MLR modeling and feature-based hypothesis testing

This PDF provides an overview of the statistical methods used within this project. The document includes details on the process of price prediction using multiple linear regression (MLR) and hypothesis testing to investigate price differences between laptops with and without touchscreen, as well as between laptops with and without IPS panels. You can use this PDF as a guide, where I explain the process with visualizations and description step by step.

Multiple linear regression

- Objective: *Estimate coefficients for different laptop features to predict price.*

- Model building: *Constructed regression models.*

- Train and test data: *Split data into 80% to train model and 20% to evaluate model performance.*

-Validation: *Validate model goodness of fit with metrics, R-squared and RMSE.*

- Model diagnostic: *Ensure MLR-assumptions.*

Hypothesis testing

-Objective: *When exploring the data, differences in prices between different laptop features were found. For example, laptops with a touchscreen were more expensive than those without. Additionally, laptops with IPS panels were more expensive than those without. This leads to two hypotheses: laptops with touchscreens are more expensive than those without touchscreens, and laptops with IPS panels are more expensive than those without IPS panels.*

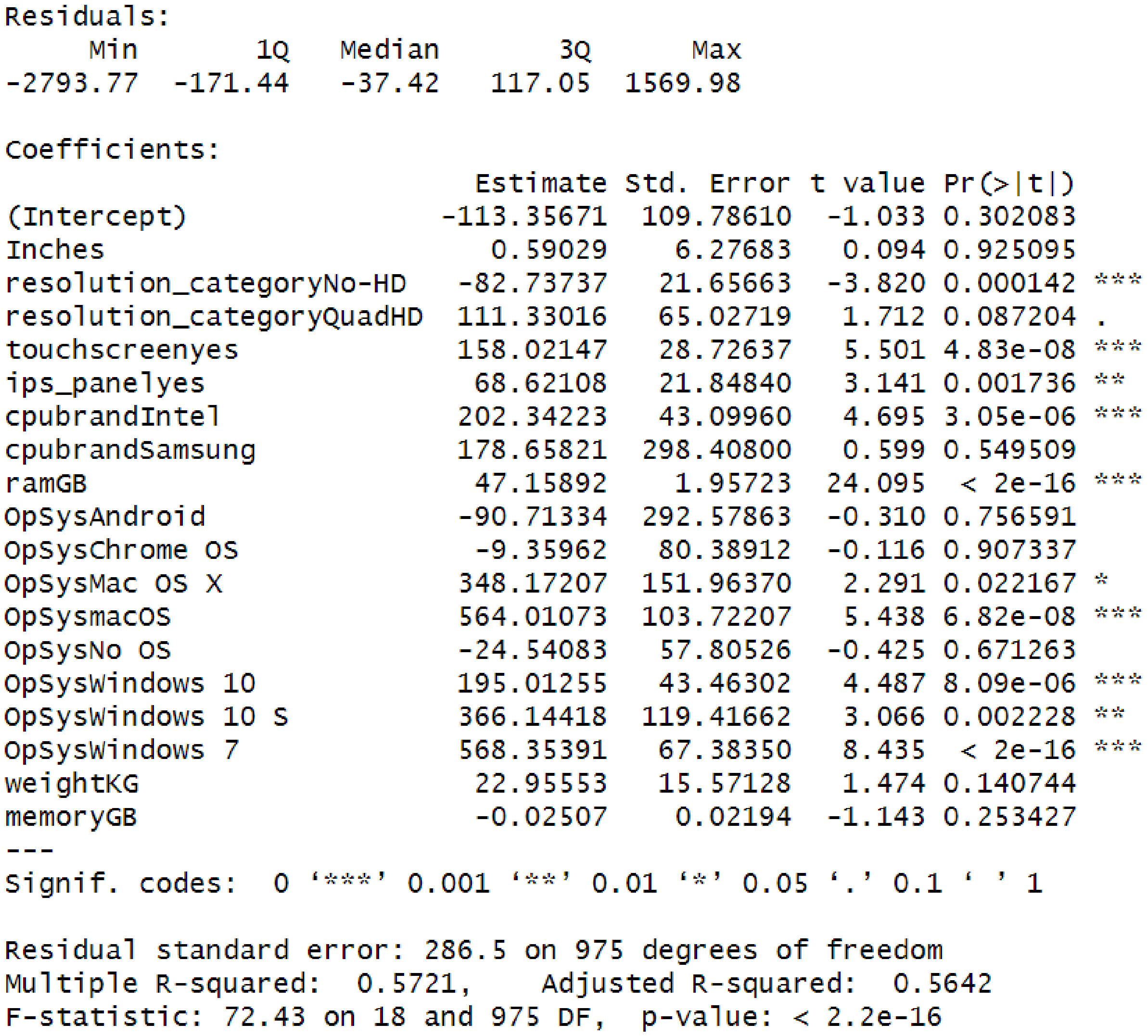
-Shapiro wilk: *Normality test of price distribution for each group.*

