Mario Kart Wii Auction Price Prediction

Phantom Squadron

Introduction

Background

Mario Kart is a series of kart racing games developed and published by Nintendo. Players compete in go-kart races while using various power-up items. Over 164.43 million copies in the series have been sold worldwide, and is available for purchase at many online retailers, including eBay. eBay is a multibillion-dollar business with operations in about 32 countries, and is an online auction and shopping website in which people and businesses buy and sell a wide variety of goods and services worldwide. The website is free to use for buyers, but sellers are charged fees for listing items after a limited number of free listings, and an additional or separate fee when those items are sold. In order to sell an item, one of the many things a seller must include are a title, photo, shipping, and listing price for the good they want to sell. Additional info may be filled by seller such as whether or not the item is a "Videogame" when determining category of listing. An interested buyer would then search for the keyword "Mario Kart" or similar keywords and look at the many listings displayed on the search engine and interact with the specific listing to purchase the product. Then when all things are settled, the seller will be notified of the purchase and send the proposed good to the buyer's address.

(sources: Super Mario Kart: Most Influential Video Game in History , eBay - Welcome to the world's online marketplace)

Dataset Description

There are several interesting features in the data. First off, note that there are two outliers in the data. These serve as a nice example of what one should do when encountering an outlier: examine the data point and remove it only if there is a good reason. In these two cases, we can see from the auction titles that they included other items in their auctions besides the game, which justifies removing them from the data set.

This data set includes all auctions for a full week in October 2009. Auctions were included in the data set if they satisfied a number of conditions. (1) They were included in a search for "wii mario kart" on ebay.com, (2) items were in the Video Games > Games > Nintendo Wii section of Ebay, (3) the listing was an auction and not exclusively a "Buy it Now" listing (sellers sometimes offer an optional higher price for a buyer to end bidding and win the auction immediately, which is an optional Buy it Now auction), (4) the item listed was the actual game, (5) the item was being sold from the US, (6) the item had at least one bidder, (7) there were no other items included in the auction with the exception of racing wheels, either generic or brand-name being acceptable, and (8) the auction did not end with a Buy It Now option.

(Credit: OpenIntro Statistics)

Variables

`1. Id (Categorical, nominal): Auction ID assigned by Ebay.`

`2. Duration (Numerical, discrete): Auction length, in days.`

`3. n_bids (Numerical, discrete): Number of bids.`

`4. Cond (Categorical, nominal): Game condition, either new or used.`

5. Start_pr (Numerical, continuous): Start price of the auction.

6. Ship_pr (Numerical, continuous): Shipping price.

7. Total_pr (Numerical, continuous): Total price, which equals the auction price plus the shipping price.

8. Ship_sp (Categorical, nominal): Shipping speed or method.

9. Seller_rate (Numerical, discrete): The seller's rating on Ebay. This is the number of positive ratings minus the number of negative ratings for the seller.

10. Stock_photo (Categorical, nominal): Whether the auction feature photo was a stock photo or not. If the picture was used in many auctions, then it was called a stock photo.

11. Wheels (Numerical, discrete): Number of Wii wheels included in the auction. These are steering wheel attachments to make it seem as though you are actually driving in the game. When used with the controller, turning the wheel actually causes the character on screen to turn.

12. Title (Categorical, nominal): The title of the auctions.

Data Exploration

We began by taking a look at the data itself as well as comparing a number of different variables to each other. First we looked at what was actually in the dataset

