Data100 Project: Diamonds Analysis

Lloyd Nsambu, Sidon Mengsteab, Kim Kiju

2025-04-04

```
# Initializers
suppressPackageStartupMessages({
library(tidymodels)
library(tidyverse)
library(dplyr)
library(readr)
library(ggplot2)
library(readxl)
library(tidyr)
library(scales)
library(ggplot2)
library(patchwork)
theme_set(theme_bw())
})
```

Introduction

We are THETA Capital, a strategic consulting and data analytics firm based in Waterloo, Ontario, Canada. Founded in 2012, we specialize in delivering high-impact insights, market intelligence, and data-driven decision-making frameworks for clients across North America, Europe, and Eastern Asia. Our expertise spans corporate strategy, financial analysis, and operational optimization, serving a diverse portfolio of Fortune 500 companies, government agencies, non-profits, and high-growth enterprises.

Context of this Analysis

THETA Capital has been engaged by a leading South African diamond jewelry firm, Le Roux International, to analyze their global sales performance, diamond quality metrics, and market trends. The dataset provided includes sales data, Diamond grades, diamond rating, years of sales, country of sales as well as amount of sales.

The Key Objectives of Analysis Target Variable -> Price - Market demand patterns: Identify top-performing regions and customer preferences - Pricing strategy: Correlate diamond grades/cuts with sales margins to optimize pricing tiers. - Sales trends: Detect seasonal spikes, anomalies, or shifts in buyer behavior.

```
diamonds_raw <- read_csv('diamond_sales.csv')
```

Data Cleaning and Tidying

Fistly, we should get a glimpse of the dataset to see what we are dealing with.

Looking at the raw dataset, there are certain columns that can be arranged better for further analysis. For example, the x,y,z columns which are the length, width, and height of each diamond should be named appropriately. Further more, the recorded years of sales and the sales in those years should be ordered properly to further better the analysis down the analysis pipeline.

```
## Lets rename the x, y, z columns to give more description to the data within them.
diamonds_tidy1 <- diamonds_raw |>
  rename(length_mm = x, width_mm = y, height_mm = z)
```

Further data cleaning.

From speaking with the jeweler firm, Le Roux International, the recording method of their sales through out the different nations is decentralized. Meaning every branch records and stores their business data in different formats. They were able to consolidate the sales for the past eight year from the worldwide branches to make this analyses.

We can see the years of sales in the different countries for the past 8 years as columns and the number of sales is under each year. For further analysis, the number of diamonds sold sales in each year should be a separate column.

Advanced Cleaning:

```
# We shall better consolidate the sales data for the previous eight years into longer format to
# facilitate better sales modelling and projections further down in the analysis.
# This will be done by pivoting the 2021- 2012 columns into longer format.
diamonds_tidy2 <- diamonds_tidy1 |>
  pivot_longer(
  cols = c("2020", "2021", "2022", "2019", "2018","2017", "2016", "2015"),
  names_to = "Year_Sold",
  values_to = "Num_sales"
  )
## We want to learn a bit more about the diamonds that the correlation they have between the cut and th

understanding_diamonds_1 <- diamonds_tidy1|> reframe(clarity, color,cut )|> distinct(clarity, cut, colo

# There is certainly a correlation. Through outside search, we realize that there needs to be better de

# The diamonds need to have better description for further analysis.
# We shall give the diamonds in the color more descriptive names and rename and convert the
# color column to color_category from an character to a factor variable and the clarity description wil
# This will be achieved with the mutate function and using the case_when() method.
diamonds_tidy2 <- diamonds_tidy2|>
  mutate(
  color = case_when(color %in% c("D","E","F") ~ "Colorless",
color %in% c("G","H","I","J") ~ "Near-colorless", color %in% c("K", "L","M") ~ "Faintly coloured",
color %in% c("N", "O", "P","Q", "R", "S", "T", "U", "W", "X", "Y", "Z") ~ "Lightly colored", T ~ "Other
  clarity_description = case_when(
   clarity == "FL" ~ "Flawless", clarity == "IF" ~ "Internally Flawless", clarity %in% c("VVS1", "VVS2"
   clarity %in% c("VS1", "VS2") ~ "Very Slightly Included", clarity %in% c("SI1", "SI2") ~ "Slightly In
   clarity %in% c("I1", "I2", "I3") ~ "Included", TRUE ~ "Unknown"
```

```
 ),
 color = as_factor(color),
 clarity = as_factor(clarity)
)|> rename(color_category = color)
```

