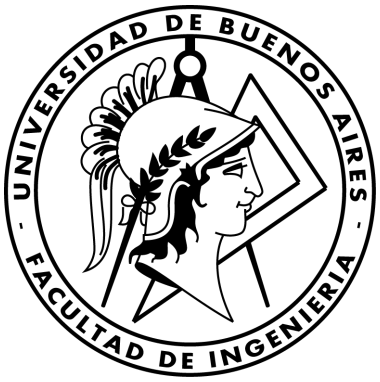

Universidad de Buenos Aires

Facultad de Ingeniería

2do Cuatrimestre de 2020

[75.06/95.58] Organización de datos

Curso 1

Trabajo práctico 1

Análisis Exploratorio

Padrón	Alumno	Email
—	—	--
102914	More, Agustín	amore@fi.uba.ar
—	—	--

Índice

1. Introducción

El trabajo práctico consiste en analizar los datos provistos por la cátedra sobre una empresa ficticia 'Frío Frío'. Los datos representan oportunidades de venta. El análisis se debe hacer con el fin de poder obtener *insights* que permitan predecir la probabilidad de éxito de la oportunidad.

2. Descripción de los datos

El *dataset* provisto especifica para cada oportunidad los sieguientes campos:

- `ID`: id único del registro (Entero).
- `Región`: región de la oportunidad (Categórica).
- `Territory`: territorio comercial de la oportunidad (Categórica).
- `Pricing, Delivery_Terms_Quote_Approval`: variable que denomina si la oportunidad necesita aprobación especial de su precio total y los términos de la entrega (Binaria).
- `Pricing, Delivery_Terms_Approved`: variable que denomina si la oportunidad obtuvo aprobación especial de su precio total y los términos de la entrega (Binaria).
- `Bureaucratic_Code_0_Approval`: variable que denomina si la oportunidad necesita el código burocrático 0 (Binaria).
- `Bureaucratic_Code_0_Approved`: variable que denomina si la oportunidad obtuvo el código burocrático 0 (Binaria).
- `Submitted_for_Approval`: variable que denomina si fue entregada la oportunidad para la aprobación (Binaria).
- `Bureaucratic_Code`: códigos burocráticos que obtuvo la oportunidad (Categórica).
- `Account_Created_Date`: fecha de creación de la cuenta del cliente (Datetime).
- `Source`: fuente de creación de la oportunidad (Categórica).
- `Billing_Country`: país donde se emite la factura (Categórica).
- `Account_Name`: nombre de la cuenta del cliente (Categórica).
- `Opportunity_Name`: nombre de la oportunidad (Categórica).
- `Opportunity_ID`: id de la oportunidad (Entero).
- `Sales_Contract_No`: número de contrato (Entero).
- `Account_Owner`: vendedor del equipo comercial responsable de la cuenta cliente (Categórica).
- `Opportunity_Owner`: vendedor del equipo comercial responsable de la oportunidad comercial (Categórica).
- `Account_Type`: tipo de cuenta cliente (Categórica).
- `Opportunity_Type`: tipo de oportunidad (Categórica).
- `Quote_Type`: tipo de presupuesto (Categórica).
- `Delivery_Terms`: términos de entrega (Categórica).
- `Opportunity_Created_Date`: fecha de creación de la oportunidad comercial (Datetime).

- `Brand`: marca del producto (Categórica).
- `Product_Type`: tipo de producto (Categórica).
- `Size`: tamaño del producto (Categórica).
- `Product_Category_B`: categoría 'B' del producto (Categórica).
- `Price`: precio (Decimal).
- `Currency`: moneda (Categórica).
- `Last_Activity`: fecha de la última actividad (Datetime).
- `Quote_Expiry_Date`: fecha de vencimiento del presupuesto (Datetime).
- `Last_Modified_Date`: fecha de ultima modificación en la oportunidad (Datetime).
- `Last_Modified_By`: usuario responsable de la última modificación en la oportunidad (Categórica).
- `Product_Family`: familia de producto (Categórica).
- `Product_Name`: nombre del producto (Categórica).
- `ASP_Currency`: moneda del precio promedio (Categórica).
- `ASP`: (Average Selling Price) precio promedio a la venta (Decimal).
- `ASP_(converted)_Currency`: moneda del precio promedio convertido en la variable (Categórica)
- `ASP_(converted)`: precio promedio a la venta convertido a otra moneda (Decimal).
- `Planned_Delivery_Start_Date`: límite inferior del rango previsto para la fecha de entrega (Datetime).
- `Planned_Delivery_End_Date`: límite superior del rango previsto para la fecha de entrega (Datetime).
- `Month`: mes-año de `Planned_Delivery_Start_Date` (Fecha).
- `Delivery_Quarter`: trimestre de `Planned_Delivery_Start_Date` (Categorica).
- `Delivery_Year`: año de `Planned_Delivery_Start_Date` (Fecha).
- `Actual_Delivery_Date`: fecha real de la entrega (Datetime).
- `Total_Power`: potencia del producto (Entero).
- `Total_Amount_Currency`: moneda del monto total (Decimal).
- `Total_Amount`: monto total (Decimal).
- `Total_Taxable_Amount_Currency`: moneda del monto gravado total (Categórica).
- `Total_Taxable_Amount`: monto gravado total (Decimal).
- `Stage`: variable target. Estado de la oportunidad (Categórica).
- `Prod_Category_A`: categoría 'A' del producto (Categórica).
- `TRF`: Toneladas de refrigeración (Entero). Es una unidad de potencia.

2.1. Análisis preliminar

Al realizar un análisis preliminar con la herramienta *Pandas Profiling* [1] se observó que algunas de las columnas no tenían datos, o el dato era el mismo para todos registros. Entre ellos:

- `Prod_Category_A`: Contenía solo la categoría `Prod_Category_A_None`.
- `Actual_Delivery_Date`: Contenía solo el valor `NaT`.
- `Last_Activity`: Contenía solo el valor `NaT`.
- `ASP` y `ASP_Currency`: Se decide no considerar estas columnas porque se cuentan con las columnas `ASP_(converted)` y `ASP_(converted)_Currency` que es el mismo monto pasado a dolares, esto permite unificar los precios y hacerlos comparables sin tener que considerar la moneda de cada territorio.
- `Submitted_for_Approval`: Contenía solo el valor `0`.

2.2. Conversión de Fechas

Para poder manipular más fácilmente los datos de tipo fecha, se convierten de `object` a `datetime` las columnas:

- `Planned_Delivery_Start_Date`
- `Planned_Delivery_End_Date`
- `Account_Created_Date`

3. Stages

El trabajo está focalizado en poder predecir esta variable, en particular cuando esta es '`Closed Won`', es decir una oportunidad ganada. Este campo puede tener varios valores: `Closed Lost`, `Closed Won`, `Negotiation`, `Proposal` y `Qualification`. Haciendo un gráfico en la figura 1 para poner en perspectiva las proporciones de cada una, se observa que la gran mayoría de oportunidades se concentran en los *Stages* '`Closed Won`' y '`Closed Lost`', en cambio los demás *Stages* son significativamente menores, esto se puede deber a algún tipo de truncamiento en los datos o bien que el *dataset* sea una *screenshot* en algún momento donde la mayoría de propuestas fueron ya cerradas.

