



# Inferencia Estadística.

# Equipo

Cinthya Yesenia López Díaz	1658003
Marco Antonio Oviedo Acevedo	1851698
Zaira Tamara Escareño Loera	1868734
Arely Yazmín Reséndiz Pinales	1869432
Eliud Narváez moreno	1863955
Luis Adrián Navarro	1866581
Yaziel Gibran Barbosa Alcocer	1851004



# INTRODUCCION

En este proyecto exploraremos los aspectos básicos del análisis estadístico y el por qué es importante utilizarlo en materias como la ciencia de datos. Utilizaremos herramientas de programación como el lenguaje de R y Python para analizar los datos de venta de la comunidad residencial de lujo dentro de Walt Disney World Resort en Lake Buena Vista, Florida, llamado Golden Oak. Tomaremos en cuenta distintas variables como la ciudad, renta, el estado, garantía hipotecarían entre muchas otras variables más que se definirá en el desarrollo del proyecto

# Investigación de Golden Oak

Golden Oak en Walt Disney World Resort es una comunidad residencial de lujo dentro de Walt Disney World Resort en Lake Buena Vista, Florida. Fue diseñado por Walt Disney Imagineering, es propiedad y está operado por una subsidiaria de Disney recién formada, Golden Oak Realty. La primera fase de desarrollo se encuentra al sureste del Parque Temático Magic Kingdom en Bay Lake. El área fue nombrada para rendir homenaje al Golden Oak Ranch de Walt Disney en California.

Golden Oaks es un conjunto residencial inspirado en los famosos personajes de los cuentos. Sí, esos que han formado parte del imaginario colectivo por generaciones y que se niegan a morir.

Se trata de una experiencia de lujo llevada a sus límites. El precio de cada una de sus viviendas empieza a partir de los 1.8 millones de dólares; un monto elevado para la mayoría, pero adecuado para los servicios y amenidades que ofrece.

# Antecedentes teóricos y metodológicos del estudio.

## **ESTADISTICA DESCRIPTIVA.**

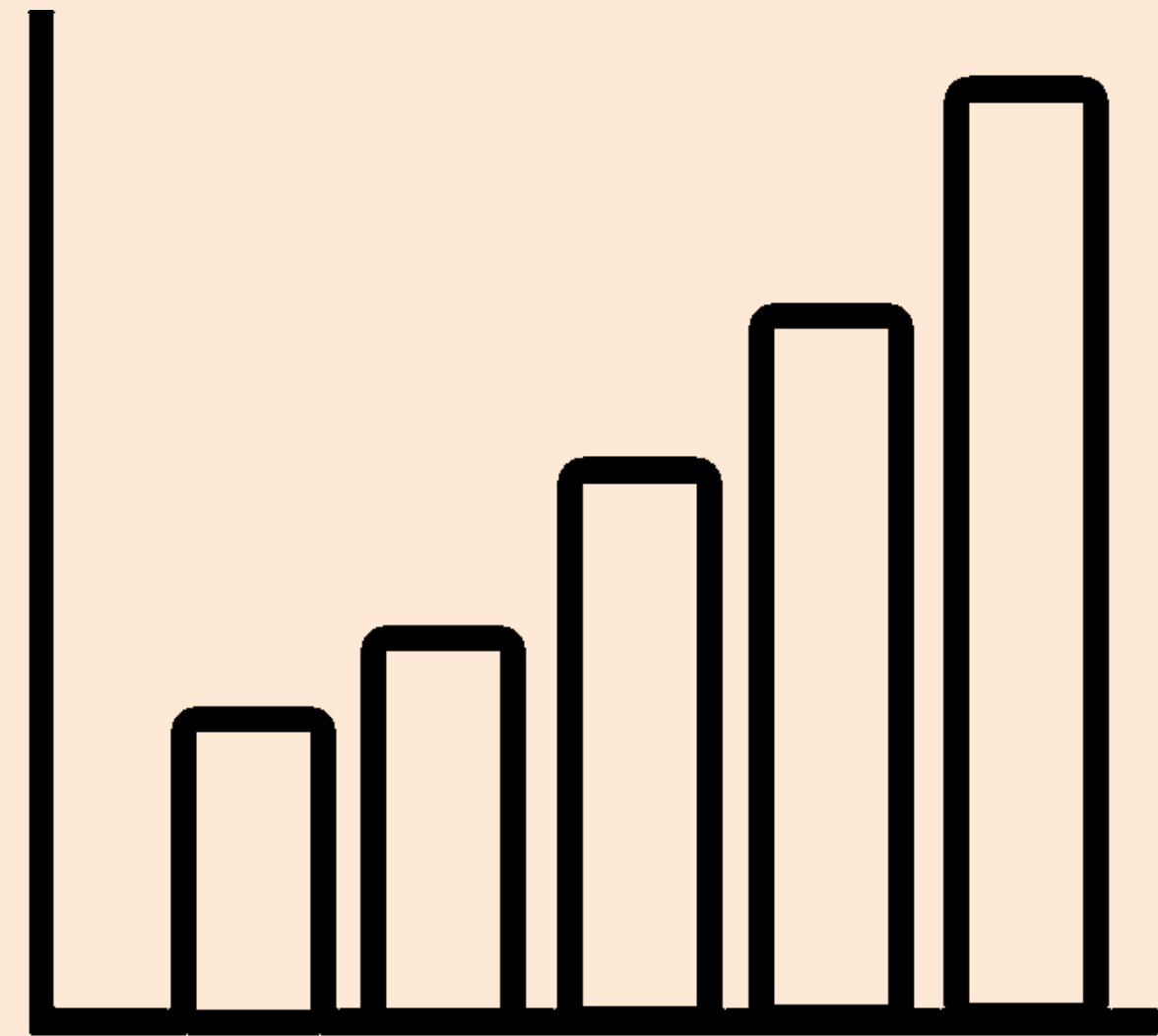
El propósito de la estadística aplicada es el de obtener conclusiones de una población en estudio, examinando solamente una parte de ella denominada muestra.

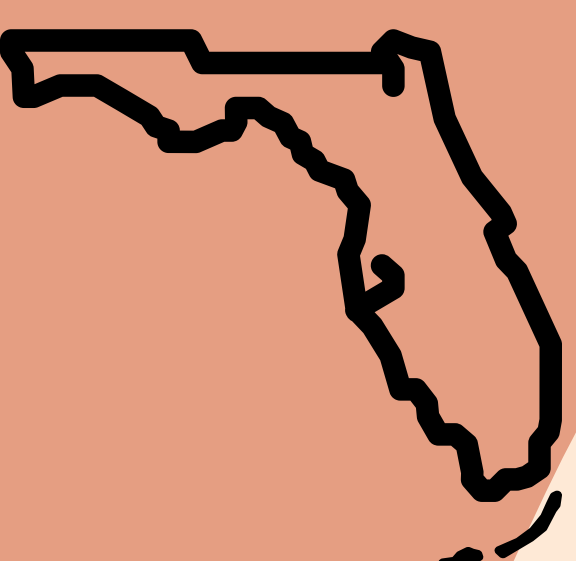


# ESTADISTICA INFERENCIAL.

La estadística es una rama de las matemáticas encargada de reunir, organizar y analizar datos generalmente numéricos, ayuda a resolver problemas y además permite luego de realizados los cálculos tomar decisiones que puedan beneficiar al contexto que las estudia.

