Análisis de Datos del Sistema de Recolección por Contenedores

Martín Plazzotta - aldo.plazzotta@fing.edu.uy Francisco González- 625francisco@gmail.com

Resumen

En este documento presentaremos el proyecto de final de curso, el cual se desarrolló sobre la plataforma Mongo Atlas y en base a datos provenientes del Sistema de Disposición Final de Residuos de la Intendencia de Montevideo (IM).

I. Introducción

Como proyecto de fin de curso nos propusimos recolectar información sobre la generación de residuos en Montevideo y como esta es atendida por el Gobierno Departamental.

Accedimos a datos provenientes del Sistema de Recolección Lateral por Contenedores. Este es uno de los tantos sistemas y servicios de recolección de residuos existentes en la ciudad, y por tanto la información y conclusiones obtenidas no pretenden reflejar la vasta realidad de la ciudad.

II. MOTIVACION

Nuestra motivación es doble. Por un lado analizar y obtener informacion acerca de la produccion de residuos en la ciudad. Por otro, investigar que posibilidades brinda Mongo Atlas para el análisis y visualización de la información.

Contamos con información de distintos aspectos respecto a la producción de residuos, e intentaremos integralos para generar nuevos datos.

En este sentido y a modo de guía para nuestro proyecto, nos hemos planteado ciertas interrogantes, las cuales intentaremos responder con la información disponible; y de este modo poder generar valor mediante la transformación e integración de los datos obtenidos para la base de datos documental generada final.

Las tareas a atender son:

- Realizar un análisis sobre la producción de residuos, contestando las siguientes interrogantes que se utilizarán como medida de la satisfacción de este análisis:
 - ¿Cómo es la generación de residuos en la ciudad a lo largo del tiempo?
 - ¿ Cuánta basura genera un ciudadano promedio?
 - ¿Distintos municipios generan basura de distinto modo? Por ejemplo, ¿la cantidad de residuos, es uniforme a lo largo del año?, ¿o existen períodos de mayor o menor generación de residuos?
 - ¿Habitantes de distintos municipios generan cantidades de basura distintas?
- Generar un indicador de cuanta demanda de levantes de contenedores genera cada municipio y si están siendo atendidos de forma equitativa (los recursos asignados, la capacidad de recoleccion dedicada a cada municipio es justa).
 - Un indicador en este sentido puede ser el atraso promedio de cada circuito, o municipio. Se puede calcular el mismo para cada fecha y luego analizarlo como varia a lo largo del tiempo.

Para probar que este punto puede ser parcialmente satisfecho por la base de datos se contestará la siguiente pregunta:

• ¿Es el tiempo promedio entre levantes de contenedores el mismo para cada municipio, o este tiempo varía dependiendo del municipio?

III. DESCRIPCIÓN Y OBTENCIÓN DE LOS CONJUNTOS DE DATOS

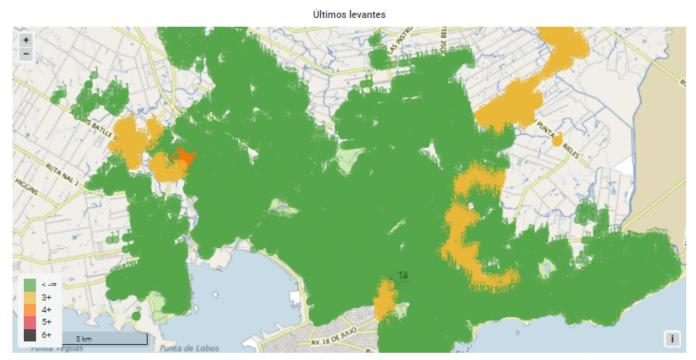
III-A. Juego de datos Últimos Levantes

La IM en su Observatorio de Datos Ambientales publica información relativa a la Limpieza y Gestion de Residuos, entre otros temas.

Allí se ofrece al público información diaria sobre cada contenedor de la ciudad, por ejemplo la cantidad de días que lleva sin ser levantado hasta la fecha actual.

1

Dicha información se presenta en formato de mapa, donde círculo corresponde a un contenedor, y su color corresponde a la cantidad de días que lleva sin ser levantado, como puede ser observado en la siguiente imagen.



