Proyecto Final Analisis de Datos

Integrantes: Nicolas Luna, Elkin Diaz,
Joseph Caza

I. Definicion del caso de estudio

El objetivo de este proyecto es aplicar los conocimientos adquiridos durante todo el semestre sobre Análisis de datos como la extracción, limpieza, transformación, y visualización de datos mediante la utilización de la herramienta Power BI.

Por medio de este proyecto, se pretende diseñar una arquitectura de datos que añadan 10 fuentes de datos diversas con los cuales nos permitirán generar dashboards explicativos.

El proyecto se apoya en la integración y análisis de datos originarios de las siguientes fuentes:

1. **Deportes y medallas olímpicas, 1896-2014. (Datos sobre los Juegos Olímpicos)**

url:https://www.kaggle.com/datasets/the-guardian/olympic-games

2. **Restaurantes en Zomato Bangalore. (Datos sobre restaurantes en Bangalore)**

url: https://www.kaggle.com/datasets/himanshupoddar/zomato-bangalore-restaurants

3. **Giras de conciertos de Taylor Swift. (Datos sobre las giras de conciertos de Taylor Swift)**

url: https://www.kaggle.com/datasets/gayu14/taylor-concert-tours-impact-on-attendance-and

4. **Precios de la Vivienda en EE. UU (Datos sobre los precios de vivienda en Estados Unidos)**

url: https://www.kaggle.com/datasets/fratzcan/usa-house-prices

5. **Costos Médicos. (Datos sobre costos médicos en distintas áreas)**

url: https://www.kaggle.com/datasets/waqi786/medical-costs

6. **Datos de temperatura diaria de 2015 a 2016. (Datos meteorológicos sobre la temperatura diaria)**

url: https://www.kaggle.com/datasets/farhakouser/temperature

II. **7. Global Electric Vehicle Sales Data (2010-2024).**

url: Global Electric Vehicle Sales Data (2010-2024) (kaggle.com)

8. **Conjunto de datos de restaurantes/trago. (Datos sobre restaurantes y bebidas)**

url: https://www.kaggle.com/datasets/ashishjangra27/swiggy-restaurants-dataset

III. **9. Paris Olympics 2024 Games Dataset (updated daily).**

url: Paris Olympics 2024 Games Dataset (updated daily) (kaggle.com)

10. **Nirvana Live Performances**

url: Nirvana Live Performances - dataset by ben-pfeifer | data.world

IV. Objetivos

A. *General*

Aplicar los conocimientos adquiridos sobre análisis de datos mediante la herramienta de Power BI, para diseñar e implementar una arquitectura de datos los cuales incorporen diversas fuentes de datos como SLQ y NoSQL

B. *Especificos*

- Consolidación de diversas fuentes de datos
- Diseño y Gestión de Base de Datos
- Análisis de Datos y Creación de Dashboards
- Generar archivos multimedia explicativas que permitan identificar patrones, tenencias en los datos.
- Documentar todo el proceso de integración, limpieza, transformación y análisis de datos.

V. Descripcion del equipo de trabajo

Nicolas Luna (Coordinador del Proyecto y Analista de Datos)

- Responsabilidades:

Coordinación del Proyecto: Supervisar y coordinar las actividades del equipo para asegurar que el proyecto avance según el cronograma establecido.

Documentación: Redactar la documentación del proceso de integración, limpieza, transformación y análisis de datos

Joseph Caza (Especialista en Bases de Datos)

- Responsabilidades:

Integración de Datos: Establecer conexiones entre bases de datos relacionales y NoSQL, y gestionar la transferencia de información entre ellos.

Transformación de Datos: Convertir los datos entre formatos CSV y JSON, y asegurar la consistencia y calidad de los datos importados.

Elkin (Desarrollador de Dashboards en Power BI)

- Responsabilidades:

Diseño de Dashboards: Crear dashboards interactivos en Power BI que permitan el análisis detallado de los datos, incluyendo gráficos, tablas y otros elementos visuales.

Análisis de Casos de Estudio: Desarrollar y analizar casos de estudio basados en los datos, y presentar hallazgos y patrones significativos.

Visualización de Datos: Generar archivos multimedia explicativos que faciliten la identificación de patrones y tendencias en los datos

Presentación de Resultados: Documentar los resultados del análisis de datos y la visualización, y preparar la presentación final para el informe.

