DATA 473 project

Attached Files:

- cubicz.csv (1.334 MB)
- cubic_DataDictionary.xlsx (16.037 KB)
- art.csv (971.955 KB)

This project assignment is worth 30% of your final mark. You will carry out a self-guided analysis of a dataset of your choice and write a report on your analyses. There are two datasets to choose from as described below. Based your project on **one** of these datasets.

2. Sculpture shipping costs

An art exhibitor plans to launch an online portal for art enthusiasts worldwide to collect art with only a click of a button. However, the logistics of selling and distributing art do not seem to be very straightforward. There are particular challenges in acquiring art effectively and shipping these artifacts to their respective destinations post-purchase. The exhibitor is interested in determining how different artifact attributes affect shipping cost. They would like to be able to predict the likely shipping cost of an artifact when they acquire it.

You are provided with a dataset of over 6000 artifacts for the analysis exercise. The dataset consists of variables such as the artist's name and reputation, dimensions, material, and price of the collectible, shipping details such as the customer information, scheduled dispatch, delivery dates, and so on.

The data were sourced from: https://www.hackerearth.com/challenges/competitive/hackerearth-machine-learning-challenge-predict-shipping-cost/ and are available in the file **art.csv.**

While there is no data dictionary available, the column names are self-explanatory.

Handling missing data

As a data science practitioner you often have to teach yourself new techniques to deal with different data issues. In this project you will learn about and implement the multiple imputation method in the **mice** package, and use it to deal with mising data in your chosen dataset, You may find the folllowing resources useful:

- University of Virginia blogpost: https://data.library.virginia.edu/getting-started-with-multiple-imputation-in-r/
- Video tutorial: https://www.youtube.com/watch?v=fsqAi9ddHU0

Project report

Write a **10-15 page (excluding appendices)** report as follows:

1. Create your report using Rmarkdown. Any output you include within the main body of the text should be in tables or graphs as appropriate, rather than a direct dump of R output.
2. While your report should be directed at the client, you must include a section that describes and discusses the methods you use, and the reasoning behind any analytical decisions you make.
3. Include all your code in an Appendix at the end and not interspersed with the text in the main body of your report.
4. I suggest you include the following sections in your report, but you don't need to stick to these:

- Introduction
- Data description, including and exploratory data analysis you undertake
- Methods - describe the approach you take
- Results
- Discussion and Conclusion

Submit both the .Rmd and .pdf files of your project report.

2. Sculpture shipping costs

An art exhibitor plans to launch an online portal for art enthusiasts worldwide to collect art with only a click of a button. However, the logistics of selling and distributing art do not seem to be very straightforward. There are particular challenges in acquiring art effectively and shipping these artifacts to their respective destinations post-purchase. The exhibitor is interested in determining how different artifact attributes affect shipping cost. They would like to be able to predict the likely shipping cost of an artifact when they acquire it.

You are provided with a dataset of over 6000 artifacts for the analysis exercise. The dataset consists of variables such as the artist's name and reputation, dimensions, material, and price of the collectible, shipping details such as the customer information, scheduled dispatch, delivery dates, and so on.

The data were sourced from: https://www.hackerearth.com/challenges/competitive/hackerearth-machine-learning-challenge-predict-shipping-cost/ and are available in the file **art.csv.**

While there is no data dictionary available, the column names are self-explanatory.

Handling missing data

As a data science practitioner you often have to teach yourself new techniques to deal with different data issues. In this project you will learn about and implement the multiple imputation method in the **mice** package, and use it to deal with mising data in your chosen dataset, You may find the folllowing resources useful:

- University of Virginia blogpost: https://data.library.virginia.edu/getting-started-with-multiple-imputation-in-r/
- Video tutorial: https://www.youtube.com/watch?v=fsqAi9ddHU0

Project report

Write a **10-15 page (excluding appendices)** report as

Sculpture shipping costs

1. Summary

An art exhibitor plans to launch an online portal for art enthusiasts worldwide to collect art with only a click of a button. The exhibitor is interested in determining how different artifact attributes affect shipping cost. They would like to be able to predict the likely shipping cost of an artifact when they acquire it.

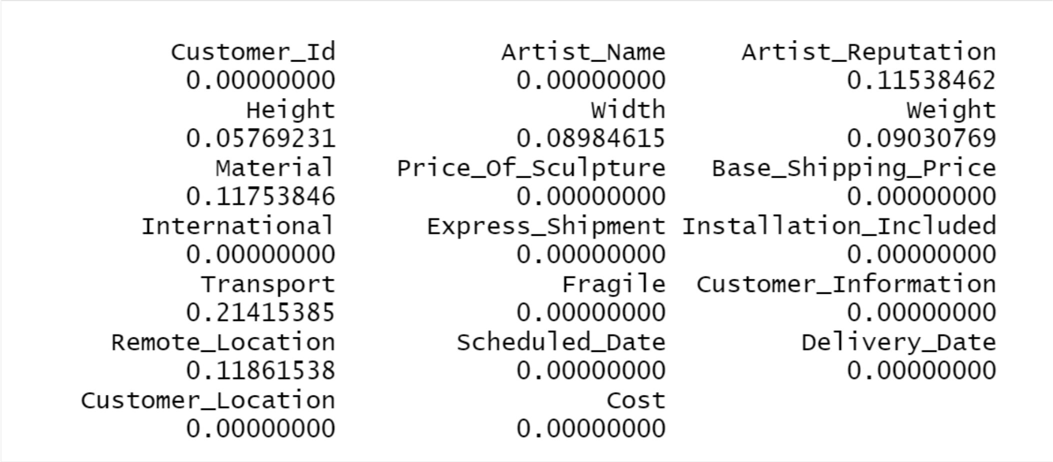
After exploratory data analysis, missing data imputation, several models were builtbuild for this project. There were investigation linear models with transformation, generalized additive models (GAM), subset selection and ridge regression.

2. EDA and data imputation

2.1. Dealing with missing data

Art set includes 6500 artifacts for the analysis exercise.   The dataset consists of “Customer Id”, “Artist Name”, “Artist Reputation”, “Height”, Width”, “Weight”, “Material”, “Price Of Sculpture” , “Base Shipping Price”   “International”, “Express Shipment”, “Installation Included”, “Transport”, “Fragile”, “Customer Information”, “Remote Location” , “Scheduled Date”, “Delivery Date”, “Customer Location”, “Cost” .

