

DESCRIPCIÓN BREVE

Este documento trata sobre solucionar un problema en lenguaje de programación R

Luis Eduardo Bahena Castillo

9°C IDyGS

P501

Ing. Maximiliano Carsi Castrejón – Extracción y Conocimiento en Bases de Datos



INTRODUCCIÓN

Práctica en Clase: Visualización de Datos del Conjunto diamonds

Objetivo:

Explorar y visualizar las relaciones entre los atributos de los diamantes y su precio utilizando **ggplot2**, con enfoque en la personalización de sistemas de coordenadas, ejes y esquemas de colores.

Instrucciones:

1. Preparación y Exploración de Datos:

- Cargar el conjunto de datos **diamonds** de **ggplot2**.
- Realizar una exploración básica para familiarizarse con las variables.

2. Visualización Básica:

- Crear un gráfico de dispersión para explorar la relación entre el precio (**price**) y el peso en quilates (**carat**). Utilizar **geom_point()** y ajustar la transparencia (**alpha**) para manejar la sobreimpresión.

3. Aplicación de Sistemas de Coordenadas:

- Experimentar con diferentes sistemas de coordenadas (por ejemplo, **coord_flip()** para invertir los ejes, **coord_fixed()** para escalas fijas) en el gráfico de dispersión creado anteriormente y discutir cómo cambia la interpretación.

4. Personalización de Ejes:

- Personalizar los ejes del gráfico de dispersión, modificando etiquetas, límites y la apariencia de las líneas de guía. Utilizar **scale_x_continuous()** y **scale_y_continuous()** para personalizar los ejes.

5. Exploración de Esquemas de Colores:

- Modificar el gráfico para incluir el corte del diamante (**cut**) como una tercera dimensión, utilizando el color para diferenciar entre las categorías. Aplicar diferentes esquemas de colores utilizando **scale_color_brewer()** y discutir cuál proporciona la mejor visualización.

6. Tarea Avanzada - Visualización Multivariable:

- Crear un gráfico que combine múltiples atributos (por ejemplo, **price**, **carat**, **cut**) utilizando facetas (**facet_wrap()** o **facet_grid()**) y color. Reflexionar sobre las historias que estos gráficos multivariables pueden contar sobre los datos.

7. Reflexión y Discusión:

- Discutir cómo la elección del sistema de coordenadas, los ejes y el esquema de colores afecta la interpretación de los datos. Reflexionar sobre la importancia de estas decisiones en la visualización de datos.

Entregables:

- Un script de R que contenga todo el código utilizado para crear las visualizaciones, debidamente comentado.
- Un breve informe escrito que incluya:
 - Descripciones de cada visualización creada, incluyendo la justificación de las decisiones de diseño tomadas.
 - Discusión sobre los insights o patrones observados en el conjunto de datos **diamonds** a través de las visualizaciones.

Criterios de Evaluación:

- Claridad y creatividad en las visualizaciones generadas.
- Profundidad en la justificación de las decisiones de diseño y en la interpretación de las visualizaciones.
- Calidad y coherencia en la presentación del informe escrito.

DESARROLLO

Introducción

En este informe se exploran y visualizan las relaciones entre los atributos del conjunto de datos **diamonds** y su precio. Utilizando la librería **ggplot2** de R, se desarrollaron varias visualizaciones que permiten un análisis detallado de cómo ciertos factores, como el peso en quilates y el corte del diamante, influyen en su precio. Este informe describe las visualizaciones creadas, justificando las decisiones de diseño tomadas, y discute los insights y patrones observados en los datos.

1. Preparación y Exploración de Datos

El análisis comenzó con la carga del conjunto de datos **diamonds**, disponible en la librería **ggplot2**. Este conjunto de datos contiene información detallada sobre aproximadamente 54,000 diamantes, incluyendo variables como precio, quilates, corte, color, claridad, entre otras. La exploración inicial de los datos, mostrando las primeras filas, permitió una rápida familiarización con la estructura y los tipos de variables disponibles.

```
# Cargar el conjunto de datos 'diamonds' de ggplot2
```

```
data("diamonds")
```

```
# Realizar una exploración básica de las primeras filas del conjunto de datos
```

```
print(head(diamonds))
```

2. Visualización Básica

La primera visualización creada fue un gráfico de dispersión que explora la relación entre el peso en quilates (**carat**) y el precio (**price**) de los diamantes. Se utilizó la función **geom_point()** para representar cada diamante como un punto en el gráfico, donde el eje **x** representa el peso y el eje **y** el precio.

