



UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE INGENIERÍA
Año 2020 - 2^{do} Cuatrimestre

ORGANIZACIÓN DE DATOS (76.05)

TRABAJO PRÁCTICO N°1

TEMA: Análisis Exploratorio

FECHA: 17 de noviembre de 2020

INTEGRANTES:

Arenas, Alejandro - 98199

<aarenas@gmail.com>

González Etchemaite, José I. - 86866

<ignacio.gonzet@gmail.com>

Rodríguez, Patricio - 103459

<parodriguez@fi.uba.ar>

Índice

1. Introducción	2
1.1. Repositorio GitHub	2
2. Limpieza de datos	2
2.1. Selección de variables	2
2.2. Transformación de variables	3
3. Análisis Exploratorios	3
3.1. Análisis 1: Territorios	4
3.1.1. Europa, Oriente Medio y África (EMEA)	4
3.1.2. Asia Pacífico (APAC)	5
3.1.3. Medio Este	6
3.1.4. Américas	7
3.1.5. Japón	7
3.2. Análisis 2: Clientes	8
3.2.1. Europa,Oriente Medio y África (EMEA)	8
3.2.2. Asia Pacífico (APAC)	9
3.2.3. Medio Este	10
3.2.4. Américas	11
3.2.5. Japón	12
3.3. Análisis 3: Vendedores	12
3.3.1. Ranking de vendedores por Oportunidades	13
3.3.2. Ranking de vendedores por Montos Totales	13
3.4. Análisis 4: Regiones	14
3.4.1. Europa, Oriente Medio y África (EMEA)	15
3.4.2. Asia Pacífico (APAC)	17
3.4.3. Medio Este)	19
3.4.4. Américas	21
3.4.5. Japón	23
4. Conclusiones	25

1. Introducción

Este trabajo consiste en realizar un análisis exploratorio sobre un set de datos que contiene toda la información de las oportunidades registradas en los últimos años por una empresa que se dedica a la venta de equipos de refrigeración, tanto de las ventas exitosas como de las que no llegaron a concretarse. Con el objeto de buscar tendencias y estructuras interesantes en la muestra disponible, se realizan distintos análisis para intentar obtener conclusiones útiles que sirvan de punto de partida para futuros trabajos.

1.1. Repositorio GitHub

El siguiente link contiene los distintos scripts para los análisis realizados en este trabajo práctico:

https://github.com/AAarenas/Data_TP1

2. Limpieza de datos

La etapa inicial del procesamiento de los datos incluye una adecuación de estos para su procesamiento subsecuente. Se vuelven a declarar algunos formatos (categorías, fechas, etc.), se completan algunos espacios en blanco (por ejemplo los territorios de la región Japón), y recién entonces se procede a eliminar los datos que contienen vacíos alguno de los campos de interés. De esta manera se reduce la cantidad de muestras descartadas.

2.1. Selección de variables

Para la realización de los análisis propuestos en este trabajo se hace una preselección de las variables con las que se trabajará, ya que muchos campos se encuentran incompletos, y otros carecen de relevancia para el estudio en cuestión. A continuación se enumeran las variables que consideradas a lo largo de este trabajo.

- **Opportunity_Name, Opportunity_ID:** nombre y número de identificación de la oportunidad. Se consideran sinónimos y se elige solo uno de ellos para diferenciar las oportunidades.
- **Opportunity_Created_Date:** fecha de creación de la oportunidad.
- **Region:** región de la oportunidad.
- **Territory:** territorio comercial de la oportunidad.
- **Opportunity_Owner:** vendedor de la oportunidad.
- **Account_Name:** nombre de la cuenta del cliente.
- **Source:** fuente de creación de la oportunidad.
- **TRF:** Toneladas de refrigeración.
- **Total_Amount:** monto total de cada ítem.

- **ASP:** precio promedio de la venta.
- **ASP_(converted):** precio promedio convertido a la moneda de la oportunidad.
- **Stage:** estado de la oportunidad.

2.2. Transformación de variables

El primer paso para poder agrupar y comparar oportunidades fue convertir los montos asociados a una unidad monetaria en común (USD). Para ello se creó una nueva columna realizando la transformación 2.1, donde el factor de conversión se computa utilizando datos locales a cada entrada, dando de esta manera una conversión dinámica que se ajusta a la fecha contemporánea a la oportunidad asociada.

$$\text{Total_Amount_Converted} = \frac{\text{ASP}(\text{converted})}{\text{ASP}} * \text{Total_Amount} \quad (2.1)$$

Luego se procedió con el agrupado de los datos por categorías, realizando las sumas de estos montos convertidos para así tener un total del monto de cada oportunidad. Del mismo modo, se sumaron las toneladas de refrigeración asociadas.

Por último, dentro de la etapa de limpieza de los datos, fueron descartadas aquellas entradas de oportunidades nunca concluidas. Esto es, se tomarán en cuenta para la exploración de los datos solo las oportunidades que terminaron en una venta exitosa o en una fallida (*Closed Won* o *Closed Lost* respectivamente).

3. Análisis Exploratorios

En esta sección del trabajo se realizarán diversos tipos de análisis exploratorios sobre el conjunto de datos. La idea es plantear distintos enfoques para captar tendencias existentes y poder elegir los más discriminantes, o bien realizar una agrupación final más provechosa de los resultados.

Previo a los análisis se realiza una separación de los datos por regiones, ya que se considera que cada una de estas puede presentar relaciones entre datos con estructuras totalmente distintas.

Métrica: φ

Sea un conjunto de oportunidades agrupadas bajo una categoría dada (e.g territorio). Se define una métrica o *score* sobre este conjunto de la siguiente forma:

$$\varphi = \frac{\text{Total_Won} - \text{Total_Lost}}{\text{Total_Won} + \text{Total_Lost}} \quad (3.1)$$

Donde:

$$\text{Total_Won} = \sum_{\forall i \in W} \text{Total_Amount}_{(i)} \quad , \quad W : \text{oportunidades ganadas} \quad (3.2)$$

$$\text{Total_Lost} = \sum_{\forall i \in L} \text{Total_Amount}_{(i)} \quad , \quad L : \text{oportunidades perdidas} \quad (3.3)$$

Este valor φ será proporcional a la cantidad de oportunidades concretadas por dicho grupo, al mismo tiempo que penalizará aquellas oportunidades no concretadas. El denominador es un

factor de normalización para adecuar el *score* a una escala donde se puedan comparar los grupos mas activos con los menos.

De esta manera cada grupo de oportunidades tendrá un valor asociado que permitirá realizar comparaciones y ordenamientos bajo este determinado criterio, el cual se utilizará durante el resto del trabajo.

3.1. Análisis 1: Territorios

El primer análisis propuesto consiste en realizar un ranking de territorios para cada región, según la cantidad total del monto de todas sus oportunidades, separando las oportunidades ganadas de las perdidas.

Para poder mostrar en una mejor escala, se optó por realizar una simple transformación: a cada monto total se le resta la media de montos totales de la población (región para este caso), y se divide por el desvío estándar muestral, estimado de los datos.

3.1.1. Europa, Oriente Medio y África (EMEA)

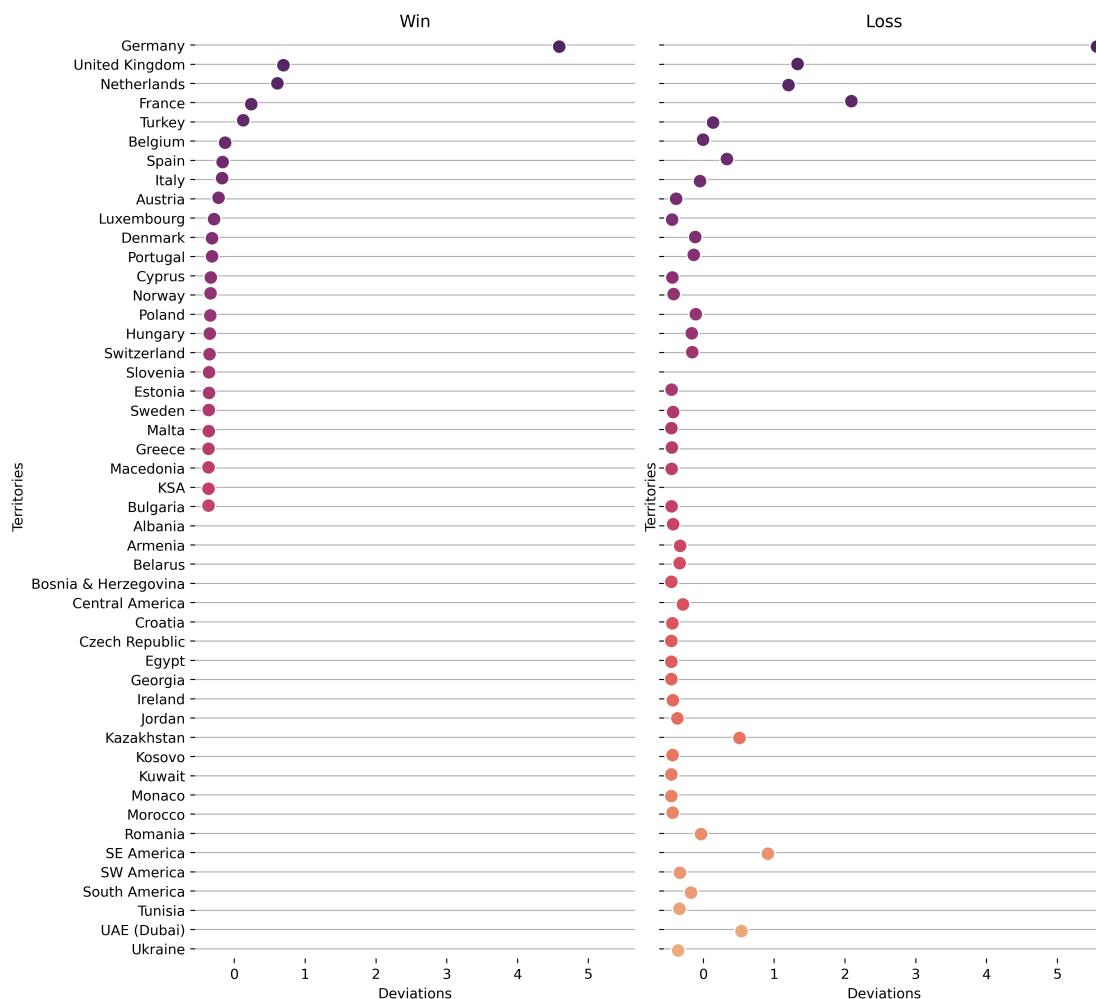


Figura 3.1: Montos totales para territorios en EMEA

En esta región se observa una dominante participación de Alemania, quien maneja cantidades de monto muy superiores al resto en la región.

