STA4813_Assignemet_01_2024

22692037_Tshepiso Mashiane

Total: 42/55 = 76.4/100

2024-04-12

Loading and examining the data.

```
library(readr)
Machinery_T <- read.csv("C:/Users/36076724/Desktop/BSC Hons Statistic/STA4813_GLM/Machinery.csv_
1.csv")
head(Machinery_T)
```

```
##   Tractor_Power Type_Tractor Type_Fruit Average_Purchase_Price Salvage_Value
## 1            48         2 WD   Vineyard                 494800         49480
## 2            36         2 WD   Orchards                 344400         34440
## 3            44         2 WD   Orchards                 433750         43375
## 4            48         2 WD   Orchards                 412902         41290
## 5            51         2 WD   Orchards                 573500         57350
## 6            52         2 WD   Orchards                 350625         35063
##   Average_Investment Depreciation_Costs Insurance_Licence_Costs Interest_Costs
## 1             222660              44.53                    2.23          28.95
## 2             154980              31.00                    1.55          20.15
## 3             195188              39.04                    1.95          25.37
## 4             185806              37.16                    1.86          24.15
## 5             258075              51.62                    2.58          33.55
## 6             157781              31.56                    1.58          20.51
##   Total_Fixed_Costs Repair_Maintenance_Costs Fuel_Costs Fuel_Usage
## 1             46.76                    59.38     181.53       8.64
## 2             32.55                    41.33     136.14       6.48
## 3             40.99                    52.05     166.40       7.92
## 4             39.02                    49.55     181.53       8.64
## 5             54.20                    68.82     192.87       9.18
## 6             33.13                    42.08     196.65       9.36
##   Tractor_Types_dummy
## 1                   0
## 2                   0
## 3                   0
## 4                   0
## 5                   0
## 6                   0
```

From the ***head(Machinery_T)*** we learn that each row in the dataset appears to represent a specific tractor with corresponding attributes, such as tractor power, type, associated fruit type, average purchase price, salvage value, average investment, depreciation costs, insurance cost, license costs, total fixed costs, repair and maintenance costs, fuel costs, fuel usage and tractor type dummy .The specific units of measurement for the average investment, depreciation costs, and insurance and license costs columns are represented in a specific South African currency (Rand).

```
str(Machinery_T)
```

```
## 'data.frame':    27 obs. of  14 variables:
##  $ Tractor_Power           : int  48 36 44 48 51 52 53 56 57 59 ...
##  $ Type_Tractor            : chr  "2 WD" "2 WD" "2 WD" "2 WD" ...
##  $ Type_Fruit              : chr  "Vineyard" "Orchards" "Orchards" "Orchards" ...
##  $ Average_Purchase_Price  : int  494800 344400 433750 412902 573500 350625 460200 847768 604
950 478500 ...
##  $ Salvage_Value           : int  49480 34440 43375 41290 57350 35063 46020 84777 60495 47850
...
##  $ Average_Investment      : int  222660 154980 195188 185806 258075 157781 207090 381496 272
228 215325 ...
##  $ Depreciation_Costs      : num  44.5 31 39 37.2 51.6 ...
##  $ Insurance_Licence_Costs : num  2.23 1.55 1.95 1.86 2.58 1.58 2.07 3.81 2.72 2.15 ...
##  $ Interest_Costs          : num  28.9 20.1 25.4 24.1 33.5 ...
##  $ Total_Fixed_Costs       : num  46.8 32.5 41 39 54.2 ...
##  $ Repair_Maintenance_Costs: num  59.4 41.3 52 49.5 68.8 ...
##  $ Fuel_Costs              : num  182 136 166 182 193 ...
##  $ Fuel_Usage              : num  8.64 6.48 7.92 8.64 9.18 ...
##  $ Tractor_Types_dummy     : int  0 0 0 0 0 0 0 0 0 0 ...
```

We can see that some of the variables, such as the Type_Tractor and Type_Fruit, are factors. Factor variables are categorical. Other variables are quantitative. Variables such as Depreciation_Costs, Insurance_Licence_Costs, Interest_Costs, Total_Fixed_Costs , Repair_Maintenance_Costs,Fuel_Costs and Fuel_Usage are continuous while the others are integer valued. We also see the sample size is 27.The expenditure on tractors purchases will vary as per specific tractor with corresponding attributes such as tractor power, type, associated fruit type, average purchase price, salvage value, average investment, depreciation costs, insurance cost, license costs, total fixed costs, repair and maintenance costs, fuel costs, and fuel usage.

Initial Data Analysis

```
summary(Machinery_T)
```

```
##  Tractor_Power   Type_Tractor        Type_Fruit         Average_Purchase_Price
##  Min.   :36.00   Length:27          Length:27          Min.   : 266650
##  1st Qu.:48.00   Class :character   Class :character   1st Qu.: 446975
##  Median :53.00   Mode  :character   Mode  :character   Median : 526400
##  Mean   :53.78                                         Mean   : 566553
##  3rd Qu.:59.00                                         3rd Qu.: 631840
##  Max.   :71.00                                         Max.   :1171456
##  Salvage_Value    Average_Investment Depreciation_Costs Insurance_Licence_Costs
##  Min.   : 26665   Min.   :119993     Min.   : 24.00     Min.   :1.200
##  1st Qu.: 44698   1st Qu.:201139     1st Qu.: 40.23     1st Qu.:2.010
##  Median : 52640   Median :236880     Median : 47.38     Median :2.370
##  Mean   : 56655   Mean   :254949     Mean   : 50.99     Mean   :2.549
##  3rd Qu.: 63184   3rd Qu.:284328     3rd Qu.: 56.87     3rd Qu.:2.840
##  Max.   :117146   Max.   :527155     Max.   :105.43     Max.   :5.270
##  Interest_Costs  Total_Fixed_Costs Repair_Maintenance_Costs   Fuel_Costs
##  Min.   :15.60   Min.   : 25.20    Min.   : 32.00           Min.   :136.1
##  1st Qu.:26.14   1st Qu.: 42.24    1st Qu.: 53.63           1st Qu.:181.5
##  Median :30.79   Median : 49.74    Median : 63.17           Median :200.4
##  Mean   :33.14   Mean   : 53.54    Mean   : 67.99           Mean   :203.4
##  3rd Qu.:36.96   3rd Qu.: 59.70    3rd Qu.: 75.82           3rd Qu.:223.1
##  Max.   :68.53   Max.   :110.70    Max.   :140.57           Max.   :268.5
##    Fuel_Usage     Tractor_Types_dummy
##  Min.   : 6.480   Min.   :0.0000
##  1st Qu.: 8.640   1st Qu.:0.0000
##  Median : 9.540   Median :1.0000
##  Mean   : 9.733   Mean   :0.5926
##  3rd Qu.:10.800   3rd Qu.:1.0000
##  Max.   :12.780   Max.   :1.0000
```

For the categorical variables, we get a count of the number of each type that occur, *e.g. Tractor that is of type "2 WD" and is used at/for Orchards*. We notice, for example, that almost half of the tractors are of type *"2 WD" and used at/for Orchards*. This will help us to estimate the effect of a particular Tractor power on the expenditure for Tractor purchase. For the numerical variables, we have eleven summary statistics that are sufficient to get a rough idea of the distributions. In particular, we notice that the tractor power ranges over orders of magnitudes. This suggests that I should consider the absolute, rather than the relative Average_Purchase_Price.

