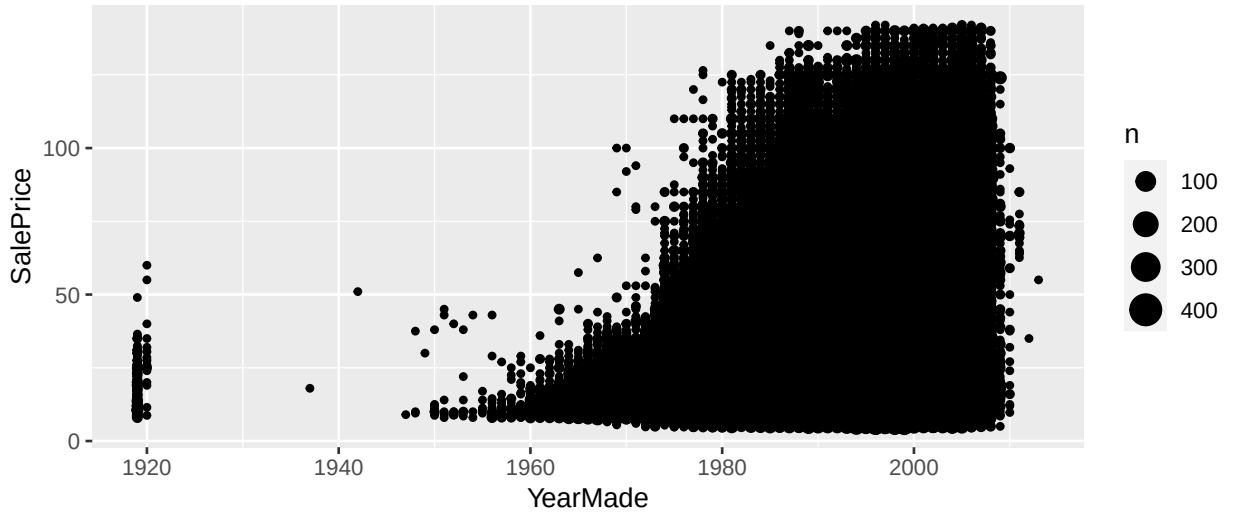
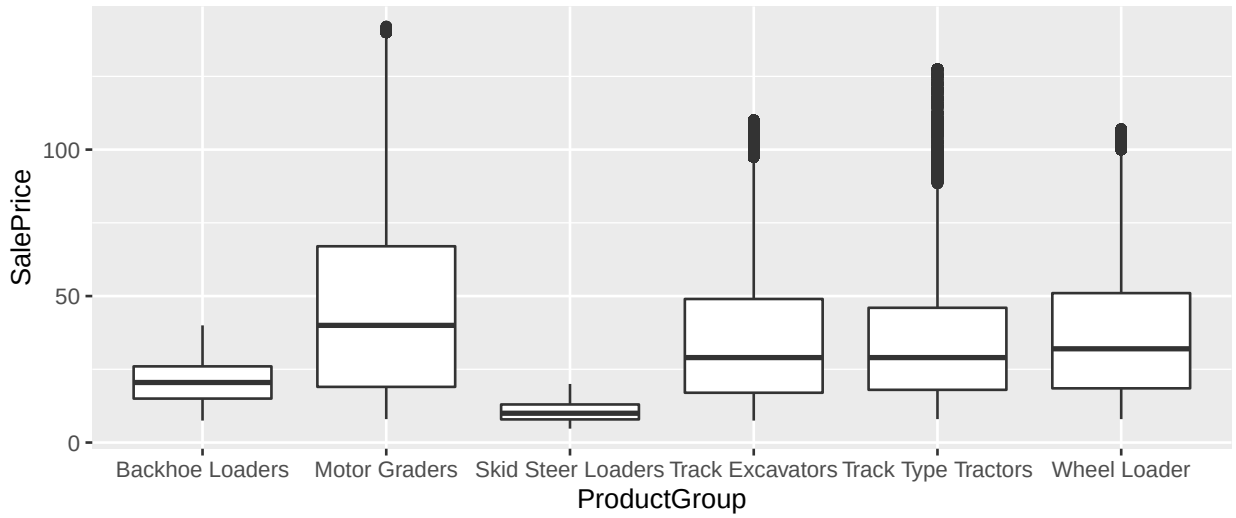
Linear Regression Analysis - Tractor Sale Price

Sivaram Ainkaran

10/08/2021

1.Plots of 'SalePrice' against 'YearMade' and 'ProductGroup':





From the 'SalePrice' vs 'YearMade' plot, we can see that vehicles that were made in more recent years tend to have more variance in their sale prices and have also become increasingly expensive. We can also see that the older vehicles sell for much cheaper on average and have much less variance in sale prices.

