# Análisis de series de tiempo sobre el subterráneo en CABA

Comparación de modelos de forecast

# Objetivo

- El alcance del trabajo comprende los siguientes objetivos
  - Describir la demanda de usuarios del subterráneo en Buenos Aires
  - Describir conceptos asociados a la serie de tiempo y modelos de forecasting
  - Utilizar distintos modelos de ***forecasting*** para generar predicciones
  - Comparar los modelos utilizados

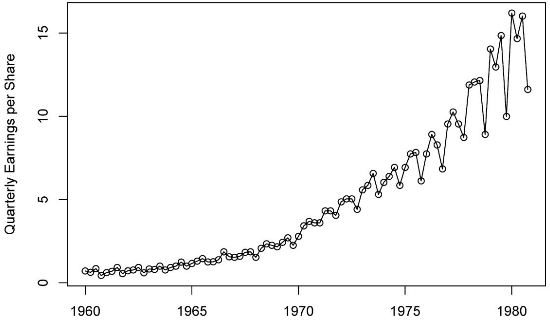
# ¿Qué es una serie de tiempo?

Una serie de tiempo es un conjunto de información tomada a **intervalos regulares**, y **ordenadas** a lo largo del tiempo.

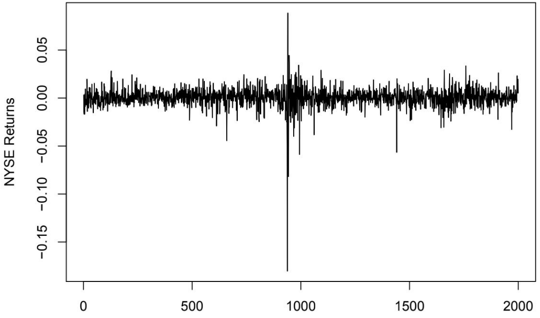
Ejemplo de series de tiempo:

- Variables macroeconómicas
- Activos financieros (balances, acciones)
- Variables sociales (pobreza, PBI per cápita, desigualdad)
- Variables experimentales (precipitaciones, velocidad del viento)

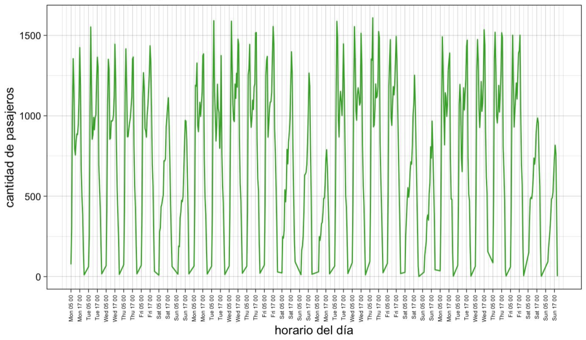
ganancias trimestrales Johnson & Johnson




Retornos diarios del NYSE.




Pasajeros por hora, estación palermo, linea D



# Datos utilizados

- Fuente: Buenos Aires Data - Molinetes 2022 (información para todo el mes de agosto de 2022)
- Variables:
  - ***fecha****:* fecha del mes de agosto de 2022
  - ***desde***: timestamp de inicio
  - ***hasta****:* timestamp de fin (**15 min** posteriores a *desde*)
  - ***linea***: línea de subte
  - ***molinete*****:** tag del molinete
  - ***estacion****:* nombre de la estación
  - ***pax_pago*****:** pasajeros que pagaron
  - ***pax_pases_pagos*****:** pasajeros que pagaron con pase
  - ***pax_franq*****:** pasajeros que pagaron franquicia
  - ***pax_total****:* sumatoria del total de pasajeros anteriores

# Análisis exploratorio y transformaciones

- ➔ Interesa estudiar la demanda a nivel **estación**:
  - ◆ Se agrupa la información de molinetes que pertenecen a la misma estación
- ➔ Se agrupa la demanda en intervalos de **1 hora** en lugar de 15min
  - ◆ El horario de apertura es 5:15 am y el horario de cierre es 12:00 am
- ➔ Interesa estudiar la demanda total
  - ◆ La variable ***pax_total*** es la que se tiene en cuenta