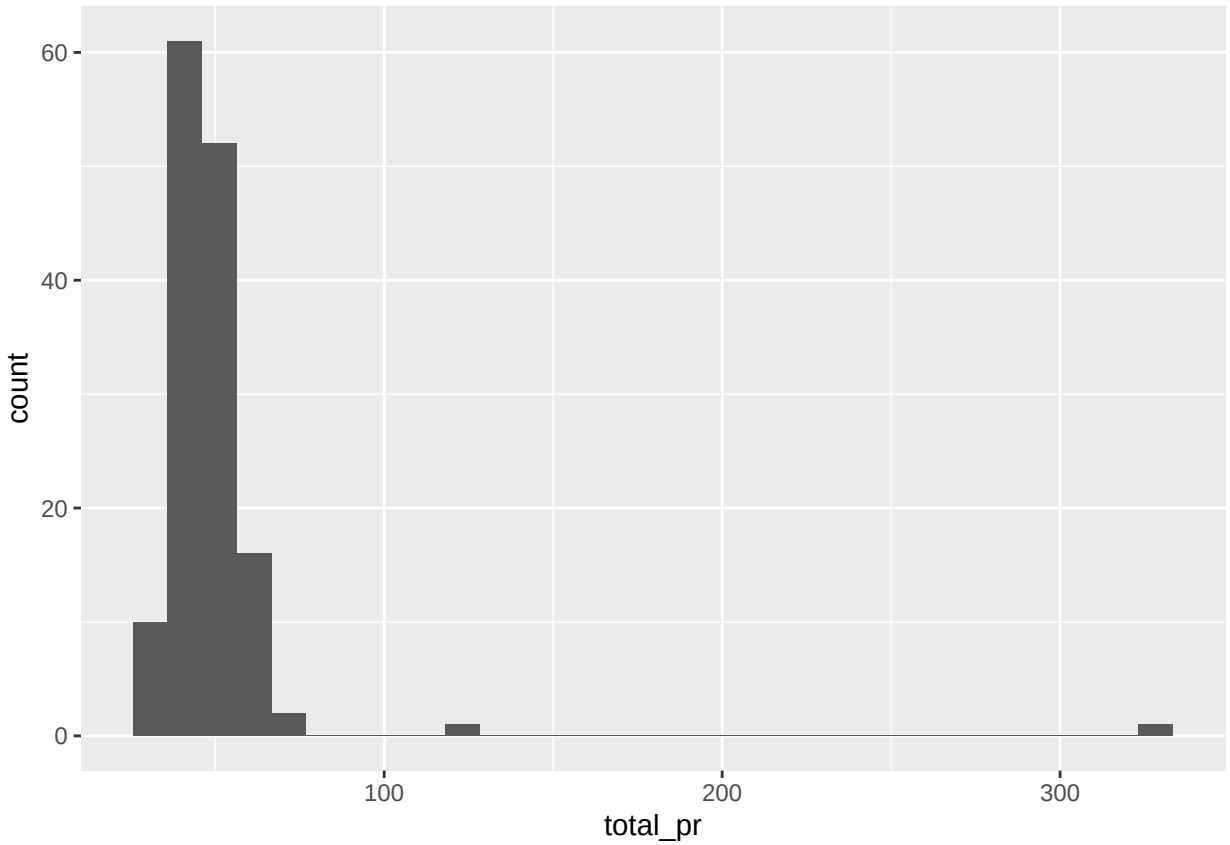
```
glimpse(mariokart)
```

```
## Rows: 143
## Columns: 12
## $ id          <dbl> 150377422259, 260483376854, 320432342985, 280405224677, 17~
## $ duration    <int> 3, 7, 3, 3, 1, 3, 1, 1, 3, 7, 1, 1, 1, 1, 7, 7, 3, 3, 1, 7~
## $ n_bids      <int> 20, 13, 16, 18, 20, 19, 13, 15, 29, 8, 15, 15, 13, 16, 6, ~
## $ cond        <fct> new, used, new, new, new, new, used, new, used, used, new,~
## $ start_pr    <dbl> 0.99, 0.99, 0.99, 0.99, 0.01, 0.99, 0.01, 1.00, 0.99, 19.9~
## $ ship_pr     <dbl> 4.00, 3.99, 3.50, 0.00, 0.00, 4.00, 0.00, 2.99, 4.00, 4.00~
## $ total_pr    <dbl> 51.55, 37.04, 45.50, 44.00, 71.00, 45.00, 37.02, 53.99, 47~
## $ ship_sp     <fct> standard, firstClass, firstClass, standard, media, standar~
## $ seller_rate <int> 1580, 365, 998, 7, 820, 270144, 7284, 4858, 27, 201, 4858,~
## $ stock_photo <fct> yes, yes, no, yes, yes, yes, yes, yes, yes, no, yes, yes, ~
## $ wheels      <int> 1, 1, 1, 1, 2, 0, 0, 2, 1, 1, 2, 2, 2, 2, 1, 0, 1, 1, 2, 2~
## $ title       <fct> "~~ Wii MARIO KART &amp;amp; WHEEL ~ NINTENDO Wii ~ BRAND NEW ~
```

The first thing we did was graph out the total prices, to see how common different prices were.

```
ggplot(data=mariokart, aes(x=total_pr)) + geom_histogram()
```

