Reto I Caso Líneas Aéreas

Nombre del participante:

Indicaciones:

- **Genera una copia** de este documento y editarla con tu nombre de la siguiente manera: **C5SC3 Reto - Nombre completo del participante**
- Una vez terminado el Reto deberás de **entregarlo** en la opción “Añadir publicación” en el apartado Reto de aplicación en el trabajo.

Los siguientes pasos te guiarán en el proceso de desarrollo del reto para que logres completarlo con éxito:

1. Descarga el archivo **pdf** llamado **Reto Material del Caso Aerolíneas**, ya que en éste se presentan estadísticas descriptivas y modelos de regresión que se deben analizar para determinar qué variables se consideran relevantes en el servicio.

2. Descarga, las siguientes bases de datos:

- **Descargar Caso Aerolíneas Datos.csv**
- **Descargar Datos Originales.xlsx**

3. Contesta cada uno de los puntos que se piden en el espacio asignado posterior a la tabla de actividades, basándote en los datos que se presentan en los recursos que previamente descargaste:

Actividad que realizar	Contesta lo siguiente
1.- Análisis descriptivo de las variables que describen el comportamiento del Servicio en las aerolíneas	Analiza la información proporcionada y destaca **al menos 5** comportamientos importantes de las variables (explícalos utilizando medidas estadísticas).

1. **Días de Compra Antes del Vuelo (DaysPurchase)**:
   - **Rango**: de 8 a 123 días.
   - **Media**: 40 días.
   - **Mediana**: 40 días.
   - La distribución de los días de compra antes del vuelo es bastante simétrica alrededor de 40 días, indicando que en promedio, las reservas se realizan con aproximadamente un mes y medio de antelación.
2. **Precio del Boleto (Ticket Price)**:
   - **Rango**: de $300 a $620.
   - **Media**: $376.
   - **Mediana**: $341.
   - El precio del boleto muestra una variación considerable, con algunos precios más altos que otros, lo cual podría ser por factores como la clase de boleto, la anticipación de la compra, y la aerolínea.
3. **Clase del Boleto (Business y First Class)**:
   - **Business**: 50% de los boletos son de clase business.
   - **First Class**: 50% de los boletos son de primera clase.
   - La distribución equitativa de las clases de boletos sugiere una división balanceada en la elección entre business y primera clase.
4. **Frecuencia de Viaje (Infrequent, Frequent, Extreme)**:
   - **Infrequent**: 31% de los viajeros.
   - **Frequent**: 13% de los viajeros.
   - **Extreme**: 12% de los viajeros.
   - La mayoría de los pasajeros son viajeros poco frecuentes.
5. **Días de la Semana (Lunes, Martes, Miercoles, Jueves, Viernes, Sabado)**:
   - Distribución bastante uniforme en los días de la semana, con un ligero aumento los viernes y sábados.

2.- Análisis de multicolinealidad	Utiliza la matriz de correlación para destacar qué variables independientes (X´s) presentan multicolinealidad. Destaca **al menos 3** de las más importantes. Además, en el **Modelo1** analiza el **VIF** que se proporciona para cada variable.

1.**First Class y Business**:

- Correlación: -0.549
- La fuerte correlación negativa sugiere que probablemente estas variables son mutuamente excluyentes, lo que tiene sentido porque por ejemplo, un boleto no puede ser tanto de primera clase como de negocios.

2. **Trips y Frequent**:
   - VIF para *Frequent*: 3.89
   - Aunque la matriz de correlación no muestra una alta correlación directa entre *Trips* y *Frequent*, el VIF elevado para *Frequent* sugiere que esta variable podría estar explicada por otras variables en el modelo, incluyendo *Trips*.
3. **Extreme**:
   - VIF para *Extreme*: 8.60
   - Un VIF tan alto indica una fuerte multicolinealidad. *Extreme* puede estar altamente correlacionada con varias otras variables, complicando su interpretación.

Además observando el **Modelo 1**:

- Extreme con un VIF de 8.60 es preocupante y podría necesitar revisión, como eliminar la variable o combinarla con otras para reducir la multicolinealidad.
- Frequent también muestra un VIF relativamente alto (3.89), lo que sugiere que esta variable también podría estar influida por otras variables del modelo.

3.- Construir la ecuación o modelo matemático (copia y pega de la información proporcionada) Indica si el modelo es congruente, es decir, no presenta efectos de multicolinealidad	**1. Estadísticas del Modelo:** • R-cuadrado: 63.94% • R-cuadrado ajustado: 57.13% • R-cuadrado predictivo: 48.05% Con estas estadísticas vemos que modelo explica aproximadamente el 63.94% de la variabilidad en el precio del boleto, lo cual es relativamente alto. Sin embargo, el R-cuadrado ajustado y predictivo más bajos muestran que algunas variables pueden no estar contribuyendo significativamente a la explicación del modelo y que se puede mejorar.
4.- Validación estadística del modelo: a) Medidas de calidad del ajuste b) Prueba de hipótesis para la ecuación (F) c) Prueba de hipótesis para cada una de las variables independientes (t)	**2. Multicolinealidad:** • Presencia de varios VIF en el modelo: • Extreme: 8.60 • Frequent: 3.89 • First Class: 2.36 • Estos valores VIF sugieren que hay un grado significativo de multicolinealidad en el modelo. Con base en las observaciones presentadas podemos concluir que en su estado actual el modelo, no es congruente debido a su alto nivel de multicolinealidad.