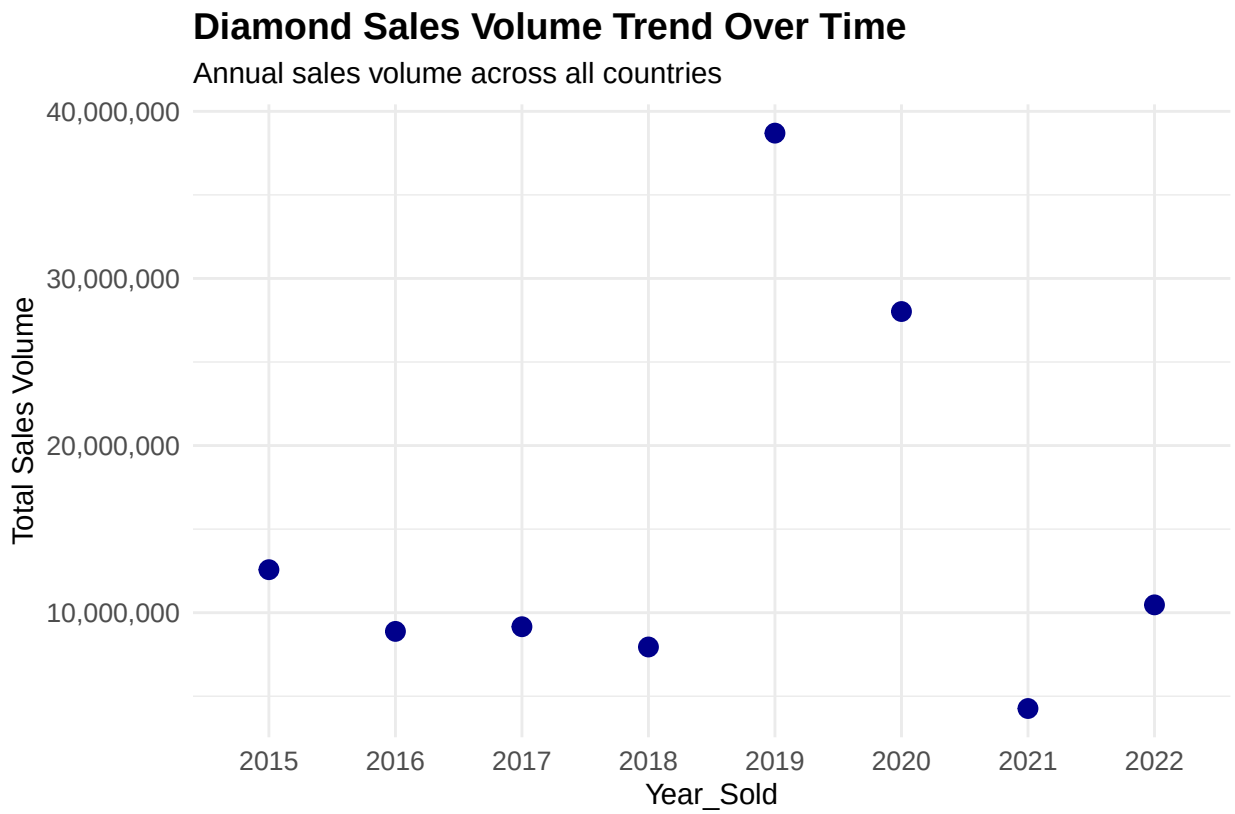
Creating Regular expressions

Exploratory Data Analysis

After carefully cleaning the data. We are aiming to explore and understand a bit more about the sales situation of the jewellery firm. More specifically,the trend of annualized sales volume across all the countries the firm operates in, the countries with the highest sales volumes and their annualized trends.

```
# Firstly, want to get an understanding of the annual sales across
# the different countries to get a broader picture. Of what's going on.
yearly_sales <- diamonds_tidy2 |>
  summarise(.by = "Year_Sold", Total_Sales = sum(Num_sales, na.rm = TRUE))

yearly_sales |>
  ggplot(mapping =  aes(x = Year_Sold, y = Total_Sales)) +
  geom_point(color = "darkblue", size = 3) +
  labs(title = "Diamond Sales Volume Trend Over Time",
       subtitle = "Annual sales volume across all countries", nx = "Year",
       y = "Total Sales Volume", caption = "Source: Diamond Sales Dataset")+
  theme_minimal() +scale_y_continuous(labels = scales::comma) + # makes the numbers have commas
  theme(plot.title = element_text(face = "bold", size = 14), axis.text = element_text(size = 10))
```



This is very interesting! In the years 2015 to 2022, The sales volume across all the different countries has been fairly consistent,maintaining an average of 9,000,0000 sales per year. However there is a significant hike in the years 2019 to 2020 and seemingly int he following year, 2021.

Let's determine which countries the jewellery firm operates in as well as the country that accounts for the most sales amount the across the eight year period. There are many other countries that contribute to the firm's sales. Let's focus the sales analysis on the top 6 countries and bottom 6 countries to really understand what's diamond product has been driving majority revenue

```
# Let's determine the specific nations that the firm sells in:
countries <- diamonds_tidy2 |> distinct(country_sold)
# Creating a vector specifc to the top 6 countries, by slicing the top 6 countries.
top_countries <- diamonds_tidy2 |> group_by(country_sold) |> summarise(total_sales = sum(Num_sales, na.
  slice_max(total_sales, n = 6)
```

Result: The country with the highest sales volume is: China, followed by the United States and India. The jewellery firms biggest market is Asia. The countries consist of China, India, Japan, and South Korea Lets go ahead and graph their annual diamond sales trend to get a better visual understanding.