Si se calcula la métrica definida antes φ sobre los territorios, los resultados son diferentes. Si bien Alemania tendrá siempre una buena aceptación en la empresa porque maneja cifras muy superiores, se observa que no tiene una buena relación entre oportunidades concretadas y descartadas. Esto es, si se quisiera predecir la concreción de una oportunidad futura discriminando solo por territorio de origen de la oportunidad, poco se podrá saber al respecto, ya que observando el ejemplo de Alemania se ven mas oportunidades rechazadas que concretadas.

3.1.2. Asia Pacífico (APAC)

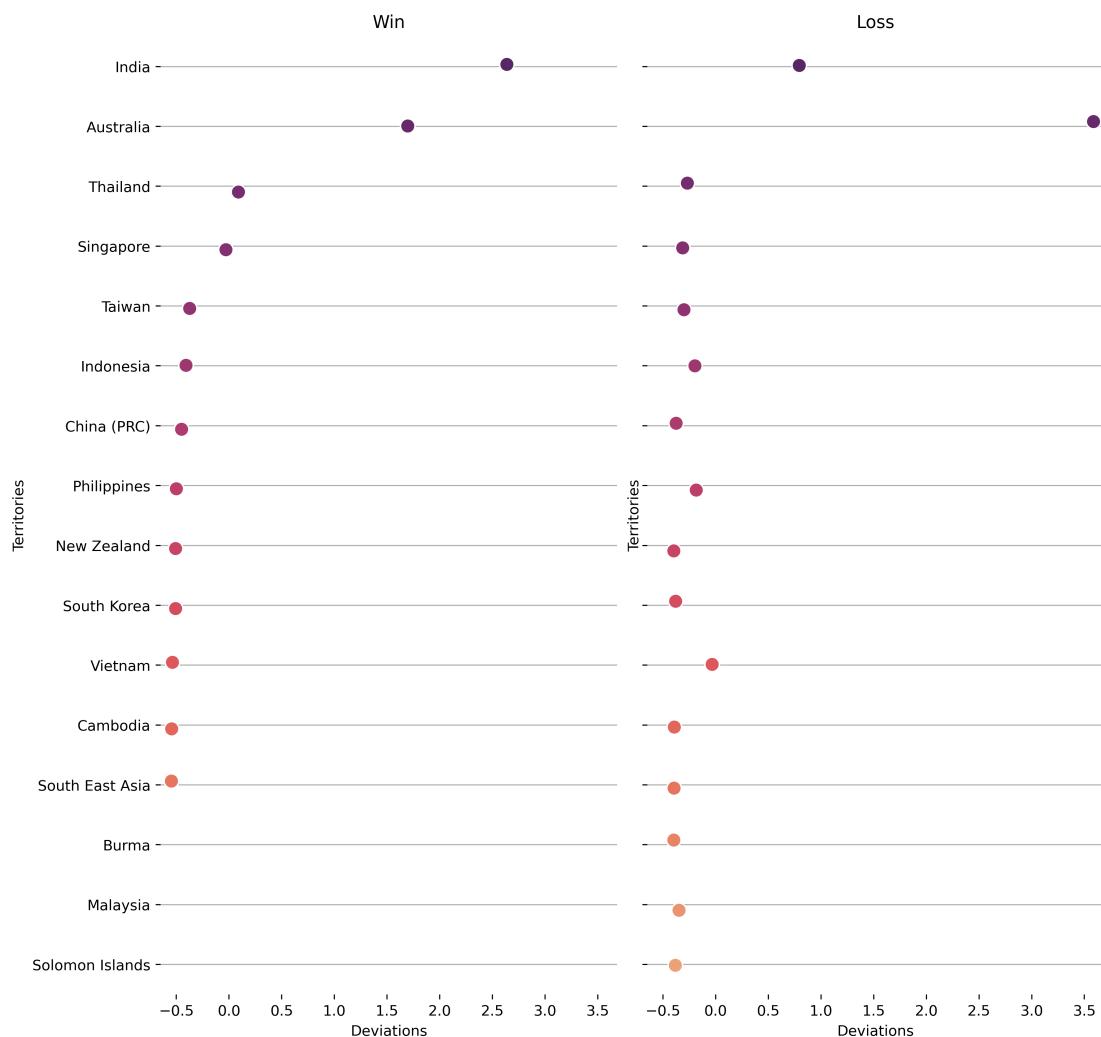


Figura 3.2: Montos totales para territorios en APAC

Siguiendo con el análisis en Asia Pacífico, se observa una cúpula de pocos territorios que concentran la mayor parte de las oportunidades: India y Australia en un primer escalón, junto a Tailandia y Singapur superan la media (valor 0.0 en el eje X del gráfico), destacándose del resto de la región.

Si bien será necesario extender este análisis a todos los demás territorios, un primer filtrado

permitiría quedarse con un subconjunto de territorios que tengan una participación mas positiva con la empresa para una determinada región.

3.1.3. Medio Este

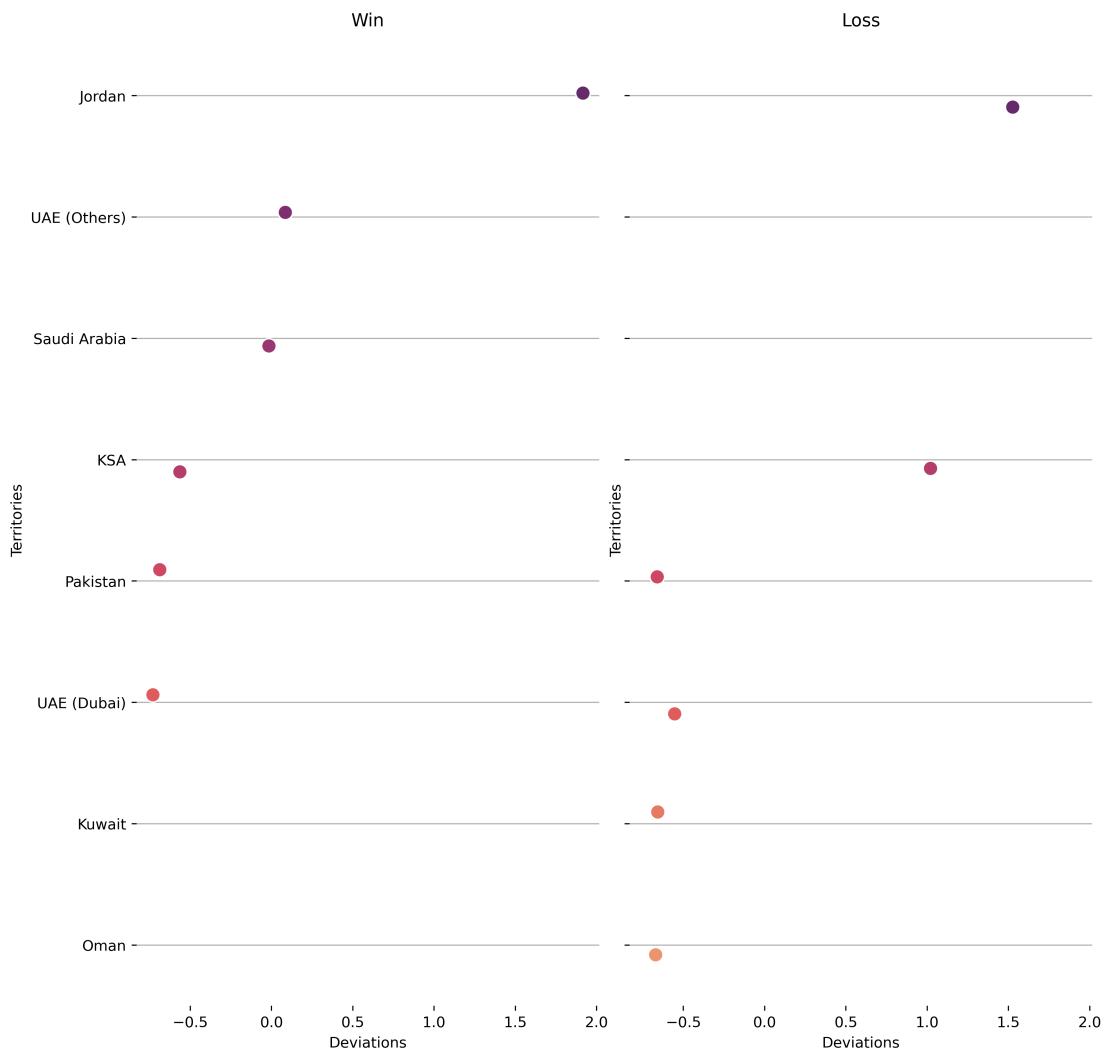


Figura 3.3: Montos totales para territorios en Medio Este

Para la región de Medio Este se observa a Jordania dominando las oportunidades, con mas oportunidades ganadas que perdidas. Solo junto a UAE y Arabia Saudí superan la media de montos en oportunidades ganadas.

