ITESO, Universidad Jesuita de Guadalajara

Maestría en Ciencia de Datos

Análisis estadístico multivariable

Práctica 1. Covarianza y Correlación

Saúl Enrique Mendoza Ruiz

Mario Alejandro Oviedo Vázquez

7 de febrero del 2023

Introducción:

Crecy es una aplicación que ofrece liquidez a los tenedores (holders) de criptomonedas (DOT, ETH, SOL, NEAR) para lo cual se hace una conexión entre alguna Cartera digital y alguna cuenta CLABE, eligiendo como colateral (token digital) alguna de las criptomonedas aceptadas.

En este documento realizaremos un análisis exploratorio utilizando estadística descriptiva en una encuesta realizada a usuarios que poseen tokens del protocolo "NEAR".

Objetivo:

El objetivo de este documento es utilizar herramientas de estadística descriptiva, sobre los datos provenientes de una encuesta de clientes; esto nos ayuda a resumir, organizar y describir las características principales de los datos, incluyendo tendencias centrales, dispersión y (posibles) relaciones entre variables.

Es conveniente mencionar que hacer inferencias sobre las distribuciones probabilísticas de los datos se encuentra fuera del alcance de este documento; si el lector lo desea puede revisar los temas $distribuciones\ continuas\ de\ probabilidad$[1] y $bondad\ de\ ajuste$[2].

Descripción de datos:

Se realizaron las siguientes preguntas:

- ¿Te gustaría contar con la opción de generar rendimiento en tus activos digitales sin la necesidad de solicitar una línea de crédito con tarjeta Crecy?
- ¿Qué valor de *NEAR en MXN *planeas poner como garantía?
- Prioridad #1
- Prioridad #2
- Prioridad #3
- ¿A cuántos meses te gustaría financiar la mayoría de las compras que hagas con tu tarjeta Crecy?

Para las prioridades se asignaron tres diferentes opciones de respuesta:

- Quiero optimizar mis impuestos.
- Me da oportunidad de acceder a liquidez sin vender mis activos digitales.
- Me gusta que puedo generar rendimientos en mis criptos.

El tamaño de la muestra fue de 126 encuestados.

ESTADÍSTICA DESCRIPTIVA

	Wish yield without CL	NEAR in MXN	Loan lenght
count	126.000000	126.000000	126.000000
mean	0.880952	13349.841270	6.579365
std	0.325137	25911.031249	3.876809
min	0.000000	0.000000	1.000000
25%	1.000000	1000.000000	3.000000
50%	1.000000	3000.000000	6.000000
75%	1.000000	10000.000000	12.000000
max	1.000000	150000.000000	12.000000

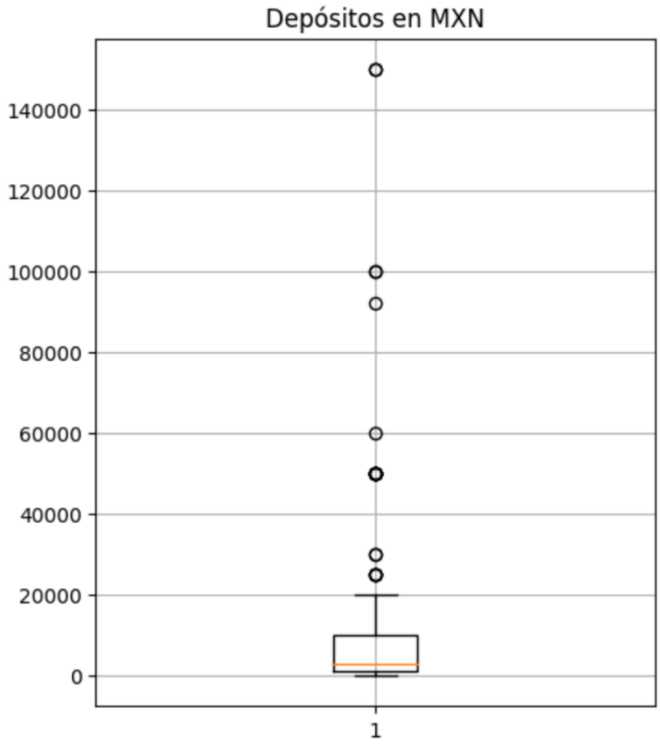
Después de limpiar los datos, se pueden observar las siguientes medidas de tendencia central:

Count: Conteo es una variable cuantitativa que, aunque no es una medida de tendencia central es necesaria para su cálculo, en esta caso es la cantidad de personas que respondieron la encuesta.