- ➔ Estación de estudio: **Palermo** (Línea D)
- ➔ Partición dataset
  - ◆ TRAIN: 3 primeras semanas de agosto
  - ◆ TEST: última semana de agosto

# Componentes de una serie de tiempo

*patrón estacional que se repite con regularidad*
**ESTACIONALIDAD**

**imprevistas, sin poder de control e impredecibles**
**FLUCTUACIONES ALEATORIAS**

$$Y_t = T_t + S_t + C_t + \varepsilon_t$$

**TENDENCIA**
*componente permanente visible a largo plazo*

**CICLOS**
*variaciones en la que opere la serie*

# ¿cómo podría modelar la tendencia?

**tendencia lineal**

$$T_t = \beta_0 + \beta_1 \times TIME_t$$

**tendencia cuadrática**

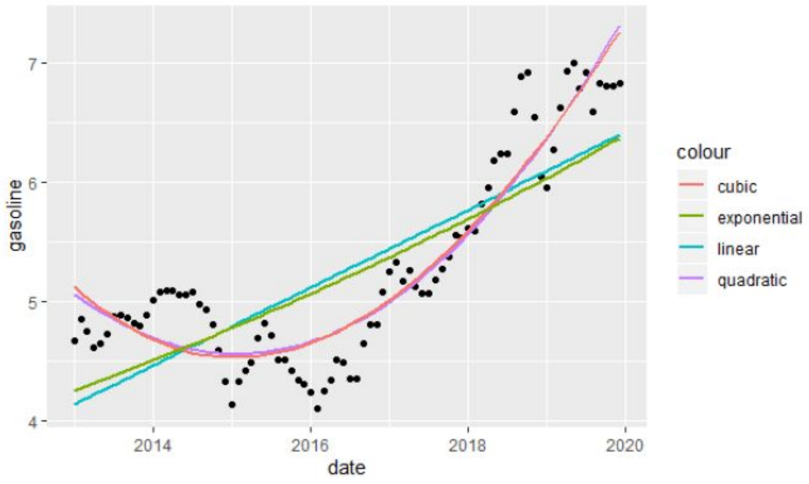
$$T_t = \beta_0 + \beta_1 \times TIME_t + \beta_2 \times TIME_t^2$$

