CIENCIA DE LOS DATOS CON PYTHON

SISTEMA DE INFORMACIÓN EN LÍNEA
SIL
COMPLEMENTO VIRTUAL

ECOE EDICIONES

Francisco J. Toro López

CIENCIA DE LOS DATOS CON PYTHON

Francisco J. Toro López

Catalogación en la publicación - Biblioteca Nacional de Colombia

Toro López, Francisco J.

Ciencia de los datos con Python / Francisco J. Toro López. -- 1a ed. -- Bogotá : Ecoe Ediciones, 2022.

201 p. – (Ingeniería y afines. Informática)

Contiene datos curriculares del autor. -- Contiene referencias bibliográficas.

ISBN 978-958-503-356-6 -- 978-958-503-357-3 (e-book)

1. Minería de datos - Técnicas 2. Búsquedas en bases de datos 3. Python (Lenguaje de programación de computadores) I. Título II. Serie

CDD: 006.312 ed. 23 CO-BoBN– a1090613

Área: *Ingeniería y afines*
Subárea: *Informática*

© Francisco J. Toro López

► Ecoe Ediciones S.A.S.
info@ecoeediciones.com
www.ecoeediciones.com
Carrera 19 # 63C 32, Tel.: 919 80 02
Bogotá, Colombia

Primera edición: Bogotá, mayo del 2022

ISBN: 978-958-503-356-6
e-ISBN: 978-958-503-357-3

Directora editorial: Claudia Garay Castro
Coordinadora editorial: Paula Bermúdez B.
Corrección de estilo: Andrés Caro
Diagramación: Magda Rocío Barrero
Carátula: Wilson Marulanda Muñoz
Impresión: Carvajal Soluciones de Comunicación S.A.S.
Carrera 69 #15-24

Prohibida la reproducción total o parcial por cualquier medio sin la autorización escrita del titular de los derechos patrimoniales.

Impreso y hecho en Colombia - Todos los derechos reservados

DEDICATORIA

A mis padres, Darío Toro Ossa y Lucila López de Toro (Q.E.P.D.), a mis hermanos Rubén e Iván por el abierto e incondicional apoyo en mi formación profesional, sentimental y personal.

Por su cariñosa y siempre amable aptitud de ayudarme a descubrir lo que es fundamental en nuestras vidas, a mi hija, Viviana Alejandra, a mi Myriam Abondano, Myriam Guzmán, Javier Llorca y esposa Jimena, a mis sobrinos Mauricio y Camila Toro; Claudio, Oscar y Nicolás Rodríguez; Julián y Paola Beltrán; Enrique, Jorge Eduardo y Andrés Abondano, y María Paula y Gaby Guzmán.

Un agradecimiento especial y cariñoso a María Constanza García y a Claudia Quiroga por su gracioso y entretenedor esparcimiento con la música en los ratos algo diferentes de las labores de investigación que exige escribir un libro sobre estos temas.

AGRADECIMIENTOS

El autor expresa formalmente sus agradecimientos a los ingenieros Fernando Salazar D., Germán Moreno y Héctor Beltrán por su valiosa cooperación y comentarios que sirvieron para dar una mejor presentación y, posteriormente, una adecuada elaboración del material textual y gráfico de la obra.

Así mismo, un especial reconocimiento a los administradores de los programas de capacitación del EdX, entre los que se distinguen IBM y la Universidad de Harvard, por el aporte que recibí al tomar varios de sus cursos y el suministro de material valioso, empleado en la presentación de diversas explicaciones y temas sobre los procedimientos y herramientas usualmente encontrados en **Python** y **Microsoft Project.**

CONTENIDO

Observaciones generales XXV
Introducción XXIX

Capítulo 1. El lenguaje Python 1
Introducción al manejo de Python 3
Instalación de Python 5
El editor IDLE 8
Tipos de datos 8
Ejercicios 10
Las cadenas de caracteres 11
Valores booleanos (bool) 13
La interfaz de Python 14
La versión de Python en uso 17
Escribir comentarios en Python 18
Errores en Python 19
¿Python sabe de un error antes de que se ejecute el código? 20
Manejo de caracteres 20
Escribir el primer programa con Python 20
Tipos de objetos en Python 21
Manejo de cadenas de caracteres 22
Listas y tuplas 24
Funciones integradas 35