Levene's test: *Test for equal variance if data is normally distributed.*

-Two sample t-test: *Conduct two sample t-tests if assumption of normal distributed data holds.*

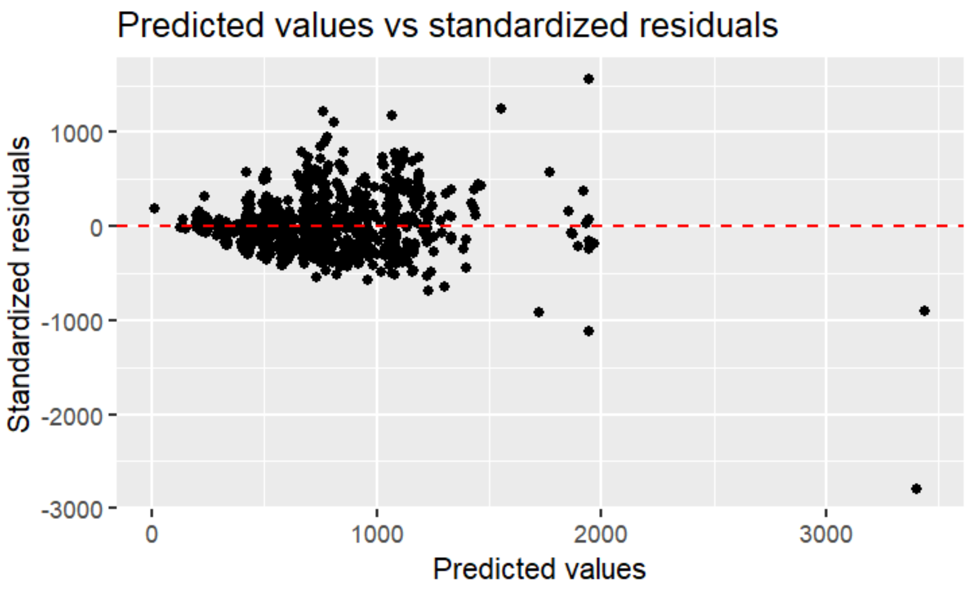
- Mann whitney u test: *Alternative test, if assumption of normal distributed data doesn't hold.*

MLR-model (1)

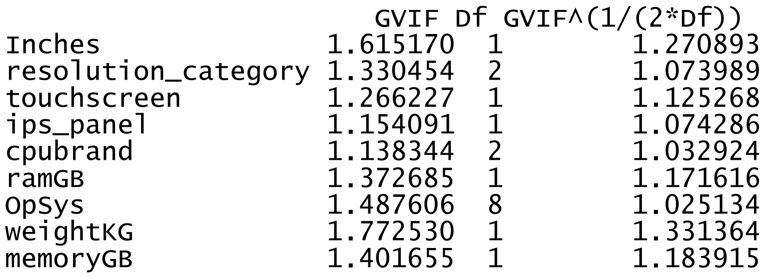


```
Residuals:
     Min       1Q   Median       3Q      Max 
-2793.77  -171.44   -37.42   117.05  1569.98 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -113.35671  109.78610  -1.033 0.302083    
Inches                       0.59029    6.27683   0.094 0.925095    
resolution_categoryNo-HD   -82.73737   21.65663  -3.820 0.000142 ***
resolution_categoryQuadHD  111.33016   65.02719   1.712 0.087204 .  
touchscreenyes             158.02147   28.72637   5.501 4.83e-08 ***
ips_panelyes                68.62108   21.84840   3.141 0.001736 ** 
cpubrandIntel              202.34223   43.09960   4.695 3.05e-06 ***
cpubrandSamsung            178.65821  298.40800   0.599 0.549509    
ramGB                       47.15892    1.95723  24.095  < 2e-16 ***
OpSysAndroid               -90.71334  292.57863  -0.310 0.756591    
OpSysChrome OS              -9.35962   80.38912  -0.116 0.907337    
OpSysMac OS X              348.17207  151.96370   2.291 0.022167 *  
OpSysmacOS                 564.01073  103.72207   5.438 6.82e-08 ***
OpSysNo OS                 -24.54083   57.80526  -0.425 0.671263    
OpSysWindows 10            195.01255   43.46302   4.487 8.09e-06 ***
OpSysWindows 10 S          366.14418  119.41662   3.066 0.002228 ** 
OpSysWindows 7             568.35391   67.38350   8.435  < 2e-16 ***
weightKG                    22.95553   15.57128   1.474 0.140744    
memoryGB                    -0.02507    0.02194  -1.143 0.253427    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 286.5 on 975 degrees of freedom
Multiple R-squared:  0.5721,	Adjusted R-squared:  0.5642 
F-statistic: 72.43 on 18 and 975 DF,  p-value: < 2.2e-16
```

