Trabajo 1 Data Visualization

Diabb Zegpi D.

1. Dependiendo del tipo de variable considerada en los conjuntos de datos, genere una representación visual adecuada y que sea representativa. Comente lo observado en la representación.

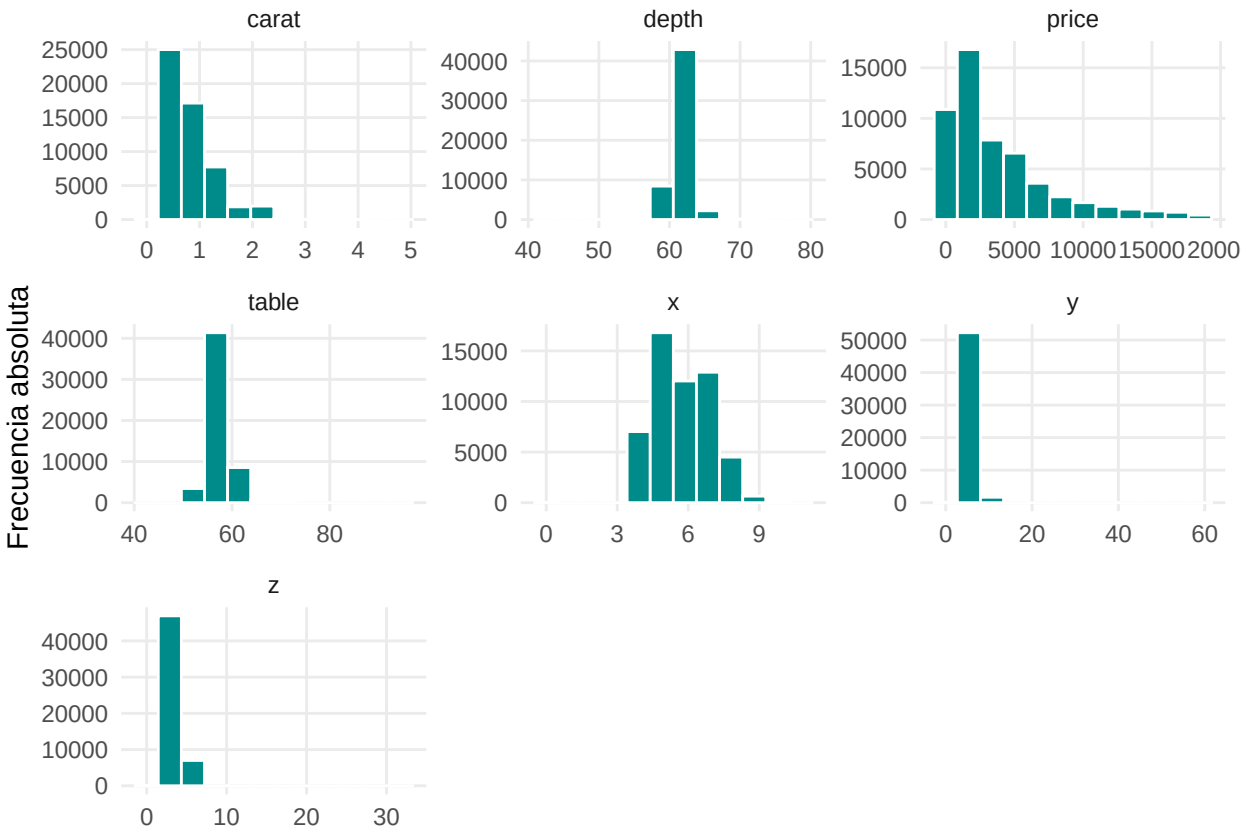

Figura 1: Diamonds | Histogramas de variables numéricas

La Figura 1 muestra la districuión de cada variable numérica en el conjunto de datos `diamonds`. Las observaciones destacables son las siguientes:

- Las variables `carat`, `table`, `y` y `z` presentan asimetría positiva y curtosis elevada.
- Tanto `depth` como `x` tienen distribuciones aparentemente simétricas.
- `price` está en una escala varios órdenes superior al resto de variables. Además, presenta asimetría positiva y un gran coeficiente de variación.