Funciones básicas matemáticas 36
Operadores de comparación 38
Operadores de identidad y pertenencia 39
Manejo de diccionarios 40
Colecciones (*set*) de datos 44
Observaciones sobre secuencias: cadenas, listas y tuplas 50
Paquetes complementarios de Python 51
Resumen de Python en la ciencia de los datos 52
Funciones 54
Apertura de archivos 58
Leyendo Archivos 60
Programas en red 61
El protocolo de transporte de hipertexto (HTTP) 61
El navegador web más sencillo 62
Análisis de html y barrido de la web 64
Análisis de HTML mediante BeautifulSoup 65
Los programas 66
Uso de objetos 67
Comenzando a Programar 68
Subdividir un problema - Encapsulación 70
Los objetos Python 71
Depuración (*debugging*) 71
Glosario 72

Capítulo 2. Términos, técnicas y tareas de las ciencias de datos 75
¿Qué es la minería de datos? 76
¿Qué es la ciencia de los datos? 77
Los científicos de datos 77
big data 80
Nuevos dispositivos 81
El *big data* requiere de un Big Analytics 81
Las limitaciones del *business intelligence* tradicional 82
La evolución del DWH al ADW 83
MAD, nueva metodología 83
Medio Magnético 83
Ágil 84
Detallado 84
Hadoop 84
HDFS 85
MAPREDUCE 85
La Ciencia de los datos 89

Adquisición de datos: paso 1 99
Manipulación de datos: paso 2 102
Aplicación de técnicas de análisis: paso 3 107
Reportar la información: paso 4 114
En acción: paso 5 117

Capítulo 3. Adquisición y manipulación de datos 121
Importación y exportación de datos con Python 122
Caso 1: análisis de datos 122
Introducción al análisis de datos con Python 125
Caso 2: el Castillo de Edimburgo 125
Inicio del análisis 126
Desmembrando todo 126

Capítulo 4. Análisis de los datos y diseño y presentación del modelo 129
Verificando la instalando Python 129
Recomendaciones para Windows 130
La integración 131
Convertir los resultados en acción 131
Las diferencias de las métricas dependiendo de las variables y tareas 142
Datos históricos + datos cercanos a la realidad → predicciones 142
Eliminar filas en el DataFrame Claudia con base a sus valores 143
Conjunto de datos de ejemplo 143

Capítulo 5. Evaluación de modelos: la vista 'minable' y la comprensión de datos 147
¿Qué es la vista 'minable'? 147
¿Para qué se utiliza una vista 'minable'? 148
Técnicas que convierten datos en modelos y procesos de predicción en temas empresariales 150
Análisis exploratorio 150
Métodos de estimación y predicción 151
Un caso manejado con tres técnicas matemáticas 153
Alisado exponencial simple 153

Capítulo 6. Proyectos de las ciencias de datos 161
Planeación y desarrollo de proyectos de la ciencia de los datos 161
Definición ampliada de las delimitaciones 163
Tiempos, costos y riesgos de proyectos 166
Las fases de un proyecto 166
Definición y creación del plan básico del proyecto 166

Ajustes e informes 170
Desarrollo y supervisión del proyecto 170
Cierre y evaluación de un proyecto 171
Metodologías empleadas para administrar proyectos 171
La asignación de recursos 174
Los costos de un proyecto 177
Caso del Castillo de Edimburgo 185

Anexos 189
Referencias 199

ÍNDICE DE FIGURAS

Figura 1. Cargue del editor IDLE de información 8
Figura 2. Tres tipos de datos en Python....8
Figura 3. Datos numéricos enteros....9
Figura 4. Datos numéricos flotantes 9
Figura 5. Tipos de datos más comunes....9
Figura 6. Datos numéricos enteros.... 10
Figura 7. Cadenas de datos flotantes.... 11
Figura 8. Convirtiendo flotantes a enteros.... 11
Figura 9. Usando el comando tipo booleano 13
Figura 10. Asignar un “0” a un dato booleano, se consigue un “False” 14
Figura 11. Un pequeño código fuente.... 15
Figura 12. Cargue del programa Python 17
Figura 13. Los caracteres de una cadena son indexados 20
Figura 14. Los elementos de una tupla son indexados.... 24
Figura 15. El tipo de una tupla es siempre TUPLA.... 25
Figura 16. Se accede al elemento de una tupla por su índice.... 25
Figura 17. Concatenando dos tuplas se puede realizar.... 26