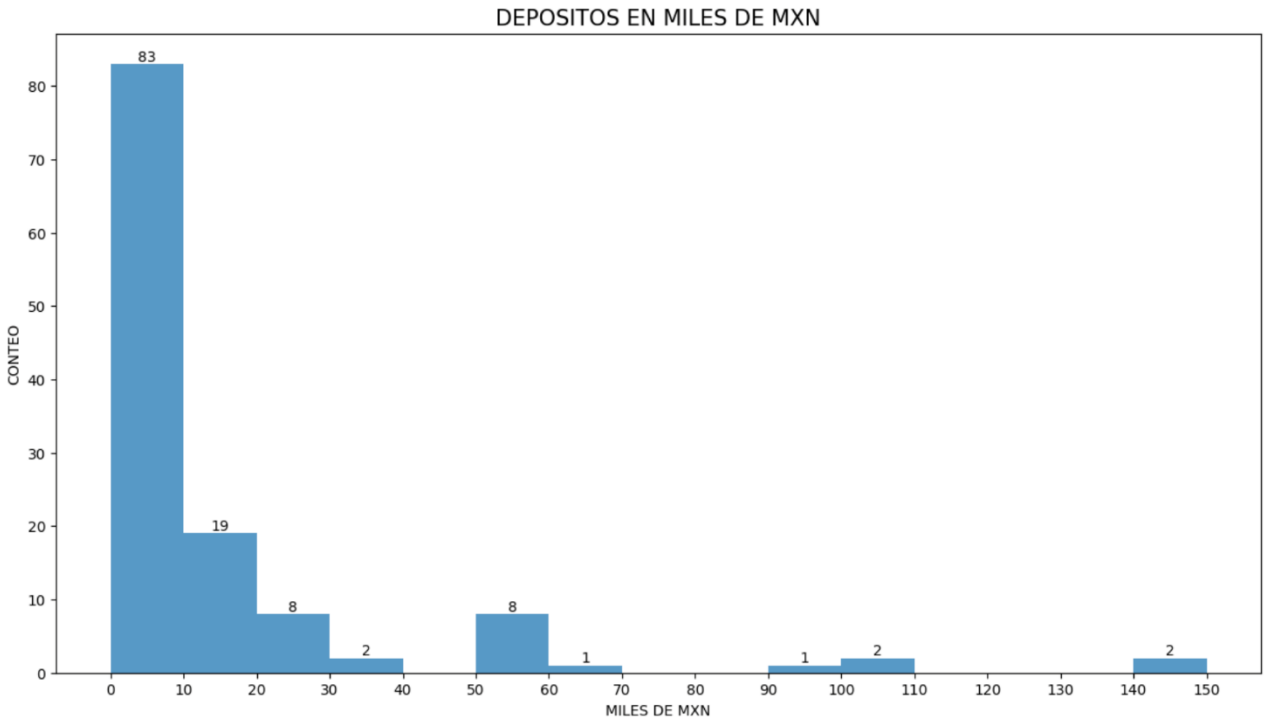
Mean: La media si es una medida de tendencial central que entre sus características se encuentra que es menos susceptible a las variaciones del muestreo y que es influenciada por valores atípicos. Podemos observar que el valor promedio que los encuestados están dispuestos a poner como garantía en pesos es de $13,349 MXN y el tiempo promedio de préstamo es de 6.57 meses.

