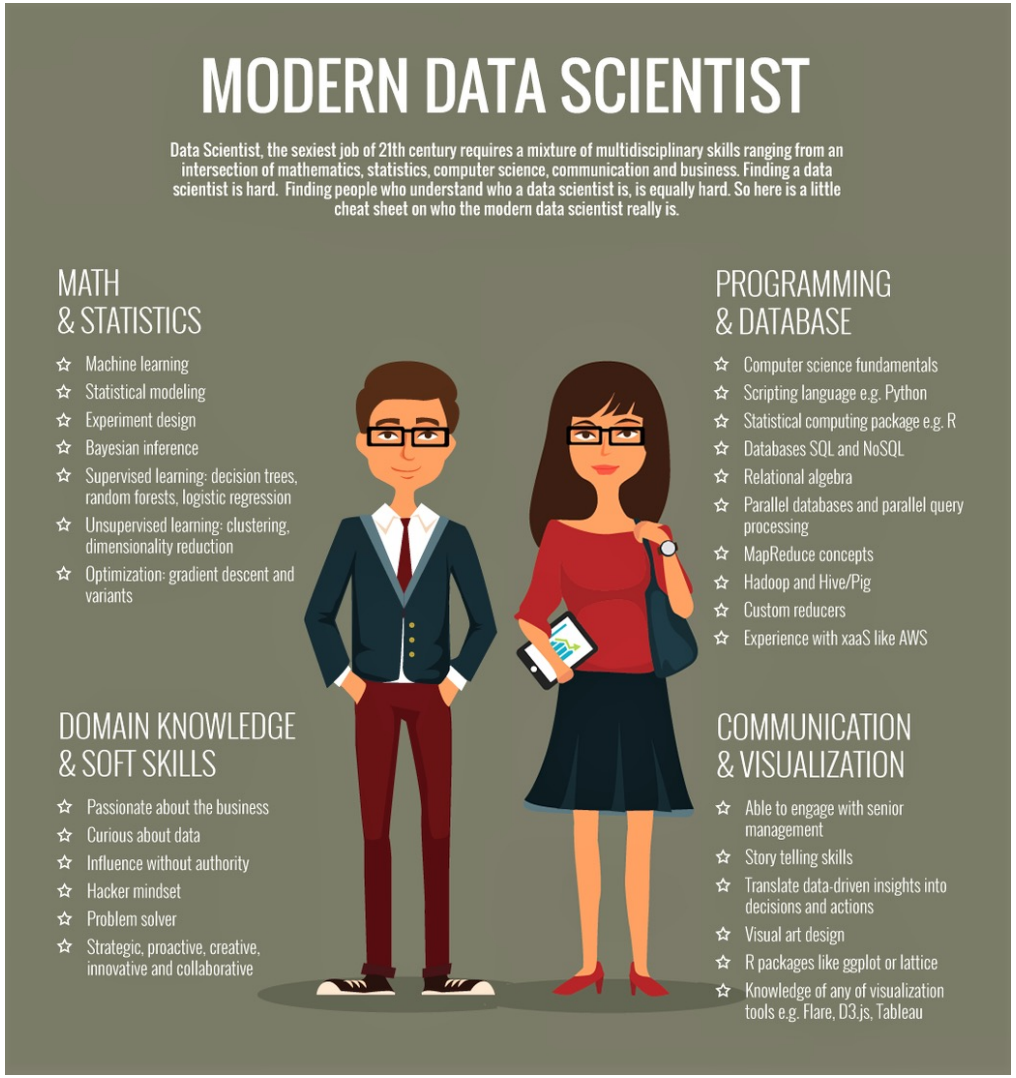
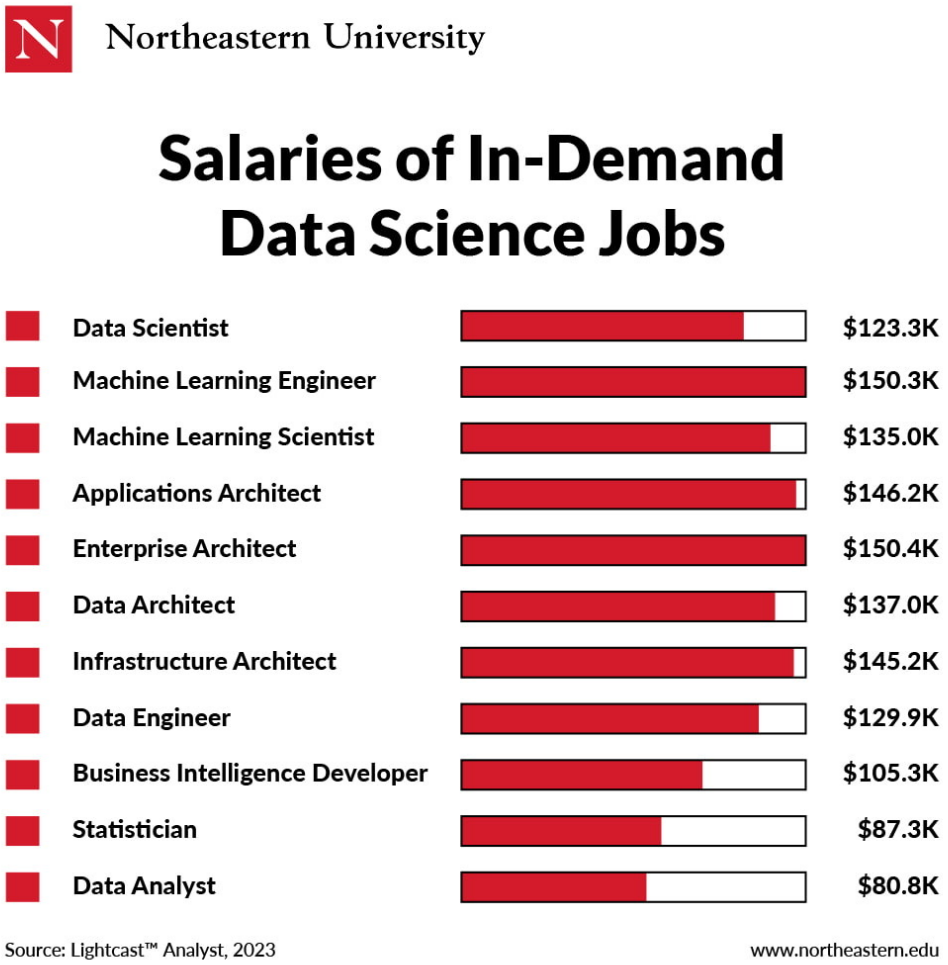
DATA JOBS

Carlos Fernando Vásquez Guerra

DATA JOBS

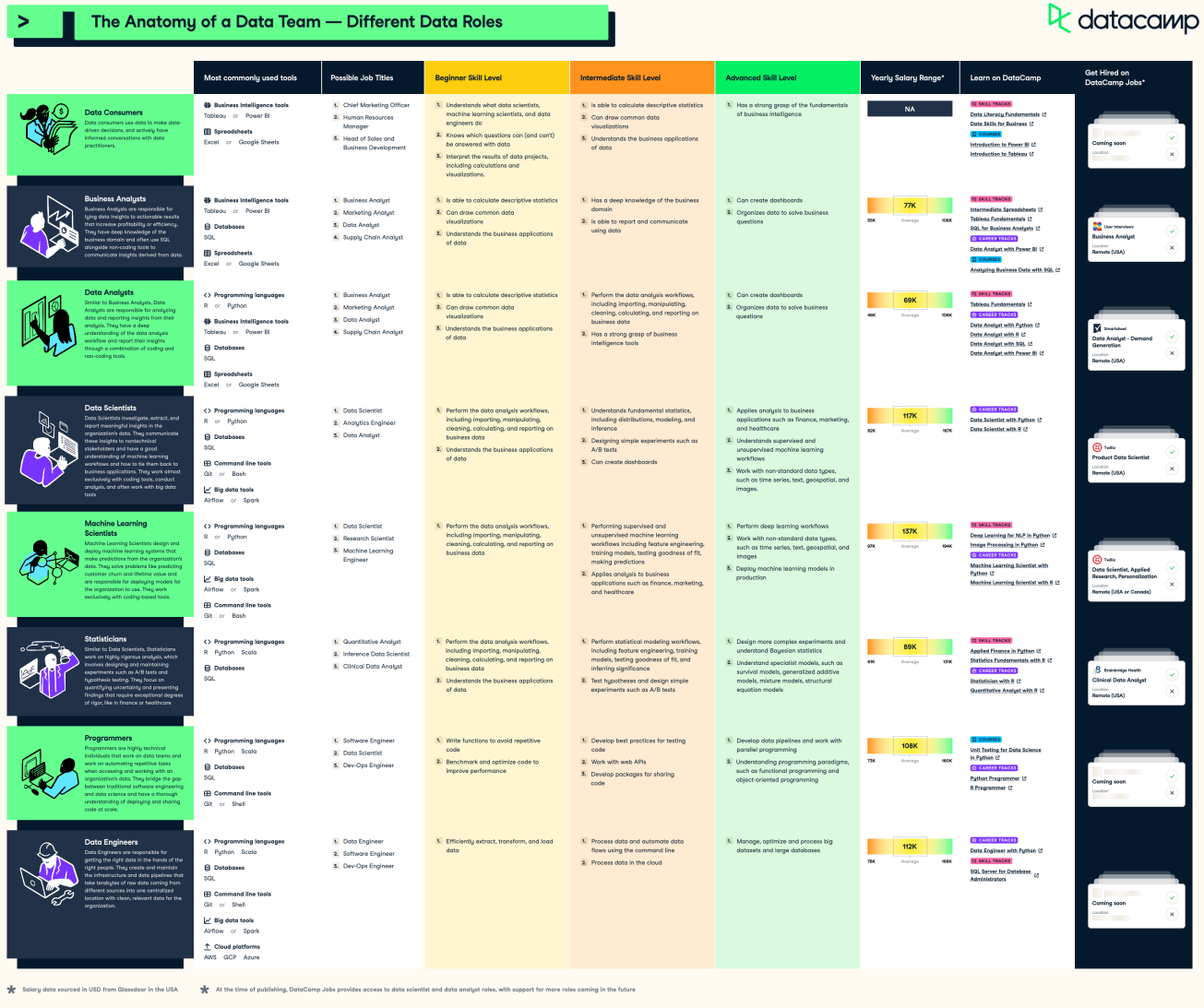


https://medium.com/@_moazzemhossain/the-10-algorithms-data-scientist-must-have-to-know-97a2c478ce94



https://graduate.northeastern.edu/knowledge-hub/data-science-careers-shaping-our-future/

DATA JOBS



- Data Analyst vs. Data Scientist: A Comparative Guide For 2025
- What's the Difference Between a Data Analyst and a Data Scientist?
- The Top 10 Data Analytics Careers For 2025: Skills, Salaries & Career Prospects
- Data Analyst vs Data Scientist vs Data Engineer: Roles, Skills & Salary Comparison
- 12 Data Science Job Titles — Which Role Is Right for You?

https://www.datacamp.com/blog/infographic-the-anatomy-of-a-data-team-different-data-roles

