FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1

[75.06] Organización de Datos
Curso 01 - Argerich
16 de Noviembre de 2020
Segundo cuatrimestre de 2020

Alumno:	Batastini, Franco Nicolás
Número de padrón:	103775
Email:	fbatastini@fi.uba.ar
Alumno:	Burman, Federico
Número de padrón:	104112
Email:	fburman@fi.uba.ar
Alumno:	Longo, Manuel
Número de padrón:	102425
Email:	mlongoe@fi.uba.ar
Alumno:	Lovera, Daniel
Número de padrón:	103442
Email:	dlovera@fi.uba.ar

Repositorio: https://github.com/federicoburman/OrgaDatos

Índice

1. Introducción **2**

2. Limpieza del set de datos **3**
2.1. Remoción de los campos nulos . . . 3
2.2. Corrección de errores en el set de datos . . . 3
2.3. Remoción de campos "None" en territorio . . . 3
2.4. Condensación de columnas redundantes . . . 3
2.5. Remoción de casos que no contaron con aprobaciones necesarias . . . 4

3. Datos considerados para el análisis exploratorio **4**

4. Cuestionamientos Iniciales **4**

5. Análisis exploratorio **5**
5.1. Probabilidad de aprobación . . . 5
5.2. Probabilidad de cerrar una venta exitosamente 'Closed Won' . . . 6
5.3. Uniendo la probabilidad de aprobación con la de Closed Won . . . 7
5.4. Fechas de oportunidades para el top 10 mas probables . . . 9
5.4.1. Closed Won . . . 9
5.4.2. Closed Lost . . . 10
5.5. Analisis Temporal de Oportunidades por Región . . . 11
5.5.1. APAC . . . 11
5.5.2. EMEA . . . 12
5.5.3. Americas . . . 12
5.6. Analisis de montos 'Amounts' . . . 13
5.6.1. Relación entre montos de venta y aplicables a impuestos . . . 13
5.6.2. Total Amount en USD para Territorios . . . 15

6. Uniendo los Análisis Individuales **16**
6.1. Top 10 interrelacionados . . . 16
6.2. Montos facturados y montos impositivos . . . 16
6.3. Fechas de Oportunidades . . . 16
6.4. Relación entre Closed Won y Facturación . . . 16

7. Conclusiones **17**

1. Introducción

En este trabajo se realiza un análisis explotario con los datos obtenidos de la empresa "Frio Frio", ya que para ellos es esencial optimizar los esfuerzos de sus representates comerciales, es decir deben priorizar las mejores oportunidades en el pipeline.

- 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 esta desarrollando.

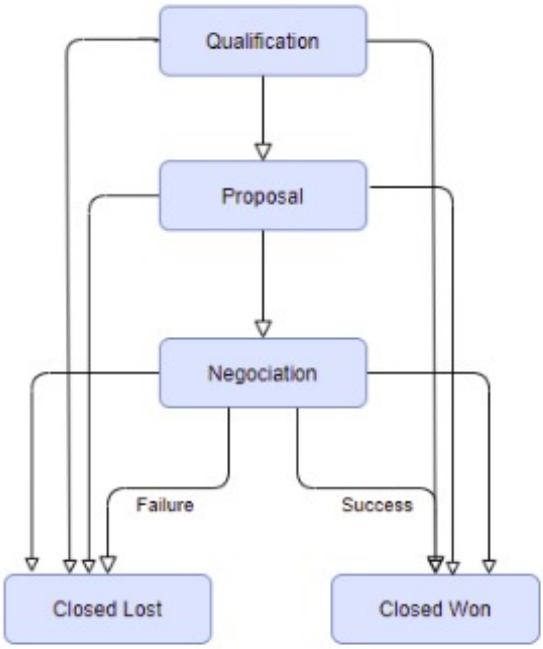

Figura 1: Pipeline de la empresa.

Por consiguiente se tratará a lo largo de este informe de predecir la 'Probabilidad de éxito' para cada oportunidad del set de datos. Al observar el pipeline de la empresa queda evidenciado que esto dependerá de las oportunidades que finalizaron 'exitosamente' y de las que no.

Para lograr esta tarea el set de datos cuenta con informacion geográfica de los clientes, fecha prevista de entrega de los equipos, presupuesto de las oportunidades, etc. Esto permitirá resolver interrogantes que se plantean a lo largo del trabajo y concluir finalmente como los representantes comerciales podrán reconocer cuales serán las mejores oportunidades en un futuro, y a su vez incrementará las ganancias de la empresa.

2. Limpieza del set de datos

2.1. Remoción de los campos nulos

Se encontraron una gran cantidad de datos faltantes en las columnas asociadas al promedio del precio de las ventas por oportunidades (ASP) y en la fecha de expiración del presupuesto (Quote expiry date) por lo cual se tomó la decisión de removerlos del análisis al no poder ser sustituidos por un dato válido para que los datos sigan teniendo sentido.

2.2. Corrección de errores en el set de datos

Al corroborar los registros de la columna "región" se detectó que se encontraron muchos que incluían los valores "Japan" y "Middle East", debido a que ya existen las regiones APAC y EMEA correspondientes a Asia-Pacífico y Europa-Oriente Medio respectivamente y que a su vez estos ya se encontraban cargados para cada registro en la columna "territorio" se concluyó que estos datos probablemente fueron mal ingresados y se cambiaron sus valores por el territorio al que pertenecen, es decir todos los registros del campo territorio correspondientes a "Japan" pasaron a contener APAC y los que incluían "Middle East" pasaron a contener EMEA.

2.3. Remoción de campos "None" en territorio

Considerando que el campo territorio es un dato geográfico importante para el análisis y que no tiene sentido que un representante comercial halla obtenido una oportunidad en un territorio inexistente pese a que si esta definida la region, se tomo la decisión de filtrarlos y que queden unicamente los campos distintos de None que tienen sentido físico.

2.4. Condensación de columnas redundantes

A continuación se presentan cuatro columnas binarias claves para el análisis exploratorio:

- 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).