STD: La desviación estándar forma parte de las medidas de dispersión, siempre tiene un valor de cero o mayor, también es afectada por los valores atípicos de la muestra y se expresa con las mismas unidades que la muestra. En este caso la desviación estándar en cuanto a monto de colateral es de $25,911 MXN y en plazo es de 3.87 meses.

GRÁFICOS DE CAJA E HISTOGRAMAS

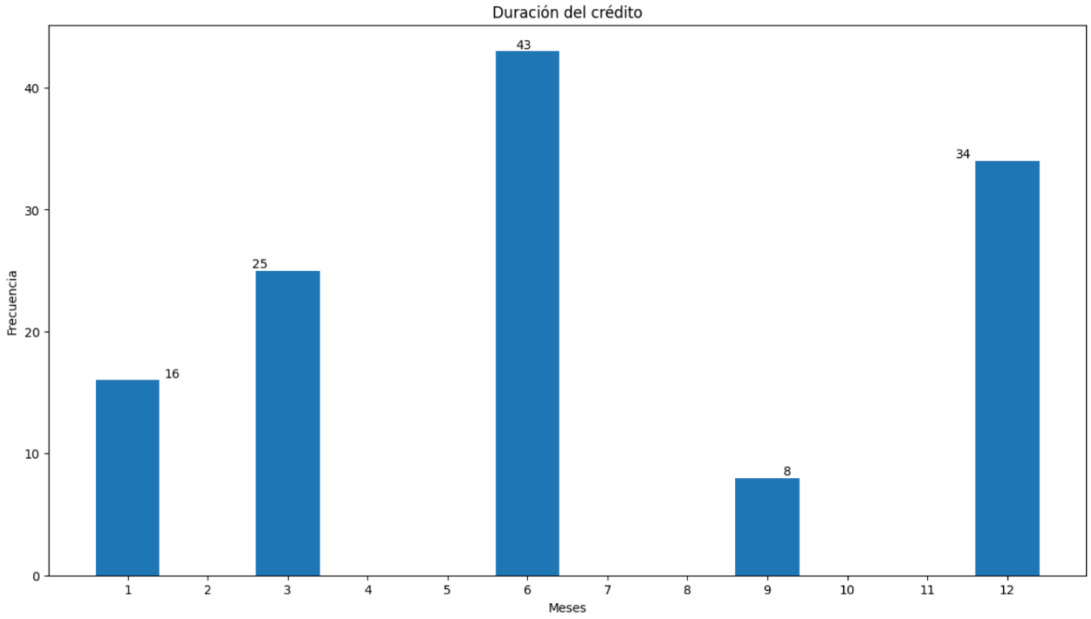


Realizamos una grafico de caja para poder visualizar como se distribuye el rango de depósitos, la base de la caja es el Cuartil 1 ($1,000 MXN), la línea anaranjada el Cuartil 2 (Mediana) y el techo de la caja el Cuartil 3 ($1,000 MXN), siendo el rango Inter cuartil de $9,000 MXN, la mediana (Q2) se encuentra muy cercano a la parte inferior del rango usando como referencia la media. Las pestañas en la parte inferior y superior de la caja representar el límite inferior y límite superior del rango y los puntos representan los datos atípicos. Para considerar un dato atípico es necesario que a una distancia 1.5 veces mayor al cuartil más próximo sin embargo los datos atípicos no necesariamente representan errores de medición o información falta también se puede explicar por una distribución de cola pesada, el análisis y discusión de los datos atípicos se encuentra fuera del análisis y objetivo de este proyecto.



En este histograma se puede visualizar de forma más clara como la mayoría de las respuestas se concentran en el rango inferior de los montos, esto se ve reflejado en la grafica de caja con la mediana (Q2) muy cercana al fondo de la caja (Q1).

HISTOGRAMA DE DURACION DE CRÉDITOS



En cuanto a la distribución de los plazos se puede considerar un poco más uniforme respecto a la distribución de los montos de la grafica anterior, se observa como los encuestados refirieron el plazo de 6 meses como el de mayor interés y el de 9 como el menos atractivo dentro de los encuestados.