3.1.4. Américas

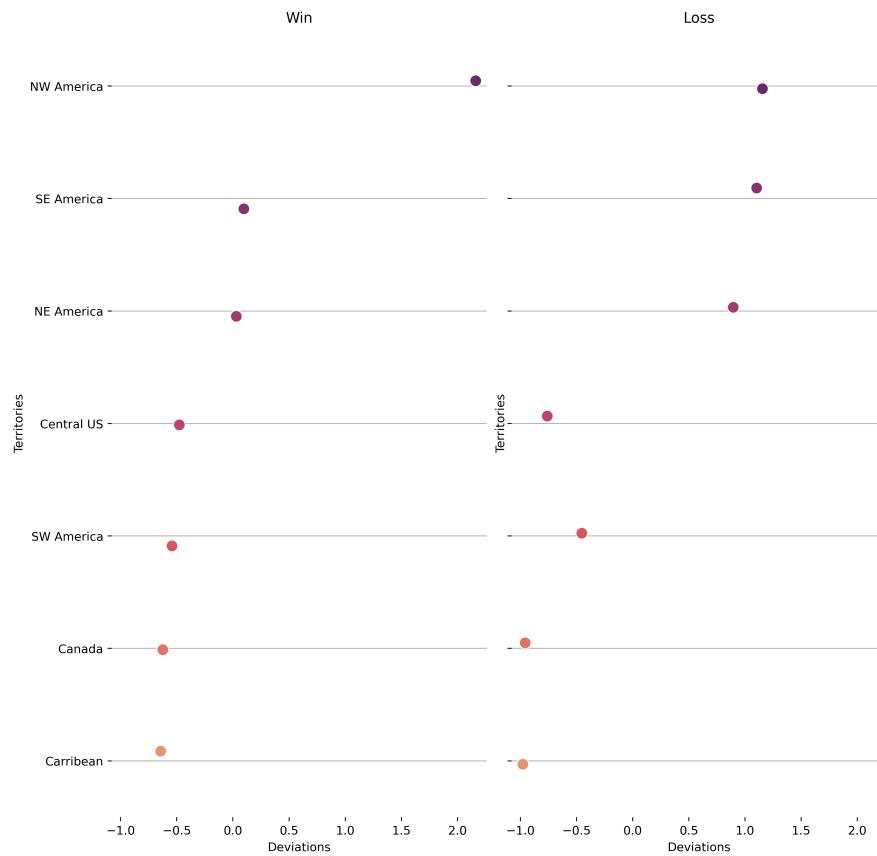


Figura 3.4: Montos totales para territorios en Américas

En la región de las Américas se observa una distribución mas uniforme de las oportunidades, teniendo a América Noroeste como dominante.

3.1.5. Japón

Como Japón se considera una región por sí sola en este análisis, en la siguiente tabla se resumen sus datos junto a los territorios cabecera de cada región, para poder obtener una idea de la magnitud del monto total de sus oportunidades.

Territorios	Closed Won	Closed Lost	φ
Japan	1.17e+08	2.41e+08	-0.347
Alemania	3.06e+08	1.05e+09	-0.548
NO America	4.00e+08	1.75e+09	-0.628
India	8.96e+07	6.99e+08	-0.773
Jordania	0.22e+07	1.11e+08	-0.961

Cuadro 3.1: Montos Totales Japón

Como se observa en el cuadro 3.1, Japón presenta un monto total de operaciones ganadas del orden de Alemania y NO América, dando al mismo tiempo una cantidad menor de oportunidades

perdidas. Este resultado se ve reflejado en el valor de la métrica φ , siendo Japón superior al resto de los territorios cabeceras de cada región.

3.2. Análisis 2: Clientes

En esta sección se realiza un análisis exploratorio de los clientes de la empresa, nuevamente haciendo una separación por regiones.

En cada gráfico se observan los clientes de mayor participación con la empresa. Esto es, se muestran los diez clientes con mayor cantidad de oportunidades (totales). Para ello se los sitúa en un sistema de ejes cartesianos con el objeto de mostrar sus respectivas métricas φ . De esta manera, su ubicación en el plano quedará determinada por la efectividad del cliente para concretar sus oportunidades.

3.2.1. Europa,Oriente Medio y África (EMEA)

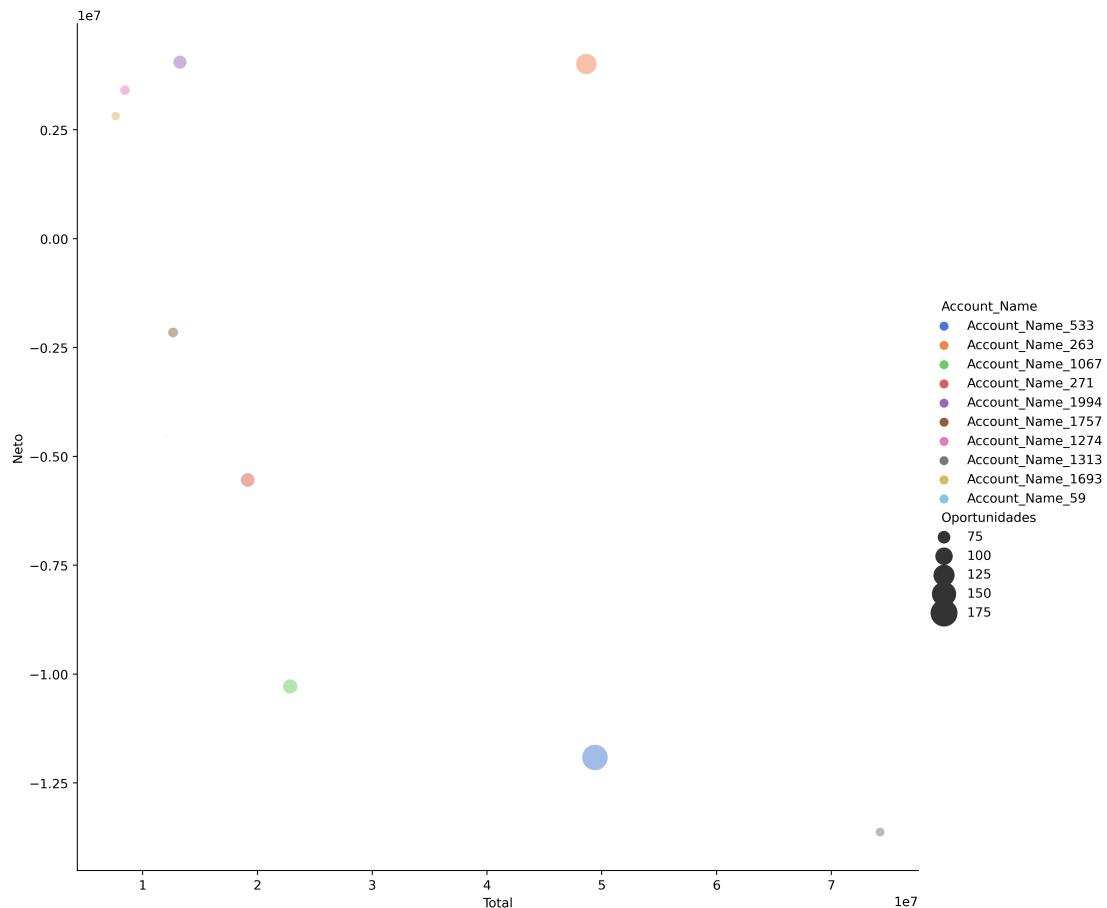


Figura 3.5: Top 10: clientes con mas oportunidades en EMEA.

Siendo EMEA posiblemente la región con mayor cantidad de oportunidades, se observa que el cliente que mayor neto aporta es *Account_Name_264*, ya que posee el mayor valor en el eje Y. Sin embargo, este cliente no tiene la mejor concreción de oportunidades, ya que aparecen tres clientes con mejor relación φ (parte superior izquierda).

3.2.2. Asia Pacífico (APAC)

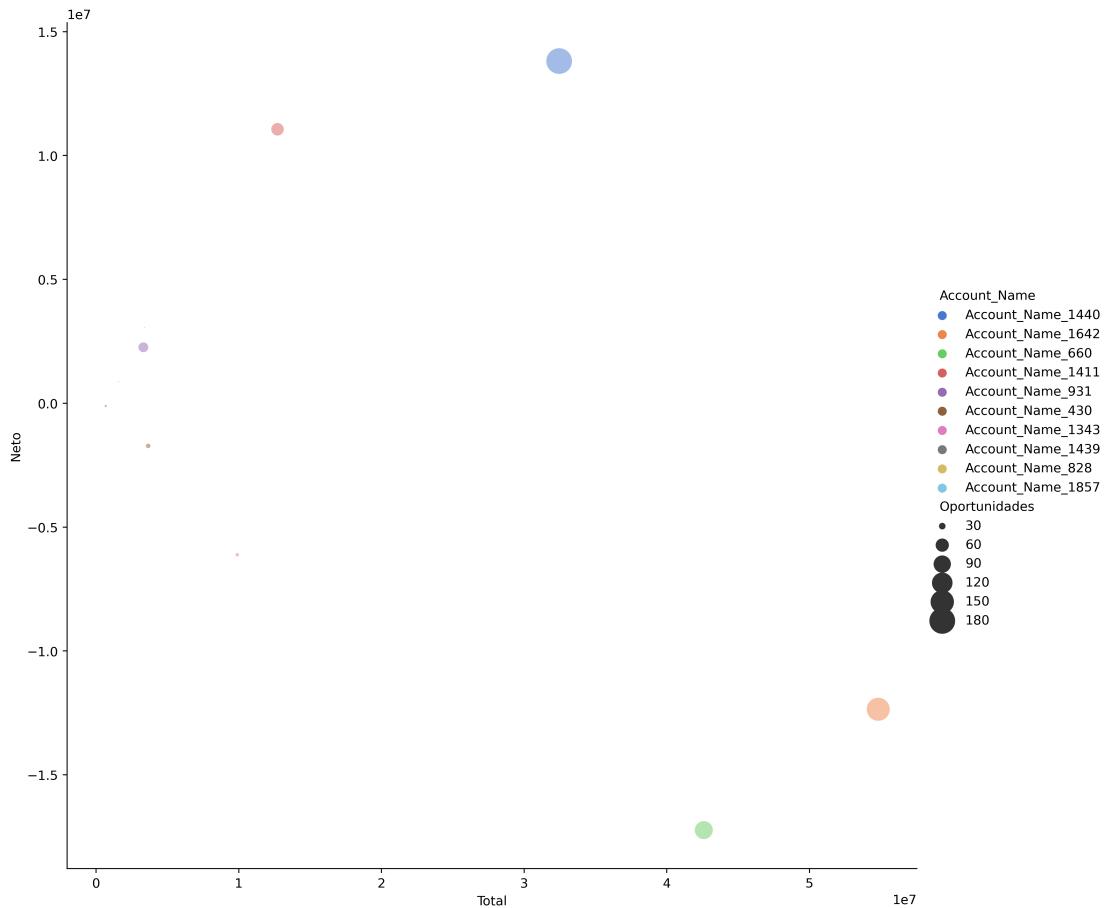


Figura 3.6: Top 10: clientes con mas oportunidades en APAC.

Para Asia Pacífico se observan tres clientes con mayor monto de oportunidades ganadas que perdidas. Dentro de ellos, el que mejor relación φ tiene es *Account_Name_1642*, mientras que *Account_Name_1857* maneja una cantidad bastante mayor de oportunidades.