Figura 18. Eliminando elementos de una tupla 26
Figura 19. Dejando a un lado elementos de una tupla..... 26
Figura 20. Dejando a un lado elementos de una tupla..... 26
Figura 21. Uso de la función "len" en una tupla 27
Figura 22. Las tuplas son inmutables 27
Figura 23. Las tuplas son inmutables 27
Figura 24. Como las tuplas son inmutables, si se quiere cambiar una, hay que crearla 28
Figura 25. Ordenando una tupla..... 28
Figura 26. Anidando una tupla en otra..... 28
Figura 27. Accesando un elemento anidado en una tupla 29
Figura 28. Accesando elementos anidados en una lista..... 29
Figura 29. Acceso a un elemento anidado en una tupla 30
Figura 30. La relación entre un índice y los elementos de una lista..... 30
Figura 31. Uso de un índice negativo en un elemento de una lista (*List*)..... 30
Figura 32. Segmentando un elemento de una lista..... 31
Figura 33. Convenciones de índice de una lista..... 31
Figura 34. Concatenar o combinar listas de datos..... 31
Figura 35. Cambiando elementos de una lista 31
Figura 36. Concatenando una lista a otra, usando el método *extend* 32
Figura 37. Cambiando listas a otra lista con *extend* 32
Figura 38. Agregando elementos a otra lista con *append* 32
Figura 39. Modificando una lista con *extend* 33
Figura 40. Modificando una cadena a una lista con *extend* y *append* 33
Figura 41. Eliminando un elemento de una Lista con "del"..... 33
Figura 42. Agregando un elemento a una lista, usando índices 33
Figura 43. Con la función Split se separan elementos de una lista 34
Figura 44. Con la función Split se cambian elementos de una lista 34
Figura 45. Con la función de alias, se cambian elementos de una lista 34
Figura 46. Con la función alias, se cambia un elemento de varias listas 35
Figura 47. Copiando elementos de una lista a otra 35
Figura 48. Con la función help se brinda información de una lista..... 35

Figura 49. La función Range() calcula el rango de valores consecutivos de una serie ... 36
Figura 50. La creación de un diccionario en Python ... 41
Figura 51. Generación de un diccionario ... 41
Figura 52. Características de un diccionario ... 41
Figura 53. Datos de un diccionario ... 42
Figura 54. Año de creación y título de un diccionario de películas ... 42
Figura 55. Verificando si el título de una película está en un diccionario de películas ... 43
Figura 56. Borrando el título de una película ... 43
Figura 57. Mostrando todos los títulos de películas ... 43
Figura 58. Muestra los contenidos de un conjunto de datos ... 44
Figura 59. Una lista conforma un conjunto de datos ... 44
Figura 60. Un conjunto de datos no repetidos y ordenados ... 45
Figura 61. Eliminando un elemento de un conjunto, usando la función remove ... 45
Figura 62. Verificando si un elemento del conjunto está, con la función in ... 46
Figura 63. Verificando que un elemento del conjunto NO está con la función *in* ... 46
Figura 64. Representación de un elemento del conjunto que sí está ... 46
Figura 65. Interceptación de un elemento en varios conjuntos ... 47
Figura 66. Interpretación de varios elementos en dos conjuntos ... 48
Figura 67. Interceptación de varios elementos con Python ... 48
Figura 68. Un nuevo álbum conteniendo varios elementos en dos albúmenes con Python ... 48
Figura 69. Unión de conjuntos conteniendo elementos comunes con Python ... 48
Figura 70. Representación gráfica de la unión de conjuntos ... 49
Figura 71. Representación gráfica de si un conjunto es subconjunto de otro ... 49
Figura 72. Representación gráfica de uniones, o no, de conjuntos ... 50
Figura 73. Python shell ... 52
Figura 74. Python shell ... 52
Figura 75. Siguiente posición ... 53
Figura 76. Python usa sangría en lugar de corchetes ... 53