Estas cuatro columnas en realidad representan pares binarios de (necesidad de aprobacion, aprobacion obtenida), como no son necesarias cuatro columnas binarias para el análisis que se lleva acabo, y es posible simplificarlas se tomo la decisión de reducirlas a una sola columna binaria que aporte la misma información contenida en las cuatro a partir del siguiente criterio:

Se consideró la intersección de los valores contenidos en necesidad de aprobación y aprobación obtenida por lo cual ante una igualdad ambos conservaban sus valores originales, pero si eran diferentes (necesidad de aprobacion verdadero y aprobacion obtenida falso, único caso) entonces el resultado era falso, este criterio se aplico para cada par binario mencionado antes y generó dos nuevas columnas binarias, estas dos nuevas columnas se volvieron a interceptar y el resultado obtenido es que la nueva serie de datos era igualmente binaria y representaba lo mismo que las cuatro. Es decir, la nueva columna, nombrada como "Es Oportunidad Posible" ya da información sobre si la oportunidad al menos tenía la posiblidad de entrar al pipeline ya que no habria tenido ningun tipo de inconveniente en el proceso.

2.5. Remoción de casos que no contaron con aprobaciones necesarias

Utilizando la columna que se generó para el set de datos en el punto anterior (`"Es_Oportunidad_Posible"`), fueron removidos todos los registros que contenían ceros. Esto es porque son ventas que no tuvieron posibilidad de seguir avanzando en el pipeline por problemas burocráticos o en las condiciones de la venta, y no se deben considerar en el análisis general.

3. Datos considerados para el análisis exploratorio

El set de datos consta de X columnas por Y registros, entre ellos hay muchas columnas que no fueron considerados por ser irrelevantes para el objetivo de predecir oportunidades de éxito, por lo cual se trabajo con el siguiente listado:

- ID: id único del registro (Entero).
- Región: región de la oportunidad (Categórica).
- Territory: territorio comercial de la oportunidad (Categórica)
- Opportunity ID: id de la oportunidad (Entero).
- Quote Type: tipo de presupuesto (Categórica).
- Opportunity Created Date: fecha de creación de la oportunidad comercial (Datetime).
- Product Family: familia de 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).
- 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).

4. Cuestionamientos Iniciales

En un principio las preguntas que nos hicimos para comenzar a analizar el set de datos fueron las siguientes:

- ¿Qué columnas son útiles a simple vista y cuáles podemos simplificar?
- ¿Qué territorios son los mas probables a ser aprobados?
- ¿Qué nos determina la aprobación de una negociación?
- ¿Hay algún año o mes en particular que favorezca a las oportunidades?
- ¿Qué territorios son en los que más se factura?

- ¿Hay alguna relación entre el monto total y el monto impositivo para closed wons y losts?
- ¿Algún mes o año hubo alguna explosión de Wons o Losts?

A partir de estas cuestiones logramos iniciar nuestro análisis exploratorio para obtener conclusiones.

5. Análisis exploratorio

5.1. Probabilidad de aprobación

Lo primero que se consideró fue la nueva columna booleana, nombrada (`"Es_Oportunidad_Posible"`), ya que esta informa si la oportunidad contemplada en el set de datos fue aprobada o no, previo a entrar en el pipeline de la empresa. Se presenta la siguiente imagen preliminar, que da una idea general de cuales son los territorios en donde las oportunidades que se registraron fueron realmente parte del pipeline.

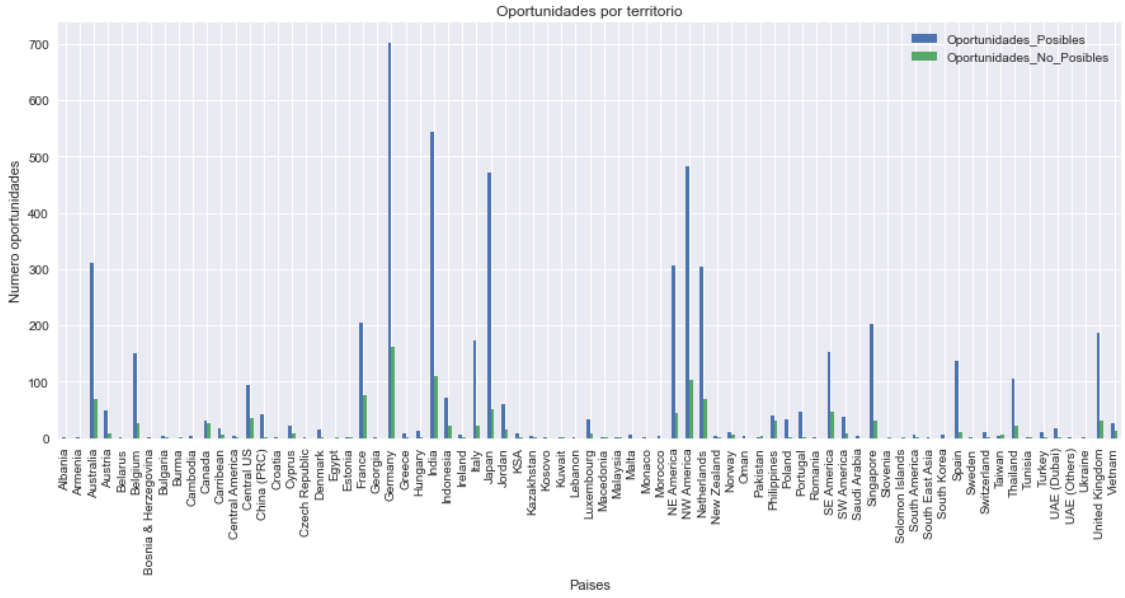

Figura 2: Imagen preliminar al análisis

Como se puede observar, territorios como **Alemania**, **India**, **Australia**, **Japón** tienen una cifra de oportunidades posibles altas, por lo cual a priorí las ventas se podrián orientar sobre esos territorios.

