Almac. y Proc. Masivo de Datos: Entrega #2

Miércoles, 25 de Octubre, 2017

Juan Pablo Arévalo, Andrés Abbott

1 Introducción

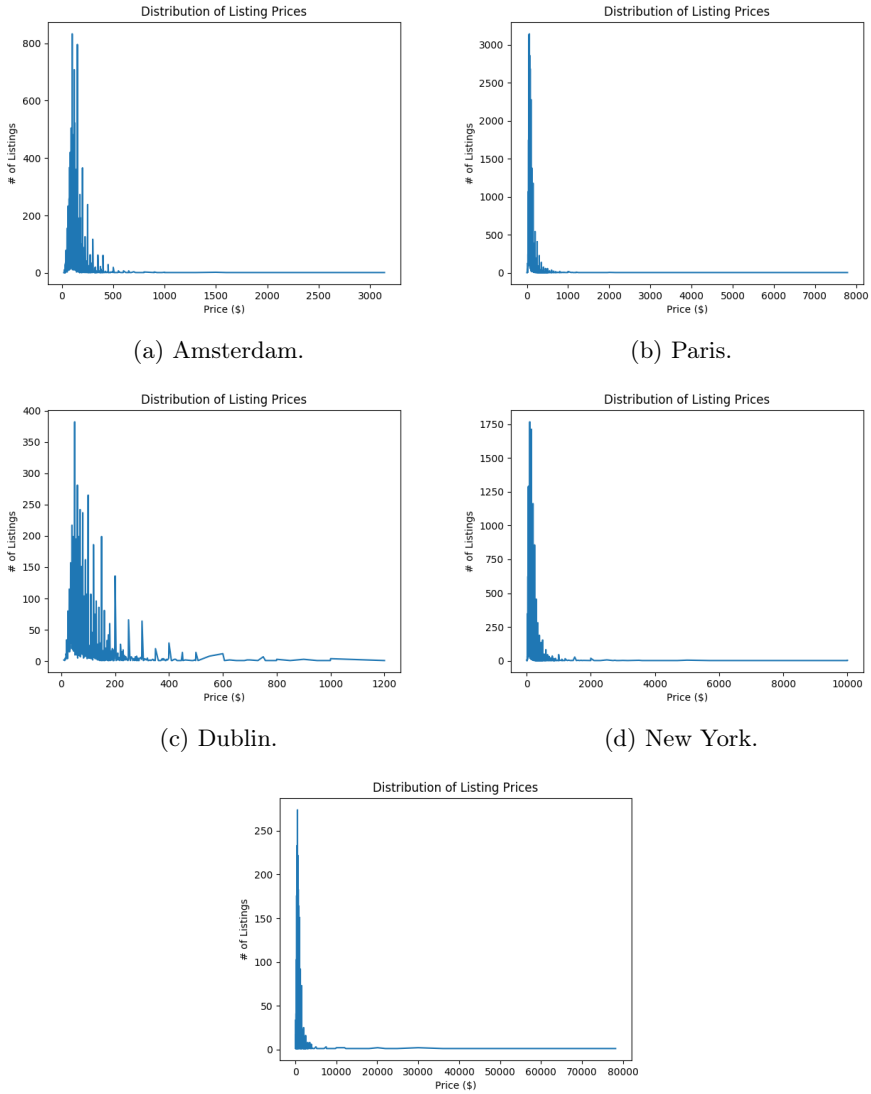
En este informe se presentaran los avances actuales del análisis de los datos de la empresa Airbnb para el ramo de Procesamiento y Almacenamiento Masivo de Datos. Se procederá a explicar las dificultades, como la limpieza de los datos, trabajar con nuevas librerías etc. También se mostrarán los avances actuales en que se encuentra el proyecto. Se describirán métricas calculadas y propuestas para desarrollar a futuro, acompañado también de un análisis preliminar de los datos, para ver las distribuciones de estos y como podrían afectar estas al desarrollo de alguna de las métricas futuras.

Para analizar los datos e implementar las métricas propuestas, se utilizó la librería MRJob[1] de Python, la cual se encarga de leer los archivos de datos linea por linea y por medio de procedimientos Map Reduce obtiene los datos de interés previamente programados por el usuario.

