Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Laboratorio de Seminario de Sistemas II
Auxiliar de cátedra: Glen Calel Robledo

DOCUMENTACIÓN

Práctica 1

Danilo Urías Coc
201314565

Contenido

Requerimientos técnicos

Requisitos del sistema:

Hardware utilizado:

Memoria RAM 8GB
Procesador Intel core i5 2.20 GHz

Software adicional utilizado:

Sistema operativo Windows 10 professional
Microsoft Office Excel
Notepad ++
Navegador web Google Chrome

Herramientas y versiones utilizadas para la solución

- **Java development kit (JDK) versión 8**

El Kit de desarrollo de Java (JDK) es un entorno de desarrollo de software utilizado para desarrollar aplicaciones y applets de Java. Incluye el Java Runtime Environment (JRE), un intérprete / cargador (java), un compilador (javac), un archivador (jar), un generador de documentación (javadoc) y otras herramientas necesarias para el desarrollo de Java.

Enlace de herramienta:
https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html

- **Apache Spark 2.4.6**

Apache Spark es una tecnología de computación en clúster ultrarrápida, diseñada para una computación rápida. Se basa en Hadoop MapReduce y amplía el modelo de MapReduce para usarlo de manera eficiente para más tipos de cálculos, que incluye consultas interactivas y procesamiento de flujo. La característica principal de Spark es su computación en clúster en memoria que aumenta la velocidad de procesamiento de una aplicación.

Spark está diseñado para cubrir una amplia gama de cargas de trabajo, como aplicaciones por lotes, algoritmos iterativos, consultas interactivas y transmisión. Además de soportar todas estas cargas de trabajo en un sistema respectivo, reduce la carga administrativa de mantener herramientas separadas.

Enlace de herramienta: http://spark.apache.org/downloads.html

- **Apache Hadoop 2.7**

Apache Hadoop es una plataforma de software de código abierto basada en Java que gestiona el procesamiento y almacenamiento de datos para aplicaciones de big data. Hadoop funciona mediante la distribución de grandes conjuntos de datos y trabajos analíticos entre nodos en un clúster informático, desglosándolos en cargas de trabajo más pequeñas que se pueden ejecutar en paralelo. Hadoop puede procesar datos estructurados y no estructurados, y escalar de manera confiable desde un único servidor a miles de máquinas.

Enlace de herramienta: https://github.com/steveloughran/winutils

Relación entre Apache Spark y Apache Hadoop

Hadoop es un framework, son librerías que hace que pueda ejecutar aplicaciones distribuidas por decenas/cientos/miles de ordenadores. Por defecto, Hadoop tiene un modelo para ejecutar todas esas tareas llamado MapReduce. Una aplicación MapReduce ocupa unas cuantas decenas de líneas de código en Java pero dado que sigue un patrón muy rígido, toda la tarea de distribución de la aplicación, ejecución, coordinación, recuperación ante caídas de uno o varios equipos, la realiza el framework por su propia cuenta.

Por lo tanto Map reduce se hace lento. Ocupa pocos recursos, pero tiene muchos pasos intermedios que graban a disco la información para, en caso de que una tarea que se esté ejecutando caiga (por ejemplo, el ordenador donde estaba ejecutándose se apague), pueda crear otra, leer la información de disco que dejó la anterior y continuar con las instrucciones.

Dadas las problemáticas de hadoop nace Spark. Es otro modelo de ejecución de procesos que se ejecuta dentro de Hadoop y viene a ser una alternativa a MapReduce donde no hay nada que se escriba en disco, todo se ejecuta en memoria y, a cambio de necesitar más RAM en los equipos el cluster, lográndose una mejor velocidad.

Python Versión 3.7

Python es un lenguaje de programación interpretado, orientado a objetos y de alto nivel con semántica dinámica. Sus estructuras de datos integradas de alto nivel, combinadas con tipeo dinámico y enlace dinámico, lo hacen muy atractivo para el Desarrollo rápido de aplicaciones, así como para usarlo como un lenguaje de secuencias de comandos o enlace para conectar componentes existentes.

La sintaxis simple y fácil de aprender de Python enfatiza la legibilidad y, por lo tanto, reduce el costo del mantenimiento del programa. Python admite módulos y paquetes, lo que fomenta la modularidad del programa y la reutilización de código. El intérprete de Python y la extensa biblioteca estándar están disponibles en formato fuente o binario sin cargo para todas las plataformas principales, y se pueden distribuir libremente.