Set summary shows that 7 columns contains missing values. Omitting them lives only half of the set, which is far more then 25%, so data imputation will be the best way On the coloured matrix, table (Fig. 1) and heatmap we can see that affected columns are Width, Height, Weight, Artist Reputation, Material, Remote Location, Transport. Missing values spread across the set and clustering in the mentioned columns.



```
          Customer_Id           Artist_Name     Artist_Reputation
           0.00000000            0.00000000            0.11538462
               Height                 Width                Weight
           0.05769231            0.08984615            0.09030769
             Material    Price_Of_Sculpture   Base_Shipping_Price
           0.11753846            0.00000000            0.00000000
        International      Express_Shipment Installation_Included
           0.00000000            0.00000000            0.00000000
            Transport               Fragile  Customer_Information
           0.21415385            0.00000000            0.00000000
      Remote_Location        Scheduled_Date         Delivery_Date
           0.11861538            0.00000000            0.00000000
    Customer_Location                  Cost
           0.00000000            0.00000000
```

Percent of missing values inside of each category shows that transport , location, weight column contain the most of missing data.

The imputation had been performed via mice (Multivariate imputation by chained equations) package which implements a method to deal with missing data based on Fully Conditional Specification (see methodology).

2.2. Further EDA

From the first look at the predictors we can see that Customer ID column and Customer location does not have repetition however there is 51 times the same author name encounter. These columns could be considered for deletion.

Date column might influence and for easier calculation it could be converted to date stamps. Also 2 other columns produced, which includes years from scheduled and delivered dated and another column added as scheduled and delivered time difference to see, how timeline influence on shipping price.

The scatterplot matrix (Fig3) shows possible non-linear relationships between the response variable Cost and predictors Artist Reputation, Hight, Width - transformations of these predictors should be considered. The response variable Cost is skewed - a transformation should be considered for dealing with non-normality. The matrix plot shows moderate pairwise correlations between predictors weight and price of sculpture, width and height. Multicollinearity should be investigated. The plot also shows that all the date variables and artist reputation have almost 0 correlation with response variables.

As part of the task scheduled and delivery date were investigated and difference applied to the dataset. Also Years from these columns were extracted to separate column for further investigation. Time Difference has very low correlation to other data.

Non-numerical values are fit in box plot vs Cost. Almost of the box “flatten” to the bottom line and multiple outliers presented. To cut outliers boxplot method was used. On figure 6 boxplot for the same predictors are paired before and after.

Looking at the Material-Cost price boxplots (Fig.4)- heavy materials (as metal, stone, marble) affect Cost the most. Waterway shipping is slightly cheaper. Fragile items are cheaper in general, however have outliers. Express shipment, international shipment, remote location, working or wealthy customer status affect cost only slightly.

As result of Exploratory data analysis for the further investigation we can exclude Customer Id, Artist Name, Times Difference. As scheduled and delivery date have correlation one of the parameters could be excluded (and also times difference comes from normal distribution). Data set was saved as art_to_research.csv

3. Methodology

3.1. Data imputation

Data imputation have been performed via mice based on “fully conditional specification” or “sequential regression multiple imputation” . There is 2 approaches: single imputation based on means and not account to uncertainty of imputation and Compare to single imputation multiple imputation involved creating multiple sets, and not performing well if observed values are not predictable. The second approach involves making several datasets. MICE uses multiple approach but it is joing mode. MICE operates under the assumption that the missing data are Missing At Random (MAR). The chained equation process includes steps which are repeated for each missing variables and imputation updated after each cycle.

1.Mean imputation. 2. One “mean” placeholder set back to missing. 3.It regressed to other variables in the regression model. 4. This missing variable replaced with prediction of this regression model

Optimal number of cycles is specified by the researcher. I have set maxit parameter equals 5. Missing value analysis (Fig1-3) showed that all missing variables are spread equally across the set. There was not found reason to exclude some of the predictors, though artist reputation could be excluded as it is correlated only with cost and price of sculpture, I left it for further investigation.

Concerining other MICE parameters used for imputation PredictorMatrix and Method, default method, which is used in the current imputation method assumes that continuous data are imputed by predictive mean and matrix tells the algorithm which variables predict missingness in which other variables by default based on correlations between variables and the proportion of usable cases.

Complete function taking previous setting and number of set created. I have chosen 2. There are also options ‘long’ and ‘board’ which could be explored.

Imputed data set saved as “art_to_research.csv”

3.2. Linear regression models

For more clear picture numerical predictors fitted to response variable in scutter plots. (Fig. 5). Predictors which demonstrated non-linearity – basic_shipping_price, height and artist reputation.

To start subset selection we are fitting dataset to linear models and implementing the recommendation from EDA paragraph 2. The first fit includes all the predictors and do not have transformations.

Cost= β_0+β_1* Price_Of_Sculpture + β_2* Base_Shipping_Price + β_3* International + β_4* Express_Shipment + β_5* Installation_Included + β_6* Fragile + β_7* Customer_Information + β_8* Scheduled_Date + β_9* Delivery_Date + β_{10}* Artist_Reputation + β_{11}* Height + β_{12}* Width + β_{13}* Weight + β_{14}* Material + β_{15}* Remote_Location+ β_{16}* Transport+ β_{17}* Year_Scheduled+ β_{18}* Year_Delivery

Plots (Fig. 6) show that there is an indication of non-linearity (some curve Residual vs Fitted plot), non-normality (points not on line QQ plot) and non-constant variance when fitted values increase(Scale-Location plot) . Shapiro test shows that there is no evidence that residuals come from normal distribution. Breush-Pagan test shows that null hypothesis of homoskedasticity is rejected and heteroskedasticity (non-constant variance) assumed. Variance Inflation factor indicates that all date predictors have severe multicollinearity which may lead to not reliable coefficients. So they should be deleted or combined to single variable. Figure 7 contains plot after removing outliers and highly influential observation 4177, 4072 which appeared appeared after on Residuals vs Leverage plot were detected. With a different scale all previous observation became more visible.