We wish to find/determine a linear model to predict the Average Purchase Price of a tractor

a.

```
library(gridExtra)
library(ggplot2)
library(dplyr)
```

4

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Average purchase price Histogram
average_purchace_price_hist<- ggplot(data=Machinery_T, aes(x=Average_Purchase_Price)) +
geom_histogram(binwidth = 100000, fill= "skyblue",color= "black") +
xlab("Average_Purchase_Price") + ylab("frequency") + labs(title = "Figure 1(a) Histogram of Aver
age_Purchase_Price ")
Tractor_power_hist <- ggplot(data = Machinery_T, aes(x = Tractor_Power)) +
  geom_histogram(binwidth = 5,  fill= "skyblue",color= "black") +
  xlab("Tractor Power") + ylab("Frequency") + labs(title = "Figure 1(b) Histogram Tractor_Powe
r")

print(average_purchace_price_hist)
```

```
print(Tractor_power_hist)
```

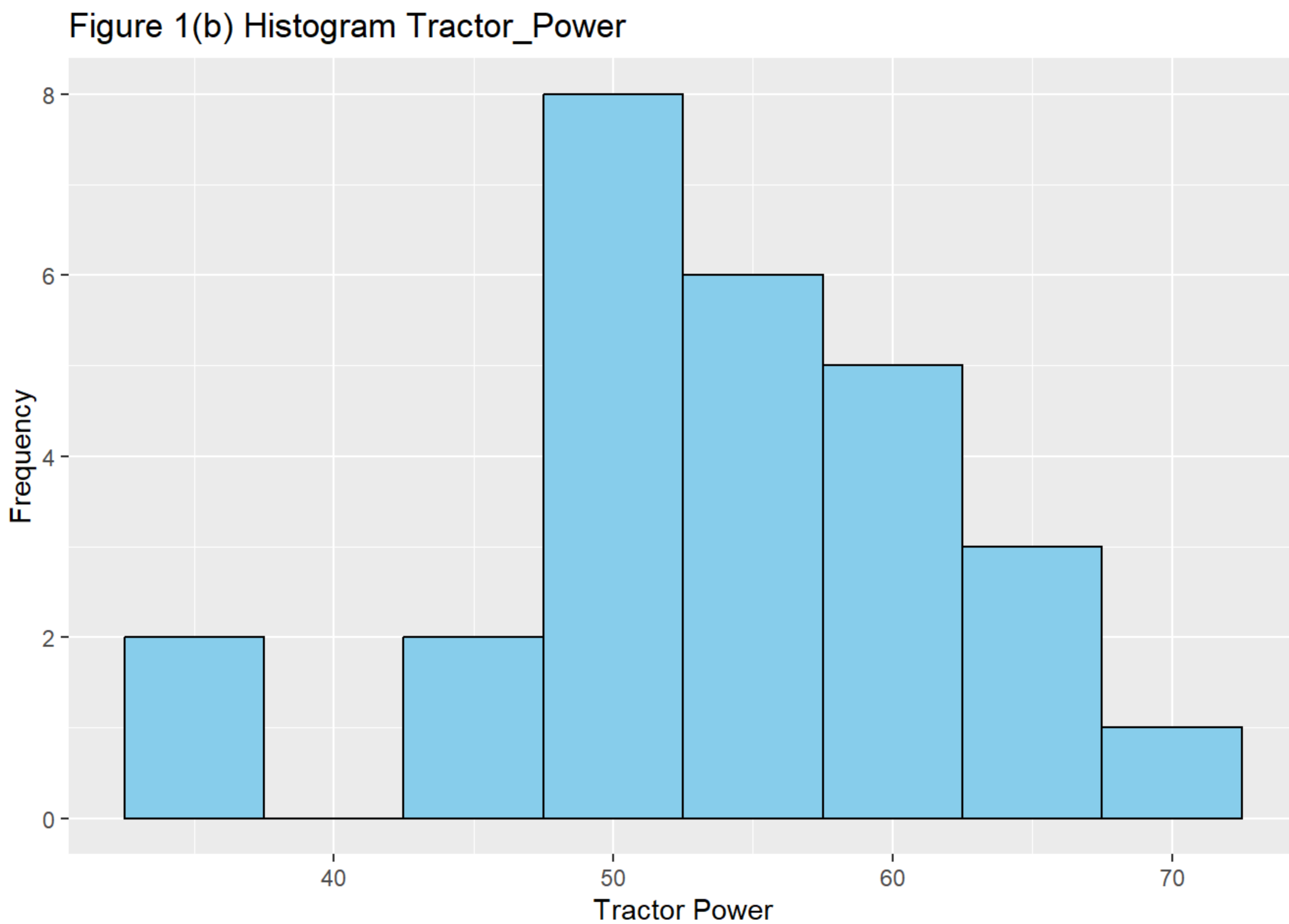


Figure 1: Histogram of Average_purchase_price and Tractor_power. Plot a. shows the average purchase price (in ***R***) of a tractor, while plot b. shows the tractor power (in ***KW***)

```
#Summary Statistics
# Average Purchase Price
summary(Machinery_T$Average_Purchase_Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  266650  446975  526400  566553  631840 1171456
```

```
# Tractor_power
summary(Machinery_T$Tractor_Power)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   36.00   48.00   53.00   53.78   59.00   71.00
```

In ***Figure 1.a*** and ***1.b***, we observe that both the *Average purchase price* and *Tractor power* exhibit a right-skewed distribution. Analyzing the summary statistics, we find that the mean values are greater than the medians, supporting the conclusion of right-skewness *[(Q3-Median)>(Median-Q1)]*. By examining the histogram in ***Figure***

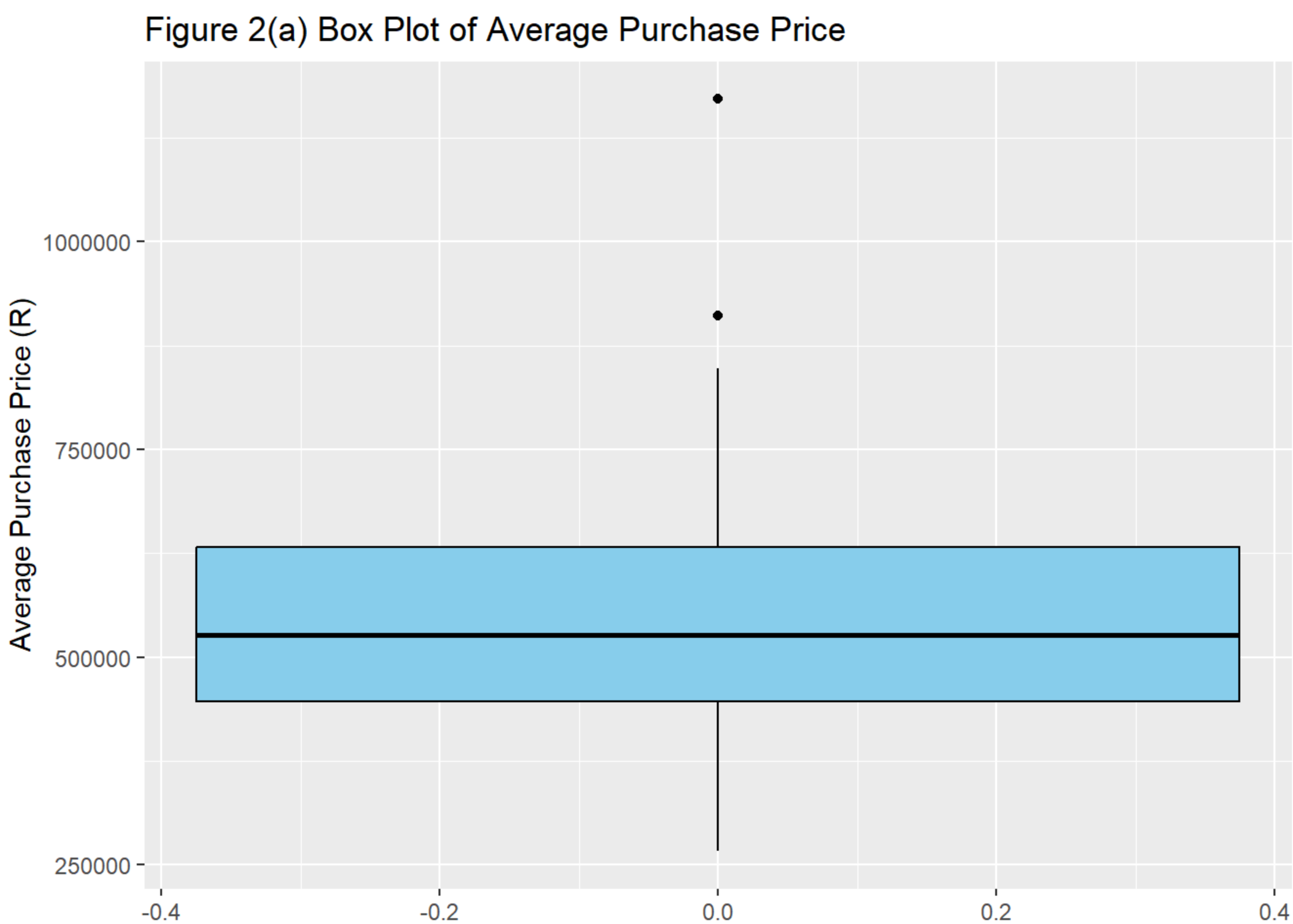
Tractor power is almost symmetrical as the mean and the median approximately equal.

1.a, it is evident that the *Average purchase price* is primarily concentrated around R500000. Similarly, in ***Figure 1.b***, a significant proportion of *Tractor powers* are distributed around 50KW, suggesting that a large number of tractor powers are approximately equal to 50KW.

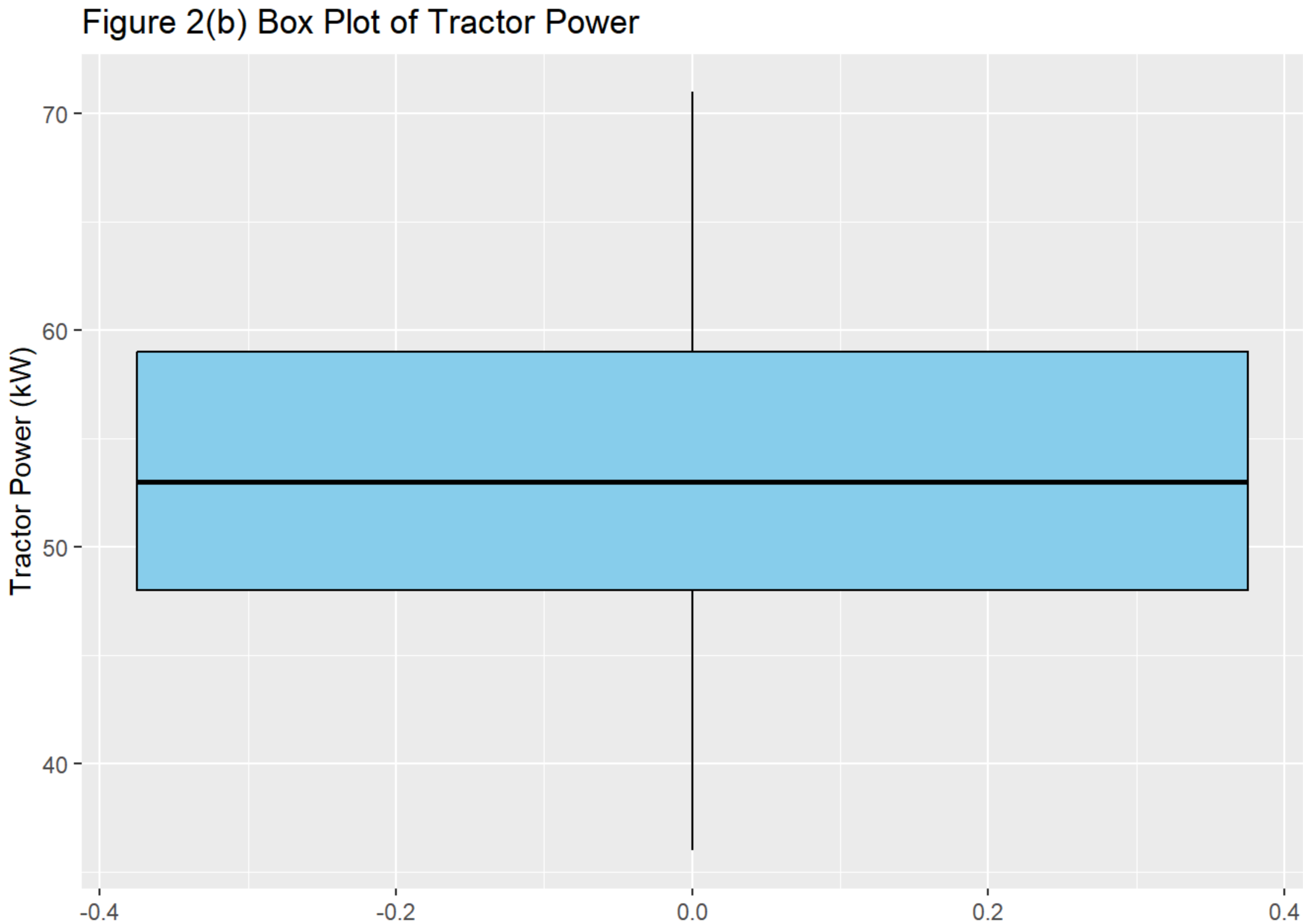
Similarly

