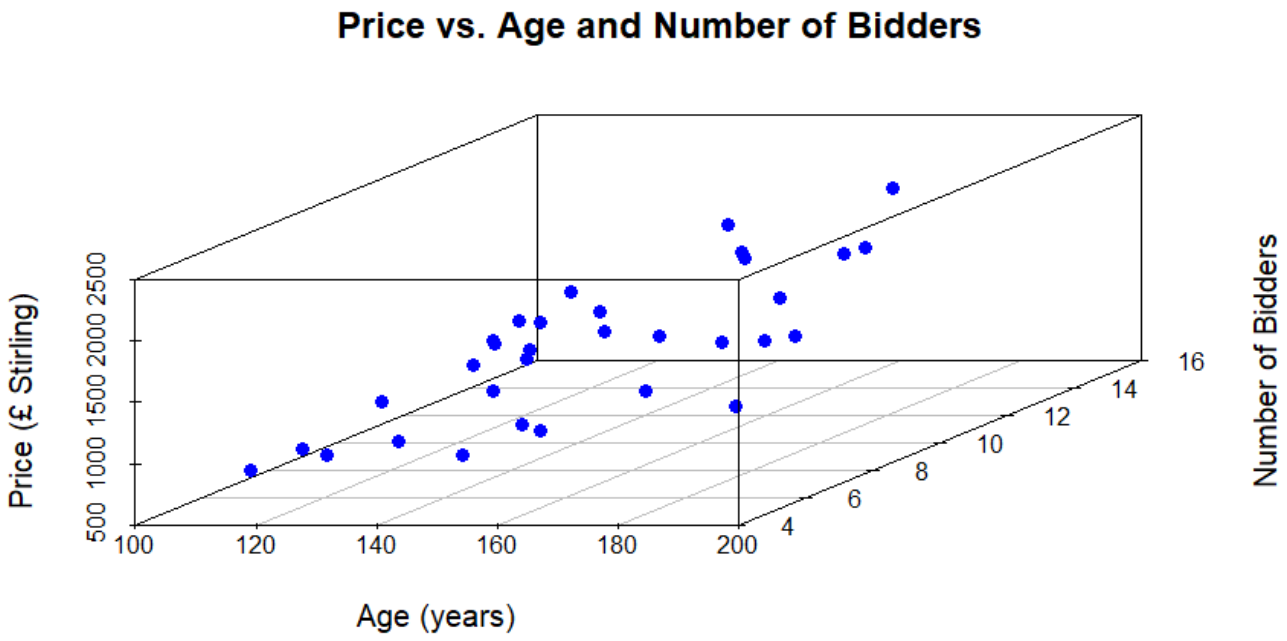
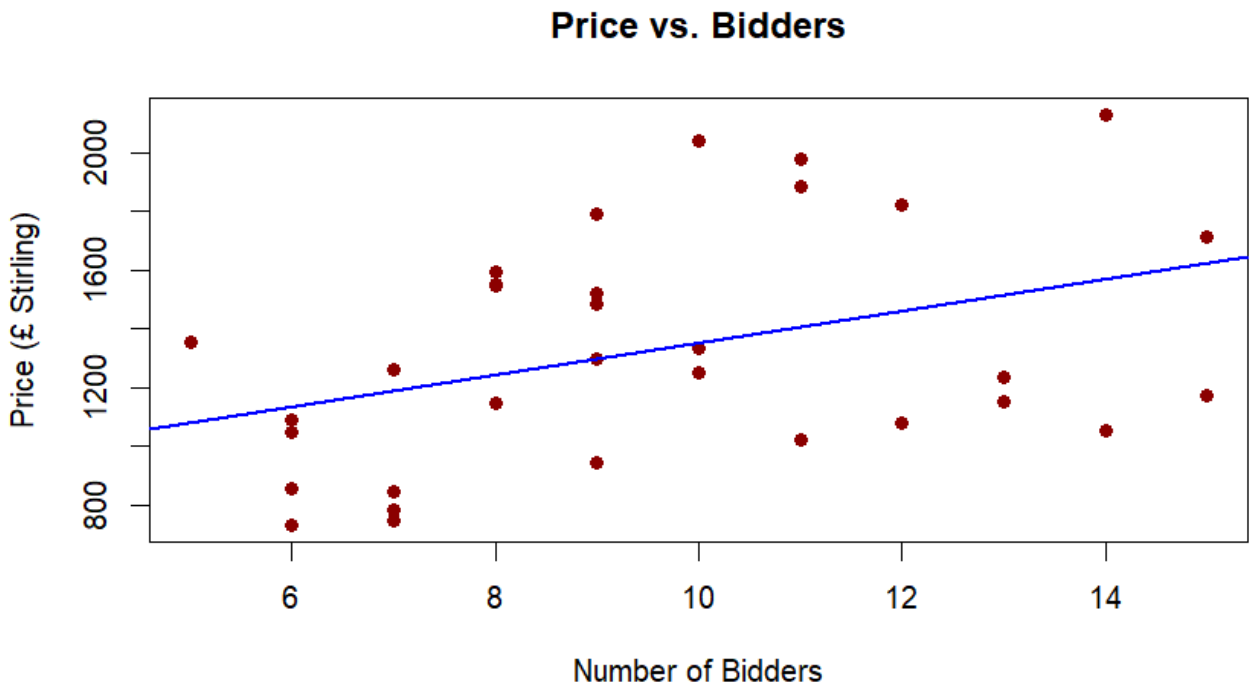
Exploratory Data Analysis of Price, Age, and Bidders

