DATOS ICFES 2017 I Y 2017II (Antioquia)

Análisis de Datos para la Toma de Decisiones

Big Data Empresarial

Melani Tirado Salas

El ICFES (Instituto Colombiano para la Evaluación de la Educación) es la entidad encargada de diseñar, aplicar y analizar las pruebas estandarizadas de educación en Colombia, siendo la más reconocida la Prueba Saber 11, comúnmente llamada “el ICFES”. Esta prueba es obligatoria para todos los estudiantes que finalizan la educación media (grado 11). En el presente informe se realiza un análisis de los resultados de la Prueba Saber 11 para el departamento de Antioquia durante los periodos 2017-I y 2017-II.

I. Datos

La base de datos corresponde a información de puntaje ICFES Periodo 2017I Y 2017II. En la ilustración I. Se cargaron las bases de datos y se unieron en una lista, con lo cual se creó la variable df.

Ilustración 1. Base de datos - Unión base de datos.

```
[3]  sb11_20171_df = pd.read_csv('SB11_20171.TXT', sep='¬', encoding='utf-8', engine='python')
     sb11_20172_df = pd.read_csv('SB11_20172.TXT', sep='¬', encoding='utf-8', engine='python')
```

```
print("Unir datasets de los 2 semestres")
lista_tablas = [ sb11_20171_df, sb11_20172_df ]
df = pd.concat(lista_tablas).reset_index(drop=True)
df.head(5)
```

En ese sentido se elige el departamento de Antioquia como objeto de estudio.

Ilustración 2. Selección departamento Antioquia

```
info_depto_df = df[df["ESTU_DEPTO_RESIDE"] == 'ANTIOQUIA'].reset_index(drop=True)
info_depto_df.head(5)
```

II. Exploración de datos

Después de unir la base de datos de Icetex 2017-I Y 2017-II, departamento de Antioquia contiene 73.675 observaciones y variables (82), se visualizan medidas de tendencia central de las diferentes variables.

Ilustración 3. Estadística descriptiva

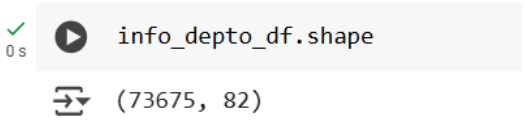
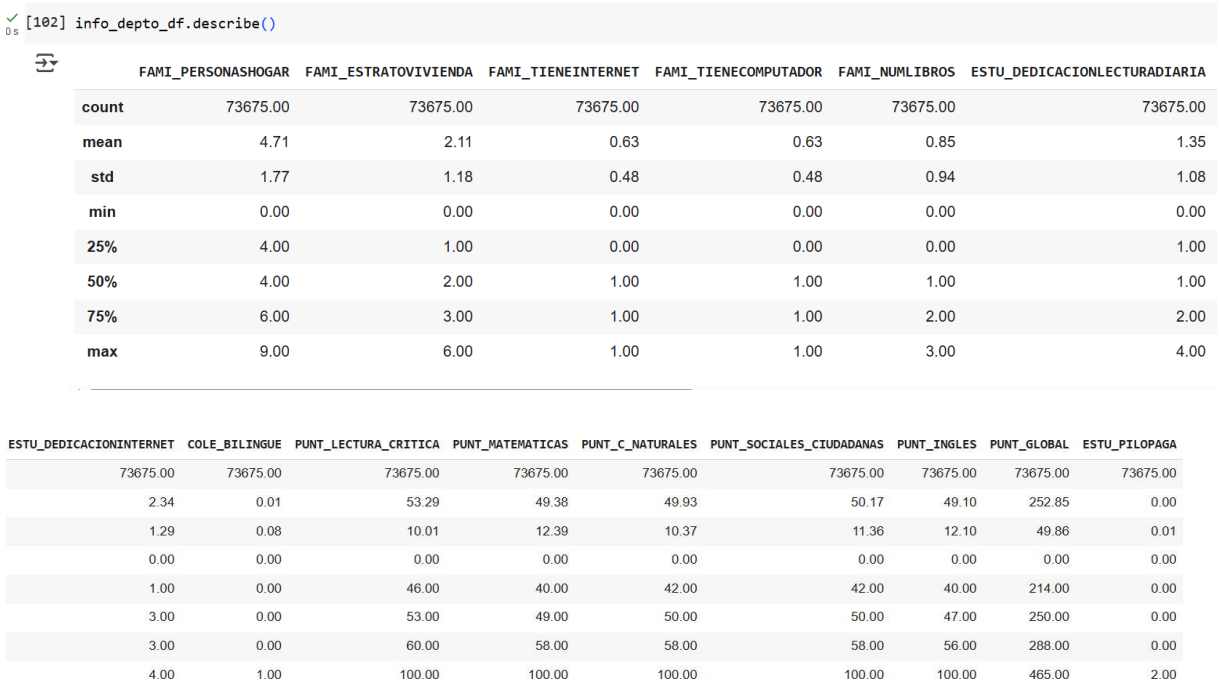