Actividades Realizadas

Nicolás Luna (Coordinador del Proyecto y Analista de Datos)

- Coordinación del Proyecto:

o Planificación y supervisión de las tareas del equipo.

o Organización de reuniones de seguimiento para revisar el progreso y resolver problemas.

o Actualización del cronograma del proyecto para reflejar avances y cambios.

- Documentación:

o Elaboración de la documentación detallada del proceso de integración de datos.

o Redacción de informes sobre las etapas de limpieza y transformación de datos.

o Creación de la sección de análisis de datos del informe final, detallando los métodos y resultados obtenidos.

Joseph Caza (Especialista en Bases de Datos)

- Integración de Datos:

o Configuración de las bases de datos relacionales (SQL Server y MySQL) y NoSQL.

o Establecimiento de conexiones entre las bases de datos relacionales y NoSQL.

o Gestión de la transferencia de datos entre diferentes sistemas y bases de datos.

- Transformación de Datos:

o Conversión de datos entre formatos CSV y JSON según las necesidades del proyecto.

o Validación de la consistencia y calidad de los datos importados para garantizar su integridad.

o Realización de pruebas para asegurar la correcta integración y transformación de los datos.

Elkin (Desarrollador de Dashboards en Power BI)

- Diseño de Dashboards:

o Creación de dashboards interactivos en Power BI para visualizar los datos de manera efectiva.

o Diseño de gráficos, tablas y otros elementos visuales para representar los datos de manera clara.

o Configuración de filtros y opciones interactivas para facilitar el análisis de datos.

- Análisis de Casos de Estudio:

o Desarrollo de casos de estudio basados en los datos, identificando patrones y tendencias relevantes.

o Presentación de hallazgos significativos y conclusiones a partir de los datos analizados.

o Preparación de informes sobre los casos de estudio para su inclusión en el informe final.

- Visualización de Datos:

o Generación de archivos multimedia explicativos, como gráficos y videos, para ilustrar patrones y tendencias en los datos.

o Diseño de presentaciones visuales para apoyar la interpretación de los datos en el informe final.

- Presentación de Resultados:

o Documentación y resumen de los resultados del análisis de datos y visualización.

o Preparación y organización de la presentación final del informe, incluyendo gráficos y hallazgos clave.

o Colaboración en la revisión y ajuste del informe final para asegurar la calidad y precisión de la presentación.

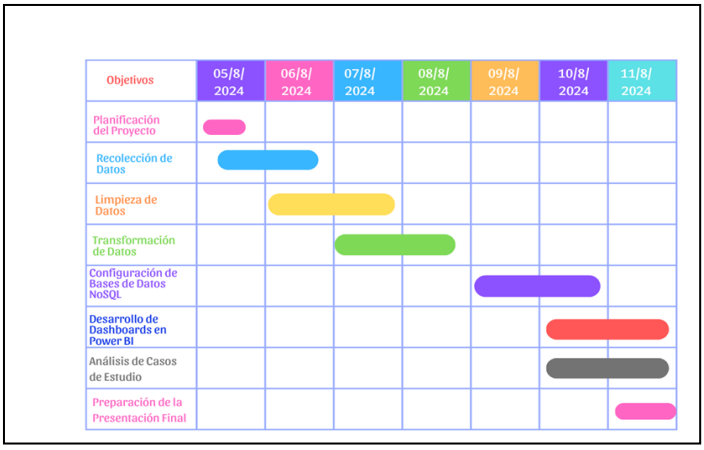
VI. CRONOGRAMA DE ACTIVIDADES

Planificación del Proyecto

- Inicio: 5 de agosto
- Fin: 5 de agosto
- Duración: 1 día
- Recolección de Datos
- Inicio: 5 de agosto
- Fin: 6 de agosto
- Duración: 2 días
- Limpieza de Datos
- Inicio: 6 de agosto
- Fin: 7 de agosto
- Duración: 2 días
- Transformación de Datos
- Inicio: 7 de agosto
- Fin: 8 de agosto
- Duración: 2 días
- Configuración de Bases de Datos NoSQL
- Inicio: 9 de agosto
- Fin: 10 de agosto
- Duración: 2 días
- Desarrollo de Dashboards en Power BI
- Inicio: 10 de agosto
- Fin: 11 de agosto
- Duración: 2 días
- Preparación de la Presentación Final

- Inicio: 11 de agosto
- Fin: 11 de agosto
- Duración: 1 días

TABLE I. DIAGRAMA DE GANTT



VII. RECURSOS Y HERRAMIENTAS UTILIZADAS

1. *Kaggle*

o Uso: Se utilizó para buscar y descargar los 10 conjuntos de datos que forman la base del análisis en este proyecto.

