Actividad 3: Transformaciones e Inferencia Estadística

1. **Problema de empleados.**
   a. **Haga un análisis exploratorio de estos datos:**
      i. **Calcular e interpretar estadísticas descriptivas de los datos: media, mediana, moda, desviación estándar, coeficiente de variación**

Statistics

Variable	Mean	StDev	CoefVar	Median	Mode	N for Mode
Salario	4812.5	183.5	3.81	4799.5	*	0
Costo de Capacitación	401.2	56.0	13.97	387.0	*	0
Producción Generada	9831.6	197.8	2.01	9793.0	*	0
Satisfacción del Cliente Intern	7.500	1.581	21.08	7.500	6, 7, 8, 9	2
Ventas Generadas	75449	3725	4.94	75750	*	0
Ausentismo	3.600	1.430	39.72	3.500	2	3

Media: Como sabemos, la media representa el valor promedio de los datos

Mediana: La mediana representa el valor de en medio de los datos, por lo que nos indica en qué valor está centrada cada variable para todos los empleados. Se puede interpretar como un promedio que descuenta outliers de cierta manera, por lo que su similitud con la media para estos datos nos indica una distribución sin outliers.

Moda: Indica el valor más repetido, o su ausencia nos enseña que ningún valor se repite para ese dato.

Desviación estándar: Es la raíz cuadrada de varianza (el promedio de las diferencias al cuadrado entre cada valor y la media), por lo que nos regresa a las unidades originales y nos enseña cuánto difieren de la media los datos.

Coeficiente de variación: es el cociente entre la desviación estándar y la media expresada como porcentaje, nos informa que tanto se parecen en promedio los datos a la media, lo cual nos da información más interpretable que la varianza.

      ii. **¿Cuál de las variables tiene mayor variabilidad? ¿Cuál tiene menor variabilidad? Explique, ¿cuáles estadísticas son relevantes para ello? y ¿por qué?**

Para hacer aserciones de la variabilidad de los datos, podemos apuntar al coeficiente de variación, ya que es una manera de normalizar nuestra información. Podríamos también ver la desviación estándar, pero como todos nuestros datos existen en unidades drásticamente distintas, esto no nos da una manera concreta de

compararlos. Observando el coeficiente de variación, podemos ver que la variable con más variabilidad es “Satisfacción del cliente interno” y el que menor varía de la media en promedio es “producción generada”. Esto nos apunta a que hay empleados que tienden a dejar a los clientes mucho más felices o infelices que la media, pero que de todas maneras todos los empleados son comparables en productividad, apuntando a que la productividad probablemente no depende de la satisfacción de los clientes (interesante).

b. **Utilizando la Técnica de Análisis Multifactor, obtener cuál debería ser el ranking de cada uno de los empleados para poder definir el reparto de los incentivos.**

Aquí se llevaron a cabo varios datos que explicaré mostrando las transformaciones “puente” desde los datos iniciales hasta el ranking.

Lo primero que se hace es estandarizar los datos dependiendo de que se necesita: un menor valor o un mayor valor. Se dividen los datos para que el valor más deseable sea 1 y los demás se escalen correspondientemente:

Num Empleado	Salario Est	CC Est	PG Est	SatC Est	VG Est	Aus Est
1	0.98485	0.93220	0.99020	0.7	1.00000	0.40000
2	0.89216	0.66132	0.97030	0.8	0.93734	0.33333
3	1.00000	0.73333	0.94059	0.6	0.86235	0.50000
4	0.95769	0.70213	0.99000	0.9	0.88734	0.66667
5	0.93853	0.86842	0.96535	0.7	0.95608	1.00000
6	0.92255	0.89189	0.95842	0.6	0.99750	0.40000
7	0.90278	1.00000	0.96891	0.8	0.97056	0.50000
8	0.97410	0.94286	0.95545	0.5	0.98108	1.00000
9	0.96829	0.79518	1.00000	0.9	0.91234	1.00000
10	0.92593	0.83756	0.99505	1.0	0.92484	0.66667

Después, estos datos se normalizan los datos dividiéndolos entre la suma de todos los datos con el fin de que cuando sumemos nuestros valores normalizados estos sean iguales a 1:

Num Empleado	Salario Norm	CC Norm	PG Norm	SatC Norm	VG Norm	Aus Norm
1	0.104031	0.111442	0.101723	0.093333	0.106051	0.061856
2	0.094240	0.079059	0.099679	0.106667	0.099405	0.051546
3	0.105632	0.087668	0.096627	0.080000	0.091453	0.077320
4	0.101163	0.083937	0.101703	0.120000	0.094104	0.103093
5	0.099139	0.103817	0.099170	0.093333	0.101394	0.154639
6	0.097450	0.106623	0.098458	0.080000	0.105786	0.061856
7	0.095362	0.119547	0.099536	0.106667	0.102928	0.077320
8	0.102895	0.112716	0.098153	0.066667	0.104044	0.154639
9	0.102282	0.095062	0.102730	0.120000	0.096755	0.154639
10	0.097807	0.100128	0.102221	0.133333	0.098080	0.103093

Después, se obtiene el ponderado de los datos multiplicándolos por el factor de su relevancia proporcionado en el ejercicio y se promedia cada fila. Este promedio representa la calificación final de cada empleado, por lo que acomodar los promedios de mayor a menor nos otorga un ranking de los empleados de mejor a peor según los criterios deseados:

Sorted Num Empleado	Sorted Salario Pond	Sorted CC Pond	Sorted PG Pond	Sorted SatC Pond	Sorted VG pond	Sorted Aus Pond	Sorted Promedio
9	0.0061369	0.0028518	0.0164368	0.0300000	0.0387018	0.0154639	0.0182652
10	0.0058684	0.0030038	0.0163554	0.0333333	0.0392320	0.0103093	0.0180171
5	0.0059483	0.0031145	0.0158672	0.0233333	0.0405574	0.0154639	0.0173808
4	0.0060698	0.0025181	0.0162724	0.0300000	0.0376415	0.0103093	0.0171352
7	0.0057217	0.0035864	0.0159258	0.0266667	0.0411713	0.0077320	0.0168006
8	0.0061737	0.0033815	0.0157045	0.0166667	0.0416177	0.0154639	0.0165013
1	0.0062419	0.0033433	0.0162757	0.0233333	0.0424204	0.0061856	0.0163000
2	0.0056544	0.0023718	0.0159486	0.0266667	0.0397622	0.0051546	0.0159264
6	0.0058470	0.0031987	0.0157533	0.0200000	0.0423144	0.0061856	0.0155498
3	0.0063379	0.0026300	0.0154604	0.0200000	0.0365812	0.0077320	0.0147902

Al observar estos promedios finales en orden, podemos ver claramente que el mejor empleado es el 9.

c. Suponga que se quiere utilizar los datos proporcionados y una regresión lineal para predecir cuáles serían las ventas generadas por 3 empleados nuevos con los siguientes valores:

Para este ejercicio, utilizamos la función MinMax Scaler en Minitab para ajustar nuestros valores de las variables independientes entre 0 y 1. Esto sirve para escalar nuestros datos de tal manera que sus diferencias en proporción por tipo de dato dejan de importar. Nuestra variable a predecir (dependiente) se queda en su escala original ya que necesitaremos respetar esa escala para nuestra predicción final:

Num Empleado	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Intern	Ventas Generadas	Ausentismo
1	0.12727	0.14201	0.83500	0.4	80014	0.75
2	1.00000	1.00000	0.50000	0.6	75000	1.00
3	0.00000	0.71006	0.00000	0.2	69000	0.50
4	0.36545	0.82840	0.83167	0.8	71000	0.25
5	0.54182	0.29586	0.41667	0.4	76500	0.00
6	0.69455	0.23669	0.30000	0.2	79814	0.75
7	0.89091	0.00000	0.47667	0.6	77658	0.50
8	0.22000	0.11834	0.25000	0.0	78500	0.00
9	0.27091	0.50296	1.00000	0.8	73000	0.00
10	0.66182	0.37870	0.91667	1.0	74000	0.25

Al hacer esto, podemos hacer una regresión más escalada que observamos a continuación:

Regression Equation

Ventas Generadas = 74820 + 6833 Salario - 5037 Costo de Capacitación
+ 11387 Producción Generada - 14830 Satisfacción del Cliente Intern
+ 1528 Ausentismo

También podemos ver el desglose de los coeficientes:

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	74820	607	123.21	0.000	
Salario	6833	898	7.61	0.002	1.96
Costo de Capacitación	-5037	832	-6.05	0.004	1.67
Producción Generada	11387	1392	8.18	0.001	4.61
Satisfacción del Cliente Intern	-14830	1704	-8.70	0.001	6.35
Ausentismo	1528	734	2.08	0.106	1.51

Finalmente, podemos observar nuestra R cuadrada:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
641.437	98.68%	97.03%	88.30%

Sabemos que el modelo es confiable porque nuestra R cuadrada se acerca a 1.