Ilustración 4



	FAMI_PERSONASHOGAR	FAMI_ESTRATOVIVIENDA	FAMI_TIENEINTERNET	FAMI_TIENECOMPUTADOR	FAMI_NUMLIBROS	ESTU_DEDICACIONLECTURADIARIA
count	73675.00	73675.00	73675.00	73675.00	73675.00	73675.00
mean	4.71	2.11	0.63	0.63	0.85	1.35
std	1.77	1.18	0.48	0.48	0.94	1.08
min	0.00	0.00	0.00	0.00	0.00	0.00
25%	4.00	1.00	0.00	0.00	0.00	1.00
50%	4.00	2.00	1.00	1.00	1.00	1.00
75%	6.00	3.00	1.00	1.00	2.00	2.00
max	9.00	6.00	1.00	1.00	3.00	4.00

ESTU_DEDICACIONINTERNET	COLE_BILINGUE	PUNT_LECTURA_CRITICA	PUNT_MATEMATICAS	PUNT_C_NATURALES	PUNT_SOCIALES_CIUDADANAS	PUNT_INGLES	PUNT_GLOBAL	ESTU_PILOPAGA
73675.00	73675.00	73675.00	73675.00	73675.00	73675.00	73675.00	73675.00	73675.00
2.34	0.01	53.29	49.38	49.93	50.17	49.10	252.85	0.00
1.29	0.08	10.01	12.39	10.37	11.36	12.10	49.86	0.01
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	46.00	40.00	42.00	42.00	40.00	214.00	0.00
3.00	0.00	53.00	49.00	50.00	50.00	47.00	250.00	0.00
3.00	0.00	60.00	58.00	58.00	58.00	56.00	288.00	0.00
4.00	1.00	100.00	100.00	100.00	100.00	100.00	465.00	2.00

III. Limpieza Dataset

Se eliminan valores nulos y se modifican variables

Ilustración 5. Modificación de nulos – se reservan las variables relevantes

```
columnas_a_conservar = [
    "ESTU_DEPTO_RESIDE", "ESTU_MCPIO_RESIDE", "FAMI_PERSONASHOGAR",
    "FAMI_ESTRATOVIVIENDA", "FAMI_TIENEINTERNET", "FAMI_TIENECOMPUTADOR",
    "FAMI_NUMLIBROS", "FAMI_SITUACIONECONOMICA", "ESTU_DEDICACIONLECTURADIARIA",
    "ESTU_DEDICACIONINTERNET", "COLE_BILINGUE", "PUNT_LECTURA_CRITICA",
    "PUNT_MATEMATICAS", "PUNT_C_NATURALES", "PUNT_SOCIALES_CIUDADANAS",
    "PUNT_INGLES", "PUNT_GLOBAL", "ESTU_PILOPAGA"
]


info_depto_df = info_depto_df[columnas_a_conservar]



print("Las columnas han sido filtradas correctamente.")
info_depto_df
```

Dado que la base de datos contenía una gran cantidad de columnas con valores nulos, se filtraron únicamente aquellas variables consideradas relevantes para facilitar la interpretación de resultados.

Del mismo modo, se transformaron los valores categóricos en valores numéricos para facilitar el análisis.

Ilustración 6. Transformación de valores categóricos a valores numéricos en la variable

```
[66] valores_unicos = info_depto_df['FAMI_PERSONASHOGAR'].unique()
     print("Valores únicos en FAMI_PERSONASHOGAR:")
     print(valores_unicos)
```

```
Valores únicos en FAMI_PERSONASHOGAR:
['1 a 2' '5 a 6' '3 a 4' '9 o más' '7 a 8' nan]
```

```
rango_personas = {
    "1 a 2": 2,
    "3 a 4": 4,
    "5 a 6": 6,
    "7 a 8": 8,
    "9 o más": 9
}


info_depto_df['FAMI_PERSONASHOGAR'] = info_depto_df['FAMI_PERSONASHOGAR'].map(rango_personas)
```

Ilustración 7.

```
[66] valores_unicos = info_depto_df['FAMI_PERSONASHOGAR'].unique()
     print("Valores únicos en FAMI_PERSONASHOGAR:")
     print(valores_unicos)
```

```
Valores únicos en FAMI_PERSONASHOGAR:
['1 a 2' '5 a 6' '3 a 4' '9 o más' '7 a 8' nan]
```

```
[67] rango_personas = {
         "1 a 2": 2,
         "3 a 4": 4,
         "5 a 6": 6,
         "7 a 8": 8,
         "9 o más": 9
     }


     info_depto_df['FAMI_PERSONASHOGAR'] = info_depto_df['FAMI_PERSONASHOGAR'].map(rango_personas)
```