For the next for fit the following modification had been done fit2 – removed date columns, fit3 – added log transformation for response variable, fit4 – added log transformation for basic shipment price, fit 5 – added log transformation for width , and polynomial transformation to Artist Reputation and Basic Shipment Price.

Assumptions check plots (Fig.8) demonstrate improvements for s non-normality, non-linearity, non-constant variance. To estimate progress R squared is used. Proportion of variance is accounted for by the model gradually increased after each modification.

Fit1	**Fit2**	**Fit3**	**Fit4**	**Fit5**
0.5341752	0.5465103	0.5671831	0.5765060	0.7177433

3.3. GAM

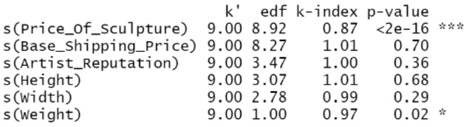
Generalised additive model could be another approach for multiple linear regression. It provides flexible smooth function smooth functions of some predictor variables, and interest focuses on inference about these smooth functions. Smoothing terms are spline which constructed as linear combinations of spline basis terms.

I included 2 fits to the project

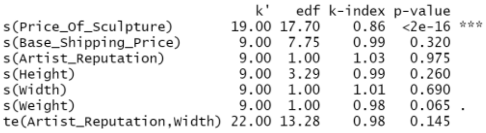
First fit includes smoothing terms applied to all numerical predictors and log transformation is left for response variable. In this model fragility seems to be less significant with p-value >0.05.

For the second GAM model interaction between Artist Reputation and Width was included.

According to gam check function, which performs residual diagnostic, first fit might require more degree of freedom for Price of Sculpture as it has low p-value k almost equal edf and k_index <0. Similarly for Weight. It was changed for the second fit.

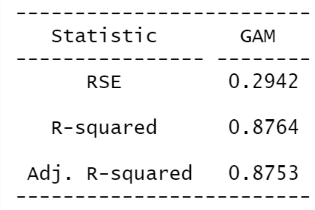


```
                          k'  edf k-index p-value    
s(Price_Of_Sculpture)   9.00 8.92    0.87  <2e-16 ***
s(Base_Shipping_Price)  9.00 8.27    1.01    0.70    
s(Artist_Reputation)    9.00 3.47    1.00    0.36    
s(Height)               9.00 3.07    1.01    0.68    
s(Width)                9.00 2.78    0.99    0.29    
s(Weight)               9.00 1.00    0.97    0.02 *  
```

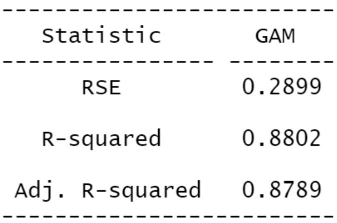


```
                               k'   edf k-index p-value    
s(Price_Of_Sculpture)       19.00 17.70    0.86  <2e-16 ***
s(Base_Shipping_Price)       9.00  7.75    0.99   0.320    
s(Artist_Reputation)         9.00  1.00    1.03   0.975    
s(Height)                    9.00  3.29    0.99   0.260    
s(Width)                     9.00  1.00    1.01   0.690    
s(Weight)                    9.00  1.00    0.98   0.065 .  
te(Artist_Reputation,Width) 22.00 13.28    0.98   0.145    
```

Models were assessed with R-squared, Adjusted R-squared and RSE. The results for the second gam fit is slightly higher and these results are is higher than for linear models from previous paragraph



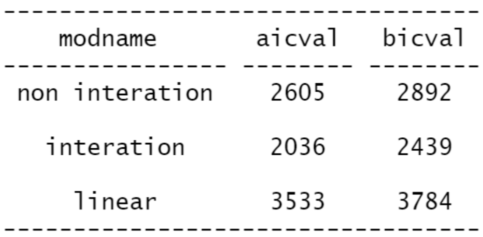
```
-------------------------
  Statistic        GAM   
---------------- --------
      RSE         0.2942 

   R-squared      0.8764 

 Adj. R-squared   0.8753 
-------------------------
```



```
-------------------------
  Statistic        GAM   
---------------- --------
      RSE         0.2899 

   R-squared      0.8802 

 Adj. R-squared   0.8789 
-------------------------
```

Another way to estimate models are Akaike Information Criterion , AIC and Bayes Information Criterion. The fit5 and both gam models were combined to the table.

According to both BIC and AIC criteria the best model is the second GAM model which included interaction of

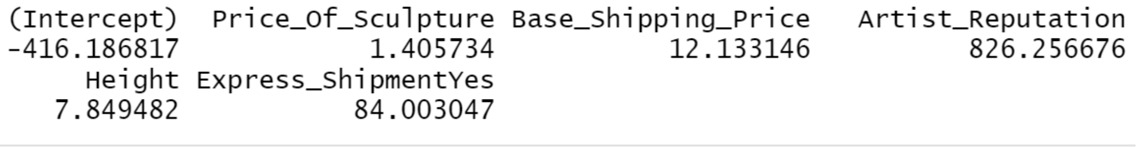


```
----------------------------------
    modname        aicval   bicval 
---------------- -------- --------
 non interation    2605     2892  

   interation      2036     2439  

     linear        3533     3784  
----------------------------------
```

3.4. Subset Selection

All the models still contain many predictors and it is not very obvious which of them are important for cost prediction. “Subset selection” methods had been used for this purpose to see how they could be reduced.

Library leaps contains “best subset”, “forward” and “backward” functions which could produce different set of predictors. For fist selection number of values was set 8 and 1 best model for each sum number of predictors. Figure 9 plots show that the best model have been chosen model with predictors and they are “Price_Of_Sculpture","Base_Shipping_Price,"Artist_Reputation","Height" and “Express Shipment”.