Entonces, como tenemos que tener en consideración el hecho de la tasa de aprobación, y no la cantidad de aprobación exacta, efectuamos un cálculo:

$$\frac{Aprobadas,100}{Totales} = \text{Tasa de aprobación}$$

Y obtuvimos graficándolo, lo siguiente:

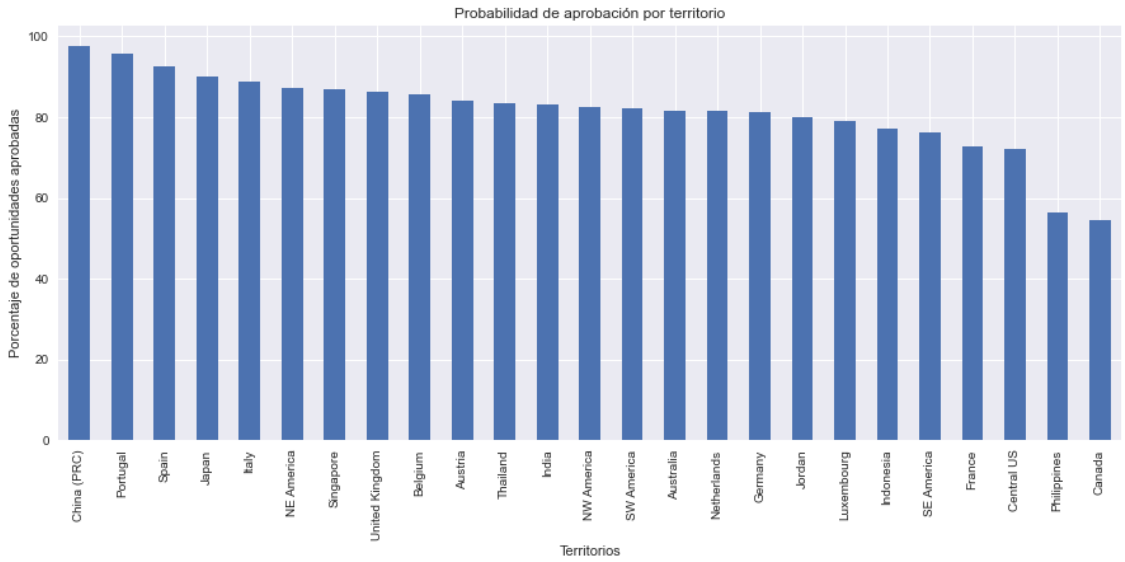

Figura 3: Probabilidad de aprobación por territorio.

Por lo tanto se ve que territorios como **China**, **Portugal**, **España**, que se podían interpretar con pocas oportunidades aprobadas anteriormente, pasan a ser los primeros en donde deberiamos enfocarnos ya que si bien en el set de datos tenemos pocas oportunidades aprobadas concretas, la relación entre aprobadas y no aprobadas termina siendo mayor para estos territorios.

5.2. Probabilidad de cerrar una venta exitosamente 'Closed Won'

Siendo que no tiene sentido que una propuesta que no fue aprobada resulte en Closed Won, ya que esta justamente no es posible en dicho territorio, entonces podemos analizar la probabilidad de que el stage resulte en Closed Won, dada la aprobación de la propuesta P(closed won|aprobación). Para esto obtuvimos el siguiente resultado:

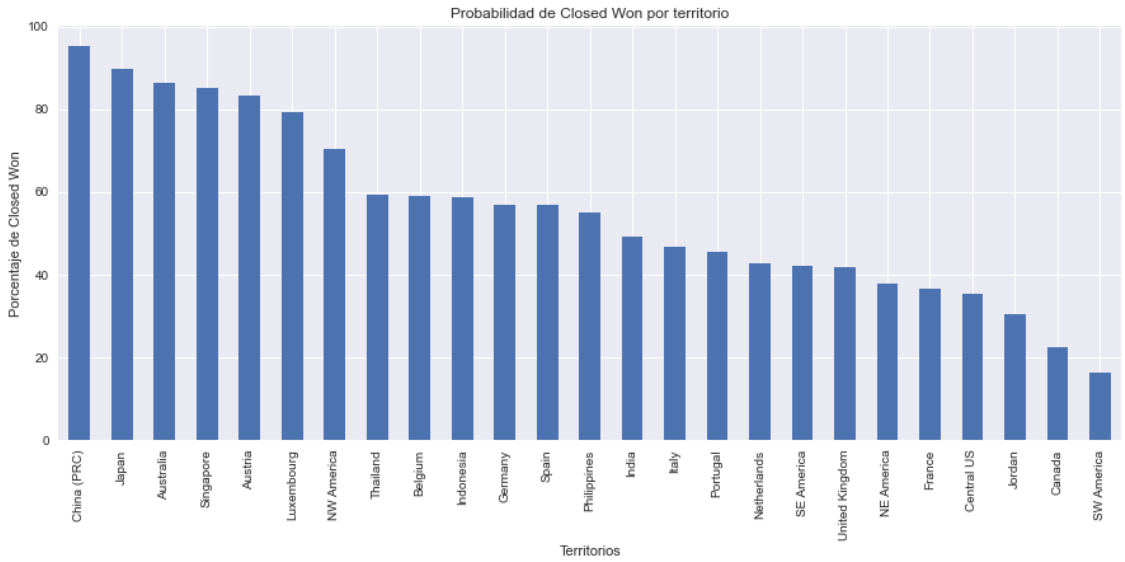

Figura 4: Probabilidad de Closed Won por territorio.

En particular se puede decir que ahora, no solo consideramos que la oportunidad de aprobada o no, sino que ademas se incluye si las ventas fueron exitosas o no. **China**, **Japón**, **Australia**, **Singapur** se perfilan como los territorios mas importantes para los vendedores en la empresa, siendo que hay mas probabilidad de que sean aprobados y que además concluyan en un closed won.

