G + P + L + AL + CC + FI + MI + MT)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8156 -1.0388  0.1277  1.1888  3.9206

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.612e+02  1.385e+02   1.164   0.2527
GMale           1.444e+00  9.597e-01   1.505   0.1416
GUnknown       -1.366e+00  1.637e+00  -0.834   0.4101
PCanaryIslands -2.280e+01  1.412e+01  -1.614   0.1157
PChina         -4.152e+01  2.539e+01  -1.635   0.1112
PEngland       -3.629e+01  2.174e+01  -1.669   0.1043
PFrance        -3.663e+01  2.238e+01  -1.637   0.1109
PIndia         -1.653e+01  1.073e+01  -1.540   0.1327
PKenya         -2.078e+01  1.966e+01  -1.057   0.2979
PMicronesia    -1.534e+01  1.182e+01  -1.298   0.2029
PScandinavia   -4.023e+01  2.166e+01  -1.857   0.0720 .
PSomalia       -1.251e+01  8.998e+00  -1.390   0.1735
PUganda        -1.694e+01  1.611e+01  -1.052   0.3004
PUSA           -3.926e+01  2.565e+01  -1.531   0.1351
L               4.142e-01  2.597e-01   1.595   0.1200
AL             -2.076e+00  2.131e+00  -0.974   0.3369
CC              6.356e-03  3.438e-03   1.849   0.0732 .
FI             -5.041e-04  3.790e-03  -0.133   0.8950
MI             -1.077e-03  1.057e-03  -1.019   0.3153
MT             -4.353e-01  5.291e-01  -0.823   0.4164
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.122 on 34 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5916,    Adjusted R-squared:  0.3633
F-statistic: 2.592 on 19 and 34 DF,  p-value: 0.007541
```

```
> ModR<-step(lm(MOV~G+P+L+AL+CC+F+MI+MT), direction = "backward")
Start:  AIC=96.91
MOV ~ G + P + L + AL + CC + F + MI + MT

       Df Sum of Sq    RSS     AIC
- P     11     54.481 209.27  91.495
- F      1      0.096 154.88  94.943
- MT     1      1.568 156.35  95.463
- AL     1      2.887 157.67  95.926
- MI     1      3.377 158.16  96.096
<none>                154.79  96.909
- L      1      9.801 164.59  98.286
- CC     1     15.039 169.82 100.009
- G      2     27.379 182.16 101.867

Step:  AIC=91.5
MOV ~ G + L + AL + CC + F + MI + MT

       Df Sum of Sq    RSS    AIC
- MI     1     0.0020 209.27 89.496
- AL     1     0.1623 209.43 89.538
- F      1     0.4953 209.76 89.625
- L      1     0.9633 210.23 89.748
- MT     1     1.6581 210.93 89.929
<none>                209.27 91.495
- CC     1    11.2715 220.54 92.381
- G      2    22.1496 231.42 93.029

Step:  AIC=89.5
MOV ~ G + L + AL + CC + F + MT

       Df Sum of Sq    RSS    AIC
- F      1     0.5330 209.80 87.636
- L      1     0.9778 210.25 87.752
- MT     1     1.8752 211.15 87.987
- AL     1     6.3246 215.59 89.133
<none>                209.27 89.496
- CC     1    11.4164 220.69 90.417
- G      2    22.1573 231.43 91.031

Step:  AIC=87.64
MOV ~ G + L + AL + CC + MT

       Df Sum of Sq    RSS    AIC
- L      1     1.4424 211.25 86.013
- MT     1     2.2350 212.04 86.219
- AL     1     6.5435 216.35 87.325
<none>                209.80 87.636
- CC     1    11.7753 221.58 88.639
- G      2    23.9787 233.78 89.588


Step:  AIC=86.01
MOV ~ G + AL + CC + MT

       Df Sum of Sq    RSS    AIC
- MT     1     2.1775 213.42 84.577
- AL     1     5.2762 216.52 85.369
<none>                211.25 86.013
- CC     1    10.9248 222.17 86.786
- G      2    24.9759 235.62 88.019

Step:  AIC=84.58
MOV ~ G + AL + CC

       Df Sum of Sq    RSS    AIC
<none>                213.42 84.577
- CC     1    10.561 223.98 85.233
- G      2    22.760 236.18 86.150
- AL     1    32.149 245.57 90.294
```

The output obtained after following the Backward Elimination technique is as follows:

1. The backward elimination process begins with the primary regression.

2. The first output reflects the lowest AIC of Population 91.495 as compared to the overall AIC of 96.91. This means that if in the next step this predictor variable was dropped then the AIC would reduce further for the model.

3. This process continues as the less significant predictor variables are dropped by comparing their individual AIC with that of the Model.
4. The final and the most significant output produced by this process eliminates most of the variable and leaves the more significant variables that is Gender, Absolute Latitude and Cranial Capacity. The model continued to run this process until all the predictor variables have an AIC higher than that of the model which means that dropping any more predictor variables would not lead to a better fitted model.

Final Model

$$Y = 17.19 + 1.57Xi1 + 0.23Xi2 + 0.45Xi3 + 0.004Xi4$$

```
> summary(final)

Call:
lm(formula = MOV ~ G + AL + CC)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0509 -1.0932  0.2345  1.2048  3.5070

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.18895    3.06962   5.600  9.1e-07 ***
GMale        1.56945    0.78435   2.001   0.0508 .
GUnknown     0.23301    0.98617   0.236   0.8142
AL           0.04497    0.01639   2.744   0.0084 **
CC           0.00390    0.00248   1.573   0.1220
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.066 on 50 degrees of freedom
Multiple R-squared:  0.4308,    Adjusted R-squared:  0.3853
F-statistic: 9.462 on 4 and 50 DF,  p-value: 8.953e-06

> anova(final)
Analysis of Variance Table

Response: MOV
          Df  Sum Sq Mean Sq F value    Pr(>F)
G          2  90.068  45.034 10.5505 0.0001504 ***
AL         1  60.915  60.915 14.2711 0.0004220 ***
CC         1  10.561  10.561  2.4742 0.1220326
Residuals 50 213.422   4.268
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This final output summarises the best fit for the data because:

a. The p-value is statistically significant for individual predictor variables and for the overall model.

b. The Adjusted R2 is higher than the original regression reflecting a better fit of the model.

Thus, the mean orbital volume is significantly impacted from Gender, Absolute Latitude and Cranial Capacity.