Realizado por:
Juan Camilo Restrepo Velez
William Leonardo Andrade Collazos
Wilder Valencia Ocampo

PRÁCTICA DE CALIDAD DE DATOS 10%

Bank Marketing

Los datos están relacionados con campañas de marketing directo de una institución bancaria portuguesa. Las campañas de marketing se basaron en llamadas telefónicas. A menudo, se requería más de un contacto con el mismo cliente, para poder acceder a si el producto (depósito bancario a plazo) sería ('sí') o no ('no') suscrito.

Información de atributos

Información Bancaria de los clientes

Age - Edad

Job - Trabajo: tipo de trabajo

Marital - Estado civil: estado civil

Education - Educación: Nivel educativo

Default - Incumplimiento: ¿tiene el crédito en mora?

Housing - Vivienda: ¿tiene un préstamo de vivienda?

Loan - Préstamo: ¿tiene préstamo personal?

Relacionado con la última llamada de la actual campaña

Contact - Contacto: tipo de comunicación

Month - Mes: último mes de contacto del año

DayofWeek - Día de la semana: último día de contacto de la semana

Duration - Duración: duración del último contacto, en segundos (numérico). Nota importante: este atributo afecta en gran medida al objetivo de salida (por ejemplo, si la duración = 0, entonces y = "no"). Sin embargo, no se conoce la duración antes de una llamada se realiza. Además, después del final de la llamada se conoce obviamente y. Por

lo tanto, esta entrada sólo debe incluirse a efectos de referencia y debe descartarse si se pretende tener un modelo predictivo realista.

Otros

Campaign - Campaña: número de contactos realizados durante esta campaña y para este cliente

Pdays - pDías: número de días que pasaron después de que el cliente fue contactado por última vez en una campaña anterior. Nota, 999 significa que el cliente no fue contactado anteriormente

Previous - Anterior: número de contactos realizados antes de esta campaña y para este cliente

Poutcome: resultado de la anterior campaña de marketing

Atributos del contexto social y económico

Emp.var.rate - Tasa de variación del empleo - indicador trimestral

Cons.price.idx: Índice de Precios al Consumidor - Indicador mensual; el Índice de Precios al Consumidor o IPC mide los cambios en los precios pagados por los consumidores por una cesta de bienes y servicios cada mes.

Cons.conf.idx: Índice de confianza del consumidor - Indicador mensual; En Portugal, el índice de confianza del consumidor se basa en entrevistas con los consumidores sobre sus percepciones de la situación económica actual y futura del país y sus tendencias de compra. Se estima utilizando la diferencia entre la proporción de respuestas de evaluación positivas y las respuestas de evaluación negativas, pero no incluye la proporción de respuestas neutras

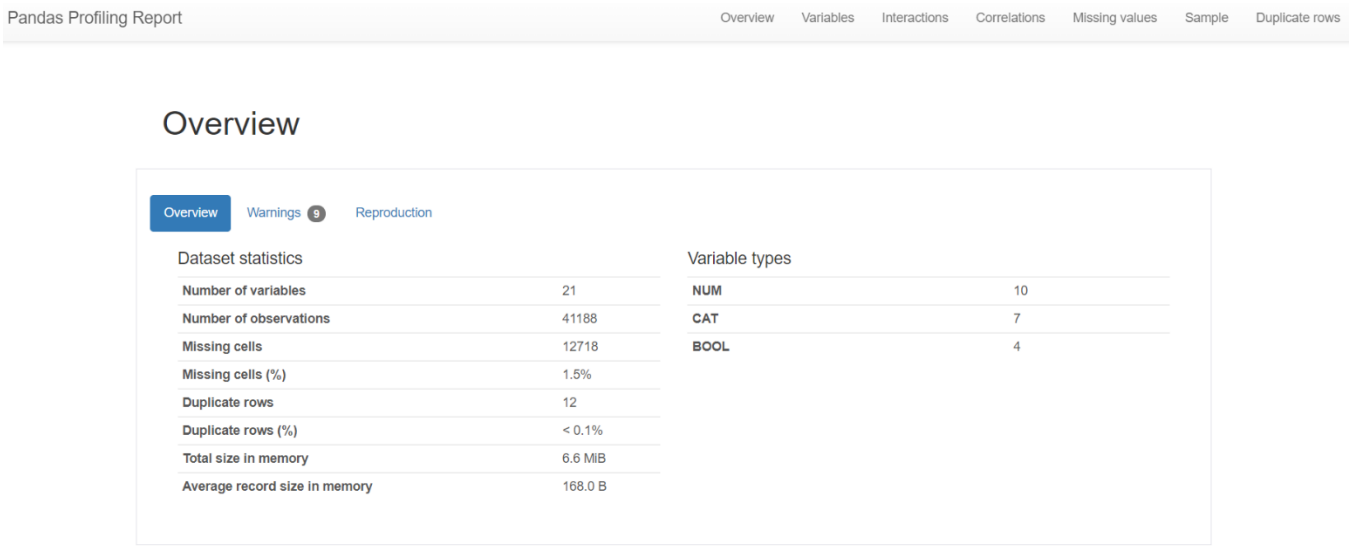
Euribor3m: euribor 3 meses - Euribor es la abreviatura de Euro Interbank Offered Rate. es un índice de referencia publicado diariamente que indica el tipo de interés promedio al que un gran número de bancos europeos dicen concederse préstamos a corto plazo entre ellos para prestárselo a terceros.