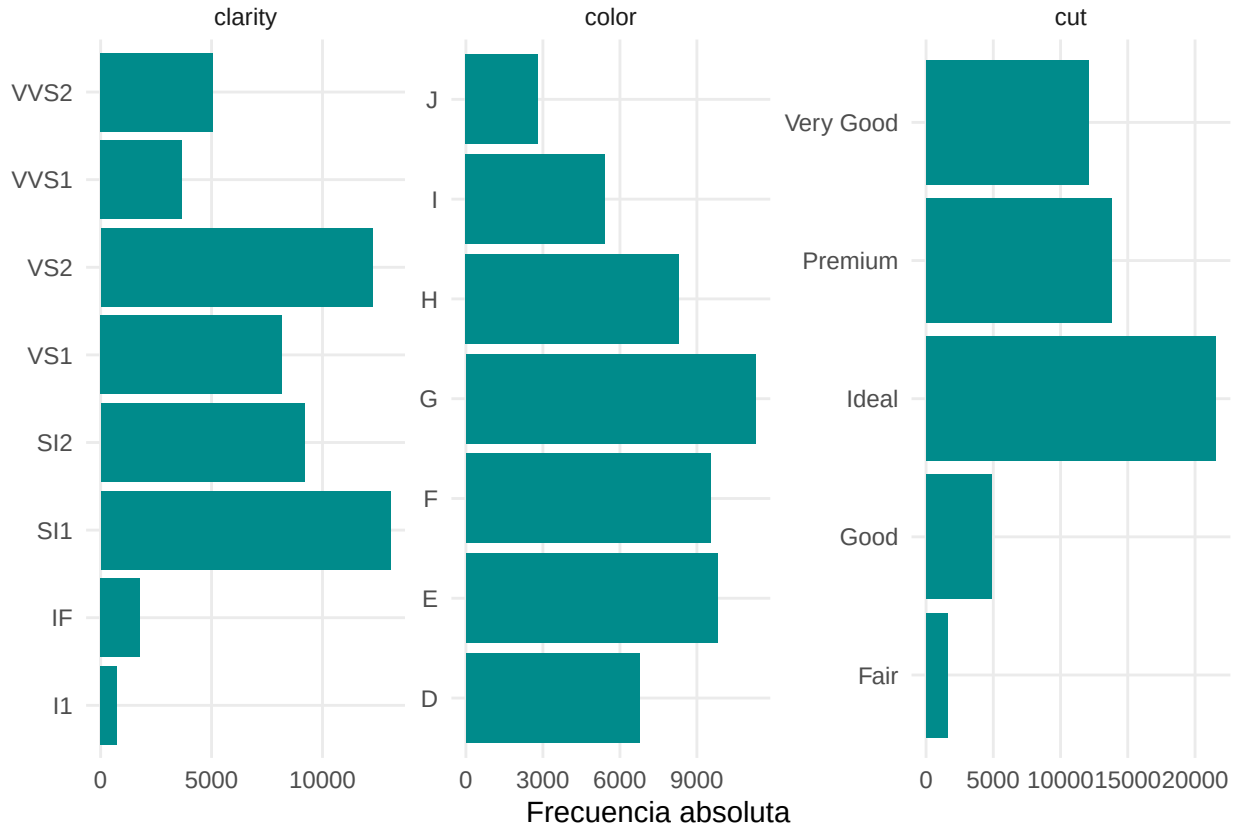

Figura 2: Diamonds | Gráfico de barras de variables categóricas

La Figura 2 muestra la frecuencia de las variables `clarify`, `color` y `cut`. Este gráfico deja en evidencia que la claridad más común es SI 1, mientras que la más rara es I1, que es la peor. Paralelamente, los colores siguen una distribución en torno a H, G y F, que son los colores de diamantes con calidad intermedia. De manera análoga, la mayoría de los cortes son intermedios, muy pocos son aceptables y muchos son premium y muy buenos.

