

**Herramientas de Software para Big Data**

**M7A ID**

**Obligatorio 2022:**

**Sistema de análisis de información de compañías bajo Big Data**

**245650 – Fernanda Secinaro**

**241896 – Juan Pablo Rodríguez Sotto**

Índice

[Introducción 2](#_Toc44772675)

[Planteamiento del Problema 2](#_Toc44772676)

[Objetivo General 2](#_Toc44772677)

[Objetivos Específicos 3](#_Toc44772678)

[Alcance 3](#_Toc44772679)

[Propuesta de Valor 3](#_Toc44772680)

[Metodología 4](#_Toc44772681)

[Ingeniería de atributos 5](#_Toc44772682)

[Análisis cualitativo 5](#_Toc44772683)

[Análisis de nulidad y componentes estadísticos 6](#_Toc44772684)

[Análisis de correlación 6](#_Toc44772685)

[Arquitectura de la Solución 7](#_Toc44772686)

[Conclusiones 8](#_Toc44772687)

[Restricciones 9](#_Toc44772688)

[Bibliografía 10](#_Toc44772689)

# Introducción

El presente documento consiste en una especificación del caso de negocio abordado para la realización del trabajo obligatorio de la materia Herramientas de Software para Big Data, más una descripción detallada de las operaciones realizadas a lo largo del análisis.

Luego de analizar varios dataset, se decidió optar por (<https://www.kaggle.com/datasets/peopledatalabssf/free-7-million-company-dataset?select=companies_sorted.csv>)

Consta de un archivo, llamado companies\_sorted.csv que cuenta con 7.173.438

|  |  |
| --- | --- |
| CAMPO | DESCRIPCIÓN |
| id | Id que identifica a la compañía |
| Name | Nombre de la compañía |
| domain | Sitio web de la compañía |
| Year\_founded | Año de fundación de la compañía |
| industry | Industria a la cual pertenece la compañía |
| Size range | Rango de tamaño |
| locality | Localidad (ciudad, estado, país) |
| country | País donde la compañía reside |
| Linkedin\_url | URL del linkedin asociado a la compañía |
| Current employee estimate | Cantidad actual de empleados estimados |
| Total employee estimate | Estimación de empleados totales |

# Planteamiento del Problema

En base a los datos disponibles se intentará analizar un volumen de aproximadamente siete millones de compañías, las cuales presentan una gran variedad de países, industrias, tamaños, etc. Analizar este volumen de información, sin un sistema de análisis como este, sería casi imposible por la cantidad de tiempo que requeriría. Pero nuestro sistema permite incrementar exponencialmente la velocidad, llegando a resultados en un tiempo razonable.

En base a esto intentaremos determinar la industria principal de cada país y de todos los países en general, analizar la palabra más repetida en los nombres, la distribución de la primera letra del nombre de la compañía, determinar las principales localidades donde residen las compañías y determinar cuáles son las indutrias de mayor tamaño. Además, analizar la tendencia de qué tipo de industria se crea por año.

Los resultados de este análisis pueden ser muy útiles para un emprendedor que quiera crear una nueva start-up o una compañía ya existente que quiera hacer un rebranding. Por ejemplo, a partir de este análisis, puede determinar cuáles son las mejores localidades o países para una industria en particular.

**volumen, variedad y velocidad.**

# Objetivo General

Desarrollar un sistema de análisis bajo big data en near to real time para determinar las principales tendencias de las compañías, y a partir de esto, poder brindarle información útil a un emprendedor o compañías ya existentes.

# Objetivos Específicos

Se plantean los siguientes objetivos específicos:

* Realizar ingeniería de atributos, analizando y determinando atributos que pueden tener mayor poder explicativo.
* Desarrollar una arquitectura de big data que permita el análisis de las compañías.
* Analizar las tendencias de las compañías.
* Proponer acciones y políticas en base a los resultado obtenidos.

# Alcance

El presente trabajo representa un ensayo teórico, basado en datos reales.