2. *MongoDB Compass y MongoDB Atlas*

o Uso: Se emplearon para gestionar bases de datos no relacionales, facilitando la carga y conversión de datos de CSV a JSON, y permitiendo el almacenamiento y gestión de estos datos en la nube.

3. *Canva*

o Uso: Se utilizó para diseñar el diagrama de Gantt que visualiza el cronograma de actividades del proyecto.

4. *Microsoft Word*

o Uso: Se empleó para redactar y estructurar el informe final del proyecto.

5. *Documentos de la Materia de Análisis de Datos*

o Uso: Se utilizaron para guiar la aplicación de conceptos y técnicas de análisis de datos a lo largo del proyecto.

6. *Power BI*

o Uso: Se utilizó para desarrollar dashboards interactivos que permiten un análisis detallado de los datos.

7. *MySQL*

o Uso: Se empleo para gestionar bases de datos relacionales

8. *Redis*

o Uso: Se empleo para gestionar bases de datos no relacionales

VIII. ARQUITECTURA DE LA SOLUCION

La solución se estructura en los siguientes componentes principales:

1. *Fuentes de Datos*

o Kaggle: Se han descargado 10 conjuntos de datos que abarcan áreas como deportes, restaurantes, y precios de vivienda.

o Proceso: Los datos son recolectados y transformados utilizando MongoDB Compass y Atlas, y almacenados en bases de datos relacionales y no relacionales.

2. *Bases de Datos*

o Relacionales: SQL Server y MySQL se utilizan para almacenar datos estructurados. Las tablas y relaciones están diseñadas para soportar el análisis de datos.

o NoSQL: MongoDB se emplea para datos no estructurados, permitiendo una mayor flexibilidad en el almacenamiento y manejo de datos en formatos JSON.

3. *Transformación de Datos*

o Conversión de Formatos: Se realiza la conversión de datos de CSV a JSON utilizando MongoDB Compass. Los datos se limpian y transforman para asegurar su calidad y consistencia.

4. *Herramientas de Análisis y Visualización*

o Power BI: Utilizado para crear dashboards interactivos que facilitan el análisis y visualización de los datos. Se conectan las bases de datos para generar informes y visualizaciones dinámicas.

5. *Flujos de Trabajo y Procesos*

o Flujos de Datos: Los datos se recolectan de Kaggle, se transforman y se almacenan en las bases de datos. Luego, se utilizan en Power BI para crear dashboards interactivos.

o Integración: La integración de datos entre bases de datos relacionales y NoSQL se gestiona mediante procesos de ETL (Extracción, Transformación y Carga).

IX. EXTRACCION DE DATOS

- **Descarga de Datos:**
  - Se accedió a las URLs proporcionadas en Kaggle para cada conjunto de datos.
  - Se descargaron los archivos en formatos compatibles como CSV, JSON, y otros según la disponibilidad de los datos.
- **Verificación de Datos:**
  - Se revisó la integridad y el formato de los datos descargados para asegurarse de que estuvieran completos y en el formato adecuado.
  - Se realizó una evaluación preliminar para identificar cualquier problema potencial con los datos, como datos faltantes o inconsistencias.

1. **Herramientas Utilizadas para la Extracción:**

- **Kaggle:**

Utilizado como la plataforma principal para la descarga de los conjuntos de datos.

Las fuentes específicas incluyen:

- Deportes y medallas olímpicas
- Reseñas de Restaurantes Europeos
- Giras de conciertos de Taylor Swift
- Precios de la vivienda en EE. UU.
- Costos médicos
- Datos de temperatura diaria
- Historia de la actuación de la Filarmónica de Nueva York
- Datos sobre restaurantes y bebidas
- Uso de bicicletas compartidas en Londres
- UEFA Euro 2024 Records

TABLE II. DATASET EN KAGGLE

Carga de Datos a las Bases de Datos:

El siguiente paso fue cargar los datos procesados en las bases de datos correspondientes para su almacenamiento y gestión. Este proceso se llevó a cabo de la siguiente manera:

Importar los datos usando los archivos csv o json, puede ser mediante Python o el gestor de base de datos.