- **Justificación del Diseño:** Se ajustó la transparencia (**alpha = 0.5**) de los puntos para manejar la sobreimpresión, dado que existen muchos diamantes con valores similares en estas variables. Esto permite visualizar mejor la densidad de puntos en ciertas áreas del gráfico, lo que es crucial para identificar patrones en datos densos.

- **Insights Observados:** Como era de esperarse, se observa una clara tendencia positiva: a medida que aumenta el peso en quilates, también lo hace el precio. Sin embargo, esta relación no es lineal, sugiriendo que otros factores podrían estar influyendo en el precio de manera significativa.

```
# Crear un gráfico de dispersión para explorar la relación entre 'price'  
y 'carat'
```

```
scatter_plot <- ggplot(diamonds, aes(x = carat, y = price)) +  
  
  geom_point(alpha = 0.5) + # Ajustar la transparencia para manejar  
  la sobreimpresión  
  
  ggtitle("Relación entre Precio y Quilates de Diamantes") +  
  
  xlab("Peso en Quilates (Carat)") +  
  
  ylab("Precio en USD") +  
  
  theme_minimal() # Aplicar un tema minimalista al gráfico  
  
# Mostrar el gráfico de dispersión básico  
  
print(scatter_plot)
```

3. Aplicación de Sistemas de Coordenadas

Para comprender cómo diferentes sistemas de coordenadas afectan la interpretación de los datos, se experimentó con `coord_flip()` y `coord_fixed()`.

- **Coord_flip():** Este sistema invierte los ejes `x` y `y`. Aunque no cambia la naturaleza de la relación entre `carat` y `price`, puede ser útil en situaciones donde se desee resaltar el eje `y` o para gráficos donde la longitud horizontal sea preferible por razones de presentación.
- **Coord_fixed():** Este sistema mantiene una relación fija entre las unidades de los ejes `x` y `y`. Al aplicar `coord_fixed(ratio = 1)`, se observa una representación equitativa de ambas variables, lo que es útil para evitar distorsiones en la percepción visual de la relación entre `carat` y `price`.
- **Justificación del Diseño:** La experimentación con diferentes sistemas de coordenadas es crucial para asegurar que las visualizaciones transmitan la información de manera precisa y sin sesgos visuales.
- **Insights Observados:** El uso de `coord_fixed` resaltó la no linealidad de la relación entre peso y precio, mientras que `coord_flip` ofreció una perspectiva alternativa que podría ser más útil en ciertos contextos de presentación.

```
# Probar con 'coord_flip' para invertir los ejes
```

```
scatter_plot_flip <- scatter_plot + coord_flip()

print(scatter_plot_flip)

# Probar con 'coord_fixed' para escalas fijas

scatter_plot_fixed <- scatter_plot + coord_fixed(ratio = 1)

print(scatter_plot_fixed)
```

4. Personalización de Ejes

El siguiente paso fue personalizar los ejes del gráfico de dispersión. Utilizando `scale_x_continuous()` y `scale_y_continuous()`, se modificaron las etiquetas, los límites y la apariencia de las líneas de guía.

- **Justificación del Diseño:** Los ejes fueron limitados a rangos específicos (x: 0-5 carat, y: 0-20000 USD) para enfocar el análisis en la mayoría de los datos y evitar la influencia de outliers extremos que podrían distorsionar la interpretación general. Además, se cambiaron los colores de los títulos de los ejes y se personalizaron las líneas de guía para mejorar la claridad y el atractivo visual del gráfico.
- **Insights Observados:** Esta personalización hizo evidente que la mayoría de los diamantes en el conjunto de datos tienen un peso inferior a 2.5 quilates, y precios que rara vez exceden los 15,000 USD. Esto sugiere que los diamantes fuera de estos rangos son menos comunes y posiblemente pertenecen a un nicho específico del mercado.

```
# Personalizar los ejes del gráfico de dispersión

scatter_plot_custom_axes <- scatter_plot +

  scale_x_continuous(name = "Peso en Quilates (Carat)", limits = c(0,
5)) + # Limitar el eje x

  scale_y_continuous(name = "Precio en USD", limits = c(0, 20000)) +
# Limitar el eje y

  theme(

    axis.title.x = element_text(color = "blue", size = 14),

    axis.title.y = element_text(color = "red", size = 14),

    panel.grid.major = element_line(size = 0.5, linetype = 'dashed')

  )
```

```
# Mostrar el gráfico con ejes personalizados
```

```
print(scatter_plot_custom_axes)
```

5. Exploración de Esquemas de Colores

Para añadir una dimensión adicional al análisis, se incorporó el corte del diamante (**cut**) como una variable categórica representada por el color en el gráfico de dispersión.