Su objetivo es validar los conceptos, metodologías y herramientas vistas a lo largo del curso, y por lo tanto, se limitará a:

* Determinar la industria principal de cada país y de todos los países en general
* Analizar la palabra más repetida en los nombres
* Analizar la distribución de la primera letra del nombre de la compañía
* Determinar las principales localidades donde residen las compañías
* Determinar cuáles son las indutrias de mayor tamaño
* Analizar la tendencia de qué tipo de industria se crea por año.

No se incluye en este enfoque:

* Evaluación de las mejores herramientas a utilizar para este tipo de análisis.
* Verificación o comprobación de los resultados obtenidos.
* Mejoramiento o enriquecimiento de los datos en caso de ser necesario.
* Análisis de inversión y propuesta de plan de acción para la implementación de las acciones sugeridas.

# Propuesta de Valor

En base a un volumen de aproximadamente siete millones, y con datos con una gran variedad, es que se va a poder determinar estadísticas varias que podrían ser cruciales para el nuevo emprendedor.

Nuestro enfoque propone un abordaje metódico y estadístico que posibilite acciones a la hora de crear o hacer un rebranding de una comañía. Sus resultados, combinados con el juicio experto de profesionales en la materia, pueden ayudar a crear nuevas y mejoras start-ups en diferentes industrias de todo el mundo. Además de que también provee información para mejorar las compañías ya existentes.

**Debe incluir volumen, variedad y velocidad**

# Metodología

El presente trabajo se realizará utilizando la metodología de desarrollo ágil denominada Scrum. La misma se estructura en base a períodos de trabajo establecidos, denominados sprints, los cuales cuentan con un conjunto de tareas y actividades a realizar dentro del mismo. En este caso el scrum master será Pedro Bonillo y se realizarán cuatro sprints:

1. Determinación de objetivos y alcance del trabajo.
2. Armado de arquitectura de big data.
3. Armado de algoritmo de análisis.
4. Análisis de resultados y armado de conclusiones.

Los entregables de este trabajo son:

* Informe de trabajo realizado.
* Documentación de ingeniería de atributos realizados de forma “externa”
* Presentación power point para defensa.
* Jupyter Notebook con implementación del modelo.

# Plan de Trabajo

Al momento de realizar la tarea se realizan las siguientes actividades:

* Documentación
* Depuración de datos (ingeniería de atributos) y selección de atributos relevantes.

Este proceso se realiza de forma independiente al streaming con un desarrollo en Python.

# Ingeniería de atributos

A continuación se realizan los siguientes análisis sobre los atributos del dataset:

* Análisis cualitativo de las variables a utilizar, en base a la presunción de impacto en el modelo.
* Análisis de campos nulos y componentes estadísticos.
* Análisis de correlación entre variables para evitar errores de ponderación en el modelo.