```
# Box plot of Average_Purchase_Price
average_purchase_box <- ggplot(data=Machinery_T, aes(y = Average_Purchase_Price)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(y = "Average Purchase Price (R)") +
  ggtitle("Figure 2(a) Box Plot of Average Purchase Price")
# Box plot of Tractor_Power
tractor_power_box <- ggplot(Machinery_T, aes(y = Tractor_Power)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(y = "Tractor Power (kW)") +
  ggtitle("Figure 2(b) Box Plot of Tractor Power")

print(average_purchase_box)
```

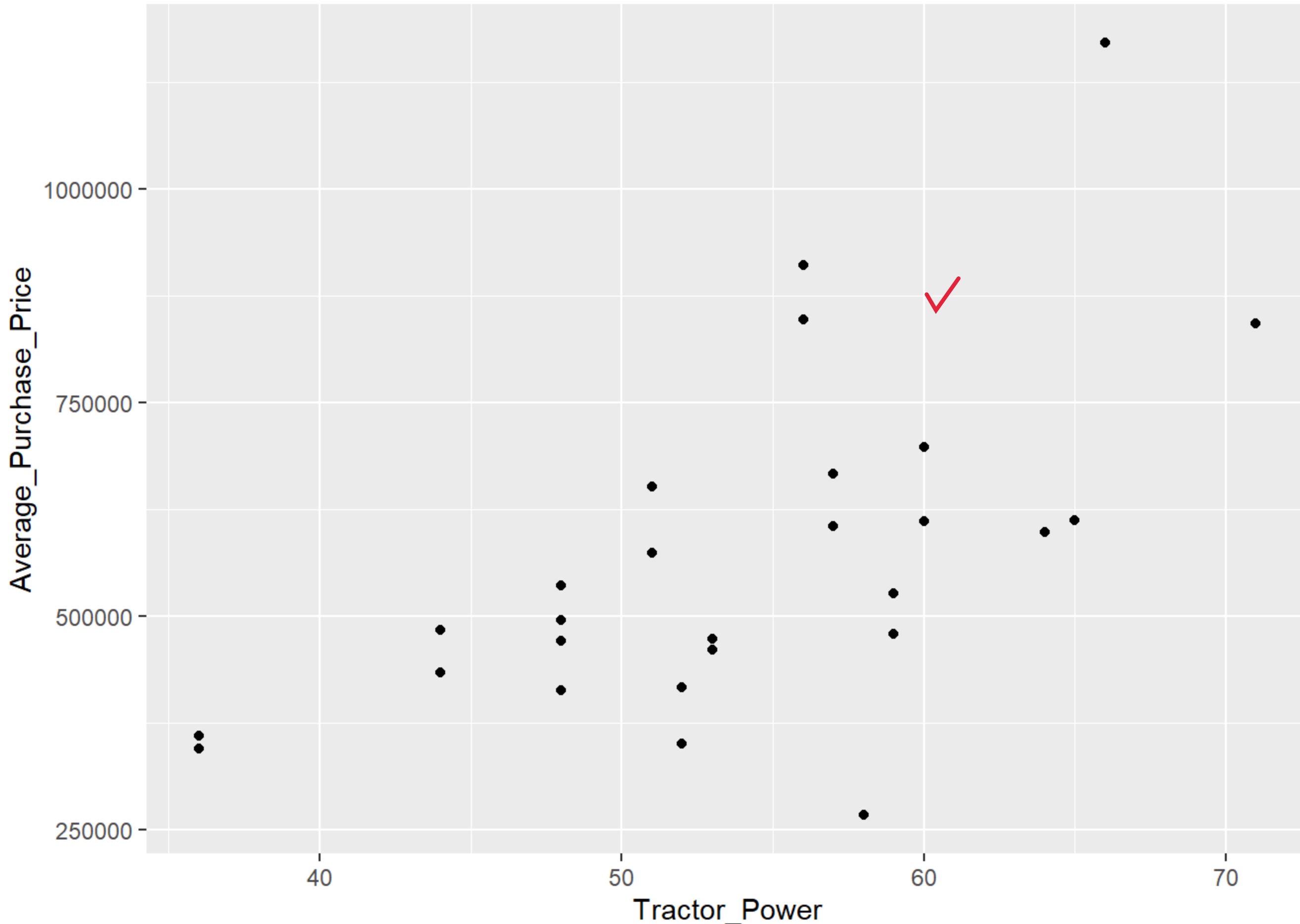



```
print(tractor_power_box)
```



In ***Figure 2.a*** and ***2.b***, we observe that both the *Average purchase price* and *Tractor power* exhibit a right-skewed distribution, as evident from the box plot analysis. By examining the box plot in ***Figure 2.a***, it becomes evident that the distribution of Average purchase price is skewed to the right. While the box plot for *Tractor power* in ***Figure 2.b*** does not clearly show the skewness, we can infer it from the summary statistics where the mean value is greater than the median value *[(Q3-Median)>(Median-Q1)]*, indicating right skewness.