Nr.employed - Número de empleados: Número de empleados - Indicador trimestral; Número de personas empleadas para el trimestre.

y - ¿el cliente ha suscrito un depósito a plazo? (Variable objetivo)

*Tomado de https://www.kaggle.com/henriqueyamahata/bank-marketing

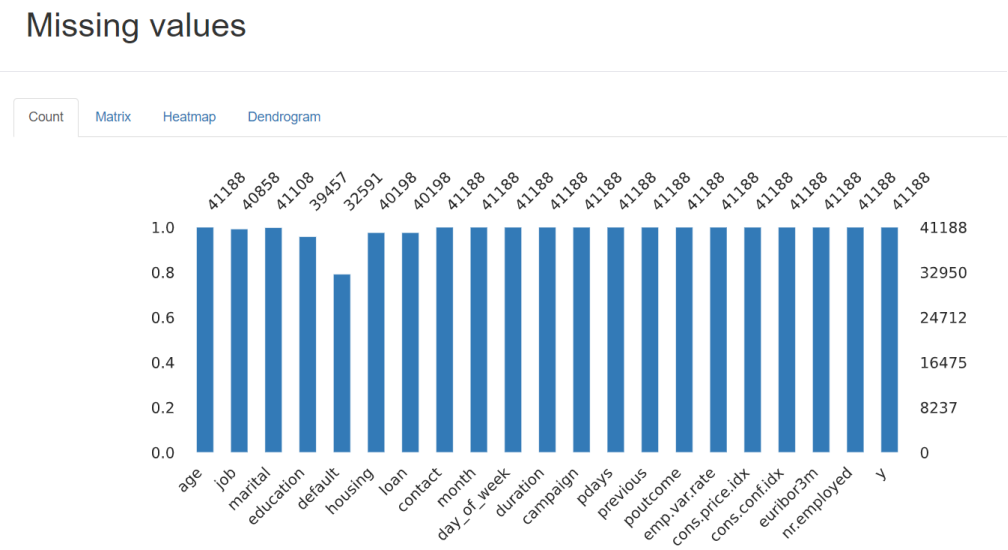
1. Genere el reporte de PYTHON sobre los datos **originales**, adjunte el resultado en HTML.



2. Evalúe cada una de las dimensiones de la calidad de datos, teniendo en cuenta los resultados de PYTHON.

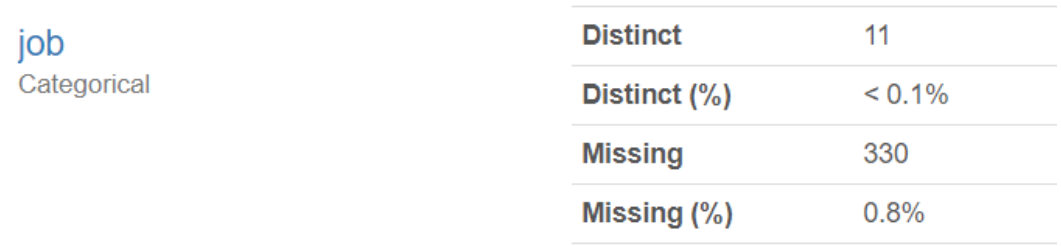
DIMENSIONES DE LA CALIDAD DE DATOS

1. **Completitud:** ¿Está toda la información disponible? ¿Hay datos faltantes o ausentes?