Model 1 shows that certain features significantly influence the price of laptops. Touchscreen, IPS panel, RAM, and OP-systems (macOS, Windows 10, etc.) are significantly different coefficients from zero. However, the intercept and some predictors like inches, CPU brand (Samsung), and some operating systems (Android, Chrome OS, no OS) were not statistically significant. Before interpreting the model coefficients, validation and diagnostics were conducted. This included visualizations such as plotting standardized residuals against predicted values. The estimated RMSE for this model is 340.39 with R-squared 0.57 which explains 57% variance in response explained by predictors. This could potentially be improved by further manipulation, diagnostics and control for outliers.

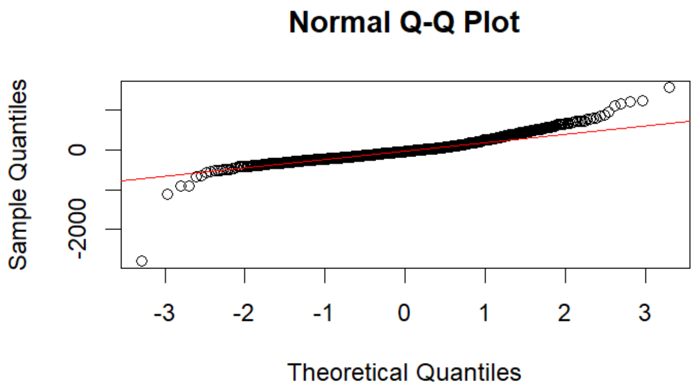


Upon inspection of the plot, heteroscedasticity was found, violating the assumption of MLR. Actions that can be taken include transforming the response variable or redefining it. Outliers can also have an impact and will be removed to train a new model.

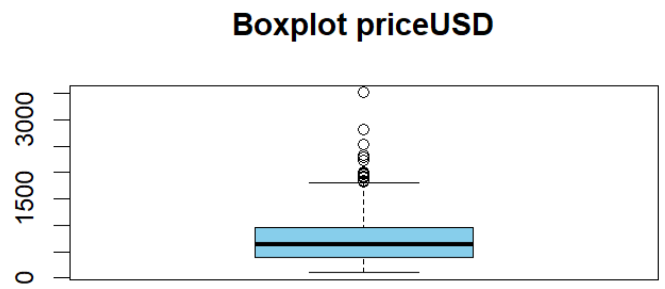


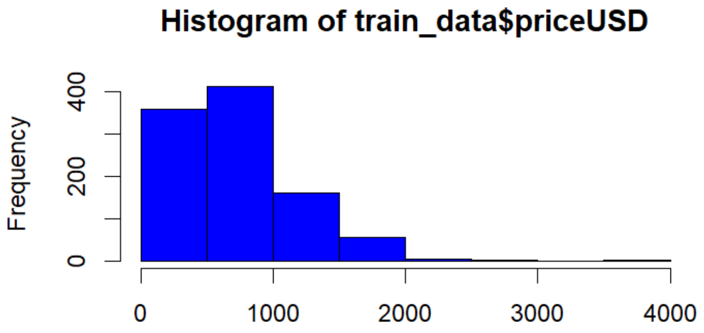
```
                        GVIF Df GVIF^(1/(2*Df))
Inches              1.615170  1        1.270893
resolution_category 1.330454  2        1.073989
touchscreen         1.266227  1        1.125268
ips_panel           1.154091  1        1.074286
cpubrand            1.138344  2        1.032924
ramGB               1.372685  1        1.171616
OpSys               1.487606  8        1.025134
weightKG            1.772530  1        1.331364
memoryGB            1.401655  1        1.183915
```

The table above shows the VIF values where no multicollinearity could be found.