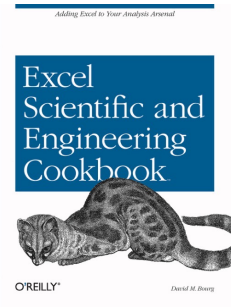
EXCEL

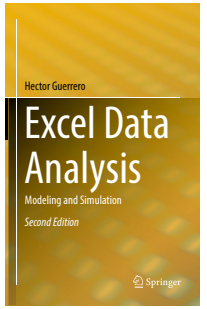
Carlos Fernando Vásquez Guerra

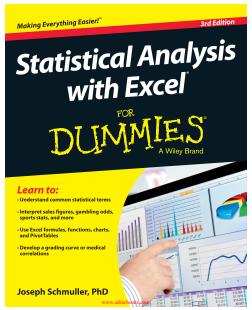
CURSOS Y MÁS

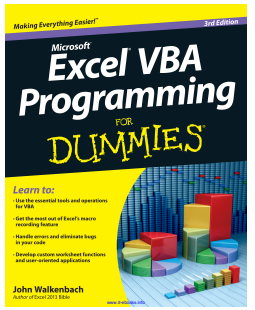
Cursos

Libros









VBA

Curso Excel Programación en Macros VBA desde cero Especial

VBA Excel: How to Get Started and Make Your Work Easier

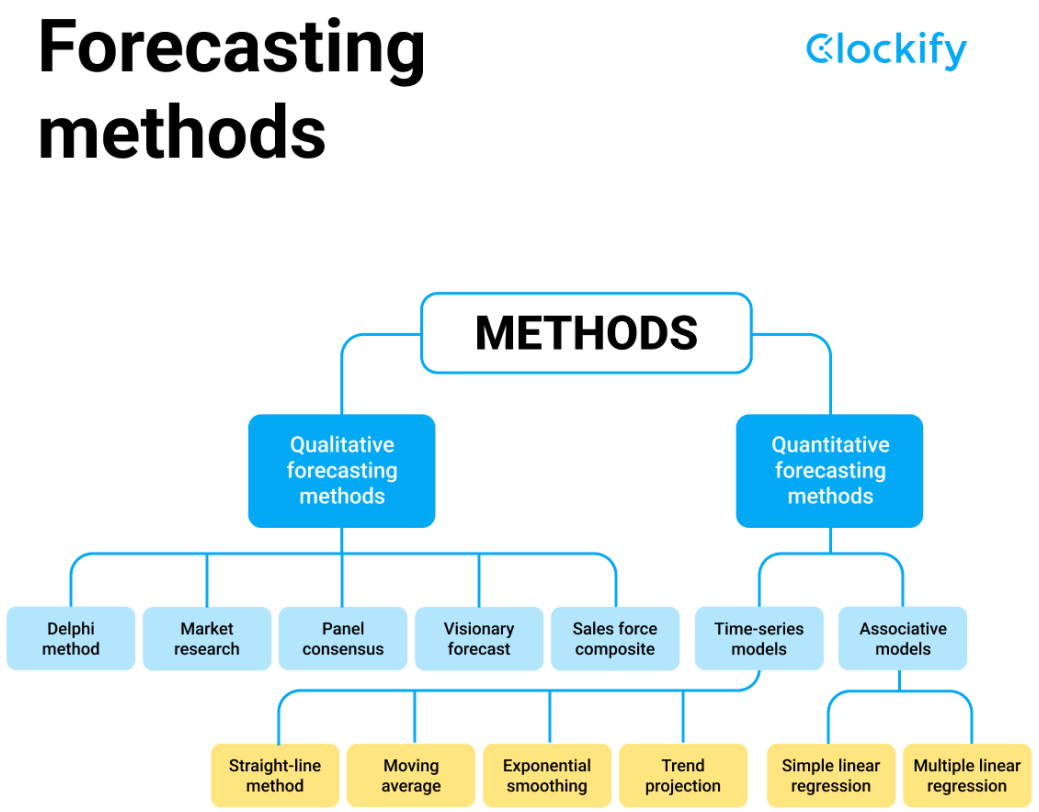
Tips

The 15 Basic Excel Formulas Everyone Needs to Know

FORECASTING

Carlos Fernando Vásquez Guerra

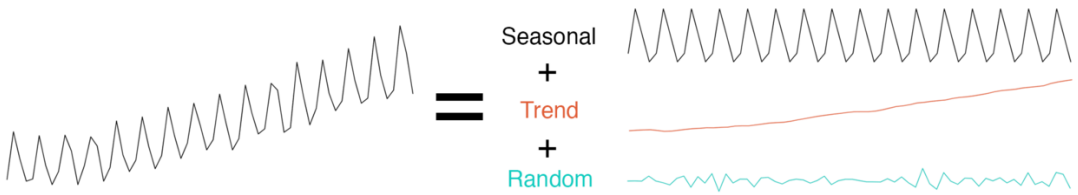
FORECASTING



https://clockify.me/forecasting-models

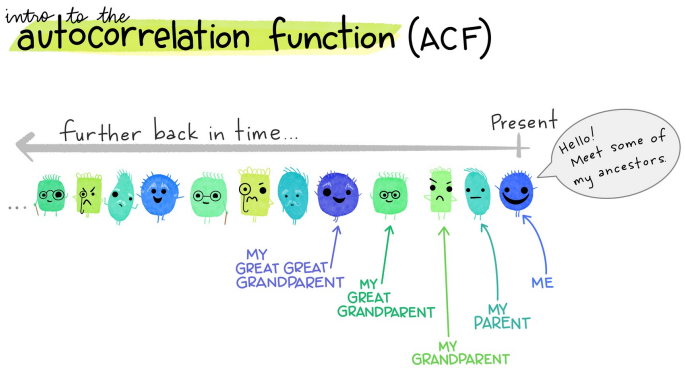
TIME SERIES

Decomposition



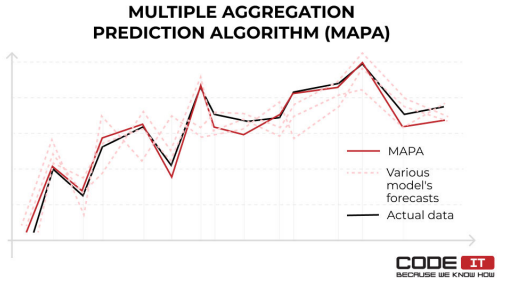
https://www.linkedin.com/pulse/time-series-decomposition-using-r-harshal-chaudhari/

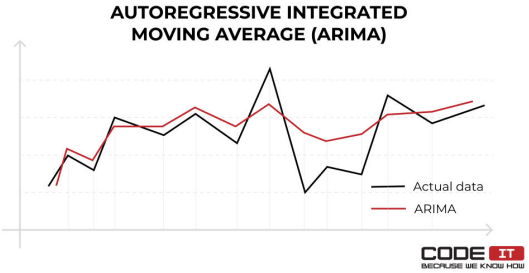
ACF & PACF

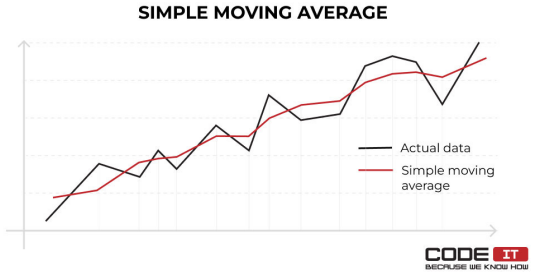


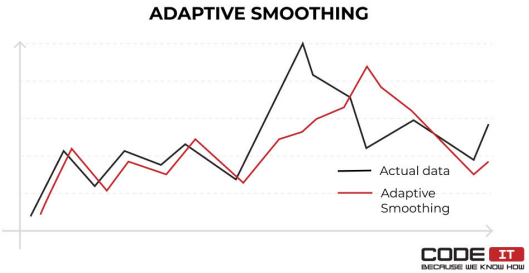
https://allisonhorst.com/time-series-acf

Models







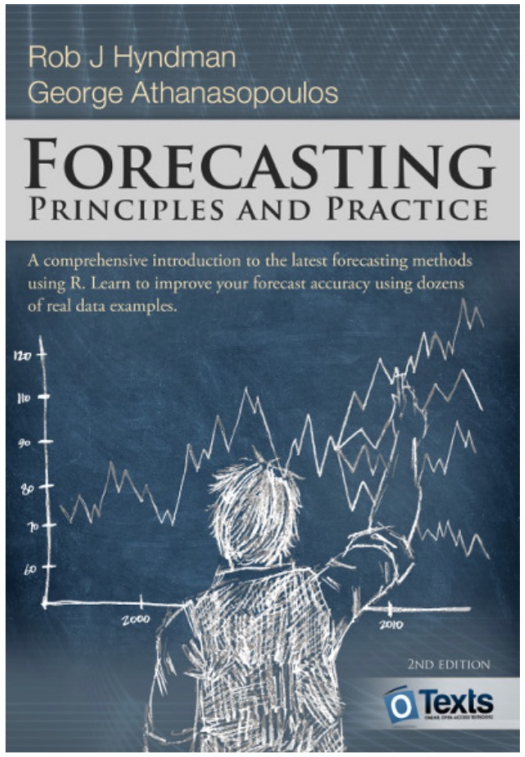


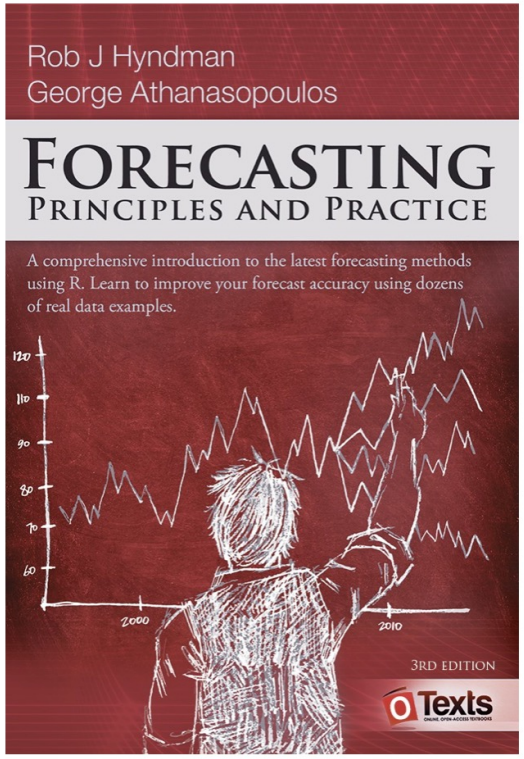
https://codeit.us/blog/forecasting-in-supply-chain

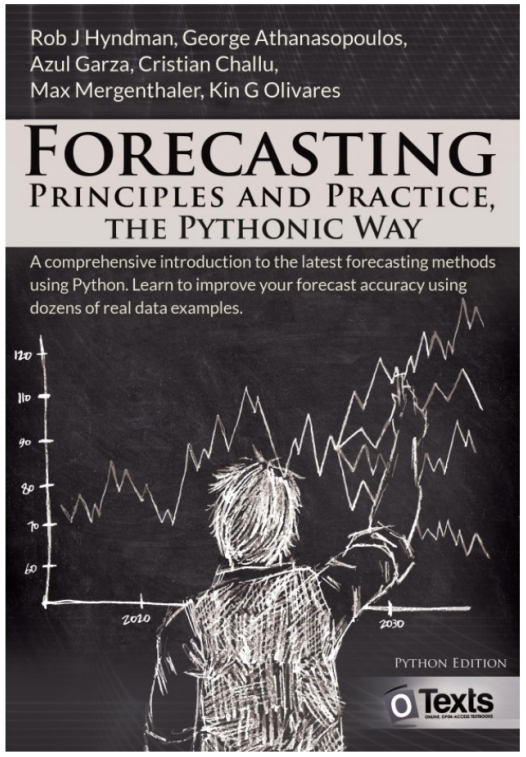
https://domino.ai/blog/time-series-with-r

TIME SERIES

Advanced methods and theory







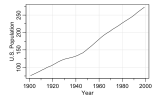
https://otexts.com

TIME SERIES

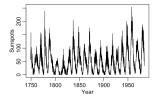
Time Series Cheat Sheet

Plot Time Series

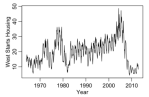
1. tsplot(x=time, y=data)



2. plot(ts(data, start=start_time, frequency=gap))



3. ts.plot(ts(data, start=start_time, frequency=gap))



Simulation

Autoregression of Order p

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + W_t$$

Moving Average of Order q

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-p}$$

ARMA (p, q)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-p}$$

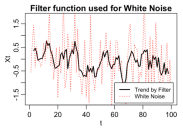
Simulation of ARMA (p, q)

arima.sim(model=list(ar=c($\phi_1, \ldots, \phi_p$), ma=c($\theta_1, \ldots, \theta_q$)), n=n)

Filters

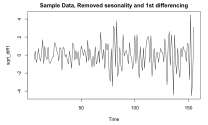
Linear Filter: filter()

filter(data, filter=filter_coefficients, sides=2, method="convolution", circular=F)