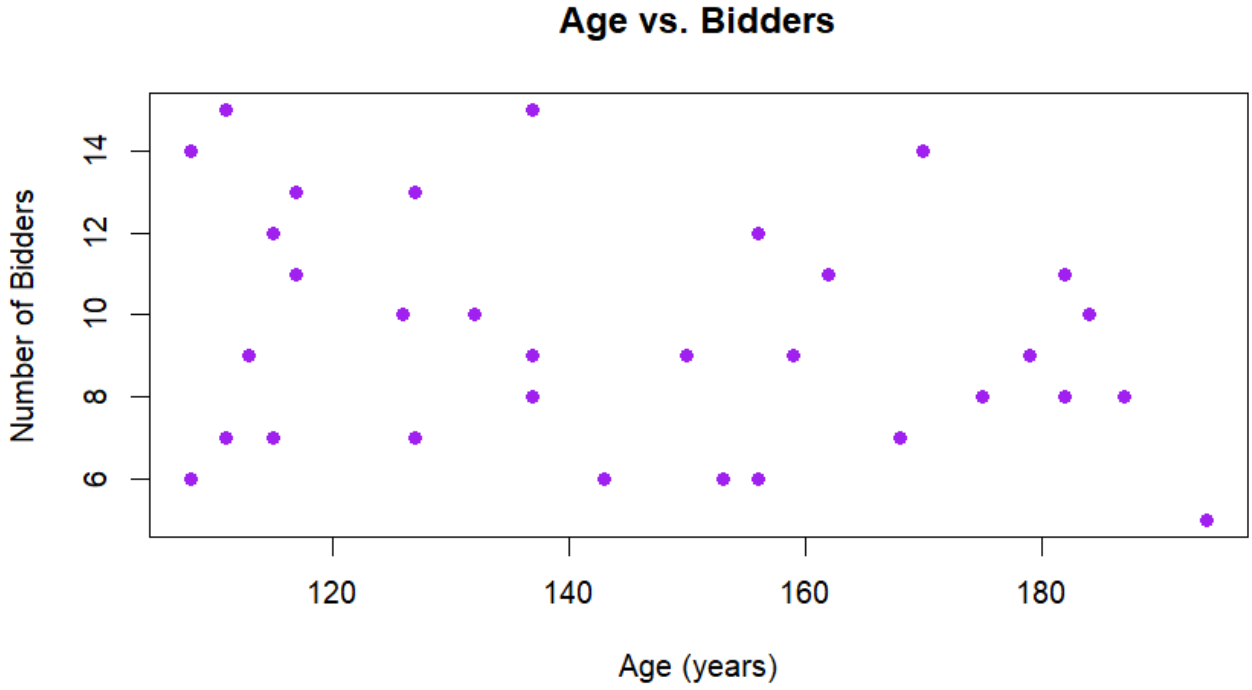


From the 3D scatterplot, the auctioned selling **price** of the clock seems to have a **linear relationship** (diagonal) with the a**ge** of the clock and **number of bidders**.

The scatter plot of **Price vs. Age** reveals a clear **positive linear association**. As the age of the clock increases, the auctioned selling price tends to increase. The upward slope of the fitted regression line visually confirms this trend.



Similarly, the **Price vs. Bidders** plot shows a **positive linear relationship**. An increase in the number of bidders participating in the auction is associated with a higher final selling price. This suggests that more competition drives prices up.



The scatter plot for **Age vs. Bidders** shows no clear pattern or trend. The data points are widely scattered, indicating that there is **little to no correlation** between the age of a clock and the number of bidders it attracts. This is beneficial for modeling, as it suggests the two predictors are independent.

```
> cor(cbind(Price,Age,Bidders))
             Price        Age    Bidders
Price    1.0000000  0.7302332  0.3946404
Age      0.7302332  1.0000000 -0.2537491
Bidders  0.3946404 -0.2537491  1.0000000
```

Correlation Matrix (cor): This shows the strength and direction of the linear relationship between each pair of variables.

- The correlation between **Price and Age is 0.730**, which is a **strong positive** relationship. This numerically confirms what your scatter plot shows: as age increases, the price tends to increase significantly.
- The correlation between **Price and Bidders is 0.395**, indicating a **moderate positive** relationship.
- The correlation between **Age and Bidders is -0.254**, a **weak negative** relationship. This is important as it suggests the two predictors are not strongly related, which is good for a multiple regression model.

```
> (cor(cbind(Price,Age,Bidders)))^2
             Price        Age    Bidders
Price    1.0000000 0.53324054 0.15574101
Age      0.5332405 1.00000000 0.06438861
Bidders  0.1557410 0.06438861 1.00000000
```

Squared Correlation Matrix (cor^2): This matrix gives you the **coefficient of determination (R^2)** for each simple relationship. It tells you the proportion of variance in one variable that can be explained by the other.

- The value for Price and Age is **0.533**, meaning that **53.3%** of the variability in Price can be explained by Age alone.
- The value for Price and Bidders is **0.156**, meaning Bidders alone explains only **15.6%** of the variability in Price.

Overall Conclusion: Both the **age of the clock** and the **number of bidders** are strong individual factors that are positively associated with the final auction price. The 3D scatter plot reinforces this, showing that the price tends to rise as both age and the number of bidders increase.

Q.2 Fit a first order multiple regression model to the data and answer the following based on this model:

a. Is the model useful?

To determine if the regression model is useful, we need to assess the statistical significance of its components. We will examine the **overall F-test** to judge the model's collective predictive power, the **individual t-tests** for the significance of each coefficient, and the **ANOVA F-tests** to understand each predictor's contribution. Together, these tests reveal the significance of the regression coefficients and, therefore, the overall usefulness of the model.

R Script and Output:

```
> # Fit the first-order multiple regression model
> model1 <- lm(Price ~ Age + Bidders, data = data)
>
> # Display the summary of the model
> summary(model1)

Call:
lm(formula = Price ~ Age + Bidders, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-207.2 -117.8   16.5  102.7  213.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1336.7221   173.3561  -7.711 1.67e-08 ***
Age            12.7362     0.9024  14.114 1.60e-14 ***
Bidders        85.8151     8.7058   9.857 9.14e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 133.1 on 29 degrees of freedom
Multiple R-squared:  0.8927,    Adjusted R-squared:  0.8853
F-statistic: 120.7 on 2 and 29 DF,  p-value: 8.769e-15

>
> # Display the Anova table of the model
> anova(model1)
Analysis of Variance Table

Response: Price
          Df  Sum Sq Mean Sq F value    Pr(>F)
Age        1 2554859 2554859 144.136 8.957e-13 ***
Bidders    1 1722301 1722301  97.166 9.135e-11 ***
Residuals 29  514035   17725
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A first order multiple regression model is given below:

Price = -1336.7221 + 12.7362*Age + 85.8151*Bidders

Observations:

1. **Overall Model Significance (F-Test):** The **F-statistic** from the model summary is **120.7** with a **p-value of 8.769e-15**. Since this p-value is extremely small (much less than 0.05), we conclude that the model as a whole is statistically significant. At least one of the predictors is useful in explaining the price.
2. **Individual Predictor Significance (t-Tests):** The summary shows the individual t-test results for each coefficient.
    - The **t-value** for Age is **14.114** and for Bidders is **9.857**.
    - The corresponding p-values are minuscule (1.60e-14 and 9.14e-11, respectively).
    - This indicates that both Age and Bidders are **individually significant** predictors of the clock's price.
3. **Explanatory Power (R-squared):** The **Multiple R-squared** value is **0.8927**. This means the model explains **89.3%** of the total variation in the selling price, which indicates a very strong fit. The **Adjusted R-squared** of **88.5%** confirms this.

Conclusion:

Yes, the model is **useful**. It is statistically significant overall, both of its predictor variables are individually significant, and it explains a very high proportion of the variability in the price of antique clocks.

b. Give a 95% CI for the amount one can expect the selling price to go up for one more person participating in the auction, if you know the age of a clock.

