Analítica de los Negocios

Explorando datos categóricos

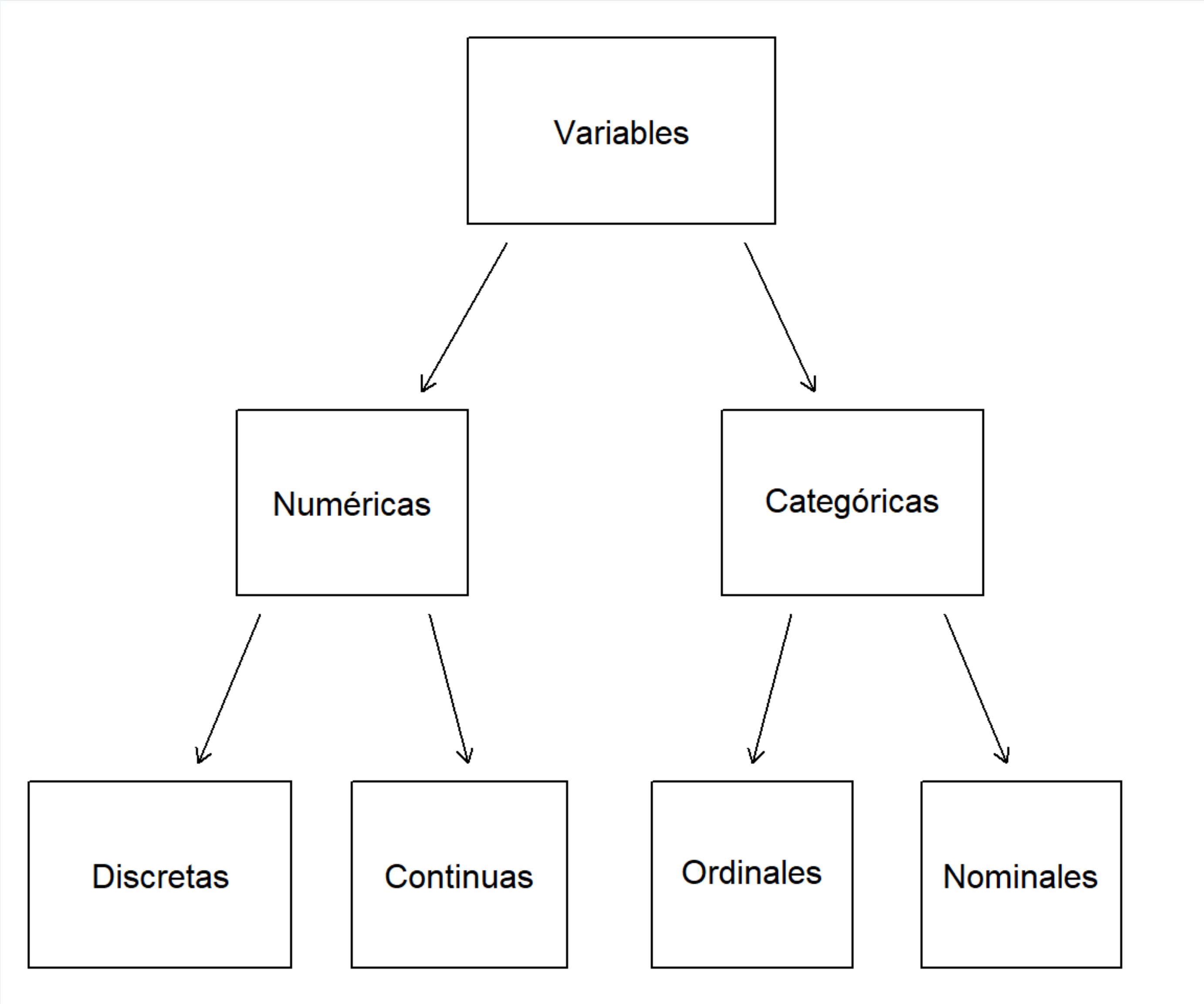
Carlos Cardona Andrade

Plan para hoy

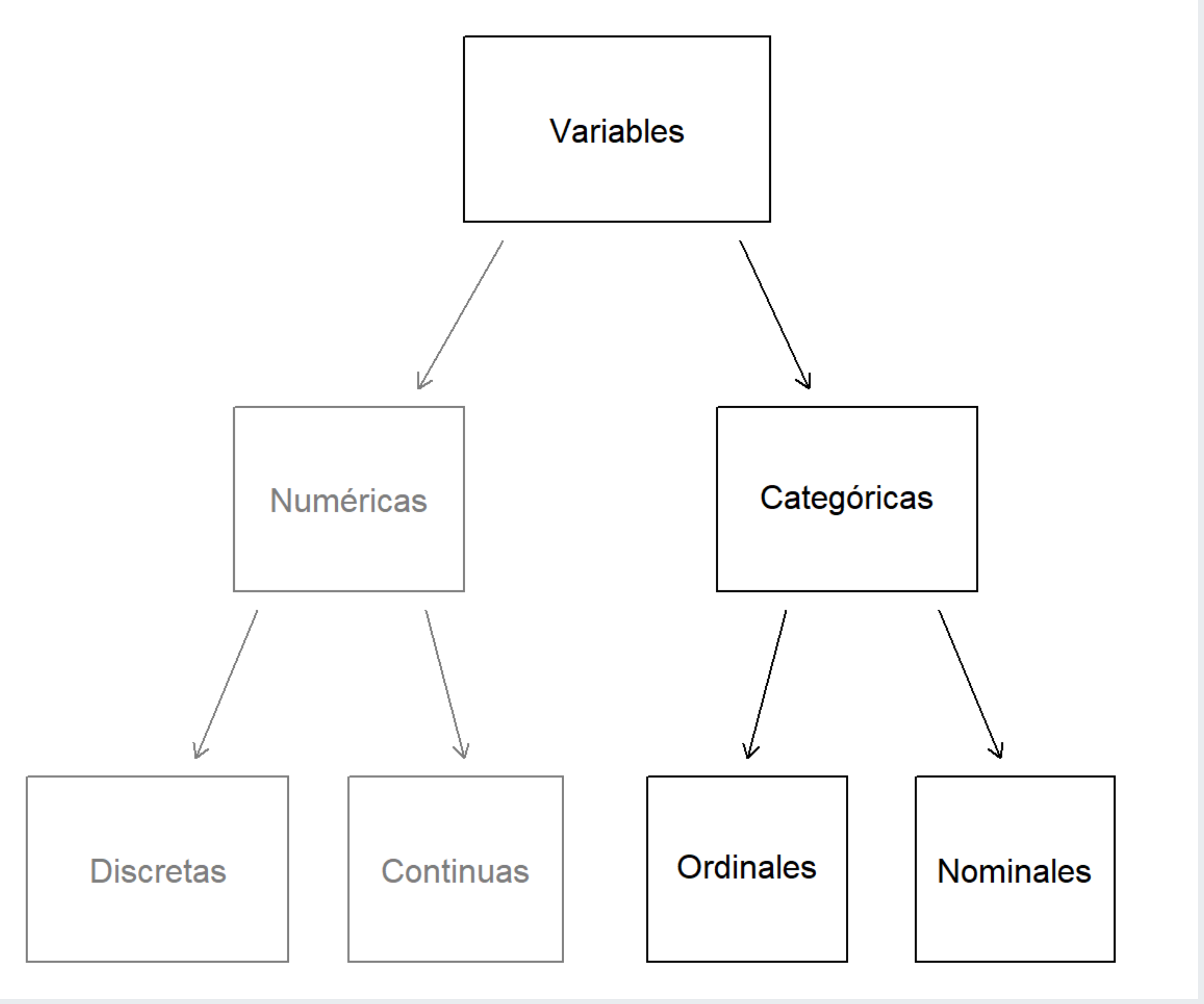
1. Datos categóricos
2. Visualizando una variable categórica
3. Visualizando dos variables categóricas
4. Comparando variables numéricas entre grupos
5. Probabilidad
6. La Distribución Normal

Datos categóricos

Tipos de variables



Variables Categóricas



Visualizando una variable categórica

Tabla de Frecuencia

- Una variable categórica se resume mediante una tabla que muestra la frecuencia o el porcentaje de casos en cada categoría
- Suele representarse mediante un gráfico de barras o un gráfico de torta

homeownership	**Frequency**
rent	3858
mortgage	4789
own	1353
Total	10000

Gráfico de barras

Un gráfico de barras es la forma más común de representar una única variable categórica.

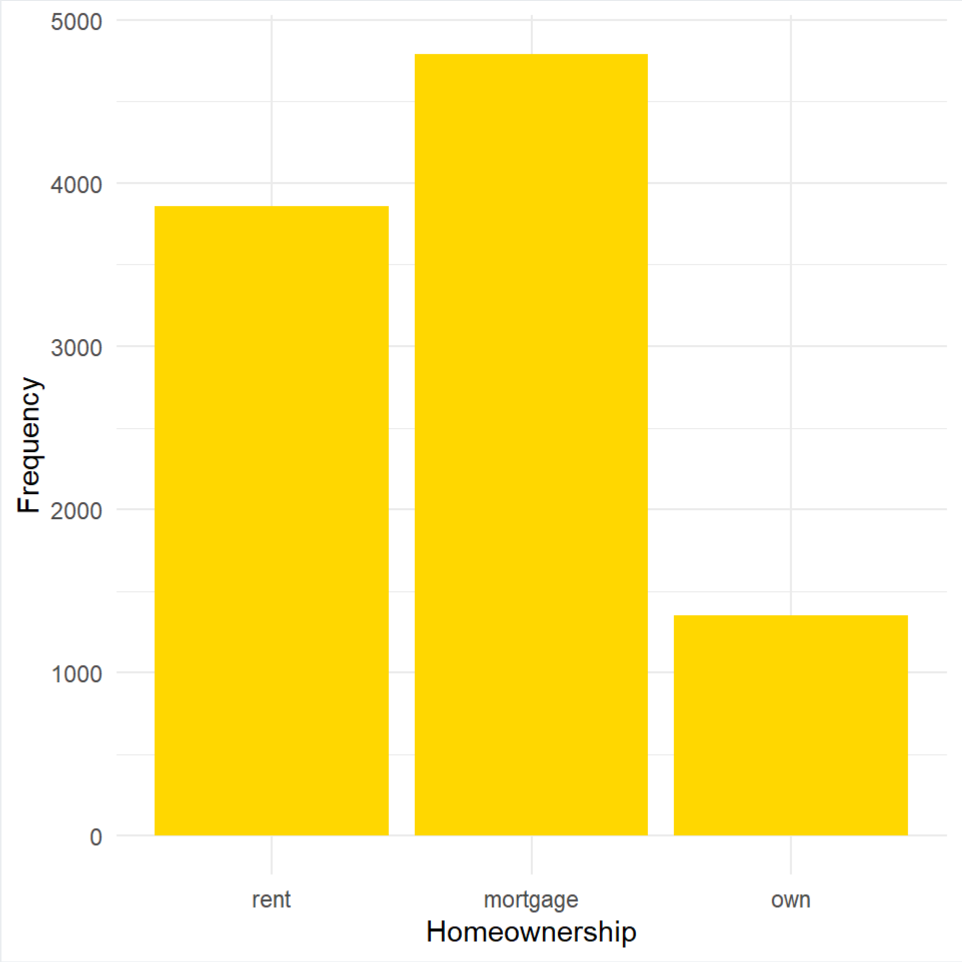


Gráfico de barras

Un gráfico de barras en el que se muestran proporciones en lugar de frecuencias se llama gráfico de barras de frecuencia relativa.

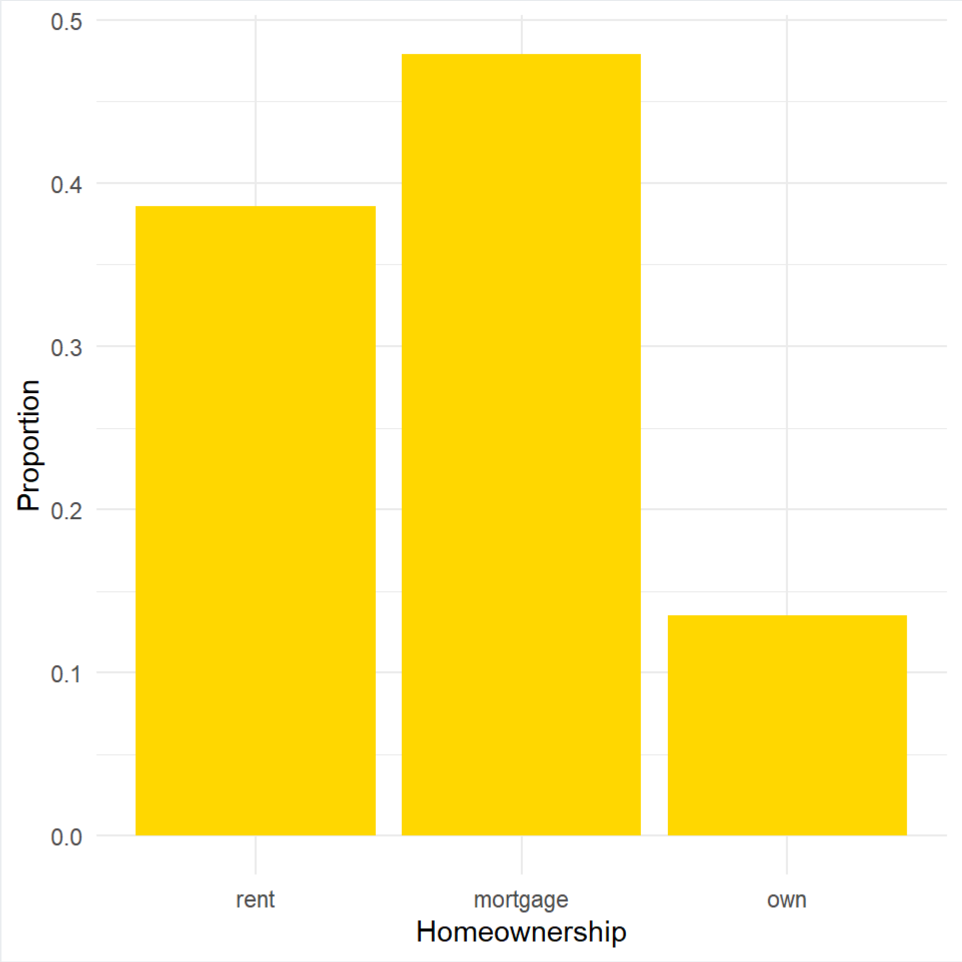


Gráfico de barras en R

```
ggplot(loans, aes(x = homeownershi
  geom_bar(fill = "gold") +
  labs(x = "Homeownership",
       y = "Frequency") +
  theme_minimal()
```

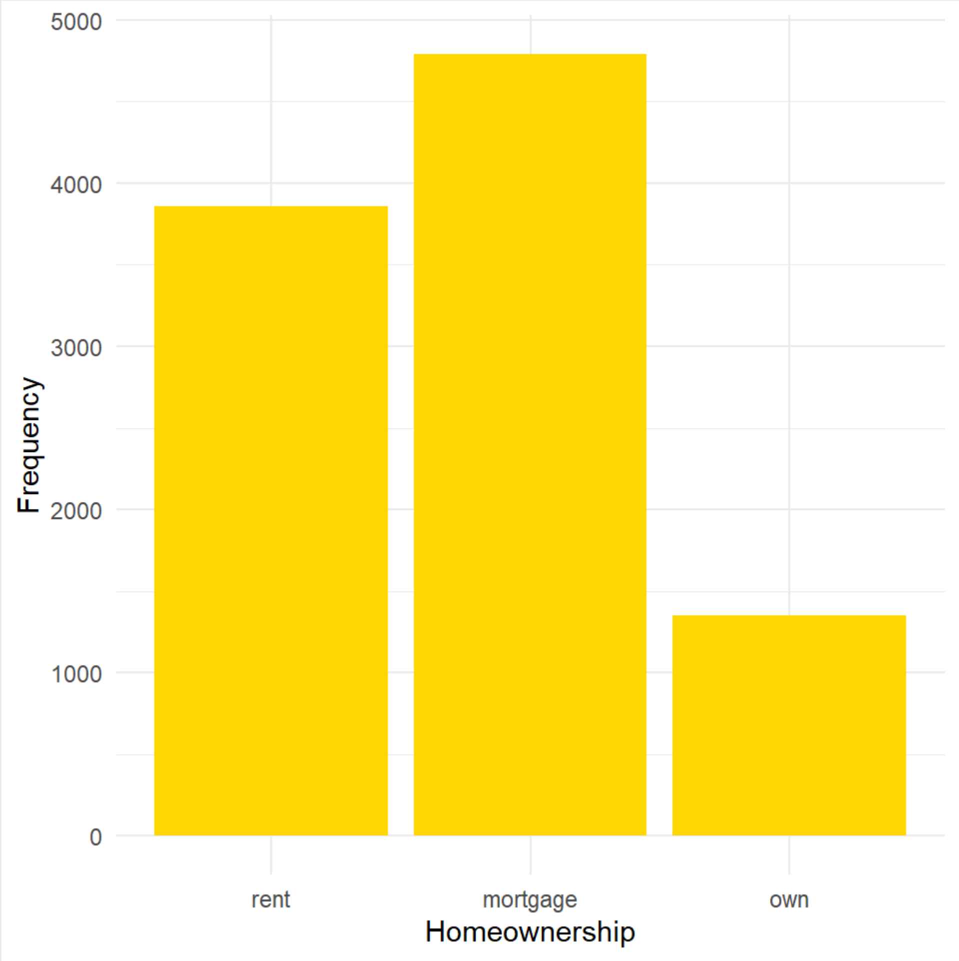


Gráfico de barras en R

```r
loans |>
  count(homeownership) |>
  mutate(proportion = n / sum(n)) |>
  ggplot(aes(x = homeownership,
             y = proportion)) +
  geom_col(fill = "gold") +
  labs(x = "Homeownership",
       y = "Proportion") +
  theme_minimal()
```