Differencing Filter: diff()

diff(data, lag=4, differences=1)

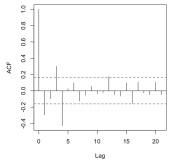


Auto-correlation

Use ACF and PACF to detect model

(Complete) Auto-correlation function: acf()

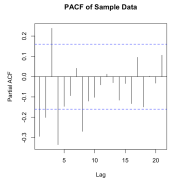
acf(data, type='correlation', na.action=na.pass)



Partial Auto-correlation function: pacf()

pacf(data, na.action=na.pass)

OR: acf(data, type='partial', na.action=na.pass)



Parameter Estimation

Fit an ARMA time series model to the data

ar(): To estimate parameters of an AR model

ar(x=data, aic=T, order.max = NULL, c("yule-walker", "burg", "ols", "mle", "yw"))

```
Call:
ar(x = sqrt1, aic = TRUE, order.max = NULL, method = c("yule-walker",     "burg", "ols", "mle", "yw"))

Coefficients:
      1       2       3       4       5       6       7       8       9
-0.3066  -0.1503   0.0793  -0.5065  -0.1673  -0.1149  -0.0588  -0.3031  -0.1207

Order selected 9  sigma^2 estimated as  1.52
```

arima(): To estimate parameters of an AM or ARMA model, and build model

arima(data, order=c(p, 0, q),method=c('ML'))

```
Call:
arima(x = sqrt1, order = c(2, 0, 6), method = c("ML"))

Coefficients:
          ar1      ar2     ma1     ma2     ma3     ma4     ma5     ma6  intercept
      -0.1658  -0.7951  -0.1536  0.7524  -0.2101  -0.7680  0.0605  -0.6812     0.0048
s.e.   0.0918   0.0754   0.0916  0.1047   0.0794   0.0743  0.0812   0.0895     0.0062

sigma^2 estimated as 1.193:  log likelihood = -230.78,  aic = 481.55
```

AICc(): Compare models using AICC

AICc(fittedModel)

Forecasting

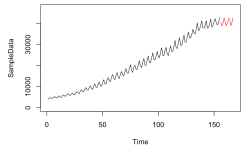
Forecasting future observations given a fitted ARMA model

predict(): Predict future observations given a fitted ARMA model

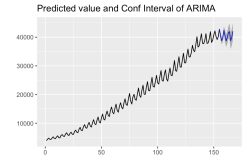
predict(arima_model, number_to_predict)

Plot Predicted values and Confidence Interval:

```
fit<-predict(arima_model, number_to_predict)
ts.plot(data,
        xlim=c(1, length(data)+number_to_predict),
        ylim=c(0, max(fit$pred+1.96*fit$se)))
lines(length(data)+1:length(data)+
      number_to_predict, fit$pred)
```



OR: autoplot(forecast(arima_model, level=c(95), h=number_to_predict))

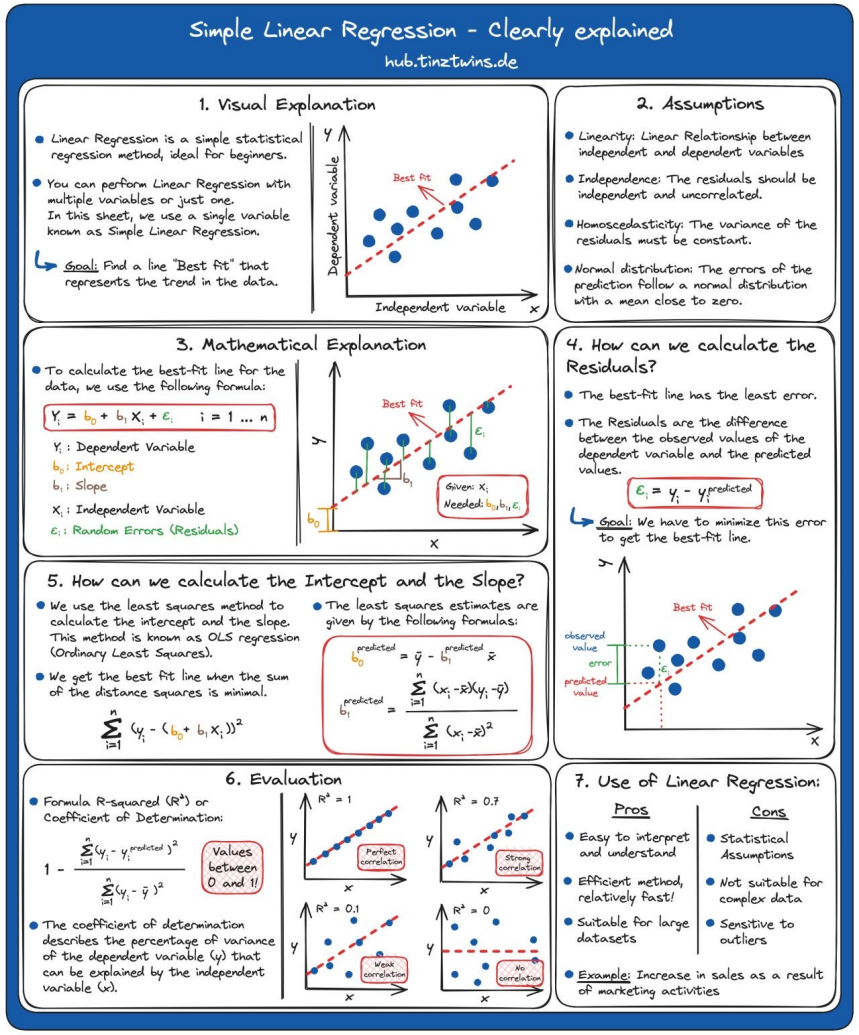


RPubs

https://rpubs.com/CarlosFVG/SeriesTCode2024

https://rpubs.com/CarlosFVG/DatesWithRCFVG

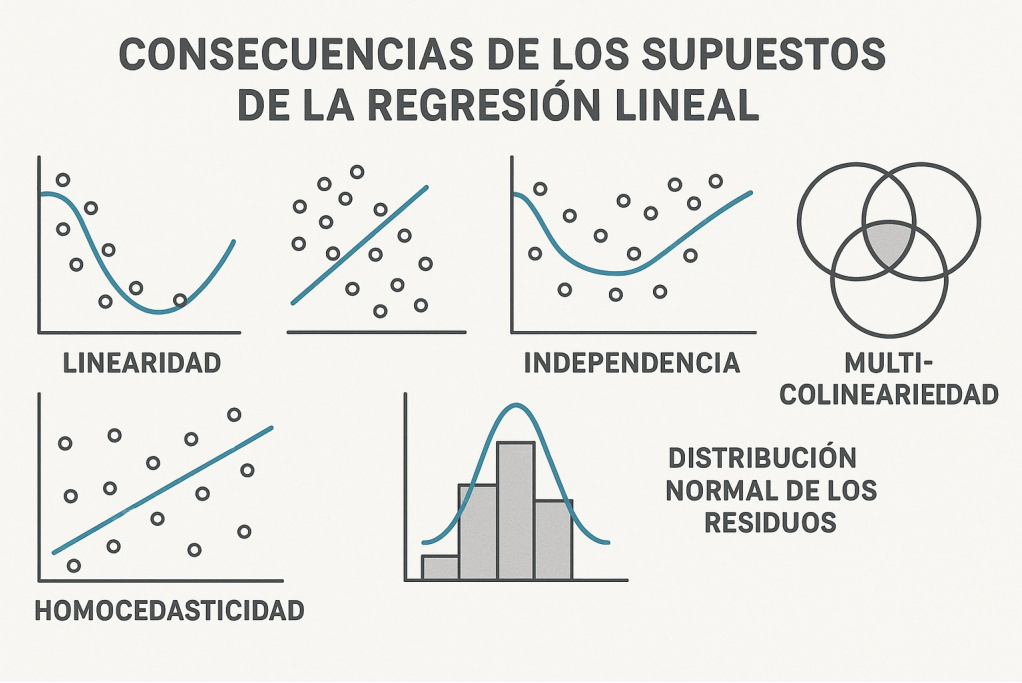
LINEAR REGRESSION

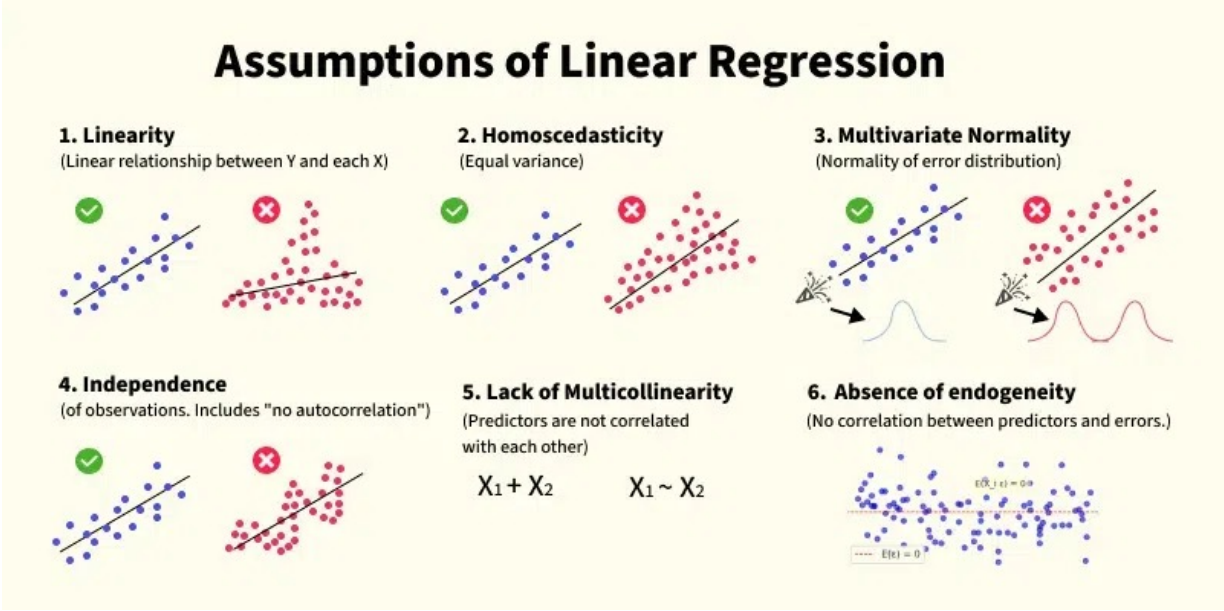


https://christophm.github.io/interpretable-ml-book/limo.html

https://faculty.nps.edu/rbassett/_book/introduction-to-linear-regression.html

LINEAR REGRESSION: WARNINGS





https://www.geeksforgeeks.org/assumptions-of-linear-regression/

LINEAR REGRESSION: WARNINGS

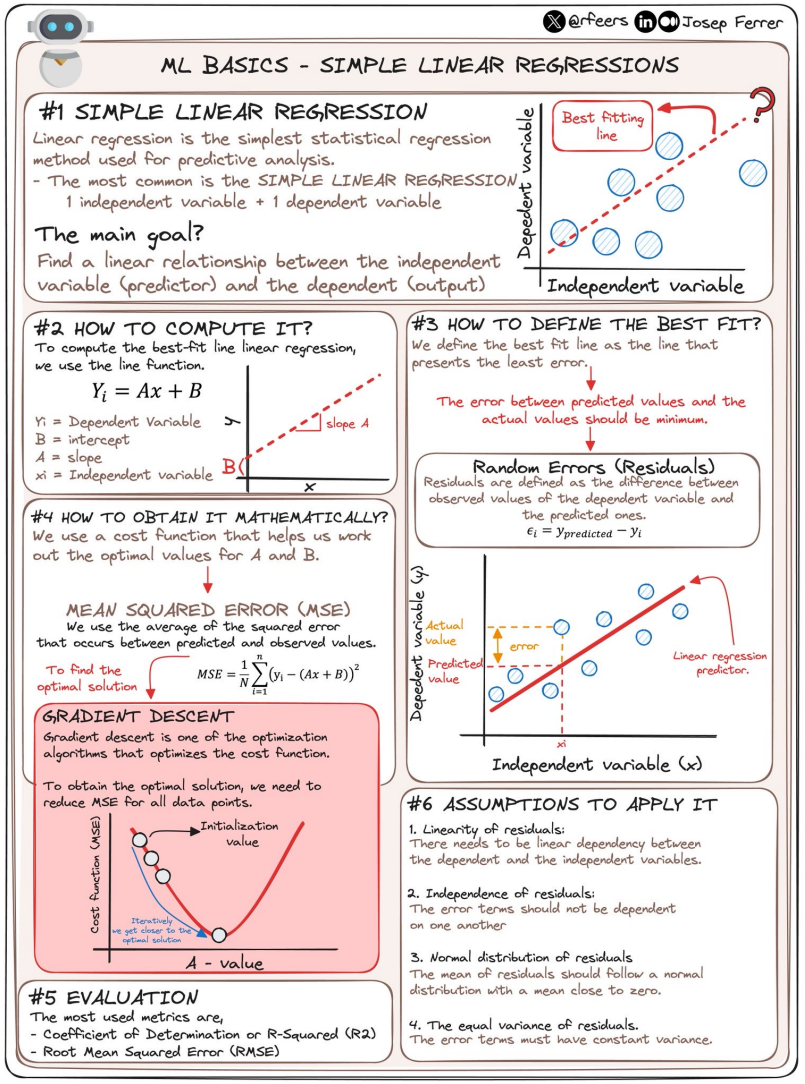
Impacto del Incumplimiento de Supuestos en Regresión ineal