```
library(ggplot2)
ggplot(Machinery_T, aes(Tractor_Power, Average_Purchase_Price)) + geom_point()
```




```
correlation<- cor(Machinery_T$Tractor_Power, Machinery_T$Average_Purchase_Price)
print(correlation)
```

```
## [1] 0.6061933
```

In this case, we observe that larger *Tractor power* is linked to higher *Average purchase prices*. In conclusion, there is a moderately strong positive relationship between the average purchase price and tractor power.

b. 6

```
Model_1 <- lm(Average_Purchase_Price ~ Tractor_Power, data = Machinery_T)
summary(Model_1)
```

```
##
## Call:
## lm(formula = Average_Purchase_Price ~ Tractor_Power, data = Machinery_T)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -360398 -104736    7289   49869  429788
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -203953     204543  -0.997 0.328262
## Tractor_Power    14328       3760   3.811 0.000804 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 161100 on 25 degrees of freedom
## Multiple R-squared:  0.3675, Adjusted R-squared:  0.3422
## F-statistic: 14.52 on 1 and 25 DF,  p-value: 0.0008037
```

Model 1: Average_Purchase_Price = -203953 + 14328(Tractor_Power)

Since the Intercept does not make sense, lets subtract the mean from both the response variable and the independent variable

```
Normalized_average_purchase_price<- Machinery_T$Average_Purchase_Price - mean(Machinery_T$Averag
e_Purchase_Price)
Normalized_tractor_power<- Machinery_T$Tractor_Power- mean(Machinery_T$Tractor_Power)
Model_1a<- lm(Normalized_average_purchase_price~Normalized_tractor_power, data= Machinery_T)
summary(Model_1a)
```

```
##
## Call:
## lm(formula = Normalized_average_purchase_price ~ Normalized_tractor_power,
##     data = Machinery_T)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -360398 -104736    7289   49869  429788
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -3.991e-11  3.101e+04   0.000 1.000000
## Normalized_tractor_power  1.433e+04  3.760e+03   3.811 0.000804 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 161100 on 25 degrees of freedom
## Multiple R-squared:  0.3675, Adjusted R-squared:  0.3422
## F-statistic: 14.52 on 1 and 25 DF,  p-value: 0.0008037
```

```
coef(summary(Model_1a))
```

```
##                             Estimate Std. Error       t value    Pr(>|t|)
## (Intercept)            -3.991025e-11   31007.43 -1.287119e-15 1.000000000
## Normalized_tractor_power  1.432759e+04    3759.52  3.811015e+00 0.000803672
```

Model 1a:

```
intercept <- coef(Model_1a)[1]
slope <- coef(Model_1a)[2]

cat("Estimated Regression Line Equation is given by: (y)i =", round(intercept, 8), "+", round(sl
ope, 8), "* x\n")
```

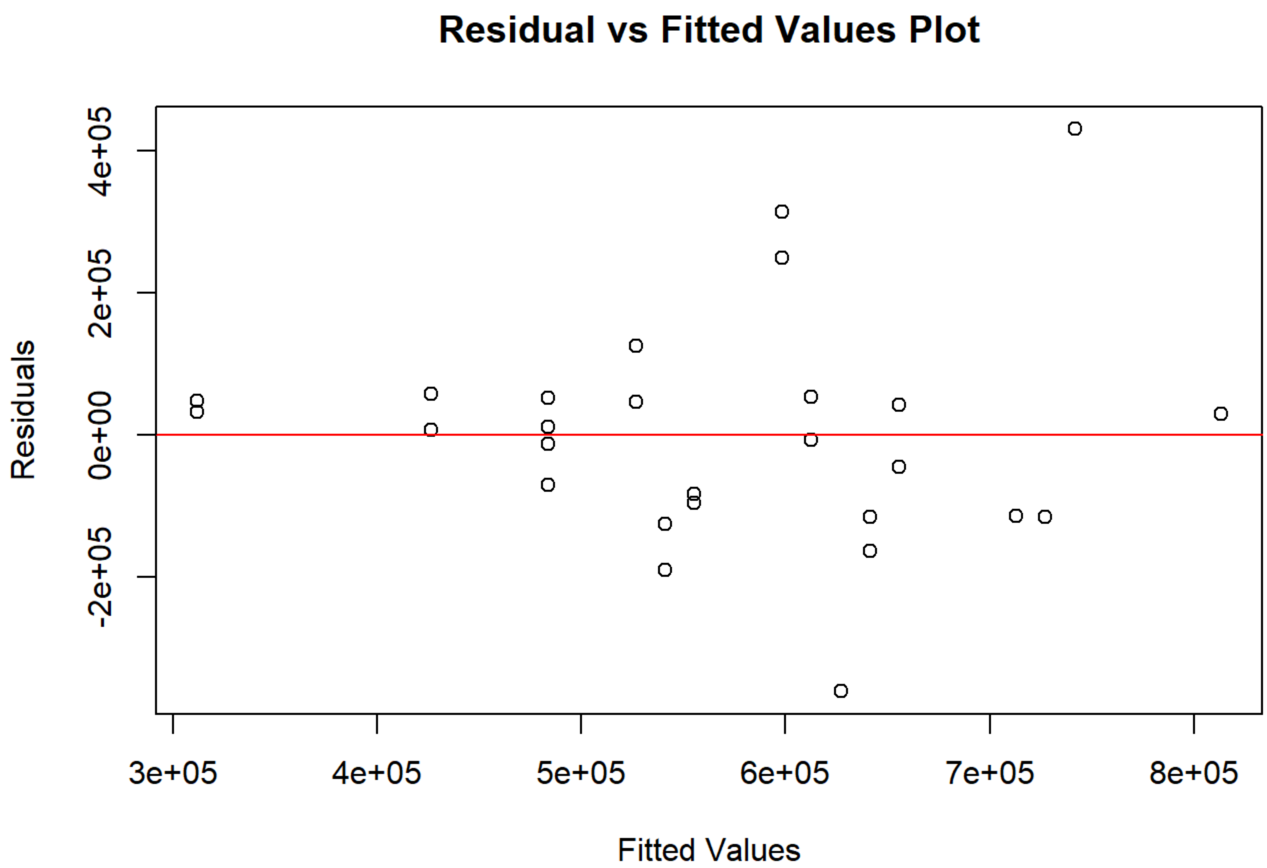
```
## Estimated Regression Line Equation is given by: (y)i = 0 + 14327.59 * x
```

Average_Purchase_Price = -3.991e-11 + 1.433e+04 (Tractor_Power)

Interpretation of β1 The coefficient for the Tractor_Power variable is 1.432759e+04 (approximately 14,327.59).The *Average purchase price* will increase by R14,327.59 when *Tractor power* is increased by 1 KW.

Assumpitions of the linear regression