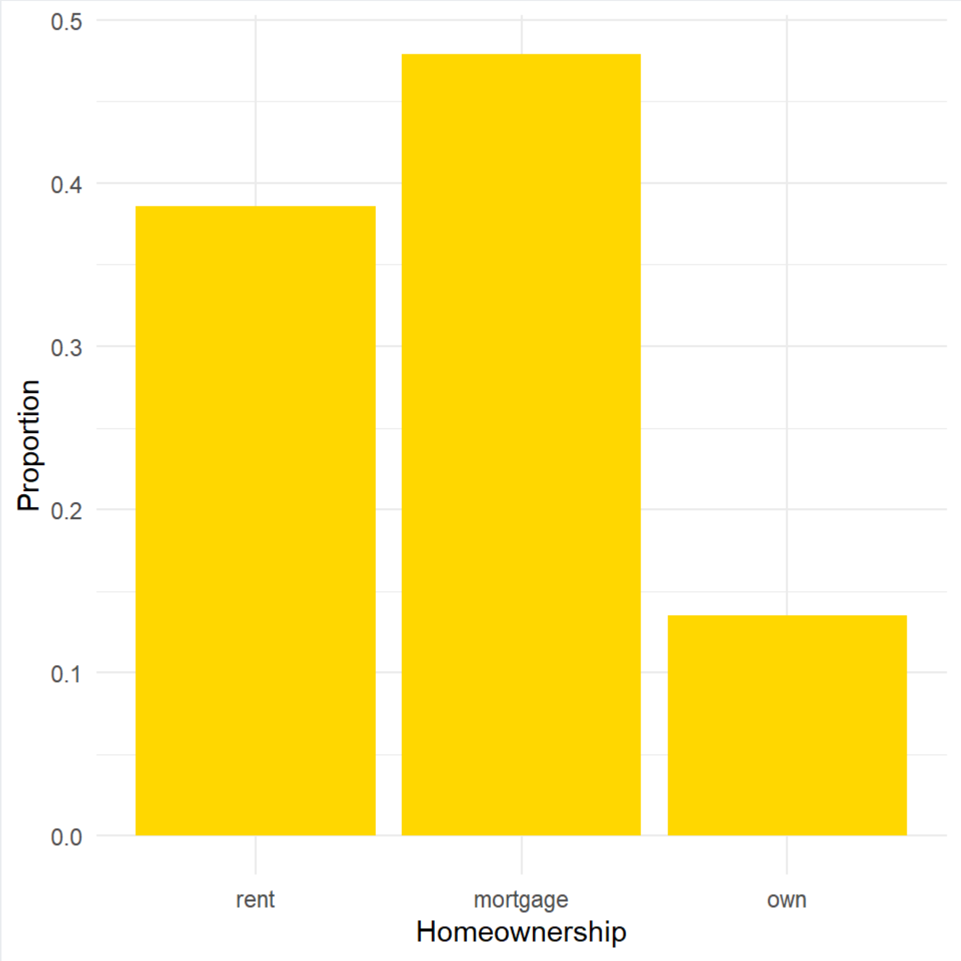
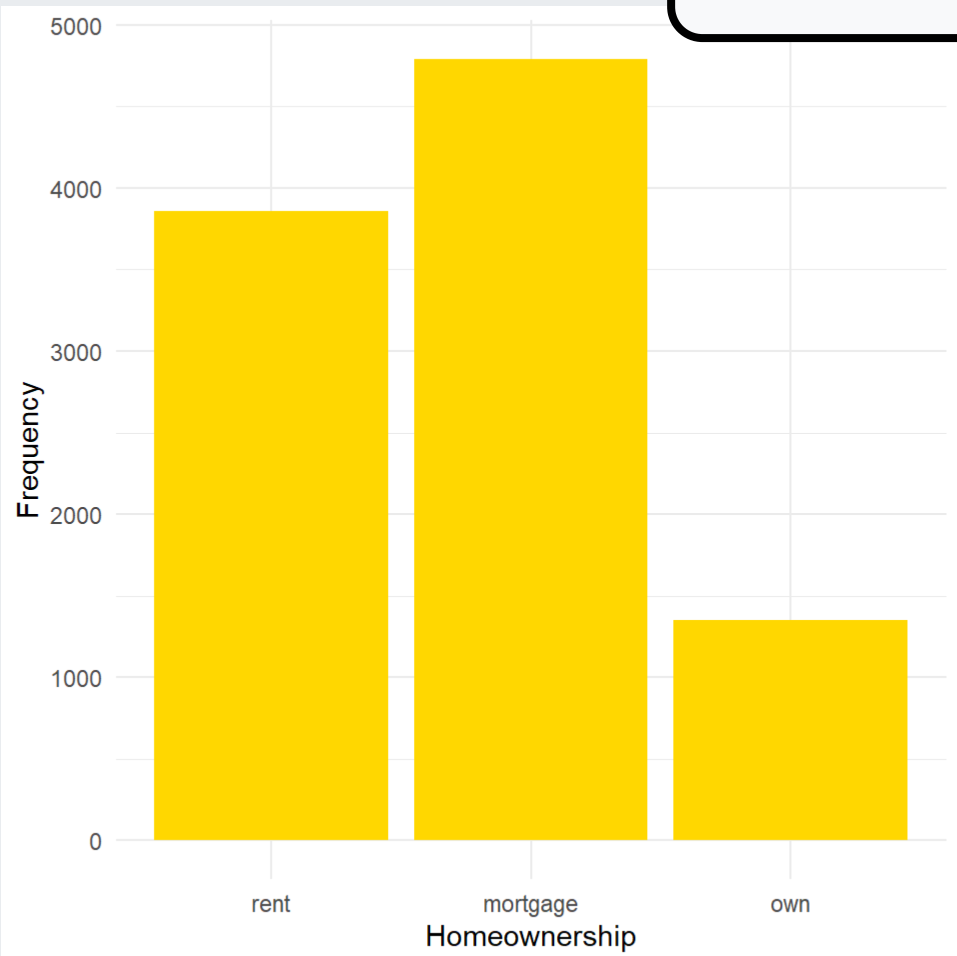
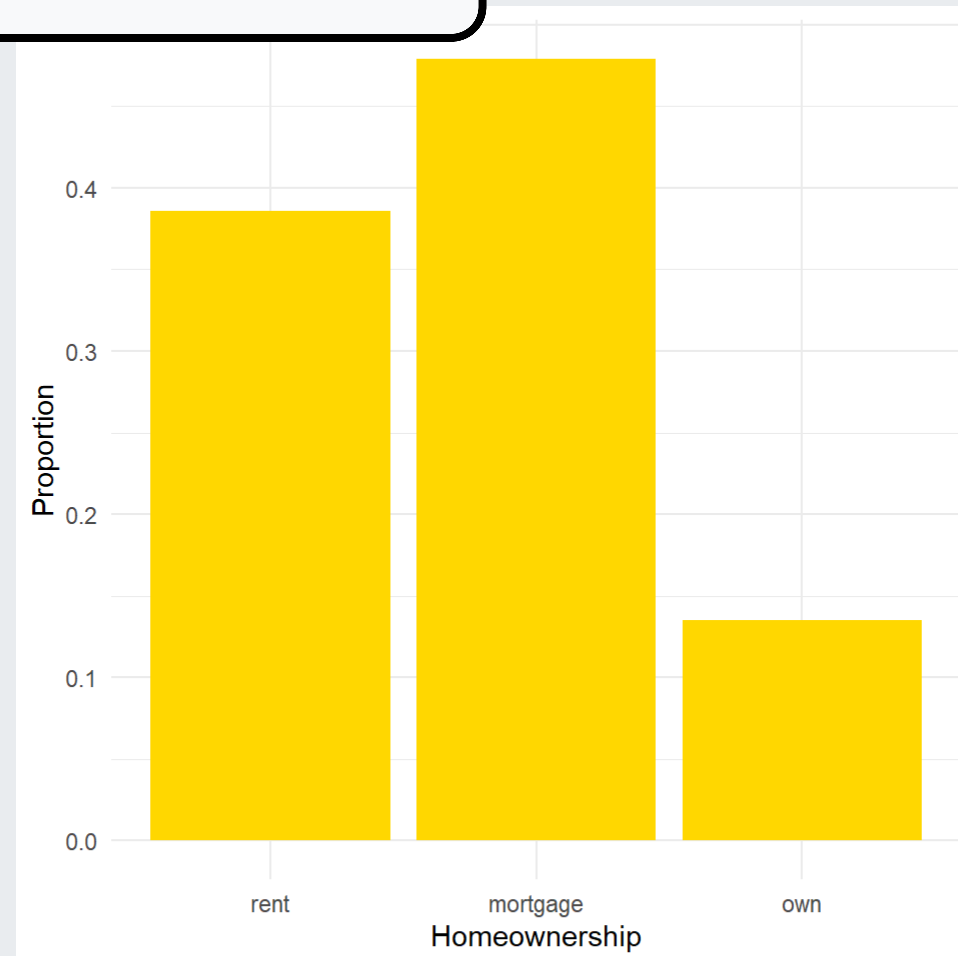


Gráfico de barras

¿Cómo sería el gráfico para una variable categórica ordinal?



Counts of homeownership.



Proportions of homeownership.

💪 Ejercicio 1

1. Usando la plantilla con la que ya hemos trabajado anteriormente, establezcan el directorio de trabajo y carguen los paquetes `tidyverse` y `janitor` (Este último instálenlo por lo que es la primera vez que lo usamos).
2. Importen los datos `credit_demographics` con el nombre `credit` usando la función `read.csv()`.
3. Exploren los datos usando la función `glimpse()`.
4. En ocasiones, algunos nombres de variables pueden ser inconsistentes o difíciles de manejar. El paquete `janitor` facilita este proceso. Ejecuten la siguiente línea de código y luego vuelvan a utilizar la función `glimpse()`. ¿Notan la diferencia en los nombres?

```
credit <- credit |>
  clean_names()
```

Ejercicio 1

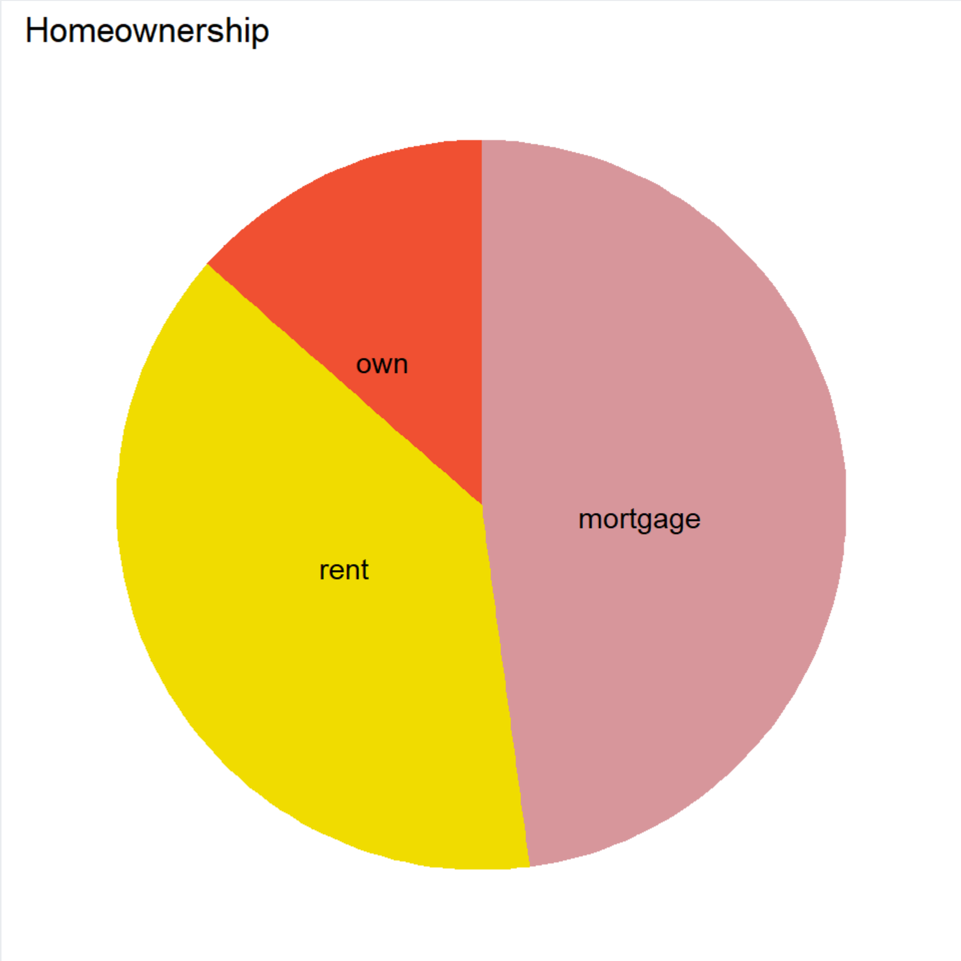
5. Como pueden notar en el punto anterior, la variable `default` contiene 0s y 1s. Vamos a convertirla en una variable de texto (`string`) para que sea más fácil de interpretar en los gráficos. Ejecuten el siguiente código para crear una nueva variable:

```r
credit <- credit |>
  mutate(
    default_string = case_when(
      default == 1 ~ "Default",
      default == 0 ~ "No Default",
      TRUE ~ NA_character_  # Assign NA for any unmatched values
    ))
```

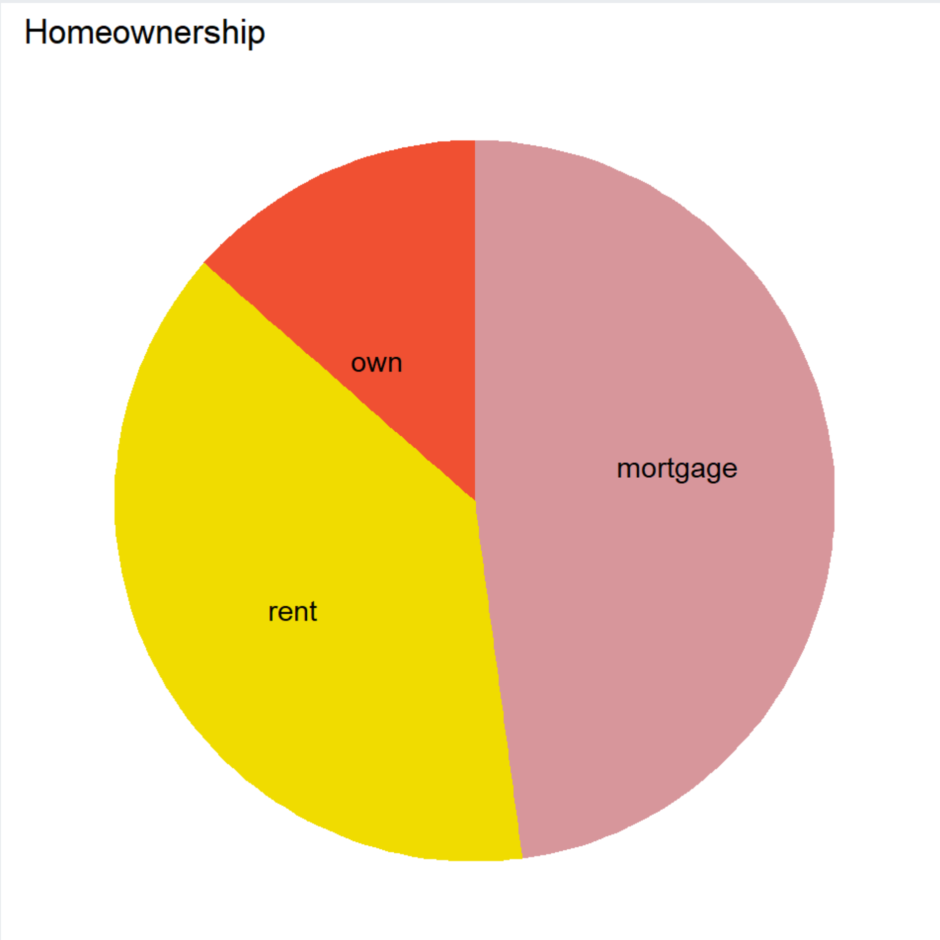
6. Usando esta nueva variable `default_string` y el paquete `ggplot`, construyan un gráfico de barras para visualizar cuántos clientes están en *default* y cuántos no. Asegúrense de incluir etiquetas y un título para hacer el gráfico más informativo.

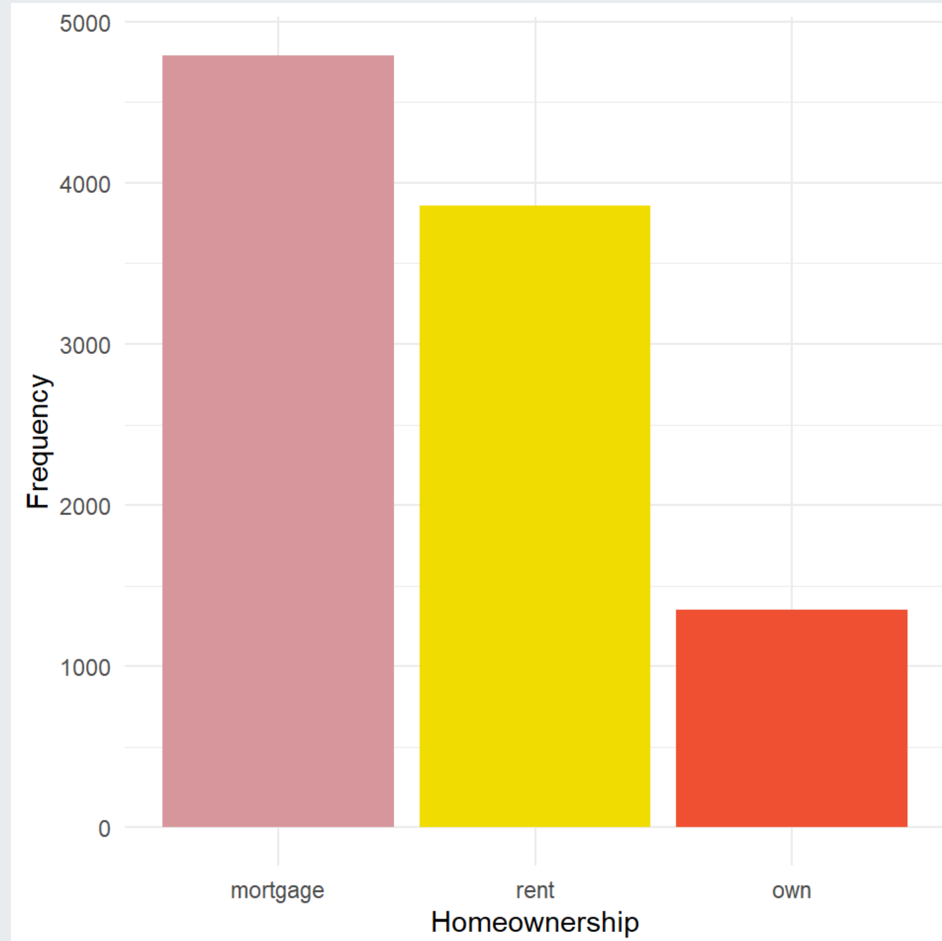
Gráficos de torta

- Las áreas de las porciones representan los porcentajes de las categorías
- Generalmente es más difícil comparar los tamaños de los grupos en un gráfico de pastel que en un gráfico de barras