## Análisis cualitativo

|  |  |
| --- | --- |
| **Atributo** | **Descripción** |
| Accident\_Index | Index - No aplica para el análisis |
| Location\_Easting\_OSGR | No aplica - Se utilizarán otros atributos para determinar la ubicación |
| Location\_Northing\_OSGR | No aplica - Se utilizarán otros atributos para determinar la ubicación |
| Longitude | Se utilizará para determinar hotspots en base a 3 puntos |
| Latitude | Se utilizará para determinar hotspots en base a 3 puntos |
| Police\_Force | No aplica - Asumimos no tiene poder explicativo |
| Accident\_Severity | No aplica - Se evaluarán los accidentes independientemente de su severidad |
| Number\_of\_Vehicles | No aplica - Se evaluarán los accidentes independientemente del número de vehículos |
| Number\_of\_Casualties | Requerido para identificar |
| Date | No aplica - Asumimos no tiene poder explicativo |
| Day\_of\_Week | Se utilizará para determinar día de la semana con mayor ocurrencia de accidentes |
| Time | No aplica - Se decide utilizar la variable condiciones de visibilidad en vez de la hora |
| Local\_Authority\_(District) | No aplica - Asumimos no tiene poder explicativo |
| Local\_Authority\_(Highway) | No aplica - Asumimos no tiene poder explicativo |
| 1st\_Road\_Class | No aplica - Asumimos no tiene poder explicativo |
| 1st\_Road\_Number | No aplica - Asumimos no tiene poder explicativo |
| Road\_Type | No aplica - Asumimos no tiene poder explicativo |
| Speed\_limit | Se utilizará para determinar si la velocidad tiene impacto en la ocurrencia de accidentes |
| Junction\_Detail | No aplica - Asumimos no tiene poder explicativo |
| Junction\_Control | No aplica - Asumimos no tiene poder explicativo |
| 2nd\_Road\_Class | No aplica - Asumimos no tiene poder explicativo |
| 2nd\_Road\_Number | No aplica - Asumimos no tiene poder explicativo |
| Pedestrian\_Crossing-Human\_Control | No aplica - Asumimos no tiene poder explicativo |
| Pedestrian\_Crossing-Physical\_Facilities | No aplica - Asumimos no tiene poder explicativo |
| Light\_Conditions | Se utilizará para determinar el impacto de las condiciones de iluminación |
| Weather\_Conditions | Se utilizará para determinar el impacto del clima |
| Road\_Surface\_Conditions | No aplica - Asumimos no tiene poder explicativo |
| Special\_Conditions\_at\_Site | No aplica - Asumimos no tiene poder explicativo |
| Carriageway\_Hazards | No aplica - Asumimos no tiene poder explicativo |
| Urban\_or\_Rural\_Area | No aplica - Asumimos no tiene poder explicativo |
| Did\_Police\_Officer\_Attend\_Scene\_of\_A | No aplica - Asumimos no tiene poder explicativo |
| LSOA\_of\_Accident\_Location | No aplica - Asumimos no tiene poder explicativo |

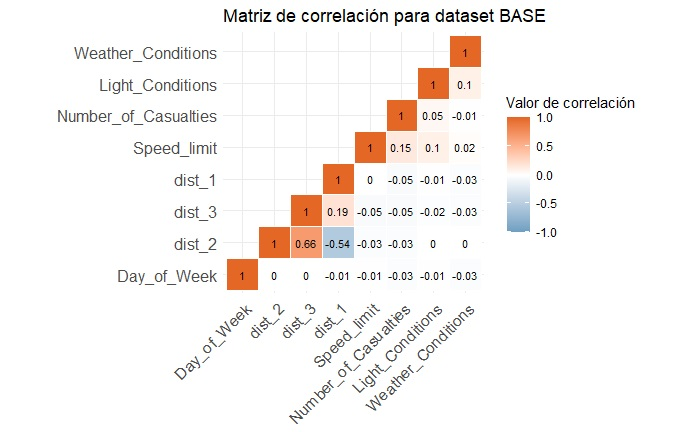
## Análisis de nulidad y componentes estadísticos

Se presentan los principales hallazgos del análisis realizado:

* Depuración de registros nulos
* Detección de outliers
* Depuración de atributos negativos (pues representan datos inexistentes)

## Análisis de correlación

Tal como se comentó previamente se presenta un análisis de correlación para las variables seleccionadas.