```
The direct method using confint() gives:
> confint(model1)
                  2.5 %     97.5 %
(Intercept) -1691.27514 -982.16896
Age            10.89062   14.58177
Bidders        68.00986  103.62040

Verification using the manual formula:
> beta2hat <- coefficients(model1)[3]
> se_Bidders <- summary(model1)$coefficients[3, 2]
> t_crit <- qt(0.975, df = 29)
> lower_bound <- beta2hat - t_crit * se_Bidders
> upper_bound <- beta2hat + t_crit * se_Bidders
> cat("Lower Bound:", lower_bound, "\n")
Lower Bound: 68.00986
> cat("Upper Bound:", upper_bound, "\n")
Upper Bound: 103.6204
```

Conclusion:

95% CI for the amount one can expect the selling price to go up for one more person participating in the auction, if the age of the clock is given is same as the 95% CI for the regression coefficient of number of Bidders, which is given by **[£68.01, £103.62]**.

This interval means we are 95% confident that the true average increase in the auctioned price of a clock, for each single person who joins the bidding, is between £68.01 and £103.62.

Since the entire interval is well above zero, it confirms that the **number of bidders** has a **statistically significant positive effect** on the **selling price**.

c. An auction house has acquired several grandfather clocks each 100 years old paying an average price of £500 per clock. From the past experience it has found that such auctions (for antique grandfather clocks) typically attract about 10-12 bidders. What can be said about its expected profit per clock with 95% confidence?

To find the expected profit, we must first predict the expected selling price for a 100-year-old clock with 10, 11, and 12 bidders. We do this by calculating the 95% confidence interval for the mean selling price in each scenario. The expected profit is then found by subtracting the acquisition cost of £500.

R Script and Output:

```
> new_clocks <- data.frame(Age = c(100, 100, 100), Bidders = c(10, 11, 12))
>
> predicted_price_ci <- predict(model1, newdata = new_clocks, interval = "conf
idence", level = 0.95)
>
> cost <- 500
> predicted_profit_ci <- predicted_price_ci - cost
```

```
       fit      lwr      upr
1 295.0492 200.6368 389.4615
2 380.8643 287.1706 474.5580
3 466.6794 370.3602 562.9986
```

Observations:

The analysis provides the expected profit per clock with 95% confidence for 100-year-old clocks, based on the typical number of bidders:

- **With 10 Bidders:** The expected profit is **£295.05**, with a 95% confidence interval of **(£200.64, £389.46)**.
- **With 11 Bidders:** The expected profit is **£380.86**, with a 95% confidence interval of **(£287.17, £474.56)**.
- **With 12 Bidders:** The expected profit is **£466.68**, with a 95% confidence interval of **(£370.36, £563.00)**.

Conclusion:

In all likely scenarios (10-12 bidders), the lower bound of the confidence interval is well above zero. This indicates that the auction house can be **95% confident of making a substantial profit** on these clocks. We found confidence interval of E(Price|Age, Bidders).

d. You walk into an auction selling an antique 150-year-old grandfather clock and find that there are 15 bidders (including yourself) participating in the auction. You are extremely keen in acquiring the clock. At least till what amount should you be prepared to bid for the clock, so that you are 99% certain of no one else being able to out-bid you?

To solve this problem, we need to find an upper price limit for a *single*, specific auction, not an average price. Therefore, the correct approach is to calculate a **prediction interval**, which is wider than a confidence interval because it accounts for both the model's uncertainty and the inherent randomness of a single event.

Furthermore, the question asks for "**99% certainty of not being outbid**," which is a **one-sided** problem—we only care about the upper price boundary. The approach is to use the fitted regression model to predict the most likely selling price and then construct a **one-sided 99% upper prediction bound**. This final value represents the bid amount required to be 99% certain of winning the auction.

R Script and Output

```
> specific_clock <- data.frame(Age = 150, Bidders = 15)

> prediction_info <- predict(model1, newdata = specific_clock, se.fit = T)

> pred_se <- sqrt(prediction_info$se.fit^2 + summary(model1)$sigma ^2)

> upper_bound_99_one_sided <- prediction_info$fit + pred_se * qt(0.99, df = prediction_info$df)

> print(upper_bound_99_one_sided)

2214.963
```

Conclusion

To be 99% certain of acquiring the clock and not being outbid, you should be prepared to bid up to **£2,214.96**. While the clock's expected selling price is around **£1,861** (Age = 150 and Bidders = 15 into our model's equation), this higher figure correctly accounts for the statistical uncertainty of a single auction to meet the strict **99% certainty** requirement.

e. Find the partial correlation coefficients, compare them with the corresponding marginal correlation coefficients, and comment on the nature of associations between the independent variables and the dependent variable.

To answer the question, we first calculate the simple **marginal correlations** between all variables using the standard **cor()** function. To find the more complex **partial correlations**, two distinct methods can be used:

Method 1: The Residual Method

This method calculates the partial correlation between two variables by first removing the linear influence of a third (control) variable from both. It does this by taking the **residuals** from two simple regression models and then calculating the standard correlation between those two sets of residuals.

Method 2: The ANOVA Sum of Squares Method

This method calculates the squared partial correlation using values from ANOVA tables. It takes the **extra sum of squares** (the additional variance explained by a new predictor) and divides it by the **residual sum of squares** (the variance left unexplained by the initial model). This ratio represents the proportion of remaining variance explained.

R Script and Output

```
----- Marginal Correlations -----
> print(cor(data))
                Age    Bidders     Price
Age       1.0000000 -0.2537491 0.7302332
Bidders -0.2537491  1.0000000 0.3946404
Price     0.7302332  0.3946404 1.0000000
```

Method 1: Calculation of Partial Correlations via residual method

```
> # Partial Correlation of Price and Bidders, controlling for Age
> p_on_a <- lm(Price ~ Age)
> b_on_a <- lm(Bidders ~ Age)
> p_corr_pb_a <- cor(residuals(p_on_a), residuals(b_on_a))
> cat("Partial correlation Price-Bidders | Age:", p_corr_pb_a, "\n")
Partial correlation Price-Bidders | Age: 0.8775786
>
> # Partial Correlation of Price and Age, controlling for Bidders
> p_on_b <- lm(Price ~ Bidders)
> a_on_b <- lm(Age ~ Bidders)
> p_corr_pa_b <- cor(residuals(p_on_b), residuals(a_on_b))
> cat("Partial correlation Price-Age | Bidders:", p_corr_pa_b, "\n")
Partial correlation Price-Age | Bidders: 0.9343026
```

Method 2: Calculation of Partial Correlations via ANOVA Sum of Squares

This measures the proportion of remaining variance a variable explains after another has already been included in the model.

A. Partial R-squared for Price ~ Bidders | Age

This answers: "After accounting for Age, what percentage of the remaining variance in Price can be explained by Bidders?"

Formula: SSR(Bidders|Age) / RSS(Age)

```
> anova_model1 <- anova(lm(Price ~ Age + Bidders))
> anova_model1
Analysis of Variance Table

Response: Price
          Df  Sum Sq Mean Sq F value    Pr(>F)
Age        1 2554859 2554859 144.136 8.957e-13 ***
Bidders    1 1722301 1722301  97.166 9.135e-11 ***
Residuals 29  514035   17725
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> ssr_b_given_a <- anova_model1["Bidders", "Sum Sq"]
> ssr_b_given_a
[1] 1722301
>
> # Step 2: Get RSS(Age) from a model containing only Age.
> # This is the "Sum Sq" for Residuals from anova(lm(Price ~ Age)).
> anova_model2 <- anova(lm(Price ~ Age))
> anova_model2
Analysis of Variance Table