From the 'SalePrice' vs 'ProductGroup' plot, we can see that Backhoe and Skid Steer Loaders tend to be cheaper, with a narrow interval and very little variance in sale price. All the other vehicles tend to have a

higher variance in sale price with multiple outliers. Track Type Tractors especially, have many outliers, with sale prices much higher than the median. Skid Steer loaders would also tend to have a lower price than all other machine types since its median and 50% interval lies below all other vehicle intervals.

2. Regression model for 'SalePrice' with the predictors as its main effects:

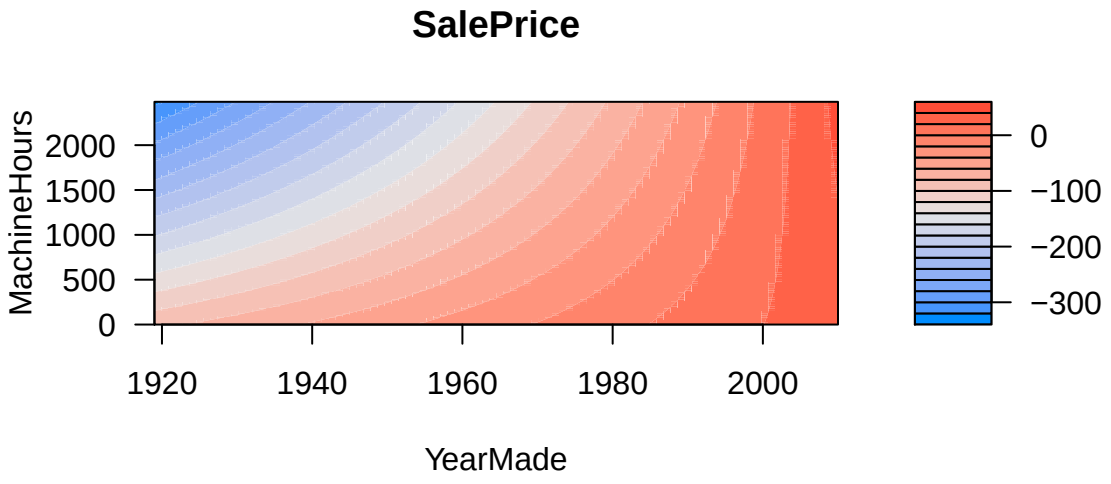
```
bds_lm <- lm(SalePrice ~ YearMade + MachineHours + SaleDate + ProductGroup + Enclosure, data=bds)
```

3. Best regression model for 'SalePrice':

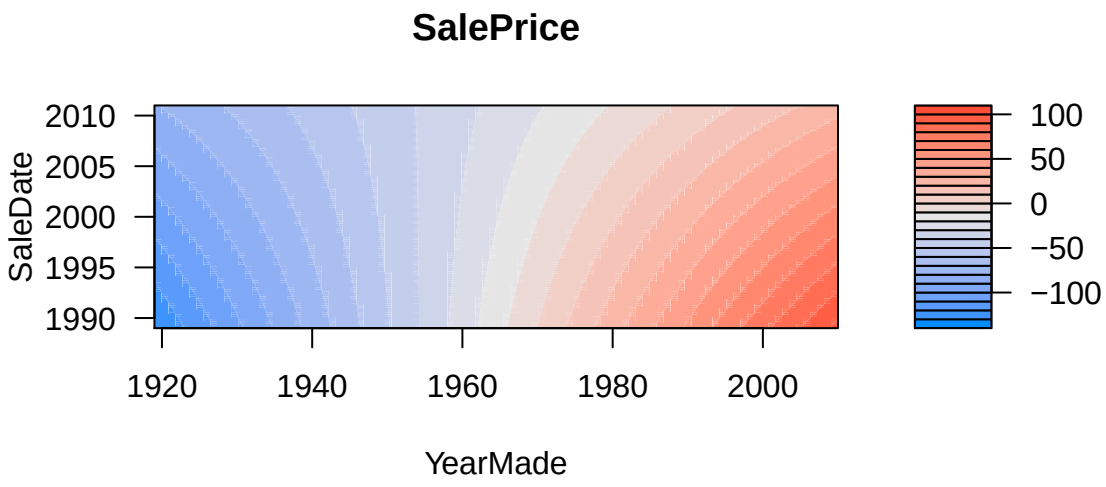
```
bds_int <- lm(SalePrice~(. - X - ID)^2, data=bds) %>%
  step(trace=FALSE)
summary(bds_int)
```