Ilustración 8

```
estrato = {
    "Estrato 1": 1,
    "Estrato 2": 2,
    "Estrato 3": 3,
    "Estrato 4": 4,
    "Estrato 5": 5,
    "Estrato 6": 6,
    "Sin Estrato": 0
}


info_depto_df['FAMI_ESTRATOVIVIENDA'] = info_depto_df['FAMI_ESTRATOVIVIENDA'].map(estrato)

info_depto_df.head(5)
```

Ilustración 9

```
[75] valores_unicos = info_depto_df['FAMI_TIENEINTERNET'].unique()

     print(valores_unicos)
```

```
['Si' 'No' nan]
```

```
[76] tiene_internet = {
         "Si": 1,
         "No": 0
     }
```

Ilustración 10

```
[78] valores_unicos = info_depto_df['FAMI_TIENECOMPUTADOR'].unique()

     print(valores_unicos)
```

```
['Si' nan 'No']
```

```
tiene_computador = {
    "Si": 1,
    "No": 0
}


info_depto_df['FAMI_TIENECOMPUTADOR'] = info_depto_df['FAMI_TIENECOMPUTADOR'].map(tiene_computador)

info_depto_df.head(5)
```

Ilustración 11

```
[81] valores_unicos = info_depto_df['FAMI_NUMLIBROS'].unique()

     print(valores_unicos)
```

```
['MÁS DE 100 LIBROS' '0 A 10 LIBROS' '26 A 100 LIBROS' '11 A 25 LIBROS'
 nan]
```

```
rango_numlibros = {
    "0 A 10 LIBROS": 0,
    "11 A 25 LIBROS": 1,
    "26 A 100 LIBROS": 2,
    "MÁS DE 100 LIBROS": 3,
}


info_depto_df['FAMI_NUMLIBROS'] = info_depto_df['FAMI_NUMLIBROS'].map(rango_numlibros)
info_depto_df['FAMI_NUMLIBROS'] = info_depto_df['FAMI_NUMLIBROS'].fillna(0)

info_depto_df.head(5)
```

Ilustración 12

```
mapa_lectura = {
    "30 minutos o menos": 1,
    "Entre 30 y 60 minutos": 2,
    "Entre 1 y 2 horas": 3,
    "Más de 2 horas": 4,
    "No leo por entretenimiento": 0,
}

mapa_internet = {
    "30 minutos o menos": 1,
    "Entre 30 y 60 minutos": 2,
    "Entre 1 y 3 horas": 3,
    "Más de 3 horas": 4,
    "No Navega Internet": 0,
}
```

Ilustración 13

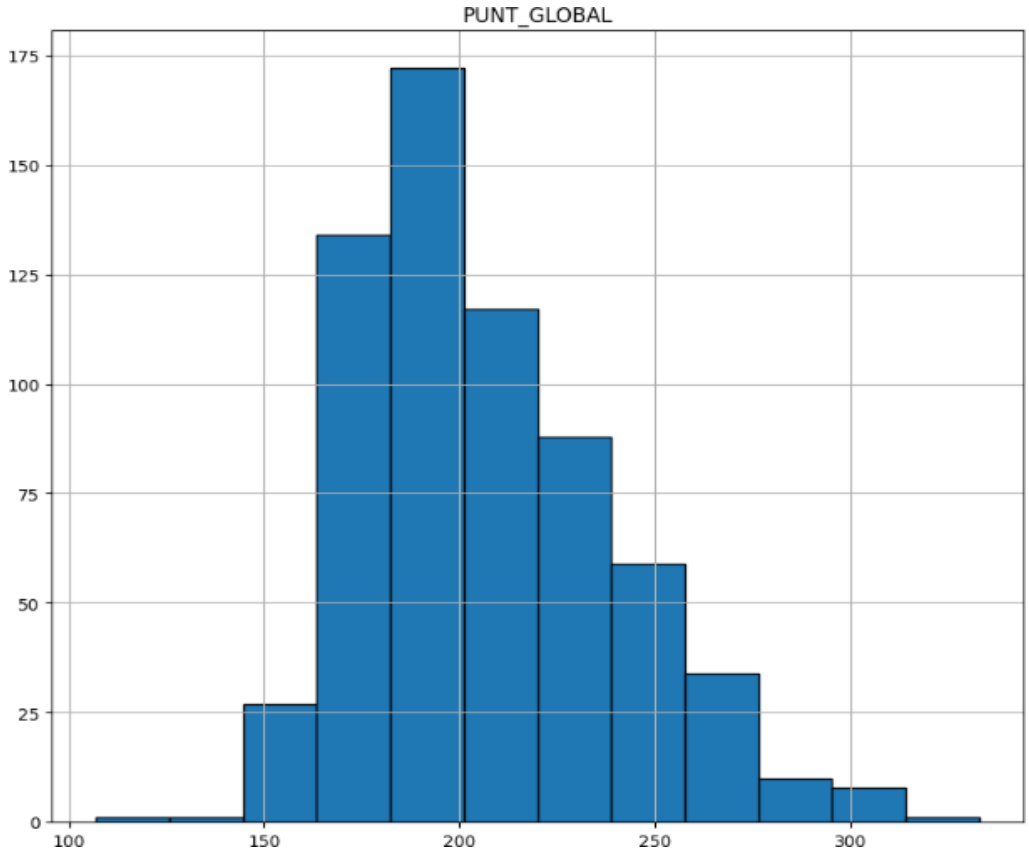
```
cole_bilingue = {
    "S": 1,
    "N": 0
}


info_depto_df['COLE_BILINGUE'] = info_depto_df['COLE_BILINGUE'].map(cole_bilingue)
info_depto_df['COLE_BILINGUE'] = info_depto_df['COLE_BILINGUE'].fillna(0)

info_depto_df.head(5)
```