Enlace Herramienta: https://www.python.org/about/

PyCharm Versión 2020.1.2 Distribución community

PyCharm es un IDE de Python extremadamente popular. Un entorno de desarrollo integrado o IDE presenta un editor de código y un compilador para escribir y compilar programas en uno o varios lenguajes de programación.

PyCharm viene con un editor de código inteligente que facilita la escritura de código Python de alta calidad. Ofrece un nivel mejorado de comprensión y legibilidad del código mediante esquemas de color distintos para palabras clave, clases y funciones, es decir, resaltado de sintaxis y error.

Además de ofrecer la función de finalización de código inteligente, el editor de código genera instrucciones para completar el código actual.

Enlace herramienta: https://www.jetbrains.com/es-es/pycharm/

Descripción de las transformaciones y acciones utilizadas

Manipulación de datos

Spark Context: Es un punto de entrada a Spark y se usa para crear mediante programación elementos llamados Spark RDD, acumuladores y variables de difusión en el clúster. Su objeto sc está disponible por defecto en spark-shell y se puede crear mediante programación usando la SparkContextclass.

Spark RDD: Los conjuntos de datos distribuidos resilientes (RDD) son la estructura de datos fundamental de Spark. Los RDD son inmutables y tolerantes a fallas por naturaleza. RDD es solo la forma de representar un conjunto de datos distribuido en múltiples nodos en un clúster, que puede funcionar en paralelo. Los RDD se denominan resistentes porque tienen la capacidad de volver a calcular siempre un RDD cuando falla un nodo.

Funciones Lambda: En Python, una función anónima es una función que se define sin un nombre. Mientras que las funciones normales se definen usando la palabra clave ***def***, las funciones anónimas se definen usando la palabra clave ***lambda***.

Por lo tanto, las funciones anónimas también se denominan funciones lambda.Las funciones de Lambda pueden tener cualquier número de argumentos pero solo una expresión. La expresión se evalúa y se devuelve. Las funciones Lambda se pueden usar donde se requieran objetos de función.

Map: Map pasa cada elemento de la una fuente a través de una función y forma un nuevo conjunto de datos distribuido.

Filter: El filtro de spark es una operación de transformación de RDD que acepta un predicado como argumento. Dicho predicado es una función que acepta algunos parámetros y devuelve un valor booleano verdadero o falso. El método ***filter*** de spark recibe este predicado como argumento y opera en el RDD de origen. Como resultado filtra todos los elementos del RDD de origen, descartando los que no satisfacen el predicado y crea un nuevo RDD con los elementos que pasa la función del predicado.

Reduce by Key: Opera en pares clave, valor (k, v), contenidos en un RDD origen, su funcionalidad es combinar los valores para cada clave, generando un nuevo RDD reuniendo los valores de las claves comunes.

Sort by: Produce un nuevo RDD ordenado a partir de un RDD de valores que recibe como parámetro y un valor que especifica si el orden es descendente o ascendente.

Reportes

Librería Plotly: Es una biblioteca de trazado interactiva de código abierto que admite más de 40 tipos de gráficos únicos que cubren una amplia gama de casos de uso estadísticos, financieros, geográficos, científicos y tridimensionales.

Construido sobre la biblioteca Plotly JavaScript (plotly.js), plotly permite a los usuarios de Python crear visualizaciones interactivas basadas en la web que se pueden mostrar y guardar en archivos HTML independientes o servir como parte de aplicaciones web puras construidas por Python.

Plotly.graph_objs: Importa los elementos necesarios para creación de objetos gráficos.

Plotly.offline: Permite la generación de reportes interactivos web, para el caso de la solución de esta práctica se utilizó para generar reportes en html.