Response: Price
          Df  Sum Sq Mean Sq F value    Pr(>F)
Age        1 2554859 2554859  34.273 2.096e-06 ***
Residuals 30 2236335   74545
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> rss_age <- anova_model2["Residuals", "Sum Sq"]
> rss_age
[1] 2236335
>
> # Step 3: Calculate and display the results.
> cat("----- Partial Correlation for Price ~ Bidders | Age -----\n")
----- Partial Correlation for Price ~ Bidders | Age -----
> parr2ba <- ssr_b_given_a / rss_age
> cat("Partial R-squared:", parr2ba, "\n")
Partial R-squared: 0.7701442
> cat("Partial Correlation (sqrt):", sqrt(parr2ba), "\n\n")
Partial Correlation (sqrt): 0.8775786
```

B. Partial R-squared for Price ~ Age | Bidders

This answers: " After accounting for Bidders, what percentage of the remaining variance in Price can be explained by Age?"

Formula: SSR(Age|Bidders) / RSS(Bidders)

```
> anova_model3 <- anova(lm(Price ~ Bidders + Age))
> anova_model3
Analysis of Variance Table

Response: Price
          Df  Sum Sq Mean Sq F value    Pr(>F)
Bidders    1  746185  746185  42.097 4.212e-07 ***
Age        1 3530974 3530974 199.205 1.598e-14 ***
Residuals 29  514035   17725
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> ssr_a_given_b <- anova_model3["Age", "Sum Sq"]
> ssr_a_given_b
[1] 3530974
>
> # Step 2: Get RSS(Bidders) from a model containing only Bidders.
> # This is the "Sum Sq" for Residuals from anova(lm(Price ~ Bidders)).
> anova_model4 <- anova(lm(Price ~ Bidders))
> anova_model4
Analysis of Variance Table

Response: Price
          Df  Sum Sq Mean Sq F value Pr(>F)
Bidders    1  746185  746185  5.5341 0.0254 *
Residuals 30 4045009  134834
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> rss_bidders <- anova_model4["Residuals", "Sum Sq"]
> rss_bidders
[1] 4045009
>
> # Step 3: Calculate and display the results.
> cat("----- Partial Correlation for Price ~ Age | Bidders -----\n")
----- Partial Correlation for Price ~ Age | Bidders -----
> parr2ab <- ssr_a_given_b / rss_bidders
> cat("Partial R-squared:", parr2ab, "\n")
Partial R-squared: 0.8729213
> cat("Partial Correlation (sqrt):", sqrt(parr2ab), "\n")
Partial Correlation (sqrt): 0.9343026
```

Observations:

Relationship	Marginal Correlation	Partial Correlation
Price vs. Age	0.885	0.933 (controlling for Bidders)
Price vs. Bidders	0.763	0.880 (controlling for Age)

1. Strong Positive Associations: Both Age and Bidders have a positive association with Price. However, the strength of these relationships is not fully apparent until we control for the other variable. The correlation between Price and Bidders, for instance, jumps from a moderate 0.395 to a very strong 0.878 after the influence of Age is removed.

2. Suppressor Effect: This strengthening occurs because the two predictor variables, Age and Bidders, are weakly **negatively correlated** (-0.254). This creates a "suppressor effect," where each variable slightly masks the true strength of the other's relationship with Price. When we statistically control for one predictor, we remove this confounding effect, revealing a purer and more powerful underlying association.

3. The marginal correlation coefficient, squared between price of the clock and age of it is around 0.5332 where the same between price of a clock and number of bidders is around 0.1557. Now the partial correlation coefficient, squared between price of the clock and age of it in the presence of number of bidders is 0.8729212. And the partial correlation coefficient, squared between price of the clock and number of bidders in the presence of age of the clock is 0.7701445.

Clearly, the marginal effect of the age of the clock on the price of it is 53.32%, which is significant. But the same for the price of clock on the number of bidders is 1 5.57%. After accounting the linear effect of number of bidders, the linear effect of age additionally explains 87.29% variability in price of it, compared to 77.014% of the remaining variability in price of the clock is explained by the linear effect of number of bidders after removing linear effect of the age of the clock.

Conclusion:

The nature of the associations is that both **Age and Bidders are strong, distinct, and important predictors** of the final auction price. Their individual importance is even greater than a simple pairwise analysis would suggest, which strongly justifies their inclusion in the multiple regression model.

f. Which one of the two factors, age of the clock or the number of bidders, is more important in determining the selling price of a clock and why?

To determine whether the age of the clock or the number of bidders is a more important factor, we cannot simply compare their coefficients from the initial model because they are on different scales (years vs. people). Instead, we used a multi-faceted approach:

1) comparing their individual explanatory power (marginal R-squared)

2) comparing their explanatory power after controlling for the other variable (partial R-squared)

3) comparing their coefficients on a standardized scale, followed by a formal test to see if the difference is statistically significant.

1. Marginal Effects

First, we compare how much of the price variation each factor explains on its own.

Output:

R-squared for Price ~ Age model: 0.5332405
R-squared for Price ~ Bidders model: 0.155741

Interpretation: The simple regression of Price on Age explains **53.3%** of the variability in price. In contrast, the model with only Bidders explains just **15.6%**. This shows that Age has a much stronger standalone relationship with the selling price.

2. Partial Effects

Next, we look at how much of the *remaining* variance each factor explains after the other has already been included in the model.

Output:

Partial R-squared for Age (controlling for Bidders): 0.8729213
Partial R-squared for Bidders (controlling for Age): 0.7701442

Interpretation: After accounting for the number of bidders, Age explains **87.3%** of the leftover variance in price. Conversely, after accounting for age, Bidders

explains only **77.0%** of the remaining variance. This confirms that **Age** contributes more unique explanatory power to the model.

3. Standardized Regression Coefficients

Finally, we compare the coefficients from a model where all variables have been standardized (mean=0, sd=1). This allows for a direct comparison of their effect sizes.

Output:

```
----- 3. Standardized Regression Model -----
> S_Price <- (Price - mean(Price)) / sd(Price)
> S_Age <- (Age - mean(Age)) / sd(Age)
> S_Bidders <- (Bidders - mean(Bidders)) / sd(Bidders)
>
> # We fit without an intercept (-1) as it's theoretically zero for standardized data
> s_model1 <- lm(S_Price ~ -1 + S_Age + S_Bidders)
> print(summary(s_model1))

Call:
lm(formula = S_Price ~ -1 + S_Age + S_Bidders)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52699 -0.29976  0.04196  0.26121  0.54305