La estadística y los procedimientos que con ella pueden realizarse han permitido de manera efectiva describir con exactitud datos de casi todas las ramas del conocimiento entre ellas: economía, psicología, política, física, biología, química, medicina e informática y ha servido como herramientas útiles para encontrarle relación a muchos datos estudiados por estas ciencias

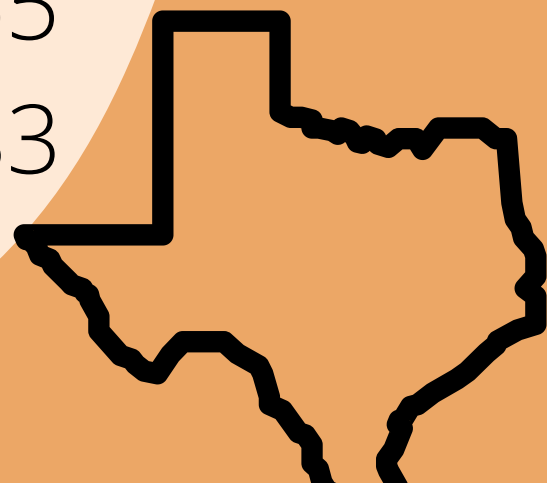




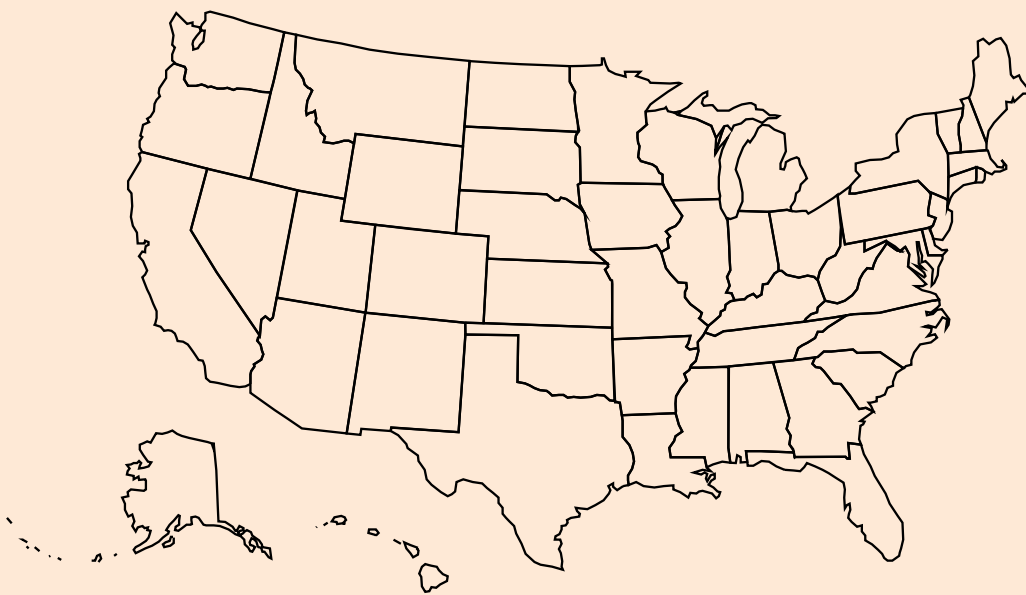
# PLANTEAMIENTO DEL PROBLEMA

Como empresa queremos analizar la información de nuestros clientes para mejorar el servicio que ofrecemos, así como nuestras campañas de marketing para incrementar nuestras ventas, por lo que revisaremos los datos de nuestros compradores de agosto del 2017 de nuestras residencias de lujo Golden Oak.

Decidimos empezar analizando los estados de vivienda donde nuestros clientes tienen una mayor compra, dado como resultado que los estados donde las personas compran más son de Florida con 2289 registros, New York con 2565 registros, Texas con 2767 registros y California con 4183 registros.



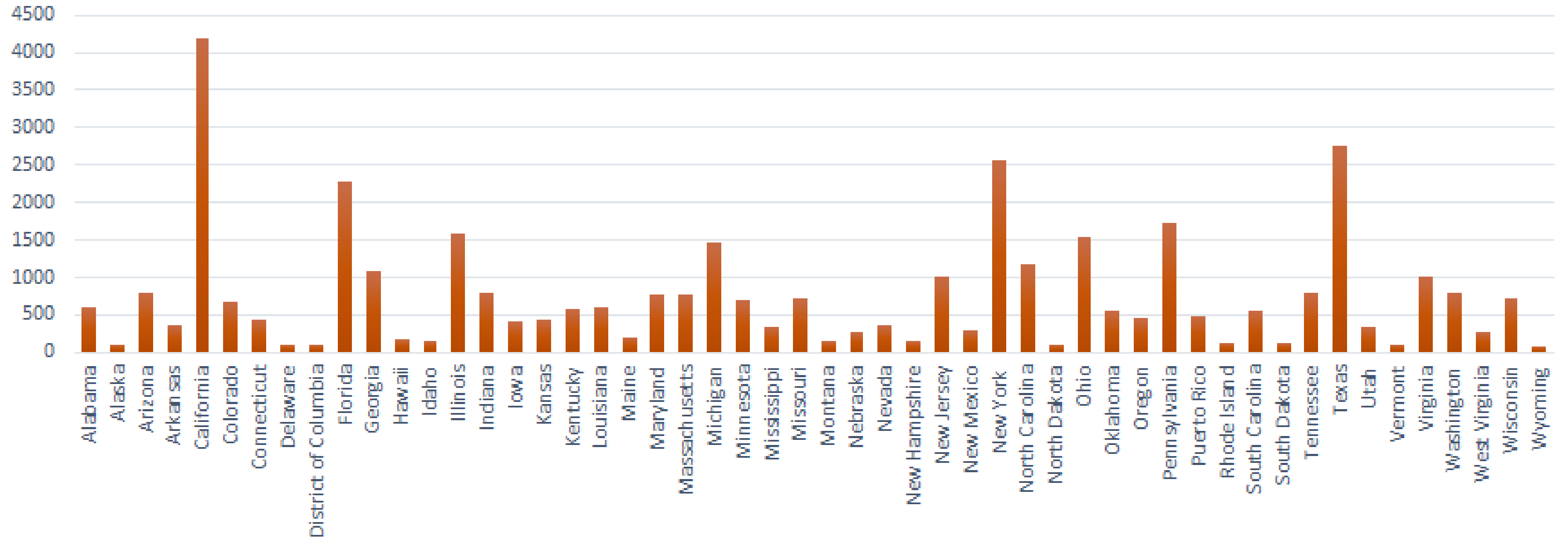
# Tabla de frecuencia de los estados



Estado	Absoluta	Absoluta Acumulada	Relativa	Relativa Acumulada	Relativa Porcentual	Relativa Acumulada Porcentual
Alabama	612	612	0.01568	0.01568	1.57%	1.57%
Alaska	105	717	0.00269	0.01837	0.27%	1.84%
Arizona	798	1515	0.02045	0.03882	2.04%	3.88%
Arkansas	363	1878	0.00930	0.04812	0.93%	4.81%
California	4193	6071	0.10743	0.15555	10.74%	15.55%
Colorado	668	6739	0.01712	0.17266	1.71%	17.27%
Connecticut	445	7184	0.01140	0.18406	1.14%	18.41%
Delaware	109	7293	0.00279	0.18686	0.28%	18.69%
District of Columbia	98	7391	0.00251	0.18937	0.25%	18.94%
Florida	2289	9680	0.05865	0.24801	5.86%	24.80%
Georgia	1078	10758	0.02762	0.27563	2.76%	27.56%
Hawaii	174	10932	0.00446	0.28009	0.45%	28.01%
Idaho	148	11080	0.00379	0.28388	0.38%	28.39%
Illinois	1593	12673	0.04081	0.32470	4.08%	32.47%
Indiana	802	13475	0.02055	0.34525	2.05%	34.52%
Iowa	415	13890	0.01063	0.35588	1.06%	35.59%
Kansas	440	14330	0.01127	0.36715	1.13%	36.72%
Kentucky	577	14907	0.01479	0.38194	1.48%	38.19%