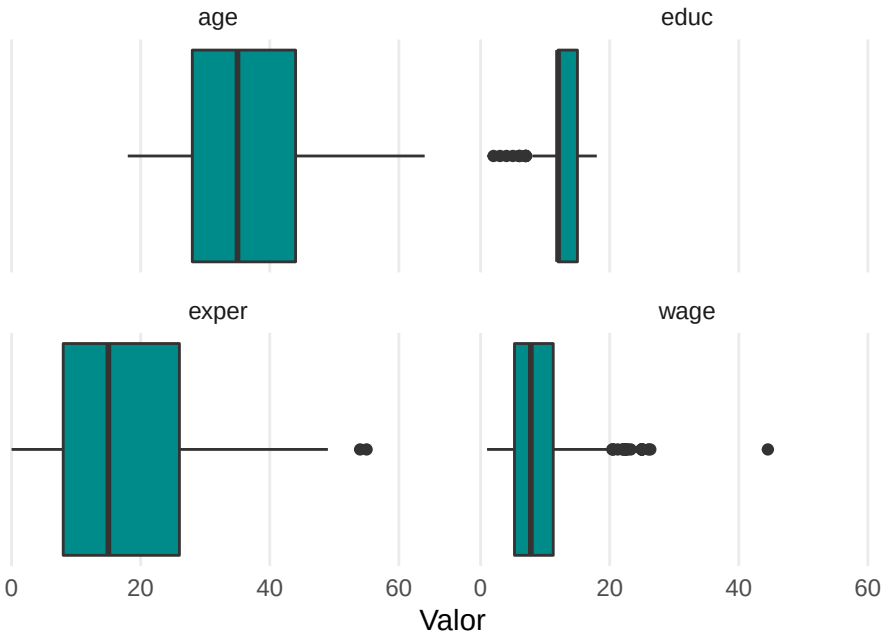

Figura 3: CPS85 | Diagramas de caja de variables numéricas

Los gráficos de caja y bigote de la Figura 3 exponen que más de la mitad de las personas tiene más de 30 años, tiene menos de 12 años de educación, con 15 o menos años de experiencia y ganan menos que 8 dólares por hora.

Por otra parte, la Figura 4 muestra la frecuencia de las variables categóricas del censo. Se obtiene que la mayoría de las personas encuestadas no son hispánicas, están casadas, son blancos, trabajan en diversos sectores, son hombres (aunque el 46% son mujeres), no viven en el sur y predomina el nivel `Not` del factor `union`.