Verde: hasta dos días sin recolección. Amarillo: hasta tres días sin recolección. Naranja: hasta cuatro días sin recolección. Rojo: hasta cinco días sin recolección. Negro: más de cinco días sin recolección.

Dichos datos también son presentados en formato tabular para su descarga, como puede ser observado en la siguiente imagen.

Capa ultimo levantes V									
Fecha	Código de recorrido	posicion	gid	Latitud	Longitud	Ultimo día de levante	historico	Días sin levantar	municipio
2022-07-02 06:06:09	G_DU_RM_CL_111	24	129352	-34.8013950	-56.2450900	2022-07-01 00:00:00.0	2022-06-30 2022-05-20 2022-05-1	1	G
2022-07-02 06:06:09	G_DU_RM_CL_108	91	129058	-34.8310930	-56.2276200	2022-07-01 00:00:00.0	2022-05-31 2022-05-24 2022-05-1	1	G
2022-07-02 06:06:09	G_DU_RM_CL_111	16	129349	-34.8032600	-56.2403200	2022-07-01 00:00:00.0	2022-05-20 2022-05-10 2022-06-0	1	G
2022-07-02 06:06:09	G_DU_RM_CL_108	41	129061	-34.8230320	-56.2362440	2022-07-01 00:00:00.0	2022-07-01 2022-06-27 2022-06-1	1	G
2022-07-02 06:06:09	G_DU_RM_CL_111	93	129359	-34.8065000	-56.2484320	2022-07-01 00:00:00.0	2022-05-11 2022-06-20 2022-05-2	1	G
2022-07-02 06:06:09	G_DU_RM_CL_108	11	129059	-34.8201300	-56.2327500	2022-07-01 00:00:00.0	2022-07-01 2022-06-27 2022-06-1	1	G
2022-07-02 06:06:09	G_DU_RM_CL_111	73	129356	-34.8035800	-56.2479600	2022-07-01 00:00:00.0	2022-07-01 2022-06-28 2022-06-2	1	G
2022-07-02 06:06:09	G_DU_RM_CL_111	18	129350	-34.8064160	-56.2440500	2022-07-01 00:00:00.0	2022-05-16 2022-05-06 2022-06-2	1	G
2022-07-02 06:06:09	A_DU_RM_CL_113	81	133872	-34.8680900	-56.2518700	2022-06-30 00:00:00.0	2022-06-30 2022-06-11 2022-06-0	2	Α
2022-07-02 06:06:09	G_DU_RM_CL_111	96	129360	-34.8084640	-56.2489800	2022-07-01 00:00:00.0	2022-06-20 2022-05-22 2022-05-1	1	G
2022-07-02 06:06:09	G_DU_RM_CL_111	97	131552	-34.8077350	-56.2505340	2022-07-01 00:00:00.0	2022-06-21 2022-05-16 2022-07-0	1	G
2022-07-02 06:06:09	G_DU_RM_CL_111	103	129364	-34.8026160	-56.2519260	2022-07-01 00:00:00.0	2022-07-01 2022-06-28 2022-06-2	1	G
2022-07-02 06:06:09	C_DU_RM_CL_104	96	134261	-34.8812750	-56.1662560	2022-07-01 00:00:00.0	2022-06-06 2022-05-06 2022-06-0	1	С

Este conjunto de datos contiene los siguientes atributos:

- Fecha: fecha de genreación de los datos (cada día a las 6 am aproximadamente).
- Código de recorrido: circuito al que pertenece el contenedor.
- Posición: posición del contenedor en su circuito.
- Longitud y Latitud: coordenadas de la posición del contenedor.
- Último día de levante: Fecha de su último levante al momento de generación de los datos.
- Días sin levantar: Cantida de días que lleva sin ser levantado a la fecha de los datos(Fecha Último día).
- Municipio: Municipio al que pertenece el contenedor.
- Zona: El centro comunal al que pertenece.

Inicialmente para obtener estos datos de últimos levantes se implemento un script python. El mismo, mediante el uso de

Selenium Webdriver, emula la acción de un usuario que descarga los datos: se conecta a la url y çlickea. el botón de descarga.

