



# Complejidad Económica Provincial

Docente: Nicolas Sidicaró

Alumnos: Nicolas Maidana y José Ignacio Rolón

Ciencia de Datos para Economía y Negocios

\*Base de datos + Metodología propia de: <https://argendata.fund.ar/>



# INDICE

1. Carga de datos
2. Limpieza de datos
3. Procesamiento de datos
4. Test ANOVA
5. Test de Hipótesis
6. Conclusiones
7. Futuras líneas de investigación





# 1. Carga de datos

- Importa las bases de **datos originales**.
- **Documenta** la fuente, el nombre y las dimensiones.
- Preparación optimizada dentro el programa. (**Ruteo y tipo de archivo**)

\*Garantiza **integridad y trazabilidad** de los datos al inicio del **flujo de trabajo**, **aislando los datos originales** de cualquier modificación posterior.

# 2. Limpieza de bases

- **Estandarización** de variables (strings y numéricas).
- **Reducción dimensional** no útiles para tratamiento posterior

\***Asegura la calidad** de los datos, creando un **conjunto de datos estandarizado y optimizado**



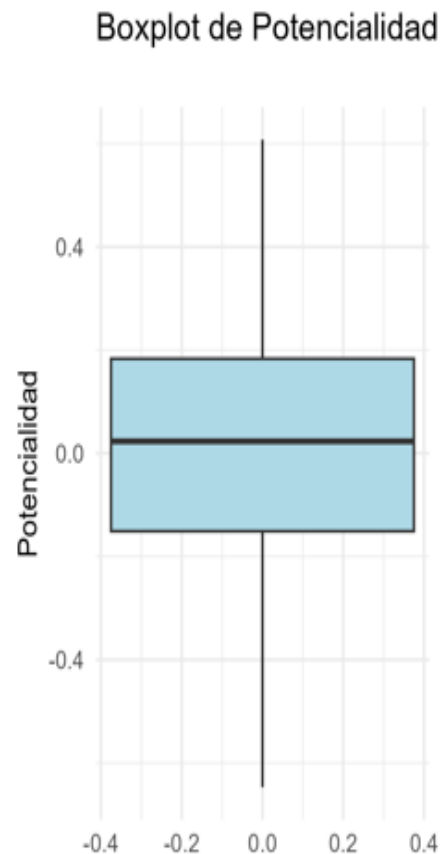
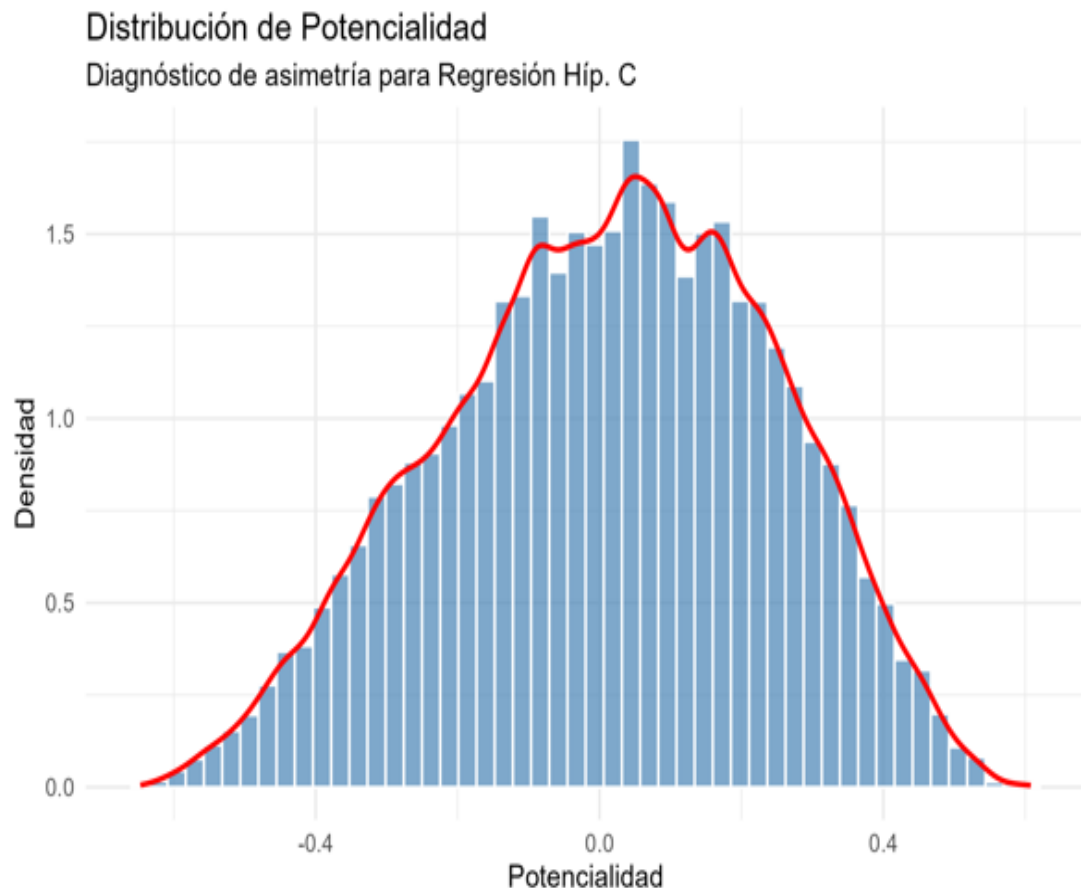
### 3. Análisis Exploratorio de Datos (EDA)