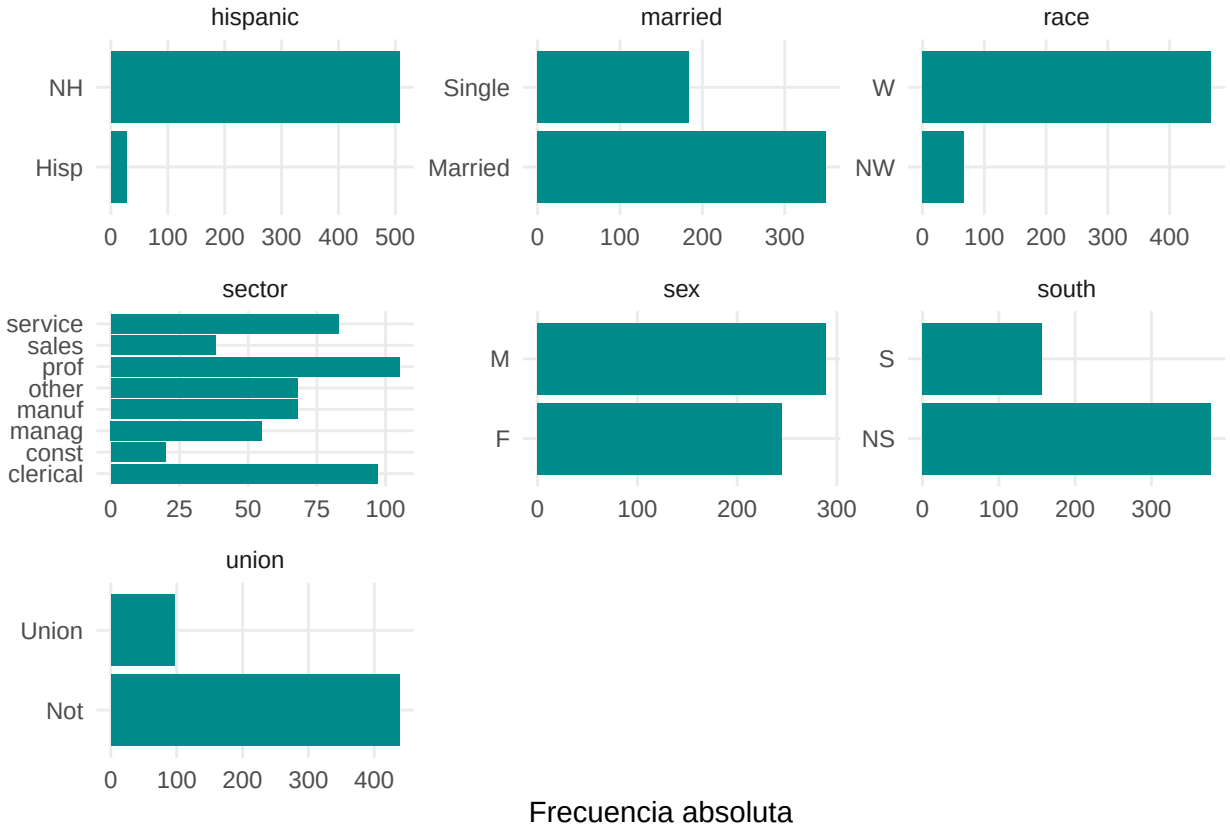

Figura 4: CPS85 | Gráfico de barras de variables categóricas

2. Considerando la función `ggplot`, genere un gráfico de dispersión para las variables carat y price de la base diamonds, realice lo mismo con la base CPS85 exper y wage comente lo que se observa.

a. De la figura creada anteriormente, genere un gráfico considerando aquellos registros con valores para la variable carat mayores o iguales a 4.

b. Repita el gráfico utilizando una escala logarítmica para ambas variables. Comente lo observado.

c. a las distintas figuras agregue un suavizamiento.

El gráfico de la Figura 5 (A) muestra una relación exponencial entre carat y price. El uso de la transformación logarítmica para uno o ambos ejes es recomendada. Por otra parte, la gran cantidad de datos hace necesaria la consideración de utilizar una muestra aleatoria o bien, graficar los puntos con transparencia.

Contrario a lo anterior, la Figura 5 (B) no presenta exceso de datos ni se observa relación entre el número de años de experiencia laboral y el salario en USD. A modo de observación estética, la variable exper es discreta, lo que provoca solapamiento entre las observaciones. Agregar un poco de ruido a la gráfica puede convertirla en una visualización más atractiva.

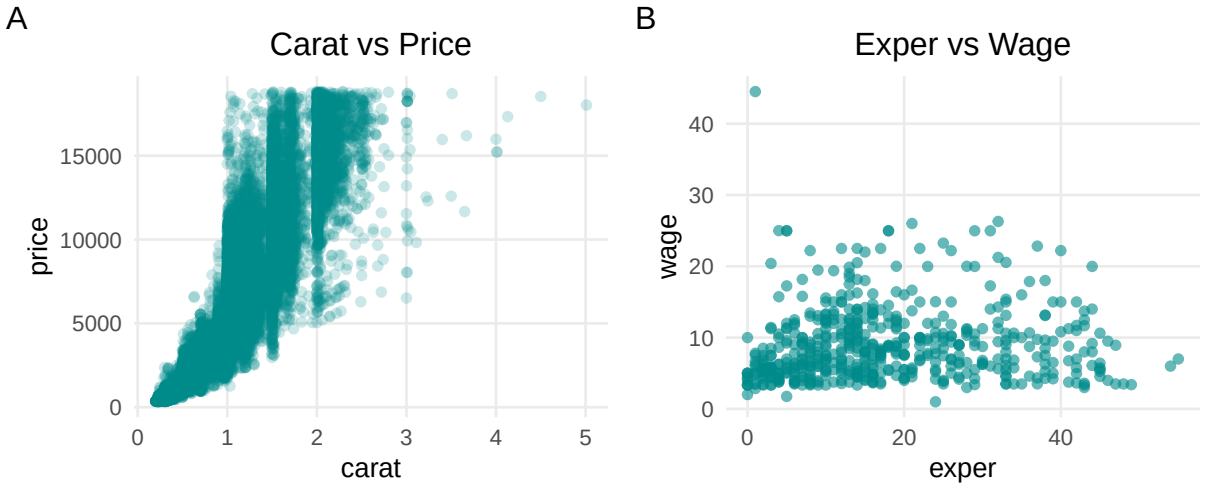

Figura 5: Diamonds & CPS85 | Gráfico de dispersión

Seleccionar solamente los diamante con `carat >= 4` deja observaciones con bajo rango de variación en ambas variables, price y carat. El resultado de aplicar escalas logarítmicas para ambos ejes y agregar una línea de regresión se observa en la Figura 6: la nube de puntos se comporta linealmente y tiene baja varianza, la que además es homogénea. Estas características son positivas para el caso en que se desee modelar una relación lineal, mejorando la bondad del ajuste y disminuyendo el error estándar de los parámetros de la regresión.