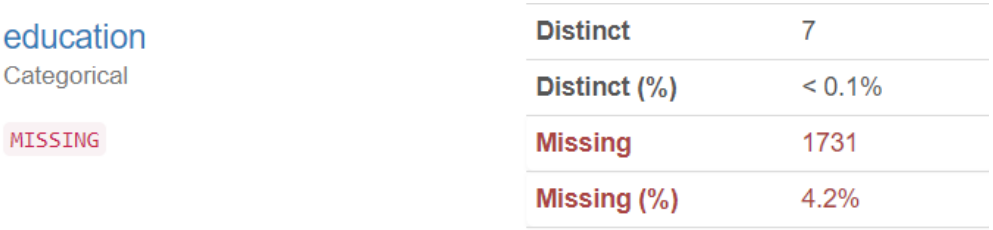


La información está disponible y contiene algunos nulos que son los siguientes:

Job: los datos faltantes 330



Education: los datos faltantes 1731



Default: los datos faltantes son 8597

default
Boolean

MISSING

Distinct	2
Distinct (%)	< 0.1%
Missing	8597
Missing (%)	20.9%

Housing: los datos faltantes son 690

housing
Boolean

MISSING

Distinct	2
Distinct (%)	< 0.1%
Missing	990
Missing (%)	2.4%

Loan: los datos faltantes son 690

loan
Boolean

MISSING

Distinct	2
Distinct (%)	< 0.1%
Missing	990
Missing (%)	2.4%

2. **Exactitud:** ¿La información es correcta y libre de error?

   Efectivamente la información es correcta y libre de error debido a que son datos reales de un banco en Portugal.

3. **Conformidad:** ¿Los valores de los datos están conformes con los formatos esperados? Ejemplo: Una fecha en formato AAAA/MM/DD cuando debería ser DD/MM/AAAA.

   Si están conformes debido a que nuestros datos no cuentan con formatos que se indican y las categorías de la variable categóricas no están duplicados.

4. **Oportunidad:** ¿La información llega cuando se necesita?

   La información llega en los tiempos establecidos debido a la importancia y que se obtienen directamente de la entidad bancaria

5. **Duplicidad:** ¿Existen múltiples instancias, innecesarias de los mismos objetos de datos en el conjunto de datos?

El método de Profile Report, reconoce unas ciertas filas como las más frecuentes por lo que se puede llegar a considerar que estas filas efectivamente son las misma, sin embargo, estas no son necesariamente instancias duplicadas porque es una base de datos amplia donde pueden a ver dichas personas que cuenten con estas similitudes.