3.2.3. Medio Este

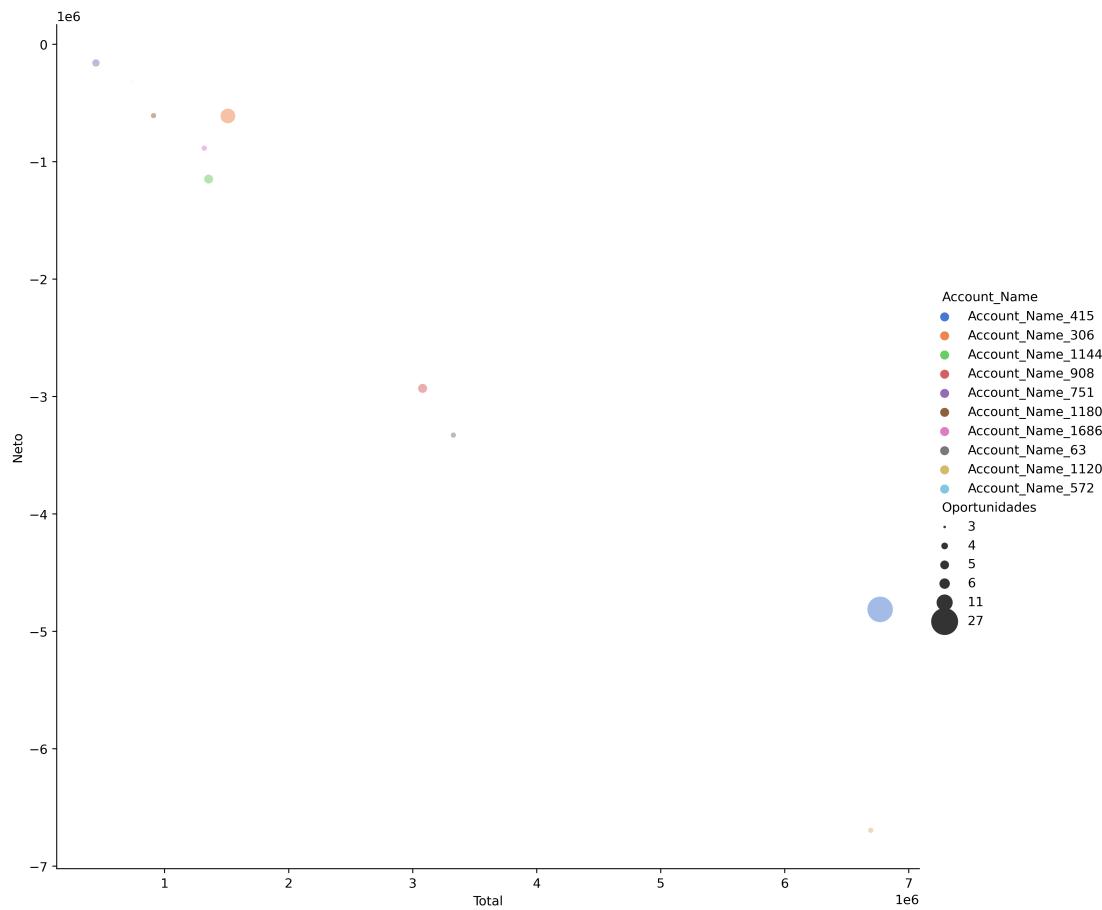


Figura 3.7: Top 10: clientes con mas oportunidades en Medio Este.

En Medio Este no se observan clientes que concreten mas dinero en oportunidades ganadas que perdidas, pero presenta un conjunto de cinco clientes con métrica φ considerablemente mejor que los demás. Se puede ver que el cliente con mayores oportunidades con la empresa, *Account_Name_572*, tiene una mala métrica, indicando que es un cliente que concreta bastante menos de lo que presupuesta.

3.2.4. Américas

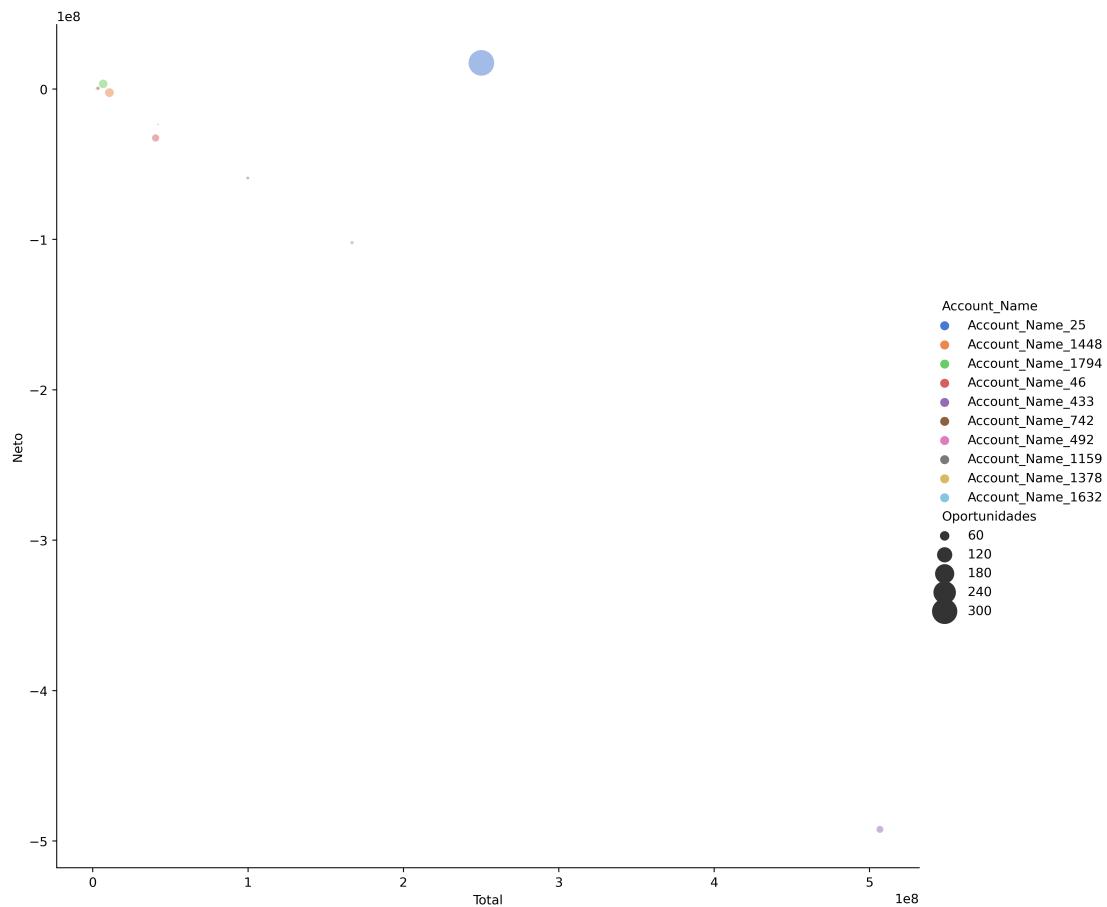


Figura 3.8: Top 10: clientes con mas oportunidades en Américas.

En la región de Las Américas se presenta un panorama similar a EMEA, habiendo un cliente con mayor participación en las oportunidades de la empresa, *Account_Name_1632*, mientras que le siguen cuatro clientes mas de cerca que se distinguen considerablemente del resto de los territorios en la región.

3.2.5. Japón

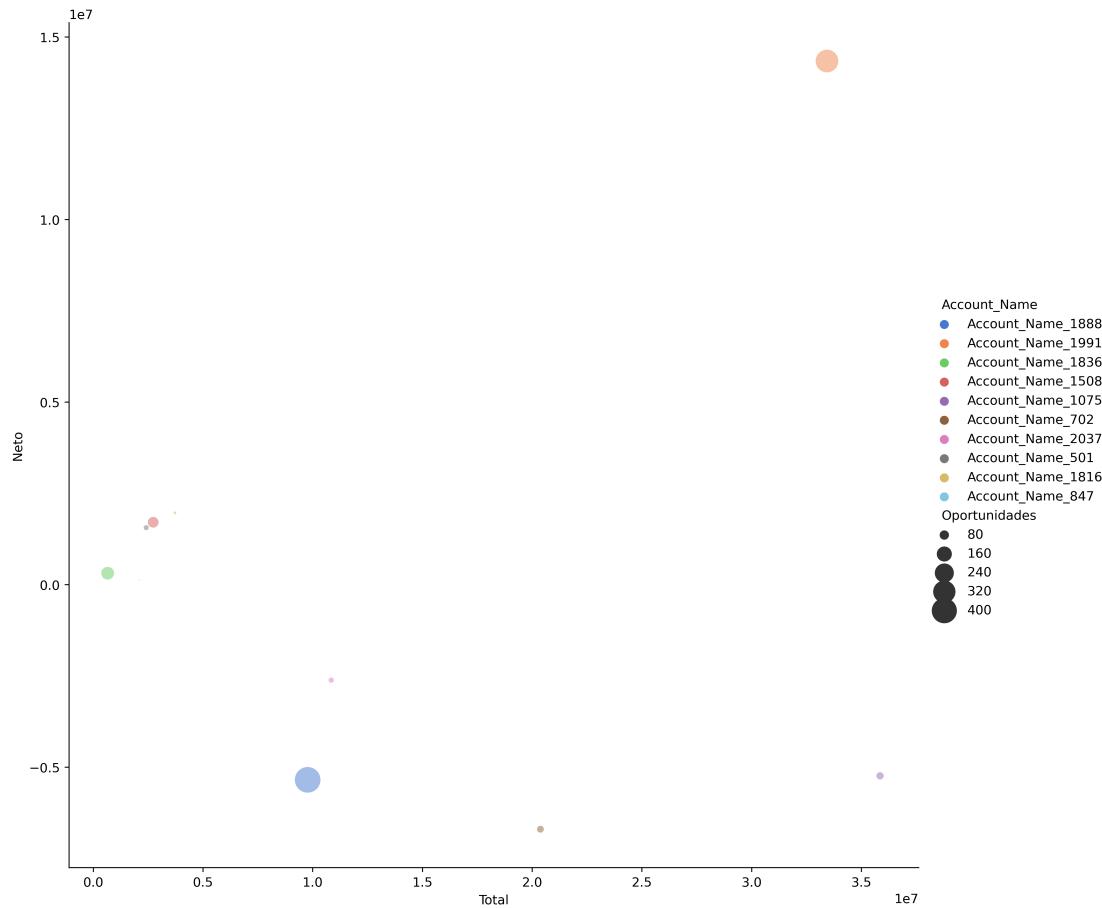


Figura 3.9: Top 10: clientes con mas oportunidades en Japón.

