

Julian Andres Ramirez Jimenez, Samuel David Villegas Bedoya, Isaac Tadina Giraldo, Juan Jose Sanchez Cortes, Jose Manuel Fonseca Palacio

Problema 1. (14pts.) Identifique los tipos de variables que aparecen en la data y determine si es categórica o cuantitativa y si es categórica, determine si es nominal u ordinal. Si es cuantitativa, determinar si es discreta o continua. Complete la tabla.

Variable	Categórica o Cuantitativa	Clasificación
Age	Cuantitativa	Discreta
Attrition	Categórica	nominal
Department	Categórica	Nominal
Education	Categórica	Ordinal
EducationField	Categórica	Nominal
EnvironmentSatisfaction	Categórica	Ordinal
Gender	Categórica	Nominal
HourlyRate	Cuantitativa	Discreta
JobRole	Categórica	Nominal
MaritalStatus	Categórica	Nominal
MonthlyIncome	Cuantitativa	Discreta
RelationshipSatisfaction	Categórica	Ordinal
WorkLifeBalance	Categórica	Ordinal
YearsSinceLastPromotion	Cuantitativa	Discreta

Cuadro 1: Problema 1

Problema 2. (30pts.) Construya tablas de frecuencias, gráficos y medidas de centralidad y dispersión apropiados para cada una de las siguientes variables Age, Education y MonthlyIncome. Comente los resultados que considere relevantes, respondiendo por lo menos a las siguientes preguntas (podrían señalar más cosas relevantes en el último ítem). No olvide respaldar los comentarios de las gráficas, tablas y medidas de centralidad y dispersión.

a. ¿Cuál es la clase modal de cada una de las variables?

Variable	Clase modal	Comentario
Age	[28.52, 32.08)	Al realizar la tabla de frecuencia por intervalos podemos decir que el rango de edad que tiene mayor frecuencia es [28.52, 32.08) siendo esta la clase modal.

Education	3 (graduado)	Al obtener la frecuencia relativa de cada nivel educativo, obtuvimos que la que tiene mayor frecuencia es el nivel educativo 3 (graduado).
MothlyIncome	[2598.91, 4198.92)	Al realizar la tabla de frecuencia por intervalos podemos decir que el rango de sueldos que tiene mayor frecuencia es [2598.91, 4198.92) siendo esta la clase modal.

Cuadro 2: Problema 2a

b. ¿Cuál es la clase con menos probabilidad de ocurrir y cuál es esa probabilidad para cada una de esas variables?

Variable	Clase	Probabilidad	Comentario
Age	[57.03,60.6)	2%	De las personas encuestadas, el intervalo de edad menos probable que pueda ocurrir es entre 57 a 60 años
Education	5 (Doctorado)	3.26%	El nivel educativo menos frecuente entre las personas censadas es Doctorado, con un 3.26%
MothlyIncome	[13798.96, 15398.97)	1%	Un salario mensual entre \$13798.96 y \$15398.97 es de menor probabilidad respecto a todos los salarios registrados.

Cuadro 3: Problema 2b

Tabla de:

- Edad

Class limits	f	rf	rf(%)
[17.82,21.39)	41	0.03	2.79
[21.39,24.95)	56	0.04	3.81
[24.95,28.52)	161	0.11	10.95
[28.52,32.08)	258	0.18	17.55
[32.08,35.64)	213	0.14	14.49
[35.64,39.21)	219	0.15	14.90
[39.21,42.77)	143	0.10	9.73
[42.77,46.34)	139	0.09	9.46
[46.34,49.91)	67	0.05	4.56
[49.91,53.47)	86	0.06	5.85
[53.47,57.03)	58	0.04	3.95
[57.03,60.6)	29	0.02	1.97

- Education:

1	2	3	4	5
0.11564626	0.19183673	0.38911565	0.27074830	0.03265306

Tabla de frecuencias relativas

- MonthlyIncome:

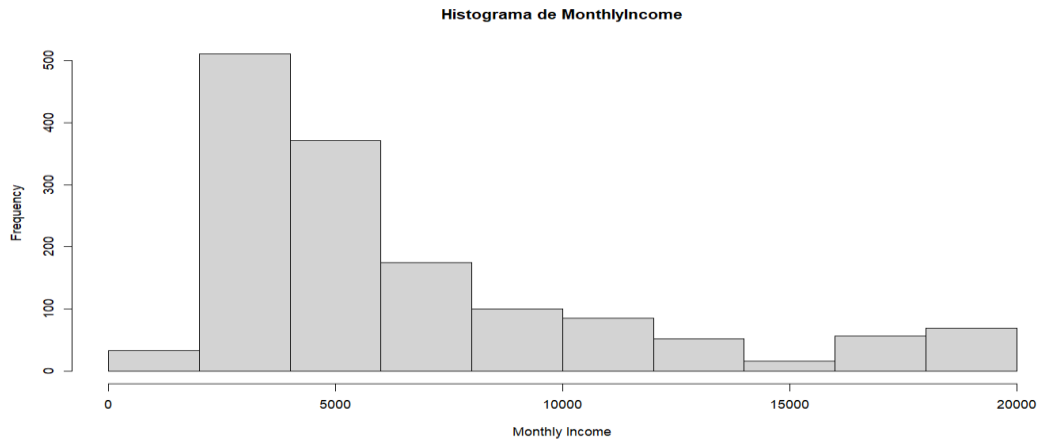
Class limits	f	rf	rf(%)
[998.91,2598.9167)	261	0.18	17.76
[2598.9167,4198.9233)	323	0.22	21.97
[4198.9233,5798.93)	309	0.21	21.02
[5798.93,7398.9367)	157	0.11	10.68
[7398.9367,8998.9433)	87	0.06	5.92
[8998.9433,10598.95)	90	0.06	6.12
[10598.95,12198.957)	52	0.04	3.54
[12198.957,13798.963)	44	0.03	2.99
[13798.963,15398.97)	16	0.01	1.09
[15398.97,16998.977)	34	0.02	2.31
[16998.977,18598.983)	37	0.03	2.52
[18598.983,20198.99)	60	0.04	4.08

c. Si suponemos forma acampanada en el histograma de las variables Age y MonthlyIncome ¿Cuál es el percentil 84 (P84) y el percentil 97.5 (P97.5) para cada una de las variables?

Variable	P ₈₄	P _{97,5}	Comentario
Age	47.64	56.225	El 84% de las personas tienen a lo sumo 47 años El 97.5% de las personas tienen a lo sumo 56 años
MonthlyIncome	10725.64	19197.22	El 84% de las personas censadas ganan un salario mensual de máximo \$10725.64 El 97.5% de las personas censadas ganan un salario mensual de máximo \$19197.22

Cuadro 4: Problema 2c

d. Si **NO** suponemos forma acampanada en el histograma de las variables Age y MonthlyIncome ¿Cuáles son los extremos del intervalo de al menos el 75% y 88,89% central de la información, para cada una de las variables?



Variable	Extremo inferior	Extremo superior	Comentario
Age Intervalo 75%	18.65306	55.19456	Los intervalos nos permiten apreciar que el 75% de la poblacion se encuentran entre los 18-19 años y los 55-56 años. Esto es como tal la edad promedio trabajadora.
MothlyIncome Intervalo 75%	-2912.982	15918.84	La cota inferior es negativa por lo cual podemos asumir que debido a que hay tantos datos extremos. El 75% de los datos se encuentran a partir de 0, hasta el extremo superior. Así infiriendo que, la mayoría de los datos se encuentran en los niveles de ganancia mensual más baja.
Age Intervalo 88,89%	9.5271689	66.32993	Los intervalos nos permiten apreciar que el 88% de la poblacion se encuentran entre los 9-10 años y los 66 años. Esto, nos permite apreciar, que tal vez, al no ser acampanados, chebychev tira un estimado que pueda abarcar el 88% a pesar de que los 9 años no es una edad trabajadora. Pero a su vez permite apreciar que a los 66 aproximadamente toda la poblacion empieza a dejar de trabajar.
MothlyIncome Intervalo 88,89%	-7620.939	20626.8	Se realiza el grafico, y podemos ver que la media queda en un punto muy bajo, se concluye y se puede observar que el límite quedaría en el negativo

Cuadro 5: Problema 2d

- e. De las tablas de frecuencias, gráficos y medidas de centralidad y dispersión apropiados para cada una de las variables Age, Education y MonthlyIncome ¿Qué más pudo observar que no se haya señalado en los anteriores items?