- Base Potencial
- Base Export

- **Verificación Estructural** sobre datos/variables.
- Manejo de **Datos Faltantes** (NAs): cuantificación y diagnóstico específico.
- **Estadísticas Descriptivas**: métricas clave (**Media**, **Mediana**, Desviación Estándar, IQR, Mín/Máx) para las **variables continuas**.
- **Visualización para Diagnóstico**: boxplots + distribución

\***Verifica** la calidad de los **datos**, confirma la necesidad de **manejar la asimetría** y el **rol de los outliers** para nuestras **variables clave**, e identifica el **patrón** de los **NAs**, **guiando** las **transformaciones** y el **modelado inferencial** subsiguiente.

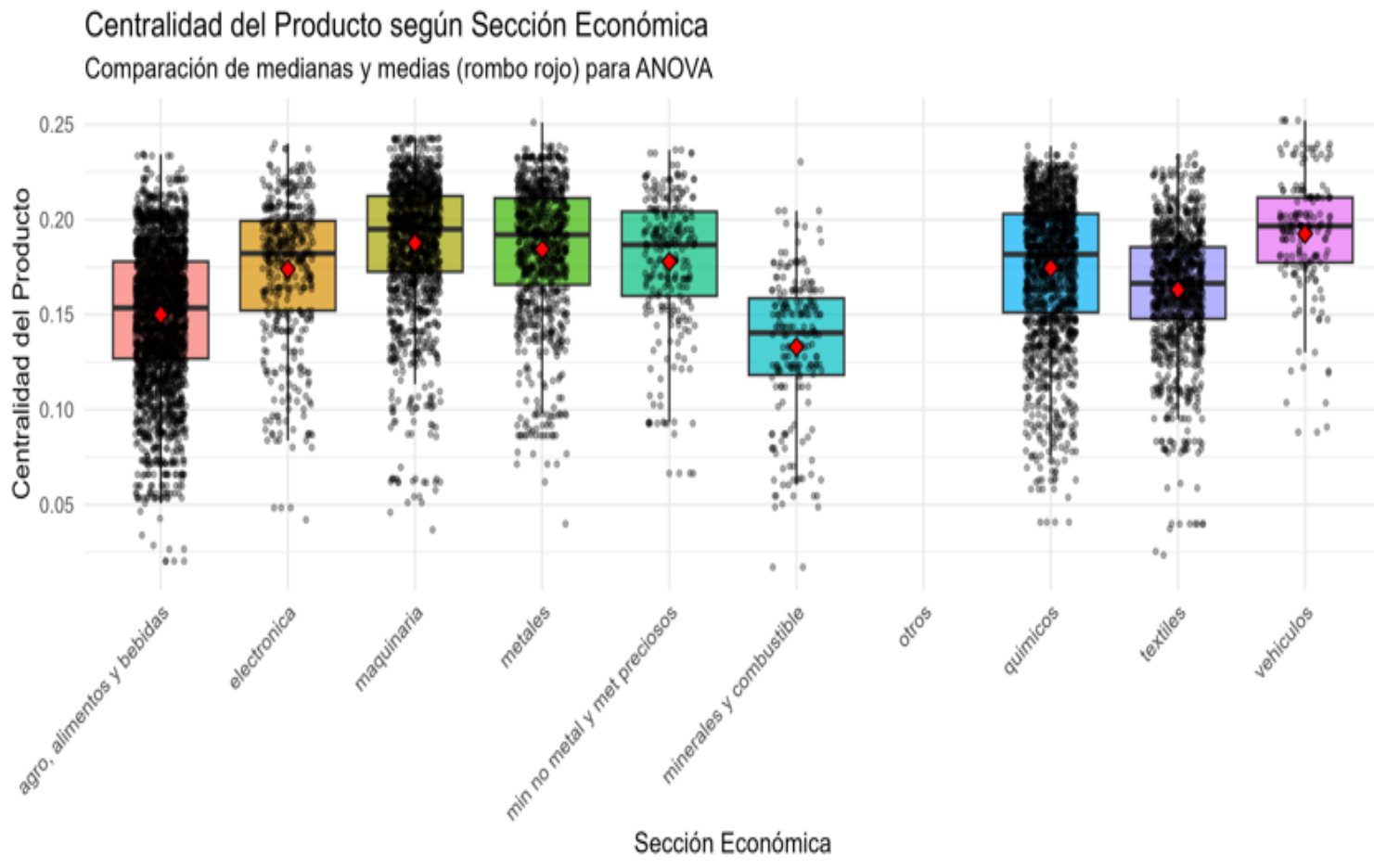
# 3. Base Pontecialidad (EDA)



```
--- DESCRIPTIVAS CLAVE (BASE POTENCIAL) ---
> print(descriptivas_potencial,)
```

	Estadística	valor
1	n	127266.000000000
2	potencialidad_Media	0.010987835
3	potencialidad_Mediana	0.023300000
4	potencialidad_Desvio_Std	0.231137929
5	potencialidad_IQR	0.334300000
6	potencialidad_Min	-0.647100000
7	potencialidad_Max	0.607900000
8	complejidad_producto_Media	0.007711314
9	complejidad_producto_Mediana	0.157700000
10	complejidad_producto_Desvio_Std	0.998865321
11	complejidad_producto_IQR	1.424600000
12	complejidad_producto_Min	-4.657100000
13	complejidad_producto_Max	2.668500000
14	distancia_Media	0.986160829
15	distancia_Mediana	0.991400000
16	distancia_Desvio_Std	0.013739186
17	distancia_IQR	0.014000000
18	distancia_Min	0.801300000
19	distancia_Max	1.000000000
20	fob_mundial_Media	3855235.626321563
21	fob_mundial_Mediana	684275.280000000
22	fob_mundial_Desvio_Std	21432366.394400638
23	fob_mundial_IQR	2133088.700000000
24	fob_mundial_Min	16.110000000
25	fob_mundial_Max	985996987.220000029
26	n_Media	127266.000000000
27	n_Mediana	127266.000000000
28	n_Desvio_Std	NA
29	n_IQR	0.000000000
30	n_Min	127266.000000000
31	n_Max	127266.000000000

