TALENTO TECH

TIC

Universidad Tecnológica de Bolívar

Preprocesamiento Datos

Ejecutor Técnico
Guillermo Bejarano Reyes

INTELIGENCIA ARTIFICIAL

Nivel Explorador

www.utb.edu.co/talento-tech

MAPA DE CONTENIDOS

PREPROCESAMIENTO DE DATOS

01

ESCALADO SIMPLE

02

MÍNIMO - MÁXIMO

03

Z-SCORE

Business Insights

Normalización de los Datos

El **escalado, normalización y estandarización** de variables numéricas, utilizadas sobre todo en Machine Learning, se utiliza para cambiar los valores de las características numéricas en el conjunto de datos a una escala común, sin distorsionar las diferencias en los rangos de valores ni perder información. Esto es especialmente útil cuando los **datos tienen variables que varían en escalas**, o cuando usas algoritmos que asumen que todos los datos están centrados alrededor de 0 y tienen una varianza en la misma escala

Importancia de la normalización de datos en el análisis estadístico

La manipulación (adición, eliminación, combinación, mutación) de su conjunto de datos para mejorar el entrenamiento del modelo de aprendizaje automático, lo que genera un mejor rendimiento y una mayor precisión

La multicolinealidad ocurre cuando las variables independientes (predictores) en un modelo de regresión están correlacionadas.

Fuente de Datos para práctica

```
#Escalado - Normalización - Estándarización
import pandas as pd

#Leer archivo excel e impirmir las primeras 5 filas
df = pd.read_excel('C:\\Otros\\UTB\\TalentoTech\\AnalisisDatos\\Material\\Semana_5\\Ejercicios\\DataNormalizacion.xlsx', she
print("\nArchivo EXCEL")
print(df.head()) #los primeros 5 registros
```

```
Archivo EXCEL
             Nombre   Salario  NroVeh
0       Luisa Lane   4500000       1
1     Maria Medina   6000000       2
2  Sergio Castillo   3900000       1
3      Pedro López  80000000       3
4   Catalina Gómez   5000000       2
```

Estadística Descriptiva – Fuente de Datos

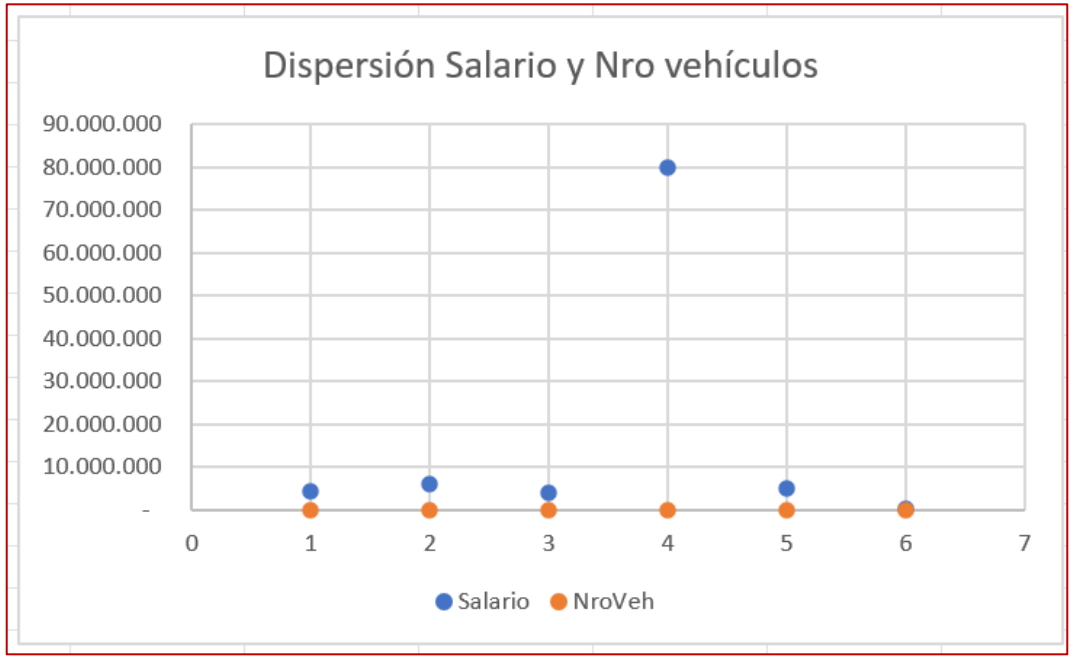
```
# Variable de Interés - Estadística Descriptiva
print("\n Variables de interés")
print("\n",df[["Salario","NroVeh"]]) # Mostrar llas variables de interés (Numéricas)

print("\n Estadística descriptiva")
print("\n",df[["Salario","NroVeh"]].describe()) # Mostrar Estadística Descriptiva
```

