**STAT 6021: Project 1**

Exploring the Relationship Between Diamond Characteristics and Price

Prepared for Professor Jeffrey Woo, PhD

University of Virginia

Group Number: 15

Group Members:

Maryam Ali,

Robert Ashby,

Angelo Orciuoli,

Molly Shand

**What Determines a Diamond's Price?**

When it comes to diamonds, everyone knows that the bigger the rock, the bigger the price tag. But just how much does size matter? What about those other factors like sparkle, color, and clarity? Our team has been approached by the diamond experts at Blue Nile to crunch the numbers and uncover the secrets behind diamond pricing.

Our analysis of 1214 diamonds confirms that carat weight, the unit used to measure a diamond's weight, is a major player in determining its price. Think of it like this: as the carat weight goes up, the price jumps up even faster. So, while a half-carat diamond might be affordable, a two-carat diamond will set you back significantly more.

But don't get too hung up on size alone! Our findings also show that a diamond's clarity and color can make a real difference in its price. Diamonds with fewer imperfections (higher clarity) and those that are more colorless tend to be more expensive. This makes sense, right? After all, who wouldn't want a diamond that's as clear and bright as can be?

Interestingly, the "cut" of a diamond, which refers to how well its facets interact with light, didn't have as big an impact on price as we expected. This could be because cut grading is complex, or simply because beauty is in the eye of the beholder. The value of type of cut may be swayed more by fashion trends than the measurable quality of the cut versus others.

So, what's the takeaway for diamond shoppers? While carat weight is definitely important, don't overlook clarity and color. And remember, the most important factor is finding a diamond that you love, regardless of its size or price. After all, a diamond is a symbol of something special, and its true value lies in the meaning it holds for you.

This analysis supports Blue Nile's emphasis on the "4Cs" (carat, clarity, color, and cut) as important factors in diamond pricing. However, our findings suggest that carat weight is the dominant factor, while the impact of cut may be less pronounced than some believe. Ultimately, choosing the perfect diamond involves balancing your preferences and budget across all four Cs.

Our findings validate the claim that carat weight is the most influential factor in diamond pricing. While clarity, color, and cut contribute, they have a comparatively smaller impact than carat. The fitted model provides a strong predictive capability for price based on carat weight alone.

**Diamonds: A Data-Driven Perspective**

This report analyzes the relationship between diamond price and various attributes, including carat, clarity, color, and cut, using a dataset sourced from Blue Nile. Through data visualizations and a simple linear regression model, we confirm that carat has the most significant impact on price. Additionally, we explore claims from Blue Nile's diamond education page, such as the influence of cut, clarity, and color on price. Our findings suggest that while these attributes contribute to price variations, carat remains the dominant factor.

Data Description and Visualizations

Dataset Overview

The dataset consists of information for 1,214 diamonds with the following attributes:

Carat: Weight of the diamond.

Clarity: A categorical variable indicating the presence of inclusions.

Color: A categorical variable ranging from D (best) to J (least desirable).

Cut: A categorical variable measuring the diamond’s cut quality.

Price: The price of the diamond in USD.

With data available from csv file “diamonds4.csv” and indices 0-1214. The 0-index row contained headers derived from the above five attributes of the dataset: "carat", "clarity", "color", "cut", and "price".

For the data reading, processing, visualization, and analysis we utilized R code R 4.4.1 in the R Studio environment.

# Install necessary packages

# install.packages(c("tidyverse", "ggplot2", "dplyr"))

# Load libraries

library(tidyverse)

library(ggplot2)

# Load the dataset

diamonds <- read\_csv("diamonds4.csv")

# View first few rows

head(diamonds)

We checked to data read:

# Check the structure of the data

str(diamonds)

Output:

spc\_tbl\_ [1,214 × 5] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ carat : num [1:1214] 0.51 0.93 0.5 0.3 0.31 1 1.03 1.31 0.5 0.24 ...