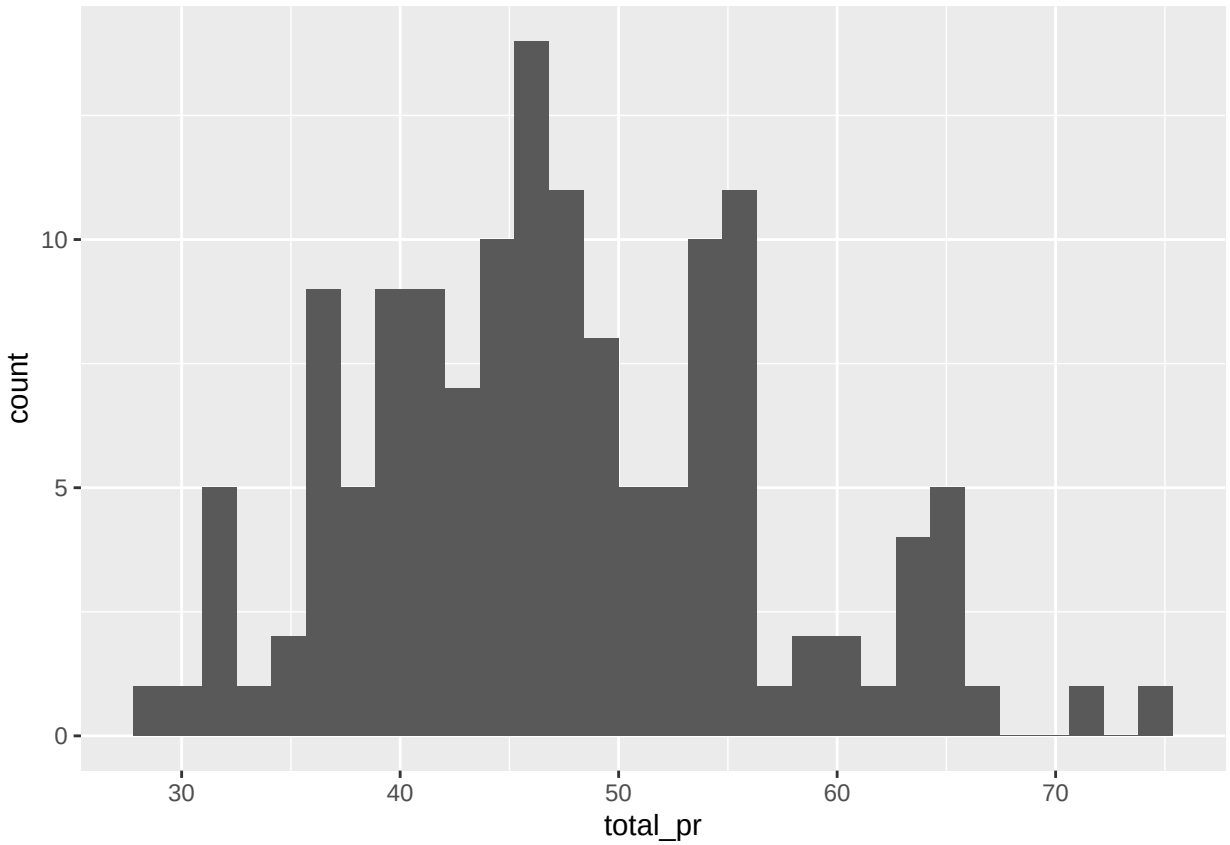
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Clearly there are some outliers, so we tried again without them