```
# Obtain the fitted values and residuals
fitted_values <- fitted(Model_1)
residuals <- residuals(Model_1)
# Create the residual vs fitted values plot to assess the assumption of the constant varience of
the residual
plot(fitted_values, residuals, xlab = "Fitted Values", ylab = "Residuals", main = "Residual vs F
itted Values Plot")
abline(h = 0, col = "red")
```



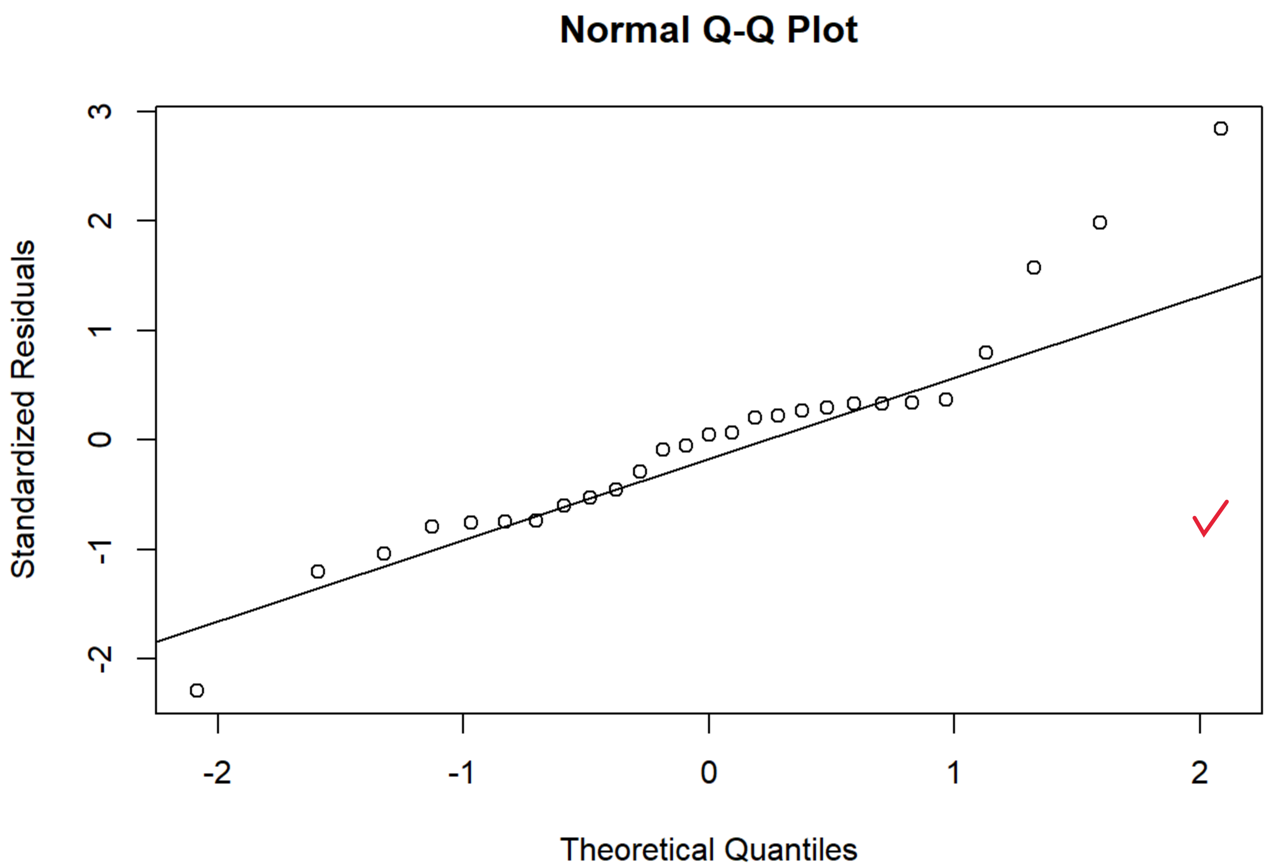
Linearity assessment: since the residuals are randomly scattered around the horizontal line at zero, it suggests that the linear regression mode_1 captures the linear relationship between the predictors and the response variable reasonably well. This indicates that the assumption of linearity is met.

Homoscedasticity/ constant Variance assessment: In the residual vs fitted values plot, we see a random scatter of points with approximately equal spread above and below zero. This indicates that the variance of the residuals is consistent across the range of fitted values/ predictor variables.

Independence of errors assessment: Observing from the Residual vs Fitted Values Plot, the error of one observation does not influence or correlate with the errors of other observations, this implies that each observation contributes unique information to the model and that the estimates of the regression coefficients are unbiased. There the assumption of independent errors is met.

Obtaining the standardized residuals for normality assumption assessment:

```
std_resid <- rstandard(Model_1)
# Create the Normal Q-Q plot
qqnorm(std_resid, xlab = "Theoretical Quantiles", ylab = "Standardized Residuals", main = "Normal Q-Q Plot")
qqline(std_resid)
```



Normality Assessment By examining the Normal Q-Q plot, we notice that the data points deviate from the expected regression line, indicating a violation of the normality assumption.

c. 4

Note:since the Normality assumption is violated, we now need to transform the data using log transformation.

```
# Create the new variable logAverage_Purchase_Price
Machinery_T$logAverage_Purchase_Price <- log(Machinery_T$Average_Purchase_Price)
print(Machinery_T$logAverage_Purchase_Price)
```

```
##  [1] 13.11191 12.74956 12.98022 12.93097 13.25951 12.76747 13.03942 13.65036
##  [9] 13.31290 13.07841 13.32285 13.19189 12.79205 13.08932 13.06178 13.38780
## [17] 12.93829 13.06636 13.72273 13.40987 12.49369 13.17382 13.45597 13.30260
## [25] 13.32396 13.97376 13.64502
```

```
#Fitting the model
Model_2 <- lm(logAverage_Purchase_Price ~ Tractor_Power, data = Machinery_T)
summary(Model_2)
```

```
##
## Call:
## lm(formula = logAverage_Purchase_Price ~ Tractor_Power, data = Machinery_T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80020 -0.13471  0.01824  0.12945  0.49020
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.918847   0.339158  35.142  < 2e-16 ***
## Tractor_Power  0.023708   0.006234   3.803  0.00082 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2672 on 25 degrees of freedom
## Multiple R-squared:  0.3665, Adjusted R-squared:  0.3412
## F-statistic: 14.46 on 1 and 25 DF,  p-value: 0.00082
```

```
coef(summary(Model_2))
```

```
##                  Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)   11.91884732 0.339158430 35.142418 8.362159e-23
## Tractor_Power  0.02370773 0.006233778  3.803109 8.200380e-04
```

```
intercept <- coef(Model_2)[1]
slope <- coef(Model_2)[2]


cat("Estimated Regression Line Equation is given by log(y)i =", round(intercept, 8), "+", round
(slope, 8), "* x\n")
```