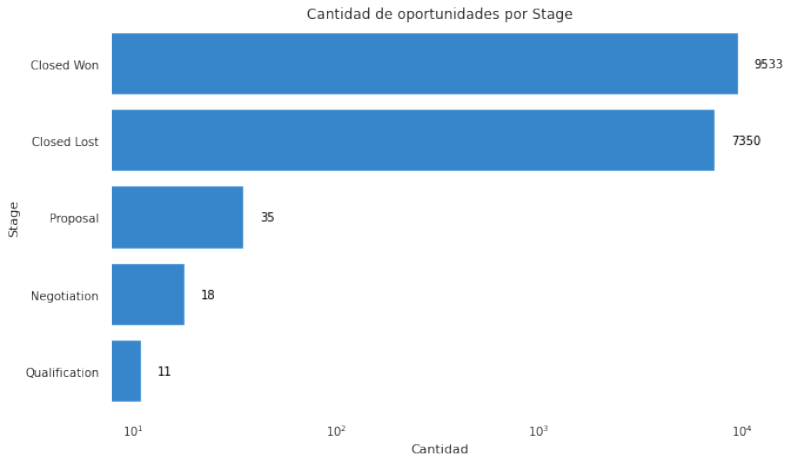

Figura 1: Cantidad de oportunidades por *Stage*

[1] Pandas Profiling https://github.com/pandas-profiling/pandas-profiling

3.1. *Stages* en el tiempo

Además contamos con información temporal de cuándo sucede la oportunidad, pudiendo así ver en el gráfico de la figura 2 cómo fue evolucionando la cantidad a lo largo de los años segmentado por cuatrimestres. Algo para considerar es que la información que se cuanta desde el primer cuatrimestre del 2019 hacia adelante es muy inferior a las años que le preceden. Esto se puede deber que se tiene un subconjunto del *dataset* original, o bien no se cuentan oportunidades posteriores al susodicho cuatrimestre. Es más probable que la primera opción sea más certera, pues los dos años posteriores se cuenta con muy poca información.

Figura 2: Cantidad de oportunidades por año-cuatrimestre.

Una vez visto las cantidades netas, en el gráfico 3 se pueden ver las oportunidades que nos interesa estudiar, que son aquellas que fueron cerradas, exitosamente o no. Era de esperar que el patrón de faltantes de datos del cuatrimestre primero de 2019 hacia delante se mantuviera. Por lo que se ve en el gráfico, solamente en los primeros dos cuatrimestre (que se tienen datos) se obtuvo un *win-ratio* ($\rho_w = \frac{|O_w|}{|O_T|}$) menor al 0.5, en los cuatrimestres posteriores, se fue superando ese número, exceptuando, nuevamente, al cuatrimestre primero de 2019, que para este caso en particular no se saca conclusiones, ya que, por lo mencionado anteriormente, no se está seguro de la completitud de los datos.

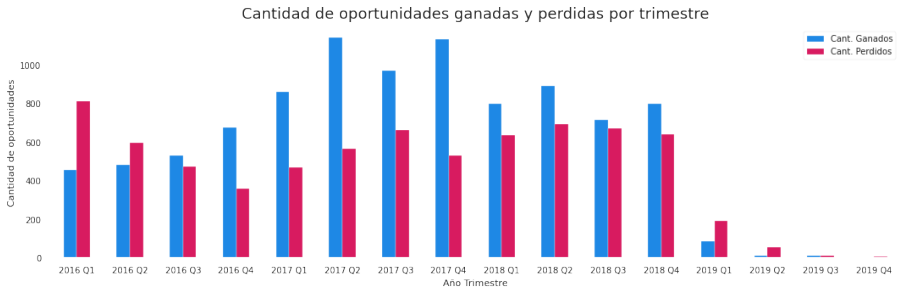

Figura 3: Cantidad de oportunidades ganadas y perdidas por año-cuatrimestre.

3.2. *Stages* por región

Para hilar más fino sobre los datos, en la figura 4 analiza la variable de cantidad de oportunidades ganadas y perdidas, segmentado por región. *Japón* cuenta con mayor *win-ratio*, es decir, las oportunidades en esa región suelen ser más exitosas comparadas con otras regiones. En contraparte, en la región Americas, se tiene el efecto inverso, aunque en menor medida. Finalmente, la región *Medio Oriente*, también cuenta con un bajo *win-rate*, puede llegar a ser apresurado sacar conclusiones en base a este dato, ya que como se ve en relación a los demás, esta región cuenta con menos oportunidades que las demás.

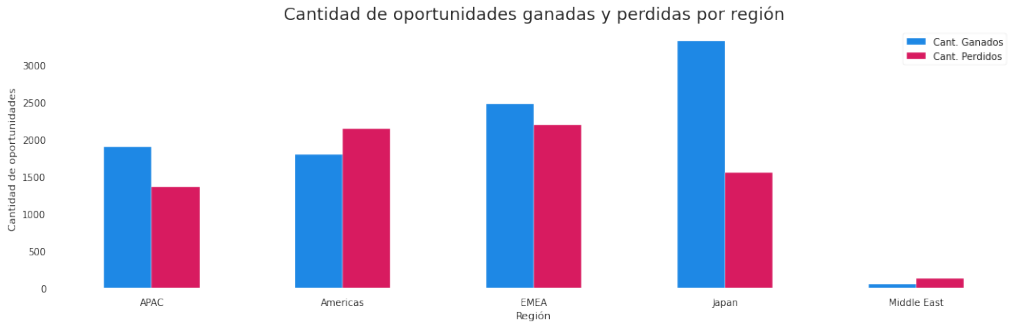

Figura 4: Cantidad de oportunidades ganadas y perdidas por región.

4. Tiempos de entrega

En base a la información provista `Planned_Delivery_Start_Date` y `Planned_Delivery_End_Date`, se calcula el tiempo (en días) que duraría la entrega. Al explorar esta nueva variable, se pudo notar una anormalidad, una de las oportunidades estaba programada con fecha `2208-12-31` que carece de sentido en este contexto.

Haciendo el primer análisis de la duración de las entregas, se calcula el promedio de la duración de envíos (para los casos `Closed Won`), como se muestra en la figura 5, *Japón* tiene el promedio más bajo, es decir realiza entregas más rápido que las demás regiones. Explorando los datos, se observó que muchas de las entregas de *Japón* están programadas para ser entregadas en el mismo día, es decir que `Planned_Delivery_Start_Date` es igual a `Planned_Delivery_End_Date`, esto puede ser efectivamente así, o puede ser un error de carga. Al no poder confirmarlo, se asumirá que esta información es correcta.

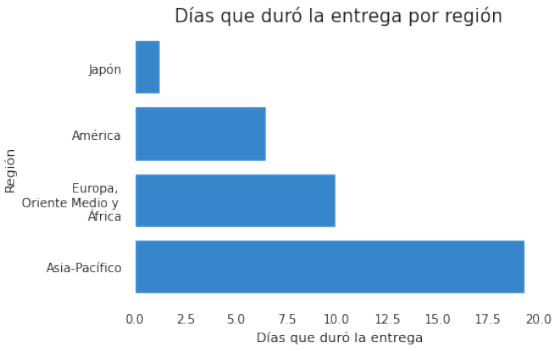

Figura 5: Tiempo promedio de entrega por región

A su vez, cada oportunidad, tiene asignado un término de entrega, este es una categoría y puede representar algún tipo de logística asociada (por ejemplo, un envío internacional o relacionado con las dimensiones del equipo en cuestión). En la figura 6 se muestran el número de oportunidades por término de envío.

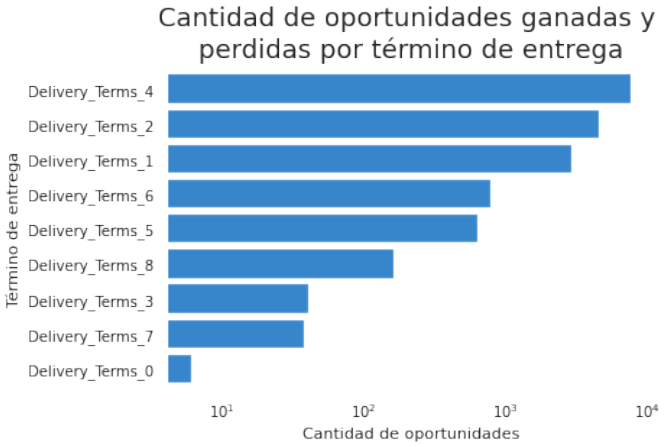

Figura 6: Cantidad de oportunidades por término de entrega.

Como se ve en la figura 7, si bien no aporta información significativa, si se puede ver que `Delivery_Terms_1` es el único con un *win-rate* menor al 0.5, aunque no por mucho. Para evitar obtener resultados que puedan llevar a conclusiones erróneas, se hace el análisis sobre aquellas que tengan por lo menos una cierta cantidad límite de oportunidades, apelando a la ecuación de *de Moivre* [1].