Luego esos datos se suben a la base de dato en Atlas usando mongo import. Este proceso lo corríamos diariamente usando CRON.

Luego observamos la necesidad de obtener estos datos para período considerable de fechas anteriores. En vista de esto, y al igual que con el data set que se explica en la siguiente sección, se solicitó dicha información a la IM.

Así, la Intendencia nos disponibilizó un archivo CSV con información diaria del atraso para cada contendedor de la ciudad.El período de fechas era desde el 16 de febrero al 29 de junio de este año.

Cada día la IM genera un reporte: para cada contenedor (aproximadamente 11 500 contenedores) se informa cuántos de días de atraso tiene al momento de su último levante. Se contaba con este reporte para ciento treinta y tres fechas hacia atrás a partir del 30 de junio. Recibimos toda esta información en un único archivo CSV el cual contiene mas de 1 500 000 registros.

El archivo en cuestión es **atrasos** y se entrega en el conjunto de datos crudos. El esquema de los datos contiene, entre otros atributos, los siguientes:

- COD_RECORRIDO y POSICION: identifican al contenedor
- DIAS ATRASO: al día de la fecha del reporte
- DIA: último día en que fue levantado
- El atributo FECHA (fecha del reporte) no se encontraba en los datos, si bien pudimos calcularlo luego en base a la relación FECHA DIA = DIAS_ATRASO.

Otro problema de calidad de datos, es que tras cargar los datos en una colección de MongoDB, notamos que varios documentos (133 para ser exactos) tenían valores extraños en sus atributos, como se ve en la imagen:

```
id: ObjectId('62c1cc95e1df9d2621c47721')
     LUGAR SALIDA: "LUGAR_SALIDA/"
 2
 3
     ID LUGAR SALIDA: "ID_LUGAR_SALIDA/"
 4
     COD RECORRIDO: "COD RECORRIDO,"
 5
     POSICION: "POSICION,"
 6
     HORA REAL SALIDA: "HORA REAL SALIDA,"
 7
     DIA: "DIA /"
     TURNO: "TURNO,"
 8
 9
     ESTADO: "ESTADO/"
10
     ULTIMA LLEGADA: "ULTIMA LLEGADA/"
11
     DIAS ATRASO: "DIAS ATRASO/"
```

III-B. Conjunto de datos de Pesadas

Los camiones del sistema de recolección lateral de residuos descargan en el cantón cuando se llenan. Cada descarga es pesada y registrada con fecha y hora. También se le asocia un circuito (esto no es del todo correcto ya que la basura descargada puede provenir de distintos circuitos, pero para la funcionalidad de la tarea y de la base de datos generada se asumirá que esta pesada asociada al circuito particular es exclusiva de ese circuito ya que no podemos distiguirla y por lo tanto no podemos dividirla para diversos circuitos).

Se cuenta con información de dichas pesadas. Para cada una se cuenta con los siguientes atributos:

- Fecha y hora de la pesada
- Peso de la basura descargada en kilos.
- Circuito o recorrido asociado.
- Municipio al que pertenece el circuito.

Estos datos fueron solicitados a la IM y amablemente nos fueron provistos en seis archivos CSV, uno para cada mes desde diciembre de 2021 a mayo de 2022.

Si bien la información de pesadas aún no figura en el observatorio de la Intendencia, se nos comunicó que existe la intención de disponibilizarla proximamente.

El data set de pesadas era de buena calidad: no figuraban valores extraños ni repetidos, estaban en formatos adecuados y su tratamiento no trajo complicaciones.

III-C. Otros datos: demográficos

Al momento de diseñar el proyecto se vio que era razonable considerar la población de cada municipio al momento de estudiar la generación de basura.

Para ello se creó en la base de datos la colección MunicipiosPoblacion, con el dato de población para cada municipio según se informa en [5]

IV. LIMPIEZA E INTEGRACIÓN DE LOS JUEGOS DE DATOS

A continuación explicaremos el tratamiento que se realizó a los datos crudos antes de poder utilizarlos para analisis y visualización.

IV-A. Datos de Pesadas

En el caso del conjunto de datos "Pesadasüsamos Mongo Compass para crear la colección "PesadasFullz cargarle los seis archivos CSV con los datos.