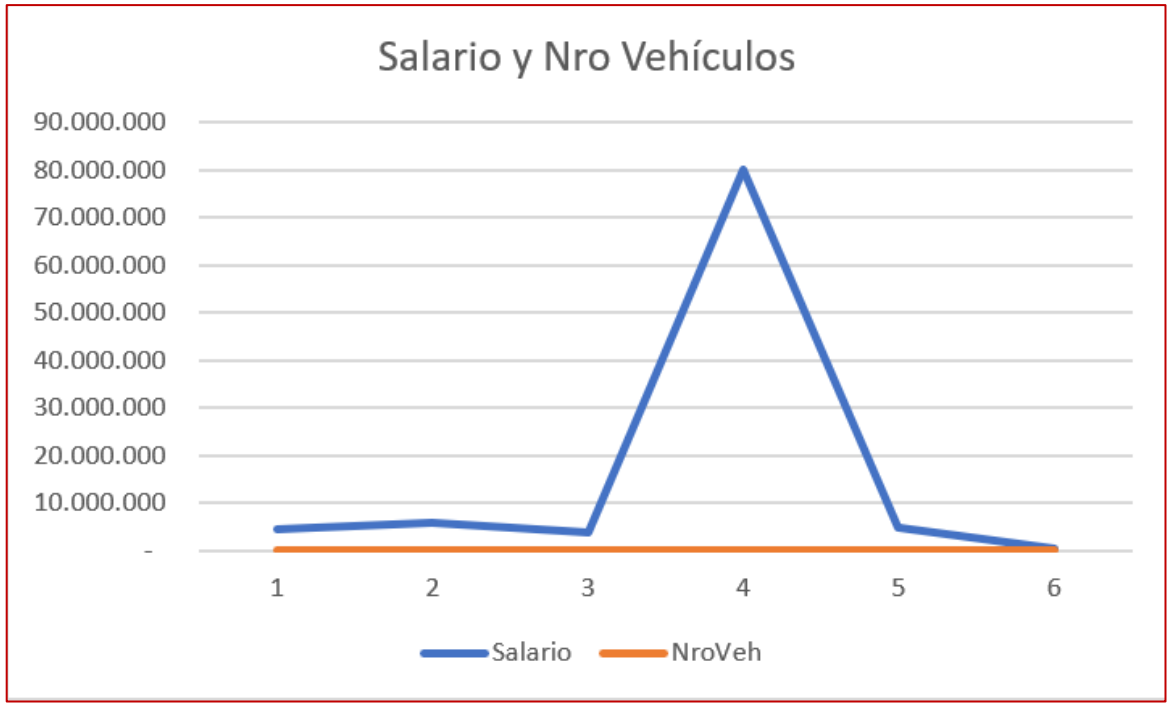
```
 Variables de interés

     Salario  NroVeh
0    4500000       1
1    6000000       2
2    3900000       1
3   80000000       3
4    5000000       2
5     400000       0
```

```
 Estadística descriptiva

              Salario    NroVeh
count   6.000000e+00  6.000000
mean    1.663333e+07  1.500000
std     3.110181e+07  1.048809
min     4.000000e+05  0.000000
25%     4.050000e+06  1.000000
50%     4.750000e+06  1.500000
75%     5.750000e+06  2.000000
max     8.000000e+07  3.000000
```

Rangos muy diferentes en escala (Mínimos y Máximos)

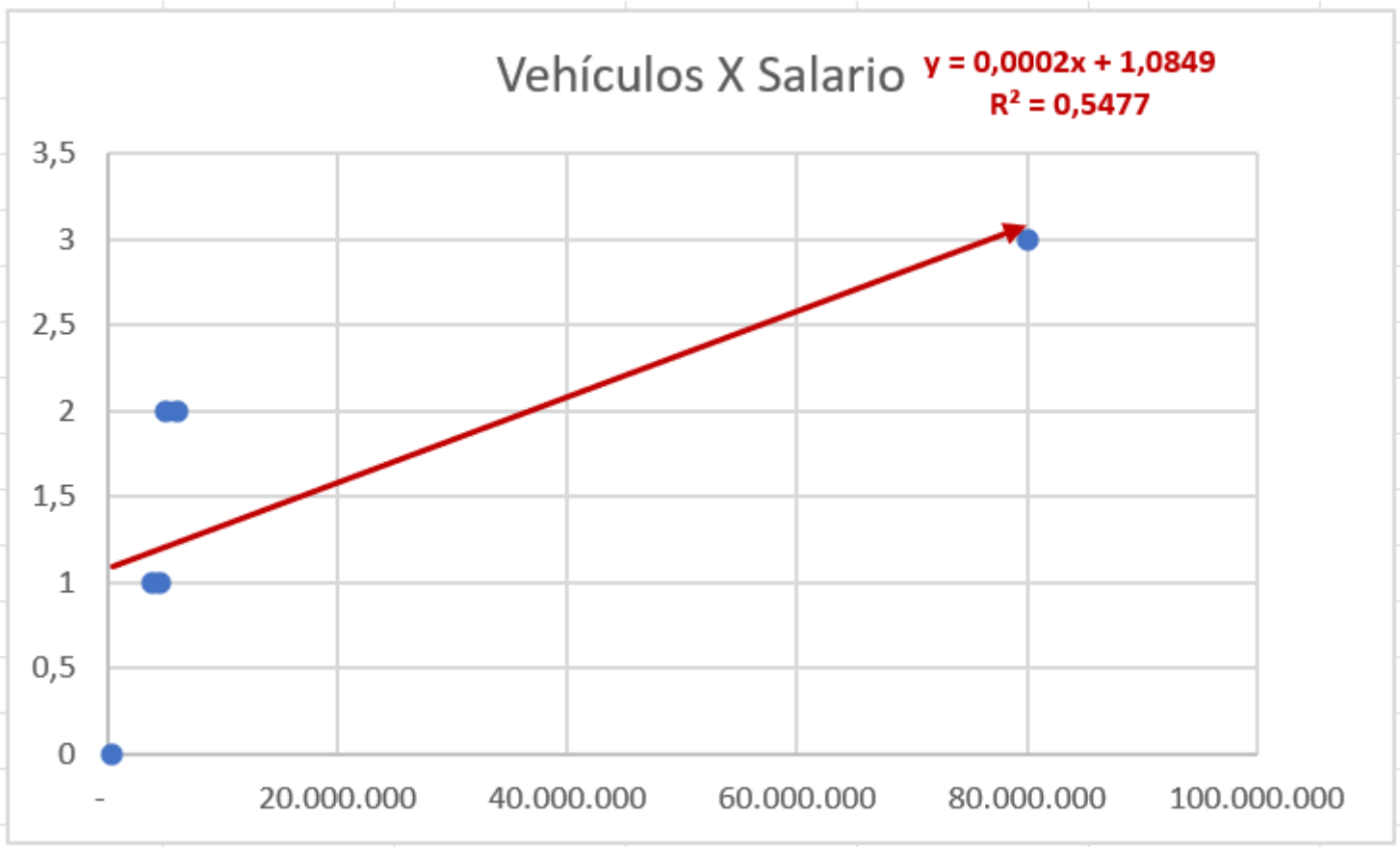
Gráfico de dispersión y de líneas – Fuente de Datos





