Trabajo Práctico 1

Organización de Datos (75.06/95.58)

Segundo Cuatrimestre de 2020

Análisis exploratorio de la empresa "Frío Frío"

GRUPO: Data Learning

INTEGRANTES:

Juan Ibarra Olalla

Natasha Tarapow

Roberto Irahola

Martín Gorsd

Pedro Casa

Repositorio GitHub:

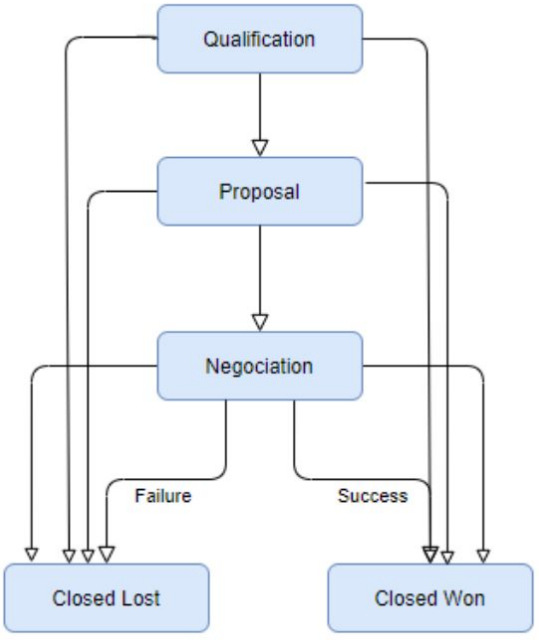
https://github.com/notkex/tp1organizaciondedatos

1. Introducción

Una empresa dedicada a la venta e instalación de equipos de aire acondicionado para grandes superficies, busca optimizar los esfuerzos de los representantes comerciales. Para ello, ha hecho un registro de las oportunidades de ventas a lo largo de su historia.

Una "*oportunidad*" consiste en un proyecto de venta o instalación de equipos para un cliente. El "*pipeline*" hace referencia al flujo de oportunidades prospecto que la empresa está desarrollando. El equipo comercial asigna a distintos momentos, para cada oportunidad, un estado en la negociación. En la siguiente ilustración se muestran los estados que las oportunidades tienen dentro del *pipeline*.

Figura 1: "*Pipeline*" de las oportunidades



El objetivo de este trabajo es realizar una *limpieza de estos datos* y, principalmente, un *análisis exploratorio* para determinar características y variables importantes, descubrir *insights* interesantes, y analizar la estructura de los mismos. Además, buscamos preparar este conjunto de datos para, posteriormente, aplicar técnicas de *machine learning* que nos permitan predecir la probabilidad de que una oportunidad sea exitosa.

El trabajo se estructura de la siguiente manera. En la sección dos, vamos a hacer una breve descripción de la limpieza de los datos. En la sección tres, se lleva a cabo el análisis exploratorio. En este apartado, tendremos en cuenta los siguientes aspectos principales:

distribución temporal de ítems, distinción geográfica, vendedores y relaciones entre las variables *TRF*, *Total Amount* y *ASP*. Por último, en la cuarta sección, se concluye.

2. Limpieza

Los datos proporcionados se estructuran tabularmente. Cada fila se corresponde con un ítem de una determinada oportunidad y, esta última, puede o no contener varios ítems. Por otro lado, las columnas representan características asociadas a cada oportunidad.

Consideramos que aquellas variables que tienen un único valor, no serán útiles para nuestro análisis; incluso si todas nuestras observaciones tienen datos para una determinada variable, si el valor asignado es siempre el mismo, no nos otorga información y, en consecuencia, dichas variables las consideramos no útiles. Este es el caso de las variables "*Submitted_for_Approval*", "*Last_Activity*", "*ASP_(converted)_Currency*", "*Actual_Delivery_Date*" y "*Prod_Category_A*". Por este motivo decidimos eliminarlas.

Por otro lado, observamos que algunas de las variables tienen una proporción importante de celdas sin datos. Específicamente, "*Brand*", "*Product_Type*", "*Size*", "*Price*" "*Product_Category_B*" y "*Currency*" que tienen datos para menos del 8% de los ítems de nuestro *dataframe*.

Tabla 1: Porcentaje de datos por variable

Variable	% de datos	Variable	% de datos
Region	100.00	Opportunity_Name	100.00
Territory	94.98	Sales_Contract_No	58.10
Pricing, Delivery_Terms_Quote_Appr	100.00	Account_Owner	100.00
Pricing, Delivery_Terms_Approved	100.00	Opportunity_Owner	100.00
Bureaucratic_Code_0_Approval	100.00	Account_Type	99.27
Bureaucratic_Code_0_Approved	100.00	Opportunity_Type	100.00
Bureaucratic_Code	100.00	Quote_Type	100.00
Account_Created_Date	100.00	Delivery_Terms	100.00
Source	52.32	Opportunity_Created_Date	100.00
Billing_Country	99.82	Brand	7.13
Account_Name	100.00	Product_Type	6.98

Variable	% de datos	Variable	% de datos
Size	6.77	Planned_Delivery_End_Date	99.48
Product_Category_B	7.02	Month	100.00
Price	2.45	Delivery_Quarter	100.00
Currency	6.32	Delivery_Year	100.00
Quote_Expiry_Date	74.26	TRF	100.00
Last_Modified_Date	100.00	Total_Amount_Currency	100.00
Last_Modified_By	100.00	Total_Amount	100.00
Product_Family	100.00	Total_Taxable_Amount_Currency	100.00
Product_Name	100.00	Total_Taxable_Amount	100.00
ASP(USD)	100.00	Stage	100.00
Planned_Delivery_Start_Date	100.00		

Tener muy pocos datos de las variables mencionadas no nos permite generar conclusiones sólidas pero, a su vez, puede ser peligroso descartarlas. Sin embargo, analizandolas con mayor detenimiento, observamos que, en todos los casos, la información disponible se corresponde exclusivamente con oportunidades perdidas. En consecuencia, decidimos eliminarlas.