- **Justificación del Diseño:** Se utilizó `scale_color_brewer()` para aplicar diferentes esquemas de colores, seleccionando paletas que maximizan la distinción entre categorías de corte. La elección de paletas como "Set1" permitió una diferenciación clara y estéticamente agradable entre los distintos niveles de corte.
- **Insights Observados:** Al observar la relación entre **carat**, **price**, y **cut**, se pudo notar que los diamantes con cortes de mayor calidad tienden a tener un precio más elevado para un mismo peso en quilates. Este patrón sugiere que el corte es un factor determinante en la valoración de los diamantes, más allá de su tamaño.

```
# Modificar el gráfico para incluir el corte del diamante (cut) como una  
tercera dimensión
```

```
scatter_plot_color <- scatter_plot +  
  
  aes(color = cut) + # Usar 'cut' para diferenciar con colores  
  
  scale_color_brewer(palette = "Set1") + # Aplicar un esquema de  
colores  
  
  ggtitle("Relación entre Precio y Quilates con Corte de Diamante")  
  
# Mostrar el gráfico con esquemas de colores  
  
print(scatter_plot_color)
```

6. Tarea Avanzada - Visualización Multivariable

Para un análisis más complejo, se creó un gráfico que combina múltiples atributos (**price**, **carat**, **cut**) utilizando facetas (`facet_wrap()`) y color.

- **Justificación del Diseño:** Las facetas permiten descomponer el gráfico según el corte del diamante, mientras que el color diferencia las categorías dentro de cada faceta. Esto facilita la comparación `facet_wrap()` directa de cómo cada nivel de corte afecta la relación entre peso y precio en los diamantes.
- **Insights Observados:** Este gráfico reveló patrones más sutiles: por ejemplo, para cortes de menor calidad, el incremento en precio por quilate es más gradual comparado con cortes de alta calidad, donde pequeños incrementos en quilates

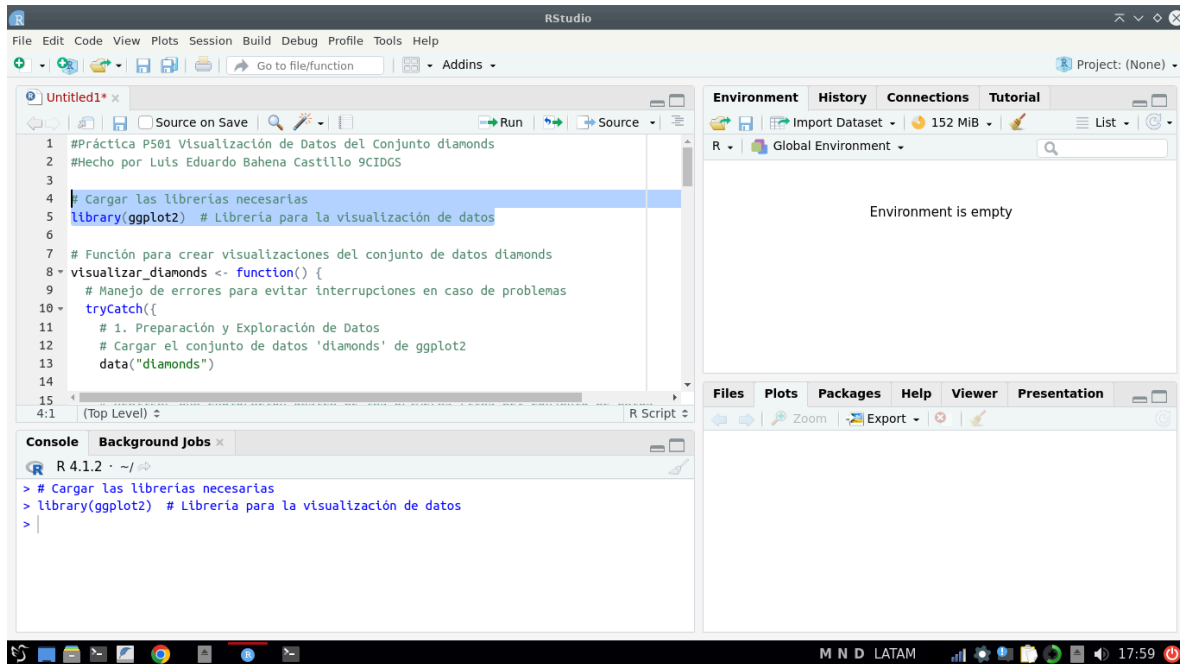
resultan en aumentos significativos en precio. Esto sugiere que los compradores valoran proporcionalmente más el tamaño en diamantes de menor calidad, mientras que en cortes superiores, la calidad del corte se convierte en el principal motor de precio.