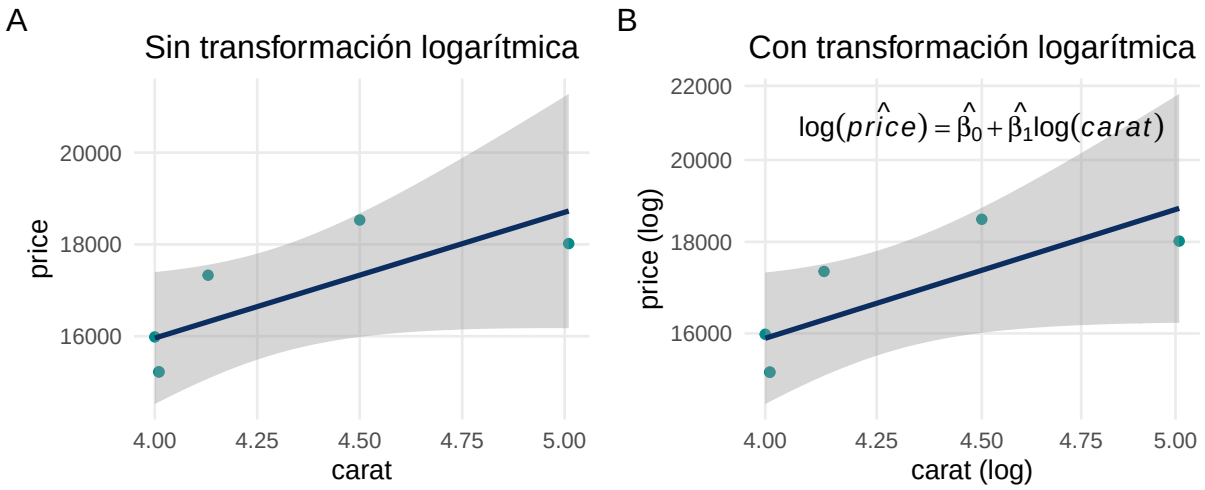

Figura 6: Diamonds | Gráfico con transformación de escala

3. Considerando la función ggplot genere, un gráfico de dispersión para las variables exper y wage de la base CPS85. Comente lo que se observa.

a. De la figura creada anteriormente genere un gráfico considerando aquellos registros con valores para la variable wage menor a 40. Por que se realizo esto, analice y comente.

b. Repita el gráfico, utilizando un color distinto dependiendo al sexo, comente lo observado.

c. A las distintas figuras agregue un suavizamiento, (recordar que solo debe agregar dos capas al gráfico)

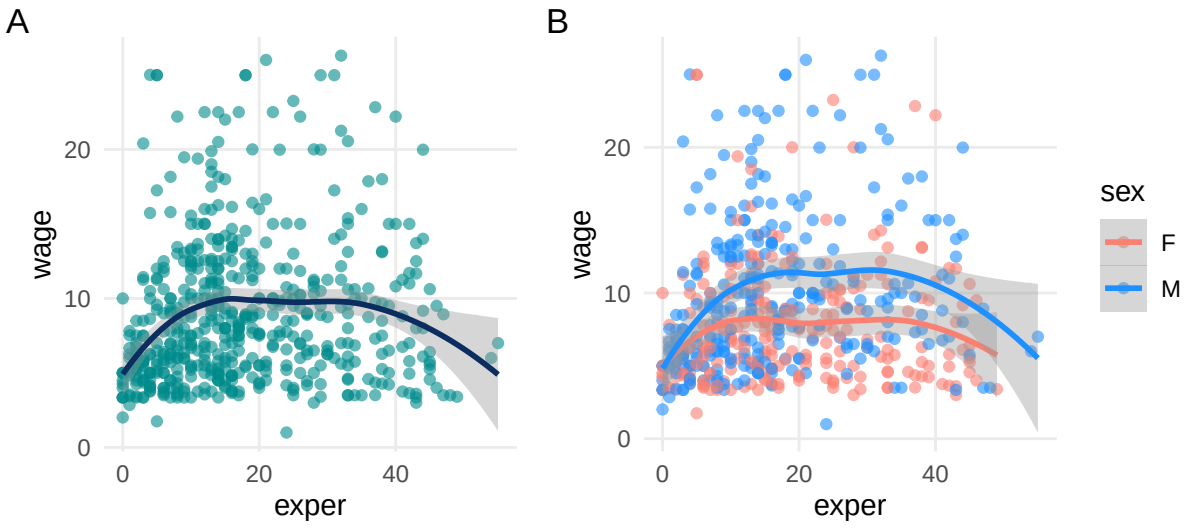

Figura 7: CPS85 | Exper vs Wage

Primero que todo, se estima que se consideraron valores para wage menores que 40 porque solamente hay un empleado con salario superior a 40 USD por hora, correspondiendo a un outlier.