# 3. Base Exportados (EDA)



```
--- DESCRIPTIVAS CLAVE (BASE EXPORT) ---  
> print(descriptivas_export,)
```

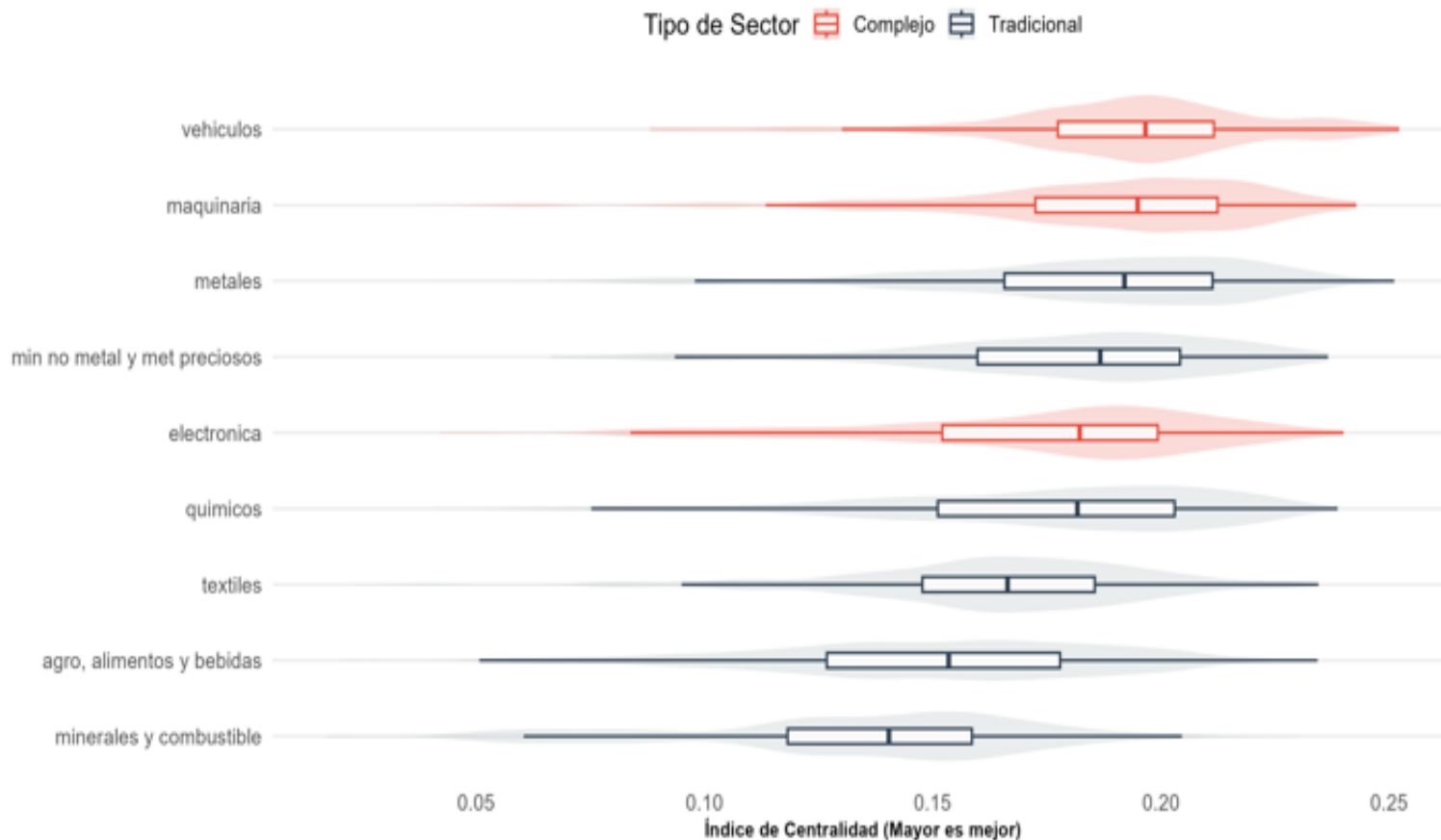
	Estadística	valor
1	n	8519.00000000
2	centralidad_Media	0.16833717
3	centralidad_Mediana	0.17430000
4	centralidad_Desvio_Std	0.03940722
5	centralidad_IQR	0.05415000
6	centralidad_Min	0.01720000
7	centralidad_Max	0.25220000
8	complejidad_producto_Media	-0.08527372
9	complejidad_producto_Mediana	0.04970000
10	complejidad_producto_Desvio_Std	0.97142310
11	complejidad_producto_IQR	1.43840000
12	complejidad_producto_Min	-3.85250000
13	complejidad_producto_Max	2.62690000
14	complejidad_provincia_Media	-0.36235559
15	complejidad_provincia_Mediana	-0.33486644
16	complejidad_provincia_Desvio_Std	0.31022827
17	complejidad_provincia_IQR	0.19610411
18	complejidad_provincia_Min	-1.33322010
19	complejidad_provincia_Max	0.08786897
20	n_Media	8519.00000000
21	n_Mediana	8519.00000000
22	n_Desvio_Std	NA
23	n_IQR	0.00000000
24	n_Min	8519.00000000
25	n_Max	8519.00000000