TABLE III. IMPORTACION DE DATASETS CON PYTHON

```
# Nombre del archivo CSV
csv_file_path = 'ny philamornic/concerts.csv'

# Leer el archivo CSV y cargar los datos en MySQ
with open(csv_file_path, 'r') as csv_file:
    reader = csv.DictReader(csv_file)

    for row in reader:
        cursor.execute(
            """
            INSERT INTO events (Date, Location,
            VALUES (%s, %s, %s, %s, %s, %s, %s,
            """,
            (row['Date'], row['Location'], row['
        )
```

X. ANALISIS DE INFORMACION

Se crearon metricas para analizar la información, para esto se utilizo la herramienta de PowerBI, en el que se crean columnas calculadas, medidas y tablas calculadas.

TABLE IV. MEDIDA EN DAX

```
Eventos en la temporada del 86 al 87 =
CALCULATE(
    COUNT(conciertos_nyphilamornic[programID]),
    conciertos_nyphilamornic[orchestra] = "New York Philharmonic",
    conciertos_nyphilamornic[season] = "1986-87"
)
```

XI. VISUALIZACION DE LA INFORMACION

Se crearon Dashboards para visualizar estadísticas de los datasets obtenidos para mostrar resultados de los casos de estudios, a continuación, se mostraran los gráficos por cada caso de estudio:

TABLE V. CONCIERTOS DE LA FILARMONICA DE NYC

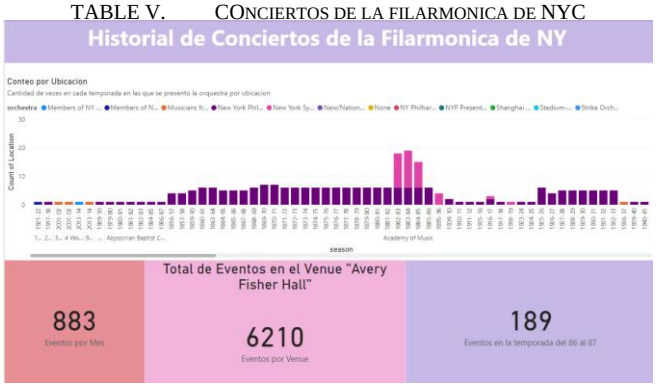


TABLE VI. OLIMPIADAS EN TEMPORADA DE VERANO DESDE 1896 A 2014

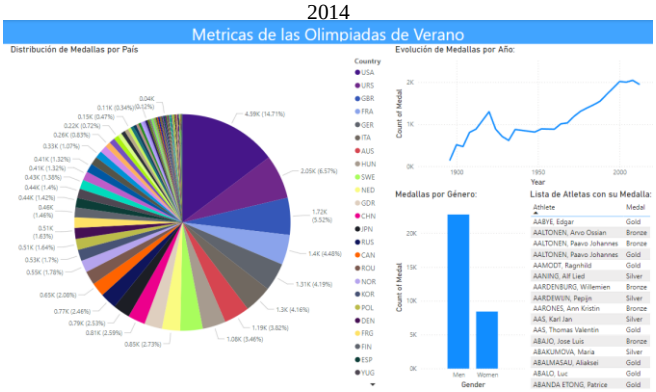


TABLE VII. TEMPERATURA DIARIA DE 2015 A 2016

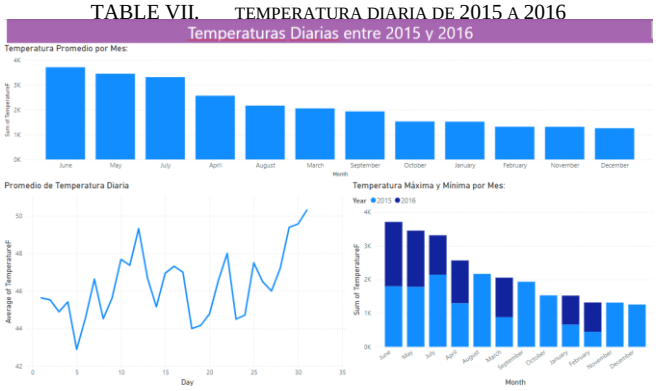


TABLE VIII. ASISTENCIAS PROMEDIO DE CONCIERTOS DE TAYLOR SIWFT

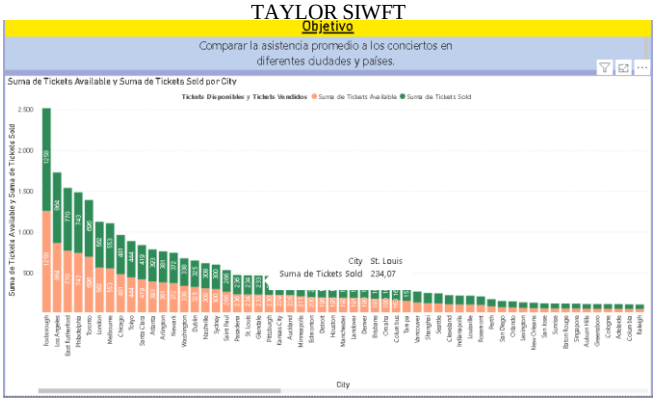


TABLE IX. ASISTENCIAS PROMEDIO CONCIERTOS DE

TABLE X.

TABLE XI. RESEÑAS NEGATIVAS Y POSITIVAS DE LOS RESTAURANTES EN EUROPA

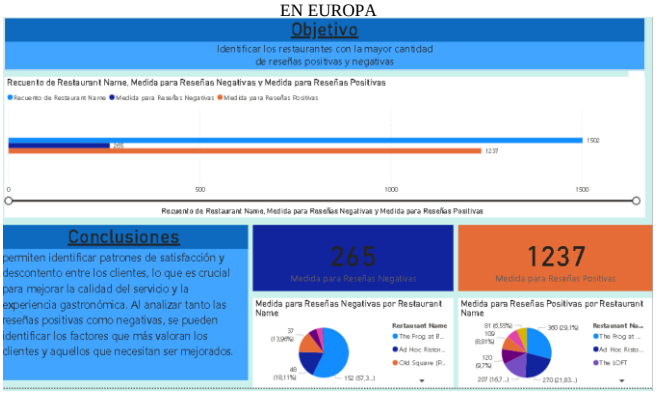


TABLE XII. ESTACIONES MAS UTILIZADAS DURANTE EL MES DE AGOSTO DE 2023

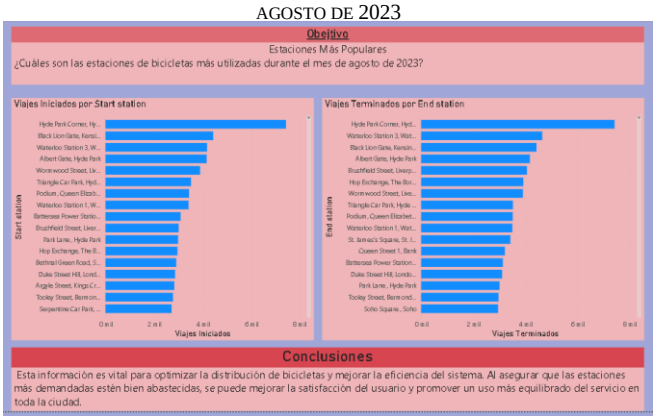
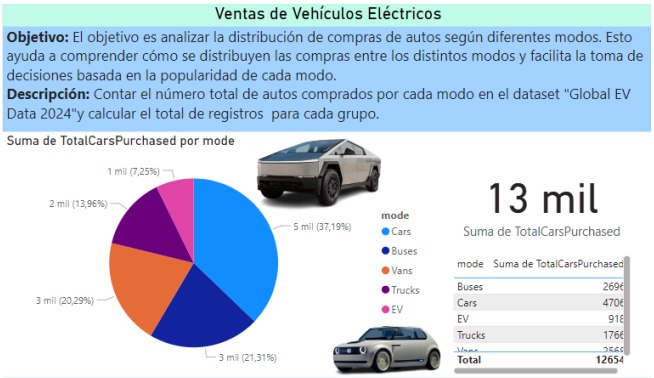


TABLE XIII. MEDALLA CON MAS ATRIBUCION EN LOS JUEGOS OLIMPICOS 2024

TABLE XIV. PARTCIPACION DE ATLETAS DEPENDIENDO CONFORME A SU PAIS.