Crear un gráfico que combine múltiples atributos utilizando facetas y color

```
multivariable_plot <- ggplot(diamonds, aes(x = carat, y = price,  
color = cut)) +  
  
  geom_point(alpha = 0.5) +  
  
  facet_wrap(~ cut) + # Usar facetas para cada tipo de corte de  
diamante  
  
  scale_color_brewer(palette = "Set2") +  
  
  ggtitle("Precio vs Quilates con Facetas por Corte de Diamante")  
  
# Mostrar el gráfico multivariable  
  
print(multivariable_plot)
```

Resultados

A continuación se proporciona los resultados del código para la Visualización de Datos del Conjunto diamonds en el que todo lo que se realiza se encapsula en una función y se maneja un error por si ocurre un error:



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source on Save Run Source

Project: (None)

Environment History Connections Tutorial

Global Environment

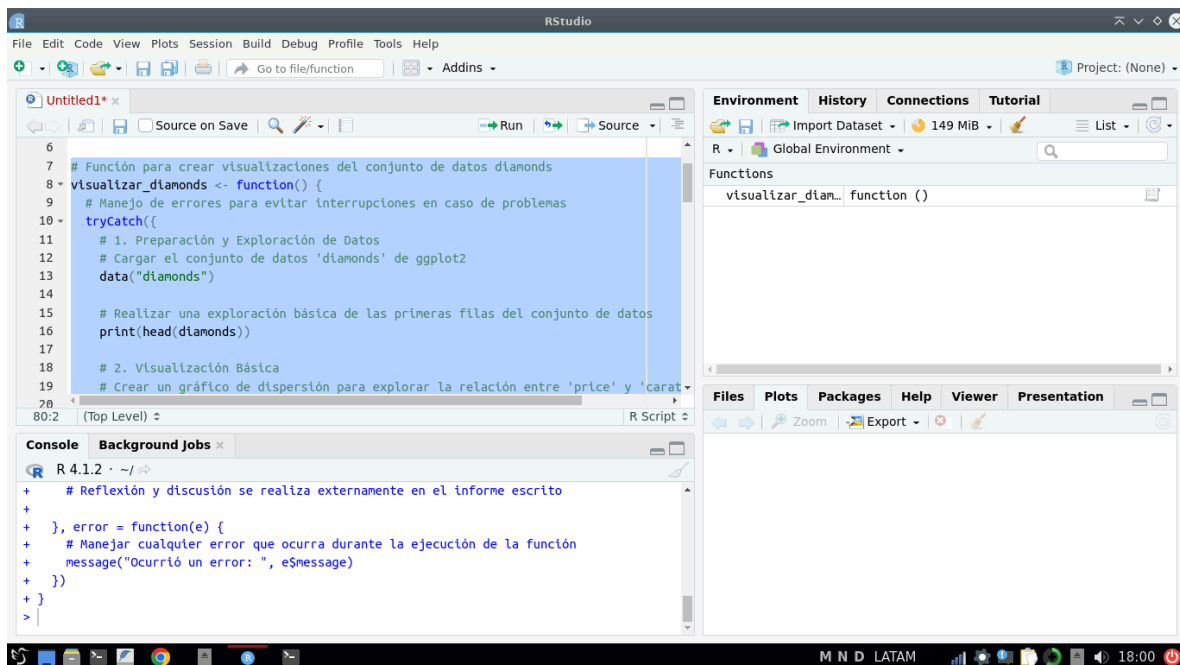
Environment is empty

Files Plots Packages Help Viewer Presentation

```

1 #Práctica P501 Visualización de Datos del Conjunto diamonds
2 #Hecho por Luis Eduardo Bahena Castillo 9CIDGS
3
4 # Cargar las librerías necesarias
5 library(ggplot2) # Librería para la visualización de datos
6
7 # Función para crear visualizaciones del conjunto de datos diamonds
8 visualizar_diamonds <- function() {
9   # Manejo de errores para evitar interrupciones en caso de problemas
10  tryCatch({
11    # 1. Preparación y Exploración de Datos
12    # Cargar el conjunto de datos 'diamonds' de ggplot2
13    data("diamonds")
14  })
15 }
16
17 Console Background Jobs
18 R 4.1.2 ~ /
19 > # Cargar las librerías necesarias
20 > library(ggplot2) # Librería para la visualización de datos
21 >
  