Ilustración 14

```
[94] pilo_paga = {
        "'SER PILO PAGA - CREDITO CONDONABLE'": 1,
        "SER PILO PAGA - ETNIA": 2,
        "NO": 0
     }


     info_depto_df['ESTU_PILOPAGA'] = info_depto_df['ESTU_PILOPAGA'].map(pilo_paga)
     info_depto_df['ESTU_PILOPAGA'] = info_depto_df['ESTU_PILOPAGA'].fillna(0)

     info_depto_df.head(5)
```

IV. Gráficos

Ilustración 15. Histograma puntaje global



La gráfica indica que la mayoría de los puntajes se concentran entre los **180 y 200 puntos**. Tal distribución revela que la mayoría de los individuos obtuvo resultados en un rango medio-bajo, con pocos casos en los extremos superiores del puntaje.

Ilustración 16

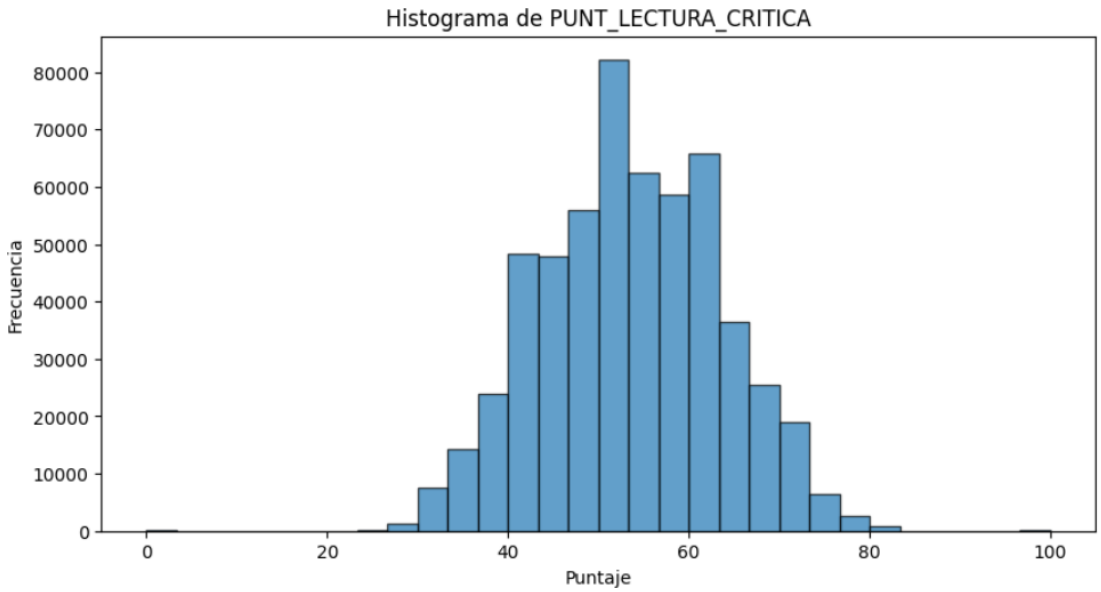
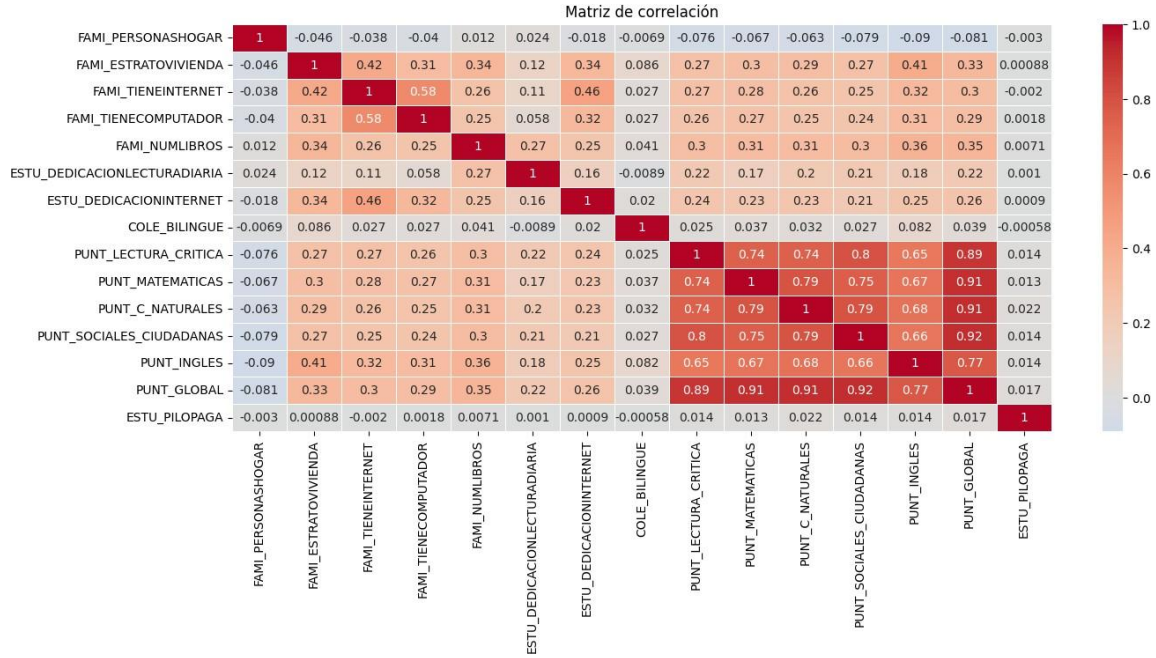
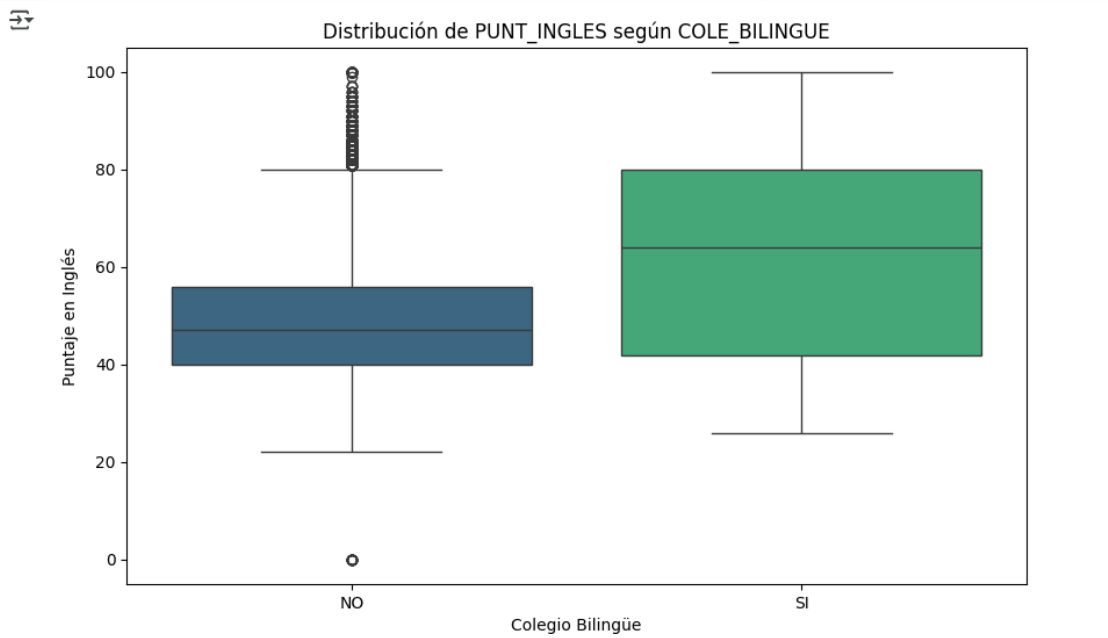


Ilustración 17. Matriz de correlación



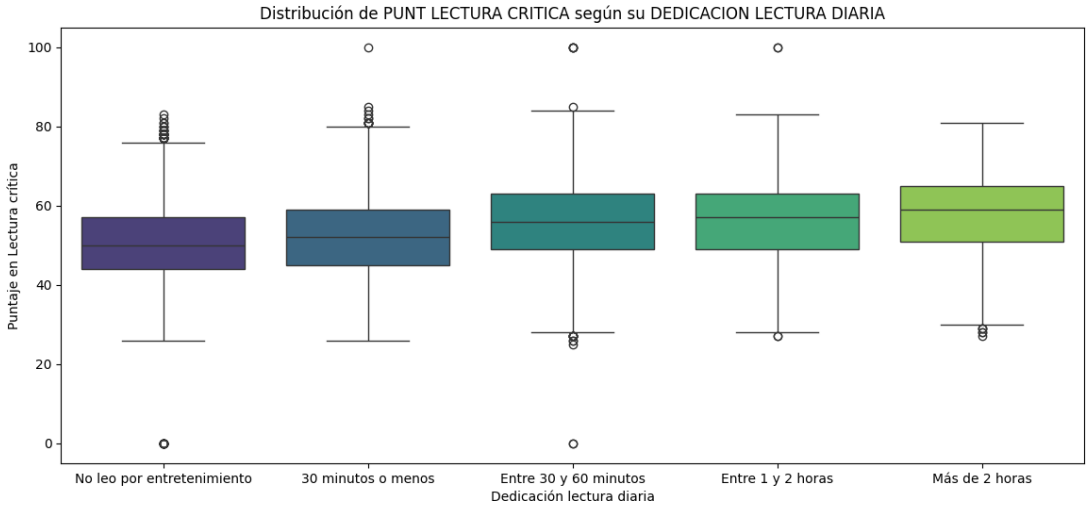
Como no se evidencia una correlación entre variables que deberían estar relacionadas se decide mirar esta relación por medio de otros gráficos.