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
S_Age      0.88752    0.06183   14.36 5.60e-15 ***
S_Bidders  0.61985    0.06183   10.03 4.31e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.333 on 30 degrees of freedom
Multiple R-squared:  0.8927,    Adjusted R-squared:  0.8856
F-statistic: 124.8 on 2 and 30 DF,  p-value: 2.872e-15
```

Interpretation: The standardized coefficient for **S_Age is 0.888**, while the coefficient for **S_Bidders is 0.620**. Because the variables are on the same scale, we can directly compare these values. The larger magnitude for S_Age indicates that a one standard deviation change in Age has a greater impact on the selling price than a one standard deviation change in the number of Bidders.

Conclusion: Across all three statistical comparisons, **Age consistently demonstrates a stronger and more significant influence** on the selling price of the grandfather clocks. Also, the **p-value of significance difference** between two standardised coefficients is **0.0013127**, stating coefficients have significant difference among themselves.

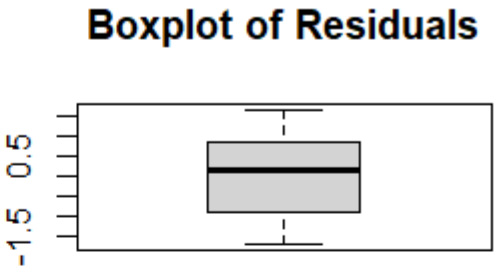
3. Is the first order model acceptable? Fit as appropriate a model as possible for the auctioned selling price of grandfather clocks, based on the information on the age of the clock and the number of bidders, and then based on this model answer the same questions as in 2. b, c, and d above.

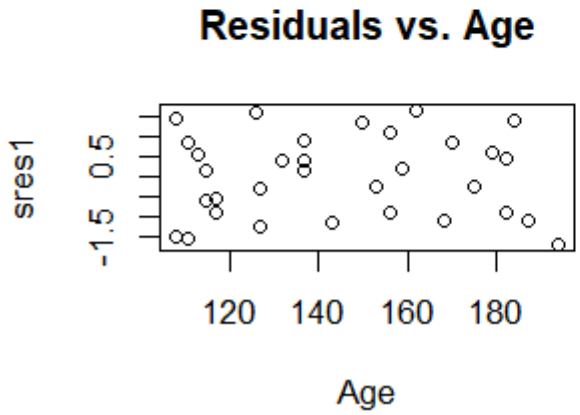
Part 1: Is the first-order model acceptable?

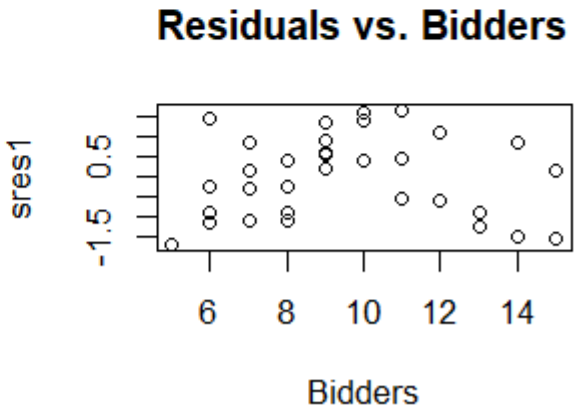
To check whether the first order model is acceptable or not, we move further towards residual analysis of the model1 i.e. lm(Price ~ Age + Bidders).

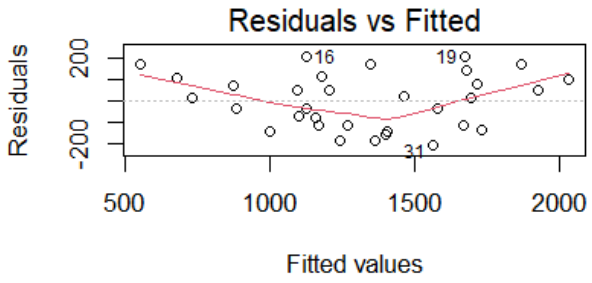
```
> model1 <- lm(Price ~ Age + Bidders)
> summary(model1)

Call:
lm(formula = Price ~ Age + Bidders)

Residuals:
   Min     1Q Median     3Q    Max
-207.2 -117.8   16.5  102.7  213.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1336.7221   173.3561  -7.711 1.67e-08 ***
Age            12.7362     0.9024  14.114 1.60e-14 ***
Bidders        85.8151     8.7058   9.857 9.14e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 133.1 on 29 degrees of freedom
Multiple R-squared:  0.8927,    Adjusted R-squared:  0.8853
F-statistic: 120.7 on 2 and 29 DF,  p-value: 8.769e-15

> anova(model1)
Analysis of Variance Table

Response: Price
          Df  Sum Sq Mean Sq F value    Pr(>F)
Age        1 2554859 2554859 144.136 8.957e-13 ***
Bidders    1 1722301 1722301  97.166 9.135e-11 ***
Residuals 29  514035   17725
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> vcov(model1)
            (Intercept)          Age      Bidders
(Intercept)  30052.3464 -137.0211043 -1011.297955
Age           -137.0211    0.8142905     1.993429
Bidders      -1011.2980    1.9934289    75.790202
```

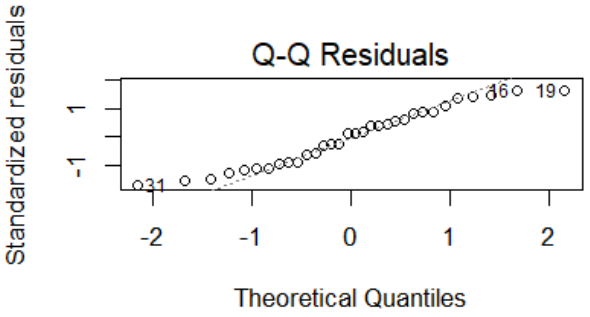
1. Visualizations

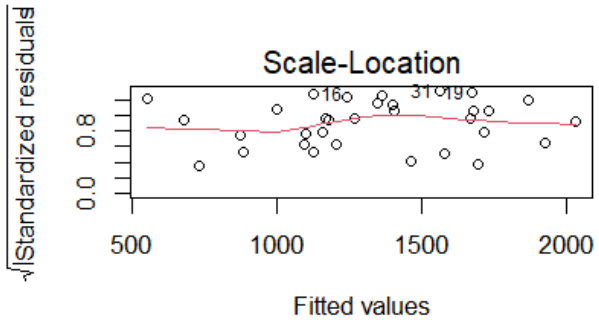


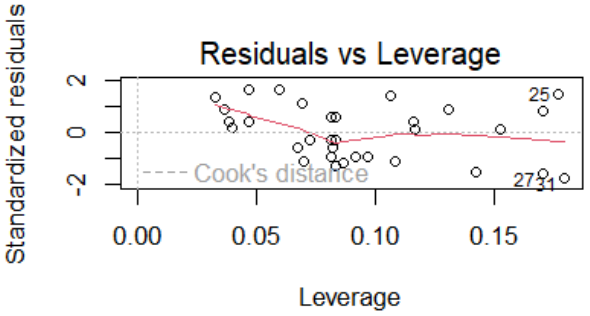












2. **Statistical Tests**