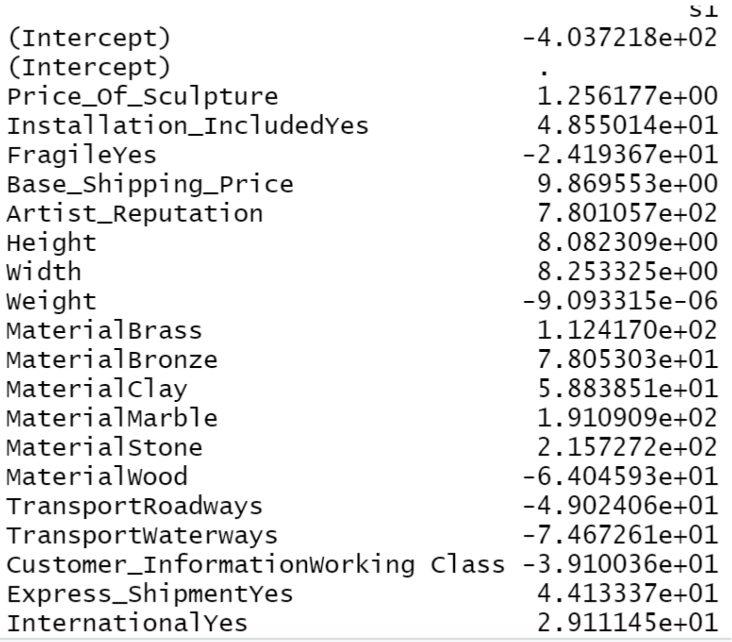


```
(Intercept)  Price_Of_Sculpture Base_Shipping_Price   Artist_Reputation
-416.186817            1.405734           12.133146          826.256676
     Height Express_ShipmentYes
   7.849482           84.003047
```

Forward and Backward selection left the same 5 predictors.

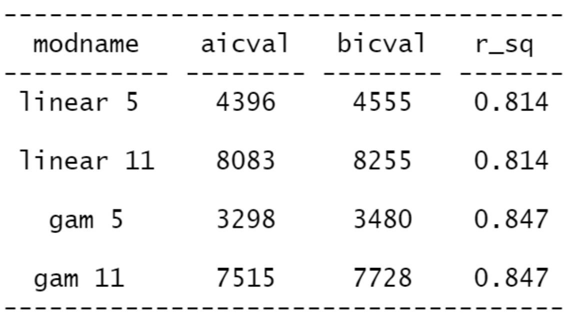
3.5. Ridge regression model

An alternative to subset selection methods ridge regression was applied as on of the shrinkage methods with glmnet method. The features fit to the model were "Price_Of_Sculpture","Installation_Included","Fragile", "Base_Shipping_Price","Artist_Reputation","Height","Width","Weight","Material","Transport","Customer_Information","Express_Shipment","International","Remote_Location" the same as for subset selection.
Best lambda =38. R-squared of predictions variables was calculated and equal 0.68. All the predictors are more then 0, however it is roughly seen that the more significant (with bigger coefficients) are the same “Price of sculpture”, “Base_Shipping_Price”, “Height”, “Express Shipment” . However “Width” coefficient is in the same group. Next follow “Material”,”Transport”,”Customer_Information”, ”Installation”



```
                                               s1
(Intercept)                         -4.037218e+02
(Intercept)                          .           
Price_Of_Sculpture                   1.256177e+00
Installation_IncludedYes             4.855014e+01
FragileYes                          -2.419367e+01
Base_Shipping_Price                  9.869553e+00
Artist_Reputation                    7.801057e+02
Height                               8.082309e+00
Width                                8.253325e+00
Weight                              -9.093315e-06
MaterialBrass                        1.124170e+02
MaterialBronze                       7.805303e+01
MaterialClay                         5.883851e+01
MaterialMarble                       1.910909e+02
MaterialStone                        2.157272e+02
MaterialWood                        -6.404593e+01
TransportRoadways                   -4.902406e+01
TransportWaterways                  -7.467261e+01
Customer_InformationWorking Class -3.910036e+01
Express_ShipmentYes                  4.413337e+01
InternationalYes                     2.911145e+01
```

3.6. Fits with reduced features



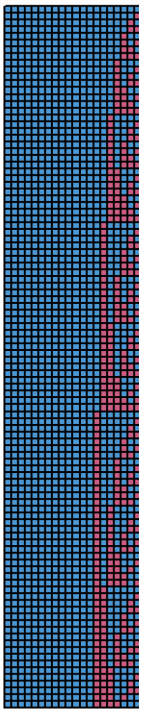
```
-------------------------------------
  modname     aicval   bicval   r_sq
----------- -------- -------- -------
 linear 5     4396     4555    0.814

 linear 11    8083     8255    0.814

   gam 5      3298     3480    0.847

  gam 11      7515     7728    0.847
-------------------------------------
```

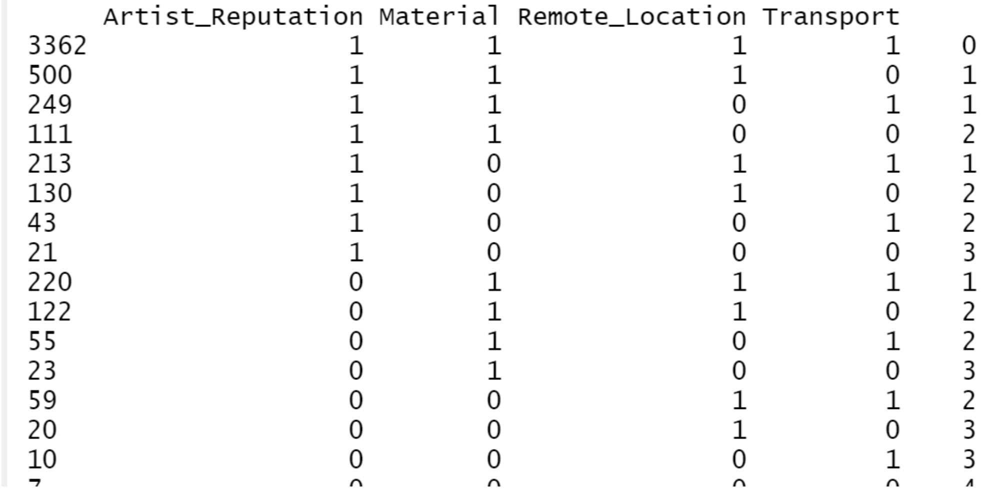
4. Results

5. Reference

1. Multiple imputation by chained equations: what is it and how does it work? - PMC (nih.gov) – MICE methods information
2. https://data.library.virginia.edu/getting-started-with-multiple-imputation-in-r/ -missing data investigation with mice.
3. Video tutorial: https://www.youtube.com/watch?v=fsqAi9ddHU0 – data imputation
4. “An Introduction to Statistical Learning with Application R” Gareth James, Daniela Witten, Trevor Hastie, Robert
5. Course materials
6. Ridge Regression in R (Step-by-Step) (statology.org) – r-squared
7. Predict in R: Model Predictions and Confidence Intervals - Articles - STHDA -prediction interval and confidence intervals,
8. autosmooth.pdf (anu.edu.au) -GAM