En la región de Japón existe un cliente dominante que contiene muchas oportunidades, y una buena métrica φ (como la tiene Japón en general según el análisis previo), *Account_Name_1991*.

Existen además tres clientes con un monto Neto (total ganado menos total perdido) positivo, seguido por un quinto (*Account_Name_847*), que si bien podría tener mejor métrica, la cantidad de oportunidades que ha tenido con la empresa a lo largo de los años registrados indican que es un cliente a tener en cuenta siempre.

3.3. Análisis 3: Vendedores

En esta sección se hace un análisis global de los vendedores de la empresa. Para ello se toma a cada *Opportunity_Owner* junto a sus oportunidades ganadas y perdidas, y para cada subconjunto se computa la suma total de distintas oportunidades. Lo mismo se hace con los respectivos montos totales.

Para una mejor interpretación visual de los resultados, aquí los vendedores no se ordenan por métrica, ya que esto puede caer en una escala visual muy pequeña y escapa de este análisis. De esta forma, se opta por ordenarlos según el rango de diferencias entre lo ganado y lo perdido.

3.3.1. Ranking de vendedores por Oportunidades

En este ranking de vendedores según sus oportunidades, se puede ver que son muy pocos los que tienen mas oportunidades ganadas que perdidas, algo que ya se podía observar en análisis anteriores donde las métricas eran mayormente negativas.

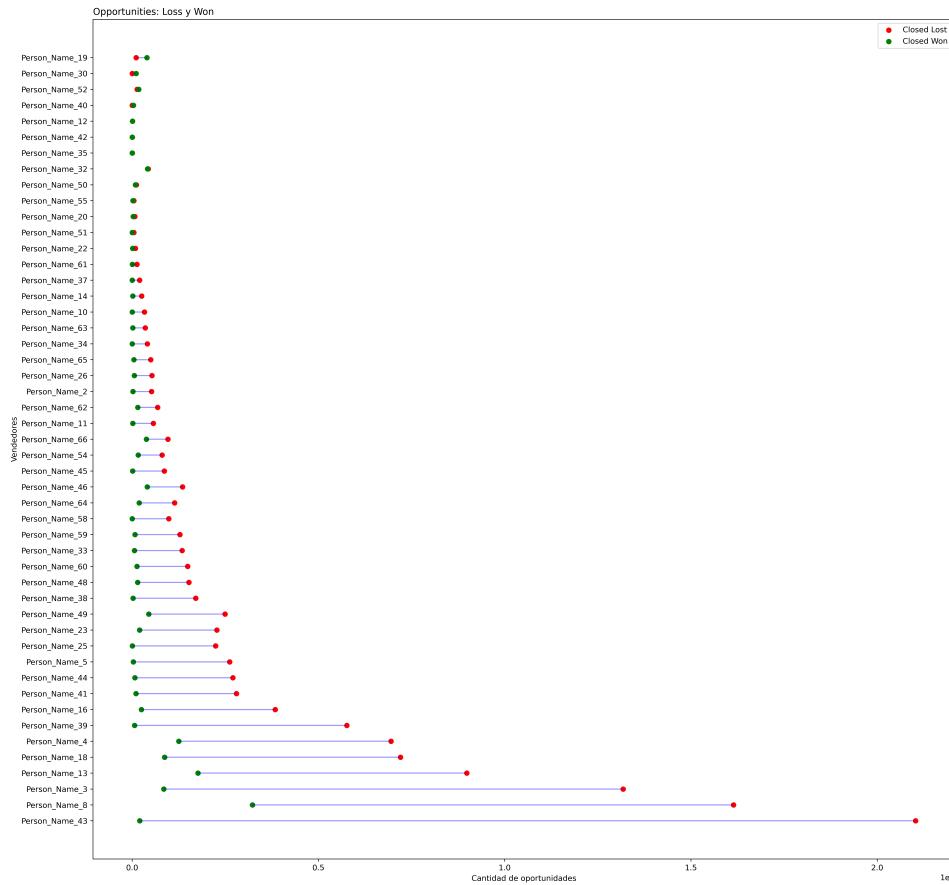


Figura 3.10: Vendedores según sus oportunidades.

Luego de una zona media donde no se encuentran grandes diferencias entre las oportunidades ganadas y perdidas, se llega a la parte baja donde se observan algunos vendedores que raramente concretan alguna oportunidad. Si bien un análisis mas detallado podría servir bastante para encontrar la razón de por que suceda esto, se puede decir *a priori* que una oportunidad futura en manos de estos vendedores tendrá muchas probabilidades de no ser concretada.

3.3.2. Ranking de vendedores por Montos Totales

Este análisis sirve para complementar el anterior, donde se muestra para cada vendedor el mismo criterio, pero esta vez analizando los montos totales de sus oportunidades. De esta

manera, se puede encontrar una relación entre el número de oportunidades que cada vendedor de la empresa ha manejado, y la magnitud de ellas.

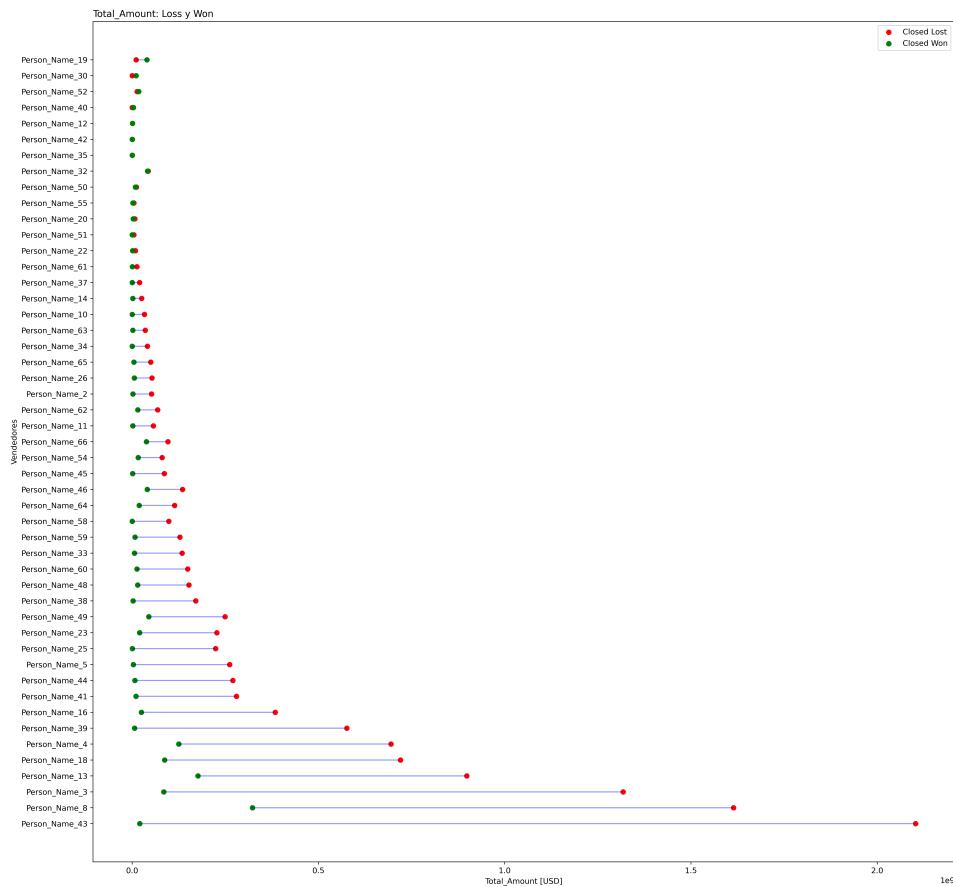


Figura 3.11: Vendedores según sus montos totales.

Se observa que se repiten varios vendedores en ambos extremos de la tabla, pudiendo concluir con mayor certeza de que los mas bajos necesitan cambiar sus estrategias. También se sugiere una redistribución de las oportunidades entrantes en la empresa, de forma de optimizar las ganancias obtenidas a través de una mejor concreción de estas.

3.4. Análisis 4: Regiones

En este último análisis se realiza una visualización de la evolución temporal de los montos totales en cada región, para poner en evidencia cuáles fueron los mejores períodos de la empresa para cada región. Por cada una de ellas se mostrarán dos gráficos diferentes, uno mostrando valores netos de montos en dólares, para poder tener así una observación directa de la magnitud del monto para cada semestre. Por otro lado, se agrega un segundo gráfico mostrando la

distribución proporcional de todo el monto presupuestado para una región dada. Esto es, para cada semestre se mostrará el porcentaje del monto histórico de la región en cuestión.

Los semestres se cuentan a partir del año 2013, siendo el semestre 1 el primero de este año, el 2 el segundo, el 3 el primer semestre del año 2014, y así sucesivamente.

