

El deserción es un problema que afecta a todas las empresas, independientemente de la geografía, la industria y el tamaño. La deserción de empleados genera costos significativos para una empresa, incluido el costo de la interrupción del negocio, la contratación de personal nuevo y la capacitación de personal nuevo. Como tal, existe un gran interés comercial en comprender los impulsores y minimizar la deserción del personal. La data `Deserción.Y_Rendimiento_Empleados.csv` contienen diferentes factores que permiten analizar la deserción y el rendimiento de los empleados de una empresa estadounidense.

- **Age:** Edad.
- **Attrition:** Deserción del empleado.
- **Department:** Departamento en el que trabaja.
- **Education:** Nivel de estudio; 1:Bachillerato, 2:Universidad, 3:Graduado, 4:Maestría, 5:Doctorado.
- **EducationField:** Campo de estudio.
- **EnviromentSatisfaction:** Nivel de satisfacción con el ambiente laboral; 1:Bajo, 2:Medio, 3:Alto, 4:Muy alto.
- **Gender:** Género.
- **HourlyRate:** Pago por hora.
- **JobRole:** Rol de trabajo.
- **MaritalStatus:** Estado civil.
- **MonthlyIncome:** Salario mensual.
- **RelationshipSatisfaction:** Satisfacción con las relaciones; 1:Bajo, 2:Medio, 3:Alto, 4:Muy alto.
- **WorkLifeBalance:** Balance entre el trabajo y la vida personal; 1:Mala 2:Buena, 3:Mejor, 4:Excelente.
- **YearsSinceLastPromotion:** Años desde el último ascenso.

**Problema 1.** (14 pts.) 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		
Attrition		
Department		
Education		
EducationField		
EnvironmentSatisfaction		
Gender		
HourlyRate		
JobRole		
MaritalStatus		
MoethlyIncome		
RelationshipSatisfaction		
WorkLifeBalance		
YearsSinceLastPromotion		

Cuadro 1: Problema 1

**Problema 2.** (30 pts.) 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		
Education		
MoethlyIncome		

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			
Education			
MoethlyIncome			

Cuadro 3: Problema 2b

c. Si suponemos forma acampanada en el histograma de las variables Age y MoethlyIncome ¿Cuál es el percentíl 84 ( $P_{84}$ ) y el percentíl 97.5 ( $P_{97,5}$ ) para cada una de las variables?

Variable	$P_{84}$	$P_{97,5}$	Comentario
Age			
MonthlyIncome			

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 %			
MonthlyIncome Intervalo 75 %			
Age Intervalo 88,89 %			
MonthlyIncome Intervalo 88,89 %			

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	
Education	
MonthlyIncome	

Cuadro 6: Problema 2e

**Problema 3.** (22 pts.) 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

- Realice un boxplot del MonthlyIncome sin los datos extremos y compárelos con los del **Problema 2**.
- Complete la siguiente tabla y comente de manera comparativa.

Variable	Sin Extremos	Con Extremos	Comentario
Min			
$Q_1$			
$Q_3$			
Max			

Cuadro 7: Problema 3b

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

Variable	Sin Extremos	Con Extremos	Comentario
Media			
Mediana			
Punto medio clase modal			

Cuadro 8: Problema 3c

- d. Señale con una  $\times$  ¿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)	Sin Extremos (B)
------------------	------------------

Cuadro 9: Problema 3d

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

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

Cuadro 10: Problema 3e

- f. Señale con una  $\times$  ¿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.

Medida estadística	Grupo A	Grupo B	Comentario
$Q_1$			
$Q_2$			
$Q_3$			
Media			
Desviación Estándar			

Cuadro 11: Problema 3f

- g. Señale con una  $\times$  ¿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
$Q_1$			
$Q_2$			
$Q_3$			
Media			
Desviación Estándar			

Cuadro 12: Problema 3g

**Problema 4.** (21 pts.) 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.

- 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.

Rol de trabajo	Frecuencia absoluta	Comentario

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

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

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

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

Cuadro 17: Problema 4e

**Problema 5.** (27 pts.) 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.

Conjunto de datos	$\mu + 2\sigma$	% salarios elevados	Comentario
Data A			
Data B			

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.

Estado civil	US	n	# de personas con salario elevado	Frecuencia relativa al estado civil	Comentario
Married					
Divorced					
Single					

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.

Género	US	n	# de personas con salario elevado	Frecuencia relativa al estado civil	Comentario
Female					
Male					

Cuadro 20: Problema 5c

**Problema 6.** (10 pts.)

- 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

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.

Deserción	Nivel de satisfacción con mayor proporción	Proporción de interés	Comentario

Cuadro 22: Problema 6b

**Entregables:** Sea muy puntual en las respuestas y comentarios, siempre evidencie con las tablas, gráficos o estadísticos encontrados en cada uno de los ejercicios. Complete las tablas de respuesta y entregue éstas en un .pdf de máximo 14 paginas. Algunas recomendaciones: Los gráficos deben ir con título y rotulados los ejes. Todos los gráficos o tablas que decida poner deben tener un comentario en las tablas dispuestas para ello. No olvide poner el grupo unión al cual pertenecen.

**Fecha limite de entrega:** Miercoles 17 de agosto antes de las 20H00. La entrega de su trabajo debe ser por :

- Adjunte el archivo en interactiva virtual.
- El nombre del archivo debe ser **grupo 0x.pdf**, donde x es en número del grupo.