6. Appendix





```
     Artist_Reputation Material Remote_Location Transport   
3362                 1        1               1         1  0
500                  1        1               1         0  1
249                  1        1               0         1  1
111                  1        1               0         0  2
213                  1        0               1         1  1
130                  1        0               1         0  2
43                   1        0               0         1  2
21                   1        0               0         0  3
220                  0        1               1         1  1
122                  0        1               1         0  2
55                   0        1               0         1  2
23                   0        1               0         0  3
59                   0        0               1         1  2
20                   0        0               1         0  3
10                   0        0               0         1  3
7                    0        0               0         0  4
```

Figure 1. Missing values pattern matrix plot and numbers (fragment)

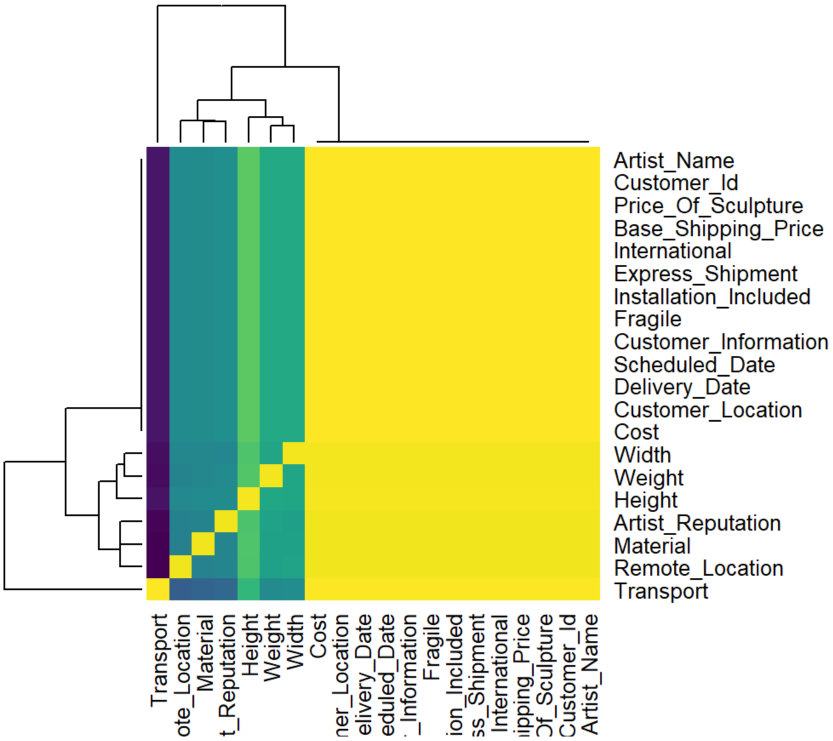


Figure 2. Missing values heatmap

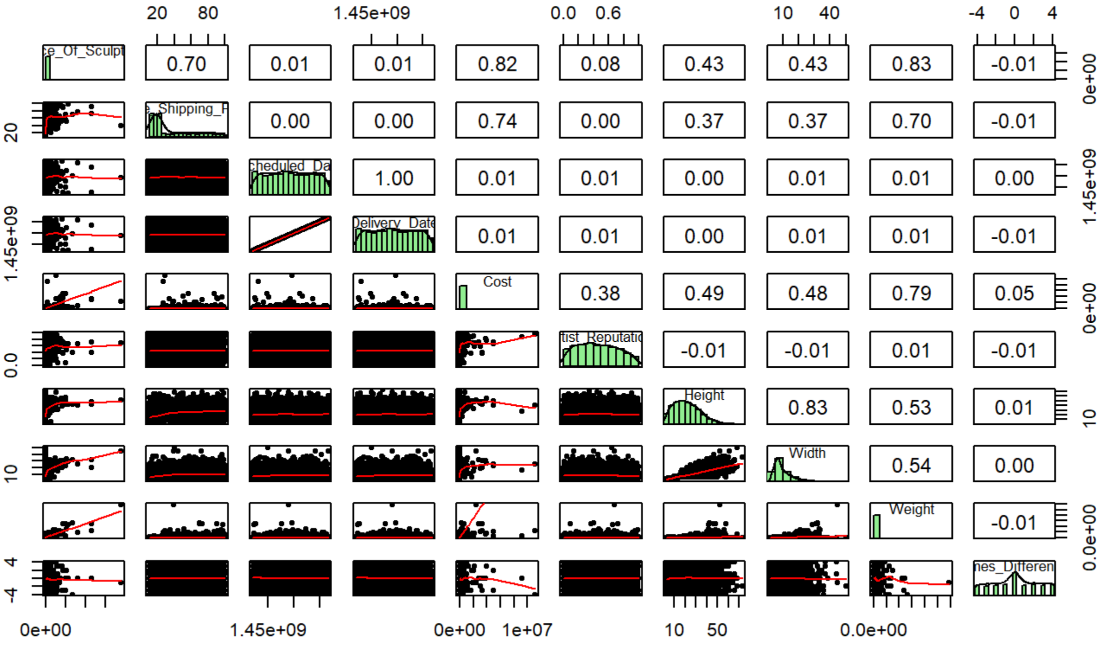
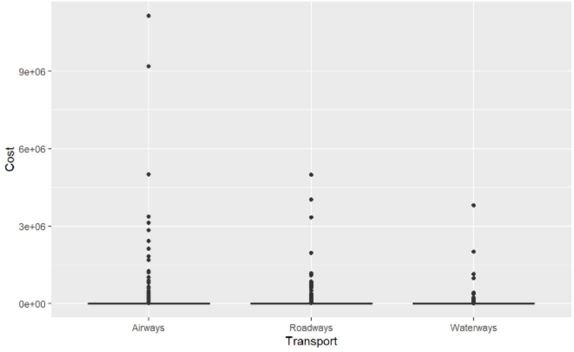
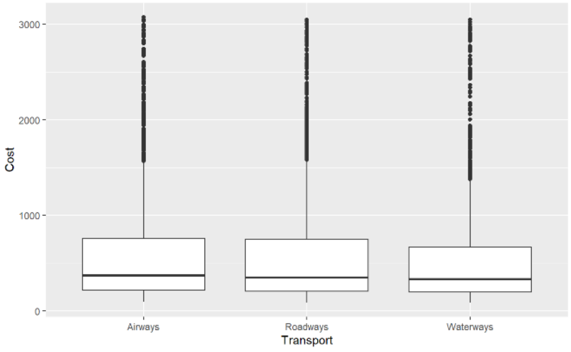
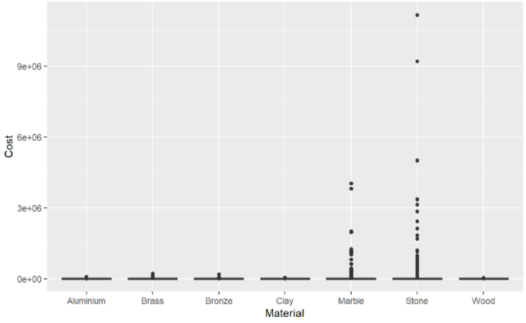
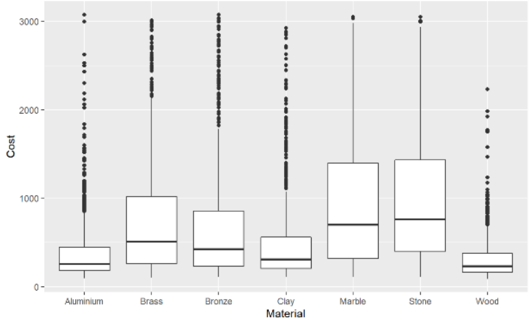