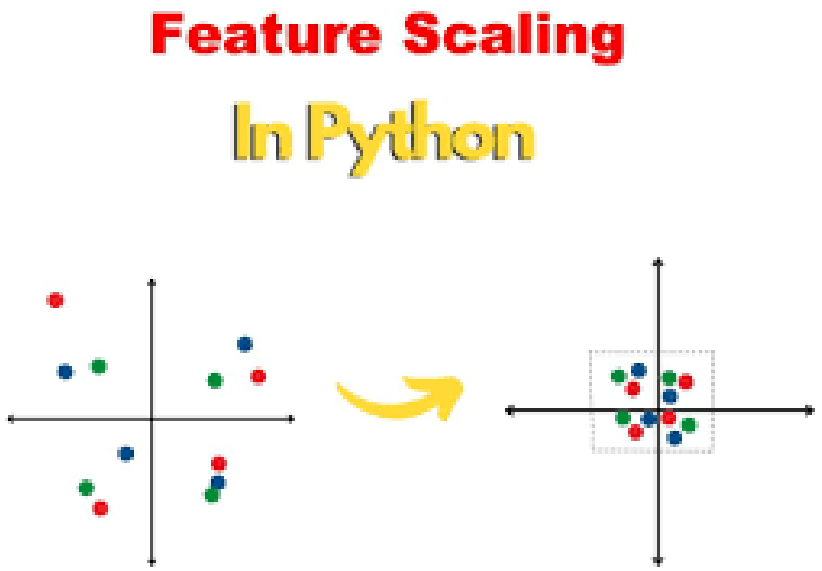
Por tener rangos tan diferentes el Salario y Nro. de vehículos, en sus gráficas no se puede apreciar la variable Nro. de vehículos

Regresión Lineal – Fuente de Datos



MÉTODO: ESCALADO SIMPLE

Método Escalado Simple



Fórmula:

$$N_i = \frac{(X_i)}{X_{max}}$$

Valor nuevo = (valor actual / Valor máximo)

Conservar la distribución de los datos en una menor escala

Método Escalado Simple - Programa

```
# Método Escalado Simple
print("\n Variables de interés")
print("\n",df[["Salario","NroVeh"]]) # Mostrar llas variables de interés (Numéricas)

print("\n Estadística descriptiva")
print("\n",df[["Salario","NroVeh"]].describe()) # Mostrar Estadística Descriptiva

print("\n Método escalado simple aplicado al salario y Número vehículos")
df["Salario"]=df["Salario"]/df["Salario"].max()
df["NroVeh"]=df["NroVeh"]/df["NroVeh"].max()
print("\n",df[["Salario","NroVeh"]]) # Mostrar llas variables de interés (Numéricas)
```

Método: Escalado simple - Resultados

Variables de interés

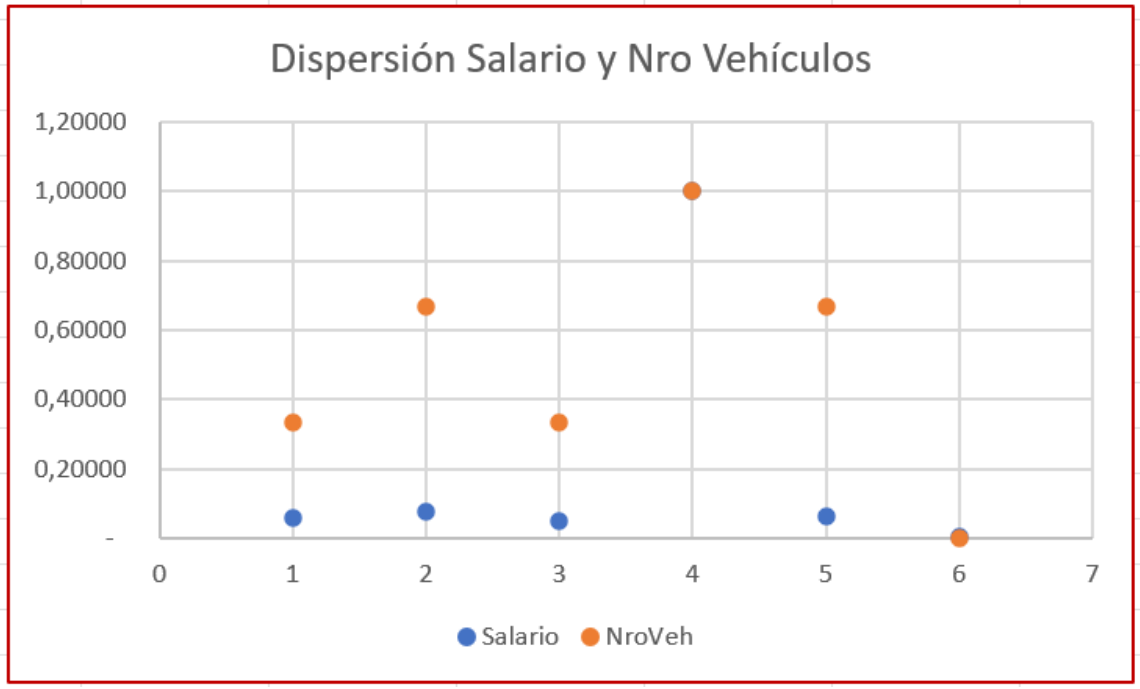
	Salario	NroVeh
0	4500000	1
1	6000000	2
2	3900000	1
3	80000000	3
4	5000000	2
5	400000	0