Gráficos de torta



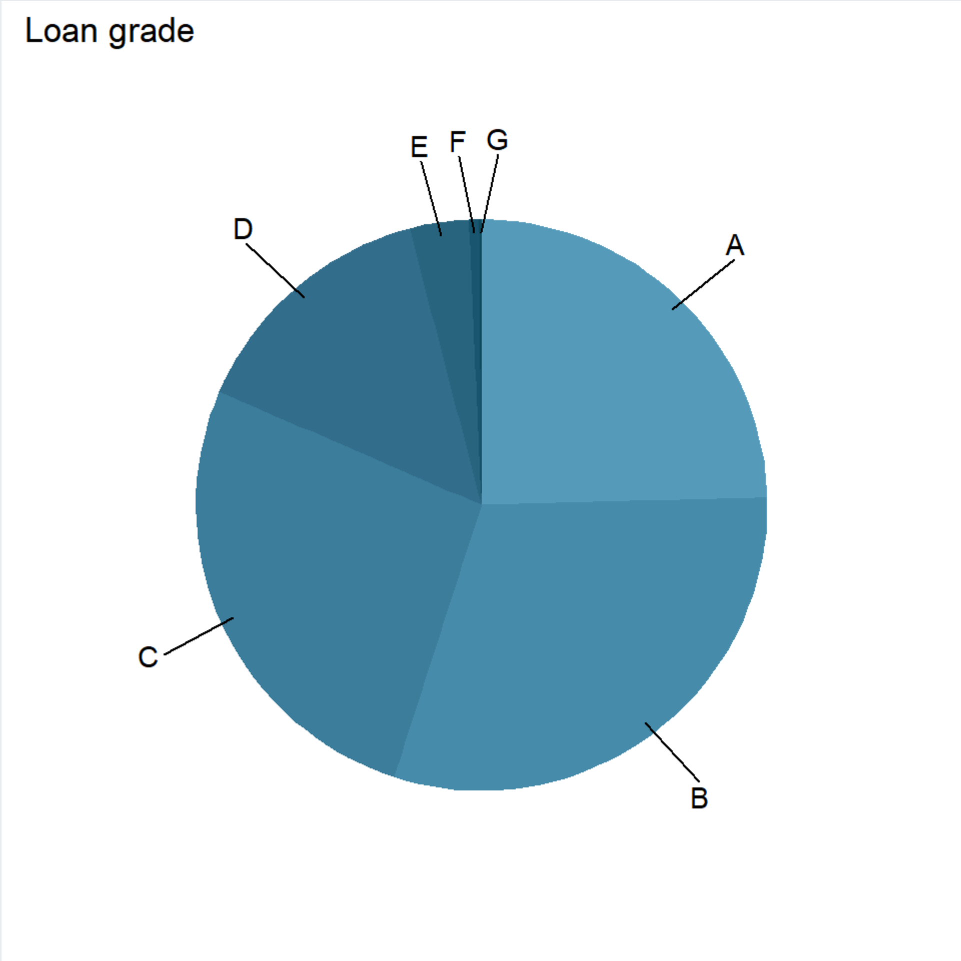


Pie chart

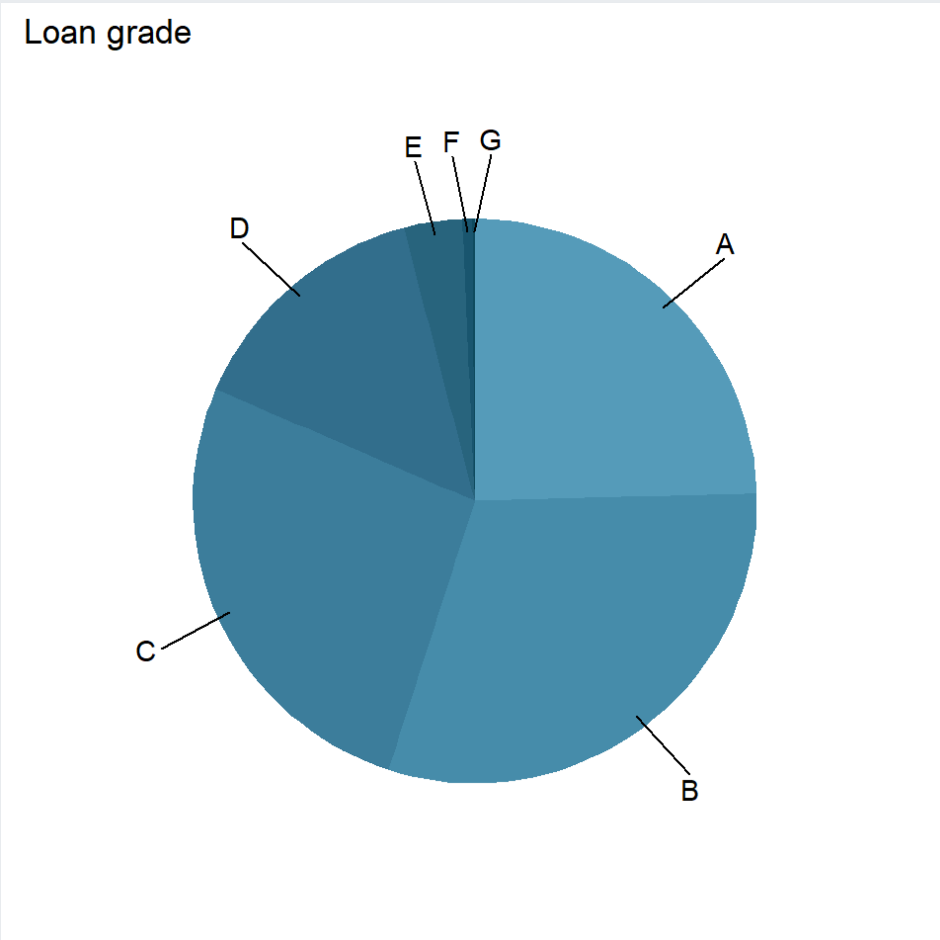
Bar plot

Gráficos de torta

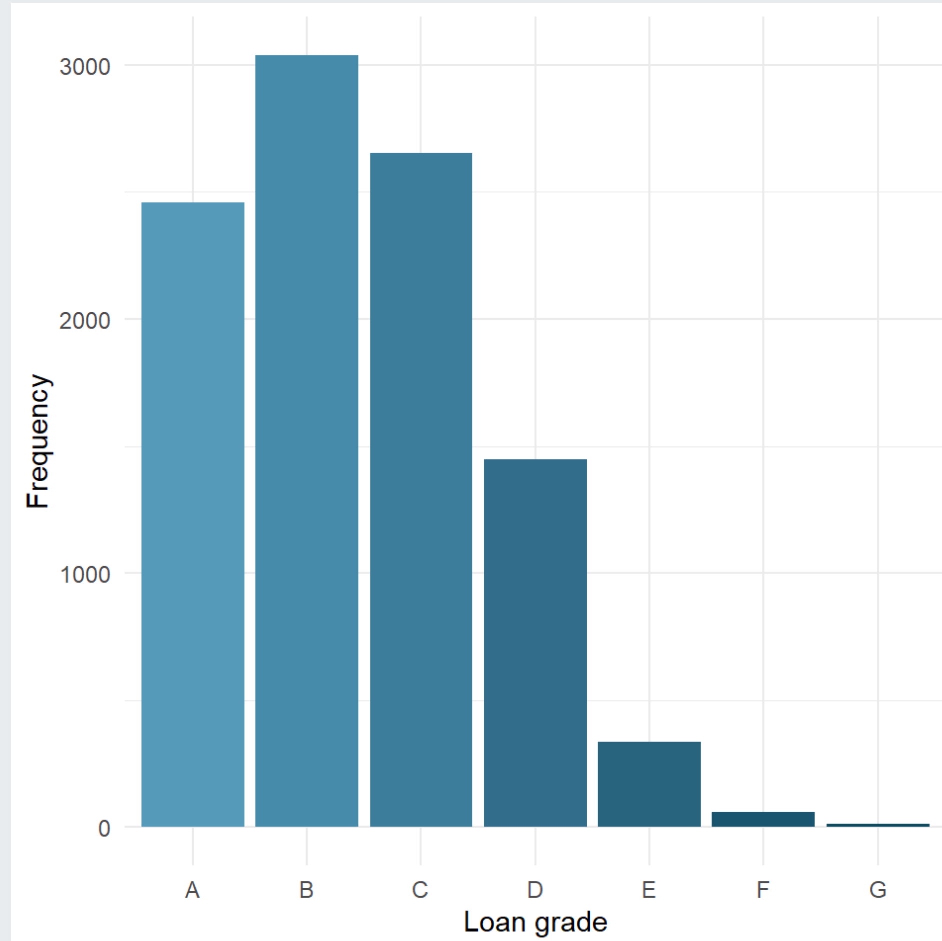
- Es mucho más fácil hacer un gráfico de pastel incorrecto que un gráfico de barras incorrecto.
- En un gráfico de pastel, las categorías deben representar un todo. No existe esta restricción para un gráfico de barras.



Gráficos de torta



Pie chart

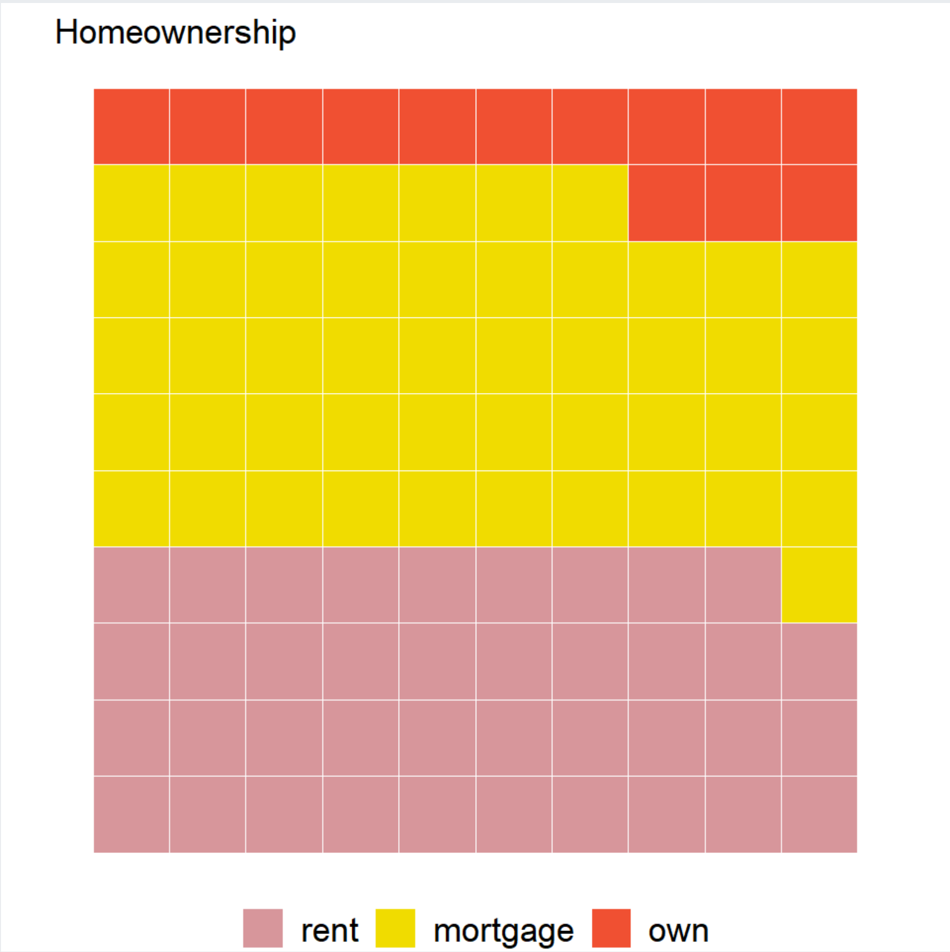


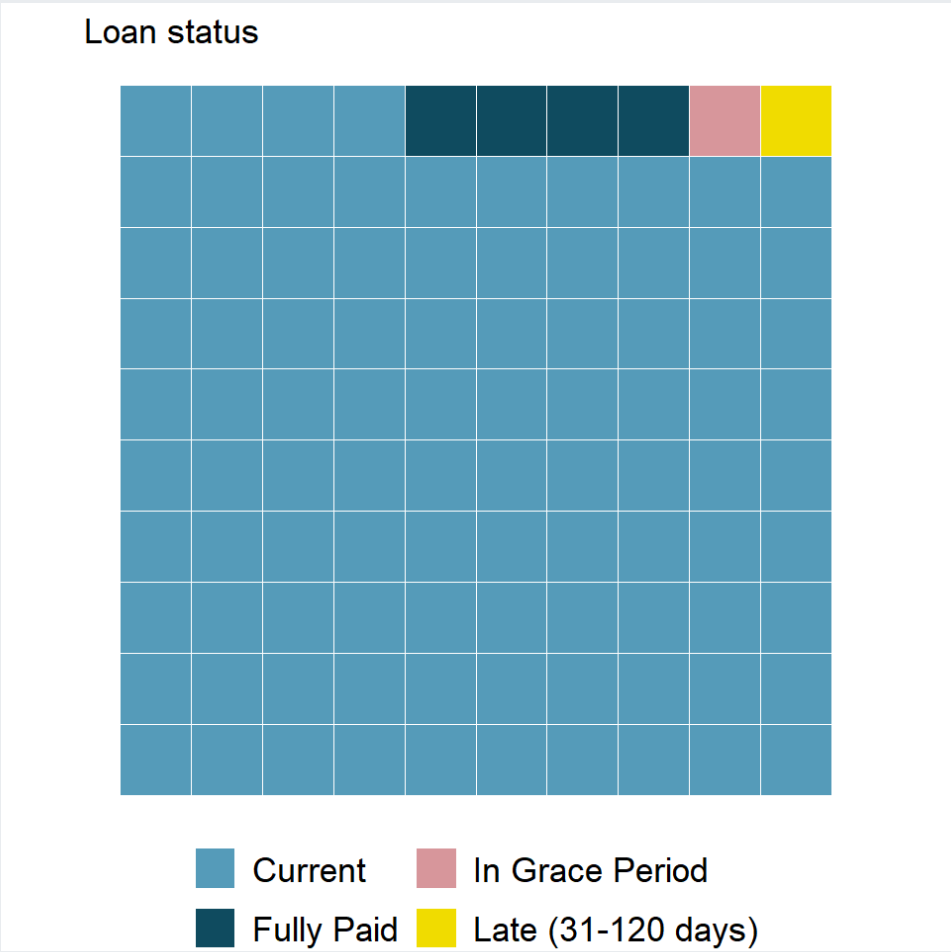
Bar plot

Gráfico de torta en R

- Existen diferentes maneras de hacer un gráfico de torta, más allá de `ggplot`
- En Pie Charts encuentran una guía explicando diferentes maneras de hacerlo en R

Gráfico de Waffle





Homeownership: rent, mortgage, and own

Loan status: fully paid, in grace period, and late

Gráfico de Waffle

- Los gráficos de waffle son otra técnica útil para visualizar datos categóricos, mostrando la proporción de cada categoría
- Al igual que los gráficos de pastel, funcionan mejor cuando el número de categorías es bajo
- A diferencia de los gráficos de pastel, facilitan la comparación de proporciones que no representan fracciones simples

Gráfico de Waffle en R

- Este tipo de gráfico va más allá de la funcionalidad de `ggplot`
- Por lo tanto no lo explicaré en clase, pero acá les dejo recursos para que aprendan por su cuenta:
  1. La página del paquete waffle
  2. Waffle Charts provee una guía de cómo crear este tipo de gráfico

Visualizando dos variables categóricas

Tablas de Contingencia

- Una tabla que resume datos para dos variables categóricas de esta manera se llama tabla de contingencia
- Cada valor en la tabla representa la cantidad de veces que ocurrió una combinación particular de resultados de las variables

	homeownership			
application_type	**rent**	**mortgage**	**own**	**Total**
joint	362	950	183	1495
individual	3496	3839	1170	8505
Total	3858	4789	1353	10000

Gráfico de barras apiladas

- Los solicitantes de préstamos viven más comúnmente en viviendas con hipoteca
- Sin embargo, basándose solo en este gráfico, es difícil determinar cómo varían los tipos de solicitud entre los niveles de tenencia de vivienda

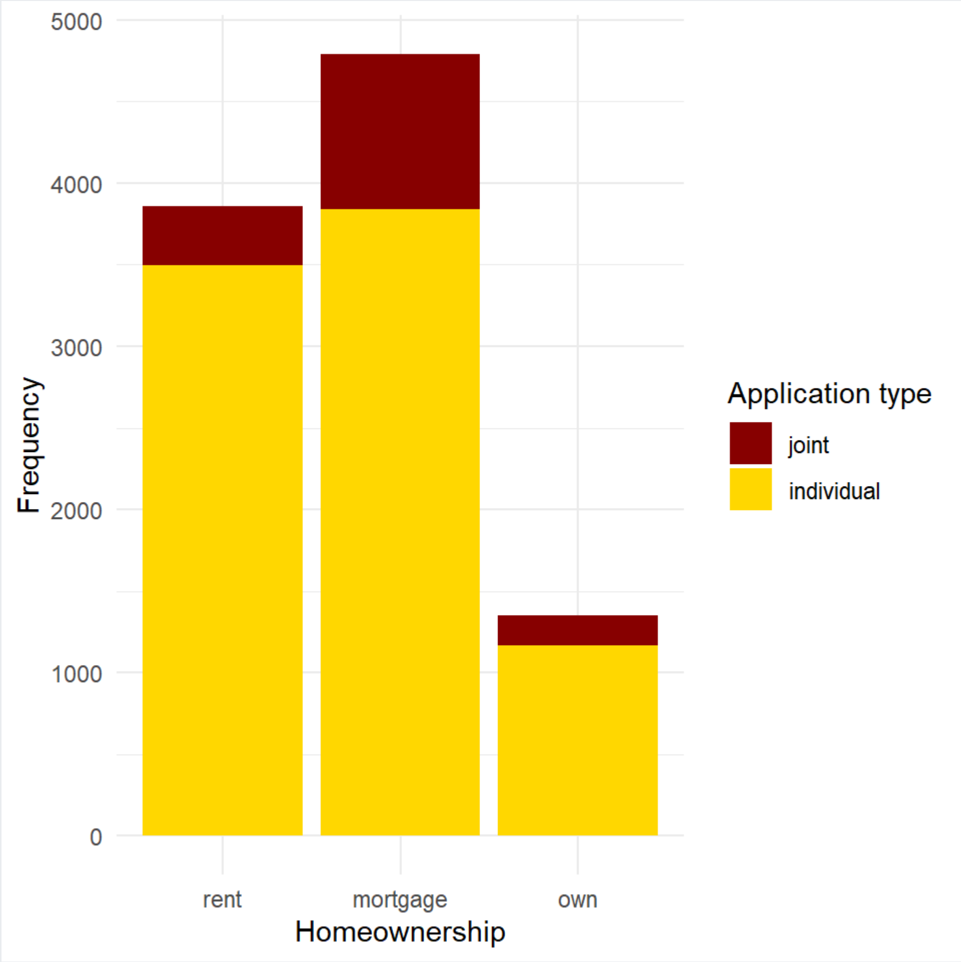


Gráfico de barras estandarizado

- Este tipo de visualización es útil para comprender la proporción del tipo de solicitudes en cada nivel de tenencia de vivienda
- Además, dado que las proporciones del tipo de préstamos varían entre los grupos, podemos concluir que estas dos variables están asociadas en esta muestra

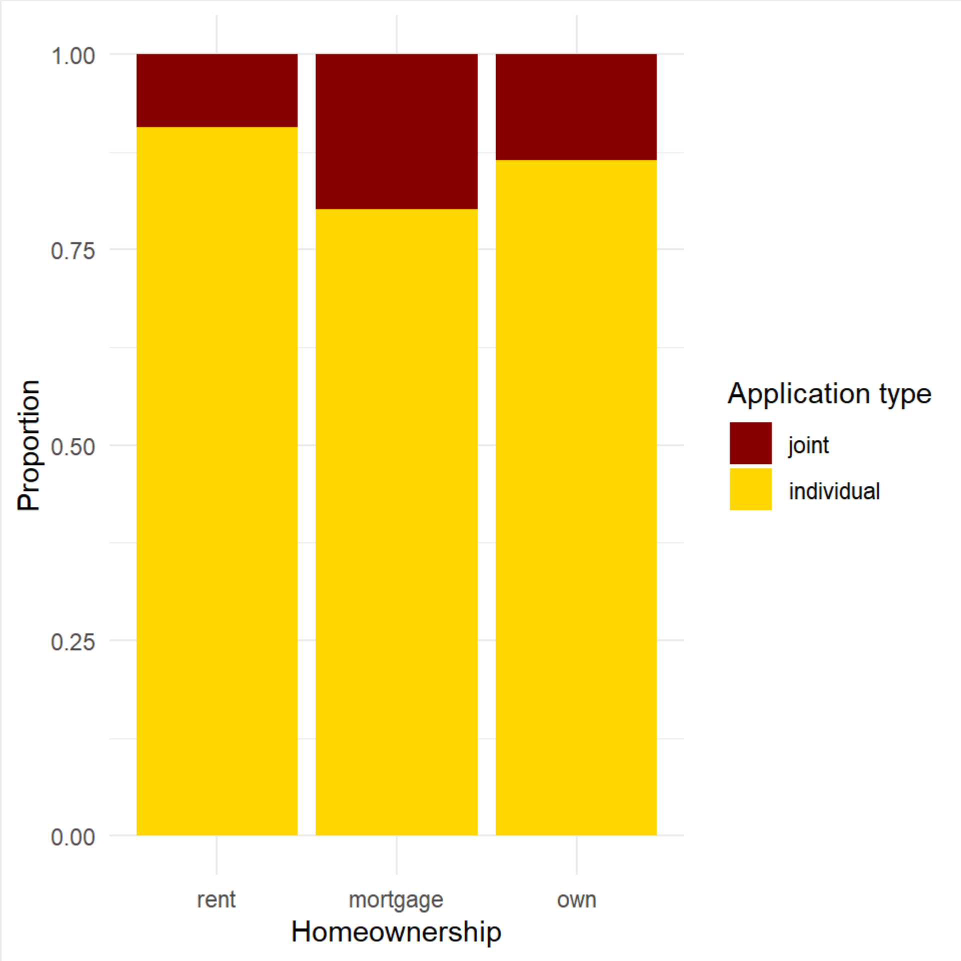


Gráfico de barras dobles

- Dentro de cada nivel de tenencia de vivienda, las solicitudes individuales son más comunes que las solicitudes conjuntas
- Las solicitudes conjuntas son más comunes entre los solicitantes con hipoteca, en comparación con los inquilinos y los propietarios.