Explatory Plot 1:

```
# Lets create a data frame that only includes the top 6 countries with the highest sales volume.
# The new data frame will be called country_sales.

top_6_countries <- head(top_countries$country_sold, 6)
```

```r
# filtering only for the top 6 countries and grouping them together with
# total sales, the average price and the most common diamond clarity
top_country_sales <- diamonds_tidy2 |> filter(country_sold %in% top_6_countries)|>
group_by(Year_Sold, country_sold) |>
summarise(num_sales = sum(Num_sales, na.rm = TRUE), avg_price = mean(price, na.rm = TRUE),
most_common_clarity =  names(which.max(table(clarity))),clarity_description = first(clarity_description)
color_distribution = paste(sort(unique(color_category)), collapse = ", "),
.groups = "drop") |> arrange(Year_Sold, desc(num_sales))

top_country_sales_plot <- top_country_sales|>
ggplot( mapping = aes(x = Year_Sold, y = num_sales, fill = country_sold)) +
geom_col(position = position_dodge2(preserve = "single"), width = 0.8) +
labs(title = "Top 6 Countries by Diamond Sales Volume", subtitle =
"Annual sales performance of highest-volume markets",
x = "Year", y = "Number of Diamonds Sold", caption = "Source: Diamond Sales Dataset") +
theme_minimal() +
theme(legend.position = "right",panel.grid.major.x = element_blank(),
plot.title = element_text(face = "bold", size = 14),
plot.subtitle = element_text(color = "gray40", size = 10), axis.text.x = element_text(angle = 45, hjust
  )+scale_y_continuous(labels = scales::comma)
```

Visualizing the sales volume this way, we can really comprehend the significant markets to the jewellery firm's sales! There is no doubt that China, India and the United States are major contributors to this. It is also important to mention that 2019 saw the highest sales record. Applying outside factors to this anaylsis, this was pre-covid. In 2020, the year of covid, sales were marginally lower than the year before but still significantly high in contrast to the 6 other years. This could be due to the fact that during covid, consumers had a lot more discretionary income due to reduced spending on other products and increased government stimulus packages.

Now, let's see how the bottomw 6 compared. We would also like to see how sales are distributed among the bottom 6 countries.

Explatory Plot 2:

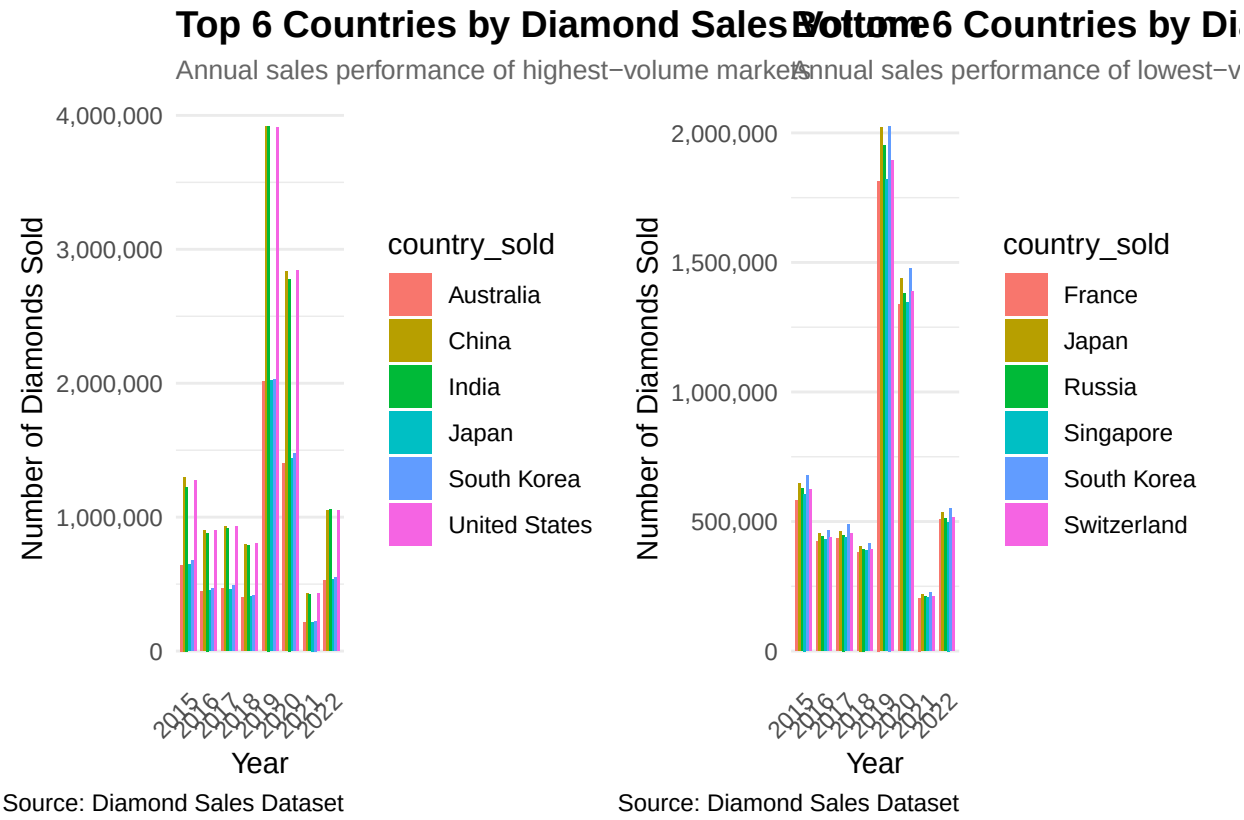
```r
# Creating a dataframe that identifies the bottom 6 diamond sellers
# The code to determine the bottom six countries, sales wise, is similair
# to determing the top 6  with some minor changes.
# using tail() instead of head() to select last 6
bottom_6_countries <- tail(countries$country_sold, 6)

bottom_country_sales <- diamonds_tidy2 |>
filter(country_sold %in% bottom_6_countries) |> group_by(Year_Sold, country_sold)|>
summarise(num_sales = sum(Num_sales, na.rm = TRUE), avg_price = mean(price, na.rm = TRUE),
most_common_clarity = names(which.max(table(clarity))),clarity_description = first(clarity_description)
color_distribution = paste(sort(unique(color_category)), collapse = ", "),
groups = "drop") |> arrange(Year_Sold, desc(num_sales))

bottom_country_sales_plot <- bottom_country_sales|> ggplot(
  mapping = aes(x = Year_Sold, y = num_sales, fill = country_sold)) +
  geom_col(position = position_dodge2(preserve = "single"), width = 0.8) +
  labs( title = "Bottom 6 Countries by Diamond Sales Volume", subtitle = "Annual sales performance of l
    x = "Year", y = "Number of Diamonds Sold", caption = "Source: Diamond Sales Dataset") +
```

```
  theme_minimal() + theme(legend.position = "right", panel.grid.major.x = element_blank(),
plot.title = element_text(face = "bold", size = 14),plot.subtitle = element_text(color = "gray40", size
axis.text.x = element_text(angle = 45, hjust = 1))+scale_y_continuous(labels = scales::comma)
```

```
top_country_sales_plot + bottom_country_sales_plot
```



