**Introducción Análisis de Datos**

**Sobre está capacitación:**

Generalmente los cursos, capacitaciones y seminarios se dan con el fin de tener una idea base en la cual el profesional se pueda desarrollar de un modo auto didáctico sobre la rama en la que tiene curiosidad o que conforma parte del desarrollo integral de su profesión.

En esta ocasión la idea es que todos entendamos y podamos crear valor como equipo de trabajo con un objetivo definido dentro del departamento.

Esto hace que el tema varié un poco en cual es el objetivo central del mismo, aportar conocimiento para el propio aprendizaje y de allí explorar o como en este caso, mostrar desde los conceptos más básicos. Hay algunas cosas que por principios estarán tentados a adelantarse, pero los vamos a cubrir todo de una manera práctica, siempre sobre los lineamientos ya establecidos del análisis de datos.

Voy a tratar la manera de ir haciendo este documento a manera de guía practica en cada una de las sesiones semanales que podremos ir teniendo.

Actualmente no tengo idea de duración, ya que los perfiles de cada uno de ustedes son distintos y el fin central será crear de cero una unidad que tenga todas las herramientas necesarias para hacer un eficaz y eficiente trabajo con los datos.

**¿El ser humano es bueno para analizar datos?**

La respuesta inicial es no, nuestro cerebro no es capaz de procesar identificar o agrupar volúmenes grandes de datos, segmentarlos, calcularlos y buscar fenómenos, patrones o inconsistencias en los mismos.

Para eso están las maquinas, sin embargo, nosotros podemos indicarles los pasos necesarios para que nos den una respuesta.

Debemos de estar conscientes que, en algunas ocasiones, que estas respuestas no son necesariamente la que nosotros esperamos y de eso se crean hipótesis, las cuales pueden ser nulas o alternativas.

Vamos a ir viendo estos conceptos de a poco y la idea medular es comprenderlos de una forma prístina, actualmente el modelo científico para analizar datos puede contener errores, pero a la larga es lo mejor que tenemos para despejar dudas o encontrar respuestas.

De antemano les pido perdón si este documento se vuelve en algunas palabras redundante o hasta cierto punto aburrido, escribo acorde a lo que en estos momentos estoy pensando, la experiencia me ha llevado a recorrer este camino de ida y vuelta varias veces en estos ya más de 25 años de trabajar con datos y aunque conozco las veredas, los atajos y los pasos difíciles en algún momento me puedo equivocar, por lo que a usted que está leyendo estas líneas en estos momentos le pido un poco de comprensión.

**¿Un analista de datos nace o se hace?**

En el año 2011 estaba preparando a un grupo de analistas de una importante corporación que en ese entonces dentro de sus datos contaba con 320 empresas registradas en Guatemala.

En esa ocasión me toco dar una introducción de la metodología de Kimball para los cubos de datos relacionales, (esto es parte de lo que en su momento era el Business Intelligence el cual no vamos a abordar), tenía allí alrededor de un mes y en la primera reunión les consulte con respecto a su tiempo en años en el trabajo, el cual variaba entre 8 y 15 años.

El porque de esta pequeña historia seria para resumir la pregunta inicial, todos a los que les estaba dando la capacitación me indicaron que eran analistas senior en información.

En la introducción después de la presentación les hice una pregunta que nadie supo contestar la cual fue la siguiente:

¿Como se llama el objeto que el dueño de la empresa tiene colgado en todas las puertas de la corporación y para qué sirve?

Claro en el tiempo que estuve previamente ya había investigado, la corporación pertenece a una familia judía y los objetos que estaban puestos en los marcos de todas las puertas de la corporación eran unos MEZUZA, que en hebreo significa “jamba de puerta” y que contiene un pergamino que tiene escrito dos versículos de la Tora. El cual esta albergado en una caja y adherido al marco derecho de las puertas con una leve inclinación y casi a la altura del hombro de un hombre promedio.

Ellos estando entre 8 y 15 años en esa empresa, nunca se les ocurrió preguntar o intentar saber que era esa cosa extraña en forma de bala de plata y oro (el dueño era bastante excéntrico y con mucho dinero) que estaba posicionado en la entrada de los 4 edificios principales en el cual pasaban todos los días.

En retrospectiva no conocen a su cliente.

Esto nos indica algo muy en claro, el cual quiero compartirles de antemano; **El análisis de datos es una capacidad difícil de adquirir.** Sin embargo, no es imposible, sino lo fundamental es irla practicando y como todo hacerla un hábito.

Como lo vamos a lograr nosotros, con pequeños ejercicios y práctica, todos comenzaremos en nivel cero y de allí ir partiendo a un ritmo suave, para ir entendiendo la base del análisis de datos.

**Nuestro Primer Ejercicio**

Vamos a ir al lugar asignado a cada compañero de los que estamos en este curso, nos paramos 5 segundos máximo enfrente de cada escritorio, no abrimos nada solo observamos.

Lo primero es que vamos a deshumanizar el lugar, hagamos de cuenta que no lo conocemos y que vamos a perfilar un lugar de un extraño, claro existe ya un sesgo previo al conocer al compañero, pero tal vez en este ejercicio notemos algo que habíamos pasado por alto.

Primera pregunta:

¿Es hombre o mujer quien se sienta en ese espacio actualmente?

Claro al ser un lugar en muchos casos compartidos si somos mas observadores podremos definir si el mismo pertenece a un solo sexo o es compartido con alguien del sexo distinto.

Segunda Pregunta:

¿Es ordenado?

Independientemente si lo comparte o no sabremos si la persona es organizada o no.

Tercera Pregunta:

¿Puede definir una personalidad?

Recordemos que son 5 segundos de observación y la idea es adquirir el habito junto con las técnicas.