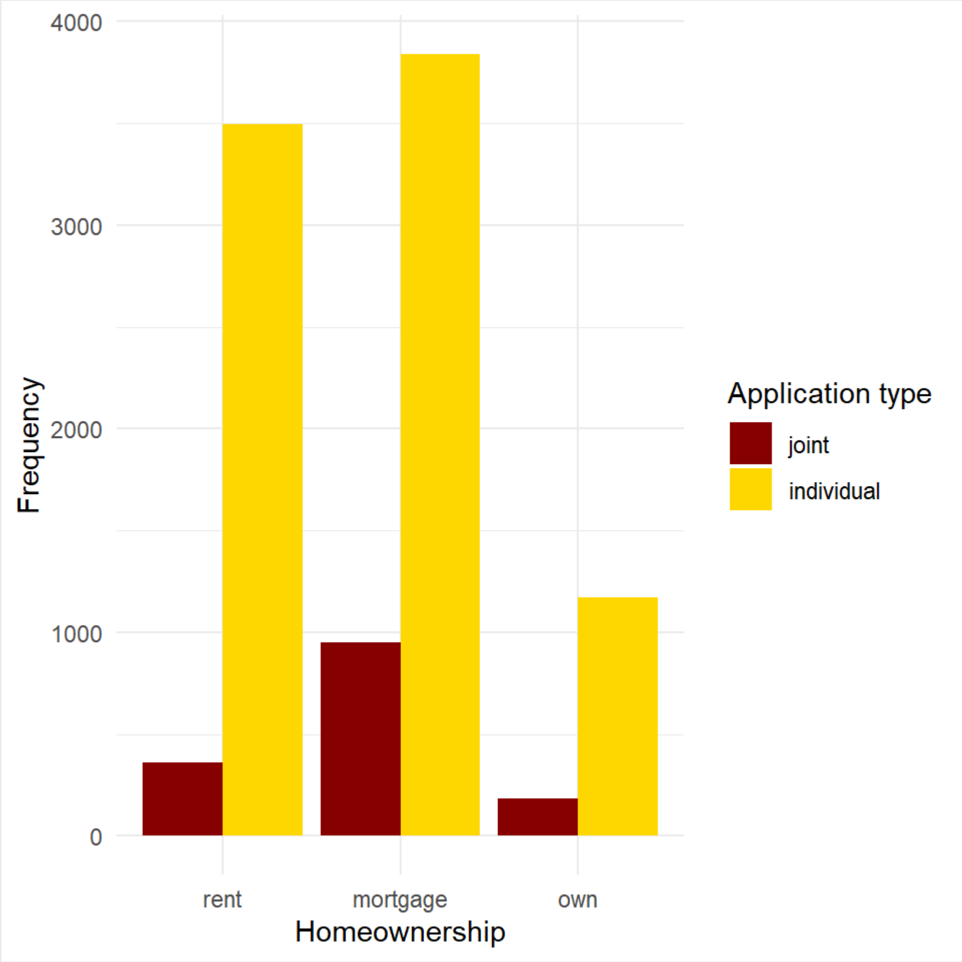


Gráfico de barras apiladas en R

```
ggplot(loans, aes(x = homeownershi
                  fill = applicati
  geom_bar() +
  scale_fill_manual(values = c("da
                               "go
  labs(x = "Homeownership",
       y = "Frequency",
       fill = "Application type")
  theme_minimal()
```

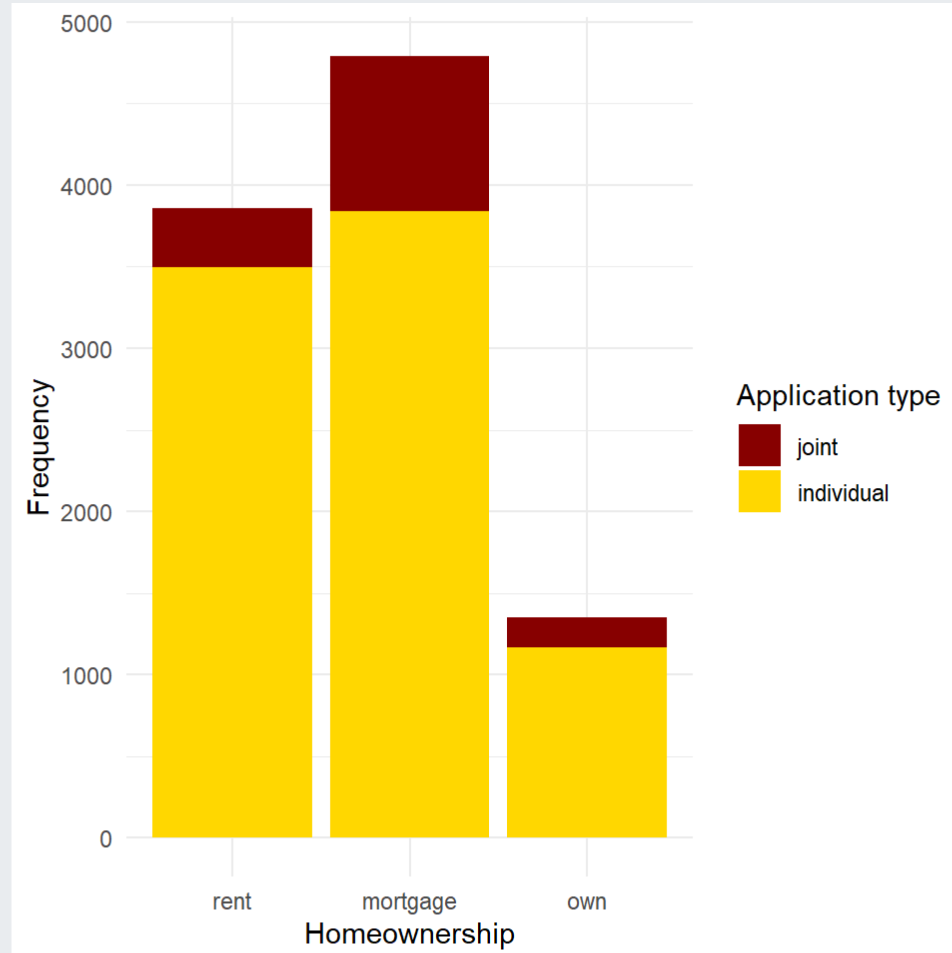


Gráfico de barras estandarizado en R

```
ggplot(loans, aes(x = homeownershi
                    fill = applicati
  geom_bar(position = "fill") +
  scale_fill_manual(values = c("da
                               "go
  labs(x = "Homeownership",
       y = "Proportion",
       fill = "Application type")
  theme_minimal()
```

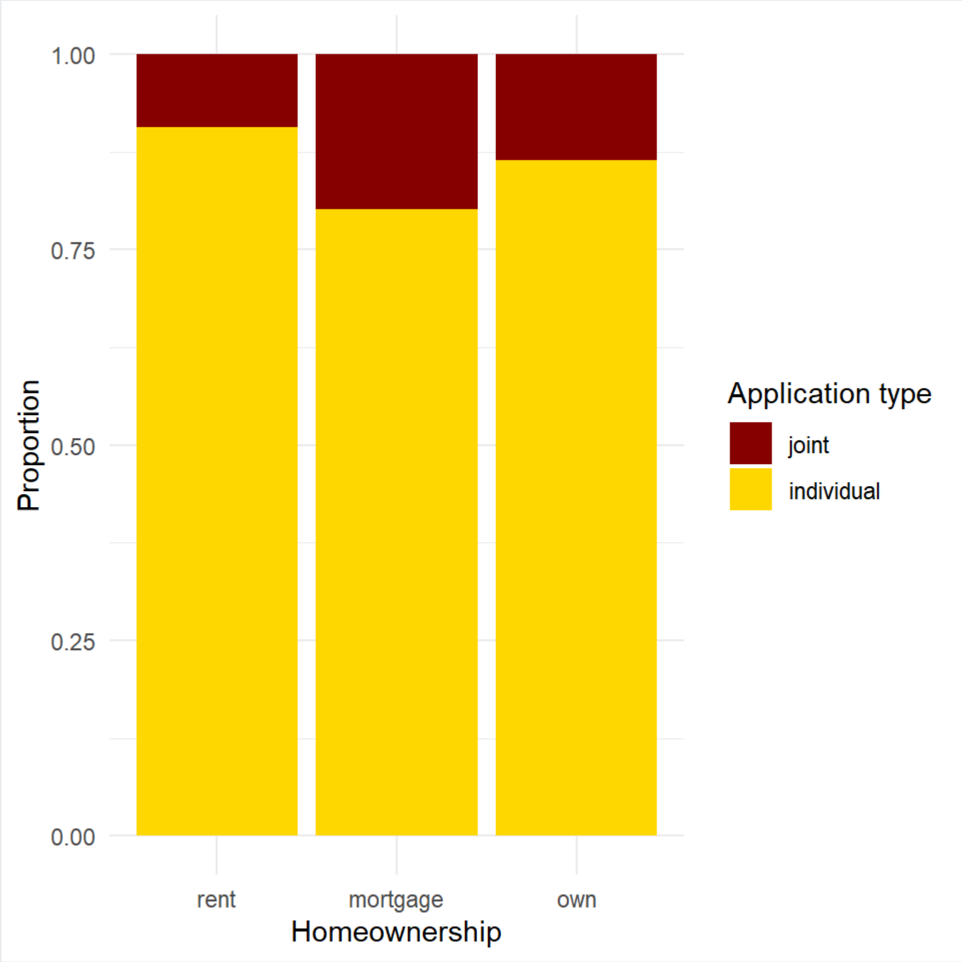
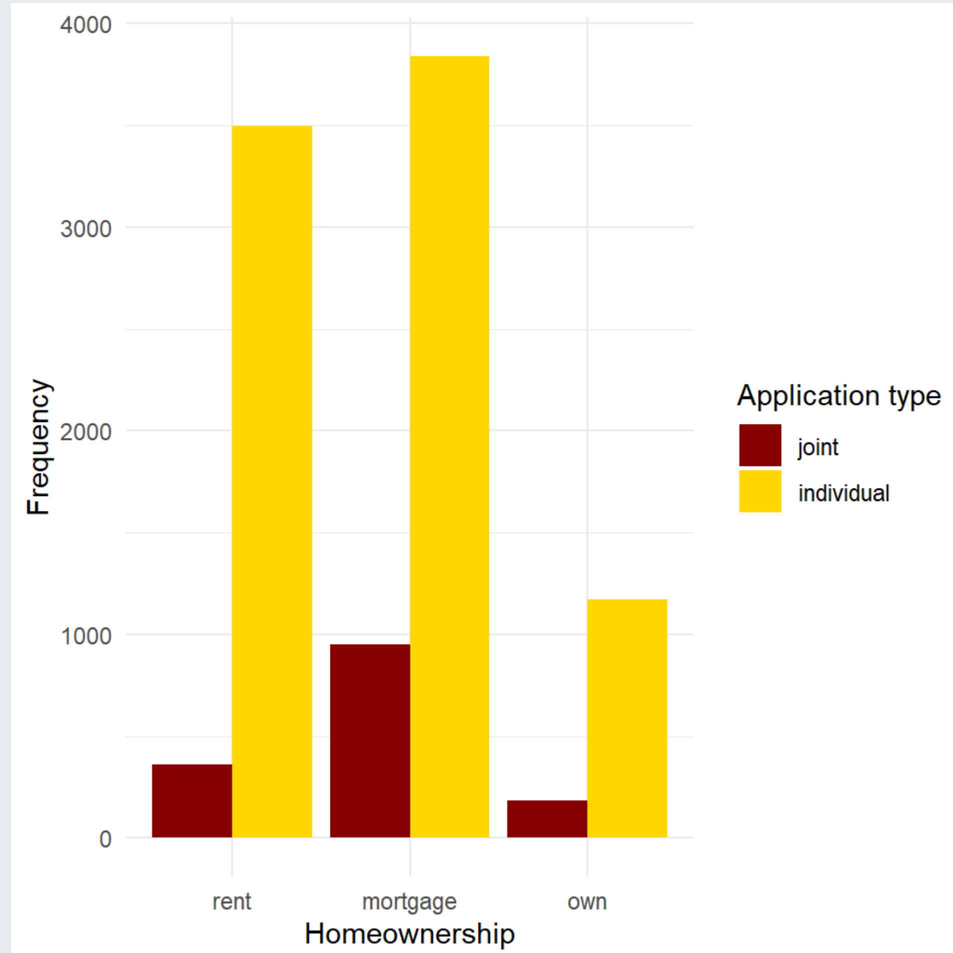
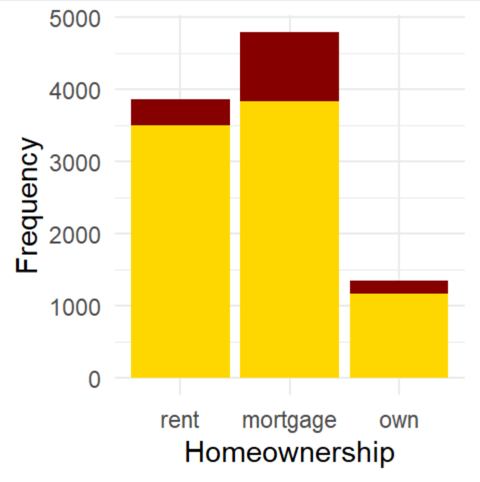


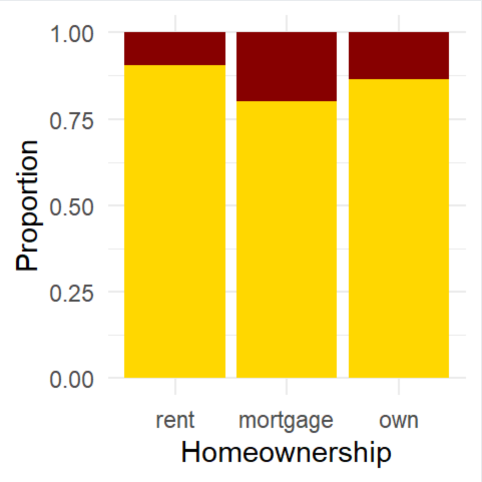
Gráfico de barras dobles en R

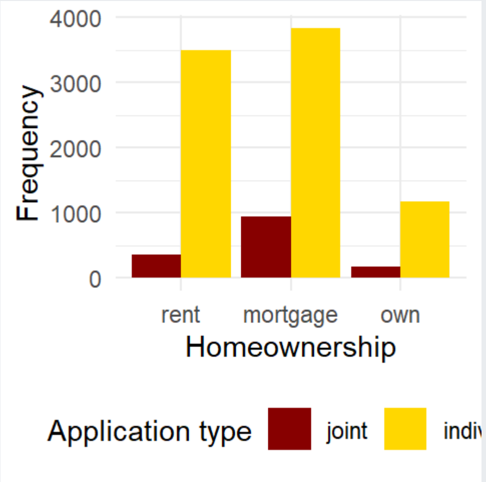
```
ggplot(loans, aes(x = homeownershi
                        fill = applicati
  geom_bar(position = "dodge") +
  scale_fill_manual(values = c("da
                               "go
  labs(x = "Homeownership",
       y = "Frequency",
       fill = "Application type")
  theme_minimal()
```



Explorando dos variables categóricas





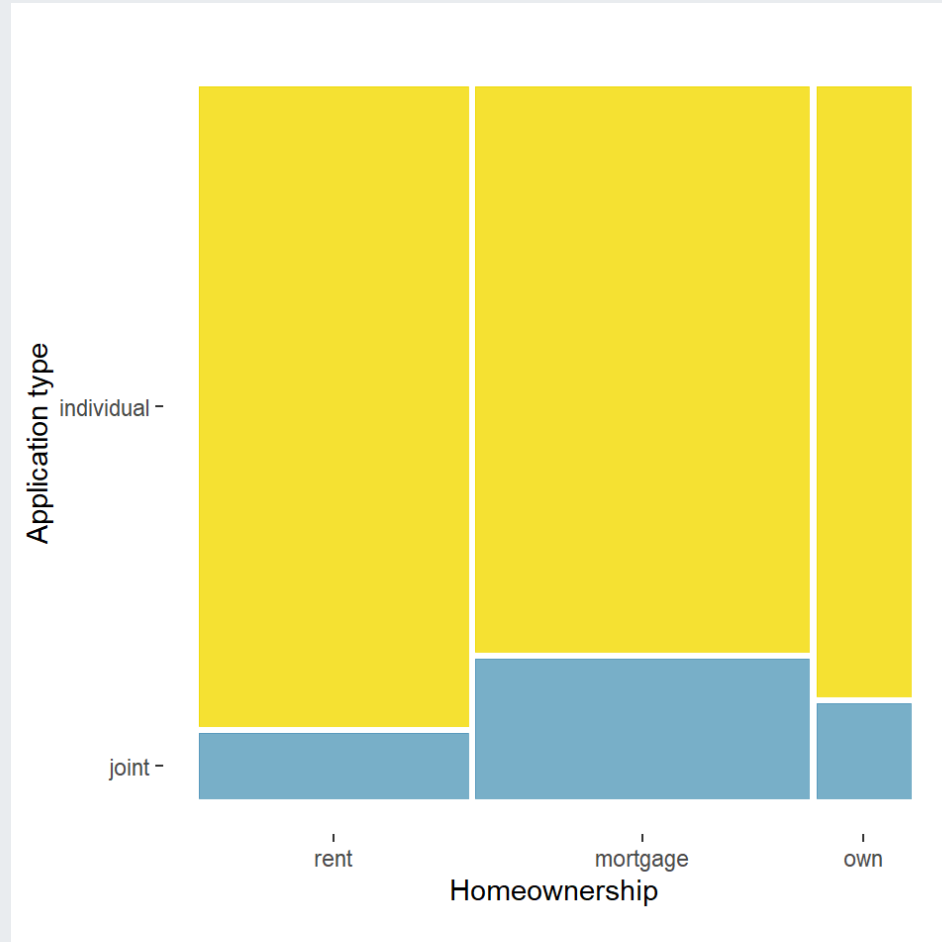


💪 Ejercicio 2

1. Con base en el código con el que crearon la variable `default_string`, generen la variable `marriage_string` según los valores de la variable `marriage`:
   - 1 $\longrightarrow$ *married*
   - 2 $\longrightarrow$ *single*
   - 3 $\longrightarrow$ *other*
2. Creen un gráfico de barras (apilados, dobles o estandarizado) para visualizar la relación entre el estado civil (`marriage_string`) y si el cliente está en default (`default_string`).
3. ¿Qué pueden concluir a partir del gráfico anterior?

Gráfico de mosaico

Homeownership.



Homeownership vs. application type.