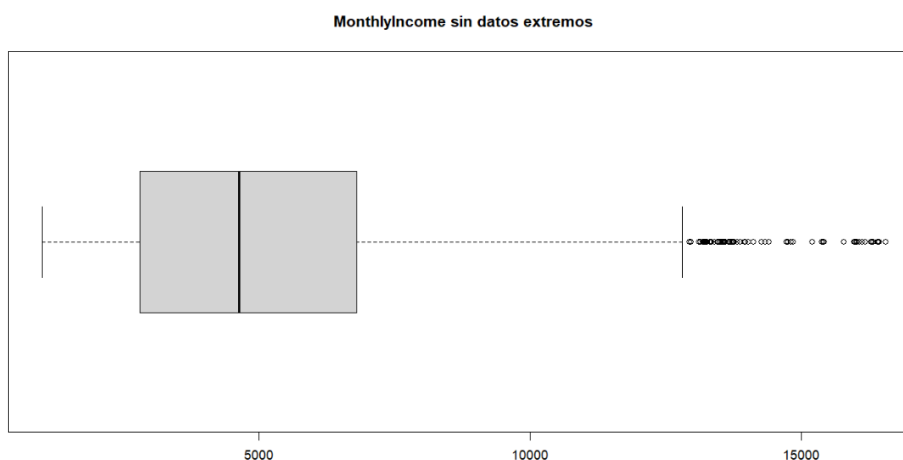
Variable	Comentario
Age	Según las edades registradas, las personas censadas son en su mayoría adultas medianamente jóvenes, el comportamiento es regular, es decir, no se registran edades muy avanzadas con respecto a la media del grupo.
Education	La tendencia es el nivel de Graduado (3), pero es importante mencionar que hay una cantidad considerable de personas que han logrado el título de maestría (4). El nivel de doctorado es el menos usual.
MonthlyIncome	La gran mayoría de los datos se encuentran en puntos muy bajos (mencionar lo de los datos extremos superiores)

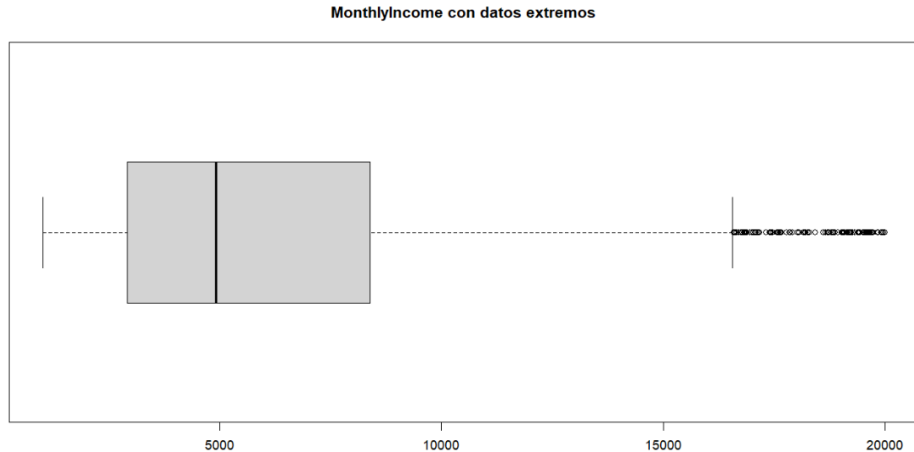
Cuadro 6: Problema 2e

Problema 3. (22pts.) Determine cuantos datos pueden ser considerados extremos para la variable MonthlyIncome. Elimine los datos atípicos (una sola vez) y no los considere para el siguiente análisis. Si al grupo de datos sin datos extremos lo llamamos B y al inicial A