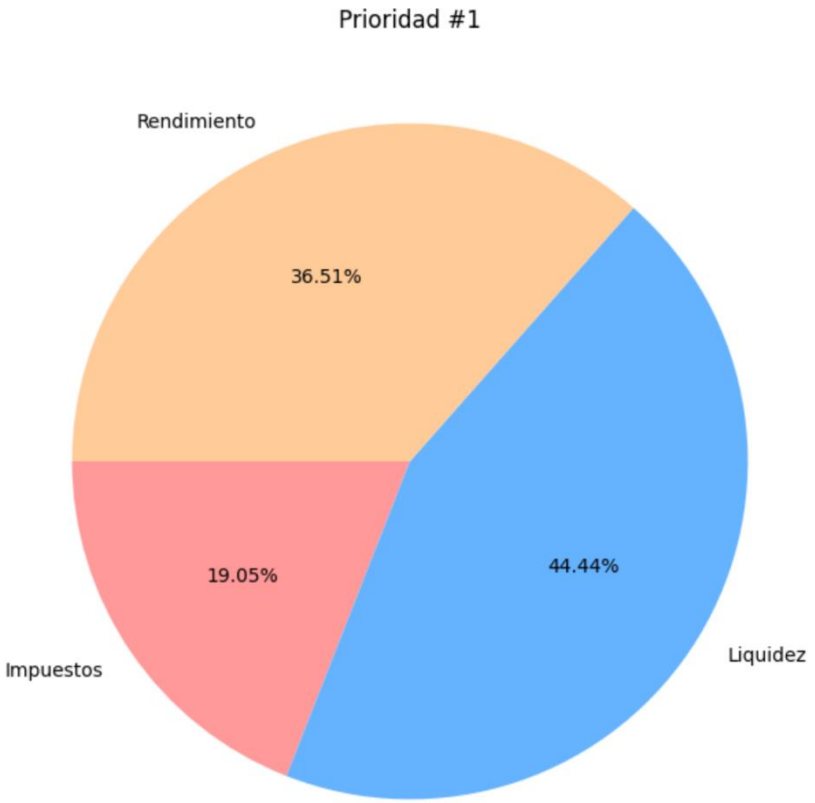
GRÁFICA DE PASTEL:

Análisis de Prioridad #1

Las gráficas $tipo\ pastel^3$ se utilizan para representar la proporción o la fracción de un total. Se utilizan principalmente para representar datos categóricos (y nominales para nuestro caso) y se usan para visualizar la distribución de los datos entre diferentes categorías o grupos.

Por la razón anterior se decidió utilizar para representar la 'PRIORIDAD 1'; abajo se muestran las proporciones de cada categoría nominal escogida.

Podemos observar como la liquidez tiene una proporción mas alta respecto al total de respuestas (44%)