[1] https://pythonhosted.org/mrjob/

2 Datos

Para poder corroborar lo obtenido con las métricas propuestas, primero deberemos saber como distribuyen los datos que vamos a utilizar, para así poder usar algoritmos adecuados a cada distribución. En un comienzo, asumiremos que los datos tienen una distribución normal, luego haremos algunos análisis de los datos e intentaremos corroborar esta hipótesis. Para comprobar si la distribución de los datos es normal, se analizó los datos de los archivos listing.csv de las distintas ciudades. Por una parte, se calculó la distribución de precios para los listings de cada ciudad. Los resultados se pueden ver expresados de manera gráfica en la Figura 1.



(a) Amsterdam.

(b) Paris.

(c) Dublin.

(d) New York.

(e) Hong Kong.

Figure 1: Distribuciones de las ciudades estudiadas.

Como se puede ver, todas las ciudades presentan una distribución similar para los precios de los listings. Pareciera existir un "right skew[2]" (sesgo estadístico por la derecha), pero esto luego de analizar los datos un poco más profundamente se ve que ocurre debido a unos pocos listings con precios excesivamente altos en comparación a la media, lo cual altera el gráfico. Dejando de lado este detalle, la distribución de los datos para todas las ciudades parece ser una distribución normal o cercana a lo que sería una distribución normal. Por último, se calculó el precio promedio de listing por ciudad, cuya representación gráfica se puede ver en la Figura 2.

[2]https://en.wikipedia.org/wiki/Skewness

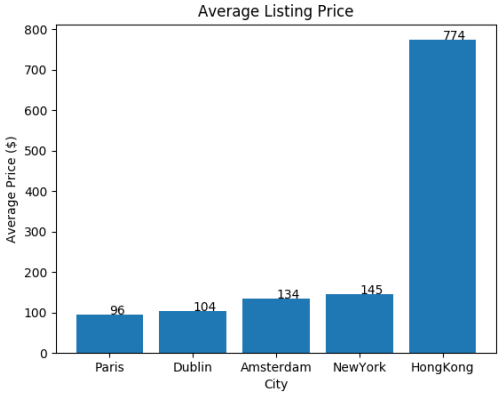


Figure 2: Promedio precios.

Podemos notar que el precio promedio de los listings para cada ciudad es bastante similar, entre los 95 y los 145, a excepción de los listings ubicados en la ciudad de Hong Kong, que alcanzan un promedio de 774 por listing.

Al notar esto, se nos ocurrió que podía existir la posibilidad de que el campo "precio" de los listings no fuera por defecto un valor en US$ (Dólares americanos), revisando nuevamente la descripción de los datos descargados, notamos que cada ciudad podía presentar un tipo de moneda distinto para el campo precio. Es por esto que investigamos en la página de AirBnb sobre el tipo de moneda correspondiente al campo precio para cada ciudad, obteniendo los siguientes resultados:

- Paris: Euro
- Dublin: Euro
- Amsterdam: Euro
- New York: USD - Dólar Americano (US$)
- Hong Kong: HKD - Dólar de Hong Kong (HK$)

Como podemos ver, estábamos bastante equivocados sobre las monedas presentes en los datos, es por esto que decidimos pasar todos esos valores a dólares americanos (US$) para representar de manera más exacta el gráfico de la Figura 2. El gráfico del precio promedio (en US$) de los listings para cada país se puede ver en la Figura 3.

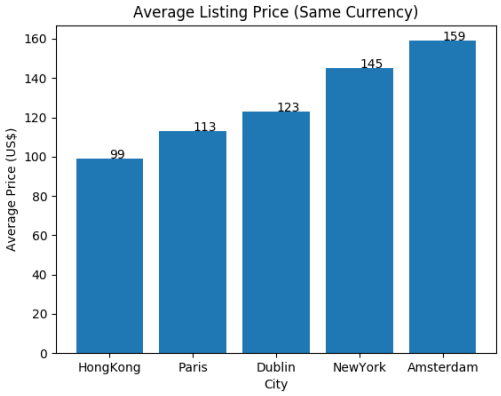


Figure 3: Promedio precios en USD.