$ clarity: chr [1:1214] "SI2" "IF" "VVS2" "VS1" ...

$ color : chr [1:1214] "I" "H" "D" "F" ...

$ cut : chr [1:1214] "Very Good" "Ideal" "Very Good" "Ideal" ...

$ price : num [1:1214] 774 6246 1146 538 502 ...

- attr(\*, "spec")=

.. cols(

.. carat = col\_double(),

.. clarity = col\_character(),

.. color = col\_character(),

.. cut = col\_character(),

.. price = col\_double()

.. )

- attr(\*, "problems")=<externalptr>

With data summary:

# Summary statistics

summary(diamonds)

Output:

carat clarity color

Min. :0.2300 Length:1214 Length:1214

1st Qu.:0.4000 Class :character Class :character

Median :0.5200 Mode :character Mode :character

Mean :0.8134

3rd Qu.:1.0000

Max. :7.0900

cut price

Length:1214 Min. : 322.0

Class: character 1st Qu.: 723.5

Mode :character Median : 1463.5

Mean : 7056.7

3rd Qu.: 4640.8

Max. :355403.0

Data Processing

With the data read in, we were able to verify and process the data.

First, we check for any missing data entries:

# Check for missing values

colSums(is.na(diamonds))

Output:

carat: 0 clarity: 0 color: 0 cut: 0 price: 0

With zeros as the measure for missing data entries in each coumn, we know we have complete data.

Next we adjusted the catergorical data into factors and double checked our data structure:

# Convert categorical variables to factors

diamonds$clarity <- as.factor(diamonds$clarity)

diamonds$color <- as.factor(diamonds$color)

diamonds$cut <- as.factor(diamonds$cut)

# Check the structure again.

str(diamonds)

Output:

spc\_tbl\_ [1,214 × 5] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ carat : num [1:1214] 0.51 0.93 0.5 0.3 0.31 1 1.03 1.31 0.5 0.24 ...

$ clarity: Factor w/ 8 levels "FL","IF","SI1",..: 4 2 8 5 3 5 5 8 3 3 ...

$ color : Factor w/ 7 levels "D","E","F","G",..: 6 5 1 3 3 3 4 4 1 4 ...

$ cut : Factor w/ 4 levels "Astor Ideal",..: 4 3 4 3 3 3 3 3 3 3 ...

$ price : num [1:1214] 774 6246 1146 538 502 ...

- attr(\*, "spec")=

.. cols(

.. carat = col\_double(),

.. clarity = col\_character(),

.. color = col\_character(),

.. cut = col\_character(),

.. price = col\_double()

.. )

- attr(\*, "problems")=<externalptr>

Data Visualizations

The price distribution is right-skewed, indicating that most diamonds are in the lower price range.

Carat values exhibit a right-skewed distribution, with most diamonds weighing under 1.5 carats.

Bivariate and Multivariate Analysis

Price vs. Carat: A strong positive correlation is observed, confirming that heavier diamonds are more expensive.

Price by Cut: The median price increases for higher quality cuts, but the variation is significant.

Price by Clarity: Higher clarity grades (e.g., IF, VVS1) tend to have higher prices.

Price by Color: Diamonds with better color grades (e.g., D, E) are more expensive on average.

These visualizations support the Blue Nile claims that carat is the primary driver of price, but cut, clarity, and color also influence pricing.

Simple Linear Regression Analysis

Model Specification

A simple linear regression model was fitted:

Intercept (): The baseline price for a zero-carat diamond.

Slope (): Indicates the expected increase in price per additional carat.

R-squared: Indicates that carat explains a large proportion of price variation.

SLR Assumption Checks

Linearity: The scatter plot and regression line suggest a linear relationship.

Normality of Residuals: A histogram of residuals shows approximate normality.

Homoscedasticity: Residuals appear randomly distributed around zero, indicating constant variance.

**References**

4Cs of Diamonds & Diamond Buying Guide. Blue Nile. (n.d.). https://www.bluenile.com/education/diamonds