|  |
| --- |
| Organización de Datos(75.06/95.58)  Segundo Cuatrimestre 2020  **Trabajo Práctico N°1**  Curso 1: Argerich Grupo Creado |

**Historial de Cambios**

{{Por cada cambio en el Informe, especificar la numeración del mismo, el autor, la fecha y la descripción del mismo.}}

| Fecha | Versión | Modificado Por | Descripción  (Nota: agregue una breve descripción de la modificación efectuada) |
| --- | --- | --- | --- |
| **16/11/2020** | **1.0** | **Grupo Creado** | **Creación** |

**Tabla de Contenidos**

1 Objetivo 4

2 Introducción 4

3 Documentos Relacionados 4

4 Participantes 5

5 presentación de problema y Consideraciones generales 6

5.1 El problema a analizar 6

5.2 Pautas Generales 6

5.2.1 Referencia de Datos 6

5.2.2 Lenguajes y Librerias Utilizados 7

5.2.3 Repositorio en GitHub. 7

6 Análisis inicial al cargar la información del DataFrame 7

6.1 Carga inicial de Datos 7

6.2 Insights - Análisis de la variable objetivo Stage 7

7 Profundizando en el Dominio y Optimizando el DataFrame 8

7.1 Análisis 8

7.2 Insigths. 8

8 Visualizaciones Exploratorias 9

8.1 Análisis 9

8.2 Insigths. 9

9 Gráfico de Oportunidades Ganadas y Perdidas por Región 10

9.1.1 Gráfico de Porcentaje de Oportunidades Ganadas y Perdidas por Región

10 conclusiones y pasos a seguir 10

# Objetivo

{{{El Informe define los detalles de las .

El documento consiste en realizar un análisis exploratorio de los datos provistos con el objetivo de determinar características y variables importantes, descubrir insights interesantes, y analizar la estructura de los mismos, teniendo en cuenta que el trabajo realizado podrá ser utilizado en el TP2 (que consistirá en predecir la probabilidad de que una oportunidad sea exitosa) y se puede considerar un paso previo al mismo.

# Introducción

{{{En esta sección deberá documentar de manera resumida el contenido del documento y su organización. Se deberá describir el propósito del documento y la audiencia a la que está dirigida.

Introducción al Análisis Exploratorio de Datos con Pandas

Análisis Exploratorio de Datos con Pandas es un enfoque que comprende un conjunto de tareas para analizar conjuntos de datos para poder encontrar sus principales características.

Estas tareas en general comprenden todo lo que tenemos que hacer desde que se formula una pregunta interesante, se reúnen los datos y se desarrolla el proceso necesario para \*\*poder responder esa pregunta\*\*.

Su Objetivo es:

- Entender los datos.

- Ver características de los datos.

- Detectar irregularidades (outliers) de los datos.

- Obtener valores estadísticos de los datos.

- Realizar visualizaciones rápidas que faciliten el proceso de exploración.

Usualmente el resultado del mismo suele ser un reporte o un notebook, que reúne código y visualizaciones para llegar a ciertas \*\*conclusiones o insights\*\*.

Es un proceso iterativo, que se retroalimenta.

En general realizaremos el análisis del mismo sobre uno o varios Dataframes.

## Temario

Estos son algunos de los temas que intentaremos revisar a lo largo del análisis propuesto

- Presentación del problema y Consideraciones generales

- Análisis inicial al cargar el dataframe.

- Anallisis propiedades especificas del set de datos

- Verificación de Calidad de Datos

- Utilizando Visualizaciones para nuestro Análisis exploratorio

# Documentos Relacionados

{{En esta sección se deberá indicar todos los documentos utilizados para construir este Informe, y los que son referenciados en alguna parte del mismo. A medida que se generen nuevos documentos que se relacionen con el presente Inoforme, se deberán agregar a la lista que se detalla a continuación.}}

|  |  |
| --- | --- |
| **Documento** | **Ubicación** |
|  | https://github.com/luzflor/DAtosTP-.git |
|  |  |
|  |  |

# Participantes

{{En esta sección se deberá indicar las personas y sus roles que participaron en la confección de del presente Informe.}}

Las siguientes personas han participado en construcción del presente documento:

|  |  |
| --- | --- |
| **Participantes** | **Rol** |
| Luz Fox | * Creación – Modificación |
| Romina Velardes | * Creación – Modificación |
| Juan M. Creado | * Creación – Modificación |

# presentación de problema y Consideraciones generales

{{En este punto se deberá dar una definición de las pautas del inofrme

## El problema a analizar

Una empresa que vende equipos de refrigeración de grandes superficies a otras empresas.

