DAB501 Final Project - Amazon Sales Dataset

Group 10: 1.Srilakshmi Gummadidala 2.Divyajot Singh Mankan 3.Jibina Francise

2023-04-21

- Student Information
- Academic Integrity
- Reading Packages
- Reading Dataset
- Data Set Description
  - Variables/Features
- MODELING: First pair of variables
- MODELING: Second pair of variables
- MODEL ASSESSMENT
- MODEL DIAGNOSTICS
- CONCLUSION
- Thank You!

Student Information

Name: Srilakshmi Gummadidala G.Srilakshmi

ID: 0803509

Name: Divyajot Singh Mankan Divyajot Singh Mankan

ID: 0822915

Name: Jibina Francis Jibina Francis

ID: 0822959

Academic Integrity

We, **Srilakshmi Gummadidala,Divyajot Singh Mankan,Jibina Francis**, hereby state that I have not communicated with or gained information in any way from any person or resource that would violate the College's academic integrity policies, and that all work is my own.

Reading Packages

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.2.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
## Warning: package 'tibble' was built under R version 4.2.3
```

```
## Warning: package 'tidyr' was built under R version 4.2.3
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
## Warning: package 'forcats' was built under R version 4.2.3
```

```
## Warning: package 'lubridate' was built under R version 4.2.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.1     ✓ purrr     1.0.1
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✗ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the �]8;;http://conflicted.r-lib.org/�conflicted package�]8;;� to force all conflicts
```

```r
install.packages("ggplot2")
```

```
## Warning: package 'ggplot2' is in use and will not be installed
```

```r
library(ggplot2)
library(dplyr)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
library(ggthemes)
```

Reading Dataset

```r
library(readxl)
installed.packages("readxl")
```

```
##      Package LibPath Version Priority Depends Imports LinkingTo Suggests
##      Enhances License License_is_FOSS License_restricts_use OS_type Archs
##      MD5sum NeedsCompilation Built
```

```
Amazon_Subset4 <- read_xlsx("C:/Users/Sreelu/OneDrive - St. Clair College/Documents/St
        Clair/DAB501/Datasets/Amazon_Subset4.xlsx")
```

Data Set Description

This actual Amazon dataset is having the data of 1K+ Amazon product’s in different categories along with ratings and reviews as per their details listed on the official website of Amazon.

We have used subset of the actual amazon sales dataset with total six categories in it and sub classified the categories column variables.Total 573 Rows and 16 Columns in the currently used subset

Variables/Features

PRODUCT_ID : PRODUCT ID

PRODUCT_NAME : NAME OF THE PRODUCT

CATEGORY : CATEGORY OF THE PRODUCT

DISCOUNTED_PRICE : DISCOUNTED PRICE OF THE PRODUCT

ACTUAL_PRICE : ACTUAL PRICE OF THE PRODUCT

DISCOUNT_PERCENTAGE : PERCENTAGE OF DISCOUNT FOR THE PRODUCT

RATING : RATING OF THE PRODUCT

RATING_COUNT : NUMBER OF PEOPLE WHO VOTED FOR THE AMAZON RATING

ABOUT_PRODUCT : DESCRIPTION ABOUT THE PRODUCT

USER_ID : ID OF THE USER WHO WROTE REVIEW FOR THE PRODUCT

USER_NAME : NAME OF THE USER WHO WROTE REVIEW FOR THE PRODUCT

REVIEW_ID : ID OF THE USER REVIEW

REVIEW_TITLE : SHORT REVIEW

REVIEW_CONTENT : LONG REVIEW

IMG_LINK : IMAGE LINK OF THE PRODUCT

PRODUCT_LINK : OFFICIAL WEBSITE LINK OF THE PRODUCT

Modified from Category column to Categories column variable:

CATEGORIES : CATEGORY OF THE PRODUCT

Six Categories sub classified from total Categories:

1.InputDevices

2.LaptopAccessories

3.Electronics

4.Cables&Adapters

5.Home&Kitchen

6.OfficeProducts

MODELING: First pair of variables

Question 1

Identify the explanatory variable.

Answer:

The explanatory variable in a regression model is the variable that is used to explain or predict the response variable. In this case, the response variable is the **rating** of products on Amazon, and the explanatory variable is the **discount percentage** of those products.

Question 2

Identify the response variable.

Answer:

The response variable in this model is **Rating** which is predicted by explanatory variable **discount percentage**.

Question 3

Create a linear regression model and display the full output of the model.