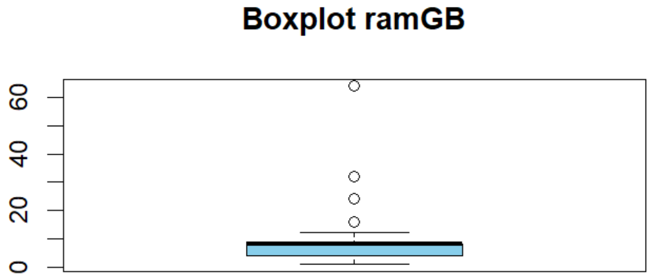


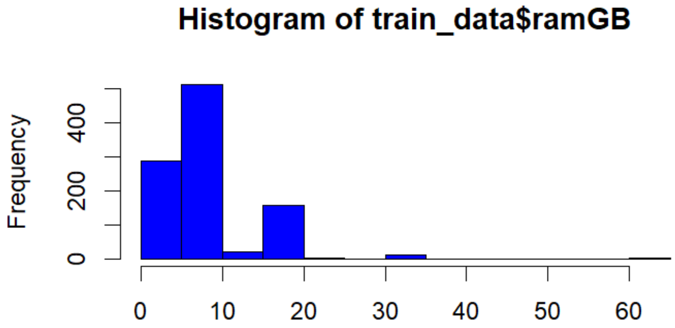
When analyzing QQ-plot, it can be concluded that residuals roughly follow a normal distribution.





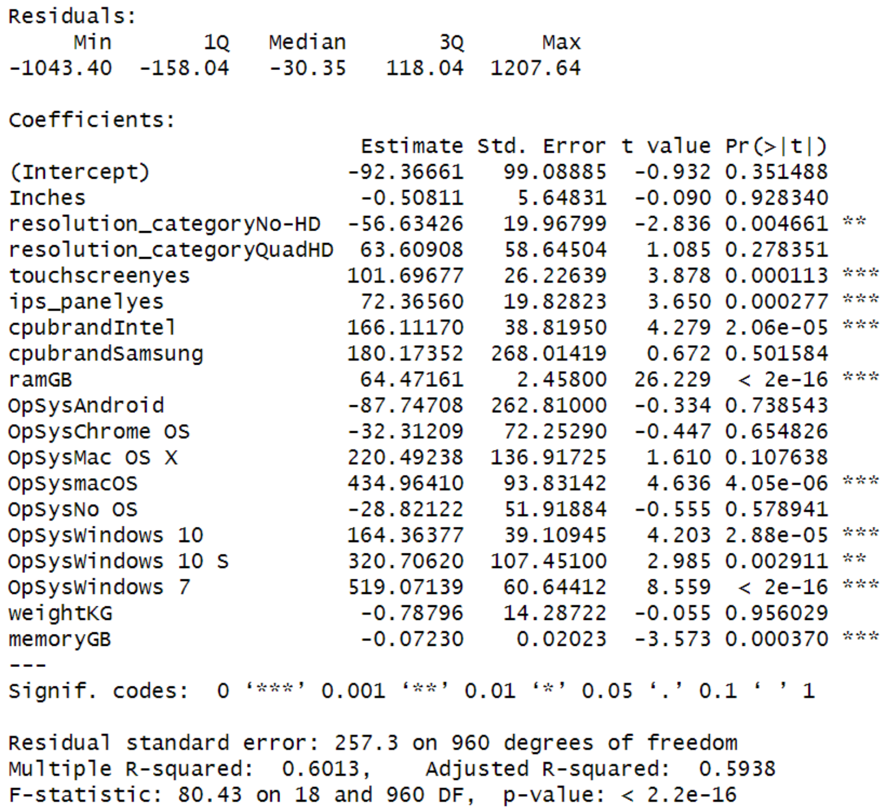
When analyzing the histogram and box plot for the response variable priceUSD, it can be concluded that outliers exist. Outliers will be removed using z-score.