### 3. Base Exportados (EDA)

#### Dime qué produces y te diré qué tan conectado estás

Distribución de la Centralidad del Producto según Sector Económico.  
Los sectores de alta complejidad técnica dominan el centro del espacio.



```
--- DESCRIPTIVAS CLAVE (BASE EXPORT) ---  
> print(descriptivas_export,)
```

	Estadística	valor
1	n	8519.00000000
2	centralidad_Media	0.16833717
3	centralidad_Mediana	0.17430000
4	centralidad_Desvio_Std	0.03940722
5	centralidad_IQR	0.05415000
6	centralidad_Min	0.01720000
7	centralidad_Max	0.25220000
8	complejidad_producto_Media	-0.08527372
9	complejidad_producto_Mediana	0.04970000
10	complejidad_producto_Desvio_Std	0.97142310
11	complejidad_producto_IQR	1.43840000
12	complejidad_producto_Min	-3.85250000
13	complejidad_producto_Max	2.62690000
14	complejidad_provincia_Media	-0.36235559
15	complejidad_provincia_Mediana	-0.33486644
16	complejidad_provincia_Desvio_Std	0.31022827
17	complejidad_provincia_IQR	0.19610411
18	complejidad_provincia_Min	-1.33322010
19	complejidad_provincia_Max	0.08786897
20	n_Media	8519.00000000
21	n_Mediana	8519.00000000
22	n_Desvio_Std	NA
23	n_IQR	0.00000000
24	n_Min	8519.00000000
25	n_Max	8519.00000000

# 4. Test ANOVA

## Centralidad por Sección

- Verificación de supuestos
  - Test de Levene (Homogeneidad de Varianzas)
- Modelo Ajustado
  - Test de Welch (Robusto ante Heterogeneidad)
- Análisis Post-Hoc
  - Test de Games-Howell (Visualización)



# 4. Test de Levene (Homogeneidad de Varianzas)

Objeto de estudio --> **CENTRALIDAD x SECCION**

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \dots = \sigma_k^2$$

$H_1$ : Al menos una de las varianzas poblacionales es diferente

Regla de decisión

# Si p-valor < 0.05: Rechazo  $H_0$  -> Varianza NO es homogénea

# Si p-valor >= 0.05: NO rechazo  $H_0$  -> Varianza es homogénea

```
--- TEST DE LEVENE (Homogeneidad de Varianzas) ---
> print(levene_test_resultado)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value      Pr(>F)
group   8  9.3721 0.0000000000005734 ***
      8470
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se **rechazó  $H_0$**  (varianzas no homogéneas).



Considerar **ANOVA de Welch**  
(Alternativa robusta a heterogeneidad)

# 4. Modelo Ajustado (Test Welch)

Objeto de estudio --> **CENTRALIDAD x SECCION**

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \dots = \sigma_k^2$$

$H_1$ : Al menos una de las varianzas poblacionales es diferente