Figura 77. Programa de valores de variables y no de tipos 53
Figura 78. Programa que imprime los 11 valores del 0 al 10 54
Figura 79. Programa definido con la función valor absoluto 54
Figura 80. ¡Programa que no devuelve nada!..... 55
Figura 81. Programa que no devuelve nada 55
Figura 82. Función *valor absoluto*..... 58
Figura 83. Uso de sockets en archivos HTTP 63
Figura 84. Estructura típica de un programa 69
Figura 85. Uso de biblioteca urllib y otros objetos en archivos HTTP..... 70
Figura 86. Confluencia de conocimientos requeridos en la ciencia de los datos..... 90
Figura 87. Extensión de conocimientos requeridos en la ciencia de los datos..... 91
Figura 88. Habilidades y conocimiento de un científico de datos 91
Figura 89. Python vs. otras herramientas de la informática para el científico de datos 92
Figura 90. Transformando los datos en conocimiento 94
Figura 91. Conjunto de datos etiquetados..... 97
Figura 92. Adquirir los datos..... 100
Figura 93. Manipulación de datos 102
Figura 94. El manejo de datos inconsistentes 103
Figura 95. Arreglando los datos inconsistentes 103
Figura 96. Arreglando datos inconsistentes 104
Figura 97. Escalando los valores de datos inconsistentes..... 104
Figura 98. Transformando valores de datos inconsistentes 105
Figura 99. Removiendo, combinando y agregando valores de datos..... 105
Figura 100. Removiendo, combinando y agregando valores de datos..... 106
Figura 101. Manipulando algunos valores de datos..... 106
Figura 102. Preparación de valores de datos inconsistentes 107
Figura 103. Construyendo el modelo de datos 107
Figura 104. Categorías de las técnicas del análisis de datos 108
Figura 105. La clasificación como técnica del análisis de datos 108

Figura 106. La regresión como técnica para predecir un valor numérico del análisis 109
Figura 107. La regresión como técnica para observar algún fenómeno por grupos de edades 109
Figura 108. La asociación como técnica para hallar asociaciones de datos 110
Figura 109. Usando gráficos como técnica para analizar conexiones entre datos 111
Figura 110. Tres pasos para construir el modelo de datos 111
Figura 111. Combinando la clasificación y la regresión como técnicas del modelo 112
Figura 112. La técnica de clúster: ¿tiene o no sentido su aplicación? 112
Figura 113. Combinando la asociación y el análisis gráfico como técnica de modelo 112
Figura 114. Evaluar los resultados obtenidos al aplicar técnicas al modelo 113
Figura 115. Reportar los resultados obtenidos al aplicar técnicas 114
Figura 116. Reportar y comunicar los resultados obtenidos 114
Figura 117. ¿Cómo se van a presentar los resultados obtenidos? 115
Figura 118. Formato de presentación de los resultados obtenidos 115
Figura 119. Las herramientas de visualización de los resultados obtenidos 116
Figura 120. Las herramientas de manejo de los resultados obtenidos 117
Figura 121. Aplicar los resultados obtenidos 118
Figura 122. La implementación de una acción 118
Figura 123. Medir y evaluar la implementación de una acción 119
Figura 124. La evaluación de una acción 119
Figura 125. Caso 1: planteamiento del problema 123
Figura 126. Caso 1: adquisición y manipulación de datos 123
Figura 127. Caso 1: adquisición de datos de precios 124
Figura 128. Caso 1: modelo de datos y procesos de datos 124
Figura 129. Instalación de Python en caso de detectar un problema 130
Figura 130. Abriendo Anaconda Promo 130
Figura 131. Ventana de Python 131
Figura 132. Convertir visiones en acciones y conectar los resultados 132
Figura 133. Herramientas tecnológicas empleadas para la toma de acciones .. 132