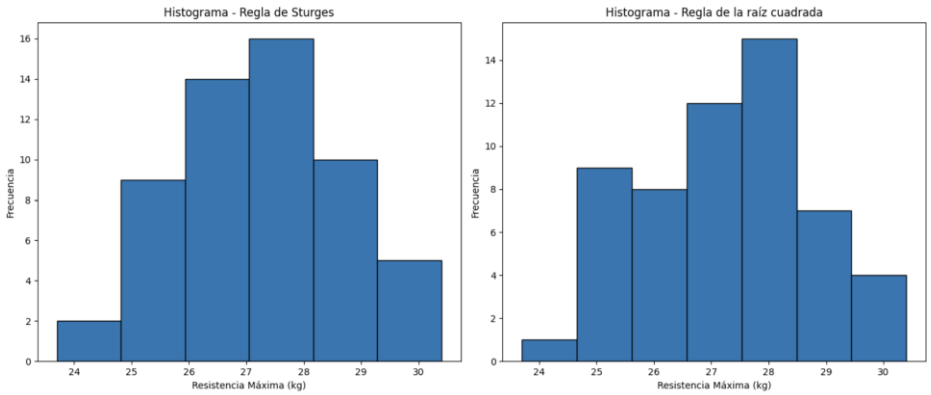
Con este modelo, ahora solamente tenemos que escalar nuestros datos nuevos, pero con los valores máximos y mínimos obtenidos en nuestros datos originales. Esto contextualiza a los datos nuevos dentro de la escala de

la regresión original, como si fueran datos de nuestro modelo inicial. Finalmente, dentro de estos datos podemos también utilizar la ecuación obtenida para predecir nuestras ventas generadas y obtener al mejor potencial en un solo paso:

Sorted Empleados Nuevos	Sorted S SC	Sorted CC SC	Sorted PG SC	Sorted SC SC	Sorted Aus SC	Sorted VG PRED
13	0.545455	0.295858	0.833333	0.6	0.50	78412.0
12	0.636364	0.710059	0.166667	0.4	0.75	72703.5
11	0.272727	0.532544	0.500000	0.6	0.25	71178.6

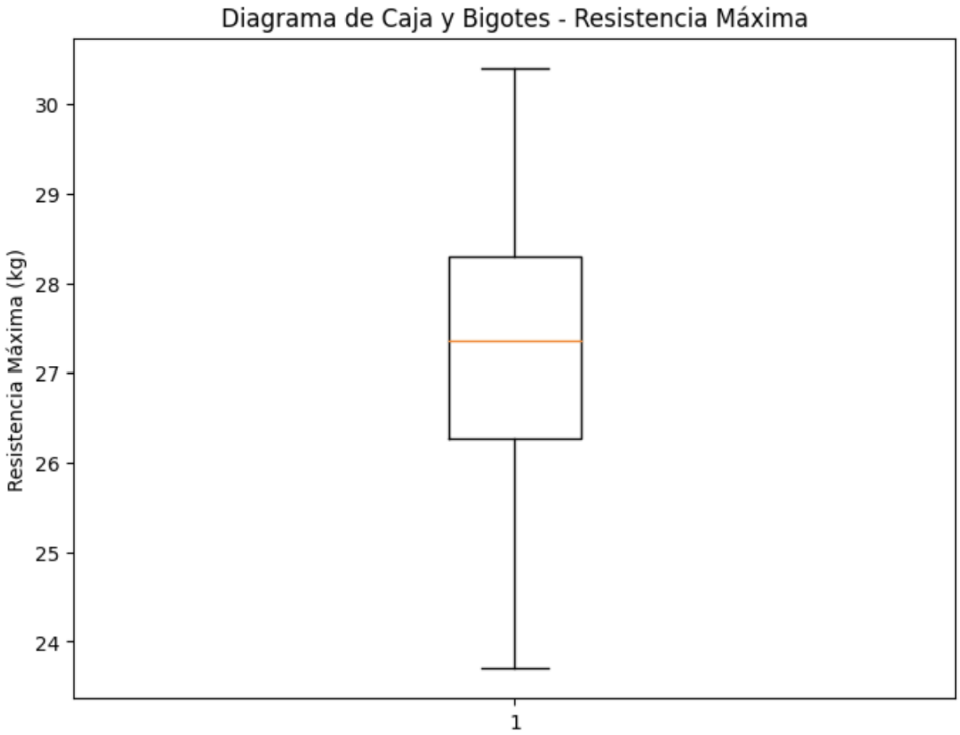
Ahora sabemos que el empleado con mayor potencial de ventas es el 13, seguido del 12 y después 11.