También desde Compass se definió el tipo de datos de cada columna: tipo Date para FECHA, tipo Number para PESO y el resto de los campos de tipo String.

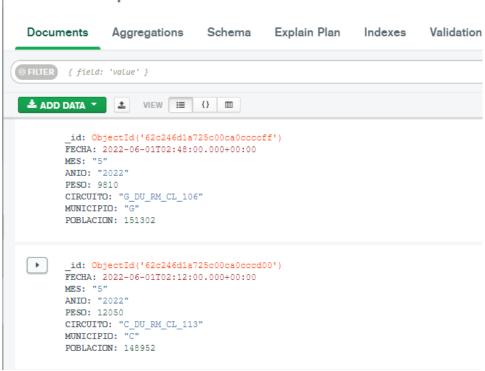
Como mencionabamos, los datos crudos eran de buena calidad, por lo cual esta etapa no presentó complicaciones.

Finalmente se agregó a esta colección la población del municipio de la pesada, trayendo dicho dato desde la colección MunicipiosPoblacion. Para ello se implemtno el siguiente agregation pipeline:

```
ſ
      '$lookup': {
         'from': 'MunicipiosPoblacion',
          'localField': 'MUNICIPIO',
          'foreignField': 'Municipio',
          'as': 'Doc'
      '$project': {
          id': 1,
          'FECHA': 1,
          'ANIO': 1,
          'CIRCUITO': 1,
          'MES': 1,
          'MUNICIPIO': 1,
          'PESO': 1,
          'POBLACION' ·
             '$first': '$Doc.poblacion'
   },
      '$out': 'PesadasFull'
]
```

La siguiente imagen muestra los documentos de la colección PesadasFull.

Recoleccion.pesadasFull



IV-B. Datos de Atrasos

Como comentabamos en la sección anterior, los datos de atrasos contenían valores extranios. Al cargar la colección atrasosFull vimos que algunos documentos tenían los nombres de los atributos como valores de atributos.

Para eliminar dichos documentos debimos utilizar la herramienta mongo shell para conectarnos a la base de datos y ejecutar una instrucción delete con clausula where del tipo DIAS_ATRASO = "DIAS_ATRASO".

Resuelto esto volvimos a cargar la colección atrasosFull y asignar los tipos adecuados a cada atributo.

```
mongosh mongodb+srv://<credentials>@cluster0.lv3pa.mongodb.net/Recolec... - 

Microsoft Windows [Versión 6.3.9600]
(c) 2013 Microsoft Corporation. Todos los derechos reservados.

C:\Users\MyLaptop>mongosh "mongodb+srv://cluster0.lv3pa.mongodb.net/Recoleccion" --apiVersion 1 --username user
Enter password: ************

Current Mongosh Log ID: 62c37a5faa4108259a19fc34
Connecting to: mongodb*srv://credentials>@cluster0.lv3pa.mongodb.net/Recoleccion?appName=mongosh*1.5.0

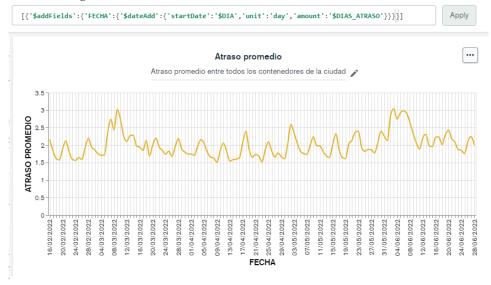
Using MongoDB: 5.0.9 (API Version 1)
Using Mongosh: 1.5.0

For mongosh info see: https://docs.mongodb.com/mongodb-shell/
Atlas atlas-e0gqkn-shard-0 [primary] Recoleccion> db.atrasosFull.deleteMany(< DIMS_ATRASO: "DIAS_ATRASO" >)
```

El otro asunto a atender con este juego de datos es que el atributo FECHA estaba perdido de los datos originales. La tarea de calcular dicho atributo sumando al atributo DIA la cantidad de días dada en DIAS_ATRASO se realizó mediante el siguiente pipeline.