Tiene varias oportunidades de negocio: algunas funcionarán, otras no.El objetivo es predecir la probabilidad de éxito para mejor ubicar el trabajo de los vendedores.

## Pautas Generales

Los datos a analizar se pueden encontrar en el archivo **Entrenamieto\_ECI\_2020.csv** que se encuentra haciendo clic en <https://drive.google.com/file/d/1dOMHOcmTuEUmN6-OwaXzxmwtrSglE4fC/view?usp=sharingí>.

La variable de interés es “Stage”, que indica en que estado se encuentra la operación (pueden ser 5: Closed Won, Closed Lost, Negotiation, Proposal, Qualification)

### Referencia de Datos

* **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 última 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.

## Lenguajes y Librerías Utilizados

Se ultizó Pytton 3 y Pandas para el Procesamiento de los Datos.

Seaborn y MatPlotLib para realizar las visualizaciones de dichos datos

## Repositorio en GitHub.

Los notebooks con el código generados se pueden encontrar en el siguiente repositorio

https://github.com/luzflor/DAtosTP-.git

# Análisis inicial al cargar la información del DataFrame

{{ En este punto se deberá dar una definición de las pautas del Iinforme.}}

## Carga inicial de Datos

En esta primera etapa se realiza una carga de datos para la comprensión de los datos, se listan los tipos de datos, se identificó valores únicos, se identificó los valores nulos, se eligieron algunas columnas para el análisis.

## Insights - Análisis de la variable objetivo Stage

### 

Si bien el enunciado del tp menciona que la variable de interés estado de la oportunidad (Stage) presenta 5 valores posibles Closed Won, Closed Lost, Negotiation, Proposal, Qualification) realizando una visualización podemos ver que la mayor cantidad de ocurrencias

Presentan los estados. Closed Won, Closed Lost. Se descartaran los datos con los otros estados ya que no se relevantes para el análisis por dos motivos están desbalanceados en el set de datos y considerando que el análisis se centra en si se ganó una oportunidad.

.

# Profundizando en el Dominio y Optimizando el DataFrame

{{ En este punto se deberá dar una definición del informe.}}

## Análisis

En este se profundiza en el análisis de dataframe. Se analizan variables categóricas, se convierten datos para la optimización de memoria. Se van pensando relaciónes entre features.

Problemas con variables numéricas

## Insigths.

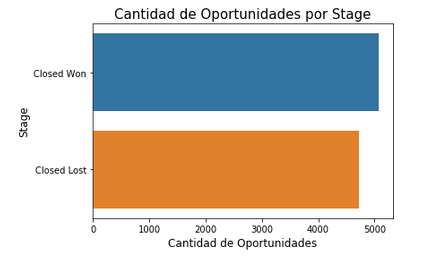
Apartar del análisis de los datos de set en genera; y del análisis de cómo se estructura una oportunidad de negocio se decide hacer un análisis a nivel macro por oportunidad de negocio ( no por el detalle de esta) en función de la variable objetivo.

Una oportunidad de negocio está compuesta por varias filas que corresponden a los ítems vendidos. Estos se repiten en función de la fecha estimada de entrega de inicio y fin.

El monto total gravado es por oportunidad de negocio y es la suma del monto total de la oportunidad.

Se procede a mostrar la cantidad de oportunidades de negocio por stage(variable de interés)

Vamos a Observas en el siguiente grafico como cambio la cantidad de oportunidades cuando se quitan el detalle de los ítems.