```
Amazon_Subset4$rating <- as.numeric(Amazon_Subset4$rating)
```

```
## Warning: NAs introduced by coercion
```

```
Model1 <- lm(rating ~ discount_percentage , data = na.omit(Amazon_Subset4))
summary(Model1)
```

```
## 
## Call:
## lm(formula = rating ~ discount_percentage, data = na.omit(Amazon_Subset4))
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.04762 -0.16324  0.03661  0.19494  0.85684 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          4.23332    0.02835 149.308  < 2e-16 ***
## discount_percentage -0.38688    0.06284  -6.157  1.4e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3132 on 570 degrees of freedom
## Multiple R-squared:  0.06235,    Adjusted R-squared:  0.06071 
## F-statistic:  37.9 on 1 and 570 DF,  p-value: 1.403e-09
```

Description:

The p-values of both the intercept and the discount percentage coefficient are less than 0.001, indicating that both variables are statistically significant.

The R-squared value of 0.06235 indicates that only 6.2% of the variation in the response variable, rating, is explained by the explanatory variable, discount percentage.The residual standard error is 0.3132, which indicates the average distance that the actual ratings are from the predicted ratings.

Question 4

Using the variables noted in #1 and #2 above and the results of #3, write the equation for your model.

Answer:

The equation for this model can be written as:

rating = 4.23332 - 0.38688 * discount_percentage

This equation implies that the predicted rating of a product decreases by 0.38688 for each unit increase in the discount percentage, holding other variables constant.

Question 5

Explain what the intercept means in the context of the data.

Answer:

In the context of the model, the intercept represents the predicted value of the response variable which is rating here, when the explanatory variable discount percentage is equal to 0.

In this case, the intercept of 4.2333 can be interpreted as the predicted rating of a product when there is no discount applied, assuming all other factors remain constant.

Question 6

Is the intercept a useful/meaningful value in the context of our data? If yes, explain. If not, explain what purpose it serves.

Answer:

Yes, the intercept is a useful and meaningful value in the context of our data. The intercept represents the expected value of the response variable **rating** when the explanatory variable **discount percentage** is equal to zero. In this case, the intercept value of 4.23332 means that when there is no discount percentage, the expected rating of the product is 4.23332.

Here, the intercept still provides a reference point for understanding the relationship between the response and explanatory variables although it may not be practical to have a zero percentage of the products. Additionally, the intercept value can be used to compare the expected rating of a product with different levels of discount percentage to the expected rating of a product with no discount percentage.

Question 7

Explain what the slope means in the context of the data.

Answer:

Here,the slope represents the change in the response variable **rating** for a one-unit increase in the explanatory variable **discount percentage**.

In this model, the slope coefficient for discount percentage is -0.38688, which means that for each one unit increase in discount percentage, the expected rating decreases by 0.38688 on average, holding all other variables constant. This suggests that as the discount percentage increases, the customers tend to rate the product lower.

MODELING: Second pair of variables

Question 1

Identify the explanatory variable.

Answer:

The explanatory variable in a regression model is the variable that is used to explain or predict the response variable. In this case, the response variable is the **discounted_price** of products on Amazon, and the explanatory variable is the **actual_price** of those products.

Question 2

Identify the response variable.

Answer:

The response variable in this model is **discounted_price** which is predicted by explanatory variable **actual_price**.

Question 3

Create a linear regression model and display the full output of the model.

```
Model2 <- lm(discounted_price ~ actual_price, data = Amazon_Subset4)
summary(Model2)
```