3.4.1. Europa, Oriente Medio y África (EMEA)

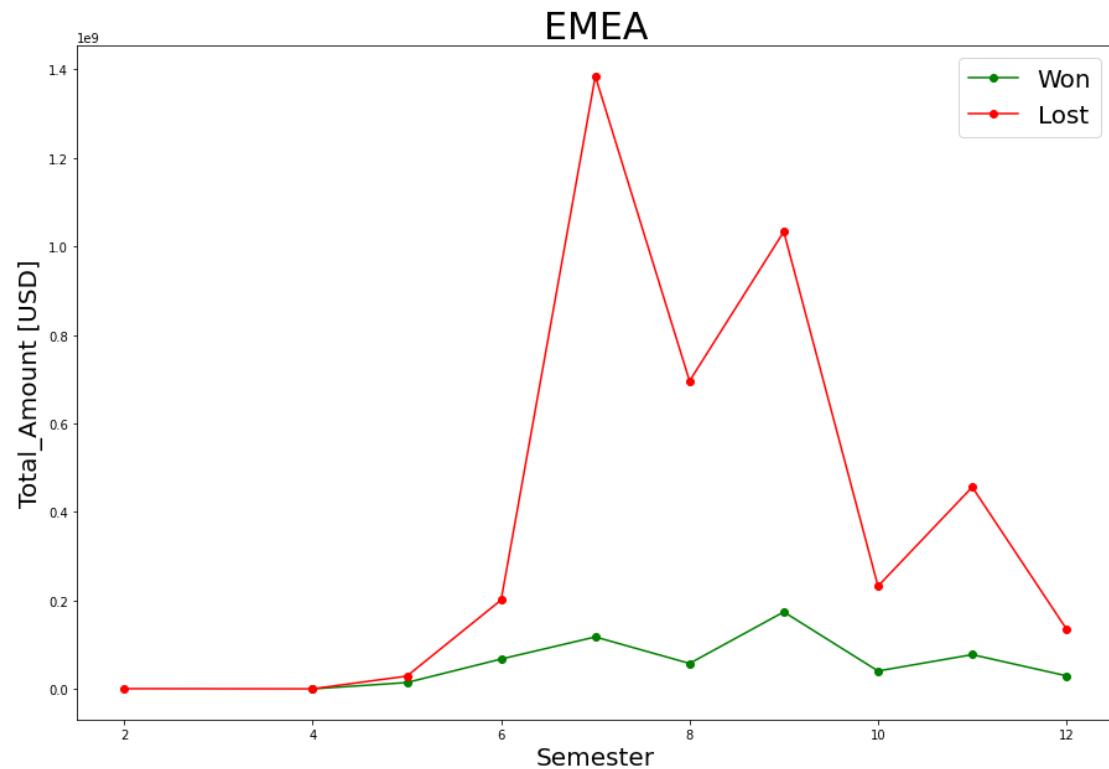


Figura 3.12: Distribución semestral de montos para EMEA.

En esta visualización se observa que a partir del segundo semestre del año 2015, los presupuestos de oportunidades perdidas son mucho mayores que los ganados para esta región.

Por otro lado, se observa que a partir del año 2016, aparece una mayor cantidad de montos totales en los primeros semestres de cada año.

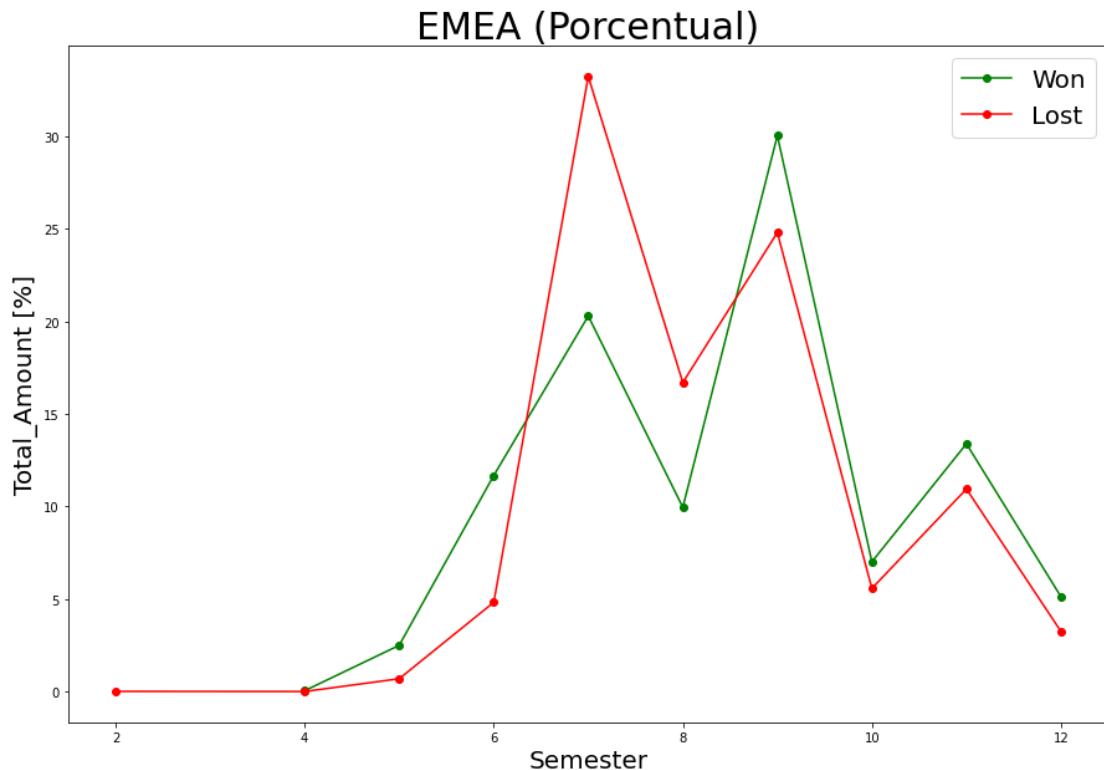


Figura 3.13: Distribución semestral porcentual de montos para EMEA.

En esta figura se pueden observar relaciones porcentuales en la variación del monto entre semestres. Por ejemplo, en el primer semestre del año 2017 se incrementaron los montos ganados en un 50 % respecto al año anterior (de 20 a 30 %), habiendo disminuido las perdidas en un 30 % (de 35 a 25 % aproximadamente).

3.4.2. Asia Pacífico (APAC)

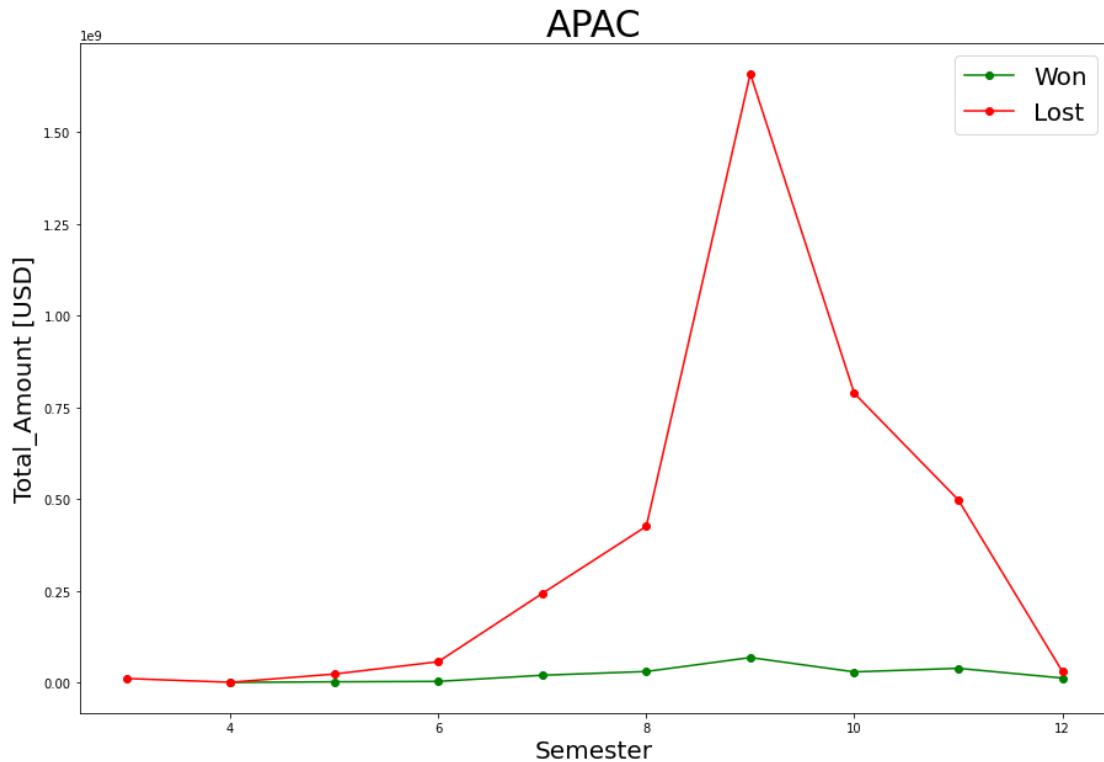


Figura 3.14: Distribución semestral de montos para Asia Pacífico.

En esta región se observa algo similar a la región anterior, donde los montos que se pierden son de mayor orden que los que se ganan. También se puede observar un pico de oportunidades no concretadas en el primer semestre del año 2017, con un valor que triplica en valor a todos los semestres anteriores.

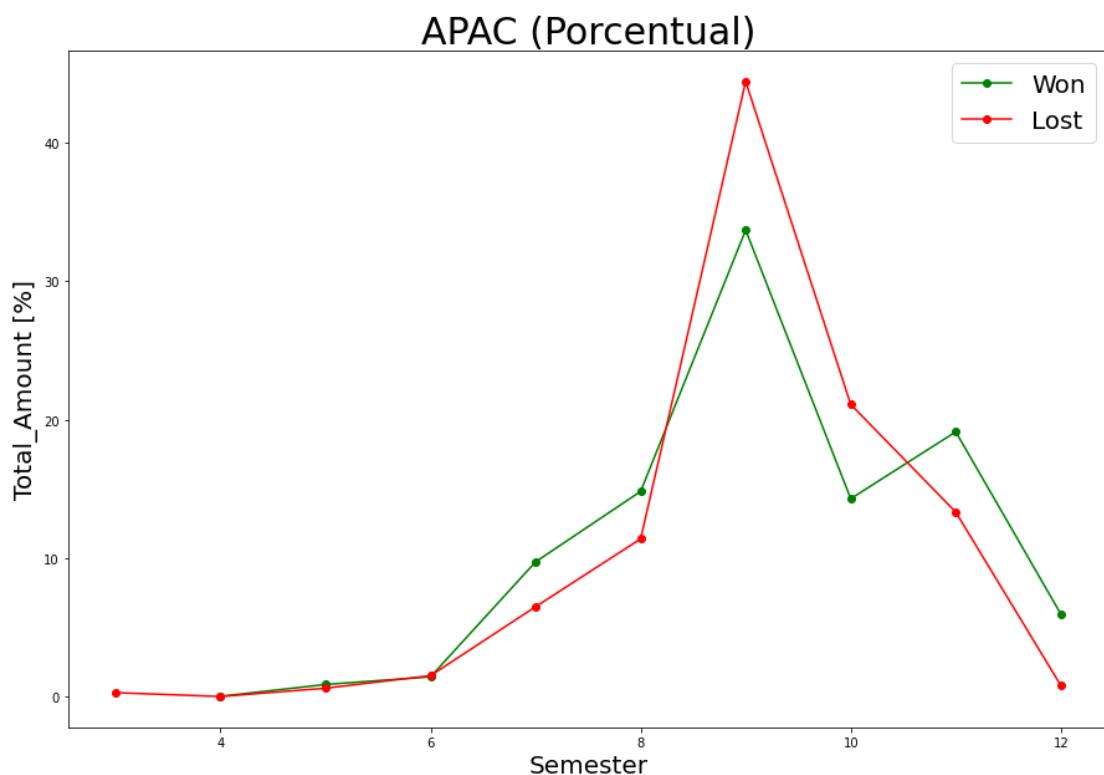


Figura 3.15: Distribución semestral porcentual de montos para Asia Pacífico.