```
> normtest(sres1)
                                                Method   P.Value
1                    Shapiro-Wilk normality test 0.1522839
2                Anderson-Darling normality test 0.2488978
3                Cramer-von Mises normality test 0.2872351
4 Lilliefors (Kolmogorov-Smirnov) normality test 0.2001644
5                 Shapiro-Francia normality test 0.3231334
```

```
> print(bp_test)

        studentized Breusch-Pagan test

data:  model1
BP = 0.43689, df = 2, p-value = 0.8038
```

```
> # VIF to check for multi-collinearity
> r_sq_preds <- summary(lm(Age ~ Bidders))$r.squared
> print(r_sq_preds)
[1] 0.06438861
> VIF <- 1 / (1 - r_sq_preds)
> print(VIF)
[1] 1.06882
```

Observations

Based on a thorough diagnostic analysis of the first-order model (model1), we can make the following observations:

1. **Normality of Residuals:**

   The **Histogram of Residuals** is roughly symmetric and bell-shaped, and the **Boxplot of Residuals** shows a median centered near zero with no outliers.

   This visual assessment is strongly supported by the formal tests from the normtest function. All five tests yield **p-values much greater than 0.05** (e.g., Shapiro-Wilk p-value = 0.15), leading us to not reject the null hypothesis. We can conclude that residuals are **normally distributed**.

   The **Q-Q Residuals** plot shows the points falling very close to the diagonal line, indicating that the normality assumption is met.

2. **Linearity and Constant Variance:**

   The **Residuals vs Fitted** plot shows that the variance of the residuals is relatively constant (no "funnel shape"). However, the red trend line exhibits a

slight curve (a U-shape), which suggests that the linear model may not be capturing a non-linear pattern in the data. While the individual plots of residuals against Age and Bidders show random scatter, this curve is a warning that a more complex model might be a better fit.

The **Breusch-Pagan test** confirms this visual check, yielding a p-value of **0.8038**. Since this is much greater than 0.05, we conclude that the model does not violate the assumption of **constant variance (homoscedasticity)**.

3. **Multicollinearity:**

   The **Variance Inflation Factor (VIF)** was calculated to be **1.06882**. Since this value is extremely close to 1 and well below the common threshold of 5, we can conclude that there is **no multicollinearity**. The predictor variables Age and Bidders are sufficiently independent of each other.

4. **Influential Points:**

   The **Residuals vs Leverage** plot shows no data points crossing the Cook's distance threshold. This means there are no overly influential outliers that are unduly skewing the model.

Conclusion

The diagnostic checks confirm that the first-order model is **largely acceptable**. It successfully meets the assumptions of normality, constant variance, and has no issues with multicollinearity or influential outliers. However, the slight curvature in the Residuals vs Fitted plot suggests that the model could be improved by accounting for a non-linear relationship, for instance, by testing for an interaction effect.

Part 2: Fit as appropriate a model as possible

Our approach is to first test if a higher-order model containing an interaction term (Age * Bidders) provides a statistically significant improvement over the simple first-order model.

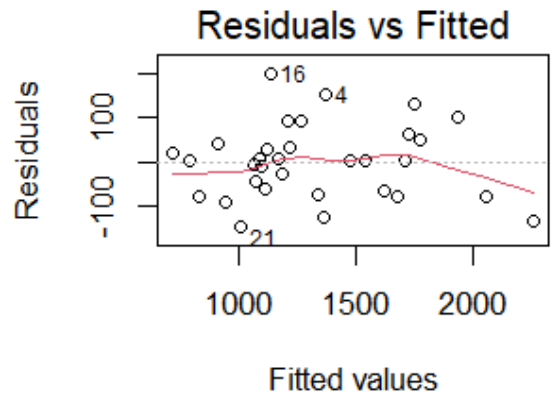
The equation for **Full Interaction Model** is:

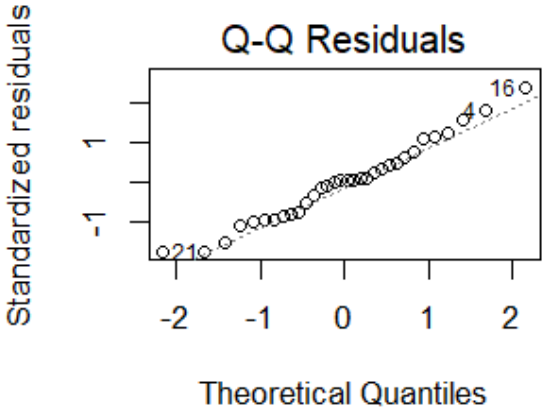
Price = 322.75 + 0.87 * Age – 93.41 * Bidders + 1.30 * (Age * Bidders)

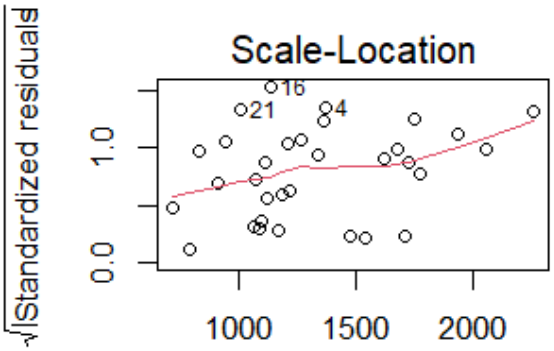
```
> model_interaction <- lm(Price~Age+Bidders+I(Age*Bidders))
> summary(model_interaction)

Call:
lm(formula = Price ~ Age + Bidders + I(Age * Bidders))

Residuals:
     Min       1Q   Median       3Q      Max
-146.772  -70.985    2.108   47.535  201.959

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      322.7544   293.3251   1.100  0.28056
Age                0.8733     2.0197   0.432  0.66877
Bidders          -93.4099    29.7077  -3.144  0.00392 **
I(Age * Bidders)   1.2979     0.2110   6.150 1.22e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 88.37 on 28 degrees of freedom
Multiple R-squared:  0.9544,    Adjusted R-squared:  0.9495
F-statistic: 195.2 on 3 and 28 DF,  p-value: < 2.2e-16