```

M N D LATAM 17:59



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source on Save Run Source

Project: (None)

Environment History Connections Tutorial

Global Environment

Functions

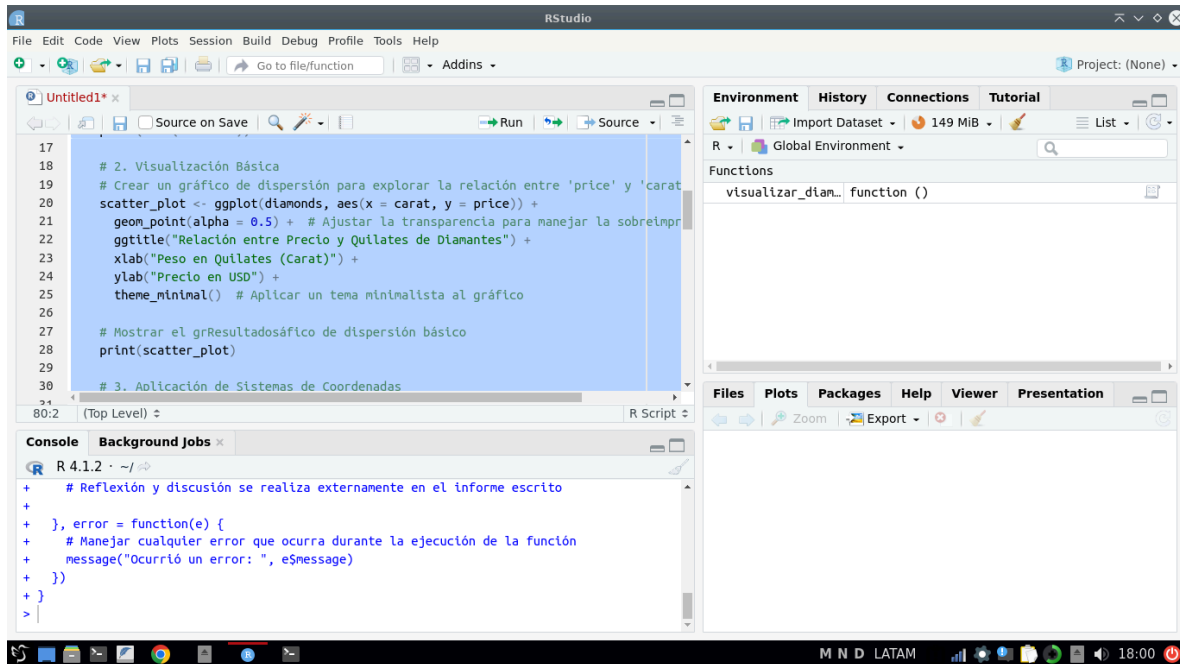
visualizar_dian... function ()

Files Plots Packages Help Viewer Presentation

```

6
7 # Función para crear visualizaciones del conjunto de datos diamonds
8 visualizar_diamonds <- function() {
9   # Manejo de errores para evitar interrupciones en caso de problemas
10  tryCatch({
11    # 1. Preparación y Exploración de Datos
12    # Cargar el conjunto de datos 'diamonds' de ggplot2
13    data("diamonds")
14
15    # Realizar una exploración básica de las primeras filas del conjunto de datos
16    print(head(diamonds))
17
18    # 2. Visualización Básica
19    # Crear un gráfico de dispersión para explorar la relación entre 'price' y 'carat'
20  })
21 }
22
23 Console Background Jobs
24 R 4.1.2 ~ /
25 + # Reflexión y discusión se realiza externamente en el informe escrito
26 + }, error = function(e) {
27 +   # Manejar cualquier error que ocurra durante la ejecución de la función
28 +   message("Ocurrió un error: ", e$message)
29 + }
30 + }
31 >
  
```

M N D LATAM 18:00



RStudio interface showing R code for creating a scatter plot and a function for error handling.