Gráfico de mosaico

- Un gráfico mosaico es otra manera de visualizar tablas de contingencia que se asemeja a un gráfico de barras apiladas estandarizado
- La ventaja consiste en aún poder ver el tamaño relativo de los grupos de la variable principal

$$
\begin{aligned}
\text{Area del segmento} &= (\text{ancho de la barra}) \times (\text{longitud del segmento}) \\
&= \text{frecuencia de la fila} \times (\text{proporcion de la fila}) \\
&= \text{frecuencia de la fila} \times \frac{\text{frecuencia de la celda}}{\text{frecuencia de la fila}} \\
&= \text{frecuencia de la celda}
\end{aligned}
$$

Gráfico de mosaico

Es importante pensar cuál variable va en el eje horizontal y cuál en el vertical. En ocasiones, una es más *explicativa* que la otra.

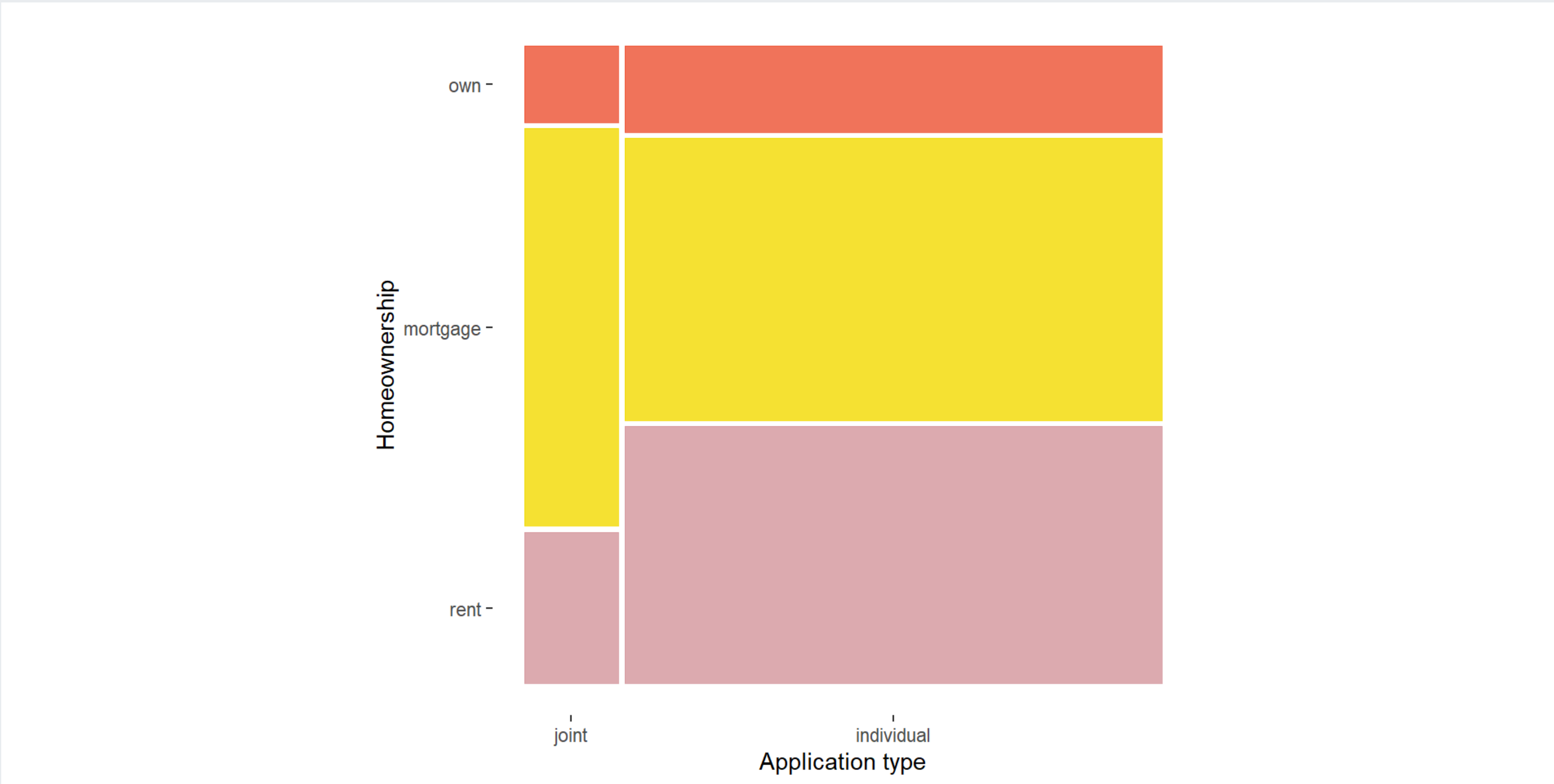


Gráfico de Mosaico en R

- Este tipo de gráfico también va más allá de la funcionalidad de `ggplot`
- La página del paquete ggmosaic es un buen sitio para empezar a practicar por su cuenta

Mosaic plots with `ggplot2`

Haley Jeppson and Heike Hofmann

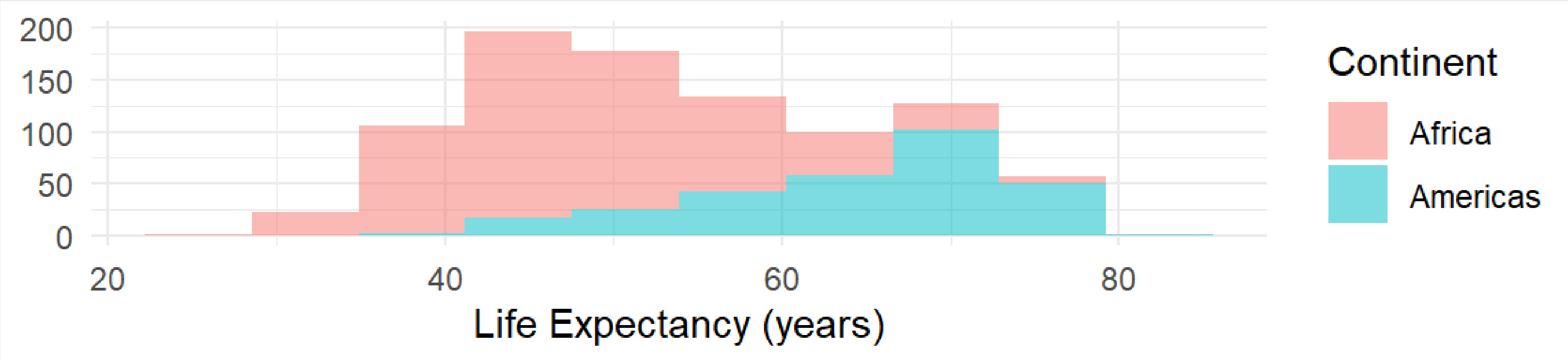
2024-09-30

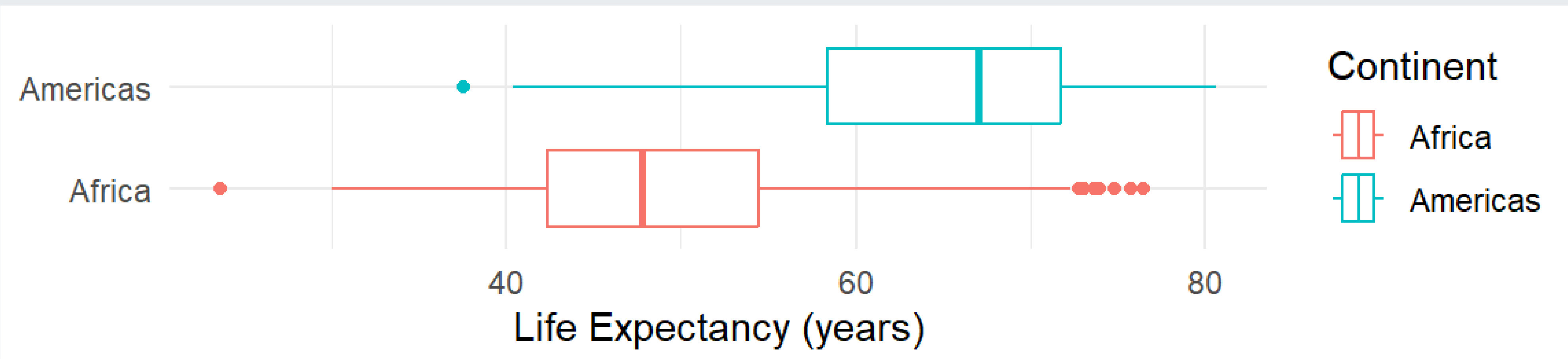
Source: `vignettes/ggmosaic.Rmd`
(https://github.com/haleyjeppson/ggmosaic/blob/master/vignettes/ggmosaic.Rmd)

Designed to create visualizations of categorical data, `geom_mosaic()` has the capability to produce bar charts, stacked bar charts, mosaic plots, and double decker plots and therefore offers a wide range of potential plots. The plots below highlight the package's versatility.

Comparando variables numéricas entre grupos

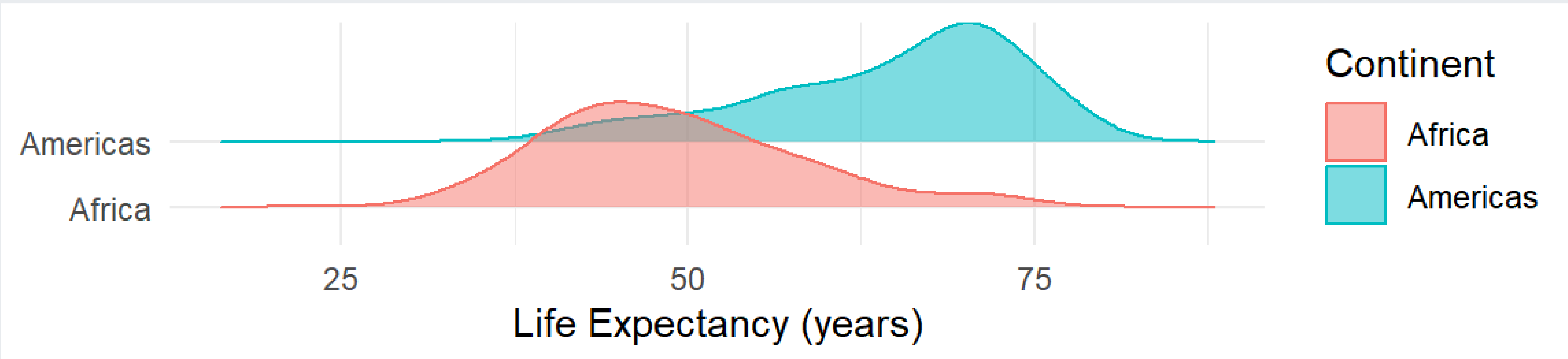
Histograma y Diagrama de Caja entre grupos





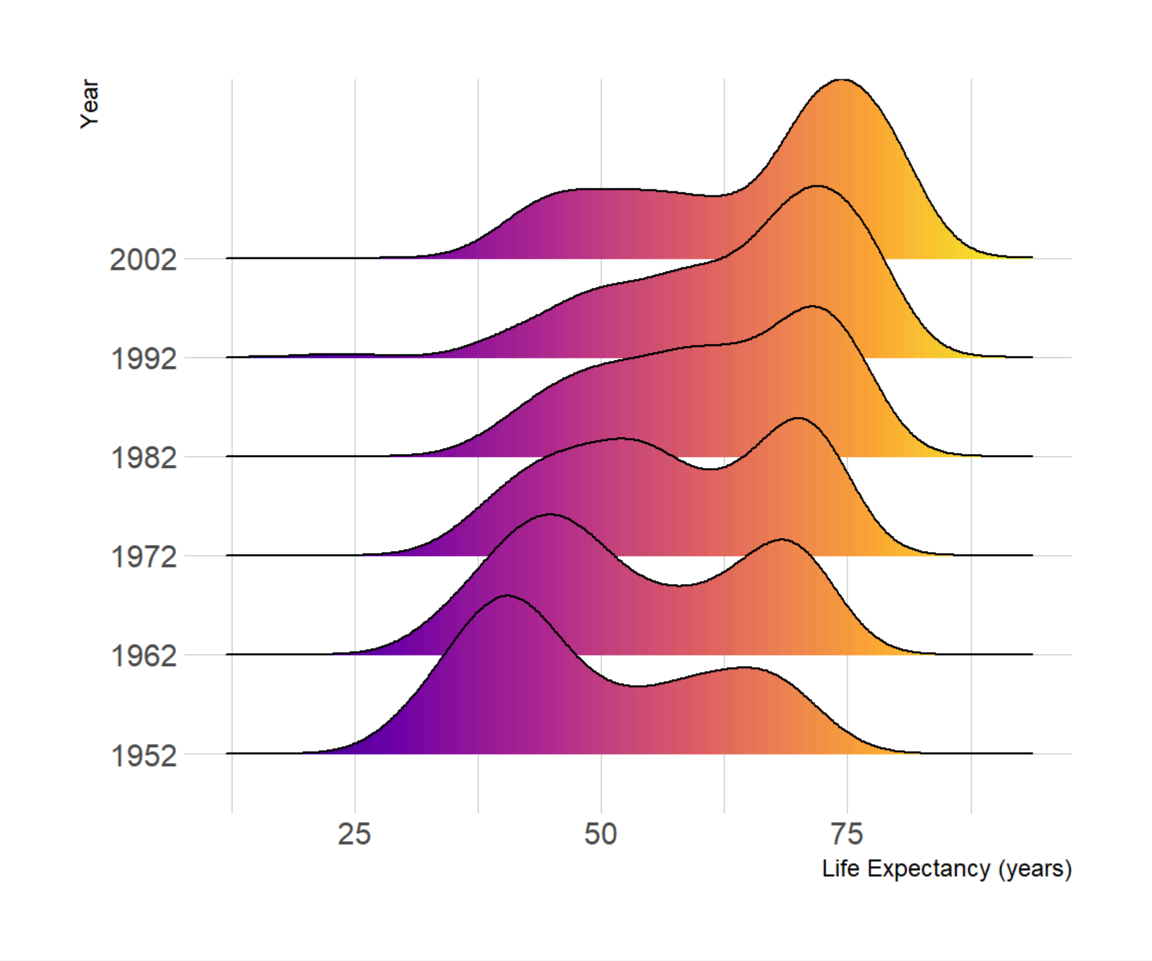
Ridge plot

Otro tipo de visualización útil para comparar datos numéricos entre grupos es el gráfico de crestas (ridge plot), que combina gráficos de densidad de varios grupos en la misma escala dentro de una única ventana de visualización



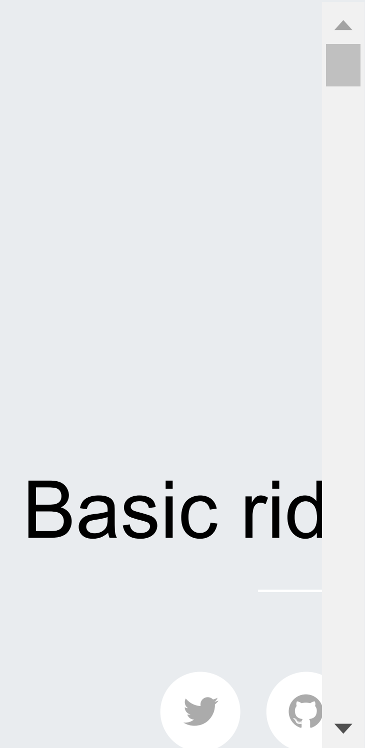
Ridge plot

Establecer el color según la variable numérica en lugar de la categórica puede ser bastante útil para contar la historia de los datos.



Ridge plot en R

- Quiero destacar la flexibilidad de R para mejorar con la incorporación de paquetes
- La página del paquete ggridges es un buen sitio para profundizar en este tipo de gráfico
- Basic ridgeline plot también explica este gráfico y tiene buenos ejemplos



Maneras de visualizar relaciones entre variables

- numérica v.s. numérica
  - Diagramas de dispersión
- categórica v.s. categórica
  - Tablas de contigencia
  - Gráficos de barra (apilados, dobles, estandarizados)
  - Gráfico de mosaico
- categórica v.s. numérica
  - Diagramas de caja entre grupos
  - Ridge plots

Probabilidad

Probabilidad

- Una probabilidad se define como la siguiente proporción:

$$P = \frac{\#\,\text{resultados deseados}}{\#\,\text{resultados posibles}}$$

- Por ejemplo, al tirar un dado la probabilidad de obtener un 2 luego de lanzar un dado es:

$$P(2) = \frac{1}{6} = 0.166 = 16.66\%$$

Probabilidad

1. Las probabilidades siempre están entre 0 y 1.

- Una probabilidad igual a 0 indica que el evento nunca va a ocurrir.
- Por otro lado, si es igual a 1 indica que con toda seguridad el evento tendrá lugar.

2. $\sum P = 1$
3. La probabilidad de que un evento **no ocurra** es igual a 1 menos la probabilidad de que el evento ocurra.

- Al tirar un dado:

$$P(\sim 2) = 1 - P(2) = 1 - \frac{1}{6} = \frac{5}{6}$$

Probabilidad

4. Si A y B son eventos alternativos (no se superponen), entonces $P(A\,o\,B) = P(A) + P(B)$

- Siguiendo con el ejemplo del dado:

$$P(2\,o\,3) = P(2) + P(3) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Probabilidad

5. Si A y B son eventos que se superponen (ocurrencia conjunta), entonces $P(A\,o\,B) = P(A) + P(B) - P(A\,y\,B)$

- ¿Cuál sería la probabilidad de sacar un número par o un 6?

$$P(Par\,o\,6) = P(Par) + P(6) = \frac{3}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3} \quad \text{Incorrecto}$$

$$P(Par\,o\,6) = P(Par) + P(6) - P(Par\,y\,6) = \frac{3}{6} + \frac{1}{6} - \frac{1}{6} = \frac{3}{6} =$$

Probabilidad

6. Si A y B son independientes, entonces $P(A\,y\,B) = P(A) \cdot P(B)$

- ¿Cuál es la probabilidad de sacar 2 luego de tirar el dado dos veces?

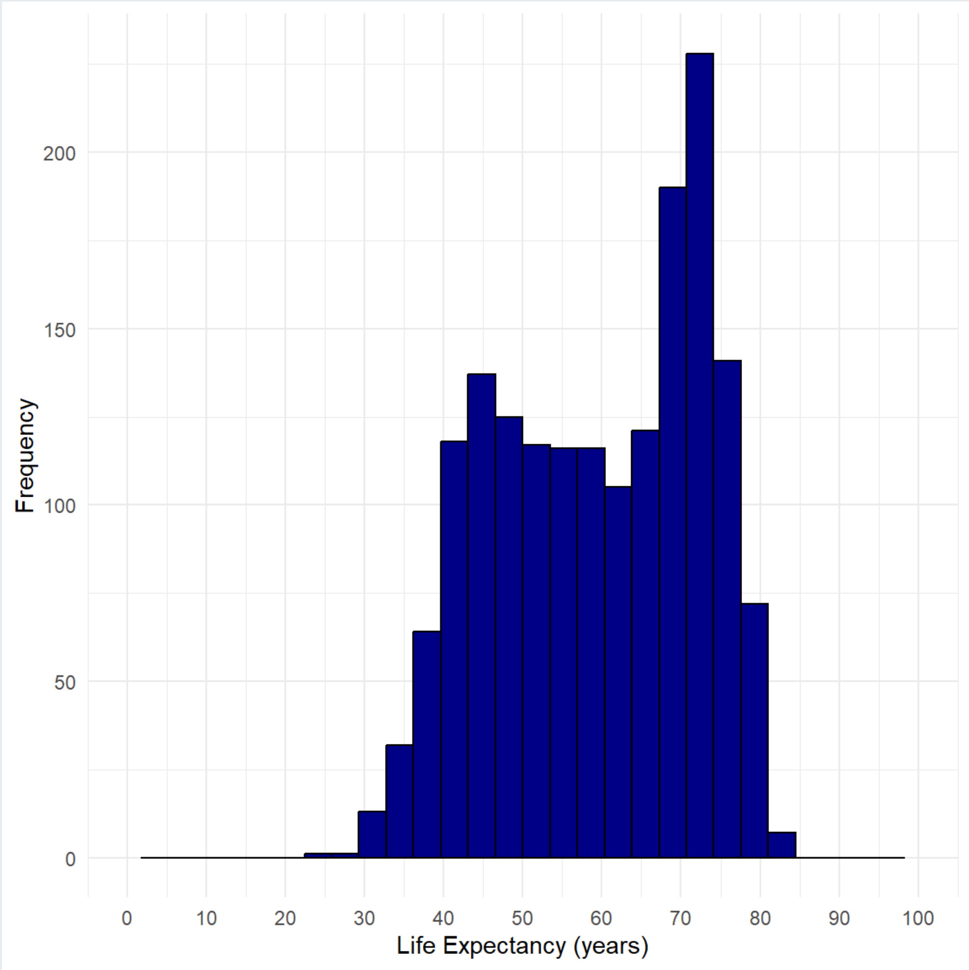
$$P(2\,luego\,2) = P(2) \cdot P(2) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

La Distribución Normal

Escala de frecuencia de un Histograma

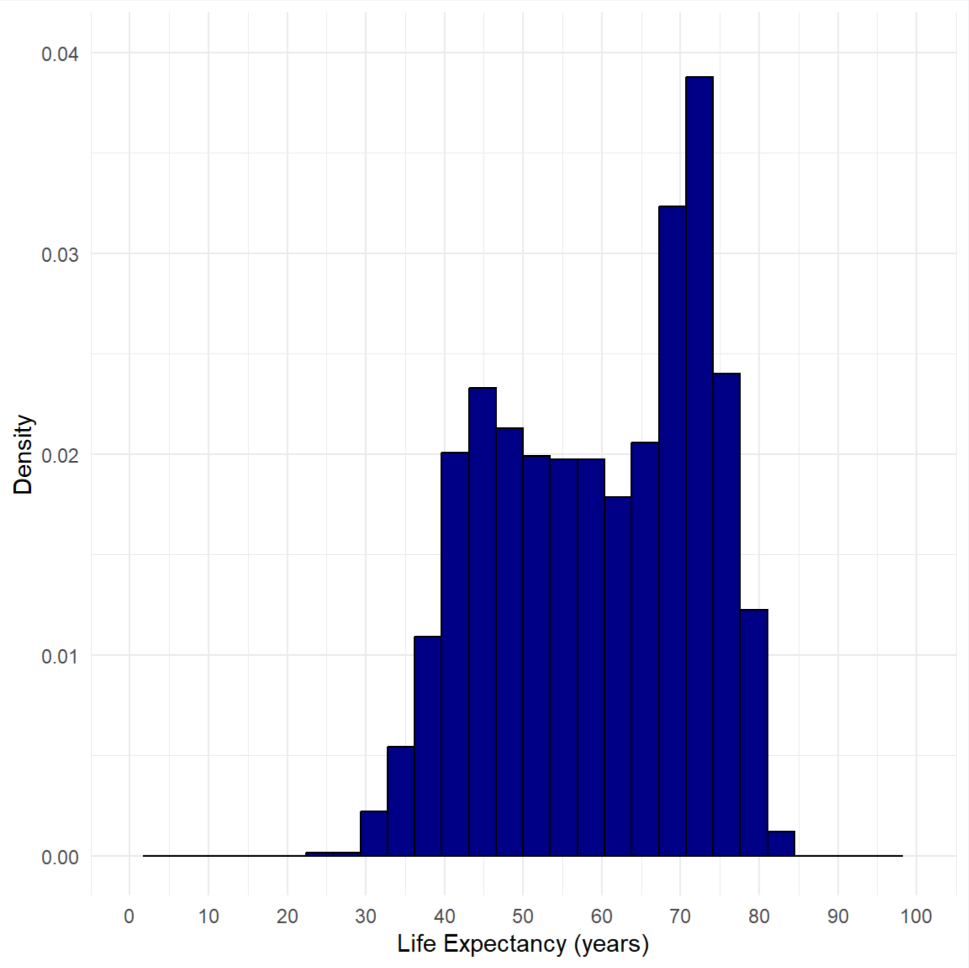
Retomemos la distribución de la expectativa de vida en los datos `gapminder`.

Para los histogramas en una escala de frecuencia, la altura de las barras = cantidad de observaciones en ese intervalo.



Escala de densidad de un Histograma

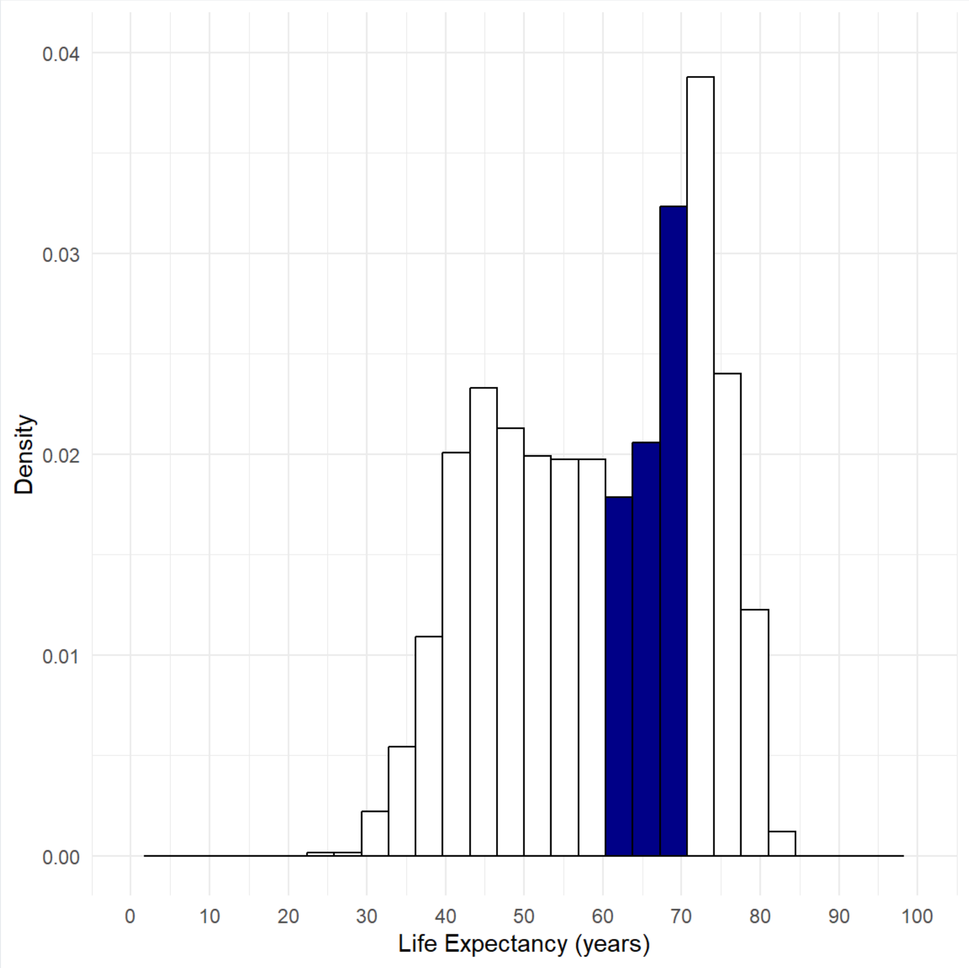
Para un histograma en una escala de densidad,
el área de la barra = proporción de observaciones en ese intervalo.



Escala de densidad de un Histograma

En una escala de densidad, la proporción de países con expectativas de vida entre 60 y 70 años = el área bajo la curva del histograma entre 60 y 70.

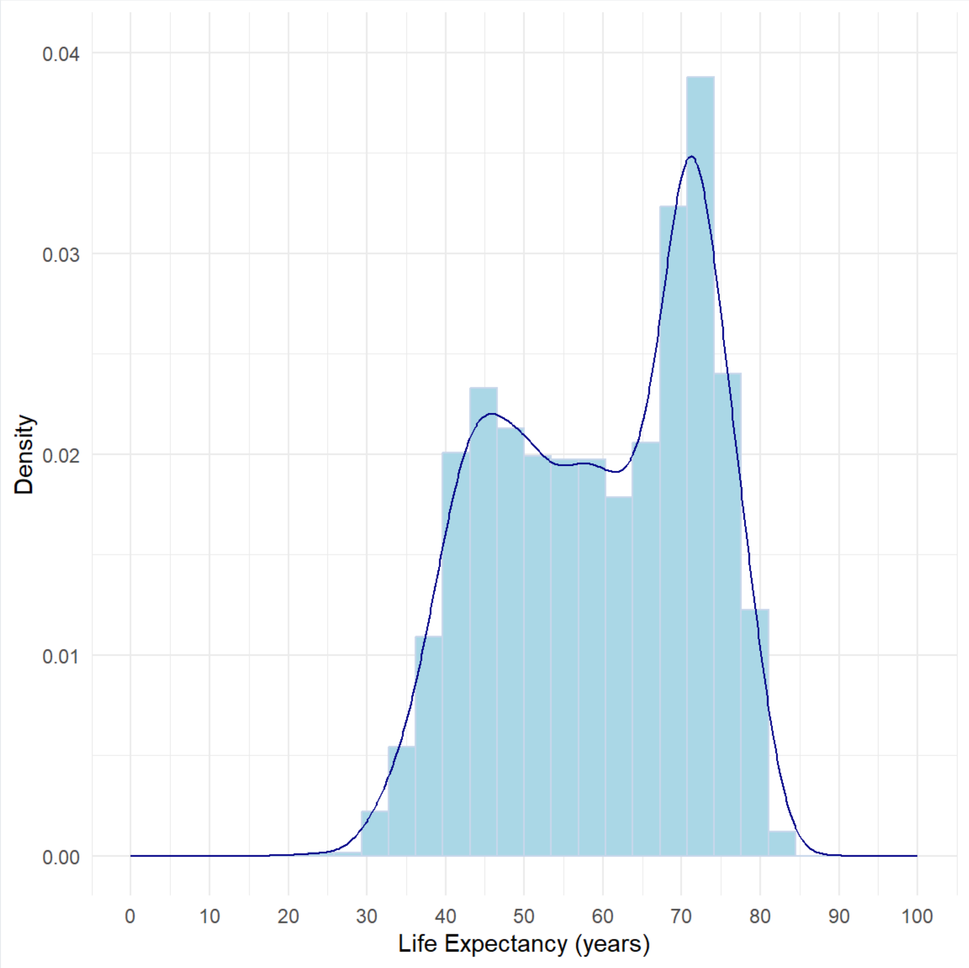
Esa proporción es básicamente la probabilidad de encontrar esos valores en nuestros datos.



Del Histograma a la Curva de Densidad

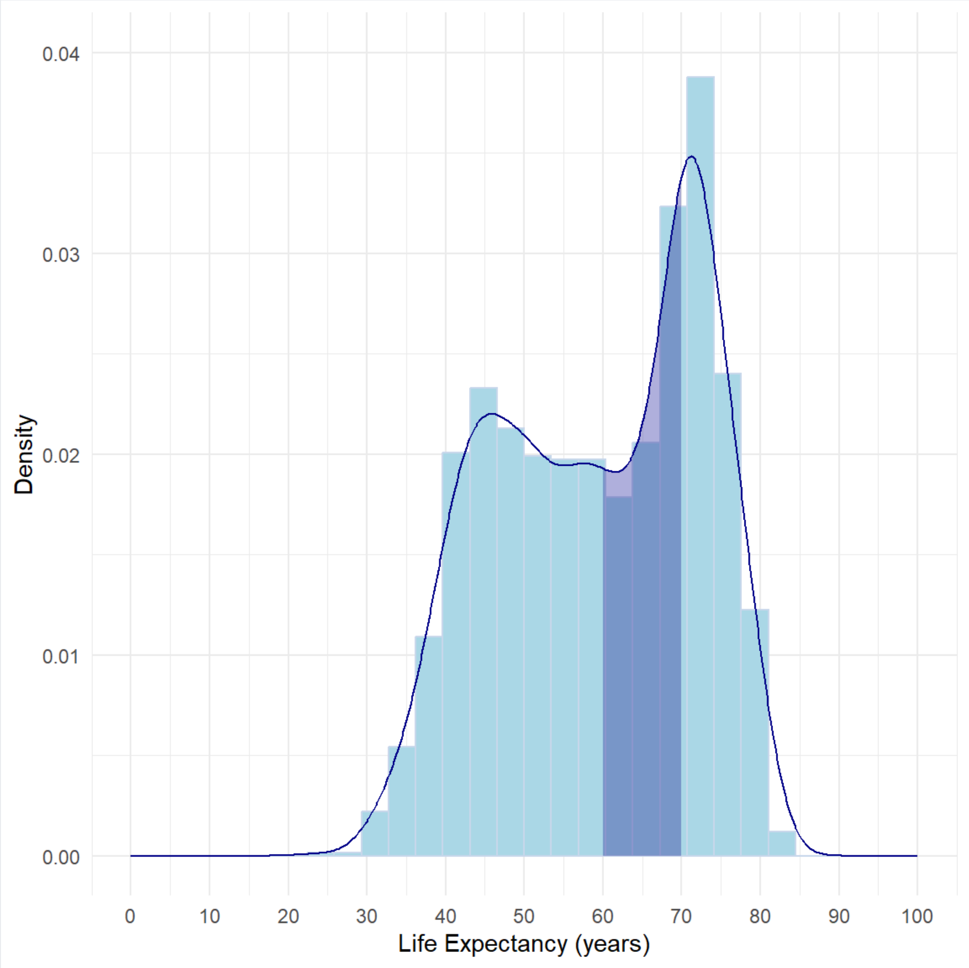
Podríamos intentar aproximar un histograma mediante una curva suave, llamada una función de densidad (probabilidad).

1. Una función de densidad nunca es negativa
2. El área total bajo la curva es siempre 1 o 100%



Del Histograma a la Curva de Densidad

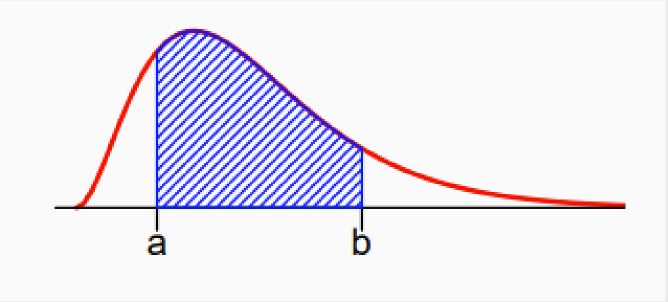
Por lo tanto, la proporción de países con expectativas de vida entre 60 y 70 años se puede estimar como el **área sombreada bajo la curva**. La proporción exacta es el área bajo el histograma.



Variables continuas y Curvas de densidad

La distribución de probabilidad de una variable aleatoria continua se describe mediante una curva de densidad.

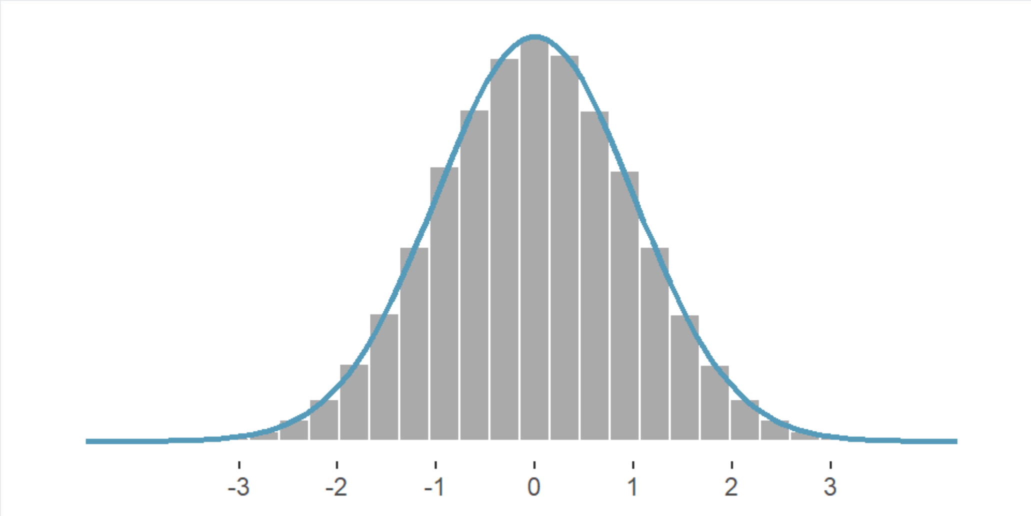
Si Y es una variable aleatoria continua, $P(a < Y < b)$ es el área bajo la curva de densidad de Y sobre el intervalo entre a y b.



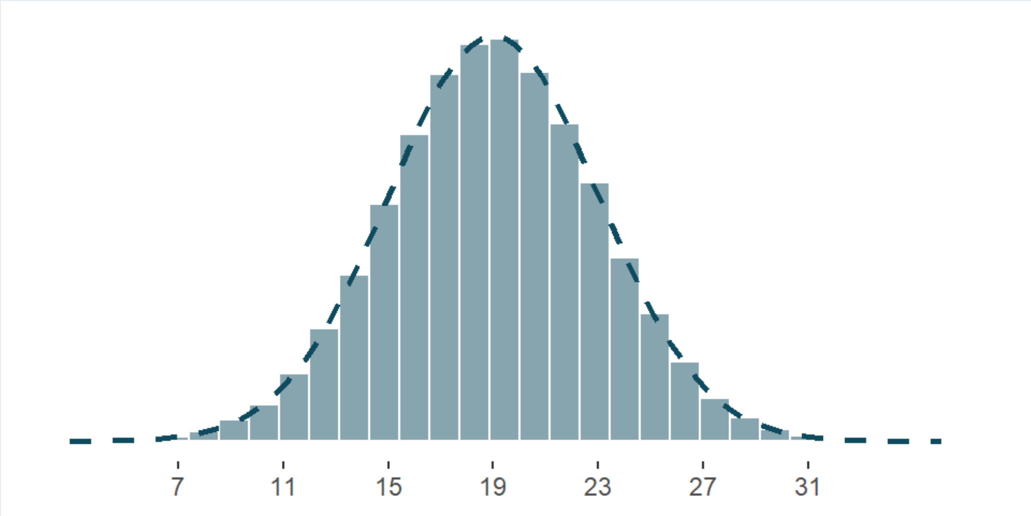
Nota: todas las distribuciones de probabilidad continuas asignan una probabilidad de cero a cada resultado individual: $P(Y = y) = 0$

Distribución Normal

La distribución normal (campana de Gauss) es una familia de curvas de densidad que son simétricas y con forma de campana. Se definen por su media μ y su desviación estándar σ, con notación $N(\mu, \sigma)$.



Mean = 0, SD = 1



Mean = 19, SD = 4

Las dos distribuciones entonces se escribirían $N(1, 0)$ y $N(19, 4)$.

Ejemplo: SAT vs. ACT

- Las puntuaciones del SAT siguen una distribución aproximadamente normal con una media de 1500 puntos y una desviación estándar de 300 puntos.