Source: Diamond Sales Dataset

Source: Diamond Sales Dataset

Modeling Diamond Quality and Price

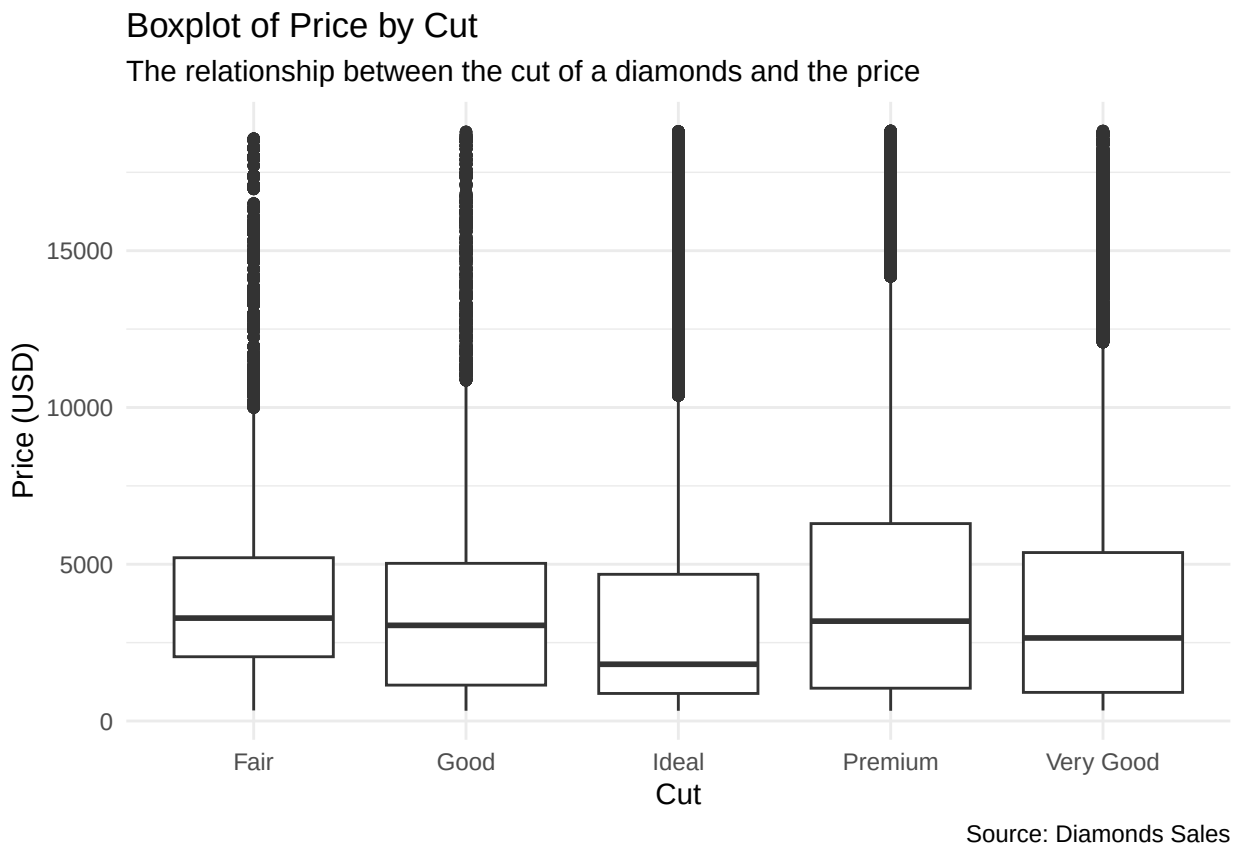
Utilizing the advanced cleaning conducted earlier

The next step of the analysis is to better understand diamonds and the quality which will better help us better understand what the different jewellery markets that firm operates in looks for when buying diamonds.

Firstly, lets model a simple relationship between the cut and price of a diamond.