From this AIC minimized linear model, we can see that 'SalePrice' is greatly affected by all the predictors ('YearMade', 'MachineHours', 'SaleDate', 'ProductGroup' and 'Enclosure'). We can also see that some of the interactions between these predictors are not significant in determining sale price of the machine. The interactions between 'MachineHours' and 'EROPs w AC' and 'NO ROPS' enclosures, the interaction between 'SaleDate' and 'OROPS' Enclosures and the interactions between many of the product groups and enclosures all have very little importance in determining sale price of the machine. The predictor with the largest effect on sale price was the Product Group Motor Graders, since it had the highest beta value at 2362.

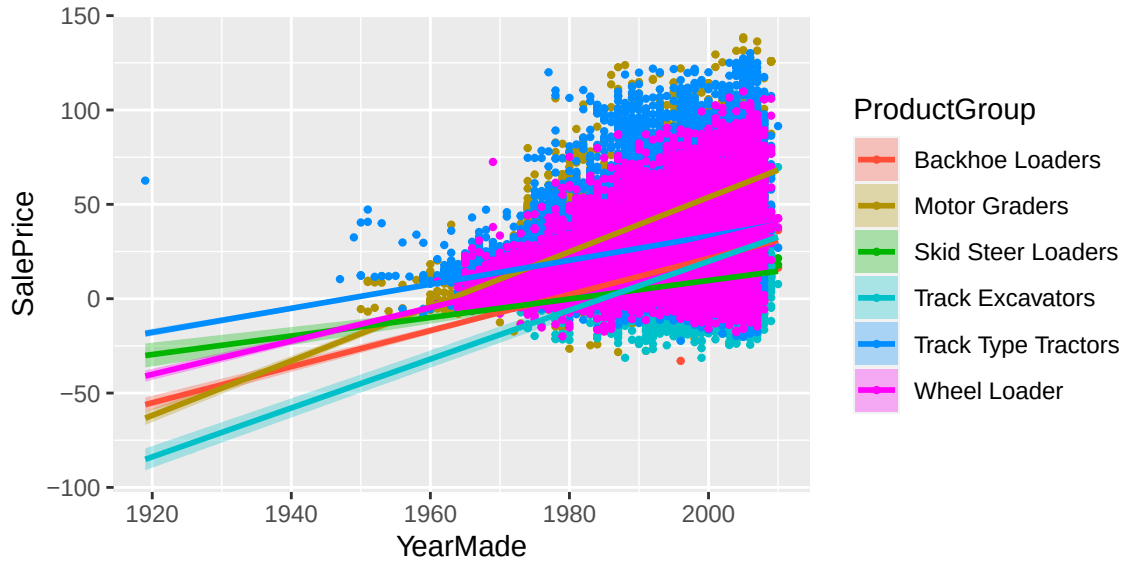
4. 'YearMade' and interactions terms visualized:



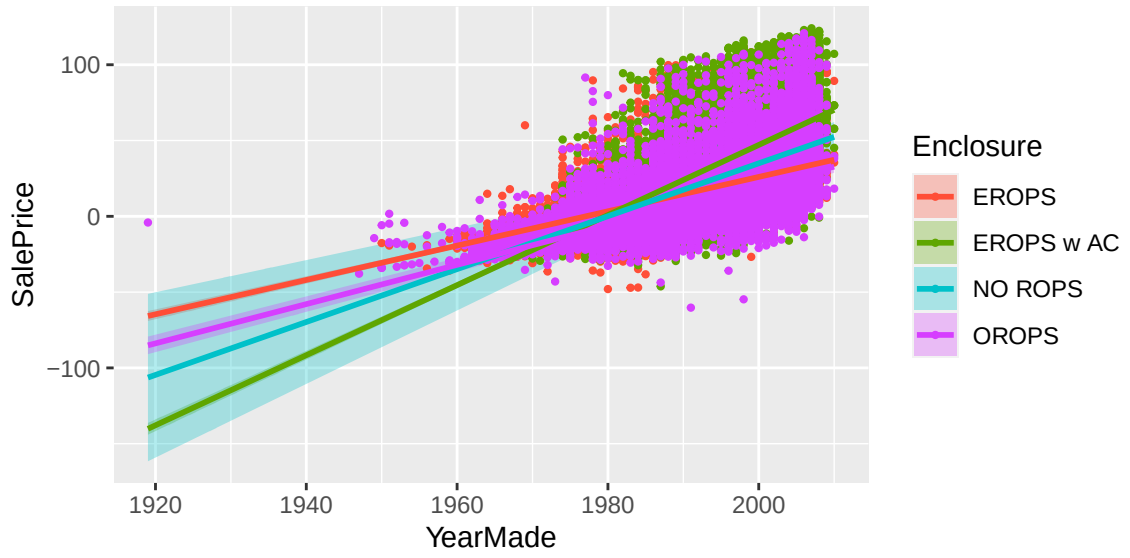
From the plot above, showing the interaction between 'YearMade' and 'MachineHours' in this regression, we can see that the more recently a machine has been produced, the higher the sale price of it tends to be. We can also see that the more hours a machine has been used for already, decreases the sale price of the machine, but the newer the machine, the less this affects the price.



From the plot above, showing the interaction between 'YearMade' and 'SaleDate' in this regression, we can see that the more recently a machine was produced, the higher the sale price will be. It can also be seen that machines made prior to the 1960s would have sold for more nowadays than 2 decades ago, even though it would have been newer then. Machines made after the 1960s however, tend to have sold for more if they were sold closer to the production date. There is some data in the plot showing some values with year of production being after year of sale, but this is impossible and should not be looked at.