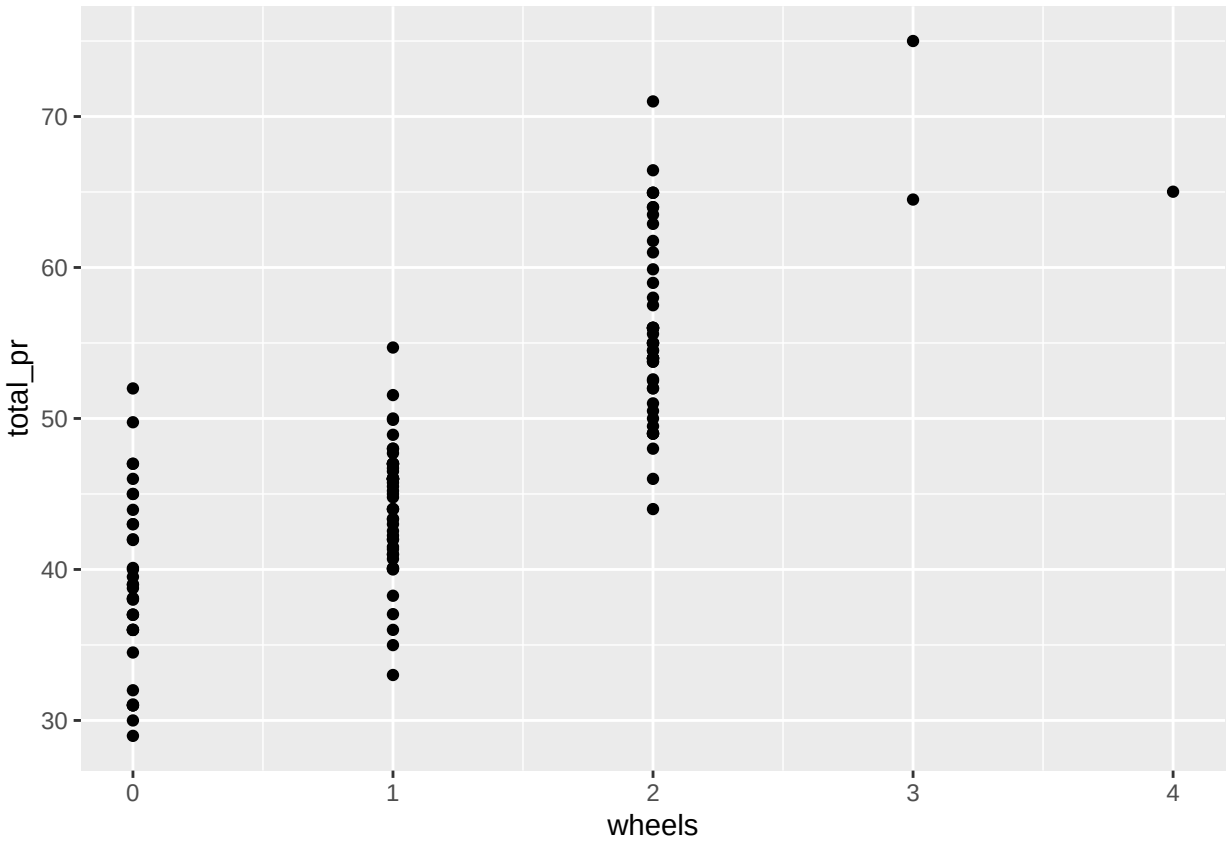
```
mkart <- mariokart %>% filter(total_pr < 100)
ggplot(data=mkart, aes(x=total_pr)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

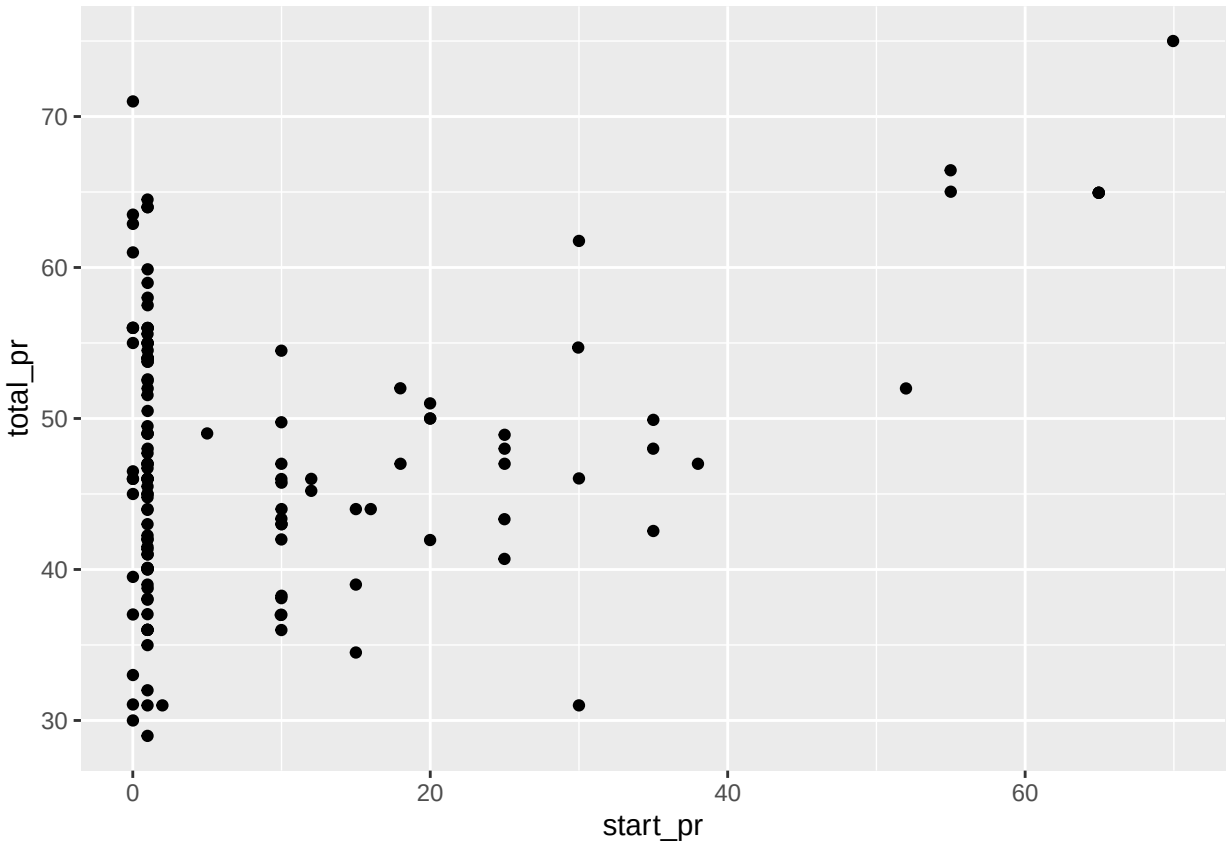


While we did look at a number of varaibles, the ones which looked like they were most likely to have a relation were: wheel count to total price and start price to total price