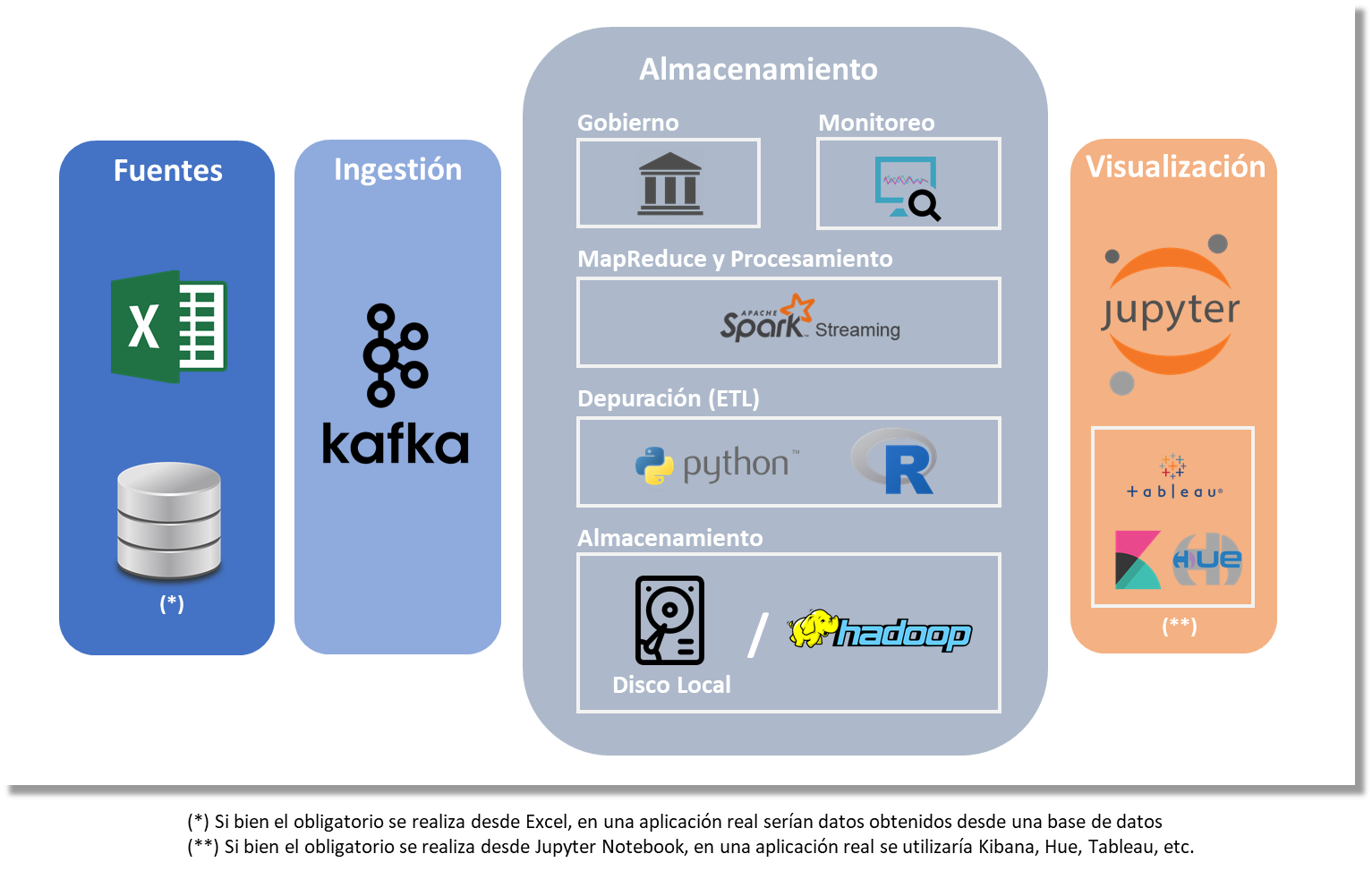


Como se puede apreciar en la imagen, no existe correlación significativa entre las variables, por lo que se procede a realizar el modelo considerando las 6 al mismo tiempo.

**Nota**: Las distancias están correlacionadas entre sí dado que efectivamente son un cálculo en base a longitud y latitud.

# Arquitectura de la Solución

Para lograr los objetivos antes mencionados se propone la siguiente:



Pasos del proceso:

1. Utilizando la herramienta R se realiza la ingeniería de atributos, seleccionando 6 atributos significativos.

Este paquete estadístico nos permite automatizar el análisis, aplicando funciones y programas pre desarrollados que facilitan sustancialmente la tarea. Adicionalmente es una herramientas abierta y robusta, que no requiere inversión en licencia.

1. Se toman los datos desde archivo formato plano y se procesan en Kafka para simular un streaming de datos. Se realiza la lectura del archivo y se publican los datos a través del productor de Kafka.