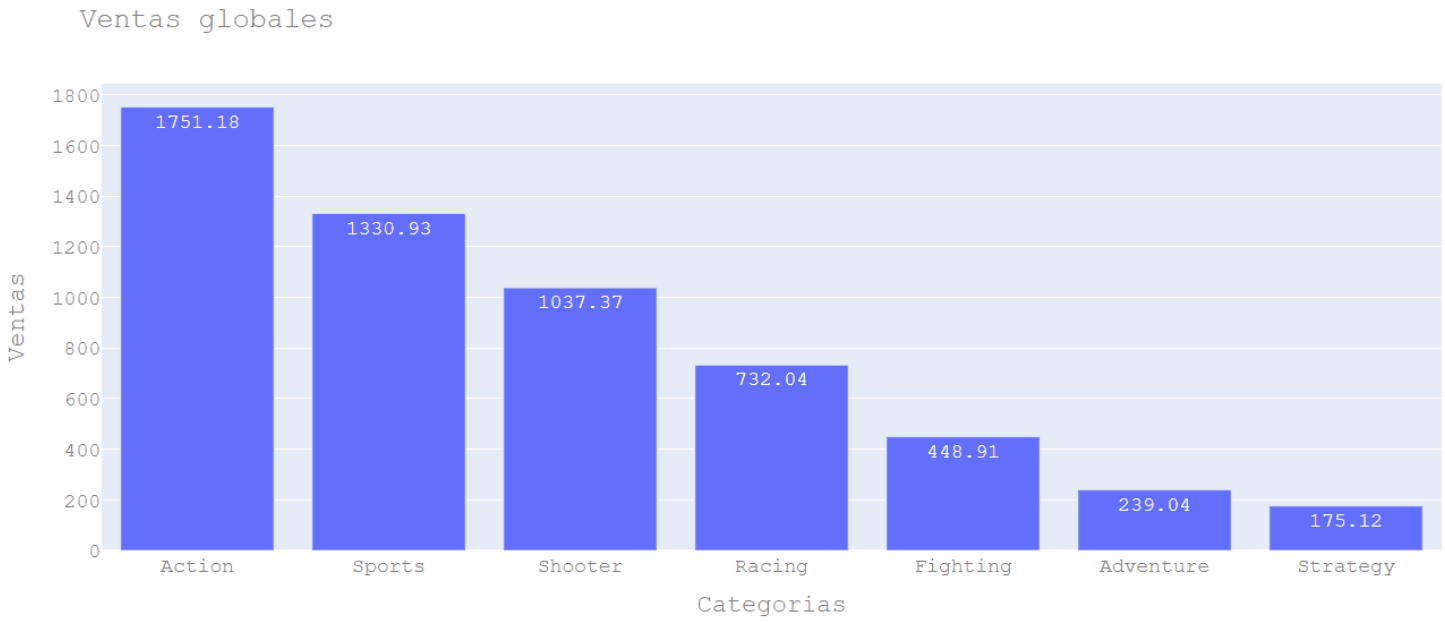
graph_objs.bar: Permite generar un gráfico de tipo barras, recibe como parámetros una estructura para los elementos del eje x y otra estructura para los elementos a reflejarse en el eje *y* del gráfico, además de un título *y*.

graph_objs.pie: Permite generar un gráfico de pie, recibe como parámetros una estructura para los elementos del eje x que sirven de etiqueta, también recibe una estructura de elementos numéricos que representarán las porciones de la totalidad del gráfico.