TABLE XV. PRESENTACIONES DE NIRVANA POR EL MUNDO

TABLE XVI. VEHICULO ELECTRICO MAS VENDIDO SEGÚN EL MODO.



XII. RESULTADOS OBTENIDOS

El resultado esperado de este proyecto es un conjunto de dashboards interactivos que permiten a los usuarios explorar los datos de manera intuitiva y obtener insights valiosos. Los dashboards deben estar diseñados de manera que faciliten la toma de decisiones y proporcionen una visión clara y concisa de los datos analizados.

Estos resultados reflejarán la capacidad de integrar, transformar y visualizar datos complejos de múltiples fuentes, demostrando habilidades avanzadas en análisis de datos y uso de herramientas como Power BI.

A. *Link de Github del proyecto:*
<link>https://github.com/EDiaz210/Proyecto-Analisis.git

B. *Link del OneDrive del proyecto:*
<link> ProyectoAnalisis-JCaza-EDiaz-NLuna

XIII. CONCLUSIONES Y RECOMENDACIONES

Diversidad de Fuentes de Datos: La integración de 10 fuentes de datos diversas permite obtener una visión más completa y holística de los datos analizados. Esto enriquece los análisis al permitir una triangulación de la información y ayuda a reducir sesgos que podrían surgir si se utilizara una única fuente.

Calidad y Limpieza de Datos: La calidad de los datos es fundamental para el éxito de cualquier proyecto de análisis de datos. Durante este proyecto, fue necesario dedicar un tiempo significativo a la limpieza y transformación de los datos, lo que demuestra que estos procesos son críticos para garantizar la fiabilidad de los resultados obtenidos en los dashboards.

Transformación de Datos: La transformación de datos fue esencial para estandarizar la información proveniente de distintas fuentes. Esto incluyó la normalización de formatos, la unificación de criterios de clasificación y la creación de métricas comunes, lo que facilitó la comparación y el análisis conjunto de los datos.

Visualización y Comunicación de Resultados: Los dashboards generados en Power BI permiten visualizar de manera clara y concisa las tendencias, patrones y anomalías en los datos. Esta visualización es clave para la toma de decisiones basada en datos, ya que facilita la comprensión de información compleja por parte de los usuarios finales.

Arquitectura de Datos: La arquitectura de datos diseñada es robusta y flexible, permitiendo la integración de múltiples fuentes de datos y la actualización automática de la información en los dashboards. Esta flexibilidad es crucial para adaptarse a cambios en las fuentes de datos o en los requerimientos de análisis.

XIV. DESAFIOS Y PROBLEMAS ENCONTRADOS

En algunos casos de estudio, los datasets que estaban en archivo JSON tenían que ser cambiados de formato a CSV a fin de poder importarlo a un gestor de base de datos como es MYSQL, ya que no suelen aceptar formato JSON.

Otro problema encontrado fue el de que algunos datasets tenian formatos que no eran correctos y codificaciones que no permitían importar los datos a PowerBI por lo que para solucionar el problema, se tuvo que limpiar el dataset, reemplazando los valores erróneos con valores que pueda interpretar el motor de PowerBI

Adicionalmente, se presentó un inconveniente al intentar cargar un dataset en MongoDB, ya que este no estaba en codificación UTF-8. Para solucionar este problema, se abrió el dataset en un editor de texto como el Bloc de Notas y luego se guardó nuevamente con la codificación UTF-8, conservando el mismo formato .CSV. Esto permitió que el dataset fuera correctamente importado en MongoDB sin problemas de codificación.

BIBLIOGRAFIA

[1] D. V. Lindberg and H. K. H. Lee, "Optimization under constraints by applying an asymmetric entropy measure," *J. Comput. Graph. Statist.*, vol. 24, no. 2, pp. 379–393, Jun. 2015, doi: 10.1080/10618600.2014.901225.

[2] B. Rieder, *Engines of Order: A Mechanology of Algorithmic Techniques*. Amsterdam, Netherlands: Amsterdam Univ. Press, 2020.

[3] I. Boglaev, "A numerical method for solving nonlinear integro-differential equations of Fredholm type," *J. Comput. Math.*, vol. 34, no. 3, pp. 262–284, May 2016, doi: 10.4208/jcm.1512-m2015-0241.