2. **Elaboración de envases de plástico.**
   a. **¿Qué tipo de variable se está midiendo? ¿Discreta o continua? Explique.**
   Se está midiendo una variable continua, ya que la resistencia máxima de una botella verdaderamente podría caer en cualquier valor entre dos valores. No existe una tabla finita de valores posibles como para las variables discretas.
   b. **Haga un análisis exploratorio de estos datos.**
      i. **Realice un histograma con al menos 2 reglas para definir el número de clases (No utilizar regla empírica). Describa la forma y analice el comportamiento de los datos.**



Al observar nuestros histogramas, podemos observar que ambos muestran una forma que se acerca a una distribución normal, quizás con un cargo ligero hacia la derecha. Eso nos indica que los datos se centran en una media específica (probablemente el valor ideal según control de calidad). La similitud de los diagramas nos indica que la distribución es marcada y no depende tanto de la granularidad del análisis.

      ii. **Realice un diagrama de caja y bigotes. Analice el comportamiento de los datos. ¿Existen datos atípicos? ¿Qué se debería hacer al respecto?**
      A continuación tenemos el diagrama de caja y bigotes de los datos de resistencia de las botellas:



Este diagrama nos indica dentro de la caja el rango entre el primer y tercer cuartil de los datos, con la línea amarilla mostrándonos la media. Podemos ver que esta se encuentra ligeramente hacia arriba correspondiendo con nuestros histogramas (que se cargaban ligeramente a la derecha). Los bigotes de la gráfica representan el rango dentro del cual existen los datos típicos del modelo, con cualquier dato fuera de este rango siendo atípico.

```
Primer cuartil (Q1): 26.275
Tercer cuartil (Q3): 28.3
Rango intercuartílico (IQR): 2.025000000000002
Límite inferior: 23.237499999999997
Límite superior: 31.337500000000006
Datos atípicos:
Empty DataFrame
```

Como podemos observar, como ningún dato cae fuera de este rango, ningún dato se puede considerar atípico. Esto sugiere que no ha habido problemas mayores en control de calidad por lo que no se debería tomar ninguna acción.

c. **Estime, con una confianza de 94%, ¿cuál sería la resistencia promedio de los envases?**

EL rango de resistencia promedio estimada con un 94% de confianza es (26.879344447464373, 27.613512695392767). Esto nos indica que la media se encuentra en este rango, que se obtiene calculando la media y restando y sumando el margen de error.

d. **Antes del estudio se suponía que la resistencia promedio era de 25kg. Dada la evidencia de los datos, ¿tal supuesto es correcto? ¿Qué tipo de prueba estadística se debe realizar? Planteé las hipótesis correspondientes y concluya adecuadamente.**
   Para este tipo de datos utilizaremos la prueba t para determinar si la hipótesis es correcta. Esta prueba se utiliza en este caso por dos razones: se desconoce la desviación estándar poblacional (ya que solo contamos con los datos de un muestreo) y estamos tratando con una cantidad baja de datos.
   Al hacer el análisis en Python obtenemos estos datos:

```
t_score: 11.752111281692763
p-value: 0.0
Rechazamos la hipótesis nula: La resistencia promedio es significativamente diferente de 25 kg.
```

   En este resultado, el valor t implica nos indica que nuestra media muestral se encuentra a 11 desviaciones estándar de la media estimada (muchísimo) y el valor p en 0 nos dice que hay un 0% de probabilidad de que un valor caiga en esa media. Por lo mismo se rechaza la hipótesis.

e. **Con los datos anteriores estime, con una confianza del 98%, ¿cuál es la desviación estándar poblacional (del proceso)?**
   El rango de resistencia promedio poblacional estimada con un 98% de confianza es: (26.7884, 27.7044). Este valor se estima con la prueba z, ya que este corresponde a poblaciones.