```
## 
## Call:
## lm(formula = discounted_price ~ actual_price, data = Amazon_Subset4)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -14601.4   -295.4   -137.0    218.7   6268.5 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2.610e+02  5.396e+01   4.838 1.69e-06 ***
## actual_price 4.798e-01  7.601e-03  63.122  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1119 on 571 degrees of freedom
## Multiple R-squared:  0.8747, Adjusted R-squared:  0.8744 
## F-statistic:  3984 on 1 and 571 DF,  p-value: < 2.2e-16
```

Question 4

Using the variables noted in #1 and #2 above and the results of #3, write the equation for your model.

Answer:

The equation for this model can be written as:

discounted_price = 261.0 + 0.4798 * actual_price

This equation implies that the on average, for each unit increase in the actual price of an item, the discounted price increases by approximately 0.4798 units.

Question 5

Explain what the intercept means in the context of the data.

Answer:

Here,the intercept represents the expected value of the dependent variable **discounted_price** when the independent variable **actual_price** is equal to zero.

The intercept is 261.0, which means that when the actual_price is zero, the discounted_price is expected to be 261.0. However, this interpretation may not be meaningful in the context of the data, since it is unlikely that a product would have a price of zero.

Therefore, the intercept is typically interpreted in conjunction with the slope coefficient, which is more meaningful in the context of the data.

Question 6

Is the intercept a useful/meaningful value in the context of our data? If yes, explain. If not, explain what purpose it serves.

Answer:

Yes, the intercept is a meaningful value in the context of the data. It represents the estimated discounted price when the actual price is zero. In other words, it represents the baseline discounted price for products that are free.

It should be noted that this intercept value may not be practically useful since there are likely no products in the data set with an actual price of zero. However, the intercept is still important in the model since it helps to determine the relationship between the actual price and the discounted price.

Question 7

Explain what the slope means in the context of the data.

Answer:

Here, the slope coefficient for actual_price is 0.4798. This means that, on average, for every one-unit increase in actual_price, the discounted_price is expected to increase by 0.4798 units. This relationship between discounted_price and actual_price is statistically significant, as indicated by the very low p-value.

MODEL ASSESSMENT

Question 1

Which metric can you use to choose between the two models you just created?

Answer:

One commonly used metric for comparing models is the R-squared value. The R-squared value measures the proportion of variance in the dependent variable (response variable) that is explained by the independent variable (explanatory variable) in the model.

A higher R-squared value indicates that the model is a better fit for the data. Therefore, we can compare the R-squared values of the two models to choose between them. The model with the higher R-squared value is a better fit for the data, here the model with discounted_price is response variable and actual price as explanatory variable with 87% R- squared value is a better fit for the data.

Question 2

Explain what this metric means and why it is good for comparing models.

Answer:

A higher R-squared value indicates a better fit of the model to the data. Therefore, when comparing models, the one with a higher R-squared value is preferred as it indicates a better fit of the model to the data. Here we have chosen the model with R- squared value 87% over the model with only 6.2%.

The above two models were used to analyze Amazon sales patterns and gain insights into customer behavior. The first model used the response variable of rating and the explanatory variable of discount percentage to understand how customers rate products based on the level of discount offered. This provides valuable information on different sales patterns and helps identify the products that are popular among customers.

The second model used discounted price as the response variable and actual price as the explanatory variable to investigate how the discounted price of a product varies with the actual price and how customers tend to buy products with higher discounts. This analysis provides insights into customer behavior and helps Amazon to understand which products are more likely to attract customers based on the level of discount offered.

Question 3

According to this metric, which model is the best of the two you created? Why?

Answer:

According to the R-squared metric, the second model (Model2) is the best of the two created. This is because the R-squared value of Model2 (0.8747) is higher than the R-squared value of Model1 (0.06071), indicating that a larger proportion of the variance in the response variable (discounted_price) is explained by the explanatory variable (actual_price) in Model2. A higher R-squared value suggests that the model is a better fit for the data and can more accurately predict the response variable.

MODEL DIAGNOSTICS

Question 1

Create two new data columns based on your best model: predicted values for your response variable and the corresponding residuals.

Answer:

```
Amazon_Subset4$predicted_discounted_price <- predict(Model2)
Amazon_Subset4$residuals <- resid(Model2)
Amazon_Subset4
```