```
ggplot(data=mkart, aes(x=wheels, y=total_pr)) + geom_point()
```




```
ggplot(data=mkart, aes(x=start_pr, y=total_pr)) + geom_point()
```



In the end, we decided on the latter, start price to total price, as it looked like it would work with a linear regression better than the wheels.

Research Question

Does the bidding start price on a copy of Mario Kart significantly affect the total/end price of the purchased items?

Hypothesis Statement

Null (H_0): There is no linear relationship between the bid start price and the final/total price the product is sold for. ($b_1 = 0$)

Alternative (H_A): There is a significant linear relationship between the bid start price and the final/total price the product is sold for. ($b_1 \neq 0$)

Where b1 represents the regression coefficient for bid start price in the linear regression model.

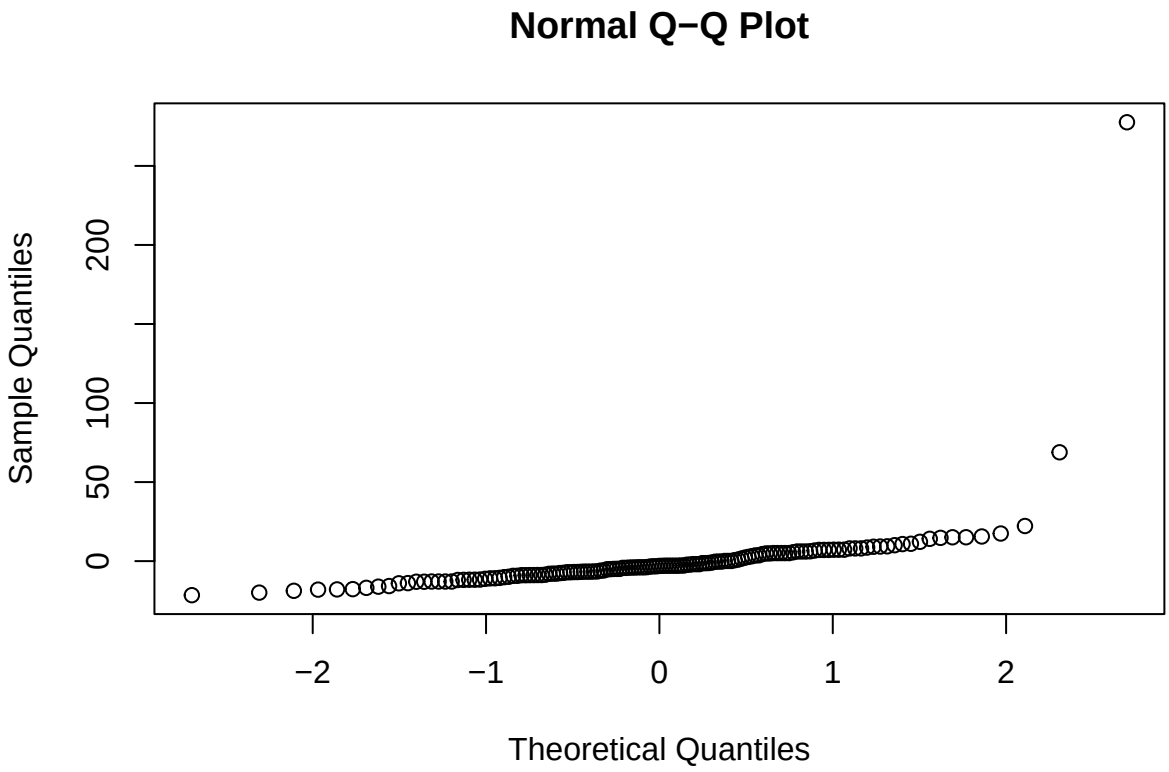
Methods

We will be testing using a linear regression of the form y=b1x+b0 where b1 is the slope and b0 is the population intercept. A linear regression is a commonly used method to investigate the relationship between two quantitative variables, and in this example it is being used to determine the relationship of the bid start price and the final/total price of the product on eBay. Linear regression can provide insights into the direction and strength of the relationship between variables, and can help identify associations between dependent/independent variables.

To better understand whether or not there is an association between the variables, we will determine whether or not the criteria for a linear regression model is satisfied. To use a linear regression we require: -

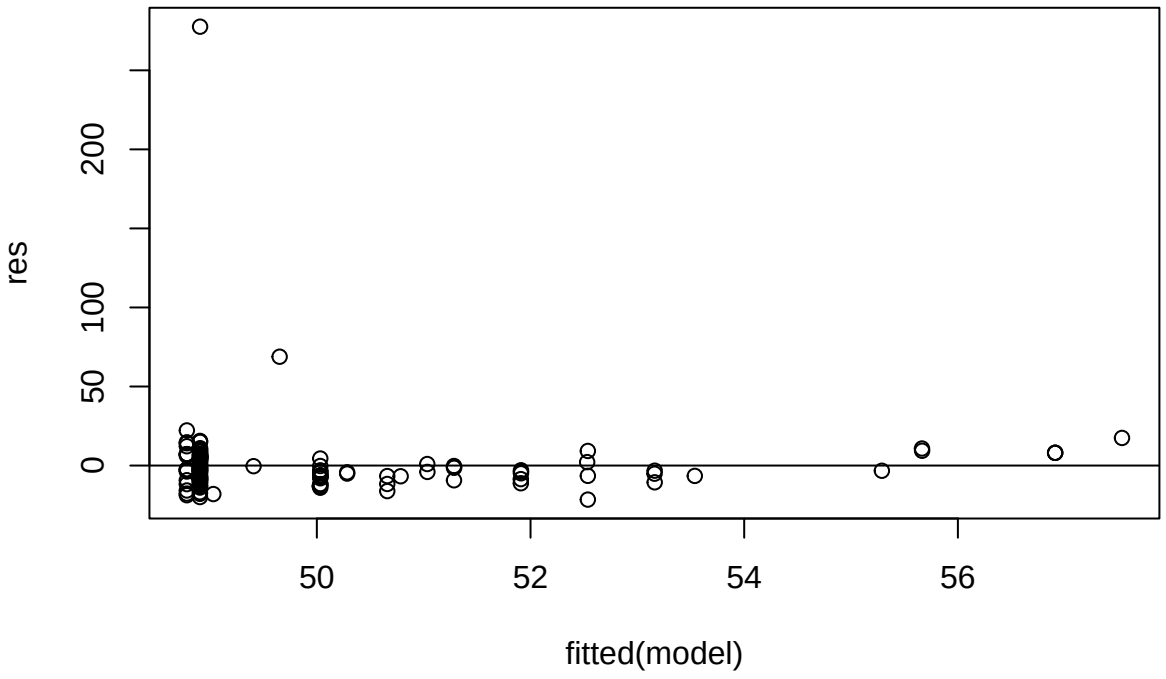
Independent random samples: The eBay auctions did not directly affect each other, and so can be understood as independent - Residuals must be Normally Distributed and Homoskedastic