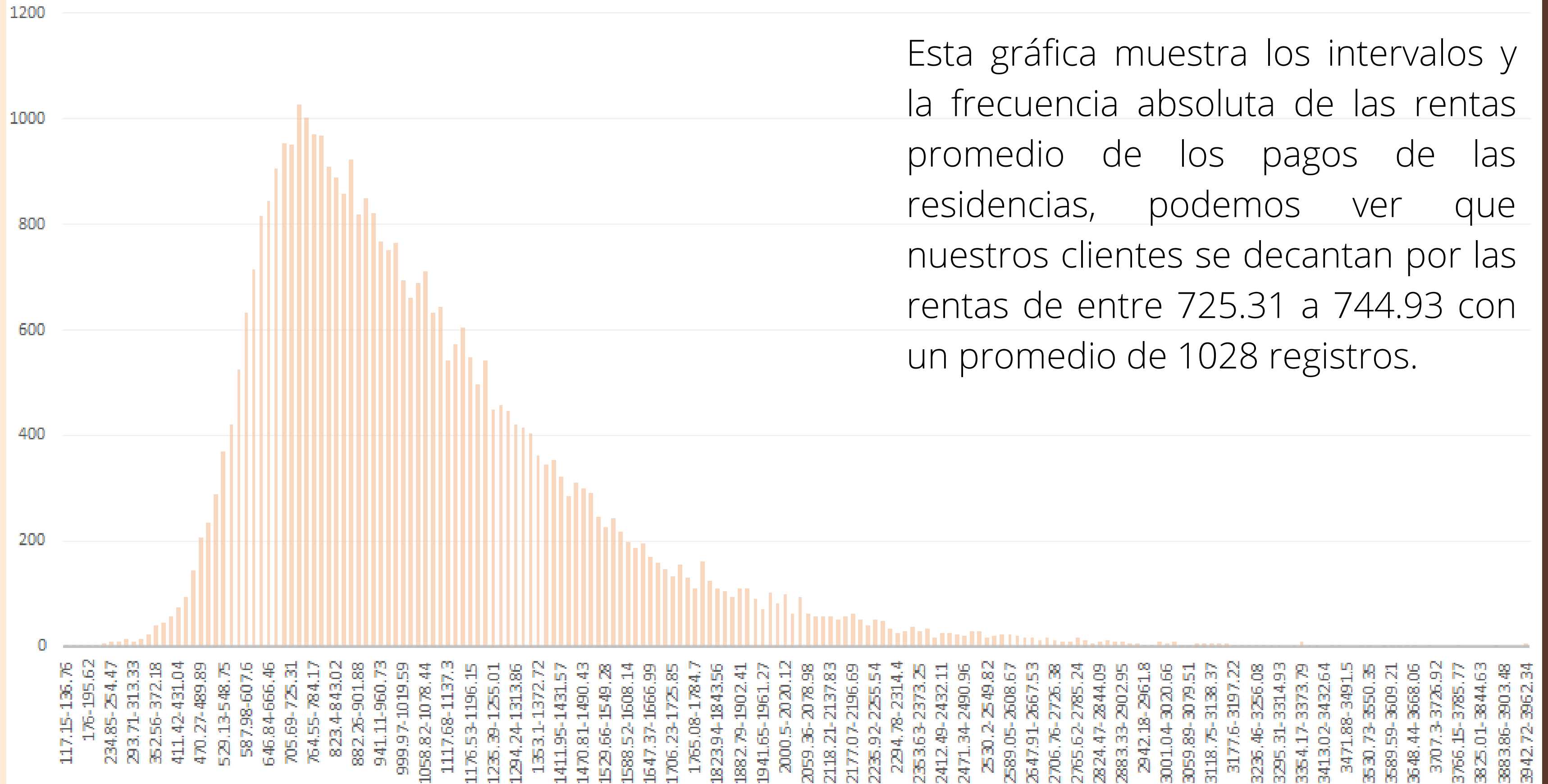
## Registros



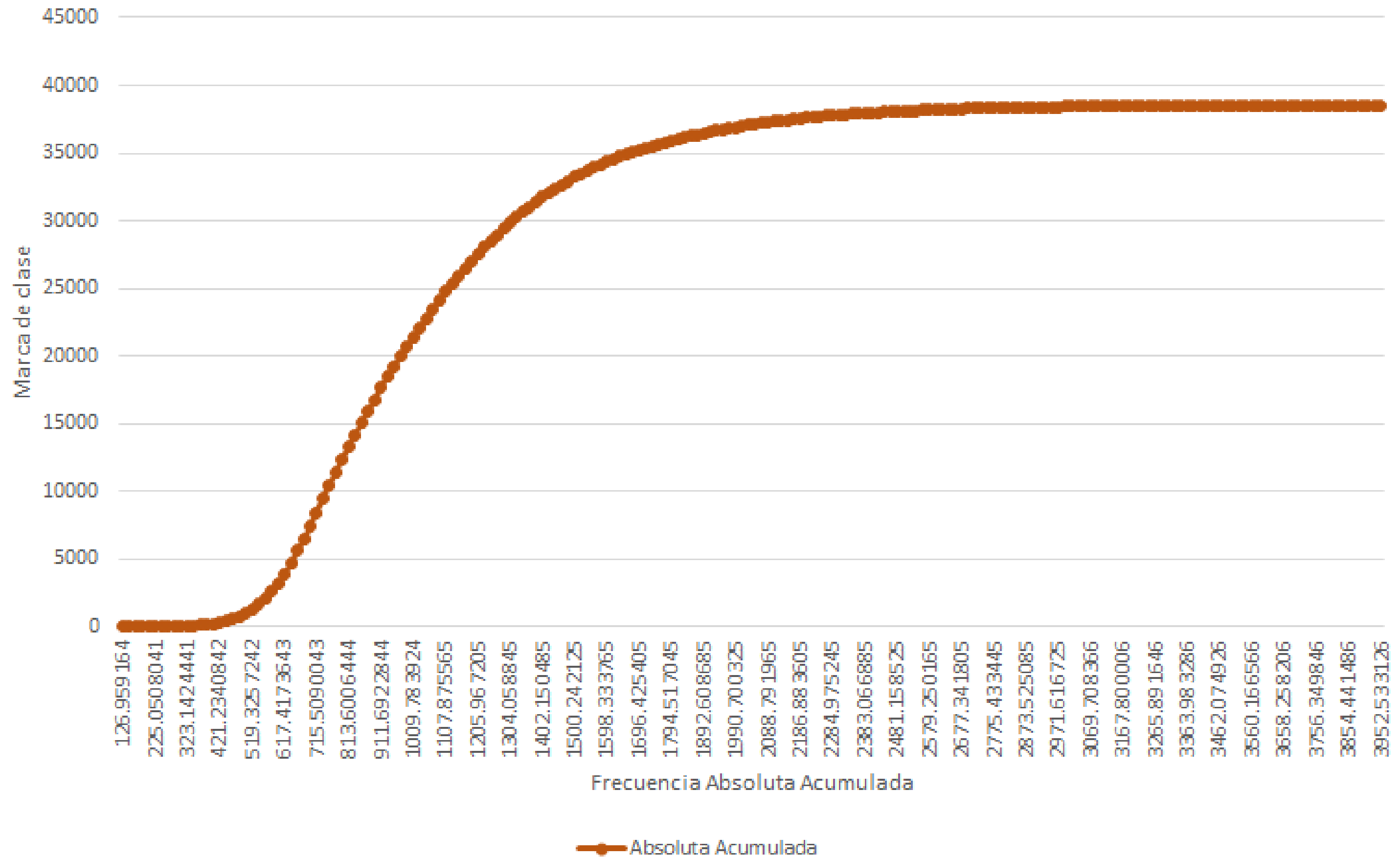
# Tabla de frecuencia de la renta

n=	38,538	*El total de datos son 39,030, pero se quitaron 492 datos atípicos						
k=	196.3109778							
Dato Min	117.15							
Dato Max	3962.34229							
Ancho de clase	19.61832801							
Intervalo		Marca de clase (xi)	Absoluta (ni)	Absoluta Acumulada	Relativa	Relativa Acumulada	Relativa Porcentual	Relativa acumulada Porcentual
117.15	136.768328	126.959164	2	2	5.18968E-05	5.18968E-05	0.005%	0.005%
136.768328	156.386656	146.577492	2	4	5.18968E-05	0.000103794	0.005%	0.010%
156.386656	176.004984	166.19582	3	7	7.78452E-05	0.000181639	0.008%	0.018%
176.004984	195.623312	185.814148	3	10	7.78452E-05	0.000259484	0.008%	0.026%
195.623312	215.2416401	205.432476	1	11	2.59484E-05	0.000285433	0.003%	0.029%
215.2416401	234.8599681	225.050804	7	18	0.000181639	0.000467071	0.018%	0.047%
234.8599681	254.4782961	244.669132	10	28	0.000259484	0.000726556	0.026%	0.073%
254.4782961	274.0966241	264.28746	8	36	0.000207587	0.000934143	0.021%	0.093%
274.0966241	293.7149521	283.905788	14	50	0.000363278	0.001297421	0.036%	0.130%
293.7149521	313.3332801	303.524116	8	58	0.000207587	0.001505008	0.021%	0.151%
313.3332801	332.9516081	323.142444	15	73	0.000389226	0.001894234	0.039%	0.189%
332.9516081	352.5699361	342.760772	23	96	0.000596814	0.002491048	0.060%	0.249%
352.5699361	372.1882641	362.3791	40	136	0.001037037	0.003528084	0.104%	0.353%

## Absoluta



Ojiva



# DESCRIPCIÓN GENERAL DE VARIABLES

- **Costos mensuales de la hipoteca y del propietario**

Suma de pagos de hipotecas, la tarifa mensual de condominio y costos de casas móviles. Los costos de propietario mensuales seleccionados se tabularon para todas las unidades ocupadas por el propietario, y se muestran por separado para las unidades "con una hipoteca" y para unidades "sin hipoteca".

- **Alquiler bruto**

Es el alquiler del contrato más el costo mensual promedio estimado de los servicios públicos y combustibles que paga el arrendatario. Los costos estimados se informan sobre una base de 12 meses, pero se convierten en cifras mensuales para las tabulaciones.

- **Ingresos del hogar y familiares**

Los ingresos del hogar incluyen la suma del jefe de hogar y las personas mayores de 15 años que residen en el hogar.