```
## Estimated Regression Line Equation is given by log(y)i = 11.91885 + 0.02370773 * x
```

alterbatively: *logAverage_Purchase_Price= 11.91884732 + 0.02370773(Tractor_Power)*

d. 2

The logAverage_purchse_price chances by R0.02370773 when tractor power increase by 1kw.

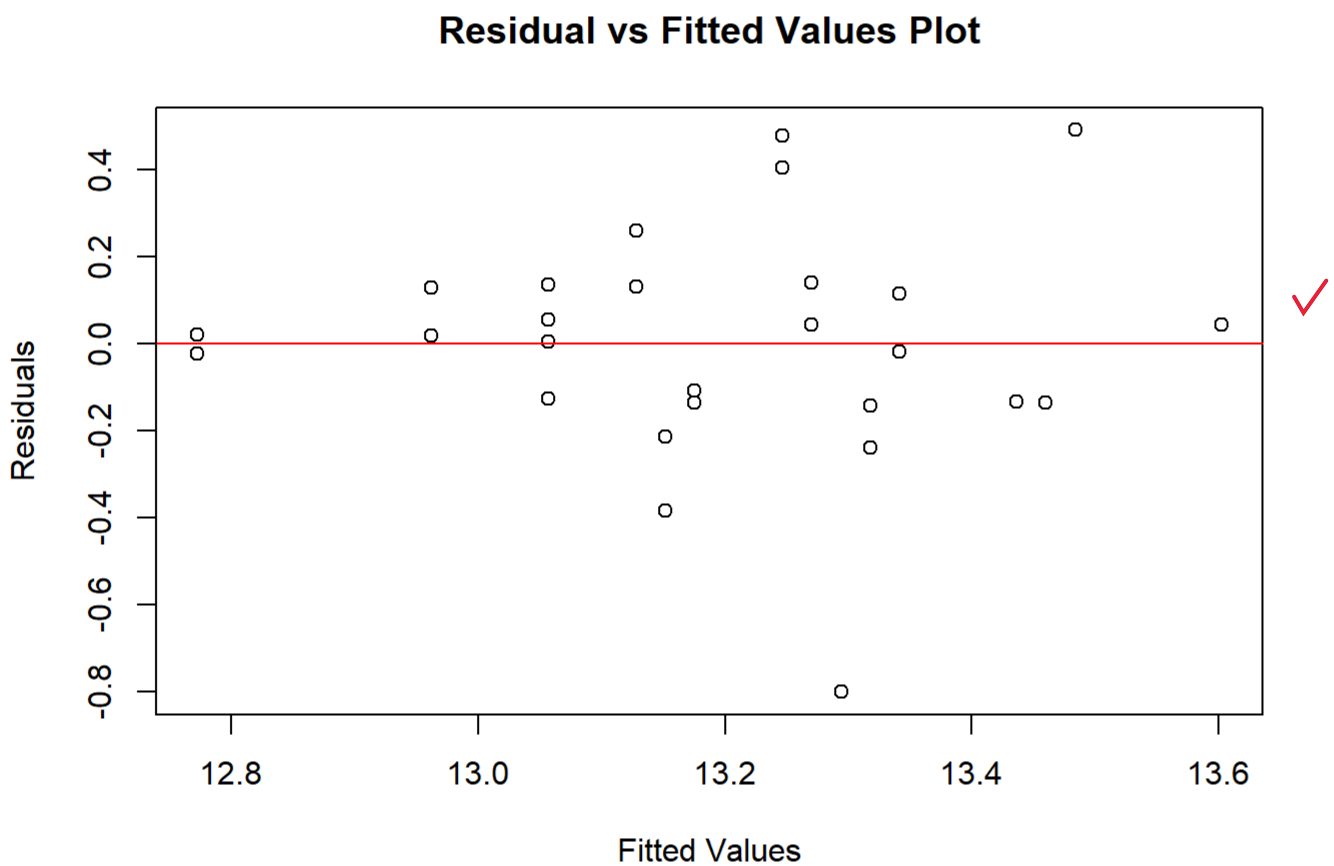
e. 3

As Tractor_Power increases by 1, the logAverage_Purchase_Price changes by log(Y)i= 11.91884732 + 0.02370773(1)= 11.94255505 .

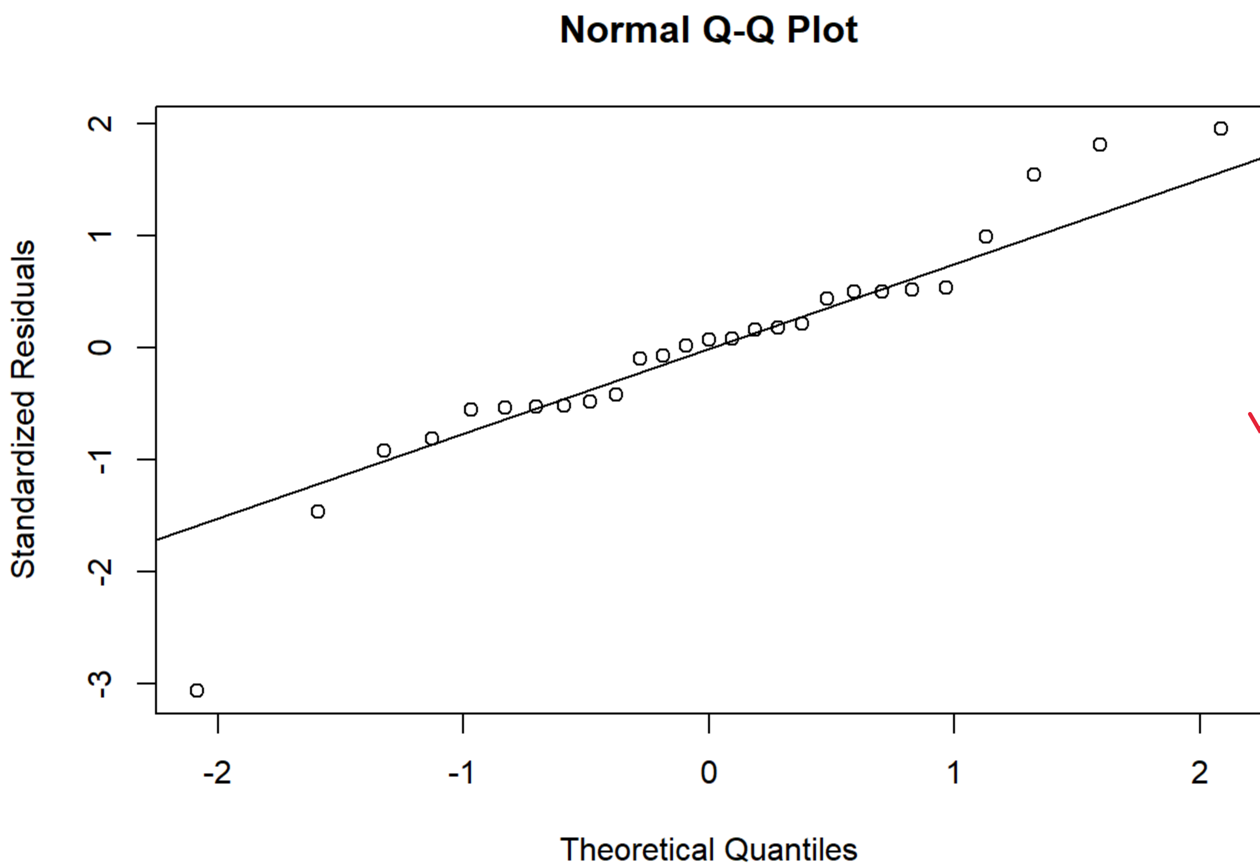
Therefore, Avarage_purchase_price changes by exp(11.94255505)=R153668.8206 when, tractor power is increase by 1 WK.

f. 3

```
# Obtain the fitted values and residuals
fitted_values <- fitted(Model_2)
residuals <- residuals(Model_2)
# Create the residual vs fitted values plot to assess the assumption of the constant varience of
the residual
plot(fitted_values, residuals, xlab = "Fitted Values", ylab = "Residuals", main = "Residual vs F
itted Values Plot")
abline(h = 0, col = "red")
```




```
# Obtain the standardized residuals for normality assumption assessment
std_resid <- rstandard(Model_2)
# Create the Normal Q-Q plot
qqnorm(std_resid, xlab = "Theoretical Quantiles", ylab = "Standardized Residuals", main = "Norma
l Q-Q Plot")
qqline(std_resid)
```



Linearity assessment: since the residuals are randomly scattered around the horizontal line at zero, it suggests that the linear regression mode_1 captures the linear relationship between the predictors and the response variable reasonably well. This indicates that the assumption of linearity is met.

Homoscedasticity/ constant Variance assessment: In the residual vs fitted values plot, we see a random scatter of points with approximately equal spread above and below zero. This indicates that the variance of the residuals is consistent across the range of fitted values/ predictor variables.

Independence of errors assessment: Observing from the Residual vs Fitted Values Plot, the error of one observation does not influence or correlate with the errors of other observations, this implies that each observation contributes unique information to the model and that the estimates of the regression coefficients are unbiased. There the assumption of independent errors is met.

Normality assessment: By examining the Normal Q-Q plot, we observe that the data points adhere closely to the expected regression line, suggesting that the assumption of normality is met.

Therefore As the diagnostic plots shows, the linear least square regression (LLSR) assumptions are satisfied in Model_2, however is seems that the presence of extreme large values may affect the constant variance assumption.

g. 9

```
as.factor(Machinery_T$Tractor_Types_dummy)
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## Levels: 0 1
```

```
Model_3<- lm(Machinery_T$Average_Purchase_Price~ Machinery_T$Tractor_Power + Machinery_T$Tractor
_Types_dummy)
summary(Model_3)
```

```
##
## Call:
## lm(formula = Machinery_T$Average_Purchase_Price ~ Machinery_T$Tractor_Power +
##     Machinery_T$Tractor_Types_dummy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -373071 -110850   22677   38741  421535
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -196061     207907  -0.943  0.35507
## Machinery_T$Tractor_Power          13775       3939   3.497  0.00186 **
## Machinery_T$Tractor_Types_dummy    36836      66127   0.557  0.58265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 163400 on 24 degrees of freedom
## Multiple R-squared:  0.3755, Adjusted R-squared:  0.3235
## F-statistic: 7.217 on 2 and 24 DF,  p-value: 0.003516
```

```
coef(Model_3)
```

wrtie the fitted model

```
##                     (Intercept)       Machinery_T$Tractor_Power
##                      -196060.95                        13774.94
## Machinery_T$Tractor_Types_dummy
##                        36835.91
```

Since the Intercept does not make sense, we can apply standardization transformation and get:

```
Machinery_T$Average_Purchase_Price_a <- Machinery_T$Average_Purchase_Price - mean(Machinery_T$Av
erage_Purchase_Price)
Machinery_T$Tractor_Power_a <- Machinery_T$Tractor_Power - mean(Machinery_T$Tractor_Power)
Machinery_T$Tractor_Types_dummy_1_a <- Machinery_T$Tractor_Types_dummy - mean(Machinery_T$Tracto
r_Types_dummy)
Model_3a <- lm(Average_Purchase_Price_a ~Tractor_Power_a + Machinery_T$Tractor_Types_dummy_1_a ,
data = Machinery_T)
summary(Model_3a)
```