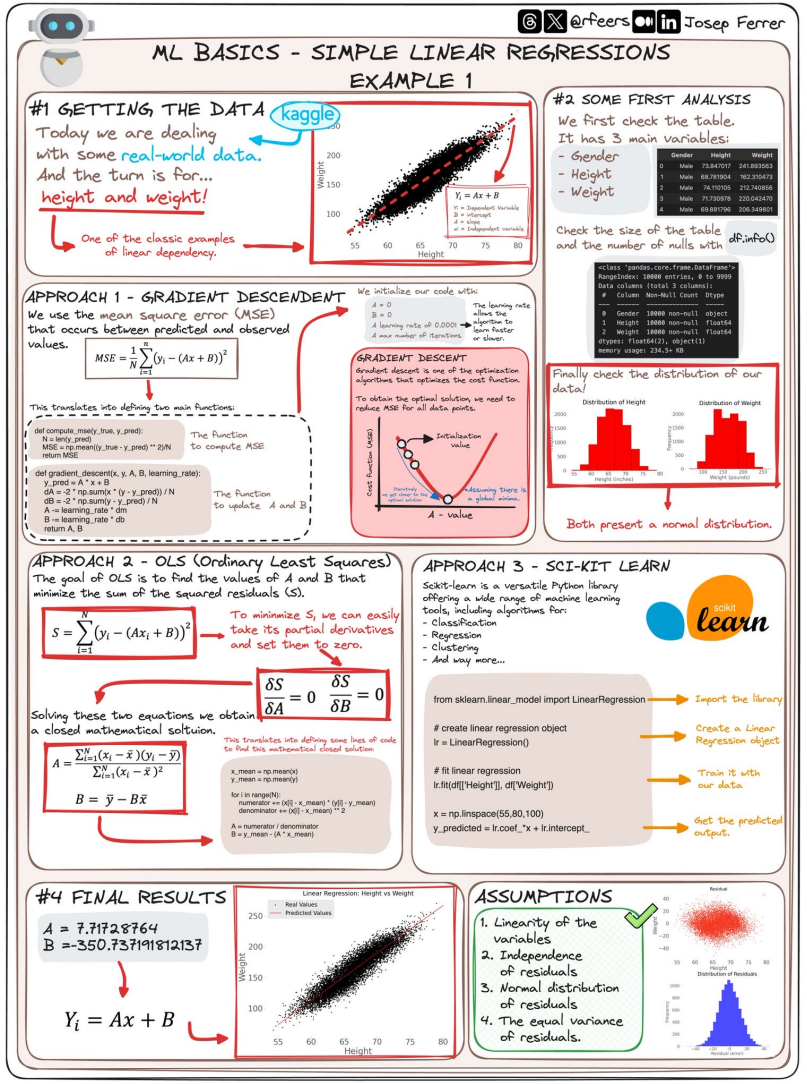
Supuesto	Consecuencias del incumplimiento	Gravedad	¿Afecta predicciones?	¿Afecta inferencias?	Acciones de contingencia	Pruebas diagnósticas / gráficas sugeridas
Linealidad	El modelo no captura la relación real. Residuos grandes, baja precisión.	**Alta**	Sí	Sí	- Transformar variables - Probar modelos no lineales (e.g., regresión polinómica, árboles)	- Gráfico de residuos vs. valores ajustados - Componentes parciales (partial residual plots)
Independencia de los errores	Subestima errores estándar, sesga intervalos de confianza y p-valores (e.g., en series de tiempo).	**Alta**	Algo	Mucho	- Modelos para datos dependientes (e.g., ARIMA) - Revisar autocorrelación (Durbin-Watson)	- Prueba de Durbin-Watson - ACF/PACF de residuos
No hay observaciones influyentes/extremas	Un outlier puede dominar los resultados.	**Alta**	Mucho	Mucho	- Diagnóstico (Cook's D, leverage) - Modelos robustos (RLM) - Tratar o justificar outliers	- Leverage vs. residuos estudentizados-Distancia de Cook - DFBETAS
Medición precisa de las variables (sin error en X)	Sesgo en coeficientes, atenuación.	**Media**	Sí	Sí	- Validar mediciones - Usar modelos con error en variables (SEM, variables instrumentales)	- Comparación con medidas repetidas - Estudios de validación externa - Análisis de sensibilidad
Homoscedasticidad (varianza constante)	Menos precisión, pruebas t y F poco confiables.	**Media**	Algo	Sí	- Usar errores robustos (White) - Transformaciones (log, sqrt) - Regresión ponderada	- Gráfico de residuos vs. valores ajustados - Prueba de Breusch-Pagan o White
Ausencia de multicolinealidad	Inestabilidad en coeficientes, errores estándar inflados.	**Media**	Poco	Sí	- Eliminar variables correlacionadas- - PCA - Ridge/Lasso	- VIF (Variance Inflation Factor) - Matriz de correlación de predictores
Normalidad de errores (para inferencia)	Afecta p-valores e IC si n es pequeño. Con n grande, el CLT lo compensa.	**Baja**	Poco	Algo	- Aumentar n - Bootstrap para IC y pruebas - Pruebas no paramétricas	- Histograma de residuos - QQ plot - Prueba de Shapiro-Wilk o Kolmogorov-Smirnov

Fuentes y referencias

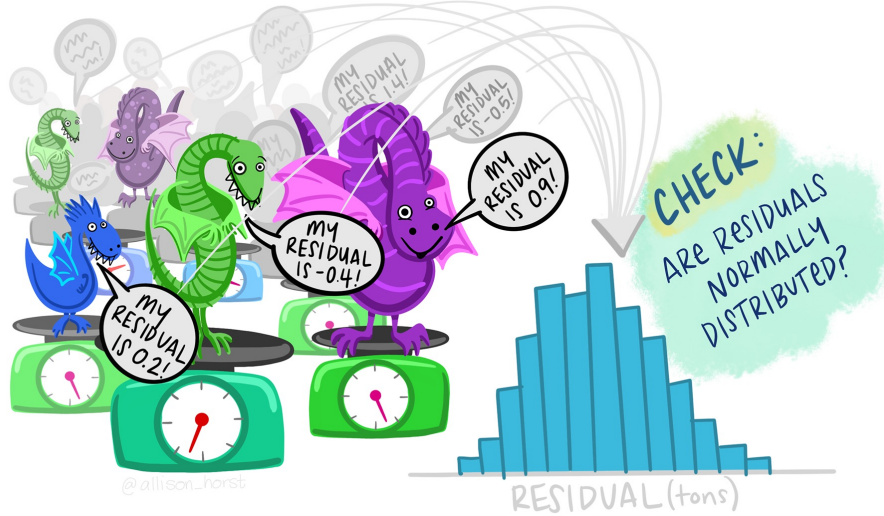
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. Wiley.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied Linear Statistical Models. McGraw-Hill.
- Fox, J. (2016). Applied Regression Analysis and Generalized Linear Models. Sage Publications.
- Wooldridge, J. M. (2013). Introductory Econometrics: A Modern Approach. Cengage Learning.

LINEAR REGRESSION: GOOD RESOURCES



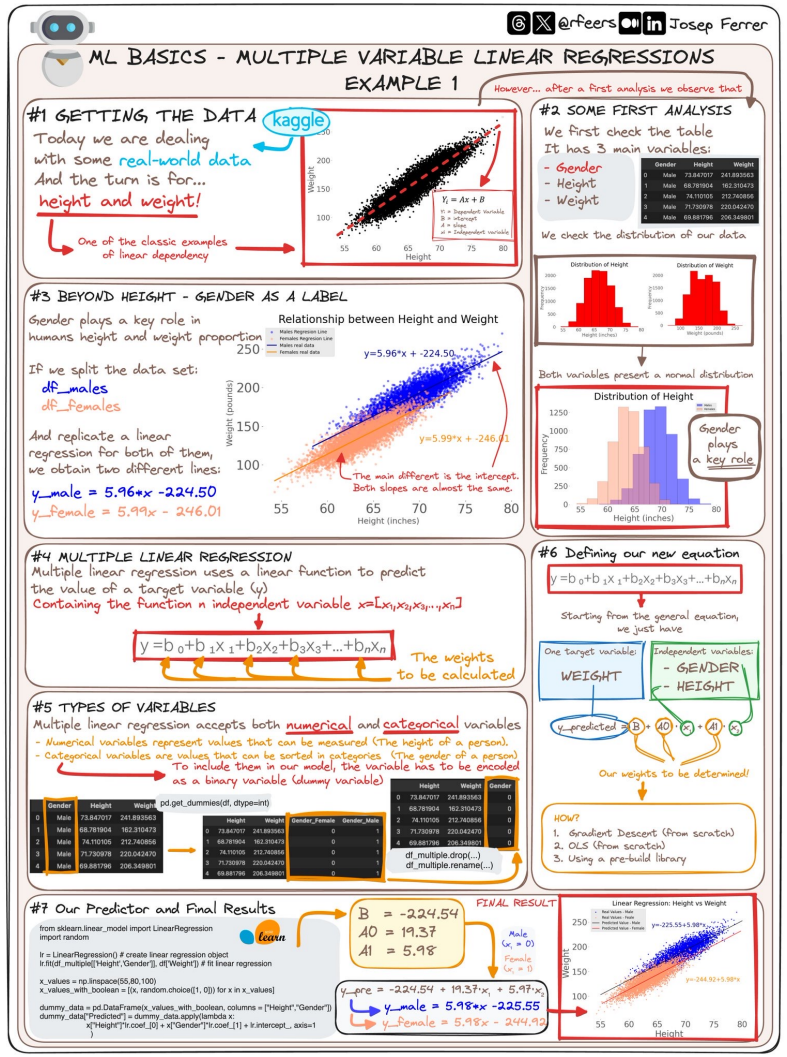


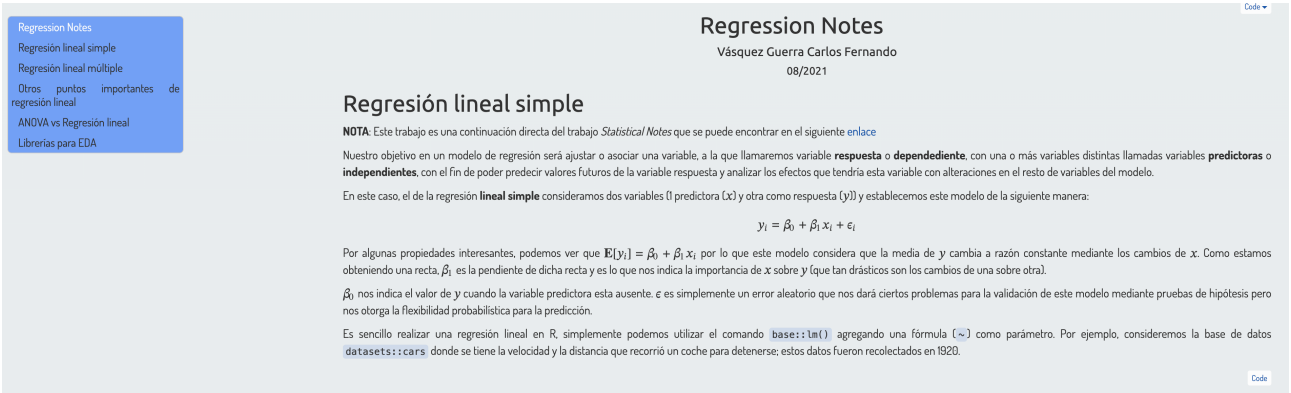
LINEAR REGRESSION: GOOD RESOURCES



https://allisonhorst.com/linear-regression-dragons

+ LINEAR REGRESSION





Regression Notes

Regresión lineal simple

Regresión lineal múltiple

Otros puntos importantes de regresión lineal

ANOVA vs Regresión lineal

Librerías para EDA

Regression Notes

Vásquez Guerra Carlos Fernando

08/2021

Regresión lineal simple

NOTA: Este trabajo es una continuación directa del trabajo *Statistical Notes* que se puede encontrar en el siguiente enlace