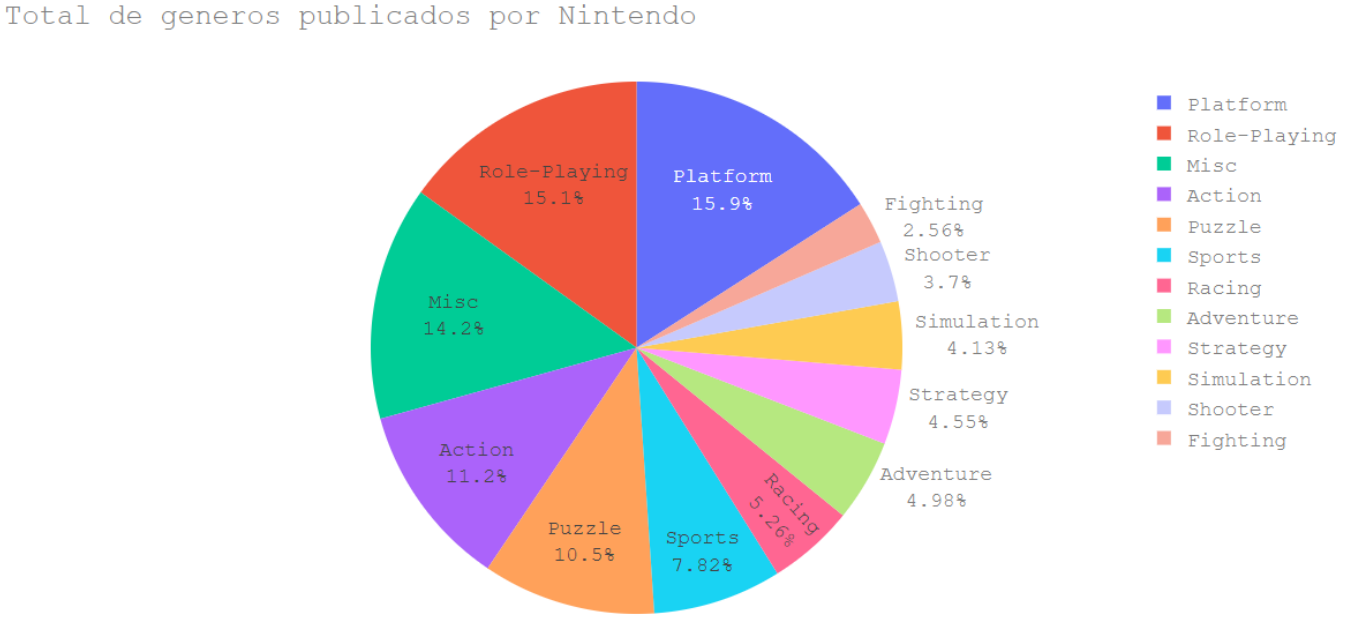
Resultados obtenidos

1. Archivo *video_games_sales*:

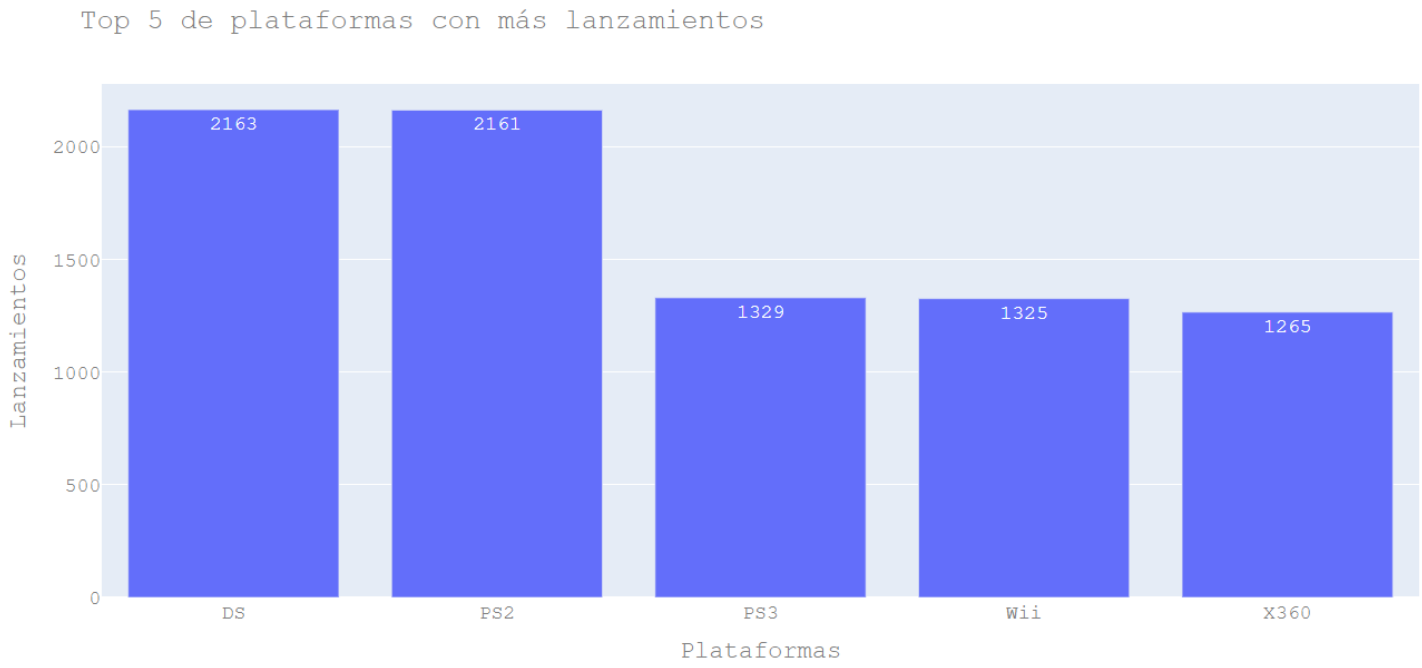
a. Gráfica de barras que muestre el total de ventas globales de las siguientes categorías: Action, Sports, Fighting, Shooter, Racing, Adventure, Strategy.



b. Gráfica de pie que muestre el total de géneros publicados por la plataforma Nintendo.

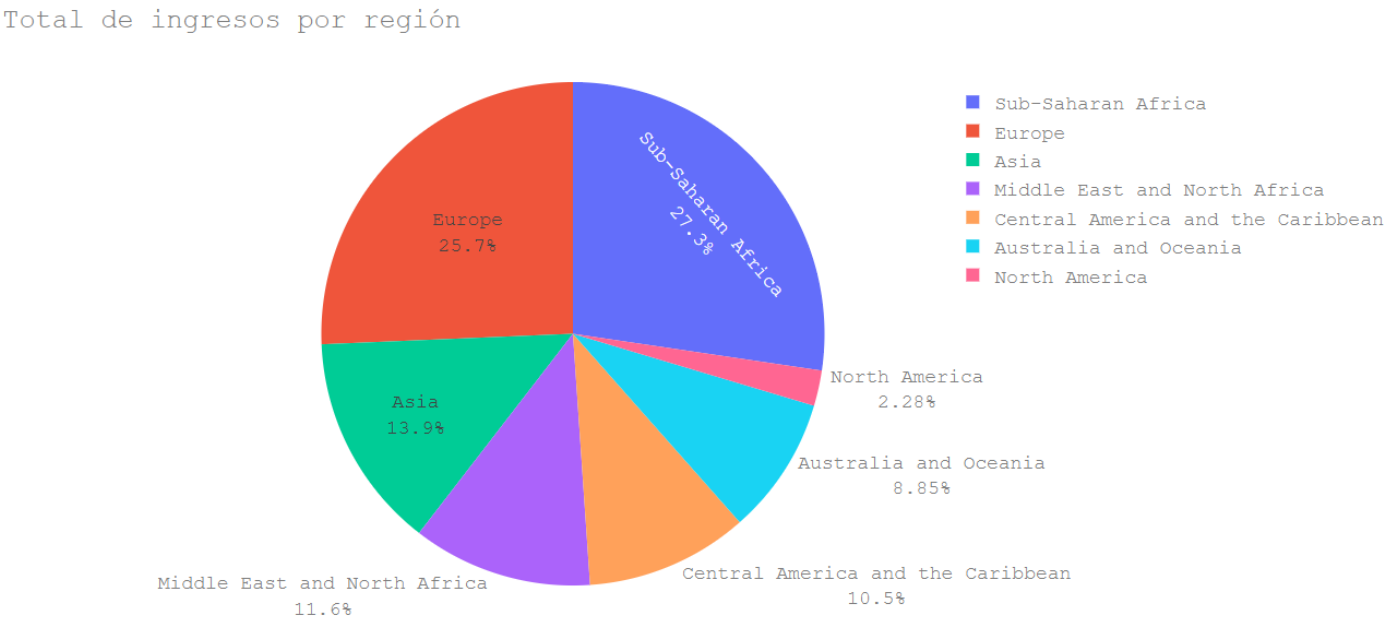


c. Gráfica de barras que muestre el top 5 de plataformas con más lanzamientos.

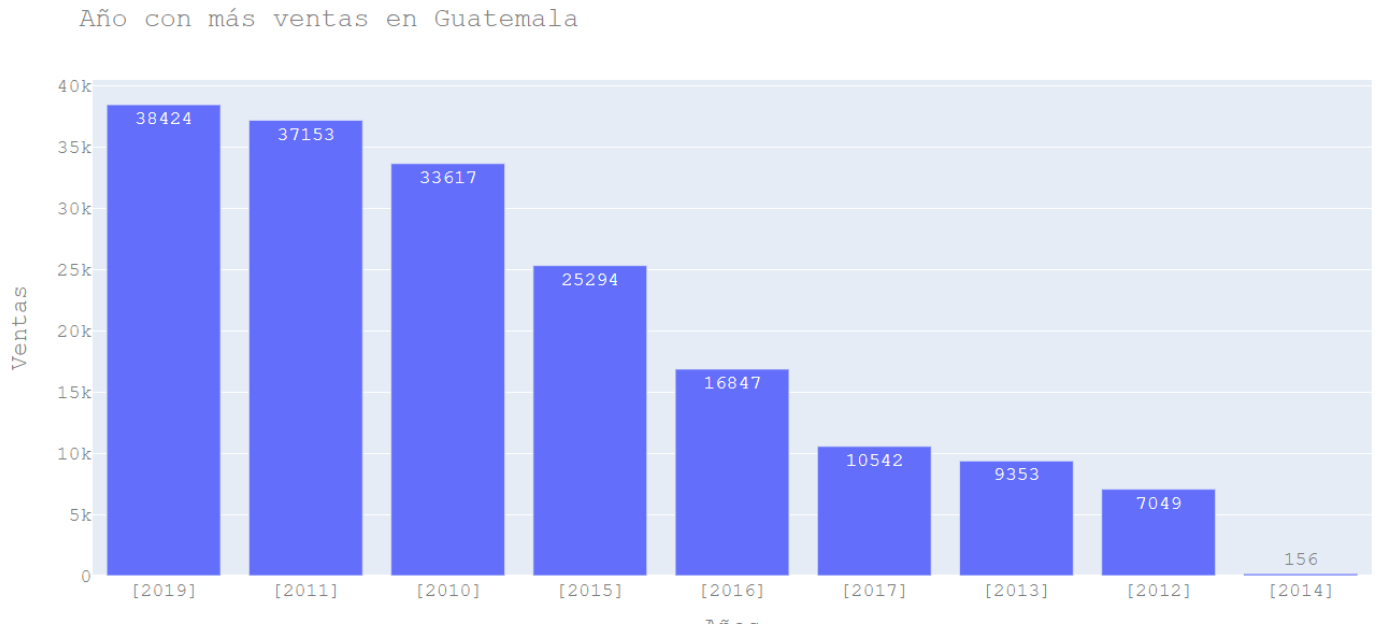


2. Archivo sales:

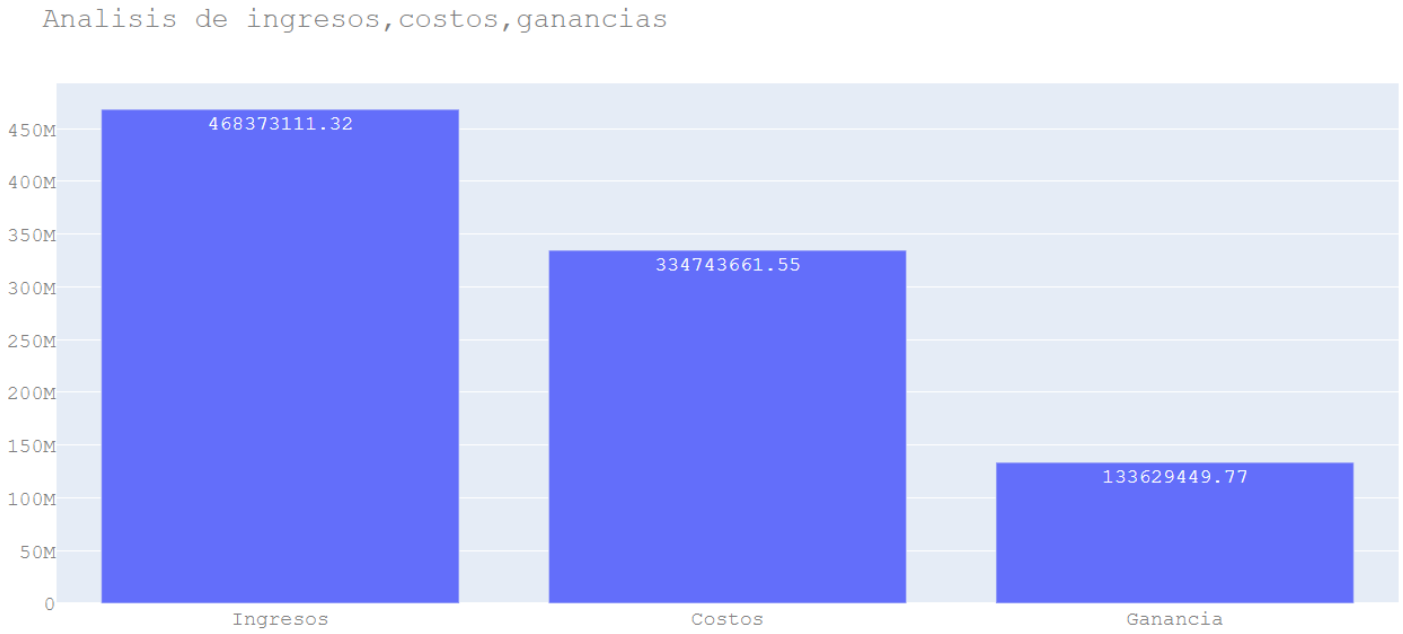
a. Gráfica de pie que muestre una comparación de los ingresos de todas las regiones (Centro América, Europa, Asia, etc).



b. Gráfica a su elección, que reporte cuál es el año con más unidades vendidas en Guatemala.

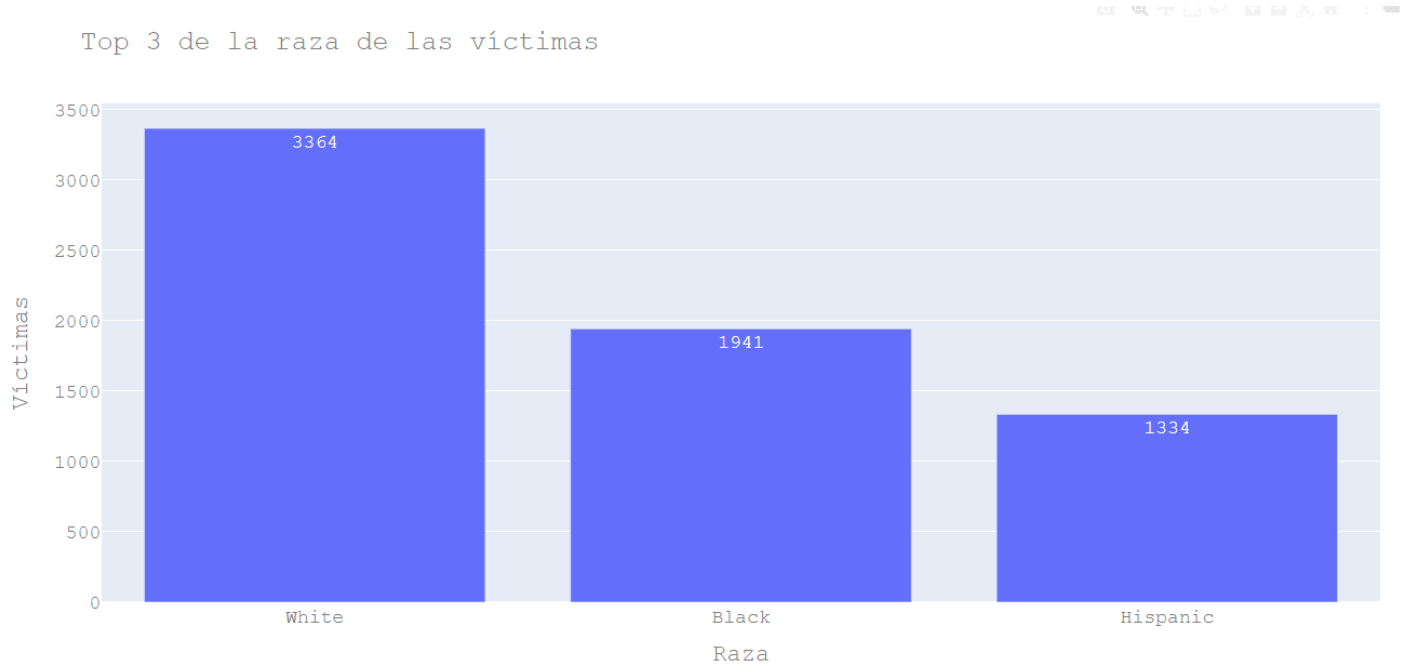


c. Gráfica a su elección, de ganancias (ingresos - costos), ingresos y costos del 2010, de las ventas en línea.



3. Archivo police_killings:

a. Gráfica de barra que muestre el top 3 de las razas de las víctimas.



b. Gráfica de barra que muestre el top 5 de años con más incidentes.