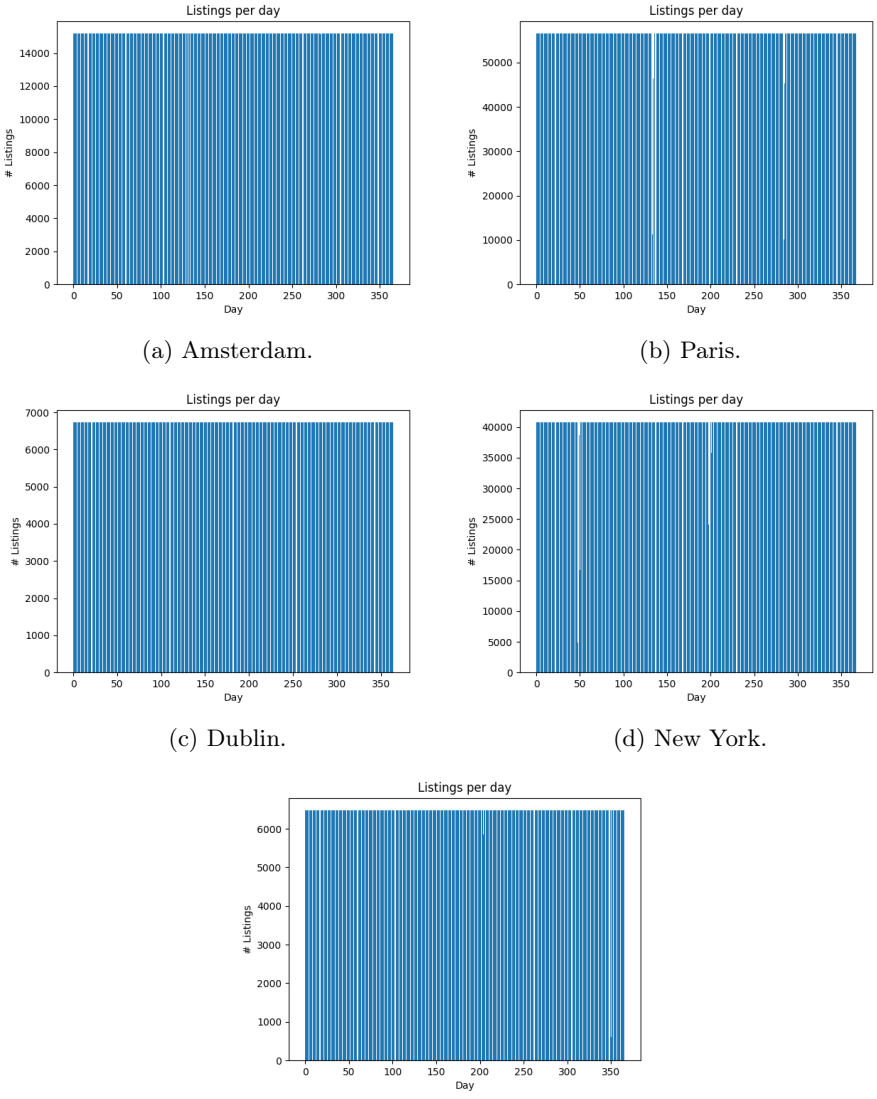
Como se puede ver en la Figura 3, ahora los promedios de precio para cada listing son bastante más similares, entre los 99 US$ y los 159 US$, lo cual deja a Amsterdam como el país con precio promedio de listings mayor y a Hong Kong con el precio promedio de listings menor.

2.1 Estacionareidad

Para comenzar a entender la distribución y características de los datos que estamos analizando, se optó por estudiar la estacionareidad de los distintos set de datos, basándonos en el modelo de Box-Jenkins. Para comenzar, nos concentramos en estudiar este factor en el set de "reviews" y el set "calendar", el primero, por su parte, contiene todos los reviews realizados a distintos listings, con su respectiva fecha de publicación, mientras el segundo contiene información de los listings en distintas fechas del año, indicando su disponibilidad y precio para dicha fecha.

2.1.1 Estacionareidad para Calendar

Comenzamos analizando la estacionareidad para el set de Calendar, para esto, graficamos el número de listings relacionados a cada día (fecha) en el set de datos, los resultados se pueden ver en la Figura 4. Como se puede ver en la Figura 4, los tiempos en el set de Calendar presentan estacionareidad



(a) Amsterdam.

(b) Paris.

(c) Dublin.

(d) New York.