}

A diferencia del pipeline aplicado a la colección PesadasFull, este pipeline no fue ejecutado sobre la colección atrasos, modificandola. En este caso, Charts permitió aplicar el pipeline a los datos unas vez dentro de la visualización, permaneciendo la colección original sin modificar.



V. ANÁLISIS DE DATOS Y RESOLUCIÓN DE LAS TAREAS PLANTEADAS

Una vez obtenidas las colecciones, y planteadas las interrogantes, se buscó generar respuestas a estas preguntas utilizando la base de datos, con ayuda de la herramienta *Charts* de MongoDB Atlas.

Esta ofrece una capa de visualización de los datos. Alli se pueden crear tableros conteniendo distintos tipos de visualizaciones (gráficos, mapas, etc) y agregar filtros en las mismas.

Algo a resaltar, es que dichas visualizaciones tienen la capacidad de calcular .ªl vuelo.ªgregaciones de los valores mostrados. De este modo nos evitamos tener que calcular y almacenar dichos valores en una etapa previa o generar pipeline no triviales facilmente.

Para la resolución de la pregunta "¿Cuánta basura genera un ciudadano promedio?" se un gráfico de barras basandonos en la colección pesadasFull, que fue generado utilizando Charts, pero a modo de ejemplo presentaremos un pipeline que nos devuelve la cantidad de basura generada por número de mes por un ciudadano promedio de Montevideo y cómo se soluciona este problema con la herramienta:

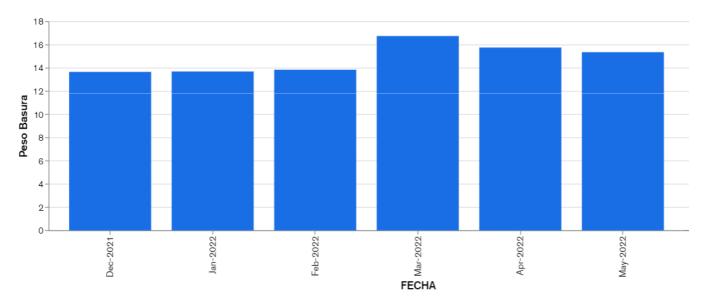
Listado 1 Cantidad de basura por habitante por número de mes.

```
},
{
    "$group":{
        "_id":"$MES","Peso_Mensual":{
        "$sum":"$PESO"
        }
},
{
        "$project":{
            "Anio":"2022","Mes":"$_id", "Peso_Mensual":{"$divide":["$Peso_Mensual",1318736]}, "_id":0
        }
},
]
```

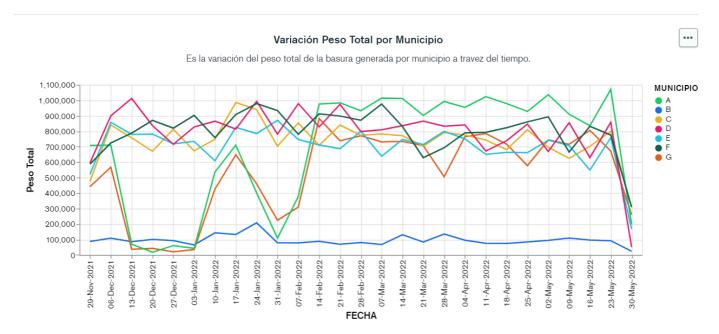
Que como podemos observar no es un pipeline trivial, mientras que con la utilización de MongoCharts, con el mismo conjunto de datos, se selecciona en el eje de las x el atributo "FECHA", luego se genera un nuevo campo llamado BasuraTotalPorHabitante eligiendo la opción de "Calculated", y pasando como valor de la expresión PESO/318736 que es el número total de la población de Montevideo, y luego se toma este nuevo campo como eje y. Se agrupan las fechas por mes y el campo calculado se agrupa como la suma(acumulación). Notese la facilidad en la utilización de la herramienta. Obtenemos como resultado el siguiente gráfico:

Peso de Basura por Mes por Habitante

Es el peso de la basura de todas las pesadas de un mismo mes dividido por la cantidad de población.