```
--- ANOVA DE WELCH (Robusto ante heterogeneidad) ---  
  
one-way analysis of means (not assuming equal variances)  
  
data:  centralidad and seccion_f  
F = 198.19, num df = 8.0, denom df = 1446.1, p-value <  
0.000000000000000022
```

## Regla de decisión

# Si p-valor < 0.05: **Rechazo  $H_0$**  -> hay una **diferencia estadísticamente significativa** entre las medias de los grupos

# Si p-valor >= 0.05: NO rechazo  $H_0$  -> No hay evidencia suficiente para rechazar  $H_0$ .



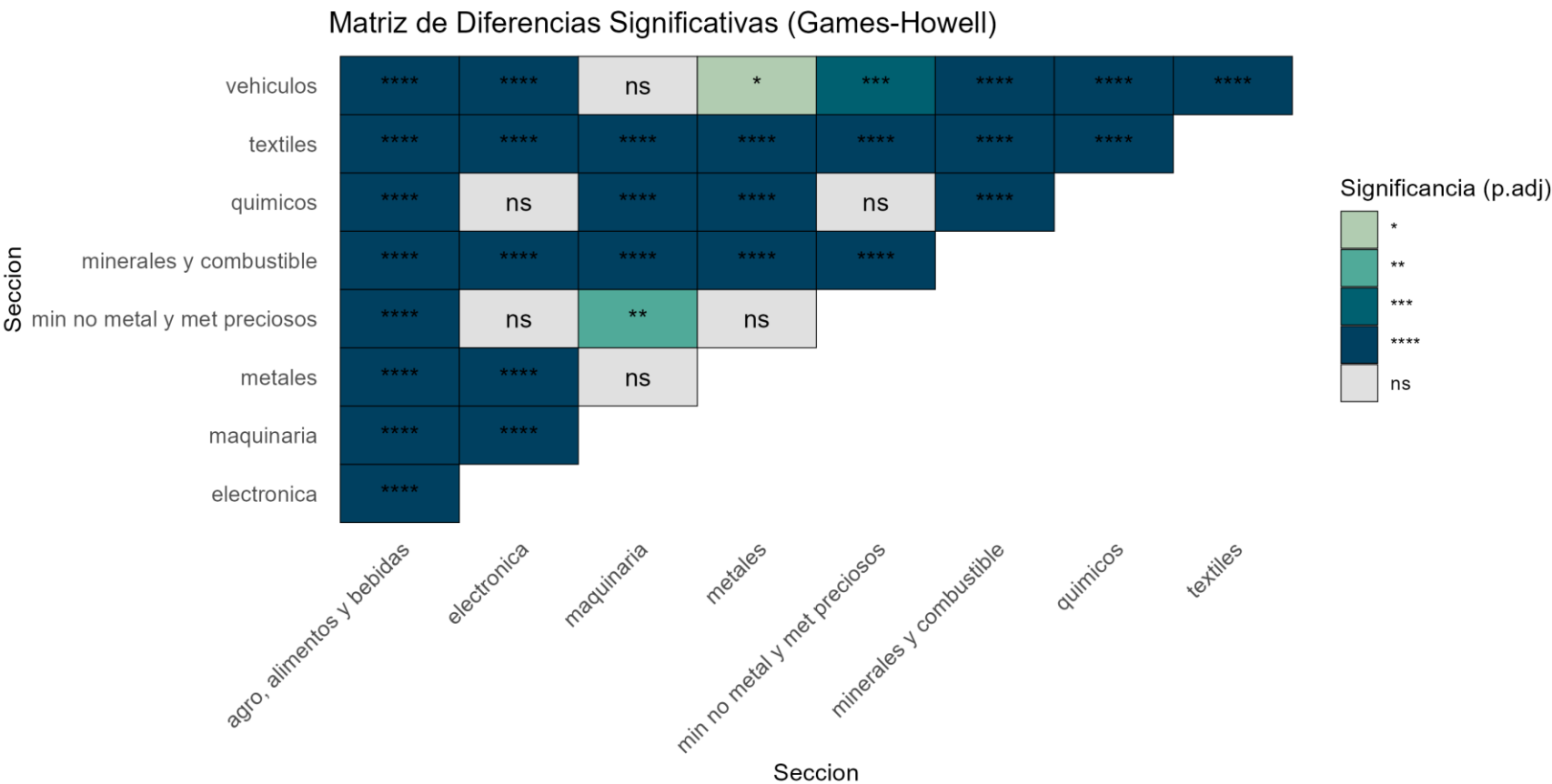
Se **rechazó  $H_0$**  (Existe diferencia significativa en Centralidad por Sector).

**Avanzamos con análisis post\_hoc para cuantificar y visualizar los resultados.**



# 4. Análisis Post-Hoc (Games-Howell)

Objeto de estudio --> **CENTRALIDAD x SECCION**



# 5. Test de Hipótesis

## Modelo de Regresión Lineal Múltiple

- Hipótesis: Presentación de Modelo
  - Método de Mínimos Cuadrados Ordinarios
- Verificación de supuestos
  - Multicolinealidad
  - Heterocedasticidad
- Resultados y Visualización



# 5. Modelo de Regresion Lineal (MCO)

Objeto de estudio --> **POTENCIALIDAD**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + C_1 + C_2$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\alpha}_1 C_1 + \hat{\alpha}_2 C_2 + \mu$$