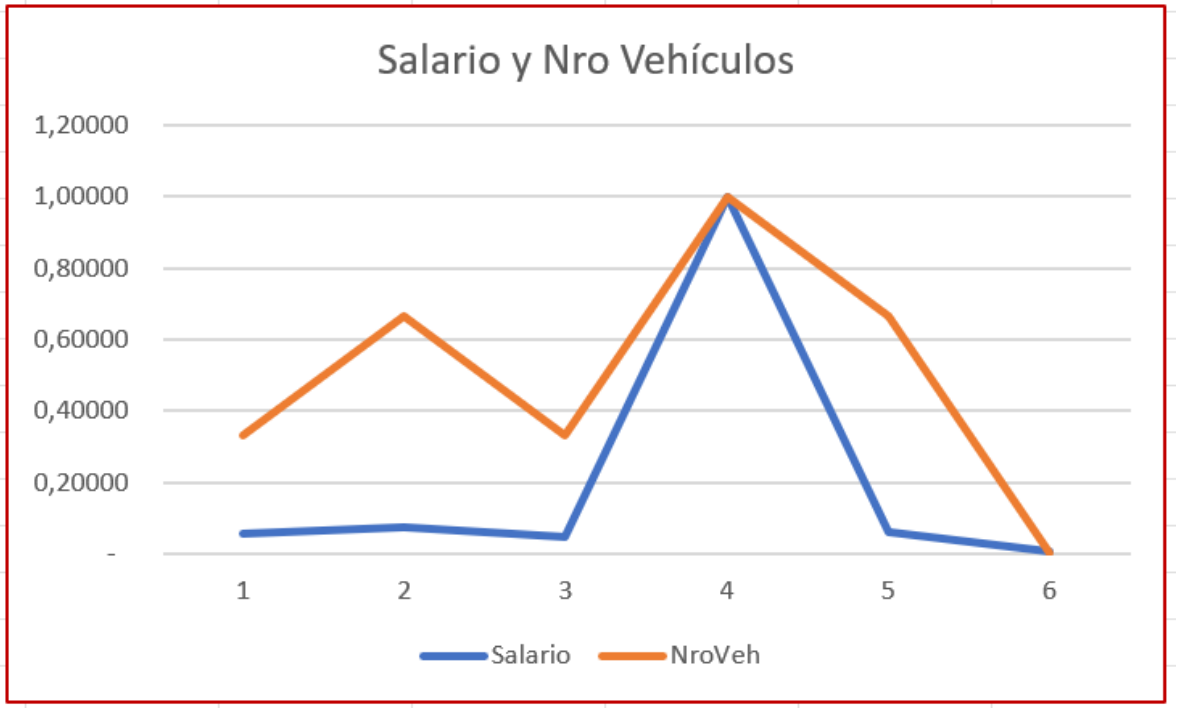
Método escalado simple aplicado al salario y Número vehículos

	Salario	NroVeh
0	0.05625	0.333333
1	0.07500	0.666667
2	0.04875	0.333333
3	1.00000	1.000000
4	0.06250	0.666667
5	0.00500	0.000000

Sin perder su distribución, los rangos para ambas variables se definen entre los valores de 0 y 1

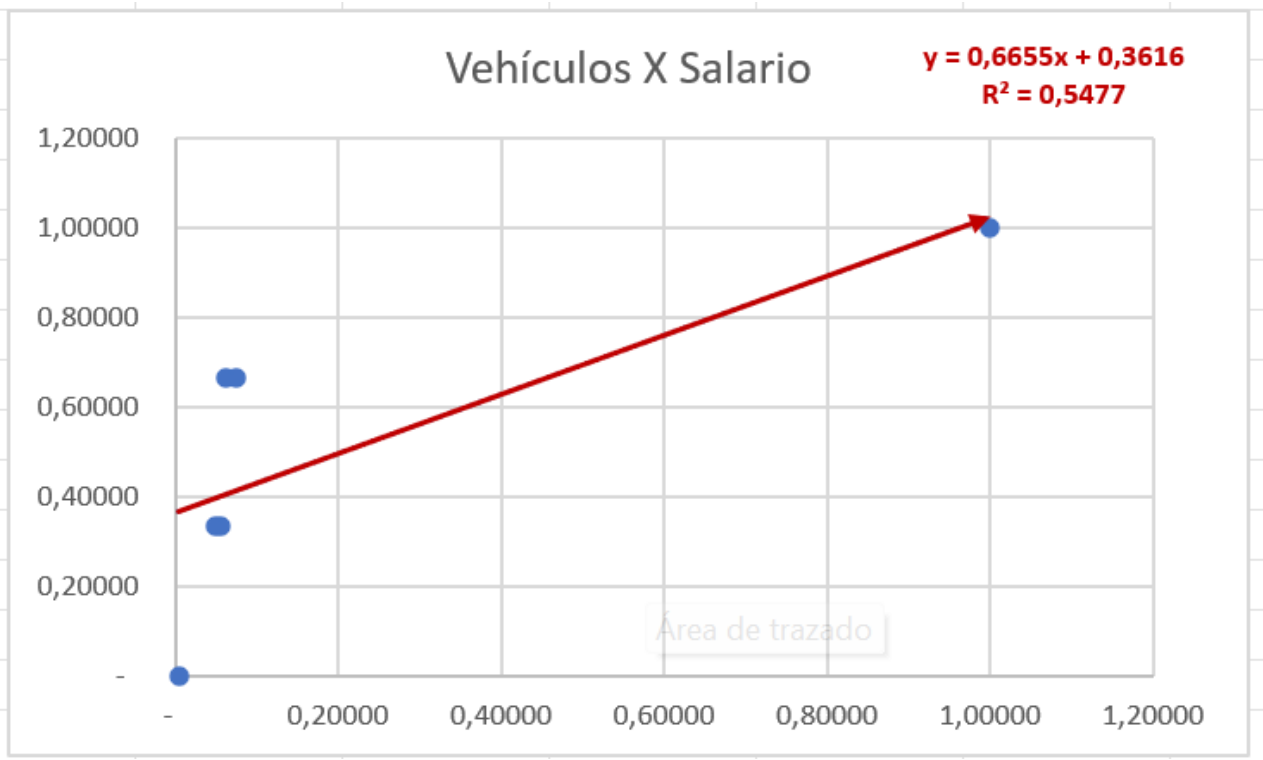
Gráfico de dispersión y de líneas – Método Escalado simple





Con los rangos entre 0 y 1 del Salario y Nro. de vehículos, en sus gráficas se aprecian mejor