- a. Realice un boxplot del MonthlyIncome sin los datos extremos y compárelos con los del **Problema 2**.

```
> nrow(data[data$MonthlyIncome > 16581,])
[1] 114
```





b. Complete la siguiente tabla y comente de manera comparativa.

```
> summary(data$MonthlyIncome) # Sin corte
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1009   2911   4919   6503   8379  19999
> summary(data[data$MonthlyIncome<=16581,]$MonthlyIncome) # Con corte
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1009   2813   4643   5503   6811  16555
> |
```

Variable	Sin Extremos	Con Extremos	Comentario
Min	1009	1009	debido a que no hay extremos inferiores.
Q1	2813	2911	A pesar de que hay poca diferencia en Q1(Alrededor de 100) para el dataset sin extremos es menor, debido a que al no tener datos extremos hay menos dispersión en los datos.
Q3	6811	8379	Para Q3 se aplica el mismo razonamiento de Q1, solo que en este caso es mucho más notoria la diferencia (Alrededor de 1500).
Max	16555	19999	Como en el dataset original tenemos extremos superiores y estos fueron los que se eliminaron, podemos observar que al realizar el corte el valor maximo en el dataset sin extremos es mucho menor al dataset con extremos.

Cuadro 7: Problema 3b

c. Calcule medidas de tendencia central para la variable y compárelos con los resultados del **Problema 2**. Comente los resultados.

```

> freq_monthly_income_corte
      Class limits    f    rf rf(%)
[998.91,2309.0467) 144 0.11 10.62
[2309.0467,3619.1833) 344 0.25 25.37
[3619.1833,4929.32) 247 0.18 18.22
[4929.32,6239.4567) 206 0.15 15.19
[6239.4567,7549.5933) 121 0.09 8.92
[7549.5933,8859.73) 69 0.05 5.09
[8859.73,10169.867) 64 0.05 4.72
[10169.867,11480.003) 66 0.05 4.87
[11480.003,12790.14) 21 0.02 1.55
[12790.14,14100.277) 47 0.03 3.47
[14100.277,15410.413) 11 0.01 0.81
[15410.413,16720.55) 16 0.01 1.18
> |
      Class limits    f    rf rf(%)
[998.91,2598.9167) 261 0.18 17.76
[2598.9167,4198.9233) 323 0.22 21.97
[4198.9233,5798.93) 309 0.21 21.02
[5798.93,7398.9367) 157 0.11 10.68
[7398.9367,8998.9433) 87 0.06 5.92
[8998.9433,10598.95) 90 0.06 6.12
[10598.95,12198.957) 52 0.04 3.54
[12198.957,13798.963) 44 0.03 2.99
[13798.963,15398.97) 16 0.01 1.09
[15398.97,16998.977) 34 0.02 2.31
[16998.977,18598.983) 37 0.03 2.52
[18598.983,20198.99) 60 0.04 4.08

```

Variable	Sin Extremos	Con Extremos	Comentario
Media	5503	6503	Gracias a la tabla con los datos, nos podemos dar cuenta que la mayoría de los datos se encuentran entre los intervalos de [998.91,623934567), esto hace que la gráfica se vaya moviendo hacia la izquierda, generando así, que la mediana quede más a la izquierda que la media.
Mediana	4643	4719	
Punto medio clase modal	2964.115	3398.92	A partir punto medio de la clase modal, se puede inferir que los datos extremos están afectando al dataset con extremos por lo que este punto es mucho mayor que con el dataset sin extremos.

Cuadro 8: Problema 3c

- d. Señale con una × ¿En qué grupo A o B es más probable encontrar profesionales que tengan ingresos superiores por encima de 6811 dólares al mes?

Con Extremos (A) X	Sin Extremos (B)
---------------------------	-------------------------

Cuadro 9: Problema 3d

- e. Señale con una × ¿En qué grupo A o B es más probable ganar menos de 2758 dólares al mes?