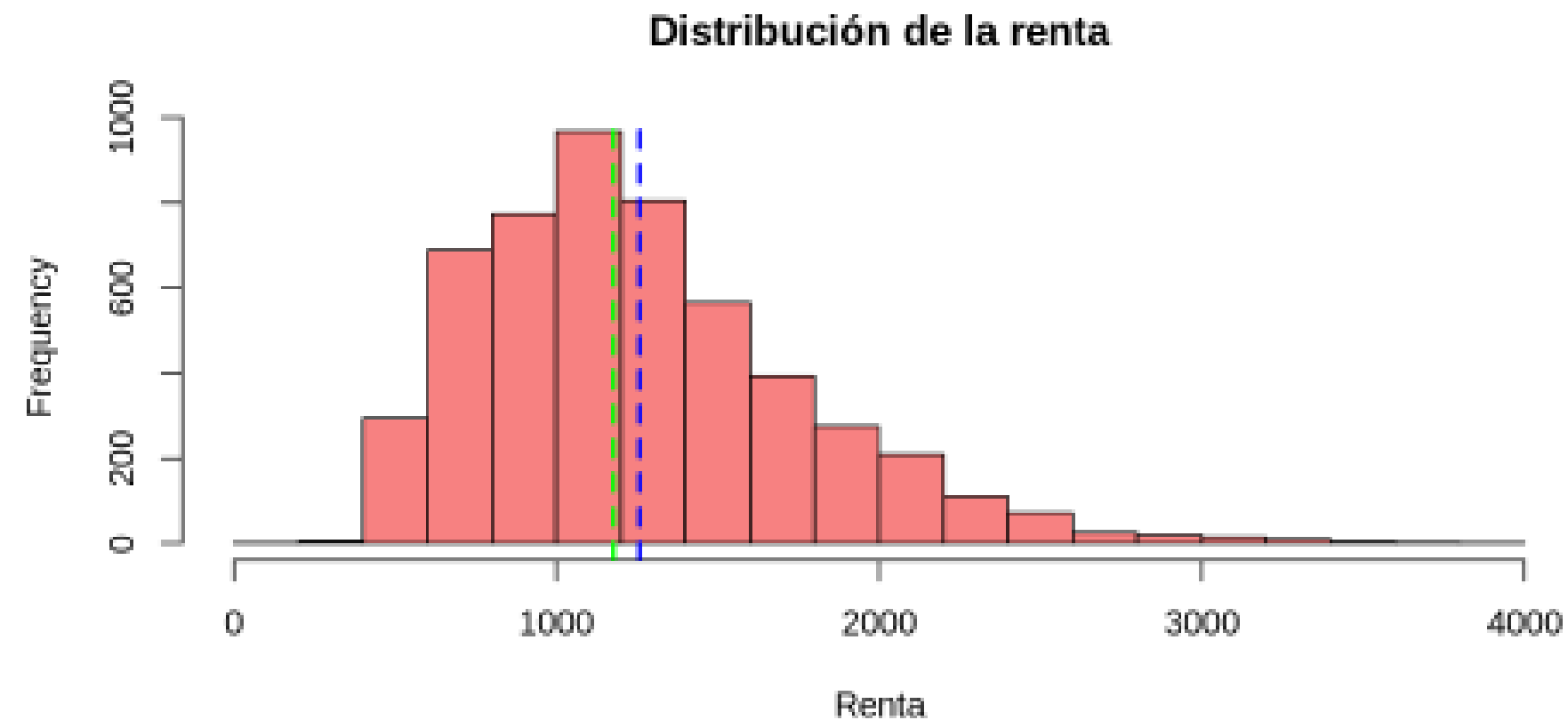


# DEFINICIONES DE CAMPO DE UBICACIÓN

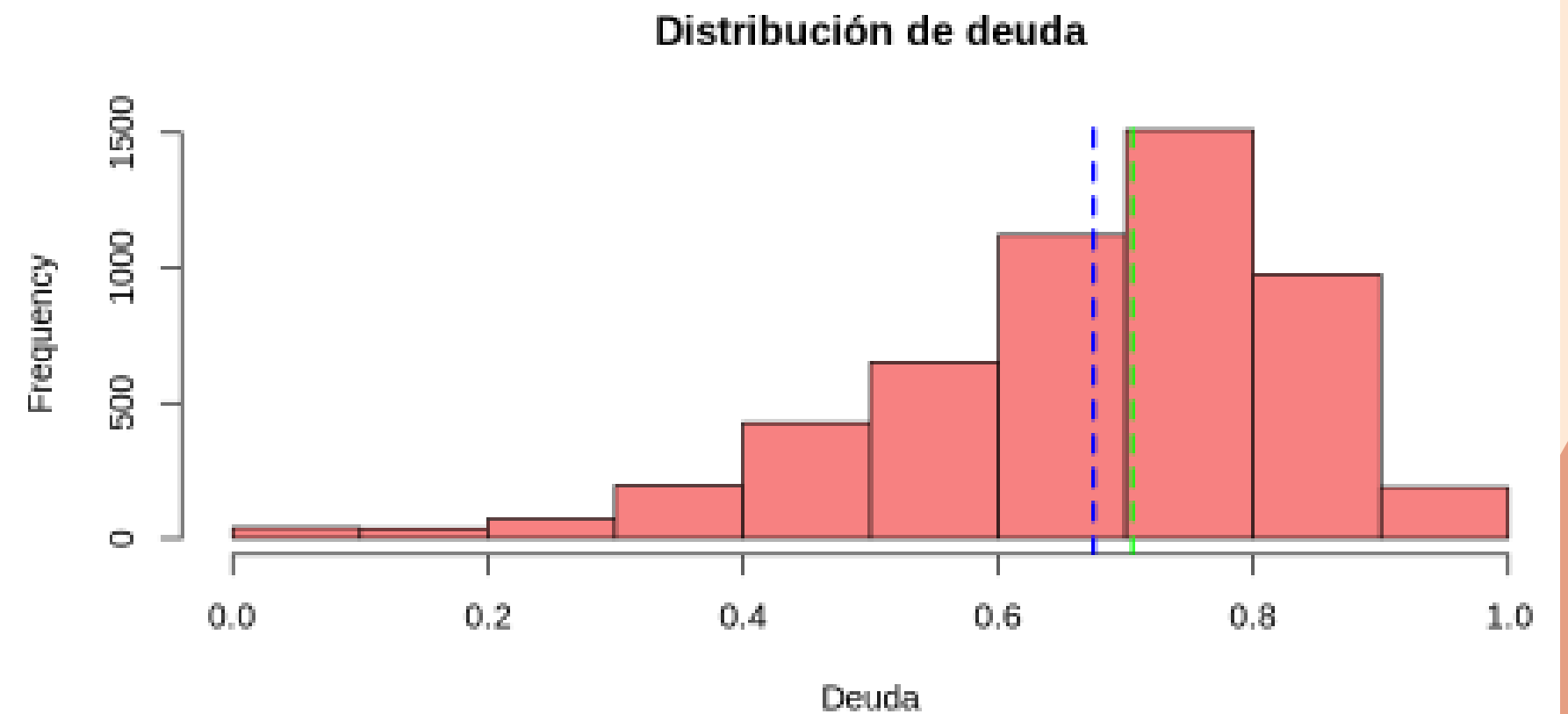
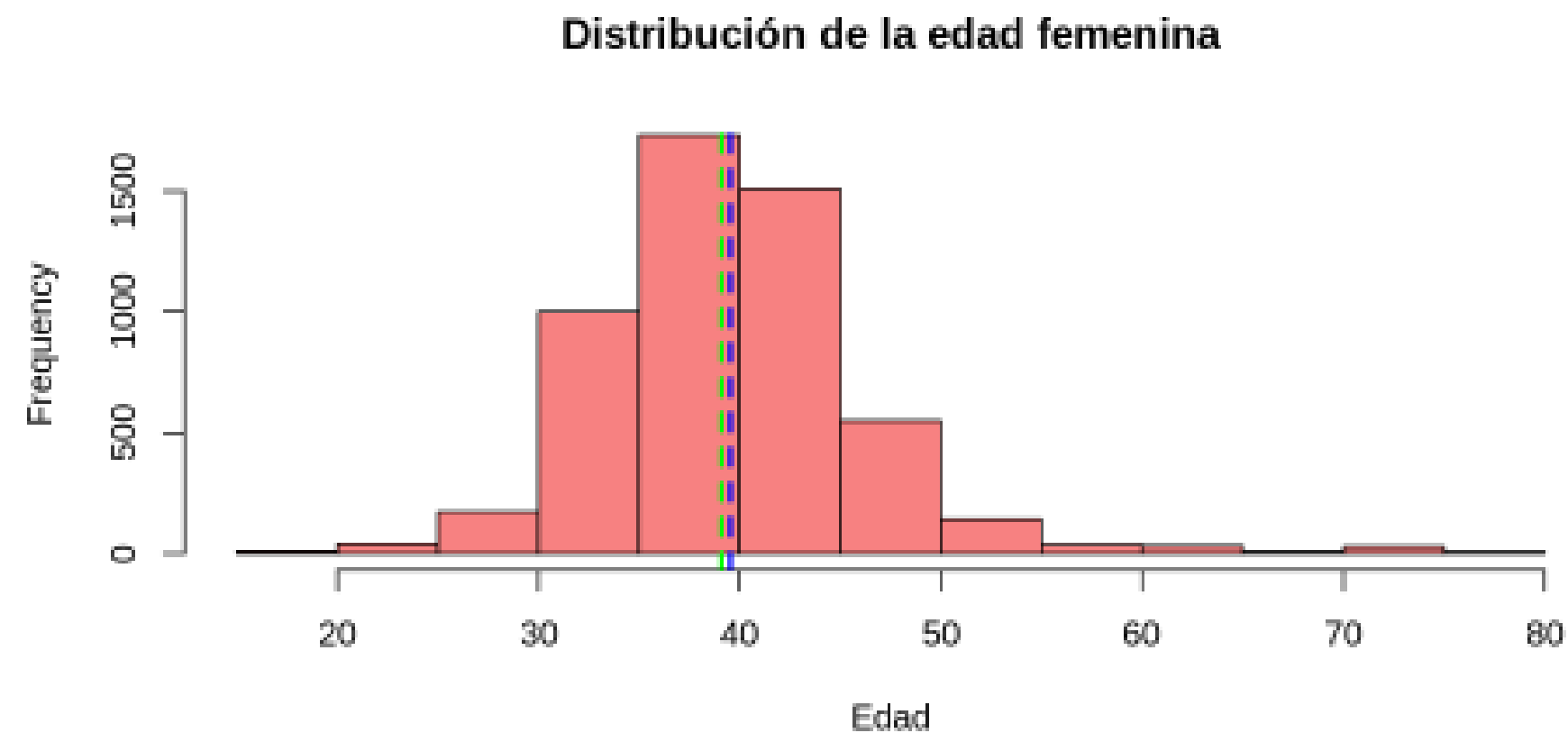
- **State:** Nombre del estado informado por la Oficina del Censo de EE. UU.
- **Male\_pop:** Población masculina en el área
- **Female\_pop:** Población femenina en el área