Medidas de Calidad del Ajuste

1. **R-cuadrado**: 63.94%
   - Este valor indica que el modelo explica aproximadamente el 63.94% de la variabilidad en el precio del boleto.
2. **R-cuadrado ajustado**: 57.13%
   - Considero que un valor de 57.13% sugiere que no todas las variables podrían ser igualmente útiles.
3. **R-cuadrado predictivo**: 48.05%
   - Un valor de 48.05% es ya considerablemente alto lo cual nos da cierta confianza en la capacidad predictiva del modelo, pero sin duda no es un valor optimo.

Prueba de Hipótesis para la Ecuación (F-Test)

- **F-Value**: 9.39
- **P-Value**: 0.000
  - Un F de 9.39 y un p-valor de 0.000 indican que podemos rechazar la hipótesis nula con un nivel de significancia alto, lo que confirma que al menos algunas de las variables independientes tienen un efecto sobre el precio del boleto.

Prueba de Hipótesis para Cada Variable Independiente (t-Test)

- Las variables *AA*, *Delta*, *United*, *Business*, y *First Class* muestran significancia estadística, indicando un efecto claro sobre el precio del boleto.
- Variables como *DaysPurchase* y *Frequent* tienen p-values al borde de la significancia convencional ($p<0.05$), lo que nos puede indicar tener cuidado con su interpretación.
- Muchas variables relacionadas con los días de la semana y otros factores como *Origin* y *Trips* no son estadísticamente significativas, lo que sugiere que podemos descartarlas.

5.- Validación de supuestos: a) Normalidad en los residuales b) Errores con varianza constante c) Independencia de los errores	Analiza los supuestos Utilizando el **Modelo 1**

1. Normalidad de los Residuales

- El **gráfico de probabilidad normal** mostrado en los diagnósticos muestra que si bien hay cierta normalidad, hay desviaciones notables, especialmente en los extremos. Esto puede indicar que los residuales no son completamente normales.
- El **histograma de los residuales** también ayuda a visualizar la distribución de los errores. Aunque parece aproximadamente simétrico, la forma no es perfectamente de campana, lo que indica posibles desviaciones de la normalidad.

2. Homocedasticidad

- **Gráfico de Residuales vs. Valores Ajustados**:
  - En el gráfico proporcionado, parece haber una dispersión aleatoria de residuales alrededor de cero sin un patrón claro de aumento o disminución en la varianza. Esto sugiere que el modelo no sufre problemas graves de heterocedasticidad, aunque la dispersión no es perfectamente uniforme.

3. Independencia de los Errores

- **Gráfico de Residuales vs. Orden de Observación**:
  - Este gráfico ayuda a detectar cualquier autocorrelación en los residuales. Si los residuales son independientes, deberíamos ver una distribución aleatoria sin patrones claros. El gráfico proporcionado parece mostrar una distribución bastante aleatoria de los residuales, sin patrones obvios de autocorrelación, aunque siempre es bueno realizar pruebas formales como la prueba de Durbin-Watson para confirmarlo estadísticamente.

El Modelo 1 parece cumplir razonablemente bien con el supuesto de homocedasticidad e independencia de los errores. Sin embargo, hay algunas dudas sobre la normalidad de los residuales, particularmente en los extremos.

6.- Predicción del precio de venta	Con el **Modelo 2** describe cómo variaría el precio de venta según las características o variables relevantes.

Influencia de Variables:

1. Aerolíneas (AA y United):
- AA: Un incremento de una unidad en la variable asociada a volar con American Airlines aumenta el precio del boleto en 104.0 unidades monetarias.
- United: Similarmente, volar con United Airlines incrementa el precio en 97.5 unidades.

Estos coeficientes sugieren que volar con estas aerolíneas, en comparación con otras no mencionadas, tiende a estar asociado con precios más altos.

2. Días de Anticipación de la Compra (DaysPurchase):
- Por cada día adicional antes del vuelo que el boleto es comprado, el precio del boleto disminuye en 0.603 unidades.

Esto sugiere que las compras anticipadas están asociadas con precios ligeramente más bajos, lo que podría reflejar descuentos por reserva anticipada.

3. Clase del Boleto (First Class):
- Viajar en primera clase incrementa el precio del boleto en 44.0 unidades en comparación con otras clases (e.g., económica o business).

Esto refleja el valor agregado y el costo adicional de los servicios y comodidades asociados con la primera clase.

4. Frecuencia de Viaje (Frequent):
- Ser un viajero frecuente disminuye el precio del boleto en 40.7 unidades.

Esto podría indicar que los viajeros frecuentes podrían beneficiarse de programas de lealtad o descuentos recurrentes ofrecidos por las aerolíneas.

Implicaciones para la Predicción de Precios:

Este modelo puede ser usado para predecir el precio de los boletos basándose en las características específicas de cada vuelo y cliente.

Cambios en cualquiera de estas variables alterarán el precio final del boleto, y el modelo proporciona una forma clara y cuantificable de estimar esos cambios.