Model Plot 1: Average and Price Range for different Diamond Cuts

```
# Create a boxplot to visualize the relationship between price and cut
(cut_vs_price_plot <- diamonds_tidy2|>
  filter(!cut == "NA"      # filtering out NA values.
)|> ggplot( mapping = aes(x = cut, y = price)) + geom_boxplot() +
  labs(title = "Boxplot of Price by Cut", subtitle = "The relationship between the cut of a diamonds an
x = "Cut", y = "Price (USD)", caption = "Source: Diamonds Sales") + theme_minimal())
```



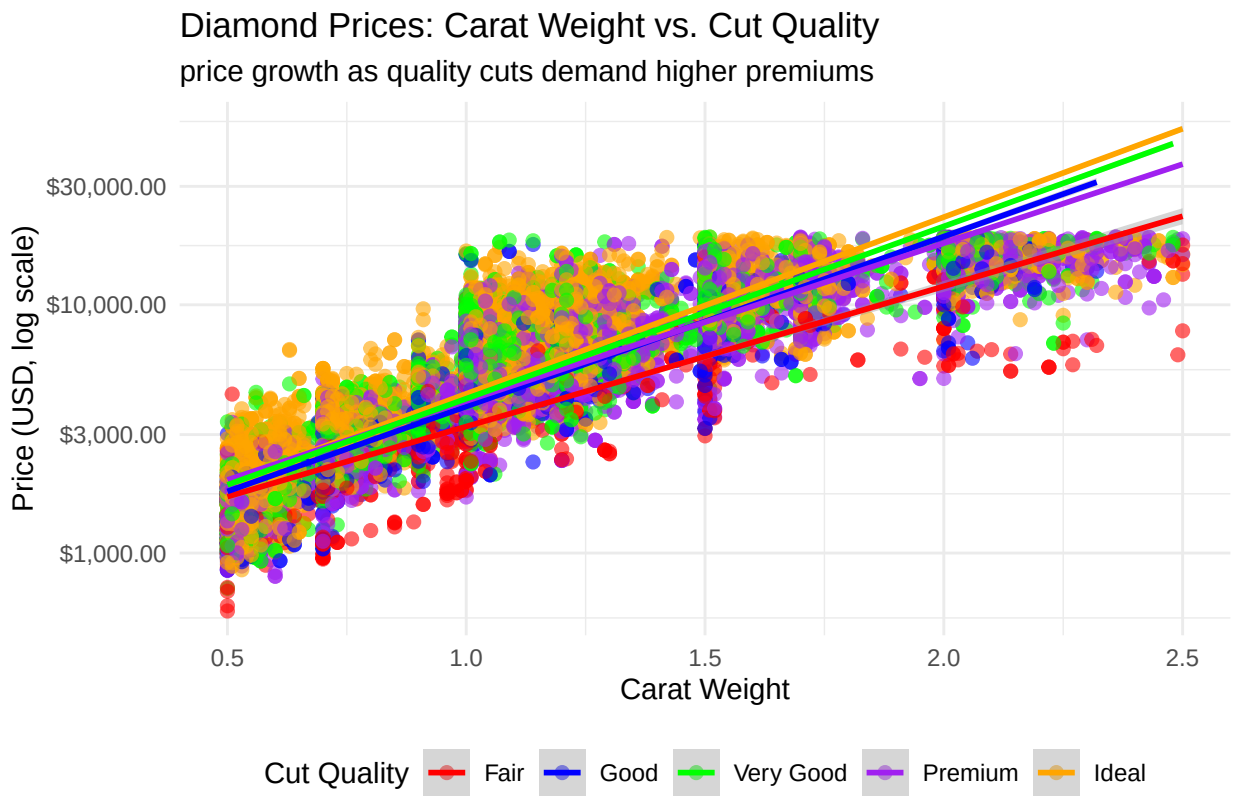
Observations: Model Plot 1

Overall, there are 5 cut categories. The line in each box is the median price that each diamond sells for. It's interesting to see that fairly cut diamonds ask for a price on par with premium diamonds and the maximum price of these diamonds maintain a maximum price of $5,000. Another observation is the outliers of each cut. All cuts seem to have had an outlier that has sold for over $15,000

Model Plot 2: Carat Weight and Price

Which Diamonds really fetch the highest price. Let's see what carat weights fetch currently fetch the highest price

```
(carat_weight_plot <- diamonds_tidy2|>
filter(between(carat, 0.5, 2.5)) |> slice_sample(prop = 0.1)|>
ggplot(aes(x = carat, y = price, color = cut)) + geom_point(alpha = 0.6, size = 2) +
geom_smooth(method = "lm", ) +
scale_y_continuous( labels = scales::dollar, trans = "log10" ) +
scale_color_manual( values = c("Fair" = "red", "Good" = "blue", "Very Good" = "green",
"Premium" = "purple", "Ideal" = "orange"), breaks = c("Fair", "Good", "Very Good", "Premium", "Ideal")
labs( x = "Carat Weight", y = "Price (USD, log scale)", title = "Diamond Prices: Carat Weight vs. Cut Qu
subtitle = "price growth as quality cuts demand higher premiums",
caption = "Source: Diamonds dataset", color = "Cut Quality" ) + theme_minimal() + theme(legend.position
```