Figure 3. Data matrix plot









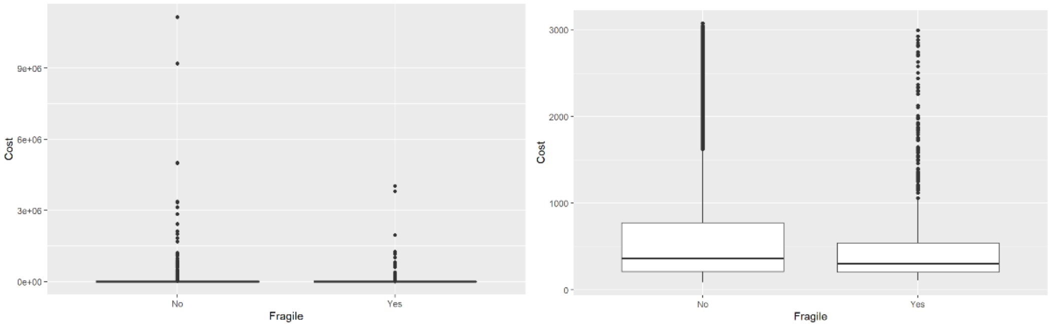


Figure 4. Boxplots for non-numeric predictors

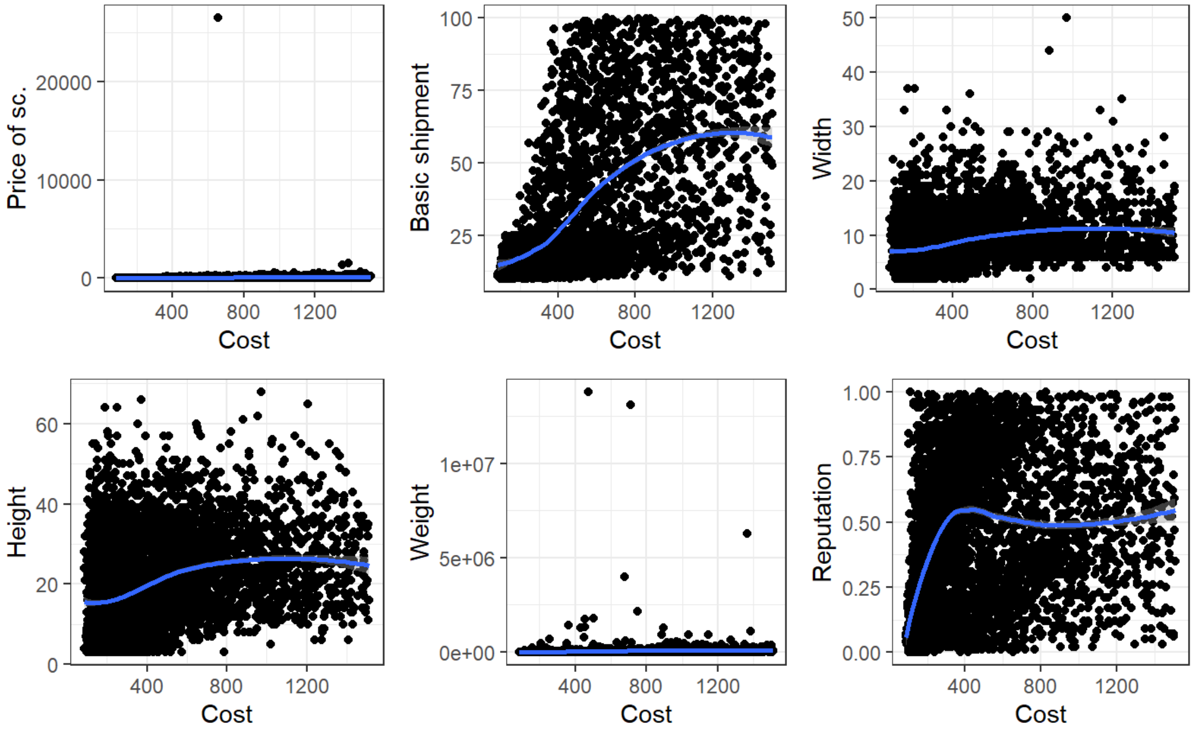
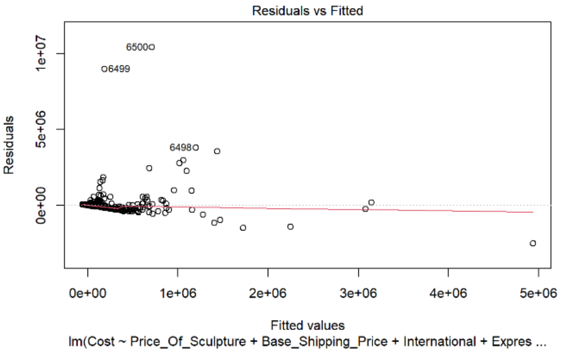
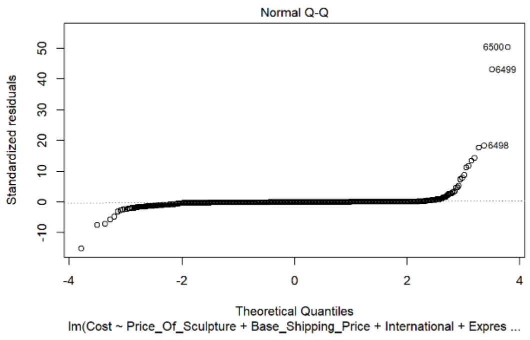