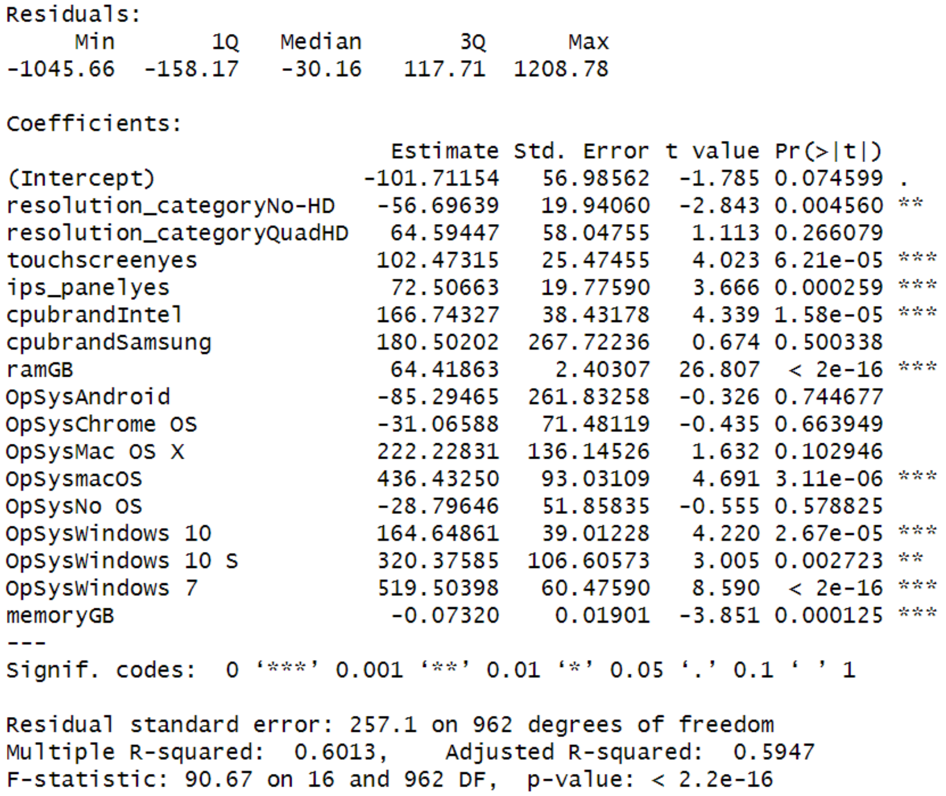
Outliers are also detected in the independent variable ramGB based on inspection of histograms and box plot. Outliers make the distributions skewed, which degrades the trained model's performance. New models are created for further validation of model performance.

MLR-model (2)