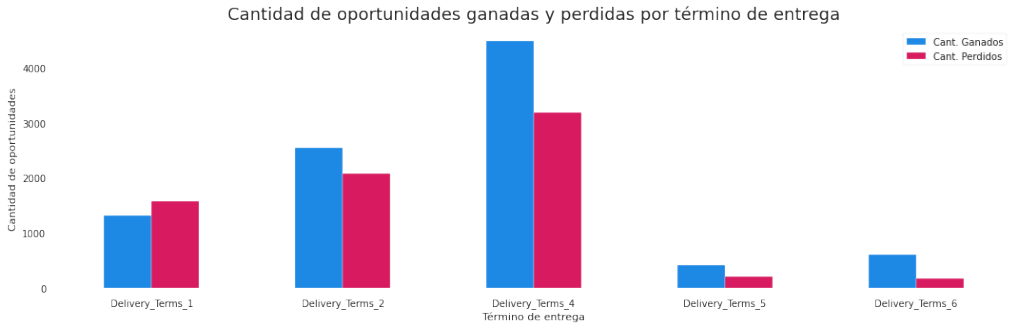

Figura 7: Cantidad de oportunidades ganadas y perdidas por término de entrega.

4.1. Tiempos de entrega a través del tiempo

Una vez obtenido este resultado, se estudia, cómo fue evolucionando este promedio en el tiempo (sería razonable que en un comienzo, por falta de logística, las entregas se demoraran más o bien, a medida que avanza el tiempo, la demanda aumente, retrasando así las entregas). Al ver el gráfico en la figura 8, se observa que tanto en *Medio Oriente* como en *América*, el promedio fue aumentando, apoyando la hipótesis del aumento en la demanda, en cambio en *Asia-Pacífico* disminuyó en el último tiempo. Por su parte, tanto *Europa, Oriente Medio y África* y *Japón* se mantuvo aparentemente constante a lo largo del tiempo. En particular, no se logra apreciar la evolución de estos últimos por el desbalance que generan las demás regiones, con lo cual, se realiza el análisis por separado en el gráfico de la figura 9

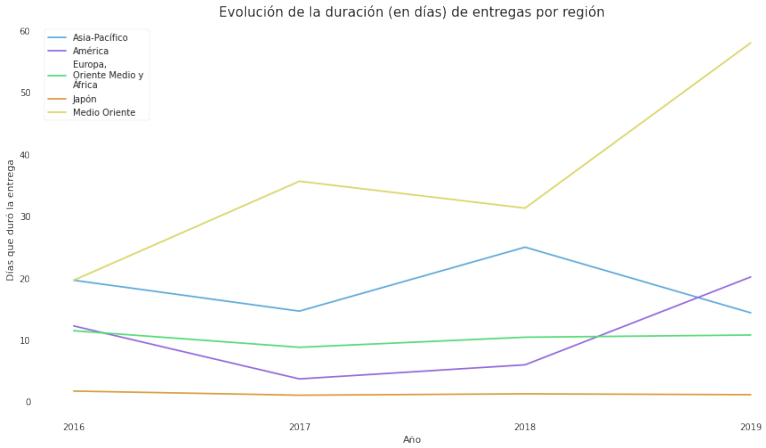

Figura 8: Tiempo promedio de entrega por región a través de los años.

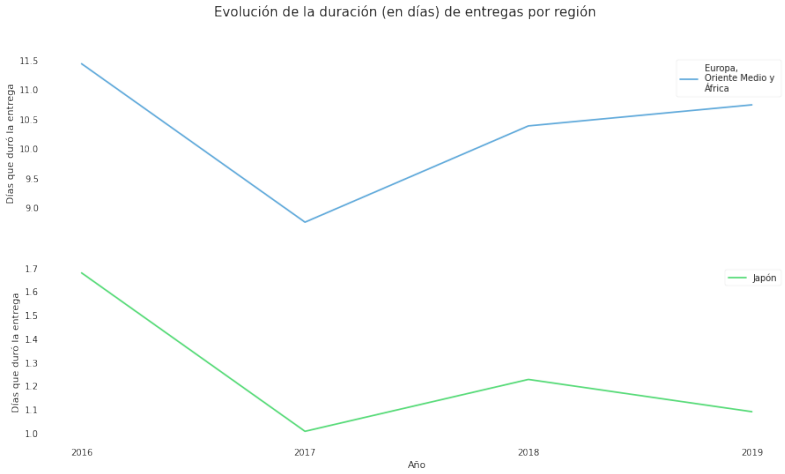

Figura 9: Tiempo promedio de entrega en *Japón* y *EMEA* a través de los años.

4.2. Causas del tiempo de entrega

Una de las razonas de la demora en la entrega puede deberse, por ejemplo, al tamaño de los equipos, para estimar esto se emplea la referencia del dato `TRF` (Toneladas de refrigeración), donde a mayor `TRF` se trata de un equipo de mayor tamaño. A su vez, esto puede estar relacionado con la probabilidad de éxito de la oportunidad, ya que en primer lugar, es difícil vender un equipo grande (con alto `TRF`), es probable que el cliente cancele una oportunidad si considera que el tiempo de entrega no le es útil para su caso de uso, y finalmente una combinación de ambas. En la figura 10, si bien, se puede ver que la mayoría de las oportunidades excediendo cierto valor de `TRF` y un valor de duración del envío, son oportunidades que fueron perdidas, esto no necesariamente es algo concluyente, ya que la mayoría del total de oportunidades se concentra en el rango opuesto.

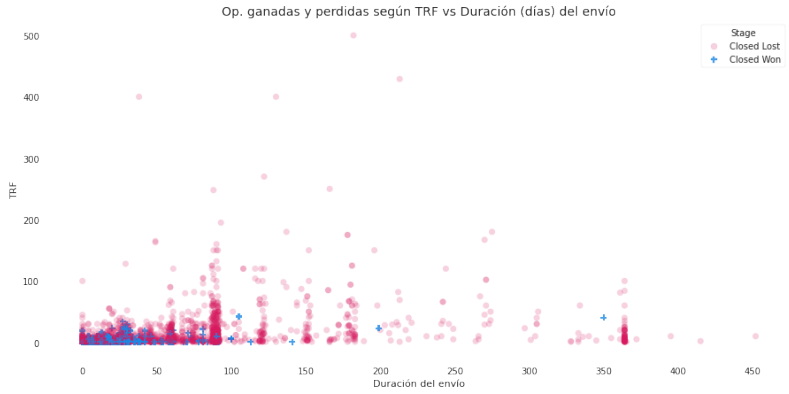

Figura 10: Tiempo de entrega contra el TRF del equipo.

Partiendo de este gráfico, se realizan los gráficos de la figura 11, realizando el mismo análisis pero sobre cada región, agregando un estudio local sobre un rango más acotado para ver los datos que no entran en el análisis anterior.

Figura 11: Tiempo de entrega contra el TRF del equipo por región.

En base al grafico de la figura 11 no se saca ninguna conclusión de la región *Medio Oriente*, pues hay pocos datos. En cuanto a *APAC* y *Americas* son las regiones donde se realizan los pedidos de equipos con mayor `TRF`, que se ve que en general, se cumple con la hipótesis de a mayor `TRF` (mayor el equipo) es más probable que la oportunidad no sea ganada. En cuanto a el tiempo de entrega,

se puede notar un quiebre, alrededor de de los 100 días, donde a partir de ahí son oportunidades perdidas.

Algo para notar, que se ve mejor en las regiones *EMEA* y *Americas* (en el gráfico aumentado), es que aproximadamente hasta los 35-40 días de entregas, se ven oportunidades exitosas, aún teniendo un `TRF` alto, desde ahí, esa tendencia tiende a invertirse.

En la figura 12, se reafirma que desde este parámetro, duración del envío, nos puede ayudar a descartar aquellos que superen cierto límite. Es notable ver, por ejemplo, en *APAC* que a partir de los 50 días, son mayoritariamente ventas perdidas. En contraposición, no es clara ninguna conclusión sobre oportunidades que estén abajo de ese límite.