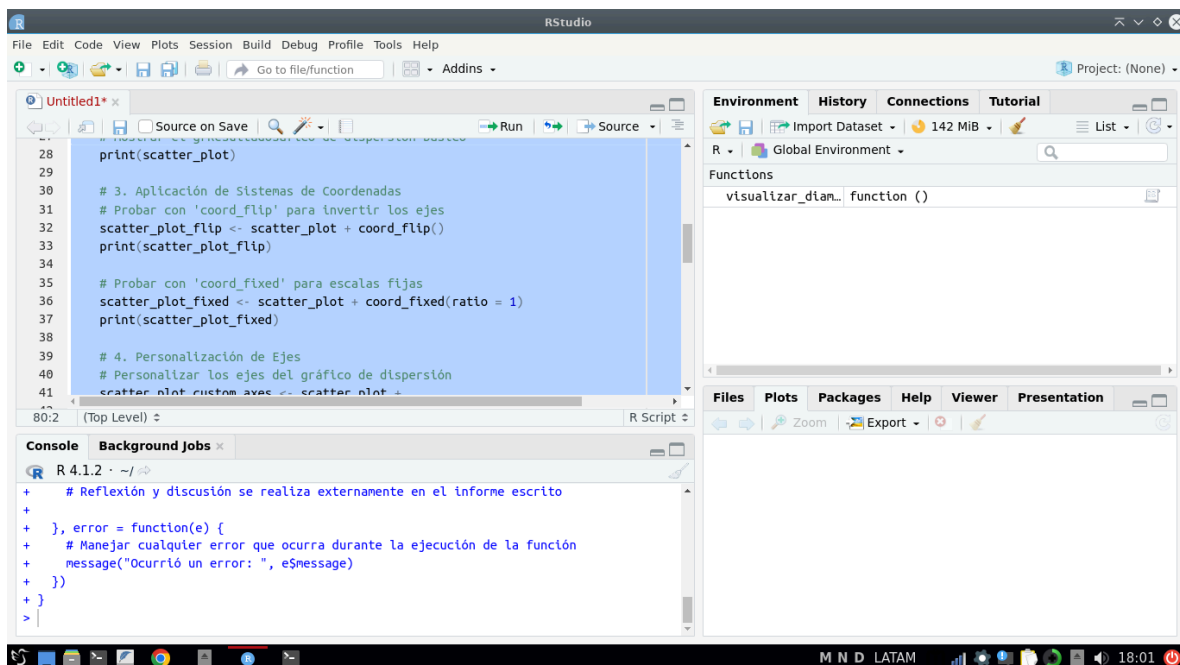
```

17 # 2. Visualización Básica
18 # Crear un gráfico de dispersión para explorar la relación entre 'price' y 'carat'
19 scatter_plot <- ggplot(diamonds, aes(x = carat, y = price)) +
20   geom_point(alpha = 0.5) + # Ajustar la transparencia para manejar la sobreimpresión
21   ggtitle("Relación entre Precio y Quilates de Diamantes") +
22   xlab("Peso en Quilates (Carat)") +
23   ylab("Precio en USD") +
24   theme_minimal() # Aplicar un tema minimalista al gráfico
25
26 # Mostrar el gráfico de dispersión básico
27 print(scatter_plot)
28
29 # 3. Aplicación de Sistemas de Coordenadas
30
31 # Reflexión y discusión se realiza externamente en el informe escrito
32 }, error = function(e) {
33   # Manejar cualquier error que ocurra durante la ejecución de la función
34   message("Ocurrió un error: ", e$message)
35 }
36 }
37
38
39
40
41
  
```

Console output:

```

R 4.1.2 ~ /
+ # Reflexión y discusión se realiza externamente en el informe escrito
+ }, error = function(e) {
+   # Manejar cualquier error que ocurra durante la ejecución de la función
+   message("Ocurrió un error: ", e$message)
+ }
+ }
+
  
```



RStudio interface showing R code for applying coordinate systems and customizing axes.