Nuestro objetivo en un modelo de regresión será ajustar o asociar una variable, a la que llamaremos variable **respuesta** o **dependediente**, con una o más variables distintas llamadas variables **predictoras** o **independientes**, con el fin de poder predecir valores futuros de la variable respuesta y analizar los efectos que tendría esta variable con alteraciones en el resto de variables del modelo.

En este caso, el de la regresión **lineal simple** consideramos dos variables (1 predictora (x) y otra como respuesta (y)) y establecemos este modelo de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Por algunas propiedades interesantes, podemos ver que $\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i$ por lo que este modelo considera que la media de y cambia a razón constante mediante los cambios de x. Como estamos obteniendo una recta, β_1 es la pendiente de dicha recta y es lo que nos indica la importancia de x sobre y (que tan drásticos son los cambios de una sobre otra).

β_0 nos indica el valor de y cuando la variable predictora esta ausente. ϵ es simplemente un error aleatorio que nos dará ciertos problemas para la validación de este modelo mediante pruebas de hipótesis pero nos otorga la flexibilidad probabilística para la predicción.

Es sencillo realizar una regresión lineal en R, simplemente podemos utilizar el comando `base::lm()` agregando una fórmula (`~`) como parámetro. Por ejemplo, consideremos la base de datos `datasets::cars` donde se tiene la velocidad y la distancia que recorrió un coche para detenerse; estos datos fueron recolectados en 1920.

https://rpubs.com/CarlosFVG/RegressionNotes

https://statisticsbyjim.com/regression/choosing-regression-analysis/

https://bookdown.org/roback/bookdown-BeyondMLR/

https://uc-r.github.io/model_selection

https://joelcarlson.github.io/2016/05/10/Exploring-Interactions/

https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html

BEYOND LINEAR REGRESSION

Cheat Sheet – Regression Analysis

What is Regression Analysis?
Fitting a function f(.) to datapoints y_i=f(x_i) under some error function. Based on the estimated function and error, we have the following types of regression

1. Linear Regression:
Fits a line minimizing the sum of mean-squared error for each datapoint.

$$min_\beta \sum_i \|y_i - f_\beta^{linear}(x_i)\|^2$$
$$f_\beta^{linear}(x_i) = \beta_0 + \beta_1 x_i$$

2. Polynomial Regression:
Fits a polynomial of order k (k+1 unknowns) minimizing the sum of mean-squared error for each datapoint.

$$min_\beta \sum_{i=0}^{m} \|y_i - f_\beta^{poly}(x_i)\|^2$$
$$f_\beta^{poly}(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k$$

3. Bayesian Regression:
For each datapoint, fits a gaussian distribution by minimizing the mean-squared error. As the number of data points x_i increases, it converges to point estimates i.e. $n \to \infty, \sigma^2 \to 0$

$$min_\beta \sum_i \|y_i - \mathcal{N}(f_\beta(x_i), \sigma^2)\|^2$$
$$f_\beta(x_i) = f_\beta^{poly}(x_i) \text{ or } f_\beta^{linear}(x_i)$$
$$\mathcal{N}(\mu, \sigma^2) \to \text{Gaussian with mean } \mu \text{ and variance } \sigma^2$$

4. Ridge Regression:
Can fit either a line, or polynomial minimizing the sum of mean-squared error for each datapoint and the weighted L2 norm of the function parameters beta.

$$min_\beta \sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2 + \sum_{j=0}^{k} \beta_j^2$$
$$f_\beta(x_i) = f_\beta^{poly}(x_i) \text{ or } f_\beta^{linear}(x_i)$$

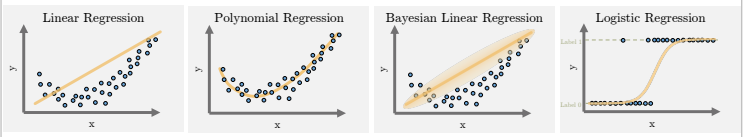
5. LASSO Regression:
Can fit either a line, or polynomial minimizing the the sum of mean-squared error for each datapoint and the weighted L1 norm of the function parameters beta.

$$min_\beta \sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2 + \sum_{j=0}^{k} |\beta_j|$$
$$f_\beta(x_i) = f_\beta^{poly}(x_i) \text{ or } f_\beta^{linear}(x_i)$$

6. Logistic Regression:
Can fit either a line, or polynomial with sigmoid activation minimizing the binary cross-entropy loss for each datapoint. The labels y are binary class labels.

$$min_\beta \sum_i -y_i log(\sigma(f_\beta(x_i))) - (1 - y_i) log(1 - \sigma(f_\beta(x_i)))$$
$$f_\beta(x_i) = f_\beta^{poly}(x_i) \text{ or } f_\beta^{linear}(x_i)$$
$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

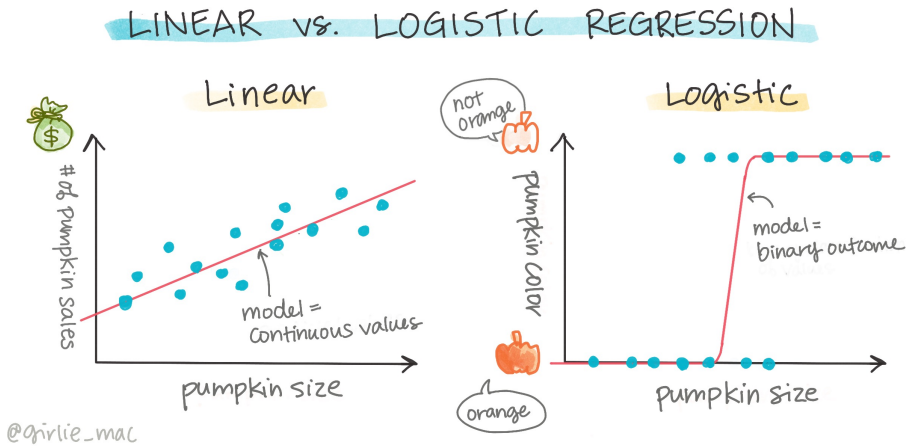
Visual Representation:



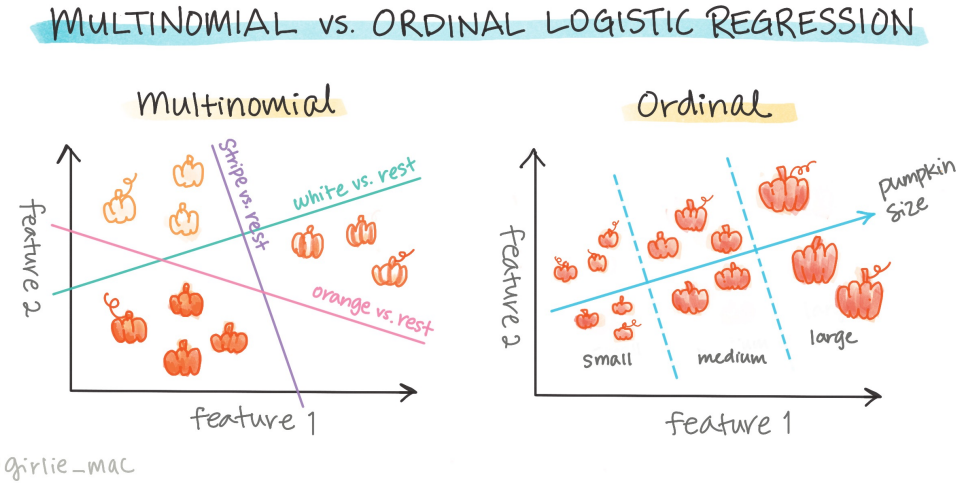
Summary:

	What does it fit?	Estimated function	Error Function
Linear	A line in n dimensions	$f_\beta^{linear}(x_i) = \beta_0 + \beta_1 x_i$	$\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2$
Polynomial	A polynomial of order k	$f_\beta^{poly}(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots$	$\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2$
Bayesian Linear	Gaussian distribution for each point	$\mathcal{N}(f_\beta(x_i), \sigma^2)$	$\sum_i \|y_i - \mathcal{N}(f_\beta(x_i), \sigma^2)\|^2$
Ridge	Linear/polynomial	$f_\beta^{poly}(x_i)$ or $f_\beta^{linear}(x_i)$	$\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2 + \sum_{j=0}^{n} \beta_j^2$
LASSO	Linear/polynomial	$f_\beta^{poly}(x_i)$ or $f_\beta^{linear}(x_i)$	$\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2 + \sum_{j=0}^{n} \|\beta_j\|$
Logistic	Linear/polynomial with sigmoid	$\sigma(f_\beta(x_i))$	$min_\beta \sum_i -y_i log(\sigma(f_\beta(x_i))) - (1 - y_i) log(1 - \sigma(f_\beta(x_i)))$

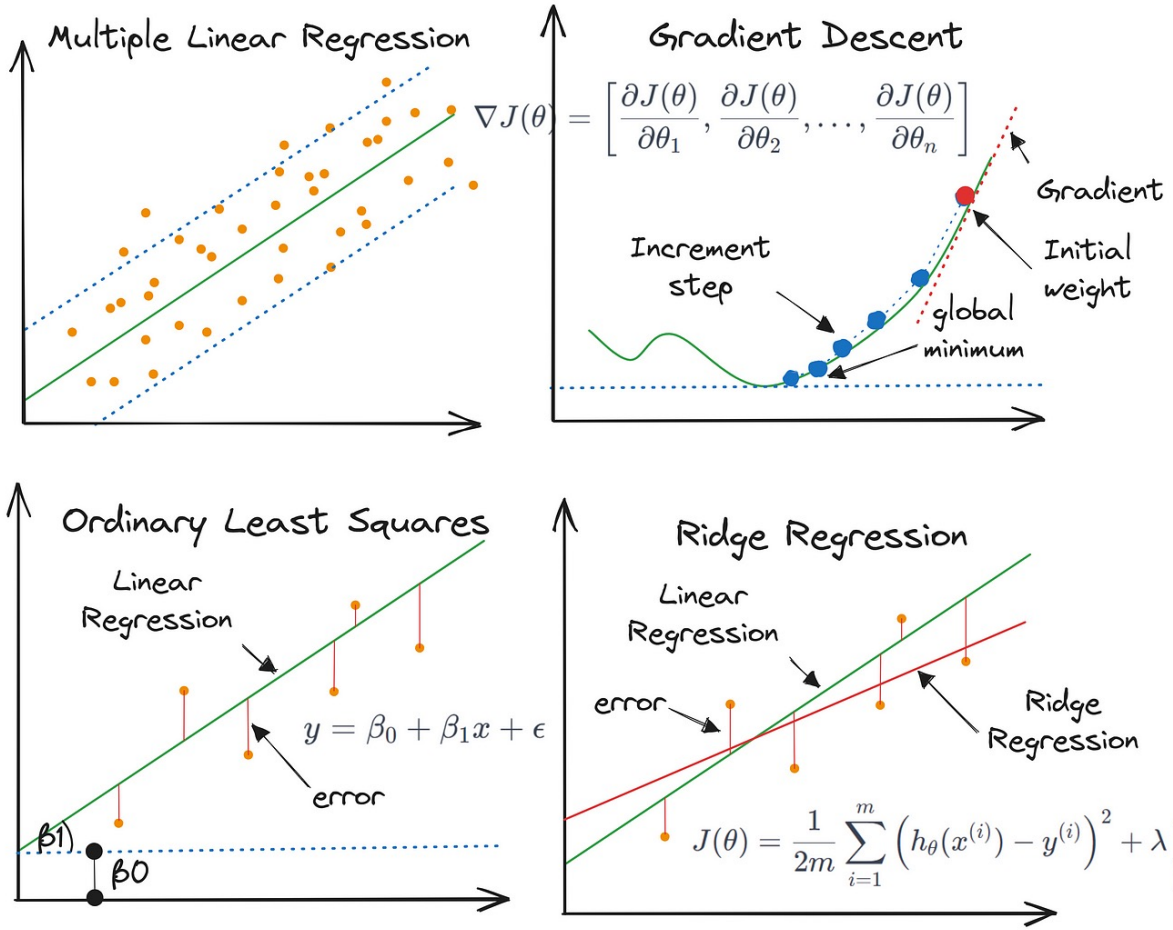
Source: https://www.cheatsheets.aqeel-anwar.com Tutorial: Click here