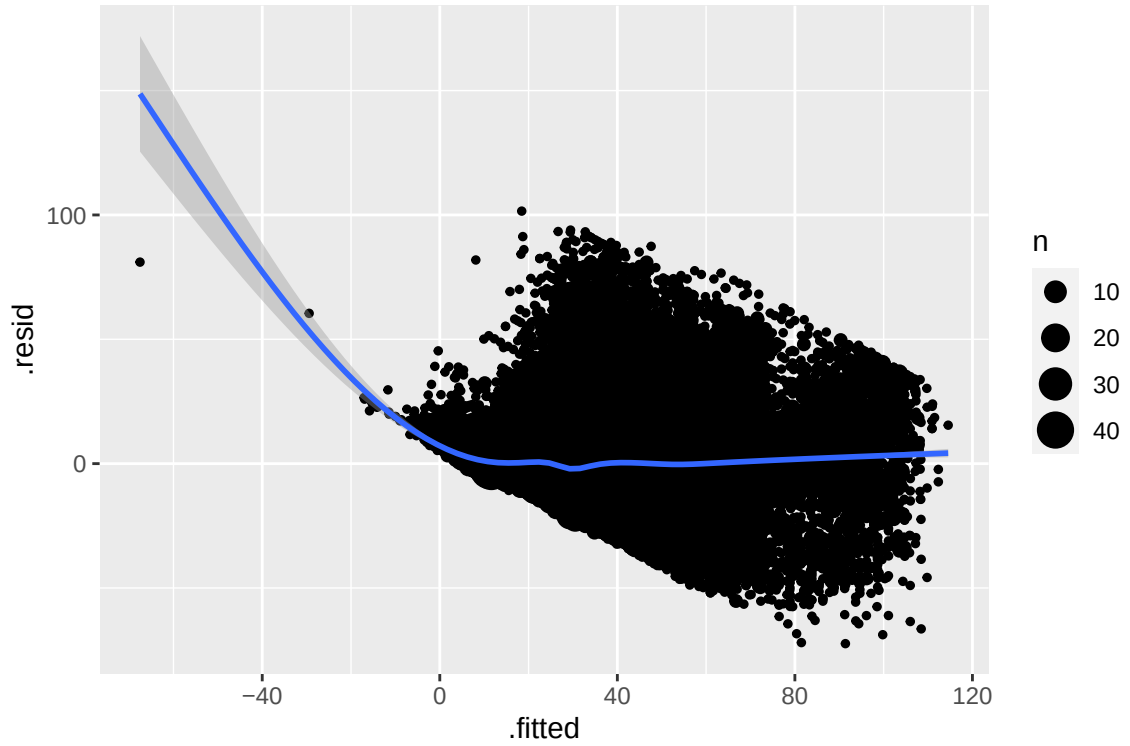


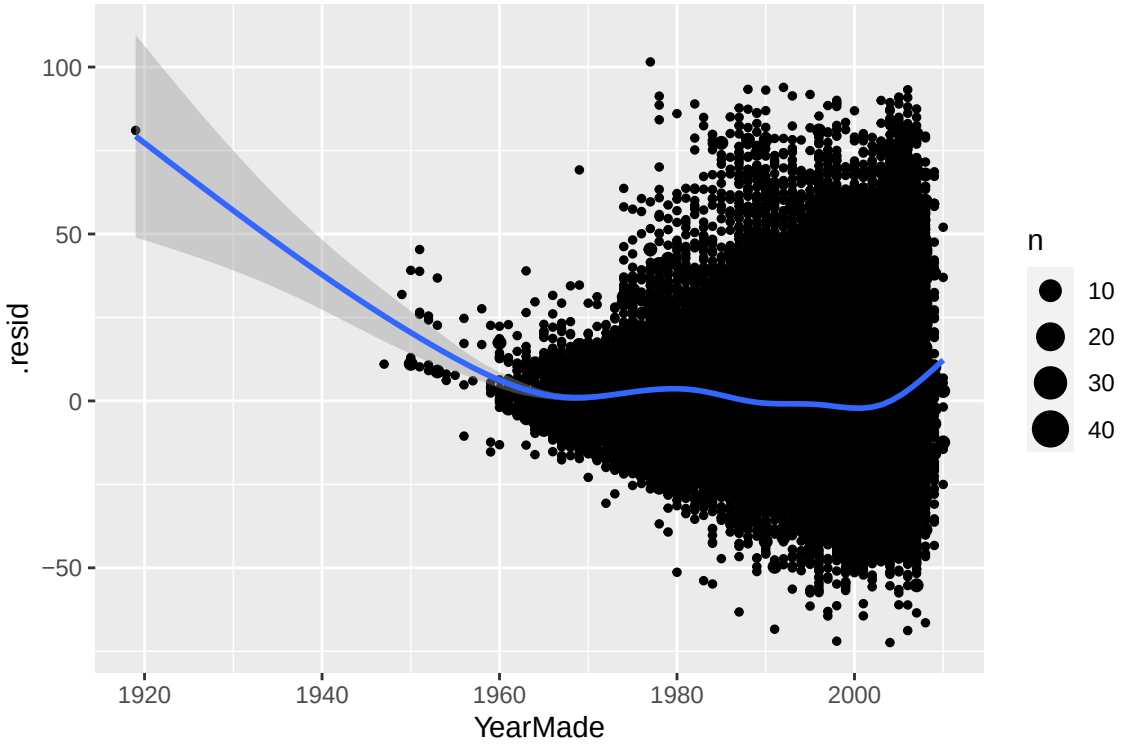
From the plot above, showing the interaction between 'YearMade' and 'ProductGroup' in this regression, we can see that the sale prices of all product groups has increased on average over their production time line. Out of these products Motor Graders have seen the greatest increase in price overall while Skid Steer Loaders have seen the lowest price increase overall. The newer the machine was, the higher the variance in sale price was as well according to this plot.