```
data(mariokart)
model <- lm(total_pr~start_pr, data=mariokart)
res <- resid(model)
qqnorm(res)
```



As you can see, apart from the two outliers on the far right, the residuals are normally distributed.

```
plot(fitted(model), res)
abline(0,0)
```



The chart above shows residuals plotted against explanatory variable, homoskedastic apart from a few outliers. In this case the criteria is also met.

Lastly, since the residuals have no relation to explanatory variable, the criteria is met.

We will also be calculating R-Squared as it is a statistical measurement that explains what proportion of variance of dependent variables can be explained by independent variables. Since this is a linear regression model, R-Squared is well-applied.

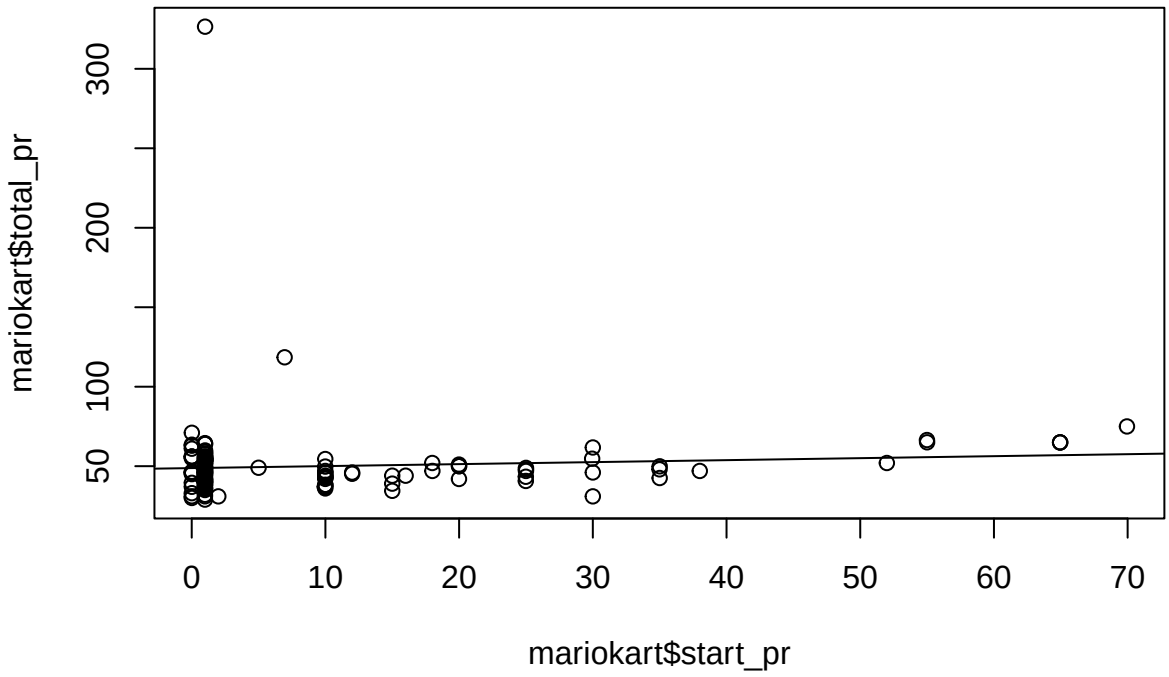
To reiterate, our hypothesis are as follows:

Null (H_0): There is no linear relationship between the bid start price and the final/total price the product is sold for. ($b_1 = 0$)

Alternative (H_A): There is a significant linear relationship between the bid start price and the final/total price the product is sold for. ($b_1 \neq 0$)

Results