En esta visualización se observa un pico de actividad en el primer semestre del 2017 para ambos tipos de oportunidades. Se observa que el 50 % del monto total histórico de oportunidades perdidas se concentra aquí, junto con un 35 % del monto histórico de oportunidades ganadas. Un análisis mas detallado sobre este semestre podría esclarecer los resultados observados.

3.4.3. Medio Este)

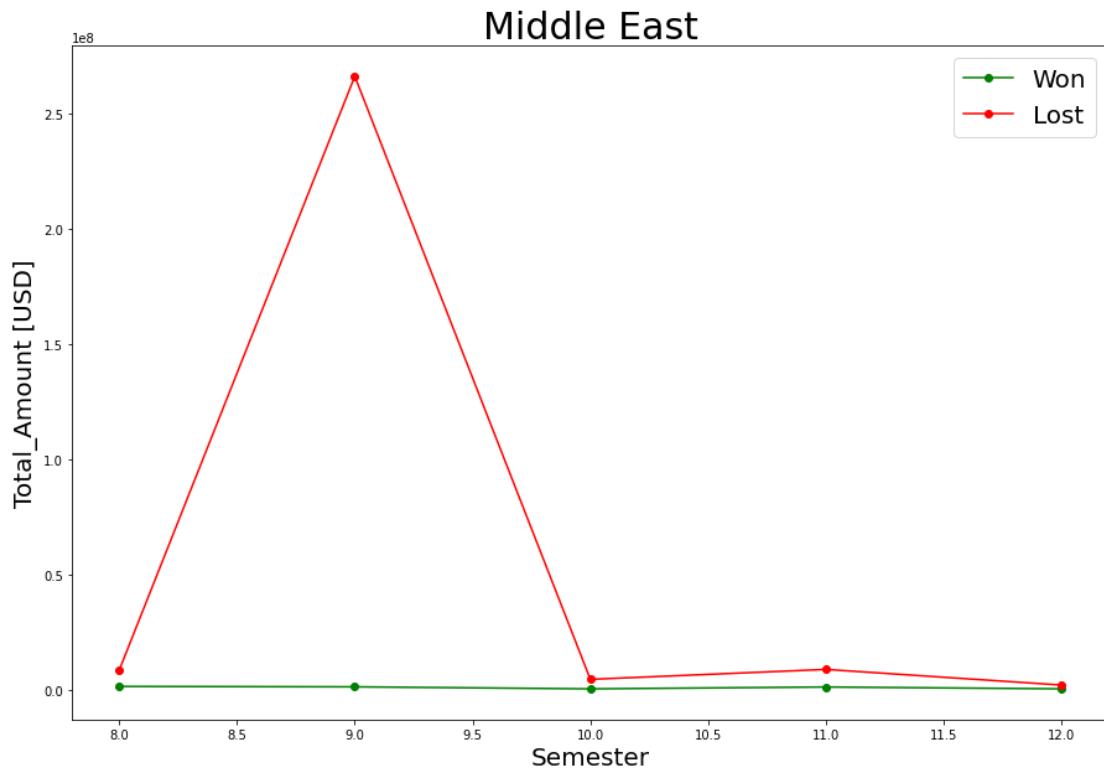


Figura 3.16: Distribución semestral de montos para Medio Este.

Esta región comenzó a tener relaciones con la empresa a partir del segundo semestre del año 2016. Teniendo un pico inicial de oportunidades perdidas en el primer semestre del año 2017, luego consigue equiparar un poco ambos tipos de oportunidades, ganadas y perdidas. Igualmente, los montos netos son de un orden menor al resto de las regiones.

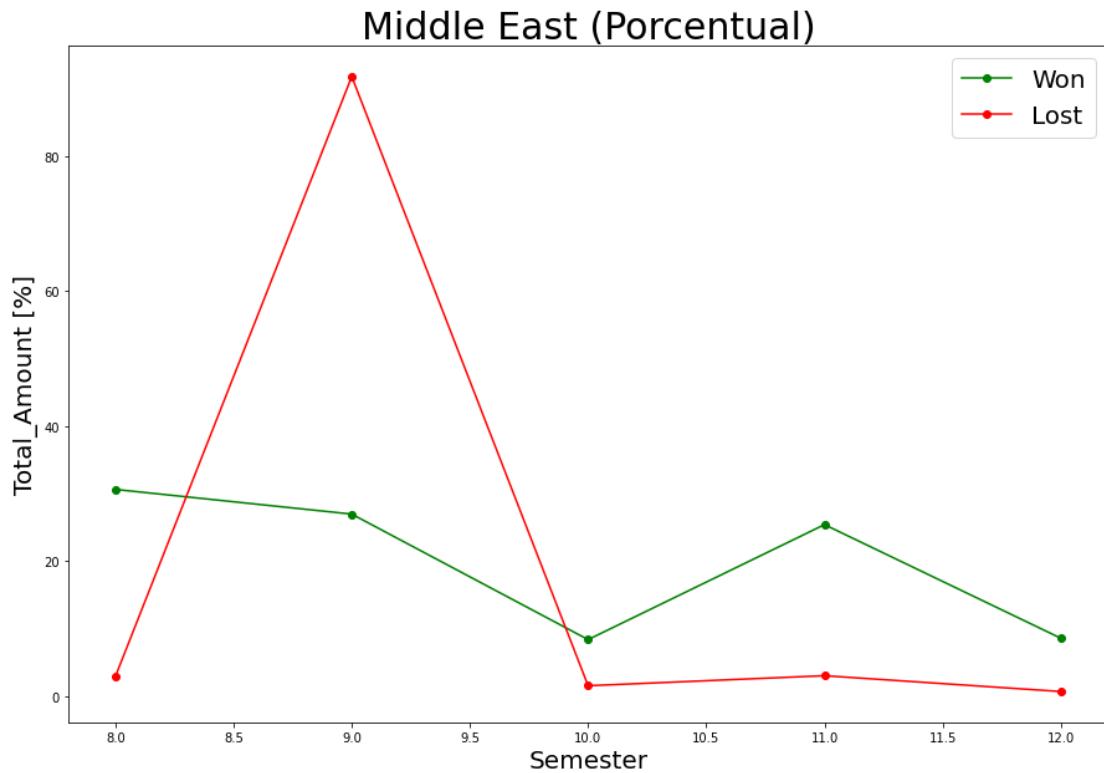


Figura 3.17: Distribución semestral porcentual de montos para Medio Este.

Algo parecido que en el gráfico anterior, se puede ver en relaciones porcentuales la evolución de las oportunidades a lo largo del tiempo. El 85 % aproximado de las oportunidades se concentró en el segundo semestre de participación de la región para con la empresa.

3.4.4. Américas

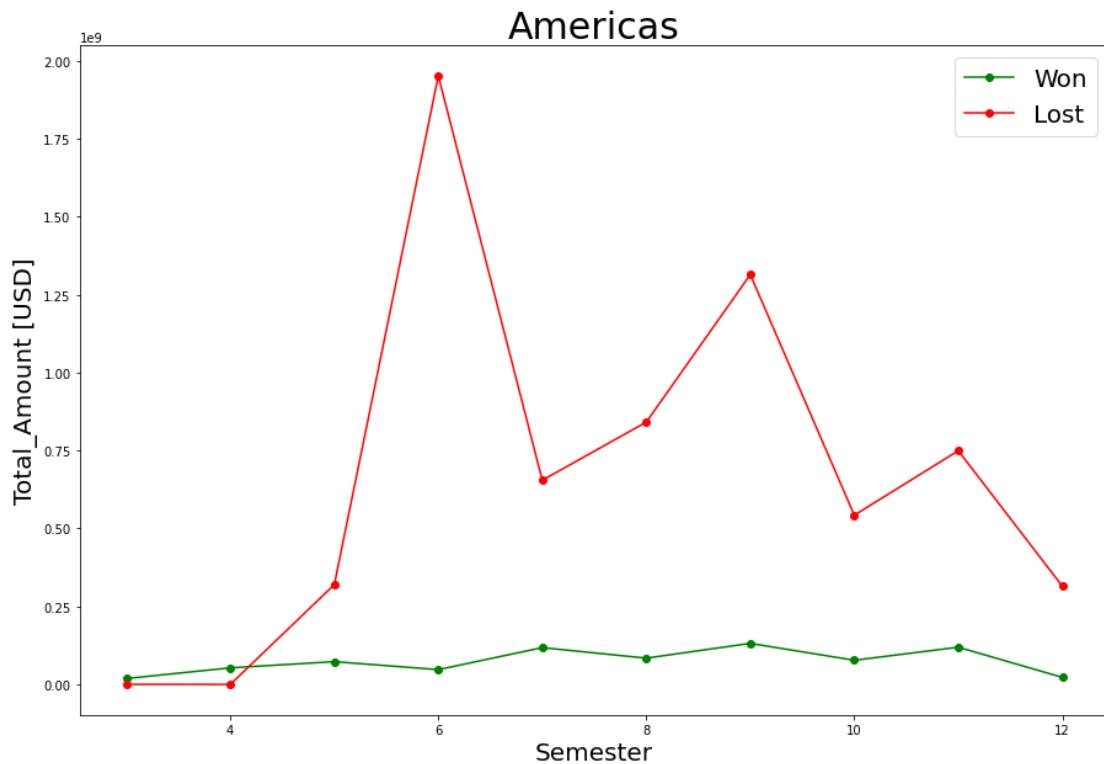


Figura 3.18: Distribución semestral de montos para Américas.

En Américas se puede ver una evolución similar a las demás regiones, con mayor participación en los primeros semestres de cada año, pero con una disminución en los montos de las oportunidades perdidas a lo largo del tiempo.