Agrupación de clientes por plazo de crédito

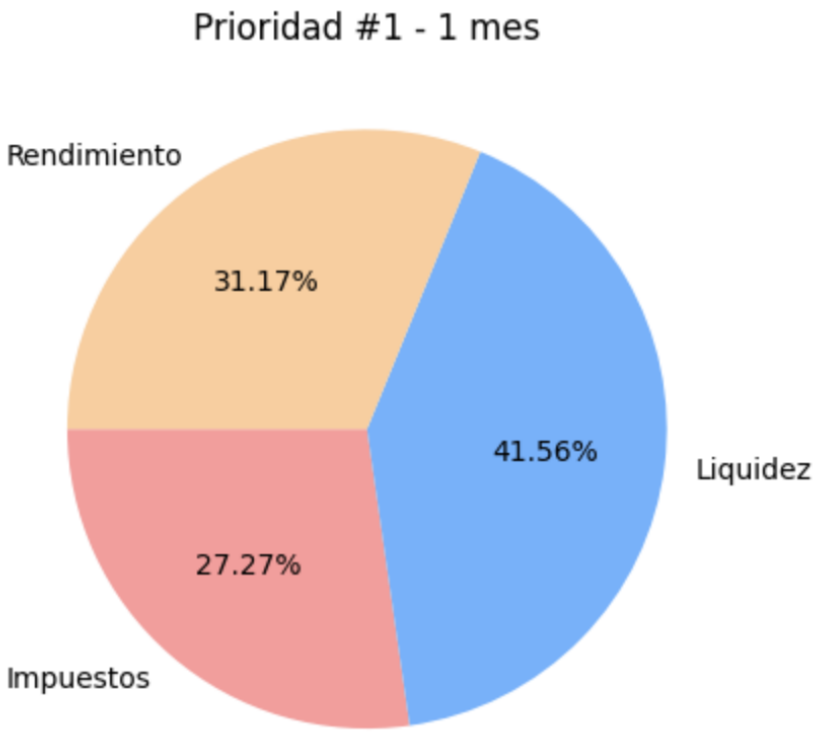
Como primer ejercicio clasificamos las respuestas de las encuestas que desean obtener un rendimiento si la necesidad de solicitar una línea de crédito con base en el plazo de los préstamos que los encuestados pretender solicitar, empezando 1 un mes y siendo el máximo plazo posible 12 meses. La mayoría de los encuestados desean préstamos a 6 meses (34%) seguido de préstamos a 12 meses (27%) y siendo 9 meses el plazo con menos interesados (6%).

	NEAR in MXN							
	count	mean	std	min	25%	50%	75%	max
Loan lenght								
1	16.0	13906.250000	12318.304469	2000.0	5000.0	10000.0	20000.0	50000.0
3	25.0	11451.600000	23056.131583	5.0	500.0	2000.0	10000.0	100000.0
6	43.0	9564.883721	17672.039575	50.0	1000.0	3000.0	8000.0	92300.0
9	8.0	22635.000000	35141.063892	80.0	3000.0	7500.0	20000.0	100000.0
12	34.0	17085.882353	37030.822798	0.0	625.0	1585.0	13750.0	150000.0

Préstamos a 1 mes:

16 encuestados mencionan estar interesados en préstamos con plazo de un mes, cuyo principal motivo es obtener liquidez (37.5%). El monto promedio del préstamo es de $13,906 MXN, siendo el máximo $50,000 MXN y el mínimo de $2,000 MXN con una desviación estándar de $12

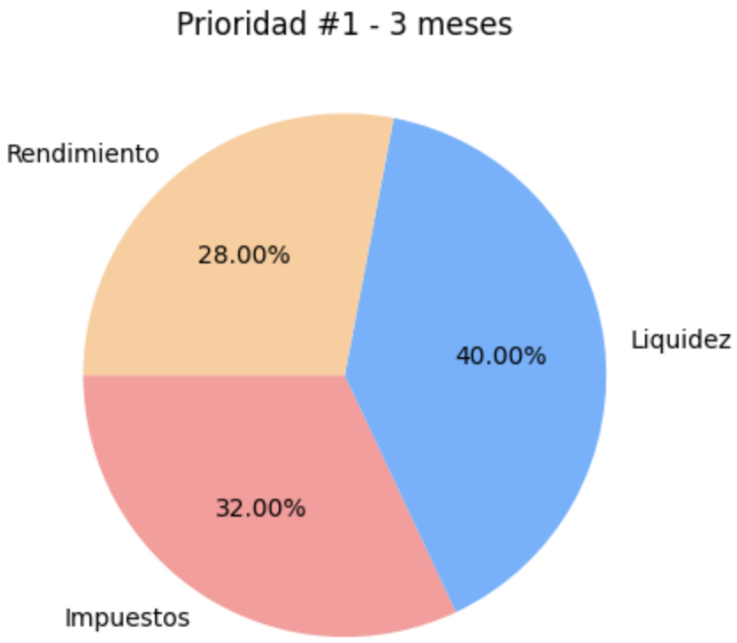
	Wish yield without CL	NEAR in MXN	Loan lenght
count	16.000000	16.000000	16.0
mean	0.812500	13906.250000	1.0
std	0.403113	12318.304469	0.0
min	0.000000	2000.000000	1.0
25%	1.000000	5000.000000	1.0
50%	1.000000	10000.000000	1.0
75%	1.000000	20000.000000	1.0
max	1.000000	50000.000000	1.0