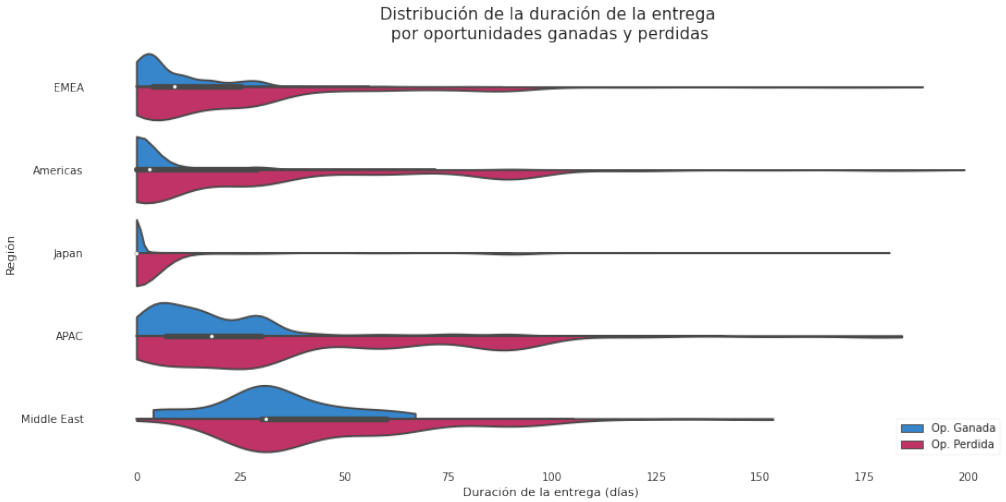

Figura 12: Distribución tiempo de entrega por región.

5. Tipo de familia de productos

Una de las características de los productos, es que cuenta con un 'tipo de familia'. Se desea hacer el análisis para ver si alguna de las familias cuanta con un mayor *win-ratio* respecto a las demás. En la figura 13 se muestran todas las familias (que superen un número de oportunidades) ordenadas según su *win-ratio*, está claro que las familias que se encuentran más arriba son aquellas con mayor ρ_w.

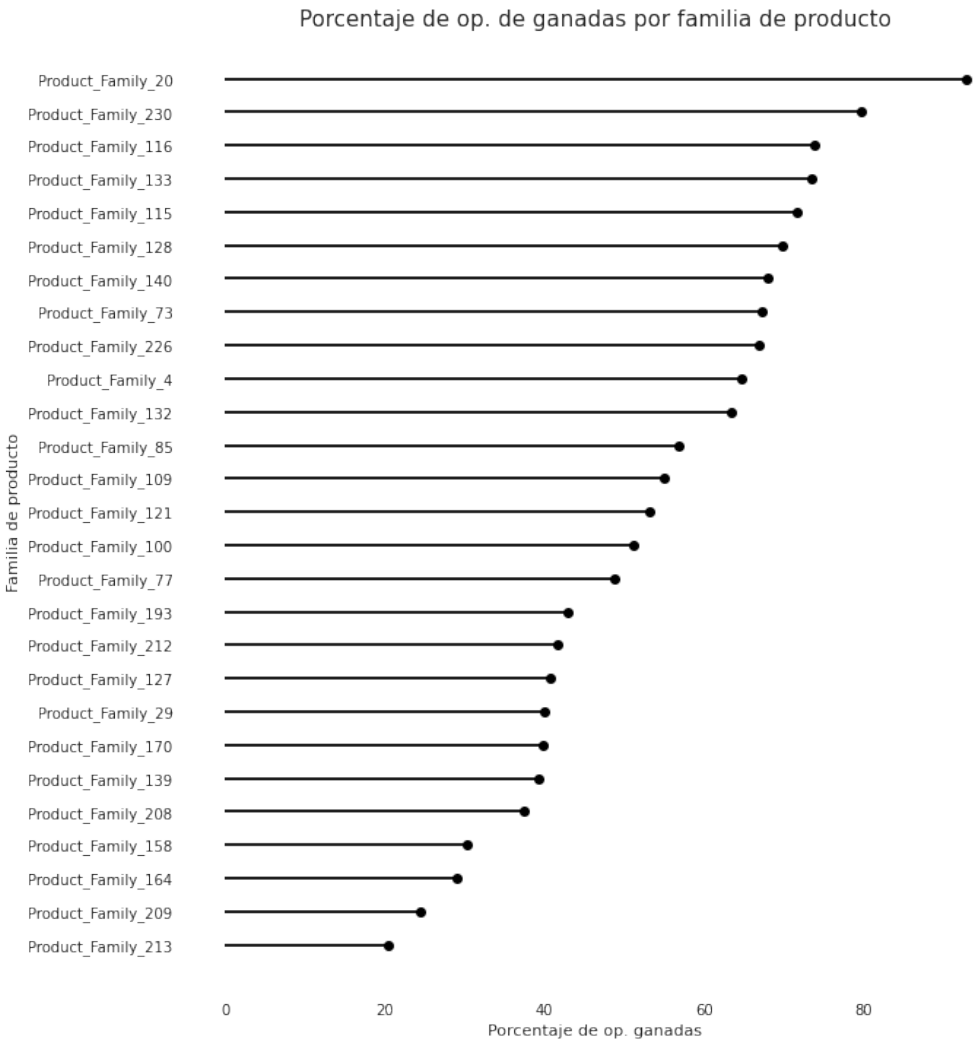

Figura 13: Familias de producto ordenadas por *win-ratio*.

Una vez obtenido este resultado general, en la figura 14 se hace foco sobre las primeras, viendo así que, no necesariamente, las que más oportunidades tienen, les va mejor. Si por el contrario, vemos en la figura 15 las familias con peor ratio. Desde acá se puede postular que dependiendo el tipo de familia de producto (y lo que ello signifique dentro de la denominación del producto) la probabilidad de éxito puede verse modificada por esta variable.

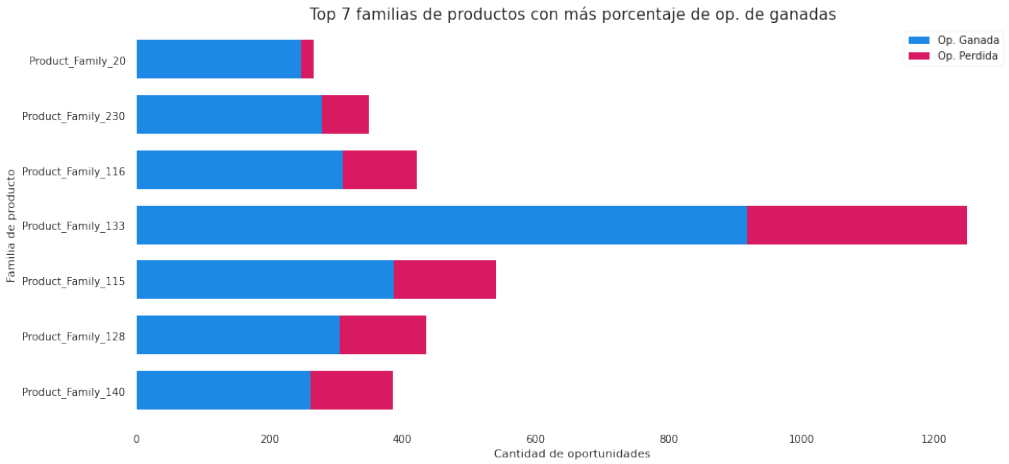

Figura 14: Top 7 familias de producto con mayor *win-ratio*.

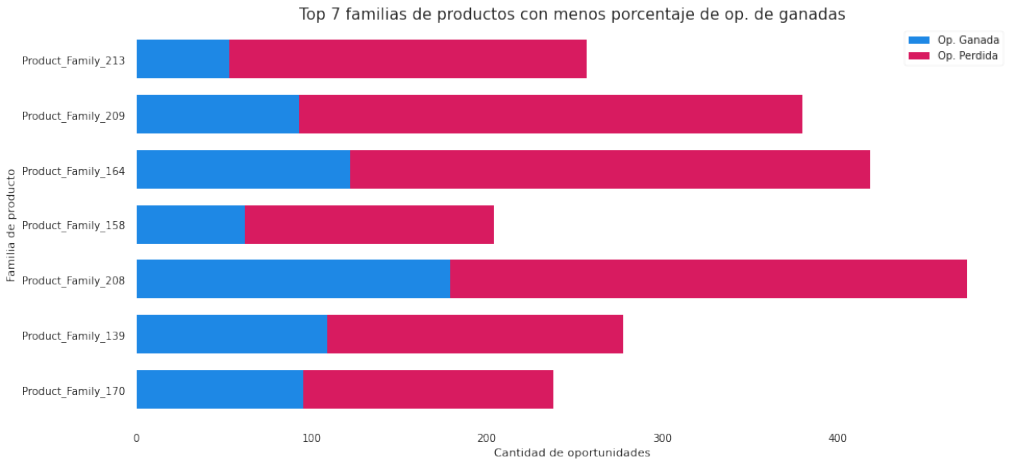

Figura 15: Top 7 familias de producto con menor *win-ratio*.