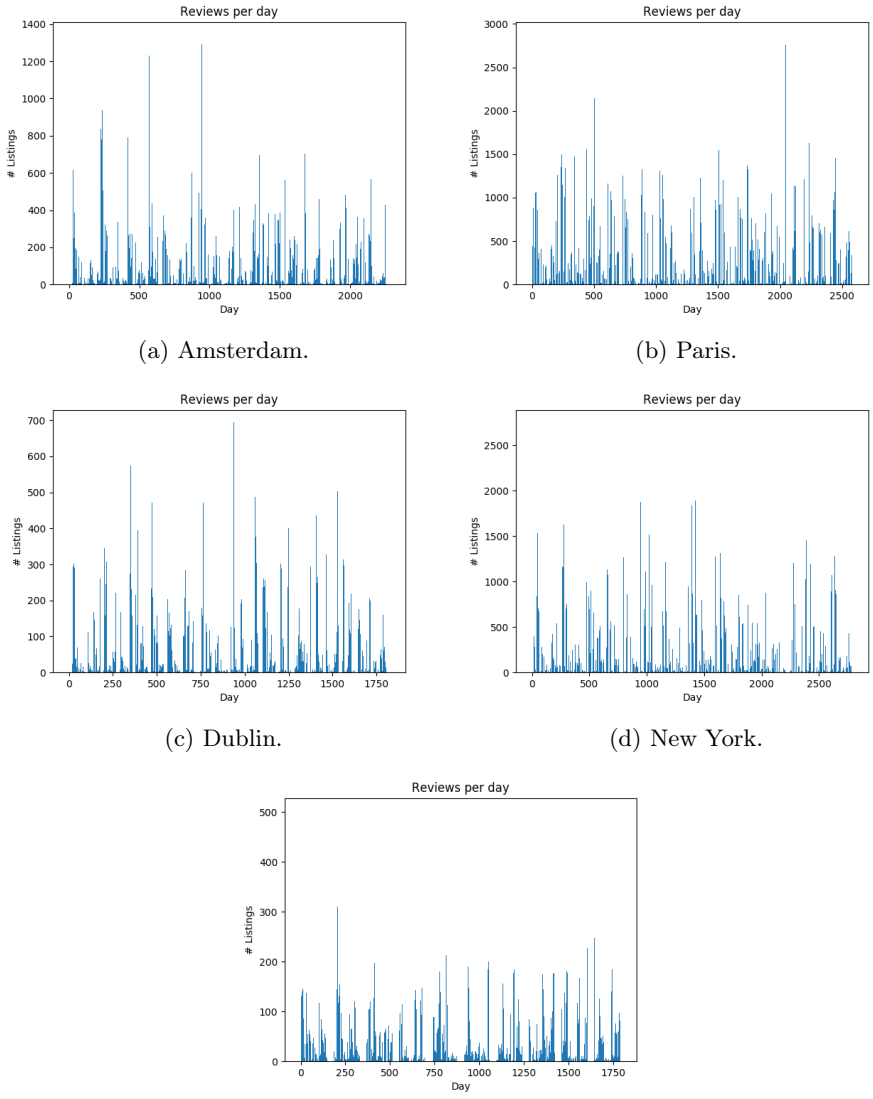
(e) Hong Kong.

Figure 4: Distribuciones de las ciudades estudiadas.

para las cinco ciudades, presentando muy pequeñas variaciones en algunos días, pero generalmente manteniendo un número similar o igual de listings en cada día. Es decir, presenta una escala constante.

2.1.2 Estacionareidad para Reviews

Luego, realizamos el mísmo cálculo para el set de Reviews, cuyos resultados se presentan en la Figura 5. Como se puede ver en la Figura 5, los tiempos en el set de Reviews no son estacionarios



(a) Amsterdam.

(b) Paris.

(c) Dublin.

(d) New York.

(e) Hong Kong.

Figure 5: Distribuciones de las ciudades estudiadas.

para ninguna de las ciudades estudiadas, presentando una gran variación en el número de reviews por cada día, es decir, la escala no es constante. Con estos resultados podemos comprender un

poco mejor las características de los datos que utilizaremos, ya que, al comprender las series de tiempo que componen nuestros sets de datos, podremos interpretar de mejor manera los resultados obtenidos y comprender el porqué algunos resultados pueden representar variaciones significativas en comparación a los otros. Para la siguiente iteración se seguirá completamentando el estudio con el modelo de Box-Jenkins, para auxiliar en la definición de los modelos de estimación de parámetros a utilizar y aplicar las técnicas de promedio autoregresivo y promedio móvil, en sus respectivos modelos.