```
##
## Call:
## lm(formula = Average_Purchase_Price_a ~ Tractor_Power_a + Machinery_T$Tractor_Types_dummy_1_
a,
##     data = Machinery_T)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -373071 -110850   22677   38741  421535
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -4.010e-11  3.144e+04   0.000  1.00000
## Tractor_Power_a                    1.377e+04  3.939e+03   3.497  0.00186 **
## Machinery_T$Tractor_Types_dummy_1_a  3.684e+04  6.613e+04   0.557  0.58265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 163400 on 24 degrees of freedom
## Multiple R-squared:  0.3755, Adjusted R-squared:  0.3235
## F-statistic: 7.217 on 2 and 24 DF,  p-value: 0.003516
```

```
intercept <- coef(Model_3a)[1]
slope <- coef(Model_3a)[2]
slope1 <- coef(Model_3a)[3]

cat("Now the Estimated Regression Line Equation is given y =", round(intercept, 8), "+", round(s
lope, 8),"* x1\n",round(slope1, 8), "*x2\n")
```

```
## Now the Estimated Regression Line Equation is given y = 0 + 13774.94 * x1
##  36835.91 *x2
```

✓

Interpretation

Model 3a:

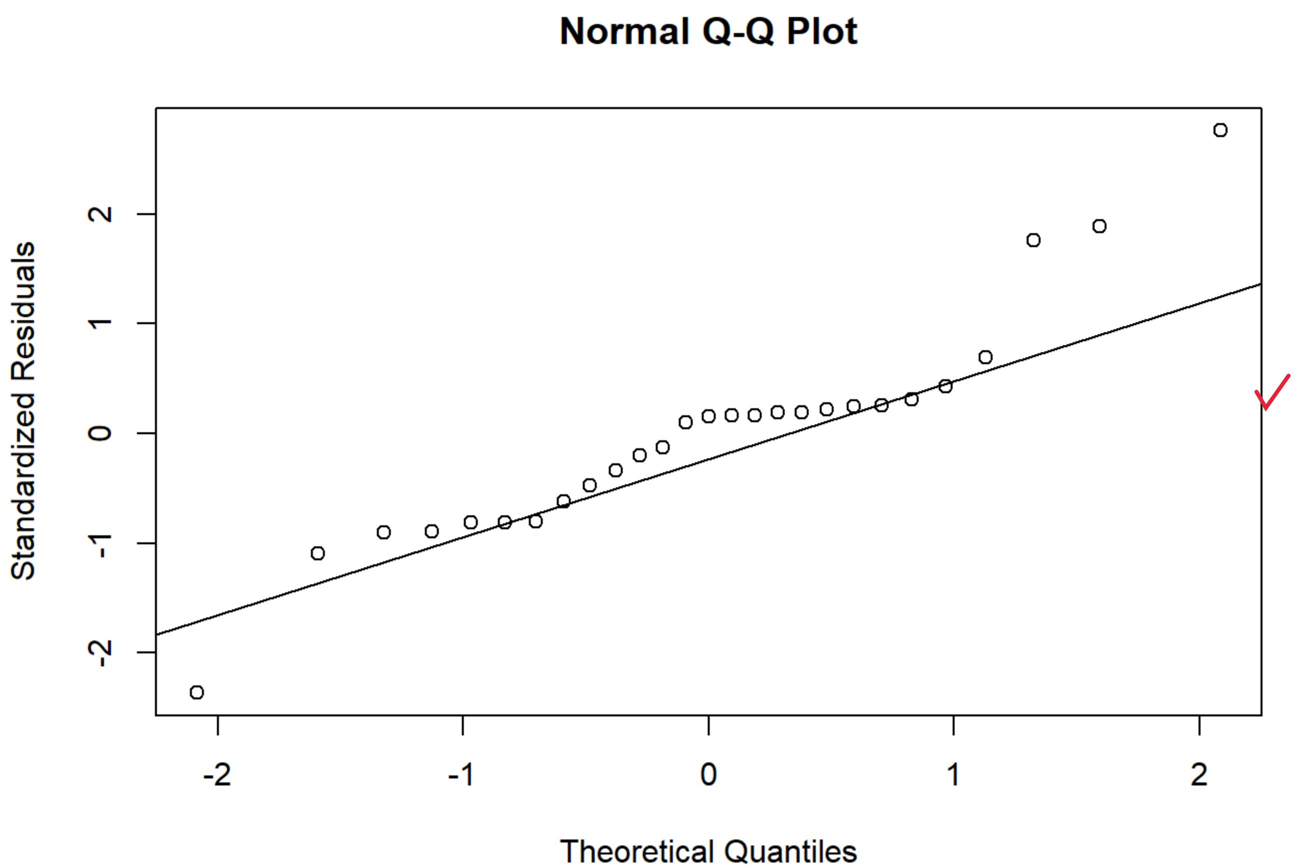
Intercept: The Average purchase price will remain at R0 when the tractor has no power and is of no type.

β1: Holding constant the effect of type_tractor, the Average purchase price is expected to increase by R13774.94 when tractor power is increased by 1KW ✓ ✓

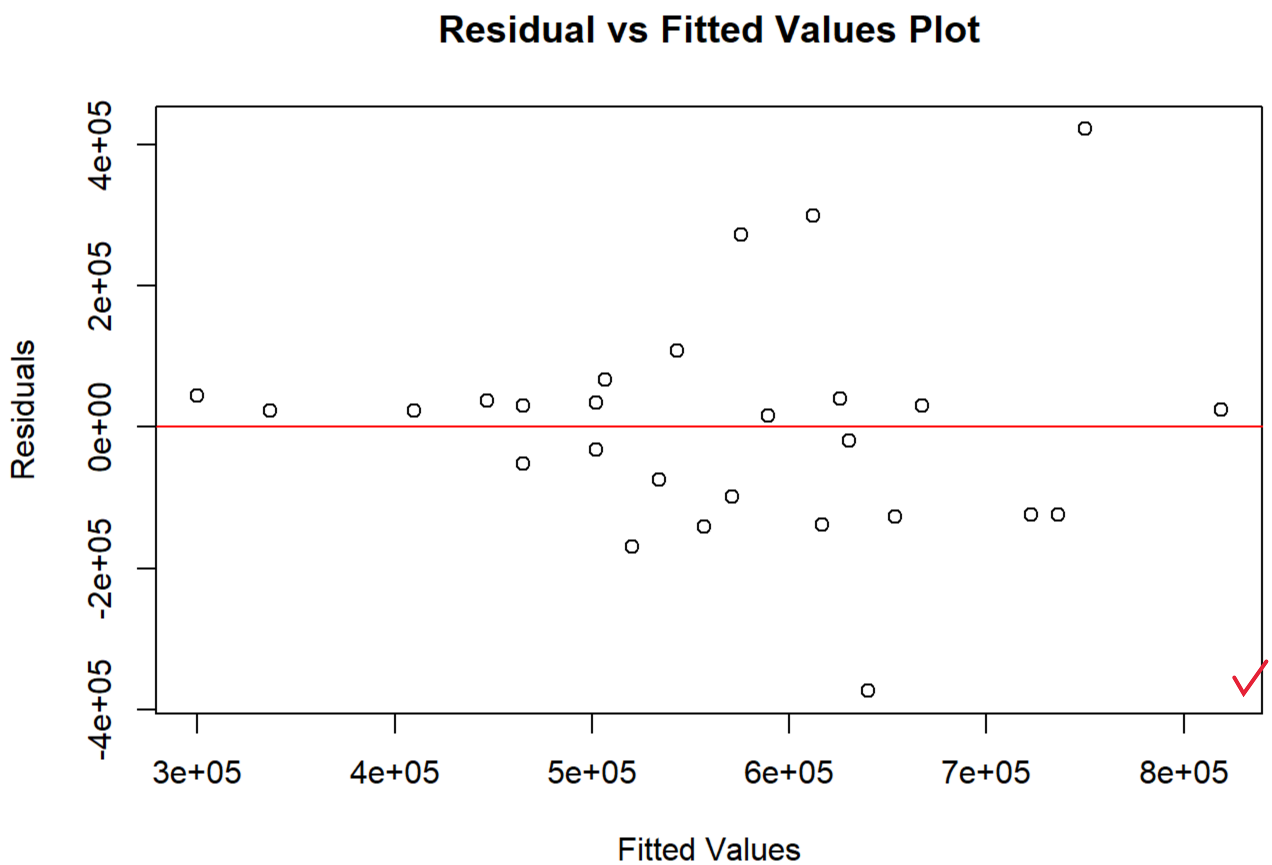
β2:Holding constant the effect of tractor power, the Average purchase price is expected to decrease by R36835.91 when type of a tractor is "2 WD" ✓ ✓

h. 2