From the plot above, showing the interaction between 'YearMade' and 'Enclosure' in this regression, we can see that the sale prices of all enclosure types has increased on average over the production of the machines. We can see that 'EROPS w AC' have seen the greatest increase in sale price overall and 'EROPS' and 'OROPS have seen similar, low changes in sale price over the production of these machines. The newer the machine was, the higher the variance in sale price was as well according to this plot.

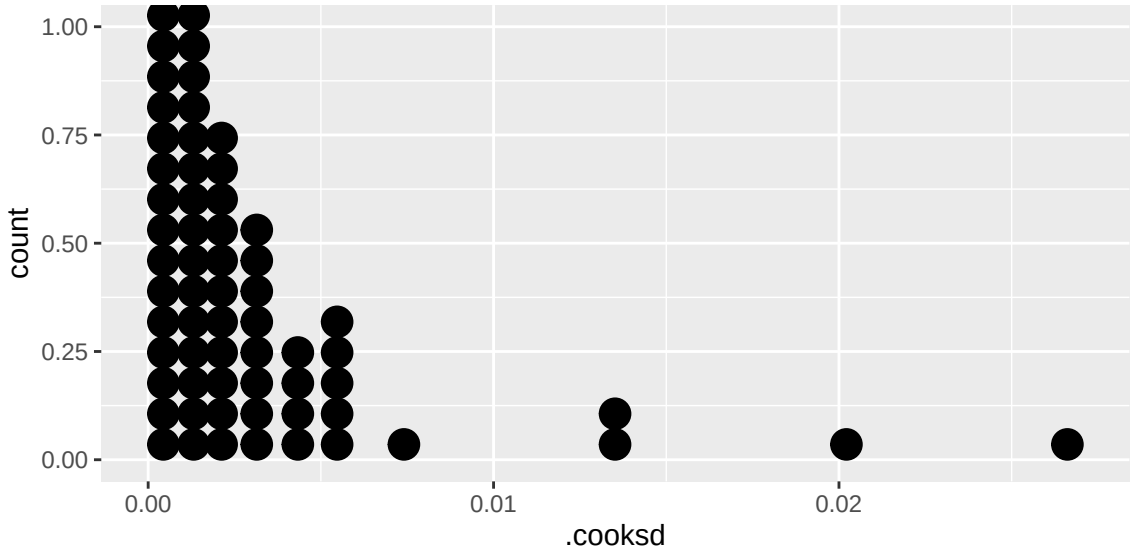
5. Diagnostic Plots to indicate heteroskedasticity and non-linearity:





We can see from the plot of residuals against fitted values that there is no constant variance in the data, showing the heteroskedasticity of the data. The points are randomly distributed around the estimated line creating points of high density and low density, with random variance. The next plot of residuals against 'YearMade' shows the non-linearity of this data. The estimated line curves down and seems to curve back up at the end of the plot. Even if outliers at the beginning are removed, the line curves downward, from around 1950 to 0, chowing no clear linearity in the data.

6.Plot displaying influential observations:




```
## Rows: 2
## Columns: 13
```

```
## $ .rownames    <chr> "308407", "369580"
## $ SalePrice    <dbl> 31.0, 16.5
## $ YearMade     <int> 1980, 1996
## $ MachineHours <dbl> 296.441, 821.747
## $ SaleDate     <int> 2007, 2010
## $ ProductGroup <chr> "Track Type Tractors", "Backhoe Loaders"
## $ Enclosure    <chr> "OROPS", "OROPS"
## $ .fitted      <dbl> -29.43186, 66.99130
## $ .resid       <dbl> 60.43186, -50.49130
## $ .hat         <dbl> 0.06269855, 0.10726263
## $ .sigma       <dbl> 16.56155, 16.56180
## $ .cooksd      <dbl> 0.02021555, 0.02661269
## $ .std.resid   <dbl> 3.768796, -3.226495
```

From this plot, we can see that Backhoe Loaders and Track Type Tractors withOROPS enclosures that were made in 1980 and 1996 and sold in 2007 and 2010 respectively were highly influential points in the model, using the cooks distance test. From the plot we can see that while most points had values below 0.01, with a few outliers between 0.01 and 0.015, these points had a cooks distance of greater than 0.02.