ITESO

Universidad Jesuita
de Guadalajara

PROPUESTA DE SEGMENTACIÓN

Elaboró:

Ariadna Galindo

Guillermo Campollo

Betsy Torres

Daniel Garcia

Índice

1. Problemática

   Hace 5 años los alumnos de la sociedad de alumnos de la carrera de Ingeniería Financiera comenzaron a hacer un concurso para la carrera en donde se busca una experiencia cercana al ámbito laboral. En su quinta edición nosotros decidimos participar y llevarlo de la mano con nuestra materia “Ciencia de Datos e Inteligencia de Negocios”.

   La división de Watson Health de IBM tiene una segmentación por proyectos y segmentos muy compleja, la cual ralentiza el proceso de análisis financiero requerido cada cierre de mes. Esta división nos proporcionó una base de datos que es llenada por personas de todo el mundo y tienen que analizar periódicamente. Hay diferentes parámetros en diferentes columnas que determinan la segmentación y están en una hoja de cálculo diferente en el mismo Excel.

   Utilizando las herramientas de nuestras clases, principalmente Ciencia de Datos, vamos a generar una propuesta para la segmentación de las bases de datos. Nuestra propuesta seguirá los lineamientos que nos ha dicho la empresa. De esta manera podremos automatizar el proceso de segmentación y las decisiones las tomará la máquina en lugar de una persona.

2. Propuesta
   2.1. Entender la base de datos

La base de datos proporcionada consta de dos hojas de cálculo que son trabajadas en Excel. En la primera hoja son datos que se llenan mes con mes y se hace la segmentación a mano. Mientras que la segunda hoja contiene los parámetros de segmentación, donde si en la columna A encuentras el dato de B entonces el segmento es C. A continuación, vamos a enlistar detalladamente los problemas por cada una de las hojas.

- Datos
  - Son ingresados por muchas personas, algunas tienen un programa, otras no. Por lo tanto, los datos no siempre son presentados igual (ej. M90 y 000M90) a parte de los errores que pueden ingresar las personas al teclear algo mal.
  - Las columnas de Cost, Maj, Minor y Maj_Minor son datos aleatorios los cuales no nos sirven para generar conclusiones.
  - Columnas idénticas o similares. Tenemos que decidir con cuáles nos quedamos.

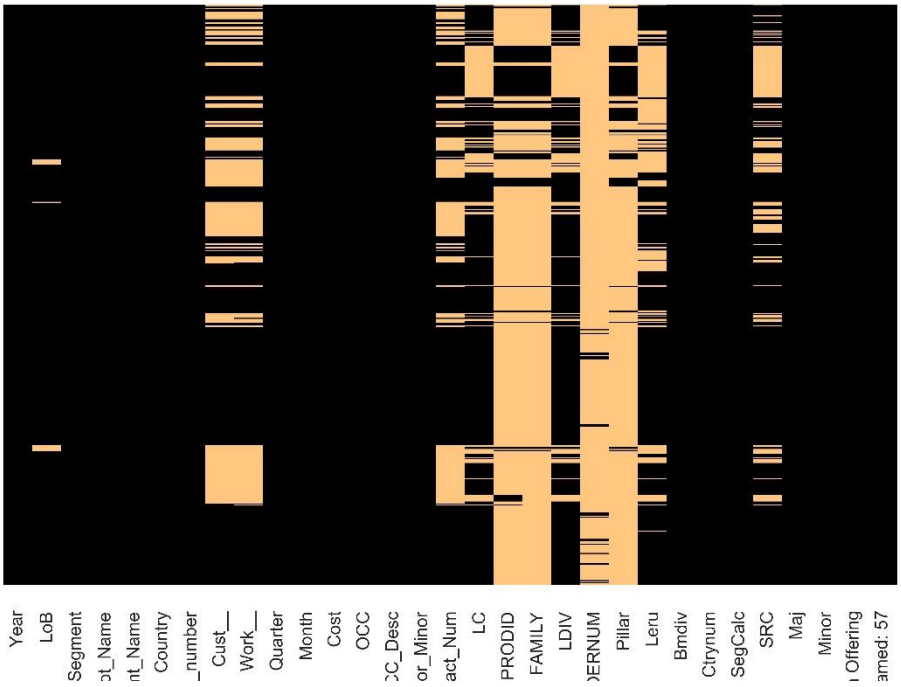
- Columnas con la misma información (ej. Quarter).
- Falta de información.
- Datos con números y letras.
- Datos sin segmentación.
- Una segmentación no necesariamente correcta.

- Parámetros
  - Truven Simpler y Truven simpler están como dos segmentos diferentes.
  - Hay muchos datos que se sobre escriben para la segmentación, ¿a cuál darle más importancia?
  - Datos con letras, números y caracteres varios.

2.2. Datos a manejar

(Justificación de los despreciados)

Para dar ilustrar rápidamente nuestro problema de falta de información presentamos la siguiente gráfica donde los colores claros son los datos faltantes de la base de datos.

