

# Fundamentos de Analítica de Datos: Caso 1

## (Sesión 7A)

Prof. Juan C. Correa, Ph.D.

Colegio de Estudios Superiores de Administración  
Bogotá - Colombia



## Objetivo de Aprendizaje

En la Sesión 6A, se enseñaron algunos de los fundamentos del pre-procesamiento y la visualización de datos. Al finalizar esta sesión y la siguiente (sesión 7A), usted estará en capacidad de profundizar aún más en la visualización de datos con Pandas, Matplotlib y ggplot2.



# 1

---

## Insumos Preliminares



Para esta sesión, usted continuará usando el mismo repositorio de GitHub que clonamos para la sesión 6A. Si no lo ha hecho aún, acá tiene el enlace.

The screenshot shows the GitHub interface for the repository 'jcorrean / WebMining-OFD'. At the top, there's a navigation bar with links for Pull requests, Issues, Marketplace, and Explore. Below this, the repository name is displayed along with 'Watch', 'Star', and 'Fork' buttons. A secondary navigation bar includes links for Code, Issues, Pull requests, Actions, Projects, Wiki, Security, Insights, and Settings. The main content area shows the 'master' branch with 1 branch and 0 tags. A list of files is displayed, including 'Domicilios.R', 'README.md', 'WebMiningOFD.R', 'getCommentsDomicilios.R', 'newdata.csv', and 'raw-data-of-a-web-mining-approach-...'. To the right, there's an 'About' section with a description of the repository's purpose and a 'Releases' section indicating no releases are published.

Search or jump to... Pull requests Issues Marketplace Explore

jcorrean / WebMining-OFD Watch 1 Star 1 Fork 1

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

master 1 branch 0 tags Go to file Add file Code

jcorrean Update WebMiningOFD.R	3ede5ec · on Dec 8, 2020	9 commits
Domicilios.R	Add files via upload	2 years ago
README.md	Update README.md	13 months ago
WebMiningOFD.R	Update WebMiningOFD.R	8 months ago
getCommentsDomicilios.R	Add files via upload	2 years ago
newdata.csv	Raw Data	13 months ago
raw-data-of-a-web-mining-approach-...	Add files via upload	3 years ago

**About**

This repository aims to facilitate the reproduction of the results reported in the published paper titled "Evaluation of collaborative consumption of food delivery services through web mining techniques"

**Releases**

No releases published  
[Create a new release](#)

<https://github.com/jcorrean/WebMining-OFD>



# 2

---

## Fundamentos de Pre-procesamiento de Datos (Continuación)



En la sesión 6A (slide 10) afirmamos que el futuro gerente de una **empresa data-driven** debe entender en qué consiste el trabajo de un experto en analítica de datos.



La razón que sustenta esa idea es que usted tendrá que decidir entre contratar empleados con experiencia o seleccionar proveedores expertos en analítica de datos. En cualquier caso, su decisión será acertada si y solo si usted sabe detectar esos criterios de experticia.



Incluso, en el ámbito del emprendimiento es fundamental entender la visión de los expertos en analítica de datos (Kotha, Kim, y Alexy, 2014). ¿Por qué?

HBR.ORG

### Idea in Brief

#### THE PROBLEM

Many inventors struggle to commercialize their inventions or discoveries successfully. All too often large companies, investors, or others walk off with the fruits of a scientist's work.

#### WHY IT HAPPENS

Commercial success with a new technology depends on the exclusive ownership of a critical asset or capability. But to create the technology, innovators draw on knowledge from many different sources. Inventors who mismanage that tension often fail to successfully commercialize their innovations.