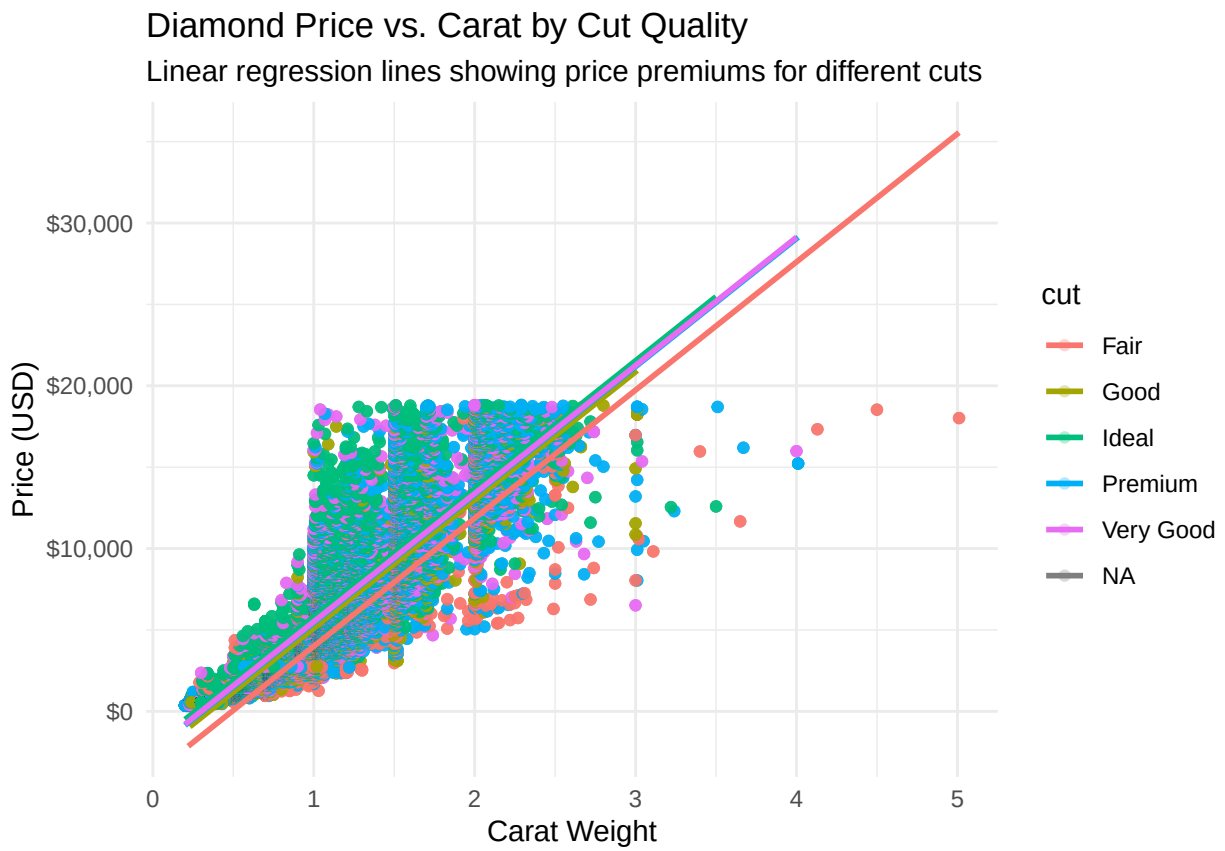


From the scatter plot above, we can see the price increasing gradually which means that carat has the highest effect. Ideal cuts command the highest premium, especially for diamonds with cuts >1 carat. Cut efficiency adds predictive power beyond basic cut grades.

Explatory Linear Model 1: Predicting the price premuim for higher cuts

Adding onto the earlier graph, we can aim to predict the premium price for each added level of cut quality, to maximize revenue on each sale and focus the jewellery firms acquisitions of diamonds on not all, but specific diamonds.

```
# Initializing the linear regression engine
cut_price_model <- linear_reg() |>
  set_engine("lm") |> ## The target variable is price, this is what we want to predict
  fit(price ~ carat + cut, data = diamonds_tidy2)   # the associated explanatory variables are carat an
fit_cut_price <- augment(cut_price_model, new_data = diamonds_tidy2)
(fit_cut_price_plot <- fit_cut_price |>
ggplot(
mapping=  aes(x = carat, y = price, color = cut)) + geom_point(alpha = 0.3) + geom_line(aes(y = .pred),
subtitle = "Linear regression lines showing price premiums for different cuts", x = "Carat Weight", y =
theme_minimal())
```