Figura 134. Algunas formas de implementar acciones 133
Figura 135. Evaluar el impacto de una acción 133
Figura 136. Determinar los siguientes pasos después de la evaluación de una acción 134
Figura 137. Tiempo de observación (min) vs. duración del video 136
Figura 138. Tiempo de observación (min) vs. duración del video 136
Figura 139. Variables, métodos y parámetros 138
Figura 140. Dos métodos de cadenas de caracteres 138
Figura 141. Dos funciones de *lower* aplicadas a cadenas de caracteres 139
Figura 142. Imprimiendo mediante un bucle de funciones de 0 a 10..... 139
Figura 143. Imprimiendo valores pares para el rango de valores 0 a 10..... 141
Figura 144. Imprimiendo valores modulo 3 para el rango de valores 0 a 10.... 141
Figura 145. Propiedades y tipos 141
Figura 146. Se usan mediante una consulta automática 148
Figura 147. La construcción de una vista 'minable' mediante la escritura de un código (¿en Python?) 149
Figura 148. Pirámide de sabiduría, conocimiento, información y datos..... 149
Figura 149. Predicción de una variable de interés 152
Figura 150. Suavizado exponencial 154
Figura 151. Suavizado exponencial: Parámetros 155
Figura 152. Suavizado exponencial: Guardar..... 156
Figura 153. Las dimensiones básicas de un proyecto..... 163
Figura 154. Procesos administrativos de un proyecto 167
Figura 155. Fases de planeación, ejecución y control de un proyecto 170
Figura 156. Desglose de un proyecto de construcción de un edificio 172
Figura 157. Diagrama de Gantt aplicado a la construcción de un edificio 172
Figura 158. La ruta crítica en la construcción de un edificio 173
Figura 159. Cantidad de recursos vs. duración de una tarea 175
Figura 160. Opciones de programación..... 176
Figura 161. Opciones de la carpeta "Avanzado" 177
Figura 162. Tipos de costos dependiendo de su comportamiento..... 177
Figura 163. Tablas de información de Project..... 182

Figura 164. Ventana del menú principal de Project 183
Figura 165. Guía de plantillas de proyectos de Project 184
Figura 166. Diagrama de Grantt del proyecto Castillo de Edimburgo 187
Figura 167. Los recursos del proyecto Castillo de Edimburgo 187
Figura 168. Los costos del proyecto Castillo de Edimburgo 187
Figura 169. Los flujos de caja del proyecto Castillo de Edimburgo 188
Figura 170. Imágenes RGB 190
Figura 171. Manejo de imágenes en PC 190
Figura 172. Los cuadros pequeñísimos en cuadros (tiles) 191
Figura 173. Un pixel 191
Figura 174. Imagen RGB en Python 192
Figura 175. Importando imagen RGB 193
Figura 176. Modelando en WIFIRE 193
Figura 177. Beneficios de la librería Pandas 194
Figura 178. Otros beneficios de Pandas 194
Figura 179. Soportes para datos en series de tiempos 195
Figura 180. Estructura de datos en Pandas 195
Figura 181. Índices, arreglos y argumentos de series y Dataframes 196
Figura 182. Pandas: soporte de muchos pasos 197

ÍNDICE DE TABLAS

Tabla 1. Los resultados de la función *And* 39

Tabla 2. Resultados del operador *Or* 39

Tabla 3. Variables y factor de ponderación de precios de coches 125

Tabla 4. Puntaje de coches usados de similares características 125

Tabla 5. El flujo de tareas del proyecto Castillo de Edimburgo escrito con Project 186

Notas:

1. Se puede llegar a pensar que la estructuración de los capítulos 3 al 5 es un tanto inversa a lo que es el proceso natural de extracción de conocimientos a partir de datos, pero se hizo pensando en la ventaja de darle una naturaleza mucho más didáctica a los contenidos de los capítulos.
2. Algunas actividades de un proyecto se pueden repetir aún después de evaluar el resultado final y leer el informe definitivo. Ante la falta de conocimientos explícitos sobre las circunstancias propias de cada proyecto, las actividades que pudieran ser repetidas, son desconocidas y solo podrán ser reconocidas después de un análisis serio; naturalmente, son desconocidas al momento de desglosar el plan original.
3. Algunos capítulos, por su extenso contenido temático, pueden traer al final un Glosario de términos. Estos se elaboraron de acuerdo al contenido de cada capítulo. Igualmente es posible citar varias referencias bibliográficas en unos capítulos.