Las variables "*Total_Amount*" y "*Total_Taxable_Amount*" son variables numéricas que expresan un determinado valor monetario; observamos que la moneda varía según la oportunidad. Esto no nos permite hacer un análisis comparativo; por este motivo decidimos homogeneizarlas. Para ello, utilizamos datos históricos de la relación entre las divisas[1] y los sincronizamos con la última fecha de modificación de la oportunidad (variable *'Last_Modified_Date'*), momento en el que asumimos se fijó el precio de la venta.

Por su parte, las variables "*Region*" y "*Territory*" tenían errores de imputación; por ejemplo, el territorio "*South America*" estaba imputado como región "*EMEA*", cuando claramente debería ser "*Americas*". Estos errores fueron revisados y corregidos. Por otra parte, observamos que "*Japan*" y "*Middle East*" se toman como "*Region*"; entendemos que

[1] Recuperado de: https://es.investing.com/

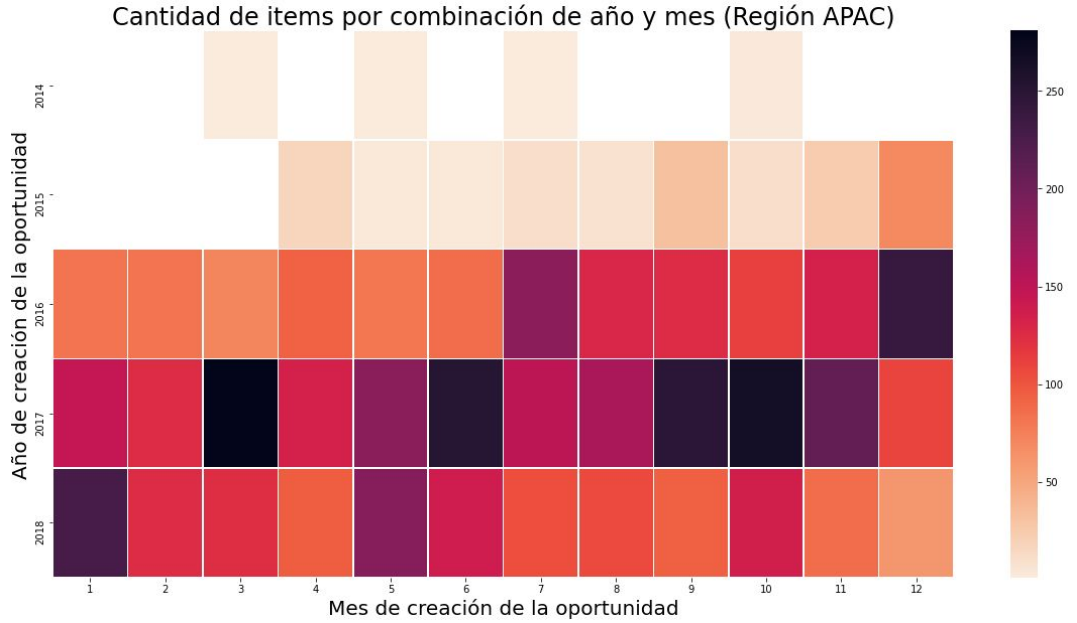
quizás esto no se trate de un error, pero de todas maneras decidimos agrupar estos dos territorios en sus regiones correspondientes ("*APAC*" y "*EMEA*" respectivamente).

3. Análisis

3.1. Distribución temporal de ítems

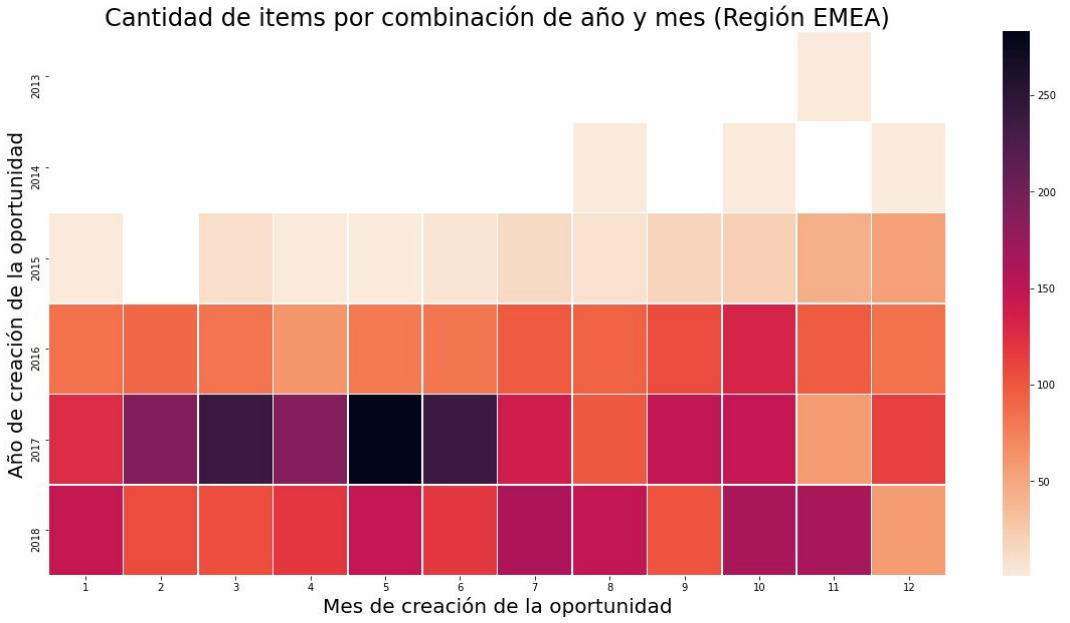
A continuación se observa la distribución temporal de la cantidad de ítems por región, tomando como fecha la creación de cada oportunidad:

Figura 2:



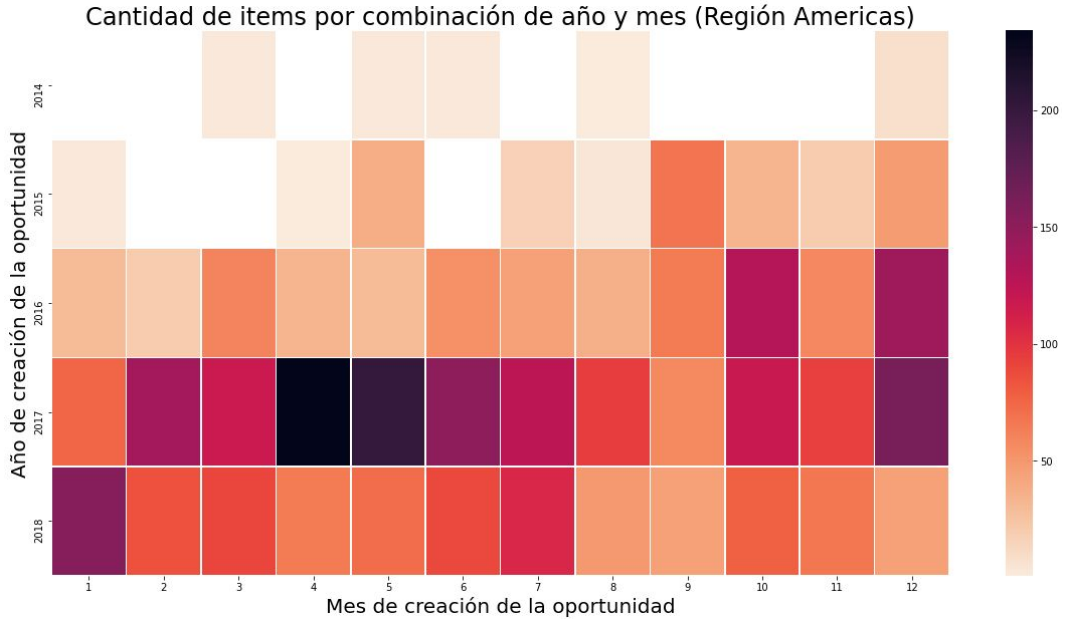
Para la región *APAC* observamos una mayor concentración de oportunidades creadas en los meses de marzo, junio, septiembre, octubre y noviembre del año 2017; sus primeras oportunidades fueron creadas en marzo de 2014.

Figura 3:



Para la región *EMEA* observamos una mayor concentración de oportunidades creadas en los meses de marzo, mayo y junio del año 2017; sus primeras oportunidades, por su parte, fueron creadas en noviembre de 2013.

Figura 4:



Para la región *Americas* observamos una mayor concentración de oportunidades creadas en los meses de abril, mayo y diciembre del año 2017. Al igual que para la región *APAC,* sus primeras oportunidades fueron creadas en marzo de 2014.

El punto en común en las tres regiones es una mayor concentración en 2017 y relativamente pocas oportunidades en los dos años siguientes, con un mayor incremento de 2016 a 2018. No existen oportunidades creadas en fechas posteriores a diciembre de 2018. Observamos un pico en los datos: hubo una gran concentración de ítems de oportunidades creadas en mayo y junio de 2017. No se observa una tendencia contundente en ningún momento en particular del año.

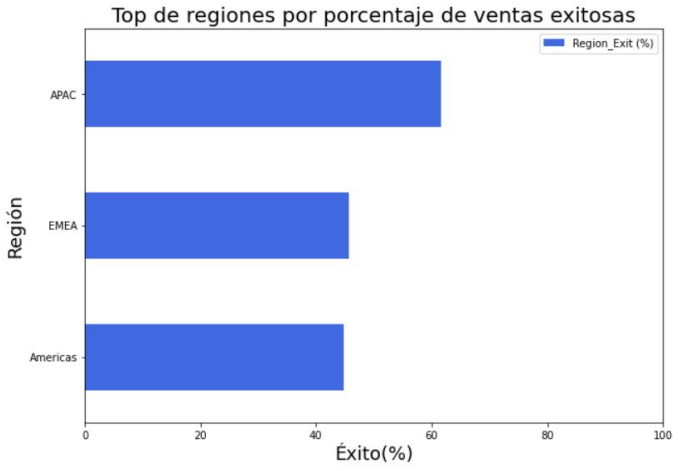
3.2. Clasificación por porcentaje de éxito y por dinero facturado

Definimos como *porcentaje de éxito* al cociente entre el número de oportunidades exitosas y la cantidad total de oportunidades.

3.2.1. Región

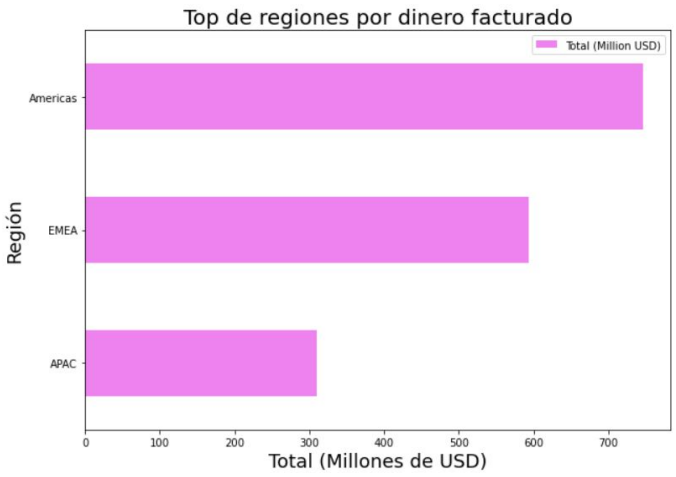
En la *figura 5* podemos ver que la región con mayor porcentaje de oportunidades exitosas es *APAC*, en segundo lugar *EMEA* y por último *Americas*.

Figura 5:



En contraste, la *figura 6* muestra que la región con mayor dinero facturado es *Americas*, en segundo lugar *EMEA* y por último *APAC*.