Figure 5. Scatter plots for response variable vs numerical predictors.





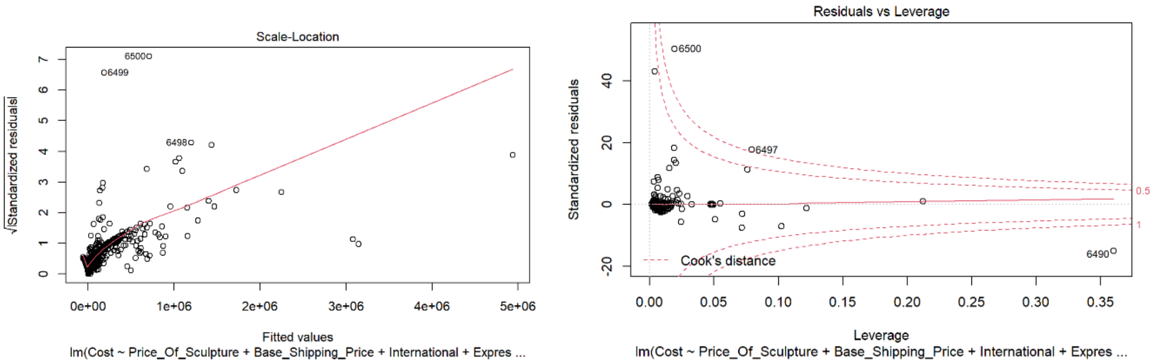


Figure 6. Initial linear model assumptions check plots

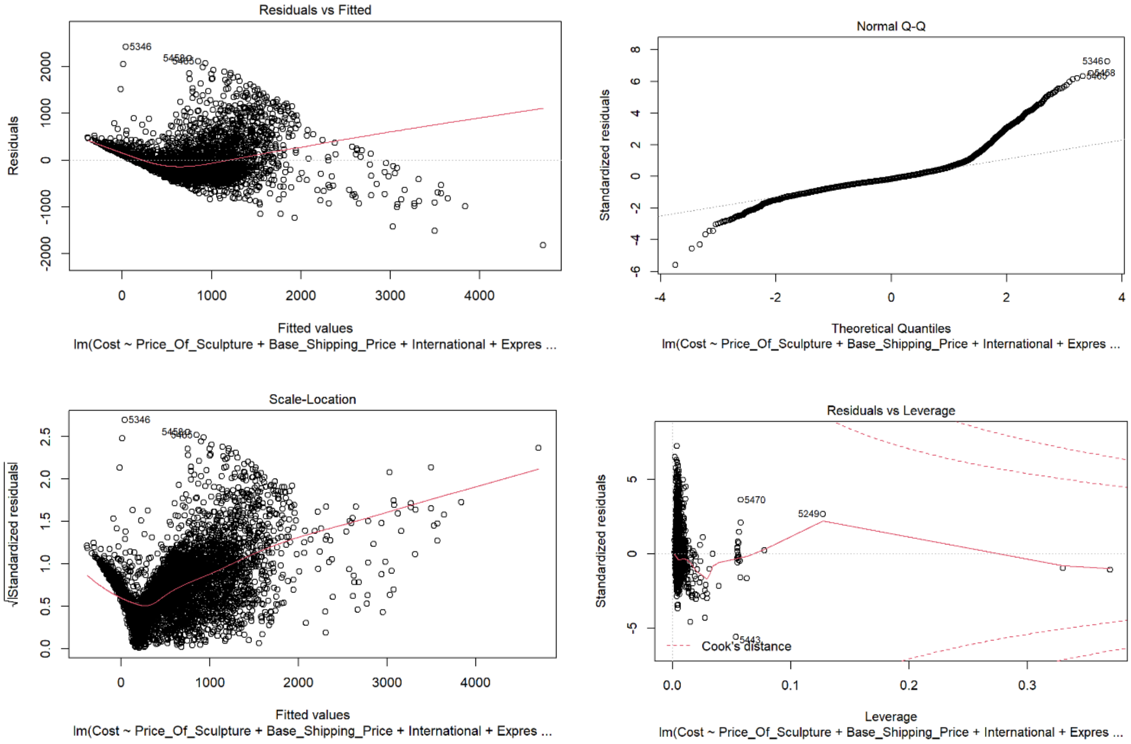
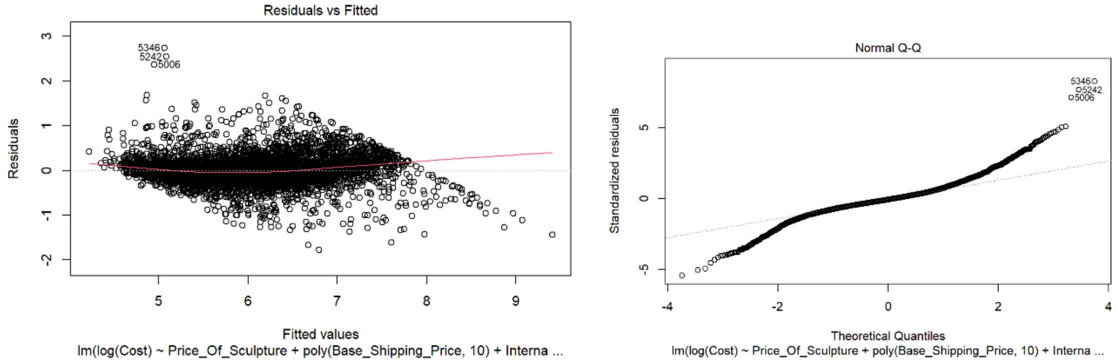