Por lo que contestando la interrogante un ciudadano promedio genera entre 13 y 17 kilos de basura por mes. Continuando con las interrogantes planteadas, se busca analizar la variación de la basura total generada por los distintos municipios a lo largo del tiempo. Para ello presentaremos la siguiente gráfica:

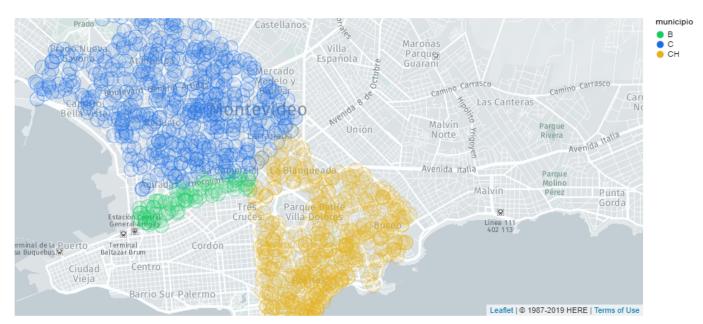


Se puede observar que no todos los municipios generan basura de la misma manera, aunque la mayoría sigue el mismo patron de generación, compartiendo los picos de crecimiento y decrecimiento en la generación de basura, como podemos ver en las fechas del 16 al 30 de Mayo. De igual manera observamos que desde el 14 de Febrero hasta el 30 de Mayo (final de los datos obtenidos) todos los municipios exceptuando el B se comportan de una forma bastante parecida, con el municipio A siendo el mayor generador.

En el caso del municipio B, vemos que su produccion de residuos es mucho menor que los demas municipios. El siguiente grafico muestra dicho municipio tiene poca cantidad de contenedores en comparacion al resto. Esto se debe a que en gran parte del municipio B (barrios Cordón, Centro y Ciudad VIeja), la recolección de residuos no es realizada por la IM sino por la empresa CAP. Por tanto los datos de recolección de dichos barrios no están incluidos en nuestras fuentes.

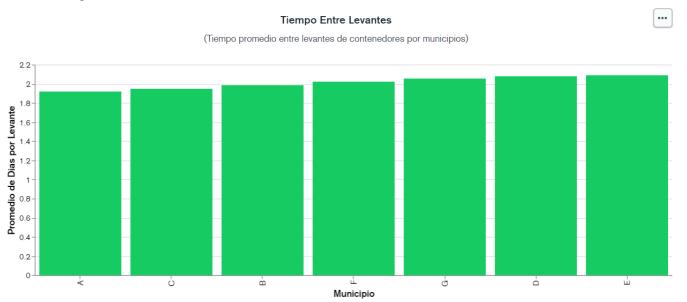


En la siguiente imagen se muestran los contenedores existentes del municipio B y sus municipios adyacentes. Podemos ver que, en el municipio B, el sistema de contenedores de la IM solo atiende las zonas de barrios Aguada y La Comercial.



En el caso del municipio B, una gran parte de la zona que abarca no tiene contenedores asociados y esto puede verse en la imagen generada por los datos obtenidos, donde parte de la zona correspondiente al municipio B no tiene circulos asociados, por lo que concluimos que no existen contenedores provistos por la IM en esa zona. Es muy posible que existan contenedores no provistos por la IM en esa zona. Concluimos entonces que si bien no todos los municipios generan basura de igual modo la mayoría si lo hace, remarcando a los municipios que no cumplen con desde el comienzo de Diciembre hasta mediados de Febrero a los municipios A y G. Además podemos agregar que luego de estas fechas todos los municipios(exceptuando el B) generan una cantidad parecida(al rededor de 700,000 kilogramos) de basura semanalmente.

Para realizar un análisis sobre la cantidad de levantes de contenedores requerida por cada municipio y la frecuencia real de los levantes se produjo la siguiente gráfica que presenta la cantidad de días de atraso promedio en levantar los contenedores de cada municipio.