Se utiliza Kafka debido a su capacidad de procesar un streaming de datos en alta performance y dada su potencialidad de integración con Python para la aplicación de modelos de Machine Learning.

1. Se utiliza Spark Streaming y MLlib para procesar el archivo, generando un modelo no estructurado de streaming y ejecutando un modelo de clusterización KMeans.

Se utilizan las herramientas de Spark dada su robustez y amplia documentación acerca de la aplicación de algoritmos de clustering (Kmeans). Existen diversos artículos al respecto de su validez para soluciones similares al caso planteado.

1. Se realiza un ploteo de los datos para visualizar los centroides obtenidos.

Se utilizan las herramientas de ploteo de Jupyter Notebook debido a su practicidad y uso extendido a lo largo de la industria.

NOTA: Se selecciona un modelo de streaming no estructurado debido a su simpleza de implementación y a que no es necesario contar con un check point en el caso de uso seleccionado.

# Conclusiones

Las conclusiones obtenidas son:

**Realizar ingeniería de atributos, analizando y determinando atributos que pueden tener mayor poder explicativo para aplicarlos al modelo.**

Luego de un análisis cualitativo y cuantitativo de los atributos se decidió seleccionar los siguientes atributos:

* Longitude
* Latitude
* Day\_of\_Week
* Speed\_limit
* Light\_Conditions
* Weather\_Conditions

Latitude y Longitude permiten ubicar geográficamente los hotspots de ocurrencia de accidentes y las demás variables nos permiten categorizar su ocurrencia en base a iluminación, clima, velocidad permitida y día de la semana.

Adicionalmente fue necesario depurar menos del 1% de los datos disponibles por casos de nulidad o valores inconsistentes (Ej: Velocidad negativa).

Finalmente, se determinó que las variables seleccionadas no cuentan con correlación significativa, por lo que se puede proseguir con el análisis.

**Desarrollar una estructura de streaming de datos no estructurado que permita procesar la información en modalidad near to real time.**

La arquitectura propuesta y las herramientas seleccionadas demostraron ser útiles para el procesamiento del streaming de datos requerido. Pese a lo anterior, existe un conjunto de herramientas adicionales (pertenecientes y no pertenecientes al ecosistema de Apache Hadoop) que pueden potenciar este tipo de trabajos y brindar mejores niveles de performance, precisión y capacidad de visualización.

Es importante destacar que algunas de las tareas de ETL y análisis de datos se realizaron por fuera de la estructura del streaming debido a la restricción de tiempo antes mencionada. Es posible mejorar sustancialmente el modelo incorporando dichos pasos al streaming y ajustando los parámetros del algoritmo en base a la totalidad de los datos.

**Determinar los hotspots de ocurrencia de accidentes, utilizando algoritmos de Machine Learning adecuados (Spark MLlib).**

Los clusters obtenidos son los siguientes:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster** | **# accidents** | **avg casualties** | **avg speed limit** | **avg light cond.** | **avg w cond.** |
| **1** | 11 | 1.18 | 10.00 | 2.00 | 1.63 |
| **2** | 7,324 | 1.18 | 19.98 | 1.72 | 1.52 |
| **3** | 667,632 | 1.28 | 30.00 | 1.84 | 1.58 |
| **4** | 86,144 | 1.44 | 40.00 | 1.94 | 1.62 |
| **5** | 287,464 | 1.53 | 61.69 | 2.23 | 1.70 |

Del análisis de los mismos se puede concluir:

1. Los clusters número 3 y 5 agrupan la mayoría de los accidentes de tránsito, representando cerca del 90% de los casos.
2. Alienado con la hipótesis previa, se detecta una tendencia de aumento de severidad de los accidentes (mayor cantidad de víctimas fatales) a medida que aumenta el promedio de velocidad de los vehículos involucrados.
3. Las condiciones de iluminación y clima tienen el mismo impacto en cada una de las zonas identificadas.