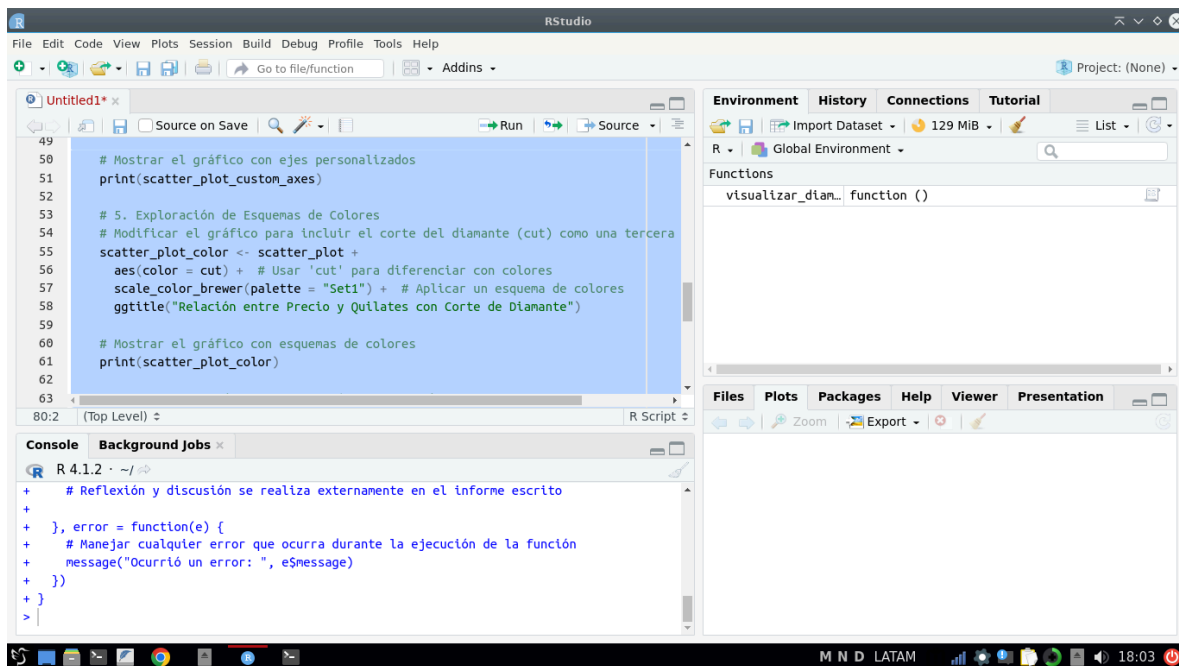
```

28 print(scatter_plot)
29
30 # 3. Aplicación de Sistemas de Coordenadas
31 # Probar con 'coord_flip' para invertir los ejes
32 scatter_plot_flip <- scatter_plot + coord_flip()
33 print(scatter_plot_flip)
34
35 # Probar con 'coord_fixed' para escalas fijas
36 scatter_plot_fixed <- scatter_plot + coord_fixed(ratio = 1)
37 print(scatter_plot_fixed)
38
39 # 4. Personalización de Ejes
40 # Personalizar los ejes del gráfico de dispersión
41 scatter_plot_custom_axes <- scatter_plot +
  
```

Console output:

```

R 4.1.2 ~ /
+ # Reflexión y discusión se realiza externamente en el informe escrito
+ }, error = function(e) {
+   # Manejar cualquier error que ocurra durante la ejecución de la función
+   message("Ocurrió un error: ", e$message)
+ }
+ }
+
  
```



RStudio

```

61 print(scatter_plot_color)
62
63 # 6. Tarea Avanzada - Visualización Multivariable
64 # Crear un gráfico que combine múltiples atributos utilizando facetas y color
65 multivariable_plot <- ggplot(diamonds, aes(x = carat, y = price, color = cut)) +
66   geom_point(alpha = 0.5) +
67   facet_wrap(~ cut) + # Usar facetas para cada tipo de corte de diamante
68   scale_color_brewer(palette = "Set2") +
69   ggtitle("Precio vs Quilates con Facetas por Corte de Diamante")
70
71 # Mostrar el gráfico multivariable
72 print(multivariable_plot)
73
74 }, error = function(e) {
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
  
```

Environment History Connections Tutorial

R - Global Environment

Functions

visualizar_dia... function ()

Files Plots Packages Help Viewer Presentation

Console Background Jobs

```

R 4.1.2 ~ /
+
+ }, error = function(e) {
+   # Manejar cualquier error que ocurra durante la ejecución de la función
+   message("Ocurrió un error: ", e$message)
+ }
+ }
+ # Función para crear visualizaciones del conjunto de datos diamonds
+ visualizar_diamonds <- function() {
  
```

M N D LATAM 18:05

RStudio

```

69 ggtitle("Precio vs Quilates con Facetas por Corte de Diamante")
70
71 # Mostrar el gráfico multivariable
72 print(multivariable_plot)
73
74 }, error = function(e) {
75   # Manejar cualquier error que ocurra durante la ejecución de la función
76   message("Ocurrió un error: ", e$message)
77 }
78 }
79
80 # Llamar a la función para ejecutar el análisis y visualización
81 visualizar_diamonds()
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
  