**tendencia exponencial**

$$T_t = \beta_0 \, e^{\beta 1 \times TIMEt}$$

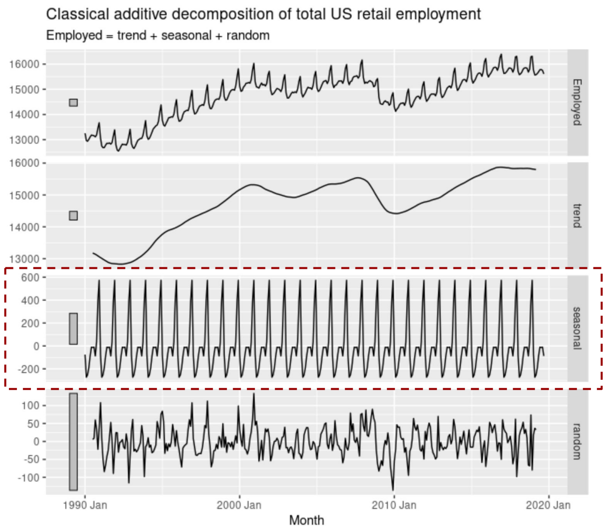
**aplicando logaritmo natural**

$$\ln(T_t) = \ln(\beta_0) + \beta_1 \times TIME_t$$



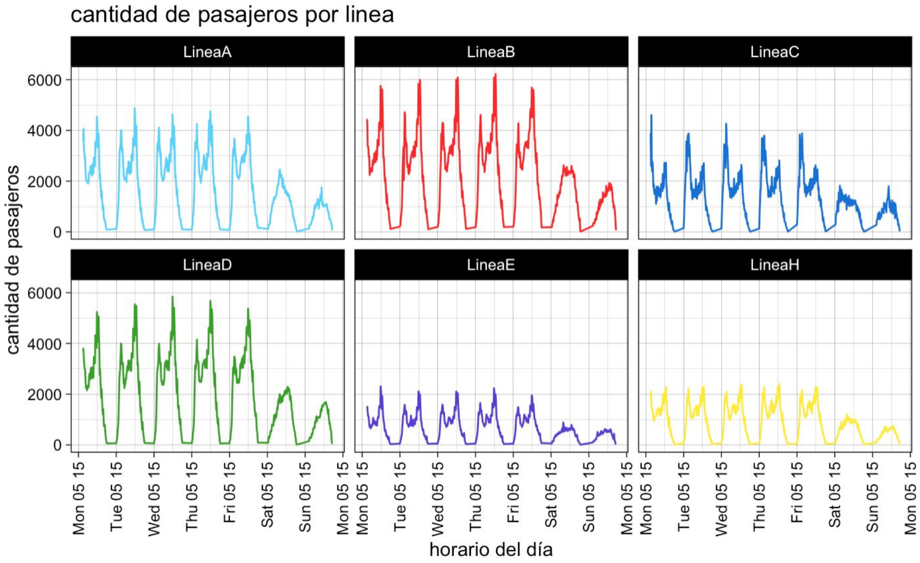
# ¿cómo podría modelar la estacionalidad?

**ej: estacionalidad mensual**




Enero = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
Febrero = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
Marzo = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)
Abril = (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
Mayo = (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
Junio = (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
Julio = (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
Agosto = (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
Septiembre = (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
Octubre = (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
Noviembre = (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
Diciembre = (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)

# Componentes de la serie de tiempo en el problema de análisis




**ESTACIONALIDAD: si observamos la serie de pasajeros de subte, se observa una fuerte estacionalidad diaria y horaria**
**TENDENCIA: para visualizarla, se debe analizar la demanda a nivel anual.**
**FLUCTUACIONES: un paro o suspensión del servicio.**

# ¿En qué consiste el componente estacional?

- Los pasajeros que ingresan al subte tiene un comportamiento cíclico **diario**
  - Los picos de demanda se ubican en el ingreso/egreso del horario laboral.
  - Entre picos se da un horario valle amesetado
  - La demanda en horarios de apertura y de cierre de las líneas es significativamente más baja.
  - Los fines de semana el comportamiento se altera, teniendo una demanda atada a fines de ocio.
  - La diferencia en el total de pasajeros transportados depende de cada línea y los nodos que conecta.

- El comportamiento estacional no sigue la misma forma, dependiendo del tipo de estación que se trate:
  - Estaciones de cabecera
  - Estaciones de combinación de líneas (hubs)
  - Estaciones regulares (el análisis se focaliza en una estación de este tipo).

# ¿en qué consiste el componente cíclico?

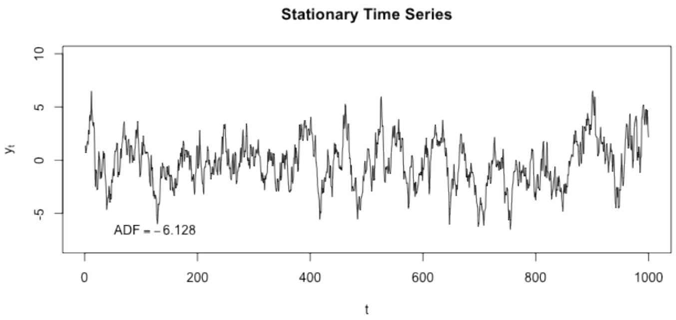
Diebold lo resume en: cualquier tipo de dinámica no capturada ni por la estacionalidad ni la tendencia.