```
# HIPÓTESIS C:  
# Y = Potencialidad (winsorizada)  
# X1 = Complejidad Producto (winsorizada)  
# X2 = Distancia  
# Control 1 = Tamaño de Mercado (Log FOB Mundial)  
# Control 2 = Sector (Sección)
```

Método de Mínimos  
Cuadrados Ordinarios



```
--- RESUMEN PRELIMINAR (MCO clásico) ---  
> print(summary(modelo_c))  
  
call:  
lm(formula = formula_hipotesis_c, data = df_pot_transf)  
  
Residuals:  
      Min       1Q   Median       3Q      Max   
-0.36219 -0.06361 -0.00462  0.05692  0.76110  
  
Coefficients:  
              Estimate Std. Error t value      Pr(>|t|)  
(Intercept)   -0.13428217  0.01883833   -7.128 0.000000000000102  
complejidad_producto_win  0.19993145  0.00032551  614.208 < 0.00000000000000002  
distancia      0.18760488  0.01898665    9.881 < 0.00000000000000002  
log_fob_mundial -0.00306745  0.00013243  -23.162 < 0.00000000000000002  
seccionelectronica  0.00007245  0.00134451    0.054      0.95702  
seccionmaquinaria  0.03116950  0.00094874   32.853 < 0.00000000000000002  
seccionmetales    0.00684403  0.00100697    6.797 0.000000000001075  
seccionmin no metal y met preciosos 0.00429619  0.00147316    2.916      0.00354  
seccionminerales y combustible  0.04560105  0.00167337   27.251 < 0.00000000000000002  
seccionquimicos   0.01530612  0.00086073   17.783 < 0.00000000000000002  
secciontextiles   -0.06885337  0.00082887  -83.069 < 0.00000000000000002  
seccionvehiculos   0.02828681  0.00172120   16.434 < 0.00000000000000002
```

Verificación de supuestos



```
(Intercept)          ***  
complejidad_producto_win ***  
distancia            ***  
log_fob_mundial      ***  
seccionelectronica   ***  
seccionmaquinaria    ***  
seccionmetales       ***  
seccionmin no metal y met preciosos **  
seccionminerales y combustible ***  
seccionquimicos      ***  
secciontextiles      ***  
seccionvehiculos     ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.09201 on 127254 degrees of freedom  
Multiple R-squared:  0.8392,    Adjusted R-squared:  0.8392  
F-statistic: 6.037e+04 on 11 and 127254 DF,  p-value: < 0.00000000000000002
```

# 5. Supuestos: Multicolinealidad y Heterocedasticidad

Objeto de estudio --> **POTENCIALIDAD**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + C_1 + C_2$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\alpha}_1 C_1 + \hat{\alpha}_2 C_2 + \mu$$

```
# HIPÓTESIS C:  
# Y = Potencialidad (winsorizada)  
# X1 = Complejidad Producto (winsorizada)  
# X2 = Distancia  
# Control 1 = Tamaño de Mercado (Log FOB Mundial)  
# Control 2 = Sector (Sección)
```

**Ajuste con Test White y test t para cada variable (Ver tabla)**

✅ CONFIRMA H1: A mayor complejidad, mayor potencialidad.

🔍 HALLAZGO (CONTRA-INTUITIVO): Relación POSITIVA y SIGNIFICATIVA.  
Interpretación: Los productos con mayor potencial estratégico son los más 'lejanos'.

## Multicolinealidad

Control vía: DIAGNÓSTICO VIF (Factor de Inflación de Varianza)