3 Métricas Implementadas

En esta sección se explicaran las métricas realizadas hasta el momento. Por alcances de esta entrega, los resultados no se explicaran en su totalidad.

3.1 Listing disponible con mayor y menor precio promedio

Estas dos métricas muestran, por una parte el arriendo con mayor precio promediando el precio total de todas las fechas en las que se encuentra disponible dicho arriendo, y por otra parte el arriendo con menor precio bajo estas mismas condiciones.

```
      (Caso Ejemplo: Amsterdam)
      "MAX"    [6335.8, "16787521"]

      (Caso Ejemplo: Amsterdam)
      "MIN"    [25.0, "10373216"]
```

3.2 Listing con mayor número de reviews

Con esta métrica, podremos saber el listing que recibió mayor numero de reviews. Esto nos podría servir a la hora de relacionar las reacciones de los usuarios a ciertos listings. Por ejemplo, si el listing que obtuvo mas reviews se encuentra en cierta ubicación, y los reviews son en general buenos, podría señalar que las personas prefieren esa ubicación a la hora de arrendar.

```
      (Caso Ejemplo: New York)
      "MAX"    [432, "903972"]
```

3.3 Usuario con mayor número de reviews

Con esta métrica, podremos saber el usuario que ha escrito el mayor numero de reviews. Esto nos podría servir a la hora catalogar usuarios en base a su actividad en la página, para poder asignar características como "Usuario Influente", "Experenciado", etc.

```
      (Caso Ejemplo: Amsterdam)
      "MAX"    [16, "10500507"]
```

3.4 Listing con mayor reservas

Esta métrica señala el listing que tuvo mayor numero de reservas en el set de datos. Esto nos ayudará a saber los listings mas populares que han existido.

```
      (Caso Ejemplo: Dublin)
      "MAX"    [365, "9995720"]
```

3.5 Listing disponibles entre un rango de fechas

Esta métrica nos indicara los listings que están disponibles entre el rango de fechas señaladas. Esto nos ayudara a saber los listings que se podrán arrendar en ciertas fechas. Por ejemplo, si me quiero ir una semana, esta métrica mostrara todos los listing que están disponibles para el arriendo en esa semana completa.

```
(Caso Ejemplo: Amsterdam, Inicio: 2017-11-07, Fin: 2017-11-14 )
"AVAILABLE"   "9104050"
"AVAILABLE"   "911536"
"AVAILABLE"   "9115414"
"AVAILABLE"   "9120754"
"AVAILABLE"   "912209"
 ...
```

3.6 Listing mas barato disponible entre un rango de fechas

Análogamente a la métrica anterior, esto se puede utilizar para realizar un filtrado por un rango de fechas, pero esta vez, nos entregara el listing mas barato disponible en ese rango. Con esto podremos saber cual es el lisitng mas barato en fechas de alta demanda, como por ejemplo navidad o año nuevo.

```
(Caso Ejemplo: Amsterdam, Inicio: 2017-11-07, Fin: 2017-11-14 )
"CHEAPEST AVAILABLE LISTING" [25.0, "10373216"]
```

3.7 Listing que generó más ingresos para su host en un año

Esta es una métrica que puede tener gran valor para alguna persona que esta pensando en entrar al negocio de los arriendos por Airbnb. La métrica nos indicara el listing que genera mas ingresos en el set de datos. Esto sera útil, ya que se podría analizar el modelo de negocios que esta llevando este listing, así como sus características y mayores atractivos para intentar replicarlo en algún otro listing y poder generar mas ganancias. Para hacer esto, se calcula el número de días que fue efectivamente arrendado el listing dentro de un determinado año y se multiplica por el precio base del listing.

```
(Caso Ejemplo: Paris, Ano: 2017 )
"GREATEST INCOME"  [218125.0, "10273521"]
```

4 Métricas a Futuro

En esta sección se explicaran las métricas que no se pudieron realizar para esta entrega, pero que se planean desarrollar en entregas futuras, para mayor información sobre cada métrica, dirigirse al documento de la Entrega 1.