Figure 7. Linear model assumptions check without outliers and highly influential points



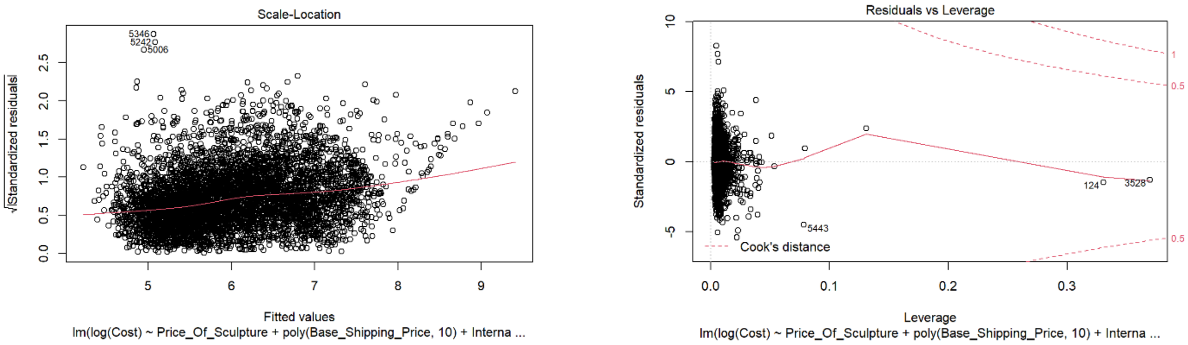


Figure 8. Modified linear model assumptions check plot

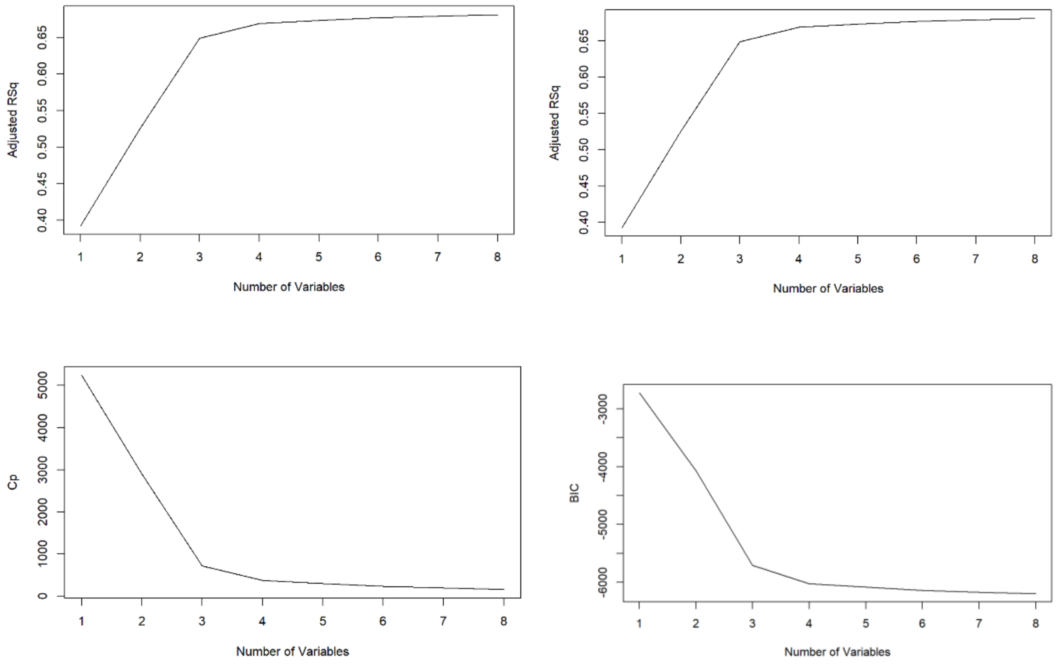


Figure 9. Best subset selection results

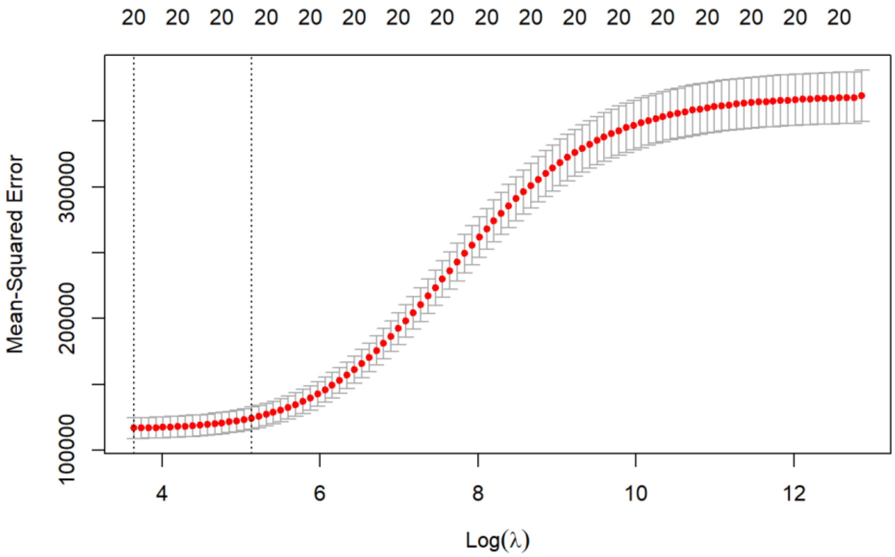


Figure 10. Best lambda selection