Figura 6:

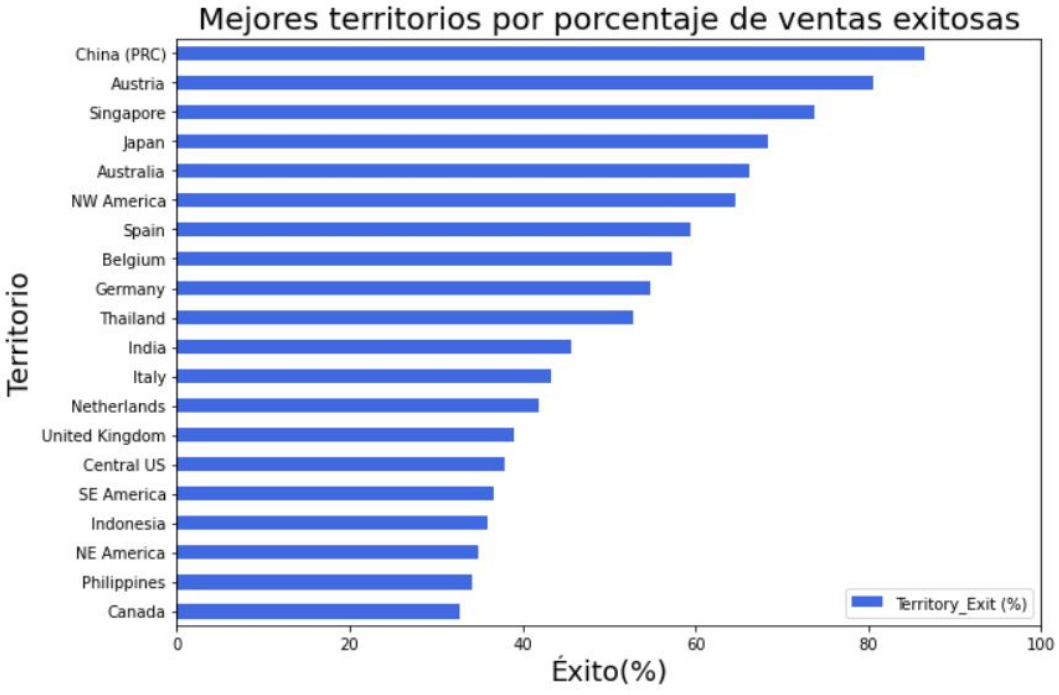


Cabe destacar que la región *Americas* es la que más factura a pesar de tener un menor éxito en cuanto a cantidad de ventas y en contraposición *APAC* es la que menos factura a pesar de tener un mayor porcentaje de ventas exitosas.

3.2.2. Territorio

En la *figura 7* vemos los 20 territorios con mayor porcentaje de ventas exitosas. En este caso, tomamos en cuenta aquéllos que cuenten con, por lo menos, 50 oportunidades.

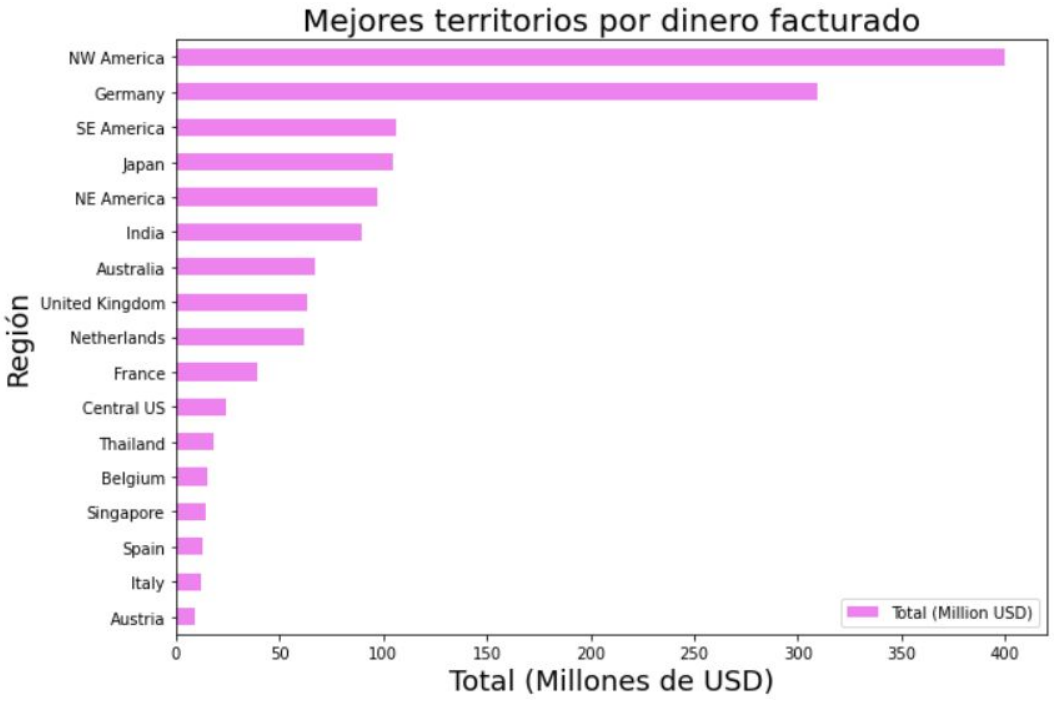
Figura 7:



Vemos una consistencia con respecto a la *sección 3.2.1*, ya que predominan con un gran porcentaje de éxito los países de la región APAC como China, Singapur, Japón y Australia.

En la *figura 8* vemos los 20 territorios con mayor dinero facturado. Para este gráfico también se tuvo en cuenta, solamente, aquellos territorios con por lo menos 50 oportunidades.

Figura 8:



Aquí también se observa consistencia respecto de la sección 3.2.1 (*Región*).

Observamos una notoria predominancia, en primer lugar de *NW America* (*región Americas*), y, en segundo lugar, de Alemania (*EMEA*). Es decir, son los dos territorios que proporcionan más ingresos.