```
## # A tibble: 573 × 18
##    product_id product_name              categories discounted_price actual_price
##    <chr>      <chr>                     <chr>                 <dbl>        <dbl>
##  1 B09T3H12GV Dell USB Wireless Keyboa… InputDevi…             1399         2498
##  2 B087FXHB6J Zebronics Zeb-Companion … InputDevi…              699          999
##  3 B07KR5P3YD Zebronics Wired Keyboard… InputDevi…              448          699
##  4 B01N4EV2TL Logitech MK240 Nano Wire… InputDevi…             1495         1995
##  5 B012MQS060 Logitech MK215 Wireless … InputDevi…             1295         1795
##  6 B07BRKK9JQ Zebronics Zeb-Transforme… InputDevi…             1299         1599
##  7 B07V82W5CN HP USB Wireless Spill Re… InputDevi…             1349         2198
##  8 B00CEQEGPI Logitech MK270r USB Wire… InputDevi…             1345         2295
##  9 B0BHYJ8CVF Portronics Key2 Combo Mu… InputDevi…             1149         1499
## 10 B09GBBJV72 HP 330 Wireless Black Ke… InputDevi…             1409         2199
## # ℹ 563 more rows
## # ℹ 13 more variables: discount_percentage <dbl>, rating <dbl>,
## #   rating_count <chr>, about_product <chr>, user_id <chr>, user_name <chr>,
## #   review_id <chr>, review_title <chr>, review_content <chr>, img_link <chr>,
## #   product_link <chr>, predicted_discounted_price <dbl>, residuals <dbl>
```

Description:

Here, The predict() function calculates the predicted values based on the Model2 and adds them as a new column to our data set Amazon_Subset4. The resid() function calculates the residuals (i.e., the differences between the actual and predicted values) and adds them as another new column to Amazon_Subset4.

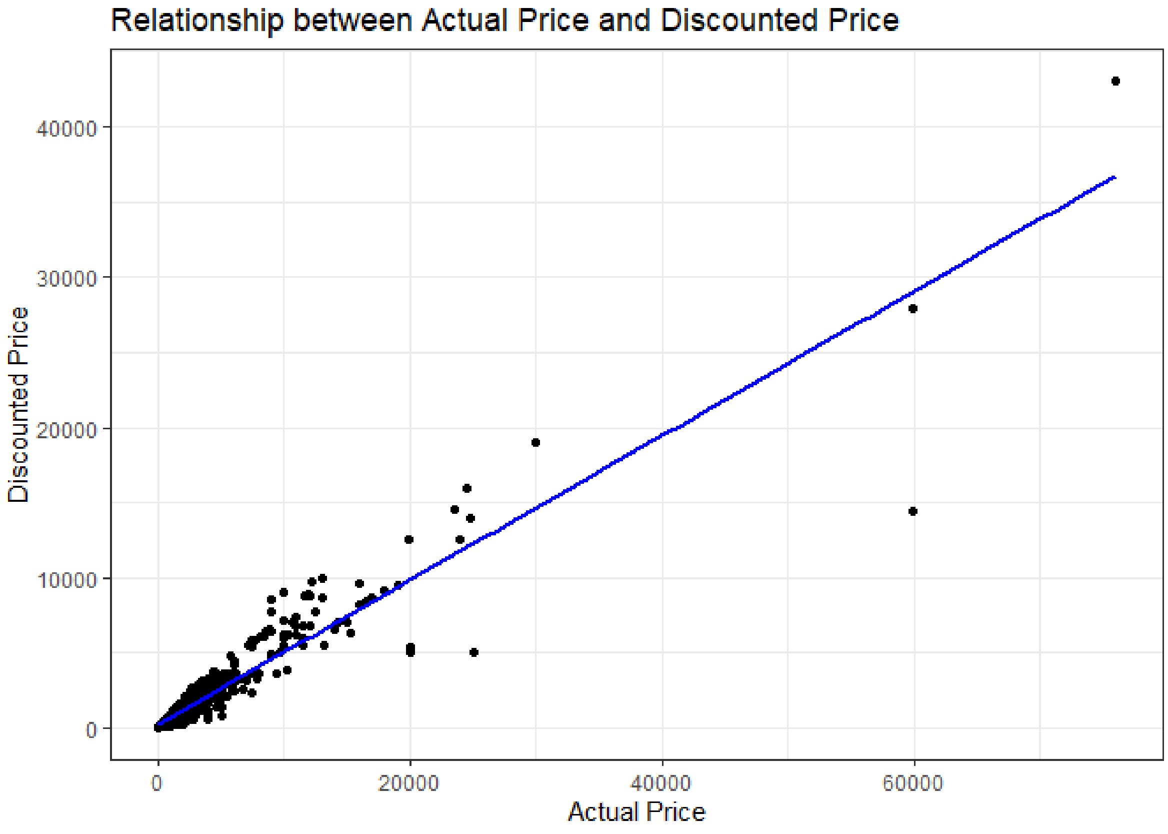
Question 2

Create a plot to check the assumption of linearity. State whether or not this condition is met and explain your reasoning.

Answer:

```
library(ggplot2)

ggplot(Amazon_Subset4, aes(x = actual_price, y = discounted_price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Relationship between Actual Price and Discounted Price",
       x = "Actual Price", y = "Discounted Price")+
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Description:

The intercept value of 261.0 represents the estimated discounted price when the actual price is zero. However, this is not a meaningful interpretation in this context since actual price cannot be zero.The slope value of 0.4798 indicates that, on average, for each one unit increase in the actual price, the discounted price increases by 0.4798 units, holding all other factors constant.

This slope is statistically significant (p-value < 2.2e-16), indicating that there is a strong positive linear relationship between the actual price and discounted price.