```
plot(mariokart$start_pr, mariokart$total_pr)
abline(model)
```




```
summary(model)
```

```
##
## Call:
## lm(formula = total_pr ~ start_pr, data = mariokart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.537  -8.693  -3.032   5.078 277.603
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.7820     2.4904  19.588   <2e-16 ***
## start_pr      0.1252     0.1432   0.874    0.384
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.71 on 141 degrees of freedom
## Multiple R-squared:  0.005388,   Adjusted R-squared:  -0.001666
## F-statistic: 0.7639 on 1 and 141 DF,  p-value: 0.3836
```

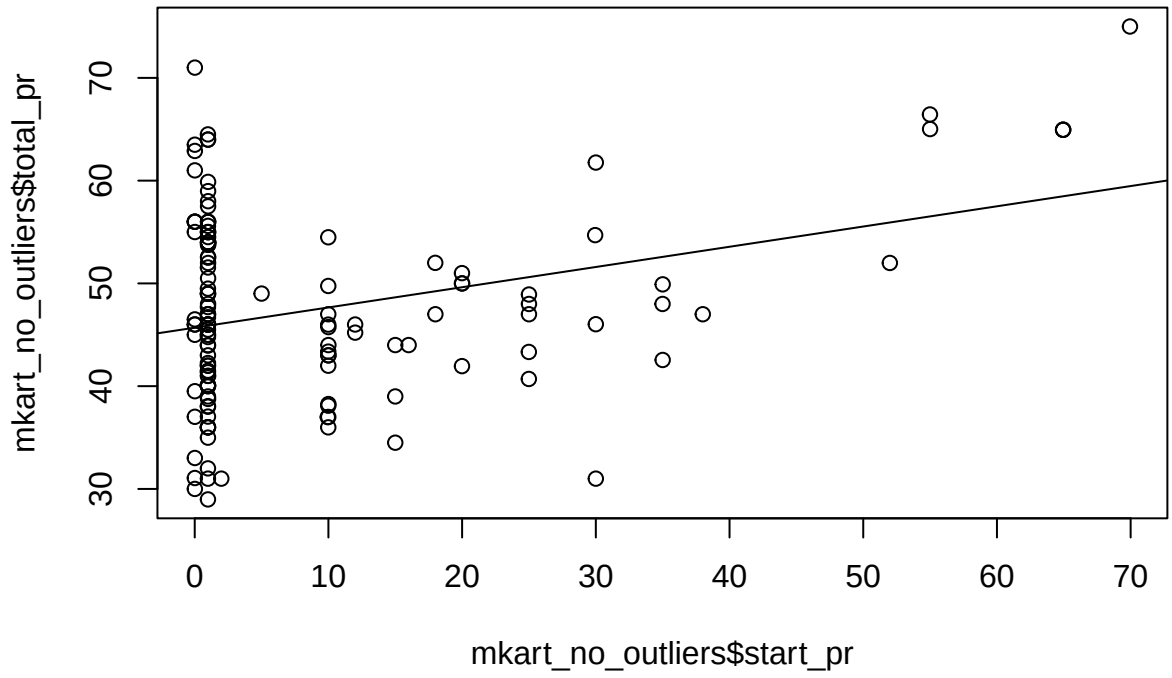
We did not find a significant linear relation between the start and end prices of the auctions. As you can see, our R^2 value is barely above 0.1, which made us conclude that a linear regression did not properly model the relation.

Further Exploration

The first thing we tried was removing the two data points which were clearly outliers. However, doing this had little to no impact on the result, as you can see below:

```
mkart_no_outliers <- mariokart %>%
  filter(total_pr < 100)
model2 <- lm(total_pr~start_pr, data=mkart_no_outliers)
plot(mkart_no_outliers$start_pr, mkart_no_outliers$total_pr)
```