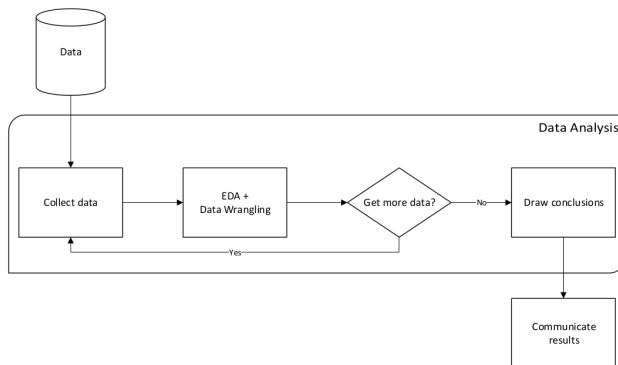
#### THE SOLUTION

To manage the tension, inventors must successfully avoid the following traps: prematurely disclosing proprietary information; neglecting policeability; failing to demonstrate originality; overrelying on known science; failing to stake out the best territory; mismanaging attribution; and falling into funders' clutches.

Porque un experto en analítica de datos es un profesional con un sólido entrenamiento en ciencias (e.g., estadística, matemáticas, física, computación).



En la sesión 6A (slide 6) afirmamos que el **el pre-procesamiento de datos** comprendía a todas las actividades que preceden al análisis de los datos.

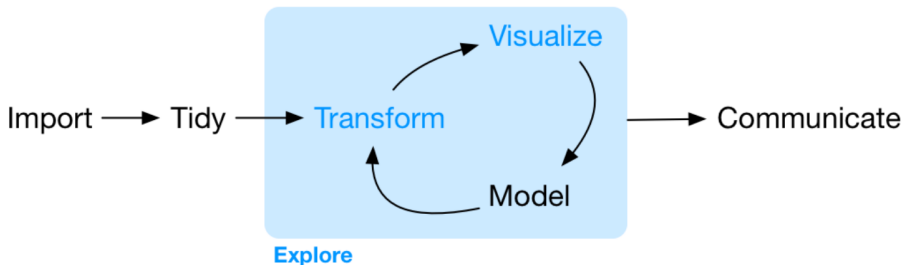


Para Molin (2021) esas actividades son Data, collect data, y EDA + Data Wrangling.

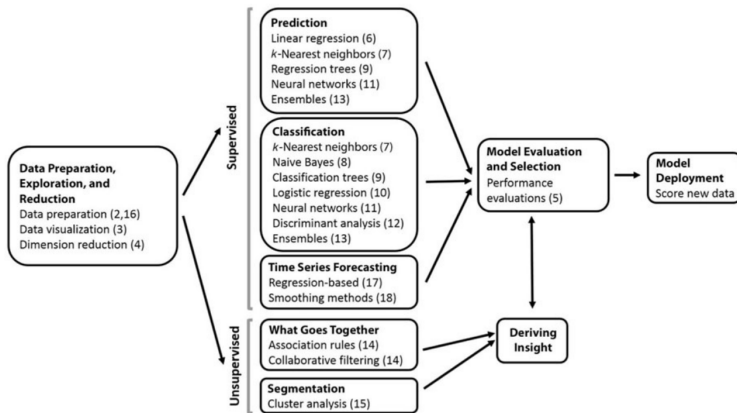




Para Wickham y Grolemund (2017) las actividades de pre-procesamiento implicarían Import, Tidy, Transform y Visualize, a excepción de model y communicate.



Para Shmueli, Bruce, Gedeck, y Patel (2020) el pre-procesamiento implica Data Preparation, Exploration, and Reduction.



## Ejercicio guiado:

- Equipo 1: Consultar en el libro de Molin (2021)
- Equipo 2: Consultar en el libro de Wickham y Grolemund (2017)
- Equipo 3: Consultar en el libro de Shmueli y cols. (2020)

Responder a las siguientes preguntas, luego de analizar concienzudamente la tabla de contenidos del libro

- ¿Cuáles y cuántos capítulos corresponden a pre-procesamiento?
- ¿Cuántas páginas tiene cada capítulo?
- ¿Cuántas sintaxis específicas aparecen en los capítulos que corresponden a pre-procesamiento?