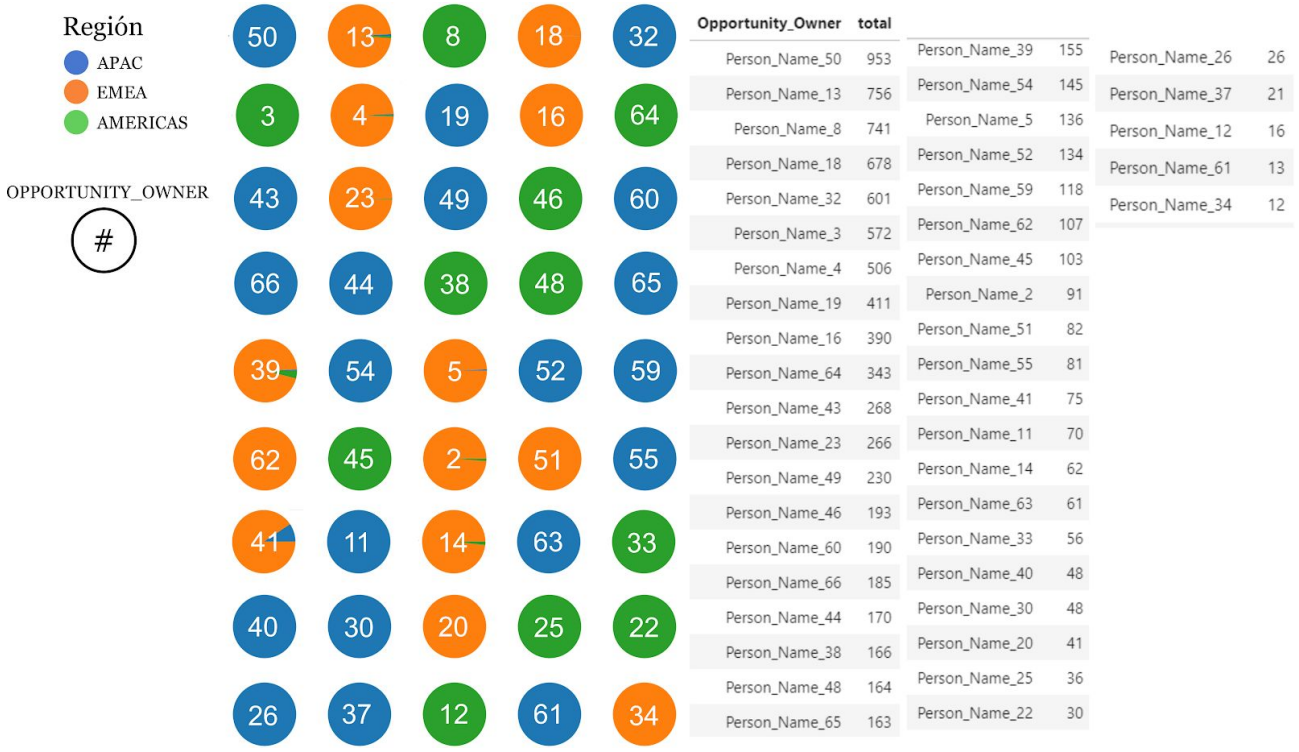
Comparando ambos gráficos podemos ver que, por ejemplo, en China, Singapur y Austria la mayoría de oportunidades creadas son exitosas pero esto no implica una gran facturación, es decir, se ganan muchas oportunidades pero son, en su mayoría, de poco valor.

3.2.3. Vendedores

Vale destacar que los vendedores se *especializan* en una región en particular, es decir, tienen la gran mayoría de sus oportunidades concentradas en una única región.

La *figura 9* hace referencia al porcentaje de oportunidades por región, donde cada *gráfico de torta* (*pie chart*) representa un vendedor. Estos están ordenados de mayor a menor según el *número total de oportunidades* (y se han considerado solamente los que tienen más de 10). A la derecha puede verse una tabla con el número total de las mismas.

Figura 9:

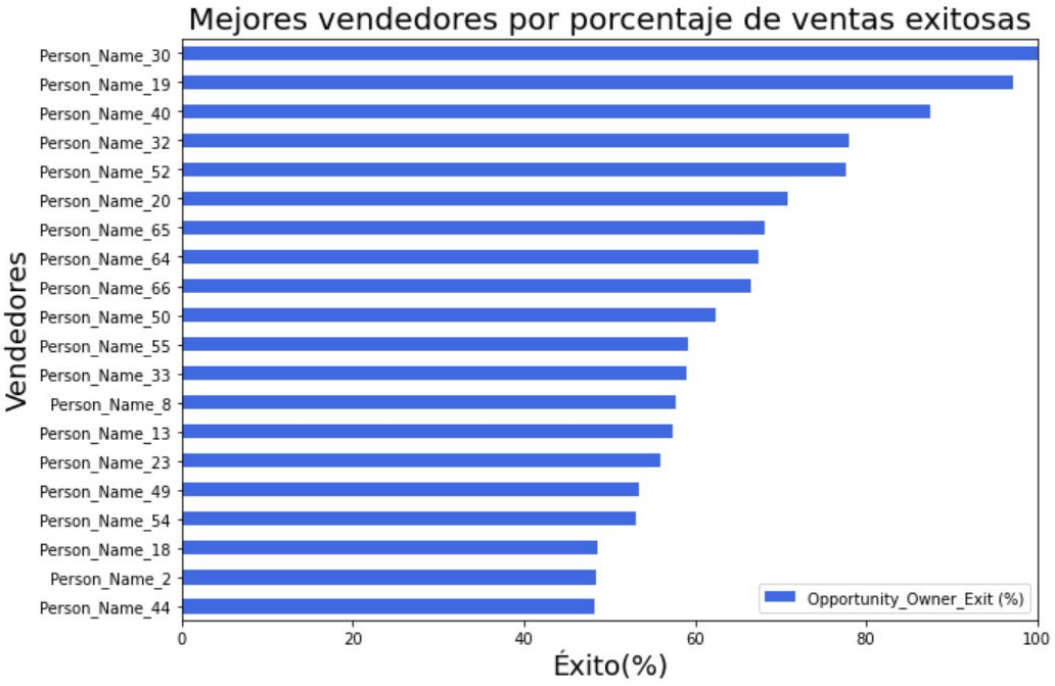


Opportunity_Owner	total
Person_Name_50	953
Person_Name_13	756
Person_Name_8	741
Person_Name_18	678
Person_Name_32	601
Person_Name_3	572
Person_Name_4	506
Person_Name_19	411
Person_Name_16	390
Person_Name_64	343
Person_Name_43	268
Person_Name_23	266
Person_Name_49	230
Person_Name_46	193
Person_Name_60	190
Person_Name_66	185
Person_Name_44	170
Person_Name_38	166
Person_Name_48	164
Person_Name_65	163
Person_Name_39	155
Person_Name_54	145
Person_Name_5	136
Person_Name_52	134
Person_Name_59	118
Person_Name_62	107
Person_Name_45	103
Person_Name_2	91
Person_Name_51	82
Person_Name_55	81
Person_Name_41	75
Person_Name_11	70
Person_Name_14	62
Person_Name_63	61
Person_Name_33	56
Person_Name_40	48
Person_Name_30	48
Person_Name_20	41
Person_Name_25	36
Person_Name_22	30
Person_Name_26	26
Person_Name_37	21
Person_Name_12	16
Person_Name_61	13
Person_Name_34	12

Claramente todos los vendedores se concentran (o especializan) en una determinada región.

En la *figura 10* vemos los 20 vendedores con mayor porcentaje de ventas exitosas. En este caso, tomamos en cuenta solo aquellos que tuvieron por lo menos 20 oportunidades.

Figura 10:



Es llamativo que *Person_Name_30* tiene un éxito del 100%, es decir, concretó todas las ventas que tuvo a cargo. También es pertinente destacar que Person_Name_19 y *Person_Name_40* tienen una efectividad mayor al 80%.

En la *figura 11* vemos los 20 vendedores que más dinero han facturado. Aquí también solo se han tenido en cuenta aquéllos que tuvieron al menos 20 oportunidades.

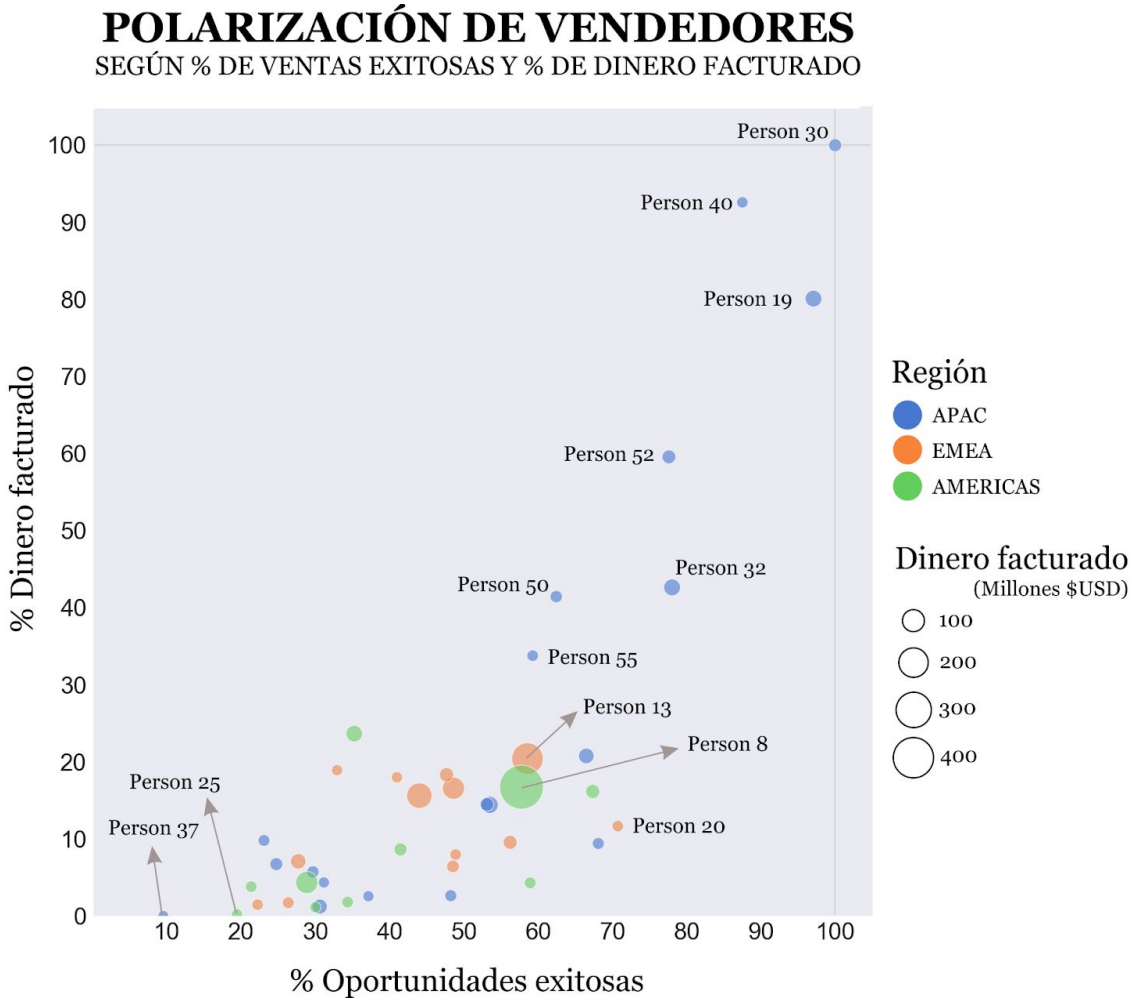
Figura 11:

Comparando con la *figura 10*, se ve que no siempre existe una *correlación* entre el porcentaje de ventas exitosas y el dinero facturado por cada vendedor. Esto se debe, potencialmente, a la diferencia de oportunidades a cargo de cada uno. Por ejemplo, *Person_Name_30* tuvo éxito en todas sus oportunidades pero, vemos en la *figura 11, que* no se encuentra entre los que más facturaron; esto se debe a que no tuvo muchas oportunidades asignadas (solamente 48). El caso opuesto es *Person_Name_8*, ya que es el vendedor con más dinero facturado pero no se encuentra dentro de los vendedores con mayor porcentaje de ventas exitosas.