Préstamos a 3 meses:

25 encuestados mencionan estar interesados en préstamos con plazo de tres meses, cuyo principal motivo es obtener liquidez (37.1%). El monto promedio del préstamo es de $11,451 MXN, siendo el máximo $100,000 MXN y el mínimo de $5.00 MXN con una desviación estándar de $23,056 MXN.

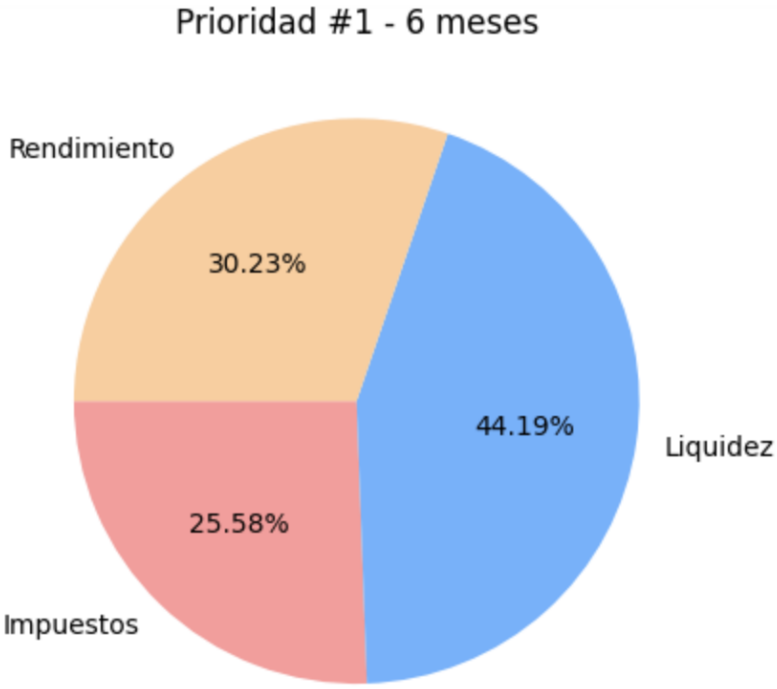
	Wish yield without CL	NEAR in MXN	Loan lenght
count	25.000000	25.000000	25.0
mean	0.920000	11451.600000	3.0
std	0.276887	23056.131583	0.0
min	0.000000	5.000000	3.0
25%	1.000000	500.000000	3.0
50%	1.000000	2000.000000	3.0
75%	1.000000	10000.000000	3.0
max	1.000000	100000.000000	3.0



Préstamos a 6 meses:

43 encuestados mencionan estar interesados en préstamos con plazo de seis meses, cuyo principal motivo es obtener liquidez (38.5%). El monto promedio del préstamo es de $9,564 MXN, siendo el máximo $92,300 MXN y el mínimo de $50.00 MXN con una desviación estándar de $17,672 MXN.