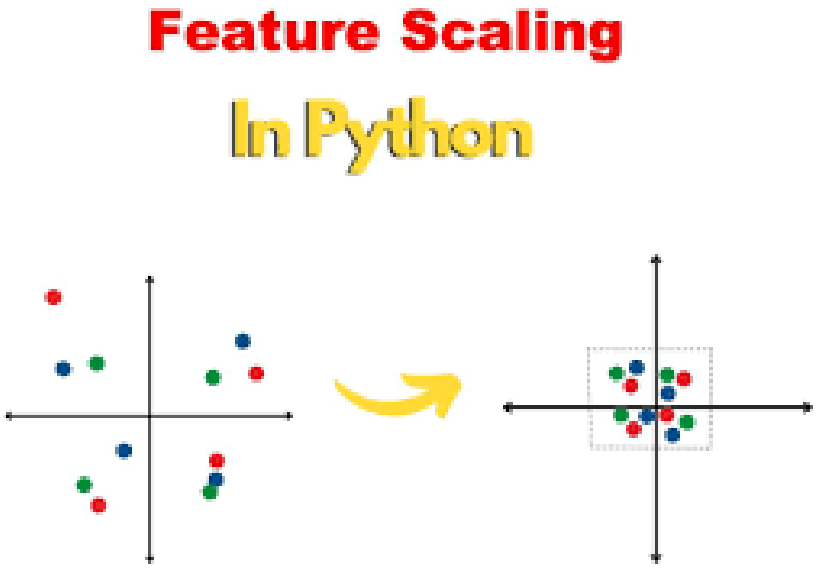
Regresión Lineal – Aplicado el Escalado simple



Con los rangos de las variables entre 0 y 1, la regresión lineal presenta un R^2 idéntico al de la regresión lineal de los datos originales

MÉTODO: MÍNIMO Y MÁXIMO

Método Mínimo y Máximo



Fórmula:

$$N_i = \frac{(X_i - X_{min})}{X_{max} - X_{min}}$$

Conservar la distribución de los datos en una menor escala

Método Mínimo y Máximo – Programa - Resultados

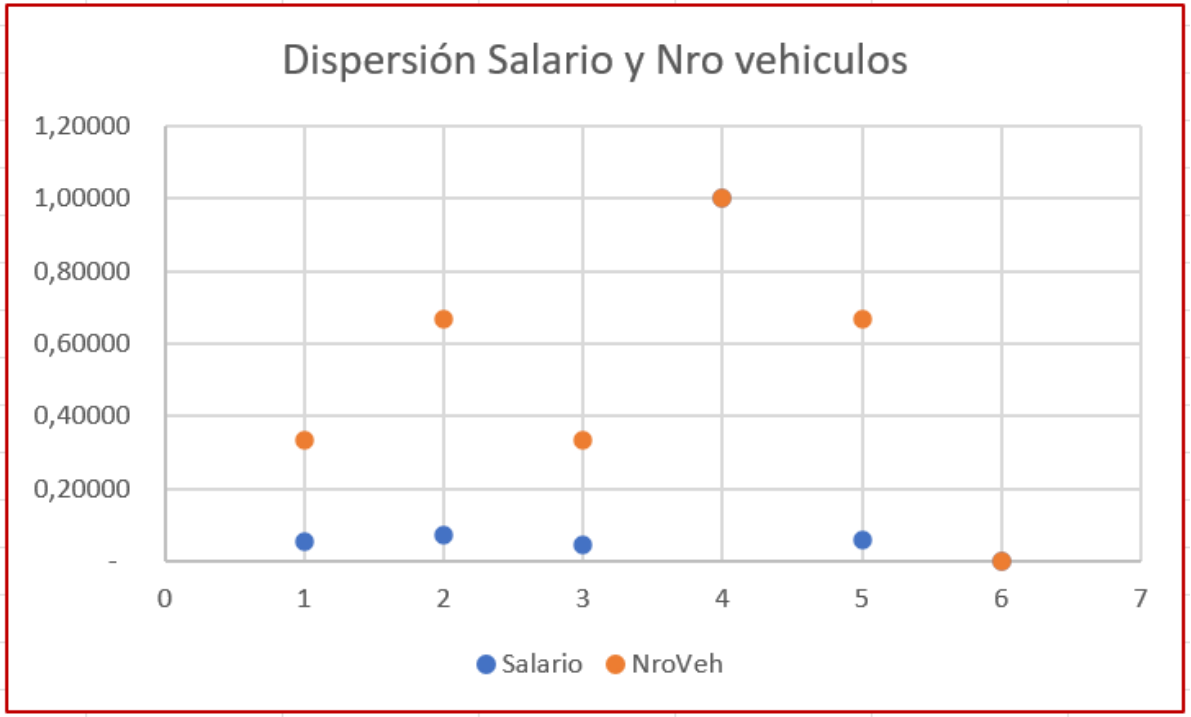
```
1  # Método Mínimo y Máximo
2  print("\n Método Mínimos y Máximos aplicado al salario y Número vehículos")
3  df["Salario"]=(df["Salario"]-df["Salario"].min())/(df["Salario"].max()-df["Salario"].min())
4  df["NroVeh"]=(df["NroVeh"]-df["NroVeh"].min())/(df["NroVeh"].max()-df["NroVeh"].min())
5  print("\n",df[["Salario","NroVeh"]]) # Mostrar llas variables de interés (Numéricas)
```