3. **Laboratorio.**
   a. **¿Las muestras son dependientes o independientes? Explique**
      En este caso las muestras son independientes, ya que ninguna variable explica ninguna otra. Como son personas distintas, cada valor encontrado es independiente de todos los demás, ya que tampoco hay un rango de espacio ni tiempo. Sin más contexto no se puede acertar nada sobre los sujetos del estudio.
   b. **¿La temperatura promedio más confortable es igual para hombre que para mujeres? ¿Qué tipo de prueba estadística se debe realizar? Planteé las hipótesis correspondientes y concluya adecuadamente.**
      En este caso, para determinar esta información, se debe llevar a cabo la prueba t para dos variables. Esta es adecuada por el tamaño pequeño de muestra y por no conocer la desviación estándar de la población. Al computar en Python obtenemos lo siguiente:
      -Estadístico t: -3.5254179083580257 (un valor se encuentra 3 desviaciones estándar a la izquierda del otro)
      -p-valor: 0.002626225071336037 (muy pequeño comparado con mi 0.05)
      -Rechazamos la hipótesis nula: La temperatura promedio más confortable es diferente entre hombres y mujeres.

c. **¿Los datos poseen la misma variabilidad? ¿Qué tipo de prueba estadística se debe realizar? Planteé las hipótesis correspondientes y concluya adecuadamente**

Para esta estimación, se debe utilizar la prueba F, que se utiliza para comparar las varianzas de dos pruebas independientes. Al realizar el código en Python, obtenemos lo siguiente:

-Estadístico F: 0.5859375 (sugiere diferencia significativa entre varianzas, ya que no es cercano a 1).

-p-valor: 0.4380709879241654 (indica que hay un 43.8% de probabilidad de obtener un estadístico F tan extremo si las dos varianzas fueran realmente iguales, muy lejano de mi alfa.)

-No rechazamos la hipótesis nula: No hay suficiente evidencia para decir que la varianza de temperatura más confortable es diferente entre hombres y mujeres, apuntando a que poseen una variabilidad similar.

4. **Prueba de un solo disco**

a. **¿Las muestras son dependientes o independientes? Explique**

En este caso, las muestras son dependientes. Esto se debe a que cada valor del método nuevo está atado directamente a un valor correspondiente del método viejo, ya que se está observando al mismo objeto de dos maneras distintas. En otras palabras: cada medición de un método tiene una contraparte específica en una medición del otro.

b. **¿Qué tipo de prueba estadística se debe realizar? Planteé las hipótesis correspondientes y concluya adecuadamente.**

Para este análisis, la prueba correcta de utilizar es la prueba t para muestras dependientes, ya que cuento con muestras dependientes que están emparejadas entre sí. Al hacer el análisis utilizando Python obtuve los siguientes datos:

-Media de diferencias: 0.0022222222222222613 (muy cercana a cero, indicando que el nuevo método en promedio genera resultados muy similares al actual).

-Estadístico t: 0.23874497225498748 (este valor está más cerca a cero que a uno, lo cual indica que la diferencia entre las medias no es lo suficientemente grande en relación con la variabilidad de los datos para considerarse estadísticamente significativa).

-p-valor: 0.8141576311659824 (lo cual indica que la probabilidad de observar un cambio como el que existe en los datos utilizando el mismo método dos veces es 81.4%).

No rechazamos la hipótesis nula: No hay suficiente evidencia para decir que el método actual y el método nuevo son diferentes.

c. **¿Recomienda la adopción del nuevo método? Argumente su respuesta.**

Al hacer el análisis correspondiente, y teniendo en cuenta el planteamiento inicial de la pregunta, recomiendo cambiar al método nuevo. Como observamos, los métodos no otorgan resultados lo suficientemente diferentes para poder diferenciar entre ambos métodos con exactitud. Como el método nuevo nos otorga beneficios importantes sin alterar los resultados de manera significativa, se debería cambiar al nuevo método.