Relationships Between the Prices of Agricultural Commodities

Jeff Shelton Okanga'a

June 13, 2020

1 Introduction

In this document shows the relationship between the prices of two commodities,'Hard log Price' and 'Plywood Price', from the Agricultural commodities prices dataset extracted from Kaggle. We employed the use of a linear regression model to effectively compare the prices of the two commodities, alongside other analytics diagrams. The linear regression model equation used to represent the relationship is summarized as follows: $y = \beta_1 + \beta_2 x + \epsilon$.

Below is a sample of the first six records of both prices and their respective summaries:
Hard Log Price:

```
[1] 161.20 172.86 181.67 187.96 186.13 185.33

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  133.3   198.0   253.0   251.0   283.0   520.8
```

Plywood Price:

```
[1] 312.36 350.12 373.94 378.48 364.60 384.92

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  312.4   442.5   505.0   508.2   570.8   751.8
```

2 Graphical Analysis

In this section, we will build a simple regression model that will be used to predict the price of Plywood by establishing a statistically significant linear relationship with Hard Log prices. Therefore for this scenario, Plywood Price will be the dependent variable and Hard Log Price the independent variable.

- Scatter plot: Scatter plots helped us to visualize the linear relationship between the response,Plywood Price, and the predictor variable,Hard Log Price

- Box plot We utilize the boxplot to show the outliers in the data set, these are datapoints outside of the 1.5 * interquartile range (1.5*IQR). Where the interquartile range is calculated as the distance between the 25th percentile and 75th percentile values for each of the variables.
- Density plot We use the density plot to show how skewed our variables are compared to a normal curve.

3 Building the Linear Regression Model

3.1 Correlation Analysis

Correlation is a statistical measure that shows the degree of linear dependence between two variables. The closer it is to one the better the claim of a linear dependence. The correlation coefficient is:

```
[1] 0.8182176
```

We can therefore proceed to build a regression model for the two variables.

3.2 The linear Regression Equation

```
> linearMod<-lm(agri$Plywood.Price~agri$Hard.log.Price, data = cars)
> linearMod

Call:
lm(formula = agri$Plywood.Price ~ agri$Hard.log.Price, data = cars)

Coefficients:
        (Intercept)  agri$Hard.log.Price
            228.809                1.113
```

The mathematical model is therefore: $PlywoodPrice = 1.113 * HardLogprice + 228.809$

4 Linear Regression Diagnostics

We use this section of the document to determine how statistically significant our linear model is.

```
> summary(linearMod)

Call:
lm(formula = agri$Plywood.Price ~ agri$Hard.log.Price, data = cars)

Residuals:
```