# DEFINICIONES DE CAMPO ESTADÍSTICO

- **rent\_mean:** La renta bruta media de la ubicación geográfica especificada.
- **rent\_stdev:** La desviación estándar del alquiler bruto para la ubicación geográfica especificada.



Tomaremos variables de nuestra base de datos para poder explicar cada una de las distribuciones, podemos observar que contamos con los tres tipo de distribución principal (normal, sesgo hacia la derecha e izquierda)



Mostramos un grafico para demostrar el sesgo de nuestras distribuciones, sabemos que si la mediana se encuentra en el mismo lugar que la media, contamos con una distribucion normal

	debt	rent_mean	female_age_mean
Min.	0.00000000	181.7723	19.76781
1st Qu.	0.5860650	878.9692	35.40263
Median	0.7054000	1177.0329	39.13049
Mean	0.6749746	1257.1211	39.58056
3rd Qu.	0.7897550	1534.8831	43.02770
Max.	1.00000000	3962.3423	76.99683



# INFERENCIA ESTADISTICA

Varianza: Este es un indicador de cómo se distribuyen nuestros datos

Desviación estándar: Es la raíz cuadrada de nuestra varianza y nos dice qué tan lejos están nuestros datos de la media.

- 1er Cuartil(Q1): Esto está compuesto por el 25% más bajo de números en nuestra distribución.
- 2do Cuartil (Q2): Compuesto por el 50% de los números más bajos hasta la mediana.
- 3rd Cuartil (Q3): Compuesto por el 75% de los números más bajos.
- Rango intercuarti (IQR): Esto nos ayuda a detectar dónde se encuentran la mayoría de los datos.

Ahora haremos una tabla que nos muestre cuartiles y rango intercuartil de cada variable

# INFERENCIA ESTADISTICA

```
[1] "Estadísticas de renta media"
```

```
[1] "-----"
```

mu	rent_med	std	rent_min	rent_max	rent_q1	rent_q3	rent_iqr
1257.125	1177.033	505.5629	181.7723	3962.342	879.2731	1535.102	655.8291

```
[1] "Estadísticas medias de edad femenina"
```

```
[1] "-----"
```

mu	fage_med	std	fage_min	fage_max	fage_q1	fage_q3	fage_iqr
39.52573	39.08989	6.635169	19.4442	76.99683	35.3303	43.00736	7.67706

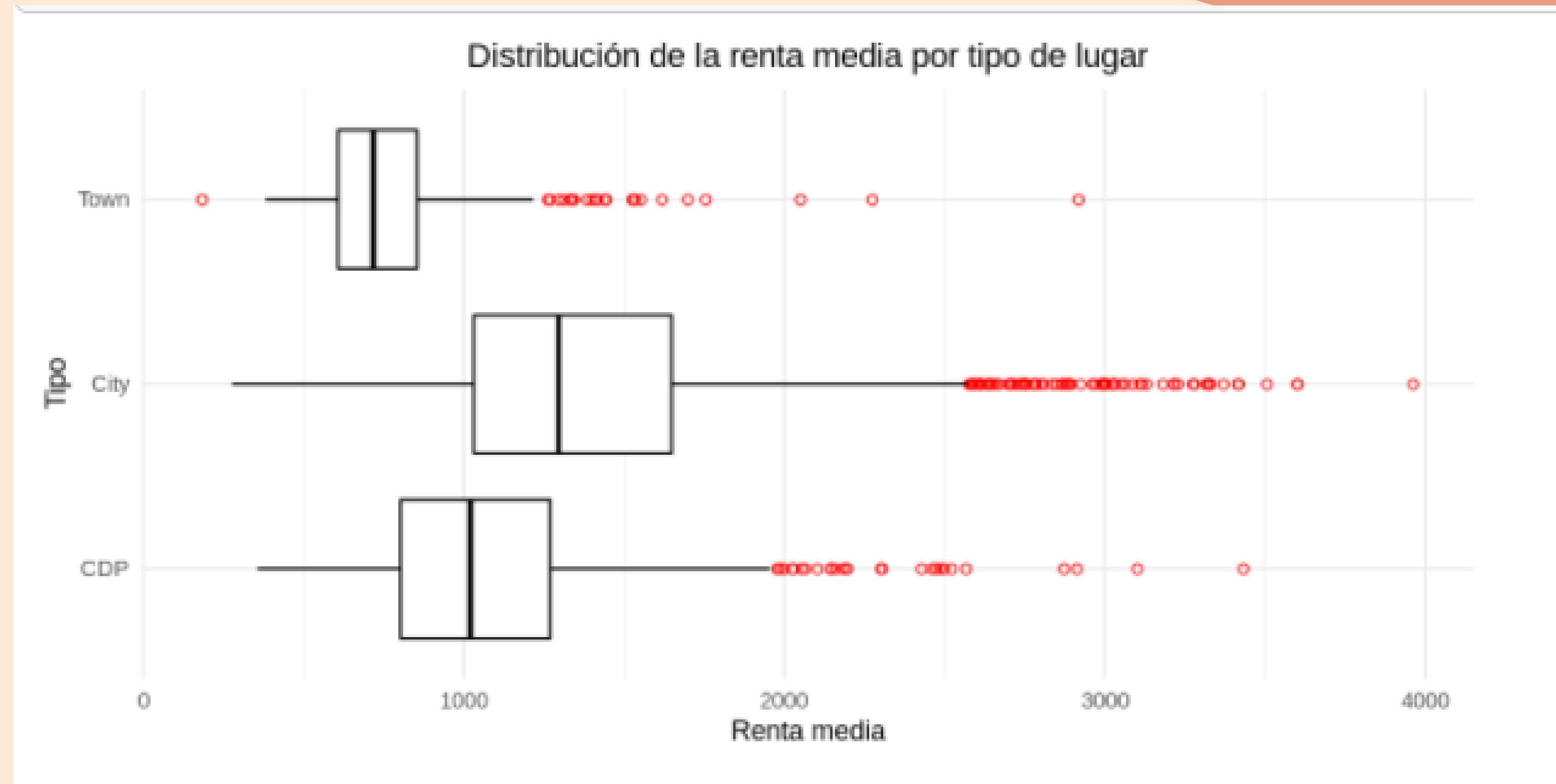
```
[1] "Estadísticas de deuda"
```

```
[1] "-----"
```