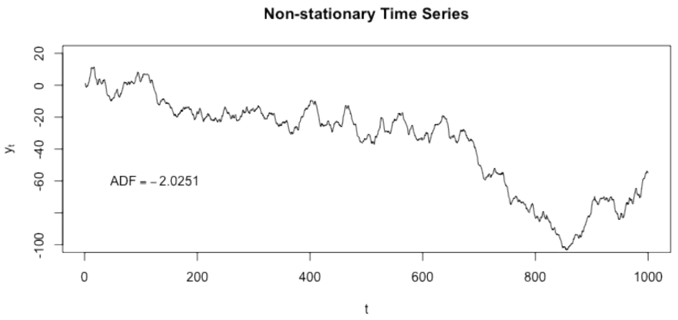
Generalmente consiste en algún efecto persistente a través del cual el presente está linkeado al pasado, y el futuro al presente.

# ¿Qué es una serie estacionaria?

- Una serie estacionaria es aquella en la que **las propiedades no dependen del momento en el que se observe la serie.** Por lo tanto, una serie con tendencias o estacionalidad no es estacionaria. Una serie con comportamiento cíclico es estacionaria, ya que los ciclos no tienen una longitud preestablecida.

- Una serie estacionaria se caracteriza por tener:
    - Una media constante a lo largo del tiempo
    - Varianza constante independiente del tiempo
    - La autocovarianza depende únicamente del desplazamiento y no de t.



# ¿Cómo convertir una serie en estacionaria?

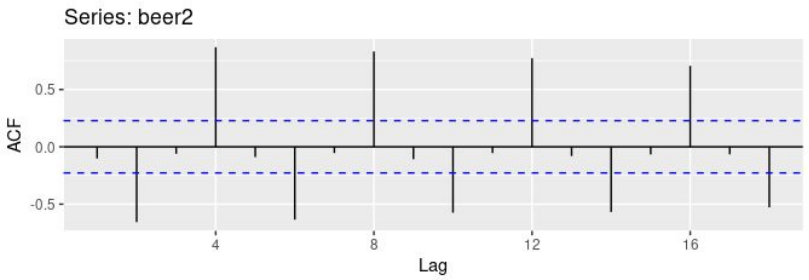
- DIFERENCIAS: computando diferencias entre observaciones consecutivas
    - Ayuda a estabilizar la **media** de una serie, eliminando o reduciendo efectos de tendencia y estacionalidad.

- TRANSFORMACIONES: aplicar logaritmo
    - Ayuda a estabilizar la **varianza** de una serie.

- DETECCIÓN: para detectar si una serie es estacionaria o no
    - Gráfico de ACF (Autocorrelation Function) debe tender a ir a 0 relativamente rápido.
    - Test  Box-Ljung

# ¿Qué es la autocorrelación?

- Los valores que toma una variable en el tiempo **no** son independientes entre sí. Existe una correlación con valores pasados de la serie




$$r_k = \frac{\sum\limits_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum\limits_{t=1}^{T} (y_t - \bar{y})^2},$$

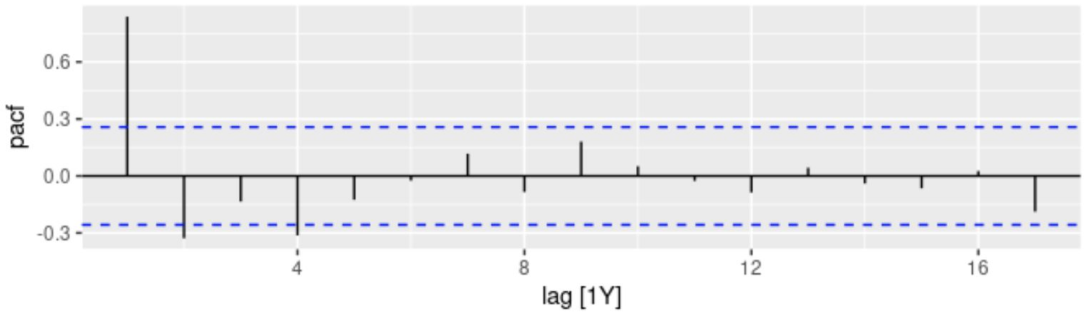
Coeficientes de autocorrelación que se grafican en un gráfico ACF con sus IC