Ilustración 18. Relacionamiento de variables



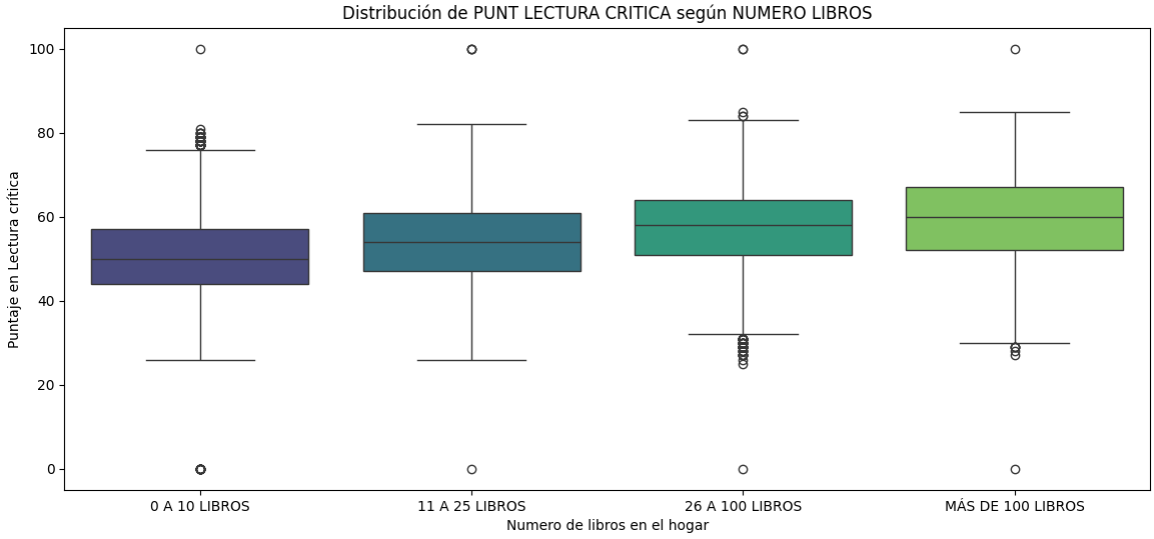
El boxplot muestra la distribución de los **puntajes en inglés según la asistencia a un colegio bilingüe**. Se observa que los estudiantes de colegios bilingües tienden a obtener puntajes más altos en inglés, con una mediana más elevada y una mayor dispersión hacia los valores altos. En contraste, los estudiantes de colegios no bilingües tienen una mediana más baja y más presencia de valores atípicos, lo cual indica un rendimiento más bajo.

Ilustración 19. Relación entre el Hábito de Lectura y el Desempeño en Lectura Crítica



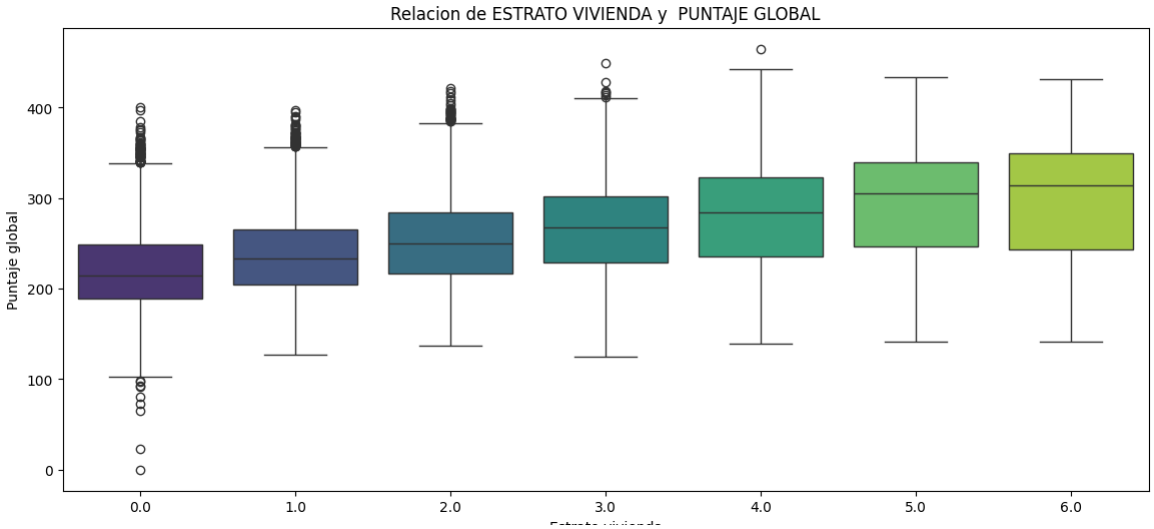
El gráfico de cajas muestra la distribución de los puntajes en Lectura Crítica según el tiempo que los estudiantes dedican diariamente a la lectura por entretenimiento. Se observa una tendencia general: a mayor dedicación diaria a la lectura, mejores son los puntajes obtenidos. Los estudiantes que leen más de dos horas al día presentan una mediana más alta y una distribución más favorable. En contraste, quienes no leen por entretenimiento tienden a obtener resultados más bajos. Esto sugiere una posible relación positiva entre el hábito de lectura y el rendimiento en lectura crítica.