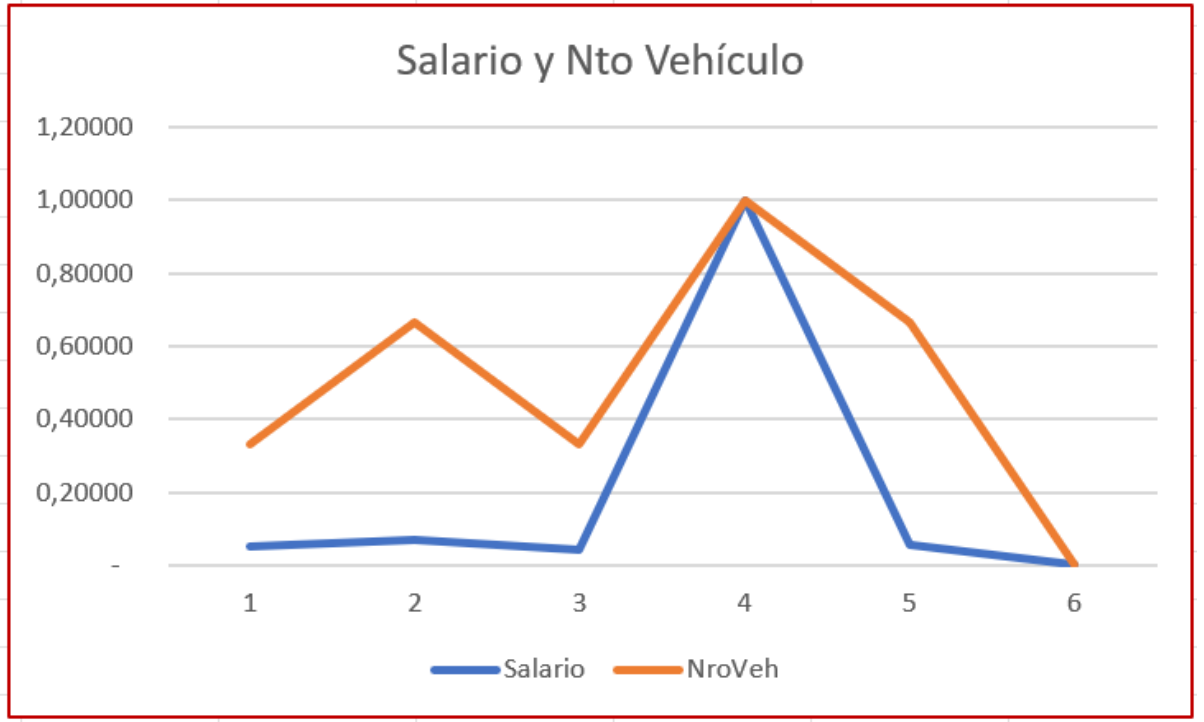
```
 Variables de interés

     Salario  NroVeh
0    4500000       1
1    6000000       2
2    3900000       1
3   80000000       3
4    5000000       2
5     400000       0
```

```
 Método Mínimos y Máximos aplicado al salario y Número vehículos

     Salario    NroVeh
0   0.051508  0.333333
1   0.070352  0.666667
2   0.043970  0.333333
3   1.000000  1.000000
4   0.057789  0.666667
5   0.000000  0.000000
```

Sin perder su distribución, los rangos para ambas variables se definen entre los valores de 0 y 1

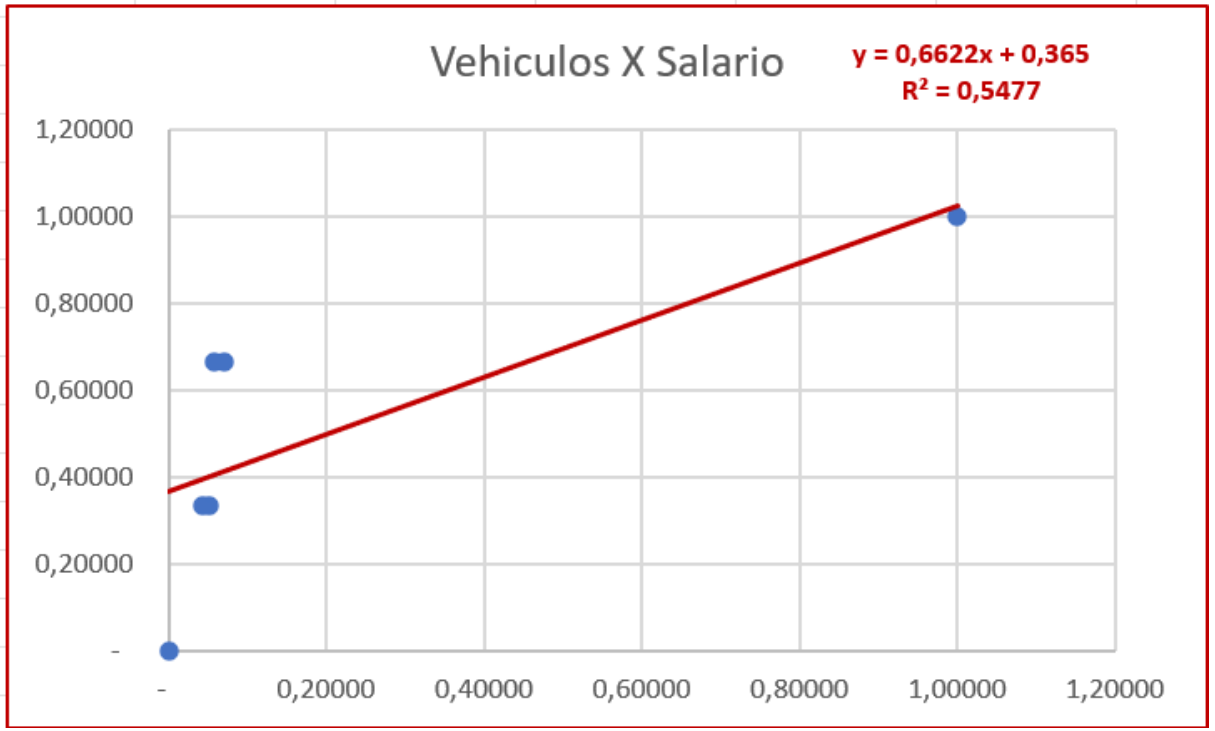
Gráfico de dispersión y de líneas – Método Mínimo y Máximo





Con los rangos entre 0 y 1 del Salario y Nro. de vehículos, en sus gráficas se aprecian mejor