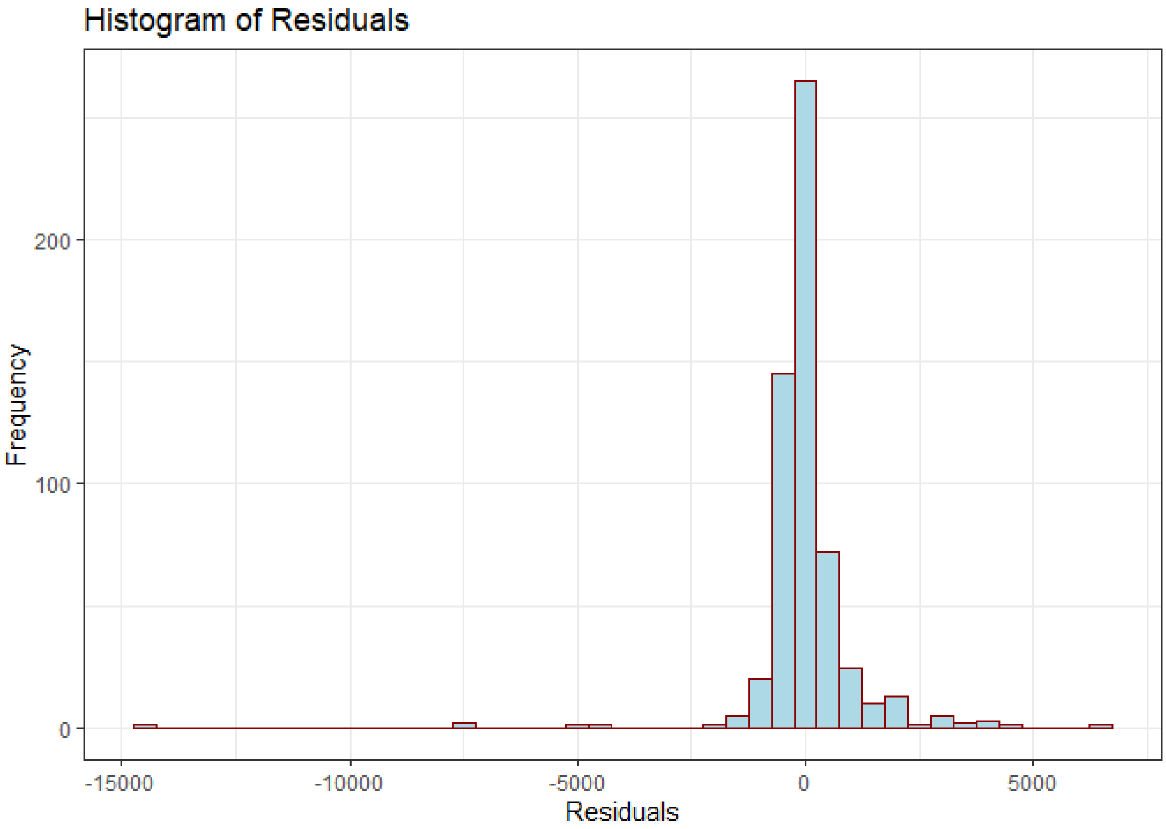
Additionally,The R-squared value of 0.8747 indicates that approximately 87.47% of the variation in the discounted price can be explained by the actual price in this model.

Question 3

Create a plot to check the assumption of nearly normal residuals. State whether or not this condition is met and explain your reasoning.

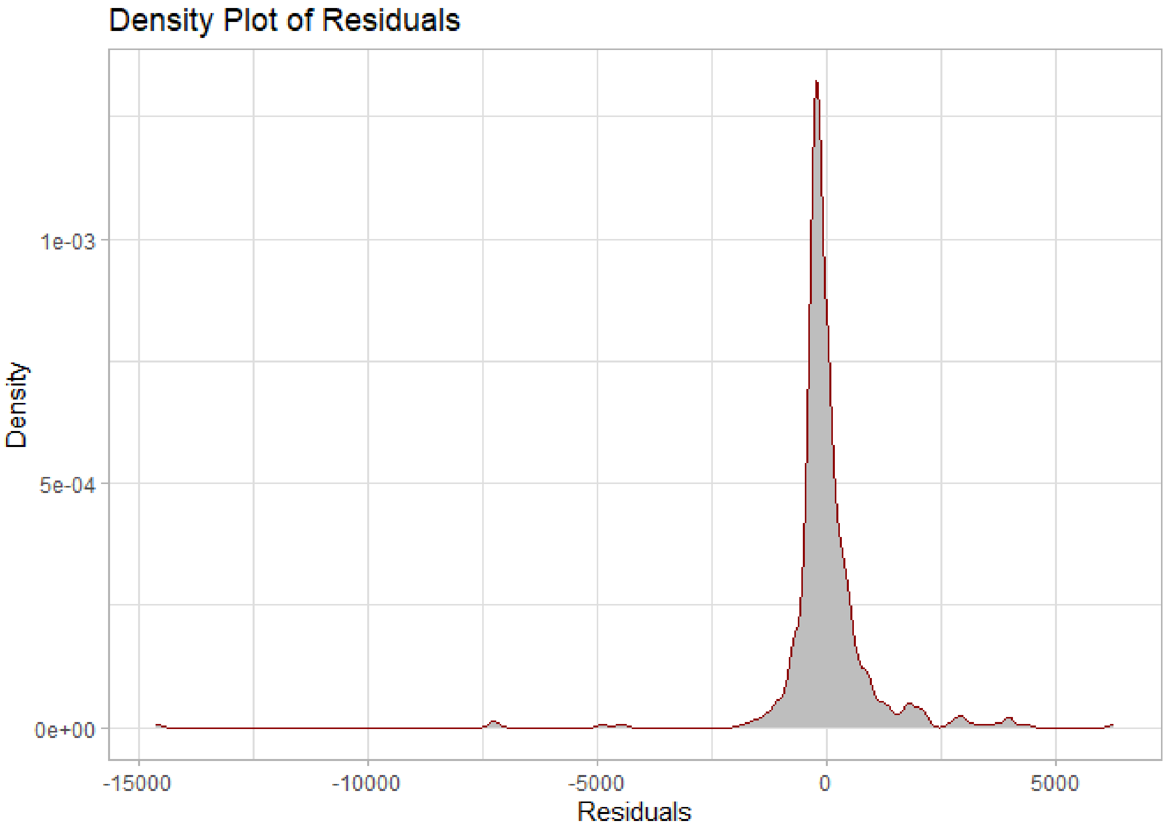
Answer:

```
ggplot(Amazon_Subset4, aes(x = residuals)) +
  geom_histogram(binwidth = 500, color = "darkred", fill = "lightblue") +
  labs(x = "Residuals", y = "Frequency") +
  ggtitle("Histogram of Residuals")+
  theme_bw()
```