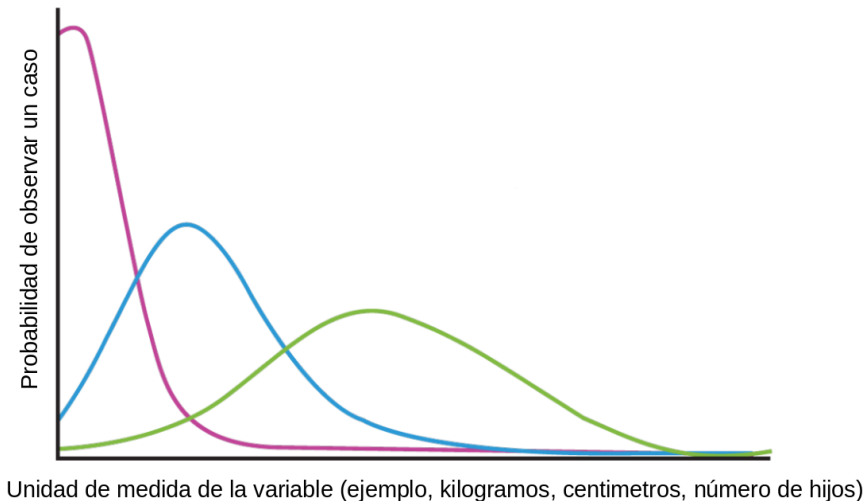
# 3

---

## Fundamentos de Análisis de Datos



Retomemos el script que habíamos estudiado en el slide 8 de Sesión 6A (cuya demostración se hizo en clase usando R). Vamos a concentrarnos en el concepto de **distribución estadística**.



Primero, debemos entender el concepto de **caso, observación, o registro**. La colección de casos es lo que se llama **distribución estadística**.



Aston Martin  
DBX 2020  
(176,900 US\$)



Ford  
Ecosport 2017  
(17,000 US\$)



Sin una distribución estadística es muy difícil afirmar ideas tales como cuál de estos dos vehículos es “*normal o típico*”



En analítica de datos, el interés siempre recae en analizar datos para extraer información a partir de ellos, y deducir conocimiento útil a partir de la información que extraigamos de los datos.

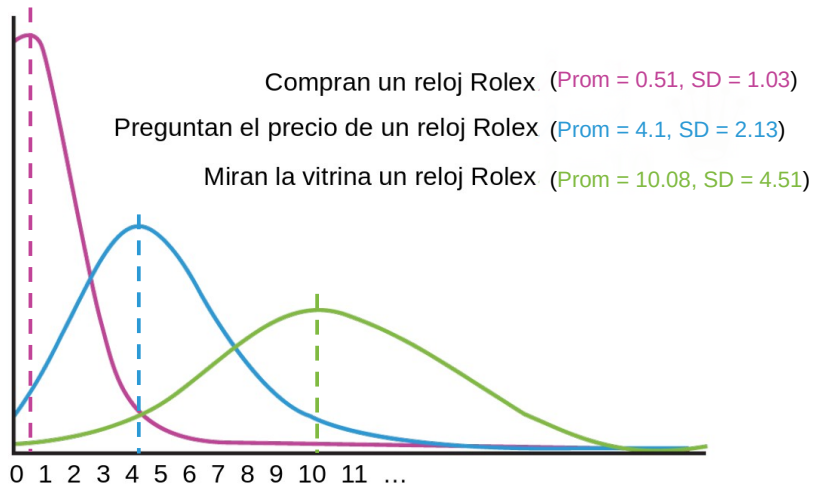
Entender cómo se describe una distribución estadística es, probablemente, lo más fundamental para un experto en analítica de datos. Para entender cómo describirla, es imprescindible apoyarse en tres características fundamentales: **tendencia**, **variación** y **forma**.