Teniendo en cuenta lo afirmado anteriormente, nos interesa analizar el rendimiento de los vendedores. Para ello definimos un nuevo parámetro que utilizaremos junto con el *porcentaje de éxito* y que llamaremos *porcentaje de dinero facturado*. Este último se calcula como el cociente entre el dinero facturado y el dinero que suman todas las oportunidades que el vendedor tuvo a cargo.

En la *figura 12* cada burbuja representa un vendedor. Solamente se han incluido aquéllos que han tenido al menos 20 oportunidades a cargo. El color asignado a cada una se corresponde con la región en la que este se especializa. El tamaño de las burbujas representa el dinero facturado por cada uno. Por último, se han etiquetado los nombres de los nueve mejores vendedores y de los dos peores.

Figura 12:



En este gráfico, vale destacar que cuanto más cerca se esté de la esquina superior derecha tanto mejor desempeño tendrá el respectivo vendedor; a su vez, cuanto más cerca se esté de la esquina inferior izquierda tanto peor será el desempeño de dicho vendedor. Por su parte, en la esquina inferior derecha se encuentran aquéllos que han tenido éxito en la mayoría de sus oportunidades, pero con un bajo porcentaje de dinero facturado. Esto implicaría que este vendedor ha perdido las oportunidades que generan mayores ingresos, por caso, *Person 20* (cuyo desempeño resulta, aproximadamente, en un 70% de éxito pero solo un 10% de dinero facturado). En conclusión: los vendedores más eficientes son *Person 30, 40, 19, 52, 32, 50 y 55*, en ese orden, mientras que, *Person 25 y 37* son los peores. Por otro lado, observando el tamaño de las burbujas, vemos que las regiones en las que más se factura son *Americas* y *EMEA*. En particular, los vendedores que más han facturado son *Person 8* y *Person 13*.

3.3. *TRF, Total_Amount y ASP*

La variable *TRF* (*toneladas de refrigeración* o *frigorías*), en principio, es una de las variables más importantes para nuestro análisis ya que entendemos que las ventas se estructura alrededor de esta. Al hacer un análisis más profundo de esta variable, observamos ciertos puntos interesantes. En primer lugar, descubrimos que, en el 61.89% de las oportunidades creadas, el valor de *TRF* es nulo. Es decir, en más de la mitad de estas, el producto o servicio solicitado no corresponde a *frigorías*. Es más, considerando solamente las oportunidades exitosas, vemos que en el 82.76% de los casos tenemos un valor nulo para la variable *TRF*. Podemos concluir entonces que esta no es la principal característica alrededor de la cual se estructuran las ventas, como suponíamos en un principio. No tenemos la información concreta en nuestros datos sobre la razón por la cual hay tantas ventas en las que el *TRF* es nulo, pero podríamos suponer que se trata de oportunidades donde se ha solicitado la instalación de equipos o algún complemento a ellos. Esta hipótesis fue comprobada verificando que los productos pueden clasificarse de forma excluyente según si la variable *TRF* es nula o no. Es decir, nunca un producto tendrá un *TRF* nulo en algunas oportunidades y en otras no.

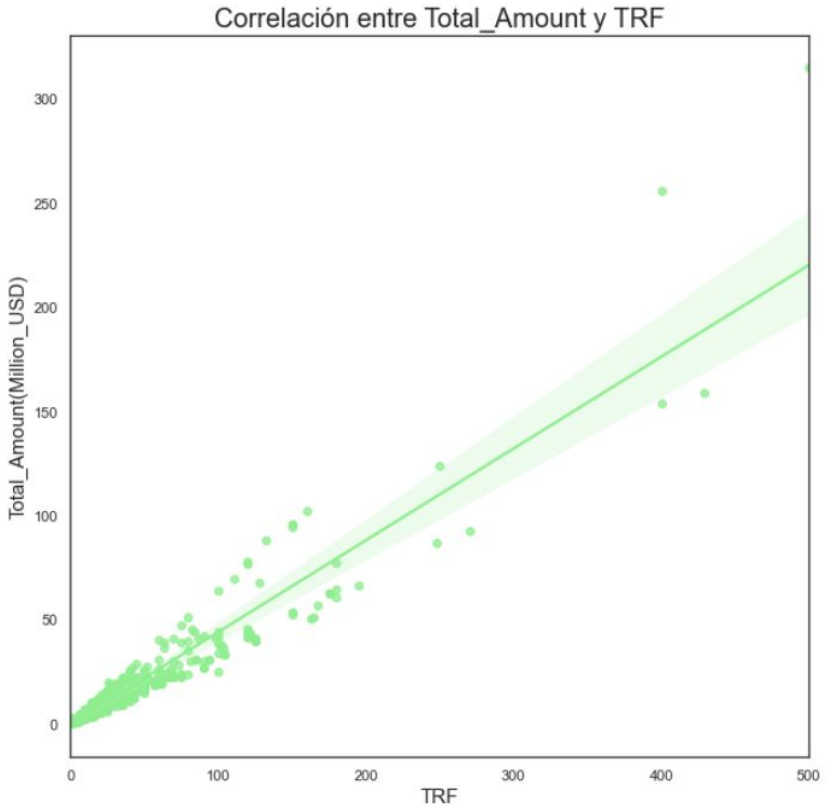
Para el caso de las oportunidades cuyo *TRF* es *no nulo* encontramos una relación lineal entre esta variable y *Total_Amount*, cuyo factor de proporcionalidad es la variable *ASP*. Es decir, *ASP* representa (para los casos con *TRF no nulo*) el *precio por gramo de refrigeración*; entonces podemos afirmar lo siguiente:

$$Total\_Amount = TRF \cdot 10^6 \cdot ASP$$

donde el factor 10^6 se debe al pasaje de toneladas a gramos.