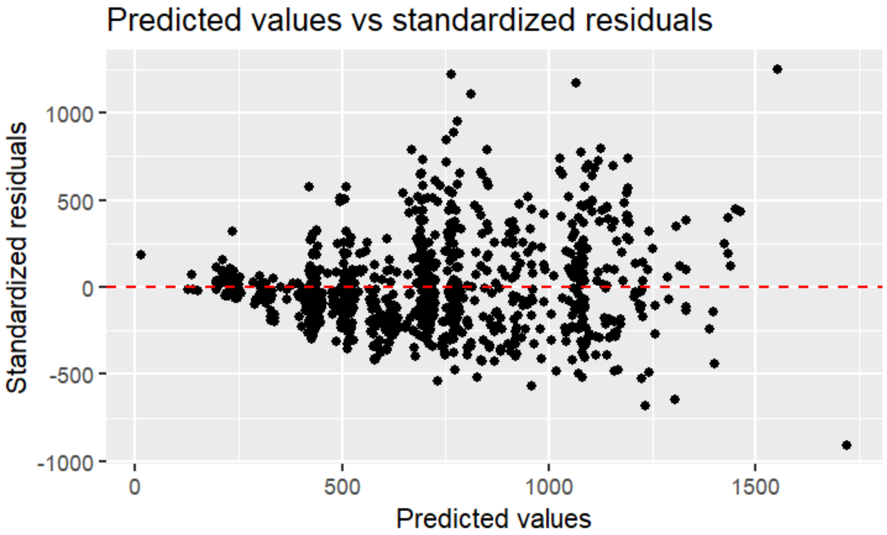


```
Residuals:
     Min       1Q   Median       3Q      Max 
-1043.40  -158.04   -30.35   118.04  1207.64 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -92.36661   99.08885  -0.932 0.351488    
Inches                     -0.50811    5.64831  -0.090 0.928340    
resolution_categoryNo-HD  -56.63426   19.96799  -2.836 0.004661 ** 
resolution_categoryQuadHD  63.60908   58.64504   1.085 0.278351    
touchscreenyes            101.69677   26.22639   3.878 0.000113 ***
ips_panelyes               72.36560   19.82823   3.650 0.000277 ***
cpubrandIntel             166.11170   38.81950   4.279 2.06e-05 ***
cpubrandSamsung           180.17352  268.01419   0.672 0.501584    
ramGB                      64.47161    2.45800  26.229  < 2e-16 ***
OpSysAndroid              -87.74708  262.81000  -0.334 0.738543    
OpSysChrome OS            -32.31209   72.25290  -0.447 0.654826    
OpSysMac OS X             220.49238  136.91725   1.610 0.107638    
OpSysmacOS                434.96410   93.83142   4.636 4.05e-06 ***
OpSysNo OS                -28.82122   51.91884  -0.555 0.578941    
OpSysWindows 10           164.36377   39.10945   4.203 2.88e-05 ***
OpSysWindows 10 S         320.70620  107.45100   2.985 0.002911 ** 
OpSysWindows 7            519.07139   60.64412   8.559  < 2e-16 ***
weightKG                   -0.78796   14.28722  -0.055 0.956029    
memoryGB                   -0.07230    0.02023  -3.573 0.000370 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 257.3 on 960 degrees of freedom
Multiple R-squared:  0.6013,	Adjusted R-squared:  0.5938 
F-statistic: 80.43 on 18 and 960 DF,  p-value: < 2.2e-16
```

MLR-model (3)



```
Residuals:
     Min       1Q   Median       3Q      Max 
-1045.66  -158.17   -30.16   117.71  1208.78 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -101.71154   56.98562  -1.785 0.074599 .  
resolution_categoryNo-HD   -56.69639   19.94060  -2.843 0.004560 ** 
resolution_categoryQuadHD   64.59447   58.04755   1.113 0.266079    
touchscreenyes             102.47315   25.47455   4.023 6.21e-05 ***
ips_panelyes                72.50663   19.77590   3.666 0.000259 ***
cpubrandIntel              166.74327   38.43178   4.339 1.58e-05 ***
cpubrandSamsung            180.50202  267.72236   0.674 0.500338    
ramGB                       64.41863    2.40307  26.807  < 2e-16 ***
OpSysAndroid               -85.29465  261.83258  -0.326 0.744677    
OpSysChrome OS             -31.06588   71.48119  -0.435 0.663949    
OpSysMac OS X              222.22831  136.14526   1.632 0.102946    
OpSysmacOS                 436.43250   93.03109   4.691 3.11e-06 ***
OpSysNo OS                 -28.79646   51.85835  -0.555 0.578825    
OpSysWindows 10            164.64861   39.01228   4.220 2.67e-05 ***
OpSysWindows 10 S          320.37585  106.60573   3.005 0.002723 ** 
OpSysWindows 7             519.50398   60.47590   8.590  < 2e-16 ***
memoryGB                    -0.07320    0.01901  -3.851 0.000125 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 257.1 on 962 degrees of freedom
Multiple R-squared:  0.6013,	Adjusted R-squared:  0.5947 
F-statistic: 90.67 on 16 and 962 DF,  p-value: < 2.2e-16
```



After removing the outliers, the analysis of heteroscedasticity was slightly improved, better than for the first model, this is something that should be considered when evaluating the trained model. R-squared increases to 0.60 which is an improvement over model 1. The RMSE has also decreased to 253.0471.