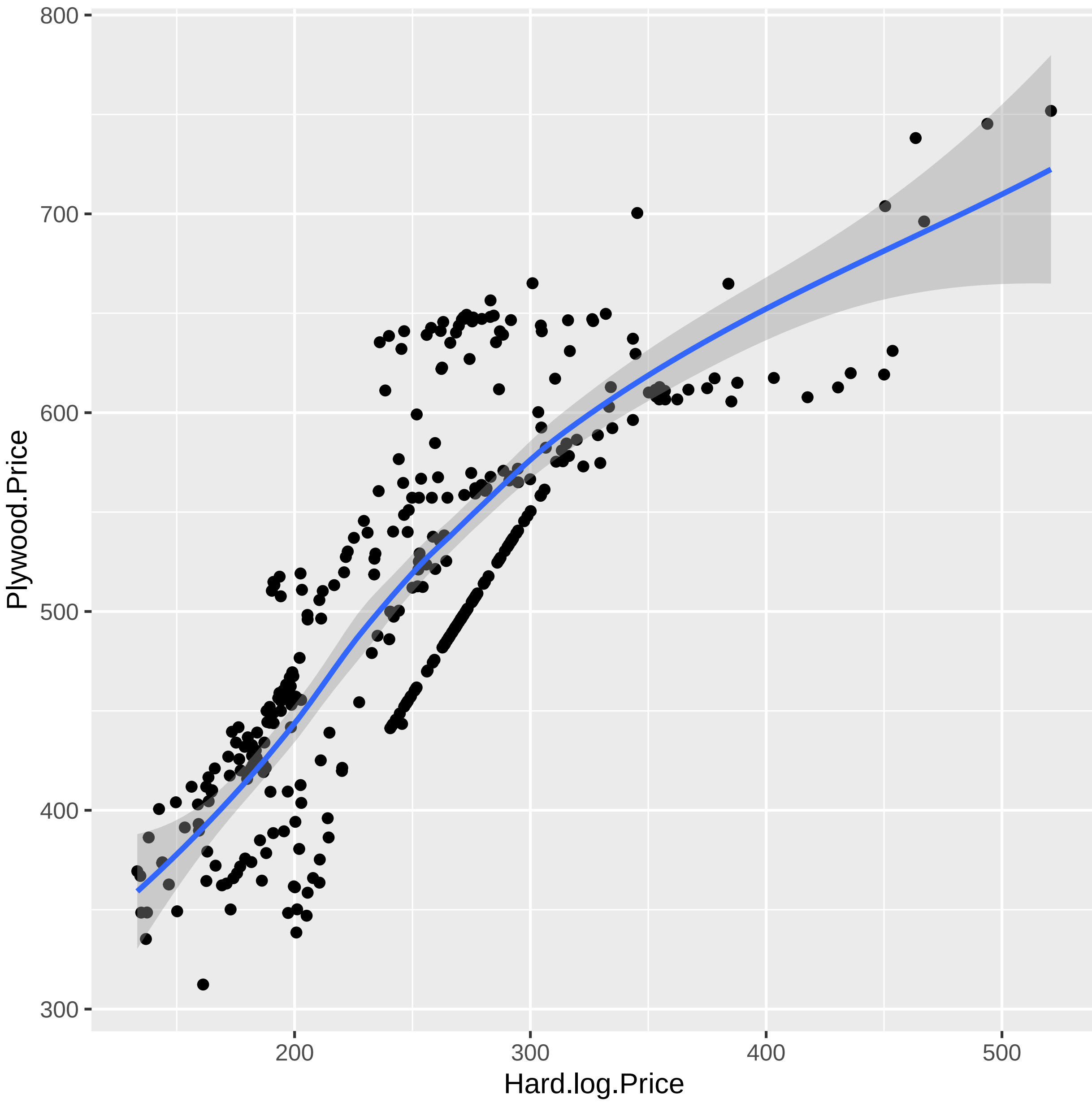

Figure 1: The Scatter plot

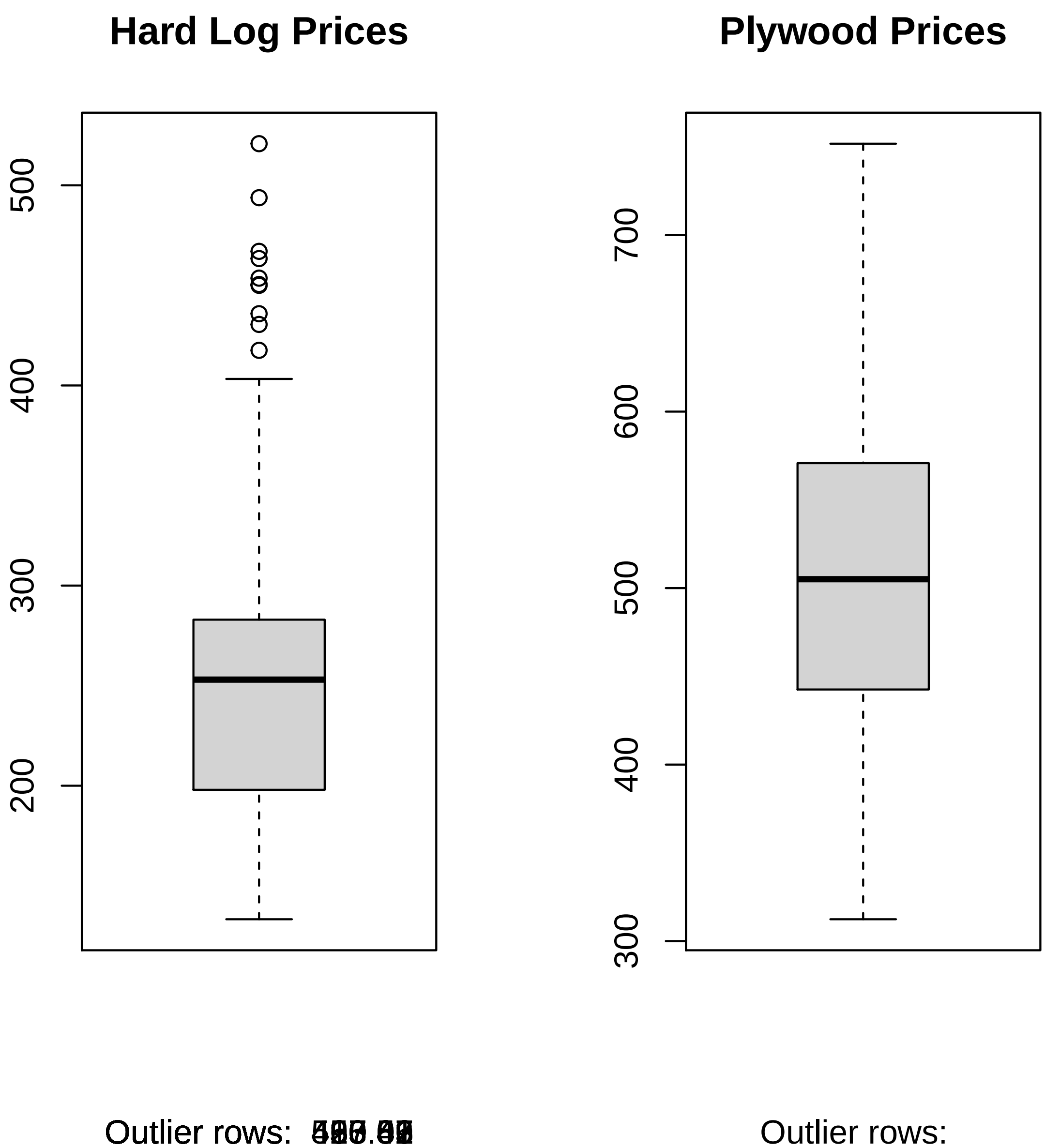

Figure 2: BoxPlot showing Outliers in the Datasets

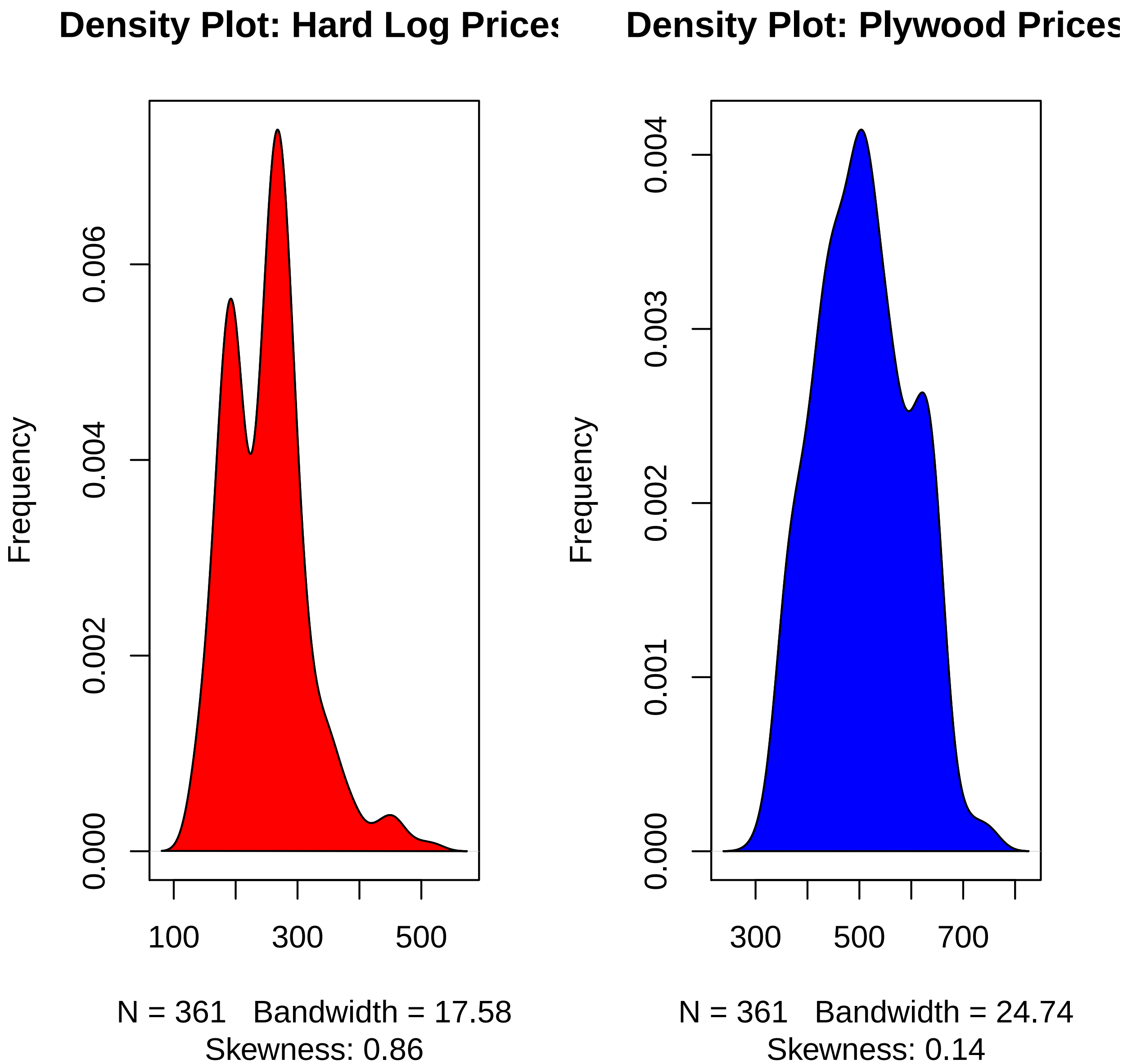

Figure 3: Density plots: Checking how skewed the dataset is

```
     Min       1Q   Median       3Q      Max
-113.729  -33.669   -7.803   20.795  143.804

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         228.80857   10.70874   21.37   <2e-16 ***
agri$Hard.log.Price   1.11303    0.04128   26.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51.4 on 359 degrees of freedom
Multiple R-squared:  0.6695,	Adjusted R-squared:  0.6686
F-statistic: 727.2 on 1 and 359 DF,  p-value: < 2.2e-16
```

The model is significant if the value of p is less than 0.05. When p Value is less than significance level (0.05), we can safely reject the null hypothesis that the co-efficient β of the predictor is zero.

5 Predicting Linear Models

To predict the values of the Plywood Price from that of the Hard Log Price we need to separate our data into training and test data. Training for our model to learn from and the test data to find out how accurate or efficiently it can learn from the provided dataset. We achieve this using the function sample().

```
> set.seed(100)
> trainingRowIndex<-sample(1:nrow(agri),0.8*nrow(agri))
> trainingData<-agri[trainingRowIndex,]
> testData<-agri[-trainingRowIndex,]
```