- **Tendencia:** Es la característica de una distribución estadística que se refiere al punto en el que se observan mayor frecuencia de casos u observaciones. La tendencia se mide a través de indicadores como **promedio**, **mediana** o **moda**.
- **Variación:** Se refiere al conjunto de observaciones que definen los límites inferiores y superiores dentro de los cuales se observan los casos. La variación de una distribución se mide con indicadores como la **desviación estándar**, la **varianza** o el **rango intercuartilar**
- **Forma:** Se refiere a la apariencia visual que adopta una colección de casos u observaciones luego de ordenarlos con base en algún criterio. La forma de una distribución se mide con indicadores como la **asimetría** y la **curtosis**.



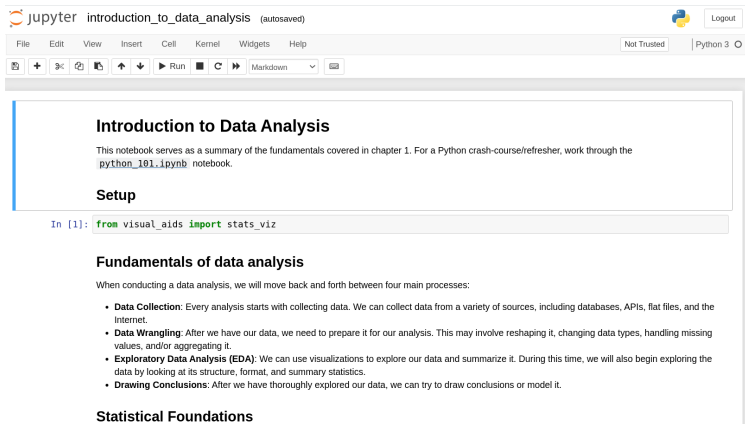




Número promedio de personas que...



Si usted abre el jupyter notebook del capítulo 1 del libro de Molin (2021) observará con detalle computacional cómo se hacen los cálculos.



The screenshot shows a Jupyter Notebook interface with the title 'jupyter introduction\_to\_data\_analysis (autosaved)'. The top bar includes a 'Logout' button and a 'Python 3' indicator. The main content area is titled 'Introduction to Data Analysis' and contains the following text:

This notebook serves as a summary of the fundamentals covered in chapter 1. For a Python crash-course/refresher, work through the [python\\_101.ipynb](#) notebook.

### Setup

```
In [1]: from visual_aids import stats_viz
```

### Fundamentals of data analysis

When conducting a data analysis, we will move back and forth between four main processes:

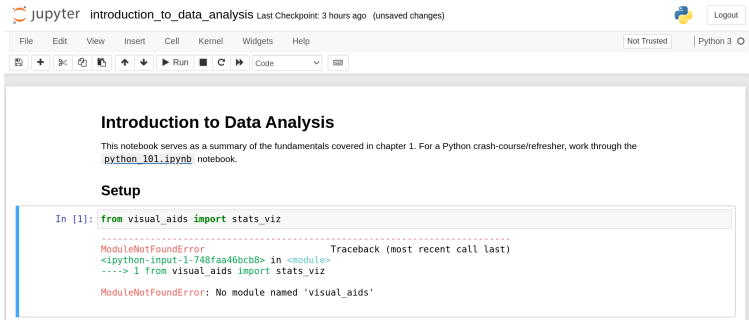
- **Data Collection:** Every analysis starts with collecting data. We can collect data from a variety of sources, including databases, APIs, flat files, and the Internet.
- **Data Wrangling:** After we have our data, we need to prepare it for our analysis. This may involve reshaping it, changing data types, handling missing values, and/or aggregating it.
- **Exploratory Data Analysis (EDA):** We can use visualizations to explore our data and summarize it. During this time, we will also begin exploring the data by looking at its structure, format, and summary statistics.
- **Drawing Conclusions:** After we have thoroughly explored our data, we can try to draw conclusions or model it.