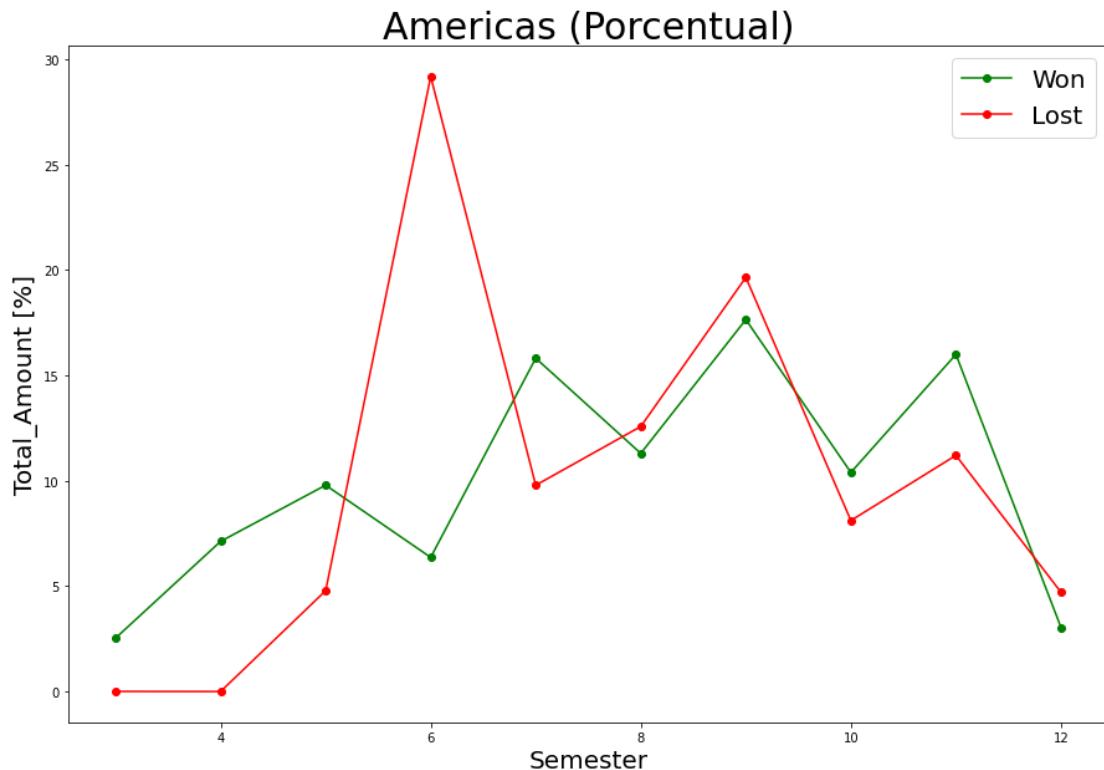


Figura 3.19: Distribución semestral porcentual de montos para Américas.

Aquí se observa que para la región Américas, la distribución del monto histórico de la empresa tiene una distribución mas estable que las demás.

3.4.5. Japón

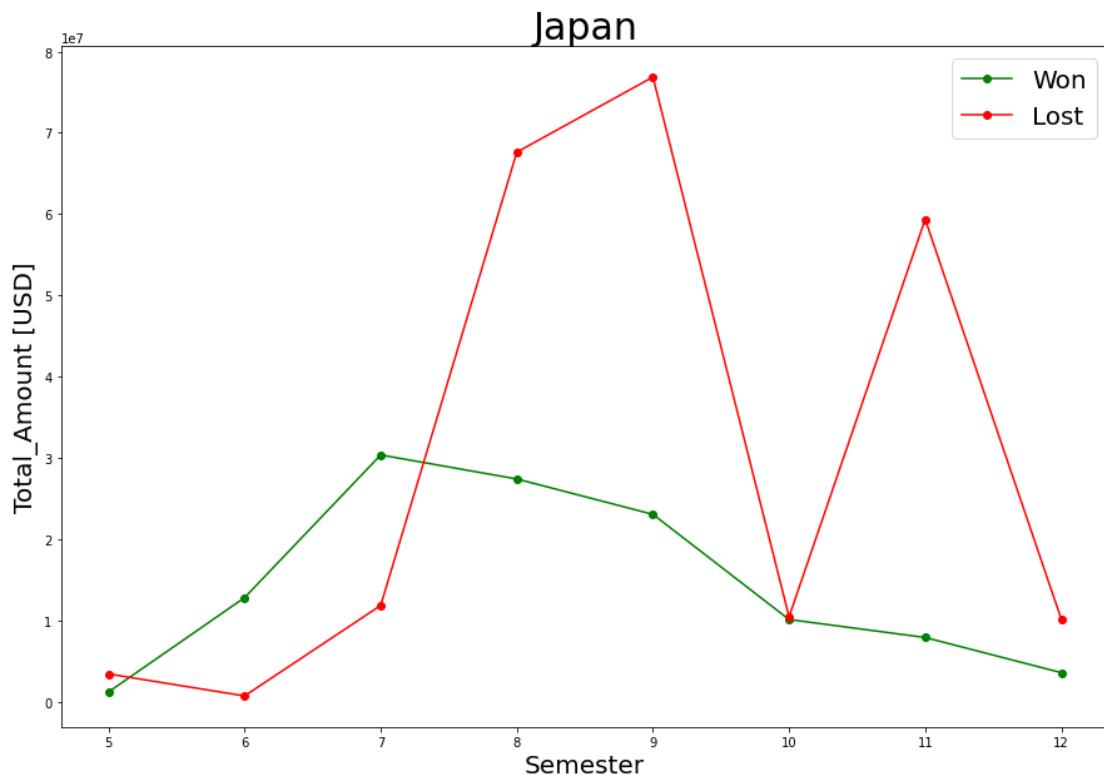


Figura 3.20: Distribución semestral de montos para Japón.

En este gráfico se puede observar que el orden de los montos perdidos es de dos órdenes menor que las regiones mas grandes (EMEA, APAC y Américas). Además, la distancia entre oportunidades ganadas y perdidas es mucho menor, lo cual se observó para esta región a lo largo de los distintos análisis y la métrica asociada a las categorías pertenecientes a esta región. Se ve que en los últimos años la participación de Japón con la empresa ha disminuido su efectividad, mediante una disminución en el monto de oportunidades ganadas.

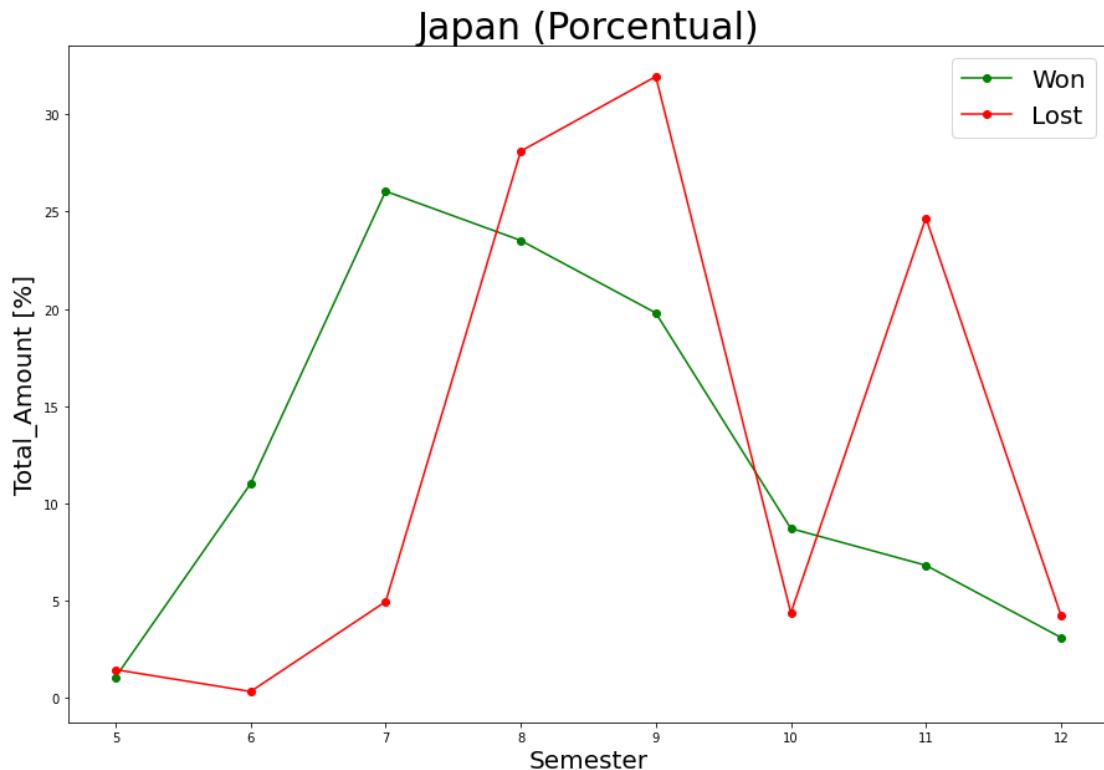


Figura 3.21: Distribución semestral porcentual de montos para Japón.

En este gráfico se observa que en los años 2016 y 2017 se concentra la mayor parte del monto total histórico ganado, mientras que el perdido se distribuye a partir del segundo semestre del año 2016 en adelante. Se ve muy poca participación en el último semestre del año 2018, pero esto es algo común en todas las regiones, probablemente debido a falta de registros en el conjunto de datos.

4. Conclusiones

En este trabajo se realizaron distintos enfoques sobre un conjunto de datos para intentar captar tendencias sobre ellos mediante supuestos simples pero directos. Si bien la complejidad de los análisis realizados no fue alta, se lograron extraer algunos resultados concluyentes respecto a distintas observaciones. Por ejemplo, se lograron captar vendedores de muy baja eficiencia dentro de la empresa. También se pudieron discernir las oportunidades según los territorios de procedencia y clientes particulares. Todos estos análisis exploratorios de los datos son de gran ayuda si se desea desarrollar estrategias para una atención más personalizada de los clientes y las oportunidades.

También debe observarse que existen en los datos un factor (siempre presente) de aleatoriedad. No deben realizarse conclusiones con demasiada autoridad de las observaciones aquí presentadas, ya que la verdadera relación de los resultados puede no verse reflejada en ellas. No debe asumirse directamente, por ejemplo, que un vendedor sea malo en su rol sin al menos contemplar otras aristas del problema, incluso aquellas que escapan de los datos registrados. Las herramientas de la estadística deben ser utilizadas siempre con un debido conocimiento del dominio de los datos para poder enriquecer sus resultados.