Ambos gráficos de la Figura 7 (A y B) muestran cómo varía el salario por los años de experiencia del empleado. El panel A indica que los empleados tienden a ganar menos en las etapas tempranas y tardías de su vida laboral, mientras que su salario aumenta entre los 10 y 32 años de experiencia. Sin embargo, la línea de regresión local muestra mucha dispersión en torno al eje de las ordenadas, indicando que el modelo ajustado a los datos no es representativo del fenómeno real de estudio.

El panel B muestra que los hombres tienden a ganar más que las mujeres, para un mismo nivel de experiencia.

4. Usando la función simplificada qplot (acrónimo para quick plot, o gráfico rápido):

a. Cree un gráfico para la variable price. Comente.

b. Repita el gráfico fijando el parámetro binwidth como 100, ¿qué observa?

c. Cree un gráfico de barras para la variable cut. Comente.

d. ¿Qué observa respecto a la función qplot? Comentarios.

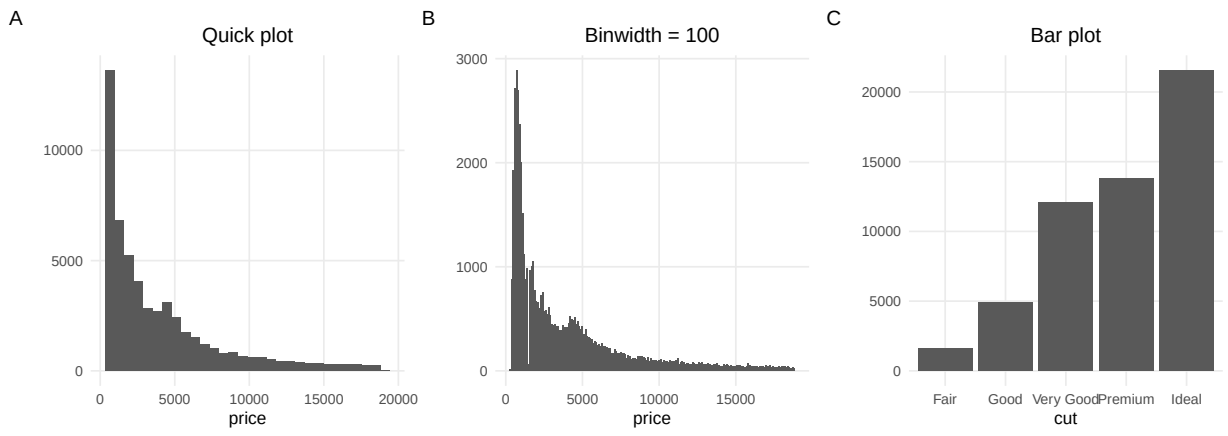

Figura 8: Diamonds | Quick plot

La función `qplot()` aplicada a price (Figura 8) forma un histograma con '`bins = 30` clases; declarar el parámetro `binwidth = 100` ajusta el tamaño de las clases para ser iguales a 100, provocando un efecto de suavizamiento y aproximándose a una curva de densidad.

Para el otro caso, qplot dibujó un gráfico de barras. En conclusión, se observa que la función `qplot()` hace gráficos predeterminados para el tipo y cantidad de variables que se le suministre.

5. Ahora bien, reúna todas las barras en una, pintando cada valor de cut de un color diferente.

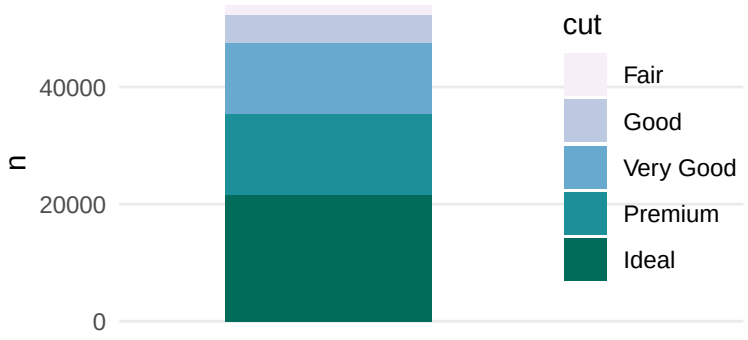

Figura 9: Diamonds | Barra de tipos de Cut

6. Los gráficos de torta corresponden simplemente a la misma representación dibujada en coordenadas polares. Genere un grafico de torta para la variable cut.