```
abline(model2)
```

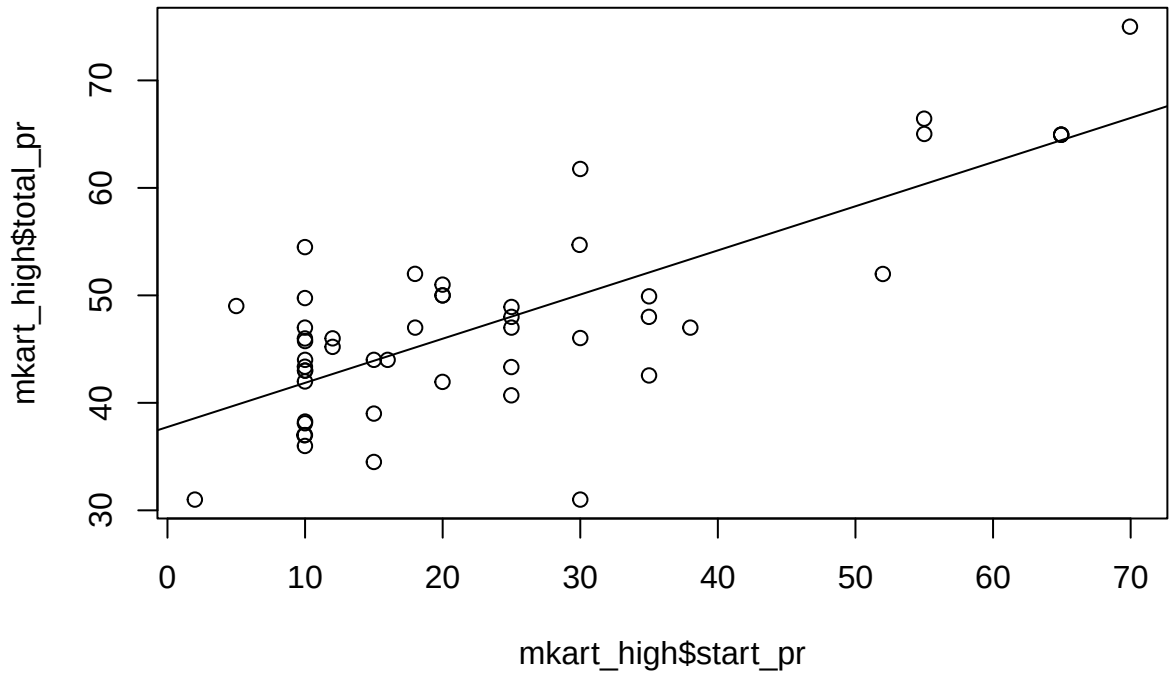



```
summary(model2)
```

```
##
## Call:
## lm(formula = total_pr ~ start_pr, data = mkart_no_outliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.5972  -5.8252  -0.6923   6.7028  25.3077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.69029    0.84337  54.176  < 2e-16 ***
## start_pr     0.19690    0.04818   4.087 7.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.642 on 139 degrees of freedom
## Multiple R-squared:  0.1073, Adjusted R-squared:  0.1008
## F-statistic:  16.7 on 1 and 139 DF,  p-value: 7.36e-05
```

The next thing we tried was to remove all data which started at a price below 1$. This is a common tactic on eBay for auctions, as it draws in bids early, and gets people to follow it looking for a good deal. Because this is so common, there is a large chunk of data with very low starting prices and extremely varied end prices. By removing those points, we are left with the following.

```
mkart_high <- mkart_no_outliers %>%
  filter(start_pr > 1)
model3 <- lm(total_pr~start_pr, data=mkart_high)
plot(mkart_high$start_pr, mkart_high$total_pr)
abline(model3)
```




```
summary(model3)
```

```
##
## Call:
## lm(formula = total_pr ~ start_pr, data = mkart_high)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.0765  -4.1264   0.5159   4.0356  12.6336
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.75243    1.47381   25.61  < 2e-16 ***
## start_pr     0.41080    0.04968    8.27 1.01e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.099 on 47 degrees of freedom
## Multiple R-squared:  0.5927, Adjusted R-squared:  0.584
## F-statistic: 68.39 on 1 and 47 DF,  p-value: 1.011e-10
```

This gives a much better R^2 value of about 0.59 (~0.58 adjusted). While the scope of this data is more narrow, it does show a clear positive correlation between start and end prices of auctions starting above 1$.

Analysis

After identifying other relationships in consideration of this one, we concluded that there is no linear relationship between bid start price and final price the product is sold for. Once we removed the auctions starting at 1$ or less, we were left with a significantly restricted data set, however we did find a correlation between start and end prices. The most interesting part of this was that the slope of the linear regression of this graph was below 1. This indicates that increasing the price as high as possible is not necessarily the best strategy, as it does not correlate to an equivalent increase in final price. With more time, we would compare the average price of products beginning above one dollar to those below it, to see if there is a significant

difference in the results of each method. Given the appropriate data, it would also be interesting to see how many auctions end without a bid compared to their starting price, as this set of data only included auctions with at least one bid. Using that we would be able to evaluate the risk compared to the reward of increasing the starting price of an item.

Conclusion

To summarize, we looked at eBay auctions of the video game Mariokart for the Nintendo Wii. We compared the start and end price of auctions, hoping to find a relationship which would be able to answer the question of whether the bidding start price on a copy of Mario Kart significantly affected the end price of the purchased item. In the end, we found no correlation across the whole data set, even after removing outliers. However, after restricting our scope to only consider items which began auction above 1$, a common tactic on eBay auctions, we found a positive correlation between start and end price of items.