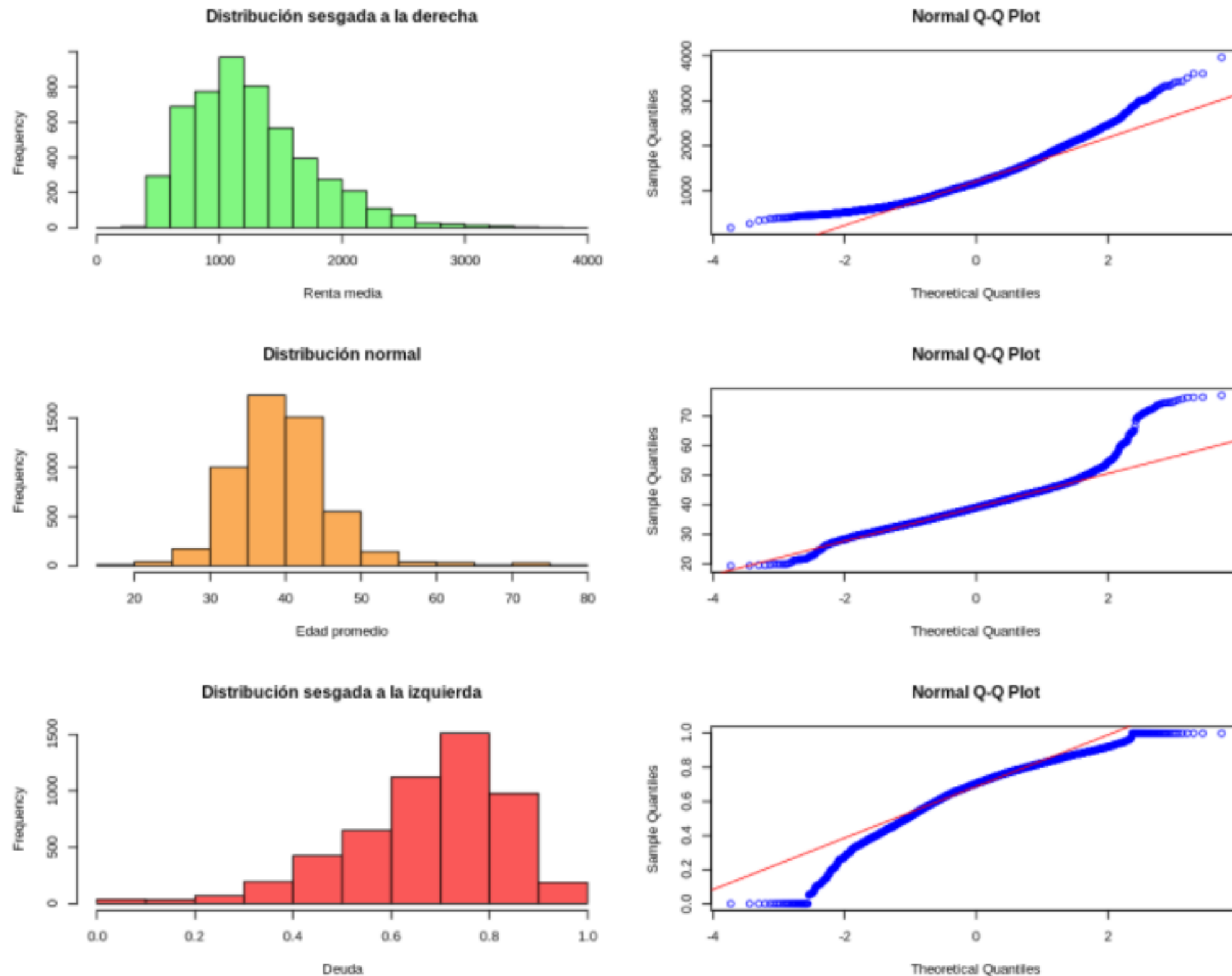
mu	debt_med	std	debt_min	debt_max	debt_q1	debt_q3	debt_iqr
0.6748094	0.70547	0.1646559	0	1	0.585815	0.789845	0.20403

# Diagramas de caja y presuntos valores atípicos.

Los valores atípicos deben analizarse cuidadosamente, cualquier valor más allá de tres desviaciones estándar debe considerarse un valor atípico. Aunque existe una pequeña probabilidad de que un valor en una distribución normal esté a 3 desviaciones estándar de la media, debemos analizar cuidadosamente por qué es así. Podría ser que los datos estuvieran mal escritos, lo que debilitaría la teoría de que una observación específica es un valor atípico.

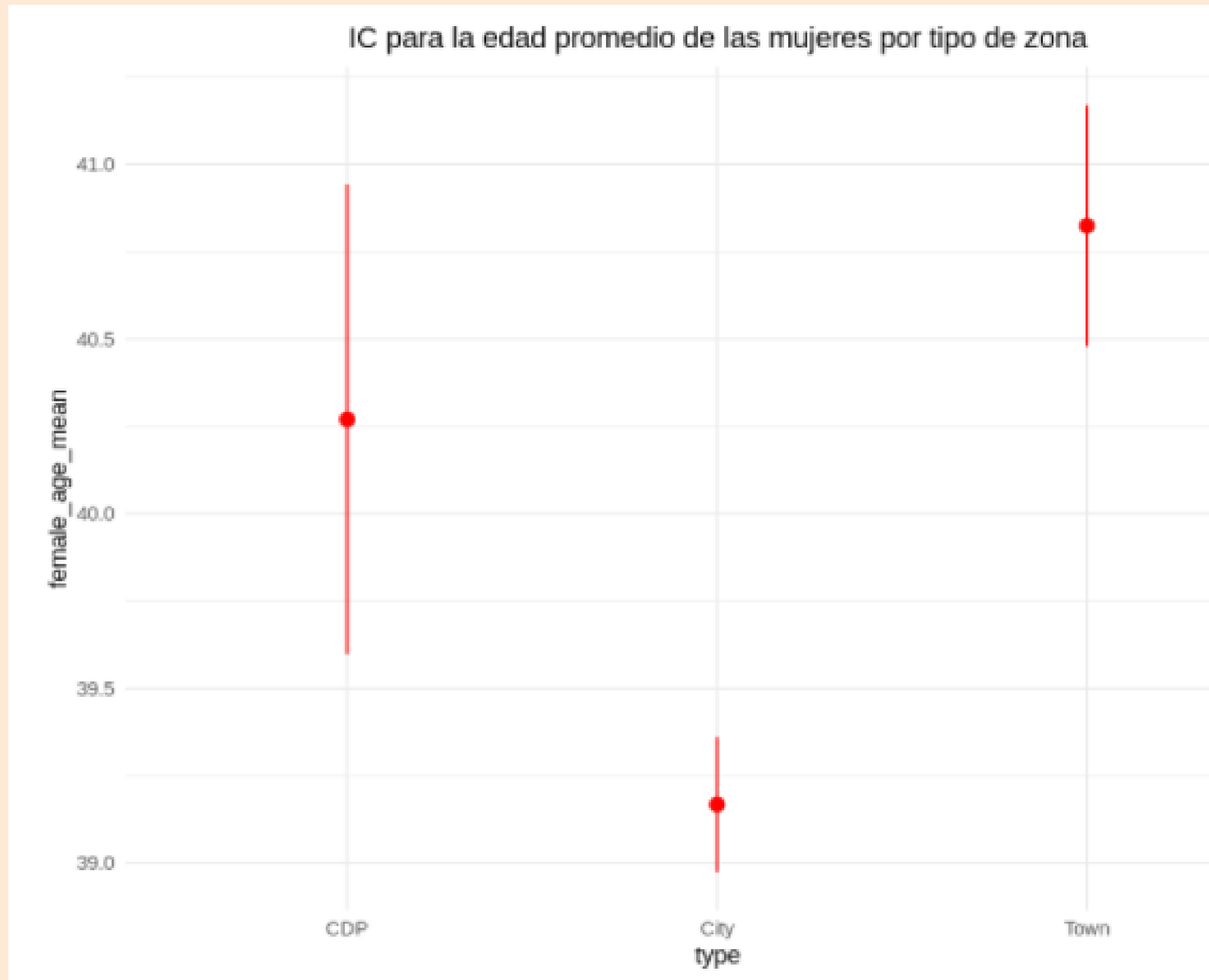


# Diagramas de caja y presuntos valores atípicos.



El uso de logaritmos naturales no necesariamente impacta en las observaciones para formar una distribución normal (ejemplo sesgado a la izquierda), la mayoría de las veces nos da una distribución normal aproximada como en el ejemplo sesgado a la derecha.