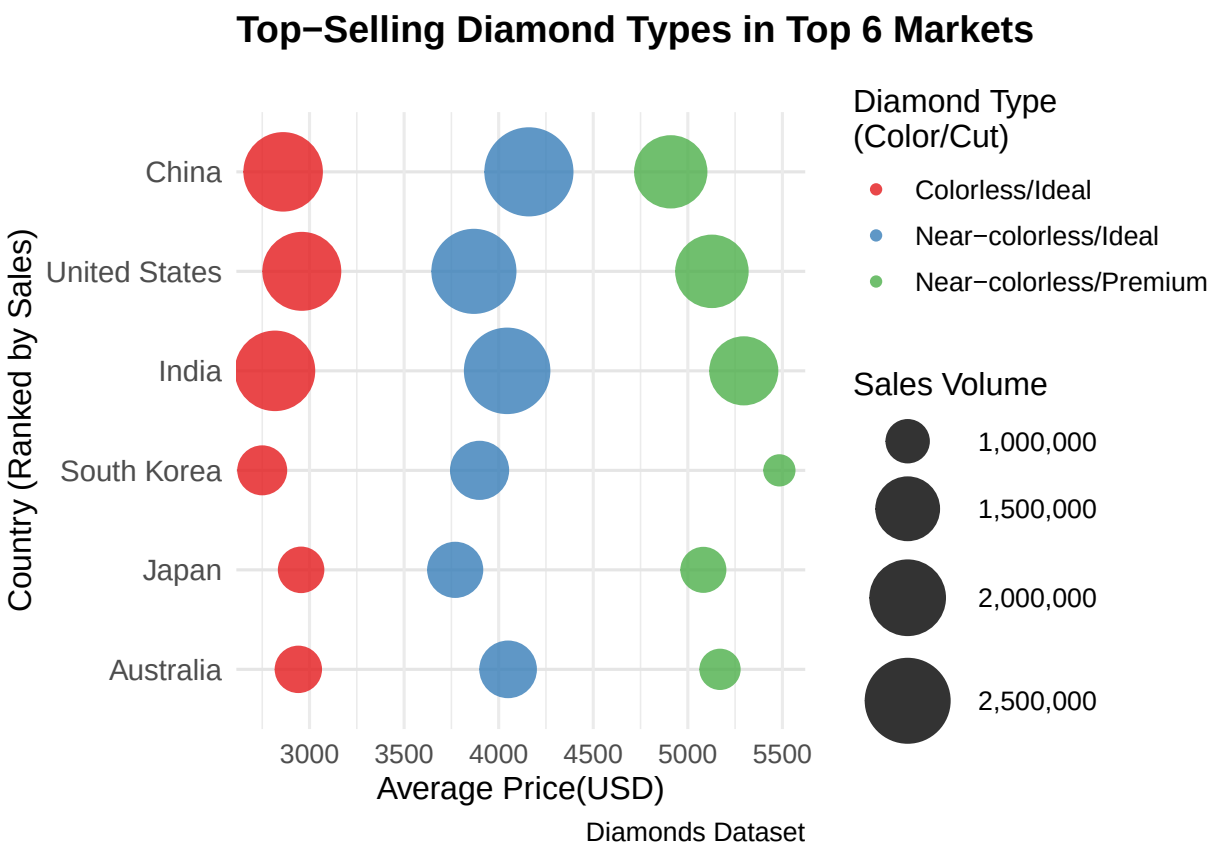
Looking at this linear regression model, it serves as a practical starting point for pricing analysis as well as making a suitable and diamond procurement method that will lead higher inventory turn over and profits. - Looking at forward diamond making, the fair cut diamonds with carat weights above 4 demand significantly higher prices than than very good and premium cut diamonds. The suggestion for the jewellery firm then would be to focus their sourcing for diamonds on those this high weights that they can design to a fair cut quality and collect larger profits.

Plot 3: Diamonds Sales Cut and Color by Country

```
top_diamonds_country <- diamonds_tidy2 |>
filter(country_sold %in% top_6_countries,   # Calling the vector we initialized earlier
!is.na(color_category), !is.na(cut))|> group_by(country_sold, color_category, cut) |>  ## Removing NAs
summarise(total_sales = sum(Num_sales, na.rm = T), avg_price = mean(price, na.rm = T),
.groups = "drop") |> group_by(country_sold) |>  # Only getting the top three per country
  # We will also create a new variable diamond_type to merge the expressions from color and cut into on
slice_max(total_sales, n = 3, with_ties = F) |> ungroup() |> mutate(diamond_type = paste(color_category

# creating a bubble chart to visualize the relationship
(top_selling_chart <- top_diamonds_country |>
  ggplot( mapping = aes(x = avg_price, y = reorder(country_sold, total_sales, sum),
           size = total_sales, color = diamond_type)) +
  geom_point(alpha = 0.8) +
  scale_size_continuous(range = c(5, 15),name = "Sales Volume", breaks = pretty_breaks(4), labels = comm
  )+scale_color_brewer(palette = "Set1", name = "Diamond Type\n(Color/Cut)"  # Updated label
  )+labs(title = "Top-Selling Diamond Types in Top 6 Markets", subtitle = "" , y = "Country (Ranked by S
```

```
  x = "Average Price(USD)",caption = "Diamonds Dataset")+
theme_minimal(base_size = 12) +
theme(legend.position = "right", panel.grid.major.y = element_line(color = "grey90"),
  plot.title = element_text(face = "bold", size = 14), plot.subtitle = element_text(color = "grey80")
  axis.text.y = element_text(size = 11)))
```



Observations: Plot 3 The right and left-skewed distribution of bubbles in between the top 6 countries clearly reflects differing wealth levels and cultural valuation of diamond characteristics. For example, in Korea, we can see that there is only a small group of people that can purchase premuim diamonds upwards of USD5,000 where as in Japan, there is an equal amount of sales activity(volume) among the varying diamond types. The United States, China and India on the other hand, being large markets, have very large sales volumes across the different diamond types.