### Statistical Foundations



## Ejercicio Guiado:

- Abrir el jupyter notebook del capítulo 1 del libro de Molin (2021).
- Intentar reproducir las sintaxis que aparecen en ese jupyter notebook.



The screenshot shows a Jupyter Notebook titled "introduction\_to\_data\_analysis" with a "Last Checkpoint: 3 hours ago (unsaved changes)" status. The interface includes a top bar with a "Logout" button and a "Not Trusted" warning. Below the top bar is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. A toolbar contains icons for file operations, navigation, and execution. The main content area displays the notebook's title "Introduction to Data Analysis" and a paragraph explaining its purpose. Below this, a "Setup" section contains a code cell with the following content:

```
In [1]: from visual_aids import stats_viz
```

The code cell has executed, resulting in a `ModuleNotFoundError`. The error message is displayed in red text:

```
ModuleNotFoundError: No module named 'visual_aids'
```

Below the error message, a traceback is shown, indicating the error occurred in the current cell during the execution of the import statement.

¿Le aparece este problema?



Observe lo que Molin (2021) comenta en la página 5 del capítulo 1.

Since the code that's used to generate the content in these notebooks is not the main focus of this chapter, the majority of it has been separated into the `visual_aids` package, which is used to create visuals for explaining concepts throughout the book, and the `check_environment.py` file. If you choose to inspect these files, don't be overwhelmed; everything that's relevant to data science will be covered in this book.

Para chequear las librerías o paquetes de software que se requieren para ejecutar las sintaxis en el jupyter notebook llamado `introduction_to_data_analysis`, basta con abrir la terminal de su computador, cambiar de directorio (Carpeta GitHub) y escribir el siguiente código o sintaxis

```
python3 check_environment.py
```



## Debería aparecerle algo así como lo siguiente

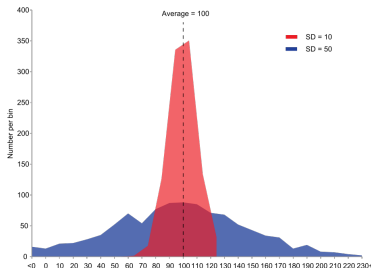
```
jc@pop-os: ~/Documents/GitHub/Hands-On-Data-Analysis-with-Pandas-2nd-edition/ch_01$ python3 check_environment.py
Using Python in /usr:
[OK] Python is version 3.9.5 (default, May 11 2021, 08:20:37)
[OK] GCC 10.3.0

[FAIL] graphviz not installed.
[FAIL] imblearn not installed.
[FAIL] ipympl not installed.
[FAIL] jupyterlab not installed.
[FAIL] matplotlib not installed.
[FAIL] numpy not installed.
[FAIL] pandas not installed.
[FAIL] requests version 2.24.0 is required, but 2.25.1 installed.
[FAIL] sklearn not installed.
[FAIL] scipy not installed.
[FAIL] seaborn not installed.
[FAIL] sqlalchemy not installed.
[FAIL] statsmodels not installed.
[FAIL] wheel not installed.
[FAIL] login_attempt_simulator not installed.
[FAIL] ml_utils not installed.
[FAIL] stock_analysis not installed.
[FAIL] visual_aids not installed.
jc@pop-os: ~/Documents/GitHub/Hands-On-Data-Analysis-with-Pandas-2nd-edition/ch_01$
```

(No se preocupe si observa alguna diferencia en la apariencia de este pantallazo con relación a lo que usted observa en su computador).



Por ahora, nos interesa comprender cómo, a partir del concepto de distribución estadística, se entiende el concepto de **análisis de varianza**.



La **varianza** indica cuánto cambian o varían los casos de la distribución. Visualmente puede entenderse cómo cuán ancha o estrecha es la base de una distribución. La varianza se calcula como el cuadrado de la desviación estándar (SD). En la imagen, la distribución roja tiene menos varianza que la azul, aunque ambas tienen exactamente el mismo promedio.