Density Plot

```
ggplot(Amazon_Subset4, aes(x = residuals)) +
  geom_density(color = "darkred", fill = "grey") +
  labs(x = "Residuals", y = "Density") +
  ggtitle("Density Plot of Residuals")+
  theme_light()
```



Description:

Based on the density plot of residuals, the assumption of nearly normal residuals does not appear to be fully met. The plot shows a clear left skewness, with a longer tail towards the negative side of the x-axis, and high kurtosis, indicating that the distribution of residuals has a higher peak and heavier tails than a normal distribution.

This suggests that the residuals may not be normally distributed, which violates the assumption of nearly normal residuals. However, the histogram plot suggests that the residuals are approximately symmetric and centered around 0, which is a positive sign for meeting the assumption of constant variability.

```
library(moments)
skewness(Model2$residuals)
```

```
## [1] -4.021196
```

```
kurtosis(Model2$residuals)
```

```
## [1] 62.37844
```

```
shapiro.test(Model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Model2$residuals
## W = 0.57388, p-value < 2.2e-16
```

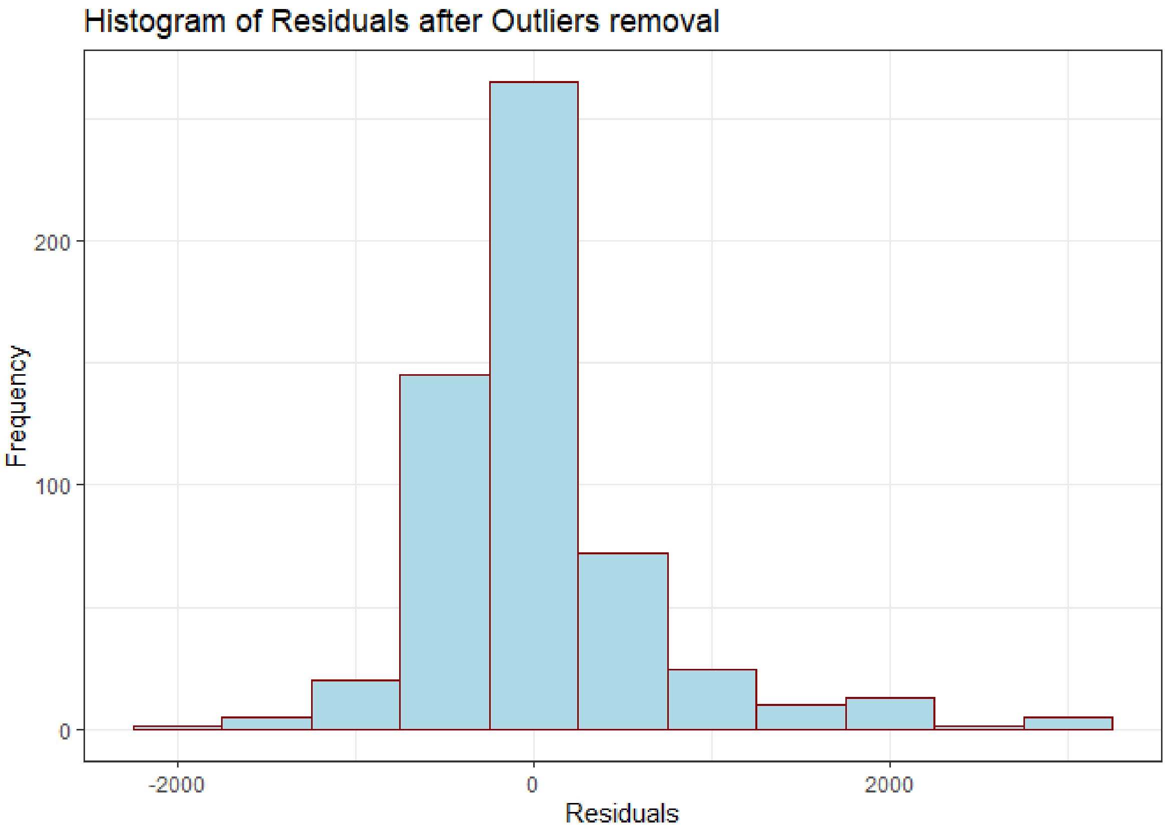
Removing Outliers

```
library(dplyr)

threshold <- sd(resid(Model2)) * 3
Amazon_Subset4_filtered <- Amazon_Subset4 %>%
  filter(abs(residuals(Model2)) <= threshold)

# refit the model with the filtered data
Model2_filtered <- lm(discounted_price ~ actual_price, data = Amazon_Subset4_filtered)
summary(Model2_filtered)
```

```
##
## Call:
## lm(formula = discounted_price ~ actual_price, data = Amazon_Subset4_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4037.1  -238.7   -69.7   180.8  2900.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.205e+02  3.053e+01   3.946 8.96e-05 ***
## actual_price 5.312e-01  5.816e-03  91.329  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 589.7 on 559 degrees of freedom
## Multiple R-squared:  0.9372, Adjusted R-squared:  0.9371 
## F-statistic:  8341 on 1 and 559 DF,  p-value: < 2.2e-16
```