5.1 Index for the Training data

The index for the training data set.

```
> trainingRowIndex
```

```
  [1] 202 358 112 206   4 311 326  98   7 183 299 307 146 281 258 324  68  48
 [19] 288 341 167 272 116  93 301 158 336 221  87  95 223 220 251  31 182 297
 [37] 171 353 191 148  88 254  47 196  12 121  16 131 133  44 156 245 348  42
 [55] 143 185 298 154 280 137 250  55 323 291  26 233 344 255 118  37 222 219
 [73] 349  91  72 194 147 151 282 261 247 305 331 327 170 230 218 216 211 361
 [91] 268 100 201  71 149  39 193  82 136 197 335 210 199 177 228 130 139 238
[109] 319 114   1 340  32 269 125 316  13  62 294 270 186 142 277 322  64 207
[127]  15  20 178 128 253 352 289 126 102 217  53  84  11 205 150 333  52 232
[145] 342 159  78 204  46 192 285 243 214 320  74 213 264 356 315  69 135 184
```

```
[163] 165 339 287   5 224 108 360 198 234  56  36  38 200  43 260  61 273 293
[181] 239  14 127  58 110   3 241 263  76 208 300 115  25 295 168 266 275 346
[199]  41 332  19 354 256 309 190 140 179 189 302 129 330 111 141 175 163 271
[217] 107 173 337 187 188 278 174 160 248 274 103 155 157  73  28 317 145 283
[235] 166 357 119 172  40 359  18 113 153  90 345  17 181 310  99  80  23  97
[253] 169  85   2  22 229 313 306 329 303  77  50 318 123 321  63  83  33 195
[271] 152 162 347 292 203  51  21 246  45 180  94 138 284  57 355  96 227  59
```

5.2 Training Data and Testing Data

Data entries cannot displayed due to length of the dataset.

6 Fit the model on training data and predict dist on test data

Build the model

```
> lmMod<-lm(Plywood.Price ~ Hard.log.Price, data = trainingData)
> summary(lmMod)

Call:
lm(formula = Plywood.Price ~ Hard.log.Price, data = trainingData)

Residuals:
     Min       1Q   Median       3Q      Max 
-114.924  -34.540   -9.268   18.394  142.667 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    230.33838   12.02594   19.15   <2e-16 ***
Hard.log.Price   1.11136    0.04633   23.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51.98 on 286 degrees of freedom
Multiple R-squared:  0.668,	Adjusted R-squared:  0.6668 
F-statistic: 575.4 on 1 and 286 DF,  p-value: < 2.2e-16

> PlywoodPricePred<-predict(lmMod,testData)
```

6.1 Calculate prediction Accuracy and error rates

```
> actuals_preds<-data.frame(cbind(actuals=testData$dist, predicteds=PlywoodPricePred))
> actuals_preds
```

	predicteds
6	436.3068
8	429.2941
9	420.5255
10	418.3694
24	454.7887
27	458.2895
29	449.5542
30	458.7563
34	508.1451
35	548.9654
49	592.9086
54	563.6576
60	544.9645
65	492.2416
66	476.7603
67	490.2411
70	490.7190
75	514.6799
79	511.1458
81	490.1300
86	510.6012
89	501.7659
92	469.0475
101	382.5281
104	407.5448
105	412.0013
106	426.1823
109	425.0265
117	452.0437
120	442.4637
122	440.9412
124	442.6860
132	414.9909
134	413.7684
144	390.1631
161	432.0058
164	449.7876
176	454.9776
209	531.4392
212	540.8636
215	550.8547
225	593.3309
226	596.6983
231	540.6858

```
235   537.7629
236   532.5951
237   524.6599
240   508.0006
242   512.2460
244   535.8625
249   571.0148
252   602.3774
257   730.4839
259   714.7359
262   661.2684
265   623.4265
267   632.9953
276   579.0610
279   570.4258
286   548.8543
290   555.4113
296   514.9911
304   498.4763
308   500.2878
312   523.3152
314   534.0398
325   530.7835
328   524.4266
334   528.5164
338   531.9393
343   523.5708
350   531.0614
351   536.3959
```

6.2 Correlational Calculations

```
> correlation_accuracy<-cor(actuals_preds)
> correlation_accuracy

           predicteds
predicteds          1
```

7 Conclusion

Using a linear regression model we have effectively shown the relationship that exists between the prices of the two agricultural commodities plywood and hard wood. The regression is satisfied with the equation: $PlywoodPrice = 1.113 * HardLogprice + 228.809$