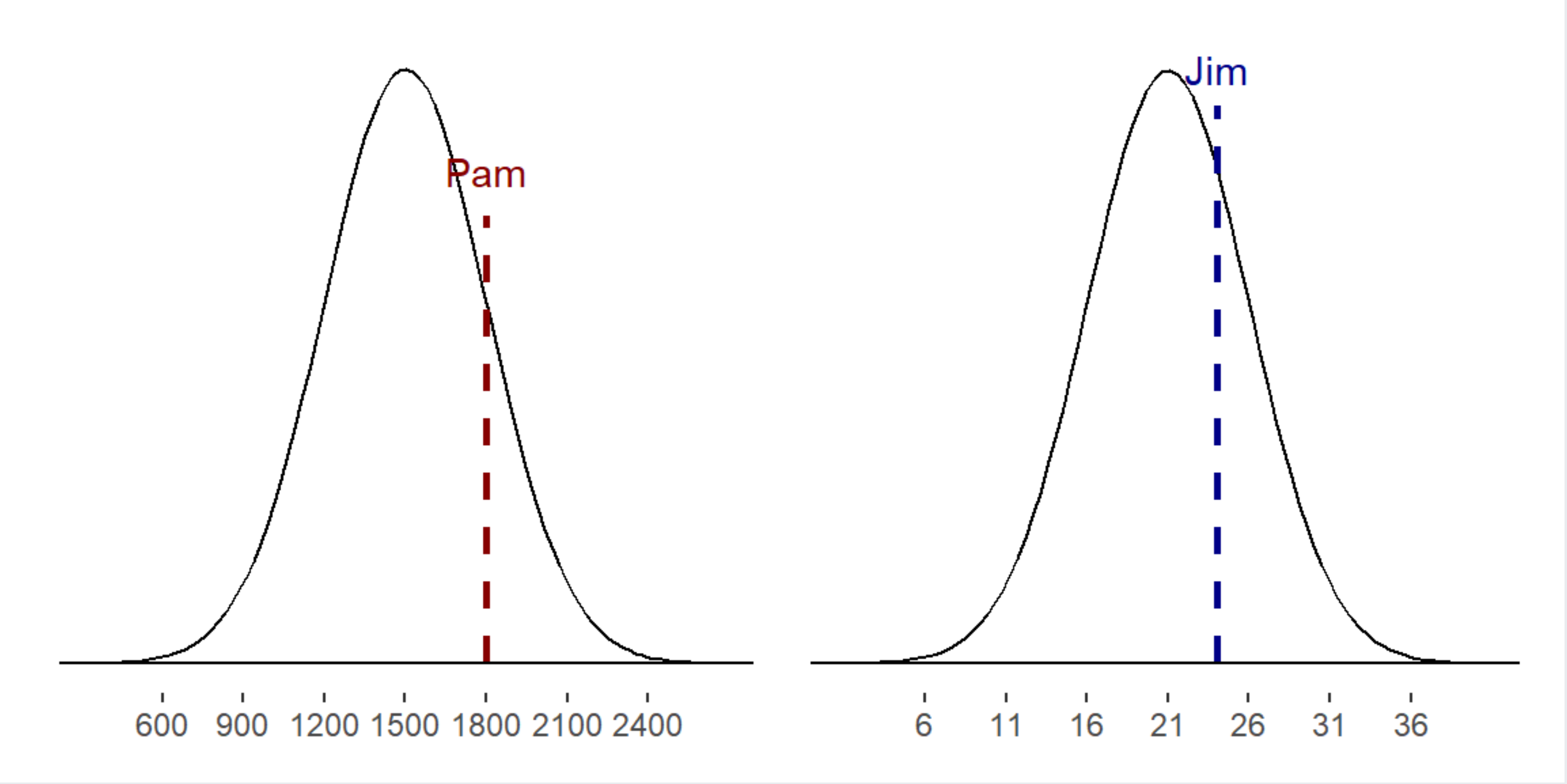
$$SAT \sim N(1500, 300)$$

- Las puntuaciones del ACT también siguen una distribución aproximadamente normal con una media de 21 puntos y una desviación estándar de 5 puntos.

$$ACT \sim N(21, 5)$$

Ejemplo: SAT vs. ACT

Supongamos que una universidad está decidiendo cuál de los dos aspirantes obtuvo un mejor puntaje en su examen estandarizado en comparación con los otros estudiantes: Pam, quien obtuvo un 1800 en su SAT, o Jim, quien obtuvo un 24 en su ACT?



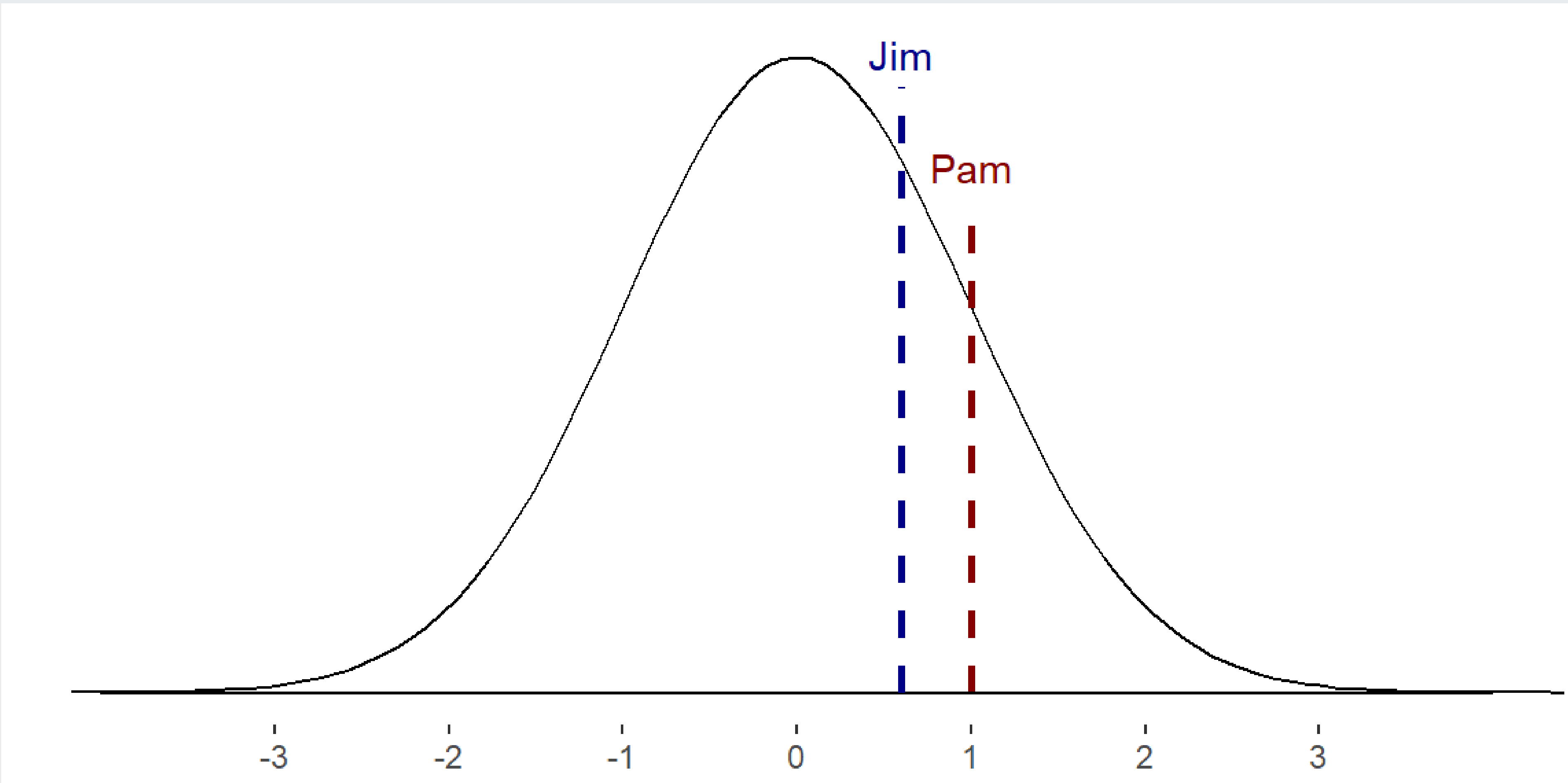
Estandarizar con el Z-Score

Dado que no podemos simplemente comparar estos dos puntajes, en su lugar comparamos cuántas desviaciones estándar por encima de la media está cada observación.

- El puntaje de Pam es $\frac{1800 - 1500}{300} = 1$ desviación estándar (SD) encima de la media
- El puntaje de Jim es $\frac{24 - 21}{5} = 0.6$ SD encima de la media

Estandarizar con el Z-Score

La siguiente gráfica visualiza la comparación que estamos haciendo al usar desviaciones estándar:



El Z-score

El Z-score de una observación representa cuántas desviaciones estándar se encuentra por encima o por debajo de la media, permitiendo comparar su posición relativa dentro de una distribución.

Si x es una observación de la distribución $N(\mu, \sigma)$, el Z-score se define:

$$Z = \frac{x - \mu}{\sigma}$$

Observaciones que estén más allá de 3 SD de la media ($|Z| > 3$) son usualmente consideradas inusuales.

Maneras de detectar *valores atípicos* (Outliers)

- $1.5 \times RIC$
- Observaciones con $|Z| > 3$
- Histogramas
- Diagramas de caja

Es importante analizar la causa de los valores atípicos antes de eliminarlos, ya que pueden contener información valiosa.

¿De dónde vienen los *valores atípicos*?

1. Errores de medición o registro de los datos

- Suelen ser valores extraños o imposibles

2. Problemas de muestreo y condiciones inusuales

- No son parte de la población que nos interesa

3. Variación natural

- Sí son parte de la población que nos interesa, por lo tanto, son informativos

Ejercicio 3

1. Completen el siguiente código que crea una variable igual a 1 si el credito está fuera del rango $1.5 \times RIC$ y 0 en caso contrario.

```r
# fivenum() devuelve: min, Q1, mediana, Q3, max
q1 <- fivenum(credit$limit_bal, na.rm = TRUE)[_]
q3 <- fivenum(_________________, na.rm = TRUE)[4]
ric <- q3-q1

credit <- credit |>
  mutate(
    # Identificar outliers por RIC
    outlier_ric = ifelse(limit_bal < (q1 - 1.5 * ___) | _________ > (q3 + 1
  )
```

Ejercicio 3

2. Completen el siguiente código que crea una variable igual a 1 si $|ZScore_i| > 3$ y 0 en caso contrario.

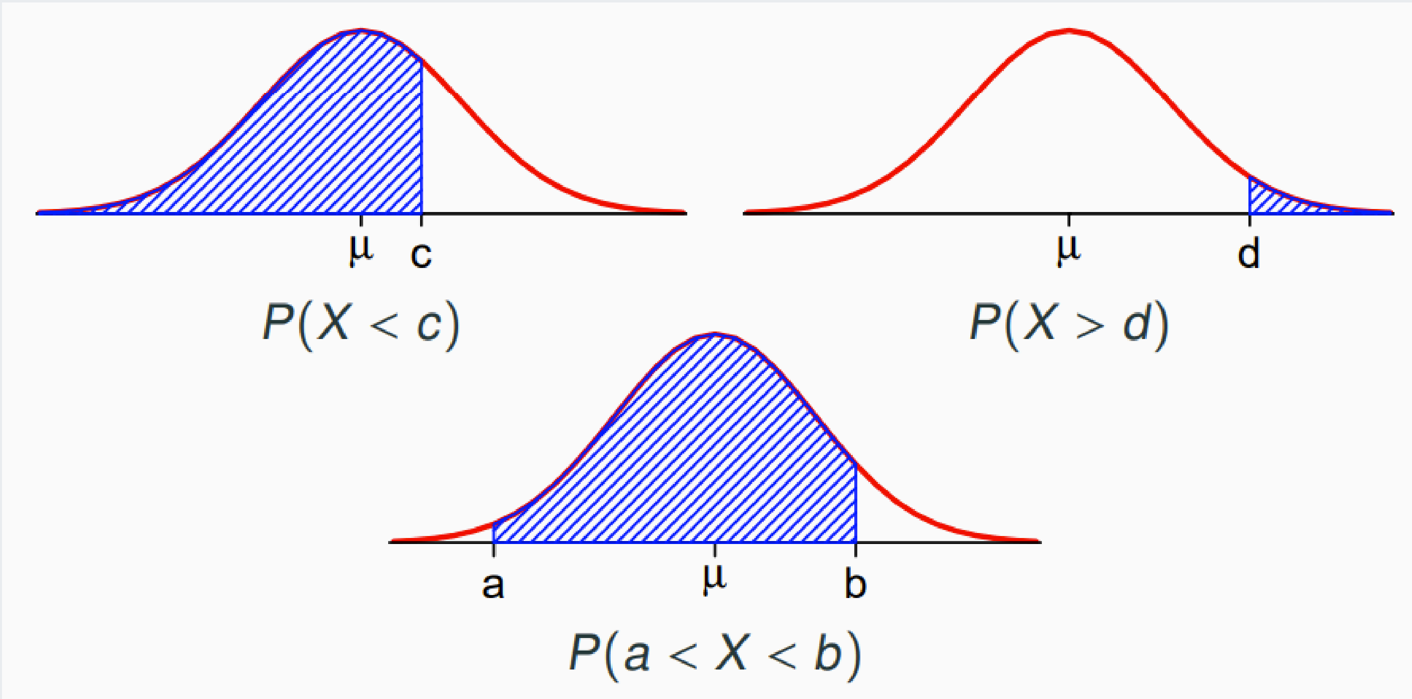
```r
credit <- credit |>
  mutate(
    z_score = (limit_bal - mean(_______, na.rm = TRUE)) / sd(_______, na.rm
    outlier_zscore = ifelse(abs(_______) > 3, 1, 0)
  )
```

3. Luego ejecuten el siguiente código. ¿Qué tan diferentes son los dos criterios?

```r
credit |>
  count(outlier_zscore, outlier_ric)
```

Distribución Normal y Probabilidad

Si X sigue una distribución normal, para encontrar probabilidades sobre X se calculan áreas bajo la curva normal $N(\mu, \sigma)$



Calculando Probabilidades: La tabla de la Normal

$$P(Z < z) = P(Z < -0.83) = 0.2033$$

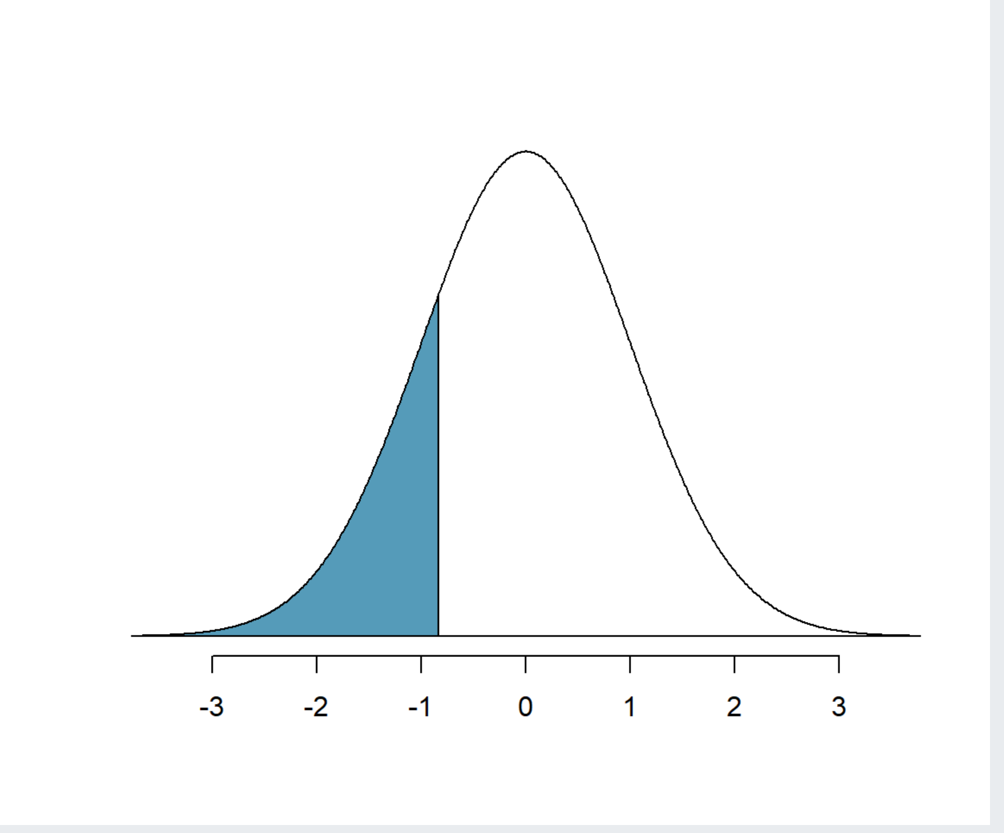
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Calculando Probabilidades en R

- Con la función `pnorm()` pueden calcular probabilidades en R:

```r
pnorm(-0.83)
```

```
[1] 0.2032694
```



Calculando Probabilidades en R

Hay dos maneras de calcular probabilidades en la parte superior de la distribución:

- P($Z > -0.83$) = 1 - P($Z < -0.83$)

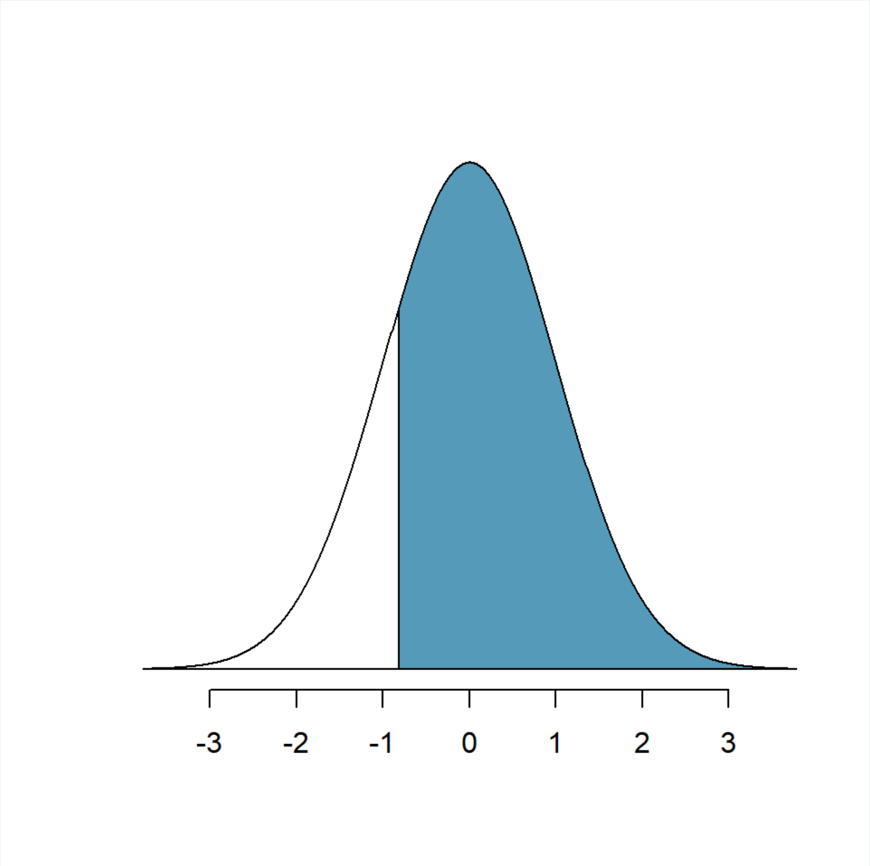
```
1 - pnorm(-0.83)
```

```
[1] 0.7967306
```

- Y la otra es cambiando las opciones de la función:

```
pnorm(-0.83, lower.tail=FALSE)
```