6. Personas

Cada oportunidad tiene asociada un *dueño de oportunidad* y *dueño de cuenta*, estos conjuntos no son necesariamente excluyente. Haciendo un análisis sobre estas dos variables, en la figura 16 vemos que hay un empleado predominantes, `Person_name_50`, con un *win-ratio* (ρ_w) superior al 0.5, seguido, se encuentran `Person_name_13`, `Person_name_8` y `Person_name_18`. Estas primeras cuatro personas, parecería que trabajan mayoritariamente solas en cada oportunidad, hay poca interacción con las demás, un excepción a esto es la persona `Person_name_43`, que cierto porcentaje de oportunidades ganadas fueron interactuando con la `Person_name_19`. Es notable destacar que la persona `Person_name_3` (dentro de las personas con mayor venta), tiene peor *win-ratio*.

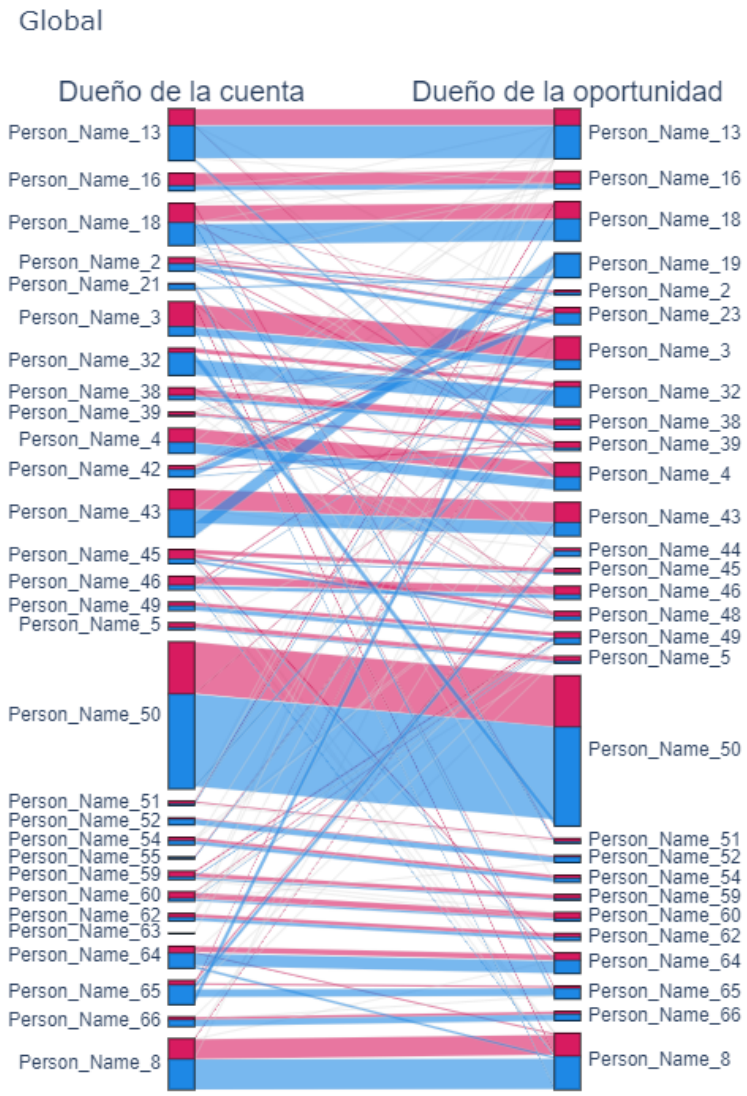

Figura 16: Relaciones entre personas diferenciando oportunidades ganadas y perdidas (global).

6.1. Relaciones de personas por región

Un problema notable del gráfico de la figura 16, es que aquellas personas con menos oportunidades que las principales, se ven opacadas. Por eso se hace un análisis segmentado por región.

En la figura 17 se muestra la cantidad de *dueños de oportunidad* que hay por cada región.

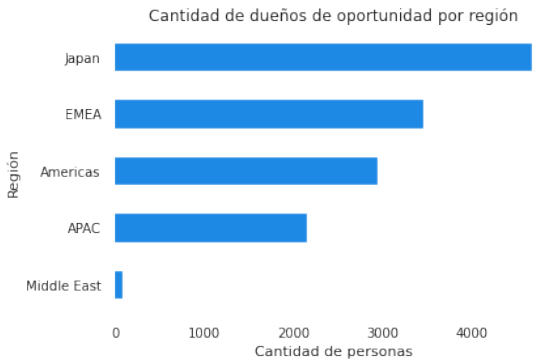

Figura 17: Cantidad de *dueños de oportunidad* por región

Aparentemente los datos en *Medio Oriente* son significativamente inferiores a las demás, con lo cual este análisis no lo va a tener en cuenta.

Una vez tenido un vistazo global, en la figura 18, se analiza cómo se distribuyen estos *dueños de oportunidad* dentro de cada una de las regiones según su cantidad de ventas. Aunque se observa una gran varianza en *Americas*, concentra el mayor promedio por persona.

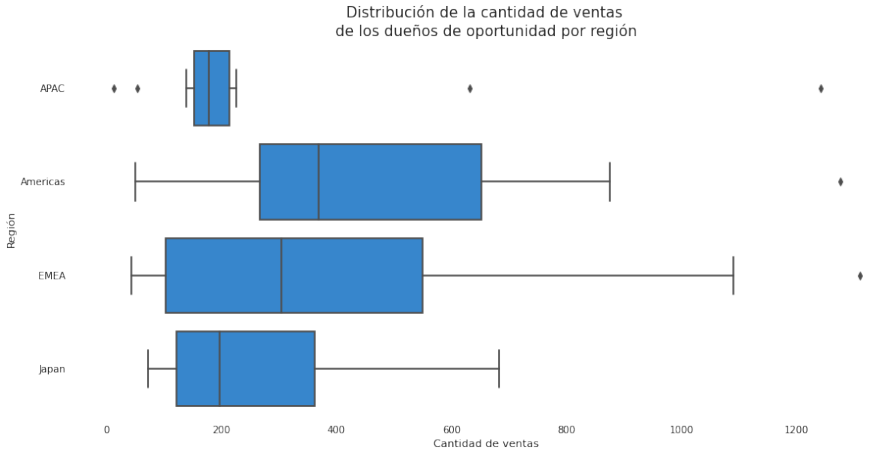

Figura 18: Distribución de las personas según su *win-ratio* por región.

En base al gráfico de la figura 18, ahora se quiere conocer cómo se distribuyen estas personas, en particular, cómo se distribuyen según su *win-ratio*, obteniendo así el gráfico de la figura 19. Se observa que las regiones con mayor *win-ratio* se concentran en *Japón* y *APAC*, mientras que en *EMEA* y Americas se encuentran un poco por debajo de las mencionadas. Nuevamente, por la falta de datos en *Middle East*, es difícil tomar conclusiones en dicha región. Algo para notar, en la región de *Japón*, no se cuenta con información significante de dueños de oportunidad con *win-ratio* bajo.