# Diagramas de caja y presuntos valores atípicos.



Con los intervalos de confianza, nos aseguramos de nuestra confianza en cuál es el promedio real de la población. Cuanto más anchas sean las barras de error, menos seguros de cuál es la media real.

# Estadísticas de inferencia

## Prueba de hipótesis (¿culpable o no culpable?)

Imagine un escenario en el que un individuo está en un juicio por cometer un asesinato en los Estados Unidos. Hasta donde sabemos, cuando un individuo es considerado "inocente" hasta que se demuestre lo contrario. A través de este breve ejemplo, me gustaría presentar el concepto de Prueba de hipótesis . Hay dos tipos de hipótesis: la hipótesis nula y la hipótesis alternativa.

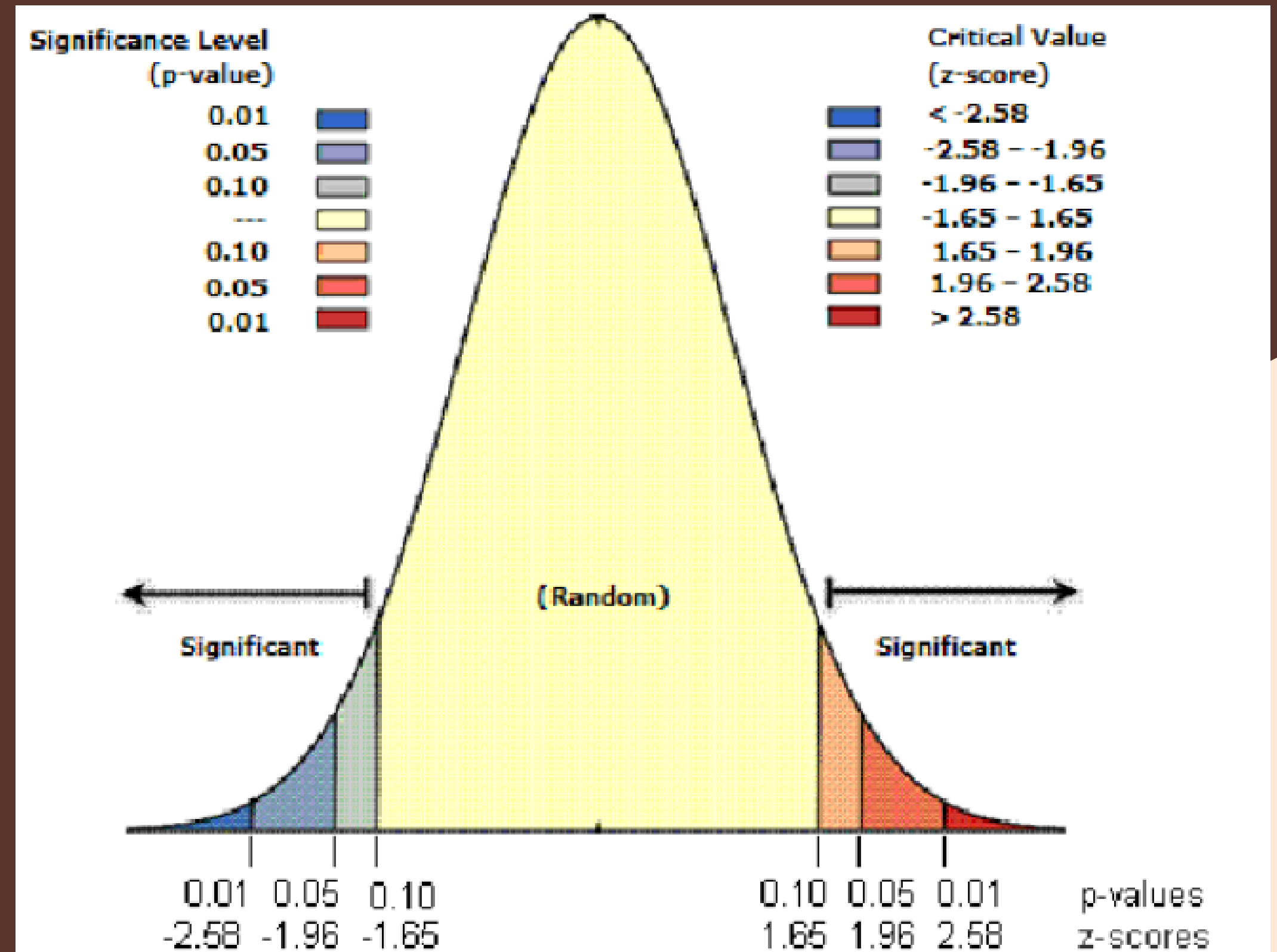
- Hipótesis nula ( $H_0$ ): Este es el "status quo", en nuestro ejemplo, el individuo que está en juicio es inocente . Cuando queremos comparar medias de dos variables, digamos el ingreso promedio de hombres y mujeres, la hipótesis Nula en este caso será que "no hay diferencia".
- Hipótesis alternativa ( $H_a$ ): Esto va en contra de lo que afirmaba la Hipótesis nula. Una persona que fue a juicio es culpable . En nuestro ejemplo de ingresos promedio por género, el ingreso promedio de los hombres no es igual al ingreso promedio de las mujeres.

# Estadísticas de inferencia

**Los intervalos de confianza (IC):** son la certeza de que un valor específico se ubicará entre dos puntos específicos. Los tipos de intervalos de confianza más habituales son los intervalos de confianza del 90%, 95% y 99% aunque el 95% es el que más se utiliza y es el que usaremos en este ejemplo.

**Valor p:** es la probabilidad de que ocurra un evento dado. Suponiendo que un intervalo de confianza es del 95%, si el valor  $p < \alpha$ , rechazamos la Hipótesis nula a favor de la Hipótesis alternativa.

**Nivel de significancia:** es la probabilidad de rechazar la Hipótesis nula (también denotada como  $\alpha$ ).





# Hallar el intervalo de confianza para la media poblacional:



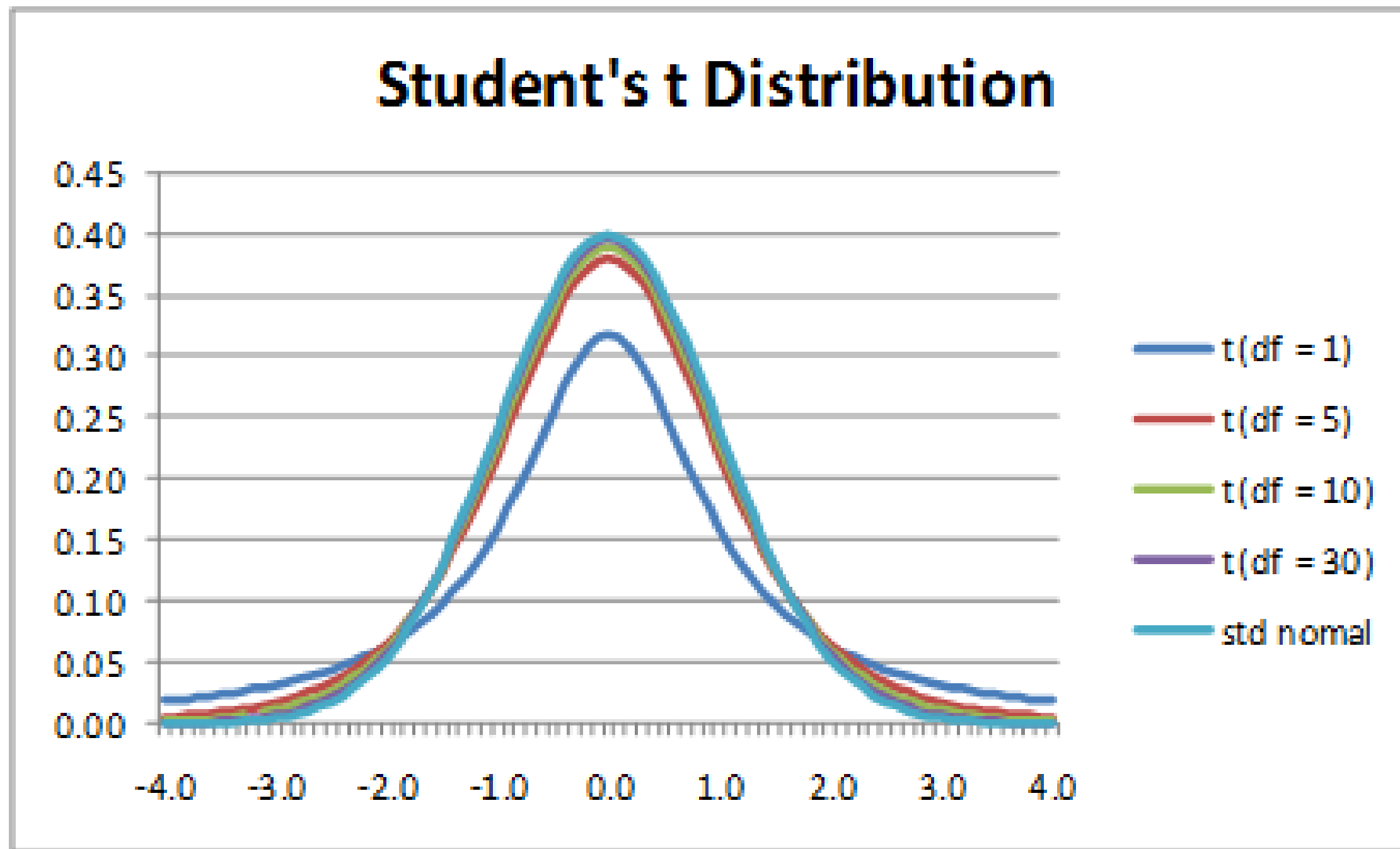
- Número de filas después de filtrar valores nulos  
5262
- El intervalo inferior es  
39.35.
- El intervalo superior es:  
39.71