- Listing con mejor rating de los listings más baratos
- Listing más barato entre los listings con mejor rating
- Listing con reviews más negativos
- Usuario (host) con más listings
- Usuario más disconforme con sus reviews
- Usuarios que se arrienden listings mutuamente
- Par de usuarios con mayor similitud en listings visitados
- Porcentaje de arriendos en fechas festivas
- Promedio del valor de cada cama
- Promedio del numero de baños en relación a los huespedes
- Número de listings por barrio
- Usuario con listings en más países
- Usuario con reviews en más países
- Comparativa del precio promedio entre países

Además de estos, se incorporaron nuevas métricas al set presentado en la Entrega 1:

- Arriendos promedio por año/mes/semana
- Reviews promedio por año/mes/semana
- Reviews promedio por usuario
- Listings promedio por usuario

5 Problemas

A continuación se procederá a explicar los problemas y dificultades mas relevantes que se tuvieron que enfrentar a la hora de desarrollar el proyecto. Estos problemas estuvieron principalmente en el formato en que venían algunos campos dentro de los archivos csv.

5.1 Formato .csv

El primer gran problema se encontró a la hora de trabajar con los archivos csv, debido a que los datos en cada campo contenían caracteres que generan distorsión a la hora de leer los archivos. Uno de los primeros caracteres que dio problemas fue el `\n` que se encontraba en la gran mayoría de los archivos con texto, como reviews, descripciones, etc. La parecencia de este carácter provocaba que a la hora de leer los archivos con la librería MRJob se dividieran las lineas en dos cada vez que se encontraba un `\n`. Esto provocaba que la información de las filas quedara separada, siendo imposible para MRJob interpretar los campos de las filas. La manera que se arreglo esto fue creando un script de python que fuera recorriendo cada linea del archivo y eliminara todos los `\n` de los textos. Una vez que eliminaba todos los `\n`, vuelve a agregar un `\n` al final de la linea para que MRJob pueda leer correctamente el archivo. Cabe destacar que la librería para leer csv de python no se ve afectada por los `\n` del texto, por lo que se puedo recorrer el archivo con normalidad.

5.2 Números

Otro gran problema fue al momento de trabajar con números. Los datos proporcionados por Airbnb eran bastante completos, no obstante, no todos los campos poseían el mismo formato, en especial los números. Por ejemplo, en la columna precio del archivo listing.csv, algunos campos contenían comas en vez de puntos, lo que provocaba que a la hora de leer el archivo como un csv, quedaran los números separados en dos. Para solucionar esto, se tomo en cuenta la posición dentro del csv de los números a analizar y extraer o reemplazar los caracteres que generan conflicto. Todo esto se realiza en tiempo real, cuando MRJob esta analizando los datos.

5.3 Comentarios

El otro gran problema que se presento fue a la hora de analizar los archivos que contenían comentarios. La gran mayoría de los campos de comentarios llevan comas, lo que dificultaba la lectura del archivo que debía estar en formato csv. Por otro lado, no todos los comentarios partían con caracteres de definición de strings, como comillas, o comillas dobles. Para solucionar esto, se toma en cuenta la posición del texto dentro de la fila. Una vez teniendo la posición relativa del texto, se procede a analizar el texto hasta que se encuentran las comillas que caracterizan a los string en python.

6 Conclusiones

En conclusión, como se puede apreciar en el informe, la gran mayoría de los problemas se originaron por el formato y el orden que tenían los datos dentro de los archivos csv. Gran parte del tiempo empleado en el desarrollo del proyecto fue para corregir estos errores y limpiar los datos para las entregas futuras.

En cuanto a los datos, podemos apreciar que la gran mayoría tiene una distribución normal o se puede aproximar a una, ya que todos concentran los datos alrededor de ciertos valores, aunque haya ciertas excepciones que se alejen de este. Estas excepciones pueden ser provocadas por hitos ocurridos en ciertos días. Por ejemplo, si en una ciudad se realiza alguna conferencia de alto impacto, es de esperarse que la demanda de alojamientos en esa ciudad aumente, y con ello el precio de los arriendos. El análisis profundo de estos hitos no fue el objetivo de esta entrega, por lo que en la próxima iteración se realizara un análisis más detallado de estos acontecimientos.

Por último, en cuanto a las métricas, si bien no analizamos en concreto los resultados de estas, serán utilizadas para la siguiente iteración y serán de gran ayuda para lograr afirmar la hipótesis planteada.