# ¿qué es la función de autocorrelación parcial?

Mide la relación entre yt y yt−k después de remover el efecto de los lags 1, 2, 3, …, k−1.

Entonces, la primera autocorrelación parcial es identica a la primera autocorrelación porque no hay nada entre ellos que eliminar.

función de autocorrelación parcial



# Modelos utilizados

A continuación, se hará una introducción a los modelos AR, MA, ARMA y ARIMA, y se realizará una comparativa con la predicción que realiza un modelo desarrollado utilizando prophet (Prophet | Forecasting at scale. (facebook.github.io))

# Aplicabilidad práctica

Para la serie de estudio, buscar predecir la demanda instantánea puede ser útil para:

- Dimensionar y verificar medios de escape y accesos
- Estudiar la congestión en hubs de la red tales como estaciones del tipo combinación

Para series de tiempo más longevas y analizadas a nivel semanal/mensual:

- Cambios en la demanda por nuevos medios de transporte o preferencias de usuarios
- Impacto de la extensión de alguna línea existente sobre la demanda total del resto de las líneas.

# AR(p)

En un modelo autorregresivo, buscamos predecir la variable de interés a partir de una combinación lineal de los **p** valores pasados de la variable.

Un modelo autorregresivo de orden p tiene la siguiente forma:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

donde

$\varepsilon_t \sim WN(0, \sigma^2)$

# MA(q)

En lugar de utilizar valores pasados como variables predictoras, un modelo MA utiliza los errores en las predicciones del pasado como variables regresoras de Yt. Se puede pensar como una media móvil ponderada de los errores de predicción del pasado.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

# ARIMA(p,d,q)

- El parámetro **d** hace referencia a la cantidad de veces que se diferencie la serie, quedando con la siguiente expresión:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

- La diferenciación de la serie se hace de manera tal de obtener una serie estacionaria o reducir los efectos de tendencia y/o estacionalidad que existan

# ¿Cómo definir p y q en modelos no estacionales?

- Gráfico ACF: autocorrelaciones para un determinado lag
- Gráfico PACF: autocorrelaciones parciales
- Los gráficos ayudan únicamente en casos donde alguno de los parámetros ***p*** o ***q*** sea nulo, viendo la cantidad de picos por fuera del intervalo de confianza.

# ARIMA(p,d,q) (P,D,Q)

## ¿Cómo definir p y q en modelos estacionales?

- Al modelo original se le agregan parámetros asociadas a la estacionalidad, que multiplican a la parte no estacional.
- La parte autoregresiva y la de media móvil quedan definidas en los gráficos PACF/ACF utilizando un lag vinculado con la estacionalidad.

# Comparativa de modelos ARIMA para selección de p y q

Se utilizan dos métricas:

**AIC (Akaike's Information Criterion):** depende de la verosimilitud L y el mejor modelo es aquel con el menor AIC. Se corrige con la siguiente expresión

$$\text{AIC} = -2\log(L) + 2(p+q+k+1),$$

$$\text{AICc} = \text{AIC} + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2},$$

**BIC (Bayesian Information Criterion):**

$$\text{BIC} = \text{AIC} + [\log(T) - 2](p+q+k+1).$$

# Bibliografía


- Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3.
- Shumway, R.H. and Stoffer, D.S. (2011) Time Series Analysis and Its Applications (With R examples), 3rd edition, Springer: New York, USA .
- Diebold, F.X. (2008) Elements of forecasting, 4th edition. Mason, Ohio: South-Western/Cengage Learning.