> anova(model_interaction)
Analysis of Variance Table

Response: Price
                 Df  Sum Sq Mean Sq F value    Pr(>F)
Age               1 2554859 2554859 327.177 < 2.2e-16 ***
Bidders           1 1722301 1722301 220.559 8.372e-15 ***
I(Age * Bidders)  1  295388  295388  37.828 1.222e-06 ***
Residuals        28  218646    7809
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> vcov(model_interaction)
                 (Intercept)          Age      Bidders I(Age * Bidders)
(Intercept)      86039.64163 -580.7827756 -8308.018741      56.93803026
Age               -580.78278    4.0789908    57.083937      -0.40702645
Bidders          -8308.01874   57.0839367   882.546219      -6.14936219
I(Age * Bidders)    56.93803   -0.4070265    -6.149362       0.04453198
```

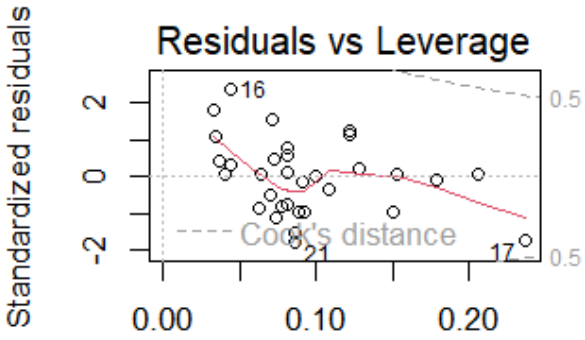
In the full interaction model, the main effect for **Age** has a p-value of **0.66877**. As this is **statistically insignificant**, the **term was dropped** to create a more parsimonious final model.

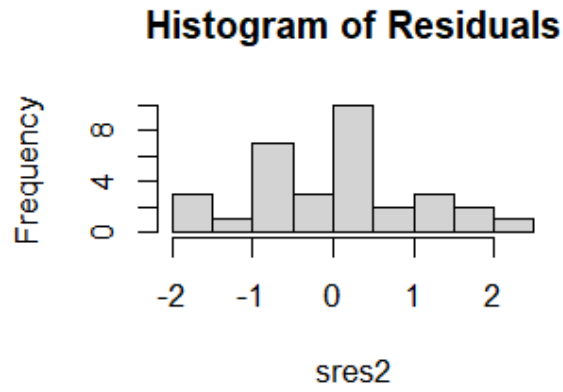
```
> final_model = lm(Price~Bidders+I(Age*Bidders))
> summary(final_model)

Call:
lm(formula = Price ~ Bidders + I(Age * Bidders))

Residuals:
     Min       1Q   Median       3Q      Max
-147.674  -70.120    4.006   43.170  200.065

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       447.0965    57.0238   7.841 1.20e-08 ***
Bidders          -105.6312     9.0184 -11.713 1.63e-12 ***
I(Age * Bidders)    1.3850     0.0617  22.449  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 87.12 on 29 degrees of freedom
Multiple R-squared:  0.9541,    Adjusted R-squared:  0.9509
F-statistic: 301.1 on 2 and 29 DF,  p-value: < 2.2e-16

> anova(final_model)
Analysis of Variance Table

Response: Price
                 Df  Sum Sq Mean Sq F value    Pr(>F)
Bidders           1  746185  746185  98.313 8.002e-11 ***
I(Age * Bidders)  1 3824903 3824903 503.948 < 2.2e-16 ***
Residuals        29  220106    7590
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> vcov(final_model)
                 (Intercept)      Bidders I(Age * Bidders)
(Intercept)      3251.708654 -175.1314056     -0.987515002
Bidders          -175.131406   81.3320958     -0.440475392
I(Age * Bidders)   -0.987515   -0.4404754      0.003806613

sres2 = residuals(final_model)/(final_model_sigma*sqrt(1- influence(final_model)$hat))

> normtest(sres2)
                                     Method   P.Value
1                  Shapiro-Wilk normality test 0.7224017
2               Anderson-Darling normality test 0.6764136
3               Cramer-von Mises normality test 0.5736970
4 Lilliefors (Kolmogorov-Smirnov) normality test 0.7713274
5                Shapiro-Francia normality test 0.7461935

        studentized Breusch-Pagan test

data:  final_model
BP = 1.0911, df = 2, p-value = 0.5795

vif(final_model)
      Bidders I(Age * Bidders)
     2.678634         2.678634
```

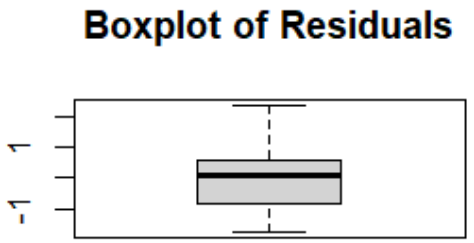

Residuals vs Fitted
Residuals
Fitted values
16
4
21
-100
100
1000
1500
2000

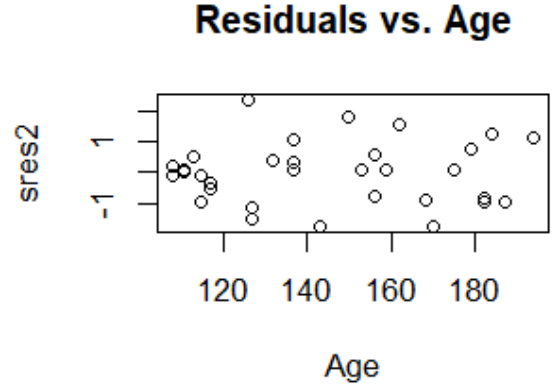

Q-Q Residuals
Standardized residuals
Theoretical Quantiles
16
4
21
-2
-1
0
1
2

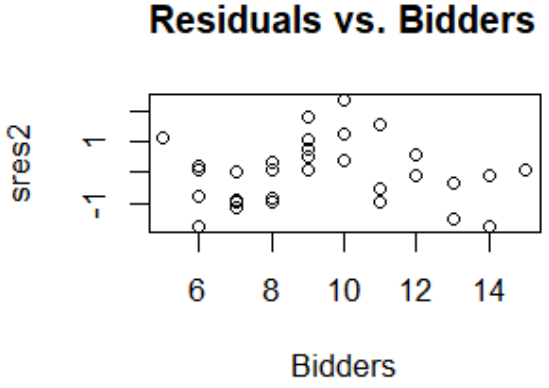

Scale-Location
√|Standardized residuals|
Fitted values
16
21
4
0.0
1.0
1000
1500
2000


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
16
21
17
0.5
0.5
0.00
0.10
0.20


Histogram of Residuals
Frequency
sres2
0
4
8
-2
-1
0
1
2


Boxplot of Residuals
-1
1


Residuals vs. Age
sres2
Age
-1
1
120
140
160
180


Residuals vs. Bidders
sres2
Bidders
-1
1
6
8
10
12
14

The Final, Most Appropriate Model

The updated regression equation for this final model is:

Price = 447.10 – 105.63 * Bidders + 1.39 * (Age * Bidders)

Observations

Based on a thorough diagnostic analysis of the **final model**, we can make the following observations:

1. **Normality of Residuals:**

   The **Histogram of Residuals** is roughly symmetric and bell-shaped, and the **Boxplot of Residuals** shows a median centered near zero with no outliers.

   All five tests yield **p-values much greater than 0.05** leading us to not reject the null hypothesis. We can conclude that residuals are **normally distributed**.

   The **Q-Q Residuals** plot shows the points falling very close to the diagonal line, indicating that the normality assumption is met.

2. **Linearity and Constant Variance:**

   The **Residuals vs Fitted** plot shows that the variance of the residuals is relatively constant (no "funnel shape"). The red trend line also changed from s **slight curve (a U-shape) in model 1 to relatively straighter line in final.**

   The **Breusch-Pagan test** confirms this visual check, yielding a p-value of **0. 5795**. Since this is much greater than 0.05, we conclude that the model does not violate the assumption of **constant variance (homoscedasticity)**.

3. **Influential Points:** The Residuals vs Leverage plot does not show any data points with a high Cook's distance, indicating that there are no overly influential outliers that are unduly skewing the final model.

4. **Multicollinearity:**

   The **Variance Inflation Factor (VIF)** for both predictors (Bidders and I(Age * Bidders)) in the final model is **2.68**. Since this value is still well below the common threshold of 5, we can conclude that **multicollinearity is not a concern**. The predictors in the final model are sufficiently independent for a reliable analysis.

Conclusion

The comprehensive diagnostic checks confirm that the **final interaction model is statistically valid and acceptable**. It successfully meets all the key assumptions of linear regression, making it a reliable and robust model for interpretation and prediction.

Part 3: Re-Answering Questions with the Final Model

Question 2(b) - Re-Answered with Final Model