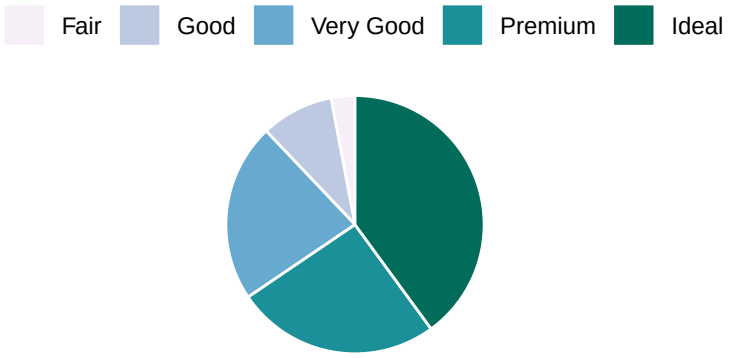

Figura 10: Diamonds | Torta de tipos de Cut

7. Genere los gráficos de dispersión para cada nivel de cut, puede considerar usar la función facet_grid.

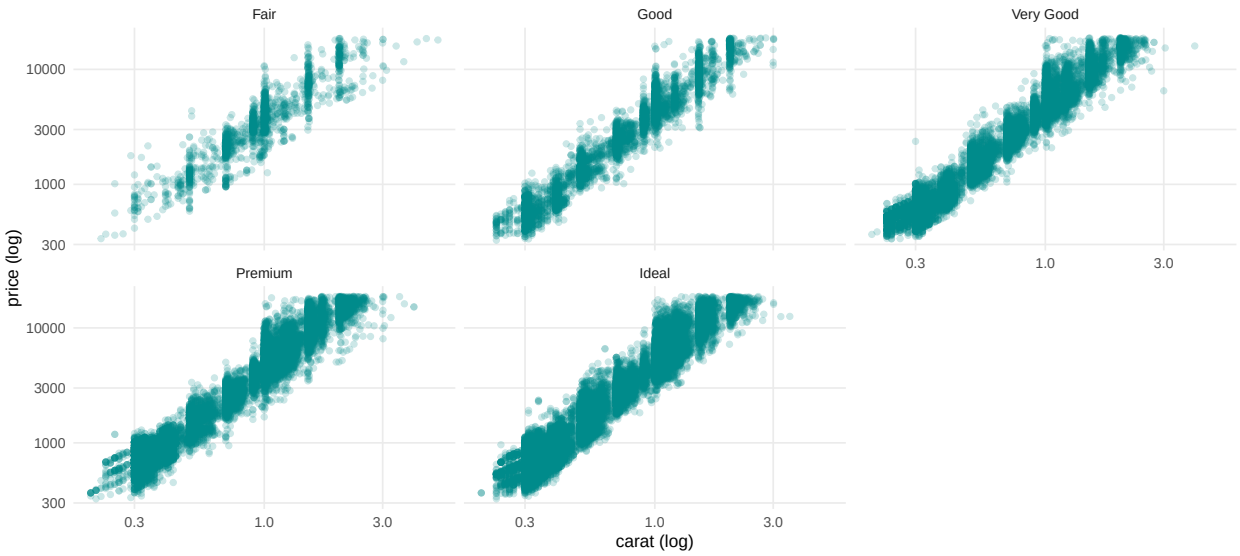

Figura 11: Diamonds | Gráfico de dispersión por Cut

8. Genere un gráfico de dispersión para las variables caret y price, para todos aquellos registros con caret mayor o igual a 3, definiendo por color mediante la variable clarity y el tamaño con la variable cut. Comente.