Ilustración 20. Relación entre la Cantidad de Libros en el Hogar y el Desempeño en Lectura Crítica



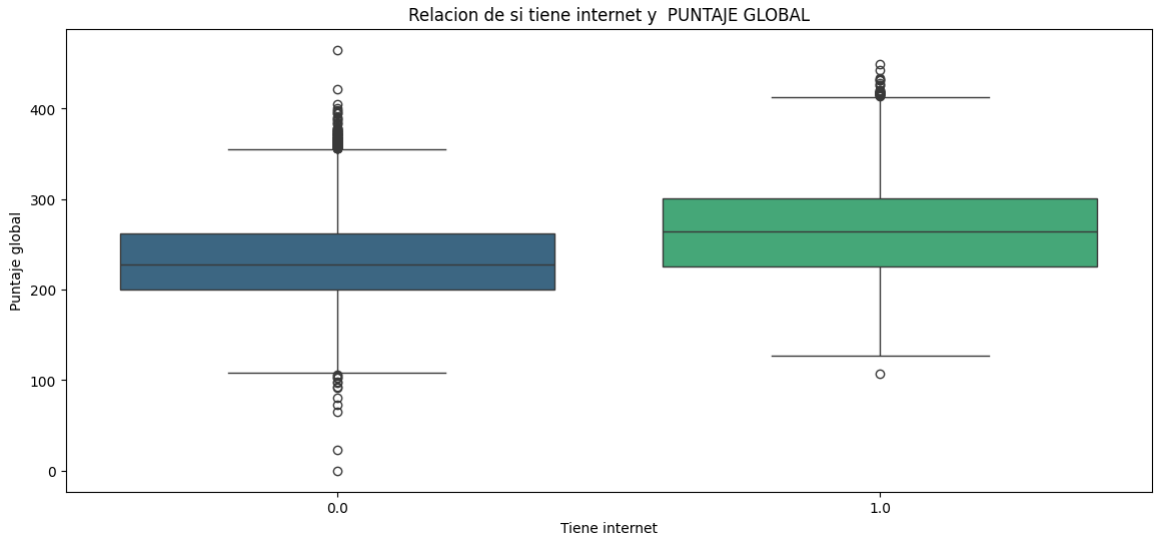
Se muestra la distribución de los puntajes en lectura crítica según el número de libros disponibles en el hogar. Se observa una tendencia ascendente en los puntajes a medida que aumenta la cantidad de libros. Los estudiantes que reportan tener más de 100 libros en casa presentan una mediana más alta y una distribución desplazada hacia valores superiores, lo cual sugiere una asociación positiva entre el acceso a libros y el rendimiento en lectura crítica. En contraste, aquellos con entre 0 y 10 libros presentan las medianas más bajas y mayor concentración de puntajes bajos.

Ilustración 21. Relación entre el Nivel Socioeconómico y el Desempeño Global



Este otro gráfico de cajas muestra la relación entre el **estrato de vivienda y el puntaje global**, evidenciando una tendencia creciente del puntaje a medida que aumenta el estrato socioeconómico. Los estratos más bajos (como el 0 y 1) presentan medianas más bajas y una mayor dispersión, lo que indica un menor desempeño global y mayor variabilidad. En contraste, los estratos más altos (como el 5 y 6) muestran puntajes globales más elevados y concentrados. Esta visualización refleja cómo las condiciones socioeconómicas podrían influir en los resultados obtenidos.

Ilustración 22. Conectividad a Internet vs. el puntaje global.



Se compara el puntaje global de personas que tienen internet versus las que no lo tienen. Se observa que la mediana del puntaje global es más alta para quienes tienen internet, y aunque ambos grupos presentan valores atípicos, el grupo con internet muestra una menor dispersión en sus puntajes centrales.