Warnings

Dataset has 12 (< 0.1%) duplicate rows — Duplicates

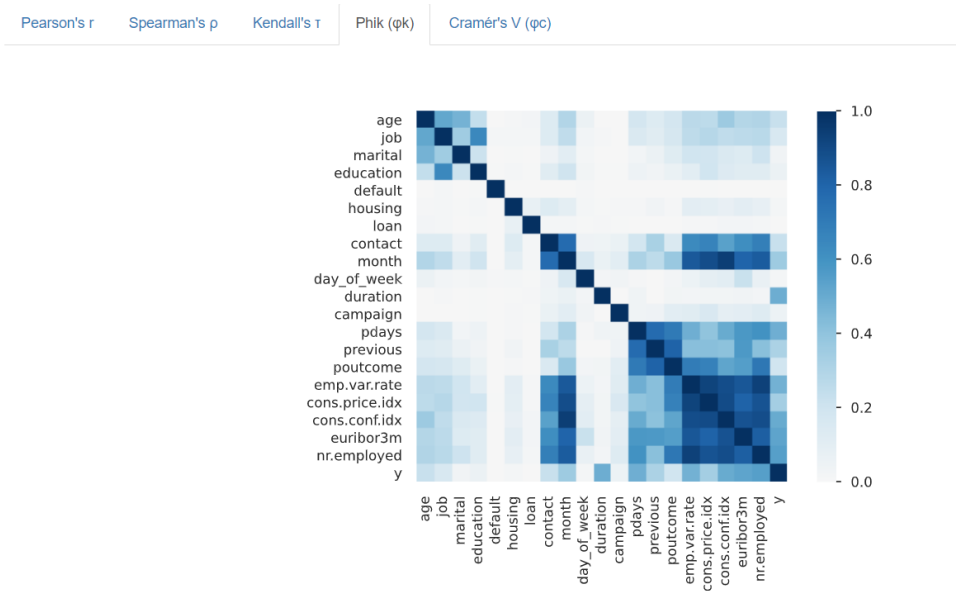
Duplicate rows

Most frequent

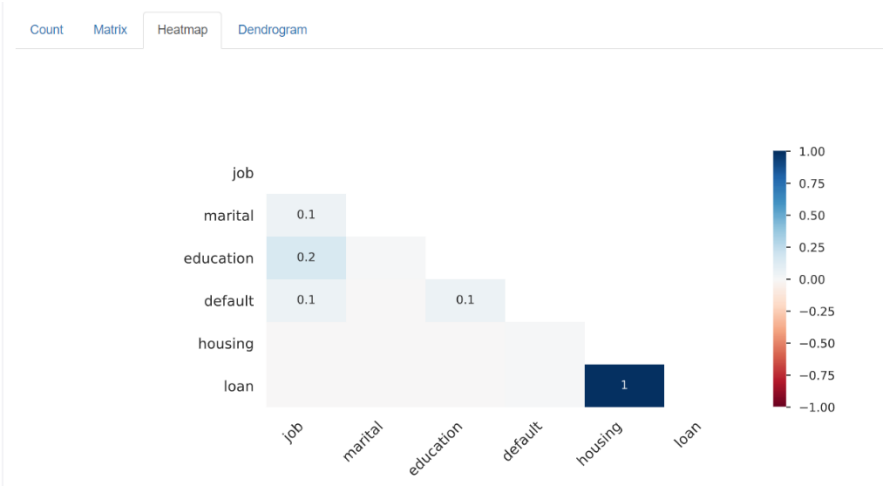
	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutco
0	24	services	single	high.school	no	yes	no	cellular	apr	tue	114	1	999	0	nonexis
1	27	technician	single	professional.course	no	no	no	cellular	jul	mon	331	2	999	0	nonexis
2	32	technician	single	professional.course	no	yes	no	cellular	jul	thu	128	1	999	0	nonexis
3	35	admin.	married	university.degree	no	yes	no	cellular	may	fri	348	4	999	0	nonexis
4	39	admin.	married	university.degree	no	no	no	cellular	nov	tue	123	2	999	0	nonexis
5	39	blue-collar	married	basic.6y	no	no	no	telephone	may	thu	124	1	999	0	nonexis
6	41	technician	married	professional.course	no	yes	no	cellular	aug	tue	127	1	999	0	nonexis
7	45	admin.	married	university.degree	no	no	no	cellular	jul	thu	252	1	999	0	nonexis
8	47	technician	divorced	high.school	no	yes	no	cellular	jul	thu	43	3	999	0	nonexis
9	71	retired	single	university.degree	no	no	no	telephone	oct	tue	120	1	999	0	nonexis

6. **Integridad:** ¿Faltan datos relacionados importantes? ¿Es clara la conectividad y las relaciones con otros datos?

No faltan datos importantes debido a que se evidencia que se procesó un dato tipo “Fecha” creando dos nuevos atributos “month” y “day_of_week”. También, se pueden evidenciar las correlaciones entre todos los atributos.



Además, se evidencia en la relación de los nulos faltantes que ‘loan’ y ‘housing’ del 100%, lo que indica que dicha información NO está presente en los mismos registros los que se cataloga como falta de datos relacionados importantes.



3. Comparar los problemas de calidad encontrados en el perfilado con la preparación de datos realizada en el proyecto de Minería. Indicar los nuevos errores encontrados en los datos.

A través del análisis se encuentran los siguientes hallazgos:

- Se encuentran los mismos datos faltantes (nulos) en ambos procesos.
- Se puede realizar un nuevo análisis para los datos duplicados por medio del apartado "Duplicate rows".
- Se obtiene una nueva matriz de correlaciones entre los atributos originales y no entre las dummies que se crea en weka.
- Se encuentran, de una forma más rápida, las variables con una alta correlación que son "emp.var.rate", "euribor3m" y "nr.employed"