El análisis de la varianza (o ANOVA: “*Analysis of variance*” como se le dice en inglés) es una técnica estadística que sirve para comprender las diferencias del promedio entre tres o más **grupos metodológicamente comparables**.

Dos o más grupos son **metodológicamente comparables** si pertenecen conceptual o empíricamente al mismo universo o población.

Por ejemplo, los estudiantes de la universidad de Los Andes y del CESA son metodológicamente comparables porque ambos grupos pertenecen a la población de instituciones de educación superior en Colombia. Los estudiantes del CESA y los de la Universidad de Hiroshima no son metodológicamente comparables.



Para interpretar los resultados de un ANOVA, hay que entender si las distribuciones de los grupos son o no diferentes desde el punto de vista probabilístico.

Conceptualmente, las diferencias probabilísticas no son iguales a las diferencias matemáticas.

En matemáticas, el número 3 es estrictamente hablando diferente del número 2.999. En probabilidades, sin embargo, podría asumirse tranquilamente que ambos números podrían ser iguales desde el punto de vista probabilístico.

¿Entonces, en qué podemos apoyarnos para decir si dos números son o no iguales desde el punto de vista probabilístico?







¿Un promedio de 3 es igual a 2.99?



Haga un ANOVA



Desde el punto de vista de cálculo, un ANOVA genera como resultado un conjunto de información referente a las diferencias probabilísticas de tres o más grupos.

```
describe.by(Morning$TimeDif, group = Morning$Typical.Traffic.Afternoon, mat = TRUE, digits = 2)
```

group1	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Green	1	1735	21.19	18.65	22	21.47	17.79	-53	89	142	-0.11	0.46	0.45
Orange	1	4246	23.80	18.58	25	24.19	17.79	-42	89	131	-0.16	0.39	0.29
Red	1	496	25.89	17.16	27	26.21	14.83	-26	88	114	-0.05	0.97	0.77

```
> summary(afternoonDTF)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Afternoon\$Typical.Traffic.Afternoon	2	4889	2444.3	7.709	0.000454 ***
Residuals	5580	1769217	317.1		

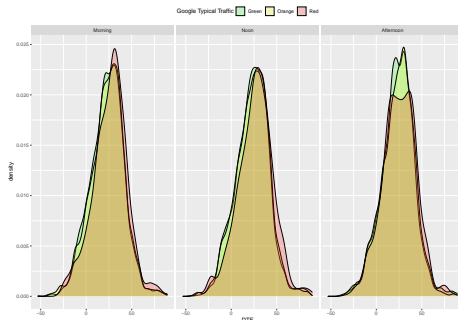
---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

424 observations deleted due to missingness



Visualmente, las diferencias entre tres o más grupos son estadísticamente significativas si sus distribuciones se separan entre sí.



Este gráfico se obtuvo luego de llegar a la línea 102 de nuestro repo de GitHub, correspondiente al artículo de domicilios de comida en Bogotá (Correa y cols., 2019).



- Correa, J. C., Garzón, W., Brooker, P., Sakarkar, G., Carranza, S. A., Yunado, L., y Rincón, A. (2019). Evaluation of collaborative consumption of food delivery services through web mining techniques. *Journal of Retailing and Consumer Services*, 46, 45–50.
- Kotha, R., Kim, P., y Alexy, O. (2014). Turn your science into a business. *Harvard Business Review*, 106–114.
- Molin, S. (2021). *Hands-On Data Analysis with Pandas* (2nd ed.). Birmingham: UK: Pakt Publishing.
- Shmueli, G., Bruce, P. C., Gedeck, P., y Patel, N. R. (2020). *Data mining for business analytics: Concepts, techniques, and applications in python*. New Jersey, USA: Wiley & Sons.
- Wickham, H., y Grolemund, G. (2017). *R for Data Science: Import, tidy, transform, visualize, and model data*. Sebastopol, CA, USA: O'Reilly.