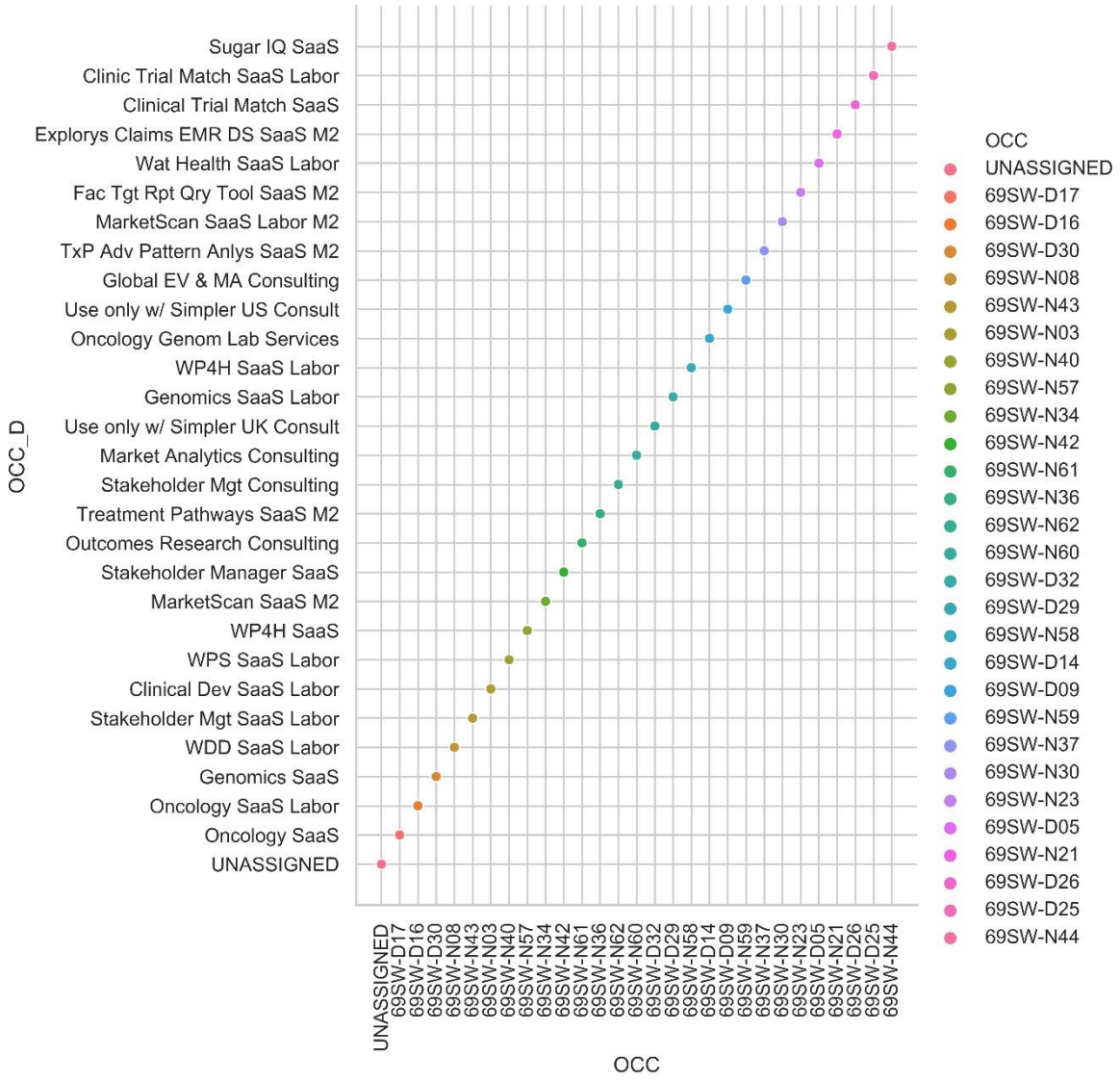


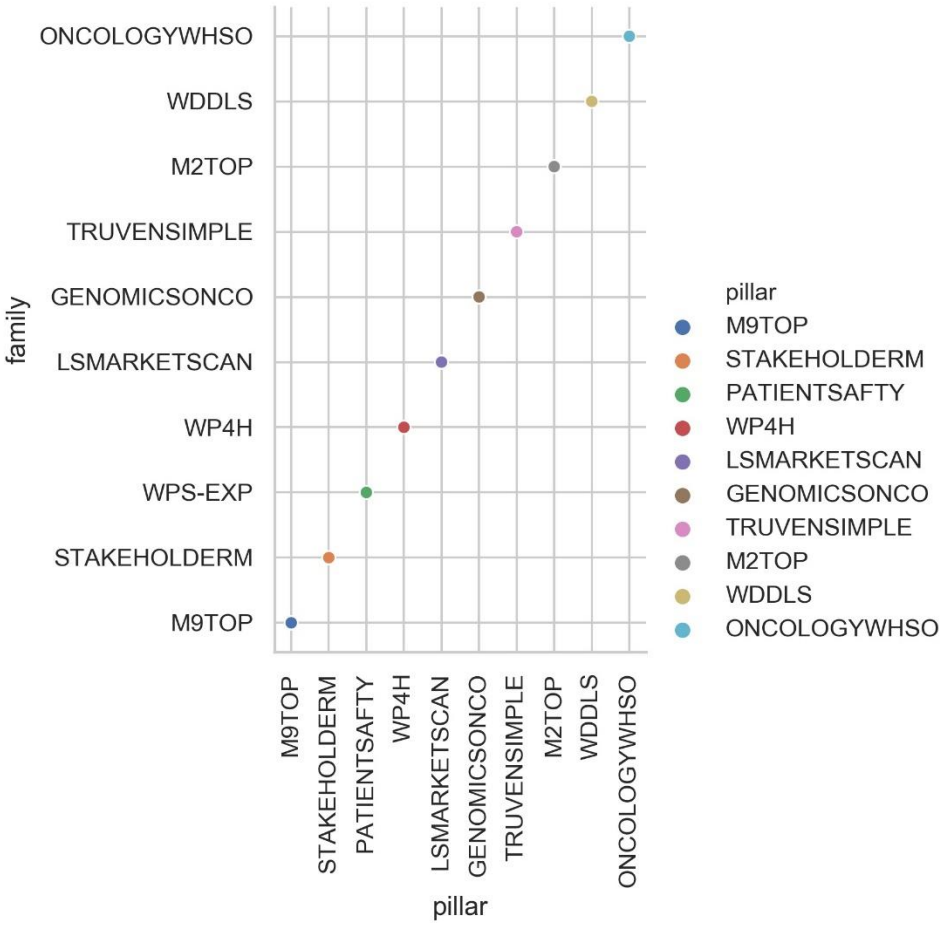
Se observa que hay columnas que no les hacen falta datos, mientras que en otras hay muy pocos datos. Lo primero que haremos será decidir las columnas que no analizaremos en este momento ya que son datos repetitivos.

Como la base proporcionada es de un solo mes, este es igual en toda la columna, también el año y el periodo. Por lo tanto, no son relevantes en nuestro análisis y las

despreciaremos. Sin embargo, si analizamos una base de datos más extensa pueden ser necesarios y tendrán que ser reintroducidos. Dentro de “Segment” solo tenemos Healthcare porque es la división de IBM, por lo tanto, no es relevante.

También encontramos columnas que tienen datos que se ven diferentes pero que al parecer significan lo mismo. Lo comprobamos al graficarlos y vemos un comportamiento lineal porque están en la misma proporción entre sí.





En general, despreciamos todas las columnas que no se encuentran en los parámetros. Aunque encontramos que OCC y OCC_Desc tienen la misma información, ambos tienen parámetros diferentes que definen la segmentación y tuvimos que dejarla para el análisis.