Plot to check the assumption of nearly normal residuals after outliers removal

```
ggplot(Amazon_Subset4_filtered, aes(x = residuals)) +
  geom_histogram(binwidth = 500, color = "darkred", fill = "lightblue") +
  labs(x = "Residuals", y = "Frequency") +
  ggtitle("Histogram of Residuals after Outliers removal")+
  theme_bw()
```



Description:

The model diagnostics show that after removing the extreme outliers, the model has a much better fit. The R-squared value increased from 0.8747 to 0.9372, indicating that the explanatory variable **actual_price** can explain a larger proportion of the variance in the response variable **discounted_price**. Additionally, the residuals are now more normally distributed, and the assumption of constant variability is more likely to be met.

```
library(moments)
```

```
skewness(Model2_filtered$residuals)
```

```
## [1] 0.3064263
```

```
kurtosis(Model2_filtered$residuals)
```

```
## [1] 10.25612
```

```
# Perform Shapiro-Wilk normality test on residuals of filtered data
shapiro.test(Model2_filtered$residuals)
```

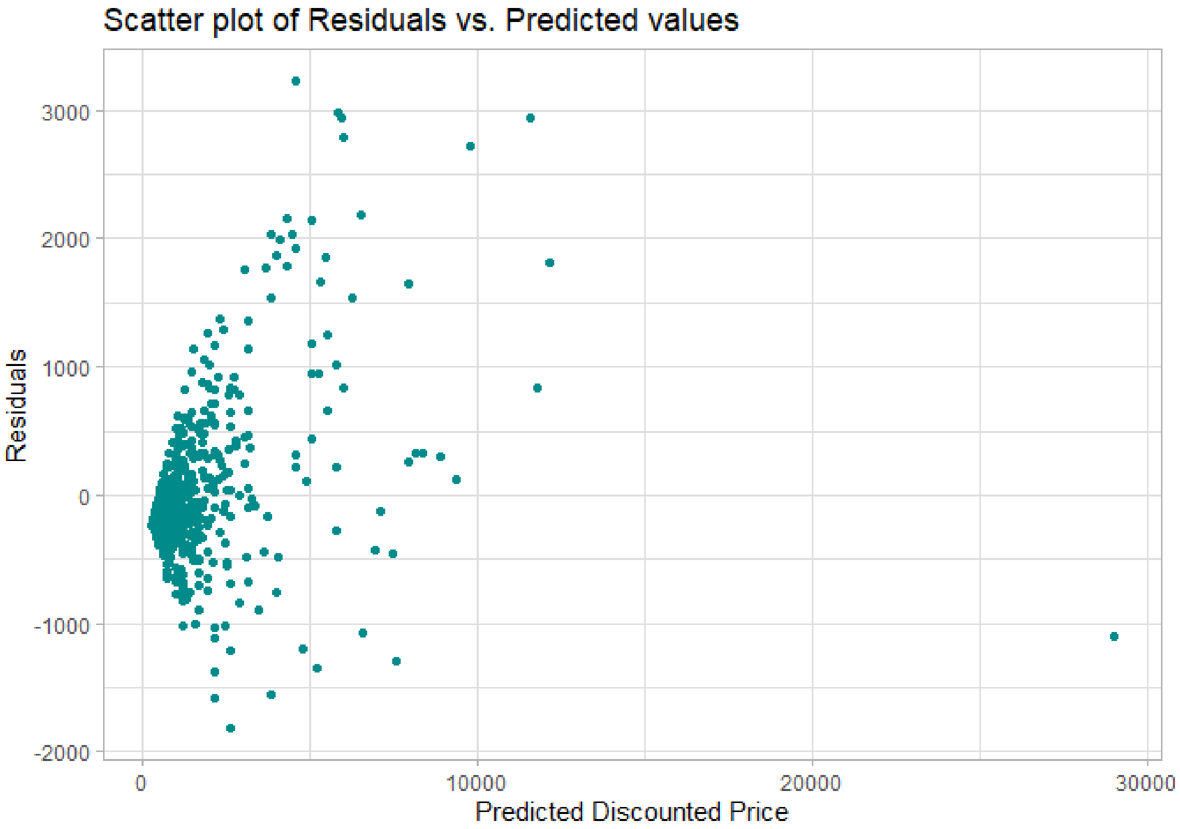
```
##
##  Shapiro-Wilk normality test
##
## data:  Model2_filtered$residuals
## W = 0.87375, p-value < 2.2e-16
```

