

Módulo: ADM1929 - Business understanding: Pensamiento analítico basado en datos. - (A51)

**Actividad: Reto de aprendizaje 15.
Entendimiento de la data**

Nombre: Roberto Mora Balderas

Asesor: José Carlos Soto Monterrubio

Fecha: 24 de julio de 2023

Objetivo:

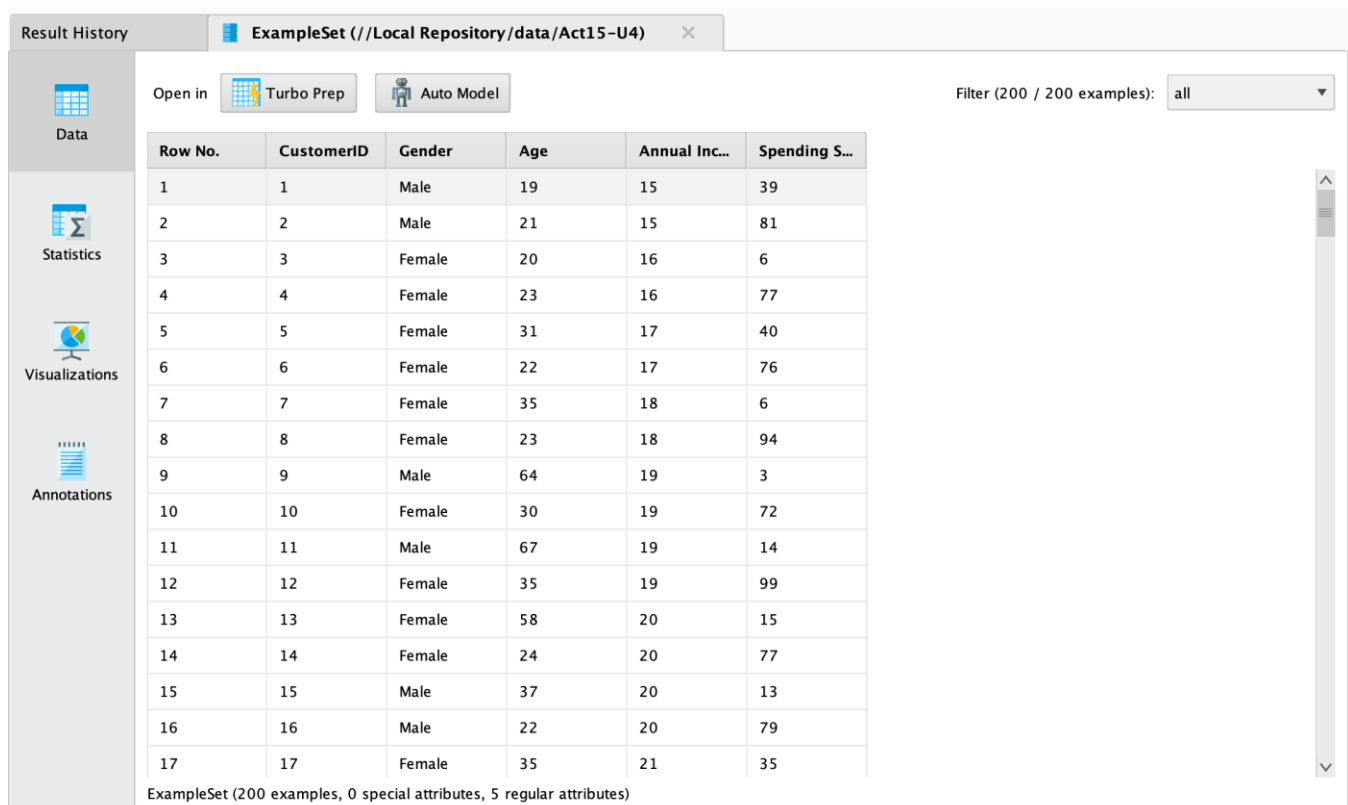
Comprender las características de la data con la que trabajará el caso de aplicación con la finalidad de definir el programa de operación de minería de datos.

Instrucciones:

1. Importa la [data proporcionada](#) a Rapid Miner, utiliza la pestaña de estadística para identificar las variables, sus valores estadísticos generales, sus niveles de correlación. En un documento de word, desarrolla los siguientes elementos:

- Número de registros de la base de datos y número de variables.
- Nombre y tipo de variables con explicación del significado de cada variable.
- Valores estadísticos descriptivos para las variables que aplique.
- Matriz de correlación con las variables vigentes.
- Identificación de variables de baja calidad con Auto Model.
- Explicación narrativa en un párrafo de la base de datos (Comenta con tus palabras un resumen de qué aspectos son los más importantes que captas en el análisis de la *example set*)

Desarrollo:

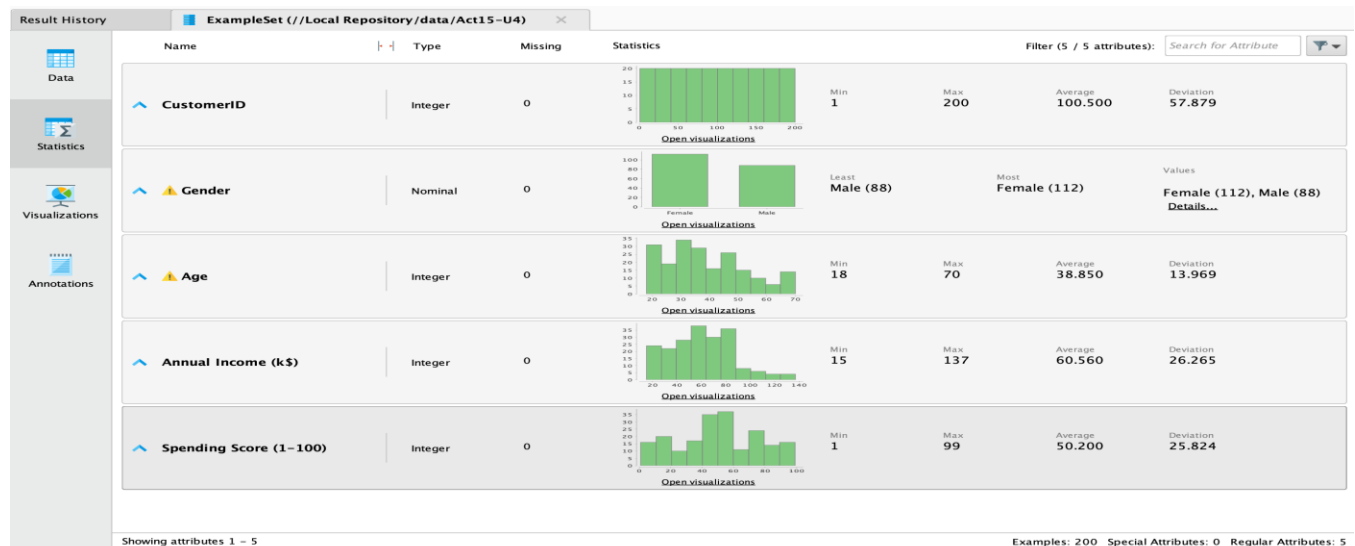


Row No.	CustomerID	Gender	Age	Annual Inc...	Spending S...
1	1	Male	19	15	39
2	2	Male	21	15	81
3	3	Female	20	16	6
4	4	Female	23	16	77
5	5	Female	31	17	40
6	6	Female	22	17	76
7	7	Female	35	18	6
8	8	Female	23	18	94
9	9	Male	64	19	3
10	10	Female	30	19	72
11	11	Male	67	19	14
12	12	Female	35	19	99
13	13	Female	58	20	15
14	14	Female	24	20	77
15	15	Male	37	20	13
16	16	Male	22	20	79
17	17	Female	35	21	35

ExampleSet (200 examples, 0 special attributes, 5 regular attributes)

Se puede apreciar, que existen 200 registros y 5 variables, CustomerID, Gender, Age, Annual Income, Spending Score.

Aquí podemos observar algunas estadísticas descriptivas de cada una de las variables.



A continuación una breve descripción del significado de cada variable.

- CustomerID: Identificación única del cliente.
- Gender: Género del cliente (por ejemplo, Masculino o Femenino).
- Age: Edad del cliente.
- Annual Income (k\$): Ingresos anuales del cliente en miles de dólares.
- Spending Score (1-100): Puntuación de gasto del cliente en una escala del 1 al 100, que refleja el comportamiento de compra.

A continuación se muestran las variables utilizadas por el predict del automodel.

Load Data Select Task Prepare Target Select Inputs Model Types Results

« RESTART < BACK > NEXT

Selected: 4 / Total: 4

✓ Select All ✗ Deselect All

Selected	Status ↑	Quality	Name	Correlation	ID-ness	Stability	Missing	Text-ness
<input checked="" type="checkbox"/>	●	<div><div></div><div></div><div></div><div></div><div></div></div>	Gender	?	1.00%	56.00%	0.00%	2.61%
<input checked="" type="checkbox"/>	●	<div><div></div><div></div><div></div><div></div><div></div></div>	Age	?	?	5.50%	0.00%	0.00%
<input checked="" type="checkbox"/>	●	<div><div></div><div></div><div></div><div></div><div></div></div>	Annual Income (k\$)	?	?	6.00%	0.00%	0.00%
<input checked="" type="checkbox"/>	●	<div><div></div><div></div><div></div><div></div><div></div></div>	Spending Score (1-100)	?	?	4.00%	0.00%	0.00%

Se muestran los resultados del agrupamiento.

k-Means - Summary

Number of Clusters: 3

Cluster 0 97

Age is on average **54.91%** larger, **Spending Score (1-100)** is on average **33.74%** smaller, **Annual Income (k\$)** is on average **0.01%** smaller

Cluster 1 63

Age is on average **65.74%** smaller, **Annual Income (k\$)** is on average **37.95%** smaller, **Spending Score (1-100)** is on average **12.82%** larger

Cluster 2 40

Spending Score (1-100) is on average **61.64%** larger, **Annual Income (k\$)** is on average **59.79%** larger, **Age** is on average **29.62%** smaller

Podemos observar como el gender es aquella variable que no puede ser significativa, además se muestran los resultados de correlaciones.

Correlations

Attribut...	Age	Annual I...	Gender ...	Spendin...
Age	1	-0.012	-0.061	-0.327
Annual I...	-0.012	1	-0.056	0.010
Gender ...	-0.061	-0.056	1	0.058
Spendin...	-0.327	0.010	0.058	1

Conclusiones

El análisis nos arroja información de la relación que tienen las variables, podemos observar como todas las variables se relacionan. El data set nos da para trabajar mucho ya sea en métodos predictivos o de agrupamiento, los cuáles pueden ser explotados por Rapid Miner.

Referencias

- Anáhuac Online. (2019). *Aplicando el caso y data understanding* [Archivo de video]. [Contenido creado para Anáhuac Online].