Con Extremos (A)	Sin Extremos (B) X
-------------------------	---------------------------

Cuadro 10: Problema 3e

- f. Señale con una × ¿Qué medidas estadísticas o medida estadística son o es suficiente para responder con precisión a la pregunta del ítem d. y para qué grupo de datos (A o B)? Comente porqué es suficiente su elección.

```

summary(data[data$MonthlyIncome<=16581,]$MonthlyIncome) # Con corte
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1009   2813   4643   5503   6811  16555

```

```
summary(data$MonthlyIncome) # Sin corte
Min. 1st Qu. Median Mean 3rd Qu. Max.
1009 2911 4919 6503 8379 19999
```

Medida estadística	Grupo A	Grupo B	Comentario
Q1		--	Para el caso del grupo B solo sería necesario el Q3 ya que este coincide directamente con el valor a partir del cual se busca encontrar los mayores o iguales. Para el grupo de datos A podríamos utilizar el Q2, ya que es un valor cercano a 6811, y además corresponde a la mediana
Q2	X	--	
Q3		X	
Media			
Desviación Estándar	--	--	

Cuadro 11: Problema 3f

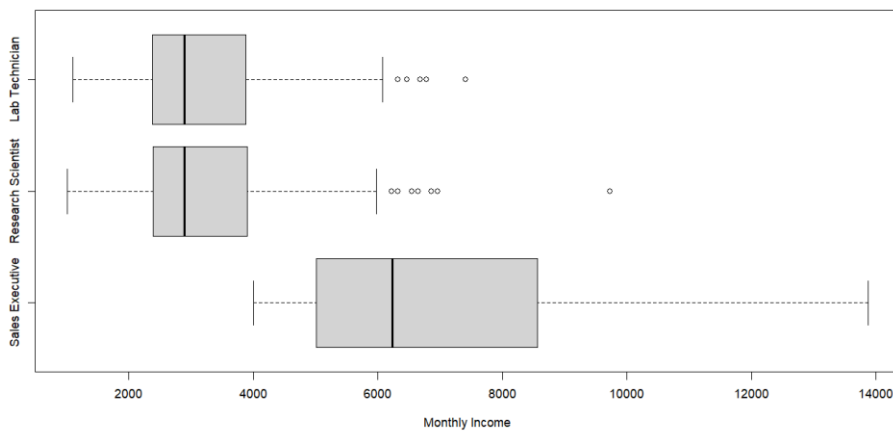
- g. Señale con una × ¿Qué medidas estadísticas o medida estadística son o es suficiente para responder con precisión a la pregunta del ítem e. y para qué grupo de datos (A o B)? Comente porqué es suficiente su elección.

Medida estadística	Grupo A	Grupo B	Comentario
Q1	X	X	Se marcaron los Q1 en ambos grupos ya que el grupo de personas que estamos buscando se encuentran dentro de ese cuartil.
Q2	--	--	
Q3	--	--	
Media	--	--	
Desviación Estándar	X	X	Como los datos están más agrupados en el dataset B, es más fácil que se puedan encontrar más datos de los que necesitamos, porque gracias a la desviación estándar vemos que estos datos están mucho cercanos/agrupados que en el dataset A.

Cuadro 12: Problema 3g

Problema 4. (21pts.) Realice la comparación gráfica más apropiada e informativa de la distribución del salario mensual de los tres roles de trabajo con mayor frecuencia y luego conteste la siguientes preguntas.

```
> prop_range_2600_7400_sales
Sales Executive
0.6503067
> prop_range_2600_7400_research
Research Scientist
0.6369863
> prop_range_2600_7400_laboratory
Laboratory Technician
0.6138996
> prop_range_9000_sales
Sales Executive
0.2116564
> prop_range_9000_research
Research Scientist
0.003424658
> prop_range_9000_laboratory
[1] 0
> |
```



a. ¿Cuál es este top tres? Complete la siguiente tabla con los tres nombres de los roles, organizados de mayor a menor frecuencia absoluta y señale su frecuencia respectiva.

Healthcare Representative	Human Resources	Laboratory Technician	Manager
131	52	259	102
Manufacturing Director	Research Director	Research Scientist	Sales Executive
145	80	292	326
Sales Representative			
83			

Rol de trabajo	Frecuencia absoluta	Comentario
Sales Executive	326	Sacando la frecuencia absoluta de cada uno de los roles como se aprecia en la tabla anterior, podemos decir que de los datos que se tienen, los roles que más repiten son Sales Executive, Research Scientist y Laboratory Technician
Research Scientist	292	

Laboratory Technician	259
-----------------------	-----

Cuadro 13: Problema 4a

- b. Según la gráfica comparativa qué rol de trabajo tiene una mayor proporción de personas con salario mensual entre 2600 y 7400 usd?