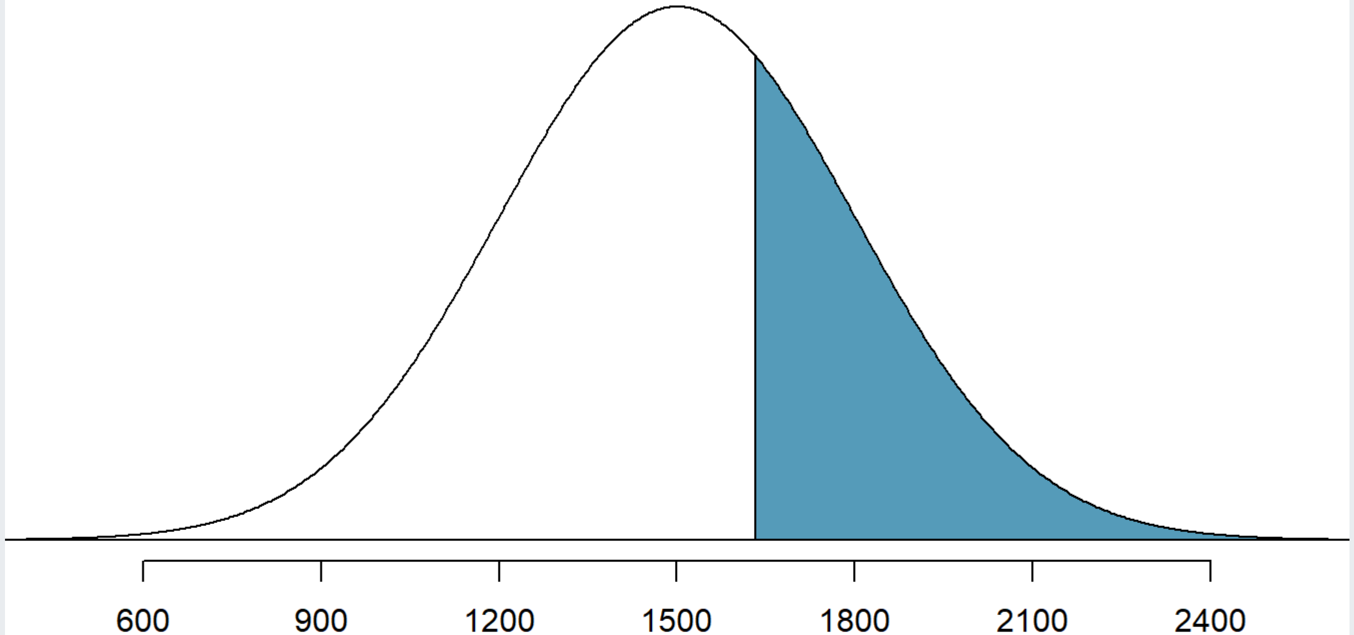
```
[1] 0.7967306
```



Calculando probabilidades para la Distribución Normal

¿Qué porcentaje de estudiantes tiene puntajes mayores a 1630 en el SAT? Recuerden que SAT$\sim N(1500, 300)$

procedimiento es más claro si
fican el área que van a calcular


600
900
1200
1500
1800
2100
2400

Calculando probabilidades para la Distribución Normal

Probabilidades con Área Complementaria

1. Calcular el Z-score:

$$Z = \frac{1630 - 1500}{300} = 0.43$$

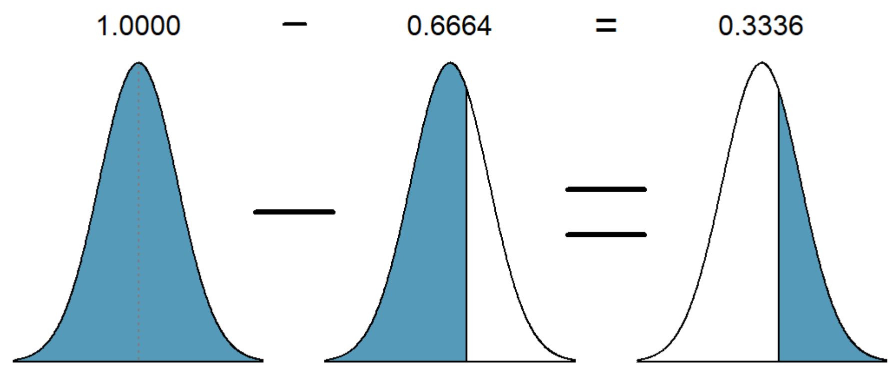
2. Calcular el área bajo la distribución normal estandarizada:

```
pnorm(0.43)
```

```
[1] 0.6664022
```

```
1 - pnorm(0.43)
```

```
[1] 0.3335978
```



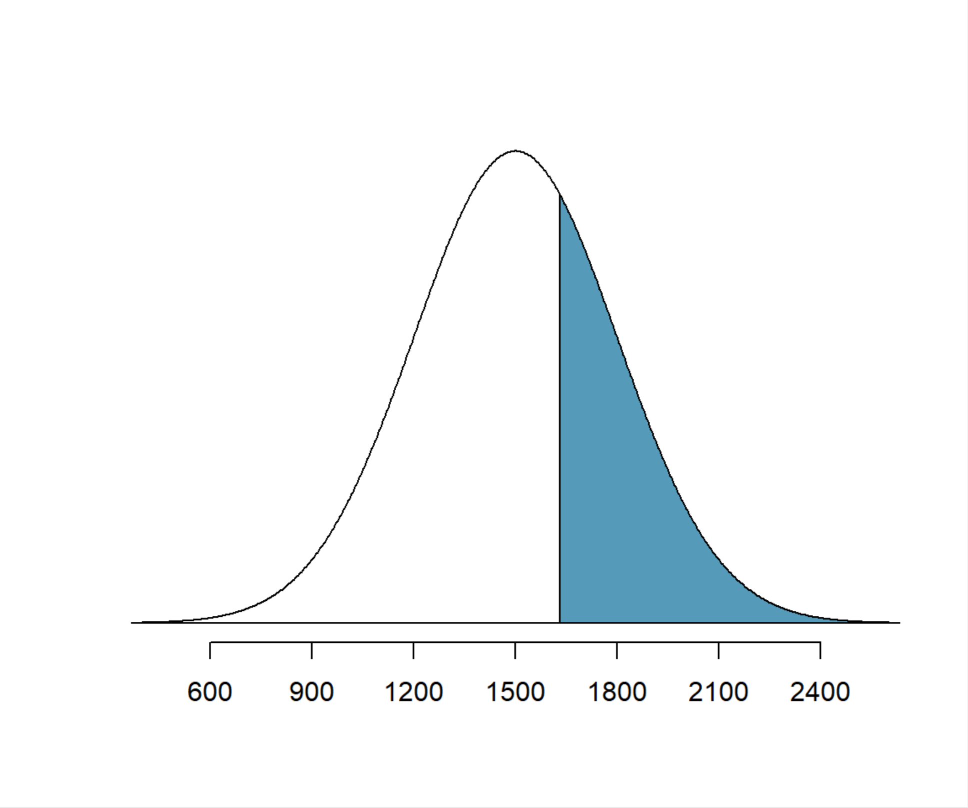
Calculando probabilidades para la Distribución Normal

Usando las opciones de pnorm()

1. Modificar los valores en la función `pnorm()`:

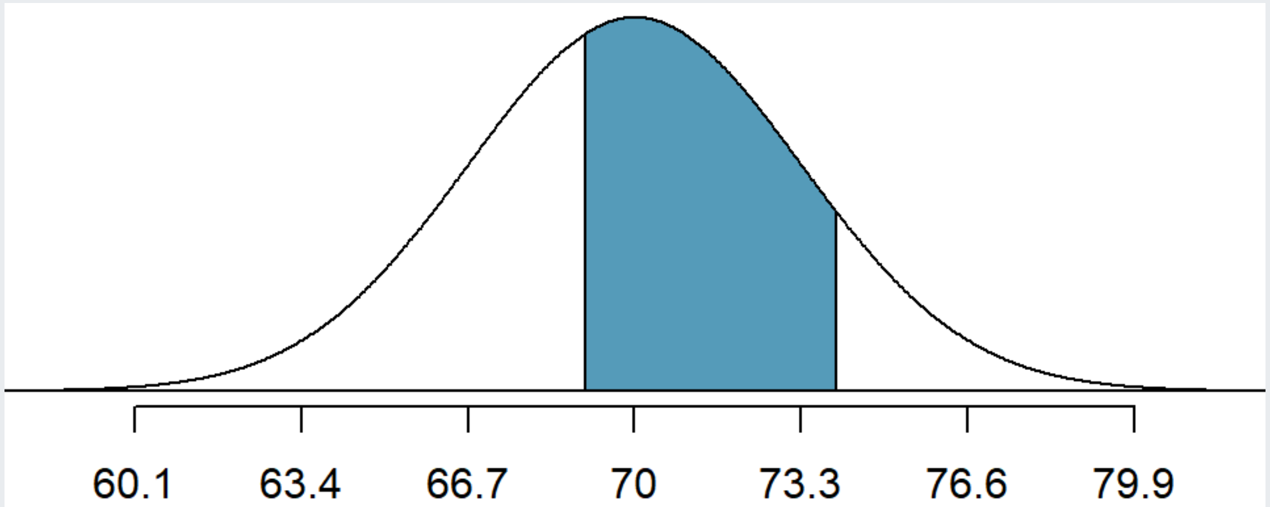
```
pnorm(1630,
      mean = 1500,
      sd = 300,
      lower.tail=FALSE)
```

```
[1] 0.3323863
```

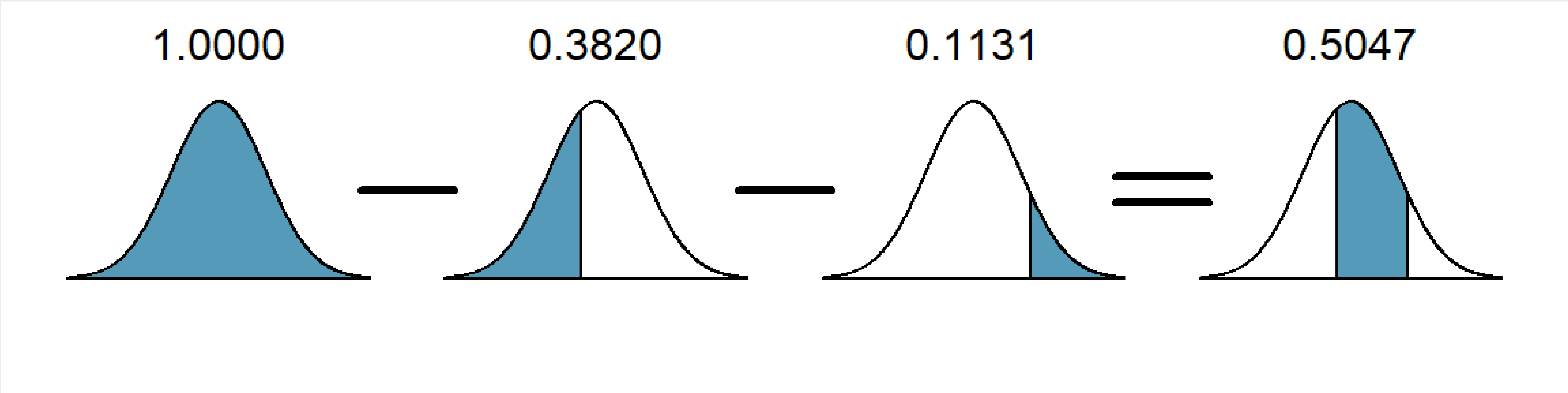


Calculando probabilidades: Otro ejemplo

Con base en una muestra de 100 transacciones, el gasto promedio mensual de los clientes en una tienda minorista estadounidense sigue una distribución casi normal con una media de \$70.00 y una desviación estándar de \$3.30. ¿Cuál es la probabilidad de seleccionar aleatoriamente una transacción entre 69 y 74 dólares?



Calculando probabilidades: Otro ejemplo



$$Z_1 = \frac{69 - 70}{3.3} = -0.30$$

$$Z_2 = \frac{74 - 70}{3.3} = 1.21$$

```r
a <- pnorm(-0.30)
a
```

```
[1] 0.3820886
```

```r
b <- pnorm(1.21, lower.tail=FALSE)
b
```

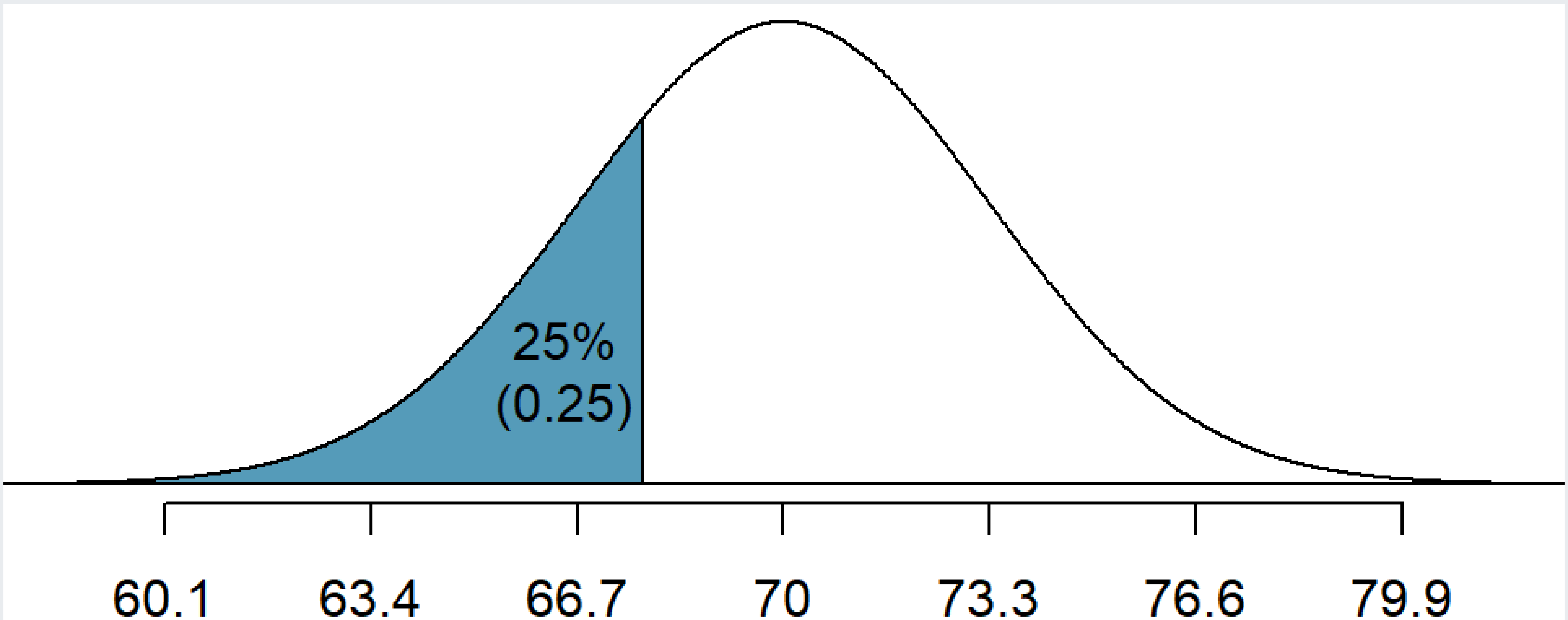
```
[1] 0.1131394
```

```r
1-a-b
```

```
[1] 0.504772
```

Calculando cuartiles/percentiles

¿Qué valor en la distribución de transacciones es el primer cuartil (Q_1)? Recuerden que 25% de los datos son menores al primer cuartil.



Calculando cuartiles/percentiles en R

La función `qnorm()` permite calcular cuartiles/percentiles para una distribución normal:

```
qnorm(0.25, mean = 0, sd = 1)
```

```
[1] -0.6744898
```

Sabiendo que el primer cuartil en la distribución normal estandarizada es $Z_{Q_1} = -0.674$:

$$-0.674 = Z_{Q_1} = \frac{x_{Q_1} - \mu}{\sigma} = \frac{x_{Q_1} - 70}{3.3}$$

Resolviendo para Z_{Q_1}, se encuentra que 67.7 dólares es el primer cuartil en la distribución original.

Calculando cuartiles/percentiles en R

Al usar las opciones de la función `qnorm()`, podemos calcular el valor exacto del primer cuartil:

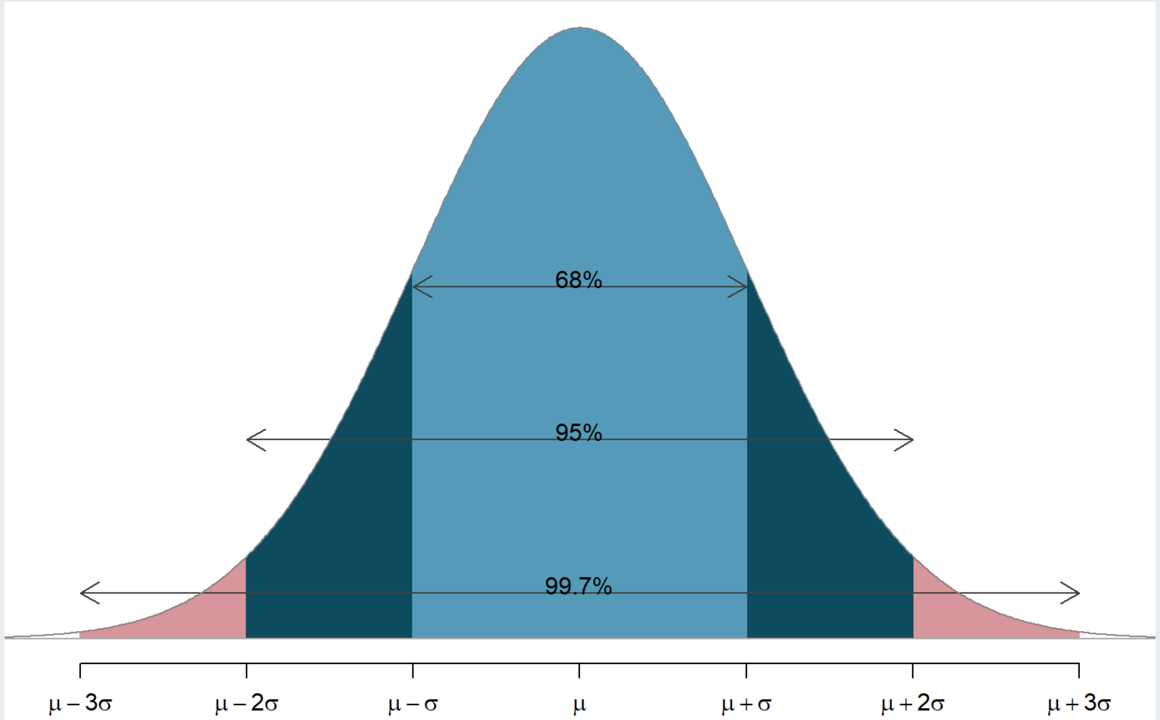
```r
qnorm(0.25, mean = 70, sd = 3.3)
```

```
[1] 67.77418
```

¿Qué es un percentil?

- Es un valor que divide un conjunto de datos ordenados en 100 partes iguales, indicando la posición relativa de un dato dentro del conjunto
- Por ejemplo, el primer cuartil es equivalente al percentil 25, lo que significa que el 25% de los datos son menores o iguales a ese valor

La regla 68-95-99.7




```r
pnorm(1) - pnorm(-1)
```

```
[1] 0.6826895
```

```r
pnorm(2) - pnorm(-2)
```

```
[1] 0.9544997
```

```r
pnorm(3) - pnorm(-3)
```

```
[1] 0.9973002
```

💪 Ejercicio 4

1. Suponga que el límite de crédito `limit_bal` $\sim N(150000, 50000)$. Calcule de dos maneras diferentes la probabilidad de que un cliente seleccionado al azar tenga un límite de crédito superior a 240000.
2. Encuentren el límite de crédito que corresponde al percentil 95.