Combining the explorator analysis of the top selling markets earlier, and the knowledge we've gained from modeling the cut and quality to the price, we can suggest strategies to the jewellery firm. From the looks all these markets are in demand for colorless diamonds, the cut of the diamonds is to some extent negligible. The highest selling colorless diamond is of the ideal cut quality which we have recognized in Model Plot 1 sells at the lowest median price out of all the diamonds.

Suggestion to the jewellery firm:

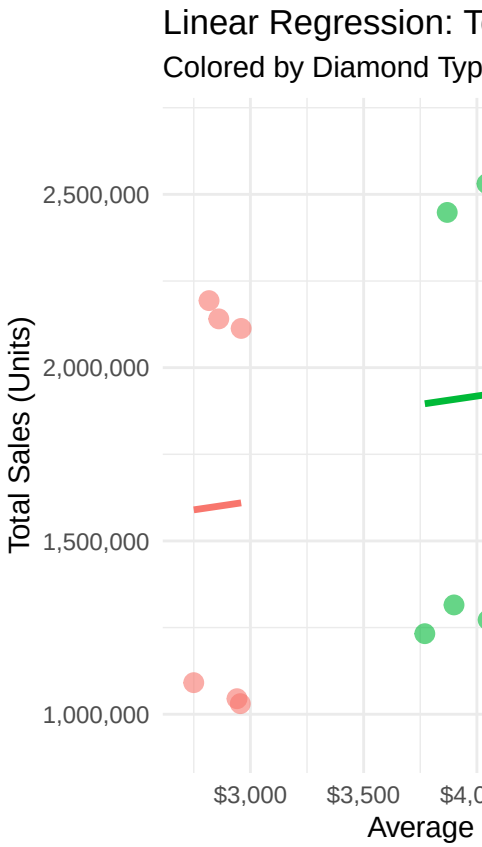
- Optimist Inventory andfocus sales to colorless are fairly cut. The box plot and the linear regression model back this statement that in the long run, fairly cut diamonds sell for even higher than premium cut diamonds. Test the sales of these diamonds in China, India and the United States first.

Therefore the last thing to do is determine a price range for the fairly cut diamonds that will be low enough to attract sales volume and high enough to maximize profits.

Let's make a prediction of the

Linear Model 3: Predicting Total sales to Average Price

```
# Fit linear regression model
sales_model <- linear_reg() %>%
  set_engine("lm") %>% # the response variable is the total sales
  fit(total_sales ~ avg_price + color_category + cut, data = top_diamonds_country) #explanatory vars ar
(top_diamonds_augmented <- augment(sales_model, new_data = top_diamonds_country)|> # once the data is a
  ggplot(mapping= aes(x = avg_price, y = total_sales, color = diamond_type)) +
  geom_point(size = 3, alpha = 0.6) +
  geom_line(aes(y = .pred), linewidth = 1.2) + # The lines were to thin when we made this.
  labs(title = "Linear Regression: Total Sales vs. Average Price", subtitle = "Colored by Diamond Type
    x = "Average Price (USD)", y = "Total Sales (Units)", color = "Diamond Type") +
  scale_x_continuous(labels = scales::dollar) + scale_y_continuous(labels = scales::comma) +
  theme_minimal(base_size = 11)) # This prevents the graph from being out of view when Rmd is knitted
```



Determining the right price to maximize the sale of fairly cut diamonds

Observation: Linear Model 3 The chart reveals that Colorless/Ideal diamonds consistently achieve high sales regardless of price, suggesting strong consumer preference and demand stability. Near-colourless/Ideal

diamonds show a slight positive relationship between price and sales, indicating that customers may associate higher prices with better quality.

Conclusion

Based of the average price prediction models vs the actual data, we are suggesting that the jewellery firm considers selling the fair cut diamonds more with a USD3750 - USD4250 price range. Additionally they should optimist diamond acquisition to Near-colorless diamonds with carat weights 2.5 to 3. Regarding the broader context of the study, sales year over year across the countries are even, which means that performance at the different branches of the jewellery firm is kept to high standards globally. There seems to be emerging markets in (see explatory plot 2) in the likes of Russia, Switzerland, and Singapore. These are markets to keep an eye on.

Analysis Limitations

There are a number of limitations to this analysis. The first being the our limited experience in the overall precious stones industry. Secondly, the data in the analysis is remote to one jewellery firm. So comparisons with rival jewellery firm prices, and market share are not taken into account. Lastly, Our analysis is limited to 12 pages....

####Notes: For some reason, we couldn't figure out why Linear Model 3 was showing out of the PDF. We tried alot of things, but ultimately could figure out what was wrong... please let us know how we could have fixed it. Thanks and have a great summer break!

References

This analysis wouldn't have been possible without the following sources:

Brilliance.com A online diamond retailer Used this site learn about the diamond color chart and making the regular color_category expressions Article Link: https://www.brilliance.com/education/diamonds/color

Brilliant Earth An online jeweller retailer Used this site to learn aboutdiamond clarities and the making of clarity_description variable Article Link: https://www.brilliantearth.com/en-ca/diamond/buying-guide/clarity/

Hubspot An online business blog Used to form a framework making sales forecasting and analysis for the project... Our first time doing this lol. https://blog.hubspot.com/sales/sales-projection

Class

Report by: Theta Capital Research analysis and documentation by: Lloyd Nsambu, Sidon Mengsteab, Kim Kiju