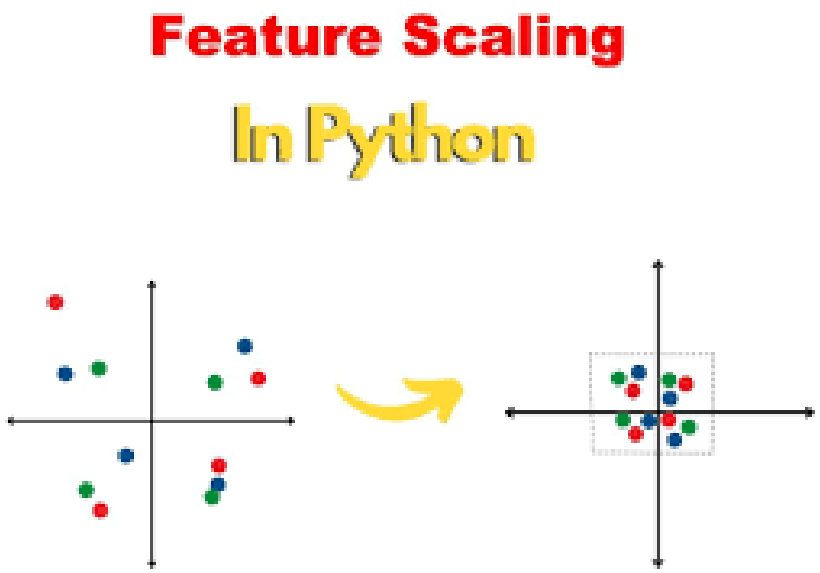
Regresión Lineal – Aplicado Mínimo y Máximo



Con los rangos de las variables entre 0 y 1, la regresión lineal presenta un R^2 idéntico al de la regresión lineal de los datos originales

MÉTODO: Z-SCORE

Método Z-Score



Fórmula:

$$N_i = \frac{(X_i - \mu)}{\sigma}$$

Media

Desviación Estándar

Conservar la distribución de los datos en una menor escala

Método Z-score – Programa - Resultados

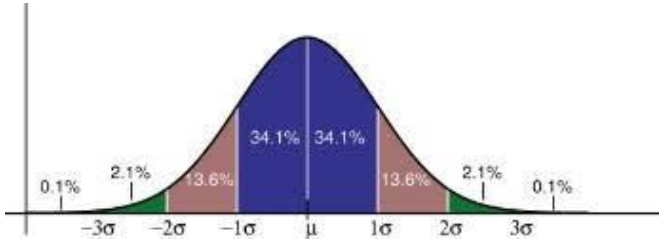
```
# Método z-score
print("\n Método z-score aplicado al salario y Número vehículos")
df["Salario"]=(df["Salario"]-df["Salario"].mean())/df["Salario"].std()
df["NroVeh"]=(df["NroVeh"]-df["NroVeh"].mean())/df["NroVeh"].std()
print("\n",df[["Salario","NroVeh"]]) # Mostrar llas variables de interés (Numéricas)
```

Variables de interés

	Salario	NroVeh
0	4500000	1
1	6000000	2
2	3900000	1
3	80000000	3
4	5000000	2
5	400000	0

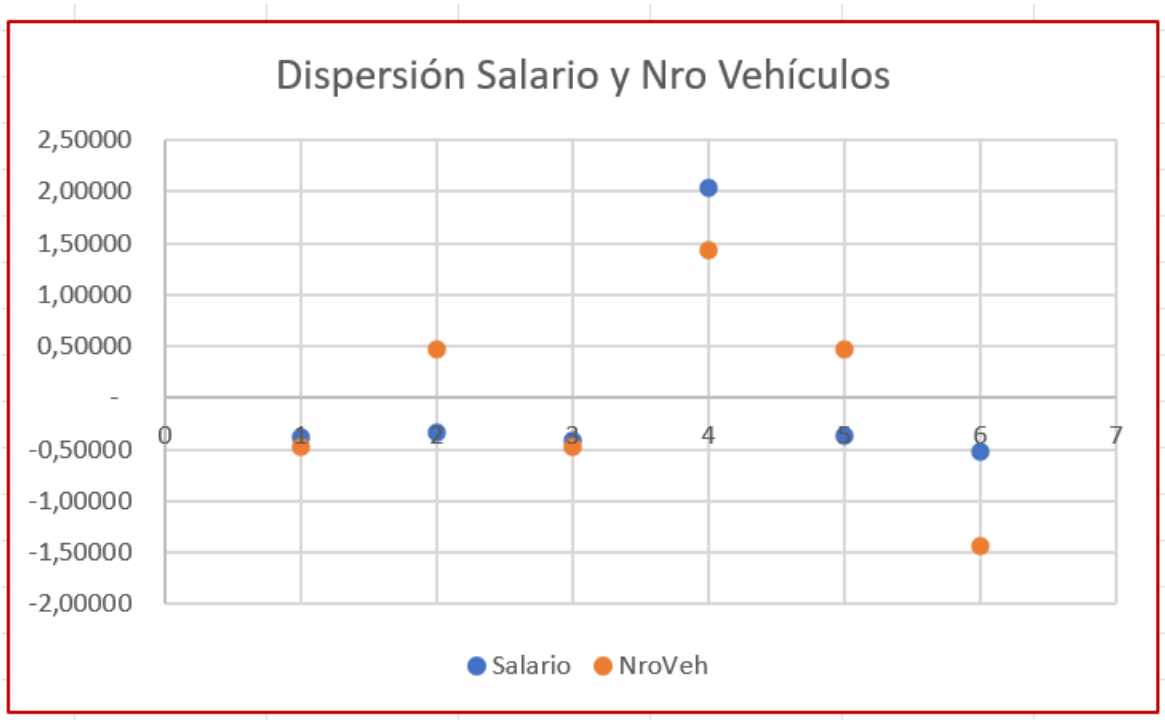
Método z-score aplicado al salario y Número vehículos