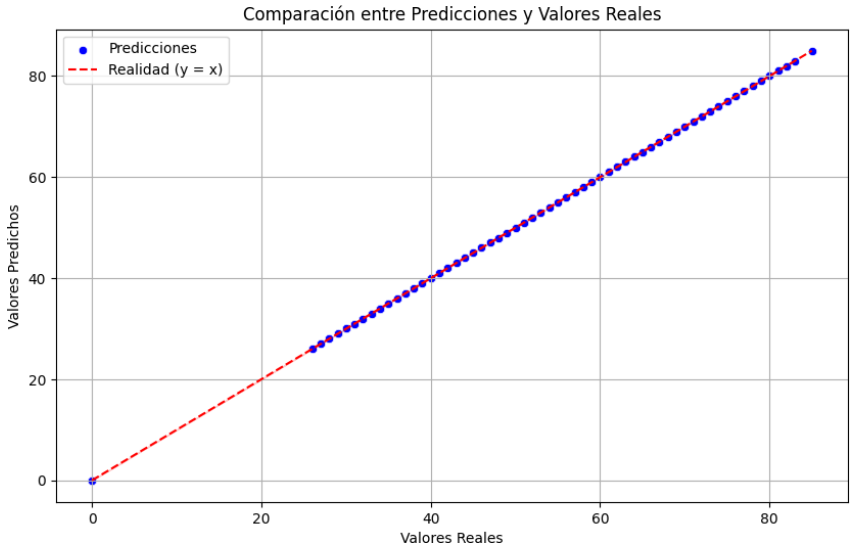
V. Regresión Lineal

Al verificar la matriz de correlación se determina que existen variables relacionadas. Por lo que se aplica el modelo de regresión lineal para predecir el puntaje de lectura crítica de los estudiantes. Primero, convierte las variables categóricas en valores numéricos usando LabelEncoder.

Ilustración 23. Modelo de regresión lineal

```python
label_encoder = LabelEncoder()

categorical_columns = info_depto_df.select_dtypes(include=['object']).columns

def label_encode_columns(df, columns):
    for column in columns:
        df[column] = label_encoder.fit_transform(df[column].astype(str))
    return df

info_depto_df = label_encode_columns(info_depto_df, categorical_columns)

X = info_depto_df.drop('ESTU_DEDICACIONLECTURADIARIA', axis=1)
y = info_depto_df['PUNT_LECTURA_CRITICA']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

print("Mean Squared Error (MSE):", mean_squared_error(y_test, y_pred))
print("Root Mean Squared Error (RMSE):", np.sqrt(mean_squared_error(y_test, y_pred)))
print("Mean Absolute Error (MAE):", mean_absolute_error(y_test, y_pred))
print("R^2 Score:", r2_score(y_test, y_pred))
```

Ilustración 24. Gráfico de líneas



En este gráfico de líneas se compara los valores reales con las predicciones del modelo de regresión lineal. Los puntos azules representan las predicciones, y la línea roja punteada indica la línea ideal donde las predicciones coinciden exactamente con los valores reales. El objetivo es visualizar la precisión del modelo: cuanto más cercanos estén los puntos a la línea roja, mejor es su desempeño.

Conclusiones

Los resultados del estudio revelan una tendencia ascendente en el puntaje global del ICFES en función del estrato socioeconómico: los estudiantes pertenecientes a estratos más altos alcanzan, en promedio, mejores desempeños, mientras que aquellos de estratos bajos presentan resultados más bajos y una mayor dispersión. Esta relación evidencia la persistencia de desigualdades en el acceso a una educación de calidad.

De igual manera, se identificó una relación positiva entre la disponibilidad de libros en el hogar y los puntajes obtenidos en lectura crítica, lo cual sugiere que un entorno familiar con hábitos lectores favorece el desarrollo de competencias fundamentales para el rendimiento académico.

La conectividad y el acceso a recursos tecnológicos también muestran una incidencia significativa: los estudiantes que cuentan con computador e internet en sus hogares obtienen puntajes superiores, en comparación con quienes carecen de estos medios. Este hallazgo subraya la importancia de cerrar la brecha digital para garantizar la igualdad de oportunidades educativas.

Adicionalmente, el tamaño del hogar y el contexto familiar también son factores relevantes. En particular, una mayor cantidad de integrantes en el hogar podría limitar el rendimiento académico debido a restricciones en el acceso a recursos, menor atención personalizada y condiciones poco adecuadas para el estudio en el ámbito doméstico.

A partir de estos hallazgos, se recomienda ampliar el análisis mediante la incorporación de nuevas variables y la aplicación de metodologías más robustas, que permitan obtener conclusiones de mayor precisión. Ello contribuiría a sustentar decisiones de política pública y el diseño de intervenciones educativas más eficaces, orientadas a mitigar las desigualdades y promover el mejoramiento del desempeño académico en la región.