5.3. Uniendo la probabilidad de aprobación con la de Closed Won

Conociendo las probabilidades de que sea closed won, y que sea una propuesta aprobada, entonces podemos obtener la probabilidad total de que un proyecto sea exitoso tanto en aprobación, como también en stage final.
Entonces, operando matemáticamente para cada territorio, obtenemos la probabilidad total de la siguiente manera:

$$\frac{ProbAprobacion.ProbClosedWon}{100} = \text{Probabilidad Total}$$

Y efectuando un gráfico de barras como los anteriores, la probabilidad total nos queda para cada territorio:

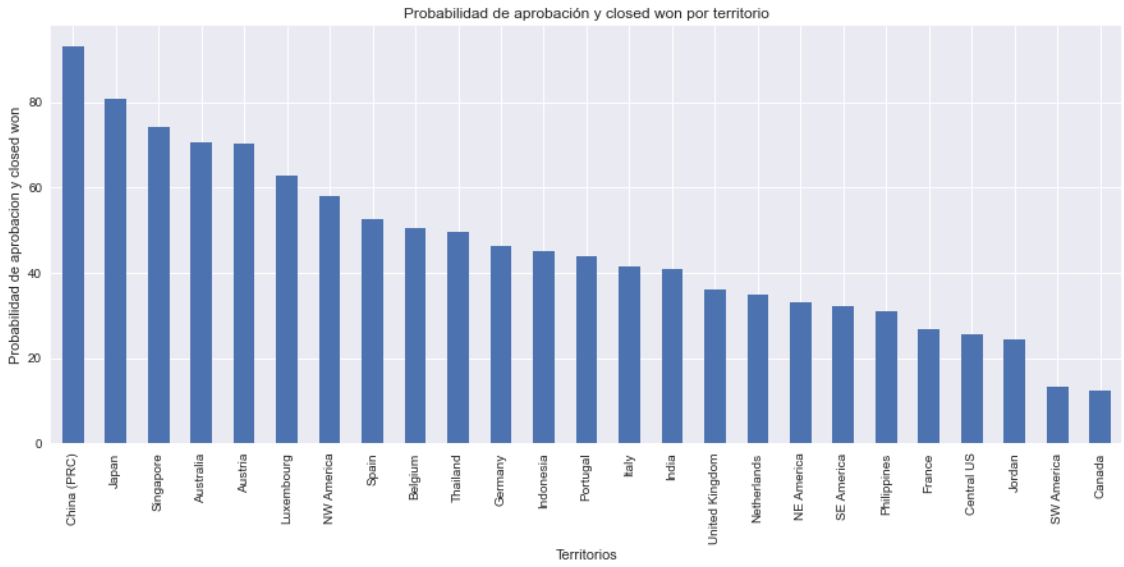

Figura 5: Probabilidad de Closed Won y Aprobación por territorio.

Ahora se resaltan a aquellos que estén en el top 10 de territorios con posibilidades de aprobación y closed won para reducir el numero de territorios. Entonces:

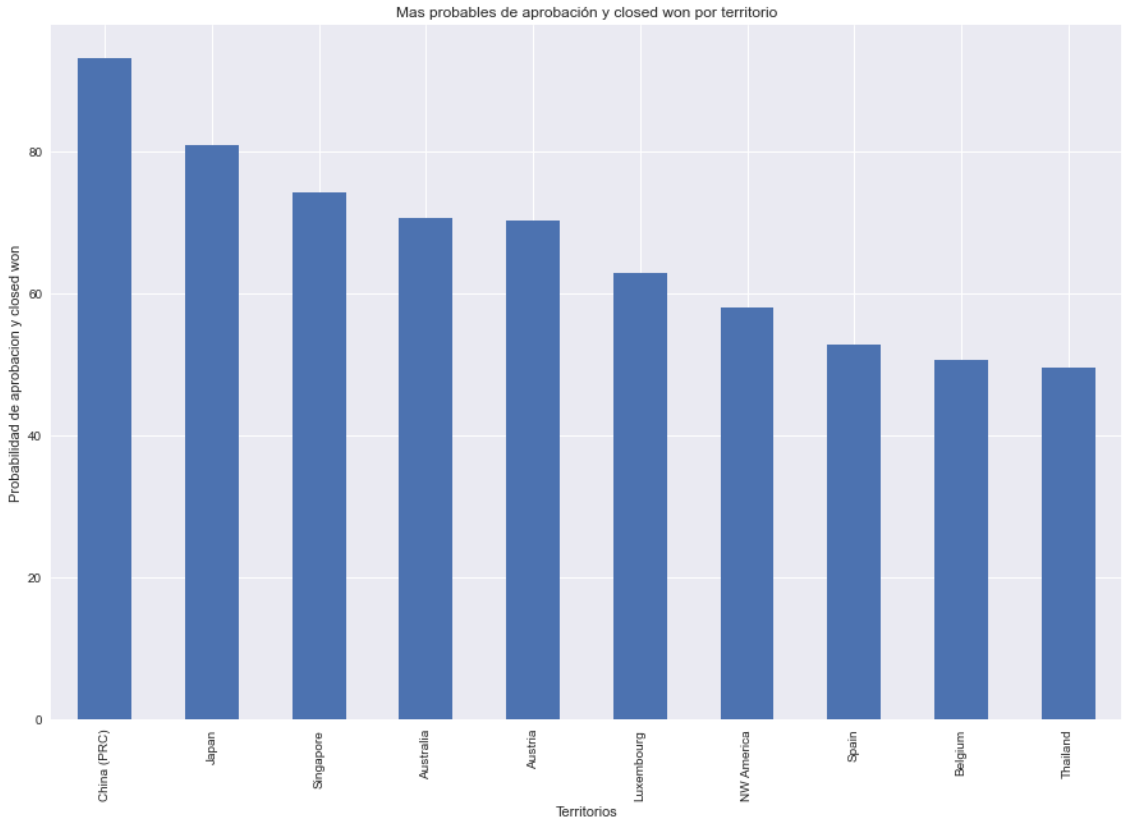

Figura 6: Top 10 Probables de Closed Won y Aprobación por territorio.