Rol de trabajo	Proporción de interés	Comentario
Sales Executive	~0.67	Según la interpretación del boxplot para Research Scientist y Laboratory Technician la proporción de los datos que se encuentran en el rango 2600 a 7400, es de un poco más del 50% Según la interpretación del boxplot para Sales Executive la proporción de datos que se encuentran en el rango 2600 a 7400, es de un poco más del 60%
Research Scientist	~0.55	
Laboratory Technician	~0.55	

Cuadro 14: Problema 4b

- c. Determine la proporción de salario mensual por cada rol de trabajo que están entre los 2600 y 7400 (incluidos). Luego, compare las tres proporciones y señale la mayor, corrobore con el resultado del item b.

Rol de trabajo	Proporción de interés	Comentario
Sales Executive	0,6503067	Según los resultados obtenidos de forma analítica el rol de trabajo que tiene mayor proporción de salarios entre 2600 a 7400 es Sales Executive, lo que coincide con nuestra interpretación de la gráfica. Con Research Scientist y Laboratory Technician nos acercamos con la estimación, pero no fue tan precisa.
Research Scientist	0,636986	
Laboratory Technician	0,613899	

Cuadro 15: Problema 4c

- d. Según la gráfica comparativa qué rol de trabajo tiene una mayor proporción de personas con salarios mayores a los 9000 usd?

Rol de trabajo	Proporción de interés	Comentario
Sales Executive	~0.25	Según la interpretación del boxplot para Research Scientist y Laboratory Technician la proporción de los datos mayores a 9000,

Research Scientist	~0.01
Laboratory Technician	~0.01

es de un poco menos del 1%, ya que estos serían datos extremos para estos dos casos.

Según la interpretación del **boxplot** para Sales Executive la proporción de datos mayores a 9000, es de alrededor del 25%

Cuadro 16: Problema 4d

- e. Determine la proporción por cada rol de trabajo de salarios mensuales por encima de los 9000 (incluidos). Luego, compare las tres proporciones y señale la mayor, corrobore con el resultado del ítem d.

Rol de trabajo	Proporción de interés	Comentario
Sales Executive	0.21	Según la solución analítica Research Scientist y Laboratory Technician tienen una proporción menor a 1% lo que coincide con nuestra interpretación de la gráfica. Asimismo, en Sales Executive realizamos una buena interpretación y estuvimos muy cerca del valor esperado.
Research Scientist	0.003	
Laboratory Technician	0	

Cuadro 17: Problema 4e

Problema 5. (27pts.) El salario mensual es considerado elevado si está por sobre 2 desviaciones estándar de la media. Considerando lo anterior como definición.

- a. Si al salario que está sobre dos desviaciones estándar de la media lo notamos por $\mu + 2\sigma$ ¿Cual es el porcentaje de salarios mensuales elevados en la data?. Utilice la data con (data A) y sin datos (data B) extremos (determinados en el problema 3). Compare cada uno de los resultados con lo visto en clase sobre el teorema de Chebyshev.

```
> percentil_75_monthly_income_sup
[1] 15918.84
> percentil_75_monthly_income_sup_corte
[1] 12142.47
> porc_mayores_sin_corte
[1] 0.08707483
> porc_mayores_corte
[1] 0.1423304
>
```

Conjunto de datos	$\mu + 2\sigma$	% salarios elevados	Comentario
Data A	15918.84	8.707483	Los salarios elevados en la Data A son muy pocos ya que el punto en el que se consideran salarios elevados se encuentra más alto debido a que, los datos extremos, mueven ese límite, por esa razón es que encontramos pocos salarios elevados.
Data B	12142.47	14.23304	

--	--	--

Mientras en la Data B, al no haber estos datos extremos, el punto en el que se consideran salarios elevados baja, y por ende es mas posible encontrar personas en esta categoría.