# Visualizaciones Exploratorias

{{ En este punto se deberá dar una definición del informe.}}

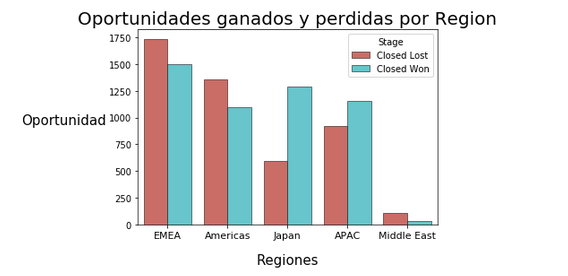
## Análisis

En esta etapa de nuestro análisis se van buscando relaciones entre las diferentes variables y visualizar mediante gráficos estas relaciones para continuar el camino de comprensión de los datos.

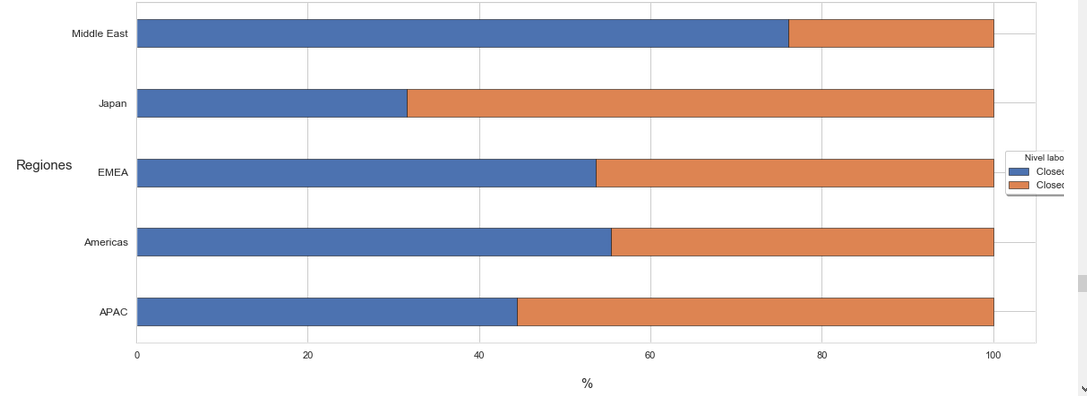
## Insigths.

Finalmente, agregado una tercera dimensión al análisis: la región. Analizamos los porcentajes de oportunidades de negocio por región. En este punto podemos empezar a generar insights interesantes sobre el problema de negocio que se plantea que es como ayudar a los vendedores a generar oportunidades de negocios, en ese sentido identificamos las regiones con sus fortalezas y desafíos(si bien no es el principal objetivo en esta etapa de proceso sino la exploración de datos),

# Gráfico de Oportunidades Ganadas y Perdidas por Región



## Gráfico de Porcentaje de Oportunidades Ganadas y Perdidas por Región



# conclusiones y pasos a seguir

Se entiende el EDA como un proceso interactivo, el cual se va retroalimentando.

en la medida que se va manipulando los datos..

En ese sentido si bien es conveniente llevar etapas que son obvias en la medida que se desarrolla el proceso de análisis se van descubriendo nuevas relaciones o desafíos en los datos que te obligan a tener nuevos perspectivas, nuevas preguntas y nuevas hipótesis.

Por ejemplo, para el análisis regional del monto total de ventas surgió el problema de las diferentes monedas.

Otro ejemplo fue encontrar en una primera etapa un set casi completo en cuantos nulos pero luego descubrir que el set estaba en muchos casos repletos de valores en cero .en variables importantes como TFR que es lo que la compañía vende. Nunca se perdió el foco que este proceso es una etapa de un proceso mayor cuya próximas etapas serian el ultima finalidad es generar insights sobre los datos para responder a una pregunta de negocio que permite tomar decisiones. En este caso esa pregunta es qué acciones se deben tomar para ayudar a los vendedores a optimizar las ventas.

Se espera en próximos pasos completar la limpieza del set, completar valores faltantes. Optimizar la menoría, completar análisis de variables