Finalmente, podemos destacar como territorios mas probables de una negociación exitosa sin tomar en cuenta los valores de facturación a **China**, **Japan**, **Singapore**, **Australia**, **Austria**, **Luxembourg**, **NW America**, **Spain**, **Belgium**, y **Thailand**. Es decir hasta este punto se tiene certeza de que se pueden enfoncar los esfuerzos de los vendedores en estos territorios para obtener una optimización válida

Si queremos una mejor visualización respecto de estos gráficos, con una visualización rápida, podemos utilizar GeoPandas para mostrar la probabilidad total en forma de mapa de calor con un mapa del mundo:

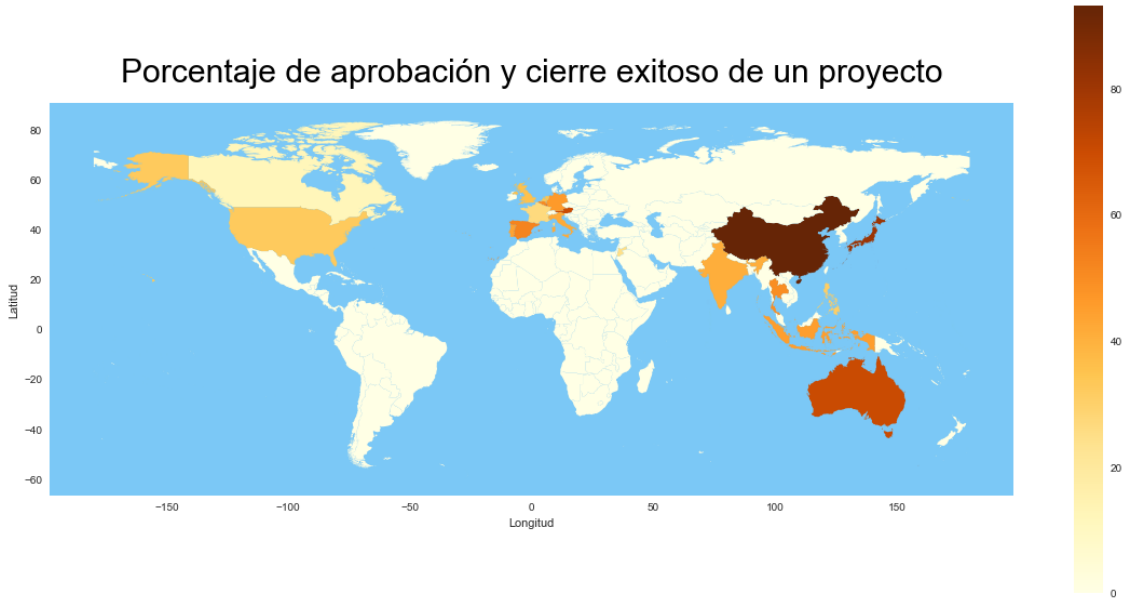

Figura 7: Mapa del mundo con su probabilidad de Aprobación y Closed Won.

5.4. Fechas de oportunidades para el top 10 mas probables

Quisimos incorporar fechas para el análisis de closed won y aprobación como dato extra. Surgieron las siguientes ideas:

5.4.1. Closed Won

Buscamos incorporar un heatmap en el cual podemos hallar en que fechas se encuentran distribuidos los closed won para el top 10 de territorios, y ver si hay algun año o mes que destaque por encima del resto. El resultado obtenido puede verse en la siguiente figura:

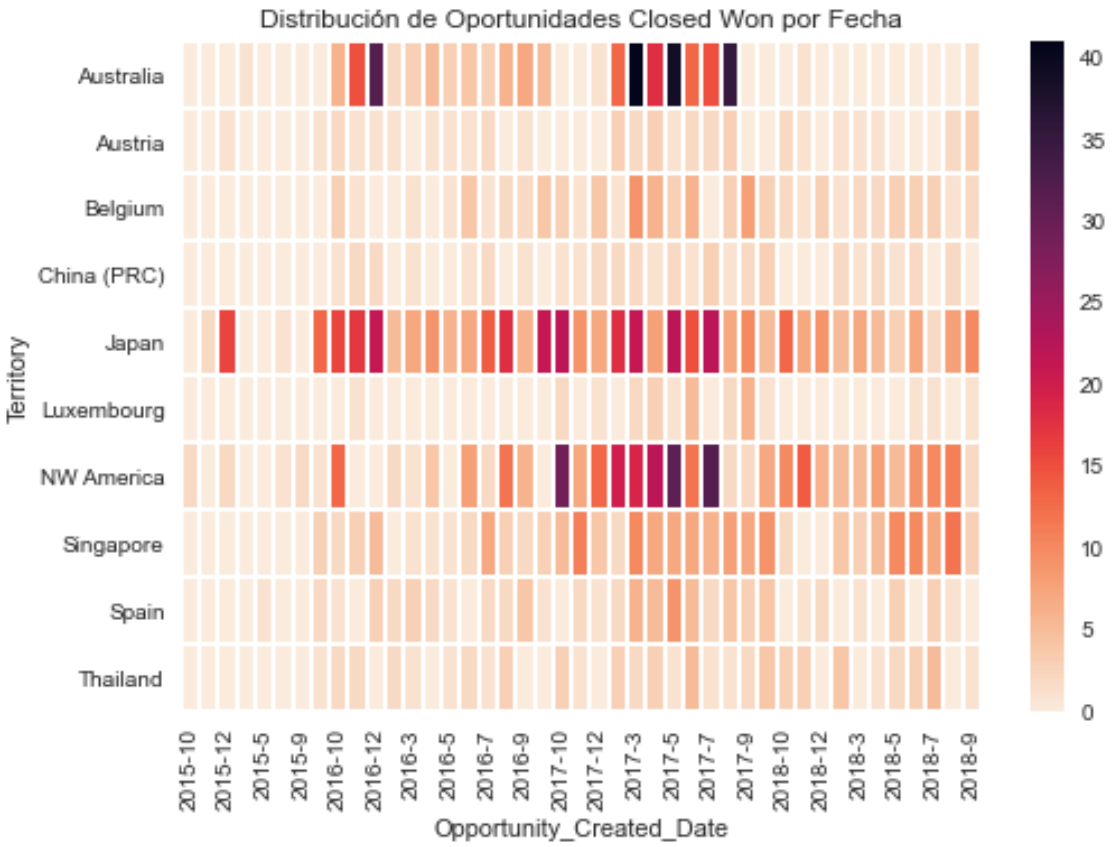

Figura 8: Heatmap de Fechas de Closed Won y Aprobación por territorio.