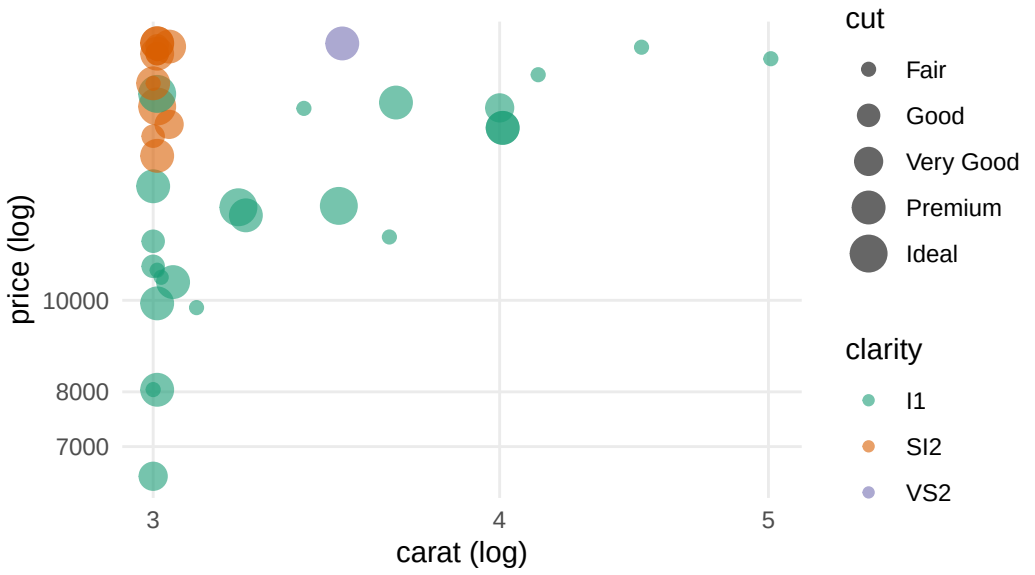

Figura 12: Diamonds | Gráfico de dispersión por Clarity y Cut

De acuerdo con la Figura 12, los diamantes con claridad SI2 son los más caros, mientras que el precio de los diamantes categoría I1 varía con gran amplitud. Además, para `carat >= 3`, la calidad del corte suele ser Very Good, Premium o Ideal.

9. Genere los gráficos de dispersión para exper y wage, dependiendo del sexo utilizando la función facet_grid para su sector laboral.

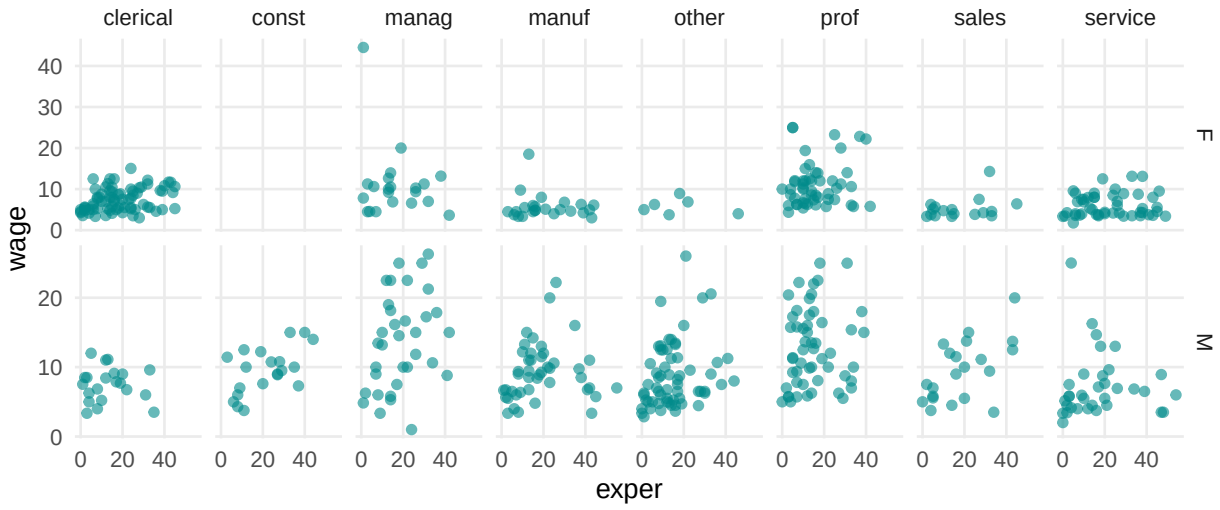

Figura 13: CPS85 | Gráfico de dispersión por Sex y Sector