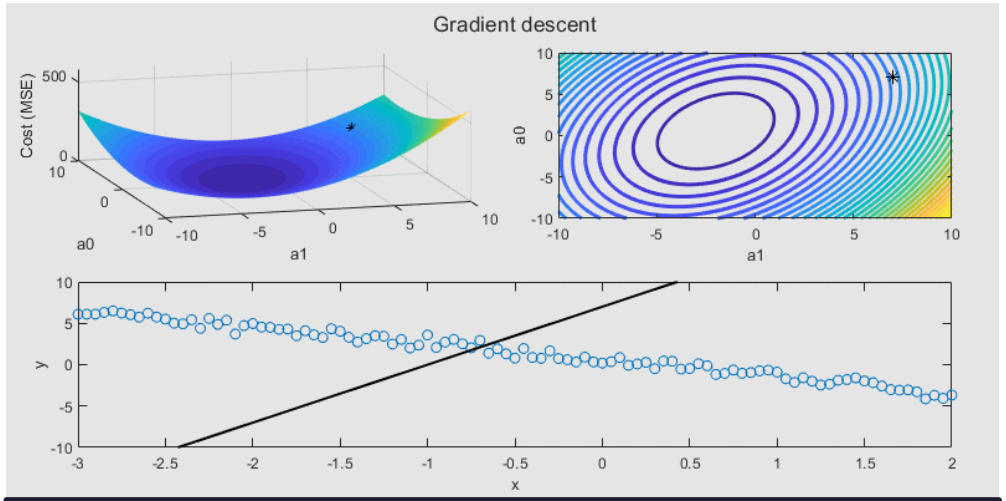
https://github.com/microsoft/ML-For-Beginners/blob/main/2-Regression/4-Logistic/solution/R/lesson_4-R.ipynb



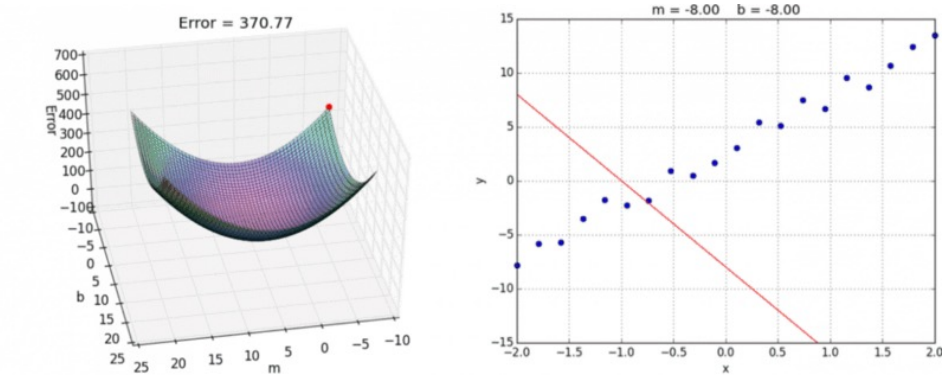
BEYOND LINEAR REGRESSION



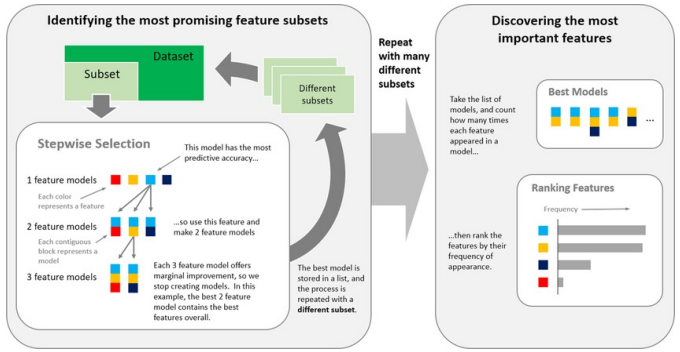
BEYOND LINEAR REGRESSION



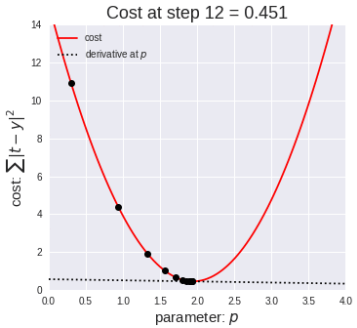
https://www.linkedin.com/pulse/linear-regressionmostly-asked-questions-manralaitop30-manral-/



https://sar-gupta.github.io/posts/2017/10/20/gradient-descent



https://www.researchgate.net/figure/Feature-selection-process-for-linear-regression-which-was-performed-separately-for-each_fig1_342116063



https://medium.com/analytics-vidhya/simple-linear-regression-cost-function-gradient-descent-50c5ed085770

BEYOND LINEAR REGRESSION

Cheat Sheet – Regression Analysis

What is Regression Analysis?
Fitting a function f(.) to datapoints y_i=f(x_i) under some error function. Based on the estimated function and error, we have the following types of regression

1. Linear Regression:
Fits a line minimizing the sum of mean-squared error for each datapoint.

$$min_\beta \sum_i \|y_i - f_\beta^{linear}(x_i)\|^2$$
$$f_\beta^{linear}(x_i) = \beta_0 + \beta_1 x_i$$

2. Polynomial Regression:
Fits a polynomial of order k (k+1 unknowns) minimizing the sum of mean-squared error for each datapoint.

$$min_\beta \sum_{i=0}^{m} \|y_i - f_\beta^{poly}(x_i)\|^2$$
$$f_\beta^{poly}(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k$$

3. Bayesian Regression:
For each datapoint, fits a gaussian distribution by minimizing the mean-squared error. As the number of data points x_i increases, it converges to point estimates i.e. $n \to \infty, \sigma^2 \to 0$

$$min_\beta \sum_i \|y_i - \mathcal{N}(f_\beta(x_i), \sigma^2)\|^2$$
$$f_\beta(x_i) = f_\beta^{poly}(x_i) \text{ or } f_\beta^{linear}(x_i)$$
$$\mathcal{N}(\mu, \sigma^2) \to \text{Gaussian with mean } \mu \text{ and variance } \sigma^2$$

4. Ridge Regression:
Can fit either a line, or polynomial minimizing the sum of mean-squared error for each datapoint and the weighted L2 norm of the function parameters beta.

$$min_\beta \sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2 + \sum_{j=0}^{k} \beta_j^2$$
$$f_\beta(x_i) = f_\beta^{poly}(x_i) \text{ or } f_\beta^{linear}(x_i)$$

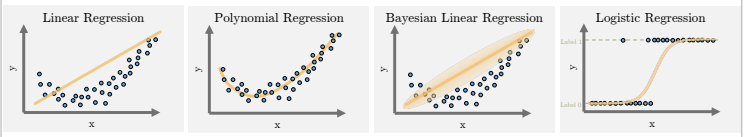
5. LASSO Regression:
Can fit either a line, or polynomial minimizing the the sum of mean-squared error for each datapoint and the weighted L1 norm of the function parameters beta.

$$min_\beta \sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2 + \sum_{j=0}^{k} |\beta_j|$$
$$f_\beta(x_i) = f_\beta^{poly}(x_i) \text{ or } f_\beta^{linear}(x_i)$$

6. Logistic Regression:
Can fit either a line, or polynomial with sigmoid activation minimizing the binary cross-entropy loss for each datapoint. The labels y are binary class labels.

$$min_\beta \sum_i -y_i log(\sigma(f_\beta(x_i))) - (1 - y_i) log(1 - \sigma(f_\beta(x_i)))$$
$$f_\beta(x_i) = f_\beta^{poly}(x_i) \text{ or } f_\beta^{linear}(x_i)$$
$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

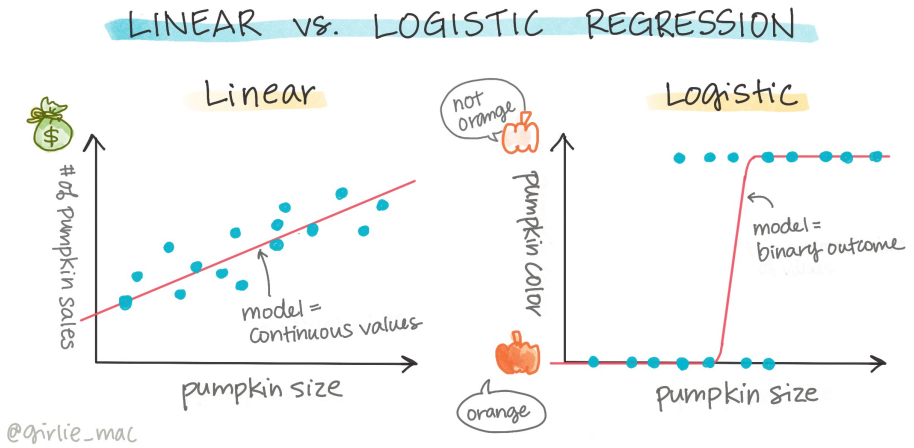
Visual Representation:



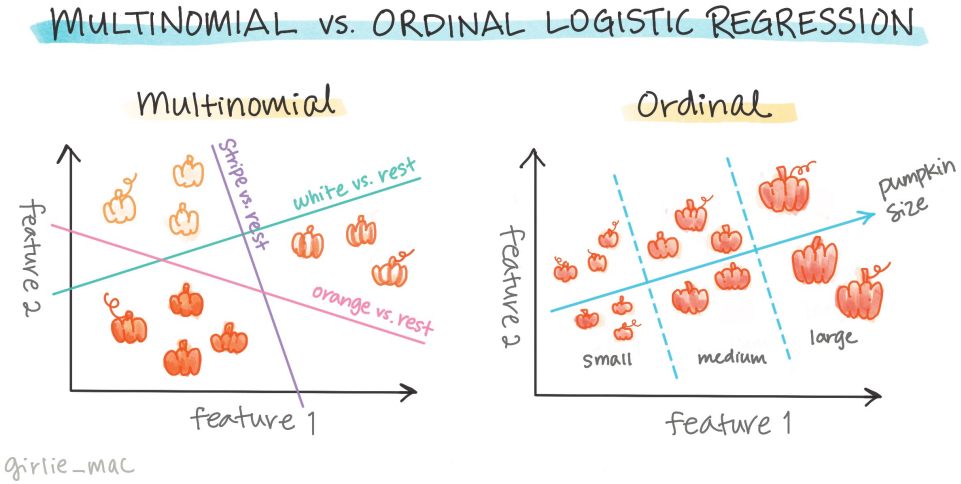
Summary:

	What does it fit?	Estimated function	Error Function
Linear	A line in n dimensions	$f_\beta^{linear}(x_i) = \beta_0 + \beta_1 x_i$	$\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2$
Polynomial	A polynomial of order k	$f_\beta^{poly}(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots$	$\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2$
Bayesian Linear	Gaussian distribution for each point	$\mathcal{N}(f_\beta(x_i), \sigma^2)$	$\sum_i \|y_i - \mathcal{N}(f_\beta(x_i), \sigma^2)\|^2$
Ridge	Linear/polynomial	$f_\beta^{poly}(x_i)$ or $f_\beta^{linear}(x_i)$	$\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2 + \sum_{j=0}^{n} \beta_j^2$
LASSO	Linear/polynomial	$f_\beta^{poly}(x_i)$ or $f_\beta^{linear}(x_i)$	$\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2 + \sum_{j=0}^{n} \|\beta_j\|$
Logistic	Linear/polynomial with sigmoid	$\sigma(f_\beta(x_i))$	$min_\beta \sum_i -y_i log(\sigma(f_\beta(x_i))) - (1 - y_i) log(1 - \sigma(f_\beta(x_i)))$

Source: https://www.cheatsheets.aqeel-anwar.com Tutorial: Click here



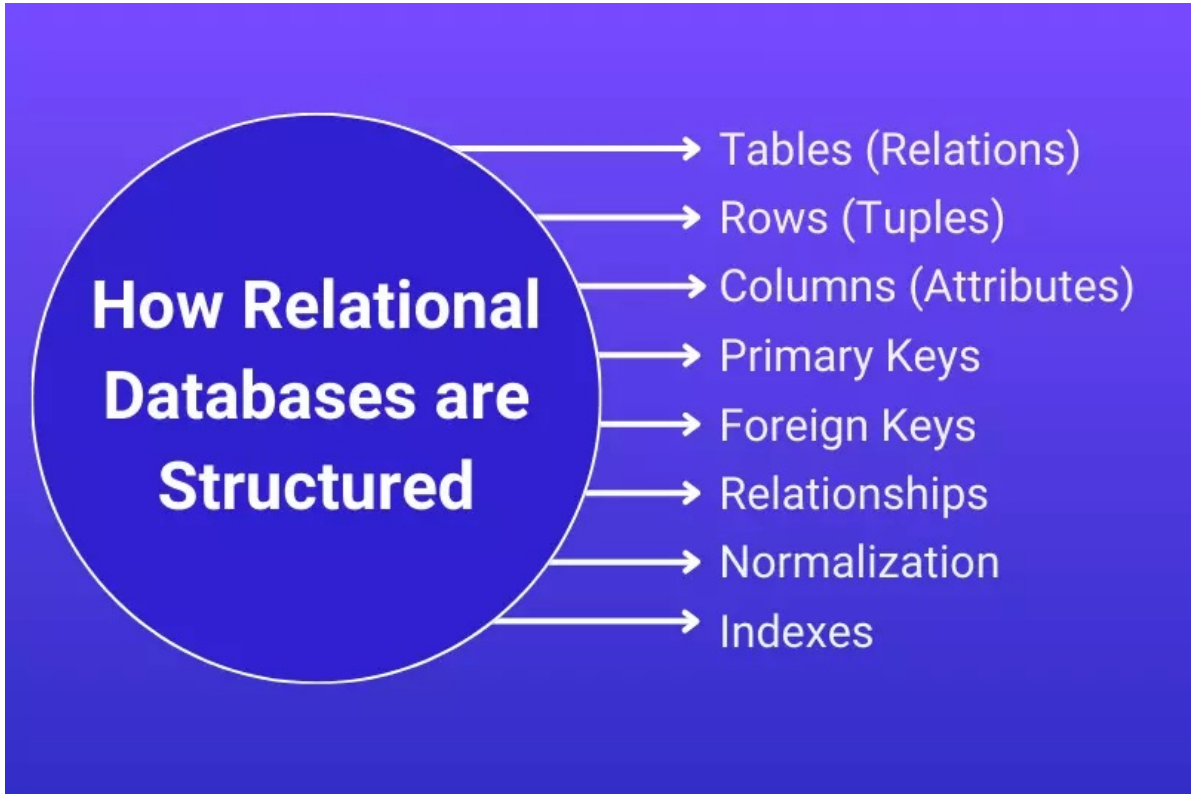
https://github.com/microsoft/ML-For-Beginners/blob/main/2-Regression/4-Logistic/solution/R/lesson_4-R.ipynb



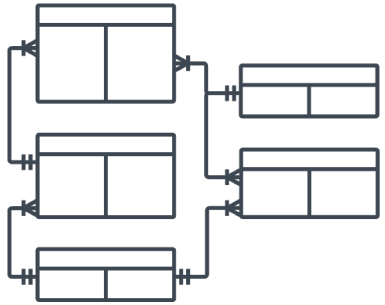
MODELADO DE DATOS

Carlos Fernando Vásquez Guerra

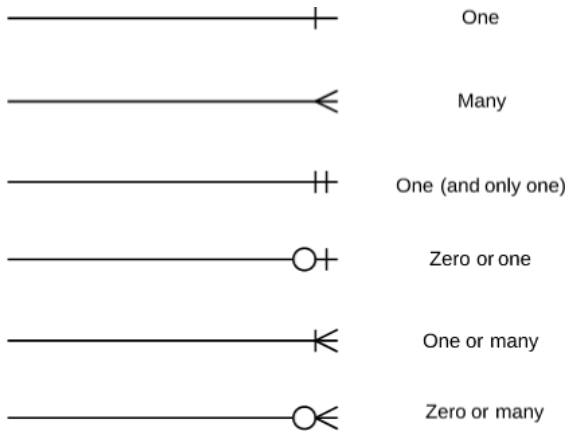
CONCEPTS



https://www.fynd.academy/blog/relational-data-model-in-rdbms



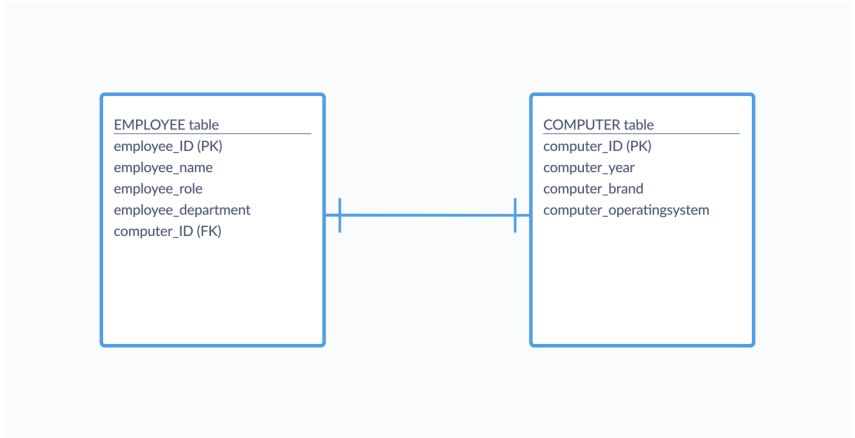
ER diagram notation



https://www.lucidchart.com/pages/ER-diagram-symbols-and-meaning

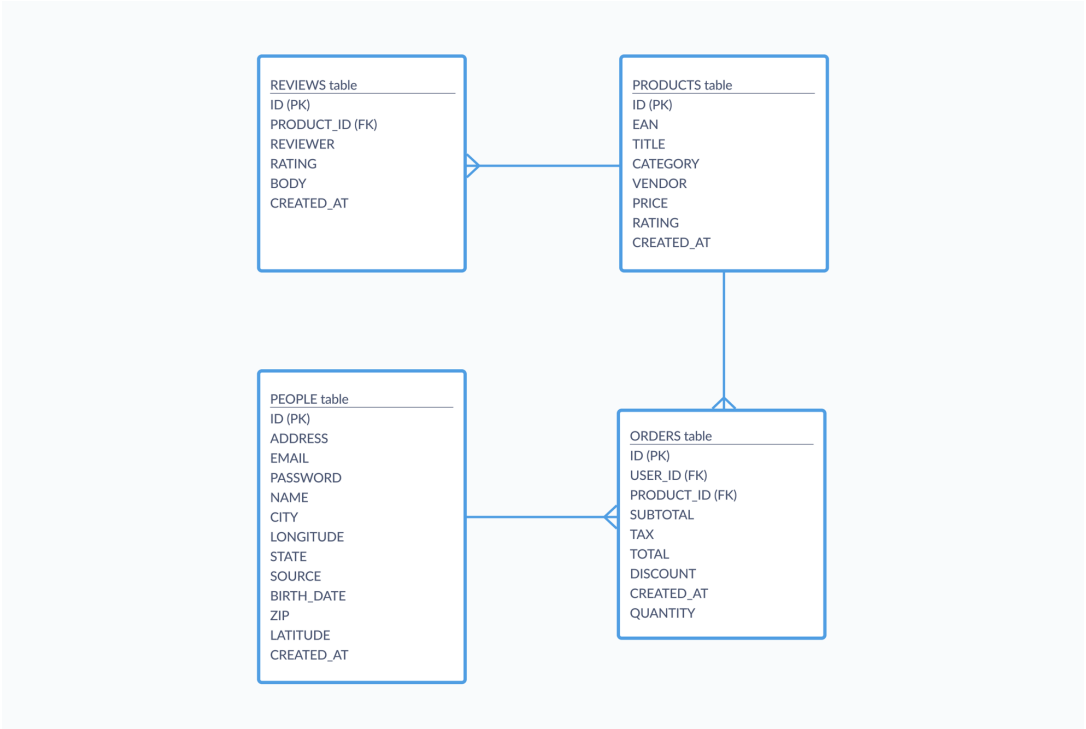
RELATIONS / CARDINALITY

One-to-one relationship



One-to-many relationship

Many-to-many relationship



https://www.metabase.com/learn/grow-your-data-skills/data-fundamentals/table-relationships

RELATIONS / CARDINALITY

NationalTeams

Country	Team
Germany	Die Mannschaft
Italy	Azzurri
Spain	La Roja

Sheet1 Sheet

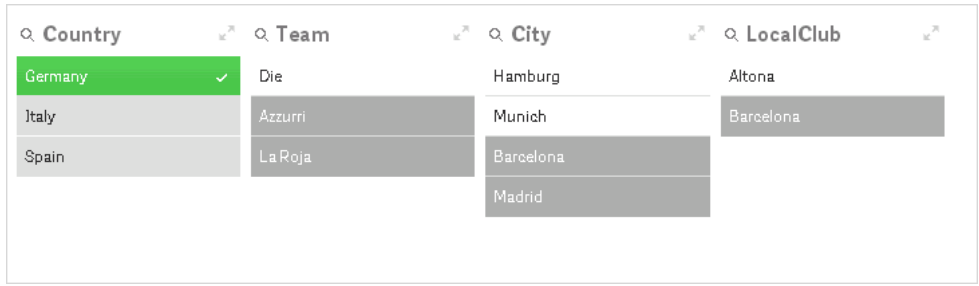
Clubs

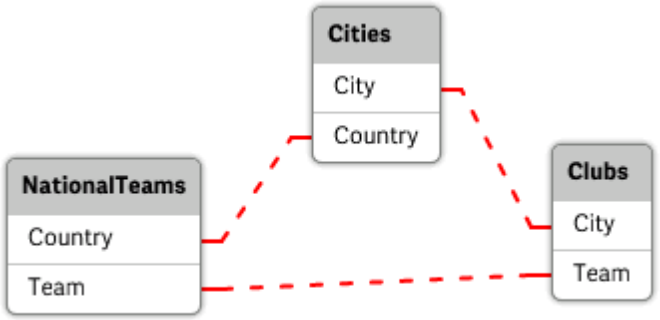
City	Team
Barcelona	Barcelona
Hamburg	Altona
Madrid	Real Madrid
Milano	Milan
Munich	Bayern München
Rome	Lazio
Turin	Juventus