5.4.2. Closed Lost

Similar al caso del closed won realizamos un heatmap de la misma estructura, pero aplicado para closed lost y se obtuvo lo siguiente:

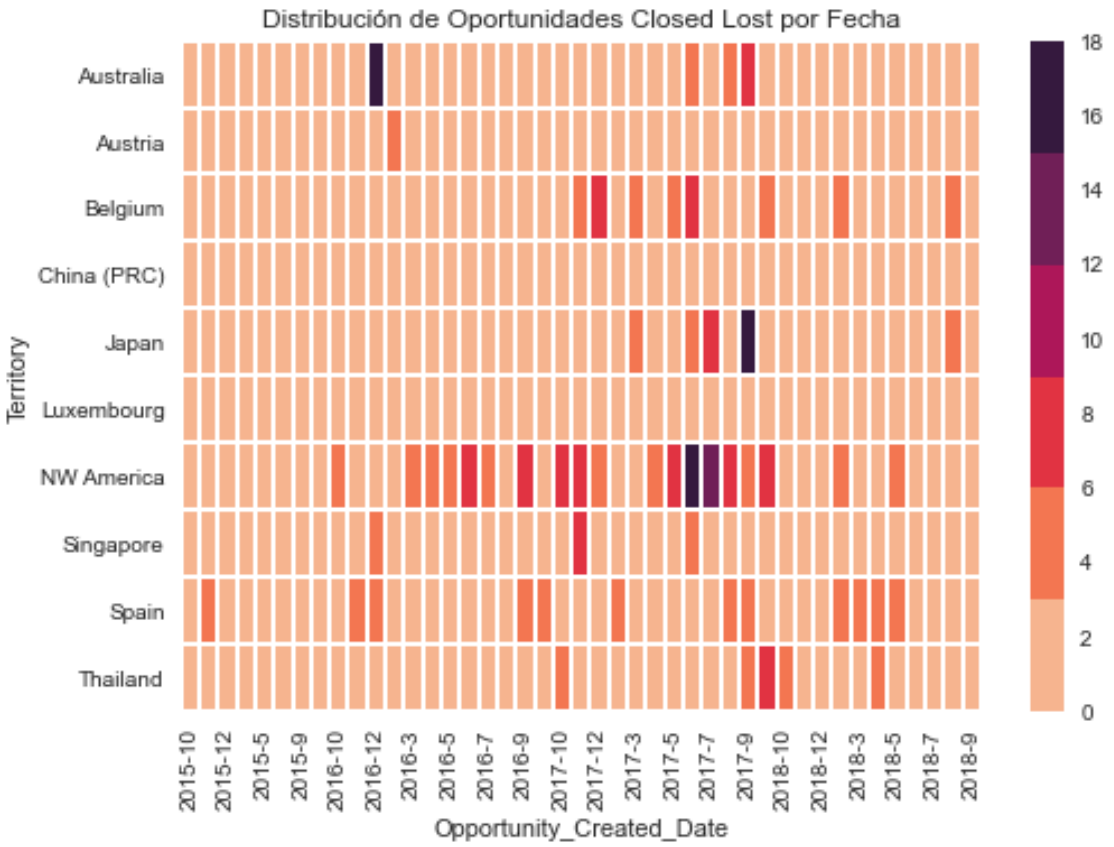

Figura 9: Heatmap de Fechas de Closed Won y Aprobación por territorio.

5.5. Analisis Temporal de Oportunidades por Región

Para poder conocer la cantidad de oportunidades dadas por año para cada región decidimos utilizar line plots que nos relacionen la cantidad de oportunidades con el año en el cual se generaron. Con este analisis pretendemos poder observar cual es la tendencia de los diferentes territorios para el futuro. De esta forma, para cada región logramos lo siguiente:

5.5.1. APAC

Para APAC obtenemos:

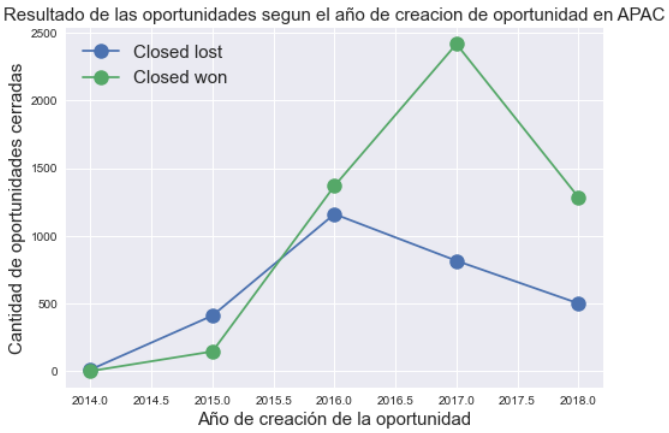

Figura 10: Line plot de oportunidades para APAC.

En la región de Asia Pácifico se puede observar como a lo largo del tiempo la diferencia entre las oportunidades Closed Won y Closed Lost cambió, especialmente fue en el año 2016 cuando hubo mas Closed Won que Lost, separandose por mucho este numero en el 2017 y disminuyendo en 2018. Por la acotada cantidad de años de la que hay información no podemos afirmar que hay una fuerte tendencia par ningún lado.

5.5.2. EMEA

Para EMEA obtenemos:

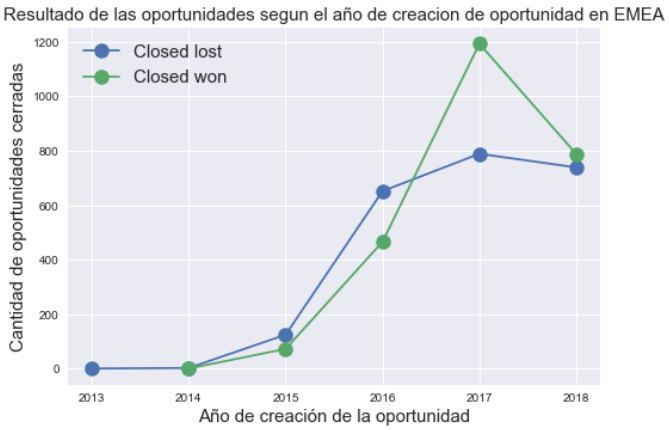

Figura 11: Line plot de oportunidades para EMEA.