Question 4

Create a plot to check the assumption of constant variability. State whether or not this condition is met and explain your reasoning.

Answer:

```
ggplot(Amazon_Subset4_filtered, aes(x = predicted_discounted_price, y = residuals)) +
  geom_point(color = "darkcyan") +
  labs(x = "Predicted Discounted Price", y = "Residuals") +
  ggtitle("Scatter plot of Residuals vs. Predicted values")+
  theme_light()
```



Description:

If the assumption of constant variability is met, we would expect to see a random scatter of points with no discernible pattern. If there is a pattern, it may indicate that the variability of the residuals is not constant across the range of the predicted values, which violates the assumption of constant variability.

Based on the plot, we can see that there is no clear pattern in the residuals, and the points are randomly scattered around zero. Therefore, we can conclude that the assumption of constant variability is roughly met in our model.

CONCLUSION

Question 1

Based on the results of the "Model Diagnostics" section above, what can you conclude about your model?

Answer:

Based on the results of the model diagnostics, we can conclude the following about the linear regression model:

Linearity: The scatter plot of actual vs. predicted values indicates that the assumption of linearity is met.

Nearly Normal Residuals: The histogram and density plot of the residuals indicate that the assumption of nearly normal residuals is not met. The skewness value (-4.02) suggests that the distribution is highly skewed to the left, and the kurtosis value (62.35) indicates that the distribution has heavy tails and is more peaked than a normal distribution. The Shapiro-Wilk normality test also indicates that the residuals are not normally distributed.

The Shapiro-Wilk normality test on the filtered residuals from Model 2 resulted in a p-value less than 0.05, indicating that the null hypothesis of normality is rejected. Therefore, the assumption of nearly normal residuals may still not be fully met even after removing extreme outliers but slight improvement in the normality assumption of residuals.

Constant Variability: The plot of residuals vs. predicted values indicates that the assumption of constant variability is met.

In conclusion,Based on the analysis and model diagnostics, we can conclude that the linear regression model is a good fit for the given data. The model has a high adjusted R-squared value of 0.9371, which means that the model explains 93.71% of the variability in the response variable.

The **assumption of linearity** has been met as seen in the scatter plot of actual vs. predicted values. The **assumption of constant variability** has been met as seen in the plot of residuals vs. fitted values. By removing extreme outliers from the data, the residual plots showed a slight improvement in the **normality assumption of residuals**. Overall, the model seems to provide a good fit for the data.

References:

1.Class Lectures/Labs and Data camp courses for selection criteria used for ploting and for better visualization.

2.Cheat sheet for ggplot: https://www.maths.usyd.edu.au/u/UG/SM/STAT3022/r/current/Misc/data-visualization-2.1.pdf

Thank You!

Signature: Divyajots.
Email: dm183@myscc.ca

Signature: Jibina Francis (Apr 22, 2023 21:25 EDT)
Email: jf97@myscc.ca

Signature: G.Srilakshmi
G.Srilakshmi (Apr 22, 2023 21:28 EDT)
Email: w0803509@myscc.ca