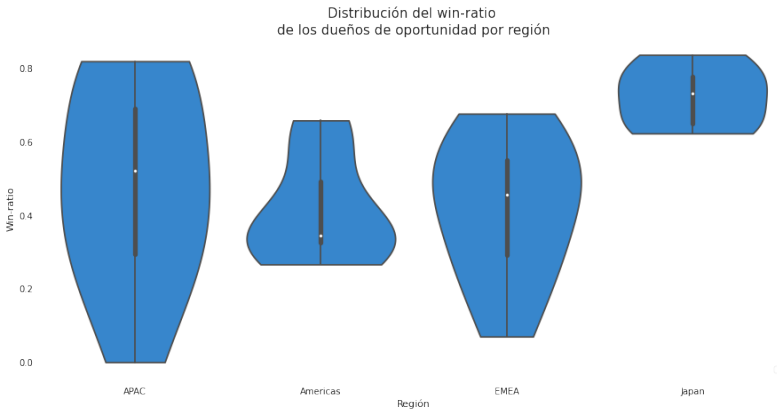

Figura 19: Distribución de las personas según su *win-ratio* por región.

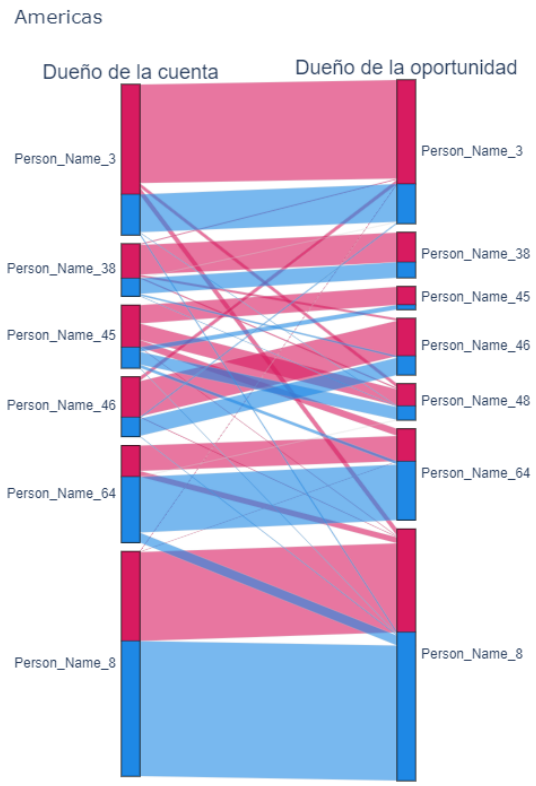

Figura 20: Relaciones entre personas diferenciando oportunidades ganadas y perdidas (Americas).

En el gráfico de *Americas* de la figura 20, se ve que las personas con mayores ventas son `Person_name_8` (uno de los principales globales), `Person_name_64` y `Person_name_3`, este último con un bajo *win-ratio*. Las demás personas en esta región no superan el *win-ratio* de 0.5.

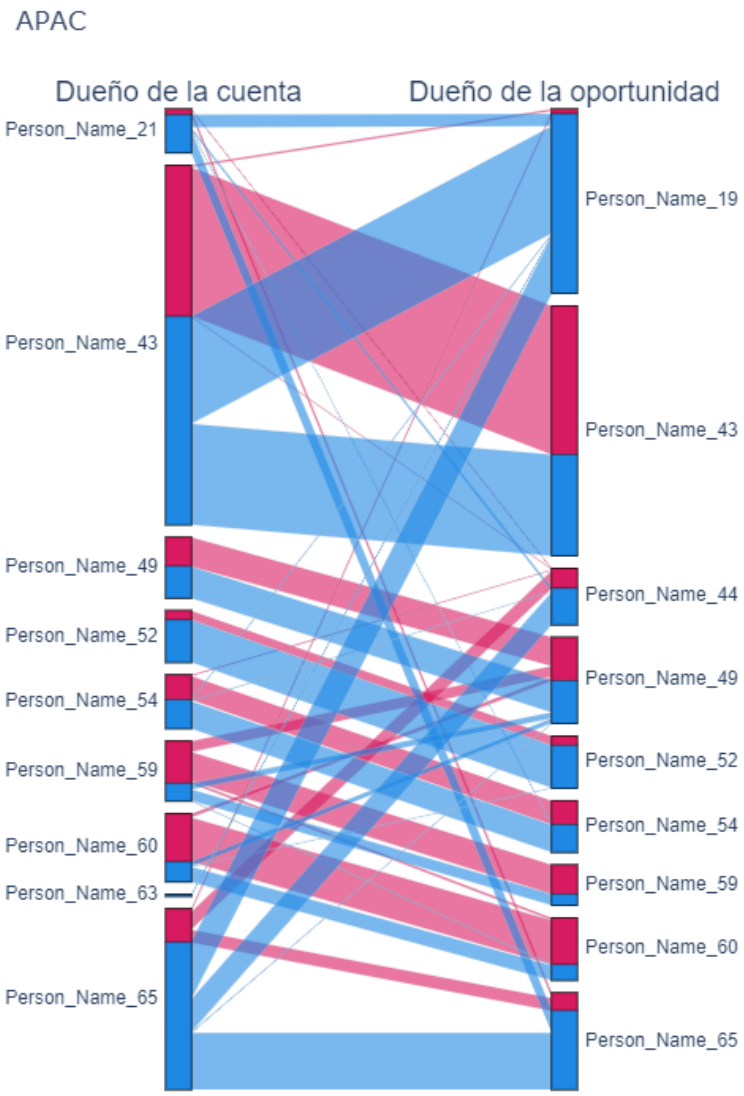

Figura 21: Relaciones entre personas diferenciando oportunidades ganadas y perdidas (APAC).

Para la región *APAC* (figura 21), se ve un *win-ratio* mayor en promedio y además se puede ver como las distintas personas tienden a cooperar más entre ellos. Esto último no es el causante del primero, ya que se puede ver que por ejemplo la persona `Person_name_59` tiene un alto grado de cooperación pero aún así un bajo *win-ratio*.

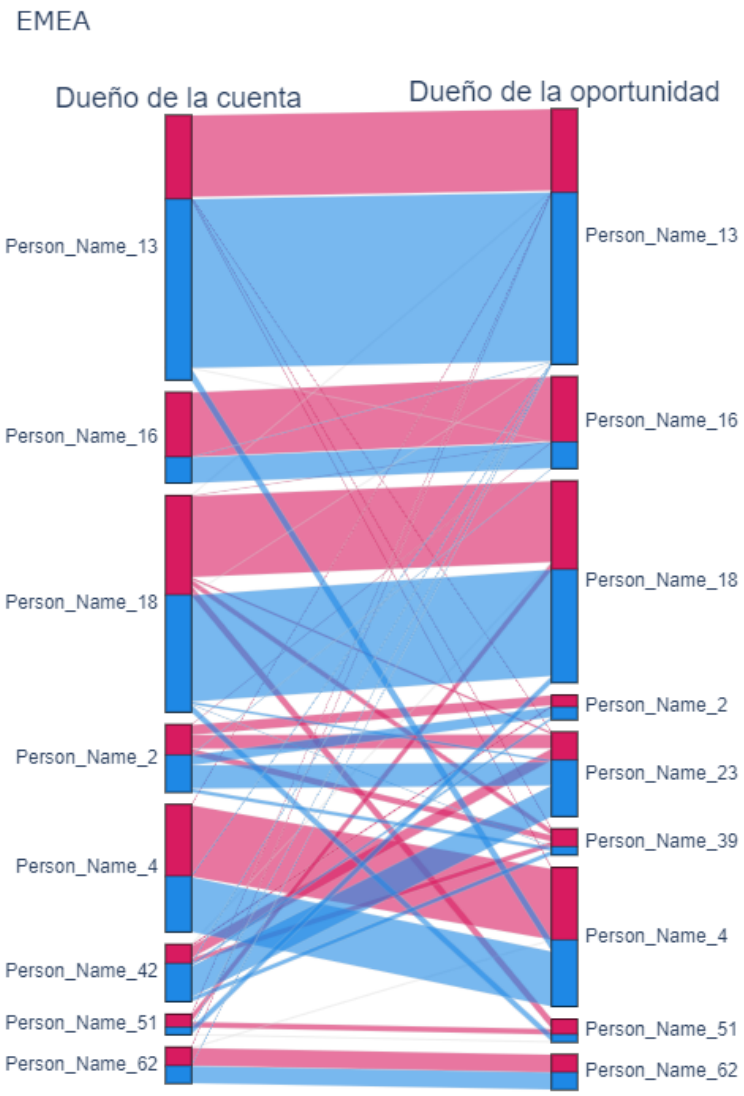

Figura 22: Relaciones entre personas diferenciando oportunidades ganadas y perdidas (EMEA).