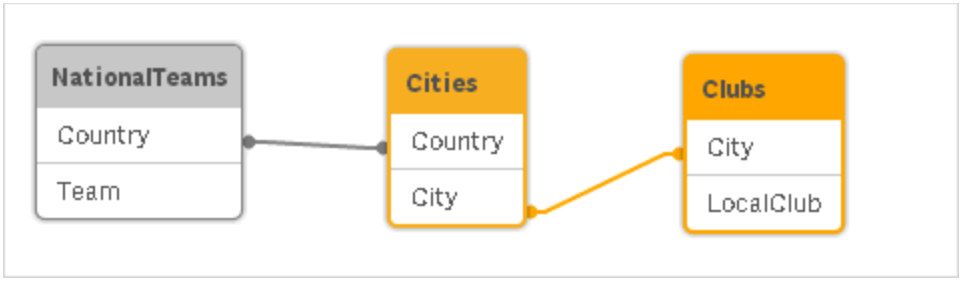
Sheet1 Sheet2

Cities

Country	City
Germany	Hamburg
Germany	Munich
Italy	Milano
Italy	Rome
Italy	Turin
Spain	Barcelona
Spain	Madrid

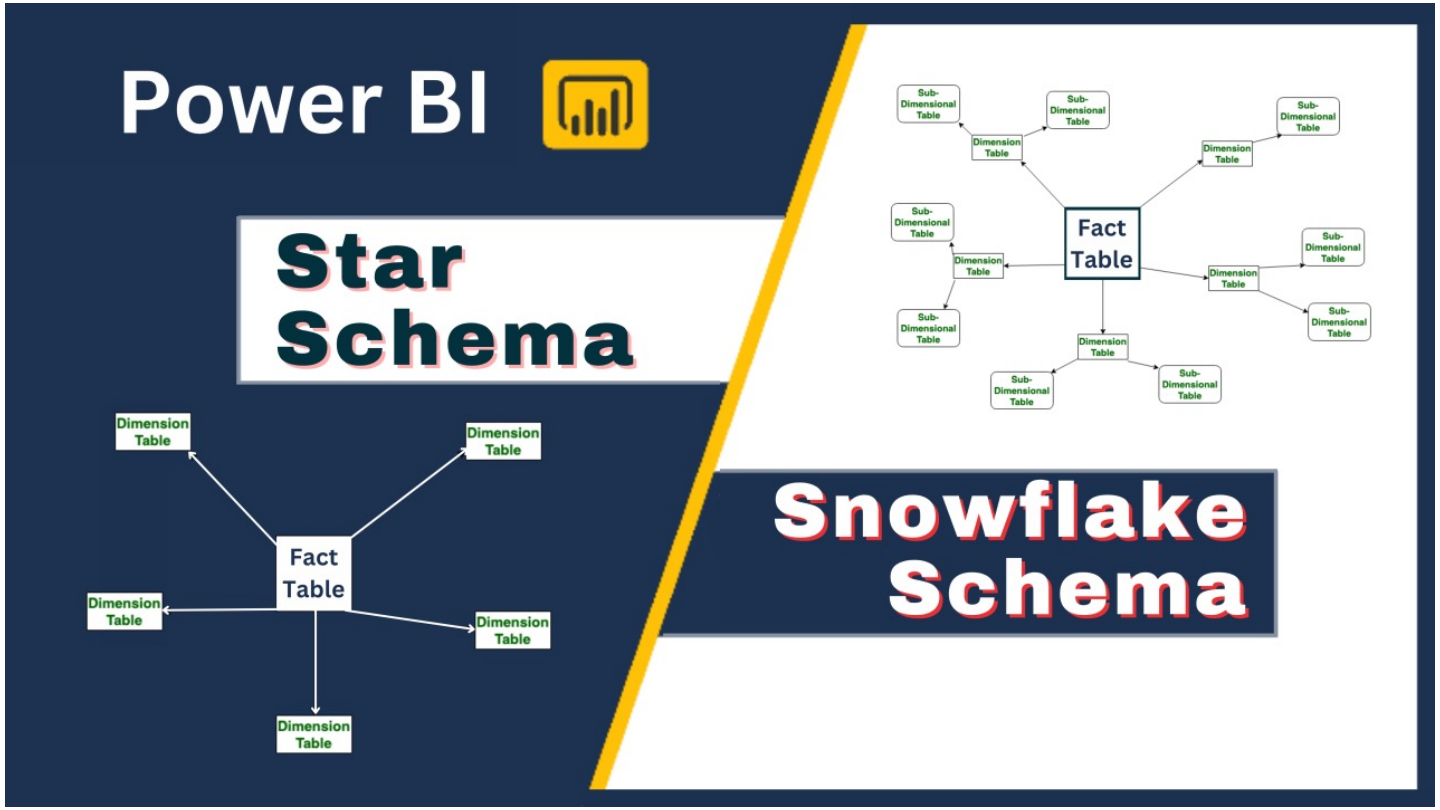






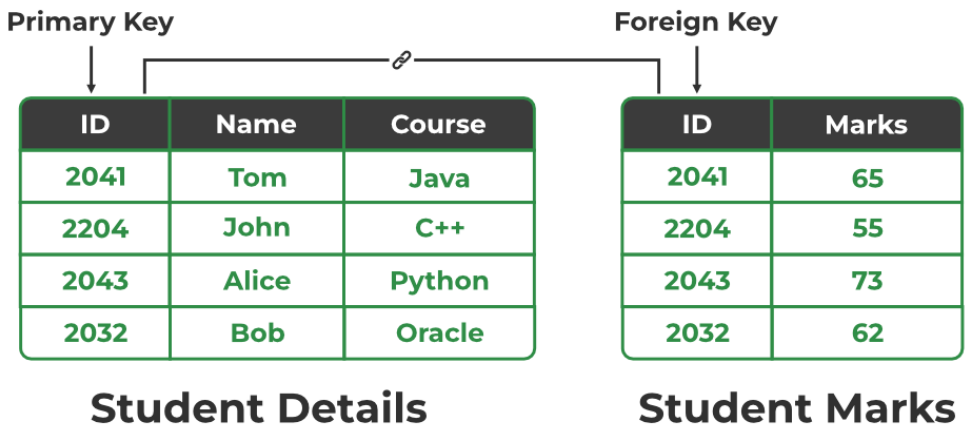
https://help.qlik.com/es-ES/sense/November2024/Subsystems/Hub/Content/Sense_Hub/LoadData/understand-circular-references.htm

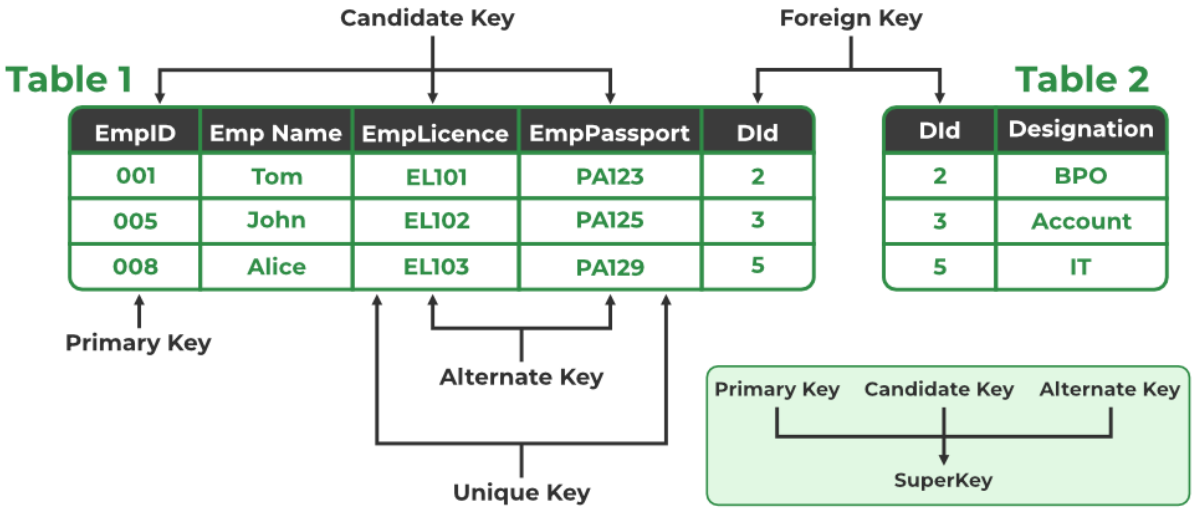
DATA MODELS SCHEMAS



https://www.linkedin.com/pulse/star-schema-vs-snowflake-data-model-power-bi-belayet-hossain--2b6rc/

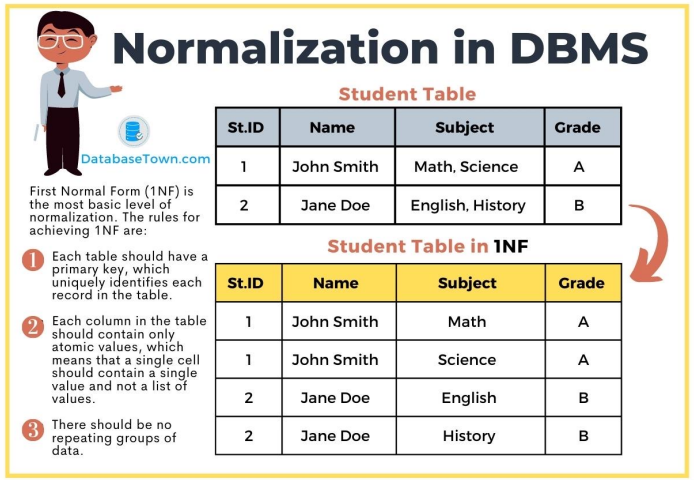
KEYS

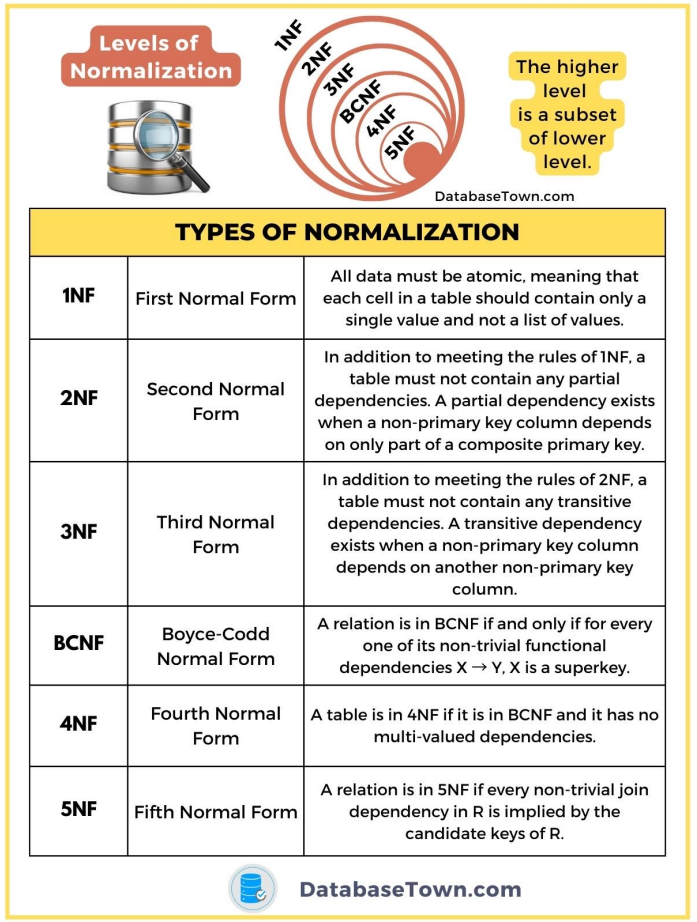




https://www.geeksforgeeks.org/types-of-keys-in-relational-model-candidate-super-primary-alternate-and-foreign/

NORMALIZATION





https://databasetown.com/types-of-normalization-in-dbms-with-examples/

https://www.metabase.com/learn/grow-your-data-skills/data-fundamentals/normalization