Model (3) key findings

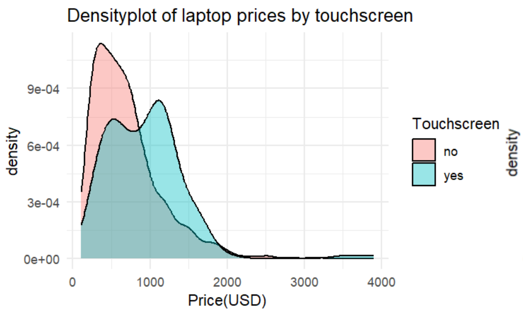
- *Touchscreen: Laptops with a touchscreen cost about $102.47 more than laptops without touchscreen.*
- *IPS panel: Laptops with an IPS panel cost about $72.51 more than laptops without IPS panel.*
- *RAM (GB): Each additional GB of RAM increases the price by $64.42.*
- *CPU Brand (Intel): Laptops with Intel CPUs cost about $166.74 more than laptops with AMD.*
- *Operating system (macOS): Laptops with macOS cost about $436.43 more than laptops with Linux.*
- *Operating system (Windows 10): Laptops with Windows 10 cost about $164.65 more than laptops with Linux.*
- *Operating system (Windows 10 S): Laptops with Windows 10 S cost about $320.38 more than laptops with Linux.*
- *Operating system (Windows 7): Laptops with Windows 7 cost about $519.50 more than laptops with Linux.*

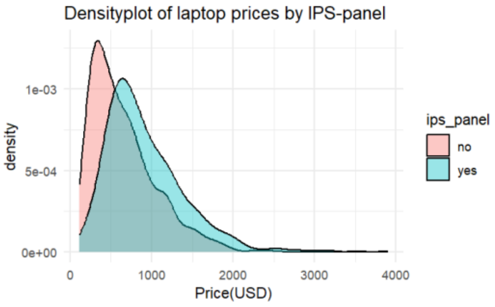
The model is statistically significant with R-squared value of 0.60 resulting in 60% of the variation in laptop prices explained by independent variables. RMSE of 253.05 shows improved accuracy after dealing with heteroscedasticity and outliers. Still, indication of heteroscedasticity exists, which should be considered when relying on model prediction.

Hypothesis testing

In exploratory data analysis, differences in prices between different laptop features were identified. Regression modeling showed significant coefficients: the average price for laptops with a touchscreen function is $102 more than for those without a touchscreen, and laptops with an IPS panel are, on average, $72 more expensive than those without an IPS panel.

Previous analysis also established that the priceUSD variable does not follow a normal distribution, a finding confirmed within the yes/no subsets for touchscreen and IPS panel features. Because the groups did not follow a normal distribution, the two-sample t-test could not be used, as it requires normally distributed data. Since subsets doesn't follows a normal distribution, test for equal variance is not necessary and instead, the man-whitney-u test was used to test the median prices.





The following two hypotheses were established:

Laptops with a touchscreen function are more expensive and have a higher median price than laptops without touchscreen functionality.

Laptops with IPS-panel are more expensive and have a higher median price than laptops without IPS-panel.

Touchscreen:

- **Null Hypothesis (H_0)**: The median price of laptops with a touchscreen function is equal to the median price of laptops without a touchscreen function.

  Mediantouchscreen = Median non-touchscreen

- **Alternative Hypothesis (H_1)**: The median price of laptops with a touchscreen function is greater than the median price of laptops without a touchscreen function. Mediantouchscreen > Mediannon-touchscreen

- **Test Statistic (W)**: 127,172
- **p-value**: 1.726e-12

Given the p-value is extremely small (much less than the significance level of 0.05), we reject the null hypothesis. This provides strong evidence to support that laptops with a touchscreen function have a significantly higher median price compared to laptops without a touchscreen function. The results of the test indicate that there is a statistically significant difference in the median prices, with laptops that have a touchscreen being more expensive than those without.

Ips-panel:

- **Null Hypothesis (H_0)**: The median price of laptops with an IPS panel is equal to the median price of laptops without an IPS panel.

  H0:Median IPS panel=Median non-IPS panel

- **Alternative Hypothesis (H_1)**: The median price of laptops with an IPS panel is greater than the median price of laptops without an IPS panel.

  H1:MedianIPS panel>Mediannon-IPS panel

- **Test Statistic (W)**: 213,261
- **p-value**: < 2.2e-16

Given the p-value is extremely small (much less than the typical significance level of 0.05), we reject the null hypothesis. This provides strong evidence to support that laptops with an IPS panel have a significantly higher median price compared to laptops without an IPS panel. The results of the test indicate that there is a statistically significant difference in the median prices, with laptops that have an IPS panel being more expensive than those without.