Literalmente Facebook necesita un promedio de 25 respuestas para saber su sexo, sus aficiones, su inclinación política y que le gusta ver y comprar.

**¿Como vamos a dividir la capacitación o en que nos vamos a centrar?**

Tal como comentamos en algunos párrafos anteriores el ser humano es extremadamente malo para analizar grandes volúmenes de datos, vamos a utilizar las maquinas y con ellas lenguajes diseñados para esta tarea.

Hoy en día existen múltiples herramientas de código abierto y con licenciamiento.

Entre las de código abierto tenemos:

R: <https://cran.r-project.org/>

RStudio: <https://posit.co/>

Python: <https://www.python.org/downloads/>

Julia: <https://julialang.org/>

Algunas con licenciamiento:

SPSS: <https://www.ibm.com/products/spss-statistics>

STATA: <https://stata.softonic.com/>

SAS: <https://www.sas.com/es_mx/software/on-demand-for-academics.html>

Primero vamos a escoger la herramienta para el desarrollo de nuestro curso, en este caso será R y RStudio y aunque tengo una certificación de hace un par de años en un “Master en Python”, sigo optando por la utilización de R, en este caso no es mas o menos que cualquiera de las otras herramientas antes presentadas, todas tienen pros y contras.

Les comento mi apreciación completamente personal e individual, tengo de trabajar con datos desde que fui DBA en ORACLE en el Ministerio de Finanzas en 1996 hasta la actualidad, mi carrera siempre a estado alrededor de los datos, pasando como comenté anteriormente desde DBA, consultor en Business Intelligence, Analista de Datos y actualmente Científico de Datos.

Python es uno de los lenguajes que esta ganando a pasos agigantados espacio dentro del análisis de datos y se debe a que conserva mucho en común con los sistemas tradicionales de programación, es decir la ciencia de datos y el análisis de datos en estos momentos está siendo tomada por aquellos que tienen una mayor relación con el lenguaje computacional y por lo mismo lo que se adapta más fácil a ellos es precisamente Python.

¿Ahora un programador puede ser analista de datos?

En este caso me viene a la mente una frase que escuche ya hace algún tiempo y queda como anillo al dedo en esta ocasión, fue en la película animada de pixar “RataTouille” en donde dice lo siguiente:

«Cualquiera puede cocinar», sostiene el gran Gusteau, quien, incluso ya fallecido, inspiró con su arte y consejo a Remy, una sencilla rata de campo, hasta convertirla en chef del más reputado restaurante de París. En un primer sentido, esta frase viene a decir que, con buena voluntad, constancia y un buen maestro cualquiera puede llegar a cocinar dignamente. Sin embargo, la frase tiene un sentido más profundo, que nos revelará el exigente crítico Anton Ego hacia el final de la película: «No cualquiera puede convertirse en un gran artista, pero un gran artista sí puede provenir de cualquier lado».

La llamada de Gusteau, por tanto, va dirigida a todos los que quieren aprender a cocinar, pero, sobre todo, va dirigida a quien siente una vocación secreta en lo más hondo de su corazón y no se atreve siquiera a pensar que su sueño pueda convertirse en realidad. «Cualquiera puede cocinar, pero sólo los audaces pueden llegar a chef», completa Gusteau en otro momento de la película.

Texto tomado de: <https://www.dialogicalcreativity.es/2014/08/ratatouille-cualquiera-puede-cocinar.html>

Es decir, el análisis de datos no está reservado únicamente para aquellos que estudian informática, sino por el contrario está dirigido a aquellos que quieren estudiar los distintos métodos de recopilación, procesamiento y análisis de datos, para que nos permita entender, interpretar y predecir fenómenos reales.

Generalmente a cualquiera que tenga estudios de programación, R se le hace muchísimo más difícil que Python y es porque sencillamente R no fue diseñado por programadores, sino por matemáticos y estadísticos que querían hacer un lenguaje fácil y comprensible para ellos, y teniendo esto en mente nace el lenguaje a partir de la base de uno anterior llamado S (de Statistics), el cual era usado para tales menesteres.

Por este motivo nos embarcaremos a aprender las técnicas del análisis de datos en este lenguaje, la ventaja es que, al haber únicamente un informático puro en el grupo, el resto tendrá muy poco que aprender de programación y nos orientaremos específicamente al análisis de datos desde una herramienta no hecha por y para programadores.

Esto lo vamos a dividir en fases que serán algo más o menos así:

**Fase I**

**Aprenderemos lo básico**, desde instalar, desinstalar el programa, actualizarlo, ver la interfase presentar y comenzar a aprender la utilización de las hojas de trucos.

Veremos como se dividen los datos, desde su base hasta sus modelos más complejos, desde allí comenzaremos con pequeñas prácticas de análisis sencillas, como uno mas uno es igual a dos y así ir poniendo complejidad de una manera escalar y comprensible.

**Fase II**

**Domado de Datos**, esta es un área compleja ya que nos dará la capacidad de extraer, transformar y cargar los datos desde cualquier fuente, ya sea interna o externa. Aquí se queda la mayoría de las personas acorde a mi experiencia, sin embargo, no podemos darnos este lujo por lo que está será la fase mas larga y delicada de todo este proceso.

**Fase III**

**Visualización de los datos**, tomaremos nuestra primera introducción a los modelos estadísticos y de que forma aplicar las geometrías primero para nuestro uso, segundo para elaborar un proyecto y tercero para presentación al cliente final.

**Fase IV**

**Estadística**, veremos los términos, leyes y todo lo referente a modelos predictivos, regresiones lineares, paradojas, distribuciones normales, bimodales, multimodales y de poisson.

**Fase V**

**Algoritmos**, machine y Deep Learning, aplicación de inteligencia artificial para el análisis de los datos.