Al final de nuestra segregación de datos podemos observar a simple vista que ahora queda más pequeña.

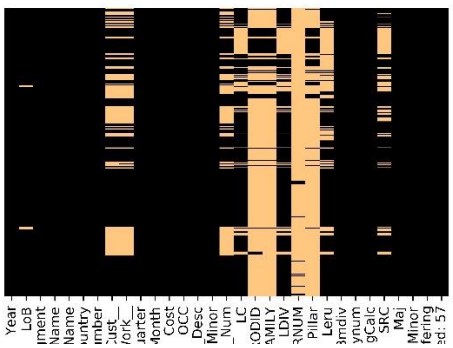


Ilustración 1: Antes de la segregación

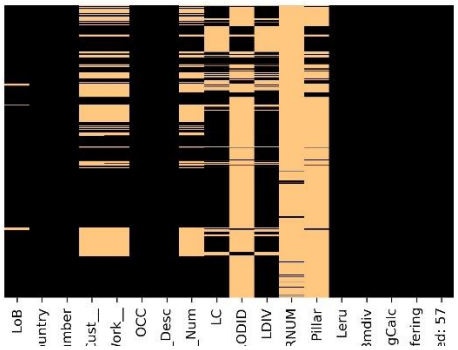


Ilustración 2: Después de la segregación

2.3. Diccionario

Una vez segregados los datos irrelevantes decidimos trabajar con los parámetros. El primer problema que nos dimos cuenta fue que hay demasiados y muy probablemente hayan repetidos o incluso se encimen los datos. Esto nos está generando errores en la segmentación.

Nuestra primera decisión fue darles importancia a los parámetros de “más tiempo” y suponemos que estos se encuentran más arriba en la tabla de los parámetros. Le dimos esta jerarquía de decisión en un diccionario que guardara los parámetros por cada columna de la base de datos.

2.4. Propuesta de segmentación

3. Análisis de resultados

Tendremos resultados diferentes porque nuestras decisiones las hace la máquina y en IBM lo hace una persona.

4. Conclusiones

Hay más parámetros de los que se utilizan y estos están revueltos.