```
# Obtain the standardized residuals for normality assumption assessment
std_resid <- rstandard(Model_3)
# Create the Normal Q-Q plot
qqnorm(std_resid, xlab = "Theoretical Quantiles", ylab = "Standardized Residuals", main = "Norma
l Q-Q Plot")
qqline(std_resid)
```




```
# Obtain the fitted values and residuals
fitted_values <- fitted(Model_3)
residuals <- residuals(Model_3)
# Create the residual vs fitted values plot to assess the assumption of the constant varience of
the residual
plot(fitted_values, residuals, xlab = "Fitted Values", ylab = "Residuals", main = "Residual vs F
itted Values Plot")
abline(h = 0, col = "red")
```



Linearity assessment: since the residuals are randomly scattered around the horizontal line at zero, it suggests that the linear regression mode_1 captures the linear relationship between the predictors and the response variable reasonably well. This indicates that the assumption of linearity is met.

Homoscedasticity/ constant Variance assessment: In the residual vs fitted values plot, we see a random scatter of points with approximately equal spread above and below zero. This indicates that the variance of the residuals is consistent across the range of fitted values/ predictor variables.

Independence of errors assessment: Observing from the Residual vs Fitted Values Plot, the error of one observation does not influence or correlate with the errors of other observations, this implies that each observation contributes unique information to the model and that the estimates of the regression coefficients are unbiased. There the assumption of independent errors is met.

Normality Assessment By examining the Normal Q-Q plot, we notice that the data points deviate from the expected regression line, indicating a violation of the normality assumption.

Therefore As the diagnostic plots shows, not all linear least square regression (LLSR) assumptions are satisfied in Model_3, however is seems that the presence of extreme large values affect the constant variance assumption.

i. 9

```
variables <- c("Average_Purchase_Price", "Fuel_Usage", "Salvage_Value", "Average_Investment", "D
epreciation_Costs", "Insurance_Licence_Costs", "Interest_Costs", "Repair_Maintenance_Costs")
correlation_matrix <- cor(Machinery_T[, variables])
cor(Machinery_T[, variables])
```

```
##                            Average_Purchase_Price Fuel_Usage Salvage_Value
## Average_Purchase_Price                 1.0000000  0.6524046     1.0000000
## Fuel_Usage                             0.6524046  1.0000000     0.6524043
## Salvage_Value                          1.0000000  0.6524043     1.0000000
## Average_Investment                     1.0000000  0.6524046     1.0000000
## Depreciation_Costs                     1.0000000  0.6524095     1.0000000
## Insurance_Licence_Costs                0.9999970  0.6517110     0.9999970
## Interest_Costs                         1.0000000  0.6523766     1.0000000
## Repair_Maintenance_Costs               1.0000000  0.6524099     1.0000000
##                            Average_Investment Depreciation_Costs
## Average_Purchase_Price             1.0000000          1.0000000
## Fuel_Usage                         0.6524046          0.6524095
## Salvage_Value                      1.0000000          1.0000000
## Average_Investment                 1.0000000          1.0000000
## Depreciation_Costs                 1.0000000          1.0000000
## Insurance_Licence_Costs            0.9999970          0.9999969
## Interest_Costs                     1.0000000          1.0000000
## Repair_Maintenance_Costs           1.0000000          1.0000000
##                            Insurance_Licence_Costs Interest_Costs
## Average_Purchase_Price                  0.9999970      1.0000000
## Fuel_Usage                              0.6517110      0.6523766
## Salvage_Value                           0.9999970      1.0000000
## Average_Investment                      0.9999970      1.0000000
## Depreciation_Costs                      0.9999969      1.0000000
## Insurance_Licence_Costs                 1.0000000      0.9999971
## Interest_Costs                          0.9999971      1.0000000
## Repair_Maintenance_Costs                0.9999971      1.0000000
##                            Repair_Maintenance_Costs
## Average_Purchase_Price                   1.0000000
## Fuel_Usage                               0.6524099
## Salvage_Value                            1.0000000
## Average_Investment                       1.0000000
## Depreciation_Costs                       1.0000000
## Insurance_Licence_Costs                  0.9999971
## Interest_Costs                           1.0000000
## Repair_Maintenance_Costs                 1.0000000
```

Average_Purchase_Price is perfectly correlated with Salvage_Value, Average_Investment, Depreciation_Costs, Insurance_Licence_Costs, Interest_Costs, and Repair_Maintenance_Costs. This high correlation indicates a strong linear relationship between these variables.

Fuel_Usage is moderately correlated (around 0.65) with Average_Purchase_Price, Salvage_Value, Average_Investment, Depreciation_Costs, Insurance_Licence_Costs, Interest_Costs, and Repair_Maintenance_Costs. This suggests a moderate linear relationship between Fuel_Usage and the other variables.

Based on the correlation matrix, if we want to add an explanatory variable to Model 3 to potentially increase the adjusted R-squared value, we should consider Fuel_Usage. It shows a moderate correlation with the target variable (Average_Purchase_Price) and other independent variables.

Explain the effects of including correlated variables with respect to LLSR assumptions.

```
Model_4<-lm(Machinery_T$Average_Purchase_Price ~ Machinery_T$Tractor_Power + Machinery_T$Tractor
_Types_dummy_1_a + Machinery_T$Fuel_Usage, data = Machinery_T)
summary(Model_4)
```

```
##
## Call:
## lm(formula = Machinery_T$Average_Purchase_Price ~ Machinery_T$Tractor_Power +
##     Machinery_T$Tractor_Types_dummy_1_a + Machinery_T$Fuel_Usage,
##     data = Machinery_T)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -351183  -88435   17233   54730  451015
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           -146333     186057  -0.786  0.43961
## Machinery_T$Tractor_Power              -65239      26731  -2.441  0.02278 *
## Machinery_T$Tractor_Types_dummy_1_a    49718      57535   0.864  0.39643
## Machinery_T$Fuel_Usage                 433696     145515   2.980  0.00669 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 141800 on 23 degrees of freedom
## Multiple R-squared:  0.5495, Adjusted R-squared:  0.4908
## F-statistic: 9.352 on 3 and 23 DF,  p-value: 0.0003143
```

```
#Normalizing Model4
Machinery_T$Average_Purchase_Price_b <- Machinery_T$Average_Purchase_Price - mean(Machinery_T$Av
erage_Purchase_Price)
Machinery_T$Tractor_Power_b <- Machinery_T$Tractor_Power - mean(Machinery_T$Tractor_Power)
Machinery_T$Tractor_Types_dummy_1_b <- Machinery_T$Tractor_Types_dummy - mean(Machinery_T$Tracto
r_Types_dummy)
Machinery_T$Fuel_Usage_b<- Machinery_T$Fuel_Usage -mean(Machinery_T$Fuel_Usage)
Model_4a <- lm(Average_Purchase_Price_a ~Tractor_Power_a + Machinery_T$Tractor_Types_dummy_1_a ,
data = Machinery_T)
summary(Model_4a)
```

```
##
## Call:
## lm(formula = Average_Purchase_Price_a ~ Tractor_Power_a + Machinery_T$Tractor_Types_dummy_1_
a,
##     data = Machinery_T)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -373071 -110850   22677   38741  421535
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -4.010e-11  3.144e+04   0.000  1.00000
## Tractor_Power_a                      1.377e+04  3.939e+03   3.497  0.00186 **
## Machinery_T$Tractor_Types_dummy_1_a  3.684e+04  6.613e+04   0.557  0.58265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 163400 on 24 degrees of freedom
## Multiple R-squared:  0.3755, Adjusted R-squared:  0.3235
## F-statistic: 7.217 on 2 and 24 DF,  p-value: 0.003516
```

```
intercept <- coef(Model_4a)[1]
slope <- coef(Model_4a)[2]
slope1 <- coef(Model_4a)[3]

cat("Estimated Regression Line Equation is given y =", round(intercept, 8), "+", round(slope,
8),"* x1\n", "+",round(slope1, 8), "*x2\n")
```

```
## Estimated Regression Line Equation is given y = 0 + 13774.94 * x1
##  + 36835.91 *x2
```

From the perspective of LLSR assumptions, adding *Fuel_Usage* to the model can be justified as long as it satisfies the assumptions of linearity, independence, homoscedasticity, normality, and no multicollinearity. While we cannot assess these assumptions solely based on the correlation matrix, evaluating them with appropriate diagnostic tests and analysis of the model residuals would be necessary.

Note: It's important to note that correlation alone does not guarantee a better model fit or higher adjusted R-squared value. Adding variables should be done cautiously, considering the theoretical relationship, domain knowledge, and statistical significance to ensure the model's interpretability and accuracy.

-