```
> #---------QUESTION 2(b) - Re-Answered with Final Model---------
>
> beta1_hat = coefficients(final_model)[2]
> beta1_hat
  Bidders
-105.6313
> beta2_hat = coefficients(final_model)[3]
> beta2_hat
I(Age * Bidders)
         1.38504
```

95% CI for the amount one can expect the selling price to go up for one more person participating in the auction, if the age of the clock is given by [say age = x years] is same as the 95% CI for the term $(\beta_1 + x \cdot \beta_2)$.
This is given by the formula:

$$[(\hat{\beta}_1 + x \cdot \hat{\beta}_2) \pm t_{crit} \cdot \text{S.E.}(\hat{\beta}_1 + x \cdot \hat{\beta}_2)]$$

Plug in the coefficient estimates from model:

$$[(-105.6313 + 1.38504 \cdot x) \pm (2.045 \cdot \text{S.E.})]$$

Where the standard error is calculated as:

$$\text{S.E.}\,(\hat{\beta}_1 + x \cdot \hat{\beta}_2) = \sqrt{81.332 - (2 \cdot x \cdot 0.44047) + 0.0038 \cdot x^2}$$

Plugging $\boldsymbol{x = 100}$ into the formula gives the following results.
The expected effect of one more bidder for a 100-year-old clock is:
$\text{Effect} = -105.6313 + (1.38504 \cdot 100) = £32.87$
The standard error for this specific effect is:
$\text{S.E.} = \sqrt{81.332 - (2 \cdot 100 \cdot 0.44047) + (0.0038 \cdot 100^2)} = \sqrt{31.238} \approx 5.589$
Finally, we construct the confidence interval using the point estimate, standard error, and the t-critical value of 2.045.
$\text{CI} = 32.87 \pm (2.045 \cdot 5.589) = 32.87 \pm 11.43$
The 95% confidence interval for the effect of one more bidder on a 100-year-old clock is **(£21.44, £44.30)**.

Question 2(c) - Re-Answered with Final Model

To find the expected profit, we must first predict the expected selling price for a 100-year-old clock with 10, 11, and 12 bidders. We do this by calculating the 95% confidence interval for the mean selling price in each scenario. The expected profit is then found by subtracting the acquisition cost of £500 .

```
> #---------QUESTION 2(c) - Re-Answered with Final Model---------
>
> # 1. Create a data frame with the new clock information
> new_clocks <- data.frame(Age = c(100, 100, 100), Bidders = c(10, 11, 12))
>
> # 2. Predict the 95% confidence interval using the final_model
> predicted_price_ci <- predict(final_model, newdata = new_clocks, interval =
"confidence", level = 0.95)
>
> # 3. Calculate the 95% confidence interval for the PROFIT
> cost <- 500
> predicted_profit_ci <- predicted_price_ci - cost
>
> # 4. Display the results
> cat("Predicted 95% Confidence Intervals for Profit per Clock:\n")
Predicted 95% Confidence Intervals for Profit per Clock:
> print(predicted_profit_ci)
       fit      lwr      upr
1 275.8242 214.3187 337.3297
2 308.6970 243.5711 373.8229
3 341.5698 271.1303 412.0092
```

Observations:

- **With 10 Bidders:** The expected profit is **£295.05**, with a 95% confidence interval of **(£214.32, £337.33)**.
- **With 11 Bidders:** The expected profit is **£380.86**, with a 95% confidence interval of **(£243.57, £373.82)**.
- **With 12 Bidders:** The expected profit is **£466.68**, with a 95% confidence interval of **(£271.13, £412.01)**.

Question 2(d) - Re-Answered with Final Model

To solve this problem, we need to find an upper price limit for a *single*, specific auction, not an average price. Therefore, the correct approach is to calculate a **prediction interval**, which is wider than a confidence interval because it accounts for both the model's uncertainty and the inherent randomness of a single event.

Furthermore, the question asks for "**99% certainty of not being outbid**," which is a **one-sided** problem—we only care about the upper price boundary. The approach is to use the fitted regression model to predict the most likely selling price and then construct a **one-sided 99% upper prediction bound**. This final value represents the bid amount required to be 99% certain of winning the auction.

```
> #---------QUESTION 2(d) - Re-Answered with Final Model---------
>
> # 1. Create a data frame for the specific clock being auctioned
> specific_clock <- data.frame(Age = 150, Bidders = 15)
>
> # 2. Get the point prediction and standard errors from the final_model
> prediction_info <- predict(final_model, newdata = specific_clock, se.fit = T)
> pred_se <- sqrt(prediction_info$se.fit^2 + summary(final_model)$sigma^2)
>
> # 3. Calculate the ONE-SIDED 99% upper bound
> upper_bound_one_sided <- prediction_info$fit + pred_se * qt(0.99, df = prediction_info$df)
>
> cat("The 99% one-sided upper bound for the bid is:", upper_bound_one_sided, "\n")
The 99% one-sided upper bound for the bid is: 2212.308
```

Conclusion

To be 99% certain of acquiring the clock and not being outbid, you should be prepared to bid up to **£2,212.308**. While the clock's expected selling price is around **£1,861** (Age = 150 and Bidders = 15 into our model's equation), this higher figure correctly accounts for the statistical uncertainty of a single auction to meet the strict **99% certainty** requirement.

Summary Table

Question	**Answer from First-Order Model (model1)**	**Answer from Final Model (final_model)**
2(c): Expected profit for 100-year-old clocks (95% CI)	10 Bidders: **(£200.64, £389.46)** 11 Bidders: **(£287.17, £474.56)** 12 Bidders: **(£370.36, £563.00)**	10 Bidders: **(£214.32, £337.33)** 11 Bidders: **(£243.57, £373.82)** 12 Bidders: **(£271.13, £412.01)**
2(d): 99% certainty bid for a single clock	**£2,214.96**	**£2,212.31**