Para la region EMEA, se puede ver que a lo largo de los primeros cuatro años de análisis la evolución de las oportunidades son bastante similares. Siempre se registraron mas oportunidades Closed Lost, que Closed Won. En el año 2017 se registro un pico en oportunidades Closed Won y en ese mismo año hubo una diferencia notoria en comparacion con las oportunidades Closed Lost. Luego en el 2018 se produjo un descenso pronunciado de las oportunidades Closed Won y se volvio a la tendencia anterior de similitud con las Closed Lost.

5.5.3. Americas

Para Americas obtenemos:

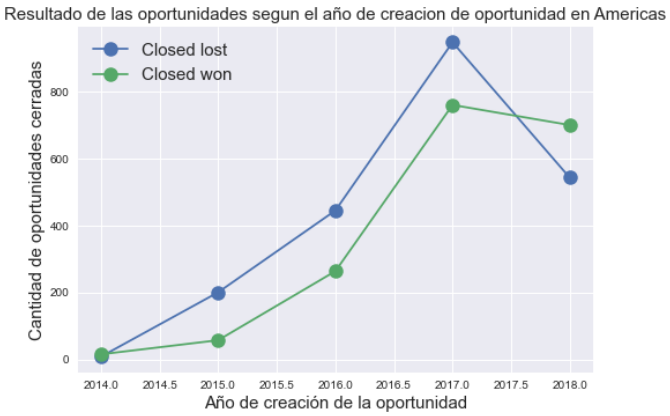

Figura 12: Line plot de oportunidades para Americas.

En la región de Americas (que solo incluye a Estados Unidos) hasta el año 2017 siempre tuvo mas oportunidades en Closed Lost que Won pero eso cambio en el año 2018, cosa que puede mantenerse a lo largo del tiempo.

5.6. Analisis de montos 'Amounts'

Como otra opción interesante decidimos analizar los montos de venta y aplicables a impuestos.

5.6.1. Relación entre montos de venta y aplicables a impuestos

Como primer idea filtramos cuales eran aquellas oportunidades que poseían total taxable amount distinto al total amount, y hicimos un pie plot para saber a que cantidad corresponden estos:

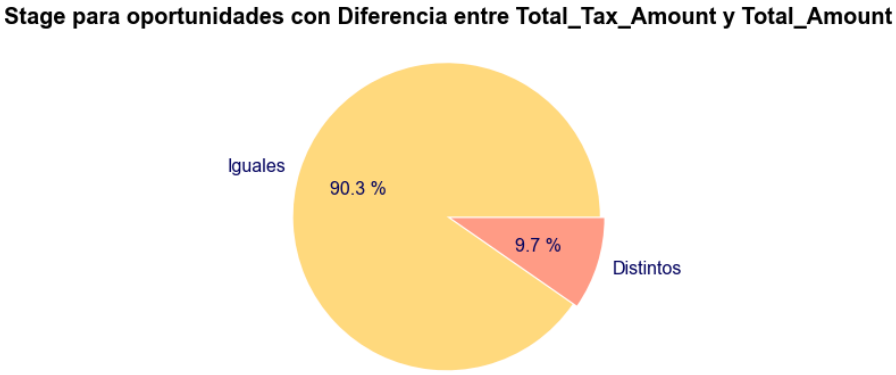

Figura 13: Pie plot de relación entre Total Taxable Amount, y Total Amount.

Tomando aquellos que difieren entre la cantidad del Total Amount y el Total Taxable Amount, decidimos analizar cuantos terminan en closed won y cuantos en closed lost, y de manera similar también realizamos otro pie plot:

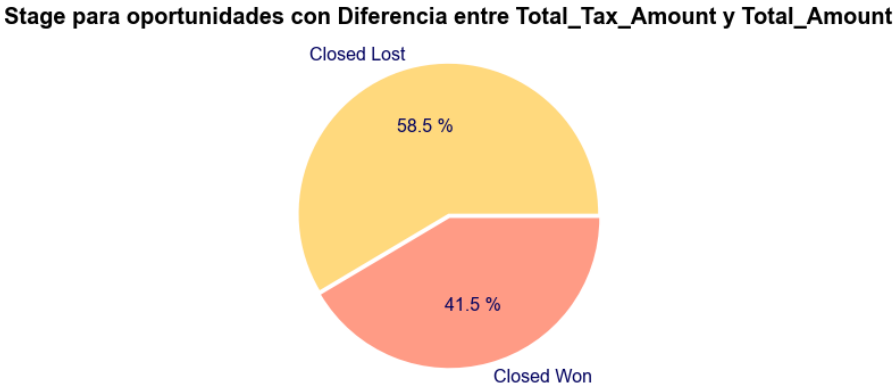

Figura 14: Pie plot del Stage para montos diferentes.

Como vemos que aún no podemos llegar a ninguna conclusión concisa, decidimos tomar un scatter plot donde aquellos puntos que se encuentran debajo de la linea roja corresponden a casos en donde el Total Amount es mayor al Total Taxable Amount; es decir, el monto impositivo es menor que el facturado, y contrariamente, los que se encuentren por encima de la linea, serán aquellos en donde el monto impositivo es mayor al monto facturado.
De esta manera, para Closed Lost obtenemos:

Figura 15: Scatter Plot de montos para Closed Lost.

Y para Closed Won de manera similar:

Figura 16: Scatter Plot de montos para Closed Won.

Inicialmente esperaríamos que los Closed Lost se encontraran por encima de la linea roja, y que los Closed Won se encontraran principalmente por debajo, pero el gráfico nos demuestra que nuestra hipótesis es errónea, y no creemos poder llegar a una conclusión útil con los datos que el mismo nos otorga, ya que la gran mayoría se encuentran cerca de la recta roja.
Sin embargo, en los Closed lost nos llama la atención la cantidad de puntos que se encuentran con un Total Taxable Amount bajo en relación con el Total Amount Sumado.