```
complejidad_producto_win      distancia      log_fob_mundial  
1.519290                      1.022990      1.095799  
seccion  
1.648143  
✅ Multicolinealidad bajo control.
```

## Heterocedasticidad

Control vía: Test de Breusch-Pagan

```
--- TEST DE BREUSCH-PAGAN ---  
> print(bp_test)  
  
studentized Breusch-Pagan test  
  
data: modelo_c  
BP = 11852, df = 11, p-value < 0.00000000000000022
```

✅ Se rechaza H0: Heterocedasticidad detectada.  
🔍 CORRECCIÓN: Se utilizarán Errores Estándar Robustos (HC1) para la inferencia final.



# 5. Modelo de Regresion Lineal

## Complejidad y Distancia: Motores de la Diversificación Productiva

Relación entre el valor de un producto (Complejidad) y la oportunidad que genera (Potencialidad), segmentada por la dificultad para alcanzarlo (Distancia).



Distancia a la Capacidad Actual    ○ 1. Baja Distancia (Cercanos)    ○ 2. Distancia Media    ○ 3. Alta Distancia (Lejanos)

Nota: La pendiente positiva confirma que la complejidad impulsa el potencial. La separación de las líneas confirma el hallazgo de la regresión: los productos 'Lejanos' tienen mayor potencial para el mismo nivel de complejidad.

## 6. Conclusiones

- Nuestros resultados **refutan** la hipótesis de que la "**lejanía**" **productiva** disminuye el valor de la oportunidad; por el contrario, la evidencia sugiere que la ganancia de **oportunidad estratégica** (opportunity gain) es una **función creciente de la distancia** a recorrer.
- El modelo econométrico planteado demuestra que la estructura de incentivos de la diversificación provincial presenta un **trade-off**: mientras que la complejidad garantiza valor agregado, las oportunidades de **mayor potencial** no se encuentran en la "zona de confort" de la baja distancia productiva, sino que **requieren saltos de capacidades significativos**.
- En pos de pensar un **desarrollo potencial** y la estructura de oportunidades actuales de cada provincia, resulta fundamental revisar las **posibilidades y estrategias**, y sus consecuentes **riesgos**. La política económica, debe **evitar fórmulas únicas**. En cambio, se vuelve crucial que cada provincia desarrolle una **hoja de ruta estratégica** basada en la **maximización de su propia función** de *opportunity gain* revelada. El análisis de complejidad debe ser la base para diseñar **paquetes de incentivos a medida** que promuevan los saltos específicos necesarios para su estructura productiva.



## 7. Futuras Lineas de Investigación

- Los resultados actuales, como el **trade-off entre Potencialidad y Distancia**, requieren una validación y profundización, especialmente en lo que respecta a la **inferencia causal**, que es inherentemente limitada en un análisis de corte transversal
- Incorporación de la **Dimensión Temporal (Datos de Panel)**: La estructura de corte transversal actual nos limita a observar las oportunidades en un solo punto en el tiempo. El análisis ganaría un valor sustancial si se migrara a una estructura donde se pudieran **observar las mismas provincias a lo largo del tiempo**. En esta línea, se podría realizar un análisis de **trayectorias de diversificación** estudiando cómo la Potencialidad (o la Centralidad) cambian en el tiempo. Esto permitiría estimar la Propensión de Largo Plazo (PLP) de la complejidad o distancia en el tiempo.
- Por último, se podría efectuar una **expansión del ANOVA y Comparaciones Múltiples**

El hallazgo de diferencias significativas en la centralidad por sección justifica un análisis más detallado. Un ANOVA de Dos Factores (Two-Way ANOVA) podría ser utilizado para un análisis más exhaustivo incluyendo otra variable categórica de interés, como la Región Geográfica (considerando que provincia es una variable categórica disponible), y evaluar la interacción entre sección y región. Esto respondería a la pregunta: "¿Varía el efecto sectorial (sección) sobre la centralidad según la región de la provincia?"