Observando la gráfica, es fácil notar que los tiempos de atraso promedio por levantes tienen una variación entre municipios que es casi nula (varían entre 1.9 y 2.1 días de atraso), por lo cual se considera que la asignación de recursos(envíos de recorridos de levantes) por municipio es justa y equitativa. La misma fue generada por el siguiente pipeline sobre la colección de atrasosFull:

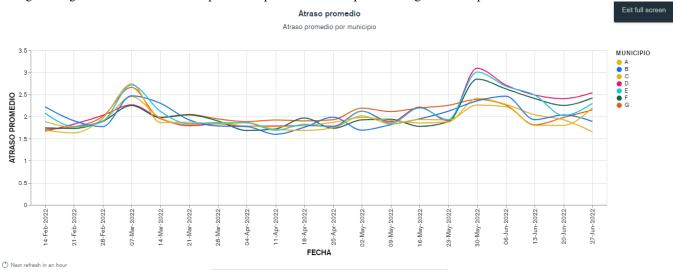
Listado 2 Pipeline del promedio de dias de atraso

```
"$group":{
    "_id":"$MUNICIPIO", "CANTIDAD_SUM":{"$sum": 1}, "DIAS_ATRASO_TOTAL":{"$sum":'$DIAS_ATRASO'}
},

{
    "$project":{
        "_id":1, "PROMEDIO":{"$divide":["$DIAS_ATRASO_TOTAL","$CANTIDAD_SUM"]}
},

{
    "$project":{
        "Municipio":"$_id", "Promedio_Dias_Levantes":"$PROMEDIO", "_id":0
}
},
```

El siguiente grafico muestra el atraso promedio para cada municpio a lo largo del tiempo.



Observamos que separando los atrasos promedio de los municipios y graficandolos a través del tiempo, todos los crecimientos y decrecimientos en atrasos se dan en las mismas fechas para cada municipio. De esto se entiende que cuando hay un aumento en los atrasos promedio de un municipio estos repercuten en los otros, por lo que los recursos de la IM estan siendo correctamente distribuidos para evitar que un municipio sufra de atrasos mayores.

VI. CONCLUSIONES Y TRABAJO FUTURO

Luego de realizada la tarea, creemos que el pasaje de los datos a la nueva base de datos, enriquecidos y tomando en cuenta estos a través del tiempo generando un histórico, que permiten la utilización de la herramienta, que como mostramos anteriormente es muy útil, es una mejora substancial sobre los datos que nos fueron disponibilizados por la Intendencia de Montevideo. Algunos de los desafíos que nos encontramos durante el desarrollo de la tarea a modo de ejemplo fueron problemas con el tipado de los datos y el uso de datos con tipos distintos, el modelado de fechas y la inseción de fechas que no cumplen con los formatos de fechas de MongoDB, la busqueda del por qué los datos de las pesadas del municipio B eran tan bajos, la modificación de los datos disponibilizados por la IM sobre los atrasos que resultaron en la colección atrasosFull, etc. Creemos que se cumplió con los objetivos planteados, lograndose una base de datos documental que permite la utilización de la herramienta MongoCharts que permite el análisis de los datos con gran facilidad, siendo el análisis presentado solo una muestra de el potencial de análisis que permite esta nueva base de datos.

No se logró realizar la visualización en los datos geográficos, con los contenedores con la esquina más cercana, pero si se logro visualizar los contenedores por municipios en un mapa, y por otro lado se logró generar una colección llamada capaContenedores, donde cada contenedor tiene una descripción con su ubicación la esquina más cercana. Es importante mencionar que no se logró presentar la bibliografía utilizada de la manera pedida, ya que al realizarlo de esa manera las citas se presentaban como un signo de pregunta, cosa que podría haberse solucionado de haber reconocido el problema con más tiempo. A futuro se podría expandir la base de datos para obtener datos de fechas anteriores y enriquecer el análisis con datos de diversos años permitiendo un análisis de la fluctuación de la producción de basura en los años. Se podría también realizar un análisis más profundo sobre la poca generación de basura en los municipios A y G en Diciembre y Enero, permitiendo emplear menos recursos en estas zonas durante esos períodos. En adición sería un gran punto de avance para la base de datos obtener los datos del proveedor de contenedores para la zona del municipio B.

VII. FORMATOS Y BIBLIOGRAFÍA REFERENCIAS

- Letra proyecto final.
 Prueba.
 Centros comunales por municipio.
 Datos de levantes.
 Datos demograficos.
 Mongodb Atlas.
 Entrega Final.