Data Understanding I

David Guzmán Leyva - A01706417
Enrique Santos Fraire - A01705746
Leonardo Alvarado Menéndez - A01705998
Oscar Enrique Delgadillo Ochoa - A01705935

Índice

Obtener los datos iniciales

Reporte inicial de recolección de datos

- Su ubicación
  - Link a data set
- Los métodos usados para obtener o adquirir los datos
  - Los datos fueron recolectados por la telefonía movistar en Santiago de chile, estos datos fueron pre-procesados por el Instituto de Data Science UC y se nos entregaron.
- Los problemas para obtener los datos
  - En nuestro caso no hubo problemas en la obtención de datos debido a que se nos entregaron completos y semilimpios.
- Las soluciones a los problemas
  - Aún no se han encontrado problemas en la recolección de datos.

Descripción de los datos

Reporte de descripción de los datos

PHONE_ID

- type……..object
- Valid……..49618132
- Mismatched…….0
- Missing…….0
- Unique…….1353435

timestamp

- type……...object
- Valid……..49618132
- Mismatched…….0
- Missing…….0
- Unique…….86400
- min……...2021-01-01 00:00:00-03:00
- max……..2021-01-01 23:59:59-03:00

bts_id

- type……..object
- Valid……..49618132
- Mismatched…….0
- Missing…….0
- Unique…….1871
- Most common….MORRF

lat

- type…….float
- Valid……49618132
- Mismatched……..0
- Missing……..0
- Unique……..1198
- min…….-3.402680*10^1
- max……-3.292550*10^1

lon

- type……...float
- Valid……..49618132
- Mismatched…….0
- Missing…….0
- Unique……..1264
- min…….-7.148880*10^1
- max……-7.005880*10^1

Exploración de los datos

Reporte de exploración de los datos

Dada las características de nuestros datos iniciales, no es posible realizar algún tipo de visualización que aporte valor de primera mano. Tenemos 3 variables de tipo objeto que representan valores únicos y las 2 variables restantes de tipo flotante son dependientes la una de la otra para tener significado, pues se trata de coordenadas, por lo que la obtención de medidas estadísticas se ve dificultada.

Verificar la calidad de datos

Reporte de la calidad de datos

Se realizaron las siguientes modificaciones en los datos:

- Cambiar el bts_id original por otro atributo llamado coordenadas, para tener un registro correcto de las antenas, debido a que en algunos registros de antenas, había más de una antena en la misma localización y había antenas con el mismo nombre pero en localizaciones distintas.
- Eliminar registros en los cuales la velocidad de movimiento que tuvieran entre antenas fuera mayor a 150 km, porque es improbable para una persona común alcanzar esa velocidad en la zona metropolitana de Santiago de Chile.
- Eliminar registros de personas en las que no se observaba ningún movimiento durante el día, debido a que las antenas también podían hacer registro de máquinas, esto crea un sesgo en los datos que tenemos.

Modelo de almacenamiento

El modelo de almacenamiento que aplicamos para guardar nuestros datos es del tipo filestore, haciendo uso de google drive.
En un entorno real con una cantidad muchísimo mayor de datos, lo ideal es usar un sistema de almacenamiento de AWS (Amazon Web Services) con AWS RDS, que nos da la opción de escalabilidad en caso aumentar o reducir la cantidad de datos, pudiendo en enlazarlo a un entorno de AWS Elastic Beanstalk.