En la región de *EMEA* (ver figura 22), se encuentra dos de las personas con mayor cantidad de ventas mundial, `Person_name_13` y `Person_name_18`. En esta región en particular, se nota que hay personas con pocas ventas asociadas, pero que intervinieron con muchas personas, y esto sucede tanto del lado *dueño de cuenta* como *dueño de oportunidad*.

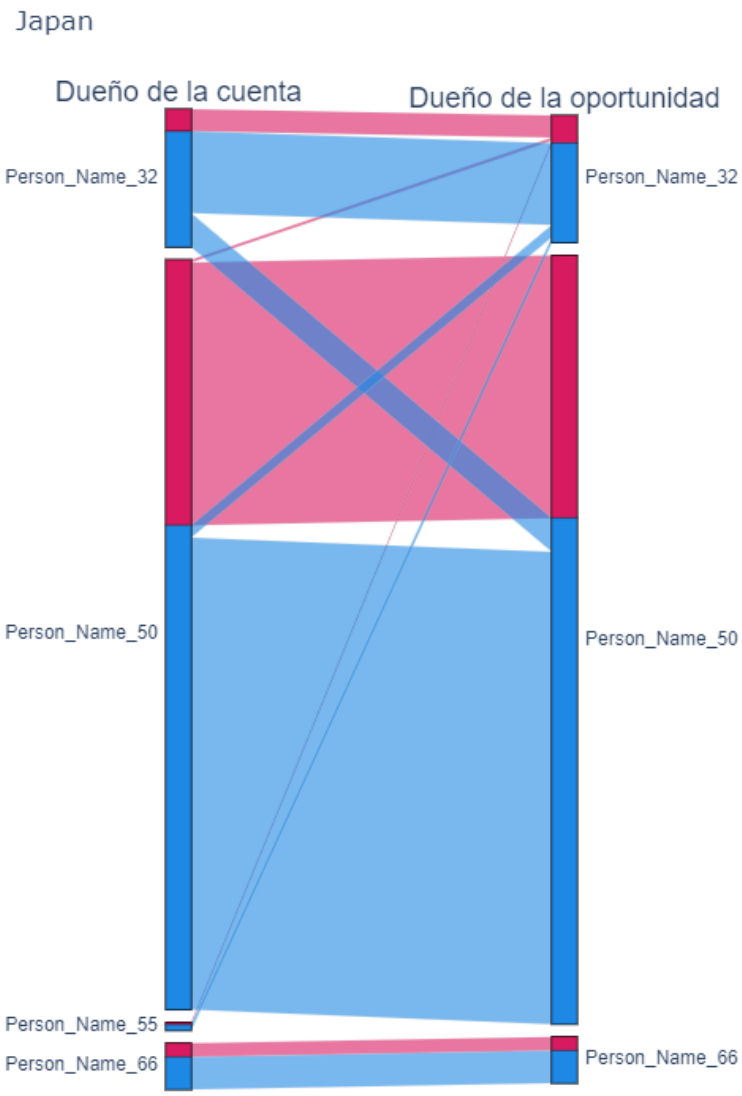

Figura 23: Relaciones entre personas diferenciando oportunidades ganadas y perdidas (Japón).

En la región de *Japón* (figura 23), se cuentan con menor cantidad de personas con ventas significativas, con lo cual es un tanto peligroso sacar conclusiones, pero algo a destacar es que tiene presente a la persona con mayores ventas asociadas mundiales, `Person_name_50`, que además cuenta con un alto *win-ratio*. Esto último es algo que se mantiene dentro de las personas que están siendo analizadas en esta región.

Finalmente, pera la región *Medio Oriente*, no se cuenta con información suficiente para realizar algún análisis de este tipo.

7. Monto Total respecto TRF

Una oportunidad puede estar formada por uno o más items, cada uno tiene asociado un *monto total*, una cantidad de *toneladas de refrigeración (TRF)*, entre otras propiedades. Haciendo el gráfico del primero en función del segundo y diferenciando cada región por colores queda como resultado la figura 24. Vemos que los valores están alineados en rectas, de las cuales en la de mayor pendiente predomina los registros pertenecientes a Japón. Esto se ve más claro en la siguiente gráfica (figura 25).

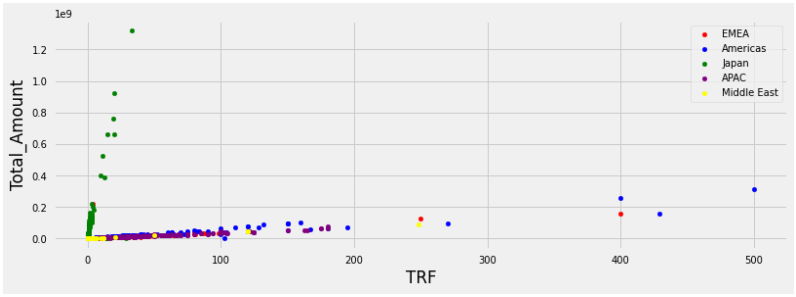

Figura 24: Relación entre monto total por item en función de TRF por cada región.

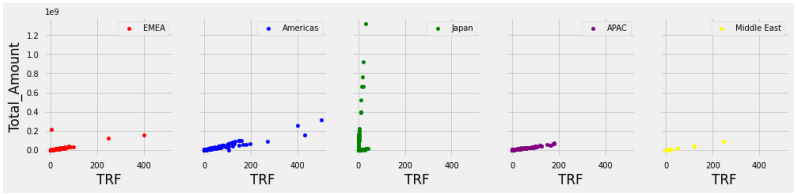

Figura 25: Relación entre monto total por item en función de TRF por cada región por separado.

Basados en esto no es difícil concluir que las diferentes rectas se deben a que están cotizados en diferentes monedas. Procedemos a repetir el gráfico, esta vez distinguiendo cada moneda por color en la figura 26. Los yenes, la moneda oficial de Japón, tiene un valor muy inferior respecto al dolar y al euro (ambos bastante parejos) por lo cual su recta tiene una pendiente mucho más alta que los otros.

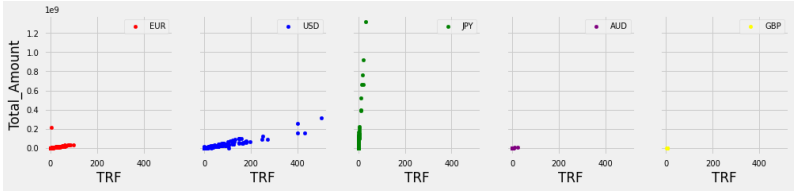

Figura 26: Relación entre monto total por item en función de TRF por cada moneda por separado.

8. Conclusiones

A medida que se fue analizando el *dataset*, se fue encontrando que ciertas categorías no se encontraban equitativamente o distribuidas racionalmente proporcional, por ejemplo, cuando se analizó la cantidad de oportunidades por años en la sección 2.2 a partir del año 2019 se ve un desbalance de información frente a los años anteriores, o bien la falta de información en la región Medio oriente como se ve en la sección 3.2. Esto se puede deber, como se explico en sus

correspondientes secciones, por un posible filtro previo en los datos. Quitando lo previamente dicho, se ve como, algunas región suelen tener en promedio más chances de éxito que otras, por ejemplo, en el gráfico 4 se observa como en *Japón*, el *win-ratio* es superior al del resto, en cambio, en el mismo gráfico, la región *América* se cuenta con más oportunidades perdidas que ganadas.

Otro *insight* interesante es lo que se explaya en la sección 4, sobre el agregado de una variable que no se tiene explícitamente en el *dataset* original, que es el tiempo que se estima que durará el envío. De este análisis, la conclusión más fuerte que se consiguió es que es muy raro que una oportunidad con un tiempo de envío muy extenso, es muy difícil que sea exitoso. El tiempo de entrega, tal como se lo explica en la sección se lo lleva aparejado a características del producto y además en el gráfico 7 se estudia la relación de la probabilidad de éxito con el término de entrega y se hipotetizan posibles causas de estos resultados.

En cuanto a características intrínsecas a los productos, no se pudo obtener una conclusión importante, ya que gran parte de los datos relacionados a este (categorías, tamaño, marcas, etc.), se encontraban principalmente con valores nulos (`None`, `NaN`, etc.). El único parámetro que sí se estudió fue a la familia que cada producto pertenece. En la figura 13 se puede ver que hay familias que cuentan con mayor probabilidad de éxito (figura 14), mientras que otras, la mayoría de oportunidades son oportunidades perdidas (figura 15). Se plantea en la sección 5 que esto se puede deber a características asociadas a los productos de cada familia, al no contar con mayor información de los productos, no se cuenta con evidencia para probar o refutar las hipótesis postuladas.

Finalmente, en la sección 6 se plantean las relaciones entre personas que intervienen en oportunidades, en el gráfico 16 y los subsiguientes, se pueden apreciar estas relaciones y además el *win-ratio* de los distintos intervinientes, en particular se puede notar que hay personas que cuentan con una mayor cantidad de oportunidades ganadas, en algunos casos muy superiores. Luego, sobre cada región se notan distintos patrones de comportamientos, en algunos, las interacciones suelen ser pluripersonales, mientras que en otras, se cuentan con personas con mayoría de interacciones consigo mismas. Esto puede tener que ver con condiciones culturales o bien, con la demografía de la empresa en las distintas regiones (sucursales con muchas o pocas personas, por ejemplo). Según la región, el win-ratio dependía de este tipo de comportamiento, pero no es un resultado global que se pueda extrapolar. Algo que se puede notar en el gráfico 19 es la 'calidad' de los empleados (estimada con el *win-ratio*), varía significativamente según la región, por ejemplo en América hay un acumulativo de personas con ρ_w de 0.5 para abajo, mientras que en el resto de las regiones la tendencia es que haya personas con un win-ratio mayor a 0.5 en general.

8.1. Oportunidades de mejoras

A lo largo del trabajo práctico fueron surgiendo inconvenientes, ya sea falta de tiempo o de conocimiento, generaron posibles oportunidades de mejoras.

8.1.1. Entendimiento mayor del modelo del negocio

Al contar con información un tanto críptica, la cual era difícil de asociar algo tangible, fue difícil interpretar los datos, desaprovechando así posibles variables importantes para la predicción de oportunidad de éxito. La mejora de la calidad del análisis se podría lograr tratando de obtener más reglas de negocios que están asociadas con los datos provistos.

8.1.2. Reducción de consumo de memoria

Cuando se empezó a analizar los datos desde una primera vista, se pudo fue viendo qué variables eran categóricas y cuales no. Las variables categóricas nos permiten ahorrar espacio en memoria que ocupa el *dataframe* al reducir la información duplicada. Los atributos que fueron pensados como categóricos fueron:

- ASP_(converted)_Currency
- Stage
- Region
- Territory
- Pricing, Delivery_Terms_Quote_Appr
- Pricing, Delivery_Terms_Approved
- Bureaucratic_Code_0_Approval
- Bureaucratic_Code_0_Approved
- Bureaucratic_Code
- Source
- Billing_Country
- Account_Owner
- Opportunity_Owner
- Account_Type
- Opportunity_Type
- Quote_Type
- Delivery_Terms
- Brand
- Product_Type
- Size
- Product_Category_B
- Currency
- Last_Modified_By
- Product_Family
- Product_Name
- Month
- Delivery_Quarter
- Total_Amount_Currency
- Total_Taxable_Amount_Currency

Que originalmente ocupaba algo así como 5.9 MB.

```
dtypes: float64(3), int64(8), object(35)
memory usage: 5.9+ MB
```

Y pasó a ocupar menos de la mitad:

```
dtypes: category(29), datetime64[ns](3), float64(3), int64(4), object(8)
memory usage: 2.9+ MB
```

Esto se logró pasando estos atributos de `object` a `pd.Categorical`. El inconveniente con esta modificación fue que generó dificultades para poder graficar, con lo cual se descartó este cambio para simplificar la codificación.

8.1.3. Mejoras de código

Para realizar cada uno de los gráficos, se fueron tomando bloques independientes en una *notebook* de *Google Colab* [2], lo que hizo que haya mucho código desordenado y repetido.

[2]Google Colab https://colab.research.google.com/

A. Ejecución del script que genera los gráficos

Para poder correr el script que genera el análisis y los gráficos, primero hay que instalar las librerías necesarias, se puede hacer mediante la herramienta `pip` [3] con el siguiente comando:

```
pip install -r requirements.txt
```

Para realizar realizar una instalación más ordenada y sin tener conflicto de versiones es recomendable usar un entorno virtual [4] Una vez finalizada la instalación, para ejecutar el programa:

```
python tp_datos_2c2020.py
```

Si se quiere evitar la parte de instalación se puede ejecutar desde la herramienta *Google Colab* [5].

B. Elección de colores

Gran parte del informe, se encuentra con la misma gama cromática, estos colores son (en formato hex) '`#1E88E5`' 🔵, generalmente asignado a oportunidades ganadas o resultados neutros, y '`#D81B60`' 🔴 generalmente usado para denotar oportunidades perdidas. Se hace esta elección, frente a las combinaciones de verde-rojo (que suele representar lo mismo), para poder facilitar el análisis para personas con problemas de daltonismo [2]. Como se puede ver en la figura 27 se presentan dos gráficos con dos paletas de colores distintas, la primera es rojo-verde, la segunda es la propuesta que se usó en el trabajo. En su contraparte, en la figura 28 se encuentra la misma imagen, pasada por un filtro de simulación de problemas de vista [3].

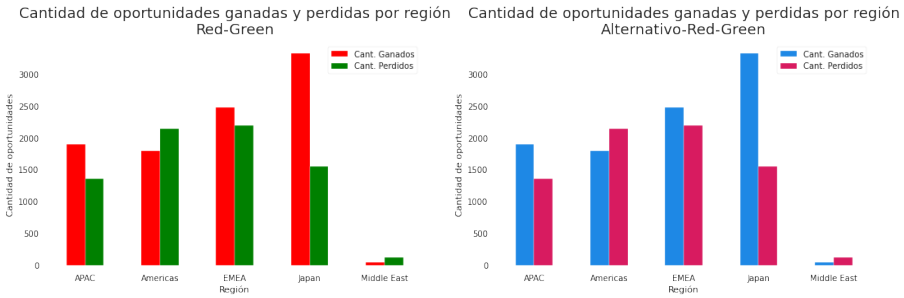

Figura 27: Barplot original.

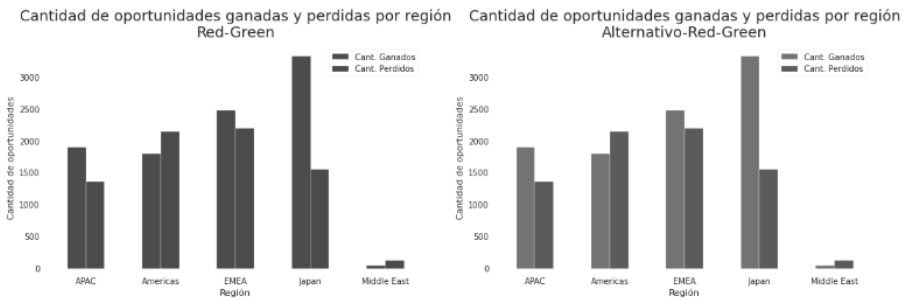

Figura 28: Barplot como lo vería una persona con vista monocromática.

[3] Herramienta **pip**: https://pypi.org/project/pip/
[4] Entorno virtual de python (venv): https://docs.python.org/3/library/venv.html
[5] Google Colab https://colab.research.google.com/

Referencias

[1] *La ecuación más peligrosa [En].* https://www.americanscientist.org/article/the-most-dangerous-equation

[2] *Alternativas de colores para rojo-verde [En].* https://www.visualisingdata.com/2019/08/five-ways-to-design-for-red-green-colour-blindness/

[3] *Simulador de distintos tipos de daltonismo.* https://www.color-blindness.com/coblis-color-blindness-simulator/