5.6.2. Total Amount en USD para Territorios

Como última instancia del análisis exploratorio buscamos ver cuales eran aquellos territorios en los que nos encontrabamos con un alto Total Amount resultante de todas las operaciones.
Para esto decidimos convertir los total amount a USD para todos los territorios, y sumar los de todas las oportunidades agrupando por territorio.
Aplicando un gráfico de barras para el top 10 de los maximos total amount convertidos obtenemos que el resultado es el siguiente:

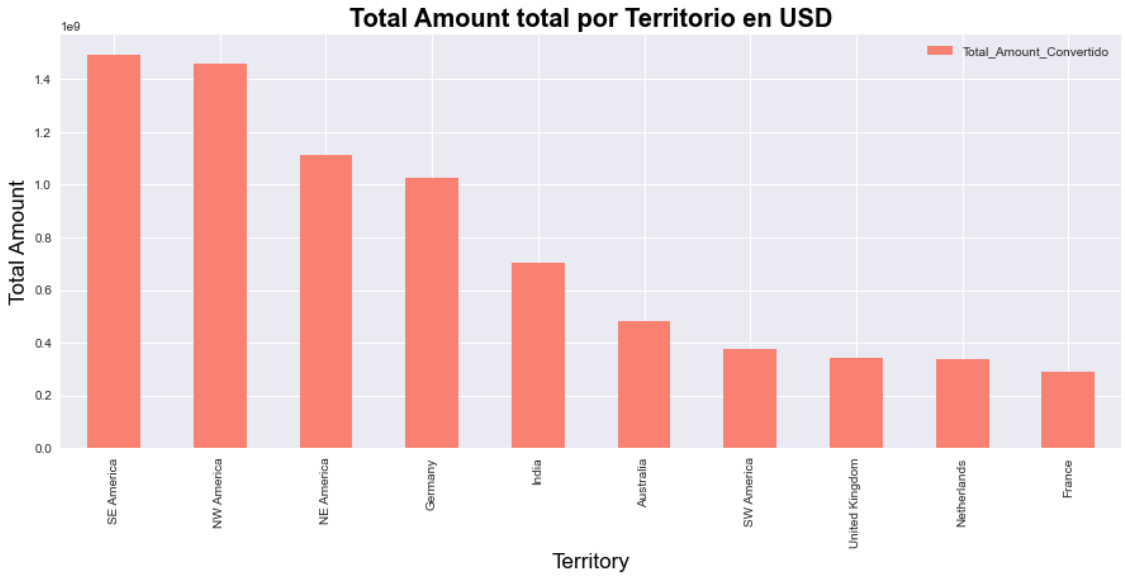

Figura 17: Top 10 Total Amount Converted por Territorio.

Donde del gráfico podemos resaltar los montos de **SE America**, **NW America**, **NE America**, **Germany**, **India**, **Australia**, **SW America**, **United Kingdom**, **Netherlands**, y **France**.

6. Uniendo los Análisis Individuales

6.1. Top 10 interrelacionados

El inicio de la correlación de datos se da viendo los territorios con mayor monto, y con mayor probabilidad de aprobación y resultante en Closed Won. A partir de esto, conocemos que por ejemplo **Australia**, y **NW America** son territorios muy probables en éxito, y con un alto monto de facturación; encontrándose ambos 2 en el top 10 de ambos casos.

6.2. Montos facturados y montos impositivos

Podemos decir que no existe una correlación clara entre el monto impositivo y el monto facturado, ya que dicho análisis nos otorgó resultados no del todo útiles como para obtener conclusiones ciertas, además de que el caso donde estos difieren es ínfimo (Menor a un `10%`.), y con poca diferenciación entre Closed Lost y Closed Won (`58.5%` y `41.5%` respectivamente.).

6.3. Fechas de Oportunidades

Tomando en cuenta los gráficos del análisis temporal por región se puede denotar un **"boom"** de oportunidades en el año 2017, y considerando el heatmap de las fechas de oportunidades de los top 10 mas probables de aprobación y closed won, podemos ver que también destaca la época de finales de 2016, y todo el año 2017 como zona mas densa en oportunidades, tanto para Closed Won como para Closed Lost.

6.4. Relación entre Closed Won y Facturación

Como dato extra del análisis, no todos aquellos territorios que tengan altas probabilidades de aprobación del proyecto, y de éxito del mismo significa que sean los mejores territorios en los

cuales invertir. Esto puede verse claramente en casos como por ejemplo China que tiene altísimas probabilidades de obtener una negociación exitosa, pero donde este mismo ni siquiera se encuentra dentro de los top 10 territorios en cuanto a facturación.
En otras palabras, un territorio puede ser muy viable para generar oportunidades, pero si el mismo no posee montos que sean representativos para los números de la empresa, entonces estos territorios pueden ser considerados como opciones pero no como las principales, ya que la economía de la empresa no se ve tan afectada por los mismos; aunque se podrían mantener propuestas constantes ya que es muy probable que estas sean aprobadas y obtengamos un cierto margen de ganancia con las mismas por resultar en closed won.

7. Conclusiones

En nuestra opinión, sería una buena opción invertir principalmente en **Australia** y en **NW America** como territorios debido a que es muy probable la aprobación, y los montos de facturación que hay en los mismos son altos para la empresa.

Sin embargo, también es interesante tener en cuenta territorios como lo pueden ser **China** o **Japon**, donde si bien sus montos no son tan altos como en los nombrados anteriormente, son muy propensos a aceptar negociaciones. De esta forma, son buenas opciones para inversiones constantes, aunque no de gran monto, que otorgarían pequeños resultados positivos a la empresa, pero que en cantidades mayores pueden resultar en un importante monto a futuro para la misma.

Por lo contrario, territorios como **Canada** o **SW America**, **no son territorios en donde debemos enfocar ventas** debido a que se requeriría de excesivo esfuerzo para una muy baja probabilidad de aprobación y exito, por lo que no deberían ser consideradas para las negociaciones. Además, como dato útil, se puede ver como si bien en SW America hay una alta facturacion (Se encuentra en el top 10 de territorios con mayor facturacion en USD), hay muy baja posibilidad de éxito, y esto resulta en hacerlo inviable para las inversiones.

Resumiendo, territorios como **China**, **Japon**, **Australia**, y **NW America**, son territorios muy viables para la inversión por lo dicho anteriormente y territorios como **Canadá** o **SW America** son inviables para el desarrollo de de negociación.