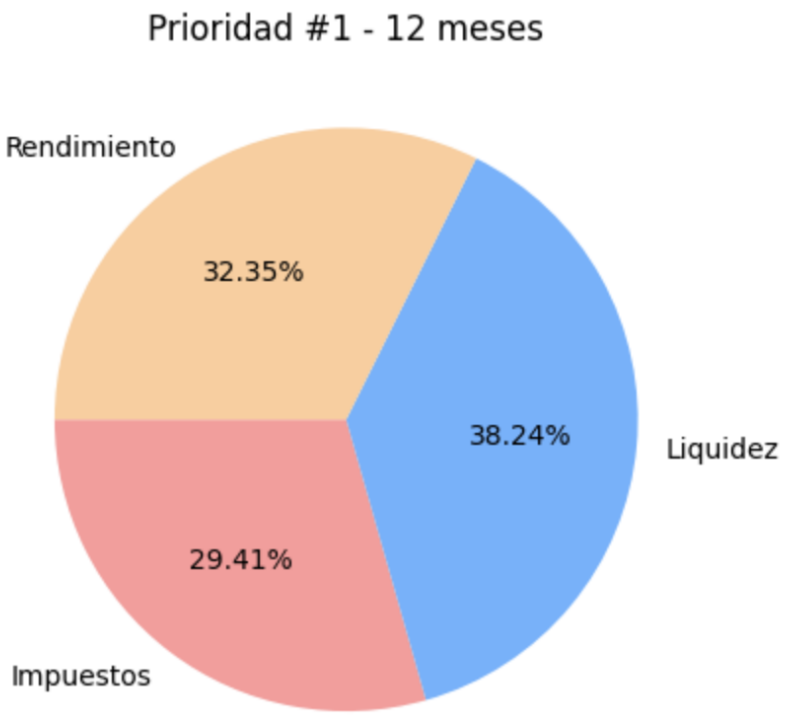
	Wish yield without CL	NEAR in MXN	Loan lenght
count	43.000000	43.000000	43.0
mean	0.930233	9564.883721	6.0
std	0.257770	17672.039575	0.0
min	0.000000	50.000000	6.0
25%	1.000000	1000.000000	6.0
50%	1.000000	3000.000000	6.0
75%	1.000000	8000.000000	6.0
max	1.000000	92300.000000	6.0



Préstamos a 12 meses:

34 encuestados mencionan estar interesados en préstamos con plazo de doce meses, cuyo principal motivo es obtener liquidez (36.1%). El monto promedio del préstamo es de $17,085 MXN, siendo el máximo $150,000 MXN y el mínimo de $0.00 MXN con una desviación estándar de $37,030 MXN.

	Wish yield without CL	NEAR in MXN	Loan lenght
count	34.000000	34.000000	34.0
mean	0.823529	17085.882353	12.0
std	0.386953	37030.822798	0.0
min	0.000000	0.000000	12.0
25%	1.000000	625.000000	12.0
50%	1.000000	1585.000000	12.0
75%	1.000000	13750.000000	12.0
max	1.000000	150000.000000	12.0



Matriz de Correlación:

Como paso intermedio se realiza una transformación de datos de tipo “One Hot Encoding” en la variable “Prioridad #1”; cabe recordar que las opciones son ‘Rendimientos’, ‘Impuestos’ o ‘Liquidez’.

El ‘one hot encoding’ se aplica a las variables nominales porque las variables nominales no tienen un orden natural o una relación numérica entre sus categorías. Sin embargo, para analizar (posibles) correlaciones es necesario que los datos de entrada sean numéricos.

El ‘one hot encoding’ convierte cada variable nominal en una columna binaria (donde un 1 indica que el registro pertenece a esa categoría y un 0 indica que no lo hace).

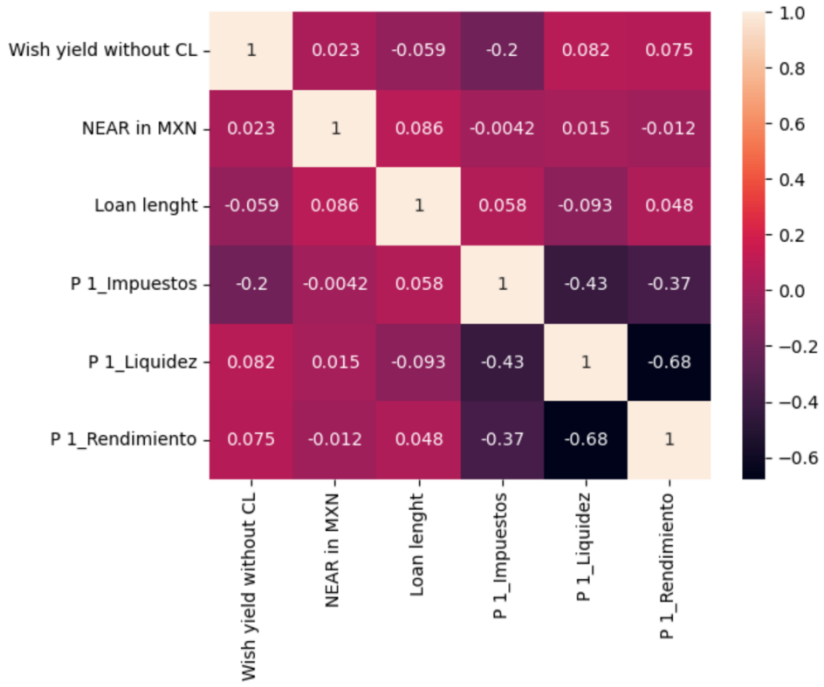
Cada registro tendrá un 1 en una sola columna y 0’s en las demás. De esta manera, se pueden representar las variables nominales de manera numérica y se evita la asignación arbitraria de valores numéricos a las categorías.

El nombre de las común de las variables binarias es ‘dummy variables’. Para explorar más se sugiere revisar $dummy\ variables^4$ de la sección de bibliografía.

La matriz queda de la siguiente manera:

	Wish yield without CL	NEAR in MXN	Loan lenght	P 1_Impuestos	P 1_Liquidez	P 1_Rendimiento
0	1	16000.0	1	1	0	0
1	1	1500.0	12	0	1	0
2	1	5000.0	6	0	0	1
3	1	150000.0	12	1	0	0
4	0	500.0	12	1	0	0
...	...	...	...	...	...	...
121	1	6000.0	6	0	1	0
122	1	5000.0	3	0	1	0
123	1	50000.0	3	0	1	0
124	1	25000.0	1	1	0	0
125	1	92300.0	6	0	1	0

Y su respectiva matriz de correlaciones es la siguiente



Conclusiones:

1. Al agrupar a los clientes a través de diferentes criterios, muestran que su mayor apetito es por la liquidez.

2. Los tenedores no están tan preocupados por los 'Impuestos', existe una correlación lineal negativa de 0.2 entre el deseo de acumular (staking) rendimiento contra el deseo de optimizar los impuestos.

3. Una correlación expresa la relación en el comportamiento entre variables y puede ir de -1 a 1, siendo 1 que se comportan de manera idéntica y -1 de manera idénticamente opuesta, si bien estas relaciones existen es necesario entender el contexto del análisis para darle significado a la relación. En este caso encontramos una fuerte correlación negativa (-0.68) entre la liquidez y el rendimiento, esto sentido si tomamos en cuenta que la liquidez y el rendimiento se comportan de manera inversa con respecto al tiempo. A menor tiempo de espera mayor liquidez, a mayor tiempo de espera mayor rendimiento mientras todo lo demás permanezca constante.

4. No se puede decir que los clientes estén segmentados por poder adquisitivo, las correlaciones lineales son cercanas a cero. La correlación es minúscula.

5. Los encuestados con tokens NEAR muestran una distribución de cola pesada (y un sesgo positivo) con respecto a los depósitos.

ANEXO:

Repositorio: https://github.com/MarioOvz/DS_AEM

Bibliografía:

1. Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). Probabilidad y estadística para ingeniería y ciencias (9th ed.). Pearson Educación. (ISBN 978-607-32-1417)

2. Huber-Carol, C., Balakrishnan, N., Nikulin, M. S., & Mesbah, M. (Eds.). (2002). Goodness-of-fit tests and model validity. New York: Springer.

3. Few, S. (2006). Now you see it: Simple visualization techniques for quantitative analysis. Analytics Press.

4. Asteriou, D., & Hall, S. G. (2015). Dummy variables. In Applied econometrics (3rd ed., pp. 209-230). London, UK: Palgrave Macmillan. ISBN 978-1-137-41546-2.