Cuadro 18: Problema 5a

- b. Determine el estado civil (MaritalStatus) cuyo salario mensual supera con mayor frecuencia relativa el umbral de las dos desviaciones estándar del promedio. Para responder este punto, utilice la data con datos extremos, donde US será el umbral en dolares desde donde se cuentan los sueldos elevados y n es el número de personas en cada estado civil.

```
> n_marital
Divorced Married Single
      327      673      470
> perc_75_sup_divorced
[1] 16477.58
> perc_75_sup_married
[1] 16513.04
> perc_75_sup_single
[1] 14540.61
> num_per_divorced
[1] 26
> num_per_married
[1] 63
> num_per_single
[1] 31
> freq_per_divorced
Divorced
0.0795107
> freq_per_married
Married
0.0936107
> freq_per_single
Single
0.06595745
> |
```

Estado civil	US	n	# de personas con salario elevado	Frecuencia relativa al estado civil	Comentario
Married	16513.04	673	63	0.093	Es más probable encontrar una persona casada en la categoría de salario elevado. Por su parte, la probabilidad de encontrar alguien soltero con salario elevado es aún mas baja que el resto.
Divorced	16477.58	327	26	0.079	
Single	14540.61	470	31	0.065	

Cuadro 19: Problema 5b

- c. Determine el genero que con mayor frecuencia supera el umbral dos desviaciones del promedio de salarios mensuales. Para responder este punto, utilice la data con datos extremos, donde US será el umbral en dolares desde donde se cuentan los sueldos elevados y n es el número de personas en cada género.

```

> n_gender
Female  Male
  588   882
> perc_75_sup_female
[1] 16077.78
> perc_75_sup_male
[1] 15810.22
> num_per_female
[1] 49
> num_per_male
[1] 75
> freq_per_female
Female
0.08333333
> freq_per_male
Male
0.08503401
> |

```

Género	US	n	# de personas con salario elevado	Frecuencia relativa al genero	Comentario
Female	16077.78	588	49	0.083	Aunque la probabilidad de encontrar a una persona con salario elevado, tiene una probabilidad muy similar para ambos generos, es visible que la probabilidad es superior para los hombres.
Male	15810.22	882	75	0.085	

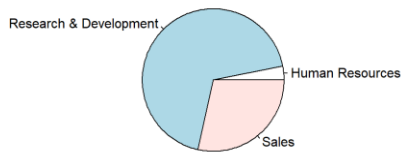
Cuadro 20: Problema 5c

Problema 6. (10pts.)

- a. Determine los dos departamentos de trabajo donde más empleados tienen un alto nivel de satisfacción con el ambiente laboral (Nivel 4).

Departamento	Proporción de interés	Comentario
Research & Development	68.39%	La población que se encuentra en un nivel de satisfacción mayor es la que trabaja en el departamento de Búsqueda y desarrollo, teniendo el 68.39% de los datos de esta población. La siguiente población serían los de ventas con un 28.48% de los datos, mostrando una clara superioridad de Búsqueda y desarrollo, y mostrando que la población de recursos humanos no se encuentra tan satisfecha con su trabajo.
Sales	28.48%	

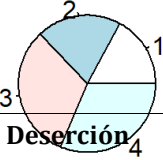
Cuadro 21: Problema 6a



- b. Por cada nivel de respuesta en la variable deserción (Attrition), determine cual es el nivel de satisfacción con el ambiente laboral con mayor proporción de empleados e indique cuál es esta proporción.

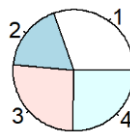
Cuadro 22: Problema 6b

No desertaron



Deserción	Nivel de satisfacción con mayor proporción	Proporción de interés	Comentario
YES	1	30.37975%	Después de ver todos los resultados podemos afirmar que las personas que más desertan son las que tienen un nivel muy bajo de satisfacción laboral. Siendo demostrado que los que desertan, la mayor proporción está en el nivel 1
NO	3	31.71127%	

Si desertaron



El nombre del archivo debe ser grupo 0x.pdf, donde x es en número del grupo.