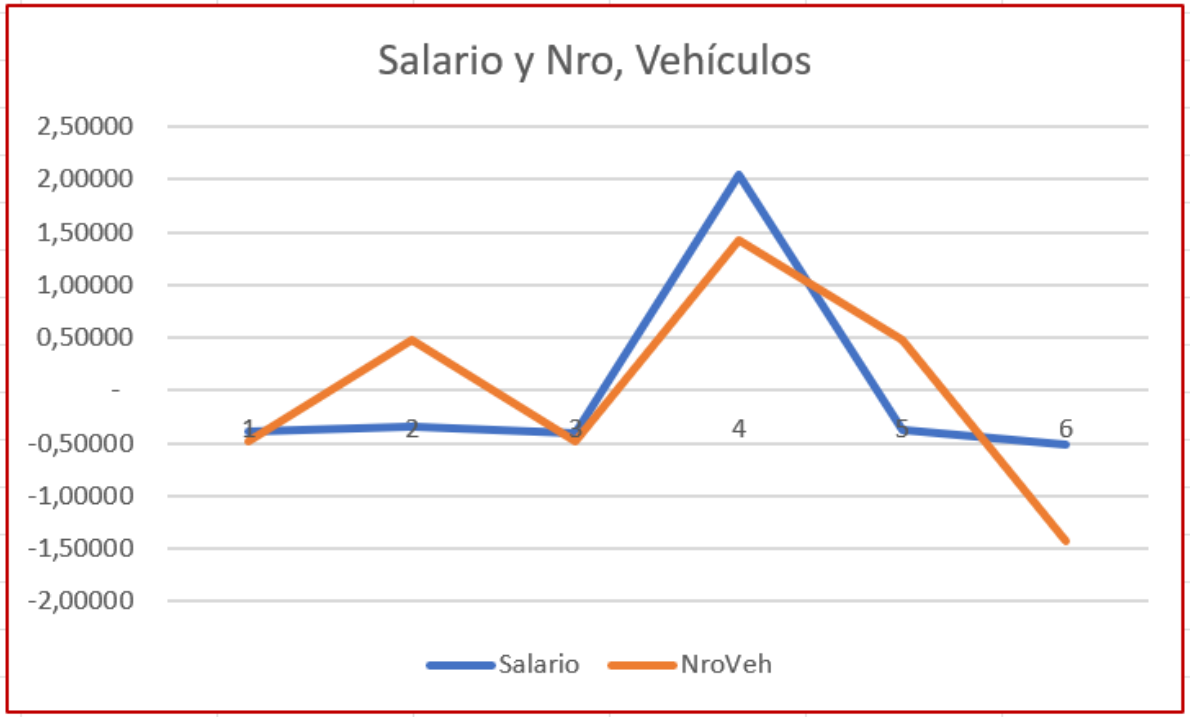
	Salario	NroVeh
0	-0.390117	-0.476731
1	-0.341888	0.476731
2	-0.409408	-0.476731
3	2.037395	1.430194
4	-0.374040	0.476731
5	-0.521942	-1.430194



Sin perder su distribución, los rangos para ambas variables son más uniformes, en una distribicon normal de -3-a 3

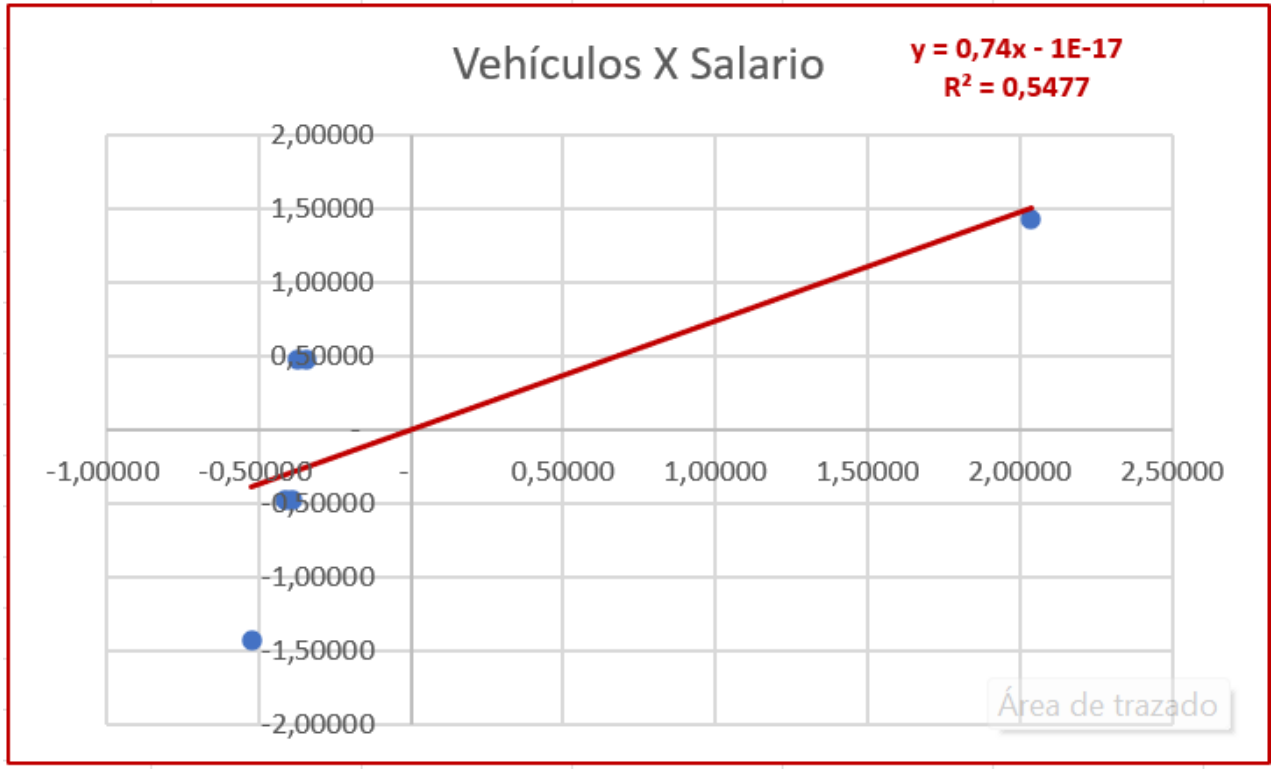
Gráfico de dispersión y de líneas – Método z-score





Con los rangos entre 0 y 1 del Salario y Nro. de vehículos, en sus gráficas se aprecian mejor

Regresión Lineal – Aplicado z-score



Con los rangos de las variables entre 0 y 1, la regresión lineal presenta un R^2 idéntico al de la regresión lineal de los datos originales

TALENTO TECH

TIC

Universidad Tecnológica de Bolívar

www.utb.edu.co/talento-tech