```

Environment History Connections Tutorial

R - Global Environment

Data


diamonds 53940 obs. of 10 variables

Functions

visualizar_dia... function ()

Files Plots Packages Help Viewer Presentation

Precio vs Quilates con Facetas por Corte de D

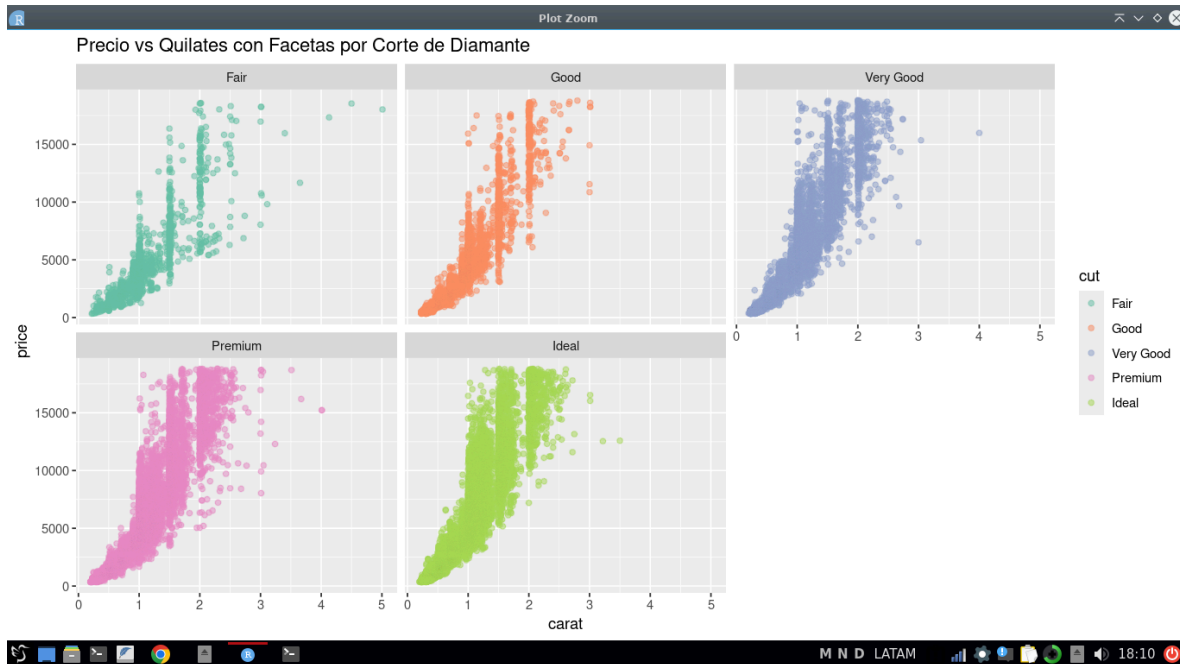


carat cut color clarity depth table price x y z

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

Warning message:

M N D LATAM 18:08



CONCLUSIONES

Reflexión y Discusión

La elección del sistema de coordenadas, los ejes y el esquema de colores tuvo un impacto significativo en la interpretación de los datos. Las visualizaciones personalizadas permitieron observar que, aunque el peso en quilates es un predictor fuerte del precio, factores como el corte del diamante pueden alterar significativamente esta relación. Las decisiones de diseño tomadas fueron guiadas por la necesidad de destacar estos patrones de manera clara y efectiva.

Este análisis subraya la importancia de elegir adecuadamente las herramientas de visualización en el análisis de datos. La capacidad de presentar la información de manera que los patrones relevantes sean fácilmente identificables es crucial para una interpretación efectiva y para la toma de decisiones basada en datos.

Conclusión

A través de las visualizaciones generadas, se pudo concluir que tanto el peso en quilates como el corte del diamante son factores determinantes en su precio. Las visualizaciones multivariantes, en particular, destacaron la complejidad de estas relaciones y la importancia de utilizar técnicas de visualización avanzadas para descomponer y entender estas complejidades en su totalidad.

Las decisiones de diseño, incluyendo la personalización de ejes y esquemas de color, jugaron un papel crucial en mejorar la claridad y utilidad de las visualizaciones, permitiendo así un análisis más profundo y matizado del conjunto de datos **diamonds**.