El 95% de las muestras aleatorias de un tamaño de muestra de 38,728 (female\_age\_mean sin Nulls) de mujeres estadounidenses producirán intervalos de confianza que capturan la media de edad real de la población de mujeres. (40.2 - 40.32)

En este caso, dado que el p-valor  $> \alpha$  no rechazamos  $H_0$



# Distribución T

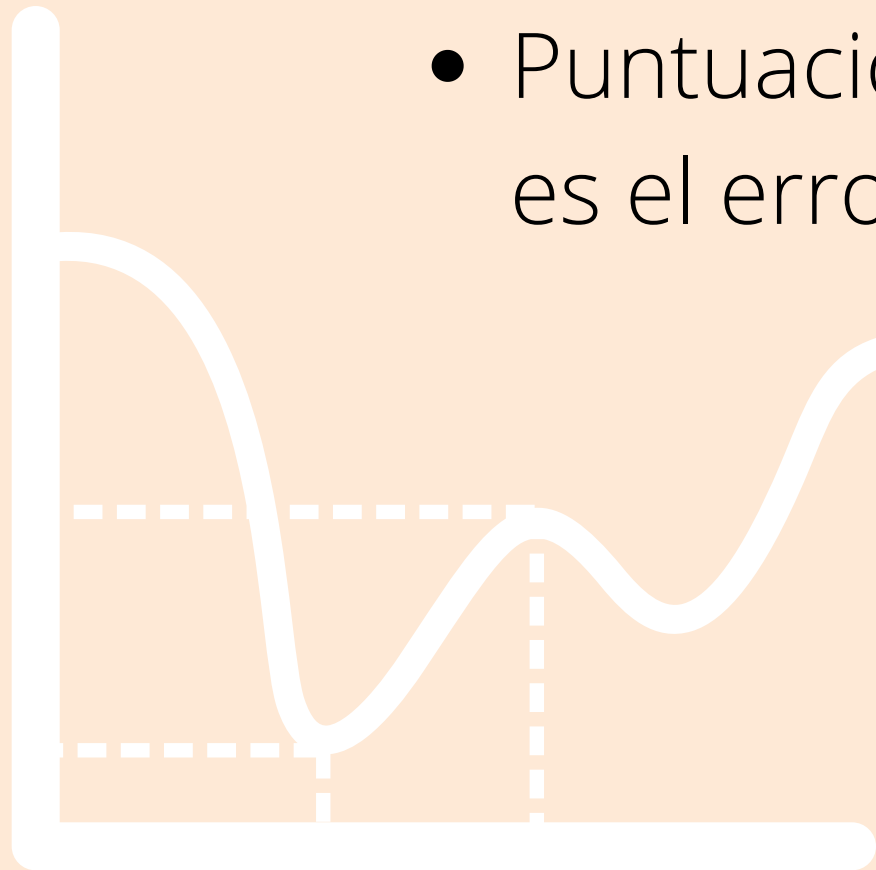


La distribución  $t$  es una distribución que solo se usa para muestras pequeñas. La primera pregunta que tuvimos al tratar con este tipo de distribución es por qué necesitamos una distribución  $t$  cuando recibimos toneladas diarias de datos, lo que hace imposible tener una muestra pequeña. Bueno, las distribuciones  $t$  se usan con más frecuencia cuando se realiza un experimento que suele tener muestras más pequeñas.

# Distribución T

Distribución t:

- Tamaño de la muestra: El tamaño de la muestra debe ser menor que 30 para ser considerado para una distribución t.
- Grados de libertad: A medida que el tamaño de la muestra se acerca a 30, la distribución t se verá exactamente como una distribución normal. Además, los grados de libertad determinan el grosor de la cola.
- Puntuación T: Para calcular la puntuación t usamos la fórmula, donde  $s$  es el error estándar y  $\mu_0$  la hipótesis nula.



# Intervalo de confianza del 95% de una muestra para alquiler en el estado de Nueva York.

**Ejercicio 1:** Hallemos el intervalo de confianza del 95% de una muestra para alquiler en el estado de Nueva York.

Tomamos los grados de libertad de nuestra  $n-1$ , nos resulta una  $t= 0.0025$  de nuestra tabla t student.

Obteniendo el promedio de alquileres del estado de Nueva York, un dicho tamaño de muestra obtenemos con un 95% de confianza en que la renta media de la ciudad de Nueva York se encuentra entre 934 y 1417.



# El alquiler promedio en Nueva York es algo diferente a 1000.

**Ejercicio 2:** Supongamos que  $\mu = 1000$  para el alquiler de Nueva York, encontremos el valor  $p$  para ver si hay suficiente evidencia de que podríamos rechazar  $H_0$  a favor de  $H_A$ , recuerde si el valor  $p$  es menor que 0.05 nivel de significancia luego rechazamos el Nulo a favor de la alternativa

Dado que nuestro valor  $p$  es mayor que 0.05, nos inclinamos a favor de la Hipótesis nula, lo que significa que no hay evidencia suficiente de que el alquiler promedio en Nueva York sea algo diferente a 1000.



# CONCLUSION

Con la estadística inferencial pudimos crear muestras aleatorias simples para poder determinar hipótesis que teníamos acerca de la base de datos.

Creemos que esta sección puede ser menos precisa a la hora de inferir o predecir, pero puede ser de gran ayuda como dato a la hora de toma de decisiones rápidas pero inteligente.

Podemos llevarnos como tarea en un futuro crear mejores técnicas de muestreo para evitar sesgos de precisión que pudieran provocar falta de precisión a la hora de modelos de predicción.

Teniendo en cuenta el punto anterior podemos aplicar más técnicas de limpieza de datos para mejorar cada inferencia para no perder mucha información pues vimos que en las fórmulas para realizar los intervalos de confianza y muestreo los datos NA simplemente fueron eliminados y no fueron transformados, como vimos en la gráfica estos mismos datos están normalmente distribuidos, pero podrían darnos información errónea el hecho de simplemente eliminarlos

# BIBLIOGRAFIA.



- Rincón, L. (2019, 5 octubre). *Una introducción a la estadística inferencial*. UNAM. Recuperado 6 de noviembre de 2021, de <https://lya.fciencias.unam.mx/lars/Publicaciones/ei2019.pdf>
- K. (2012, 28 julio). *Estadística inferencial*. Slideshare. Recuperado 6 de noviembre de 2021, de <https://es.slideshare.net/katemora/proyecto-estadistica-inferencial-13786749>
- Moreno, J. (2020, 14 septiembre). *Golden Oaks, Florida. ¡Tu casa en Walt Disney World!* El Souvenir. Recuperado 5 de noviembre de 2021, de <https://elsouvenir.com/golden-oaks-florida-tu-casa-en-walt-disney-world/>
- Wikipedia contributors. (2021, 12 agosto). *Golden Oak at Walt Disney World Resort*. Wikipedia. Recuperado 5 de noviembre de 2021, de [https://en.wikipedia.org/wiki/Golden\\_Oak\\_at\\_Walt\\_Disney\\_World\\_Resort](https://en.wikipedia.org/wiki/Golden_Oak_at_Walt_Disney_World_Resort)