**Proponer acciones y políticas a aplicar para la disminución de la ocurrencia en base a los resultados obtenidos.**

En base a los resultados obtenidos previamente se proponen las siguientes acciones:

* Disminuir los límites de velocidad en ubicaciones con alta concentración de accidentes.
* Asegurar disponibilidad de servicios médicos de forma rápida en ubicaciones donde no se pueda disminuir el límite de velocidad.
* Realizar campañas de concientización en medios para sensibilizar a la población, en especial a los que transitan en esas ubicaciones (cartelería, realidad aumentada, etc.).

Al ajustar estos parámetros se deberá analizar si los resultados varían o no, lo cual generará uno de dos escenarios:

* No existe variación: Esto implicaría que las variables seleccionadas no son las correctas, por lo que se deberá realizar nuevamente la ingeniería de atributos, seleccionando nuevos atributos para ejecutar el modelo y analizar sus resultados.
* Existe variación: Esto implicaría que las variables eran acertadas y se propone continuar con el análisis. La mejoría de las cifras en los centroides establecidos generará variaciones en los datos que permitirán obtener nuevos centroides sobre los cuales se deberá aplicar acciones correctivas.

La conclusión general del trabajo es que aplicando las herramientas y metodologías antes mencionadas es posible determinar los posibles hotspots, basados únicamente en datos, que permitan a las autoridades definir políticas y procedimientos que tiendan a disminuir el número de accidentes y consecuentemente el número de víctimas de los mismos.

# Restricciones

El presente análisis representa un trabajo de laboratorio, con el objetivo de validar los conceptos, metodologías y herramientas vistos en el curso. Sobre este trabajo aplican las siguientes restricciones:

* Tiempo: Existe escasa disponibilidad de tiempo tanto por contar con una fecha de entrega preestablecida, como por no contar con dedicación total a la tarea por parte de los realizadores.
* Teletrabajo: Dada la situación particular generada por la pandemia el curso completo y las sesiones de scrum se realizaron de forma virtual. Se requirió una alta dedicación de tiempo para la instalación y configuración de las herramientas requeridas para el análisis.
* Equipos disponibles: Se trabajó de forma local, con equipos de escasos recursos, lo que limitó la capacidad de procesamiento y aumentó los tiempos considerablemente.
* Experiencia: El equipo de trabajo asignado no cuenta con experiencia relevante en este tipo de análisis, siendo su primera aproximación al concepto práctico de Big Data.

En particular se deberá profundizar en la utilización de algoritmos complejos de Machine Learning sobre un streaming de datos.

# Bibliografía

Materiales de clase ORT:

* <https://aulas.ort.edu.uy/course/view.php?id=3894&section=0>
* <https://aulas.ort.edu.uy/course/view.php?id=3939>
* <https://aulas.ort.edu.uy/course/view.php?id=4303>

Libros

* [Shumway, Stoffer, Time Series Analysis and its Applications (4th ed.)](https://aulas.ort.edu.uy/mod/url/view.php?id=232570)
* Tom M. Mitchell. McGraw-Hill, 1997

Artículos

* <https://www.linkedin.com/pulse/propuesta-de-una-arquitectura-gesti%C3%B3n-grandes-datos-la-bonillo-ramos/>

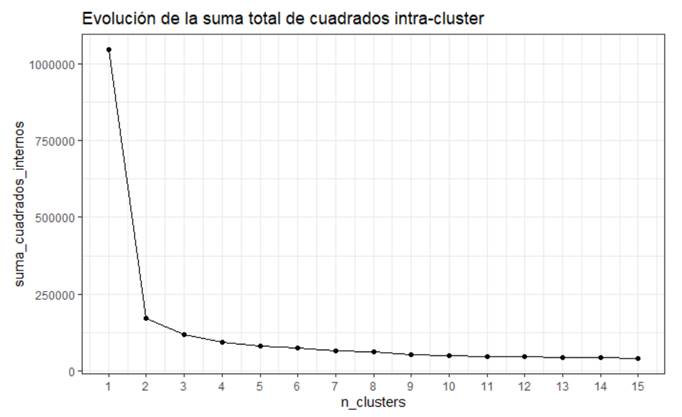
# Anexo 1

Se presentan en este anexo algunos de los resultados del procesamiento de datos realizado:

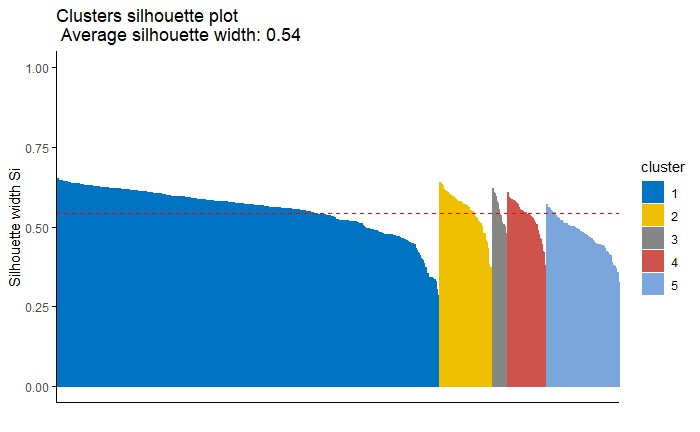
1. Los puntos seleccionados para el KMean son:

* lat\_1 <- 49.91294 - lon\_1 <- -7.516225
* lat\_2 <- 60.75754 - lon\_2 <- 1.762010
* lat\_3 <- 60.75754 - lon\_3 <- -7.516225

1. Evaluación por herramienta del codo:



1. Evaluación por herramienta de silueta:



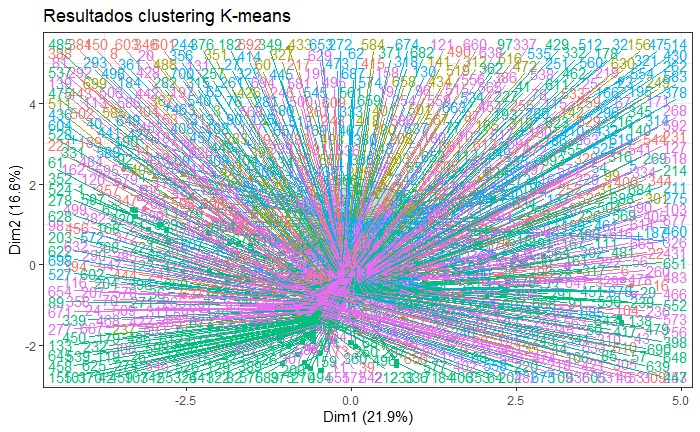
1. Centros obtenidos:

|  |
| --- |
| **CENTROS** |
| x1 = c(3.675580,  5.49205,  5.590802,  4.00,    10.00,     1.00,     1.00,      1.00) |
| x2 = c(6.875082,  8.31552,  9.613966,  4.26293, 20.00,     1.771551, 1.484913,  1.165948) |
| x3 = c(6.925740,  8.64506, 10.088608,  4.14049, 30.00,     1.858326, 1.442504,  1.292282) |
| x4 = c(6.749553,  8.41970,  9.634581,  4.16604, 40.00,     1.934323, 1.550542,  1.441462) |
| x5 = c(6.743433,  8.19988,  9.285921,  4.04869, 61.723326, 2.190364, 1.574882,  1.594636) |

1. Resumen de resultados sobre datos etiquetados:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **# accidents** | **avg casualties** | **avg speed limit** | **avg light cond.** |
| **1** | 11 | 1.18 | 10.00 | 2.00 |
| **2** | 7,324 | 1.18 | 19.98 | 1.72 |
| **3** | 667,632 | 1.28 | 30.00 | 1.84 |
| **4** | 86,144 | 1.44 | 40.00 | 1.94 |
| **5** | 287,464 | 1.53 | 61.69 | 2.23 |

1. Ploteo de resultados:



Nota: Si bien los centros existen, esta herramienta de ploteo no es adecuada debido a que con esta cantidad de datos, no se pueden visualizar los mismos.

A modo de ejemplo, se presenta el mismo gráfico presentando menor cantidad de datos:

