UNIVERSIDAD DE PIURA

Entrada, Sistematización y Visualización de datos en lenguaje Python

Docente: Pedro Rotta

Asignatura: Fundamentos de Python 1

Alumnos: Daniel Cobos Sanchez

Allison Maldonado Vargas

John Hubert Aquise Pari

Piura, 23 de Enero del 2022

Resumen

Actualmente para los científicos de datos representa un reto la adquisición, sistematización y el muestreo gráfico de los resultados del análisis de datos, regularmente los científicos de datos son expertos en programación y en estadística, pero no en técnicas de sistematización y diseño de gráficos, por esta razón se propone el uso de algoritmos para tales motivos.

El alcance del siguiente trabajo fue hasta las representaciones graficas de los datos adquiridos.

Con los datos obtenidos se hizo una sistematización y clasificación de los mismos, con los resultados obtenidos de la programación de los mismos patrones y sus métricas, pero con los resultados programados en Python, y a partir de esto llegar a las conclusiones de que un lenguaje es más recomendable para hacer ciencia de datos y visualización de patrones con software de calidad.

Palabras clave: Patrón, librería, paquete, visualización de datos.

1. Descripción del problema

En este capítulo se presenta una breve descripción del contenido de este trabajo. El objetivo es dar a conocer la motivación que llevamos en la implementación de adquisición, sistematización y de visualización de datos y sus respectivos casos prácticos, desarrollados en un lenguaje de programación python.

1.1. Motivación y descripción del problema

En la actualidad uno de los principales problemas que abordan los científicos de datos es encontrar la mejor manera para recopilar y representar gráficamente los resultados que dan respuesta a una pregunta que ha requerido análisis de información, y que dicha respuesta, sea fácil de entender e interpretar por el público o la audiencia a la que se quiere presentar ya sean del campo de la ciencia de datos o no.

En el intento de dar respuesta al problema de la recopilación y representación gráfica de datos, estadistas, artistas y diseñadores gráficos han incursionado en el mundo de la ciencia de datos en una nueva disciplina denominada “Visualización de Información” o también conocida como “Visualización de Datos”; sin embargo, la convergencia entre personas con un perfil de artes con personas con un perfil en ciencias, no ha sido fácil. Como una propuesta para introducir a los profesionales del diseño en el mundo de la visualización de información.

La importancia de la aplicación de patrones de recopilación y visualización de datos radica en mostrar la interpretación de los datos a partir de gráficos, dado que cualquier cosa de la vida real de manera sencilla y entendible para el publico en general. Es necesario el uso de patrones de visualización de datos para exponer diferentes tipos de datos. Por ejemplo, para una empresa es más sencillo hacer una comparación de sus ganancias por mes en su año actual frente al año pasado con un gráfico de multilíneas que lo muestre todo de forma compacta, con alzas y bajas a usar un gráfico de barras simples o barras múltiples que muestra aumentos puntuales en sus barras.

1.2. Descripción general de la solución

Tomando en cuenta la problemática generada en la visualización grafica de los datos, basados en el trabajo realizado, se tomaron los patrones de despliegue propuestos por el mismo Behrens y se desarrollaron en el lenguaje de programación R. A dichos patrones se les emplearon métricas de calidad de software

seleccionadas a partir de la técnica AHP (Analytic Hierarchy Process), y con eso dar respuesta a la siguiente pregunta:

¿Es R mejor lenguaje estadístico que Python para realizar visualización de datos?

2. Estado de la práctica

En este capítulo se presentan los trabajos previamente revisados, los cuales podrían presentar una respuesta a la pregunta: ¿Es R mejor lenguaje estadístico que Python para realizar visualización de datos?, también se mostrará una breve descripción de la literatura referente a la visualización de datos, su aplicación en el área software y la calidad del mismo, los cuales son los elementos más relevantes en que se basa este trabajo.

2.1. Marco Teórico

2.1.1 Visualización de datos y patrones de visualización

Hoy en día estamos viviendo una época donde las formas de enseñar, aprender, de trabajar e incluso de interactuar son distintas. La revolución tecnológica que ha traído el uso de computadora, internet, teléfonos inteligentes, aplicaciones y distintos tipos de multimedios ha llevado tanto la industria, la academia, al estado y a cualquier persona civil a utilizar programas de software. Tan solo en el Perú según la investigación (2019) en 2011 se estimaba que había 19 millones de dispositivos móviles, que se traduce a 68.5 dispositivos por cada 100 habitantes, quienes hacen descargas de aplicaciones de las cuales 35% son redes sociales. Esto quiere decir que tan solo en redes sociales se manejan miles de millones de datos personales que en algún lado están guardados y que también se debería de tener acceso a todos ellos. Aquí es donde reside la importancia de la visualización de datos, en hacer muestras gráficas de los datos que se tiene guardados.

La visualización de datos permite a los mismos ser expresados de forma rápida, clara y sencilla a través de gráficos y símbolos, gracias a ello podemos describir la visualización de datos como un lenguaje universal. El objetivo principal de la visualización de datos es comunicar la información de forma clara y eficaz a través de medios gráficos. No significa que la visualización de datos tiene que parecer aburrida para ser funcional o muy sofisticada para verse hermosa. Para transmitir ideas eficazmente, tanto la forma estética y funcionalidad necesitan ir de la mano Friendly (2008). Lo cierto es que aunque para algunas investigaciones científicas, el muestreo de los datos puede no ser estrictamente visual, es una extensión adicional incorporar las visualizaciones, ilustraciones y gráficas visuales de ideas abstractas con expresión simbólica previa.

Cualquier objeto del mundo real puede ser traducido a datos estadísticos, y estos mismos se pueden representar de forma gráfica, El proceso de representación y del cual se encarga el científico de datos que según Josh Willls “es una persona que sabe más de estadística que cualquier programados y que a la vez sabe más de programación que cualquier estadista". Suele ser una tarea donde se emplea recolección de datos, estadística, ingeniería y diseño. Puesto que no todos los datos son iguales tampoco se deberían representar igual gráficamente, pero tampoco puede haber una gráfica por cada tipo de objetos, es aquí donde se recomienda el uso de patrones, estos contribuirán a tomar la decisión de qué forma gráfica se podrían representar mejor los datos.

Del mismo modo que en la arquitectura, los patrones son métodos que describen el proceso de planeación y construcción de un producto. De hecho, el término diseño de patrones fue implementado por primera vez por Christopher Alexander en arquitectura urbana. Behrens (2008) propone el uso de patrones para la visualización de datos dividiéndolos en categorías, las cuales a su vez tienen meta patrones que contienen los patrones de diseño. Si bien los patrones propuestos por Behrens son un gran avance, aún le falta la automatización de patrones a partir del software (Figura 2.1).

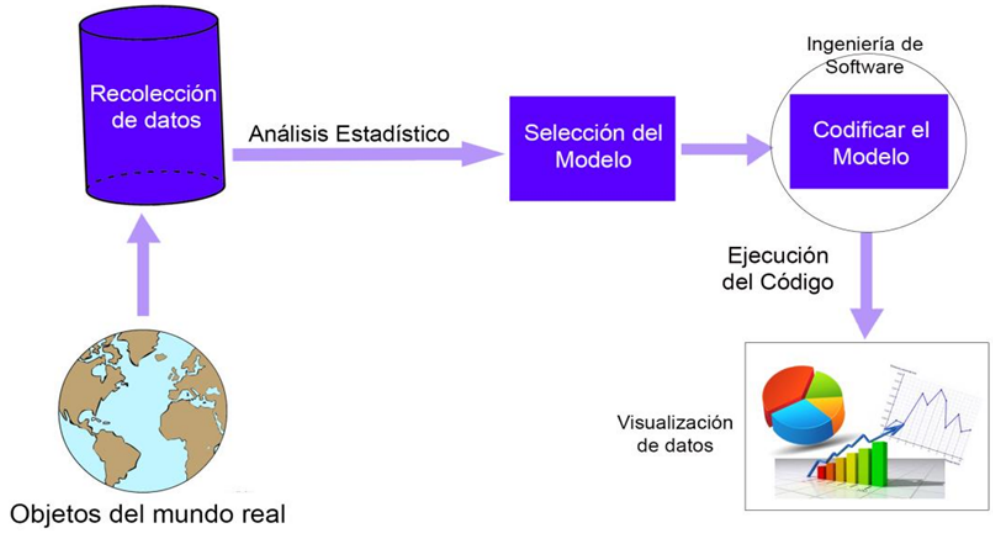


Figura 2.1 Análisis y visualización de datos

2.1.2 Lenguajes de programación y la visualización de datos

Los lenguajes de programación juegan un papel muy importante a la hora de hacer el análisis y la graficación de los datos, es imprescindible seleccionar el lenguaje conveniente para hacer el desarrollo. Existen muchos lenguajes de programación e incluso programas de software con entornos de desarrollo integrado, que permiten hacer uso de manipulación de datos. Según el sitio KDnuggets (2017), los lenguajes más utilizados por los científicos de datos, del año 2012, 2013, 2014 son R, SAS,

Python y SQL, también figuran algunos programas con entorno de desarrollo como Matlab y GNU Octave.

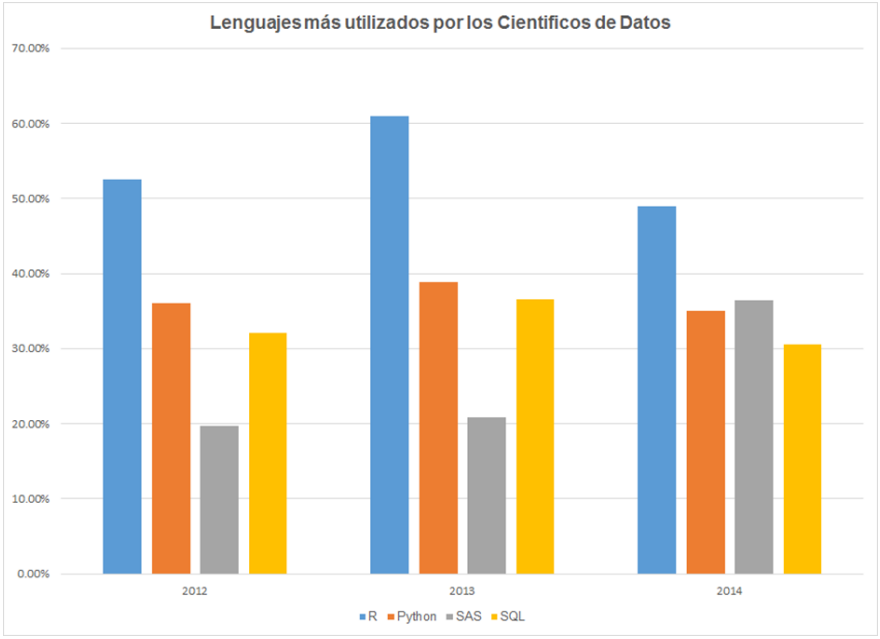


Figura 2.1.2 Lenguajes más utilizados por los científicos de datos según KDnuggets

Para este trabajo se utilizará el lenguaje Python para la implementación de patrones puesto que cuenta con las siguientes ventajas frente a los anteriores 3 más utilizados.

- Es uno de los lenguajes más utilizados por los científicos de datos
- Tiene una de las comunidades más grandes de desarrollo y de soporte
- Es un lenguaje de código libre
- Es sencillo de aprender
- Existe un gran número de librerías de código abierto para el uso de gráficos

2.2. Soluciones existentes

Actualmente existen programas de software, lenguajes de programación, paquetes para lenguajes y aplicaciones para la visualización de los datos, estas herramientas proporcionan distintos gráficos para proporcionar de forma amigable la visualización de los datos. Pero, ¿Realmente existe una aplicación, programa, paquete o lenguaje en el cual se puedan implementar todos los patrones propuestos por Behrens? La respuesta en definitiva es no, dado que el uso específico de patrones no es del todo común en la visualización de datos programadores y diseñadores se han enfocado a los gráficos más usuales en sus aplicaciones. Algunos programas como Matlab y Octave son frecuentemente utilizados para el desarrollo e implementación de gráficos. Sin embargo, no todo está perdido lenguajes de programación como Python y su amplia comunidad han desarrollado un gran número de librerías para la

visualización de datos, dado que Python es un lenguaje de sintaxis simple, es sencillo de aprender y se pueden generar casi cualquier tipo de gráficos o patrones.

Soluciones con Python

Si bien Python no es un lenguaje enfocado totalmente a la estadística, es un lenguaje que también es muy popular entre los científicos de datos. Por el mismo motivo que Python no es un lenguaje totalmente enfocado para la manipulación de datos, los códigos elaborados en el lenguaje para algunos patrones suelen ser más extensos que los códigos desarrollados en R. El lenguaje cuenta con una gran cantidad de librerías para hacer visualización de datos, pero igual que con R no existe una aplicación o librería que contenga todos los elementos para la visualización de los patrones propuestos por Behrens. En el siguiente enlace, en la pestaña Python se encuentran las librerías que se podrían utilizar para la implementación de patrones de despliegue.

2.3 Conclusiones acerca del estado de la práctica

2.3.1 Soluciones en el mercado

Tomando en cuenta lo importante que se ha vuelto la manipulación de los datos, y la visualización gráfica de los mismos, no existen herramientas que den solución total al problema de la visualización. En el mercado existen algunos programas que haciendo uso de ellos se pueden generar algunos gráficos predeterminados en el mismo programa, pero ninguno sigue la especificación de patrones ni mucho menos cuenta con todos, puesto que la mayoría de esos programas tiene el uso de gráficos más comunes y no aquellos que se utilizarían en casos específicos.

Softwares comerciales y con entornos de desarrollo como Matlab y Octave suelen ser muy utilizados para el análisis de datos a un bajo nivel y no tanto en la ciencia de datos, no es posible la generación de gráficos avanzados ni la personalización de todas las gráficas.

2.3.2 Soluciones desarrollo de software

Lenguajes de programación orientados a la manipulación de datos como R y Python tienen un gran número de paquetes y librerías que se utilizan en el campo del análisis y visualización de datos, pero para poder programar todos y cada uno de los patrones se tendría que hacer una búsqueda detallada de las librerías y paquetes para conocer cuales podrían programar ciertos patrones. Una de las desventajas es que no todas las librerías son de código libre y se tiene que hacer un pago para el uso de ellas, la ventaja de que existan muchas librerías y paquetes es que buscando detenidamente se pueden encontrar algunas que son de código libre y pueden ser utilizadas con facilidad.

El lenguaje Python cuenta con una sintaxis sencilla de comprender y aprender y eso aminora la curva de aprendizaje a la hora de hacer manipulación de datos utilizando los paquetes para Python, no tiene un paquete donde se puedan graficar cada uno de los patrones, pero, facilita el trabajo el poder conocimiento del lenguaje y los ejemplos de propietarios de paquetes.

3. Análisis del programa

Una vez iniciado el programa, este mostrará dos opciones, podrás pedirle que lea un documento de Excel(xlsx) preexistente a partir de una ruta o crear uno nuevo con el nombre “lista.xlsx”. Si se elige crear un nuevo documento se deberá ingresar el nombre de las columnas.
Luego el programa mostrará un menú con diferentes opciones que te permitirán manejar los datos, añadir nuevas características, graficar y relacionar columnas, filtrar datos, etc. El menú se mostrará continuamente después de cada acción, cada vez que vuelva aparecer se guardarán los cambios en el documento de Excel.
La primera opción del menú permite ingresar nuevas filas, si no es la primera entrada el programa condiciona al usuario para ingresar el mismo tipo de datos de la columna.
La segunda opción buscará las filas que coincidan con un dato ingresado dentro una columna específica y mostrará una tabla con los mismos.
La tercera opción permite filtrar los datos de una columna para que cumplan con una característica (mayor a, menor a) condicional con respecto a un dato ingresado, mostrará estos datos dentro de una tabla.
La cuarta opción mostrará una tabla con datos estadísticos básicos de todas las filas que contengan números.
La quinta y sexta opción mostraran una gráficas de pastel y de barras, respectivamente, a partir de los datos de una columna.
La séptima opción graficara la relación entre dos características de los datos mediante una gráfica de puntos.

- **Menú**

```
Bienvenido
0: Usar un excel ya existente
1: Crear uno nuevo
Ingrese: 0
Porfavor ingrese la ruta del archivo(.xlsx): /content/lista.xlsx
0: Ingresar datos(fila)
1: Buscar datos(igual a)
2: Filtrar por caracteristica(mayor a,menor a)
3: Ingresar caracteristica(columna)
4: Datos estadisticos basicos
5: Grafica pastel
6: Grafica de barras
7: Grafica de puntos
8: Mostrar tabla
Ingrese culaquier otra cosa para salir
--------------------------------
Ingrese:
```

- **Nueva fila**

```
Ingrese: 0
Porfavor ingrese:
Id: 52
DNI: 75165552
Nombre: Daniel
Género: Masculino
Tit.: Ingenieria
Movil/Pag.: 969854123
Cumpleaños(AAAA/MM/DD): 1990/08/25
Fecha Inicio(AAAA/MM/DD): 2000/01/02
Añadir: a
Tel.Oficina: numero
El dato Tel.Oficina debe ser un numero de tipo float64: 123456
Dpto.: Finanzas
Administrador: S
Nacionalidad: Perú
CardNo: 123
Privilegio: no
Se añadio:
{'Id': 52, 'DNI': 75165552, 'Nombre': 'Daniel', 'Género': 'Masculin
S para volver al menu:
```

- **Filtrar**

```
6: Cumpleaños
7: Fecha Inicio
8: Añadir
9: Tel.Oficina
10: Dpto.
11: Administrador
12: Nacionalidad
13: CardNo
14: Privilegio
Buscar en: 0
El tipo de dato para Id es int64
Buscar por:
0: Mayor a
1: Menor a
2: Dentro del rango
Ingrese: 0
Valores mayores a: 1025
```

	Id	DNI	Nombre	Género	T
25	1026	1217286	FLORES ORDONO Elba Noemí	Femenino	Tercero Da

- **Buscar**

```
0: Id
1: DNI
2: Nombre
3: Género
4: Tit.
5: Movil/Pag.
6: Cumpleaños
7: Fecha Inicio
8: Añadir
9: Tel.Oficina
10: Dpto.
11: Administrador
12: Nacionalidad
13: CardNo
14: Privilegio
Buscar en: 0
El tipo de dato para Id es int64
¿Que desea buscar?: 1008
```

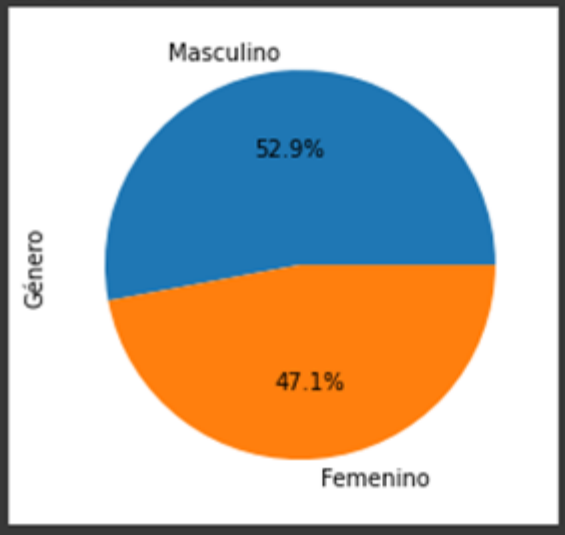
	Id	DNI	Nombre	Género	Tit.	Movil/Pag.	C
7	1008	1503997	CAMACHO FIGUEROA, Maryoris Camacho	Femenino	SECRETARIA	933298790	

- **Nueva columna**

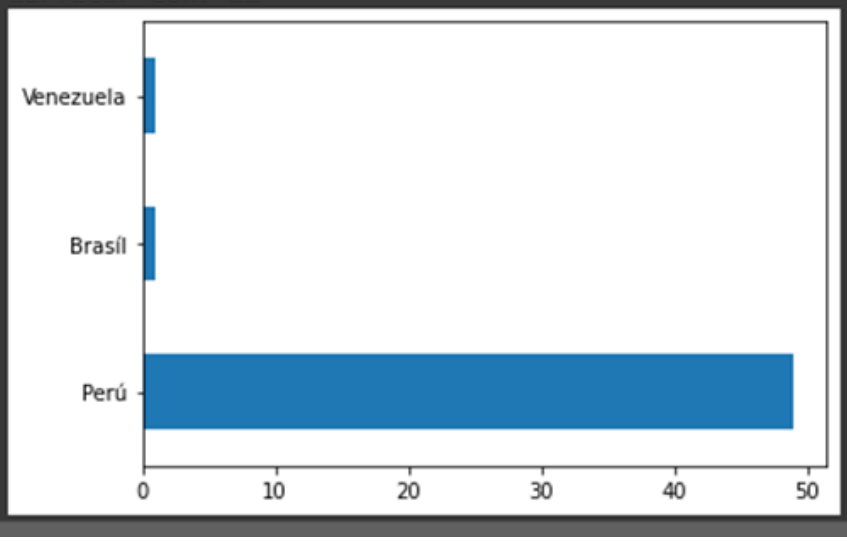
```
Ingrese el nombre de la nueva columna: abc
Actual mente se cuenta  con 51 filas, debera ingresar 51 datos
S para iniciar: s
0: a
1: b
2: c
3: d
4: e
5: f
6: g
7: h
8: i
9: j
10: k
11:
```

- **Gráficas**

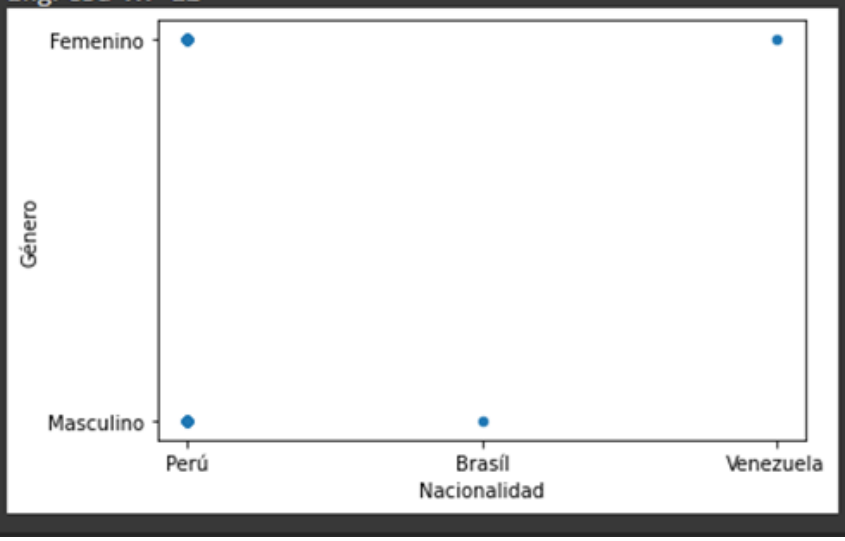
```
7: Fecha Inicio
8: Añadir
9: Tel.Oficina
10: Dpto.
11: Administrador
12: Nacionalidad
13: CardNo
14: Privilegio
Graficar con: 3
```



```
9: Tel.Oficina
10: Dpto.
11: Administrador
12: Nacionalidad
13: CardNo
14: Privilegio
Garficar con: 12
```




```
9: Tel.Oficina
10: Dpto.
11: Administrador
12: Nacionalidad
13: CardNo
14: Privilegio
Ingresa Y: 3
Ingresa X: 12
```



4. Conclusiones

Conocer la sintaxis, los equivalentes y la manera en la que un comando regresa datos es extremadamente importante para poder escribir códigos coherentes y simples. Por esto es necesario siempre estar al día con las nuevas librerías y las actualizaciones de las ya existentes.

Google Colab, gracias a sus diferentes herramientas y que no es necesario poseer un dispositivo suficientemente potente para programar es un sistema que facilita en gran medida el aprendizaje de un nuevo lenguaje de programación.

Con el programa que hemos desarrollado se podrán organizar mucho mejor los datos para administrar una empresa con mucho más facilidad, y con un programa amigable para que todo el mundo pueda entenderlo y ejecutarlo.

Bibliografía y Anexos

1. Librerías seleccionadas para el lenguaje de programación R y Python: https://docs.google.com/spreadsheets/d/1xcyRa6fG5uvYYa4FQLtY-1y9JNoVtxSJSbUuE2AeCFI/edit#gid=0