Esto puede apreciarse en la *figura 13*.

Figura 13:



C*orrelación lineal* entre *Total_Amount* y *TRF*

Por definiciones de las variables "*Total_Amount*" y "*Total_Taxable_Amount*", y entendiendo que cada fila del *dataframe* corresponde a un ítem de una oportunidad, si agrupamos por "*Opportunity_Name*" y sumamos los "*Total_Amount*" de cada ítem deberíamos obtener como resultado "*Total_Taxable_Amount*". Verificamos que esto se cumple en un 80.3% de los casos tomando como margen de tolerancia una diferencia de USD 10.000.

En las *figura 14* podemos ver los casos en los que se cumple dicha relación, en el eje horizontal se ha representado el monto total por oportunidad (la suma de los montos de los ítems dentro de cada oportunidad):

Figura 14:

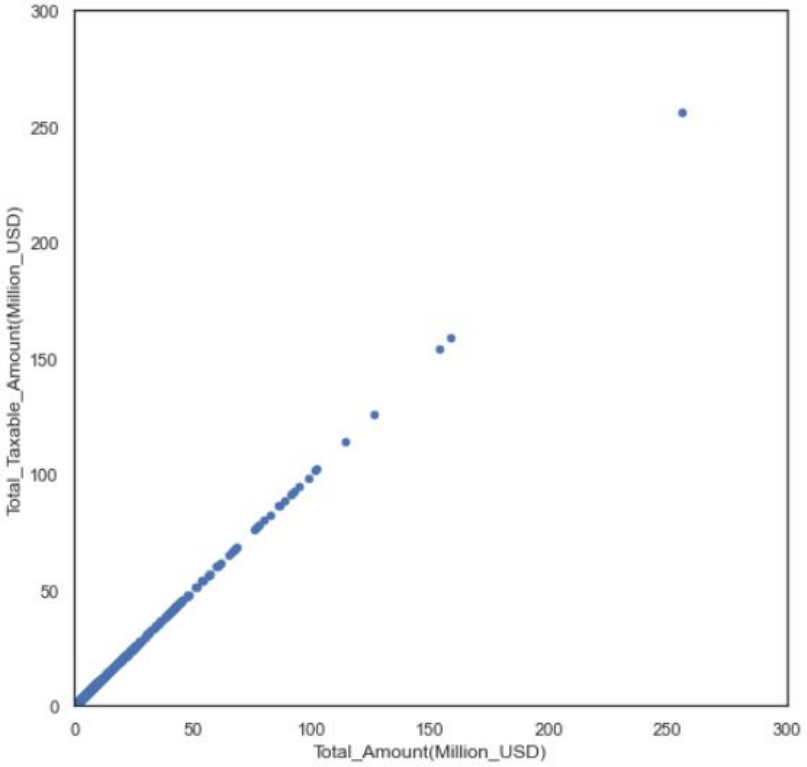
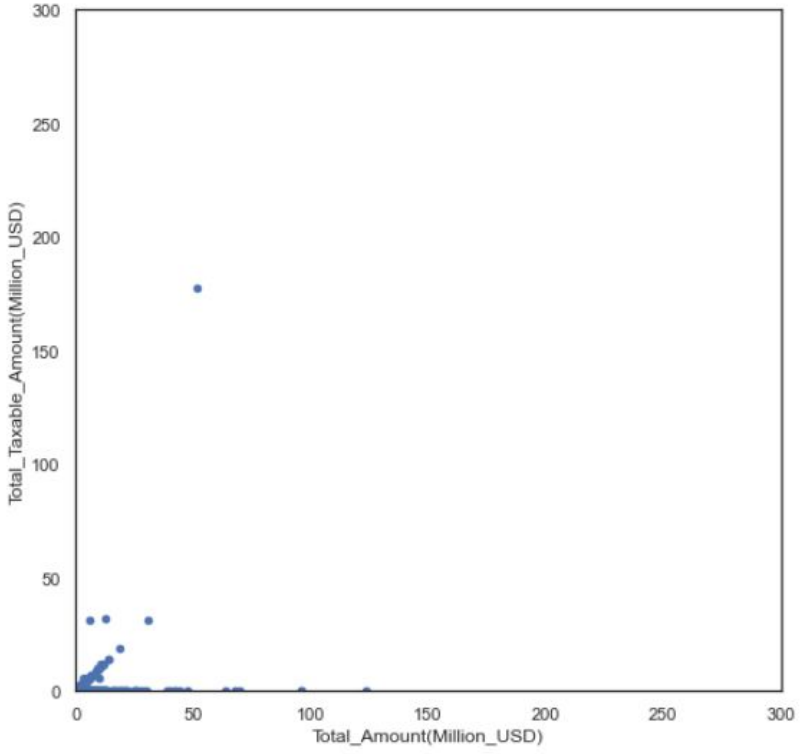


Figura 15:



En la *figura 15* están representadas las oportunidades en las que no se cumple la relación.

4. Conclusiones

El *dataset* bajo análisis contiene información faltante para varias características de las oportunidades que, potencialmente, podrían ser importantes. Estas contienen menos del 8% de las observaciones y son: "*Brand*", "*Product_Type*", "*Size*", "*Price*" "*Product_Category_B*" y "*Currency*". A su vez, notamos que la mayoría de las oportunidades fueron creadas en el año 2017.

El 82.76% de las *ventas exitosas* corresponden a productos cuyo *TRF* es nulo, por lo que podemos afirmar que dicha variable no es realmente la característica alrededor de la cual se estructuran las ventas.

La mayor proporción de oportunidades exitosas se encuentran en la región *APAC*, mientras que en la región *Americas* es donde se factura más.

Por otro lado, se buscó destacar el rol del vendedor a cargo de cada *oportunidad*. Por lo que analizamos, los vendedores están *especializados (o concentrados)* en una región en particular, es decir tienen todas (o la gran mayoría de) sus oportunidades en una única región. Además, encontramos que los mejores vendedores se especializan en la región *APAC*.