# Tasca PF1

## Presentació del projecte

**1. Presentació del conjunt de dades escollit:** explicació i observacions generals del conjunt de dades a utilitzar per a fer el projecte final.

- M'he decantat per una BBDD del kaggle per fer análisis de sentiment sobre les ressenyes a un establiment de Starbucks de l'any 2020 al 2023. (adjunto document csv).

- Te 851 observacions i 6 columnes com a punt de partida

*https://www.kaggle.com/datasets/harshalhonde/starbucks-reviews-dataset*

**2. Característiques generals:** sense necessitat d'entrar molt en detall, explicar les principals característiques que defineixen aquest conjunt de dades. Tipologia, sector, tipus de dades, font, context, etc.

- Estem parlant d'un conjunt de dades obtingut a partir de webscraping per l'autor des de la página web consumeraffairs.

- L'establiment en qüestió és la famosa cadena de café Starbucks, i les ressenys son de diferents llocs d'Estats Units d'Amèrica. (segons he fet una ullada ràpida al dataset).

- Tenim dades númeriques que clasifiquen, dades de dates, podem convertir alguna columna en dummies (si tenen imatges o no) i de text.

**3. Definició de les variables:** explicació teòrica de les principals variables que conté el conjunt de dades.

Les dades que tenim son:

1. Name -> Nom del ressenyador (en cas que el tinguem)
2. Location -> el lloc o ciutat associada a la ressenya si es dona
3. Date -> Data quan va ser publicada la review
4. Rating -> Puntuació de la ressenya que dona el ressenyador a Starbucks de l'1 al 5
5. Review -> La ressenya en qüestió
6. Image Links: links de la imatge si son proveïdes.

**4. Presentació dels objectius:** detallar els objectius inicials marcats de cara a extreure informació rellevant del conjunt de dades.

Els objectius principals del projecte serà fer una neteja del dataset, i amb l'ajuda de SPARK poder fer un anàlisis de sentiment de les ressenyes, a partir d'aqui he pensat accions complementaries que es poden fer i faré i altres que em requereixen més temps però les deixo plantejades:

1. Comprovar si té lògica el resultat de l'analisis de sentiment amb la puntuació donada

2. Fer un anàlisi sobre si afecta el dia de la setmana al tipus de ressenya que es fa (dilluns a divendres) o també per trimestres / estació de l'any, etc..

3. També comprovar o analitzar si pujar imatges acostuma anar lligat a les ressenyes amb millor o pitjor puntuació?

---

Ampliació projecte (preguntes que em faig però no se si serè capaç de resoldre per falta de temps, dades o habilitats)

1. Intentar mirar si amb el text es pot saber quants dies han passat des de que s'ha fet la ressenya i es va anar a l'estiblement -> Predicció ML de quants dies passen a fer la ressenya? relacionar-ho amb el sentiment.

2. També fer un anàlisis per zones (Location)

3. També seria interessant poder veure per gènere biologic (nom del ressenyador/a)